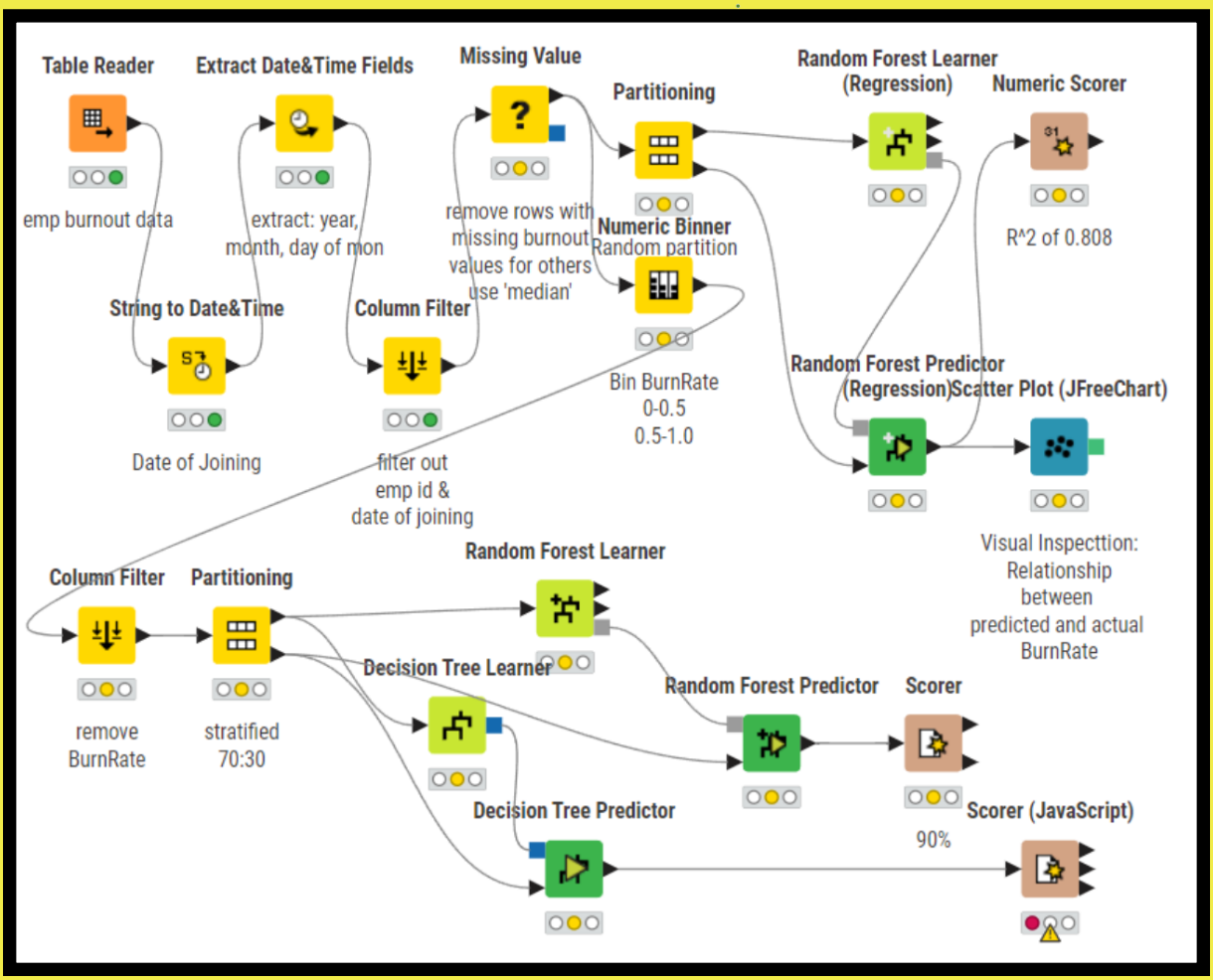
Strojno učenje u ekonomiji i turizmu (s primjerima u programu KNIME Analytics Platform)

Đokić, Kristian

STROJNO UČENJE U EKONOMIJI I TURIZMU

(sa primjerima u programu KNIME Analytics Platform)



Kristian Đokić

Strojno učenje u ekonomiji i turizmu

(s primjerima u programu **KNIME Analytics Platform**)

Kristian Đokić

Nakladnik

FAKULTET TURIZMA I RURALNOG RAZVOJA U POŽEGI

Autor

doc. dr. sc. Kristian Đokić

Recenzenti

prof. dr. sc. Ivana Đurđević Babić

doc. Dr. sc. Mirjana Jeleč Raguž

Lektura

dr.sc. Vesna Vlašić, pred. v. š.

Grafička priprema

doc. dr. sc. Kristian Đokić

Odobreno odlukom Senata Sveučilišta Josipa Jurja Strossmayera u Osijeku na 11. sjednici održanoj 30. rujna 2024.g. klasa:611-01/24-01/32; ur.broj: 2158-60-01-24-2

ISBN 978-953-7744-46-5

https://doi.org/10.62598/ftrr.002

PREDGOVOR

Strojno učenje, a zatim i duboko učenje, kao središnji dio umjetne inteligencije, danas se ne javljaju samo unutar područja poslovanja, nego se svakodnevno pojavljuju u medijima te izazivaju različite reakcije. S jedne strane su oni koji su oduševljeni razvojem umjetne inteligencije s obzirom da određene tehnike strojnog učenja već danas pomažu u svakodnevnom životu i radu, kao npr. u medicinskoj dijagnostici, prepoznavanju govora i kreiranju preporuka u online knjižarama. Na drugoj strani, ovi pojmovi i mogućnosti izazivaju određenu nelagodu, jer se tehnike strojnog učenja mogu i zlorabiti na različite načine, kao npr. za korištenje lica osobe bez dopuštenja, izradu autonomnog naoružanja i prekomjernu kontrolu stanovništva od strane vlada. Isto tako sve češće se pojavljuju tekstovi ili objave u kojima se najavljuje doba u kojemu će roboti, koje pokreću modeli strojnog učenja, učiniti cijeli niz zanimanja nepotrebnima. Takve najave nestanka pojedinih zanimanja vjerojatno su istinite, no ti se procesi odvijaju već stoljećima. Svaka industrijska revolucija rezultirala je nestankom pojedinih zanimanjima i nastankom novih, ali evidentno je da su nestajala zanimanja koja uključuju naporan rad. Povijest nas uči kako razvoj umjetne inteligencije, uključujući i implementaciju strojnog učenja, nije moguće zaustaviti. No, ono što je moguće jest pokušati razumjeti i primijeniti u području kojim se bavimo, na opću dobrobit.

Ovaj priručnik pisan je za studente i stručnjake s polja ekonomije i turizma, ali i ostalih društvenih područja, koji žele naučiti kako primijeniti strojno učenje koristeći vizualno programiranje. Iako pozadinu strojnog učenja čine kompleksni matematički i statistički modeli, ovaj pristup omogućuje primjenu metoda, ne ulazeći dublje u njihovu matematičku pozadinu. S druge strane za primjenu ovih metoda, potrebno je osnovno poznavanje načina funkcioniranja, uvjeta, ograničenja i osnovnih principa, kako bi se odabrala odgovarajuća metoda i dobili realni rezultati. U priručniku je opisano sedam osnovnih metoda, a nakon svake je naveden primjer iz područja struke. U opisu pojedinih metoda, na apstraktnoj su razini objašnjeni temeljni principi funkcioniranja istih, dok su jednostavnije metode objašnjene primjenom matematičkih izraza te bi trebale biti razumljive i široj publici.

Priručnik se sastoji od 12 poglavlja. U prva tri poglavlja obrađuju se pojmovi vezani uz strojno učenje i vrste strojnog učenja te se kreira prvi hodogram pri čemu se koristi vizualno programiranje. Četvrto i peto poglavlje odnose se na učitavanje, pripremu i analizu podataka potrebnih u procesima strojnog učenja. Od šestog do dvanaestog poglavlja prikazane su najčešće korištene metode kao što su linearna regresija, logistička regresija, bayesov naivni klasifikator, stablo odlučivanja, metoda slučajnih šuma, metoda potpornih vektora i metoda k-srednjih vrijednosti.

Posebna pozornost posvećena je odabiru programa s kojim čitatelji mogu testirati opisane tehnike. Prvi uvjet je bio da program bude **besplatan**, da **ne zahtijeva poznavanje programiranja** te bude **jednostavan za upotrebu**. Odabran je program *KNIME*, čiji puni naziv glasi *Konstanz Information Miner*. S obzirom da se većina čitatelja vjerojatno prvi puta susreće s programom KNIME, u trećem poglavlju priručnika navedene su upute vezane uz preuzimanje i instalaciju, a u istom poglavlju su opisani prvi koraci korištenja programa. U priručniku je korištena verzija programa KNIME 4.x, iako je u međuvremenu izdana verzija 5 koja koristi drugačije sučelje. No, i u toj novoj verziji je i dalje dostupno sučelje prethodnih verzija, a u trećem poglavlju je opisano kako se može izabrati klasično korisničko sučelje na kojem je temeljen priručnik. Dodatne informacije, datoteke i eventualne korekcije bit će dostupne na: https://github.com/kristian1971/knimeprirucnik_v1 . Nadam se da će vam ovaj priručnik biti koristan te će vam opisane tehnike pomoći u poslu i profesionalnom razvoju, kao i potaknuti na dalje proučavanje strojnog učenja.

Autor

TERMINOLOGIJA I KONVENCIJE U PRIRUČNIKU

U ovom je priručniku napravljen niz kompromisa vezanih uz hrvatski prijevod izvornih termina na engleskome jeziku. Za tehnike strojnog učenja korišteni su hrvatski nazivi uz prijevode na engleski, jer često prijevod ukazuje i na neko svojstvo tehnike ili na način funkcioniranja. Modeli strojnog učenja, kreirani primjenom programa KNIME i opisani u priručniku, sastoje se od niza povezanih čvorova koji čine radni tijek ili hodogram, za što se u programu koristi naziv *workflow.* Pojam *hodogram* izabran je jer se radi o poznatom terminu koji označava slijed radnji nekog organiziranog procesa, za razliku od doslovnog prijevoda „tijek rada“ koji je preopćenit i zvuči pomalo nespretno. Za čvorove koji čine *hodogram*, u programu KNIME koristi se naziv *node*, dakle u tom slučaju je u priručniku korišten doslovan prijevod - *čvor*. Pojam *confusion matrix* je u priručniku preveden kao *matrica konfuzije*, iako se u domaćoj literaturi može naići na pojmove kao *matrica zabune* i *matrica zbunjenosti*.

Osim toga, još jedan od problema jest različito nazivlje za neke pojmove koji ovise o struci i kontekstu. Za ono što matematičari nazivaju varijablom, u strojnom učenju ponekad se koristi termin značajka, dok će se u računovodstvu kod primjene tabličnog kalkulatora to nazvati zaglavljem stupca. U priručniku će se za takve pojmove najčešće koristiti terminologija s područja strojnog učenja, ali će biti navedeni i drugi nazivi kako bi sadržaj bio razumljiviji većem broju čitatelja.

Autor

KAZALO KRATICA I AKRONIMA

KNIME – Konstanz Information Miner
BPB - broj prodanih bočica
BPN - broj prijavljenih noćenja
SSE - Sum of Square Errors
RMSE - Root Mean Square Error
TP - True Positives
TN – True Negatives
FP – False Positives
FN – False Negatives
CSV - Comma Separated Values
ARFF - Attribute-Relation File Format
REST - Attribute-Relation File Format
XML - Extensible Markup Language
PMML - Predictive Model Markup Language
SVM - Support Vector Machines

Sadržaj

1. Uvod

Za strojno učenje enciklopedija Brittanica navodi kako se radi o „disciplini koja se bavi implementacijom računalnog softvera koji se može učiti autonomno" (Hosch, 2023). Softver „uči" iz dostupnih podataka koji su donedavno bili potpuno neiskorišteni, ali korištenjem strojnog učenja ti isti podaci omogućuju bolje poznavanje kupaca i konkurencije, prepoznavanje objekata na slikama, generiranje novinskih tekstova, prepoznavanje neželjene pošte i prijevara pri korištenju kreditnih kartica i cijeli niz drugih primjera. Druga definicija, prema Sahay, pojam strojno učenje opisuje kao „metodu dizajniranja sustava koji mogu učiti, prilagođavati se i poboljšavati na temelju podataka koji im se šalju bez da su eksplicitno programirani" (Sahay, 2021).

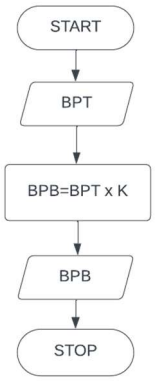
Strojno učenje može se koristiti praktički na svim područjima, a neka područja primjene su koeficijent K (Probyto Data Science and Consulting Pvt. Ltd., 2020):

a) Problemi predikcije – jedan od primjera je hoće li potencijalni korisnik kredita biti u stanju otplatiti kredit (Bussmann, et al., 2021; Khandani, et al., 2010; Lee, et al., 2006).
b) Prepoznavanje slika – prepoznavanje osoba na slikama je jedan od problema koji je rješiv uz pomoć strojnog učenja (Chen & Jenkins, 2017; Damale & Pathak, 2018; Wei, et al., 2011).
c) Prepoznavanje govora i prevođenja – problem prepoznavanja govora zajedno s prevođenjem i generiranjem govora na drugim jezicima problemi su koje također rješava strojno učenje (Ganapathiraju, et al., 2004; Ganapathiraju, et al., 2000; William, et al., 2022).
d) Problemi vezani uz medicinu – pomoć u ranom otkrivanju bolesti iz rendgenskih slika pacijenata (Choy, et al., 2018; Kohli, et al., 2017; Wang & Summers, 2012).
e) Financijska industrija i prodaja – strojno učenje koristi se za prepoznavanje obrazaca koji se pojavljuju kod prijevara (Kovač, et al., 2022; Xu, et al., 2022; Seemakurthi, et al., 2015).

Da bi bilo jasno kako to softver „uči", potrebno je krenuti od klasičnog načina funkcioniranja softvera. Kao primjer, može se uzeti jednostavan slučaj programa koji će izračunavati očekivani broj prodanih bočica osvježavajućeg pića na samoposlužnom aparatu u ovisnosti o prijavljenom broju noćenja turista, u mjestu gdje se nalazi samoposlužni aparat. Zdrava logika ukazuje da su dvije varijable povezane – prijavljen broj noćenja i broj prodanih bočica osvježavajućeg pića. Kako bi se te dvije varijable povezale, koristit će se koeficijent *K* i on će služiti za predikciju broja prodanih bočica na način da se njegova vrijednost pomnoži s prijavljenim brojem noćenja. Koeficijent *K* dobiva se dijeljenjem broja noćenja turista s brojem prodanih bočica. Ako je broj prodanih bočica jednak broju prijavljenih noćenja tada će vrijednost koeficijenta *K* biti jednaka broju 1. Ako je broj prodanih bočica manji od broja noćenja, vrijednost koeficijenta *K* bit će manja od 1, a u slučaju da je broj prodanih bočica veći od broja noćenja turista, vrijednost koeficijenta K bit će veća od broja 1. Može se pretpostaviti da bi navedeni program bio koristan tvrtki koja puni samoposlužne aparate kako bi bolje upravljala zalihama. Izraz kojim bi mogli izračunati povezanost glasi:

broj prodanih bočica (BPB) = broj prijavljenih noćenja (BPN) x koeficijent (K)

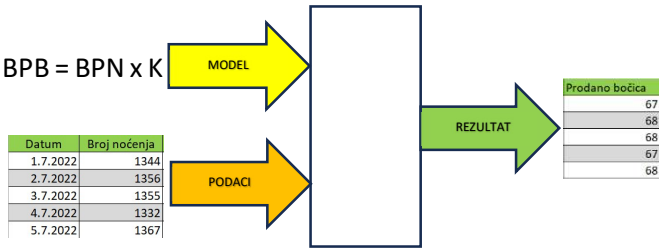
Programer koji bi pisao program, trebao bi na osnovu iskustva osobe koja puni samoposlužni aparat definirati koeficijent *K* s kojim se množi broj prijavljenih noćenja turista kako bi dobili očekivani broj prodanih bočica. Koeficijent *K* služi da bi iz dvije varijable za koje se vjeruje da su linearno povezane dobili vrijednost nepoznate, koristeći vrijednost poznate varijable. Programeri za grafički prikaz problema koji rješavaju često koriste dijagrame tijeka koji prikazuju ulaz, obradu i izlaz podataka. Slika1 prikazuje dijagram tijeka za navedeni primjer.



Slika1. Primjer dijagrama tijeka

Nadalje, da bi „klasični" algoritam funkcionirao, potrebni su ulazni podaci i model kreiran od strane programera u obliku programa koji obrađuje ulazne podatke. Na osnovu unaprijed definirane obrade dobiva se rezultat.

Slika 2 prikazuje shemu klasičnog pristupa rješavanja problema uz pomoć algoritama (Mehra & Hasanuzzaman, 2020).



Slika 2. Klasičan „algoritamski" pristup rješavanja problema

Slika 2 s lijeve strane prikazuje ulaz podataka pri čemu se radi o tablici s dva stupca. Jedan stupac je datum koji algoritam ne uzima u obzir, a u drugom stupcu je broj prijavljenih noćenja turista na osnovu kojih algoritam predviđa potrošnju osvježavajućih pića. Osim podataka s lijeve strane je i model u obliku programskog koda u nekom programskom jeziku koji u suštini samo izvršava funkciju koja je navedena – množi broj noćenja turista sa zadanim koeficijentom. Slika 2 s desne strane prikazuje rezultat u obliku stupca pri čemu se radi o predviđenoj potrošnji osvježavajućih pića koji su izračunati od strane programa.

Treba obratiti pažnju na način kako će programer implementirati algoritam. U formuli koja povezuje broj prodanih bočica i broj noćenja turista koristi se koeficijent *K*. Taj koeficijent programer ne zna, ali saznat će ga u razgovoru s osobom koja ima iskustva s punjenjem konkretnog samoposlužnog aparata. Pri izradi bilo kojeg programa po narudžbi kupca, vrlo bitno je komunicirati s kupcem da bi on bio zadovoljan i kako bi program radio točno ono što kupac očekuje. Kupac je u pravilu osoba s iskustvom koja će programeru dati informaciju o prosječnoj potrošnji, a u navedenom primjeru to je oko 50 bočica u slučaju kada je prijavljeno oko 1000 noćenja. Programer će na osnovu toga izračunati koeficijent *K* koji iznosi 0,05. Slika 3 prikazuje programski kod napisan u programskom jeziku Python koji rješava navedeni problem.

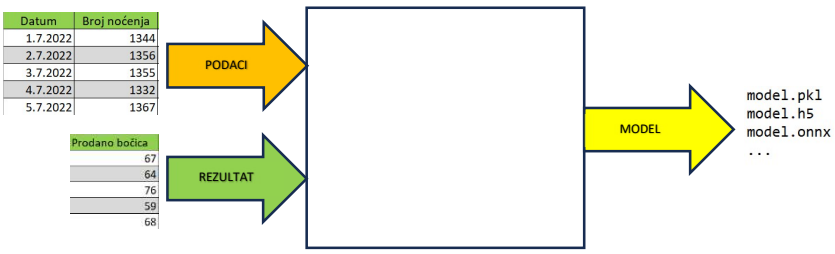
```
X=float(input('Unesite broj noćenja:'))
Y=X*0.05
print(Y)
```

Slika 3. Programski kod u Python-u

Najvažniji dio programa je u drugom redu u kojem je programer implementirao „znanje" osobe s iskustvom i pretvorio ga u programski kod. Radi se o umnošku broja noćenja i koeficijenta 0,05 pri čemu se u Python-u za množenje koristi znak "*". Taj rezultat se pohranjuje u varijablu Y i njezina vrijednost se ispisuje u trećem redu. Navedeni primjer vrlo je jednostavan, ali zorno ukazuje na činjenicu da kod klasičnog pristupa programiranju na programeru ostaje prijenos i implementacija „znanja" u program. U ovakvim jednostavnim primjerima to i nije problem, ali za rješavanje zahtjevnijih problema poput prepoznavanja osobe na slici, klasičan algoritamski pristup u kojem programer implementira kompletan algoritam, jednostavno nije primjenjiv.

Navedeni primjer zbog svoje jednostavnosti ima niz ograničenja. Prije svega ako se zamisli malo mjesto na Jadranskom moru u kojem je postavljen samoposlužni aparat, jasno je da će se osvježavajuća pića prodavati i kada u njemu nema turista. Razlog tome jest što i stanovnici toga mjesta kupuju osvježavajuća pića izvan turističke sezone. Osim toga, predloženi algoritam ima ograničenje i u situaciji kada je znatno povećan broj turista. Ako je moguće puniti aparat za prodaju pića više puta dnevno, ipak postoji maksimalan broj bočica koje samoposlužni aparat može isporučiti kupcima. Ni ovo ograničenje nije implementirano u algoritam. U nastavku je predstavljen način kako bi se to moglo riješiti strojnim učenjem.

Za razliku od „klasičnog" pristupa programiranju, kod strojnog učenja potrebni su ulazni i izlazni podaci, kako bi algoritam mogao „naučiti" na koji način ulazne podatke transformirati u izlazne. Klasičan pristup programiranju i dalje će se koristiti u razvoju aplikacija, ali vremenom se zaključilo kako se kompleksniji problemi teško mogu rješavati algoritmima koje implementiraju programeri. S druge strane razvoj procesora omogućio je da se procesorska snaga koristi za izradu kompleksnih modela transformacije ulaznih podataka u izlazne rezultate. Ti modeli postali su toliko kompleksni da ih se često naziva „crnim kutijama" jer je prilično teško razjasniti zašto i kako su parametri algoritma postavljeni tako kako jesu. Slika 4 prikazuje proces treniranja modela strojnog učenja, pri čemu treba naglasiti da je na slici prikazan model nadziranog učenja (Mehra & Hasanuzzaman, 2020). O različitim vrstama strojnog učenja bit će riječi kasnije, a sad je bitno razumjeti da su za treniranje jedne vrste modela potrebni podaci i rezultati. Ključna razlika jest ta da se „znanje" implementira u model korištenjem poznatih podataka iz procesa koji se modelira, a uloga programera u transferu „znanja" je izmijenjena. Naravno, i dalje je potreban programer koji će pripremiti ulazne i izlazne podatke, odabrati adekvatan model, kreirati sučelje, itd.

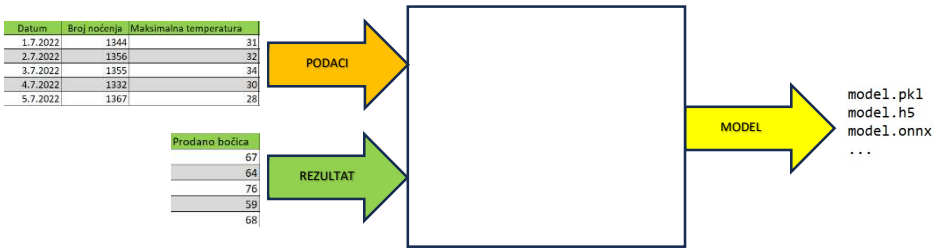


Slika 4. Proces treniranja modela strojnog učenja

Da bi u rješavanju prethodno opisanog problema predviđanja potrošnje osvježavajućih pića mogli koristiti strojno učenje, potrebni su prethodno poznati podaci koji se odnose na broj kupljenih bočica i

broj prijavljenih noćenja turista u mjestu. Vrlo vjerojatno će ovi podaci biti dostupni te bi mogli izabrati model nadziranog učenja i „istrenirati" ga koristeći postojeće podatke. Nakon treniranja dostupan je model koji se može koristiti na način da mu se dostavi prijavljeni broj noćenja, a model predviđa broj prodanih osvježavajućih pića sljedećeg dana. Model može predvidjeti prodaju i ako u podacima za „treniranje" nisu bile navedene sve mogućnosti broja noćenja, što znači da ima mogućnost predviđanja i za situacije koje nisu postojale u podacima za treniranje. Ipak, treba biti svjestan da takav model nikada neće biti savršen i ma koliko točno predviđao potrošnju, u pravilu će predviđanje više ili manje odstupati od stvarnog stanja.

S druge strane, moguće je povećati točnost modela na način da se razmisli o čemu sve ovisi potrošnja osvježavajućih pića na samoposlužnom aparatu. Osim navedenog broja prijavljenih noćenja turista, vjerojatno temperatura zraka utječe na potrošnju. Što je temperatura zraka viša, vjerojatno se kupi i više bočica osvježavajućeg pića. Isto tako, s padom temperature zraka potrošnja osvježavajućih pića vjerojatno opada. Prema tome, točnost modela mogla bi se povećati na način da model „treniramo" s više značajki, odnosno varijabli koje utječu na prodaju. Slika 5 prikazuje prethodno opisan model kojem je dodan još jedan stupac podataka i koji bi nakon ponovnog „treniranja" vjerojatno bio točniji (Mehra & Hasanuzzaman, 2020). Međutim, za primjenu modela bit će potrebna i prognozirana temperatura za sutrašnji dan, kako bi model funkcionirao i da bi predvidio potrošnju. Što se više značajnih varijabli implementira u model on će vjerojatno biti točniji, a vrijednosti ovih varijabli će biti potrebne pri korištenju modela za predikciju.



Slika 5. Proces treniranja kompleksnijeg modela strojnog učenja

U prethodnim odlomcima korišten je glagol „trenirati" i to pod navodnicima. Iako se ovaj glagol veže uz sport, u području strojnog učenja koristi se za proces „treniranja" modela uz korištenje poznatih podatka. Navedeni proces treniranja na primjeru potrošnje bočica jest približavanje što točnijoj vrijednosti koeficijenta *K* koji povezuje broj prijava noćenja turista i potrošnju bočica, kroz veći broj iteracija. Na kraju treniranja dobiva se optimalna vrijednost koeficijenta *K*. U nastavku priručnika uz glagol „trenirati" neće biti postavljeni navodnici.

Nakon upoznavanja s načinom rada jedne tehnike strojnog učenja (nadzirano učenje) i opisa razlika između „klasičnog" programiranja i korištenja strojnog učenja, slijedi podjela tehnika strojnog učenja.

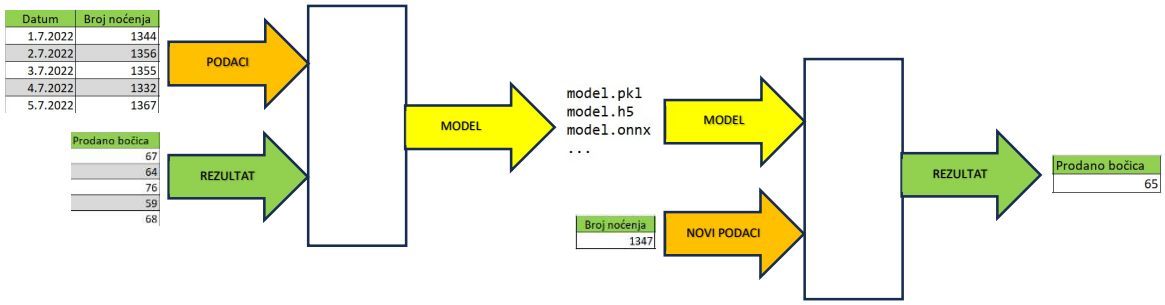
1.1. Tehnike strojnog učenja

Začetak strojnog učenja seže u prvu polovicu dvadesetog stoljeća, točnije 1943. godine kada su McCulloch i Pitts predložili model umjetnog neurona (McCulloch & Pitts, 1943). Od tada je predloženo više različitih tehnika strojnog učenja, koje se najčešće dijele u tri grupe ovisno o načinu funkcioniranja i pripremi podataka na (Probyto Data Science and Consulting Pvt. Ltd., 2020):

a) Nadzirano učenje (eng. *Supervised Learning*)
b) Nenadzirano učenje (eng. *Unsupervised Learning*)
c) Podržano učenje (eng. *Reinforcement Learning*)

1.1.1. Nadzirano učenje

Nadzirano učenje je specifično po tome što vrijednosti ulaznih varijabli i željena ciljana vrijednost služe za treniranje modela. Treniranjem se kreira funkcija koja preslikava nove podatke u očekivane izlazne vrijednosti (Mohri, et al., 2018). Nadzirano učenje opisano je prethodno u primjeru s osvježavajućim pićima, u kojem su kao ulaz korišteni podaci o broju prijavljenih noćenja turista i broju prodanih bočica osvježavajućeg pića. Za nadzirano učenje karakteristično je da su za izradu modela potrebni poznati podaci koji uključuju ulazne varijable (značajke) i izlaznu varijablu (ciljna vrijednost). U opisanom primjeru sa samoposlužnim aparatima za osvježavajuća pića, poznati podaci su broj prijavljenih noćenja u prošlom periodu i broj prodanih bočica u istom tom periodu. Pri tome, za svaki dan u navedenom periodu moraju biti dostupna oba podatka. Broj prijavljenih noćenja se smatra ulaznom varijablom ili značajkom (eng. *Feature*), a broj prodanih bočica izlaznom ili ciljnom varijablom (eng. *Target*). Korištenjem nekog od algoritama nadziranog strojnog učenja i poznatih podataka moguće je istrenirati model koji će za neku novu vrijednost broja prijavljenih noćenja (ulaznu varijablu) predvidjeti potrošnju osvježavajućih pića (izlaznu varijablu). Slika 4 na kojoj je prikazan proces treniranja modela nadziranog učenja bit će nadograđena na način da će biti prikazana i primjena modela.



Slika 6. Kompletan prikaz treniranja i primjene modela nadziranog učenja

Slika 6 prikazuje treniranje i primjenu modela nadziranog učenja (Mehra & Hasanuzzaman, 2020). S lijeve strane je prikazano treniranje modela pri čemu se koriste poznati podaci, a s desne strane slike je vidljiva primjena modela s novim podacima. Pri korištenju modela za ulaz se koriste novi podaci iz kojih model vrši predikciju vrijednosti ciljne varijable. Bitno je naglasiti da je dijagram korištenja modela praktički isti kao i dijagram „klasičnog" programiranja koji je opisan na početku poglavlja. Ključna je razlika što kod klasičnog programiranja programer implementira model, odnosno „znanje" iskusnog korisnika u obliku funkcije Y=X*0.05, dok se kod strojnog učenja „znanje" dobiva iz poznatih podataka o broju prijavljenih noćenja turista i kupljenih bočica osvježavajućih pića. Zanimljivo je da se implementacija „znanja" kod jednostavnih algoritama strojnog učenja (linearna regresija) neće gotovo uopće razlikovati od klasične implementacije, osim u vrijednosti koeficijenta i konstante koja se pojavljuje u modelima linearne regresije. Upotrebom koeficijenata dobivenih strojnim učenjem, vjerojatno će greška biti manja, što znači da će rezultat biti točniji, a koeficijent bi za ovaj primjer vjerojatno imao nešto više decimala i tako dosegnuo veću preciznost. Funkcija u programskom kodu bit će ista – umnoškom broja prijavljenih noćenja i koeficijenta *K* dobiva se predikcija prodaje osvježavajućih pića.

Kao što je navedeno na početku opisa nadziranog učenja, da bi se provelo „treniranje" modela strojnog učenja potrebni su prethodno poznati podaci - ulazne i izlazne varijable. U navedenom primjeru postoji samo jedna ulazna (broj noćenja turista) i jedna izlazna varijabla (broj prodanih bočica osvježavajućeg pića). Ta jedina ulazna varijabla bi trebala biti dostupna u turističkoj zajednici mjesta ili kroz informacijski sustav www.evisitor.hr. Izlazna varijabla je broj prodanih bočica što bi također

trebalo biti dostupno u izvješću samoposlužnog aparata. Kako bi model bio što pouzdaniji, takvih podataka bi trebalo biti što više.

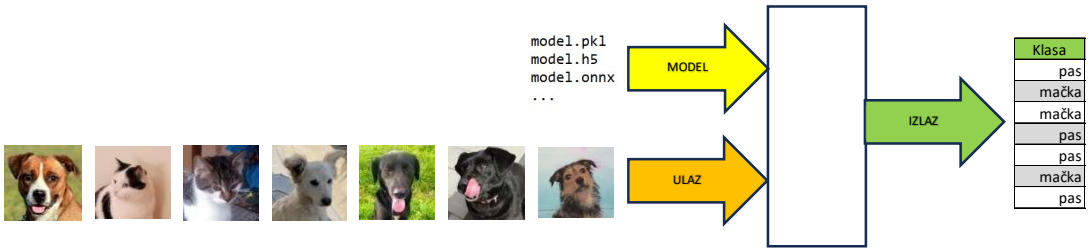
Za razliku od navedenog primjera, kompleksniji algoritmi nadziranog učenja koriste se primjerice za kategoriziranje životinja na digitalnim fotografijama. U tom slučaju ulazna varijabla je sama fotografija, odnosno pikseli na fotografiji, dok je izlazna varijabla naziv životinje na fotografiji. Za primjer, to mogu biti fotografije pasa i mačaka koje treba kategorizirati. Kod takvih kompleksnijih primjera potrebna je veća količina fotografija koja se mjeri u desecima i stotinama tisuća fotografija. Problem koji nastaje je definiranje vrijednosti ciljne varijable za svaku fotografiju, odnosno obilježavanje fotografija, kao što je u spomenutom primjeru kategorizacija - pas ili mačka. Taj dio posla trebao bi odraditi čovjek, a to u slučaju većeg broja fotografija predstavlja vremenski zahtjevan zadatak. Osim što je posao dugotrajan, prilično je i monoton pa se u zadnje vrijeme sve više istraživanja bavi algoritmima koji mogu postići visoku točnost i s malim brojem slučajeva koji se koriste za treniranje. Bez obzira na to, važno je još jednom naglasiti da je kod nadziranog učenja potrebno imati ulazne i izlazne podatke. Postojanje izlaznih varijabli koje nerijetko ručno definira čovjek kroz obilježavanje podataka ciljnom vrijednošću, temeljno je svojstvo grupe tehnika koji se nazivaju nadziranim učenjem.

Drugi primjer vezan uz prepoznavanje pasa i mačaka, naveden u prethodnom odlomku, razlikuje se od primjera sa samoposlužnim aparatom koji je prethodno detaljnije obrađen. U drugom primjeru spominje se podjela fotografija u dvije grupe, dok u prvom primjeru ne postoji podjela u grupe već model koji služi za predikciju i njime se izračunava pozitivna cjelobrojna vrijednost – broj bočica. Razlog tome je što navedeni primjeri pripadaju različitim grupama nadziranog učenja, a te grupe su:

a) regresija
b) klasifikacija.

Regresijski modeli kao rezultat daju brojčane vrijednosti na kontinuiranoj skali (Tatsat, et al., 2020). To može biti masa izražena u kilogramima, valutne vrijednosti, vodostaj rijeke izražen u cm, itd. (Khairi & Darmawan, 2021; Novarlia, 2022; Maulana, et al., 2021). U primjeru s osvježavajućim pićima, izlazna vrijednost je brojčana vrijednost, a za izradu mogu se koristiti tehnika regresije. Jedna od najjednostavnijih tehnika koja pripada u ovu grupu je Linearna regresija koja će biti obrađena u sljedećim poglavljima.

Modeli klasifikacije kao rezultat daju kategorijalnu varijablu (Tatsat, et al., 2020). Slika7 prikazuje korištenje prethodno opisanog kompleksnog modela prepoznavanja slika u kojem su postojale dvije kategorije: pas i mačka. Ulaz modela je fotografija životinje, a izlaz je jedna od dvije moguće kategorije. Broj kategorija je najčešće relativno mali, iako postoje kompleksni modeli koji prepoznaju tisuće kategorija. Jedan od primjera s velikim brojem kategorija jest prepoznavanje prometnih znakova na fotografijama kod autonomne vožnje. U ovoj grupi također postoji niz različitih tehnika koje će u sljedećim poglavljima biti obrađene, a to su Metoda potpornih vektora, Stabla odlučivanja i Slučajne šume. Važno je još spomenuti kako se veći dio tehnika koje će biti obrađene u priručniku mogu koristiti i za regresiju i za klasifikaciju , ali prije opisa konkretne tehnike bit će naznačeno za što će se koristiti.



Slika7. Primjena modela klasifikacije

1.1.2. Nenadzirano učenje

Za razliku od nadziranog, tehnike koje pripadaju u nenadzirano učenje ne zahtijevaju postojanje ciljnih varijabli (Tatsat, et al., 2020). U prethodno opisanom primjeru klasifikacije fotografija s obzirom na sadržaj (psi i mačke), podatke je potrebno ručno klasificirati. S druge strane, iste te fotografije životinja mogu se analizirati tehnikama nenadziranog učenja, pri čemu te tehnike funkcioniraju bez izlaznih varijabli. Tehnike nenadziranog učenja jednostavno „pokušavaju" pronaći sličnosti u podacima i na osnovu sličnosti nešto učiniti s podacima. Kao rezultat nenadziranog strojnog učenja, može biti podjela u grupe koje imaju slična svojstva, ali može biti i smanjenje broja ulaznih varijabli ukoliko pojedine ulazne varijable nemaju bitan utjecaj.

Kod ovih tehnika najčešće nema točnog rješenja s kojim se može uspoređivati rezultat, iz jednostavnog razloga što nema ciljnih varijabli. U primjeru s prepoznavanjem životinja na fotografijama pasa i mačaka, vjerojatno bi očekivali da tehnike nenadziranog učenja podjele slike pasa i slike mačaka u dvije grupe. Možda bi kao rezultat dobili model koji bi podijelio fotografije u dvije grupe, a u jednoj grupi bi možda bili bijeli psi i mačke, dok bi u drugoj bili crni psi i mačke. Boja životinje bi vjerojatno bila „lakše" prepoznata nego razlike u strukturi glave i tijela pasa i mačaka.

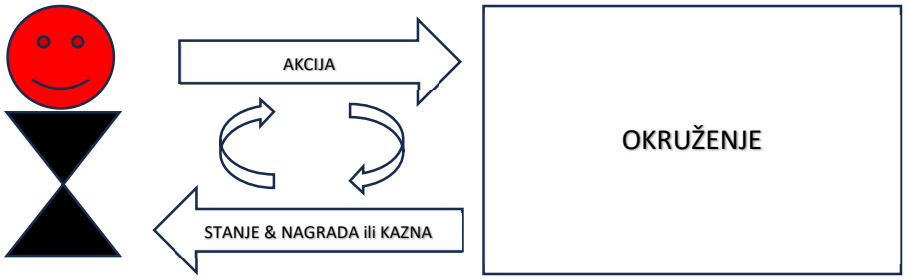
Slika 8. Primjer aktivnosti modela nenadziranog učenja

Slika 8 prikazuje aktivnost modela nenadziranog učenja pri čemu su u dva klastera ili grupe podijeljeni slučajevi koji imaju slične vrijednosti ulaznih varijabli. Ulazne varijable su nosivost, masa i pogon. Radi se o vozilima, pri čemu kao rezultat klasteriranja nastaju dva klastera ili grupe koje su očigledno automobili i kamioni. Treba obratiti pažnju da tehnike za klasteriranje uzimaju u obzir sve varijable i traže sličnosti pojedinih slučajeva uzevši u obzir vrijednosti svih dostupnih varijabli.

Neki primjeri tehnika nenadziranog učenja su tehnika K-srednjih vrijednosti (eng. *K-means Clustering*), Analiza glavnih komponenti (eng. *Principal Component Analysis*) i Hijerarhijsko klasteriranje (eng. *Hierarchical Clustering*). Potrebno je naglasiti kako osim klasteriranja, tehnike nenadziranog učenja mogu poslužiti i za smanjivanje dimenzionalnosti podataka (eng. *Dimensionality Reduction*) te za pronalaženje asocijativnih pravila (eng. *Association Rules*) (Sterne, 2017; Egger, 2022; Tatsat, et al., 2020).

1.1.3. Podržano učenje

Podržano učenje (eng. *Reinforcement Learning*) često se koristi pri objašnjavanju strojnog učenja i umjetne inteligencije jer na neki način rade na istom principu na kojem funkcioniraju živa bića kada uče metodom pokušaja i pogrešaka (Copeland, 2023). Trenirani model djeluje kao samostalni agent koji reagira na podražaj ili novu informaciju iz okruženja. Na osnovu ishoda reakcije, koji može biti pozitivan ili negativan, algoritam ažurira svoju strategiju. Za pozitivan ishod agent dobiva nagradu, a za negativan ishod dobiva kaznu. Jednostavan primjer u kojem se koristi tehnika podržanog učenja jest kretanje autonomnog robota u prostoriji s preprekama od neke početne do završne točke. Autonomni robot mora imati informaciju na kojoj se lokaciji nalazi i bira smjer u kojem ide. Za svaki sudar s preprekom dobiva kaznu u obliku negativnih bodova, dok za svaki uspješan prolaz dobiva nagradu u obliku pozitivnih bodova. Autonomni robot će uz podržano učenje vremenom „naći" najoptimalniji put od početne do završne točke.



Slika 9. Primjer modela podržanog učenja

Slika 9 prikazuje interakciju agenta s okolinom gdje se interakcija ostvaruje kroz akciju agenta i reakciju okruženja na način da se stvara novo stanje i dobiva nagrada ili kazna za prethodnu akciju. Nakon toga, agent ažurira strategiju i poduzima novu akciju. Radi se o ponavljajućem procesu koji dovodi do strategije koja je sve optimalnija.

Jedan od problema koji se pojavljuje s tehnikama podržanog učenja jest taj što je stvarno okruženje u kojem bi agent mogao učiti ponekad prilično skupo, a posebno kada bi agent trebao uništavati dijelove okruženja da bi optimizirao strategiju (Coad, 2021). Iz tog razloga je daleko jednostavnije koristiti simulirano okruženje, odnosno okolinu (Coad, 2021). S druge strane, simulirana okolina u kojoj su uključeni svi detalji i moguće situacije iz stvarnog svijeta ponekad predstavlja veliki problem za kreiranje. Iz ovog razloga, tehnike podržanog učenja još se ne koriste u većoj mjeri, ali se svakako radi o tehnikama koji će se sve više koristiti u budućnosti.

2. Provjera performansi modela

Jedan od važnih elemenata u primjeni strojnog učenja jest provjera performansi modela. Ako istrenirani model nije dovoljno dobar, neće se koristiti i pustiti u produkciju. Kod tehnika podržanog učenja model bi trebao pregledati čovjek i utvrditi je li dovoljno kvalitetno odrađuje zadatak za koji je treniran. Slična situacija je i kod tehnika nenadziranog učenja čije izlaze također treba pregledati i analizirati karakteristike klastera koje je tehnika generirala. S druge strane, modele nadziranog učenja lakše je kontrolirati jer su poznate vrijednosti ciljne varijable. S obzirom da se u ovom priručniku najviše obrađuju tehnike nadziranog učenja, posebna pažnja bit će posvećena provjeri performansi za tehnike nadziranog strojnog učenja. U nastavku će biti opisane osnovne metode provjere performansi za regresijske i klasifikacijske modele. Za regresijske modele to su: zbroj kvadrata pogreške, korijen srednje kvadratne pogreške i koeficijent determinacije. Za klasifikacijske modele to su: matrica konfuzije, točnost, preciznost, odziv i F1-mjera.

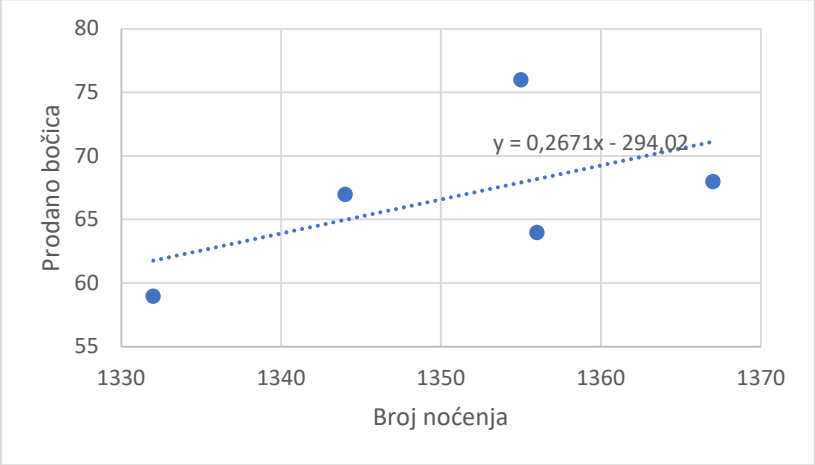
2.1. Provjera performansi regresijskih modela

Regresijski modeli predviđaju vrijednosti na kontinuiranoj numeričkoj skali tako da se provjera performansi svodi na usporedbu stvarnih vrijednosti i vrijednosti koje je predvidio model. Prethodni primjer s osvježavajućim pićima može poslužiti za prikaz provjere. Tablica 1 prikazuje podatke prodaje za pet dana srpnja, osim toga prikazan je i broj noćenja.

Tablica 1. Podaci prodaje i broja noćenja za 5 dana srpnja

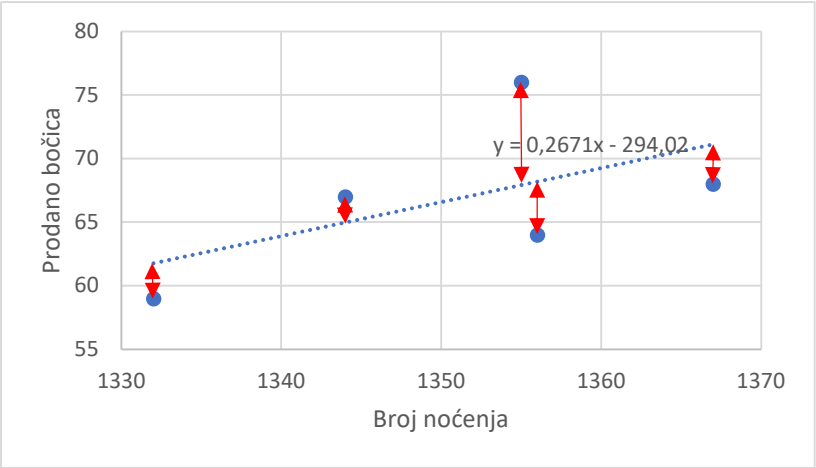
Datum	Broj noćenja	Prodano bočica
1.7.2022	1344	67
2.7.2022	1356	64
3.7.2022	1355	76
4.7.2022	1332	59
5.7.2022	1367	68

Slika 10 prikazuje tablične podatke grafički, a na osnovu podataka korištenjem tabličnog kalkulatora, linearnom regresijom dobivena je funkcija i pravac koji predstavlja model. Sama linearna regresija bit će opisana detaljnije u sljedećim poglavljima. Vidljivo je kako nagib pravca ukazuje da porastom broja noćenja raste i potrošnja osvježavajućih pića na samoposlužnom aparatu.



Slika 10. Grafički prikaz prodaje za 5 dana srpnja

Isto tako je vidljivo da vrijednosti koje se očitavaju na pravcu, odstupaju od izmjerenih vrijednosti označenih plavim kružićima koje su služile za izradu modela. Slika 11 prikazuje navedene razlike.

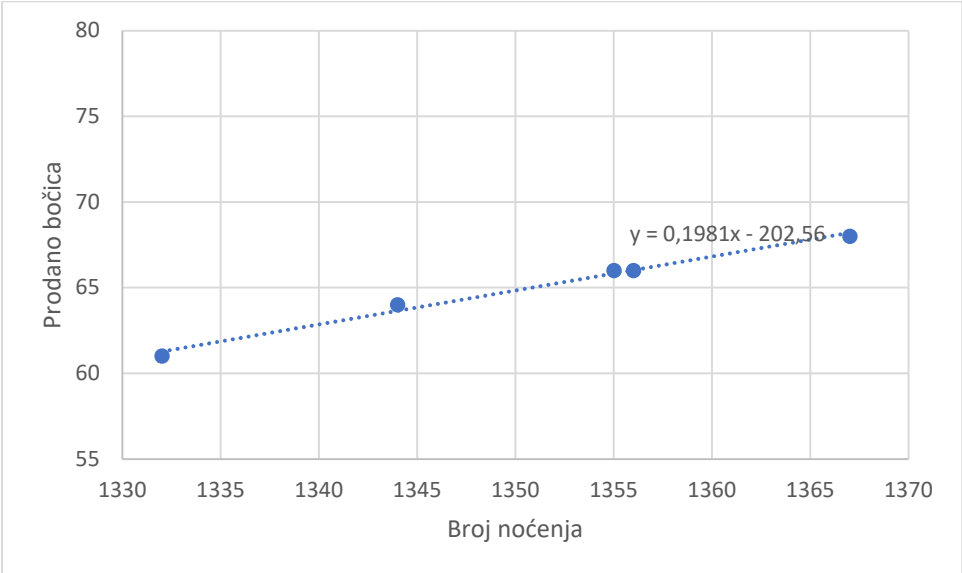


Slika 11. Odstupanja stvarnih podataka od modela

Što su razlike veće, model je lošiji, a što su razlike manje, model je bolji. Ove razlike između zavisne varijable i modelom predviđenih vrijednosti nazivaju se **rezidualnim odstupanjima** (Horvat & Mijoč, 2019), a mogu se izračunati prema (Horvat & Mijoč, 2019):

$$SSE = \sum_{k=0}^{n} (y - y)^2$$

Kraticom SSE (eng. *Sum of Square Errors*) – ***zbroj kvadrata pogreške*** označavaju se rezidualna odstupanja. Pogreškom se smatra razlika između modelom pretpostavljene i stvarne vrijednosti na okomitoj osi, a kvadriranjem se gubi predznak te razlike. Slika 12 prikazuje regresijski model na kojem je zbroj kvadrata pogreške znatno manji od prethodno prikazanog modela, što znači da model na toj slici znatno bolje opisuje pojavu. Rezidualna odstupanja predstavljaju varijabilnost modela koji nije moguće predvidjeti modelom (Horvat & Mijoč, 2019).



Slika 12. Regresijski model s malim odstupanjima stvarnih podataka od modela

Ako se navedena mjera rezidualnih odstupanja podijeli s brojem uzoraka i izvadi se kvadratni korijen iz rezultata, dobiva se popularna mjera ***korijen srednje kvadratne pogreške*** ili RMSE (eng. *Root Mean Square Error*). Izraz za izračunavanje korijena srednje kvadratne pogreške je (Géron, 2022):

$$RMSE = \sqrt{\frac{SSE}{n}}$$

Pod pretpostavkom da su rezidualna odstupanja normalno distribuirana, vrijednost RMSE je standardna devijacija rezidualnih odstupanja te se može koristiti kao interval pouzdanosti. Vrijednost RMSE je najveća pogreška koju se može očekivati u 68 % slučajeva, a RMSE pomnožen s brojem dva predstavlja najveću pogrešku koju se može očekivati u 95 % slučajeva (Kelley & Lai, 2011). Treba spomenuti kako je model bolji što je RMSE manji, kao i kod mjere zbroja kvadrata pogreške.

Da bi se dokazala razlika standardnih devijacija rezidualnih odstupanja za oba primjera, može ih se izračunati. Tablica 2 sadrži podatke koji su korišteni za izradu prvog grafikona s većom pogreškom. Tablica 3 sadrži podatke koji su korišteni za izradu drugog grafikona sa znatno manjom pogreškom.

Tablica 2. Predviđanje prvog modela

Datum	Broj noćenja	Prodano bočica	Predviđanje modela
1.7.2022	1344	67	64,9624
2.7.2022	1356	64	68,1676
3.7.2022	1355	76	67,9005
4.7.2022	1332	59	61,7572
5.7.2022	1367	68	71,1057

Slika 10 prikazuje funkciju i pravac s većim RMSE prema podacima iz tablice 2 i on iznosi 4,57.

Tablica 3. Predviđanje drugog modela

Datum	Broj noćenja	Prodano bočica	Predviđanje modela
1.7.2022	1344	64	63,6864
2.7.2022	1356	66	66,0636
3.7.2022	1355	66	65,8655
4.7.2022	1332	61	61,3092
5.7.2022	1367	68	68,2427

Slika 12 prikazuje funkciju i pravac s manjim RMSE prema podacima iz tablice 3 i on iznosi 0,23. Takav rezultat je očekivan jer su rezidualna odstupanja manja na drugom grafikonu na slici 12.

Još jedna mjera kojom se mjere performanse regresijskih tehnika jest **koeficijent determinacije**, a označava se s R^2 (Horvat & Mijoč, 2019). Izraz kojim ga se može izračunati je (Horvat & Mijoč, 2019):

$$R^2 = 1 - \frac{\sum(y - \tilde{y})^2}{\sum(y - \bar{y})^2}$$

Koeficijentom determinacije dobiva se udio varijacije koji je objašnjen modelom. Što je koeficijent determinacije bliže broju 1, to model bolje opisuje pojavu (Horvat & Mijoč, 2019). U gornjem izrazu varijabla y označava izmjerenu vrijednost, odnosno broj kupljenih osvježavajućih pića. Varijabla $\tilde{y}$ označava broj predviđen modelom, a $\bar{y}$ srednju vrijednost broja kupljenih osvježavajućih pića. Da bi se utvrdilo kako koeficijent determinacije stvarno ukazuje na udio varijacije objašnjen modelom, može se

izračunati za prethodno prikazana dva modela sa slika 10 i 12. Slika 10 prikazuje model kod kojeg bi koeficijent determinacije trebao biti manji od koeficijenta za model sa slike 12. Slika 12 prikazuje model kod kojeg bi koeficijent determinacije trebao biti blizu 1. Ispod su tablice i izračuni za oba modela.

Tablica 4. Ukupni podaci prvog modela

Datum	Broj noćenja (x)	Prodano bočica (y)	$\tilde{y}$	$(y-\tilde{y})^2$	$\bar{y}$	$(y-\bar{y})^2$
1.7.2022	1344	67	64,9624	4,15181376	66,8	0,04
2.7.2022	1356	64	68,1676	17,36888976	66,8	7,84
3.7.2022	1355	76	67,9005	65,60190025	66,8	84,64
4.7.2022	1332	59	61,7572	7,60215184	66,8	60,84
5.7.2022	1367	68	71,1057	9,64537249	66,8	1,44
			SSE	104,3701281		154,8

$$R_a^2 = 1 - \frac{104,37}{154,80} = 0,33$$

Tablica 5. Ukupni podaci drugog modela

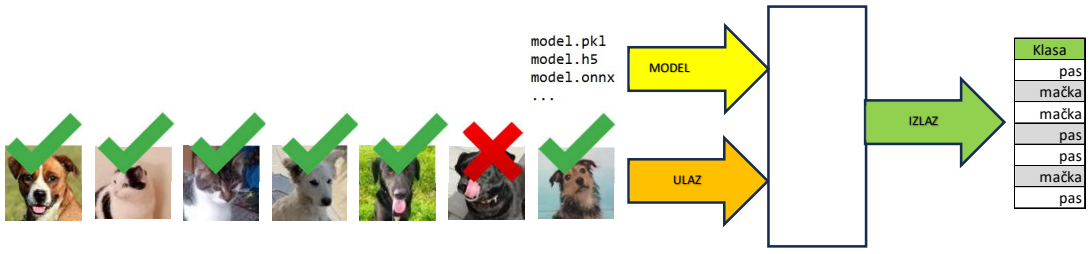
Datum	Broj noćenja (x)	Prodano bočica (y)	$\tilde{y}$	$(y-\tilde{y})^2$	$\bar{y}$	$(y-\bar{y})^2$
1.7.2022	1344	64	64,9624	0,92621376	65	1
2.7.2022	1356	66	68,1676	4,69848976	65	1
3.7.2022	1355	66	67,9005	3,61190025	65	1
4.7.2022	1332	61	61,7572	0,57335184	65	16
5.7.2022	1367	68	71,1057	9,64537249	65	9
			SSE	19,4553281		28

$$R_b^2 = 1 - \frac{0,27}{28} = 0,99$$

Slika 10 prikazuje model s koeficijentom determinacije koji iznosi 0,33, dok za drugi model on iznosi 0,99. Iz grafičkih prikaza vidljiva je razlika između dobivenih vrijednosti. Model s koeficijentom determinacije od 0,99 gotovo bez greške opisuje pojavu.

2.2. Provjera performansi klasifikacijskih modela

Za razliku od regresijskih modela, kod klasifikacije se može reći kako je provjera performansi jednostavnija jer je potrebno samo pobrojiti i usporediti točno i netočno klasificirane slučajeve. Prethodno je opisan model klasifikacije fotografija na kojima su psi ili mačke. Takvom istreniranom modelu dovoljno je dostaviti određeni broj fotografija na kojima su psi i mačke i usporediti izlaz modela sa stvarnim stanjem.



Slika 13. Prikaz klasifikacijskog modela u primjeni

Slika 13 prikazuje 7 fotografija koje je model klasificirao, a dobiveni je izlaz prikazan tablično na izlazu modela. Na ulazu, od ukupno 7 fotografija, 4 su fotografije pasa i 3 fotografije mačaka. Model je dobro klasificirao 6 fotografija, a jedna fotografija nije dobro klasificirana. Da bi se provela analiza performansi modela koristit će se matrica konfuzije (eng. *Confusion Matrix*) (Liu, 2017). Tablica 6 prikazuje matricu konfuzije s dvije kategorije. Značenja kratica su (Liu, 2017):

a) TP (eng. *True Positives*) – podaci koji pripadaju pozitivnoj kategoriji i ispravno klasificirani kao takvi.
b) TN (eng. *True Negatives*) – podaci koji pripadaju negativnoj kategoriji i ispravno klasificirani kao takvi.
c) FP (eng. *False Positives*) – podaci koji pripadaju negativnoj kategoriji i netočno klasificirani kao pozitivni.
d) FN (eng. *False Negatives*) – podaci koji pripadaju pozitivnoj kategoriji i netočno klasificirani kao negativni.

Tablica 6. Matrica konfuzije

		PREDIKCIJA	
		1	0
STVARNO STANJE	1	TP	FN
	0	FP	TN

Tablica 7 prikazuje matricu konfuzije za konkretan primjer s psima i mačkama. Model je obradio 7 fotografija od čega je 5 fotografija pasa. Model je 4 fotografije označio kao fotografije pasa, što znači da se u 4 slučaja stvarno stanje i predikcija poklapaju. Jednu fotografiju psa model je označio kao fotografiju mačke i to se vidi u matrici konfuzije. Ćelija s brojem 0 označava kako niti jednu fotografiju mačke model nije označio kao fotografiju psa, a ćelija s brojem 2 označava da je dvije fotografije mačke model označio ispravno, kao fotografije mačke. Valja spomenuti kako u primjeru nema pozitivne i negativne kategorije, nego kategorije psi i mačke. Zbog jednostavnosti primjera pretpostavit će se da su psi pozitivna kategorija, a mačke negativna te da model služi za izradu uređaja koji psima omogućuje pristup hrani, dok mačkama ne dozvoljava.

Tablica 7. Matrica konfuzije s unesenim podacima

		PREDIKCIJA	
		Pas	Mač
STVARNO STANJE	Pas	4	1
	Mač	0	2

Na osnovu matrice konfuzije moguće je izračunati nekoliko mjera na osnovu kojih je moguće vrednovati model (Liu, 2017). To su redom:

a) Točnost
b) Preciznost
c) Odziv
d) F1-mjera.

Točnost (eng. *Accuracy*) je omjer ispravno klasificiranih slučajeva i ukupnog broja slučajeva (Liu, 2017). Računa se kao:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

U primjeru s fotografijama izračun točnosti izgleda ovako:

$$Acc = \frac{4+2}{4+2+\ 0+1} = \frac{6}{7} = 0{,}86$$

Kod ove mjere bitno je pripaziti na ravnomjernu raspoređenost veličine kategorija. Ako u primjeru s dvije kategorije postoji situacija da je broj fotografija jedne kategorije znatno veći od broja fotografija druge kategorije, moguće je dobiti veliku točnost, a da model možda uopće ne prepoznaje fotografije kategorije koje ima malo.

Bitno je naglasiti da ista vrijednost koja se izračuna za model može biti izvrsna za neko područje primjene, dok je za drugo niska. Drugim riječima kod procjene dobivene točnosti treba biti oprezan i uzeti u obzir područje primjene (Artasanchez & Joshi, 2020).

Preciznost (eng. *Precision*) je mjera koja označava koliko je klasifikator pouzdan kad označi slučaj kao pozitivan, odnosno od koliko je svih pozitivnih predviđanja klasifikator bio točan (Liu, 2017; Powers, 2020). Računa se kao:

$$P = \frac{TP}{TP + FP}$$

U primjeru za fotografije pasa dobivamo:

$$P = \frac{4}{4+0} = 1{,}00$$

Odziv (eng. *Recall*) ili **osjetljivost** (eng. *Sensitivity*) je mjera koja prikazuje koliko je od svih pozitivno označenih slučajeva klasificirano ispravno, a računa se kao mjera istinito pozitivnih i sveukupnog broja slučajeva koji su klasificirani kao pozitivni (Liu, 2017; Powers, 2020). Formula glasi:

$$R = \frac{TP}{TP + FN}$$

U primjeru za fotografije pasa odziv ili osjetljivost iznosi:

$$R = \frac{4}{4+1} = 0{,}80$$

Posljednja mjera koja se obrađuje je **F1-mjera**(eng. *F1-measure, F1-score*). Radi se o harmonijskoj sredini između preciznosti i odziva (Liu, 2017; Powers, 2020), a računa se na sljedeći način:

$$F1 = \frac{2PR}{P+R}$$

U primjeru F1-mjera za kategoriju pasa iznosi 0,89, a izračun je u nastavku.

$$F1 = \frac{2 * \ 1 * \ 0{,}80}{1 + 0{,}80} = 0{,}89$$

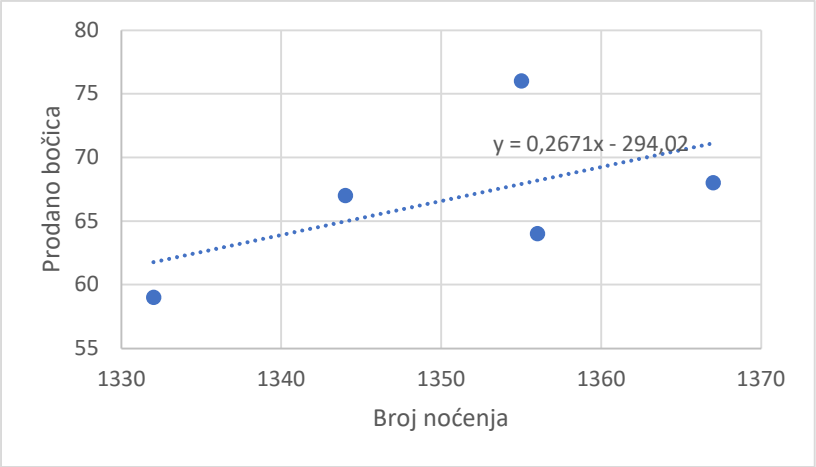
2.3. Prenaučenost i podnaučenost

Prenaučenost (eng. *Overfitting*) i podnaučenost (eng. *Underfitting*) su pojmovi koji se spominju u kontekstu strojnog učenja u procjeni generalizacije modela.

Da bi se stekla ideja o čemu se radi, dovoljno je zamisliti kako imaginarni ulagač ulaže u nove *startup* tvrtke i zanima ga na koji način bi prepoznao koja tvrtka će mu donijeti višestruki povrat ulaganja. Za taj analitički posao potrebno je osigurati snažno računalo i podatke o svim dostupnim *startup* tvrtkama koje su u godinu dana postojanja ostvarile promet veći od milijun dolara. Analiza

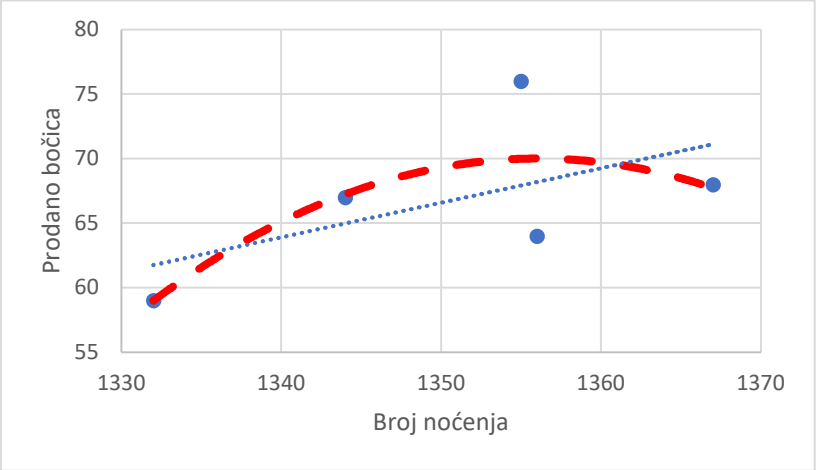
može početi nakon pribavljanja podataka u obliku velike tablice sa stotinama tisuća redaka podataka o uspješnim tvrtkama iz kojih snažno računalo omogućuje pronalaženje skrivenih pravila koja povezuju te uspješne tvrtke. Nakon detaljne automatske analize dostupnih podataka, dobiva se odgovor na postavljeno pitanje. „Najveći povrat uloženog ostvarit će se ako se novci ulože u tvrtku koja se bavi poluvodičima, sa sjedištem u Taipeiu (Taiwan), u tvrtku koja je osnovana u petak, a treće slovo imena tvrtke je H!". Nakon zbunjenih pogleda u rezultat, imaginarni ulagač shvaća kako je moćno računalo našlo poveznice zajedničke najuspješnijim tvrtkama, ali jasno je i da neki elementi prijedloga nemaju veze s uspješnošću u poslovanju. Očigledno je model prenaučen na osnovu dostupnih podataka. No, ako se izbace druga dva uvjeta (dan osnivanja tvrtke i treće slovo imena), rezultat zvuči puno logičnije.

Drugi primjer za prenaučen/podnaučen model moguće je vidjeti na već poznatom primjeru prodaje osvježavajućih pića. Slika 14 prikazuje linearni model koji je u nekim situacijama prejednostavan da bi opisao neke pojave pa se može reći kako je model podnaučen (eng. *Underfitting*).



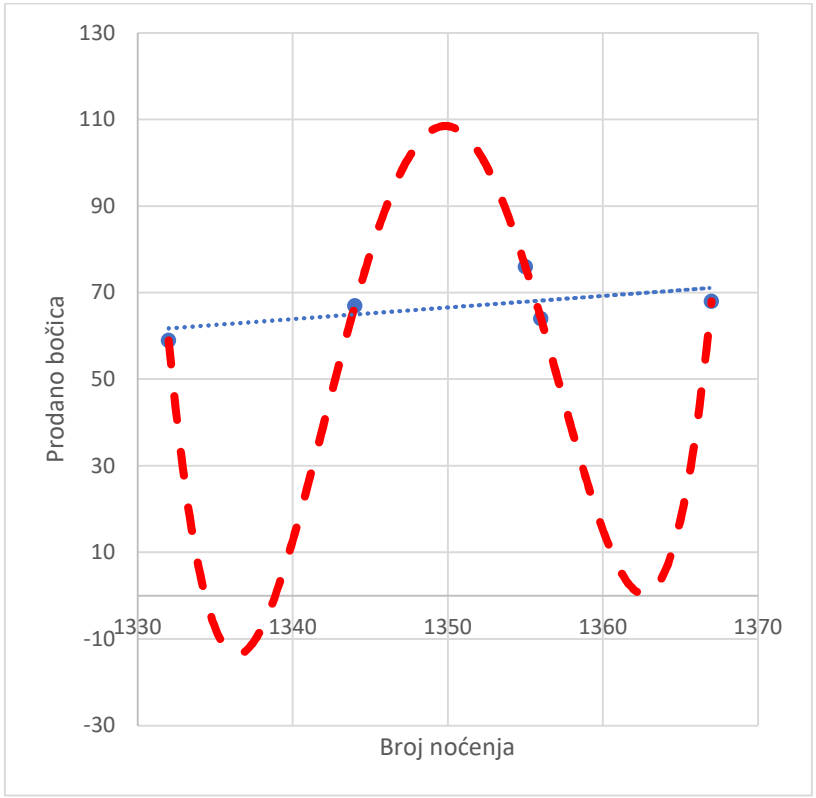
Slika 14. Linearni model

Slika 15 odnosno crvena isprekidana crta prikazuje graf polinomske regresije drugog stupnja nastao na osnovu pet dostupnih redova podataka. Polinomska regresija koristi se kada postoji razlog vjerovati kako je odnos između dvije varijable krivolinijski (Ostertagová, 2012). Za ovaj graf može se reći da za nijansu bolje opisuje pojavu, a ako bi računali RMSE, vjerojatno bi se dobila manja vrijednost u usporedbi s vrijednošću koja je dobivena za linearnu regresiju, što pokazuje da bi polinomska regresija bila bolji izbor za izradu modela povezanosti potrošnje pića i broja noćenja. Koeficijent determinacije koji ukazuje na udio varijance koji je objašnjen modelom bio bi svakako veći od prethodno prikazane linearne regresije.



Slika 15. Graf polinomske regresije drugog stupnja

Slika 16 odnosno crvena isprekidana crta prikazuje graf polinomske regresije četvrtog stupnja. Za njega se može reći da je RMSE jednak nuli što označava da praktički nema razlike između stvarnih vrijednosti i vrijednosti dobivenih modelom, a koeficijent determinacije je vrlo blizu jedinici. To se vidi jer isprekidana crvena krivulja gotovo savršeno prati pet točaka na grafu na osnovu kojih je model generiran. Za model se može reći da je izuzetno dobro istreniran na dostupnim podacima, ali ako se očita potrošnja za 1335 noćenja vidi se da je ona -10. Drugim riječima model previđa da će za 1335 noćenja turisti u samoposlužni aparat doslovno ugurati još 10 bočica osvježavajućeg pića, jer to je manje nevjerojatan način da se opiše negativna potrošnja. Drugi više nevjerojatan način bi bio da samoposlužni aparat sam proizvede još 10 bočica pića. U svakom slučaju radi se o modelu koji je prenaučen i potpuno neupotrebljiv za primjenu s nekim drugim podacima koji nisu podaci za treniranje.



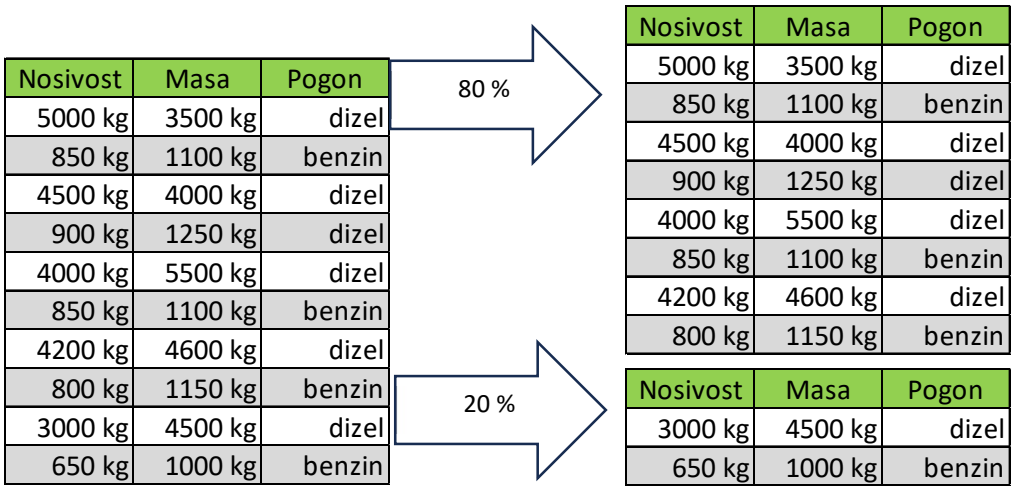
Slika 16. Graf polinomske regresije četvrtog stupnja

Tablica 8 prikazuje tipične karakteristike podnaučenog, balansiranog i prenaučenog modela (Watt et al., 2020). Podnaučen model je u pravilu prejednostavan i zbog svoje jednostavnosti ima lošije rezultate te se u pravilu ne koristi u produkciji. Prenaučen model je dosta kompleksan s visokom točnošću kada se testira s podacima na kojima je treniran, ali kada se testira s novim podacima rezultati su loši. Balansiran model je optimalan jer ima prihvatljivu i približno istu razinu točnosti i s nepoznatim podacima kao i s onim podacima na kojima je treniran.

Tablica 8. Karakteristike modela

	Podnaučen model	**Balansiran model**	**Prenaučen model**
KOMPLEKSNOST MODELA	Niska	Srednja	Visoka
PERFORMANSE – PROŠLI SLUČAJEVI	Niske	Srednje	Visoke
PERFORMANSE – BUDUĆI SLUČAJEVI	Niske	Srednje	Niske

Iz tablice je vidljivo da podnaučeni i prenaučeni modeli imaju loše performanse s podacima na kojima nisu trenirani, odnosno s budućim slučajevima. Jednostavna metoda da se dođe do tzv. budućih slučajeva jest priprema podataka prije no što se krene trenirati model i to na način da se skup podataka podijeli na dva dijela. Jedan dio podataka čine podaci za treniranje, a drugi dio podaci za testiranje. Model se trenira isključivo s podacima za treniranje. Omjer ove dvije grupe podataka može biti proizvoljan, ali u pravilu je skup za treniranje znatno veći od skupa podataka za testiranje. Često se susreću omjeri 70 %:30 % ili 80 %:20 %. Osim toga važno je naglasiti kako bi podjela u ove dvije grupe trebala biti temeljena na slučajnom izboru. Slika 17 prikazuje primjer podjele. Kod podjele je najvažniji detalj da se podaci za testiranje nikada ne koriste za treniranje modela.



Slika 17. Podjela skupa podataka na dio za treniranje i dio za testiranje

Osim podjele na podatke za treniranje i testiranje, u literaturi se spominju i podaci za validaciju. Radi se o trećem skupu podataka koji se koristi prilikom treniranja modela kako bi se tokom samog treniranja provjerilo koliko kvalitetno je model nešto naučio. S obzirom da će se u priručniku koristiti jednostavnije tehnike strojnog učenja, skupovi za validaciju se neće koristiti, a neke tehnike ih u KNIME-u ni ne podržavaju.

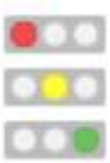
3. Program KNIME Analytics Platform

Program *KNIME Analytics Platform* je besplatan program za analizu podataka čija osnovna prednost je što omogućuje kompleksne transformacije i analize podataka bez ijednog reda programskog koda, no svakako zahtijeva poznavanje svojstava i načina funkcioniranja tehnika, kako bi rezultati bili ispravni i valjani. Program *KNIME Analytics Platform* temelji se na vizualnom programiranju kroz izradu hodograma (eng. *Workflow*) sa sastavnim dijelovima, što ga čini dobrim izborom za cijeli niz struka. U nastavku priručnika ovaj će se program ukratko nazivati KNIME. Sam program se razvija od 2004. godine (KNIME AG, 2023).

Rad s podacima u programu KNIME je intuitivan i temelji se na postavljanje čvorova na bijelu površinu koja će se u nastavku nazivati uređivač hodograma. Čvorovi predstavljaju određenu aktivnost nad podacima, a ovisno o željenoj aktivnosti postavlja se i odgovarajući čvor. U pravilu čvorovi se postavljaju od lijeve strane uređivača hodograma prema desnoj strani. Detaljnije o postavljanju čvorova bit će riječi u nastavku ovog poglavlja. Kada se uz pomoć čvorova opišu sve transformacije s podacima ili kada se generira model temeljen na odabranim podacima, taj niz čvorova može se spremiti kao cjelina u obliku hodograma (eng. *Workflow*).

Nakon instalacije programa KNIME u repozitoriju čvorova dostupno je na stotine čvorova s različitim funkcionalnostima. U pravilu bi za jednostavnije zadatke svi čvorovi trebali biti dostupni već u osnovnoj verziji koja je instalirana. Ako čvor sa željenom funkcionalnošću nije vidljiv, dostupan je repozitorij naziva *KNIME Hub* koji nudi pretraživanje čvorova koji se nalaze na poslužitelju proizvođača programa i koje je potrebno naknadno instalirati. Osim dodatnih čvorova u sklopu repozitorija *KNIME Hub*, dostupni su i različiti hodogrami s objašnjenjima koji mogu pomoći u rješavanju nekog problema, a ponekad se može naći i kompletno rješenje problema.

S obzirom da čvor predstavlja aktivnost potrebno je naglasiti kako čvor može biti u različitim stanjima. Kada se čvor postavi, potrebno ga je povezati i konfigurirati. Čvor „daje do znanja" u kojem je stanju kroz semafor s tri stanja, odnosno svjetla. Prvo je crveno i označava da čvor još nije konfiguriran. Drugo je žuto i označava da je čvor konfiguriran i spreman za pokretanje, odnosno aktivaciju. Nakon što se čvor pokrene, aktivnost za koju je čvor namijenjen se izvršava, a na semaforu se pali zeleno svjetlo. Zeleno svjetlo označava da je čvor uspješno završio zadanu aktivnost. Slika 18 prikazuje sva tri navedena stanja čvora.



Slika 18. Stanja na vodoravnom semaforu ispod čvora

Na vodoravnom semaforu ispod čvora mogu se pojaviti još dvije oznake. Slika 19 prikazuje oznaku upozorenja i radi se o žutom trokutu s uskličnikom koji se pojavljuje kad čvor djelomično odradi aktivnost, ali zbog nekih okolnosti tu aktivnost nije moguće odraditi u potpunosti. Primjer u kojem se pojavljuje upozorenje jest kada čvor može obrađivati samo stupce s brojčanim vrijednostima, a jedan od stupaca sadrži tekstualne vrijednosti. U tom slučaju čvor obradi stupce koje može, a stupac s tekstualnim vrijednostima ignorira i pri tom prikazuje upozorenje.

Slika 19. Oznaka upozorenja

Slika 20 prikazuje oznaku greške te se radi o crvenom krugu s bijelim slovom „x“ na sredini kruga. Oznaka greške pojavljuje se kada čvor iz nekog razloga nije u mogućnosti odraditi aktivnost za koju je predviđen. Primjer u kojem se pojavljuje oznaka greške je kad se čvoru dostavi prazna tablica bez podataka te se nakon toga čvor izvršava. Detalji o problemu mogu se pročitati u konzoli ili ako se nakratko postavi pokazivač miša iznad čvora.

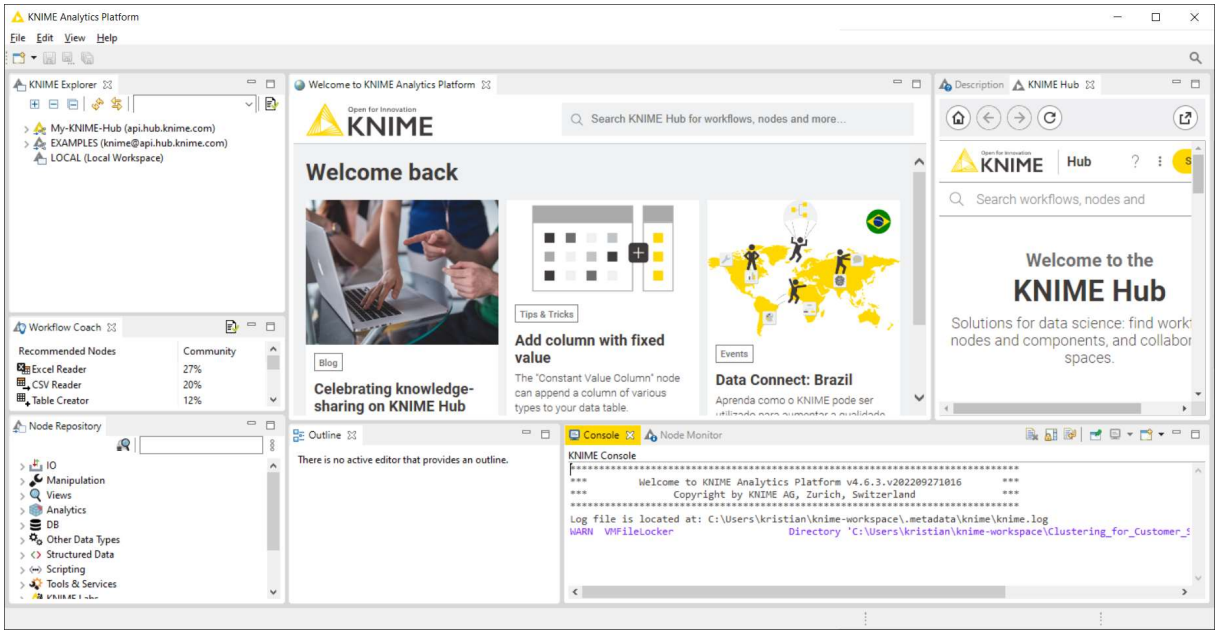
Slika 20. Oznaka greške

Kratke napomene o čvoru moguće je umetnuti umjesto zadanog teksta ispod čvora.

Osim programa KNIME koji je potpuno besplatan i čiji kod je otvoren, tvrtka *KNIME AG* iz Švicarske nudi i komercijalnu serversku verziju programa koja može biti zanimljiva većim tvrtkama jer omogućuje suradnju u e-okruženju među korisnicima u tvrtki.

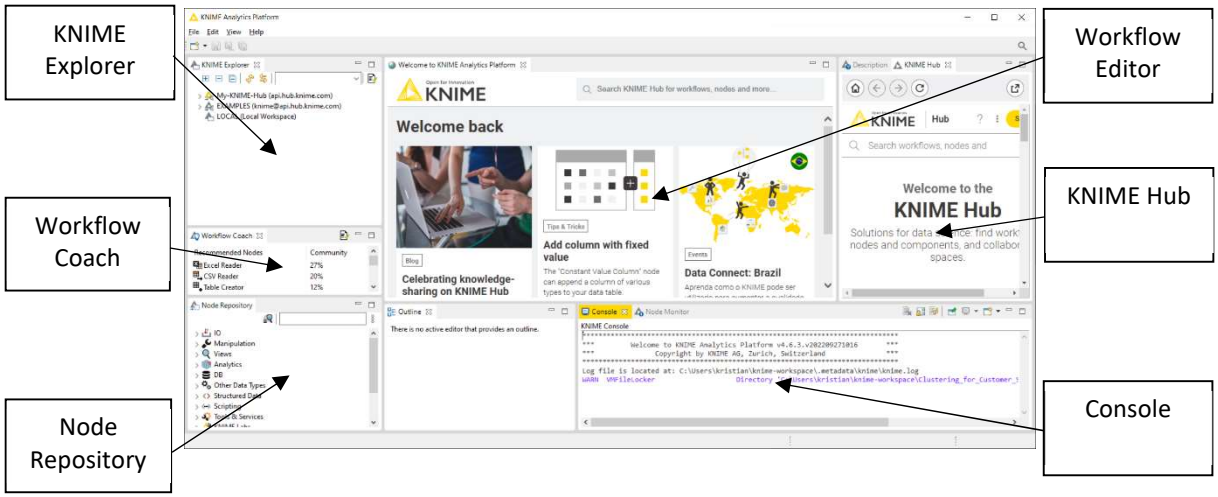
Od 2023. godine tvrtka *KNIME AG* je od verzije 5 izmijenila sučelje programa i učinila ga jednostavnijim. U trenutku pisanja ovoga priručnika za preuzimanje je dostupna verzija 5.1 pri čemu je instalacija gotovo ista kao i kod verzija 4.x (slovo x označava broj podverzije). Sučelje verzije 5.1 je značajno izmijenjeno, iako se osnovna logika postavljanja čvorova na hodogram nije izmijenila. Nova verzija nudi mogućnost prebacivanja na sučelje iz starije verzije što se naziva *KNIME classic user interface*, odnosno klasično korisničko sučelje. Ovaj priručnik temeljen je na verziji KNIME 4.x, odnosno na klasičnom korisničkom sučelju verzije KNIME 5.x. U narednim potpoglavljima opisano je i klasično i djelomično novo sučelje programa KNIME, kao i postupak prelaska s novog na klasično korisničko sučelje.

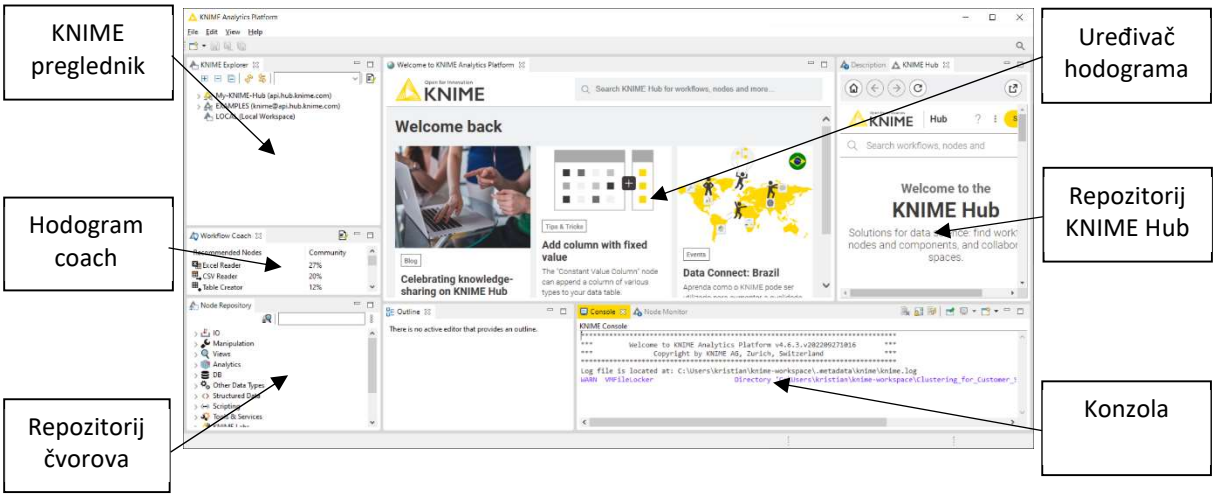
3.1. Sučelje programa KNIME 4.x



Slika 21. Izgled sučelja

Sučelje programa KNIME na engleskom je jeziku. Da bi probleme vezane uz prevođenje dijelova sučelja sveli na najmanju moguću mjeru, u nastavku će biti prikazani nazivi sučelja na hrvatskom i na engleskom jeziku. Kod prevođenja pojam „*node*“ je preveden kao „*čvor*“ što je doslovan prijevod, a pojam „*workflow*“ je preveden kao „*hodogram*“. Slika 21 prikazuje originalne nazive dijelova sučelja. Slika 22 prikazuje djelomičan prijevod na hrvatski jezik.



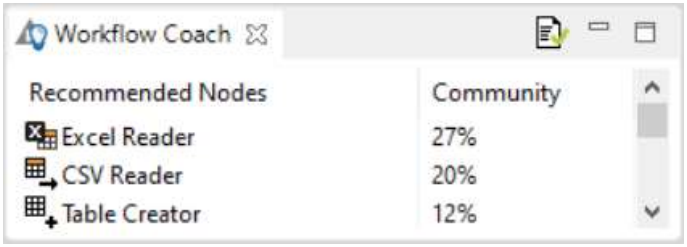


Slika 22. Originalni nazivi i prijevod dijelova sučelja

Nakon što je dio pojmova preveden u nastavku je navedeno čemu služe pojedini dijelovi sučelja.

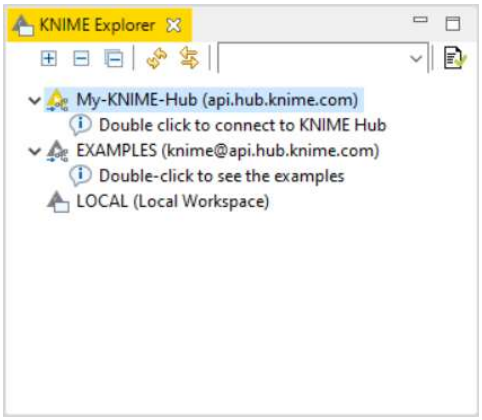
a) KNIME preglednik (eng. *KNIME Explorer*) – pregled dostupnih hodograma.
b) Hodogram coach (eng. *Workflow Coach*) – predlaže sljedeći čvor u hodogramu na osnovu prikupljenih statističkih podataka od strane korisnika.
c) Repozitorij čvorova (eng. *Node Repository*) – pregled svih dostupnih čvorova s pretraživačem.
d) Uređivač hodograma (eng. *Workflow Editor*) – prostor u kojem se uređuje hodogram.
e) KNIME hub (eng. *KNIME Hub*) – pretraživač javno dostupnih čvorova.
f) Konzola (eng. *Console*) – ispis grešaka i upozorenja.

Slika 23 prikazuje *Hodogram coach* koji služi za preporuku jednoga od mogućih idućih čvorova u nizu. Preporuke se kreiraju na osnovu statističkih podataka prikupljenih od korisnika programa KNIME.



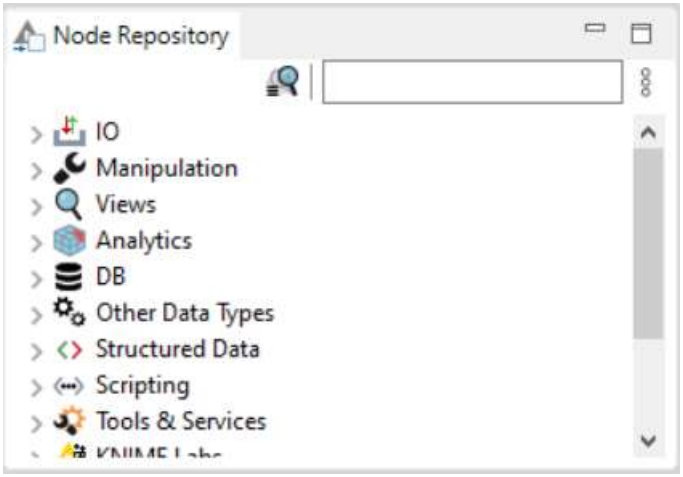
*Slika 23. Hodogram coach (*Workflow Coach*)*

Slika 24 prikazuje KNIME preglednik koji služi za osnovni rad s datotekama koje se koriste u programu. Nakon instalacije, u pregledniku su vidljive sljedeće mape: *My-KNIME-HUB, EXAMPLES i LOCAL* mape. Prve dvije mape se nalaze na serverima tvrtke i omogućuju udaljenu pohranu, odnosno pregledavanje primjera. Treća mapa *LOCAL* je, kako joj samo ime kaže, smještena lokalno i to u mapu koja je izabrana prilikom prvog pokretanja programa KNIME. Do te mape može se doći koristeći i preglednik datoteka operativnog sustava.



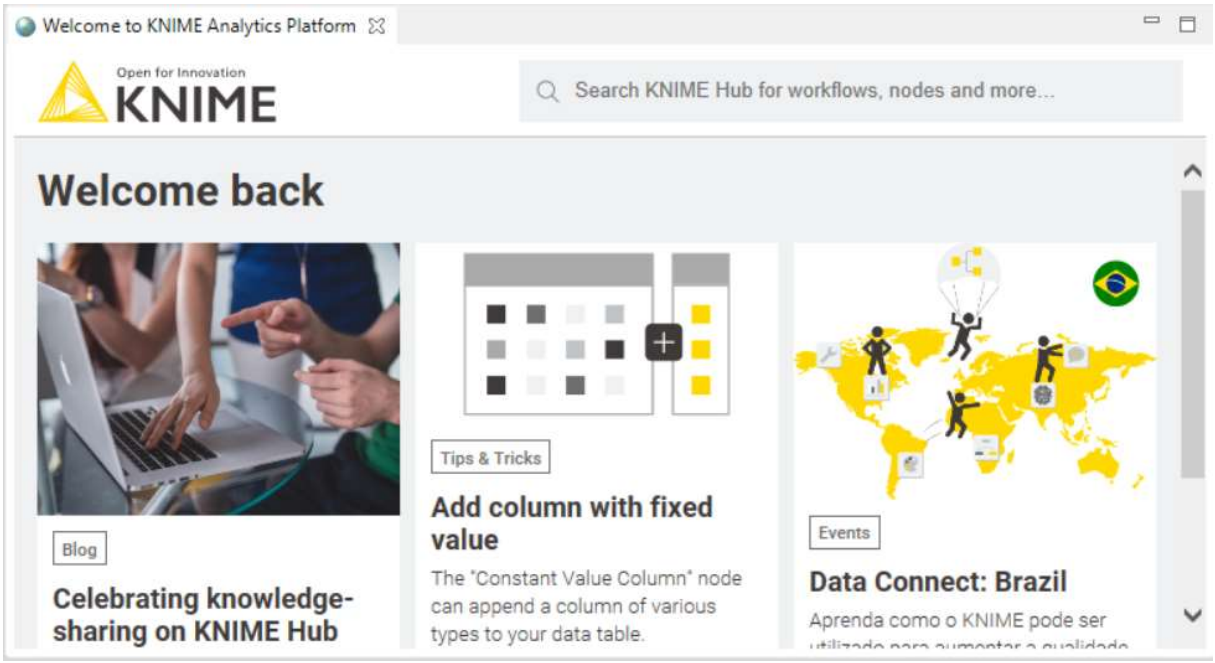
Slika 24. KNIME preglednik (KNIME Explorer)

Slika 25 prikazuje repozitorij čvorova. U ovom se repozitoriju nalaze stotine čvorova razvrstanih u grupe, no u gornjem desnom kutu je dostupna tražilica koju se također može koristiti ako je poznat dio imena čvora.



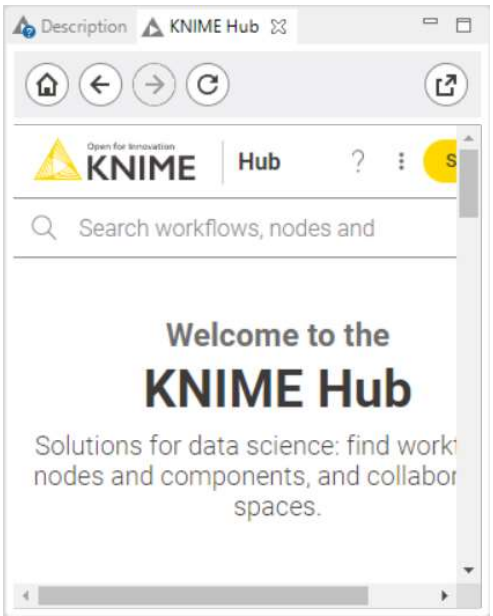
*Slika 25. Repozitorij čvorova (*Node Repository*)*

Slika 26 prikazuje uređivač hodograma. To je prostor na koji se postavljaju čvorovi i koji se povezuju stvarajući hodograme. Radi se o dijelu prozora programa *KNIME Analytics Platform* koji predstavlja „glavni prozor“, odnosno najviše se koristi.



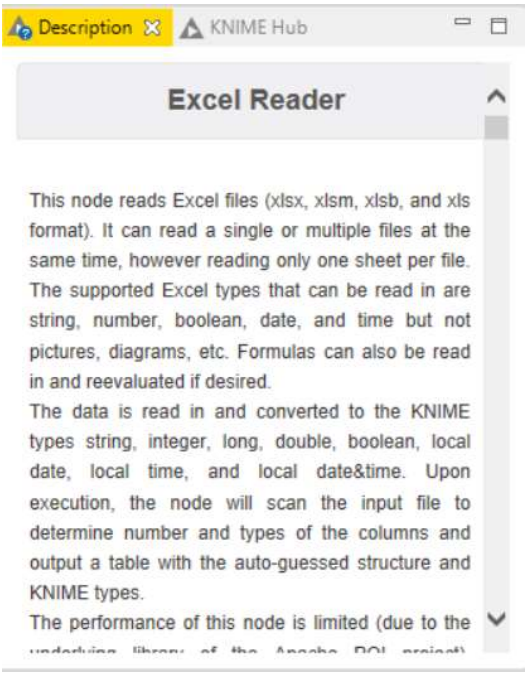
Slika 26. Uređivač hodograma (eng. Workflow Editor*)*

Slika 27 prikazuje repozitorij *KNIME Hub,* alat koji omogućuje pretraživanje čvorova i već gotovih hodograma. Ukoliko čvor nije inicijalno dostupan u repozitoriju čvorova (eng. *Node Repository*), moguće da je dostupan u repozitoriju *KNIME Hub* i tada je dovoljno povući ga u uređivač hodograma (eng. *Workflow Editor*), čime se pokreće instalacija toga istog čvora. Nakon nekoliko klikova, čvor je instaliran i možete ga se koristiti. Isto tako, mogu se pretraživati i kompletni hodogrami, ali oni se povlače u željenu mapu KNIME preglednika. U pravilu se radi o mapi *LOCAL*.



Slika 27. Repozitorij KNIME hub *(KNIME Hub)*

Slika 28 prikazuje opis izabranog čvora koji je dostupan na istom mjestu gdje i repozitorij *KNIME Hub*, ali se koristi druga kartica toga dijela prozora.

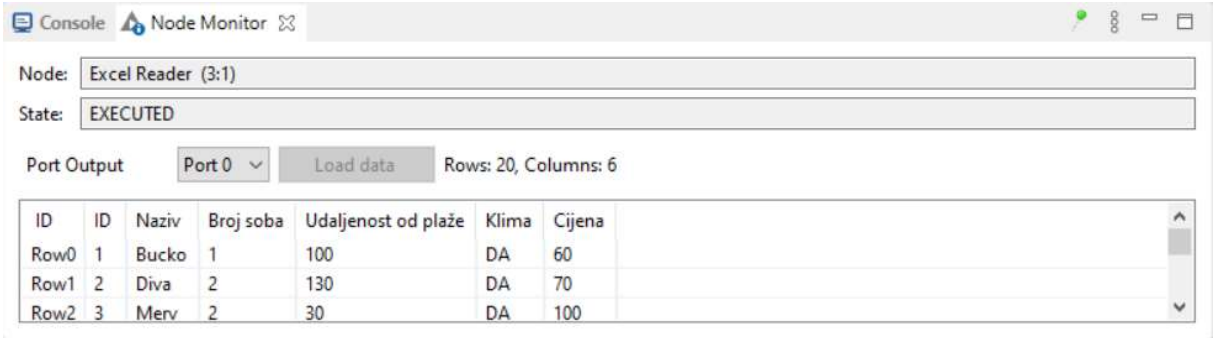


Slika 28. Opis izabranog čvora

Slika 29 prikazuje konzolu koja se nalazi u donjem dijelu prozora programa KNIME. To je mjesto na kojem će se ispisivati poruke o greškama ili upozorenjima.

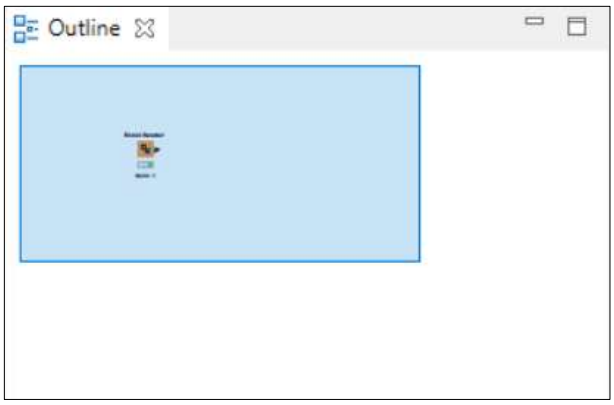
Slika 29. Konzola

Slika 30 prikazuje monitor čvorova (eng. *Node Monitor*). Radi se o dijelu prozora programa koji se preklapa s konzolom i u gornjem dijelu se izabire željeni dio. Monitor čvorova daje osnovne informacije o čvoru, kao što su vrsta čvora, stanje i slično.



Slika 30. Monitor čvora

Slika 31 prikazuje shemu hodograma. Radi se o umanjenom uređivaču hodograma koji služi za lakše snalaženje u samom uređivaču. U nekim situacijama kada postoji puno čvorova, ova shema služi da bi imali pregled nad cijelim hodogramom (Silipo, 2011).

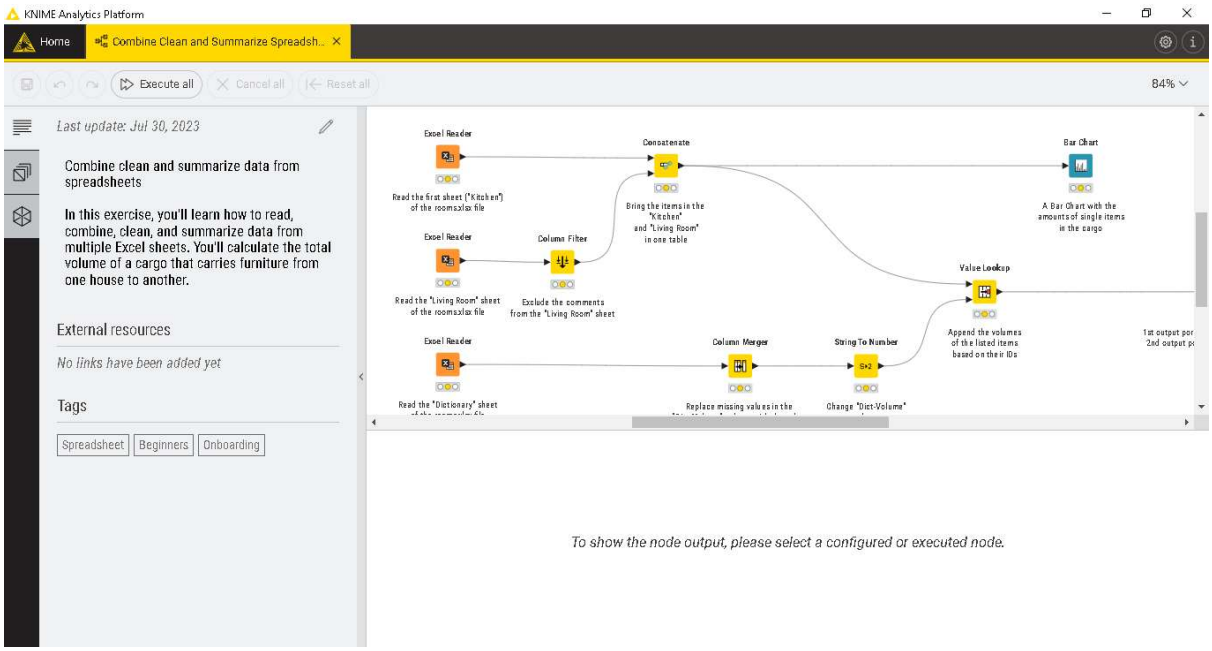


Slika 31. Shema hodograma

3.2. Sučelje programa KNIME 5.x i prelazak na klasično sučelje

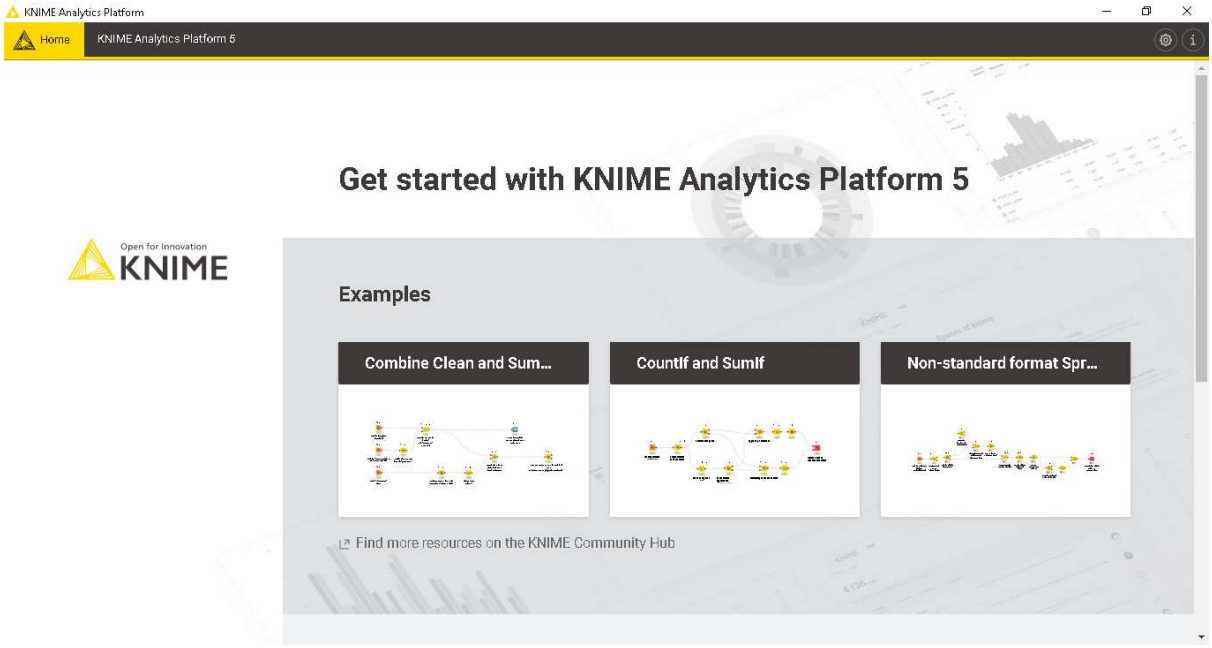
Godine 2023. KNIME AG je od verzije 5.x izmijenila sučelje programa KNIME. Instalacija programa je ostala ista, ali nakon pokretanja program izgleda drugačije. S obzirom da je ovaj priručnik pisan za verziju programa KNIME 4.x, u ovom dijelu će kratko biti opisano sučelje verzije 5.x i način prelaska na tzv. klasično korisničko sučelje koje je jednako sučelju iz verzija 4.x. Osnovna logika izgradnje hodograma korištenjem čvorova nije se promijenila, tako da se svi primjeri iz priručnika mogu napraviti koristeći novo sučelje.

Opis novoga sučelja kreće od pokretanja programa. Slika 32 prikazuje primjer hodograma koji je prikazan nakon pokretanja programa KNIME 5.x.



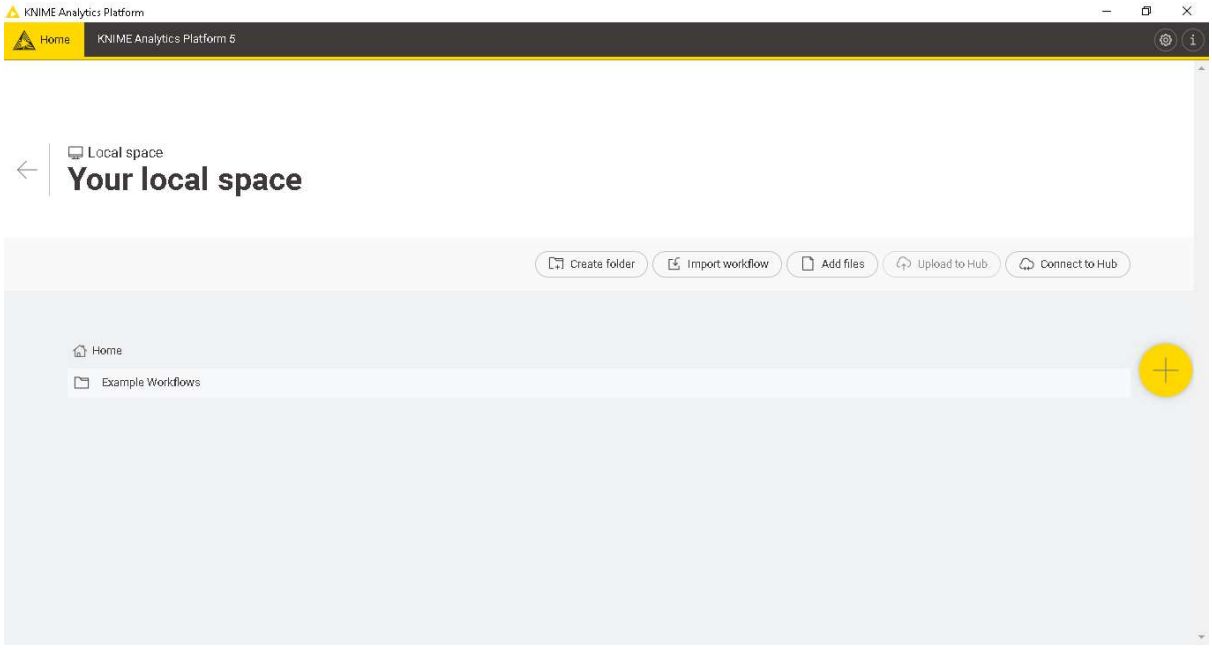
Slika 32. Primjer hodograma prikazan nakon pokretanja programa KNIME 5.x

Klikom na karticu *Home* u gornjem lijevom kutu programa dobiva se početno sučelje kao što prikazuje slika 33.



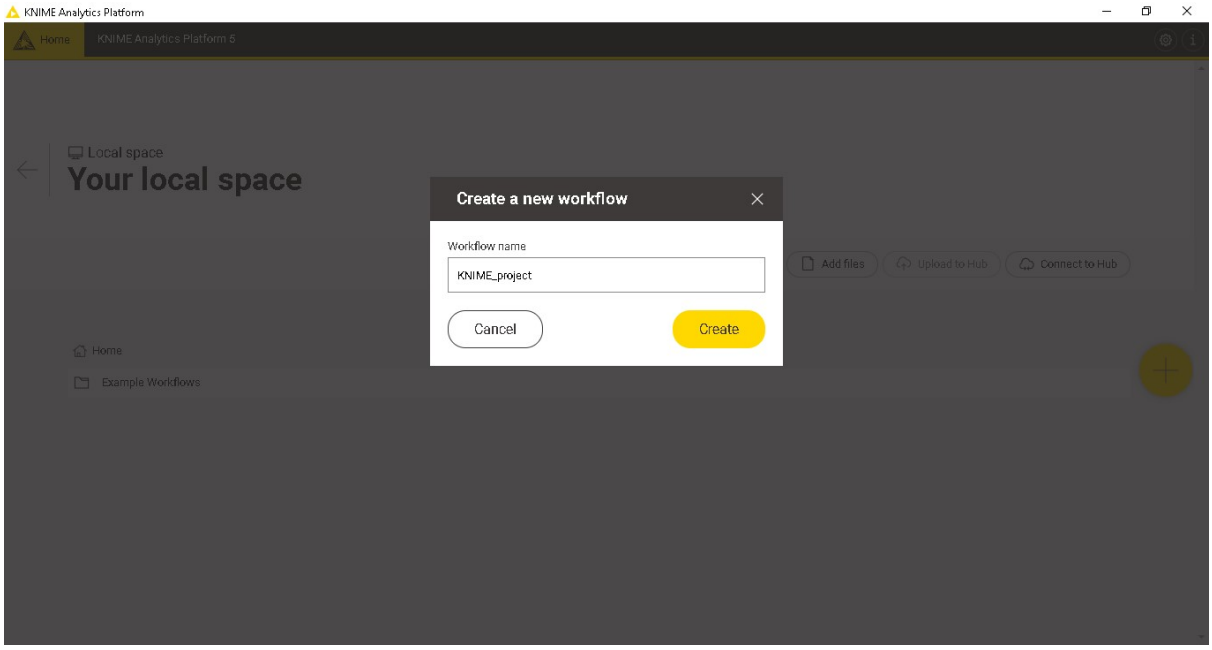
Slika 33. Mogućnosti na početnom sučelju programa KNIME

Da bi se kreirao hodogram korištenjem klizača potrebno se spustiti u donji dio prozora. Slika 34 prikazuje taj donji dio prozora.



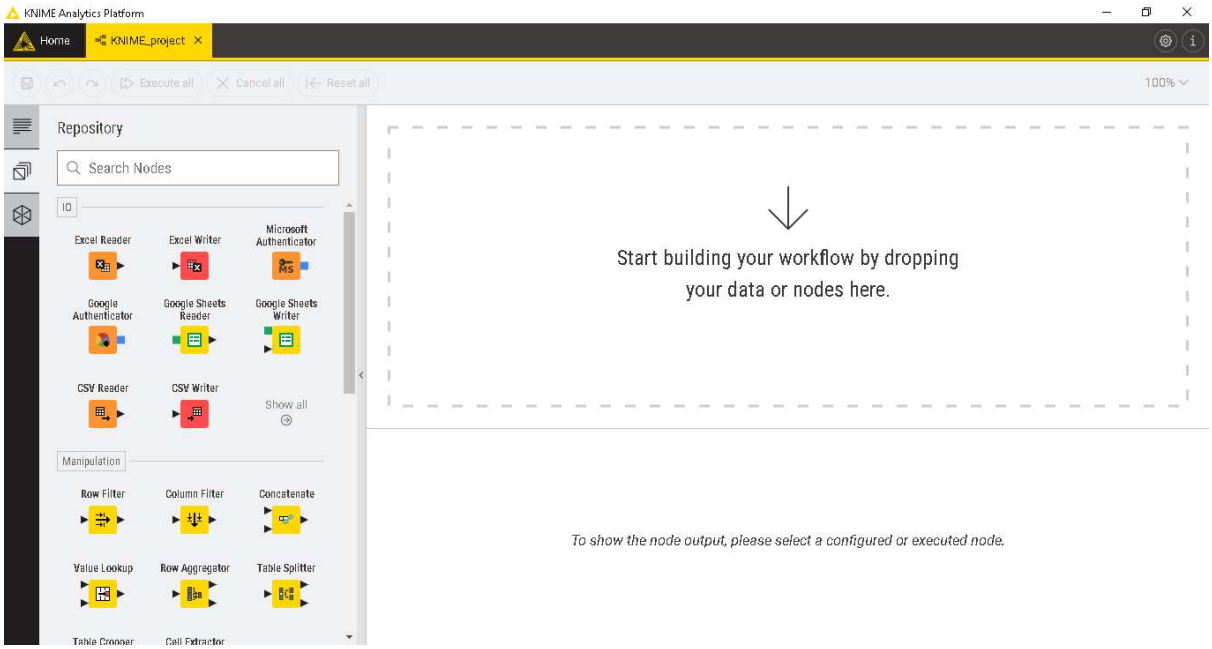
Slika 34. Dio mogućnosti iz donjeg dijela početnog prozora

U donjem dijelu prozora potrebno je kliknuti na žuti krug s desne strane na kojem se nalazi znak +. Time se dodaje novi hodogram, ali prije toga potrebno je dati naziv hodogramu. Slika 35 prikazuje izgled dijaloga u koji se unosi ime novog hodograma.



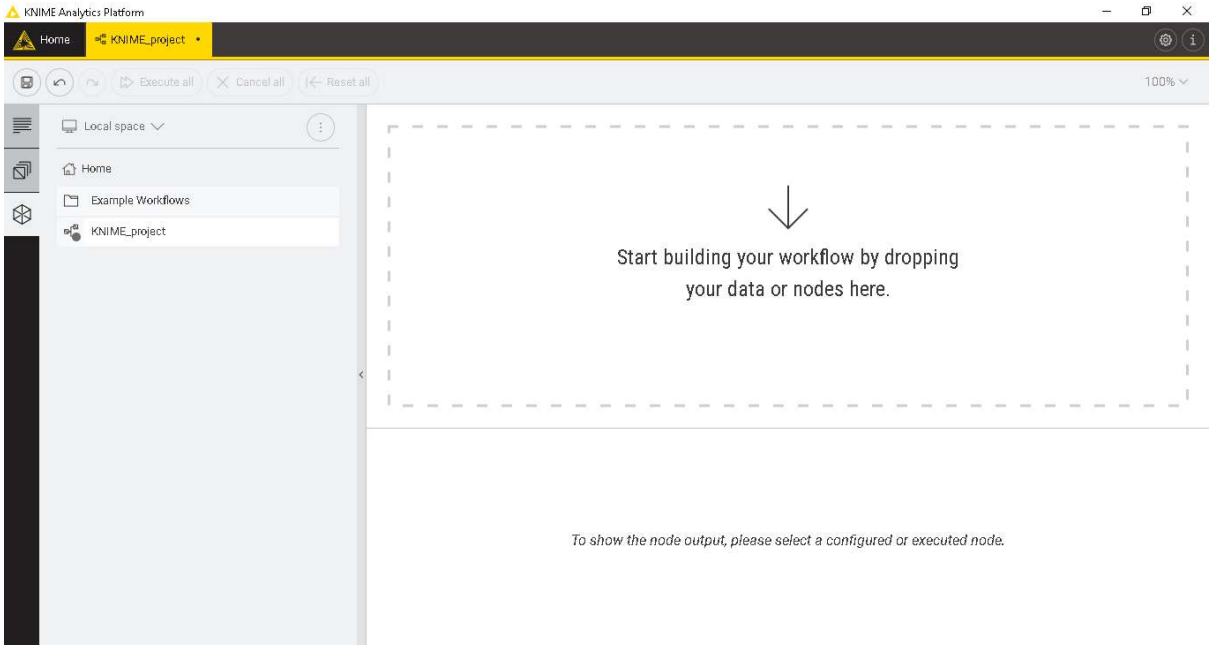
Slika 35. Definiranje imena novom hodogramu

Nakon kreiranja praznog hodograma prikazuje se novo radno sučelje programa KNIME 5.x. Novo sučelje na slici 36 sadrži manje elemenata od sučelja iz prethodnih verzija, ali osnovna logika izgradnje hodograma se nije promijenila.



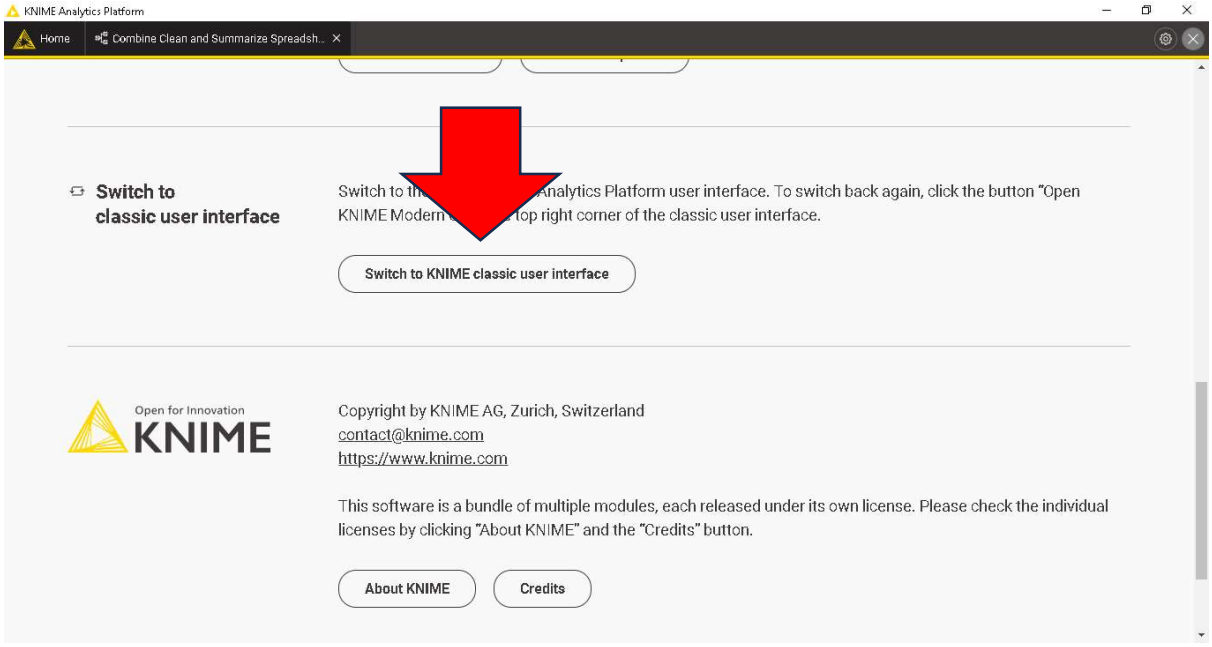
Slika 36. Novo sučelje KNIME 5.x

Datoteke dostupne u radnom okruženju dobivaju se klikom na šesterokut s lijeve strane prozora. Slika 37 prikazuje pogled na datoteke, a trenutno se vidi samo prazan hodogram po imenu *KNIME_project*.



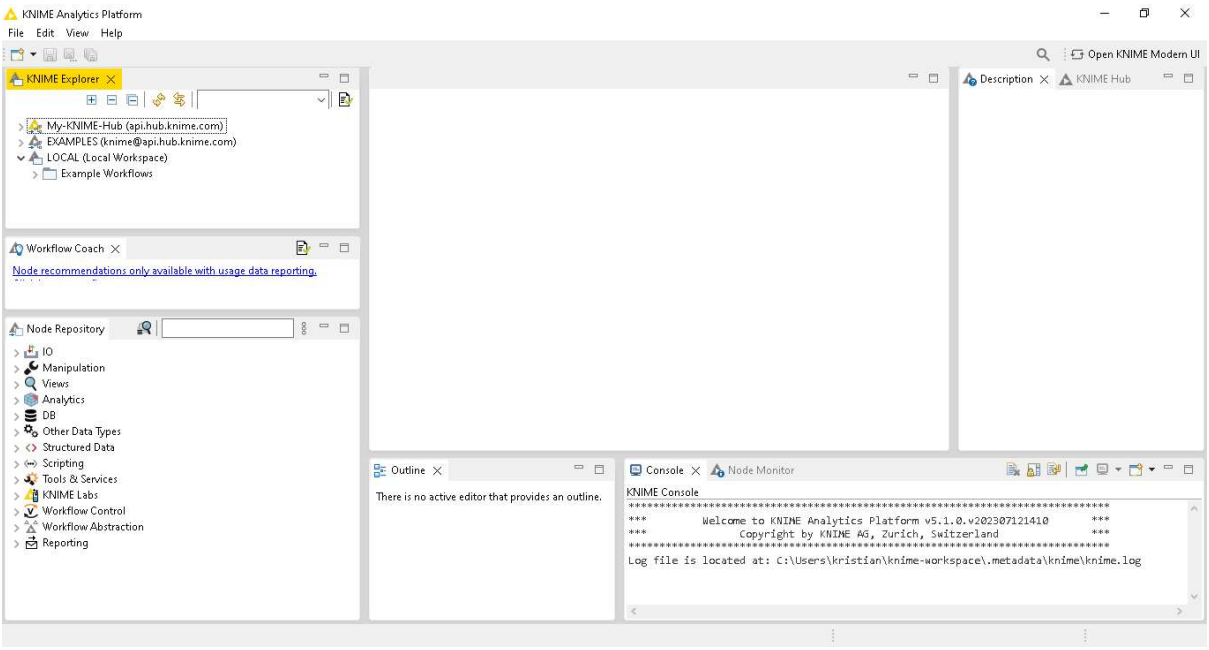
Slika 37. Pristup datotekama radnog okruženja

Konačno se dolazi do načina prelaska na klasično korisničko sučelje. Ono je dostupno klikom na veliko slovo *I* u gornjem desnom kutu prozora, nakon čega se treba spustiti u donji dio ekrana koristeći klizač i kliknuti na dugme *Switch to classic user interface,* što bi u prijevodu značilo „Prebaci na klasično sučelje“. Slika 38 prikazuje izgled tog ekrana, a crvena strelica ukazuje na položaj gumba.



Slika 38. Dugme za prelazak na klasično korisničko sučelje

Transformacija u klasično korisničko sučelje se odvija trenutno. Slika 39 prikazuje klasično korisničko sučelje koje se koristi u priručniku. Za povratak u novo sučelje dovoljno je kliknuti na gumb u gornjem desnom kutu “*Open KNIME Modern UI”, što bi značilo „Otvori moderno KNIME korisničko sučelje“.*



Slika 39. Klasično korisničko sučelje na verziji KNIME 5.x

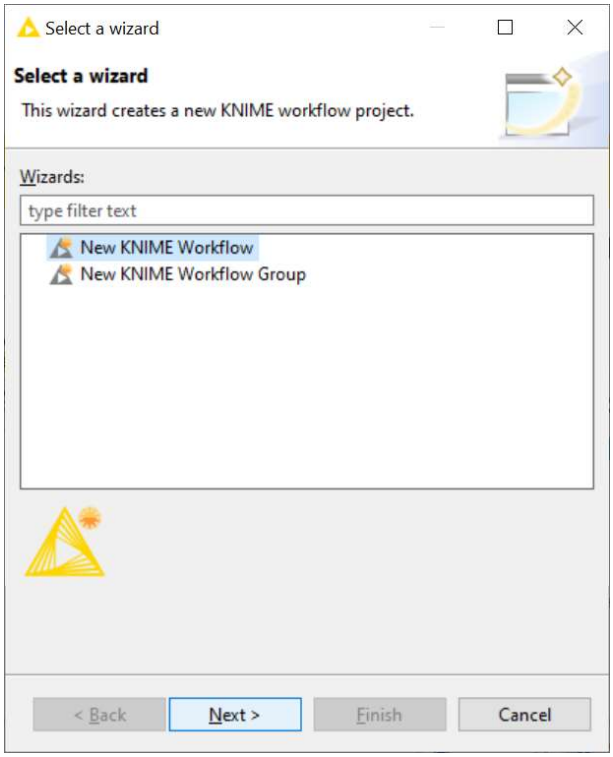
3.3. „Zdravo svijete“ u KNIME 4.x

Prvi korak u učenju nekog programskog jezika često počinje s izradom programa koji se naziva „Hello world!” ili u prijevodu „Zdravo svijete!”. Radi se o vrlo jednostavnom programu kojim se testira osnovna funkcionalnost programskog jezika, uređivača i prevodioca, a funkcija mu je da na zaslonu ispiše tekst „Hello world!” ili „Zdravo svijete!”. U uređivaču teksta ili tabličnom kalkulatoru nešto slično nema previše smisla, ali u programiranju da bi napisali takav jednostavan program potrebno je poznavati minimum pravila, kao što je naredba za ispis na ekran te način na koji se program prevodi u strojni jezik i pokreće, odnosno interpretira ako je u pitanju interpreter.

Da bi se upoznali s programom KNIME, u nastavku će biti opisan način na koji će se kreirati jednostavan hodogram. Konkretno, kreirat će se model linearne regresije koristeći podatke iz primjera sa samoposlužnim aparatom za osvježavajuća pića i izračunati kolika će biti potrošnja ako je prijavljeno samo 1000 noćenja. Navedeni primjer je vrlo jednostavan i u praksi neupotrebljiv jer se model generira iz samo pet redova podataka iz prethodnog primjera, ali je sasvim dovoljan za osnovno razumijevanje načina generiranja hodograma.

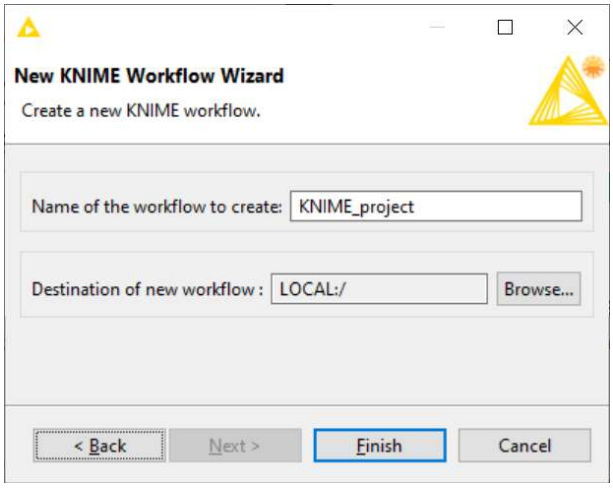
3.3.1. Kreiranje novog hodograma

Očekivano, kao i u većini programa pod operativnim sustavom *Microsoft Windows,* nova datoteka se kreira u izborniku *File* (Datoteka) i klikom na *New* (Nova). Nakon toga se otvara dijaloški okvir u kojem se bira kreiranje novog *KNIME* hodograma (*New KNIME Workflow*) ili grupe novih KNIME hodograma (eng. *New KNIME Workflow Group*). Taj dijaloški okvir prikazuje slika 40.



Slika 40. Čarobnjak za kreiranje KNIME projekta

Nakon odabira novog *KNIME* hodograma (eng. *New KNIME Workflow*) potrebno je kliknuti na *Next>* i otvara se novi dijaloški okvir. Slika 41 prikazuje drugi dijaloški okvir u kojem je potrebno definirati ime hodograma i mjesto pohrane. Za sada se odabire ponuđena lokacija: LOCAL:/.



Slika 41. Dijaloški okvir za definiranje imena i lokacije hodograma

Ukoliko se ostavlja zadano ime hodograma, u prostoru uređivača hodograma pojavljuje se bijeli prazan prostor sa zaglavljem *KNIME_project* u kojem će se postavljati čvorovi, koji će se povezivati i gdje će se na kraju kreirati modeli. Ovaj dio se opisuje u sljedećoj sekciji.

3.3.2. Uređivanje hodograma

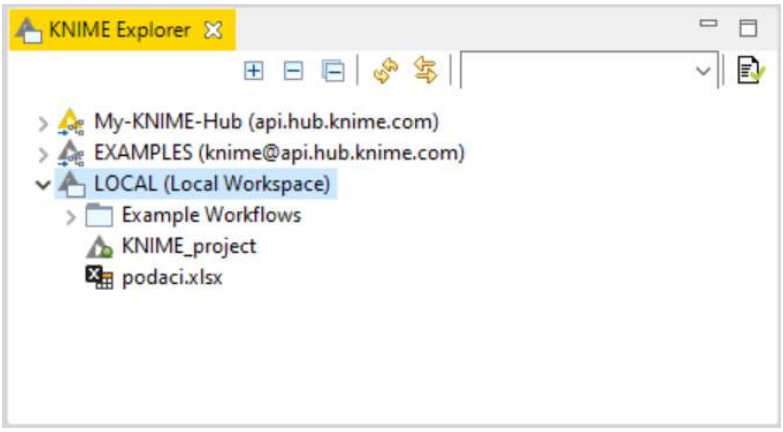
Da bi se generirao bilo koji model koristeći odabranu tehniku strojnog učenja, najprije je potrebno učitati podatke za treniranje. Nakon toga se treniranom modelu dostavljaju novi podaci, kako bi model

mogao predvidjeti vrijednost ciljne varijable. Primjer koji se odnosi na broj prodanih bočica, a koji se koristio u prvom dijelu, koristit će se i u nastavku, što znači da su potrebni prethodno korišteni tablični podaci. Tablica 9 sadrži podatke koji su korišteni u prethodnim primjerima, a potrebno ih je prepisati u radni list programa *Microsoft Excel* i pohraniti kao XLSX datoteku u radni prostor programa *KNIME*.

Tablica 9. Podaci broja noćenja i prodanih bočica

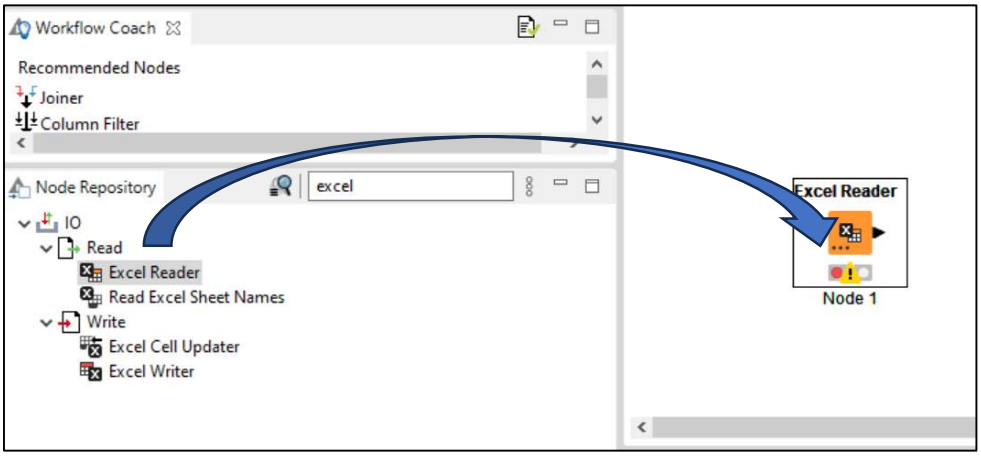
Datum	Broj noćenja	Prodano bočica
1.7.2022	1344	67
2.7.2022	1356	64
3.7.2022	1355	76
4.7.2022	1332	59
5.7.2022	1367	68

Slika 42 prikazuje *KNIME* preglednik nakon spremanja podataka iz tablice u XLSX datoteku pod imenom *podaci.xlsx*. Za naglasiti je da podatke treba unijeti u prvi radni list novootvorene Excel radne knjige i da se osim tih podataka u radni list ne unose nikakvi drugi podaci. Važno je naglasiti i da korištene boje u tablici (zelena i siva) nisu bitne, odnosno da se mogu zanemariti.



Slika 42. KNIME preglednik s novim hodogramom i datotekom podaci.xlsx

Nakon toga odabire se prvi čvor koji služi za učitavanje podataka. Čvor se nalazi u repozitoriju čvorova i povlači se u uređivač hodograma. Slika 43 prikazuje ovu aktivnost. Najprije je potrebno u repozitoriju čvorova pronaći odgovarajući čvor, a to je odrađeno na način da se u polje za traženje, koje se nalazi u vrhu okvira repozitorija čvorova, upisuje za pretragu riječ „excel". Na taj se način prikazuju čvorovi koji u imenu sadrže riječ „excel", a među njima je i čvor **Excel Reader** ili Excel čitač. Ovaj čvor služi za učitavanje podataka iz radne knjige programa Microsoft Excel. Čvor se jednostavno povlači na prostor uređivača hodograma kao što je prikazano na slici 42.



Slika 43. Postupak umetanja čvora iz repozitorija čvorova

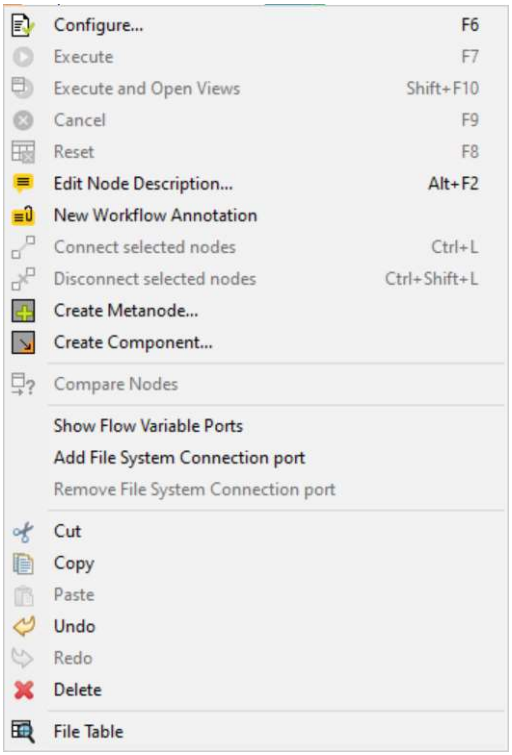
Na isti način se mogu umetnuti daljnji čvorovi, kao što je čvor naziva **Scatter Plot** ili Dijagram raspršenosti**,** desno od postojećeg čvora **Excel Reader**. Slika 44 prikazuje odnos između umetnutih čvorova. Nije važno jesu li su čvorovi poravnani, ali je vizualno prihvatljivije ako se o tome vodi računa.

Slika 44. Izgled uređivača hodograma nakon umetnuta dva čvora

Nakon umetanja, oba je čvora potrebno povezati. To se čini tako da se klikne na crni trokutić s desne strane čvora **Excel Reader** i ne puštajući tipku miša povući od jednog do drugog crnog trokutića koji se nalazi s lijeve strane čvora **Scatter Plot**. Na taj se način povezuju čvorovi, što znači da izlaz čvora **Excel Reader** postaje ulaz čvora **Scatter Plot**. U nastavku priručnika elementi koji služe za povezivanje čvorova nazivat će se *priključcima* (eng. *Port*). Slika 45 prikazuje vezu između dva čvora.

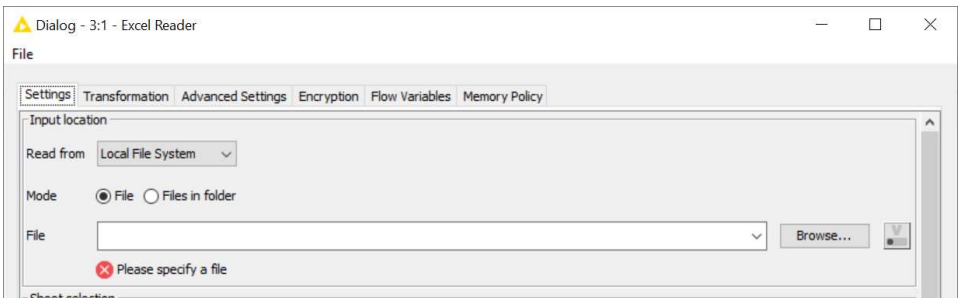
Slika 45. Izgled uređivača hodograma nakon spajanja dvaju čvorova

Povezivanje čvorova je vrlo važna radnja pri izradi hodograma i ukazuje na smjer kojim prolaze tablični podaci kroz različite čvorove. Sljedeći korak je u postavkama čvora **Excel Reader** definirati radnu knjigu iz koje će čvor učitati podatke. Ovdje je najjednostavnije kliknuti desnom tipkom miša na čvor kojem se žele mijenjati postavke. Slika 46 prikazuje kontekstni izbornik koji se prikazuje.



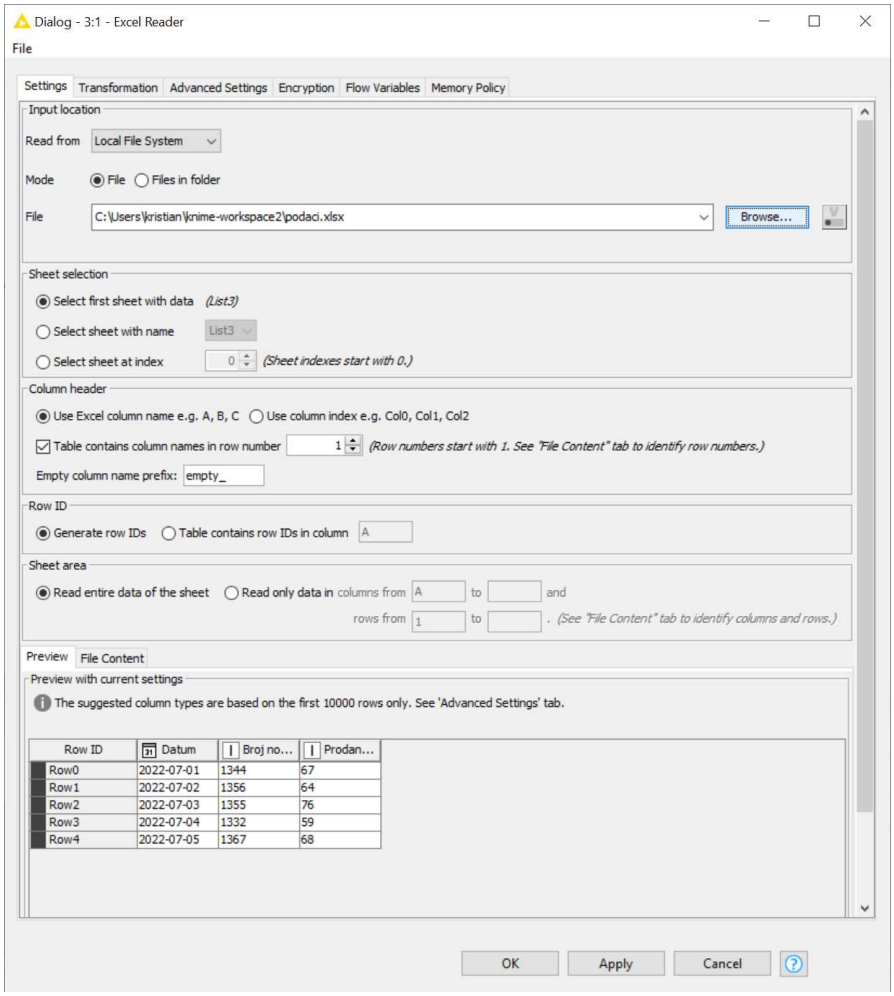
Slika 46. Kontekstni izbornik dobiven desnim klikom miša na čvor „Excel Reader"

Postavke se mijenjaju klikom na prvu ponuđenu aktivnost konfiguracije (eng. *Configure*), ali isti rezultat može se postići i pritiskom na tipku F6 na tipkovnici. Može se uočiti kako su neke aktivnosti nedostupne i samim time sive boje. Kada budu dostupne bit će i objašnjene, a za početak potrebno je obratiti pažnju na donji dio izbornika *Edit* u kojem se može izrezivati (*Cut*), kopirati (*Copy*) i brisati (*Delete*) izabrani čvor. Isto tako mogu se pogledati učitani podaci klikom na aktivnost *„File Table"*. S obzirom da nije definiran radni list iz kojeg se učitavaju podaci, klikom na *„File Table"* neće se dobiti ništa. Ipak je nužno izabrati radnu knjigu pa se odabire *Configure* ili pritisne tipka F6 na tipkovnici. Slika 47 prikazuje gornji dio prve kartice konfiguracijskog dijaloškog okvira čvora **Excel Reader** koja se naziva *Settings*. Ključna aktivnost je očekivano izbor *Microsoft Excel* radne knjige, što se dobiva klikom na gumb *Browse* i traženjem datoteke *podaci.xlsx*. Moguće je kako će se trebati otići jednu razinu prema gore u strukturi mapa i potražiti mapu *knime-workspace* u kojoj je spremljena datoteka *podaci.xlsx*.



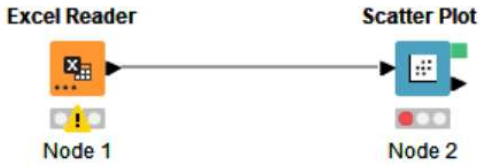
Slika 47. Gornji dio prve kartice konfiguracijskog dijaloškog okvira čvora „Excel Reader"

Slika 48 prikazuje prvu karticu konfiguracijskog dijaloškog okvira čvora **Excel Reader** nakon izbora radne knjige. Potrebno je uočiti kako su crveni krugovi s križićem nestali, a podaci su vidljivi u donjem dijelu dijaloškog okvira.



Slika 48. Prva kartica postavki čvora „Excel Reader" nakon izbora radne knjige

Slika 49 prikazuje oba čvora nakon izbora radne knjige. Važno je uočiti kako je boja na vodoravnom semaforu u donjem dijelu čvora **Excel Reader** izmijenjena i sada je žuta, iako se ne vidi od žutog trokuta upozorenja. Sljedeći korak je izvršavanje čvora, a izvodi se klikom na *Execute* nakon što se desnim klikom na čvor dobije kontekstni izbornik ili pritiskom na tipku F7 nakon što se izabere čvor lijevim klikom.



Slika 49. Izgled uređivača hodograma nakon izbora radne knjige

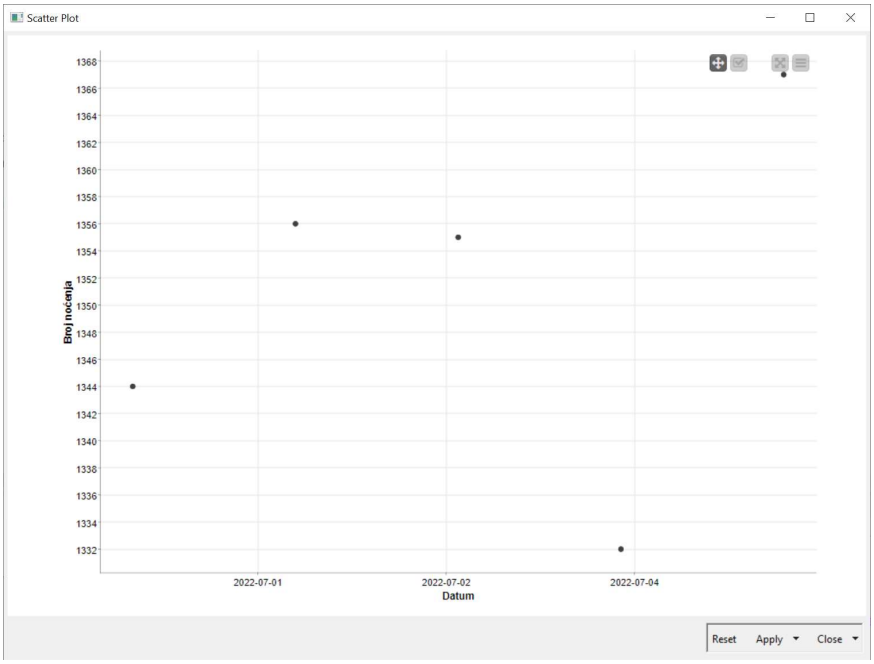
Slika 50 prikazuje oba čvora nakon izvršavanja lijevog čvora. Na vodoravnom semaforu čvora **Scatter Plot** boja je izmijenjena iz crvene u žutu. To je rezultat izvršavanja prethodnog čvora koji je čvoru **Scatter Plot** omogućio pristup podacima iz tablice.

Slika 50. Izgled uređivača hodograma nakon izvršavanja čvora „Excel Reader"

Sljedeći korak je izvršavanje čvora **Scatter Plot** označavanjem istog i pritiskom na tipku F7 na tipkovnici. Moguće je izvršiti čvor i klikom na *Execute* na kontekstnom izborniku koji dobije desnim klikom miša na sam čvor. Slika 51 prikazuje izgled uređivača hodograma nakon izvršavanja obaju čvorova.

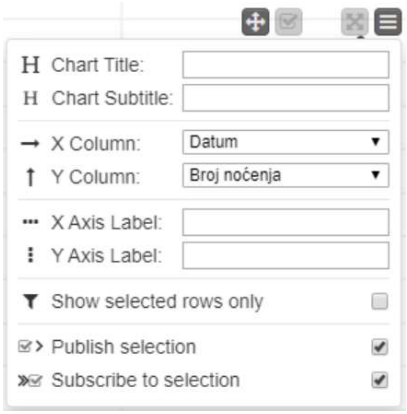
Slika 51. Izgled uređivača hodograma nakon izvršavanja obaju čvorova

Prethodnim aktivnostima učitani su podaci iz radne knjige i dostavljeni čvoru za grafički prikaz **Scatter Plot**. Ovo omogućuje grafički prikaz ovisnost potrošnje osvježavajućih pića o broju noćenja. Da bi se on dobio, izabire se *Interactive View: Scatter Plot* u kontekstnom izborniku čvora **Scatter Plot**. Slika 52 prikazuje dobiveni grafikon. S obzirom da *KNIME* ne može znati koje podatke se želi prikazati na osima, automatski na os X postavlja podatke iz prvog stupca, dok je na os Y postavio podatke iz drugog stupca. Sada grafikon prikazuje broj noćenja u prvih pet dana srpnja 2022. godine, no to nije predmet analize. Da bi se izmijenili prikazani podaci, lijevom tipkom miša potrebno je kliknuti na gumb s tri vodoravne crte u gornjem desnom kutu.



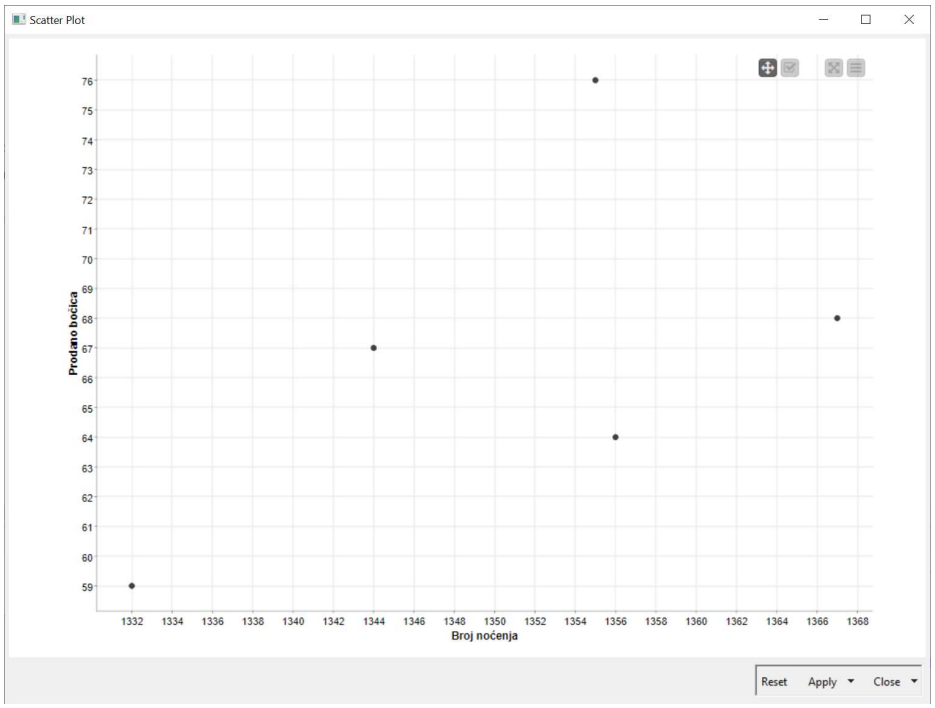
Slika 52. Grafički prikaz ovisnosti prvog i drugog stupca podataka

Slika 53 prikazuje dijaloški okvir za konfiguraciju grafikona u kojem je potrebno izmijeniti stupce koji se prikazuju na osi X i Y. U padajućem izborniku *X Column* izabire se „Broj noćenja“, a u padajućem izborniku *Y Column* izabire se „Prodano bočica“.



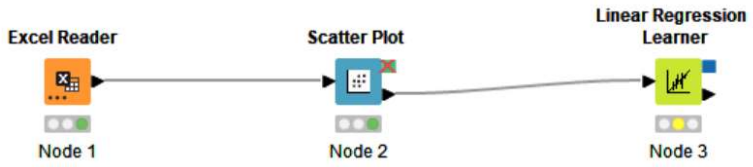
Slika 53. Dijaloški okvir za konfiguraciju grafikona

Slika 54 prikazuje odgovarajući grafikon na kojem se vidi ovisnost broja prodanih bočica i broja noćenja turista.



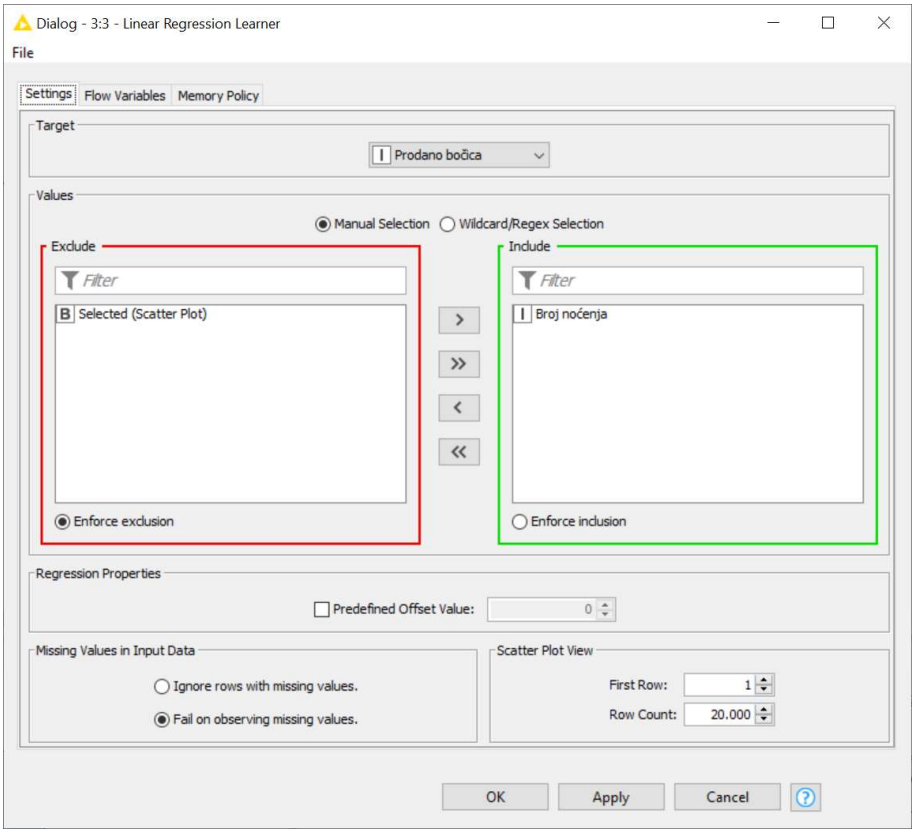
Slika 54. Grafički prikaz ovisnosti prodanih bočica o broju noćenja

Sljedeći korak je treniranje modela linearne regresije s podacima iz učitane radne knjige. Za to će se koristiti čvor **Linear Regression Learner(** što bi se na hrvatski jezik pomalo nespretno prevelo s Učenik linearne regresije) koji se može naći u repozitoriju čvorova, nakon čega se čvor povlači u uređivač hodograma i spaja s prethodnim čvorom. Slika 55 prikazuje izgled uređivača hodograma na kojem se nalaze sva tri do sada postavljena čvora. Slijedi izmjena postavki čvora **Linear Regression Learner**.



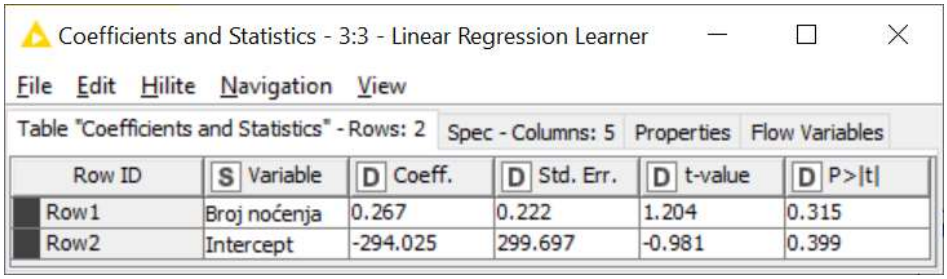
Slika 55. Izgled hodograma sa sva tri čvora

Slika 56 prikazuje kako treba konfigurirati čvor **Linear Regression Learner**. U padajućem izborniku *Target* treba biti izabran stupac, odnosno ciljna varijabla „Prodano bočica“. U zelenom pravokutniku treba biti izabran stupac, odnosno ulazna varijabla „Broj noćenja“. Čvor je nakon ovih promjena potrebno pokrenuti tj. izvršiti.



Slika 56. Postavke čvora Linear Regression Learner

Nakon izmjena postavki i izvršavanja čvora, potrebno je izabrati *Coefficients and Statistics* iz kontekstnog izbornika čvora *Linear Regression Learner* kao na slici 57.



Slika 57. Dijaloški okvir Coefficients and Statistics

U ovoj fazi završeno je treniranje modela linearne regresije i dobiveni su koeficijenti 0,267 i -294,025. Ove vrijednosti se mogu promatrati kao koeficijenti jednadžbe pravca. Eksplicitni oblik jednadžbe pravca glasi (Dakić & Elezović, 2019):

$$y = ax + b$$

Koeficijent ***a*** se naziva nagib ili koeficijent smjera pravca, a koeficijent ***b*** odsječak na Y osi pravca. Umjesto varijable *x* uvrštava se vrijednost ulazne varijable koja je poznata i iz jednadžbe se izračunava ciljna vrijednost *y* (Dakić & Elezović, 2019). Ako se uvrste dobiveni koeficijenti, dobiva se:

$$y = 0{,}267\ x - 294{,}025$$

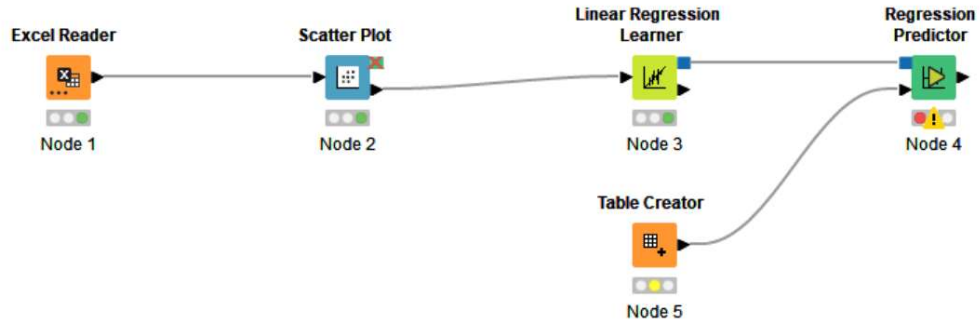
Ovo je model linearne regresije sa samo jednom ulaznom i jednom izlaznom varijablom. Na osnovu ovog modela može se izračunati potrošnja osvježavajućih pića u ovisnosti o broju noćenja, na način da se uvrsti broj noćenja umjesto varijable x i izračuna se y. Na početku ovog primjera definirano je da se želi izračunati potrošnja osvježavajućih pića ako je u mjestu zabilježeno 1000 noćenja. Uz pomoć ručnog kalkulatora to se može jednostavno izračunati, a očekivana potrošnja je oko -27 bočica. Rezultat je pomalo zbunjujući i model ukazuje kako bi kod navedenog broja noćenja turista svakog dana bilo u samoposlužnom aparatu dodatnih 27 bočica osvježavajućih pića, odnosno da bi turisti punili samoposlužni aparat umjesto da ga prazne! Problem je u premalom broju podataka koji su se koristili za treniranje modela pri čemu je dobiven nepouzdan model. Na samom početku priručnika je opisano kako se taj problem rješava na klasičan način koristeći iskustvo stručnjaka pri čemu je procijenjen koeficijent K iznosio 0,05, a radi se o procijenjenom nagibu pravca linearne funkcije – koeficijentu **a**. Model generiran iz samo pet redova podataka očigledno je puno lošiji od procjene, ali kao što je navedeno da bi se dobio pouzdan model nužno je imati što veću količinu podataka za treniranje.

U nastavku preostaje završiti hodogram na način da se ubace čvorovi koji će omogućiti korištenje modela i izračun potrošnje na osnovu broja noćenja. Za to su potrebna dva čvora: **Regression Predictor i Table Creator**. Slika 58 prikazuje hodogram koji uključuje čvor **Regression Predictor**.

Slika 58. Hodogram koji uključuje i Regression Predictor

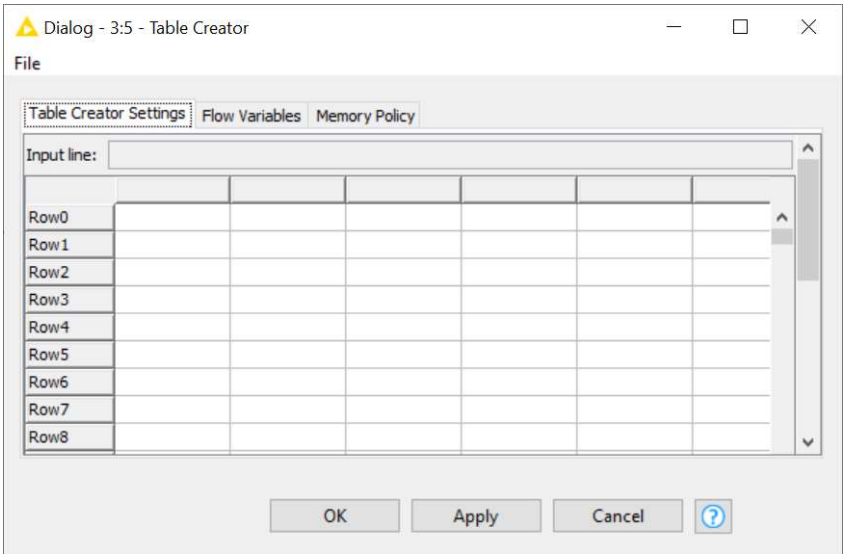
Kako mu samo ime kaže, čvor **Regression Predictor** ili Prediktor regresije, predviđa vrijednosti koristeći regresijski model. Ono što treba naglasiti jest da taj čvor treba regresijski model, a on je istreniran od

strane prethodnog čvora **Linear Regression Learner**. Model se prebacuje između dva čvora vezom koja povezuje dva plava pravokutnika pri čemu je uvijek jedan čvor onaj u kojem se trenira model dok je drugi čvor onaj koji model primjenjuje. Može biti i više čvorova koji koriste model, ali samo jedan ga može generirati. Te veze treba razlikovati od podatkovnih veza između čvorova koji su spojeni između crnih trokutića. Konačno, potrebno je čvoru **Regression Predictor** dostaviti i podatke, a to je najjednostavnije odraditi korištenjem čvora **Table Creator** ili Tvorac tablice. Na slici 59 prikazan je kompletan hodogram.



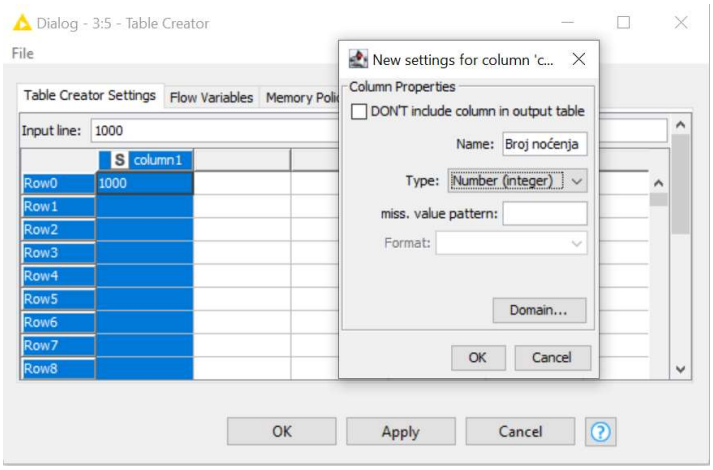
Slika 59. Kompletan hodogram „Zdravo svijete!"

Čvor **Table Creator** služi, kako mu samo ime kaže, za kreiranje tablice. Ovdje je potrebna tablica sa samo jednom ćelijom u kojoj je vrijednost 1000. Slika 60 prikazuje konfiguraciju čvora **Table Creator**.



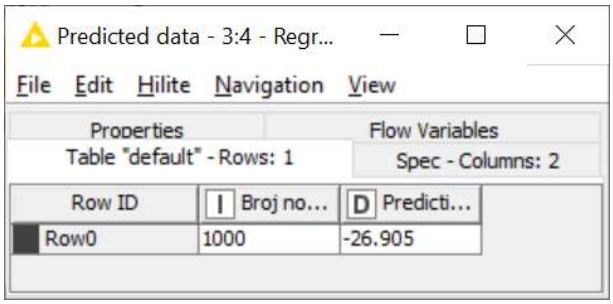
Slika 60. Konfiguracija čvora Table Creator

U prvu ćeliju unosi se vrijednost 1000, ali potrebno je duplim klikom miša na naziv stupca *column1* otvoriti svojstva te ulazne varijable. Pri tom se otvara mali dijaloški okvir u kojem treba promijeniti ime ulazne varijable, odnosno stupca u „Broj noćenja". Osim toga, potrebno je promijeniti tip podatka u *Number (integer)*. Radi se o brojčanom podatku. Nakon toga potrebno je potvrditi promjene klikom na gumb OK i izvršiti čvor. Slika 61 prikazuje potrebne postavke čvora **Table Creator**.



Slika 61. Potrebne postavke čvora Table Creator

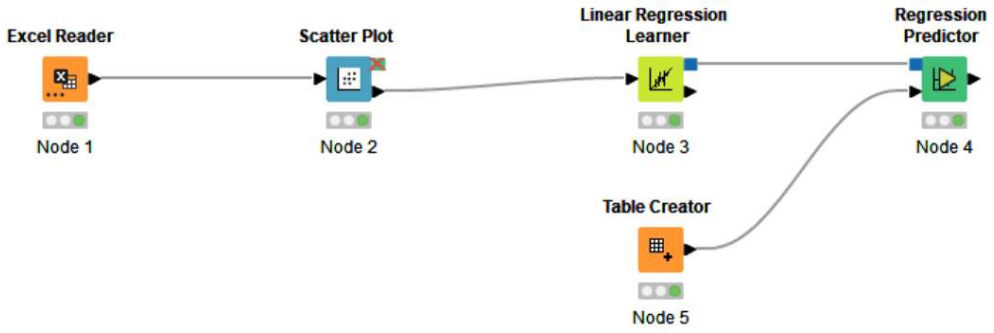
Konačno, potrebno je izvršiti i čvor **Regression Predictor** i iz kontekstnog izbornika izabrati posljednju aktivnost *Predicted Data* čime se otvara dijaloški okvir s predviđenom potrošnjom bezalkoholnih pića na samoposlužnom aparatu. Rezultat je isti kao i kod izračuna s ručnim kalkulatorom, oko 27 boca bezalkoholnog pića. To potvrđuje očekivanje da je čvor **Regression Predictor** koristio koeficijente koji su očitani iz tablice na slici 56 te je na osnovu njih izračunat očekivan broj prodanih boca pića. Slika 62 prikazuje konačan rezultat.



Slika 62. Predviđena potrošnja pića na samoposlužnom aparatu za 1000 noćenja

3.3.3. Spremanje hodograma

Nakon što je hodogram završen, model istreniran i željeni rezultat dobiven, preostaje spremanje hodograma. Očekivano, potrebno je iz izbornika *File* izabrati *Save*. Time se završava izrada jednostavnog hodograma, a uz pažljivo praćenje izrade hodograma moglo se usvojiti osnove funkcioniranja programa KNIME. Slika 63 prikazuje kompletan hodogram s izvršenim svim čvorovima.

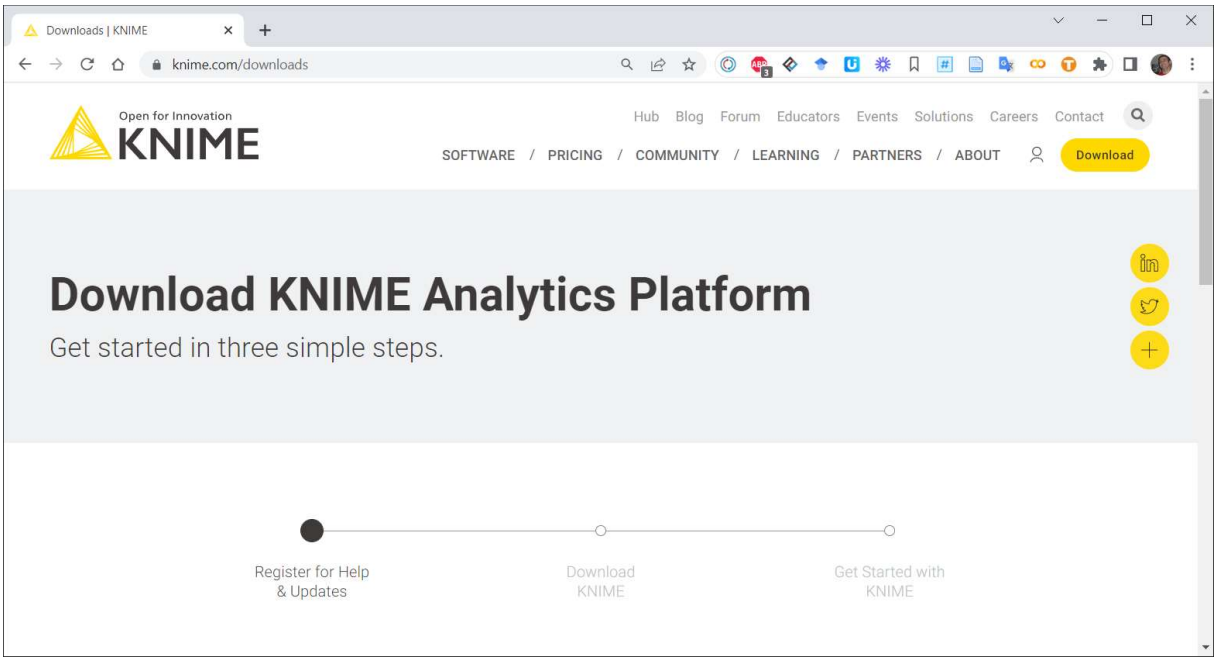


Slika 63. Kompletan hodogram s izvršenim svim čvorovima

3.4. Instalacija programa KNIME verzije 4.x i 5.x

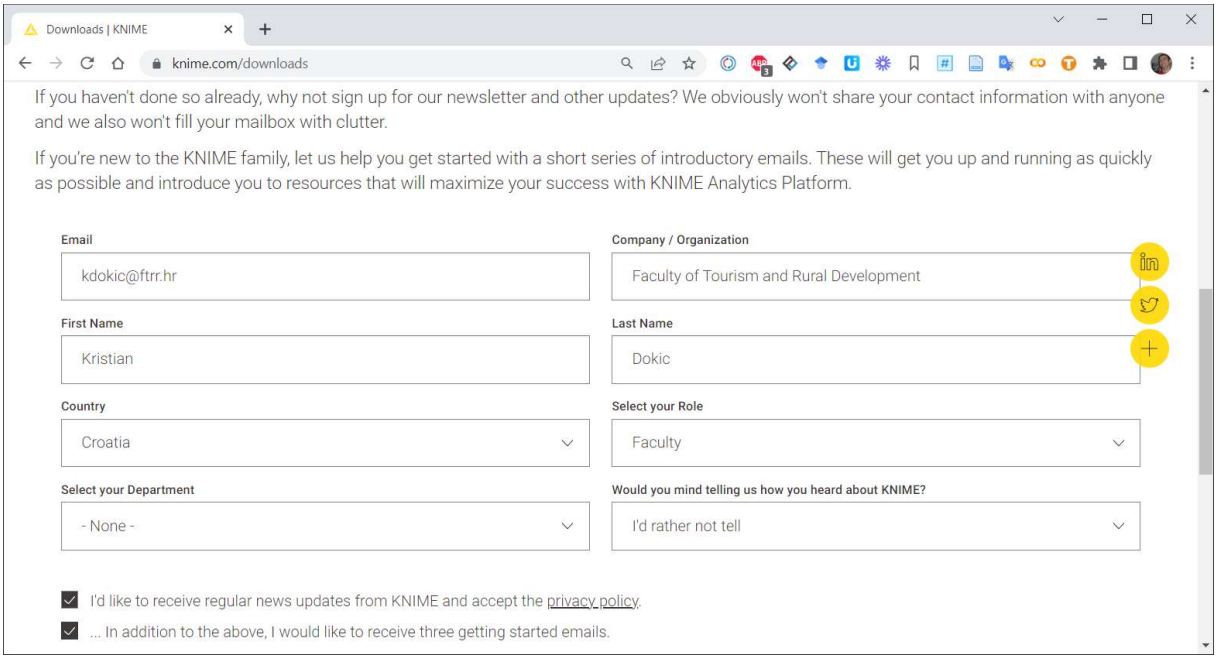
U ovom dijelu priručnika opisano je kako se instalira program KNIME verzija 4.x i 5.x

KNIME se može preuzeti s adrese https://www.knime.com/downloads pri čemu se treba registrirati sa svojim podacima te prihvatiti politiku privatnosti. Dostupne su verzije za operativne sustave Windows, Linux i Mac, a veličine instalacijskih datoteka su po nekoliko stotina megabajta. Prije instalacije potrebno je pogledati minimalne specifikacije za računalo na koje ga se planira instalirati. Slika 64 prikazuje web stranicu od koje kreće proces preuzimanja.



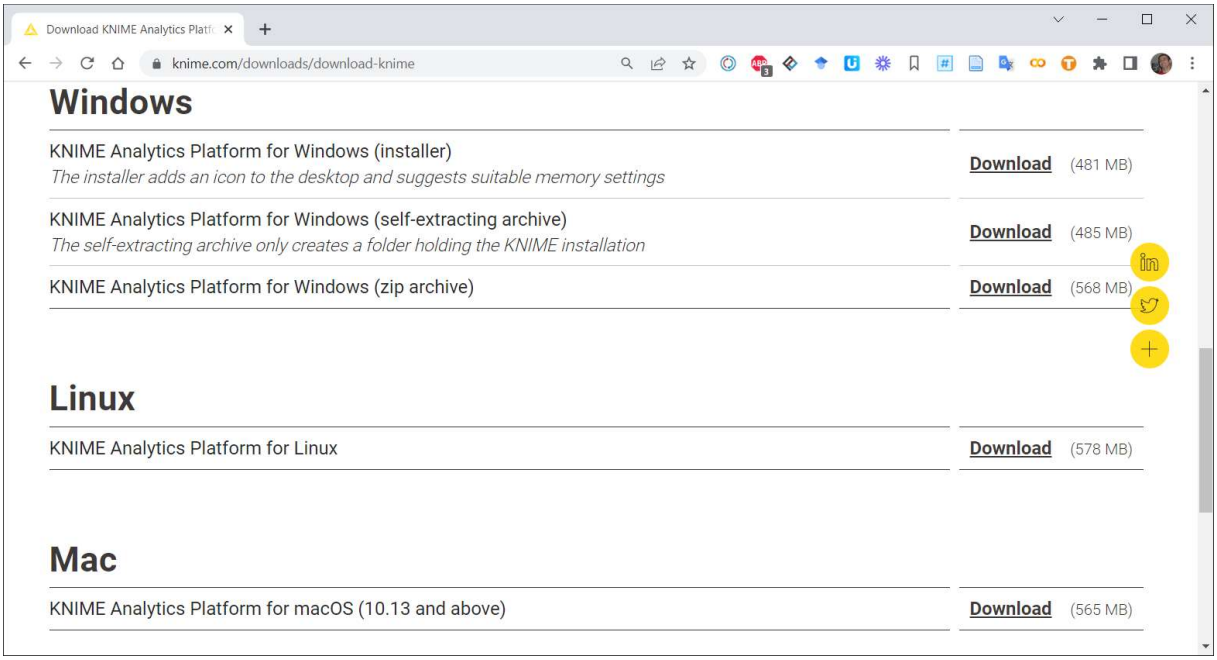
Slika 64. Web stranica na kojoj kreće proces preuzimanja programa KNIME

U donjem dijelu stranice su polja u koja se unose podaci korisnika. Na Slika 65 prikazuje se taj dio stranice.



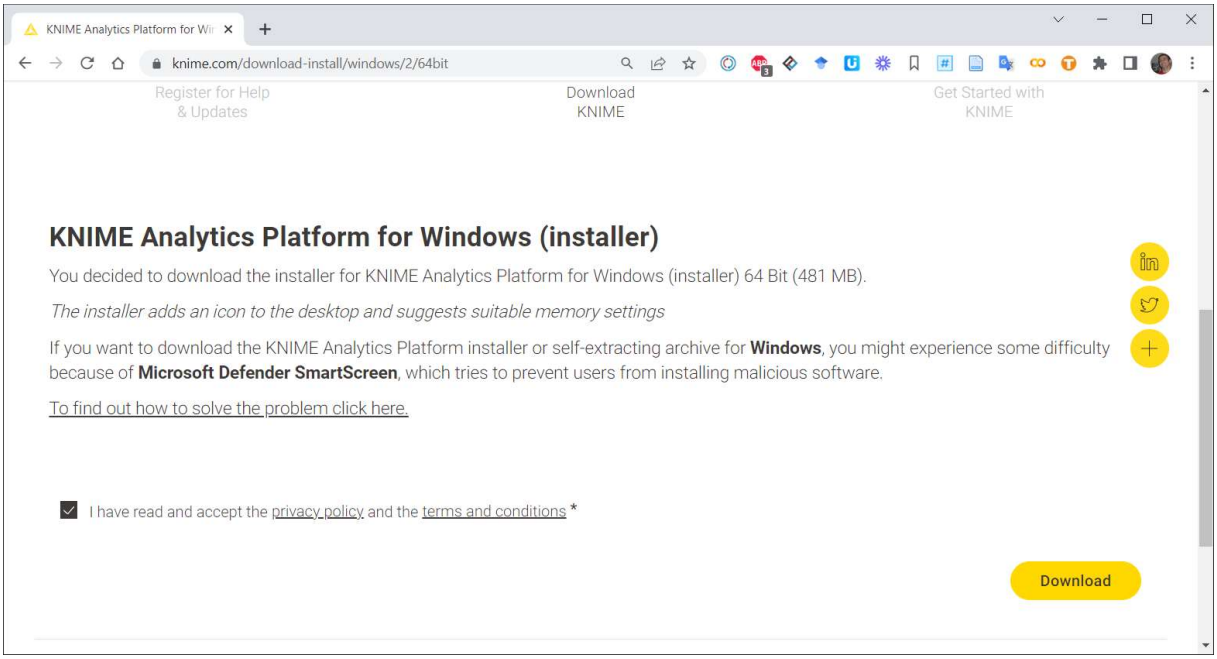
Slika 65. Donji dio web stranice u kojoj unosite podatke

Nakon unosa podataka i potvrde istih potrebno je izabrati operativni sustav na koji se planira instalirati KNIME Analytics Platform. Slika 66 prikazuje stranicu.



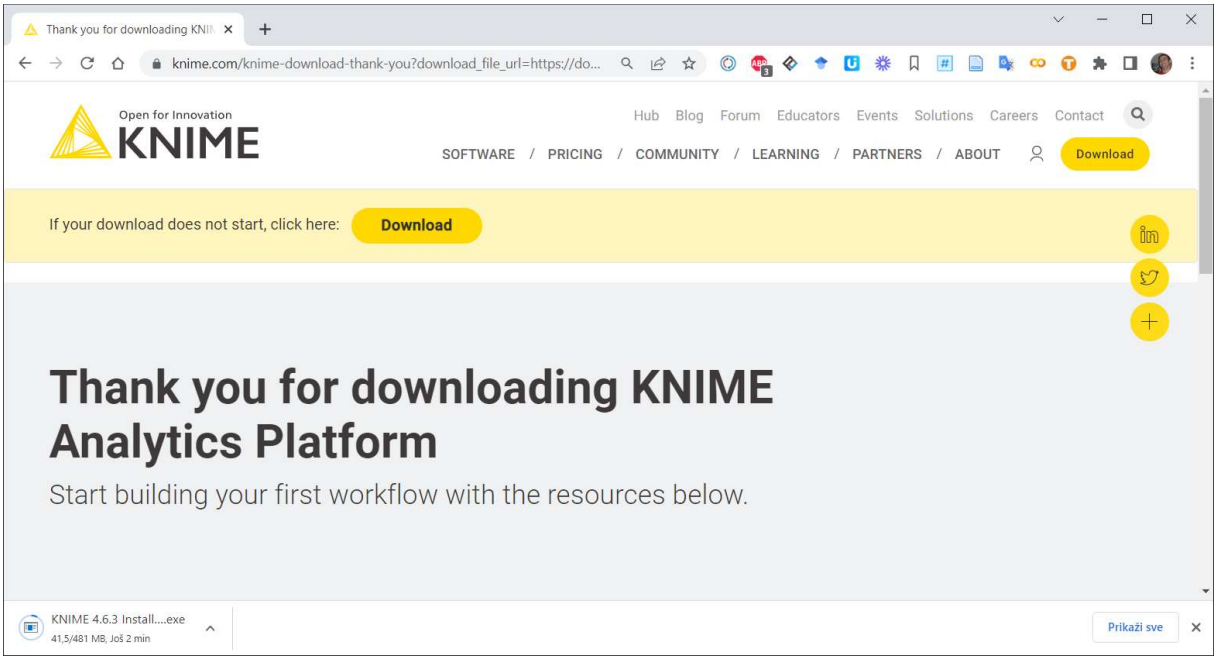
Slika 66. Web stranica na kojoj se bira operativni sustav svog računala

S obzirom da je na većini osobnih računala u Hrvatskoj instalirana neka verzija operativnog sustava *Microsoft Windows*, vjerojatno će biti izabrana ta verzija. Potrebno je složiti se s uvjetima korištenja pri čemu treba „kliknuti" na kvadratić ispred teksta „*I have read…*". Slika 67 prikazuje web stranicu.



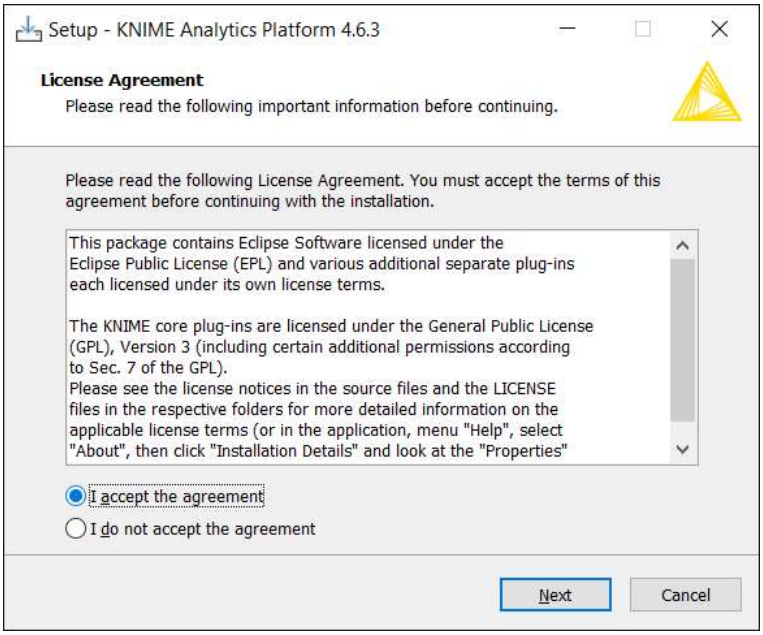
Slika 67. Web stranica na kojoj se pokreće preuzimanje instalacijske datoteke

Nakon pokretanja procesa preuzimanja, dobiva se poruka zahvale i preuzimanje traje određeno vrijeme ovisno o brzini veze na internet. Slika 68 prikazuje web stranicu.



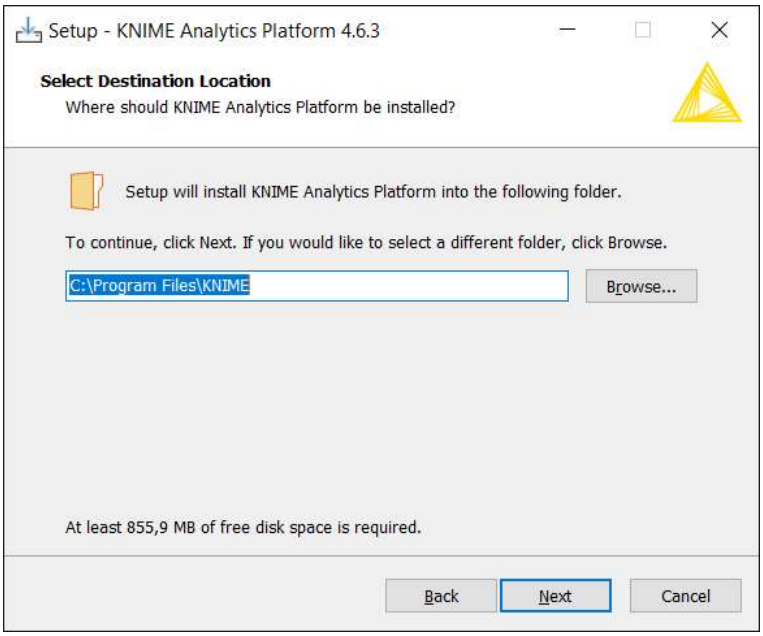
Slika 68. Web stranica sa zahvalom

Nakon preuzimanja instalacijske datoteke, potrebno je pokrenuti instalaciju duplim klikom na samu ikonu datoteke u mapi Preuzimanja (*Downloads*). Time se pokreće instalaciju i otvara se dijaloški prozor u kojem treba potvrditi slaganje s licencom pod kojom je program izdan. Slika 69 prikazuje dijaloški prozor.



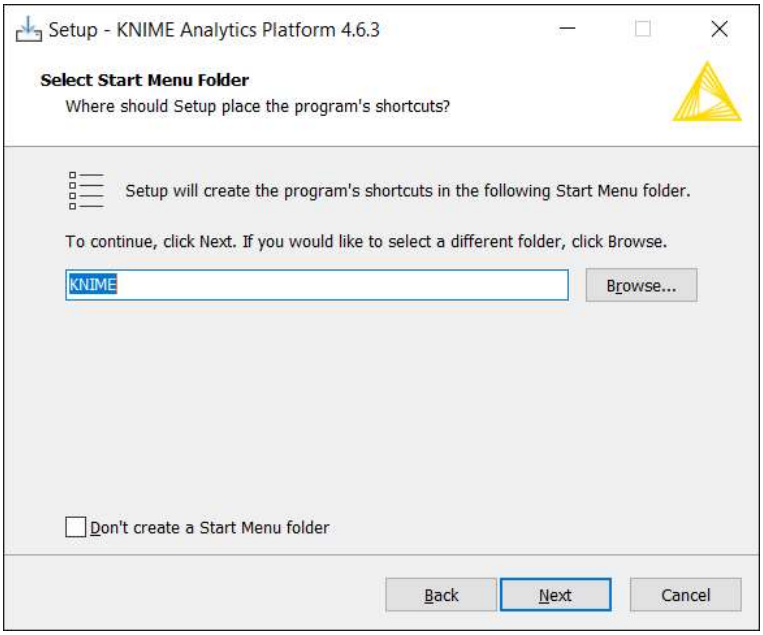
Slika 69. Dijaloški prozor u kojem se korisnik slaže s uvjetima korištenja

Nakon klika na dugme *Next* potrebno je definirati putanju instalacije programa. Preporuka je ostaviti predloženo. Slika 70 prikazuje dijaloški prozor.



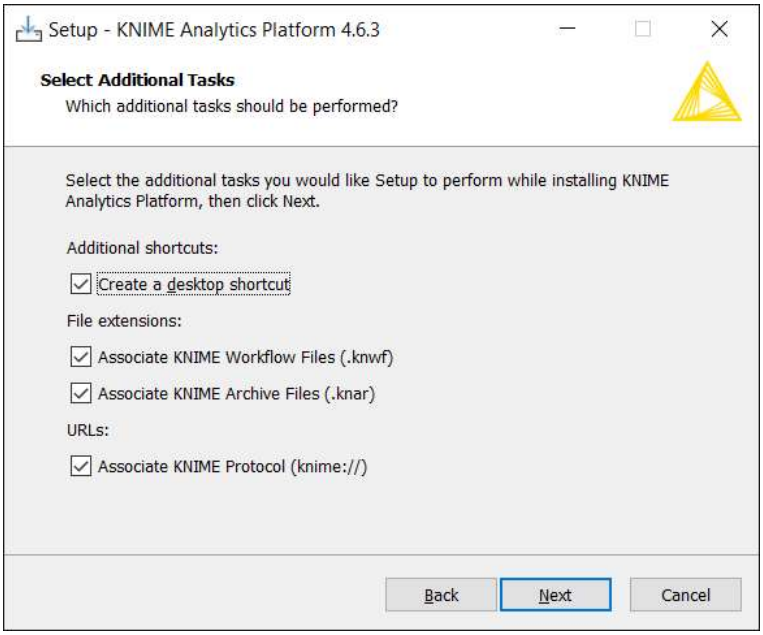
Slika 70. Dijaloški prozor u kojem se definira putanju instalacije programa

Nakon klika na dugme *Next* potrebno je definirati naziv foldera koji će biti umetnut među ostale foldere za pokretanje programa korištenjem gumba *Start* u donjem lijevom kutu ekrana, na operativnim sustavima *Microsoft Windows*. Taj dijaloški okvir prikazuje slika 71.



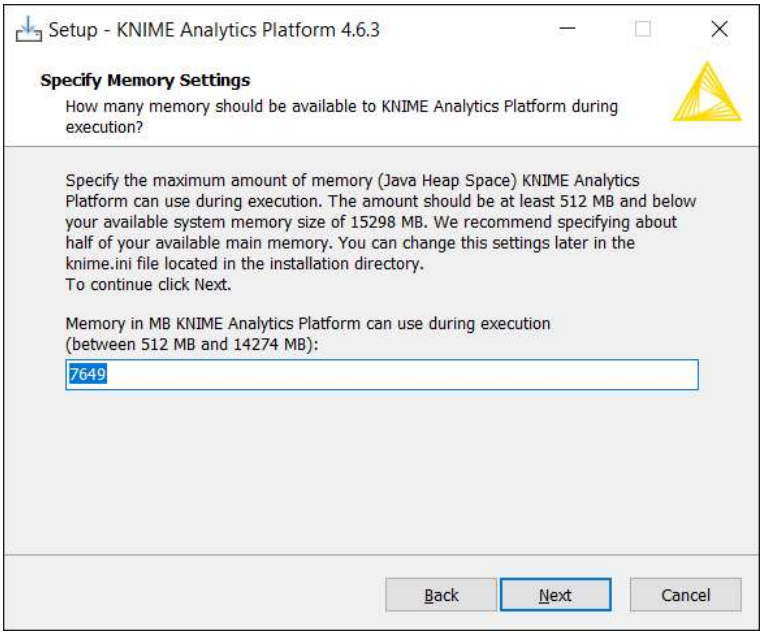
Slika 71. Dijaloški okvir kojim se bira naziv foldera

Sljedeći dijaloški okvir nudi mogućnost postavljanja prečice na radnu površinu i pokretanje programa *KNIME* dvostrukim klikom na datoteku s nastavcima *.knfw* i *.knar*. Osim toga, može se povezati protokol „knime" s programom KNIME. Preporuka je ostaviti sve uključeno. Slika 72 prikazuje dijaloški okvir.



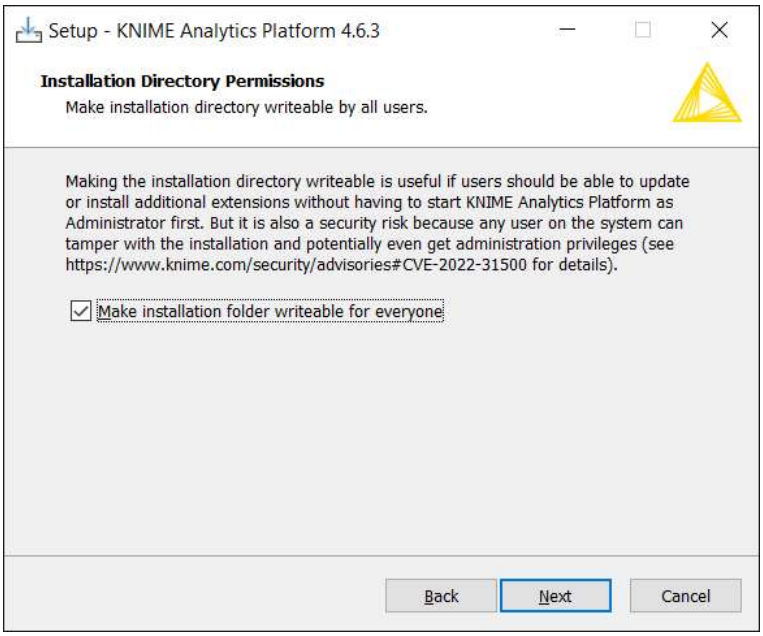
Slika 72. Dijaloški okvir s izborom više mogućnosti

U sljedećem dijaloškom okviru bira se maksimalna količina radne memorije koja je dostupna programu KNIME. Preporuka je ostaviti pola radne memorije za potrebe programa, a tako je i u zadanim postavkama definirano. Slika 73 prikazuje taj dijaloški okvir.



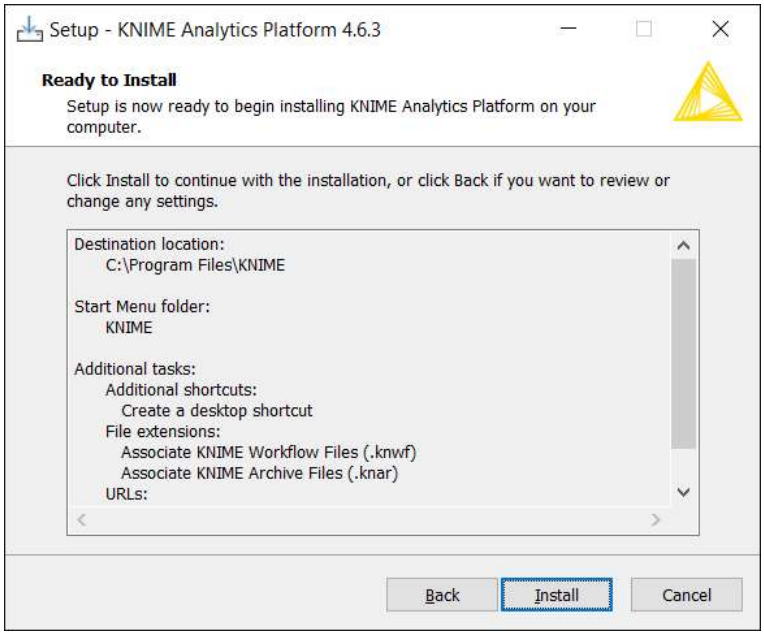
Slika 73. Dijaloški okvir za izbor količine programu dostupne radne memorije

Sljedeći dijaloški okvir omogućuje da se izabere hoće li će u instalacijski folder svi korisnici računala moći spremati datoteke. Preporuka je ostaviti postavljeno. Slika 74 prikazuje dijaloški okvir.



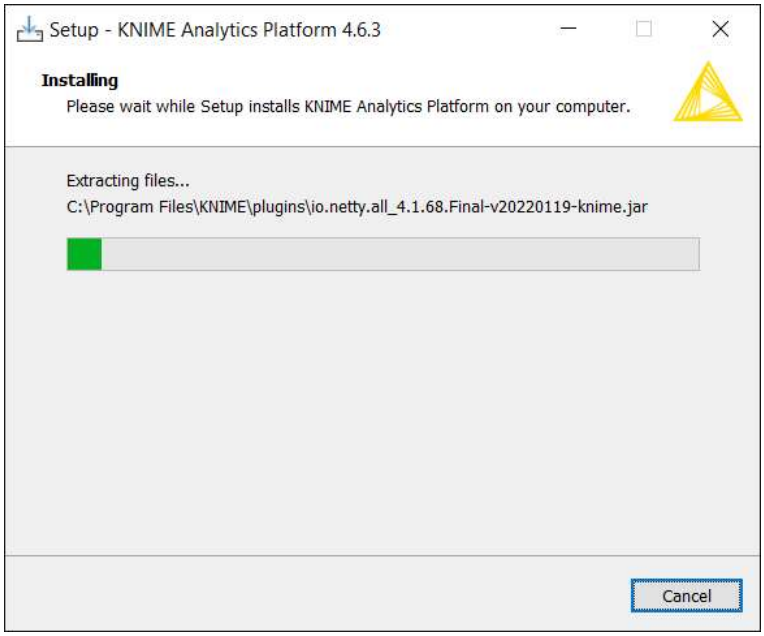
Slika 74. Dijaloški okvir koji omogućuje svima zapisivanje u instalacijski folder

Na sljedećem dijaloškom okviru vidi se rezime do sada izabranog. Slika 75 prikazuje dijaloški okvir.



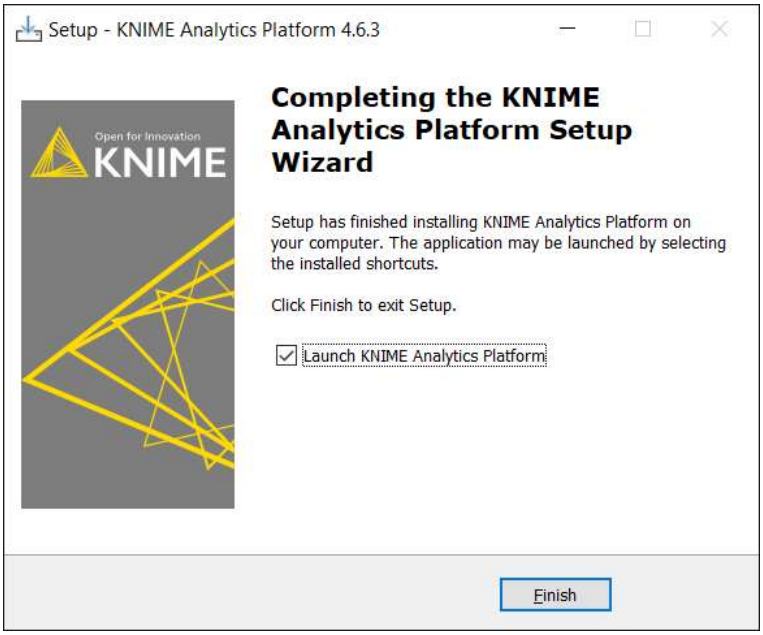
Slika 75. Dijaloški okvir u kojem je rezimirano sve do sada izabrano

Nakon klika na gumb *Install* pokreće se instalacija. Slika 76 prikazuje dijaloški okvir.



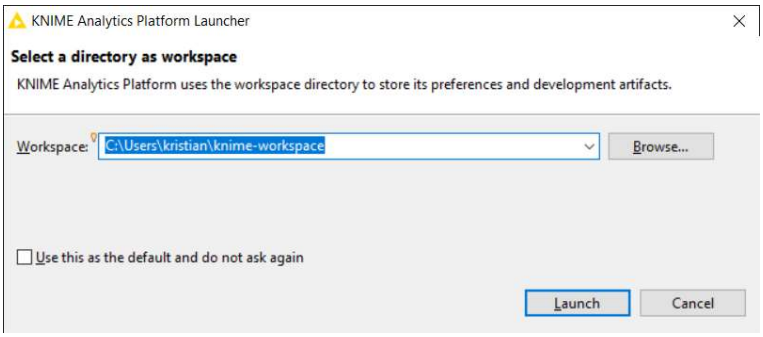
Slika 76. Dijaloški okvir u kojem je prikazan tijek instalacije

Nakon nekoliko desetaka sekundi ili nekoliko minuta instalacija se završava. Slika 77 prikazuje dijaloški okvir, u njemu je zadano pokretanje programa *KNIME* nakon što se klikne na gumb *Finish*.



Slika 77. Završni dijaloški okvir

Prilikom pokretanja *KNIME* traži da se izabere putanju do radnog okruženja. Preporuka je ostaviti navedeno. Klikom na dugme *Launch* konačno se pokreće program.

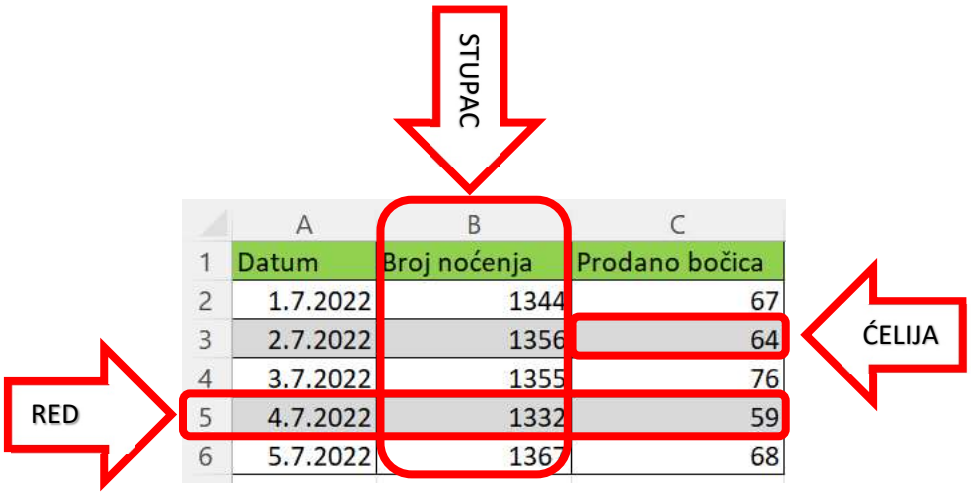


Slika 78. Dijaloški okvir za izbor radnog okruženja

4. Učitavanje i priprema podataka

4.1. Učitavanje podataka

Podaci koji se koriste za analizu i izradu modela strojnog učenja najčešće su pohranjeni na računalu ili dostupni na internetu. U pravilu se radi o tabličnim podacima, sastavljenih od stupaca, redova i ćelija. Slika 79 prikazuje terminologiju za datoteku *podaci.xlsx* *Microsoft Excel* ili *LibreOffice Calc*.



Slika 79. Terminologija vezana za tablice poznata iz tabličnih kalkulatora

U stupcu se nalaze istovrsni podaci, osim zaglavlja koje se u pravilu nalazi u prvom redu tablice. Slika 79 prikazuje brojčane i datumske podatke u stupcima, dok su zaglavlja tekstualna. U redovima se nalaze podaci pojedinih slučajeva, s različitim svojstvima navedenima u zaglavlju stupaca. Slika 79 prikazuje „Broj noćenja“ i broj prodanih bočica u „Prodano bočica“ osvježavajućih pića na dan 4. srpnja 2022. U ćeliji se nalaze pojedinačne vrijednosti, odnosno podaci, a svojstvo ćelije je da ima jedinstvenu adresu u tablici.

Osim u radnim knjigama tabličnih kalkulatora (*Microsoft Excel* ili *LibreOffice Calc*), podaci se često spremaju i u CSV formatu. Kratica CSV dolazi od engleskih riječi *Comma Separated Values*, a doslovan prijevod bi bio „vrijednosti odvojene zarezom“. Ovakve datoteke mogu se otvoriti i najjednostavnijim uređivačima kao što je Blok za pisanje (eng. *Notepad*) pod operativnim sustavom *Windows*. Bez obzira na drugačiji format datoteke, podaci su i dalje spremljeni u stupcima i recima. Slika 80 prikazuje sadržaj CSV datoteke koji je isti sadržaju datoteke tabličnog kalkulatora koji prikazuje Slika 79. Vidljivo je zaglavlje u prvom retku, a podaci su razdvojeni zarezom umjesto granicama pojedinih stupaca.

```
Datum,Broj noćenja,Prodano bočica
1.7.2022,1344,67
2.7.2022,1356,64
3.7.2022,1355,76
4.7.2022,1332,59
5.7.2022,1367,68
```

Slika 80. Sadržaj CSV datoteke

Kao što je prethodno navedeno kod nadziranog strojnog učenja varijable u stupcima koje su poznate nazivaju se ulaznim varijablama ili značajkama (eng. *Features*), a varijabla čiju vrijednost se želi

predvidjeti modelom, naziva se izlazna varijabla ili ciljna vrijednost (eng. *Target*). U nastavku priručnika koristit će se za ulazne varijable naziv značajka, a za izlazne varijable naziv ciljna vrijednost.

Podaci mogu biti različitog tipa pa tako razlikujemo:

a) Decimalni broj dvostruke točnosti (eng. *Double*)
b) Cjelobrojna vrijednost (eng. *Integer*)
c) Polja znakova (eng. *String*)
d) Datum i vrijeme (eng. *Date & Time*)
e) Nepoznato (eng. *Unknown*)
f) Ostali tipovi (eng. *Other types*) (Silipo, 2011).

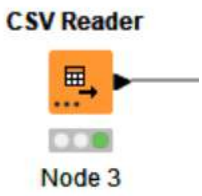
Da bi učitali podatke u *KNIME*, potrebno je umetnuti početni čvor (eng. *Node*) u hodogram (eng. *Workflow*) koji služi za učitavanje podataka. Na raspolaganju je cijeli niz formata za koje su dostupni različiti čvorovi, u koje se podaci mogu učitati. Neki od njih su:

a) **Excel Reader** (Excel čitač)
b) **File Reader** (Čitač datoteka)
c) **ARFF Reader** (ARFF čitač)
d) **CSV Reader** (CSV čitač)
e) **Line Reader** (Čitač redka)
f) **Table Reader** (Čitač tablice)
g) **Fixed Width File Reader** (Čitač datoteka fiksne širine)
h) **Read Excel Sheet Names** (Čitač Excel naziva radnih listova)
i) **Read Images** (Čitač slika).

Navedeni popis uključuje samo dio čvorova, jer osim toga postoje čvorovi za učitavanje/spajanje na baze podataka, REST (eng. *representational state transfer*) servise ili servise za prijenos reprezentacijskog stanja, *big data* poslužitelje, audio datoteke, slikovne datoteke, web logove itd. Ako se ne pronađe odgovarajući čvor za učitavanje podataka, potrebno ga je potražiti u repozitoriju *KNIME Hub*.

U nastavku su prikazani neki od češće korištenih čvorova za učitavanje i pripremu podataka. Aktivnosti pripreme podataka u programu *KNIME* slične su kao i u nekome od tabličnih kalkulatora.

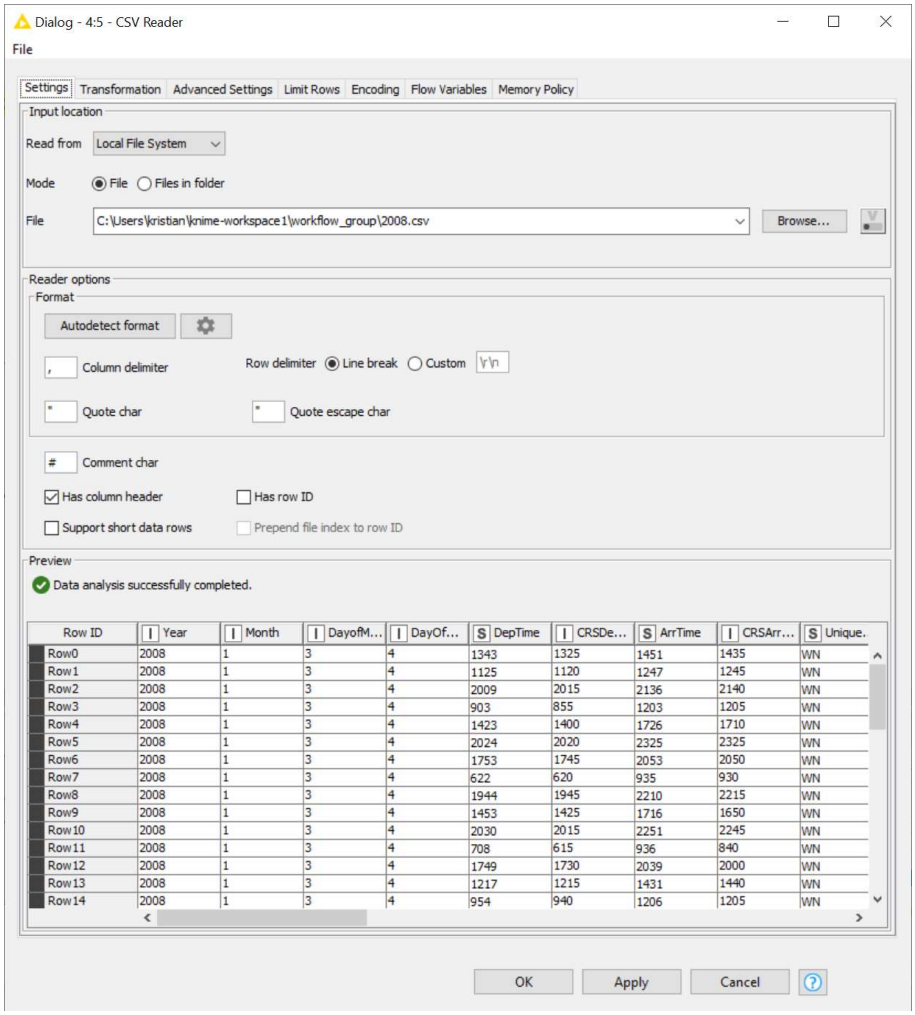
Jedan od najkorištenijih čvorova je svakako **CSV Reader** koji omogućuje učitavanje podataka spremljenih u datoteci pri čemu su podaci odvojeni zarezom (eng. *Coma Separated Values*). Slika 81 prikazuje izgled čvora.



Slika 81. Čvor CSV Reader

Navedeni čvor postavljen na hodogram potrebno je konfigurirati, a sama konfiguracija uključuje nekoliko dijaloških okvira. Kako bi se došlo do konfiguracije, potrebno je desnim klikom miša kliknuti na čvor te iz kontekstnog izbornika izabrati *Configure*. Na prvoj kartici naziva *Settings* podešava se

putanja do CSV datoteke gdje se nalaze podaci. Klikom na gumb *Browse* iza okvira *File* izabire se putanja i sama datoteka. Slika 82 prikazuje dijaloški okvir i na njemu odabranu karticu *Settings*.



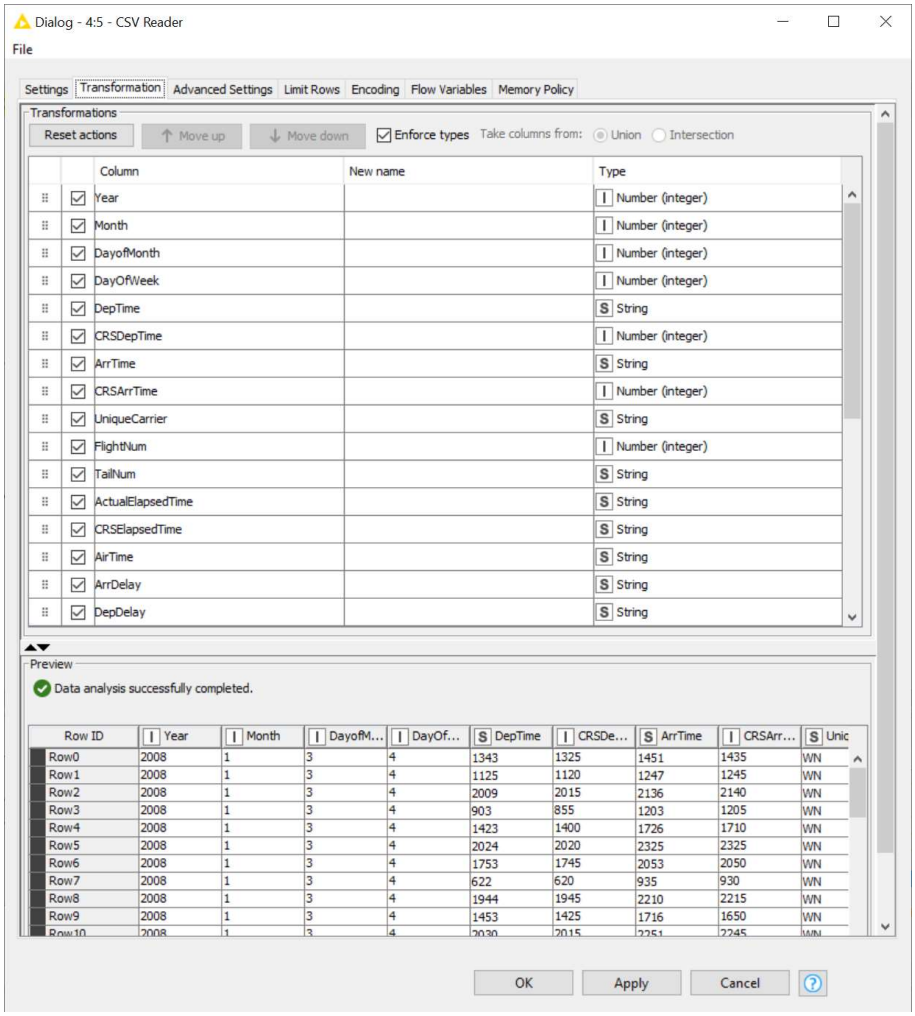
Slika 82. Prva kartica konfiguracije čvora CSV Reader

U primjeru se koristi skup podataka *Data Expo 2009: Airline On Time Data* koji je dostupan na adresi: https://www.kaggle.com/datasets/wenxingdi/data-expo-2009-airline-on-time-data?select=2008.csv Skup podataka uključuje podatke o svim dolaznim i odlaznim letovima zračnih prijevoznika u Sjedinjenim Američkim Državama od listopada 1987. do travnja 2008. godine. U primjeru se koriste samo podaci iz 2008. godine. Za preuzimanje je potrebna prijava na poslužitelj koju je moguće odraditi korištenjem Google računa.

Osim naziva i putanje do CSV datoteke u dijaloškom okviru *Settings* bitno je da znak koji razdvaja stupce bude odgovarajući, to je najčešće zarez. U nekim slučajevima za razdvajanje stupaca koristi se točka-zarez, ali to je odmah vidljivo na pretpregledu podataka u donjem dijelu dijaloškog okvira.

Čest je slučaj da prvi red CSV datoteke čini zaglavlje s nazivima varijabli, a označavanjem polja *Has column header* ova opcija je uključena. Osim toga, neke CSV datoteke imaju i komentare u kojima se nalaze opisi, ali ne i podaci. Za komentare se koristi znak # te ga je potrebno unijeti u polje *Comment char*.

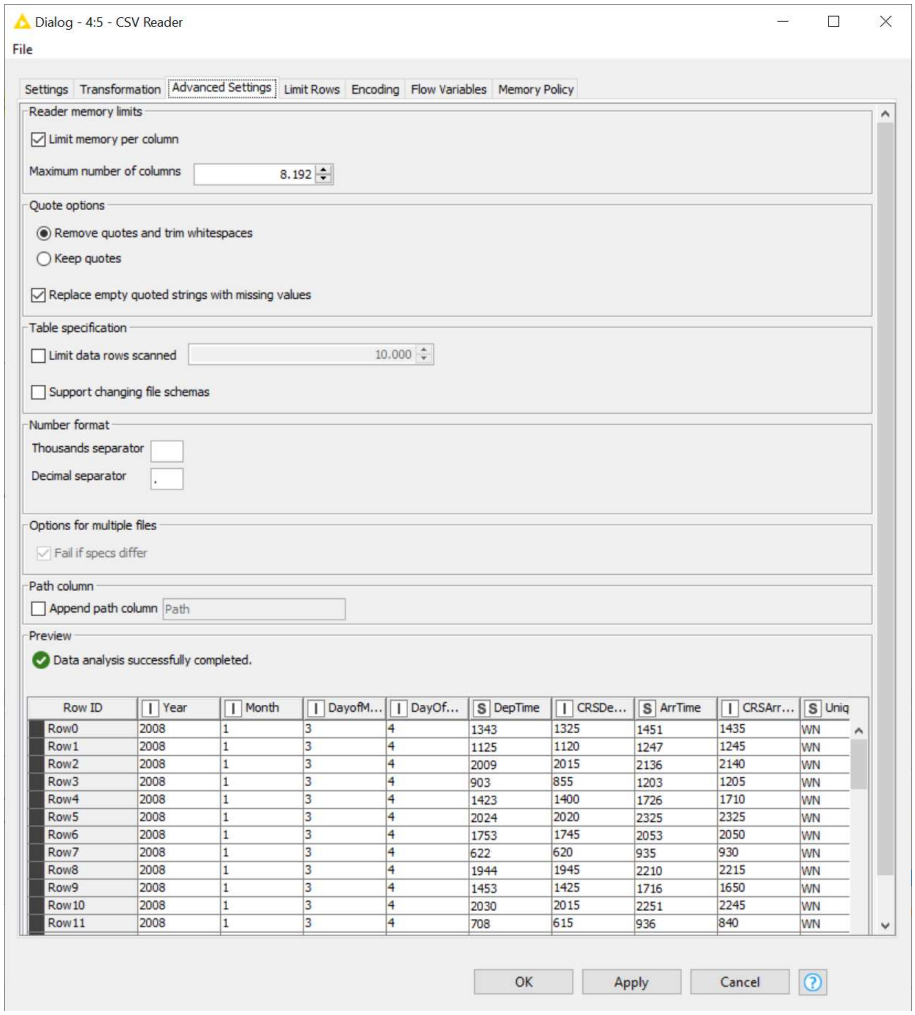
Na sljedećoj kartici u dijaloškom okviru naslova *Transformation* potrebno je definirati u koju vrstu podatka će se svaka pojedina vrijednost konvertirati. Slika 83 prikazuje primjer za konverziju skupa podataka *Data Expo 2009: Airline On Time Data*.



Slika 83. Druga kartica konfiguracije čvora CSV Reader

Pri toj konverziji u stupac *New name* moguće je unijeti novi naziv stupca i tako ga preimenovati, a u stupcu *Type* moguće je unijeti vrstu podataka u koju će podaci iz određenog stupca biti konvertirani. Pri tome je dostupno više različitih formata, a KNIME u pravilu nudi formate u koje može konvertirati postojeće podatke. Nekada će to biti različiti oblici brojčanog zapisa (*cjelobrojni, realni*), a ponekad tekstualnog (*TXT, XML zapis,…*). Izmjenu vrste podatka moguće je izvršiti tako da se desnom tipkom miša klikne na željeno polje koje se namjerava izmijeniti. Gumb *Reset actions* omogućuje vraćanje na početne postavke koje su predložene od samog programa KNIME Analytics Platform nakon analize podataka u datoteci.

U određenim situacijama, ako *KNIME Analytics Platform* javlja grešku pri učitavanju podataka potrebno je izmijeniti zadane postavke na dijaloškom okviru *Advanced Settings*.

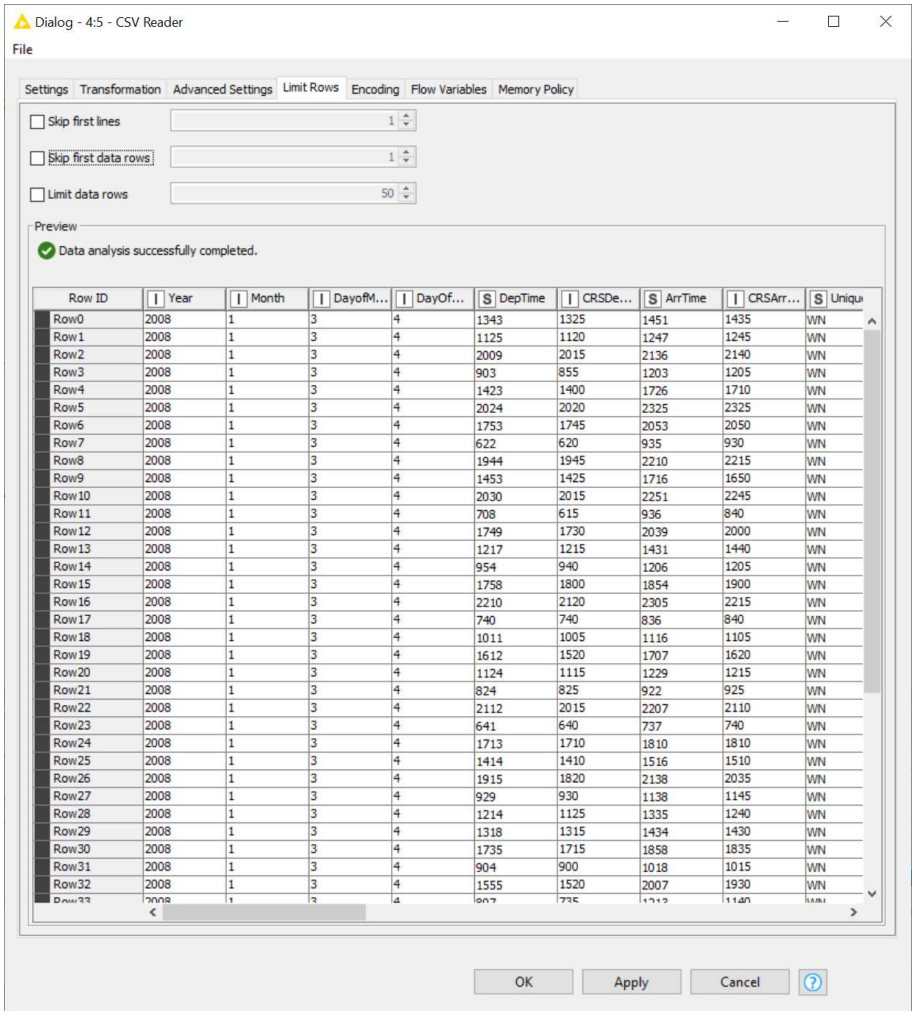


Slika 84. Treća kartica konfiguracije čvora CSV Reader

Opcija *Remove quotes and trim whitespaces* uključit će se ako je potrebno ukloniti navodnike i prazna mjesta između zareza u CSV datoteci, a opcija *Replace empty quoted strings with missing values* uključit će se ako je potrebno zamijeniti duple navodnike praznim mjestom. Opcija *Limit data rows scanned* isključit će se ako se želi da *KNIME* pregleda sve podatke prije nego što izabere u koju vrstu podatka će konvertirati vrijednosti iz CSV datoteke. Ova je opcija uključena kroz zadane postavke i ponekad uzrokuje da KNIME prijavljuje grešku pri konverziji.

U srednjem dijelu dijaloškog okvira bira se znak koji služi za odvajanje decimala, koji razlikuje decimalni zarez ili decimalnu točka. Osim toga, u tom dijelu može se izabrati znak za razdvajanje tisućica.

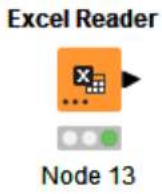
Na četvrtoj kartici dijaloškog okvira pod nazivom *Limit Rows* može se ograničiti broj redova koji *KNIME Analytics Platform* učitava i s kojim se dalje radi (*Limit data rows*), a moguće je i definirati i broj reda od kojeg počinje učitavanje podataka (*Skip first lines* i *Skip first data rows*).



Row ID	Year	Month	DayofM...	DayOf...	DepTime	CRSDe...	ArrTime	CRSArr...	Uniqu
Row0	2008	1	3	4	1343	1325	1451	1435	WN
Row1	2008	1	3	4	1125	1120	1247	1245	WN
Row2	2008	1	3	4	2009	2015	2136	2140	WN
Row3	2008	1	3	4	903	855	1203	1205	WN
Row4	2008	1	3	4	1423	1400	1726	1710	WN
Row5	2008	1	3	4	2024	2020	2325	2325	WN
Row6	2008	1	3	4	1753	1745	2053	2050	WN
Row7	2008	1	3	4	622	620	935	930	WN
Row8	2008	1	3	4	1944	1945	2210	2215	WN
Row9	2008	1	3	4	1453	1425	1716	1650	WN
Row10	2008	1	3	4	2030	2015	2251	2245	WN
Row11	2008	1	3	4	708	615	936	840	WN
Row12	2008	1	3	4	1749	1730	2039	2000	WN
Row13	2008	1	3	4	1217	1215	1431	1440	WN
Row14	2008	1	3	4	954	940	1206	1205	WN
Row15	2008	1	3	4	1758	1800	1854	1900	WN
Row16	2008	1	3	4	2210	2120	2305	2215	WN
Row17	2008	1	3	4	740	740	836	840	WN
Row18	2008	1	3	4	1011	1005	1116	1105	WN
Row19	2008	1	3	4	1612	1520	1707	1620	WN
Row20	2008	1	3	4	1124	1115	1229	1215	WN
Row21	2008	1	3	4	824	825	922	925	WN
Row22	2008	1	3	4	2112	2015	2207	2110	WN
Row23	2008	1	3	4	641	640	737	740	WN
Row24	2008	1	3	4	1713	1710	1810	1810	WN
Row25	2008	1	3	4	1414	1410	1516	1510	WN
Row26	2008	1	3	4	1915	1820	2138	2035	WN
Row27	2008	1	3	4	929	930	1138	1145	WN
Row28	2008	1	3	4	1214	1125	1335	1240	WN
Row29	2008	1	3	4	1318	1315	1434	1430	WN
Row30	2008	1	3	4	1735	1715	1858	1835	WN
Row31	2008	1	3	4	904	900	1018	1015	WN
Row32	2008	1	3	4	1555	1520	2007	1930	WN

Slika 85. Četvrta kartica konfiguracije čvora CSV Reader

Čvor **Excel reader** sličan je čvoru **CSV Reader**, ali se razlikuje zbog strukture radne knjige programa *Microsoft Excel* koja se može sastojati od više radnih listova, odnosno tablica. Slika 86 prikazuje čvor **Excel Reader**.



Slika 86. Čvor Excel Reader

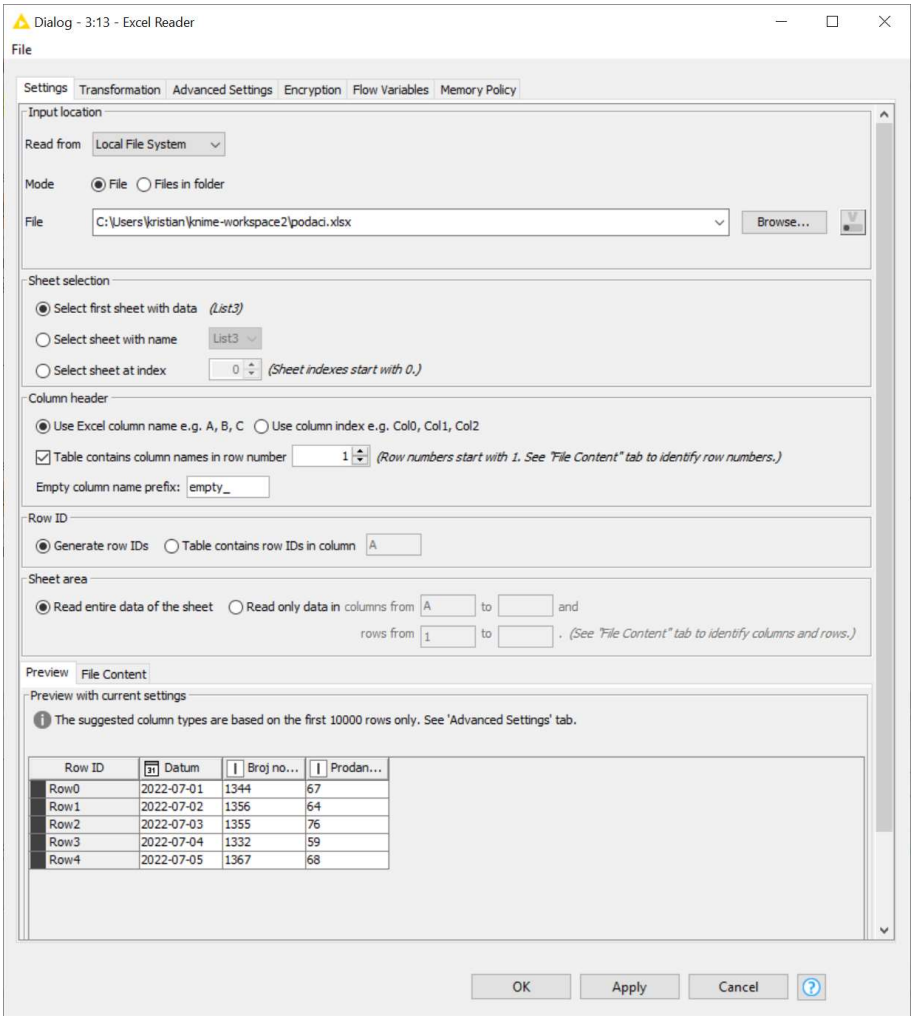
Slika 87 prikazuje prvu karticu dijaloškog okvira gdje se nalaze osnovne konfiguracijske postavke. Prvi korak je izbor radne knjige odabirom putanje, kojoj prethodi klik na gumb *Browse*. Nakon tog koraka moguće je u pravokutnom području nazvanom *Sheet selection* izabrati radni list s kojeg će se učitati podaci. To može biti prvi radni list, radni list koji se izabire po samom nazivu s time da su nazivi dostupni kroz padajući izbornik, a posljednja mogućnost je izabrati redni broj radnog lista s tim da prvi radni list ima redni broj 0.

U pravokutnom području *Column header* izabiru se nazivi stupaca pri čemu je moguće uzeti nazive koje *Microsoft Excel* zadano dodjeljuje stupcima (A, B, C,…), ali je moguće postaviti nazive stupaca s prefiksom *Col* i brojem stupca (*Col0, Col1, Col2,…*). Ako se u prvom redu radnog lista nalazi zaglavlje s nazivima stupaca što je dosta čest slučaj, tada je potrebno označiti *Table contains column names in row number* te definirati u kojem redu se nalazi zaglavlje.

U pravokutnom području *Row ID* odabire se da li će *KNIME* generirati brojeve redova u stupcu *Row ID* ili će ih preuzeti iz nekog stupca u radnom listu. Ako se odabere druga mogućnost, potrebno je definirati stupac u kojem se nalaze redni brojevi podataka.

U pravokutnom području *Sheet Area* odabire se područje u radnom listu s kojeg će se preuzeti podaci. U pravilu se ostavlja zadani izbor, osim ako se ne žele preuzeti podaci samo s dijela radnog lista. Treba obratiti pažnju ako na rubnim dijelovima radnog lista postoje neki izračuni da ih se ne preuzima, a za to se koristi ova mogućnost izbora područja.

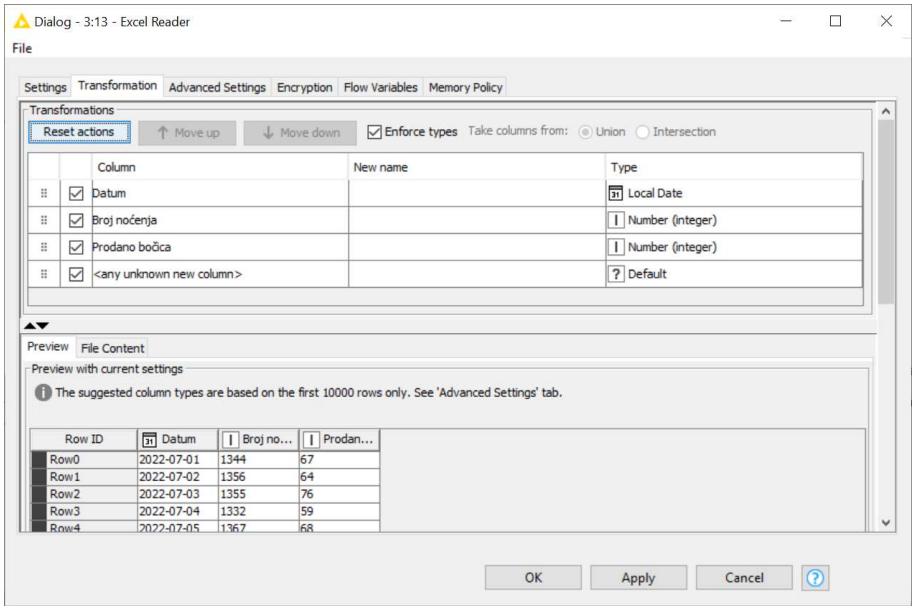
Konačno, na dnu dijaloškog okvira nalazi se pretpregled koji prikazuje učitan sadržaj s prethodno unesenim postavkama.



Slika 87. Prva kartica konfiguracije čvora Excel Reader

Na drugoj kartici postavki čvora **Excel Reader** može se definirati u koji format podataka će se učitati podaci iz radnog lista. Najčešće je prijedlog programa *KNIME* prihvatljiv, ali postoji mogućnost i ručne

prilagodbe po stupcu. Format se mijenja u zadnjem stupcu tablice koja se nalazi u gornjem dijelu dijaloškog okvira. Srednji stupac u toj tablici služi za izmjenu naziva stupca koji je definiran u zaglavlju radnog lista. Slika 88 prikazuje drugu karticu postavki čvora.



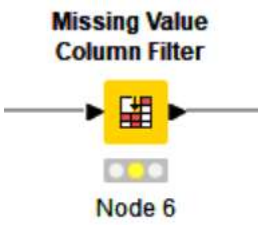
Slika 88. Druga kartica konfiguracije čvora Excel Reader

Kao što je navedeno, postoji cijeli niz drugih formata koji se mogu učitati u program *KNIME*, a za to postoje odgovarajući čvorovi. U daljem tekstu priručnika neki od tih čvorova bit će obrađeni, a za početak su dovoljni **Excel Reader** i **CSV Reader**.

4.2. Priprema podataka

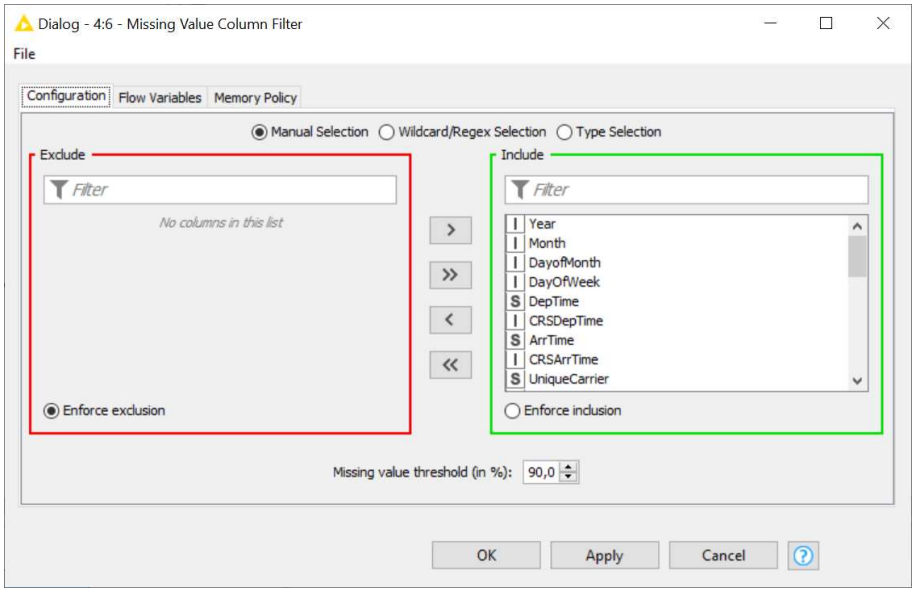
Nakon učitavanja podataka često postoji potreba za dodatnim aktivnostima vezanima uz pripremu podataka. U nastavku je opisano osam čvorova za pripremu podataka i to: **Missing Value Column Filter** (Filtar stupaca s vrijednostima koje nedostaju)**, Missing Value** (Vrijednost koja nedostaje)**, Constant Value Column Filter** (Filtar stupaca s konstantnim vrijednostima)**, Column Filter** (Filtar stupca)**, Row Filter** (Filtar retka)**, Sorter** (Razvrstavač)**, Duplicate Row Filter** (Filtar dupliciranog reda)**, Column Rename** (Preimenovanje stupca) **i Column Resorter** (Razvrstavač stupaca). Radi se o čvorovima koji služe za filtriranje, sortiranje, izmjenu naziva stupaca i manipulaciju praznim ćelijama.

Nakon pregleda učitanih podataka često se ukaže potreba za uklanjanjem određenih stupaca zbog prevelikog broja praznih ćelija. Čvor koji briše takve stupe naziva se **Missing Value Column Filter**.



Slika 89 Čvor Missing Value Column Filter

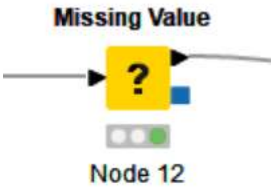
Za razliku od čvorova za učitavanje podataka, ovaj čvor ima znatno manje postavki koje se mogu izmijeniti. Označavanjem čvora i pritiskom tipke F6 na tipkovnici otvara se dijaloški okvir za postavke. Slika 90 prikazuje prvu karticu tog dijaloškog okvira.



Slika 90. Prva kartica čvora Missing Value Column Filter

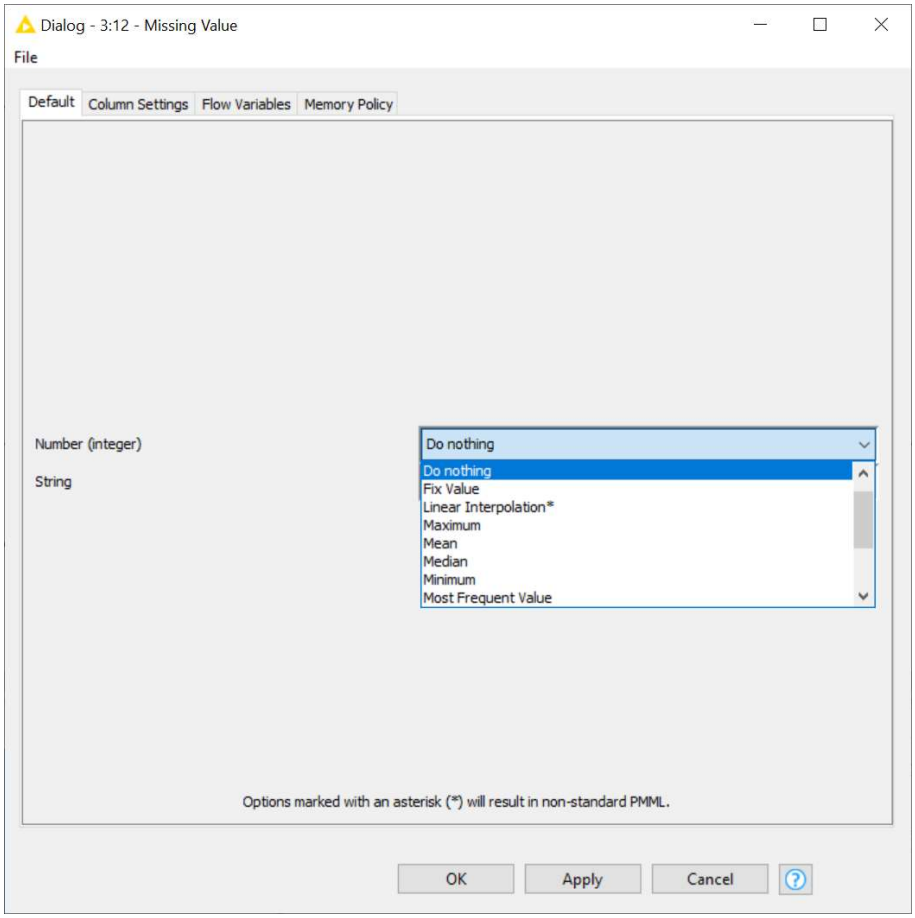
S desne strane u zelenom okviru su navedene varijable, odnosno nazivi stupaca na koje se može primijeniti uklanjanje stupaca. Granična vrijednost pri kojoj će stupac biti uklonjen definira se u polju iza teksta *Missing value threshold (in %).* Ako se uklanjanje ne želi primijeniti na nekim stupcima dovoljno ih je premjestiti u lijevi dio dijaloškog okvira koji je uokviren crveno. Premješta se na način da ih se odabere i klikne na gumb <, odnosno na > ovisno o željenom smjeru premještanja. Gumb s duplim znakovima premješta sve varijable, odnosno nazive stupaca u željenom smjeru.

Ako i dalje u tablici s podacima postoje vrijednosti koje nedostaju, na raspolaganju je čvor **Missing Value**. Čvor je prikazan na slici 91.



Slika 91. Čvor Missing Value

Ovaj čvor služi za popunjavanje praznih mjesta u stupcima, ukoliko postoje. U prvoj konfiguracijskoj kartici zadaju se vrijednosti kojima se popunjavaju prazne ćelije, ali na razini vrste podataka u stupcima cijele tablice. Tu je moguće posebno zadati što napraviti s praznim mjestima u stupcima u kojima se nalaze brojčane vrijednosti, a što napraviti u stupcima s tekstualnim vrijednostima. Prva kartica postavki takvoga čvora prikazana je na slici 92.



Slika 92. Prva kartica postavki čvora Missing Value

Mogućnosti koje su ponuđene za popunjavanje brojčanih praznih mjesta su sljedeće:

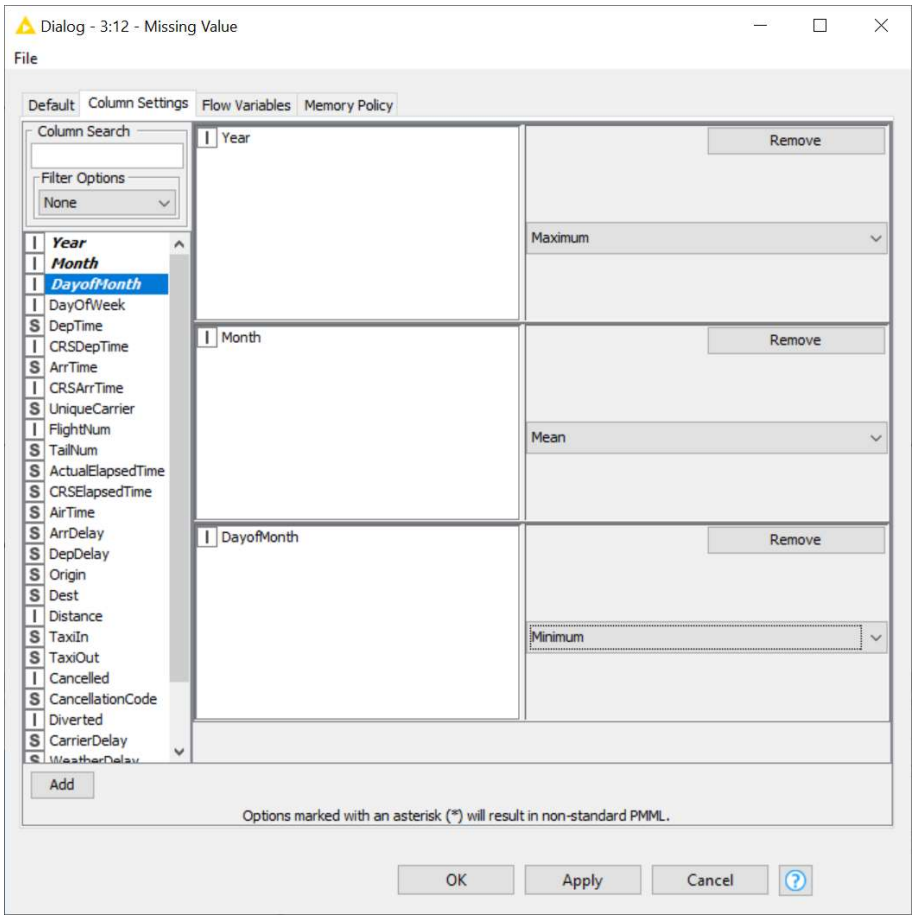
a) *Do nothing* – ne činiti ništa
b) *Fix Value* –fiksna vrijednost
c) *Linear Interpolation* –linearna interpolacija prethodne i sljedeće vrijednosti
d) *Maximum* – najveća vrijednost u stupcu
e) *Mean* – srednja vrijednost stupca
f) *Median* – medijan stupca
g) *Minimum* – najmanja vrijednost u stupcu
h) *Most Frequent Value* – najčešća vrijednost u stupcu
i) *Moving Average* – srednja vrijednost zadanog prozora oko vrijednosti
j) *Next Value* – sljedeća vrijednost
k) *Previous Value* – prethodna vrijednost
l) *Remove Row* – uklanjanje cijelog reda
m) *Rounded Mean* – zaokružena srednja vrijednost.

Mogućnosti koje su ponuđene za popunjavanje tekstualnih praznih mjesta su skromnije i tu se nudi sljedeće:

a) *Do nothing* – ne činiti ništa
b) *Fix Value* – fiksna vrijednost
c) *Most Frequent Value* - najčešća vrijednost u stupcu
d) *Next Value* – sljedeća vrijednost

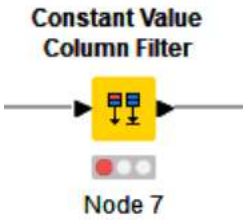
e) *Previous Value* – prethodna vrijednost
f) *Remove Row* – uklanjanje cijelog reda.

U drugoj konfiguracijskoj kartici *Custom Settings* moguće je definirati pravilo za prazne vrijednosti na razini pojedinačnog stupca. Potrebno je kliknuti na naziv stupca i nakon toga na gumb *Add*. Time se naziv stupca prebacuje u desni dio dijaloškog okvira i iz padajućeg izbornika bira se vrijednost koja će se umetnuti umjesto praznog polja. Ponuđene opcije su iste kao i gore navedene, ovisno o tome radi li se o stupcu s brojčanim ili tekstualnim podacima. Slika 93 prikazuje drugu karticu postavki čvora.



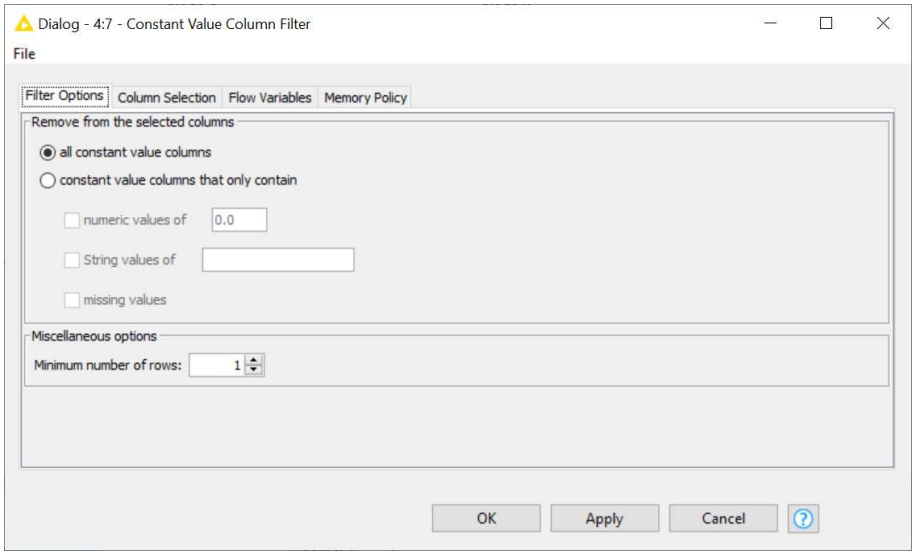
Slika 93. Druga kartica konfiguracije čvora Missing Value

U slučaju da u tablici s podacima postoji stupac koji sadrži konstantu, čvorom **Constant Value Column Filter** taj stupac se uklanja. Slika 94 prikazuje izgled čvora.



Slika 94. Čvor Constant Value Column Filter

Prečac za dobivanje postavki je pritiskom na tipku F6 na tipkovnici. Slika 95 slici 95 prikazan je izgled dijaloškog okvira postavki čvora.

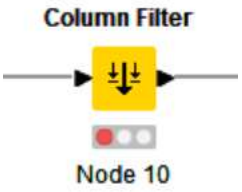


Slika 95. Postavke čvora Constant Value Column Filter

Ako se primjeni čvor **Constant Value Column Filter** sa zadanim postavkama, svi stupci s konstantnim vrijednostima će biti izbrisani. Ako se želi obrisati samo stupce koji imaju neki konkretan brojčani ili tekstualni podatak u svim redovima, onda se ta vrijednost unosi u za to predviđeno polje ispod teksta *constant value columns that only contain*. Prije unosa potrebno je označiti tu mogućnost, kao i željenu opciju *numeric values of* ili *String values of*.

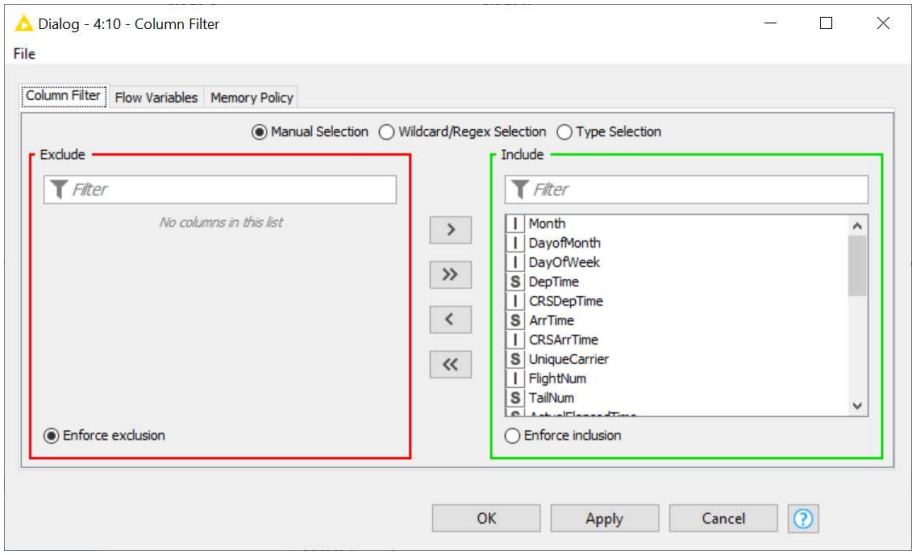
Izbor stupaca na kojem se navedeni čvor želi primijeniti može se mijenjati tako da se izabere druga kartica *Column Selection* u dijaloškom okviru postavki.

Čvor za filtriranje stupaca **Column Filter** jedan je od jednostavnijih. On služi za ručno isključivanje pojedinih stupaca iz dalje obrade. Na taj način se štedi memorijski prostor, a i jednostavnije je raditi s manje stupaca. Slika 96 prikazuje čvor **Column Filter**.



Slika 96. Čvor Column Filter

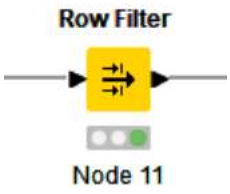
Na slici 97 prikazan je dijaloški okvir postavki čvora **Column Filter**.



Slika 97. Postavke čvora Column Filter

Premještanje varijabli, odnosno naziva stupaca iz zelenog okvira u crveni okvir, rezultira uklanjanjem tih stupaca iz dalje obrade.

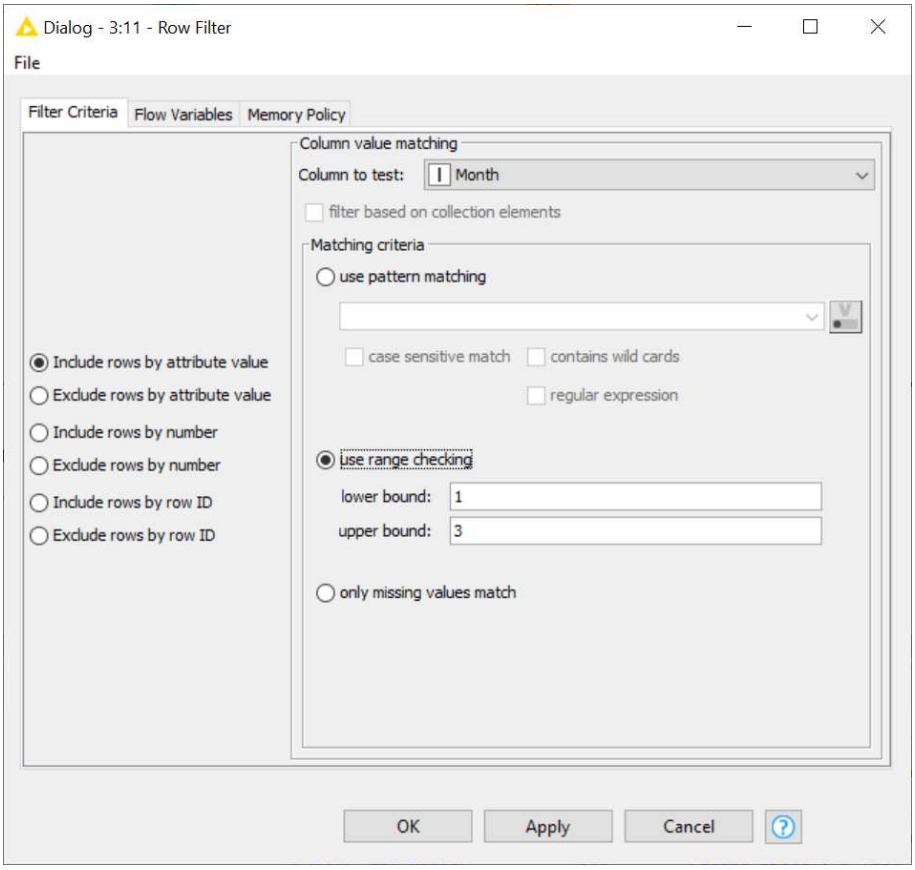
Čvor kojim je moguće filtrirati redove očekivano se naziva **Row Filter**. Slika 98 prikazan je izgled toga čvora.



Slika 98. Čvor Row Filter

Postavke za ovaj čvor ovise o tome što se želi isključiti, odnosno ostaviti. Slika 99 prikazan je dijaloški okvir u kojem je izabrana prva od ponuđenih mogućnosti koja uključuje redove ovisno o vrijednosti ćelija izabranog stupca. U gornjem desnom dijelu bira se stupac čije se vrijednosti testiraju, a nude se tri mogućnosti. Može se unijeti vrijednost u ćeliju ako se želi da red ostane u tablici, pri čemu se mogu koristiti i tzv. „regularni izrazi“ (eng. *Regular Expressions*). Radi se o načinu zadavanja pravila kojima se opisuju i generiraju drugi nizovi znakova, no ovdje se neće posebno obrađivati. Nagy (2018) detaljno opisuje to područje (Nagy, 2018). Druga mogućnost je navesti donju i gornju granicu nekog niza, a treća mogućnost je uključiti samo prazne ćelije. Iz skupa podataka *Data Expo 2009: Airline On Time Data* izabran je stupac *Month* te su ostali samo redovi u kojima je vrijednost ćelije toga stupca broj između 1 i 3. Krajnje vrijednosti su uključene.

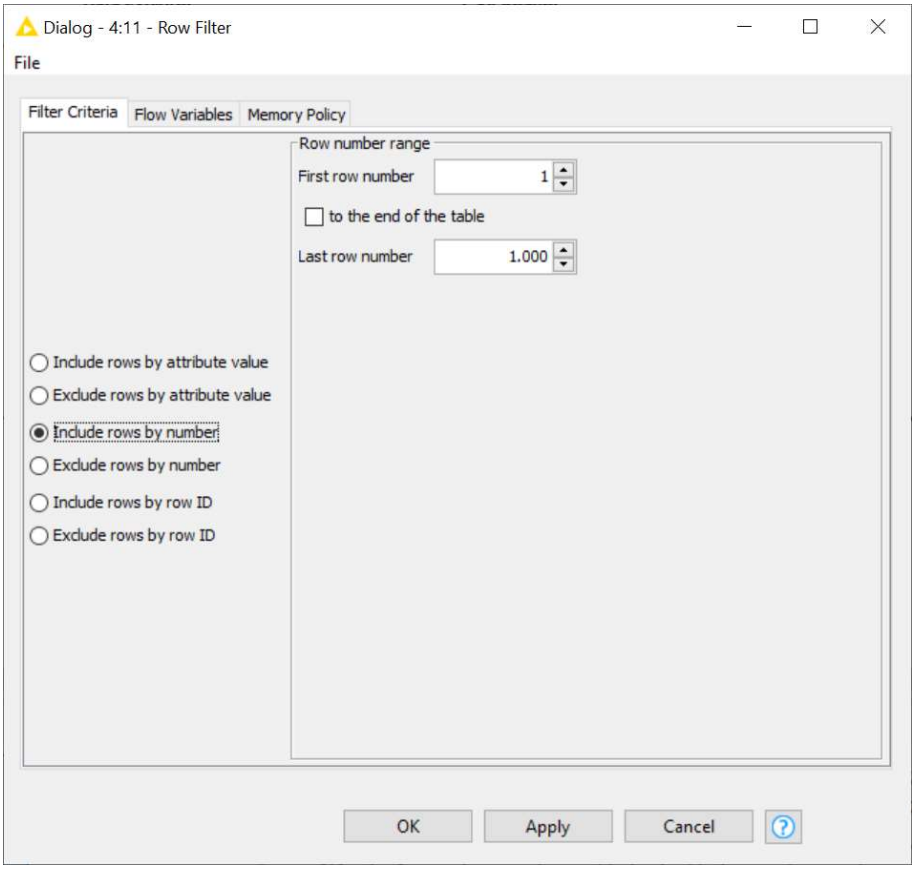
Druga ponuđena mogućnost isključuje redove ovisno o vrijednosti ćelija izabranog stupca. Jedina razlika jest u tome što se redovi isključuju, umjesto uključuju (ostavljaju), ovisno o zadanom kriteriju.



Slika 99. Postavke filtriranja redova tako da ostanu prva 3 mjeseca

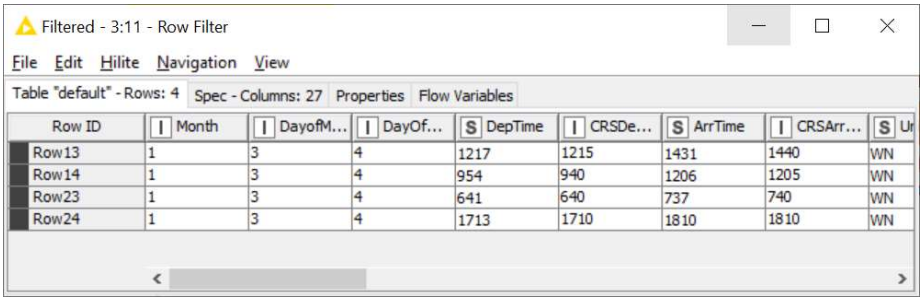
Treća mogućnost filtriranja redova prikazana je na slici 100 pri čemu se definiraju početni i završni redovi koji u navedenom slučaju ostaju u tablici za dalju obradu. Postoji mogućnost da se zada početni red i izabere opcija *to the end of the table* pri čemu će biti uključeni svi redovi do kraja tablice. Pri tom nije potrebno znati koliko redova sadrži tablica. U konkretnom primjeru će u daljnju obradu proći samo prvih 1000 redova jer je zadan početni red 1 i krajnji red 1000.

Četvrta mogućnost je vrlo slična trećoj s razlikom da se isključuju redovi koji zadovoljavaju zadani uvjet vezan uz početni i završni broj reda.



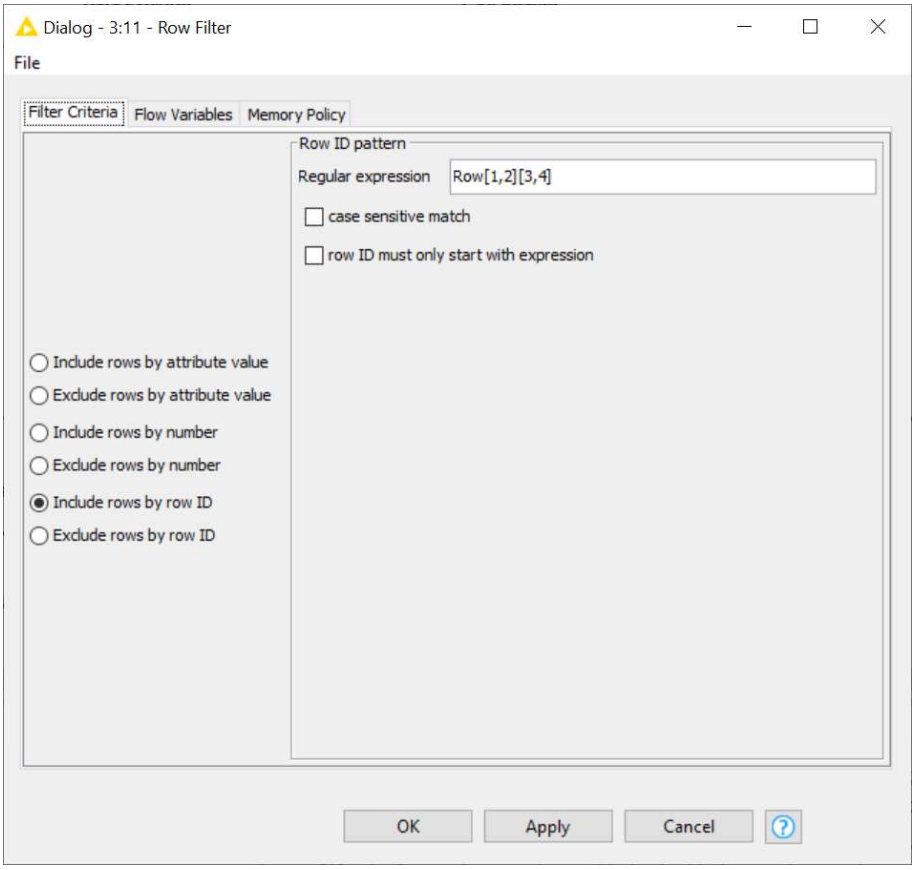
Slika 100. Postavke filtriranja redova pri čemu ostaju samo redovi od 1. do 1000.

Peta ponuđena mogućnost uključuje redove čiji prvi stupac **Row ID** zadovoljava kriterij definiran regularnim izrazom. S obzirom na prethodno spomenutu kompleksnost i obim materije vezane uz temu regularnih izraza, navest će se jednostavan primjer regularnog izraza. Unesen je tekst *Row[1,2][3,4]* koji označava sve mogućnosti koje počinju s tri slova Row, a iza toga je četvrti znak 1 ili 2 te peti znak 3 ili 4. Time se dobivaju sljedeće mogućnosti: Row13, Row14, Row23 i Row24. Navedeni rezultat u stupcu *Row ID* može se vidjeti i nakon izvođenja čvora klikom na *Filtered* iz kontekstnog izbornika čvora. Slika 101 prikazuje rezultat izvođenja čvora. Slika 102 prikazuje konfiguraciju prije izvođenja.



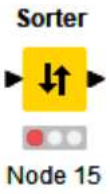
Row ID	Month	DayofM...	DayOf...	DepTime	CRSDe...	ArrTime	CRSArr...	Ur
Row13	1	3	4	1217	1215	1431	1440	WN
Row14	1	3	4	954	940	1206	1205	WN
Row23	1	3	4	641	640	737	740	WN
Row24	1	3	4	1713	1710	1810	1810	WN

Slika 101. Rezultat filtriranja primjenom regularnog izraza



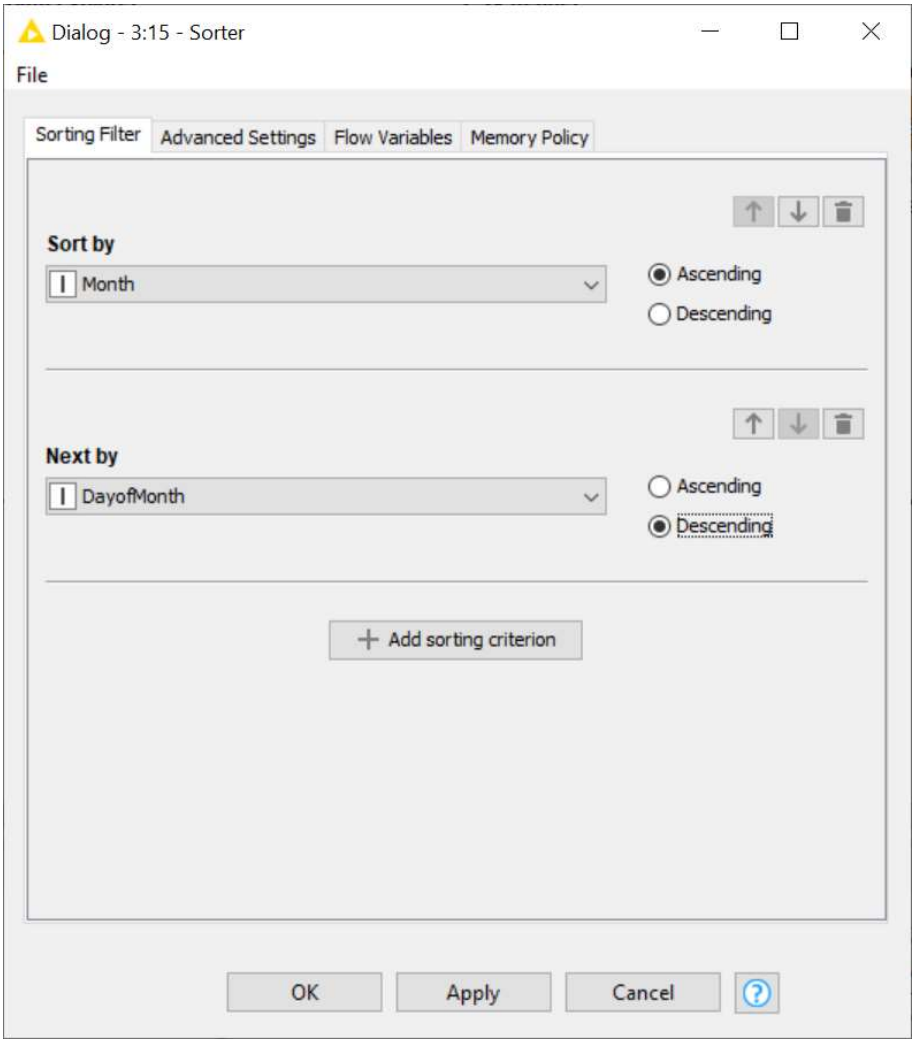
Slika 102. Postavke filtriranja redova pri korištenju regularnih izraza

Čvor **Sorter** služi za sortiranje. U nekim situacijama ulazne podatke potrebno je sortirati prema određenom kriteriju, zbog čega se koristi ovaj čvor. Slika 103 prikazuje izgled čvora.



Slika 103. Čvor Sorter

Postavke ovoga čvora su relativno jednostavne i slične sortiranju u tabličnim kalkulatorima. Jednostavno se izabere stupac po kojem se želi sortirati, kao i vrsta sortiranja (rastuće ili padajuće). Rastuće sortiranje bira se klikom na *Ascending*, a padajuće sortiranje opcijom *Descending*. Ako je potrebno višerazinsko sortiranje jednostavno se doda niža razinu ispod više razine klikom na *Add sorting criterion*, a razine sortiranja se mogu izmjenjivati korištenjem okomitih strelica. Na slici 104 prikazane su postavke čvora **Sorter**.

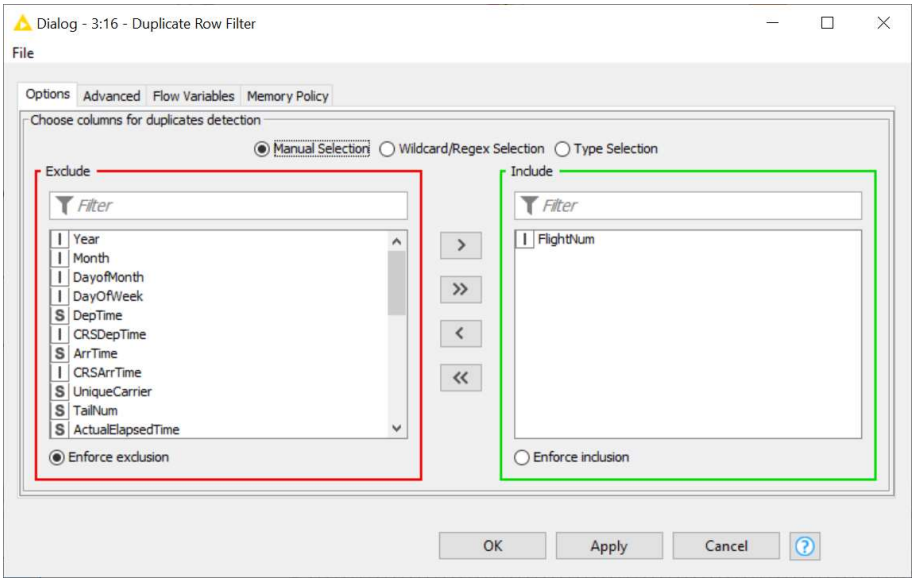


Slika 104. Postavke čvora Sorter

Čvor **Duplicate Row Filter** služi za pronalaženje i isključivanje redova koji u određenom stupcu imaju istu vrijednost. Nalazi se na slici 105.

Slika 105. Čvor Duplicate Row Filter

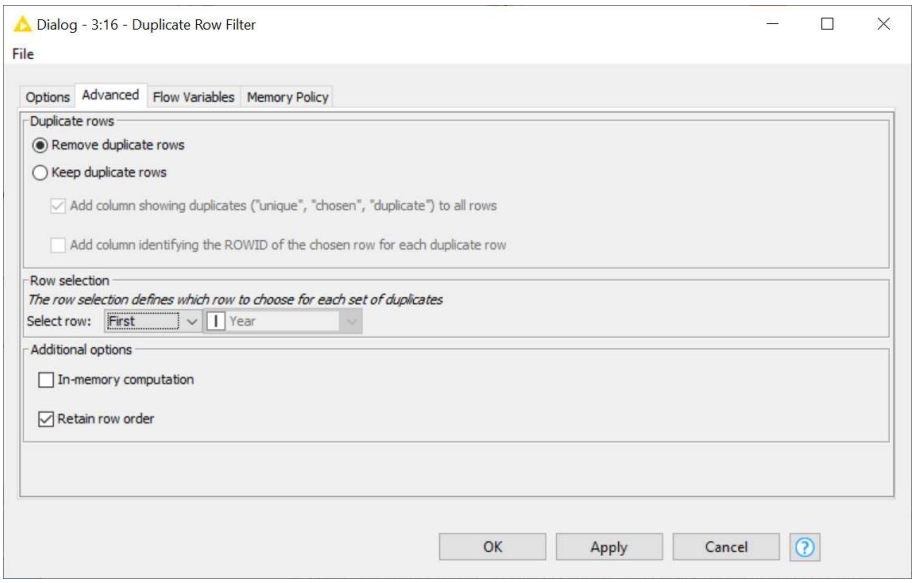
Postavke čvora **Duplicate Row Filter** su na slici 106 i svode se najčešće na izbor jednog stupca koji ostaje u zelenom okviru i u tom stupcu traže se duplikati, odnosno vrijednosti koje se pojavljuju više puta.



Slika 106. Prva kartica postavki čvora Duplicate Row Filter

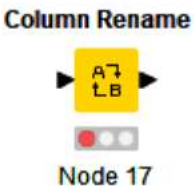
U slučaju da se pronađu barem dvije ćelije s istom vrijednošću, tada se izvršava aktivnost zadana u drugoj kartici postavki. Prva ponuđena mogućnost je brisanje redova u kojima se nalaze iste vrijednosti. S obzirom da jedan red s tom jedinstvenom vrijednosti treba ostati, u dijelu *Row selection* bira se red koji ostaje. Zadano je da prvi red ostaje, ali osim te mogućnosti može se izabrati i zadnji red, odnosno red u kojem je vrijednost pojedinog stupca najveća ili najmanja, pri čemu se može izabrati red koji se analizira.

Osim brisanja redova moguće je definirati da redovi s istim vrijednostima u izabranom stupcu ostaju u tablici, ali pri tom se tablici dodaje novi stupac u kojem se za svaki postojeći red podataka postavlja vrijednost „*unique*", „*chosen*" ili „*duplicate*". Vrijednost „*unique*" postavlja se ako je vrijednost ćelije u promatranom stupcu jedinstvena. U slučaju da postoje duplicirane vrijednosti, za izabranu se vrijednost postavlja opcija „*chosen*", a za ostale duplikate postavlja se vrijednost „duplicate". Dijalog za navedene postavke je na slici 107.



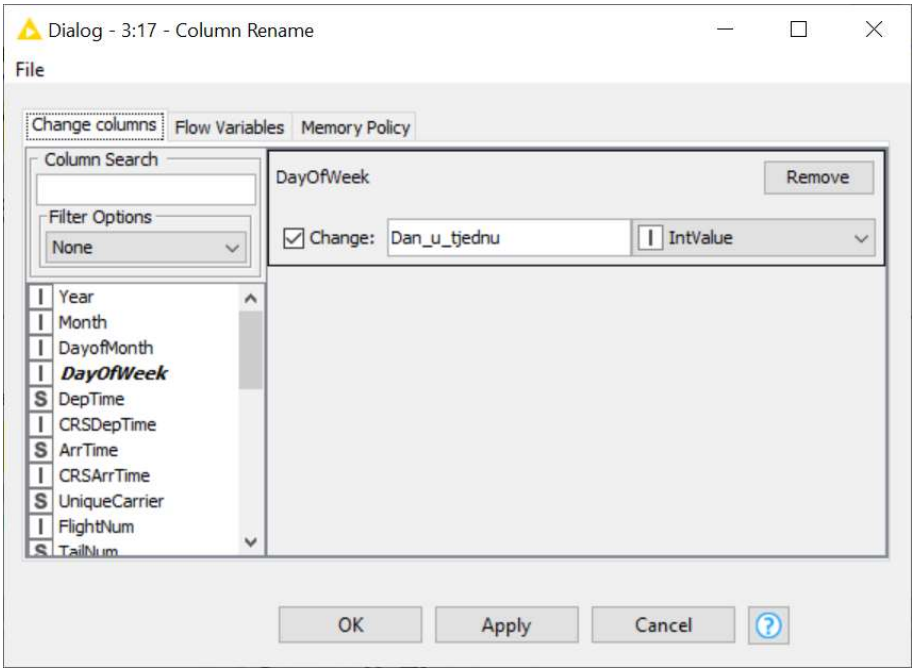
Slika 107. Druga kartica postavku čvora Duplicate Row Filter

Ako se ukaže potreba za promjenom imena stupca, tada se koristi čvor **Column Rename**. Slika 108 prikazuje čvor.



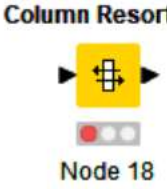
Slika 108. Čvor Column Rename

Postavke čvora **Column Rename** svode se na izbor naziva stupca i unosa novog imena stupca u za to predviđeno polje. Osim toga, moguće su i promjene vrste podatka u padajućem izborniku iza novog imena stupca. Ovaj čvor omogućuje izmjenu više naziva stupaca pri čemu se u dijaloškom okviru postavki stvara lista izmjena. Slika 109 prikazuje dijaloški okvir u kojem se mijenjaju postavke čvora.



Slika 109. Postavke čvora Column Rename

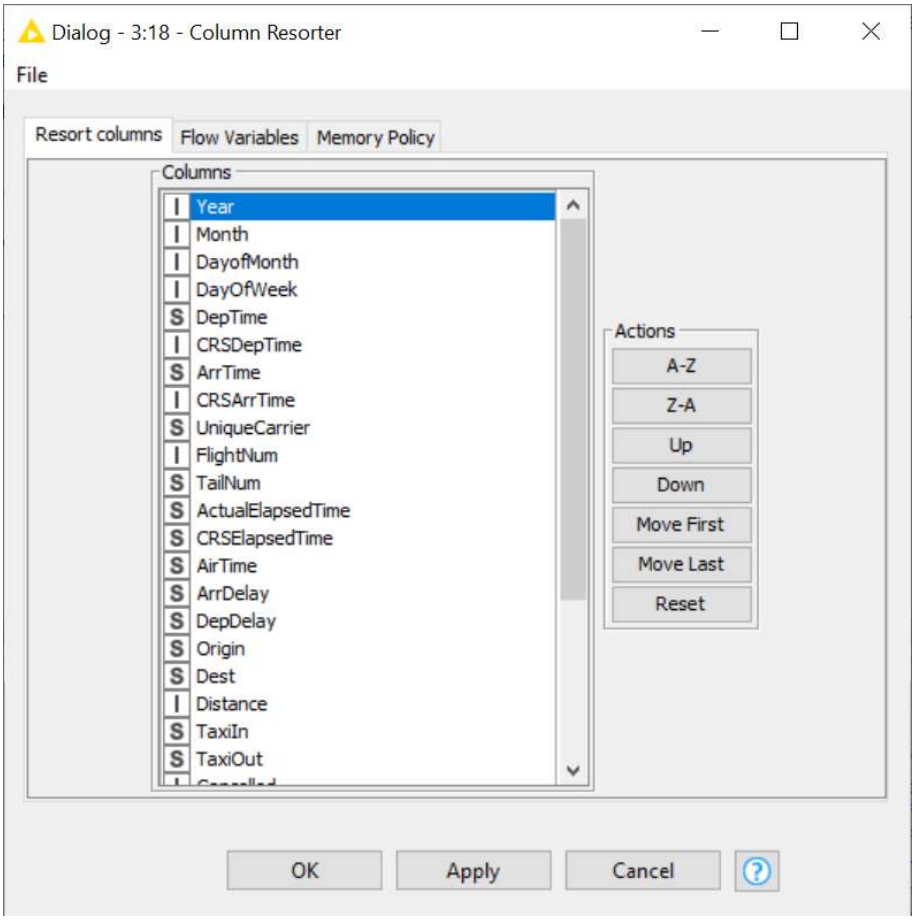
U slučaju da postoji potreba za izmjenom redoslijeda stupaca, koristi se čvor **Column Resorter** koji se vidi na slici 110.



Slika 110. Čvor Column Resorter

Taj čvor omogućava preslagivanje redoslijeda stupaca po želji. Za kreiranje modela strojnog učenja redoslijed stupaca je nebitan, ali KNIME se može i koristiti samo za sređivanje i konverziju podataka, u

tom su slučaju ovakvi čvorovi korisni. Rezultat hodograma ne mora biti kreiranje modela, nego i stvaranje nove datoteke u formatu Microsoft Excel radne knjige, kada se koristi čvor **Column Resorter**. Slika 111 prikazuje dijaloški okvir u kojem se mogu mijenjati postavke čvora.



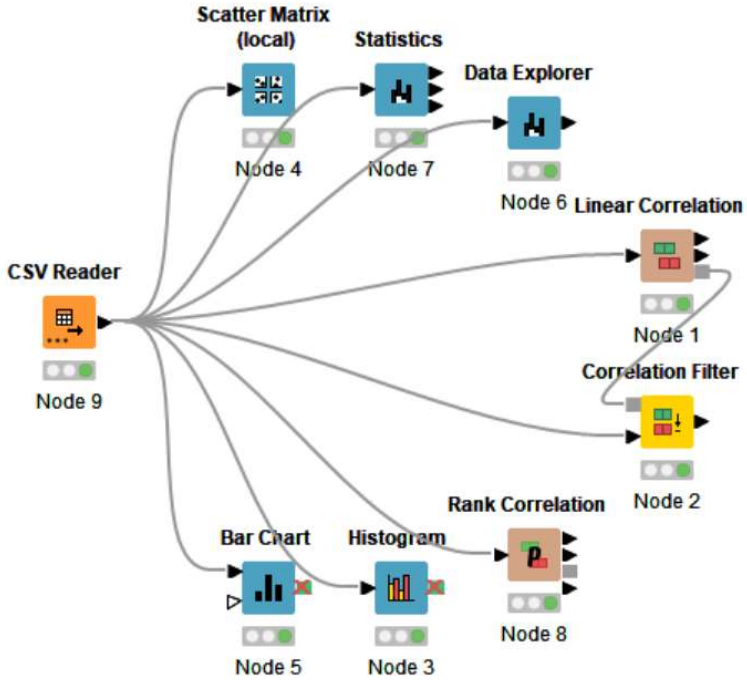
Slika 111. Postavke čvora Column Resorter

Akcije koje su na raspolaganju su sortiranje po abecedi uzlazno (*A-Z*) i silazno (*Z-A*), zatim gumbi *Up* i *Down* koji omogućuju izmjenu redoslijeda gore ili dolje u listi naziva stupaca. Izabrani stupac može se prebaciti na prvo mjesto (*Move First*) ili na zadnje mjesto (*Move Last*), a moguće je i vratiti redoslijed stupaca na početni (*Reset*).

Nakon upoznavanja s čvorovima za učitavanje i pripremu podataka može se krenuti i s analizom učitanih podataka.

5. Analiza podataka

Nakon što su učitani podaci i odrađena je osnovna priprema za izradu modela, bitan korak je upoznavanje sa samim podacima na osnovu kojih se želi izgraditi model. Slika 112 prikazuje hodogram s osam čvorova koji mogu pomoći u analizi i upoznavanju s podacima. U nastavku će prikazani čvorovi biti opisani, a isto tako, vidjet će se kako se njihove funkcije u nekim slučajevima preklapaju.



Slika 112. Hodogram za analizu podataka

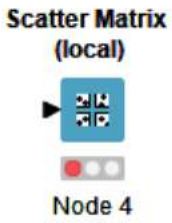
Da bi se vidjelo kako čvorovi funkcioniraju na konkretnim podacima, koristit će se podaci o stotinjak tisuća registriranih korisnika francuske C2C web trgovine. Osoba koja je postavila podatke na poslužitelja www.kaggle.com navodi kako su podaci iz baze u kojoj je oko 10 milijuna korisnika, ali ovo je samo dio tih podataka. U analizi će biti vidljivo da podaci nisu izabrani nasumično, nego je izabrana samo jedna selektirana skupina korisnika, o tome će biti govora u narednim poglavljima. Podaci su dostupni u CSV formatu u datoteci *6M-0K-99K.users.dataset.public.csv* na adresi: https://www.kaggle.com/datasets/jmmvutu/ecommerce-users-of-a-french-c2c-fashion-store . Da bi se podaci preuzeli, potrebna je registracija na poslužitelj, prijava je moguća i uz Google korisnički račun. Nakon preuzimanja skupa podataka, najčešće je potrebno raspakirati preuzetu datoteku i iz nje preuzeti navedenu datoteku. Za početak treba se upoznati sa značajkama, odnosno nazivima stupaca. U nastavku slijede nazivi i opisi.

- *identifierHash* - hash ID-a korisnika
- *type* - vrsta korisnika
- *country* - zemlja korisnika (napisano na francuskom jeziku, bitno za kasniju analizu)
- *language* - korisnikov preferirani jezik
- *socialNbFollowers* - broj korisnika koji su se pretplatili na aktivnost korisnika
- *socialNbFollows* - broj korisničkog računa koji korisnik prati
- *socialProductsLiked* - broj proizvoda koji su se svidjeli korisniku
- *productsListed* - broj trenutno neprodanih proizvoda koje je korisnik učitao

- *productsSold* - broj proizvoda koje je korisnik prodao
- *productsPassRate* - % proizvoda koji zadovoljavaju opis proizvoda (prodane proizvode pregledava tim trgovine prije nego što se pošalju kupcu.)
- *productsWished* - broj proizvoda koje je korisnik dodao na svoju listu želja
- *productsBought* - broj proizvoda koje je korisnik kupio
- *gender* - spol korisnika
- *civilityGenderId* – titula kao cijeli broj
- *civilityTitle* - titula
- *hasAnyApp* - korisnik je barem jednom koristio bilo koju službenu aplikaciju trgovine
- *hasAndroidApp* - korisnik barem jednom koristio službenu Android aplikaciju
- *hasIosApp* - korisnik barem jednom koristio službenu iOS aplikaciju
- *hasProfilePicture* - korisnik ima prilagođenu profilnu sliku
- *daysSinceLastLogin* - broj dana od posljednje prijave
- *seniority* - broj dana od registracije korisnika
- *seniorityAsMonths* - broj mjeseci od registracije korisnika
- *seniorityAsYears* - broj godina od registracije korisnika
- *countryCode* - zemlja korisnika (ISO-3166-1).

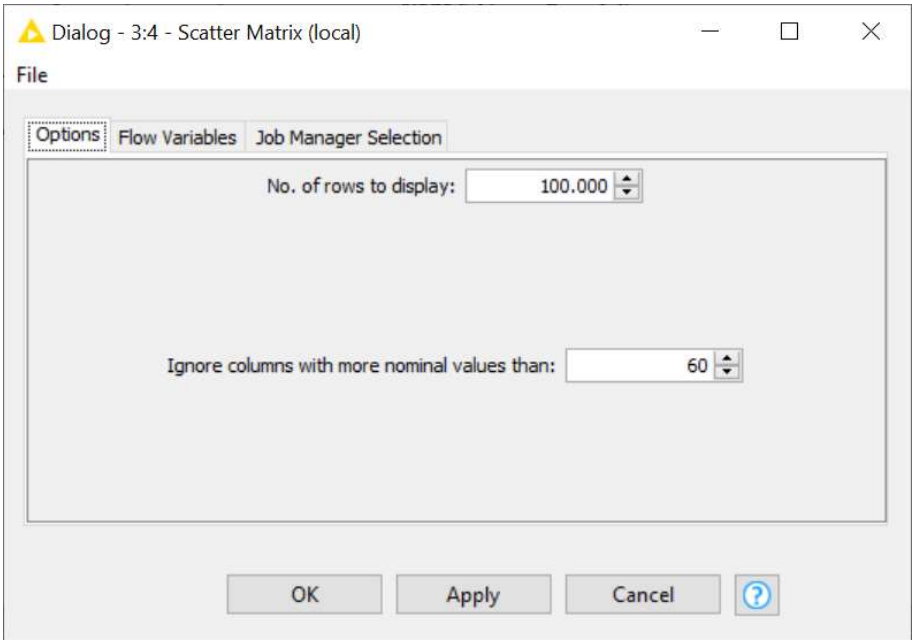
Prvi čvor na hodogramu je **CSV Reader** ili CSV čitač i služi za učitavanje podataka iz CSV datoteke. Potrebno je učitati podatke iz datoteke *6M-0K-99K.users.dataset.public.csv.*

Sljedeći čvor koji do sada nije opisan je čvor **Scatter Matrix (local)** ili Matrica raspršenosti (lokalna). **Pogreška! Izvor reference nije pronađen.** prikazuje navedeni čvor. Radi se o matrici raspršenja u kojoj je svaki element matrice dijagram raspršenja stupaca i i j, gdje su vrijednosti i-tog stupca prikazane na x osi, a vrijednosti j-tog stupca na y-osi, dok su koordinate prikazane na svim stranama dijagrama.



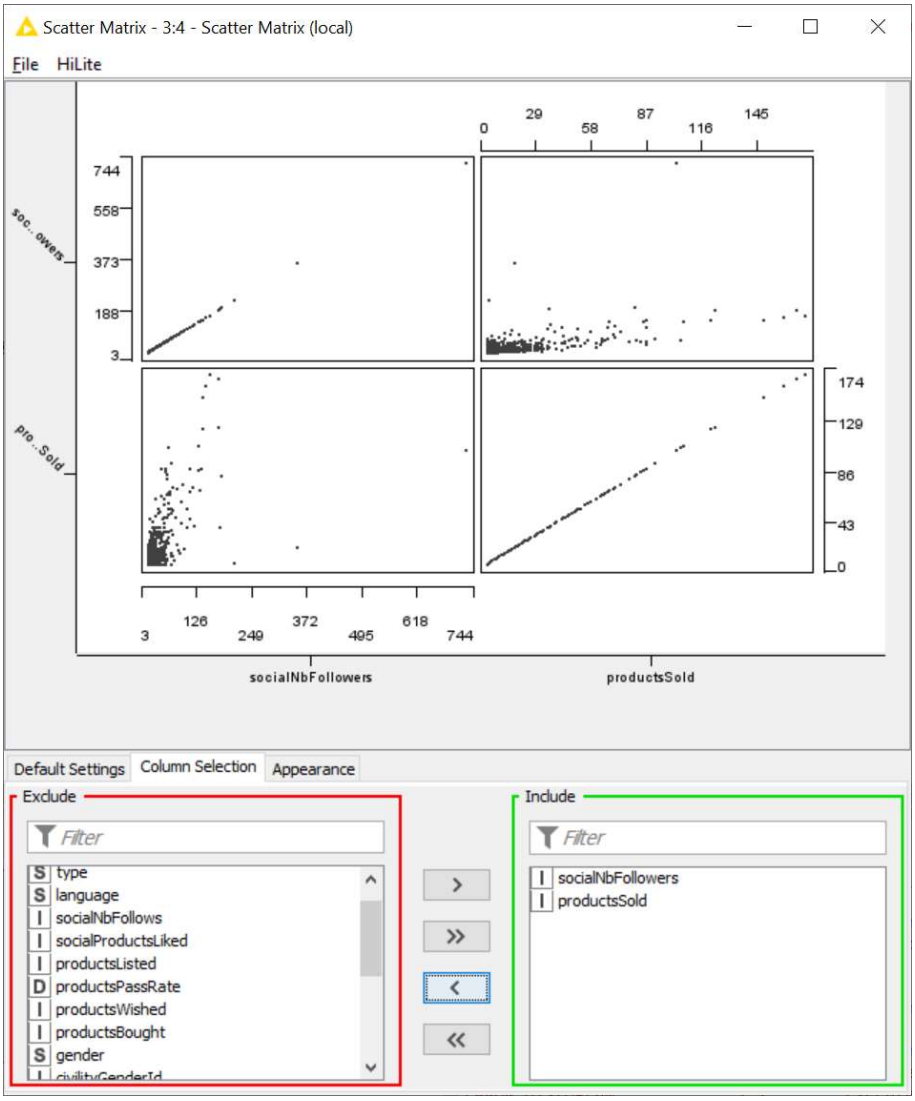
Slika 113. Čvor Scatter Matrix (local)

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Scatter Matrix (local)** u kojem se mogu podešavati samo dvije vrijednosti. Prva je broj redaka, odnosno podataka koji se prikazuje. Zadano je 2500 redova, ali s obzirom da datoteka sadrži blizu 100.000 redova u CSV datoteci unesen je taj broj. U pravilu bi program trebao dobro raditi i s 100.000 redova podataka , mada to ovisi dobrim dijelom o računalu na kojem je instaliran KNIME i na kojem se radi. Druga vrijednost je broj nominalnih vrijednosti nakon koje se stupac ignorira, odnosno ne prikazuje se. Nominalne vrijednosti pripadaju nominalnoj mjernoj ljestvici, a njihovo svojstvo je da se ne mogu uređivati redoslijedom, ali se mogu prebrojiti (Horvat & Mijoč, 2019). Primjer za nominalnu varijablu je spol. Ta postavka služi da se izbace značajke s puno jedinstvenih nominalnih vrijednosti što najvjerojatnije ne doprinose modelu. U učitanoj CSV datoteci takav slučaj je značajka *identifierHash*, koja sadrži tzv. *hash* zbrojeve korisničkih imena i za svakog korisnika ta vrijednost je jedinstvena. S obzirom da se radi o vrijednostima na nominalnoj ljestvici koje su jedinstvene, taj stupac je za model potpuno beskoristan i ako se ostavi postavka *Ignore columns with more nominal values than* na zadanoj vrijednosti (60), taj stupac će biti isključen iz prikaza.



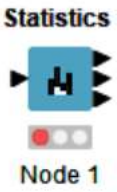
Slika 114. Postavke čvora Scatter Matrix (local)

Pogreška! Izvor reference nije pronađen. prikazuje rezultat čvora **Scatter Matrix (local)** nakon obrade podataka, ali taj prikaz omogućuje interakciju, odnosno izbor varijabli koje se želi prikazati. U donji desni dio dijaloškog okvira, odnosno u zeleni okvir, treba postaviti varijable za koje se želi provjeriti postojanje povezanosti među odabranim varijablama. U ovom slučaju može se pretpostaviti kako postoji povezanost između broja korisnika koji su se pretplatili na aktivnost korisnika i broja proizvoda koje je korisnik prodao. Ako je veći broj korisnika pretplaćen na aktivnosti korisnika, svakim novim postavljanjem predmeta na prodaju, veći broj korisnika o tome biva informiran. Time se povećava i vjerojatnost prodaje predmeta. Donji lijevi, odnosno gornji desni dijagrami to i potvrđuju odnosno ukazuju na tu povezanost. Gornji lijevi i donji desni dijagram treba zanemariti.



Slika 115. Rezultat čvora Scatter Matrix (local)

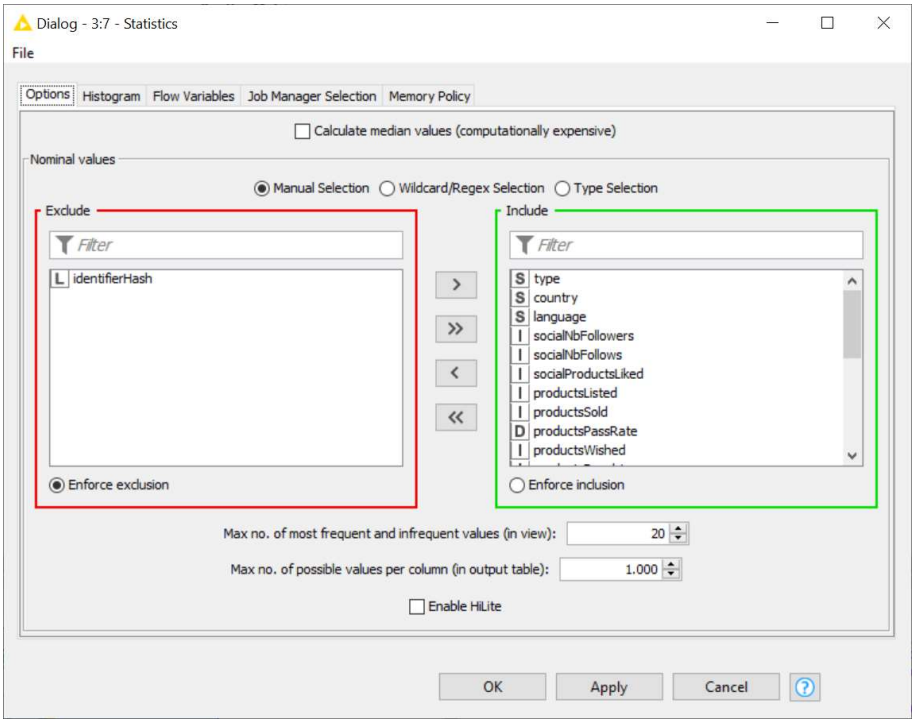
Pogreška! Izvor reference nije pronađen. prikazuje čvor **Statistics** ili Statistika. On je sljedeći u nizu na hodogramu i izračunava statističke pokazatelje kao što su minimum, maksimum, srednja vrijednost, standardna devijacija, varijanca, medijan, ukupni zbroj, broj nedostajućih vrijednosti i broj redaka u svim numeričkim stupcima te broj svih kategorijalnih vrijednosti zajedno s njihovim pojavljivanjima.



Slika 116. Čvor Statistics

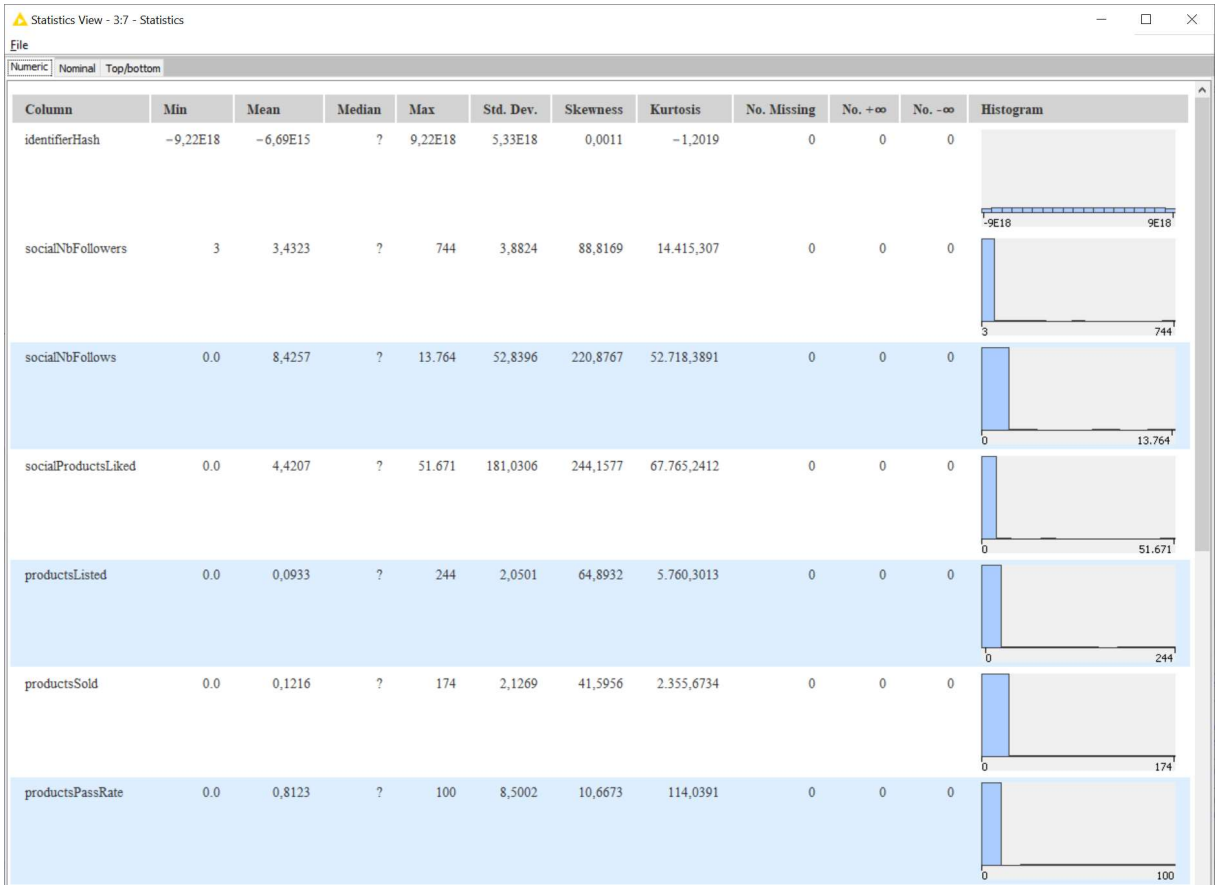
Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Statistics**. U vrhu dijaloškog okvira nalazi se opcija za izračunavanje medijana za sve numeričke značajke. U zagradi je upozorenje kako to može biti zahtjevno što znači da može potrajati nešto duže zbog većeg broja redova i stupaca. Ispod se nalaze crveni i zeleni okviri u kojima se definiraju značajke koje će se statistički analizirati.

Ispod toga je postavka kojom se definira broj najčešćih i najrjeđih pojavljivanja kategorijalnih vrijednosti u trećoj kartici rezultata (*Top/bottom*), a postavka se naziva *Max no of most frequent and infrequent values (in view)*.



Slika 117. Postavke čvora Statistics

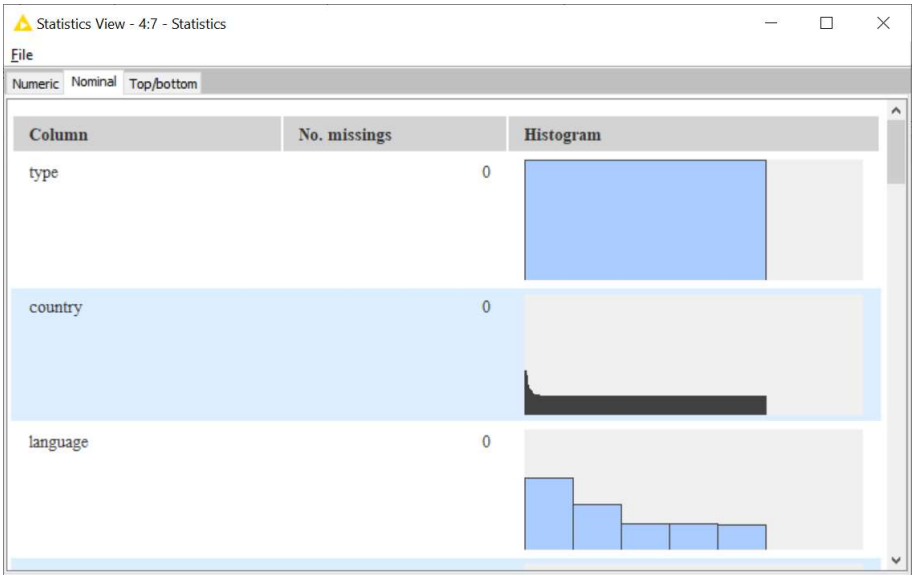
Pogreška! Izvor reference nije pronađen. prikazuje rezultate analize podataka čvora **Statistics**. Ovaj dijaloški okvir dobiva se na način da se izabere *View: Statistics view* iz kontekstnog izbornika čvora. U dijaloškom okviru postoje tri kartice. Prva kartica se naziva *Numeric* i prikazuje statističke podatke za brojčane varijable. Osim očekivanih statističkih parametara podataka na kraju svakog retka je i mali histogram koji prikazuje distribuciju vrijednosti.



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
identifierHash	−9,22E18	−6,69E15	?	9,22E18	5,33E18	0,0011	−1,2019	0	0	0	-9E18 9E18
socialNbFollowers	3	3,4323	?	744	3,8824	88,8169	14.415,307	0	0	0	3 744
socialNbFollows	0.0	8,4257	?	13.764	52,8396	220,8767	52.718,3891	0	0	0	0 13.764
socialProductsLiked	0.0	4,4207	?	51.671	181,0306	244,1577	67.765,2412	0	0	0	0 51.671
productsListed	0.0	0,0933	?	244	2,0501	64,8932	5.760,3013	0	0	0	0 244
productsSold	0.0	0,1216	?	174	2,1269	41,5956	2.355,6734	0	0	0	0 174
productsPassRate	0.0	0,8123	?	100	8,5002	10,6673	114,0391	0	0	0	0 100

Slika 118. Numeric rezultat

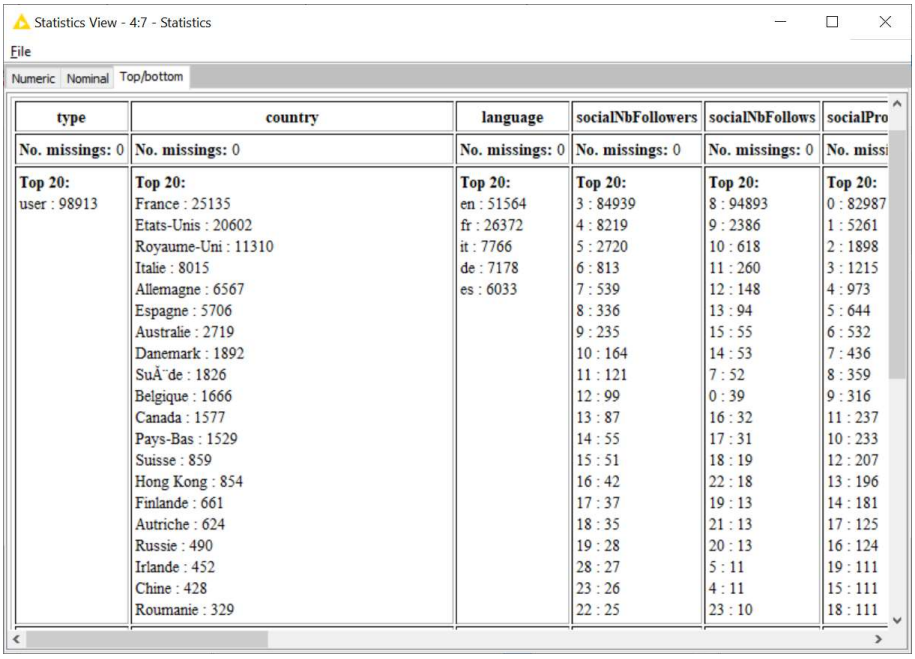
Druga i treća kartica prikazuju kategorijalne podatke koji pripadaju nominalnoj ljestvici na sljedeći način. Kartica *Nominal* prikazuje tri stupca: naziv varijable/značajke, broj elemenata koji nedostaju i histogram. **Pogreška! Izvor reference nije pronađen.** prikazuje *Nominal* karticu.



Column	No. missings	Histogram
type	0	
country	0	
language	0	

Slika 119. Nominal rezultat

U kartici *Top/bottom* prikazane su najčešće i najrjeđe vrijednosti. Zadano je prikazivanje 20 najčešćih i 20 najrjeđih nominalnih vrijednosti koje se pojavljuju u svakom stupcu, ta opcija se može promijeniti na postavkama čvora. **Pogreška! Izvor reference nije pronađen.** prikazuje tu karticu.



type	country	language	socialNbFollowers	socialNbFollows	socialPro
No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missi
Top 20: user : 98913	Top 20: France : 25135 Etats-Unis : 20602 Royaume-Uni : 11310 Italie : 8015 Allemagne : 6567 Espagne : 5706 Australie : 2719 Danemark : 1892 SuÄ¨de : 1826 Belgique : 1666 Canada : 1577 Pays-Bas : 1529 Suisse : 859 Hong Kong : 854 Finlande : 661 Autriche : 624 Russie : 490 Irlande : 452 Chine : 428 Roumanie : 329	Top 20: en : 51564 fr : 26372 it : 7766 de : 7178 es : 6033	Top 20: 3 : 84939 4 : 8219 5 : 2720 6 : 813 7 : 539 8 : 336 9 : 235 10 : 164 11 : 121 12 : 99 13 : 87 14 : 55 15 : 51 16 : 42 17 : 37 18 : 35 19 : 28 28 : 27 23 : 26 22 : 25	Top 20: 8 : 94893 9 : 2386 10 : 618 11 : 260 12 : 148 13 : 94 15 : 55 14 : 53 7 : 52 0 : 39 16 : 32 17 : 31 18 : 19 22 : 18 19 : 13 21 : 13 20 : 13 5 : 11 4 : 11 23 : 10	Top 20: 0 : 82987 1 : 5261 2 : 1898 3 : 1215 4 : 973 5 : 644 6 : 532 7 : 436 8 : 359 9 : 316 11 : 237 10 : 233 12 : 207 13 : 196 14 : 181 17 : 125 16 : 124 19 : 111 15 : 111 18 : 111

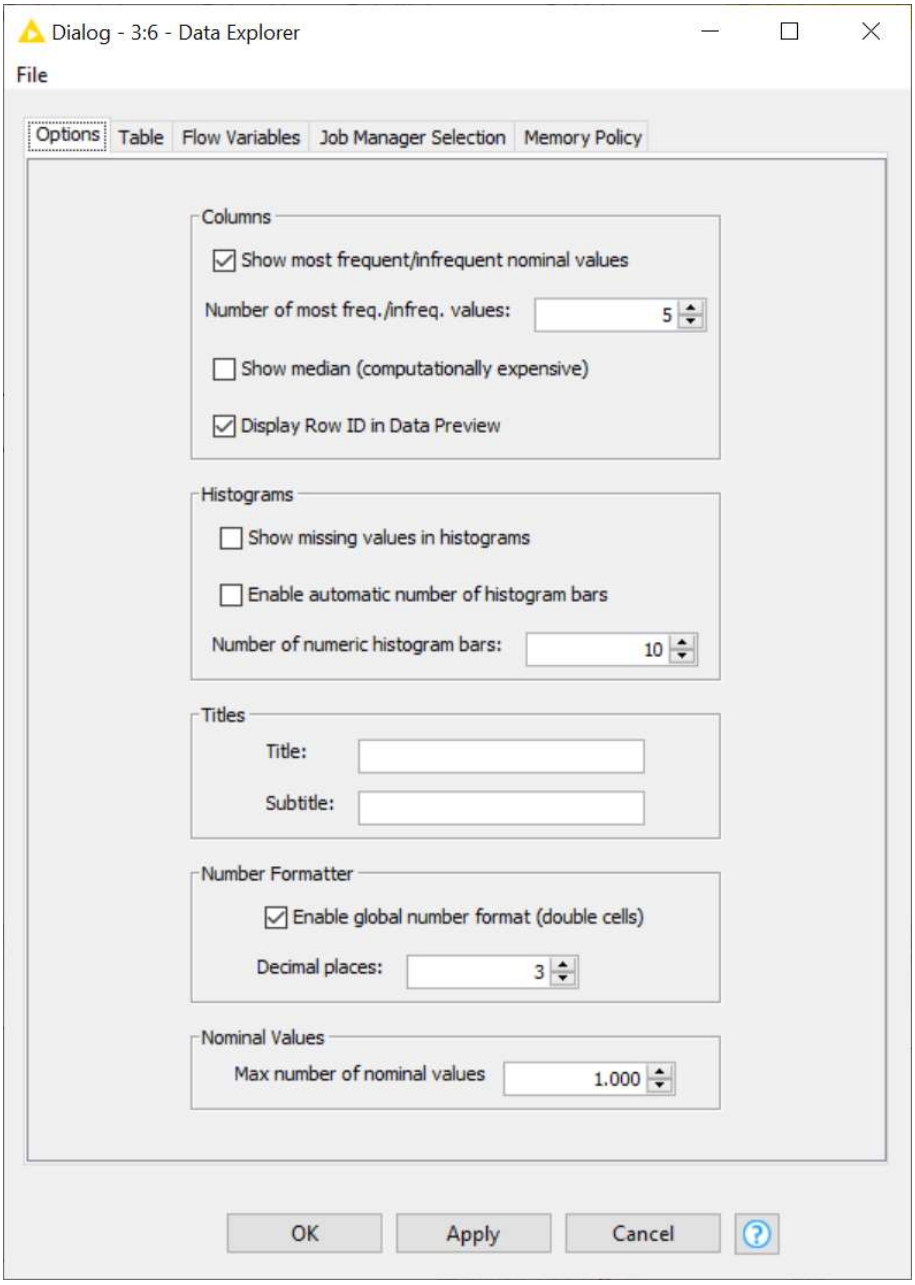
Slika 120. Top/bottom rezultat

Pogreška! Izvor reference nije pronađen. prikazuje sljedeći čvor hodograma, to je **Data Explorer** ili Istraživač podataka. Radi se o čvoru koji, kao i prethodni, daje niz statističkih podataka o varijablama/značajkama.

Slika 121. Čvor Data Explorer

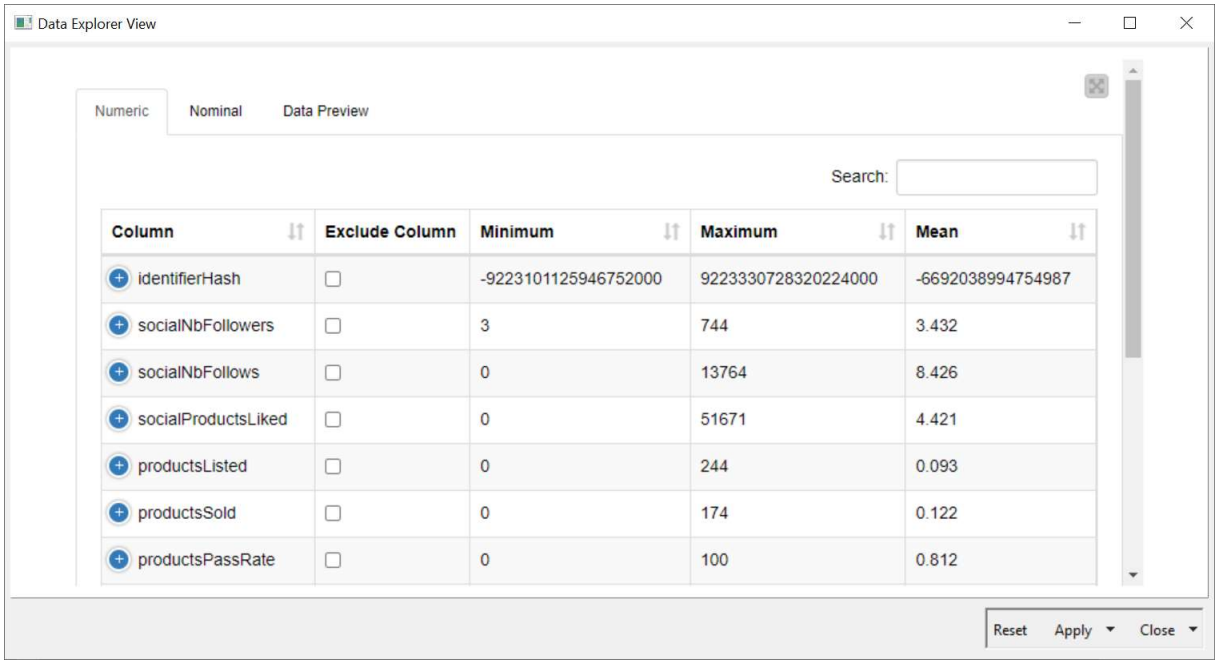
Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Data Explorer**. Prva postavka (*Show most frequent/infrequent nominal values*) omogućuje da kartica *Nominal* u interaktivnom prikazu stvori stupac s n najčešćih nominalnih vrijednosti i drugi stupac s najrjeđim nominalnim vrijednostima, za neki broj *n* koji se odabire u susjednom polju. Prikazane vrijednosti su navedene u padajućem redoslijedu prema učestalosti. Prikaz medijana je sljedeća opcija (*Show median (comutationally expensive)*) uz koju se nalazi upozorenje kako može biti zahtjevna za računalo i samim time računanje medijana za sve stupce može nešto duže trajati. Opcija *Disply Row ID in Data Preview* stvara stupac s identifikatorima redaka. Sljedeća opcija (*Show missing values in histogram*) omogućuje da se na kartici *Nominal* u interaktivnom prikazu prikažu nedostajuće vrijednosti kao dodatne trake u histogramima. Nakon toga je opcija za automatsko prilagođavanje prikaza broja stupaca histograma ovisno o vrijednostima koje se pojavljuju u stupcu. Ta vrijednost može se unijeti i ručno. Ispod toga postoji mogućnost unošenja naslova (*Title*) i podnaslova (*Subtitle*). Na kraju se nalazi opcija za

usklađivanje numeričkih vrijednosti i definiranje broja znamenaka iza decimalnog zareza te najveći broj jedinstvenih vrijednosti koje se uzimaju u obzir u jednom stupcu s kategorijalnim vrijednostima.



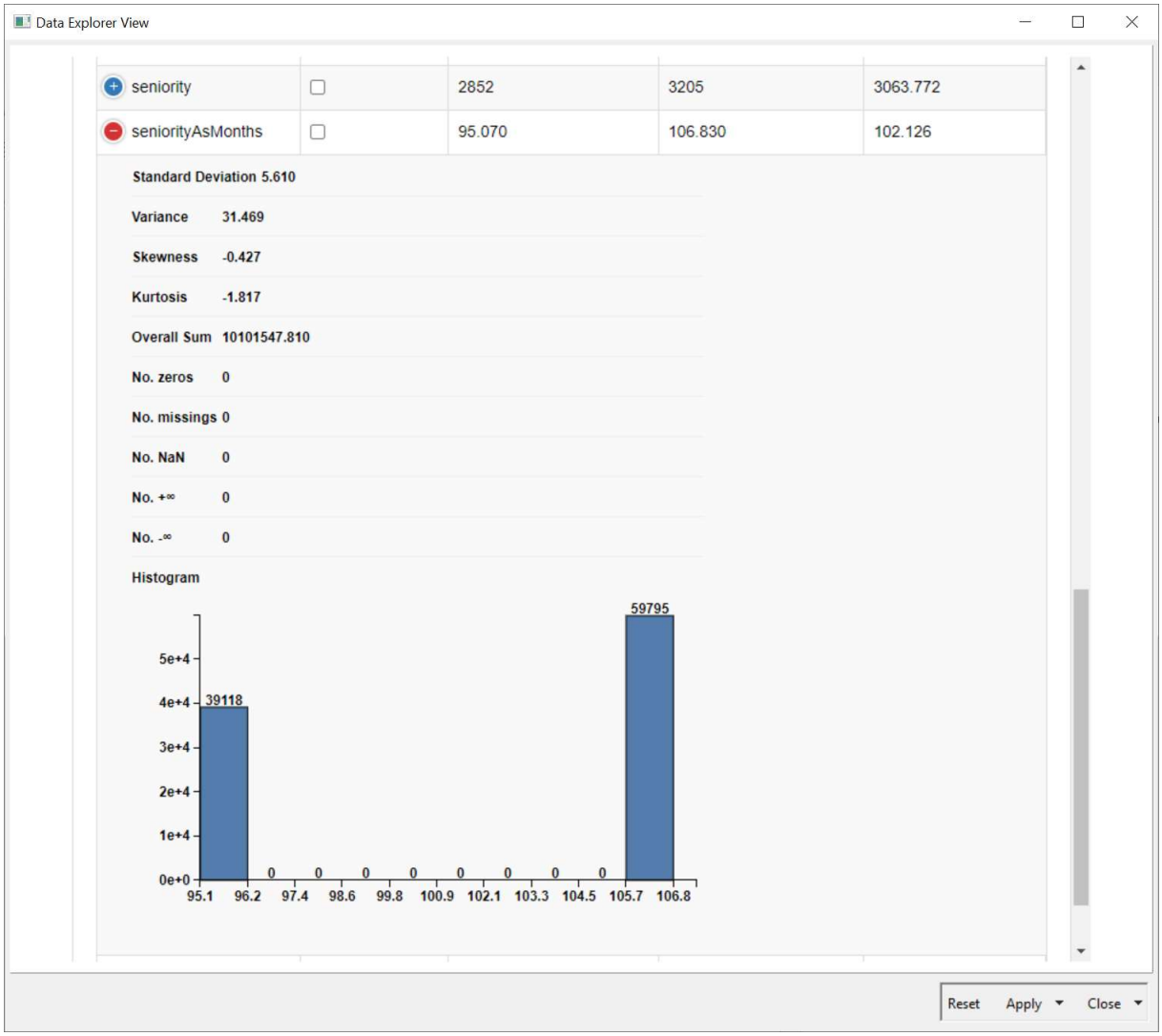
Slika 122. Postavke čvora Data Explorer

Pogreška! Izvor reference nije pronađen. prikazuje rezultate analize podataka čvora **Data Explorer** za karticu *Numeric*. Klikom na plavi krug na početku svakog retka otvorit će se dodatne opcije za izabranu značajku. Zadani prikaz uključuje minimum, maksimum i srednju vrijednost za numeričke podatke.



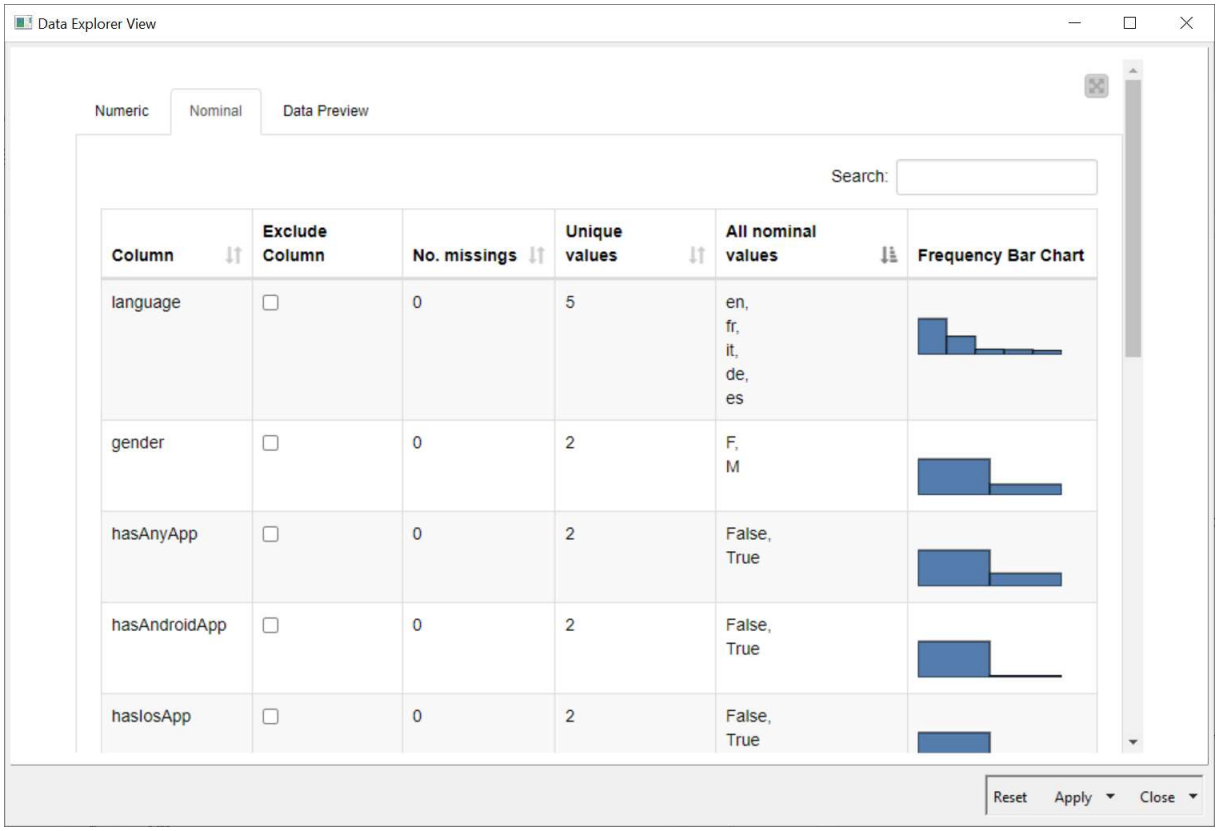
Slika 123. Rezultati čvora Data Exlorer, kartica Numeric

Pogreška! Izvor reference nije pronađen. prikazuje rezultate čvora **Data Explorer** za karticu *Numeric*, s detaljnim prikazom značajke *SeniorityAsMonths* koji označava broj mjeseci od registracije korisnika. Na dnu je histogram,a to je naziv za površinski grafikon predviđen za kontinuirane numeričke varijable, mada se isti grafikon može koristiti i za prikaz distribucije frekvencija kategorijalnih varijabli (Horvat & Mijoč, 2019). Čvor **Data** Explorer zadano prikazuje histogram s deset razreda, odnosno stupaca, koji su dobiveni na način da se razlika najveće i najmanje vrijednosti varijable *SeniorityAsMonths* podijeli u deset razreda iste širine te se nakon toga pobroji koliko vrijednosti pripada u pojedini razred. U primjeru najveći broj vrijednosti varijable *SeniorityAsMonths* (59795) pripada desetom razredu, odnosno stupcu, koji je u granicama od 105,7 i 106,8 mjeseci. Kada se pogleda histogram, vidljivo je kako podaci nisu distribuirani ravnomjerno, odnosno osam od deset razreda su prazni. Ovo su podaci korisnika koji su se registrirali u dva kratka vremenska perioda od oko mjesec dana i to prije osam i nešto manje od devet godina. Osoba koja je postavila podatke na poslužitelja www.kaggle.com ne navodi iz kod razloga je dostupan samo dio podataka korisnika servisa. Ono što je bitno je da se na osnovu ovih podataka ne može kreirati model koji bi predstavljao sve korisnike servisa. Na ovakvom se primjeru prepoznaje vrijednost analize podataka i čvorova koji za to služe jer je ovakvu anomaliju u podacima vrlo teško otkriti analizirajući tablicu sa sirovim podacima jer se radi o skoro 100.000 redova. Uz pomoć histograma mogu se uočiti slične anomalije u podacima.



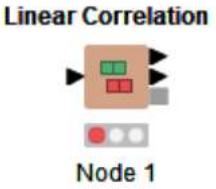
Slika 124. Rezultati čvora Data Exlorer, kartica Numeric, značajka SeniorityAsMonths

Pogreška! Izvor reference nije pronađen. prikazuje drugu karticu rezultata čvora **Data Explorer**. Za razliku od numeričkih stupaca ovdje postoji manje podataka o svakome stupcu. Osim naziva stupca tu je broj praznih ćelija, broj jedinstvenih vrijednosti te na kraju histogram s frekvencijama pojedinih vrijednosti. U tim histogramima na osi X nema numeričkih vrijednosti koje služe za podjelu u razrede, nego su razredi definirani samim kategorijalnim vrijednostima varijable te se prikazuju frekvencije pojavljivanja pojedine vrijednosti.



Slika 125. Rezultati čvora Data Exlorer, kartica Nominal

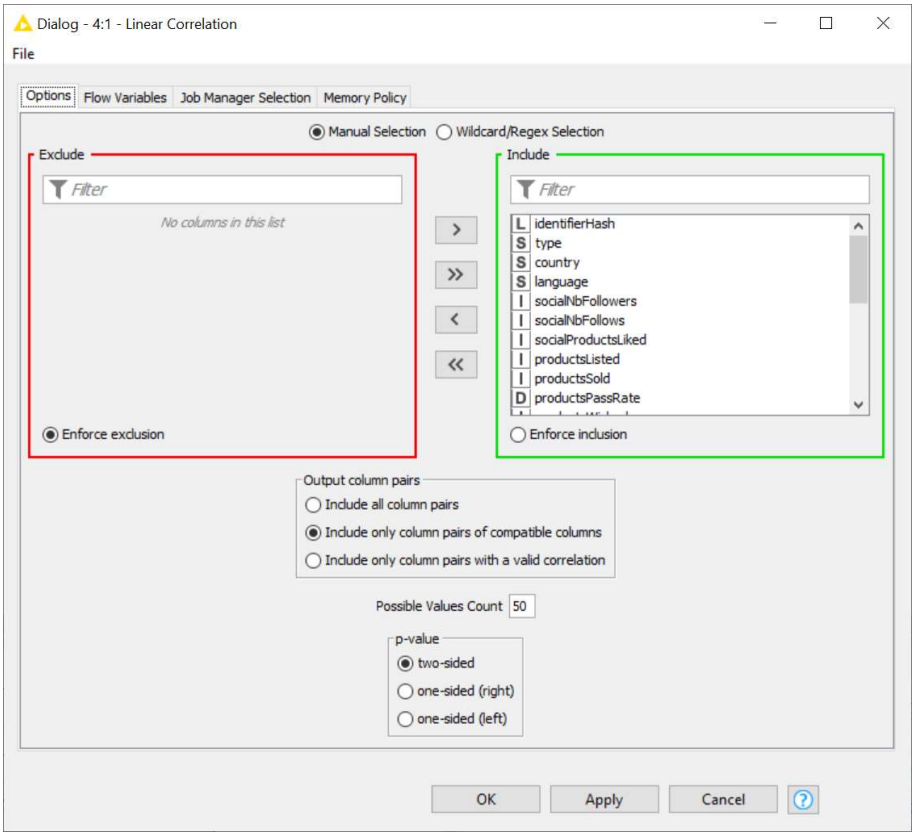
Pogreška! Izvor reference nije pronađen. prikazuje čvor **Linear Correlation** ili Linearna korelacija. On izračunava koeficijent povezanosti za svaki par izabranih stupaca. Koeficijent povezanosti može biti od koristi iz više razloga. Ako između nekih stupaca postoji jaka povezanost, moguće je da ti stupci sadrže istu informaciju prikazanu na drugi način. U tom slučaju jedan od stupaca može biti uklonjen iz daljnje analize i kreiranja modela. Povezanost je korisna kako bi se otkrila povezanost varijabli, odnosno značajki, što doprinosi razumijevanju podataka i pojava koje proučavamo.



Slika 126. Čvor Linear Correlation

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Linear Correlation**. U vrhu dijaloškog okvira biraju se stupci za koje se želi izračunati povezanost koristeći zeleni i crveni okvir. Nakon toga opcija *Output column pairs* nudi tri mogućnosti. Prva je da se povezanost računa za sve stupce, druga je da se povezanost računa samo između kompatibilnih stupaca (numerički-numerički i kategorijalni-kategorijalni) dok se međusobne kombinacije isključuju, dok je treća mogućnost da se povezanost računa samo među stupcima za koje ju je moguće izračunati. Vrijednosti koje nedostaju u stupcima se zanemaruju na način da se za izračun povezanosti između dva stupca uzimaju u obzir samo potpuni redovi. Pri dnu dijaloškog okvira se nalazi polje *Possible Values Count* koje se koristi za testiranje broja jedinstvenih vrijednosti u stupcima za koje se računa povezanost. Ako je broj

jedinstvenih vrijednosti u jednom od stupaca veći od zadane, povezanost se ne računa. Ovaj čvor izračunava korelaciju numeričkih varijabli koristeći Pearsonovu metodu, a za nominalne varijable radi Pearsonov χ^2 test na tablici kontigencije.

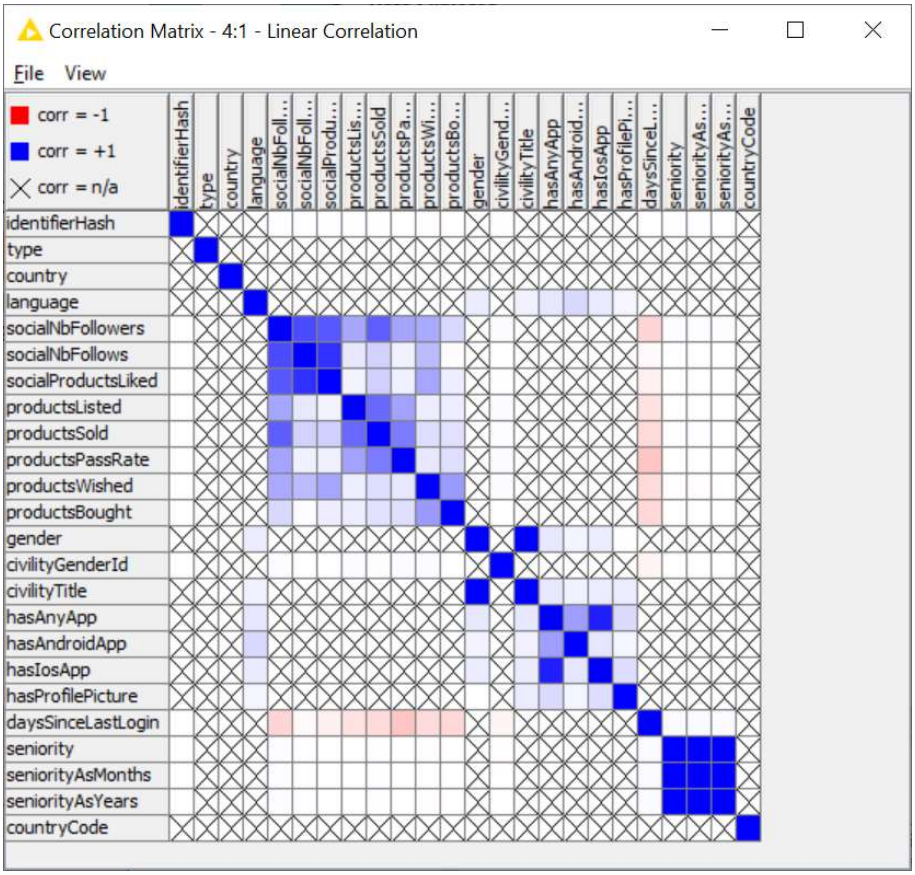


Slika 127. Postavke čvora Linear Correlation

Pogreška! Izvor reference nije pronađen. prikazuje rezultate izračuna koeficijenata povezanosti između izabranih stupaca prikazane grafički. Tamno plava ukazuje na korelaciju +1 što je najveća moguća povezanost između dvije varijable. Tamno crvena boja prikazuje negativnu korelaciju koja ima vrijednost -1. Bijela boja prikazuje korelaciju koja ima vrijednost 0, odnosno u tom slučaju nema povezanosti između varijabli. Polja u kojima je znak X su na sjecištu stupaca za koje se povezanosti ne računaju zbog postavki.

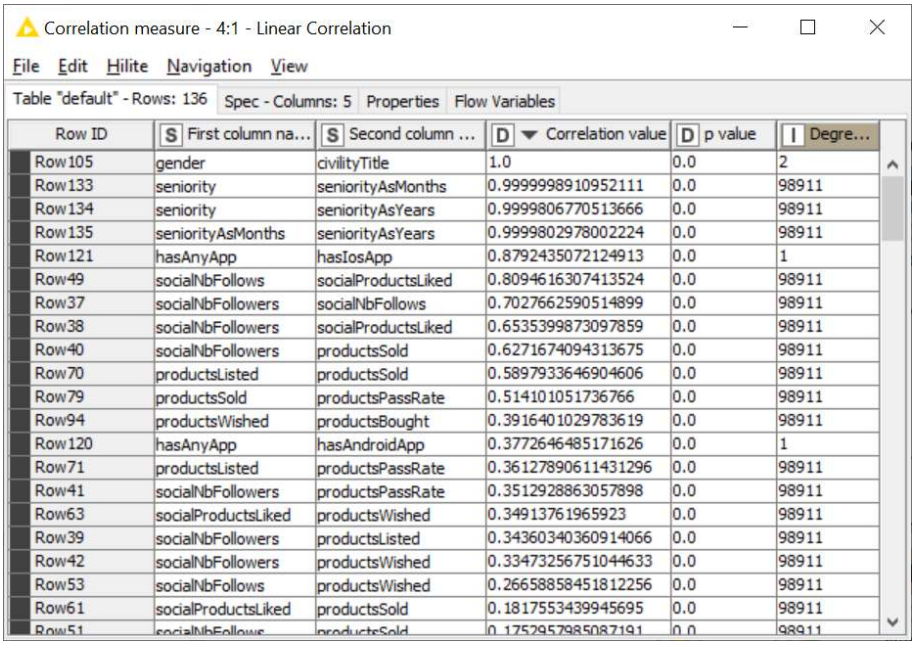
U grafičkom prikazu na obje osi su prisutne sve varijable. Na glavnoj dijagonali su povezanosti jednake 1 jer su to vrijednosti povezanosti varijable sa samom sobom. Iz tog razloga na glavnoj dijagonali od gornjeg lijevog do donjeg desnog kuta postoji plava crta.

Grafički prikaz je izuzetno koristan jer se jednostavno mogu uočiti povezanosti između varijabli. U konkretnom primjeru vidi se jaka povezanost između varijabli *seniority, seniortyAsMonths i seniorityAsYears.* To je očekivano jer se vremenski period u ova tri stupca prikazuje na različite načine, ali radi se o istim vremenskim periodima. Prikazani su kao dani, mjeseci i godine. Jaka povezanost se još može uočiti i kod varijabli vezanih uz aplikacije za mobilne platforme, za spol i prefikse koji se koriste za označavanje spola. Zanimljivo je kako postoji i slaba negativna povezanost vezana uz stupac *daysSinceLastLogin* i nekoliko stupaca vezanih uz prodaju i društvene mreže. Da bi vidjeli koliko iznose te povezanosti, koristit će se još jedan prikaz koji ovaj čvor nudi.



Slika 128. Rezultati čvora Linear Correlation, izbornik View: Correlation Matrix

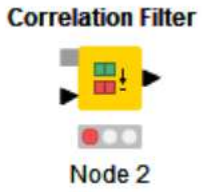
Pogreška! Izvor reference nije pronađen. također prikazuje povezanosti među parovima stupaca, ali je povezanost izražena numerički kao vrijednost između -1 i 1. S obzirom da ovaj prikaz omogućuje sortiranje po stupcima, klikom na zaglavlje tablice četvrtog stupca (*Correlation Value*) koeficijenti povezanosti se mogu sortirati od najveće do najmanje vrijednosti.



Row ID	First column na...	Second column ...	Correlation value	p value	Degre...
Row105	gender	civilityTitle	1.0	0.0	2
Row133	seniority	seniorityAsMonths	0.9999998910952111	0.0	98911
Row134	seniority	seniorityAsYears	0.9999806770513666	0.0	98911
Row135	seniorityAsMonths	seniorityAsYears	0.9999802978002224	0.0	98911
Row121	hasAnyApp	hasIosApp	0.8792435072124913	0.0	1
Row49	socialNbFollows	socialProductsLiked	0.8094616307413524	0.0	98911
Row37	socialNbFollowers	socialNbFollows	0.7027662590514899	0.0	98911
Row38	socialNbFollowers	socialProductsLiked	0.6535399873097859	0.0	98911
Row40	socialNbFollowers	productsSold	0.6271674094313675	0.0	98911
Row70	productsListed	productsSold	0.5897933646904606	0.0	98911
Row79	productsSold	productsPassRate	0.514101051736766	0.0	98911
Row94	productsWished	productsBought	0.3916401029783619	0.0	98911
Row120	hasAnyApp	hasAndroidApp	0.3772646485171626	0.0	1
Row71	productsListed	productsPassRate	0.36127890611431296	0.0	98911
Row41	socialNbFollowers	productsPassRate	0.3512928863057898	0.0	98911
Row63	socialProductsLiked	productsWished	0.34913761965923	0.0	98911
Row39	socialNbFollowers	productsListed	0.34360340360914066	0.0	98911
Row42	socialNbFollowers	productsWished	0.33473256751044633	0.0	98911
Row53	socialNbFollows	productsWished	0.26658858451812256	0.0	98911
Row61	socialProductsLiked	productsSold	0.1817553439945695	0.0	98911
Row51	socialNbFollows	productsSold	0.1752957985087191	0.0	98911

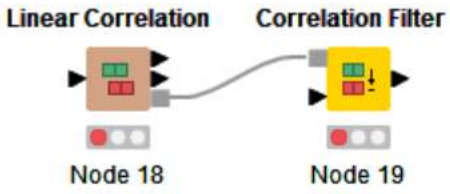
Slika 129. Rezultati čvora Linear Correlation, izbornik Correlation Measure

Pogreška! Izvor reference nije pronađen. prikazuje čvor **Correlation Filter** ili Filter korelacije. Taj čvor se koristi u kombinaciji s čvorom **Linear Correlation** jer mu je potreban model na osnovu kojeg će se filtriranje provesti. Potrebno je uočiti kako čvor **Linear Correlation** ima izlazni priključak u obliku sivog kvadratića koji se treba spojiti s ulaznim priključkom čvora **Correlation Filter.** Tek nakon toga se mogu dovesti podaci na drugi ulazni priključak čvora **Correlation Filter**.



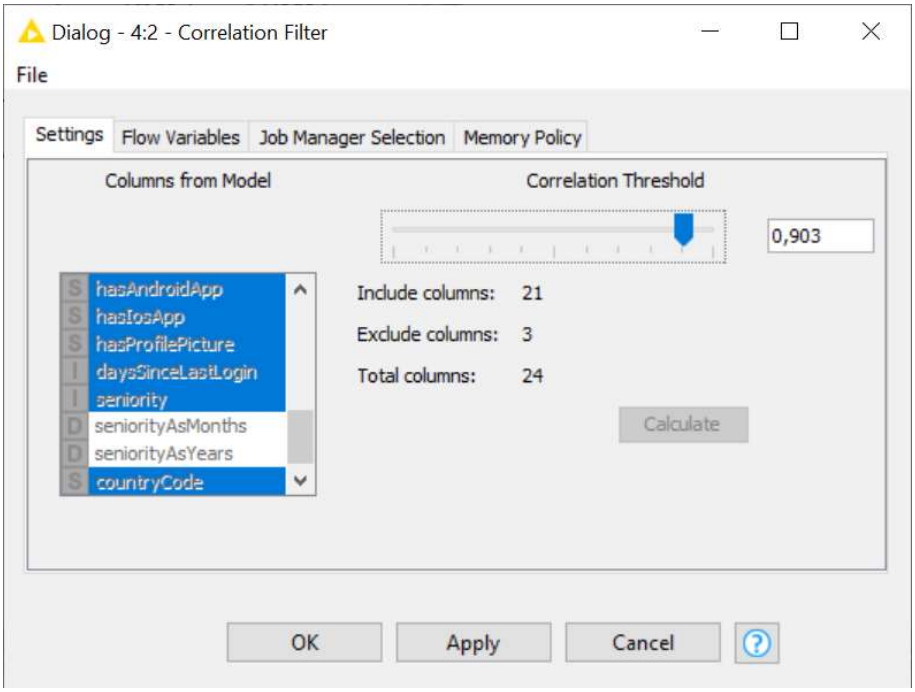
Slika 130. Čvor Correlation Filter

Pogreška! Izvor reference nije pronađen. prikazuje spoj između čvorova **Linear Correlation** i **Correlation Filter**.



Slika 131. Spojeni čvorovi Linear Correlation i Correlation Filter

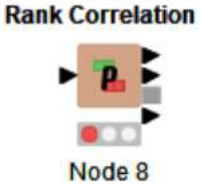
Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Correlation Filter**. U njemu se zadaje prag povezanosti, odnosno vrijednost iznad koje će se stupci odbacivati. Klikom na gumb *Calculate* računa se koliko stupaca će biti odbačeno. S obzirom da je povezanost vezana za dva stupca, pitanje je koji stupac odbaciti. Za svaki stupac u korelacijskom modelu određuje se broj koreliranih stupaca s obzirom na vrijednost praga koji se zadaje. Stupac s najviše koreliranih stupaca se propušta dalje, a svi ostali stupci kod kojih je povezanost veća od navedenog praga se odbacuju. Ovaj postupak se ponavlja sve dok je povezanost veća od definiranog praga.



Slika 132. Postavke čvora Correlation Filter

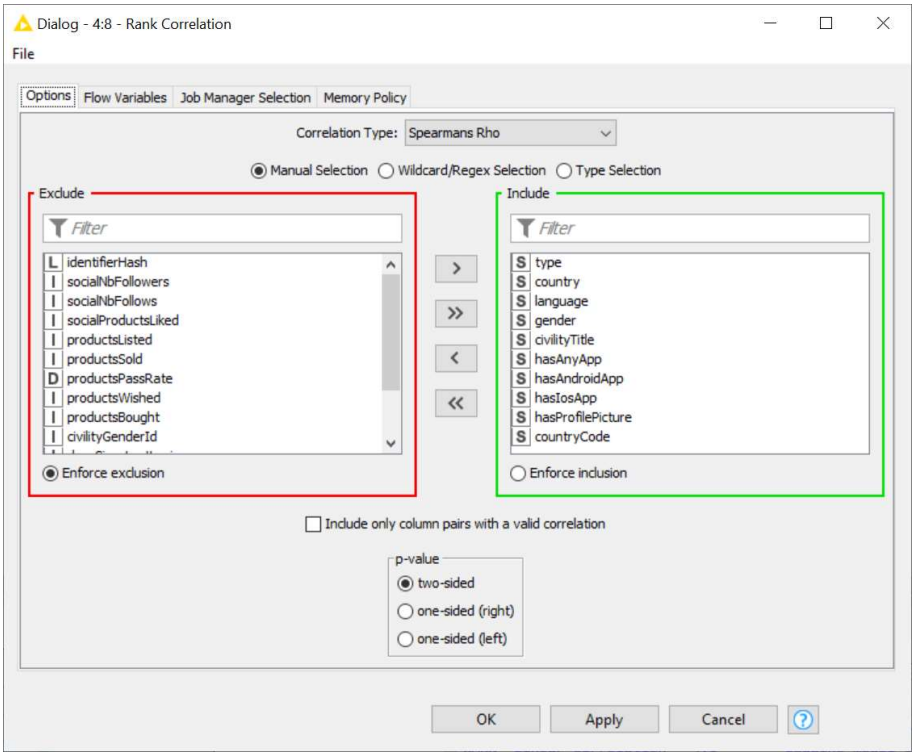
Nakon filtriranja na izlaznom priključku čvora **Correlation Filter** dostupni su filtrirani podaci bez odbačenih stupaca.

Pogreška! Izvor reference nije pronađen. prikazuje čvor **Rank Correlation** ili Korelacija ranga. Njegova funkcija je ista kao i funkcija čvora **Linear Correlation** s tom razlikom da je ova povezanost namijenjena kategorijalnim varijablama kod kojih se zadano primjenjuje *Spearmanova* povezanost. Osim nje na raspolaganju su sljedeći koeficijenti povezanosti: *Kendalls Tau A, Kendalls Tau B* i *Goodmans and Kruskal's Gamma*. Izbor koeficijenta povezanosti moguć je u postavkama čvora.



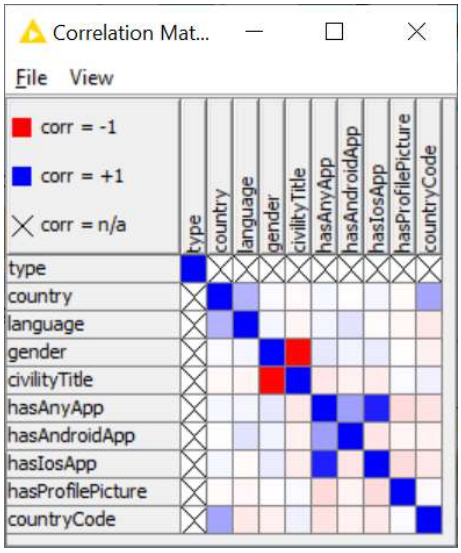
Slika 133. Čvor Rank Correlation

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Rank Correlation**. U vrhu se bira željeni koeficijent povezanosti, a nakon toga stupci nad kojima se povezanost računa.



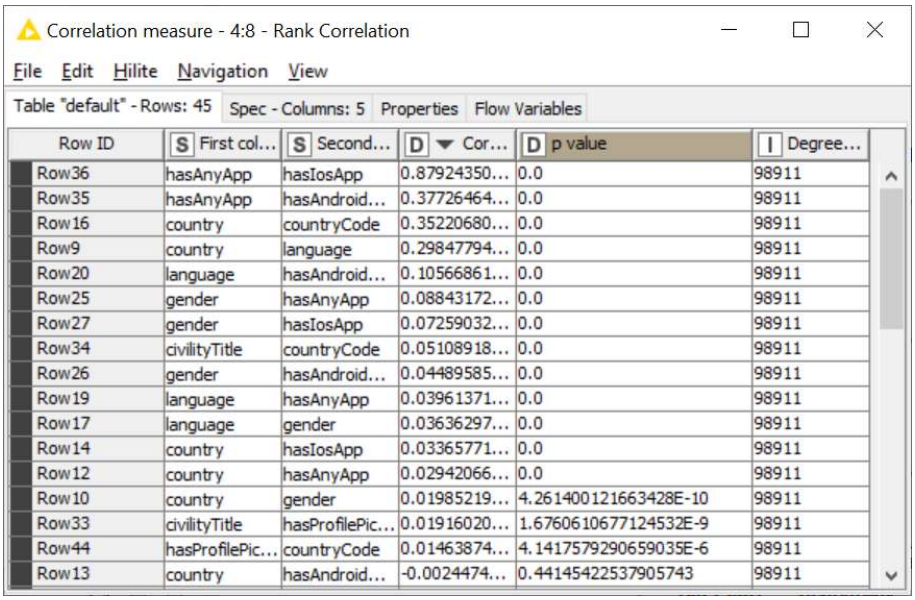
Slika 134. Postavke čvora Rank Correlation

Pogreška! Izvor reference nije pronađen. prikazuje matricu povezanosti uz grafički prikaz vrijednosti povezanosti između stupaca. U prikazu je bitno uočiti kako postoji jaka negativna povezanost između spola i prefiksa koji se dodjeljuju pripadnicima određenog spola. Razlog jake negativne povezanosti leži u tome što su nominalne vrijednosti za spol (F, M) rangirane drugačije od prefiksa (mr, mrs, miss). Npr. ako je oznaci F dodijeljen broj 0, a oznaci M dodijeljen broj 1, dok su prefiksima *mr, mrs* i *miss* dodijeljeni rangovi 0,1 i 2, dobiva se situacija da je za broj 0 u stupcu spola, vrijednost u stupcu prefiksa uvijek 1 ili 2. S druge strane za broj 1 u stupcu spola, vrijednost u stupcu prefiksa je uvijek 0. Na taj način dobiva se maksimalna negativna povezanost.



Slika 135. Rezultati čvora Rank Correlation, matrica povezanosti

Pogreška! Izvor reference nije pronađen. prikazuje tablične vrijednosti povezanosti među stupcima koje su sortirane po vrijednosti povezanosti. Najjača povezanost postoji između značajki *hasAnyApp i hasIosApp* i iznosi 0,879. Za prikaz najjače negativne povezanosti koja iznosi -0,98 između stupaca *gender* i *civilityType* potrebno je sortirati stupac s vrijednostima rastuće povezanosti .



Row ID	First col...	Second...	Cor...	p value	Degree...
Row36	hasAnyApp	hasIosApp	0.87924350...	0.0	98911
Row35	hasAnyApp	hasAndroid...	0.37726464...	0.0	98911
Row16	country	countryCode	0.35220680...	0.0	98911
Row9	country	language	0.29847794...	0.0	98911
Row20	language	hasAndroid...	0.10566861...	0.0	98911
Row25	gender	hasAnyApp	0.08843172...	0.0	98911
Row27	gender	hasIosApp	0.07259032...	0.0	98911
Row34	civilityTitle	countryCode	0.05108918...	0.0	98911
Row26	gender	hasAndroid...	0.04489585...	0.0	98911
Row19	language	hasAnyApp	0.03961371...	0.0	98911
Row17	language	gender	0.03636297...	0.0	98911
Row14	country	hasIosApp	0.03365771...	0.0	98911
Row12	country	hasAnyApp	0.02942066...	0.0	98911
Row10	country	gender	0.01985219...	4.261400121663428E-10	98911
Row33	civilityTitle	hasProfilePic...	0.01916020...	1.6760610677124532E-9	98911
Row44	hasProfilePic...	countryCode	0.01463874...	4.1417579290659035E-6	98911
Row13	country	hasAndroid...	-0.0024474...	0.44145422537905743	98911

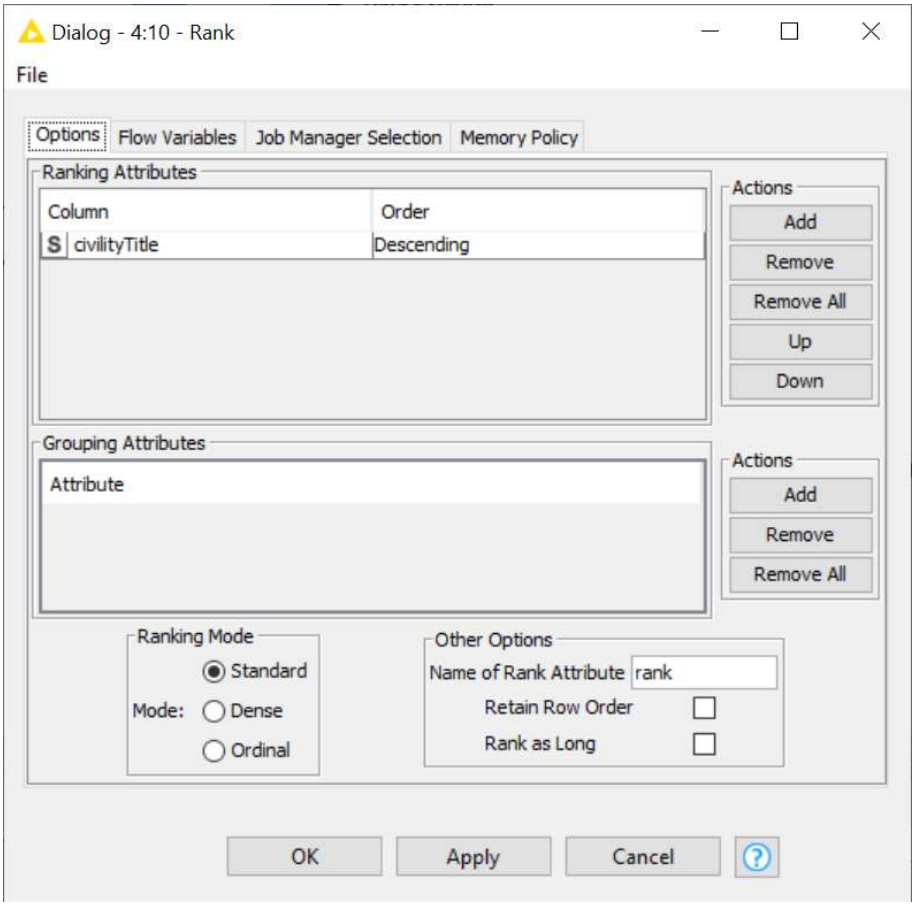
Slika 136. Rezultati čvora Rank Correlation, tablica s povezanostima

Ako se želi prikazati vrijednost povezanosti kao pozitivnu dovoljno je presložiti podatke u jednom od stupaca u čvoru **Rank Correlation**. To je moguće koristeći čvor **Rank** ili Rang. **Pogreška! Izvor reference nije pronađen.** prikazuje taj čvor.

Slika 137. Čvor Rank

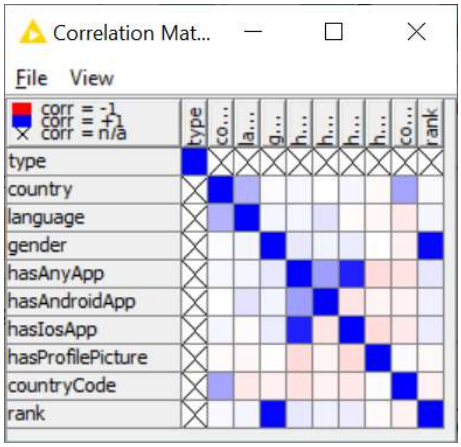
Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Rank**. Čvor za svaku grupu izračunava pojedinačni rang na temelju odabranih atributa rangiranja i načina rangiranja. Potrebno je navesti barem jedan atribut na temelju kojeg bi se rangiranje trebalo provesti. Grupe su određene jedinstvenim kombinacijama vrijednosti atributa grupiranja. Ako atribut grupiranja nije naveden, jedno rangiranje će se izračunati za cijelu tablicu. U postavkama se definiraju atributi rangiranja i grupiranja, kao i mod rangiranja. Izlazna tablica uključuje još jedan stupac koji je numeričkog tipa (*Integer*) u kojem se nalazi vrijednost ranga.

U primjeru je dovoljno izabrati stupac *civilityTitle* i sortirati ga padajući. Preporuka je pogledati podatke u dodanom stupcu. S obzirom da je stupac *civilityTitle* uključivao samo tri tekstualne vrijednosti (mrs, mr i miss), rangovi su samo tri broja i to 1, 75.685 i 98.477. Radi se o početnim ćelijama svake od tri navedene tekstualne vrijednosti, kad su sortirane na izabran način.



Slika 138. Postavke čvora Rank

Nakon čvora **Rank** ubačen je čvor **Column Filter** koji je obrađen u prethodnom tekstu, a služi za filtriranje stupaca. Iz tablice je obrisan stupac *CilivityTitle*. **Pogreška! Izvor reference nije pronađen.** prikazuje rezultat čvora **Rank Correlation** nakon izmjena. Podaci iz stupca *CivilityTitle* sada se nalaze u stupcu *rank*, samo su presloženi na način da je čvor **Rank Correlation** prepoznao korelaciju između stupaca *gender* i *rank* kao pozitivnu.



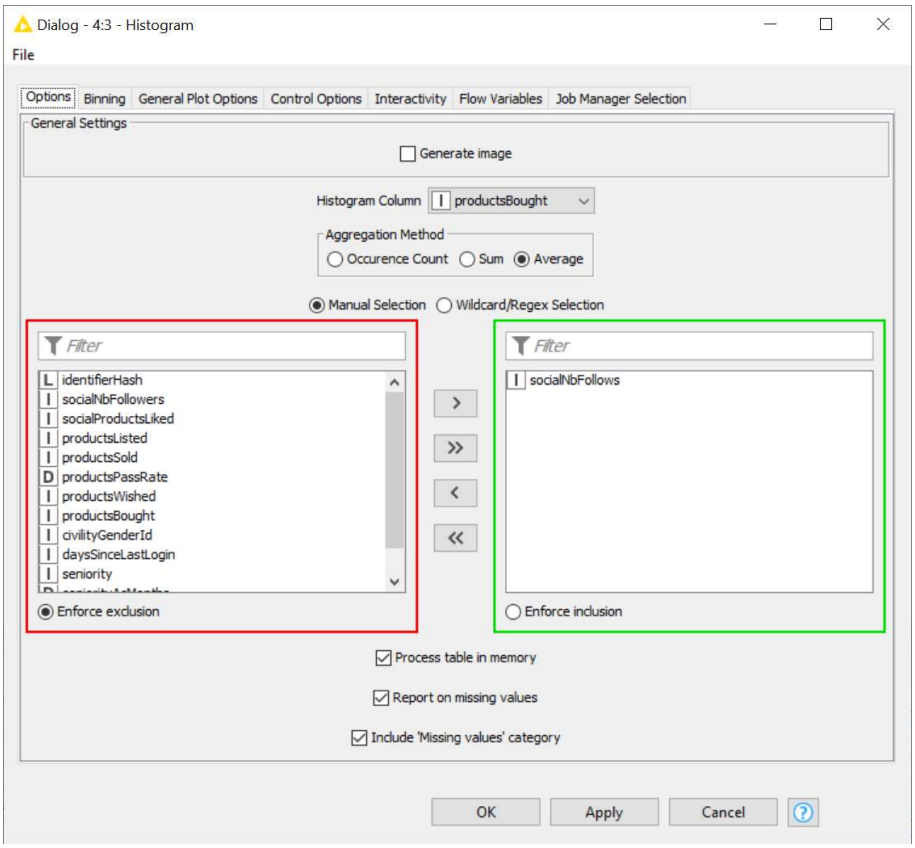
Slika 139. Rezultati čvora Rank Correlation, matrica povezanosti nakon izmjena

Pogreška! Izvor reference nije pronađen. na kraju poglavlja prikazuje izmijenjen hodogram s umetnuta dodatna dva čvora.

Pogreška! Izvor reference nije pronađen. prikazuje čvor **Histogram** koji se i na hrvatskom jeziku naziva Histogram. Taj čvor služi za grafički prikaz nekog stupca u obliku histograma.

Slika 140. Čvor Histogram

Pogreška! Izvor reference nije pronađen. prikazuje prvu karticu dijaloškog okvira s postavkama čvora **Histogram**. Opcija *Generate Image* uključuje se ako se želi histogram izvesti kao sliku. U slučaju da se samo želi pogledati histogram, ta opcija se ne uključuje. Opcija *Histogram Column* omogućuje izbor stupca čije vrijednosti se prikazuju na osi X, odnosno dijele se u zadani broj stupaca. Opcija *Aggregation Method* omogućuje izbor načina prebrojavanja, pri čemu se može brojati, zbrajati i računati prosjek. Nakon toga je moguće izabrati stupac na koji se primjenjuje zadani način prebrojavanja. Sučelje je poznato, koriste se zeleni i crveni pravokutnici. Ispod toga su još tri postavke vezane uz obradu u memoriji računala, izvještavanje o praznim ćelijama i kreiranje kategorije s praznim ćelijama.

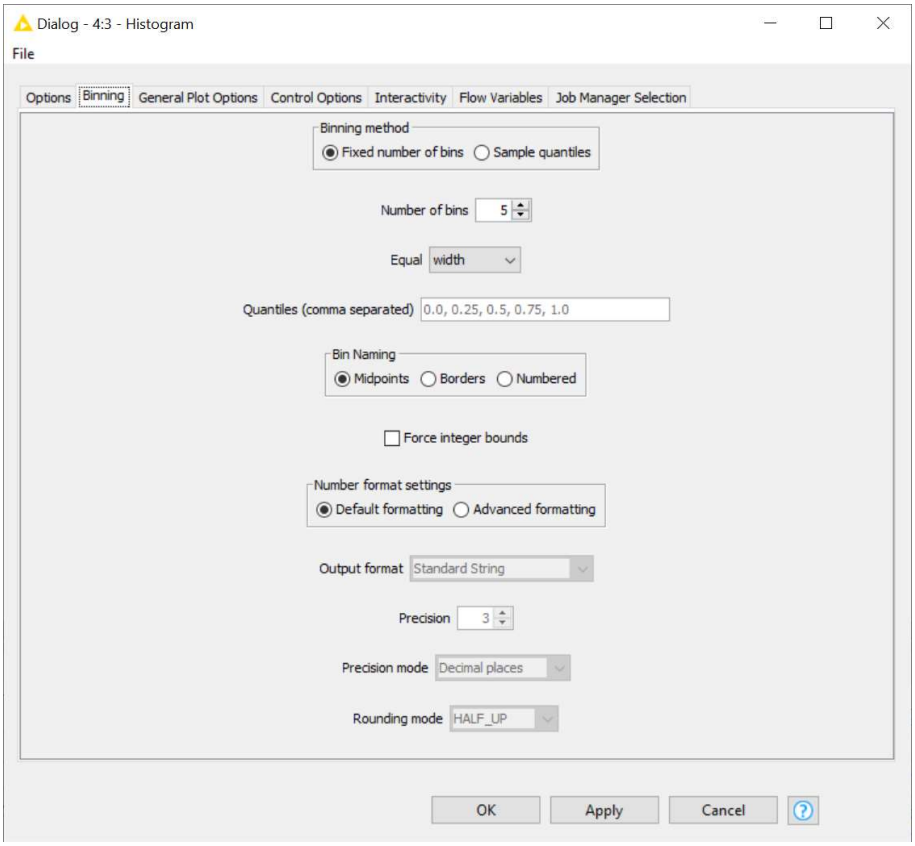


Slika 141. Postavke čvora Histogram

Pogreška! Izvor reference nije pronađen. prikazuje drugu karticu dijaloškog okvira postavki čvora **Histogram** s nazivom *Binning*. Na početku se odabire metoda grupiranja, a ponuđene su dvije: *Fixed*

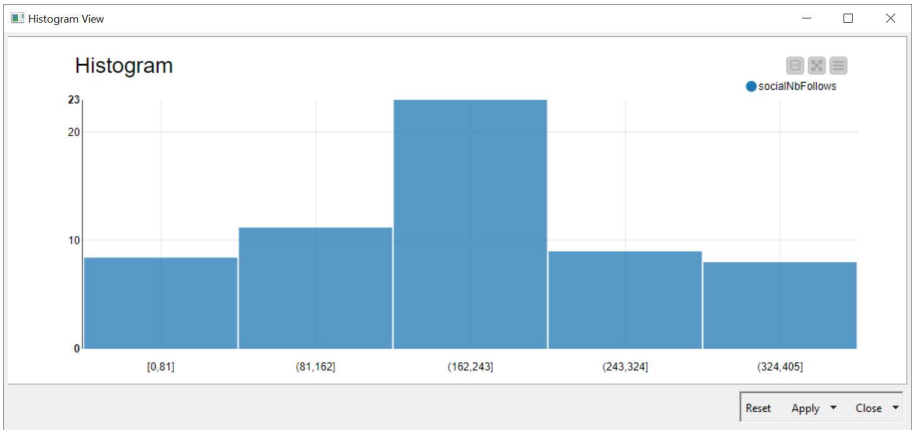
number of bins i *Sample quantiles*. Prva metoda omogućuje definiranje fiksnog broja razreda/grupa, dok druga omogućuje korištenje kvantila za izradu razreda/grupa. Ako je izabrana prva opcija, tada je potrebno definirati broj razreda/grupa (*Number of bins*) i izabrati jesu li razredi/grupe jednake širine (*Equal width*) ili jednakih učestalosti (*Equal frequencies*). U slučaju izbora druge opcije, mogu se definirati granice kvantila.

Nakon izbora metode spajanja bira se način imenovanja razreda/grupa. Ponuđene su tri opcije: točke za oznake koje pokazuju sredinu intervala, granice za oznake koristeći „(a,b]"' zapis koji prikazuje granice razreda i nazivi za razrede označene rastućim cijelim brojem s prefiksom „Bin". Osim toga dostupne su opcije vezane uz zaokruživanje i formatiranje brojeva.



Slika 142. Postavke čvora Histogram

Pogreška! Izvor reference nije pronađen. prikazuje histogram sa zadanim postavkama. Na osi X su brojevi kupljenih predmeta od strane korisnika grupirani u pet razreda/grupa pri čemu se vide granice, dok je na okomitoj osi prosječan broj osoba koje korisnik prati na društvenim mrežama.

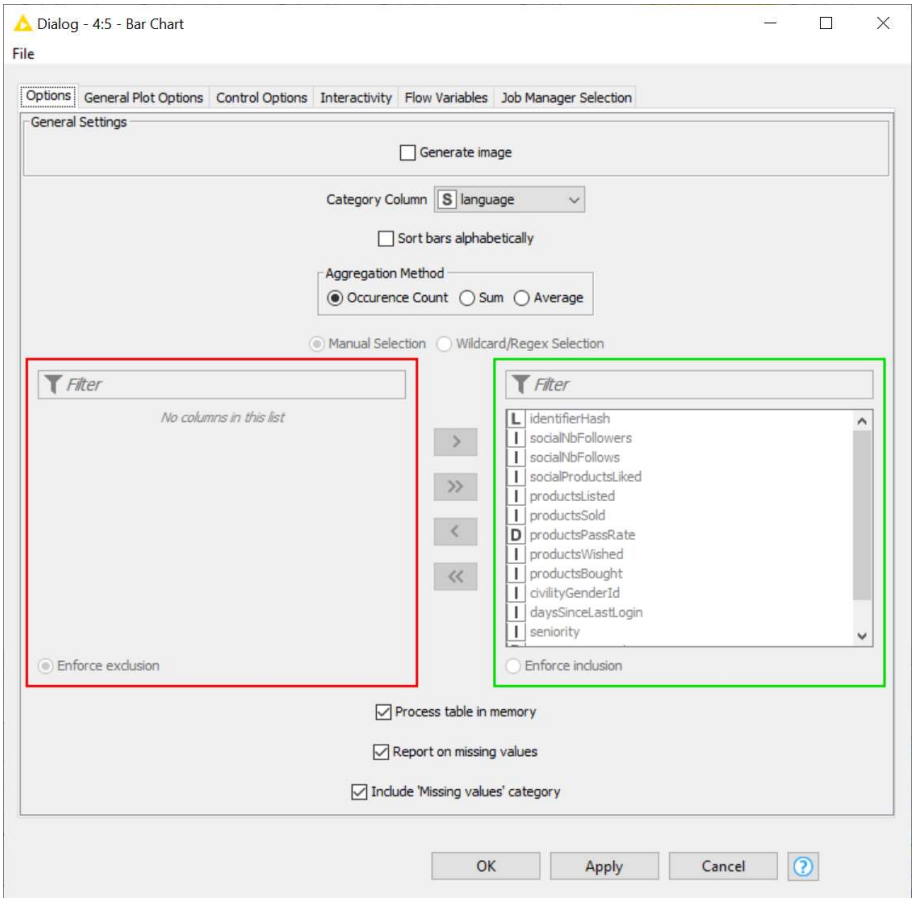


Slika 143. Izgled histograma

Pogreška! Izvor reference nije pronađen. prikazuje čvor **Bar Chart** ili Trakasti grafikon, to je posljednji čvor na hodogramu u ovome poglavlju koje se bavi analizom podataka. Radi se o čvoru koji služi za izradu stupčastog grafikona.

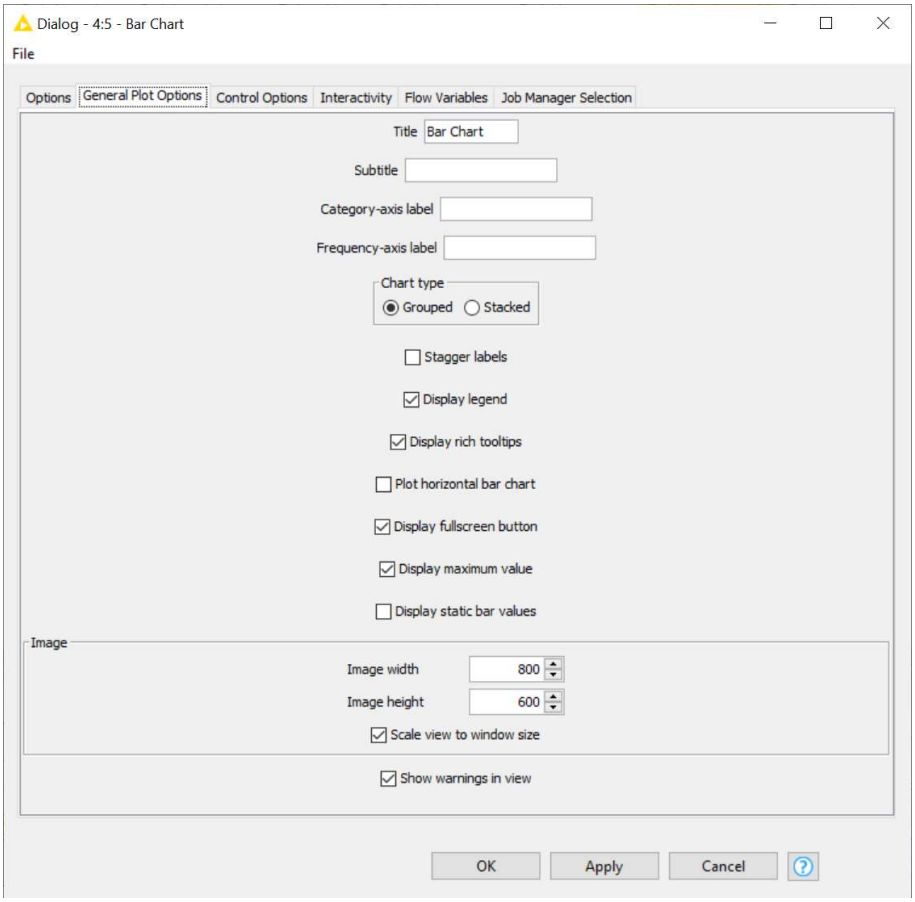
Slika 144. Čvor Bar Chart

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Bar Chart** i to karticu *Options*. Tu se može uključiti izvoz slike, ako se ju želi izvesti kao datoteku u nekom od podržanih formata. Ispod toga može se izabrati kategorijalna vrijednost koja se želi prikazati stupčastim grafikonom, a nakon toga se odabire metoda prebrojavanja pri čemu je ponuđeno brojanje, zbrajanje i računanje prosjeka. Na samom dnu dijaloškog okvira nalaze se postavke vezane uz obradu podataka u memoriji te izvještavanje i kategoriziranje praznih ćelija.



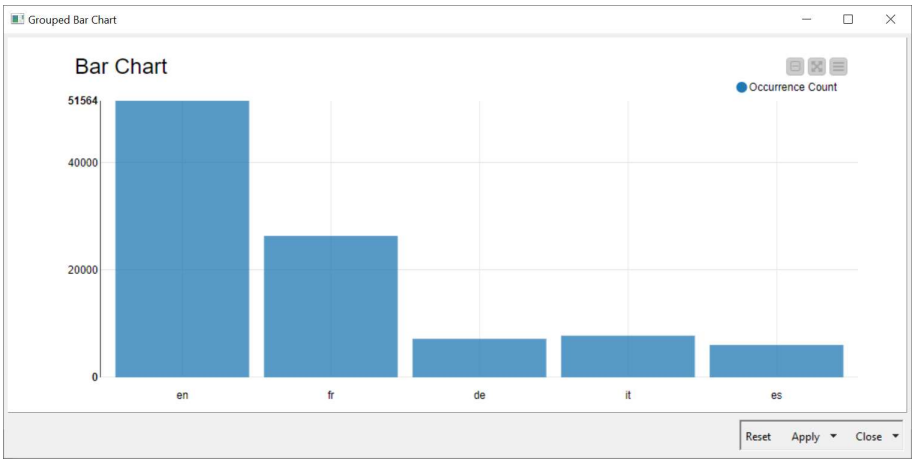
Slika 145. Postavke čvora Bar Chart, kartica Options

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Bar Chart** i to karticu *General Plot Options*. Većina opcija koje se nalaze na toj kartici vezane su uz izgled grafikona. Od naslova i podnaslova, preko natpisa na osima pa do same razlučivosti slike koju čvor može generirati. Jedna od opcija je i izrada vodoravnog stupčastog grafikona.



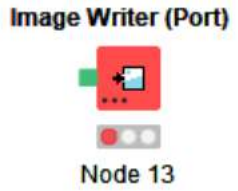
Slika 146. Postavke čvora Bar Chart, kartica General Plot Options

Pogreška! Izvor reference nije pronađen. prikazuje stupčasti grafikon koji je rezultat unesenih postavki. Prikazan je broj korisnika s obzirom na jezik koji preferiraju. Vidi se da su na osi X kratice za engleski, francuski, njemački, talijanski i španjolski jezik te da najveći broj korisnika preferira engleski jezik.



Slika 147. Izgled stupčastog grafikona

Ako se želi ovaj stupčasti grafikon spremiti kao sliku, koristi se čvor **Image Writer (Port)**. **Pogreška! Izvor reference nije pronađen.** prikazuje taj čvor.

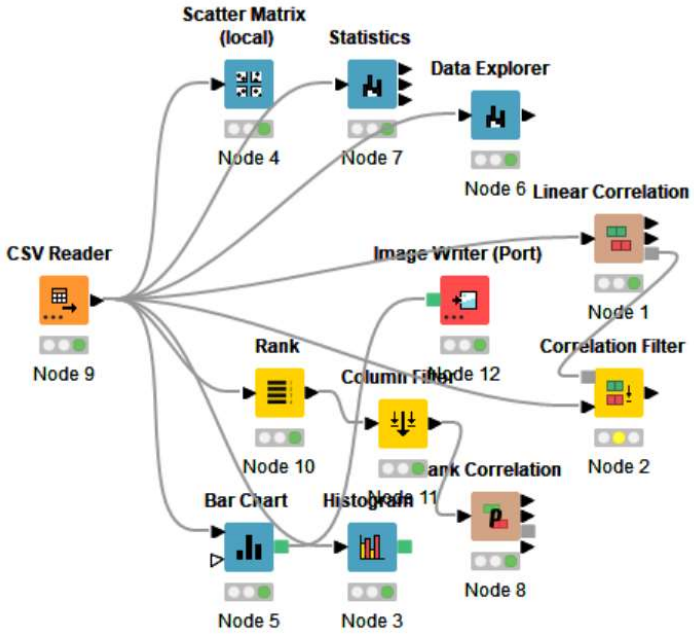


Slika 148. Čvor Image Writer (Port)

Pogreška! Izvor reference nije pronađen. prikazuje postavke čvora **Image Writer (Port)** ili Zapisivač slika**.** Ključni detalji su definiranje imena datoteke, formata datoteke i lokacije gdje se datoteka sprema u slučaju da je spremamo na lokalno računalo. Ako se izabere neka druga lokacija, postavke su drugačije. **Pogreška! Izvor reference nije pronađen.** prikazuje način povezivanja čvorova **Bar Chart** i **Image Writer (Port)**. Bitno je povezati priključke čvorova pri čemu se radi o zelenim kvadratićima koji označavaju kako se tom vezom prenosi slikovna datoteka.

Slika 149. Postavke čvora Image Writer (Port)

Pogreška! Izvor reference nije pronađen. prikazuje konačan izgled hodograma sa svim izmjenama.



Slika 150. Konačan izgled hodograma

6. Linearna regresija

Regresijska analiza je jednostavna i često korištena tehnika u strojnom učenju. Temelji se na povezanosti nezavisne (x) i zavisne varijable (y), pri čemu se promjenom vrijednosti nezavisne varijable mijenja vrijednost zavisne. Sama regresijska analiza ne objašnjava povezanost nezavisne i zavisne varijable na što treba obratiti pažnju. Regresija se može podijeliti na linearnu i nelinearnu, a osim toga na jednostavnu i višestruku. Ovo poglavlje pokriva jednostavnu i višestruku linearnu regresiju. Primjer za jednostavnu linearnu regresiju je povezanost trajanja punjenja spremnika vode i napunjenosti spremnika ili povezanost broja noćenja u mjesecu i prihoda lokalne zajednice od turističke takse. Linearna regresijska analiza omogućuje da se kvantitativno povežu različite varijable (Brownlee, 2016).

6.1. Jednostavna linearna regresija

Regresija je je jedna od najčešće korištenih tehnika u strojnom učenju. Koristi se kada se želi predvidjeti vrijednost zavisne varijable uz pomoć dostupnih poznatih podataka. Najjednostavniji regresijski model je jednostavna linearna regresija, a može se izraziti kao (Buglear, 2010):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Zavisna varijabla označena je slovom y i nju se želi predvidjeti. Nezavisna varijabla označena je slovom x i ta vrijednost je poznata. β_0 i β_1 su parametri populacije, dok je ε pogreška modela (Su, et al., 2012). Bitno je uočiti kako je u prethodnim poglavljima već djelomično obrađena linearna regresija, ali su se koristili parametri a i b, a pogreška modela se nije spominjala uobličena u varijablu. Ima više načina kako odrediti parametre β_0 i β_1, a jedan od njih je metoda najmanjih kvadrata. Ne ulazeći u matematički izvod same metode, formule za izračun β_0 i β_1 tom metodom glase (Buglear, 2010):

$$\beta_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

S obzirom da su dostupni podaci iz primjera s osvježavajućim pićima, mogu se uvrstiti u formule. **Pogreška! Izvor reference nije pronađen.** sadrži dostupne podatke.

Tablica 10. Podaci iz primjera s osvježavajućim pićima

Datum	Broj noćenja	Prodano bočica
1.7.2022	1344	67
2.7.2022	1356	64
3.7.2022	1355	76
4.7.2022	1332	59
5.7.2022	1367	68

Nezavisna varijabla je broj noćenja (x), a zavisna broj prodanih bočica (y). Prije uvrštavanja podataka u formule, izračunat će se sljedeće:

ZBROJ	REZULTAT
$\sum x_i^2$	$=1344^2+1356^2+1355^2+1332^2+1367^2=1806336+1838736+1836025+1774224+1868689=$ **9124010**

$\sum y_i$	=67+64+76+59+68= **334**
$\sum x_i$	=1344+1356+1355+1332+1367= **6754**
$\sum x_i y_i$	=1344*67+1356*64+1355*76+1332*59+1367*68= **451356**
$(\sum x_i)^2$	=(1344+1356+1355+1332+1367)^2=6754^2= **45616516**

Izračun slijedi:

$$\beta_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{9124010 * 334 - 6754 * 451356}{5 * 9124010 - 45616516} = -294{,}0249$$

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{5 * 451356 - 6754 * 334}{5 * 9124010 - 45616516} = 0{,}267119411$$

U formulama se još tražio i n, a on je broj redaka podataka koji su dostupni i iznosi 5. Konačna formula modela jednostavne linearne regresije koja je dobivena računski iznosi:

$$y = 0{,}267\ x - 294 + \varepsilon$$

Upravo te vrijednosti dobivene su i u prethodnim poglavljima priručnika gdje je primjer detaljno i opisan, samo što se prethodno koristio tablični kalkulator i program KNIME za izračun parametara β_0 i β_1. Očito je kako je sam postupak dosta mukotrpan ako se za izračun koristi obični kalkulator, čak i za pet redova. Primjena tabličnih kalkulatora znatno olakšava posao, a specijalizirani programi kao što je KNIME još i više.

6.2. Izrada modela jednostavne linearne regresije u programu KNIME

Da bi se izračunali parametri β_0 i β_1 u programu KNIME potrebno je prije svega otvoriti novi hodogram. Može se zvati jednostavna_linearna_regresija.knwf. Nastavak „knwf“ iza naziva datoteke dolazi od „knime workflow“ i s tom ekstenzijom ili nastavkom se hodogram sprema na računalu. Nakon toga potrebno je izabrati datoteku s podacima koji će služiti za izradu modela. Za jednostavnu linearnu regresiju koristit će se podaci dostupni na internetu, a radi se o podacima o prodaji 5891 stana u gradu Daegu u periodu od 2007. do 2017. godine. Podaci su dostupni na adresi: https://www.kaggle.com/datasets/gunhee/koreahousedata. Za preuzimanje je potrebna prijava koja je moguća uz Googleov račun. Podaci su u CSV formatu, što znači da je potreban čvor **CSV Reader**. Podaci uključuju značajke prodanih stanova i ciljnu varijablu – cijenu po kojoj je stan prodan. Očekivano, trenirani model će predviđati cijenu stana na osnovu unesenih vrijednosti značajki.

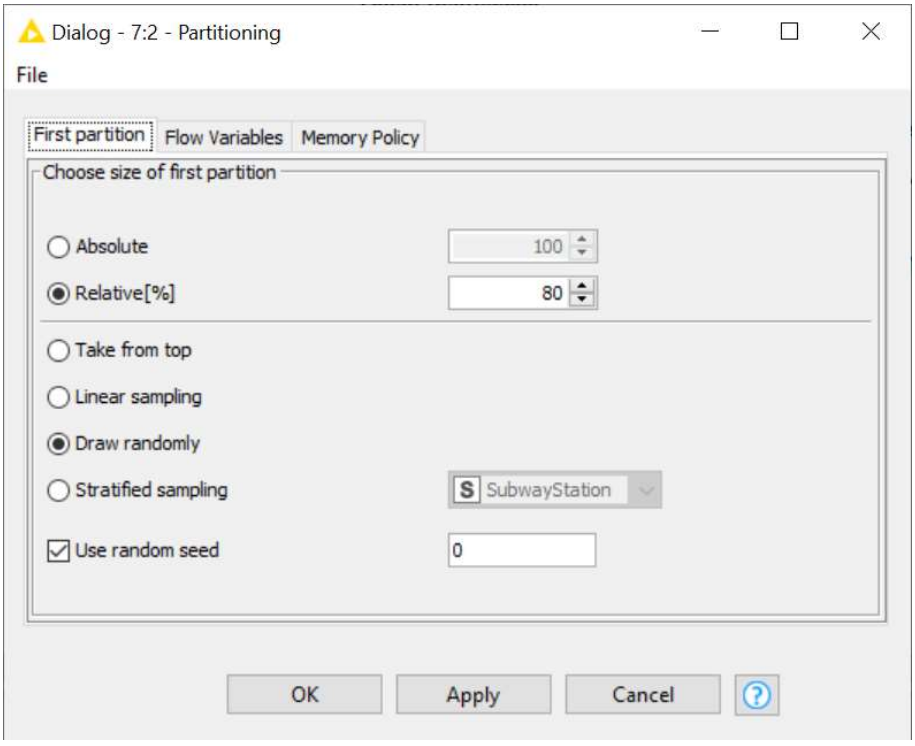
Nakon umetanja **CSV Reader** čvora koji je obrađen u jednom od prethodnih poglavlja, podatke treba podijeliti na dio za treniranje modela i dio za testiranje modela. Za treniranje može biti 80 % podataka, dok se testiranje može obaviti s preostalih 20 %. Razlog podjele je prethodno objašnjen pa slijedi upoznavanje s čvorom koji odrađuje dijeljenje podataka za treniranje i testiranje. Slika 151 prikazuje čvor **Partitioning** ili Particioniranje.

Slika 151. Čvor Partitioning

Postavke čvora **Partitioning** svode se na definiranje postotka podataka za treniranje, a ostatak ostaje za testiranje. Postoji mogućnost i zadavanja apsolutnog broja redova za treniranje (*Absolute*), ali to nije uobičajena praksa. U donjem dijelu dijaloškog okvira definira se na koji način će se izabrati prva grupa podataka za treniranje. Mogućnosti su:

a) Zadani postotak od početka (*Take from top*).
b) Linearna podjela podataka s obzirom na zadani postotak, ali u manjim skupinama (*Linear sampling*).
c) Slučajan izbor (*Draw randomly*).
d) Metoda stratificiranog uzorkovanja (*Starified sampling*).

Na kraju postoji mogućnost zadavanja tzv. „sjemena“ (*Use random seed*) što omogućuje generiranje slučajnih brojeva za podjelu na skupove za treniranje i testiranje, ali tako da ti skupovi budu uvijek isti. To omogućuje optimizaciju modela izmjenom parametara, ali bez utjecaja koji mogu nastati randomizacijom tih skupova. Na slici Slika 152 prikazuju se postavke čvora **Partitioning**.

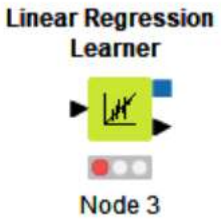


Slika 152. Postavke čvora Partitioning

Na slici Slika 153 prikazuju se spojeni čvorovi **CSV Reader** i **Partitioning**.

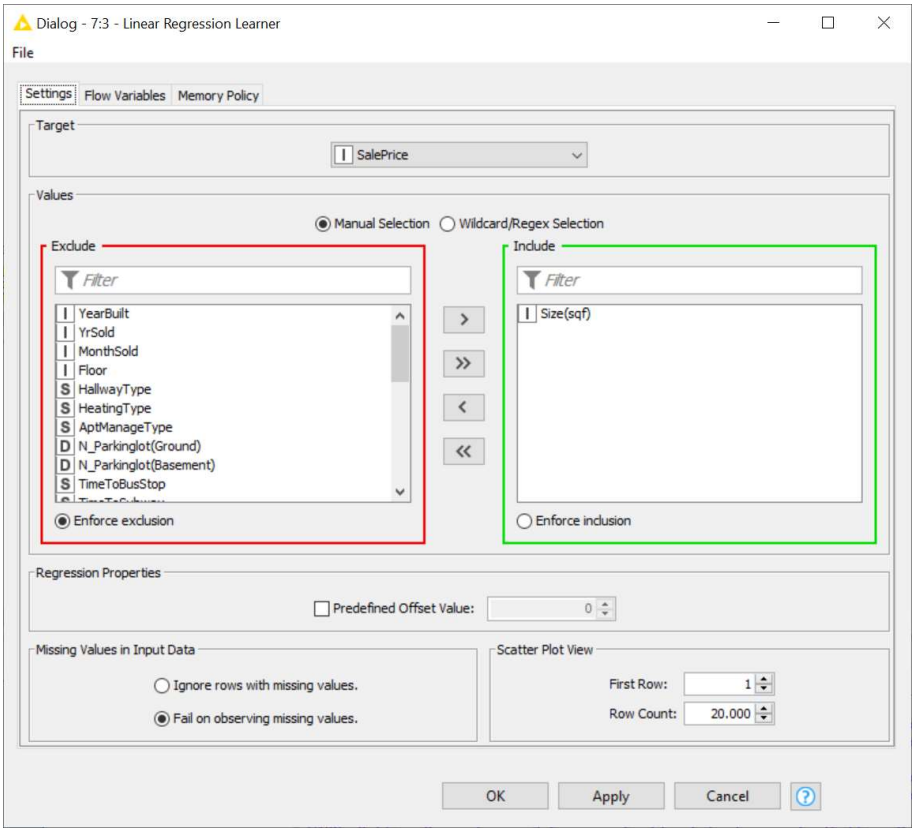
Slika 153. Izgled hodograma s čvorovima CSV Reader i Partitioning (1. faza)

Slijedi najvažniji čvor u modelu linearne regresije, a to je **Linear Regression Learner**. Već je djelomično opisan u hodogramu „Zdravo svijete!“. Taj čvor prikazan je na sliciSlika 154.



Slika 154. Čvor Linear Regression Learner

Navedeni čvor ima zadaću istrenirati model na osnovu podataka koji su mu dostavljeni. Postavke čvora uključuju definiranje značajki i ciljne varijable (*Target*). Postavke čvoraSlika 155 prikazuje slika 155.



Slika 155. Postavke čvora Linear Regression Learner

U primjeru sa stanovima u korejskom gradu Daegu izabrana je kvadratura stana (*Size(sqf)*) kao jedina značajka koja utječe na prodajnu cijenu stana. Potrebno je obratiti pažnju da je jedinica kvadratna

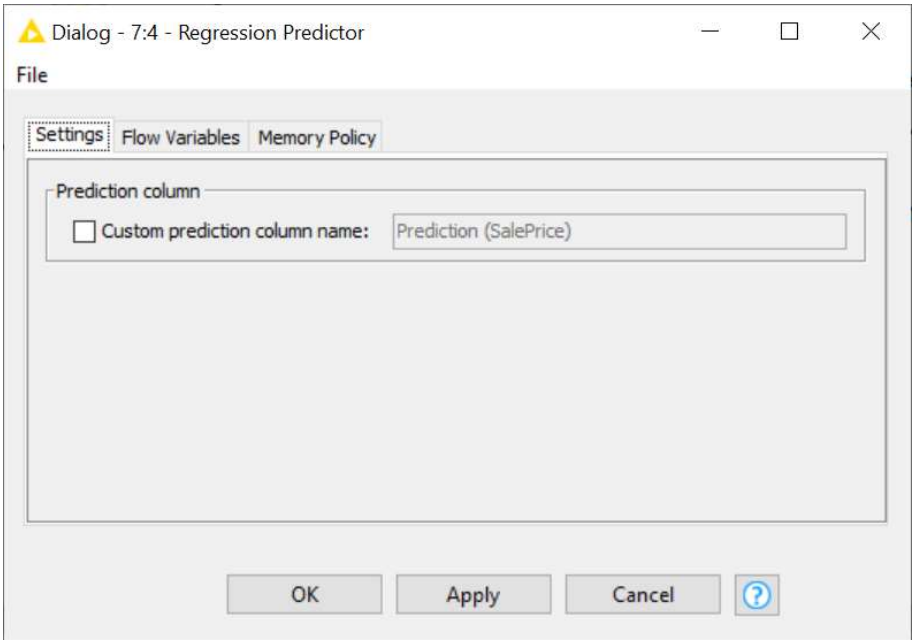
stopa, a ne kvadratni metar. Omjer konverzije je 1:10,76 što znači da jedan kvadratni metar ima 10,76 kvadratnih stopa. Ciljna varijabla je cijena stana koja je u dolarima. Jasno je kako je to samo jedna značajka koja utječe na cijenu te da se izborom samo jedne značajke neće dobiti model velike točnosti, ali postepenim dodavanjem značajki, model će biti sve točniji.

Potrebno je obratiti pažnju na izlazne priključke čvora koji se nalaze s desne strane čvora **Linear Regression Learner**. Osim crnog trokutića tu je i plavi kvadratić. Njegova funkcija je generirani model linearne regresije prenijeti dalje drugim čvorovima koji ga mogu koristiti. U daljnjim primjerima kod svakog čvora koji generira model na bazi neke tehnike strojnog učenja bit će prisutan taj plavi kvadratić. Naravno, mora postojati čvor koji može koristiti trenirani model, a to je sljedeći čvor u nizu.

Regression Predictor je čvor koji osim podataka treba i trenirani regresijski model koji će koristiti za predikciju. S lijeve strane čvora vidljivo je kako postoje dva priključka i to jedan za podatke (crni trokutić), a drugi za istrenirani model (plavi kvadratić). Taj čvor prikazuje slika Slika 156.

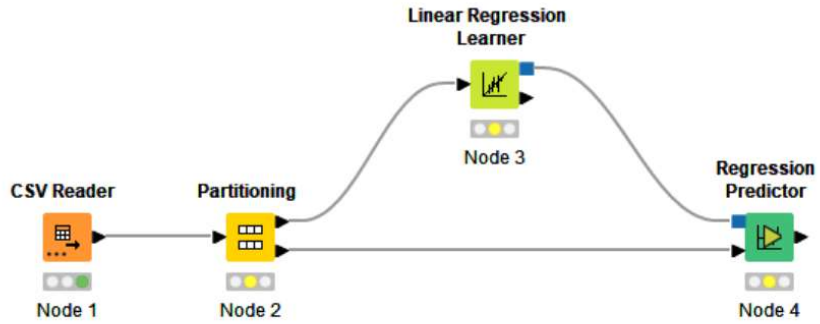
Slika 156. Čvor Regression Predictor

Čvor zahtijeva vezu s čvorom s istreniranim modelom koji mu prethodi, a veza između njih mora povezivati plave kvadratiće. Način povezivanja bit će vidljiv na sljedećoj slici hodograma koji se nadograđuje u koracima. Slika 157 prikazuje postavke čvora, koje su skromne. Jedina mogućnost je izbor druge varijable za predikciju.



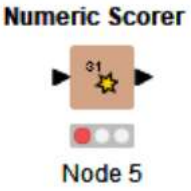
Slika 157. Postavke čvora Regression Predictor

Slika 158 prikazuje sljedeću fazu razvoja hodograma. Osim čvorova za učitavanje podataka i podjelu na 80 % podataka za treniranje i 20 % podataka za testiranje, sada su dodani čvorovi **Linear Regression Learner** i **Regression Predictor**. Važno je uočiti kako su s desne strane čvora **Partitioning** dva crna trokutića. Iz gornjeg se povlači 80 % podataka (tako je u postavkama čvora u primjeru), a iz donjeg 20 % podataka. 80 % podataka ide prema čvoru za treniranje modela, a 20 % ide prema čvoru za testiranje modela. Čvor **Linear Regression Learner** spaja se na čvor **Regression Predictor** i na taj način se prenosi model, a čvoru **Regression Predictor** se prebacuje onih 20 % podataka iz donjeg priključka čvora **Partitioning**.



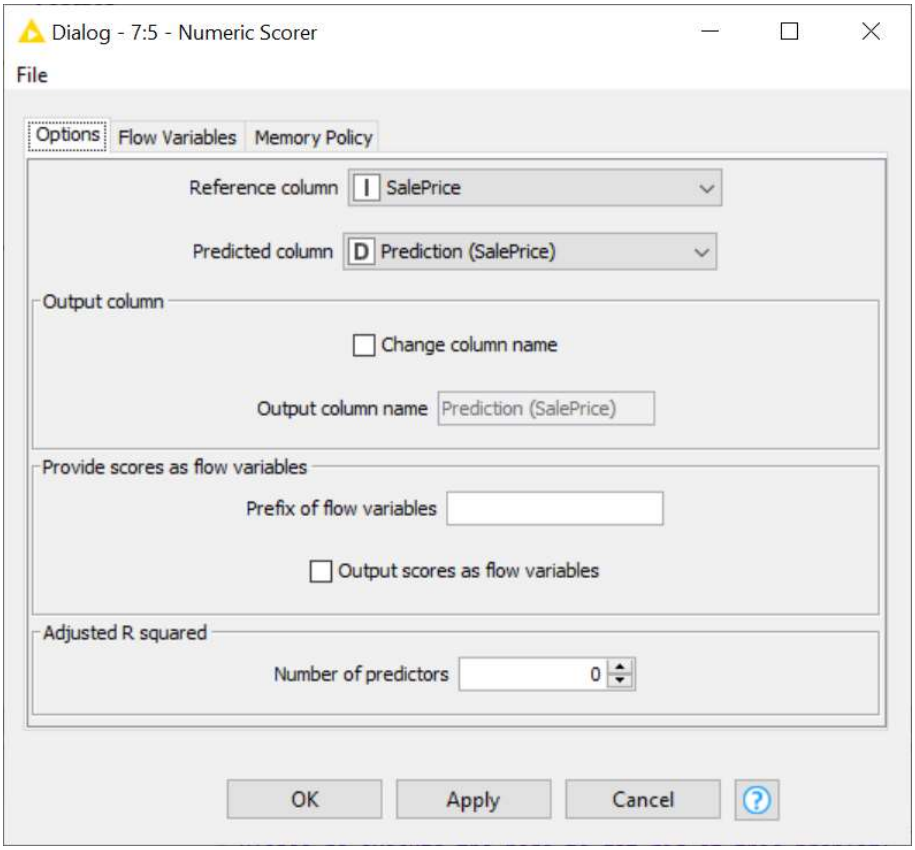
Slika 158. Izgled hodograma s dodanim Linear Regression Learner i Regression Predictor

Sljedeći čvor u nizu je **Numeric Scorer** ili na hrvatskom jeziku pomalo nespretan naziv bio bi Numerički bilježnik. Taj čvor izračunava već spomenute pokazatelje kvalitete modela na osnovu stvarnih vrijednosti i vrijednosti dobivenih modelom. Izračunava koeficijent determinacije (R^2), srednju apsolutnu pogrešku, srednju kvadratnu pogrešku, korijen srednje kvadratne pogreške, srednju razliku, srednju apsolutnu postotnu pogrešku i prilagođeni koeficijent determinacije (R^2). Izračunate vrijednosti mogu se pregledati iz kontekstnog izbornika čvora (*Statistics*) i/ili dalje obraditi korištenjem tablice koja je dostupna na izlaznom priključku čvora. Taj čvor prikazan je na sliciSlika 159.



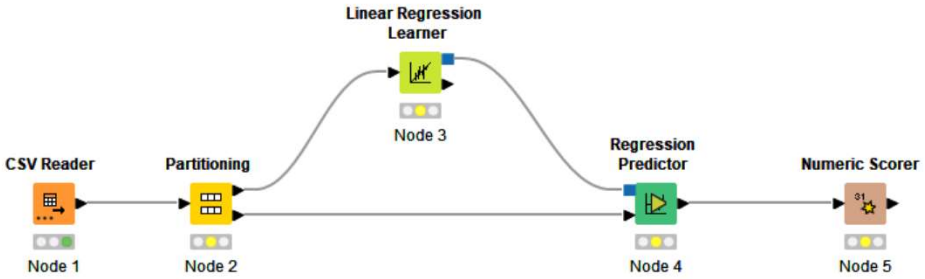
Slika 159. Čvor Numeric Scorer

Slika 160 prikazuje dijaloški okvir s postavkama čvora, a na njemu je najbitnije postaviti odgovarajuće vrijednosti u padajuće izbornike na vrhu okvira. U padajućem izborniku *Reference column* bira se ciljna varijabla (*SalePrice*), a u *Predicted column* bira se ciljna varijabla koja je izračunata modelom (*Predicted(SalePrice)*). Osim toga u postavkama se mogu mijenjati naziv stupca u izlaznoj tablici s pokazateljima točnosti modela, omogućiti izvoz statističkih pokazatelja kao varijabli i mijenjati broj prediktora kod prilagođenog koeficijenta determinacije.



Slika 160. Postavke čvora Numeric Scorer

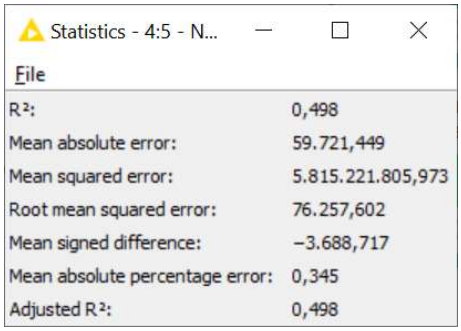
Slika 161 prikazuje hodogram u trećoj fazi u kojoj mu je dodan i čvor **Numeric Scorer**.



Slika 161. Izgled hodograma s dodanim čvorom Numeric Scorer

U uređivaču hodograma dostupne su još neke informacije koje olakšavaju razumijevanje postupka treniranja modela, kao i analize rezultata. Ako se pokazivač miša postavi na donji izlazni trokutić čvora **Partitioning**, dobit će se informacija o tome što je izlaz tog čvora i konkretno tog izlaznog priključka. Radi se o tablici od 1130 redova i 30 stupaca, a to je 20 % podataka koji su učitani čvorom **Excel Reader**. Nakon čvora **Regression Predictor** na izlaznom priključku dobiva se informacija kako je izlaz tog čvora tablica od 1130 redova i 31 stupac. Očekivano, dodan je stupac s vrijednostima stana koji je izračunao model. Te tablice mogu se pogledati u kontekstnom izborniku čvorova s tim da se za čvor **Partitioning** to dobiva na *Second Partition (Remaining Rows)*, a za čvor **Regression Predictor** na *Predicted Data*.

Posebnu pažnju treba obratiti na koeficijent determinacije koji je postignut s modelom, a koji je dostupan kao izlazni podatak na čvoru **Numeric Scorer.** Slika 162 prikazuje taj podatak.

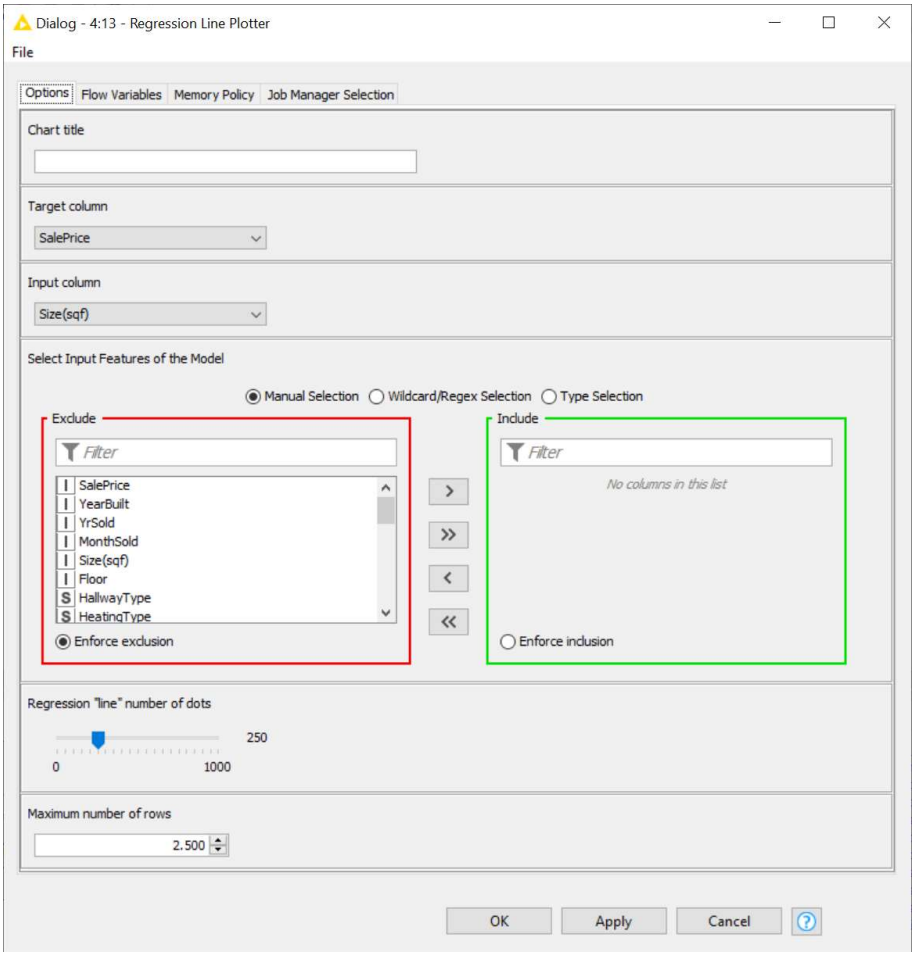


Slika 162. Podaci o modelu generirani na osnovu testnih podataka

Koeficijent determinacije iznosi 0,498 što i nije tako loše s obzirom da je model kreiran koristeći samo jednu značajku. Osim toga može se pogledati kako izgleda pravac linearne regresije. Za to se može koristiti čvor **Regression Line Plotter** ili Crtač regresijske linije, koji je dostupan u repozitoriju KNIME Hub, a nakon što je pronađen jednostavno se prevuče u hodogram što pokreće kratku instalaciju istog. Treba ga povezati s modelom koji je istreniran (plavi kvadratić) i s testnim podacima (crni trokutić).

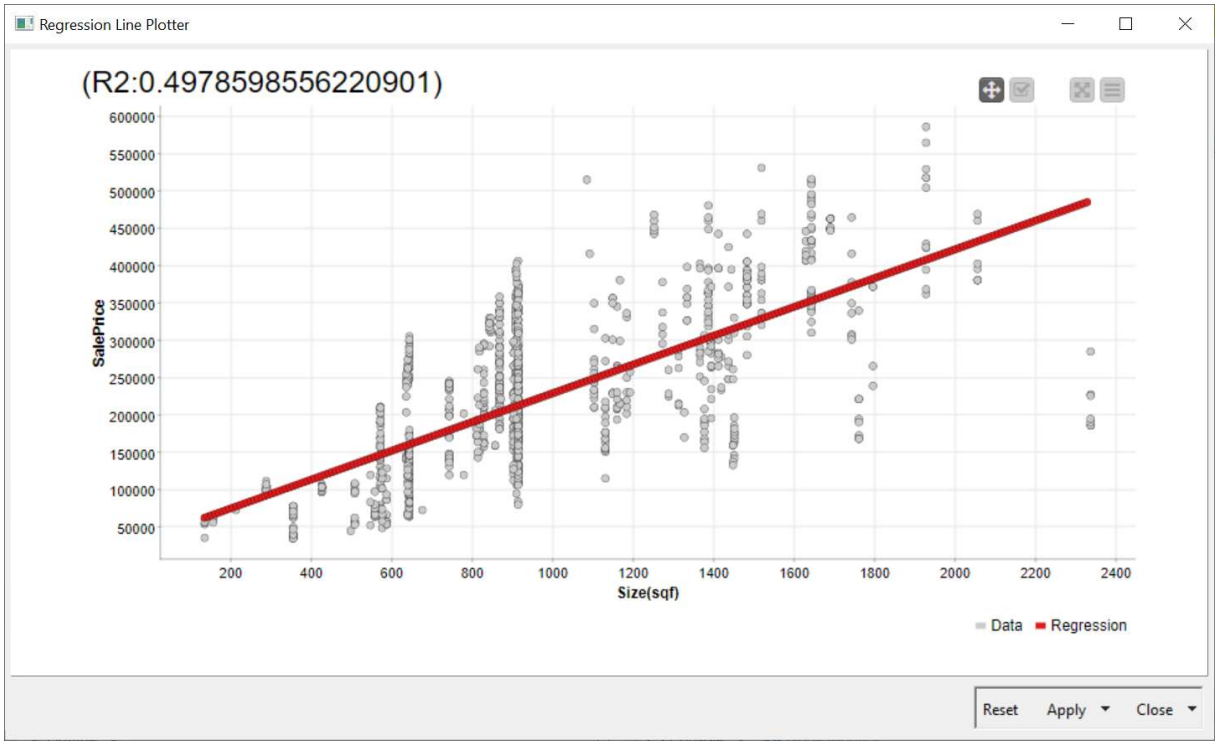
Slika 163. Čvor Regression Line Plotter

Slika 164 prikazuje postavke čvora **Regression Line Plotter**. Može se definirati naslov grafikona, a dva glavna polja koja je potrebno definirati su *Target Column* i *Input Column*. Tu je postavljeno *TargetPrice* i *Size (sqf)*. Opcija *Regression „line" number of dots* služi za definiranje broja točaka u pravcu regresije, dok *Maximum number of dots* služi za izbor maksimalnog broj točaka koje će biti prikazane u grafikonu. S obzirom da je broj redaka u testnom dijelu podataka manji od tog broja, ta vrijednost u ovom primjeru nema utjecaja na izgled grafikona.



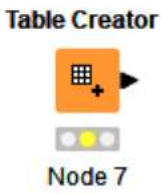
Slika 164. Postavke čvora Regression Line Plotter

Kao izlaz čvora **Regression Line Plotter** dobiva se pravac linearne regresije i testni podaci koji su izabrani za prikaz. Slika 165 prikazuje pravac i koeficijent determinacije u gornjem lijevom kutu grafikona.



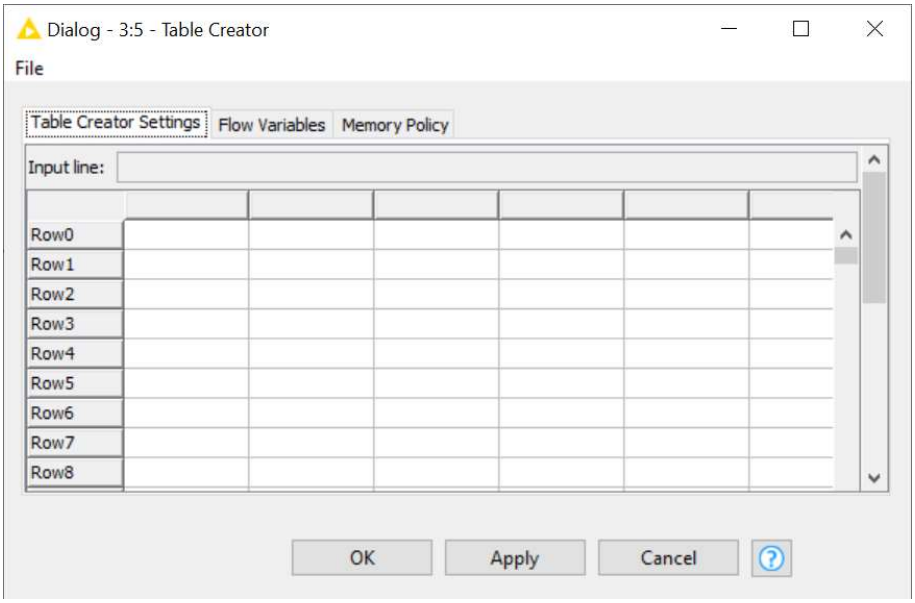
Slika 165. Pravac linearne regresije istreniranog modela

Sljedeći korak je primjena istreniranog modela linearne regresije. Da bi se istrenirani model koristio ubacit će se još jedan čvor **Regression Predictor** i spojiti ga s čvorom **Linear Regression Learner** da bi mogao koristiti model koji je prethodno istreniran. Čvor **Regression Predictor** osim modela na ulaznim priključcima treba i podatke i to onih značajki na osnovu kojih je treniran model. Na osnovu tih podataka će čvor **Regression Predictor** izračunati vrijednost ciljne varijable. U ovom primjeru model je treniran samo s jednom značajkom i to je kvadratura stanova, a ciljna varijabla je cijena stana. Model je treniran s 4519 redova, odnosno parova podataka. Taj broj dobije se tako da se stavi pokazivač na izlazni gornji priključak čvora **Partitioning**. Za predikciju cijene stana potreban je broj kvadratnih metara stana, a taj broj se dostavlja čvoru **Regression Predictor** koristeći čvor **Table Creator** koji se koristio u primjeru „Zdravo svijete!“. Slika 166 prikazuje taj čvor.



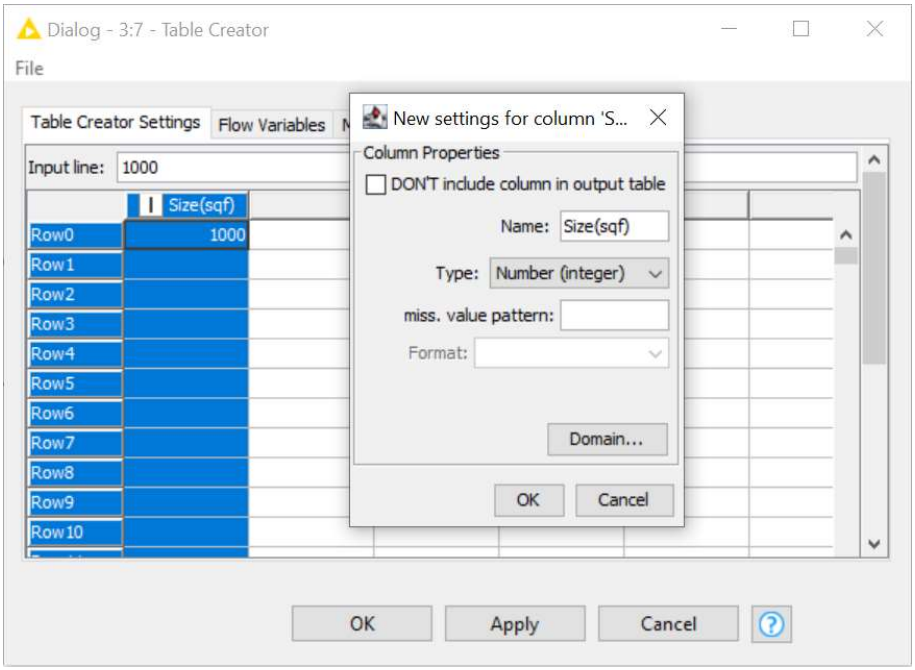
Slika 166. Čvor Table Creator

Postavke čvora **Table Creator** svode se na popunjavanje tablice, pri čemu se treba popuniti sadržaj i zaglavlja stupaca. Slika 167 prikazuje postavke čvora **Table Creator**.



Slika 167. Postavke čvora Table Creator

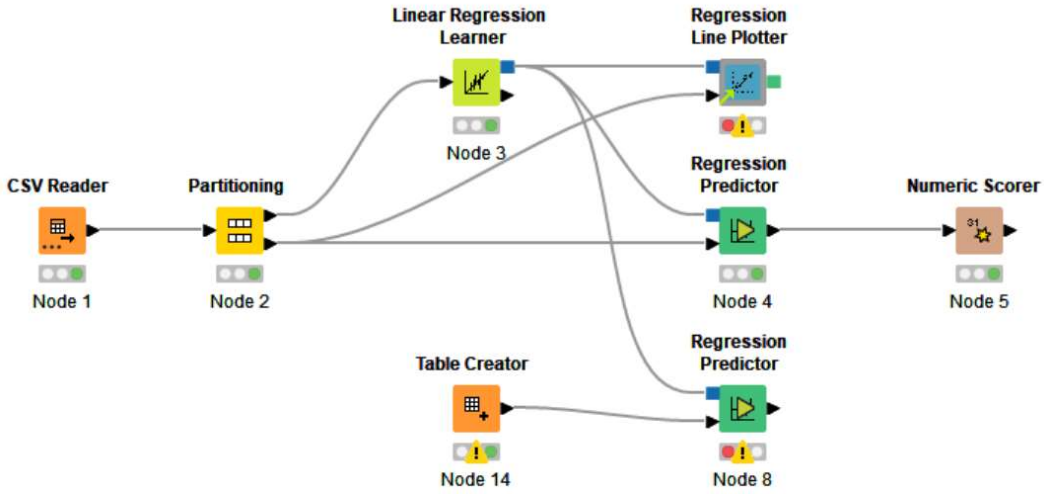
Ćelije se popunjavaju klikom na samu ćeliju i unosom sadržaja, dok se naziv stupca mijenja na način da se klikne na pravokutnik zaglavlja stupca pri čemu se otvara mali dijaloški okvir. U njemu se popunjava naziv stupca (*Name*), kao i vrsta podataka koji se nalaze u stupcu (*Type*). Bitno je da naziv stupca bude točno napisan ako se tablica koristi za predikciju modela jer čvor **Regression Predictor** očekuje da naziv stupca, odnosno značajke bude isti kao i u modelu. Slika 168 prikazuje popunjavanje zaglavlja čvora **Table Creator**. Treba uočiti kako je prva ćelija popunjena i to s vrijednošću 1000 što znači da će se uz pomoć modela izračunati cijena stana od 1000 kvadratnih stopa.



Slika 168. Postavke čvora Table Creator – popunjavanje zaglavlja

Slika 169 prikazuje hodogram s čvorovima koji omogućuju učitavanje podataka, izradu i testiranje modela, a na kraju i korištenje modela na način da se u čvor **Table Creator** unese kvadratura stana za

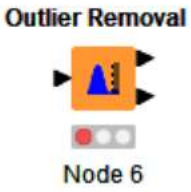
koju se želi predikcija cijene. Nakon što se izvrše svi čvorovi iz kontekstnog izbornika donjeg čvora **Regression Predictor** klikom na Predicted Data dobiva se cijena stana od 229041,454 dolara.



Slika 169. Izgled hodograma s dodanim čvorom Table Creator

6.3. Optimizacija modela jednostavne linearne regresije

Postoji niz načina da se model poboljša,a u nastavku će se koristiti čvor **Outlier Remover** ili Uklanjanje odstupanja koji služi da se iz učitanih podataka isključe ekstremne vrijednosti koje su često pogrešno unesene. Čvor je najbolje postaviti odmah nakon učitavanja podataka. Slika 170 prikazuje izgled čvora.

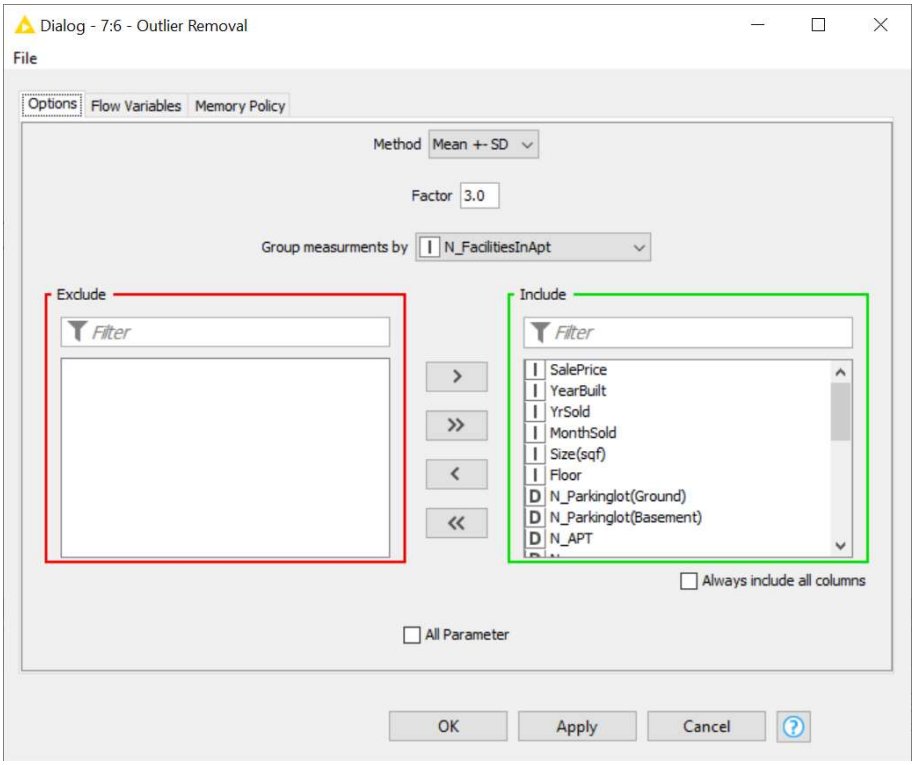


Slika 170. Čvor Outlier Removal

Čvor **Outlier Remover** ne dolazi sa zadanom instalacijom programa KNIME nego ga je potrebno dodatno instalirati. Za instalaciju ga je nužno pronaći koristeći repozitorij KNIME Hub u gornjem desnom dijelu prozora programa KNIME te ga prevući na uređivač hodograma. To pokreće instalaciju čvora pri čemu je neophodno složiti se s uvjetima korištenja i nekoliko puta kliknuti na OK. Ponekad je potrebno kod instalacije novih čvorova ponovo pokrenuti KNIME. Jednom instaliran čvor dostupan je trajno.

Kao što je navedeno, čvor uklanja ekstremne vrijednosti tako da definira gornju i donju vrijednost raspona podataka, a u postavkama može se izabrati jednu od dvije metode na osnovu koje se ekstremne vrijednosti uklanjaju. Metoda *Mean +- SD*, odnosno Srednja vrijednost +- standardna devijacija će definirati gornju granicu kao srednju vrijednost zbrojenu s umnoškom standardne devijacije s faktorom koji posebno zadajemo. Druga metoda zvana *Boxplot* definira gornju granicu kao 75 % kvantil zbrojen s međukvantilnim rasponom pomnoženim sa zadanim faktorom.

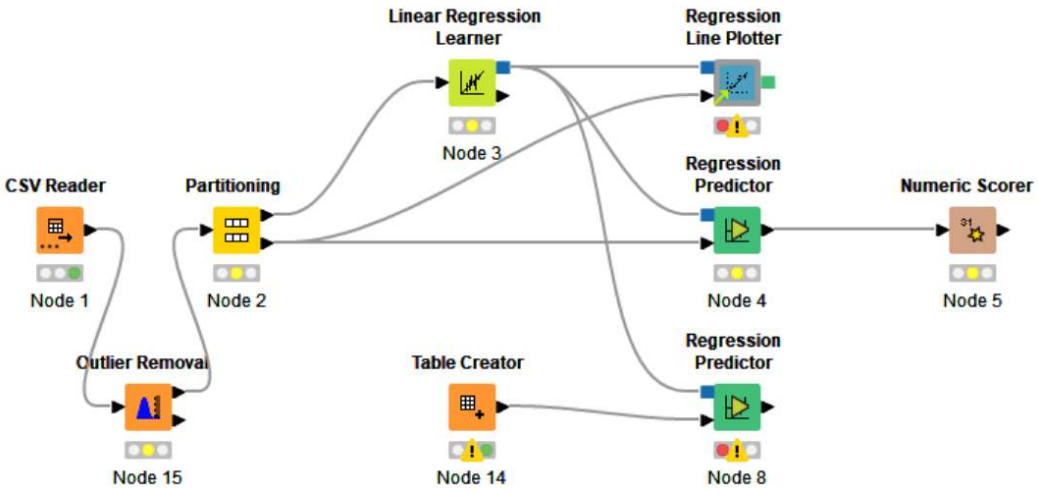
Slika 171 prikazuje postavke čvora Outlier Removal pri čemu se u desnom zelenom pravokutniku ostavljaju varijable na koje se želi primijeniti opisana metoda uklanjanja ekstremnih vrijednosti.



Slika 171. Postavke čvora Outlier Removal

Kao što je navedeno čvor **Outlier Removal** umetne se odmah nakon čvora za učitavanje podataka. Slika 172 prikazuje konačan izgled hodograma. Neke od veza su obrisane, a brisanje se obavlja tako da se klikne na vezu i pritisne tipku *Delete* na tipkovnici. Kod ovakvih promjena u kojima se umetne čvor na početak hodograma najjednostavnije je izvršiti zadnji čvor u nizu pri čemu se izvršavaju svi čvorovi prije njega.

Konačno se može provjeriti koliko je model točniji nakon umetanja čvora **Outlier Removal**. Prije umetanja čvora koeficijent determinacije u čvoru **Numeric Scorer** iznosio je 0,498, a nakon umetanja čvora koeficijent determinacije se povećao na 0,554. Ipak, treba biti svjestan da su možda te ekstremne vrijednosti bile ispravne te da se na ovaj način dobila veća vrijednost koeficijenta determinacije, ali da je model manje precizan za ekstremne vrijednosti. Ako postoji sumnja, preporuka je pregledati te ekstremne vrijednosti i vidjeti koliko imaju smisla prije umetanja samog čvora.



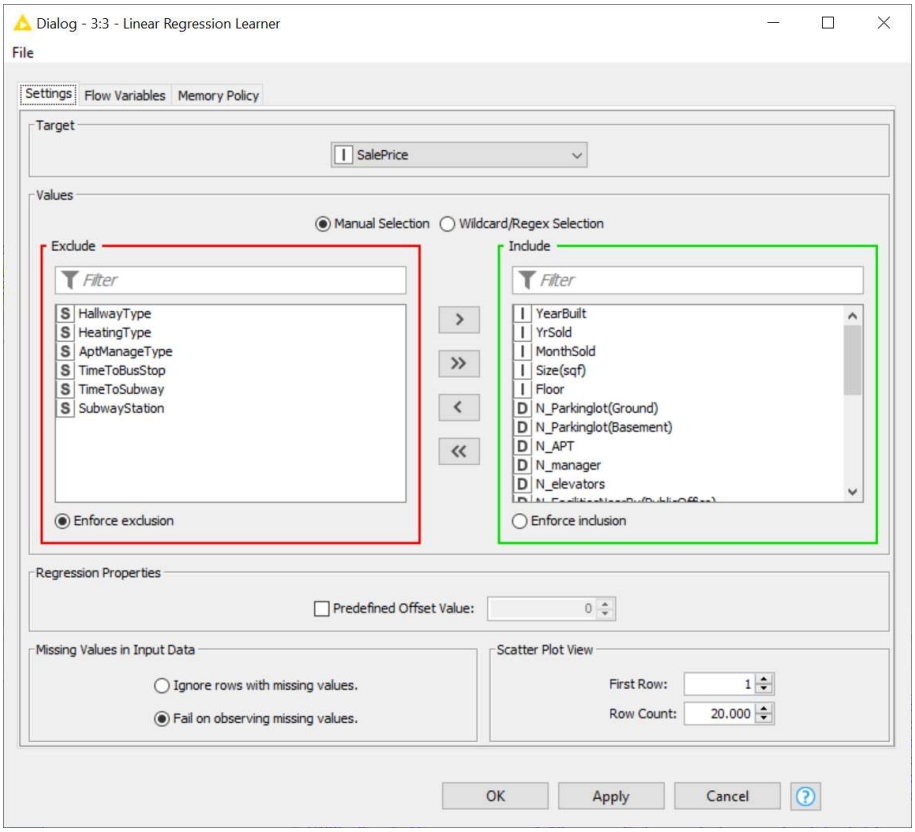
Slika 172. Izgled hodograma s dodanim čvorom Outlier Removal

6.4. Višestruka linearna regresija

Skup podataka koji se koristio u prethodnom primjeru za jednostavnu linearnu regresiju ima 30 stupaca, od kojih su 29 stupci s nekim značajkama stanova koji su prodani, a jedan stupac je ciljna varijabla – cijena stana po kojoj je prodan. S obzirom da se radi o jednostavnoj linearnoj regresiji u prethodnom primjeru korištena je samo jedna značajka i to kvadratura stana za predviđanje cijene stana. Iz stvarnog života je poznato kako to nije jedini čimbenik koji utječe na cijenu stana nego ih ima više. Jedan od čimbenika je svakako lokacija, a bitan je i kat na kojemu se stan nalazi. Blizina javnog prijevoza je također važna. Uključivanjem tih čimbenika u model za očekivati je da će model biti točniji. Uključivanjem dodatnih značajki prelazi se iz jednostavne linearne regresije u višestruku linearnu regresiju koju se može opisati izrazom (Eberly, 2007):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \varepsilon$$

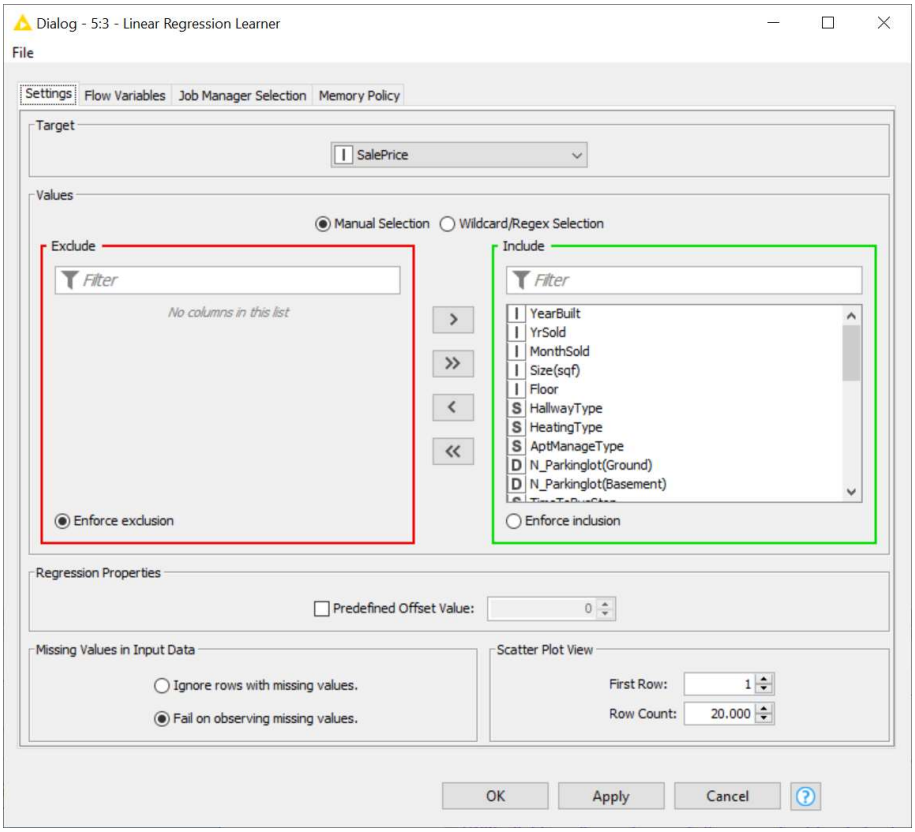
Kod jednostavne linearne regresije varijabla ili značajka x u primjeru bila je kvadratura stana, a čvor **Linear Regression Learner** je izračunao varijablu β_1. Kod višestruke linearne regresije postoji više varijabli x i svaka predstavlja neku od značajki stana, kao što su broj soba, udaljenost od javnog prijevoza, kat itd. U ovom slučaju koristit će se čvor **Linear Regression Learner** koji će izračunati pripadajuće varijable β. Jedina razlika bit će u tome što će se čvoru **Linear Regression Learner** omogućiti pristup svim značajkama i pustiti ga da generira model višestruke linearne regresije.



Slika 173. Postavke čvora Linear Regression Learner kod višestruke regresije

Treba uočiti kako se u ovom slučaju model trenira koristeći sve numeričke značajke jer su u zelenom okviru sve varijable ispred kojih je oznaka I ili D, dok su u crvenom okviru varijable ispred kojih je slovo S. Oznake I i D dolaze od *Integer* i *Double*, dok oznaka S dolazi od *String*. *Integer* je oznaka za cjelobrojne varijable, a *Double* za varijable s decimalnim brojevima duple preciznosti. Nakon te izmjene izvrše se svi čvorovi i pogleda se koeficijent determinacije u čvoru **Numeric Scorer**. Iznosi 0,808! Novi model je znatno bolji od modela koji je generiran koristeći samo jednu značajku – kvadraturu stana. Može se probati izbaciti čvor Outlier Removal da se vidi koliki će u tom slučaju biti rezultat. Rezultat je još bolji i iznosi 0,868.

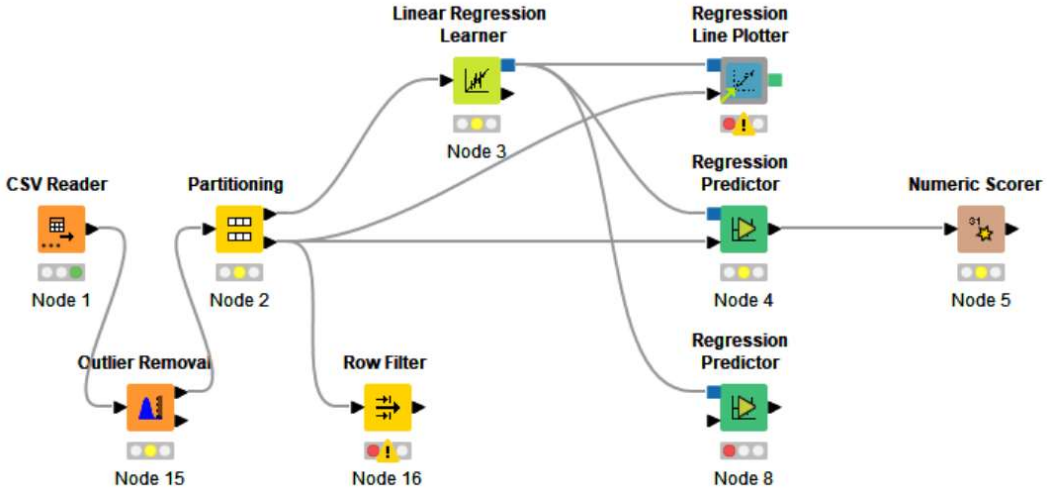
Postoji još jedan način kako se može postići veća točnost, a to je uključenjem kategorijalnih varijabli, odnosno značajki. Čvor **Linear Regression Learner** ima mogućnost transformacije kategorijalnih varijabli u tzv. *dummy* odnosno lažne varijable. Ako su dostupne tri vrijednosti jedne kategorijalne varijable (*mr, mrs i miss*), pretvaranjem u *dummy* ili lažne varijable dobivaju se dvije varijable, a vrijednosti se kodiraju na način da za prvu vrijednost (*mr*) prva *dummy* ili lažna varijabla ima vrijednost 1, za drugu vrijednost (*mrs*) druga *dummy* ili lažna varijabla ima vrijednost 1, a za treću (*miss*) obje *dummy* ili lažne varijable imaju vrijednost 0. Slika 174 prikazuje postavke čvora s uključenim svim varijablama.



Slika 174. Postavke čvora Linear Regression Learner sa svim varijablama

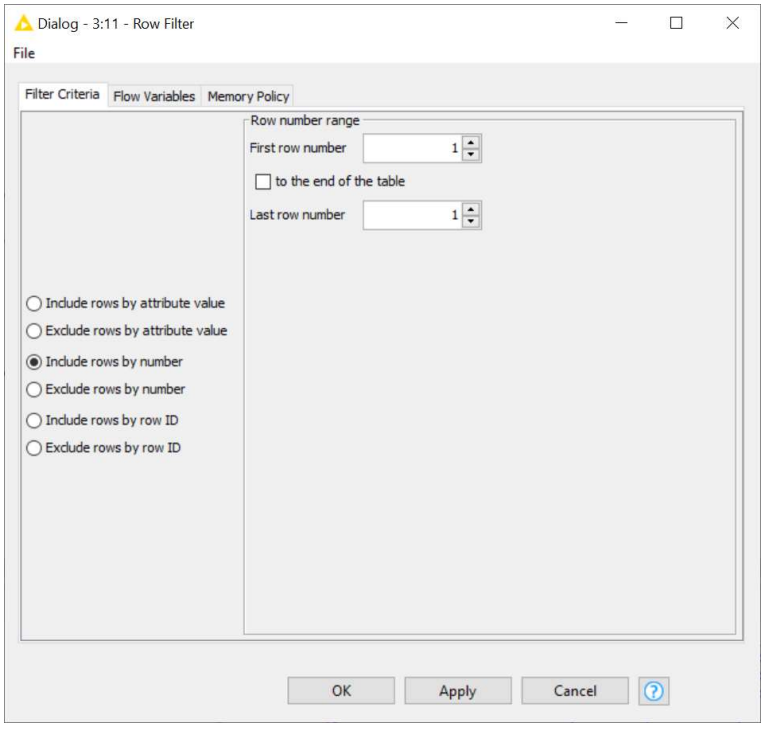
Nakon ubacivanja kategorijalnih varijabli koeficijent determinacije iznosi 0,879, što je najbolji rezultat do sada.

Nažalost, donji čvor **Regression Predictor** sad nije od neke koristi jer čvor **Table Creator** ima samo jednu značajku, a trebalo bi ih unijeti 29 da bi se mogao isprobavati model sa stvarnim podacima. Drugim riječima, potrebno je prepisati 29 naziva stupaca u čvor **Table Creator** da bi se mogao testirati model s raznim vrijednostima značajki. Da bi to ipak mogli bez prepisivanja naziva stupaca, izvest će se trik. Za donji čvor **Regression Predictor** potrebna je tablica sa samo jednim redom, ali zaglavlje treba biti isto kao i zaglavlje podataka koje je učitano čvorom **CSV Reader**. To nije problem i koristeći čvor **Row Filter** kreirat će se tablica sa samo prvim redom iz bilo koje dostupne tablice. Može se preuzeti tablicu iz donjeg priključka čvora **Partitioning** i u njoj ostaviti samo prvi red. Prije toga se čvor **Table Creator** treba obrisati, tako da se označi i pritisne tipka *Delete* na tipkovnici. Nakon toga, hodogram se izmijenio. Slika 175 prikazuje izmijenjeni hodogram.



Slika 175. Izgled hodograma s dodanim čvorom Row Filter

Nakon umetanja čvora **Row Filter** potrebno je podesiti postavke čvora tako da ostane samo jedan red podataka. Slika 176 prikazuje te postavke. Dovoljno je izabrati *Include rows by number* čime se definiraju brojevi redova koji se propuštaju kroz filter te nakon toga s desne strane dijaloškog okvira definirati da se propuštaju redovi od prvog (*First row* number) do posljednjeg (*Last row number*).

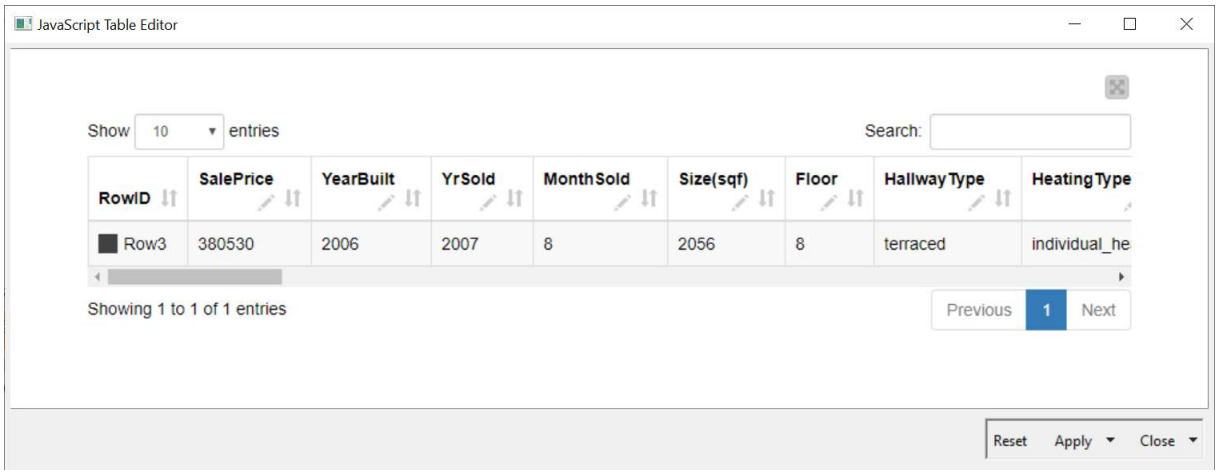


Slika 176. Postavke čvora Row Filter pri čemu ostaje samo prvi red

Nakon filtriranja na izlaznom priključku čvora **Row Filter** dobiva se tablica s 30 zaglavlja i jednim redom podataka. Kako bi se mogao uređivati taj jedan red podataka i unositi različite vrijednosti značajki, postavit će se čvor **Table Editor** ili Uređivač tablice. Na Slika 177 prikazan je čvor.

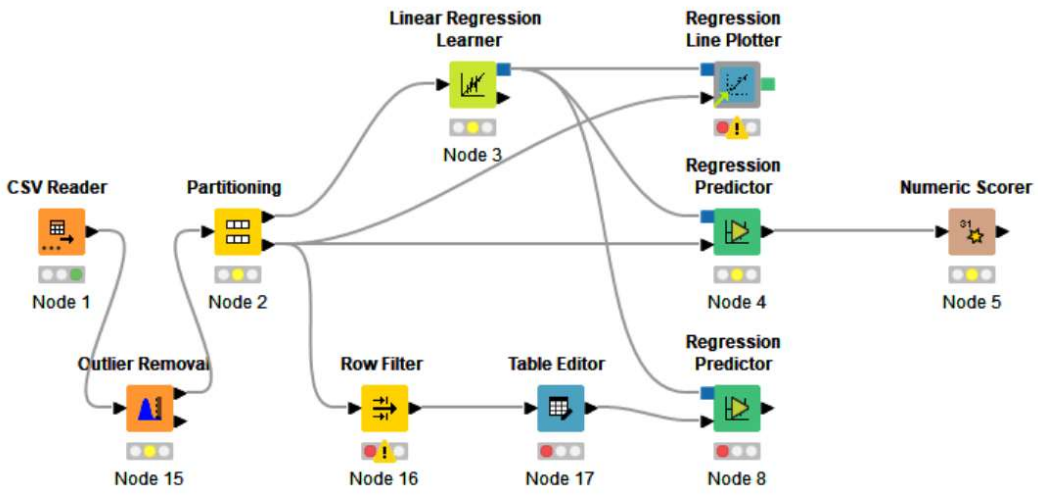
Slika 177. Čvor Table Editor

Kako mu samo ime kaže, čvor **Table Editor** omogućuje uređivanje tablice koja mu se isporuči na ulazni priključak s lijeve strane. Nakon uređivanja, tablica je dostupna na priključku s desne strane čvora. Slika 178 prikazuje dijaloški okvir u kojemu se mogu mijenjati vrijednosti značajki u tablici. Moguće je prethodnim čvorom omogućiti prolaz i više redova podataka, koji bi kasnije bili korišteni za izračun više vrijednosti ciljne varijable (cijene stana) korištenjem modela višestruke linearne regresije. Nakon prilagodbe značajki unesenu vrijednost je potrebno potvrditi pritiskom na tipku Enter nakon promjene te na kraju treba kliknuti na gumb *Apply* u donjem desnom kutu dijaloškog okvira.



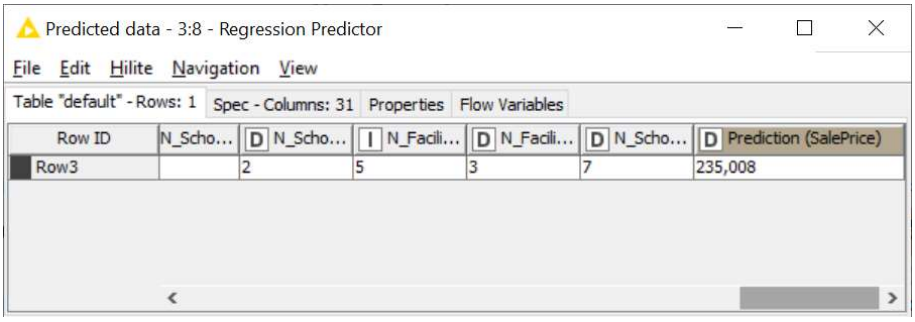
Slika 178. Uređivanje tablice u čvoru Table Editor

Slika 179 prikazuje kompletan hodogram sa svim čvorovima.



Slika 179. Izgled kompletnog hodograma

Konačna vrijednost ciljne varijable izračunata korištenjem modela višestruke linearne regresije uz podatke koji su uneseni u čvor **Table Editor** može se pročitati iz donjeg čvora **Regression Predictor**. Iz kontekstnog izbornika treba izabrati *Predicted data*.



Slika 180. Izračun vrijednosti ciljne varijable korištenjem modela

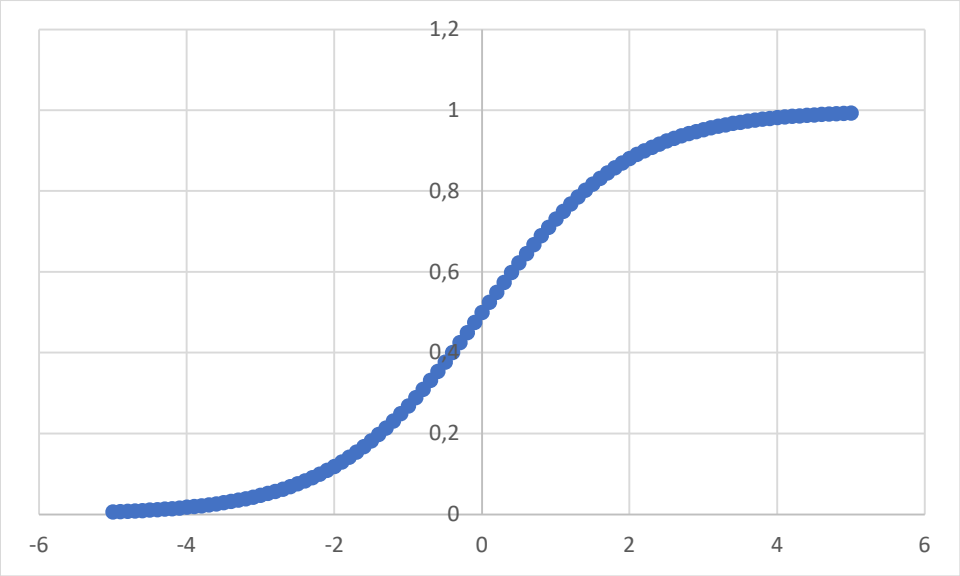
Predikcija prodajne cijene stana nalazi se u zadnjoj ćeliji tablice ispod zaglavlja *Prediction (SalePrice).* Time se obradio kompletan proces od učitavanja podataka, treniranja modela i primjene istog koristeći nove podatke s kojima model nije bio treniran.

7. Logistička regresija

Logistička regresija koristi se za klasifikaciju. Logistička regresija bazirana je na logističkoj funkciji koju su uveli statističari, a za nju je karakteristično da svaki prirodni broj x pretvara u vrijednost između 0 i 1, ali nikada ne postiže te krajnje granice. Evo logističke funkcije (Brownlee, 2016):

$$y = \frac{1}{1 + e^{-x}}$$

Slovo „e" u funkciji je baza prirodnog logaritma. Slika 181 prikazuje grafikon logističke funkcije. Na grafikonu se vidi funkcija s granicama na -5 i 5 na osi X, ali bez obzira koliko se ide u beskonačnost na jednu ili drugu stranu osi X, funkcija nikada ne dostiže granične vrijednosti 0 i 1.



Slika 181. Logistička funkcija

Logistička regresija za model koristi funkciju iz koje se izračunava y slično kao i kod linearne regresije. Opet se susrećemo koeficijenti β_0, β_1 … β_n ovisno o tome koliko ima značajki za treniranje modela. Svaka značajka nakon treniranja modela ima svoj koeficijent β. Funkcija je nešto drugačija od funkcije linearne regresije i za samo jednu značajku glasi (Brownlee, 2016):

$$y = \frac{e^{\beta 0 + x * \beta 1}}{1 + e^{\beta 0 + x * \beta 1}}$$

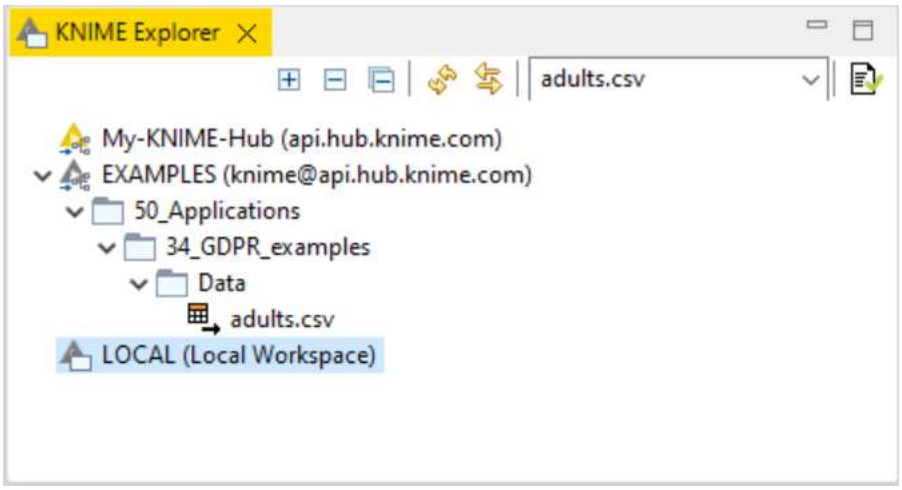
U slučaju da postoji više značajki, funkcija je kompleksnija i ima onoliko varijabli x i β koliko ima značajki te još i β_0. Drugim riječima potencije iznad prirodnog logaritma u brojniku i nazivniku bi bile duže. Zbog kompleksnosti neće se objašnjavati način treniranja modela, ali u zadnjem dijelu ovog poglavlja bit će pokazano na jednostavnom primjeru sa samo jednom značajkom da su ključni elementi istreniranog modela koeficijenti β_0 i β_1 kao i kod linearne regresije.

Prednosti logističke regresije su mala zahtjevnost što se tiče računalnih resursa za treniranje i kreiranje jednostavnog modela koji se lako interpretira. Nedostaci su osjetljivost na podatke koji nisu normalizirani i primjenjivost samo kad je prikladan linearni model (Egger, 2022).

Na kraju teorijskog uvoda treba spomenuti kako logistička regresija izračunava vjerojatnost zadane kategorije, a ta vjerojatnost se onda pretvara u vrijednosti 0 ili 1 kod jednostavne klasifikacije s dvije moguće kategorije (Brownlee, 2016).

7.1. Izrada modela logističke regresije u programu KNIME

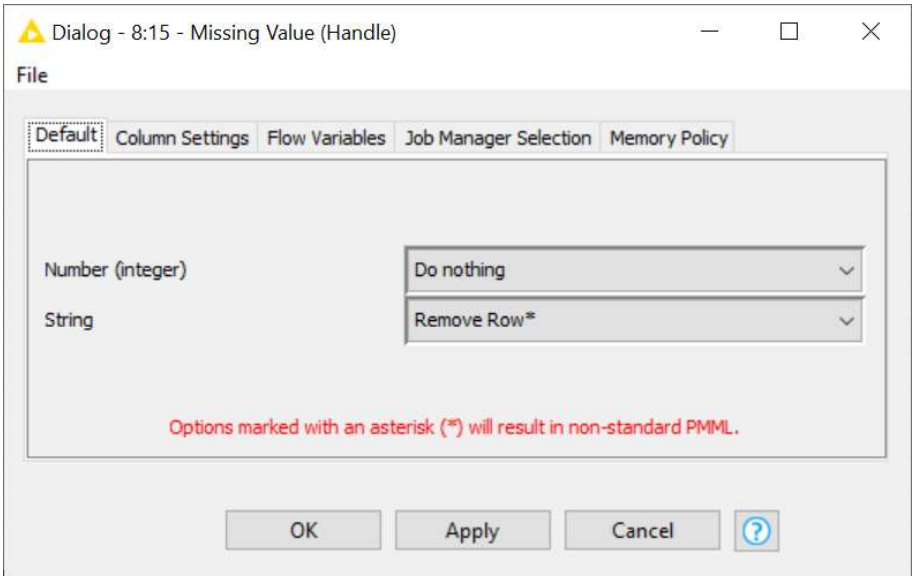
Hodogram za model logističke regresije ne razlikuje se bitno od hodograma linearne regresije. Prvi čvor služi za učitavanje podataka. Što se tiče samih podataka koristit će se CSV datoteka koja dolazi s instalacijom programa KNIME i nije ju potrebno posebno preuzimati s interneta. Datoteci je ime *adults.csv,* a najjednostavnije ju je pronaći tako da se u KNIME preglednik (*KNIME explorer*) upiše „adults.csv" (bez navodnika) i datoteka će se pojaviti među primjerima. Prebacuje se u željenu mapu tako da se kopira koristeći kontekstni izbornik i zalijepi na isti način u mapu LOCAL ili neku podmapu.



Slika 182. Dohvat datoteke adults.csv u KNIME pregledniku

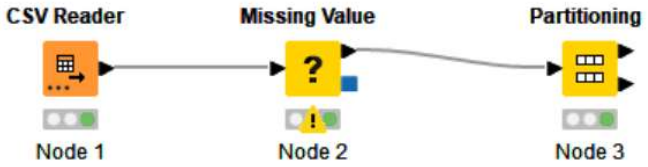
Potrebno je reći i nekoliko riječi o samim podacima u datoteci *adults.csv*. Radi se o podacima iz 1994. godine gdje su dane značajke nekoliko desetaka tisuća građana SAD-a, a u posljednjem stupcu svatko je klasificiran u jednu od dvije kategorije s obzirom na godišnje prihode. Jedna kategorija ima prihod koji je manji ili jednak $ 50000 na godišnjoj razini, dok druga kategorija ima prihod koji je veći od $ 50000. Datoteka koja dolazi s instalacijom programa KNIME je pročišćena i preostao je 32561 red podataka. Značajke koje su dostupne su broj godina, spol, bračno stanje, tjedni broj radnih sati, nacionalnost, obrazovanje, rasa, zanimanje, itd. Detaljni podaci su dostupni na adresi: http://archive.ics.uci.edu/dataset/2/adult

Podaci se učitavaju koristeći čvor **CSV Reader**, a nakon njega se umetne čvor **Missing Values** kojim se isključuju svi redovi kojima nedostaju nenumeričke vrijednosti. Bez tog brisanja čvor **Logistic Regression Learner** bi prijavljivao grešku. Slika 183 prikazuje te postavke u dijaloškom okviru postavki čvora **Missing Values**. Čvorove koji su prethodno opisani u priručniku neće se ponovno opisivati, a **CSV Reader** i **Missinga Values** spadaju među njih. Nakon njih, dodaje se čvor **Partitioning** koji dijeli podatke u dio za treniranje i dio za testiranje. Taj omjer može biti 80 %:20 %.



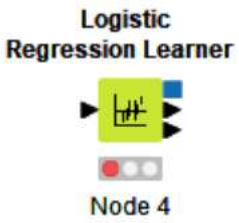
Slika 183. Postavke čvora Missing Values za uklanjanje redova

Upozorenje koje se pojavljuje vezano uz nestandardni PMML (eng. *Predictive Model Markup Language*) ukazuje da uklanjanje redova s praznim ćelijama u stupcima gdje treba biti tekst (*String*) ne pripada u standardni PMML format. Radi se o posebnoj vrsti zapisa koji služi za spremanje modela ili u ovom slučaju pravila filtriranja. Iz čvora **Missing Values** taj zapis je moguće izvesti iz izlaznog priključka koji ima izgled plavog kvadratića i može ga se spremiti u datoteku. Ako se ta datoteka koristi u nekom drugom programu, može doći do problema s kompatibilnošću pri korištenju nestandardnog PMML formata. Ako se povezuju plavi kvadratići, odnosno modeli i upute za filtriranje samo unutar KNIME hodograma, problema s kompatibilnošću ne bi trebalo biti.



Slika 184. Hodogram s čvorovima za učitavanje, filtraciju praznih ćelija i podjelu skupa

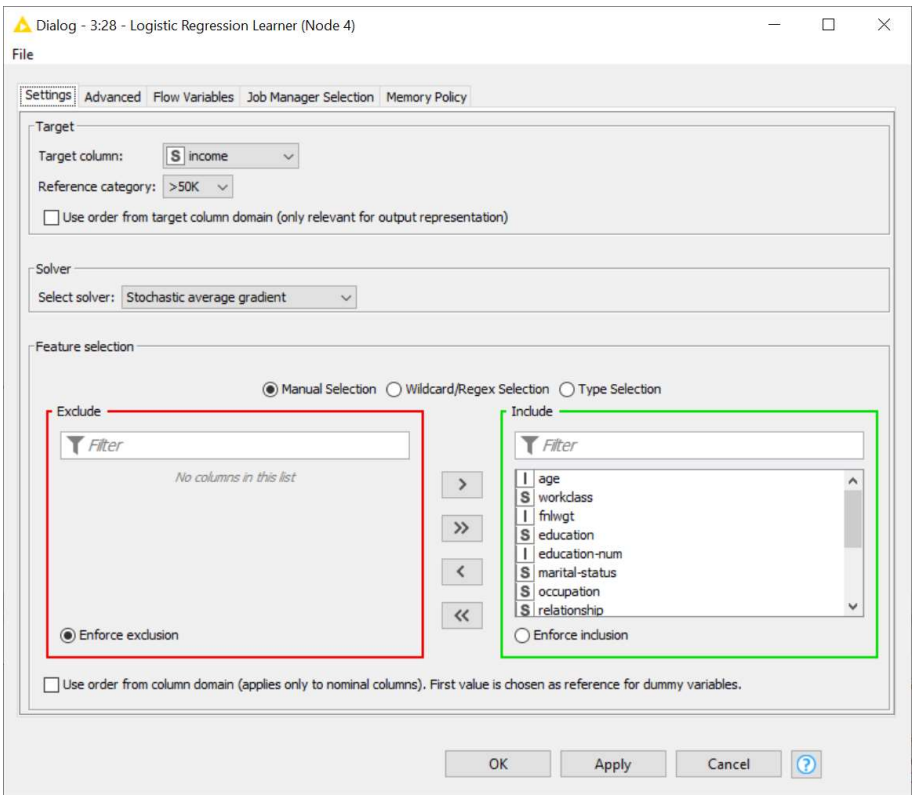
Sljedeći čvor je **Logistic Regression Learner** ili pomalo nespretan hrvatski naziv Učenik logističke regresije. Slika 185 prikazuje izgled čvora.



Slika 185. Čvor Logistic Regression Learner

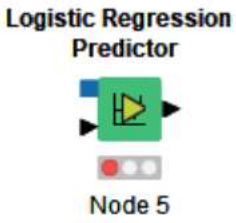
U postavkama čvora **Logistic Regression Learner** u vrhu se definira naziv ciljne varijable. S obzirom da se kod logističke regresije radi o klasifikaciji tu će se u pravilu postavljati stupac koji ima konačan broj numeričkih ili tekstualnih vrijednosti (*String*). Nakon definiranja ciljne varijable, odnosno stupca, u

padajućem izborniku ispod se pojavljuju vrijednosti iz tog stupca te se bira jedna kao referentna vrijednost. Izbor metode (*Solver*) nalazi se ispod u okviru *Solver*. Ponuđene su dvije metode: *Iteratively rewighted least squares* i *Stohastic average gradient*. Bez ulaženja u analizu tih metoda, preporuka je testiranje obje, a izbor ovisi o rezultatima. U ovom primjeru najprije će se koristiti metoda *Stohastic average gradient.* Očekivano, u donjem dijelu dijaloškog okvira u zelenom pravokutniku su značajke koje se uključuju pri treniranju modela, dok su u crvenom pravokutniku značajke koje se isključuju.



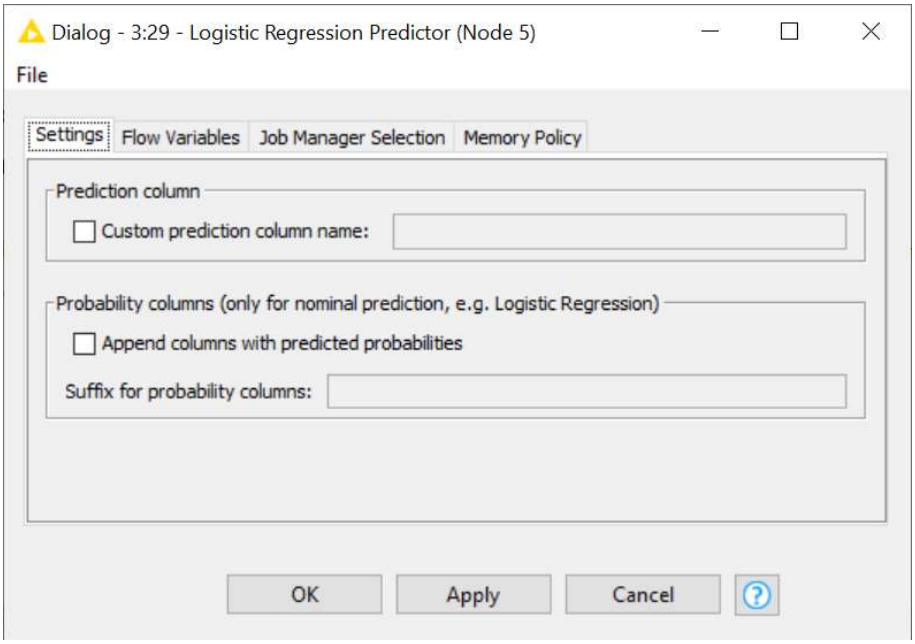
Slika 186. Postavke čvora Logistic Regression Learner

Kao i kod linearne regresije, ako postoji čvor **Logistic Regression Learner**, postoji i **Logistic Regression Predictor** ili Prediktor logističke regresije. Slika 187 prikazuje čvor **Logistic Regression Predictor**.



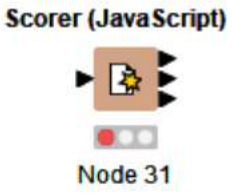
Slika 187. Čvor Logistic Regression Predictor

Postavke čvora **Logistic Regression Predictor** su relativno skromne i svode se na izmjenu stupca ciljne varijable i uključivanja stupaca vjerojatnosti za vrijednosti izračunate modelom. Slika 188 prikazuje postavke čvora **Logistic Regression Predictor.**



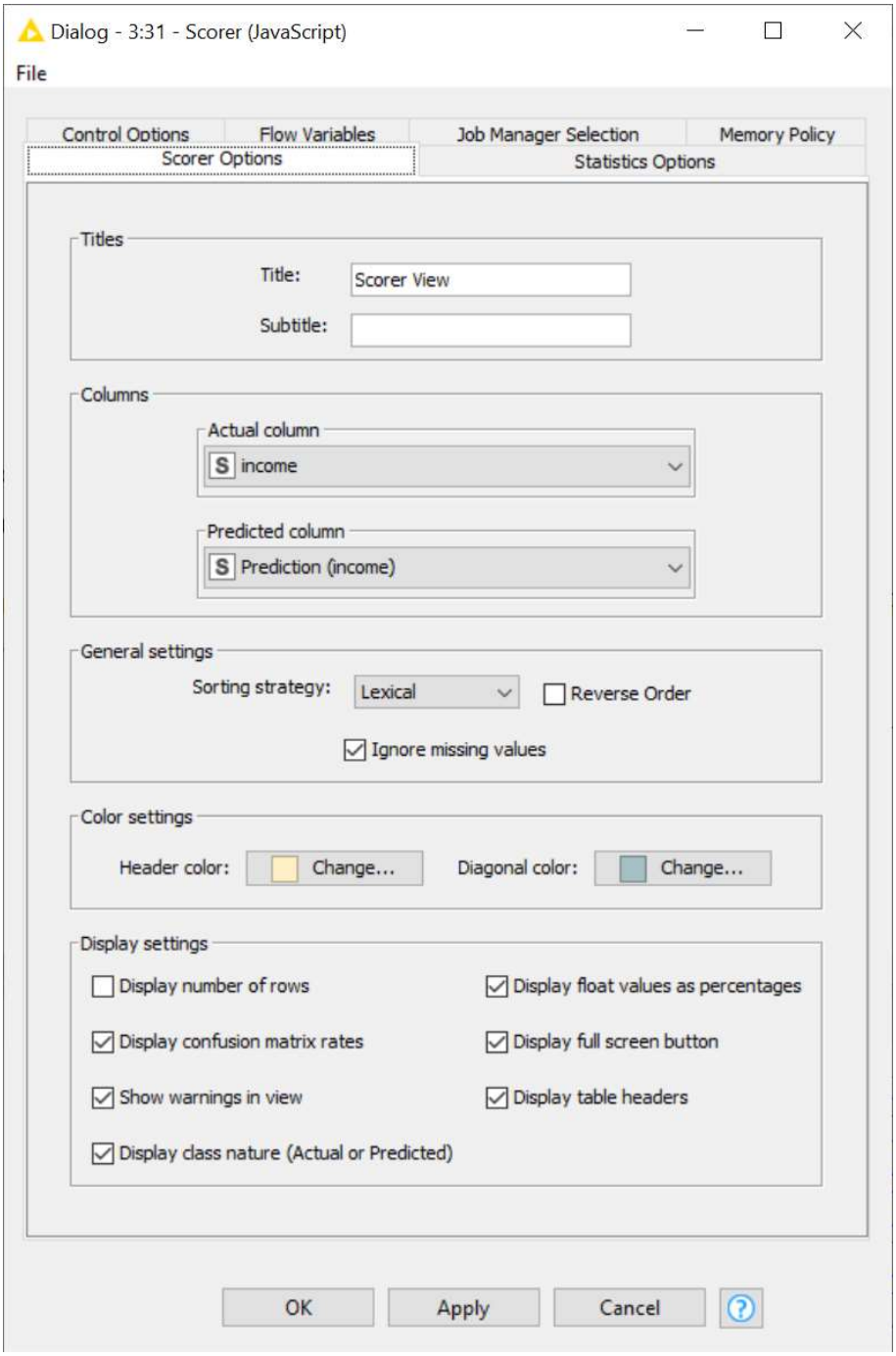
Slika 188. Postavke čvora Logistic Regression Predictor

Zadnji čvor u osnovnom hodogramu je **Scorer (JavaScript)** ili opet nespretan naziv Bilježnik (JavaScript)**.** Kod linearne regresije zadnji čvor bio je **Numeric Scorer**, dok se ovdje koristi sličan čvor, ali s donekle drugačijim karakteristikama. Slika 189 prikazuje čvor.



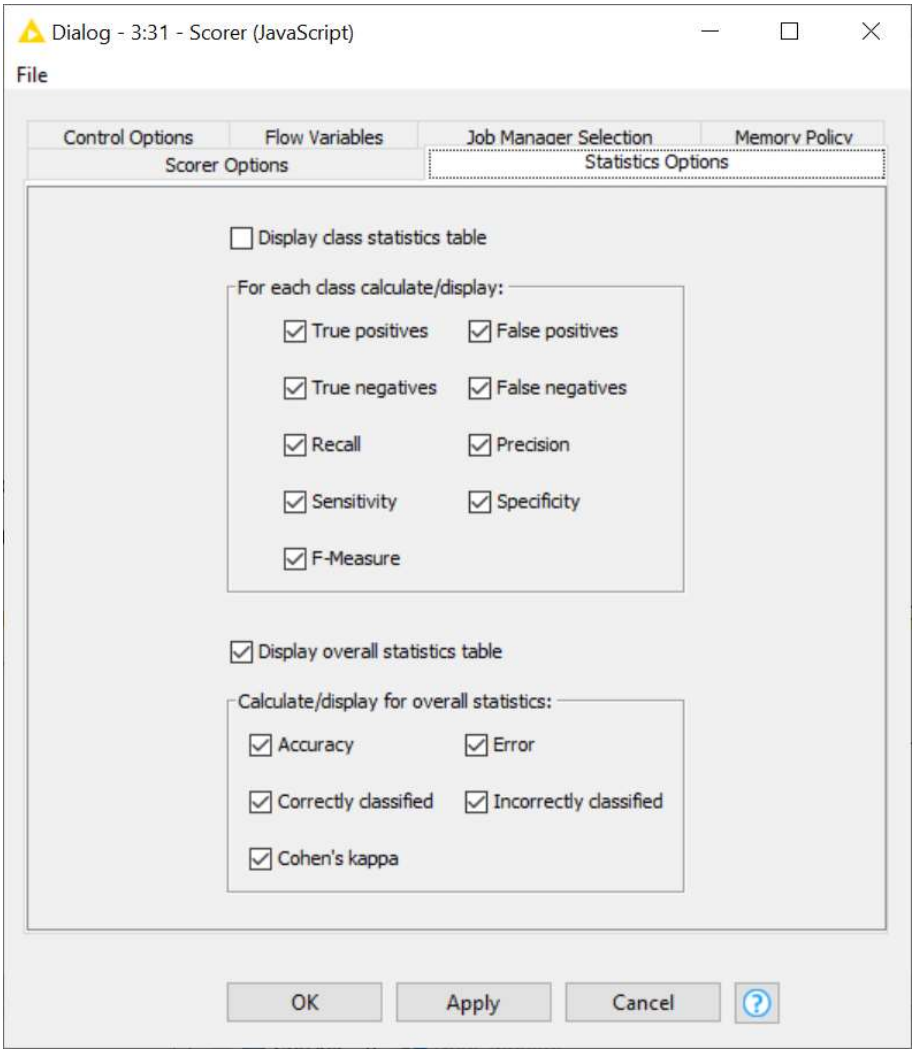
Slika 189. Čvor Scorer (JavaScript)

Postavke čvora omogućuju postavljanje naslova i podnaslova. Nakon toga postoji mogućnost izbora varijable čije vrijednosti su bile u tablici i varijable čije vrijednosti su izračunate modelom. Ispod toga su postavke vezane uz izgled izvještaja pri čemu se može mijenjati način sortiranja, boje i elementi koji se prikazuju u izvještaju. Slika 190 prikazuje izgled dijaloškog okvira na kojem se podešavaju postavke čvora **Scorer (JavaScript).**



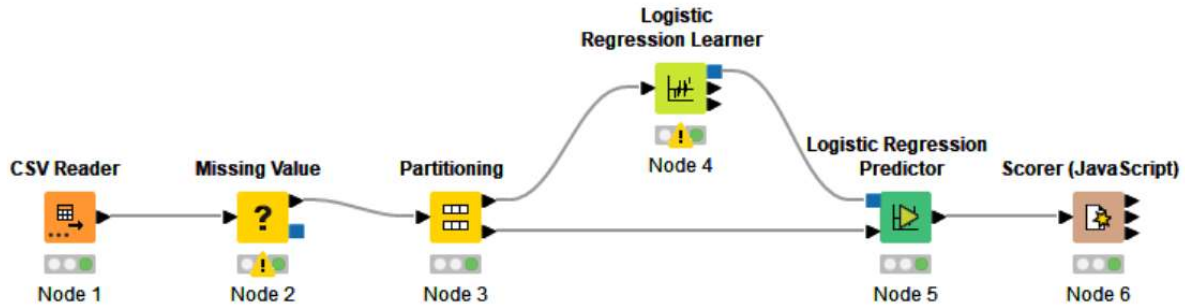
Slika 190. Postavke čvora Scorer (JavaScript)

Osim osnovnih postavki na drugoj kartici (*Statistics Options*) su postavke vezane uz statističke pokazatelje koji će biti dostupni u izvještaju. Slika 191 prikazuje postavke vezane uz pokazatelje performansi modela na čvoru **Scorer (JavaScript).**



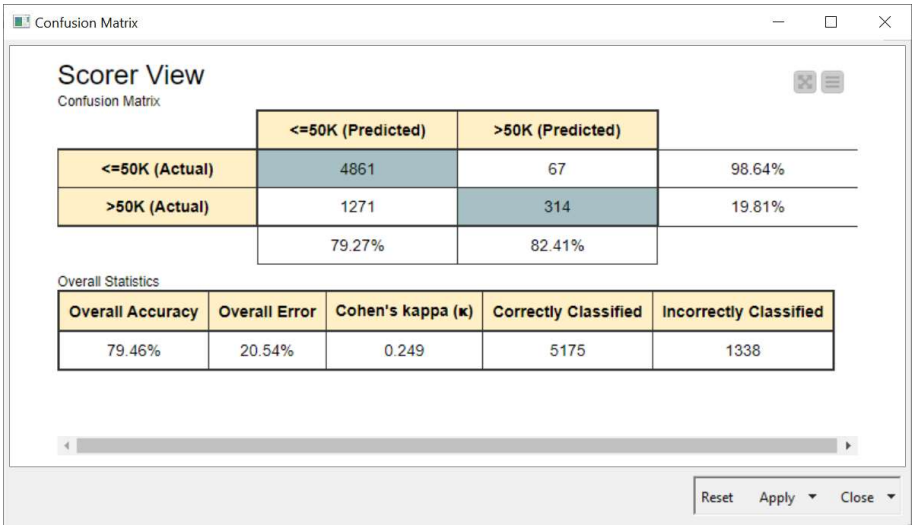
Slika 191. Postavke čvora Scorer (JavaScript) vezane uz statistiku

Slika 192 prikazan je hodogram funkcionalnog modela logističke regresije.



Slika 192. Hodogram modela logističke regresije

Nakon izvršavanja svih čvorova rezultati su dostupni u zadnjem čvoru i to klikom na *Interactive View: Confusion Matrix* u kontekstnom izborniku čvora. Slika 193 prikazuje rezultat.

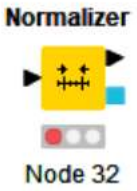


Slika 193. Matrica konfuzije i ostali statistički podaci čvora Scorer (JavaScript)

Ukupna točnost modela je skoro 80 %. Ipak, ako se malo detaljnije pogleda matrica konfuzije vidi se kako je točnost kod predikcije za osobe koje imaju godišnje prihode više od $ 50000 ispod 20 %. Na sreću postoje metode kojima se može poboljšati točnost modela.

7.2. Optimizacija modela logističke regresije

Jedna od metoda koja se često koristi za optimizaciju je normalizacija. Radi se o postupku koji je koristan kod nekih tehnika ako se vrijednosti značajki razlikuju za nekoliko redova veličine. U podacima koji se koriste u primjeru postoji značajka naziva *fnlwgt* čija vrijednost ide do milijun i pol, dok kod značajki *age* vrijednost ide do 90. Ta razlika nekim tehnikama zna predstavljati problem pa se svaki stupac za sebe mora normalizirati. Treba naglasiti da se normalizacija koristi u ovom primjeru, ali može biti korisna i kod drugih tehnika. Slika 194 prikazuje čvor **Normalizer** ili Normalizacija.

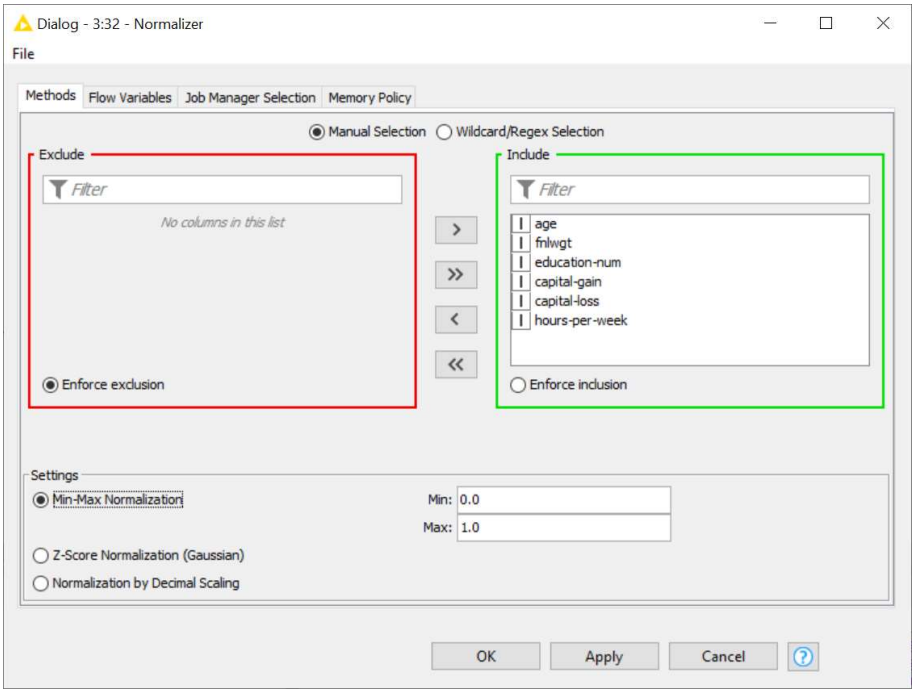


Slika 194. Čvor Normalizer

Slika 195 prikazuje dijaloški okvir za postavke čvora **Normalizer**. U gornjem dijelu dijaloškog okvira za postavke su zeleni i crveni pravokutnik u koje se postavljaju značajke koje se žele, odnosno ne žele normalizirati. U donjem dijelu okvira su ponuđene tri metode normalizacije i to:

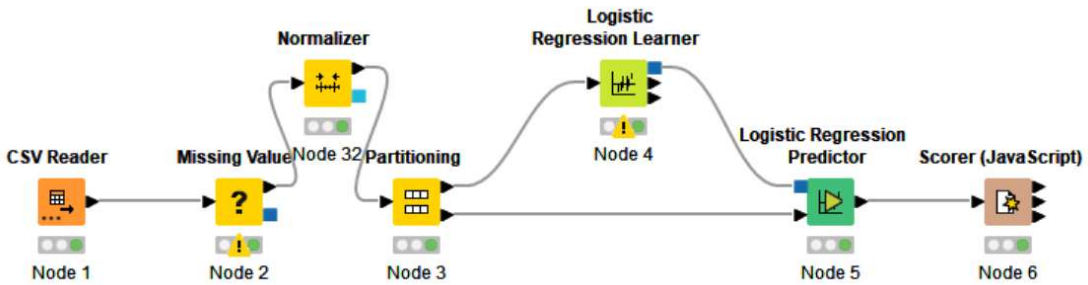
a) Min-max normalizacija - linearna transformacija svih vrijednosti tako da su minimum i maksimum u svakom stupcu definirani s dva polja u koja se unose vrijednosti. U pravilu se postavljaju vrijednosti 0 i 1.
b) Z-score normalizacija - linearna transformacija svih vrijednosti na način da se utvrdi koliko standardnih devijacija je vrijednost udaljena od srednje vrijednosti skupa i u kojem smjeru. Ta vrijednost se upisuje u stupac umjesto postojeće nakon normalizacije s tim da je predznak negativan ako je vrijednost manja od srednje vrijednosti skupa, odnosno pozitivan ako je vrijednost veća od srednje vrijednosti skupa.

c) Decimalna normalizacija - najveća vrijednost u stupcu (i pozitivnom i negativnom) dijeli se j puta s 10 sve dok njegova apsolutna vrijednost ne bude manja ili jednaka 1, da bi se nakon toga sve vrijednosti u stupcu podijelile s 10 na potenciju j.



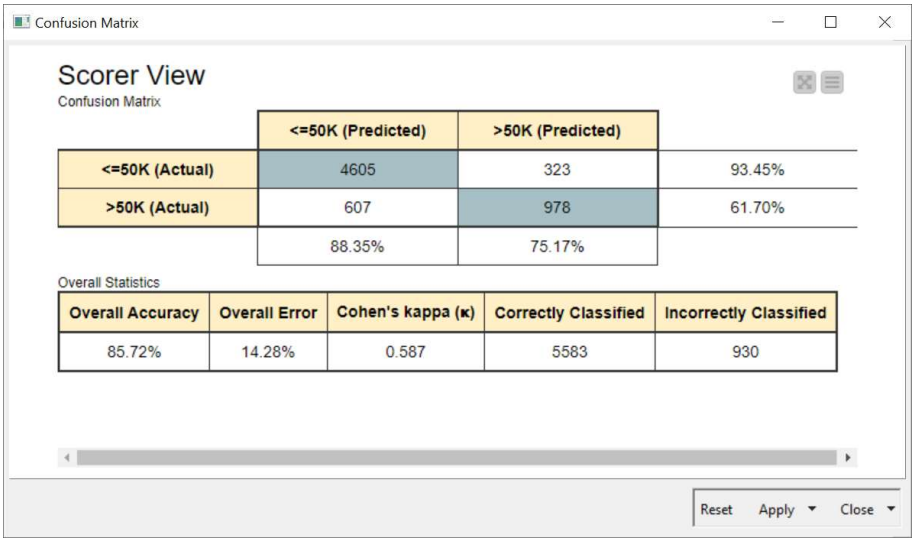
Slika 195. Postavke čvora Normalizer

Čvor za normalizaciju može se umetnuti iza čvora **CSV Reader** ili iza čvora **Missing Values**. Slika 196 prikazuje hodogram s umetnutim čvorom **Normalizer**.



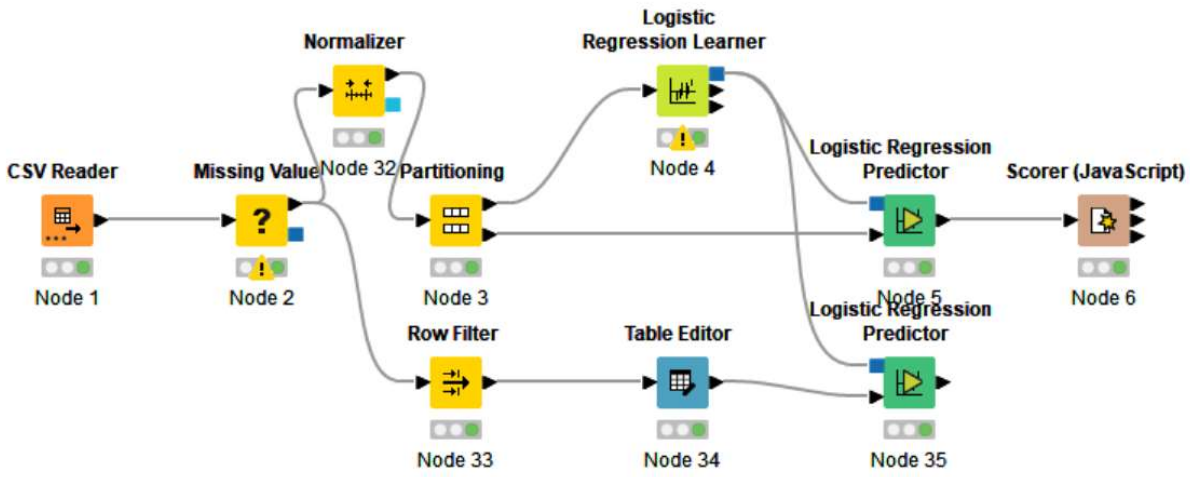
Slika 196. Hodogram s umetnutih čvorom Normalizer

Nakon ponovnog izvršavanja svih čvorova rezultati su dostupni u zadnjem čvoru i to klikom na *Interactive View: Confusion Matrix* u kontekstnom izborniku čvora. Očekivano, vrijednost *Overall Accuracy* je narasla na 85,72 % što znači da se s normalizacijom povećala točnost modela. Prethodna vrijednost bez normalizacije je bila 79,46 %. Značajno je da se normalizacijom povećao postotak točnosti kod predikcije osoba koje imaju godišnje prihode više od $ 50000 s manje od 20 % na više od 60 %.



Slika 197. Matrica konfuzije i ostali statistički podaci nakon normalizacije

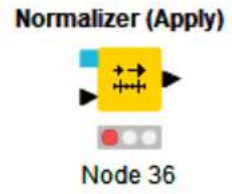
Nakon izrade i treniranja modela mogu se dodati još neki čvorovi koji omogućuju testiranje modela s različitim vrijednostima. Može se dodati čvor **Row Filter** da bi se iz cjelokupnog skupa podataka izlučio samo jedan red. Nakon toga može se umetnuti čvor **Table Editor** za uređivanje vrijednosti u tom jednom redu podataka i može se dodati još jedan čvor **Logistic Regression Predictor** koji se veže za **Logistic Regression Learner**. Čvor **Logistic Regression Predictor** služi za predikciju kategorije ovisno o vrijednostima koje su unesene u red podataka u čvor **Table Editor**, a bazirano na modelu logističke regresije koji je pohranjen u čvoru **Logistic Regression Learner**. Slika 198 prikazuje histogram.



Slika 198. Hodogram nakon umetanja čvorova za testiranje modela

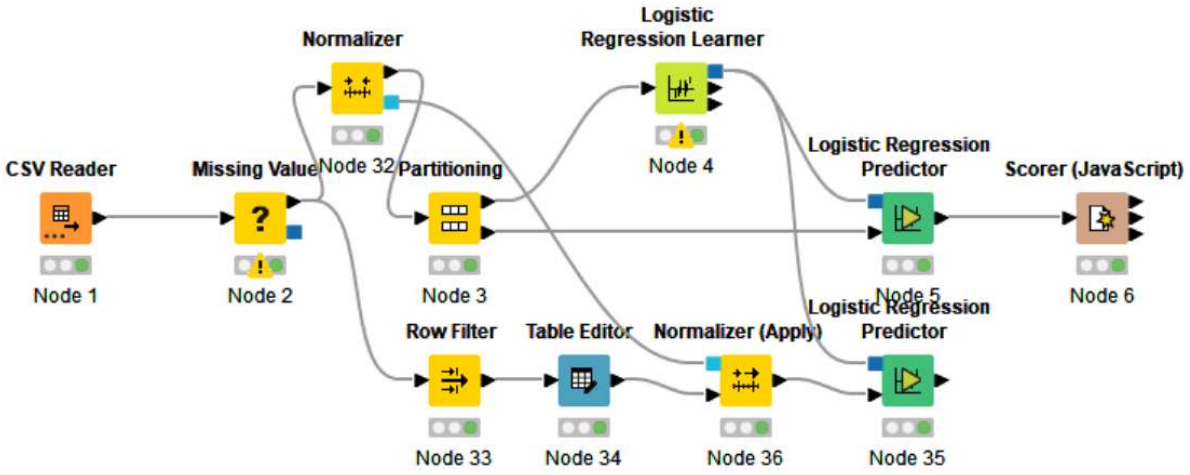
Pažljiviji čitalac će gledajući ovaj hodogram uočiti problem. Model je treniran s normaliziranim podacima, a ako se želi testirati za konkretan slučaj koji se unosi u **Table Editor**, unosit će se stvarne vrijednosti koje nisu normalizirane. To se može testirati tako da se pogleda sadržaj stupaca u tablici iza čvora **Partitioning** i iza čvora **Table Editor**. Dovoljno je pogledati prvi stupac kojem je navedena starost iza čvora **Partitioning**. Tamo su vrijednosti između 0 i 1, pri čemu brojka 1 označava starost od 90 godina. S obzirom da je model treniran s podacima koji su normalizirani na određeni način, jedan red podataka koji se unosi u donji čvor **Logistic Regression Predictor** također mora biti normaliziran pod

istim uvjetima. Za to služi čvor **Normalizer (Apply)** ili Normalizacija (Primjena). Slika 199 prikazuje taj čvor.



Slika 199. Čvor Normalizer (Apply)

Taj čvor nema značajnijih postavki koje treba mijenjati, a ključan detalj je spojiti čvor koji je normalizirao sve podatke skupa i čvor koji normalizira, na osnovu modela normalizacije čvora **Normalizer**. Čvor **Normalizer (Apply)** samo primjenjuje normalizaciju koja je definirana vrstom, ali i vrijednostima koje su se pojavile u svakom stupcu podataka koji su normalizirani u čvoru **Normalizer**. Za spajanje ta dva čvora treba napraviti vezu između svjetlo plavih kvadratića izlaznog priključka čvora **Normalizer** i ulaznog priključka čvora **Normalizer (Apply).**



Slika 200. Hodogram nakon umetanja čvora Normalizer (Apply)

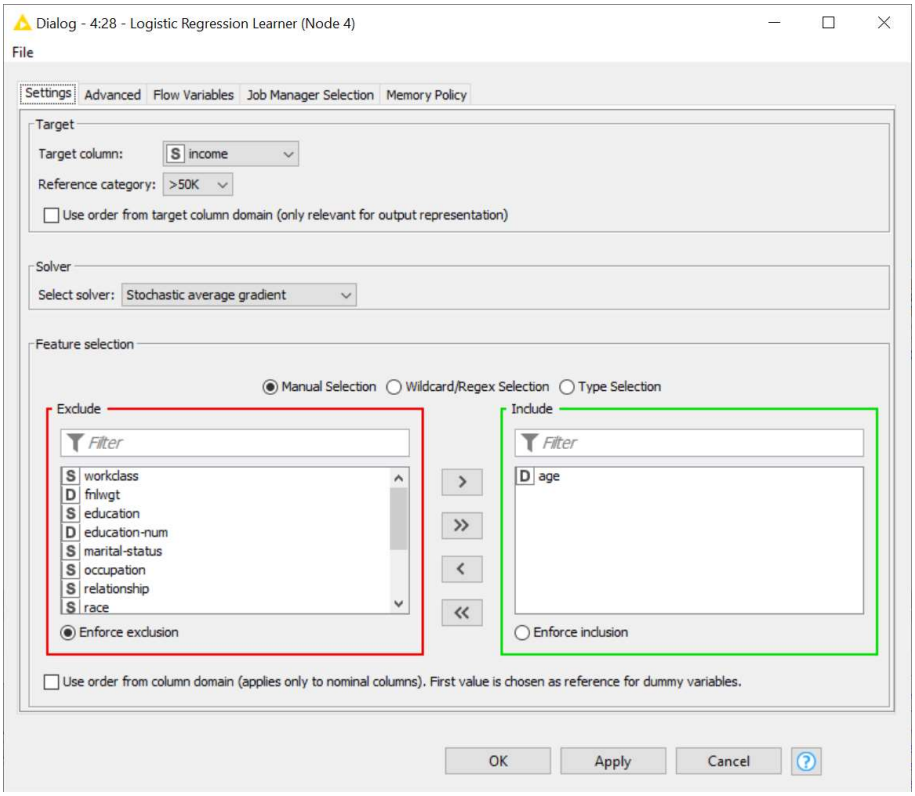
Slika 200 prikazuje hodogram nakon što je umetnut i čvor **Normalizer (Apply)**. Nakon toga se model može testirati s proizvoljnim vrijednostima koje se mijenjaju u čvoru **Table Editor**.

7.3. Trenirani model i logistička funkcija

U teorijskom dijelu na početku poglavlja je navedeno da je temelj modela logističke regresije funkcija, odnosno parametri funkcije koji definiraju model. Funkcija za jednu značajku i jednu ciljanu varijablu glasi ovako:

$$y = \frac{e^{\beta 0 + x * \beta 1}}{1 + e^{\beta 0 + x * \beta 1}}$$

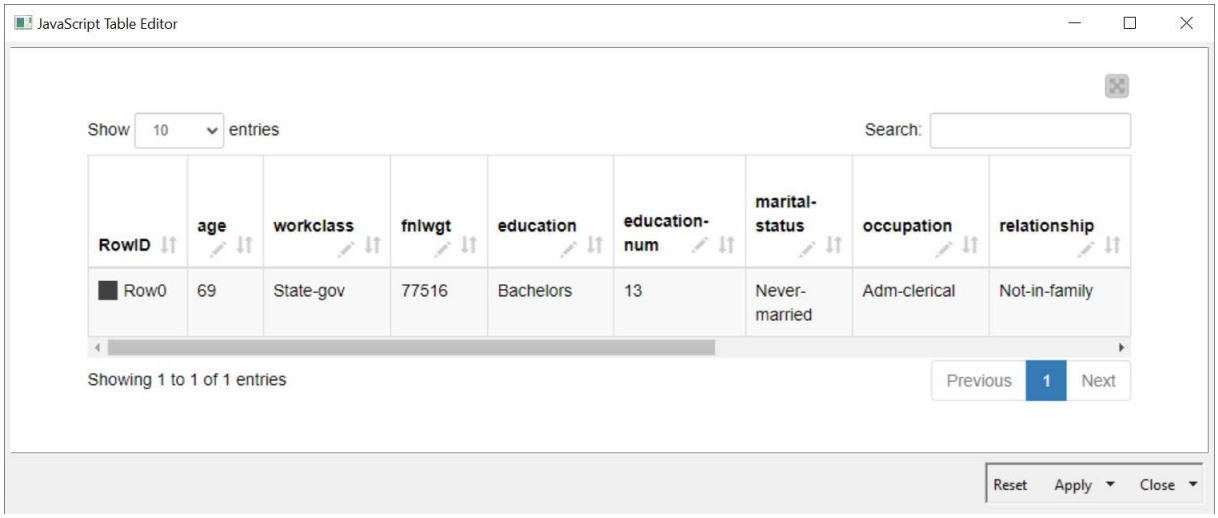
Kako bi se to pokazalo, iskoristit će se prethodno trenirani model, isključit će se sve značajke pri treniranju osim jedne (*age*) i pogledat će se koje su vrijednosti parametara β_0 i β_1. Te vrijednosti unijet će se u gornju funkciju te izračunati y s nekom vrijednošću x te usporediti dobiveni y s onim što je izračunao model. Slika 201 prikazuje postavke čvora **Logistic Regression Learner** gdje se vidi kako je za treniranje modela ostavljena samo značajka *age*, a referentna kategorija je *>50k*.



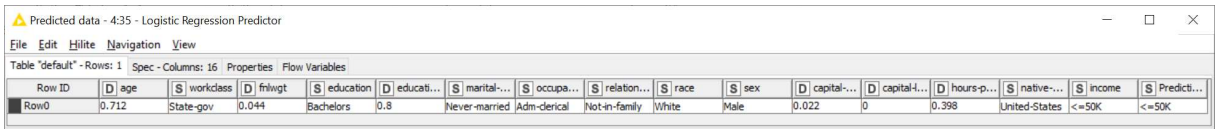
Slika 201. Postavke čvora Logistic Regression Learner za testiranje jednostavnog modela

Nakon izmjena opet se trenira model i provjeravaju rezultati. Očekivano sveukupna točnost je pala na oko 75 % što je i dobro s obzirom da je model treniran sa samo jednom značajkom. Model je opet jako loš s predikcijom osoba s prihodima većim od $ 50000, ali to sada nije važno jer je cilj usporediti vrijednosti koje se dobivaju iz programa KNIME s ručno izračunatim vrijednostima.

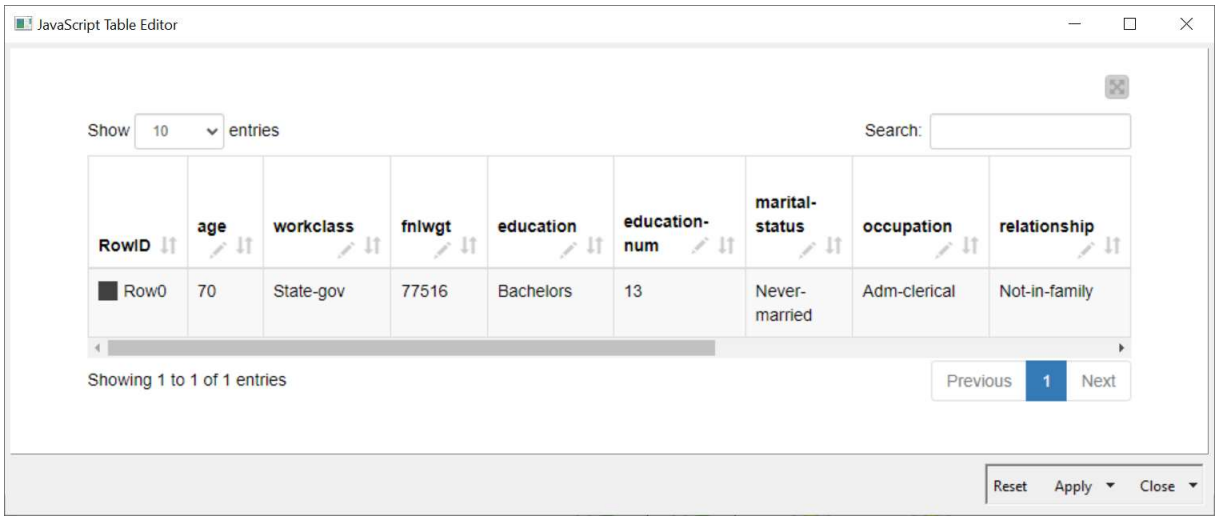
Prvi korak je pronaći vrijednost značajke starost (*age*) koja je granična, odnosno za koju se prihod mijenja iz vrijednosti *<=50k* na *>50k*. Treba uočiti da normalizacijom dobivamo vrijednost varijable starost 0, ako je stvarna starost 17 godina. U slučaju da je stvarna starost 90 godina, onda se normalizacijom ta vrijednost pretvara u 1. To su granične vrijednosti kompletnog skupa podataka, tako da se u ovom slučaju unose razne vrijednosti u čvor **Table Editor** i prate vrijednost nakon normalizacije, kao i vrijednost ciljne varijable. Nakon više pokušaja u ovom slučaju je zaključeno da se za vrijednost varijable starost (*age*) od 69 godina nakon normalizacije dobiva 0,712, a za vrijednost 70 godina dobiva 0,726. Model očigledno između te dvije vrijednosti ima granicu jer za 0,712 modelom predviđeni prihodi pripadaju kategoriji <=50k, dok za 0,726 modelom predviđeni prihodi pripadaju kategoriji >50k. Slika 202 prikazuje unos vrijednosti 69 godina u čvoru **Table Editor**. Slika 203 prikazuje normaliziranu vrijednost starosti u prvom stupcu i modelom predviđenu vrijednost prihoda <=50k u zadnjem stupcu. Slika 204 prikazuje unos vrijednosti 70 godina u čvoru **Table Editor**. Slika 205 prikazuje normaliziranu vrijednost starosti u prvom stupcu i modelom predviđenu vrijednost prihoda >50k u zadnjem stupcu.



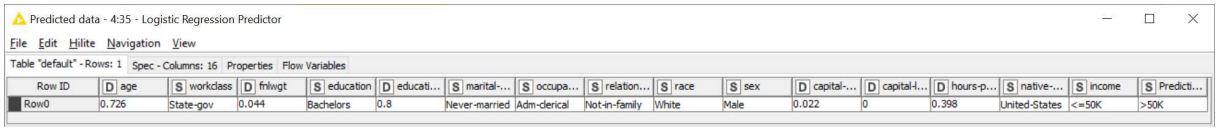
Slika 202. Izmjena vrijednosti varijable age na 69



Slika 203. Predikcija prihoda za vrijednost varijable age od 69 godina



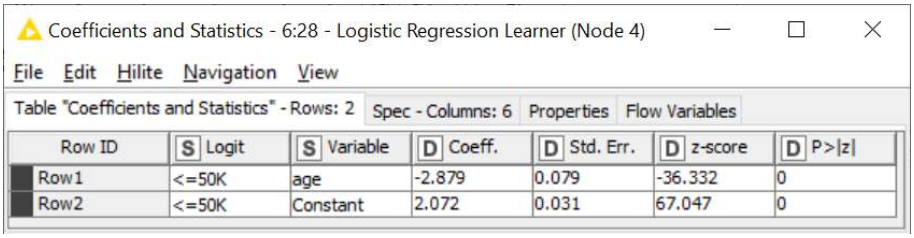
Slika 204. Izmjena vrijednosti varijable age na 70



Slika 205. Predikcija prihoda za vrijednost varijable age od 70 godina

Vratimo se opet funkciji za logističku regresiju koja osim značajke x i ciljne varijable y uključuje varijable β_0 i β_1 za primjer sa samo jednom značajkom. U uvodnom dijelu je navedeno da model mora uključivati varijable β_0 i β_1 te da se na osnovu njih i vrijednosti značajke x koja je u ovom slučaju normalizirana starost, može izračunati y, koji je u ovom slučaju kategorijalna varijabla koja može imati vrijednost <=50k ili >50k. Najlogičnije mjesto gdje bi se mogle potražiti vrijednosti varijabli β_0 i β_1 je u

čvoru **Logistic Regression Learner**. Ako se desnim klikom na taj čvor iz kontekstnog izbornika izabere *Coefficients and Statistics* dobit će se tablica s nizom podataka o modelu. Slika 206 prikazuje tu tablicu.



Row ID	S Logit	S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
Row1	<=50K	age	-2.879	0.079	-36.332	0
Row2	<=50K	Constant	2.072	0.031	67.047	0

Slika 206. Podaci o modelu iz čvora Logistic Regression Learner

U četvrtom stupcu pod nazivom *Coeff.* su dva broja, s tim da je do gornje vrijednosti s lijeve strane riječ *age*, dok je do donje vrijednosti s lijeve strane u ćeliji riječ *Constant*. Očigledno je da je broj -2,879 uz riječ *age* vrijednost varijable β_1, dok je broj 2,072 uz riječ *Constant* vrijednost varijable β_0. Time je dostupno sve potrebno za funkciju modela logističke regresije za ovaj primjer. Ona glasi:

$$y = \frac{e^{2,072-\mathrm{x}*2,879}}{1+e^{2,072-\mathrm{x}*2,879}}$$

Konačno se može izračunati y s tim da se za x moraju unijeti normalizirane vrijednosti starosti. Za x=0,712 se dobiva:

$$y = \frac{e^{2,072-0,712*2,879}}{1+e^{2,072-0,712*2,879}} = 0,5055$$

Pokušat će se i za x=0,726.

$$y = \frac{e^{2,072-\ ,726*2,879}}{1+e^{2,072-0,726*2,879}} = 0,4954$$

S obzirom da je zadana kategorija ciljne varijable <=50k, iz tog razloga za ekvivalent starosti od 69 godina dobiva se 0,5055, dok se za ekvivalent starosti od 70 godina dobiva 0,4954. S obzirom da je granica na 0,5 dobilo se da za starost 69 godina prihod vjerojatnije pripada u zadanu kategoriju (<=50k), dok je za starost od 70 godina prihod vjerojatno pripada alternativnoj kategoriji (>50k). Logistička regresija izračunava vjerojatnost zadane kategorije, a ta vjerojatnost se onda pretvara u vrijednosti 1 ili 0 kod jednostavnog binarnog klasifikatora, odnosno u ovom primjeru u <=50k ili >50k (Brownlee, 2016).

8. Bayesov naivni klasifikator

Bayesov teorem je osnova Bayesovog naivnog klasifikatora, a on se svodi na sljedeće izraze (Buglear, 2010):

$$\text{ako } P(A \text{ i } B) = P(A) * P(B|A)$$

$$\text{onda } P(B|A) = P(A \text{ i } B) / P(A)$$

Drugim riječima, ako je vjerojatnost zbivanja događaja A i B jednaka je vjerojatnosti zbivanja događaja A, pomnožena s vjerojatnošću događaja B uz prethodno zbivanje događaja A, onda je i vjerojatnost zbivanja događaja B pod uvjetom da se dogodio događaj A, jednaka vjerojatnosti događaja A i B podijeljena s vjerojatnošću zbivanja događaja A. Nadalje, ako se zamjene događaji A i B vrijedi i sljedeće:

$$\text{ako } P(B \text{ i } A) = P(B) * P(A|B)$$

$$\text{onda } P(A|B) = P(B \text{ i } A) / P(B)$$

Prethodna dva reda su praktički ista kao prva dva, osim što su zamijenjeni događaji A i B u izrazima. Osim toga može se tvrditi sljedeće:

$$P(A \text{ i } B) = P(B \text{ i } A)$$

Vjerojatnost događaja A i B jednaka je vjerojatnosti događaja B i A. Ovaj izraz je nužan da se umjesto svake strane jednadžbe umetnu vrijednosti iz prvog i trećeg reda navedenih izraza. Tako se dobiva sljedeće:

$$P(A) * P(B|A) = P(B) * P(A|B)$$

Ako se u prethodnom redu s lijeve strane izraza ostavi samo P(A|B), dobiva se sljedeći izraz (Buglear, 2010):

$$\mathbf{P(A|B) = P(A) * P(B|A) / P(B)}$$

Konačno, može se zaključiti: Vjerojatnost događaja A pod uvjetom da se zbio događaj B jednaka je umnošku vjerojatnosti događaja A i događaja B pod uvjetom da se zbio događaj A, podijeljen s vjerojatnošću zbivanja događaja B.

Evo jednostavnog primjera vezano uz ponašanje kupaca. U trgovini je bilo 654 kupca. Od toga su 443 žene i 211 muškaraca. Muškarci su platili 151 put u gotovini, dok su žene platile 53 puta u gotovini. Ostale uplate bile su s karticom. Kolika je vjerojatnost da je osoba koja je uplatila karticom muško ?

Plaćanje karticom se označava s P(B) i iznosi (654-(151+53))/654 = 450/654. Vjerojatnost da je osoba muško je P(A)=211/654, a vjerojatnost da je muška osoba platila karticom, odnosno P(B|A) je (211-151)/211=60/211. Kada se cijeli izraz ispiše dobijemo:

$$P(A|B) = (211/654) * (60/211) / (450/654)$$

$$P(A|B) = (60/654)/(450/654)$$

$$P(A|B) = 60/450$$

$$P(A|B) = 0,13$$

Odgovor glasi, vjerojatnost da je osoba koja je uplatila karticom muško iznosi 13 %.

Dalja transformacija Bayesovog teorema u naivni Bayesov klasifikator izlazi iz okvira ovog priručnika, a fokus će biti na kreiranju i korištenju modela koristeći naivni Bayesov klasifikator.

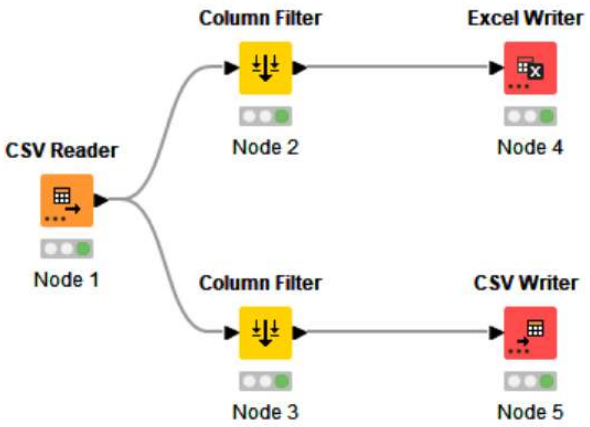
Na kraju upoznavanja s tehnikom treba spomenuti prednosti ove tehnike, a to su mala zahtjevnost što se tiče računalnih resursa za treniranje i mogućnost rada s kategorijalnim i numeričkim značajkama. Osnovni nedostatak je što ova tehnika traži neovisnost značajki, a to se rijetko susreće u skupovima podataka (Egger, 2022).

8.1. Priprema podataka

Nakon kratkog uvoda i objašnjenja Bayesovog teorema prelazi se na primjenu naivnog Bayesovog klasifikatora. Primjena će biti prikazana kroz rješavanje jednog od problema tvrtki koje se bave uslugama, a to je odljev korisnika. Tvrtke koje se bave telekomunikacijskim uslugama često pokušavaju pridobiti korisnike da se pretplate na njihovu uslugu. Nakon isteka pretplate dio korisnika ne obnavlja pretplatnički ugovor, a tvrtkama je važno prepoznati te korisnike, unaprijed se pripremiti na takvu situaciju i tim korisnicima ponuditi neku dodatnu pogodnost kako bi produžili pretplatu. Pretpostavka je da bi se ti korisnici mogli prepoznati i klasificirati na osnovu prethodno dostupnih podataka. Potrebni podaci za navedeni primjer su dostupni na adresi: https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset . Potrebna je prijava za preuzimanje skupa podataka što se može odraditi s Google računom.

Nakon preuzimanja komprimirane datoteke istu je potrebno dekomprimirati te se dobije jedna CSV datoteka. Da bi se kroz ove primjere steklo što više znanja i vještina, za rješavanje ovog problema kreirat će se dodatni hodogram kojim će se učitane podatke podijeliti u dvije tablice te ih spremiti u različitim formatima. Nakon toga će se kod kreiranja glavnog hodograma spojiti podatke iz te dvije tablice i kreirati model naivnog Bayesovog klasifikatora. Treba obratiti pažnju da je ovaj dodatni hodogram potpuno suvišan i da bi se u glavnom hodogramu jednostavno moglo samo učitati podatke iz CSV datoteke, ali za navedeni primjer je čest slučaj da se osnovni podaci o korisnicima telekomunikacijskih usluga i podaci o njihovom ponašanju čuvaju u odvojenim datotekama, odnosno tablicama u bazi. Baš iz tog razloga se razdvajaju podaci, kako bi se u stvarnoj situaciji znali povezati. Pri izradi dodatnog hodograma koristit će se dva nova čvora te će i oni biti opisani.

Slika 207 prikazuje strukturu dodatnog hodograma. Pažljivijim čitateljima već je vjerojatno jasno čemu služe prikazani čvorovi, a i hodogram u cjelini. Čvor **CSV Reader** je već opisan prethodno i on učitava podatke iz CSV datoteke. Ti podaci se u cijelosti dovode do dva čvora pri čemu svaki filtrira dio stupaca. Jedan dio stupaca se sprema kao XLSX datoteka korištenjem čvora **Excel Writer**, dok se drugi dio stupaca sprema u CSV datoteku korištenjem čvora **CSV Writer**. Treba naglasiti da su dva stupca prisutna u obje izlazne datoteke, a ti stupci će poslužiti u glavnom hodogramu za spajanje dvaju datoteka stvorenih u ovom dodatnom hodogramu.



Slika 207. Struktura dodatnog hodograma

Prije analize dva nova čvora koji su prikazani na dodatnom hodogramu, preporučljivo je pogledati podatke koji su učitani korištenjem **CSV Reader** čvora. Dovoljno je kroz kontekstni izbornik dobiven desnim klikom na čvor **CSV Reader** izabrati *File Table*. Dobiva se tablica s nizom stupaca od kojih su neki važni za predikciju, dok su neki potpuno nevažni. Stupci *area code* i *phone number* služe za identifikaciju korisnika jer su jedinstveni kad ih se koristi u kombinaciji. Osim toga u nekoliko stupaca su navedene karakteristike pretplate koju korisnik koristi, a zatim i brojčane vrijednosti koje opisuju navike korisnika pri korištenju telefonske usluge. Posljednji stupac pod nazivom *churn* označava je li korisnik prethodno pripadao u kategoriju korisnika koji nisu produžili pretplatu. Taj stupac bit će ciljna varijabla, dok su ostali stupci značajke koje će se koristiti za treniranje modela.

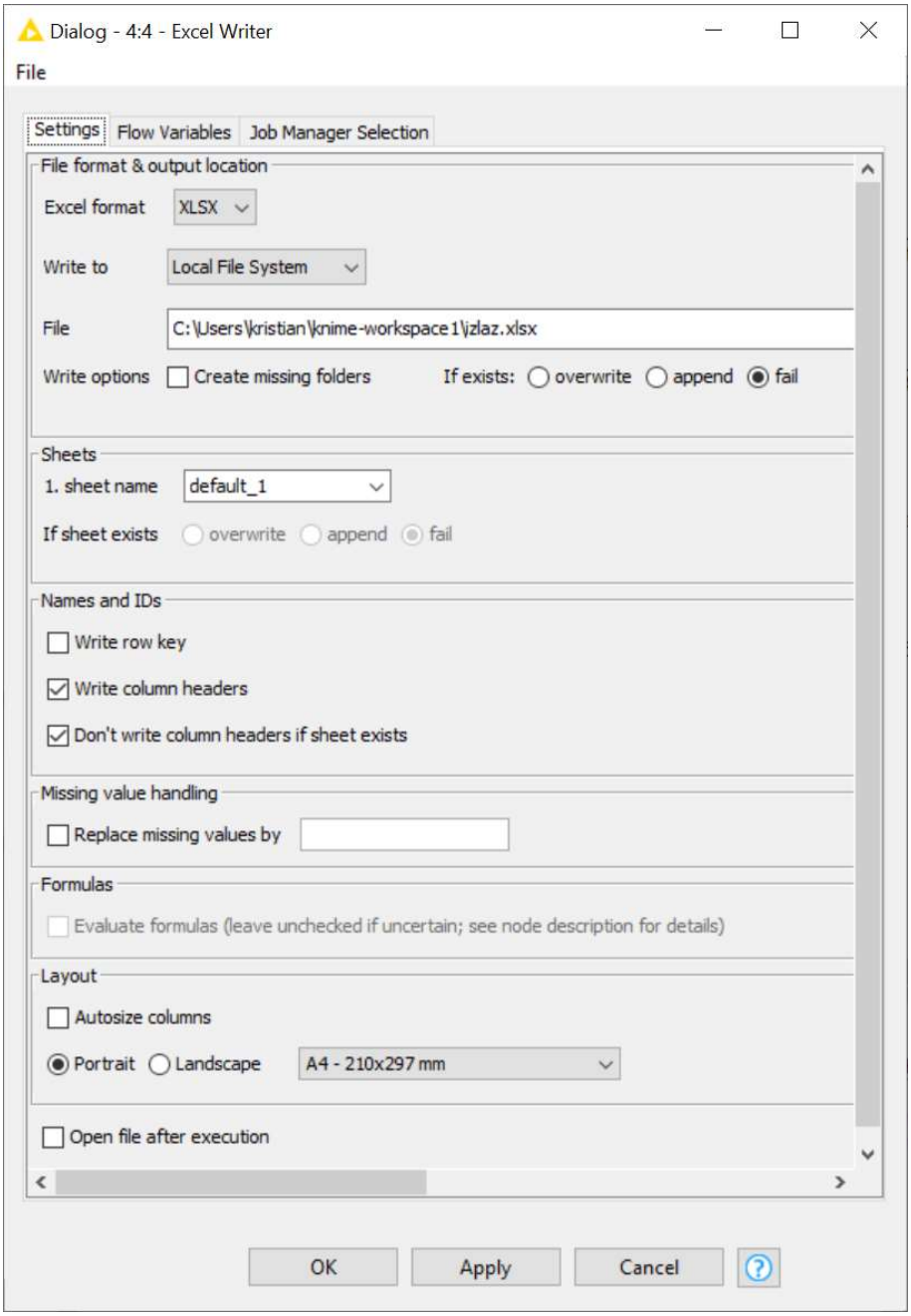
U nastavku će biti opisani novi čvorovi koji se pojavljuju u dodatnom hodogramu. Slika 208 prikazuje čvor **Excel Writer** ili Excel pisač koji služi za spremanje tabličnih podataka u radnu knjigu u MS Excel formatu.

Slika 208. Čvor Excel Writer

Slika 209 prikazuje postavke čvora **Excel Writer**. Prvi padajući izbornik omogućuje izbor formata u kojem će tablični podaci biti spremljeni. Ponuđena su dva formata, XLS i XLSX. Stariji XLS format je još uvijek podržan od novih verzija MS Excel-a, a noviji XLSX format je uveden od MS Excel-a 2007. Ispod tog izbornika bira se lokacija i naziv izlazne datoteke te opcije vezane uz prepisivanje radne knjige s istim imenom.

Izbor imena radnog lista omogućen je u okviru *Sheets* prve kartice postavki (eng. *Sheet Name*), kao i opcije vezane uz prepisivanje podataka ako navedeno ime radnog lista već postoji. Okvir *Names and IDs* omogućuje izbor vezan uz zapisivanje brojeva redova (eng. *Write Row Key*), zaglavlja (eng. *Write Column Headers*) i opcije vezane uz zapisivanje zaglavlja u slučaju postojanja radnog lista navedenog imena.

U donjoj polovici prve kartice postavki nalaze se mogućnosti zadavanja sadržaja u ćelije koje su prazne, evaluacije formula, formata radnog lista i otvaranja radne knjige odmah nakon izvršavanja čvora.

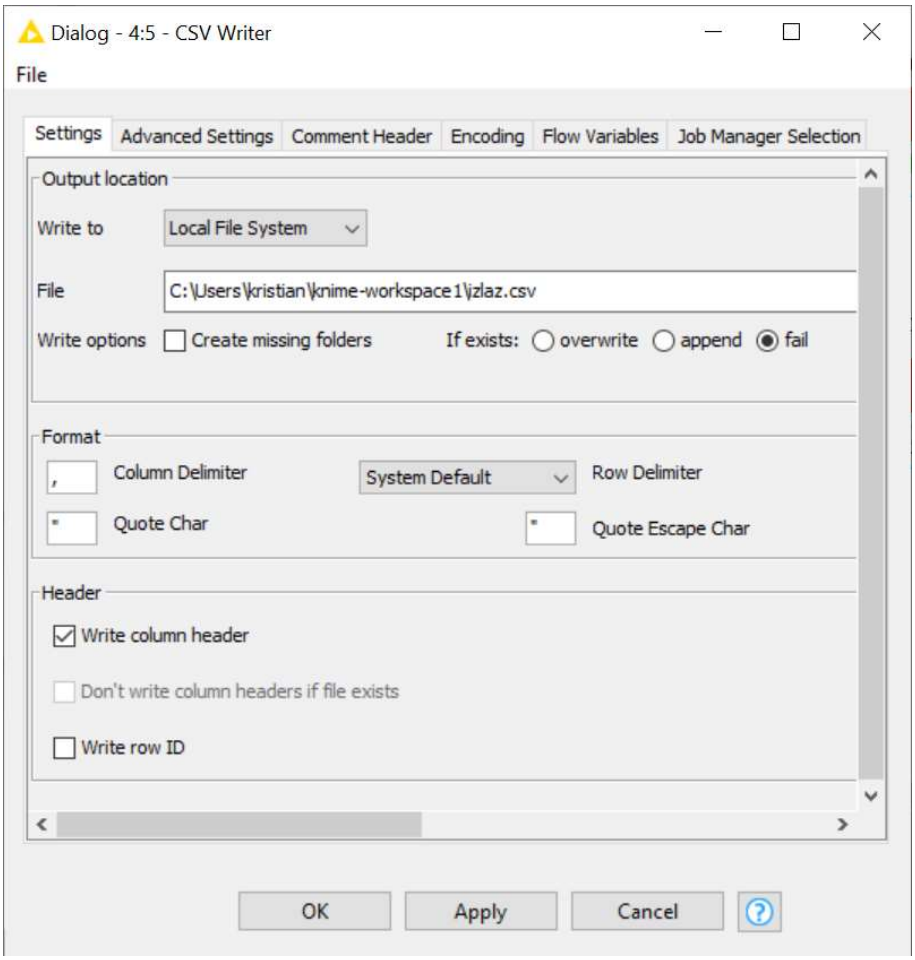


Slika 209. Postavke čvora Excel Writer

Slika 210 prikazuje čvor **CSV Writer** ili CSV pisač. Navedeni čvor služi za zapisivanje tabličnih podataka u CSV formatu.

Slika 210. Čvor CSV Reader

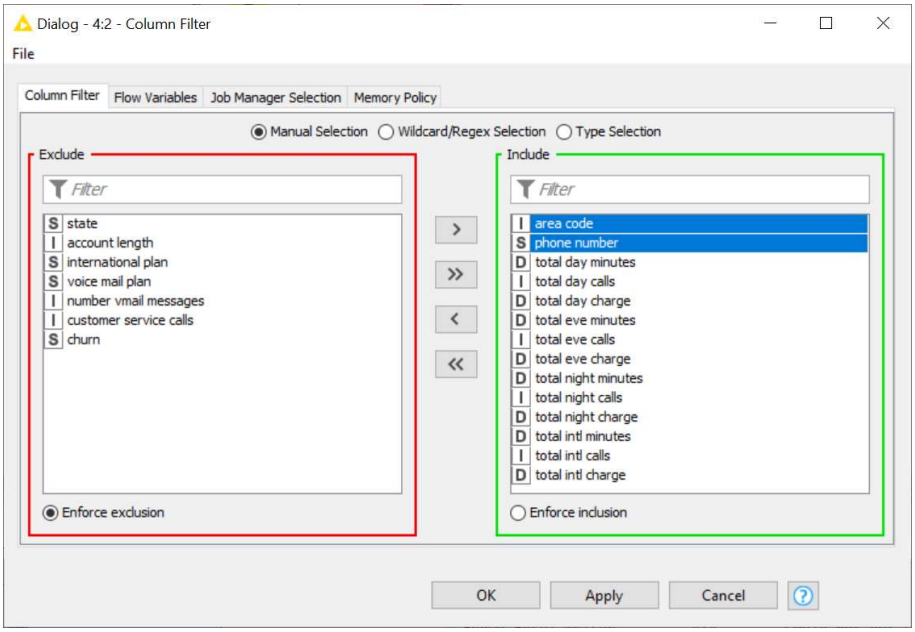
Slika 211 prikazuje postavke čvora CSV Writer. Prvi dio (*Output location*) vrlo je sličan postavkama čvora Excel Writer osim što ne dozvoljava izbor formata zapisa, koji je zadano CSV. U postavkama formata najvažnije su stavke *Column Delimiter* kojom se definira znak koji dijeli sadržaj unesen u jednom redu po pripadajućim stupcima i Row Delimiter kojom se definira niz koji označava prelazak u sljedeći red. U donjem dijelu moguće je izabrati zapisivanje zaglavlja i identifikatora redova.



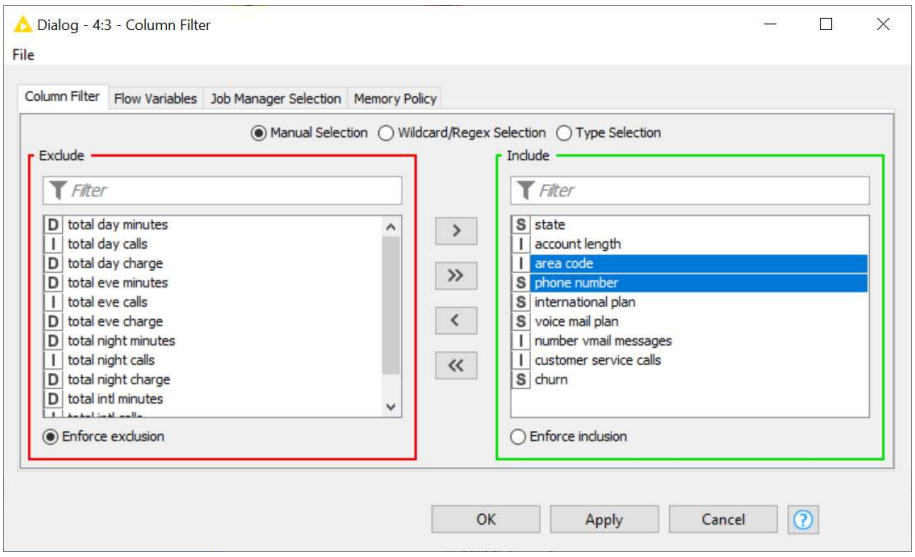
Slika 211. Postavke čvora CSV Writer

Prije izrade glavnog hodograma važno je naglasiti na koji način su podijeljeni stupci izvorne datoteke s podacima. Kao što je navedeno, dio podataka je zapisan u XLSX datoteku, dok je dio zapisan u CSV datoteku. Izbor stupaca definiran je s dva čvora **Column Filter**. Slika 212 prikazuje postavke filtriranja stupaca prije čvora **Excel Writer**, pri čemu se vidi kako su uključeni stupci *area code* i *phone number* koji identificiraju pretplatnika u obje izlazne tablice. Osim ta dva stupca u XLSX datoteku izvezeni su svi stupci čiji nazivi počinju s riječi *total*. Slika 213 prikazuje postavke filtriranja stupaca prije

čvora **CSV Writer**. Tu se opet vide stupci *area code* i *phone number* za identifikaciju pretplatnika, osim toga uključeni su svi ostali stupci čije ime ne počinje s riječju *total*.



Slika 212. Postavke čvora Column Filter prije čvora Excel Writer

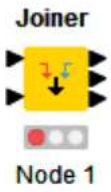


Slika 213. Postavke čvora Column Filter prije čvora CSV Writer

Pokretanjem svih čvorova dodatnog hodograma dobivaju se dvije nove datoteke, jedna je u formatu XLSX, a druga u formatu CSV. Obje će biti potrebne u glavnom hodogramu.

8.2. Izrada modela temeljenog na Bayesovom naivnom klasifikatoru

Hodogram za model baziran na Bayesovom naivnom klasifikatoru ne razlikuje se bitno od prethodno kreiranih hodograma linearne i logističke regresije. Na početku se učitavaju podaci, ali s obzirom da su u dodatnom hodogramu podaci podijeljeni u dvije datoteke, moraju se učitati iz te dvije datoteke i spojiti u jednu tablicu. Pri tom se koristi dosad nepoznat čvor koji se zove **Joiner** ili Sastavljač.



Slika 214. Čvor Joiner

Slika 214 prikazuje čvor **Joiner** koji služi za spajanje dvije tablice u jednu. Slika 215 prikazuje postavke čvora **Joiner**. U gornjem dijelu postavki biraju se stupci na osnovu kojih se spajaju redovi iz dvije tablice. U ovom slučaju moraju se poklapati pozivni brojevi (*Area Code*) i brojevi telefona (*Phone*). Treba uočiti kako je u prethodnoj rečenici između naziva stupaca veznik „i". Osim njega na tom mjestu može biti i veznik „ili", pri čemu bi se spojili redovi koji sadrže bar jedan isti par podataka iz navedenih stupaca. Je li se koristi logično „ili" ili logično „i" bira se na vrhu kartice postavki u nastavku riječi *Match* pri čemu *all of the following* pretpostavlja korištenje logičkog „i" u navedenim uvjetima, dok *any of the following* pretpostavlja korištenje logičkog „ili" u navedenim uvjetima. Za dva reda koja se spajaju, red iz lijeve tablice mora imati istu vrijednost u izabranom stupcu/stupcima kao red iz desne ulazne tablice u stupcu/stupcima izabranom s desne strane. Treba obratiti pažnju kako je u ovom primjeru izabrana mogućnost *all of the following*, jer bi druga mogućnost generirala ogroman broj kombinacija koje su potpuno bespotrebne.

Pri usporedbi vrijednosti kod spajanja u gornjem dijelu kartice mogućnosti čvora i pri dnu okvira *Join columns* može se izabrati uvjet da sadržaj ćelije mora biti isti po vrijednosti i tipu (*Compare values in join columns by value and type*). Drugim riječima, ako postoji brojčana vrijednost 1 spremljena kao tekstualni i cjelobrojni podatak u dvije ćelije koje se uspoređuju uz izabran navedeni uvjet, KNIME će ignorirati taj par i neće sadržaj te dvije ćelije prepoznati kao iste. Druga mogućnost je usporedba nakon konverzije u tekstualnu vrijednost (*string representation*), a za navedeni primjer s brojem 1 KNIME bi prepoznao te dvije ćelije kao iste. Treća mogućnost (*making integer types compatible*) služi za ignoriranje različitih formata zapisa brojčanih vrijednosti i usporedbu bez obzira na format.

U srednjem dijelu svojstava čvora (*Include in output*) dostupne su tri mogućnosti i kombinacije među njima. To su:

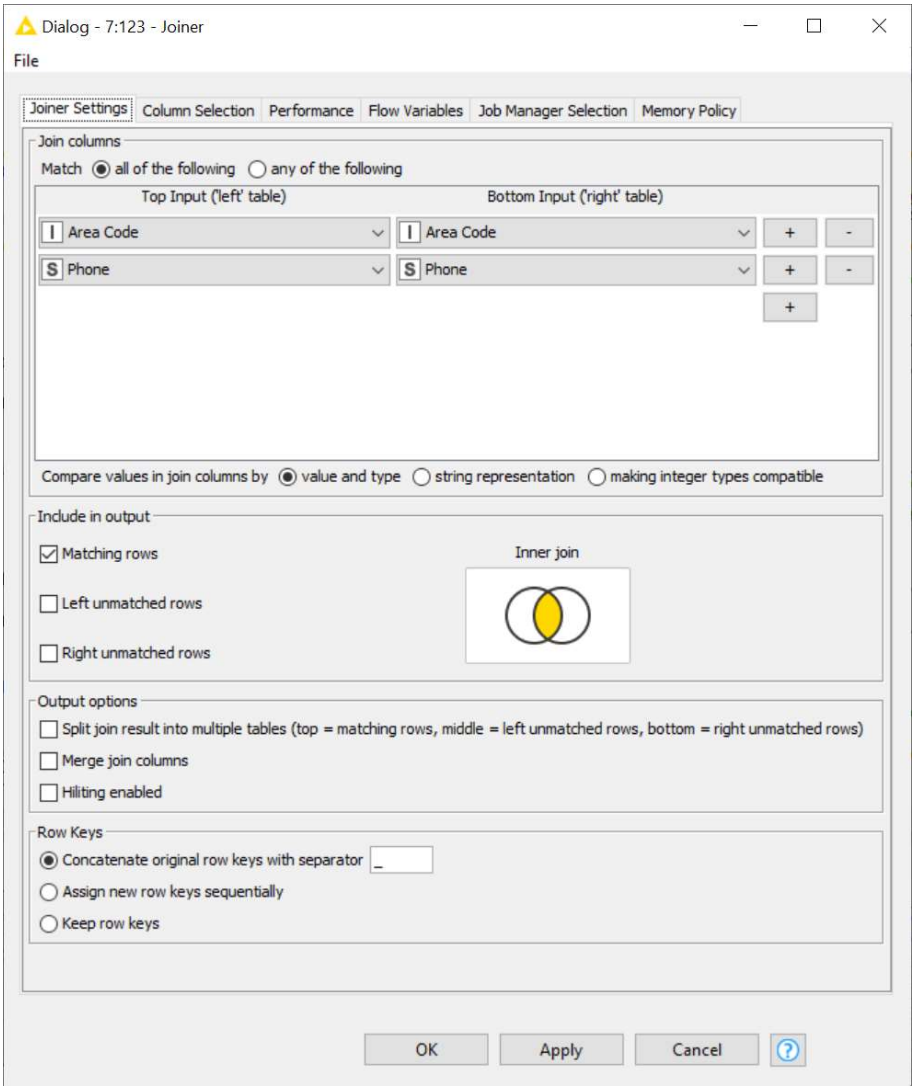
a) *Matching rows* – uključenjem ove opcije u izlaznu tablicu uključeni su redovi čiji sadržaj je u zadanim stupcima isti, a isključenjem se mogu dobiti raspareni redovi koji su preostali nakon uspoređivanja.
b) *Left unmatched rows* – uključenjem ove opcije u izlaznu tablicu uključuju se redovi iz lijeve tablice koji nisu upareni s redovima iz desne tablice.
c) *Right unmached rows* – uključenjem ove opcije u izlaznu tablicu uključuju se redovi iz desne tablice koji nisu upareni s redovima iz lijeve tablice.

U donjem dijelu svojstava čvora (*Output options*) dane su mogućnosti:

a) *Split join result in the multiple tables* – uključenjem ove opcije čvor će kreirati tri izlazne tablice umjesto jedne. Gornji izlazni port uključit će spojene retke koji zadovoljavaju zadani uvjet, srednji izlazni port uključivat će rasparene retke iz lijeve ulazne tablice, a donji port uključivat će rasparene retke iz desne ulazne tablice.
b) *Merge join columns* – uključenjem ove opcije dobit će se novi nazivi stupaca za uparene vrijednosti.

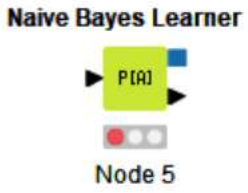
c) *Hiliting enable* – uključenjem ove opcije omogućuje se brisanje redova koje se proteže kroz sam čvor. Ako nije nužno ovu opciju je bolje ostaviti isključenom iz razloga što se njenim uključivanjem dodatno troše memorijski resursi računala.

Donji dio svojstava čvora u zadnjem okviru (*Row keys*) omogućuje izmjenu naziva ključeva za redove. Izabrana je prva opcija (*Concatenate original row keys with separator*) koja spaja ključeve iz prethodnih tablica dok između njih stavlja znak „_”. Druga opcija (*Assign new row keys sequentially*) je dodavanje novih ključeva redova, a treća (*Keep row keys*) je ostavljanje postojećih.



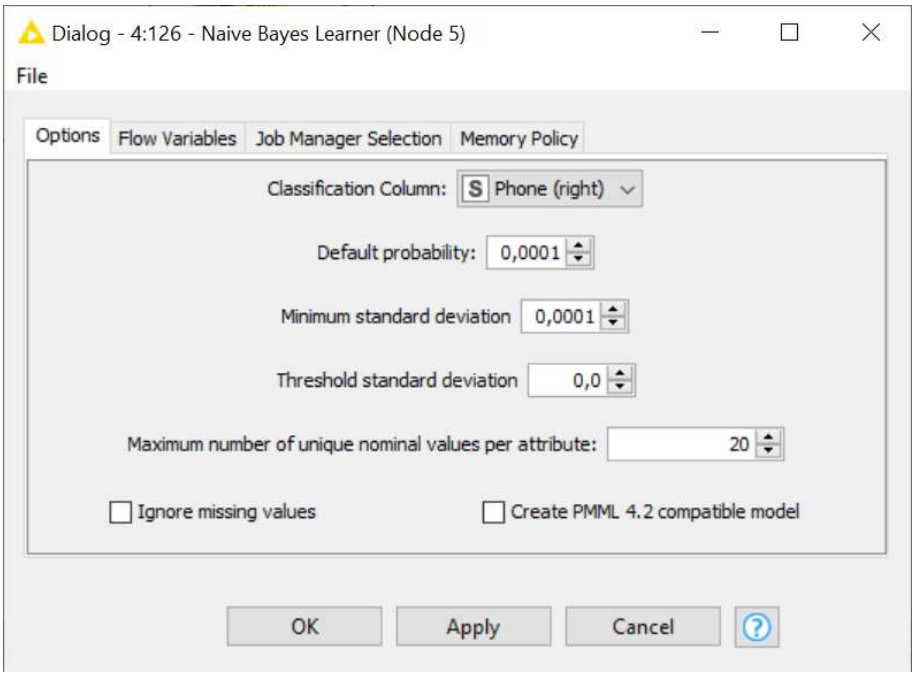
Slika 215. Postavke čvora Joiner

Nakon spajanja podataka iz dvije tablice korištenjem čvora **Joiner** hodogram se završava sljedećim čvorovima: **Partitioning**, **Naïve Bayes Learner, Naïve Bayes Predictor** i **Scorer (JavaScript)**. Od navedenih, dva čvora su nova. Slika 216 prikazuje čvor **Naïve Bayes Learner** ili Naivni Bayesov učenik.



Slika 216. Čvor Naïve Bayes Learner

Slika 217 prikazuje postavke čvora. Prvi padajući izbornik nudi moguće stupce za ciljnu varijablu. U ovom primjeru ciljna varijabla je stupac *Churn* koji označava je li korisnik spada u skupinu koja u prošlosti nije produžavala ugovor nakon isteka. Problem koji se ovdje pojavljuje je taj da stupac Churn nije ponuđen u padajućem izborniku. Razlog je, što naivni Bayesov klasifikator traži za ciljanu vrijednost kategorijalnu varijablu, a stupac *Churn* je cjelobrojna varijabla. Nakon opisa postavki čvora taj problem se može riješiti dodavanjem još jednog čvora za konverziju vrijednosti stupca iz cjelobrojnih u tekstualne vrijednosti koje tehnika prihvaća.

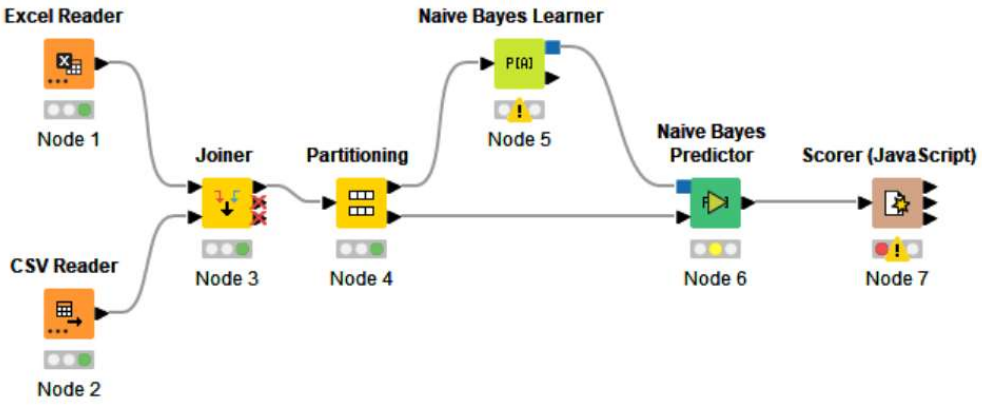


Slika 217. Postavke čvora Naïve Bayes Learner

Vrijednost *Default probability* mora biti mala i veća od nule i koristi se umjesto nulte vrijednosti za zadani par atribut/kategorija. Preporuka je ostaviti zadano. Vrijednost *Minimum standard deviation* koristi se za opažanja bez dovoljno različitih podataka. Vrijednost mora biti najmanje 1E-10, a preporuka je ostaviti zadano. Vrijednost *Maximum number of unique nominal values per attribute* definira vrijednost nakon koje se stupci s više nominalnih vrijednosti od zadanog broja preskaču kod treniranja modela. Posljednje dvije opcije tiču se ignoriranja redova s praznim ćelijama i kreiranje PMML 4.2 kompatibilnog modela. Kompatibilnost modela je važna ako se model izvozi u druge programe, inače se može zanemariti.

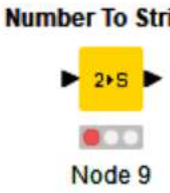
Preostaje rješavanje problema postavljanja numeričke varijable za ciljnu varijablu u postavkama čvora **Naïve Bayes Learner**. Radi se o stupcu *Churn*. Spomenuti stupac nije bio ponuđen, a razlog je što su vrijednosti u navedenom stupcu cjelobrojne. Da bi se te vrijednosti konvertirale u tekstualne

(*String*), jer tehnika to zahtjeva, koristi će se čvor **Number To String**. Slika 218 prikazuje hodogram koji treba nadograditi sa spomenutim čvorom.

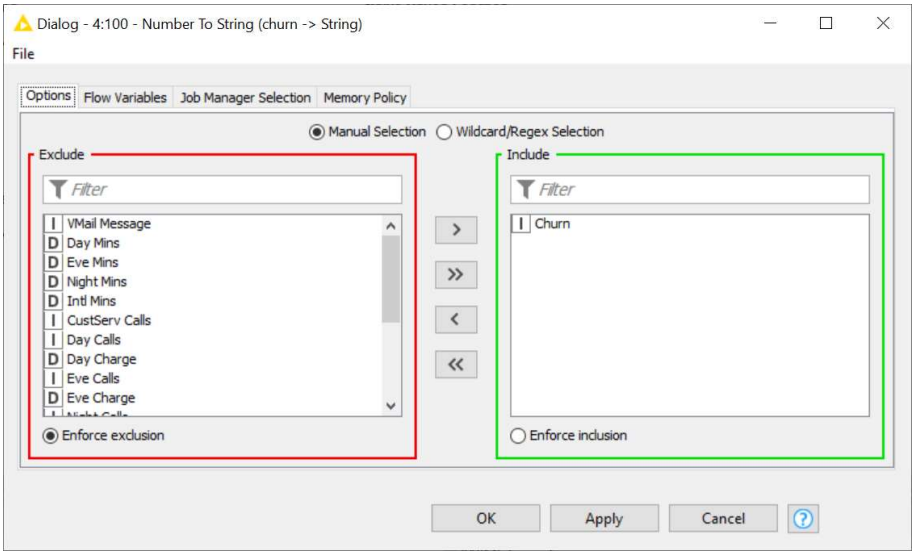


Slika 218. Hodogram naivnog Bayes-ovog klasifikatora

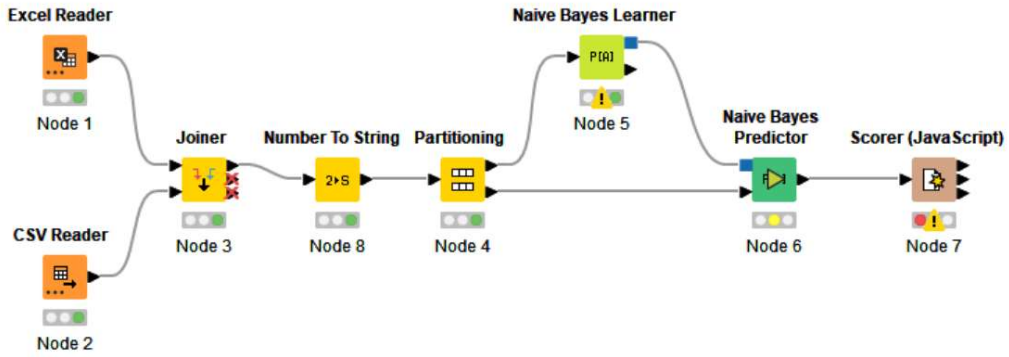
Slika 219 prikazuje čvor **Number To String** ili Broj u tekstualni zapis. Slika 220 prikazuje postavke čvora koje su jednostavne. Potrebno je samo u zelenom okviru ostaviti nazive stupaca koji se žele konvertirati iz brojčanih vrijednosti u tekstualne, odnosno kategorijalne. U ovom primjeru radi se samo o stupcu *Churn* koji je ciljna varijabla.



Slika 219. Čvor Number To String



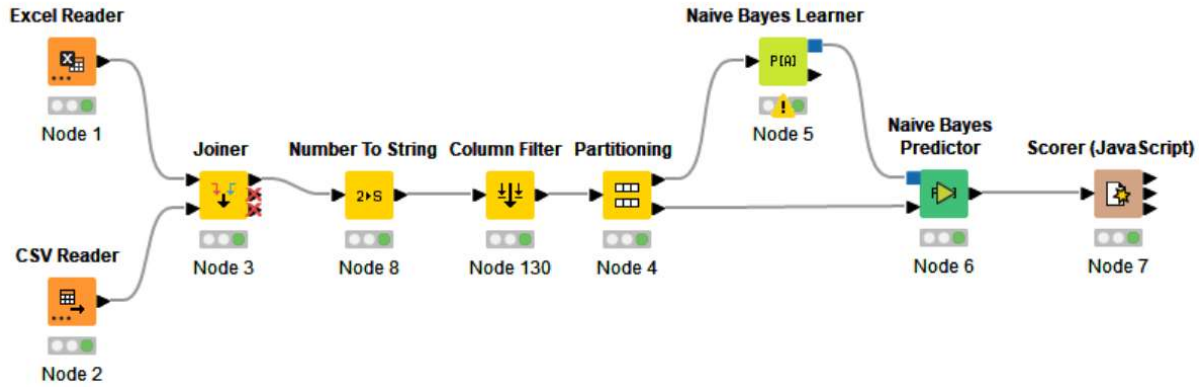
Slika 220. Postavke čvora Number To String



Slika 221. Hodogram naivnog Bayesovog klasifikatora s dodanim čvorom Number To String

Slika 221 prikazuje hodogram s dodanim čvorom **Number To String**.

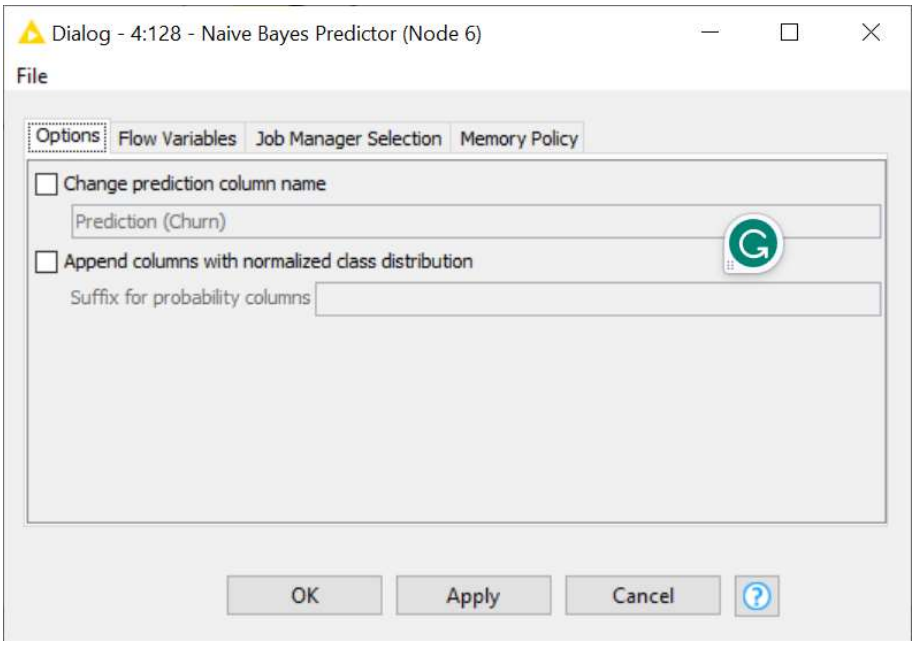
Nakon pokretanja svih čvorova hodograma čvor **Naïve Bayes Learner** i dalje prijavljuje upozorenje vidljivo kao žuti trokut na čvoru, a u konzoli se ispisuje sljedeći tekst: `WARN Naive Bayes Learner 4:12 The following attributes are skipped: Phone/Too many values, Phone (right)/Too many values, State/Too many values.` Analizom teksta upozorenja očito je kako čvor **Naive Bayes Learner** ignorira stupac *Phone* i *Phone(Right)* jer imaju previše jedinstvenih vrijednosti. Navedeno upozorenje nema utjecaj na konačan rezultat i karakteristike modela, ali ako se želi navedeno upozorenje ukloniti, dovoljno je korištenjem čvora **Column Filter** isključiti ta dva navedena stupca u nekom od prethodnih koraka, ali se ostavlja stupac *State*. Čvor **Column Filter** je opisan u prethodnom tekstu.



Slika 222. Hodogram naivnog Bayesovog klasifikatora sa svim potrebnim čvorovima

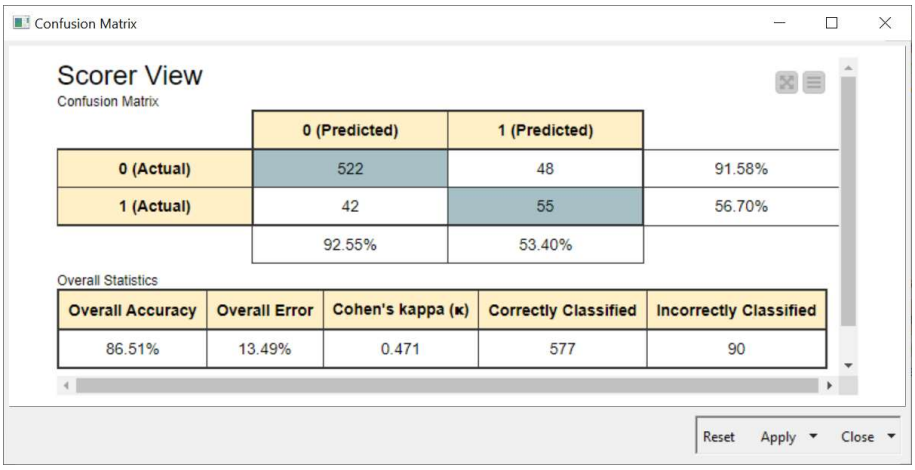
Slika 222 prikazuje konačni hodogram sa svim potrebnim čvorovima. Preostalo je samo još ukratko opisati čvor **Navie Bayes Predictor** te prikazati i komentirati rezultate. Slika 223 prikazuje čvor **Naive Bayes Predictor** ili Naivni Bayes-ov prediktor. Slika 224 prikazuje postavke čvora **Naive Bayes Predictor**. Postavke su praktički iste kao i kod prethodno opisanih prediktorskih čvorova, a u pravilu ih se ne mijenja.

Slika 223. Čvor Naïve Bayes Predictor



Slika 224. Postavke čvora Naïve Bayes Predictor

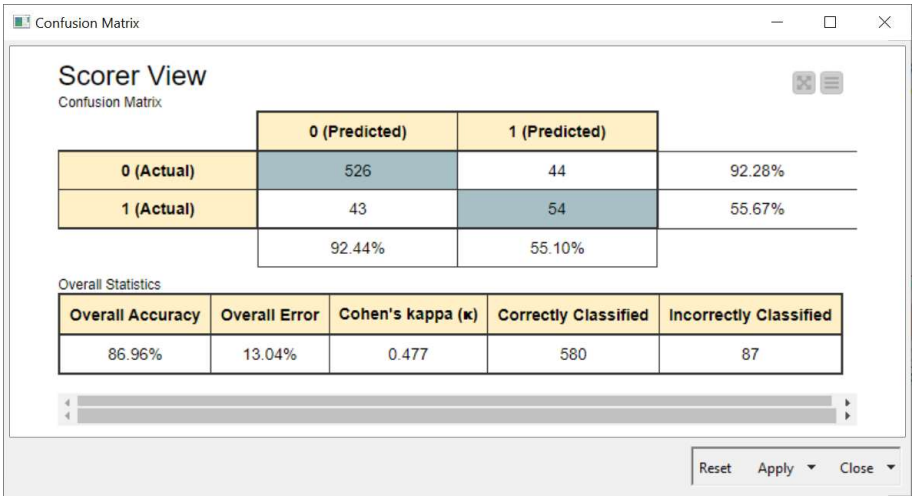
Pokretanjem svih čvorova hodograma iz čvora **Scorer (JavaScript)** dobivaju se podaci o modelu. Slika 225 prikazuje podatke.



Slika 225. Podaci o modelu

Upozorenje je i dalje prisutno na čvoru **Naïve Bayes Learner** pa slijede upute kako se rješava taj problem. U konzoli se pojavljuje upozorenje sa sljedećim tekstom: `WARN  Naive Bayes Learner 4:12 The following attributes are skipped: State/Too many values`. Očigledno je da je broj jedinstvenih vrijednosti u stupcu *State* veći od zadanih 20 u čvoru Naïve Bayes Learner i taj stupac se

ignorira. Rješenje je u izmjeni vrijednosti tog parametra na 100 tako da model ne zanemari podatke iz spomenutog stupca. Nakon toga se mogu pokrenuti svi čvorovi i ponovo pogledati podatke o točnosti modela.

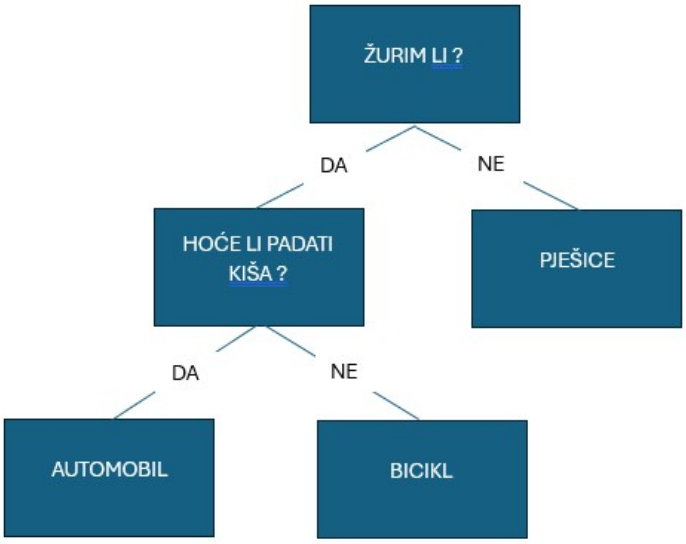


Slika 226. Podaci o modelu

Zanimljivo je kako je model neznatno točniji (*Overall Accuracy*) kad su u model uključene i države iz kojih su korisnici usluga. Navedena značajka ima mali doprinos točnosti modela, ali na neki način ukazuje da se trebaju iskoristiti sve dostupne značajke. Doduše moguće je da bi se to izgubilo drugačijim preslagivanjem skupa podataka za treniranje i testiranje. Za kraj treba prokomentirati i činjenicu kako ovaj skup podataka nije razmjeran kada se pogleda broj slučajeva koji pripadaju svakoj kategoriji pri čemu jedna kategorija ima oko pet puta više slučajeva, odnosno pretplatnika. Iz tog razloga, ako se promatra ukupnu točnost može se reći kako je rezultat zadovoljavajući, ali ako se pažljivije pogleda matrica konfuzije, vidi se da je za kategoriju 1 točnost nešto više od 50 %. Za ovu tehniku je u uvodu naveden jedan nedostatak, a to je preduvjet neovisnosti značajki, tako da je moguće kako je ovako loš rezultat uzrokovan postojanjem povezanosti među varijablama. Za sada će ostati ta vrijednost, ali u sljedećim poglavljima se obrađuju još neke metode optimizacije koje mogu poslužiti i za optimizaciju ovog primjera. Osim toga, mogu se primijeniti i druge tehnike s ovim skupom podataka s ciljem povećanja točnosti.

9. Stablo odlučivanja

Jedna od osnovnih prednosti ove metode je njena jednostavnost i razumljivost. Na primjer, osoba koja vozi automobil i bicikl vjerojatno je ponekad nesvjesno primijenila ovu metodu prije nego što se zaputila na odredište i to na način da se u sebi pitala dva pitanja. Pitanja su: „Žurim li ?“ i „Hoće li danas padati kiša?“. Ako je odgovor na prvo pitanje NE, osoba odlazi pješice. Ako je odgovor na prvo pitanje DA, onda si osoba postavlja drugo pitanje vezano uz kišu. Ovisno o odgovoru koji se može pročitati iz vremenske prognoze, osoba bira automobil ili bicikl kao prijevozno sredstvo. Slika 227 prikazuje grafički stablo odlučivanja iz primjera, a mogućnost grafičkog prikazivanja modela je još jedna prednost ove metode (Géron, 2022).



Slika 227. Primjer stabla odlučivanja

Na osnovu stabla odlučivanja može se pojedini slučaj dodijeliti nekoj kategoriji, a taj proces bi u stvari bila primjena modela koji je prethodno istreniran. Model se kao i u prethodnim tehnikama stvara na osnovu dostupnih poznatih podataka, a kod primjene stabla odlučivanja za klasifikaciju se koriste tablični podaci. Jednostavno je zamisliti kako svaki slučaj (red u tablici) na osnovu značajki (vrijednosti u pojedinom stupcu) pripada u konačnici određenoj kategoriji (ciljna varijabla). Nakon što se na osnovu podataka kreira stablo odlučivanja, ono je primjenjivo na nove slučajeve za klasifikaciju. Dovoljno je ići od vrha stabla i odgovarati na pitanja u svakom čvoru, kako bi se ovisno o značajkama slučaja došlo na kraju do kategorije kojoj pripada. Sam grafički prikaz stabla odlučivanja sastoji se od nekoliko elemenata, a to su (Géron, 2022):

a) Korijenski čvor – čvor na vrhu hijerarhijske strukture koji nema nadređenog čvora te predstavlja jednu ulaznu varijablu, a na granama čvora su moguće vrijednosti (tamno plavi).
b) Čvor – predstavlja jednu ulaznu varijablu, a na granama čvora su moguće vrijednosti (plavi)
c) List – vrijednost ciljne varijable (svjetlo plavi).

Nakon što je pojašnjen način primjene stabla odlučivanja, preostaje objašnjenje na koji način ga se može kreirati. Postoji više metoda, ali najjednostavnija verzija tehnike za izradu stabla odlučivanja počet će od korijena stabla i kroz više iteracija će provjeravati koje je „najbolje moguće" sljedeće dijeljenje kako bi se kategorije razlikovale na najmanje dvosmislen način. Tablica 11 sadrži podatke za deset osoba i to dvije kategorijalne značajke (sportaš i uporan) i jednu kategorijalnu ciljnu varijablu

(završio ekonomski fakultet). Iz ovih podataka kreirat će se stablo odlučivanja. S obzirom da postoje samo dva stupca sa značajkama, stablo odlučivanja imat će dvije razine. Ključno pitanje je koju značajku staviti u korijenski čvor i zašto.

Tablica 11. Podaci za model stabla odlučivanja

SPORTAŠ	UPORAN	ZAVRŠIO EKONOMSKI FAKULTET
NE	NE	NE
DA	DA	DA
NE	DA	DA
NE	NE	NE
NE	NE	NE
DA	DA	DA
DA	NE	NE
DA	DA	DA
NE	DA	NE
DA	DA	DA

Iz ovih tabličnih podataka moguće je napraviti dvije kombinirane statističke tablice koje mogu pomoći kod odlučivanja koja značajka je bitnija. Radi se o tablicama koje služe za prikaz podataka promatranih prema dvije ili više kategorijalnih varijabli (Horvat & Mijoč, 2019). U ovom slučaju jedna tablica (Tablica 12) prikazuje povezanost ciljne varijable *završio ekonomski fakultet* i značajke *sportaš*, dok druga (Tablica 13) prikazuje povezanost ciljne varijable *završio ekonomski fakultet* i značajke *uporan.*

Tablica 12. Tablica povezanosti značajke "sportaš" i ciljne varijable

	SPORTAŠ			
ZAVRŠIO EKONOMSKI FAKULTET	DA	DA (%)	NE	NE (%)
DA	4	80%	1	20%
NE	1	20%	4	80%
UKUPNO	5	100%	5	100%

Tablica 13. Tablica povezanosti značajke "uporan" i ciljne varijable

	UPORAN			
ZAVRŠIO EKONOMSKI FAKULTET	DA	DA (%)	NE	NE (%)
DA	5	83%	0	0%
NE	1	17%	4	100%
UKUPNO	6	100%	4	100%

Koja od ove dvije značajke više pomaže u procjeni je li osoba završila fakultet? Intuitivno bi to bila ona koja je više deterministička ili određena, a to bi bila značajka „upornost". Jedan od kvantitativnih načina da se to ispita je korištenjem tzv. Gini indeksa. Gini indeks se računa za obje značajke u ovom primjeru,

odnosno za oba načina dijeljenja na razini čvora. Ako pojedina značajka ima manju vrijednost Gini indeksa, ona se izabire za čvor. Skupni Gini indeks se računa po formuli (Géron, 2022):

$$Gini\ (D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2)$$

pri čemu se za svaki dio podataka Gini računa kao (Géron, 2022):

$$Gini(D_j) = 1 - \sum_{J=1}^{c} p_j^2$$

Slovo D označava kompletan skup podataka, koji se dijeli na osnovu neke značajke na D1 i D2. Broj elemenata skupa podataka D je označen slovom n sa brojčanim indeksom, a n_1 i n_2 zbrojeni daju n. Broj elemenata skupa D1 je n_1, a skupa D2 je n_2. Slovo c označava broj kategorija (u ovom slučaju c=2), a slovo p označava vjerojatnost da uzorci pripadaju kategoriji j u danom čvoru.

Izračun Gini indeksa za značajku „sportaš“ i to za oba skupa podataka te na kraju skupno, slijedi:

$$Gini(D_1) = 1 - \left(\frac{4}{1+4}\right)^2 - \left(\frac{1}{1+4}\right)^2 = 1 - 0{,}64 - 0{,}04 = 0{,}32$$

$$Gini(D_2) = 1 - \left(\frac{1}{1+4}\right)^2 - \left(\frac{4}{1+4}\right)^2 = 1 - 0{,}04 - 0{,}64 = 0{,}32$$

$$Gini\ (D) = \frac{5}{10} 0{,}32 + \frac{5}{10} 0{,}32 = 0{,}32$$

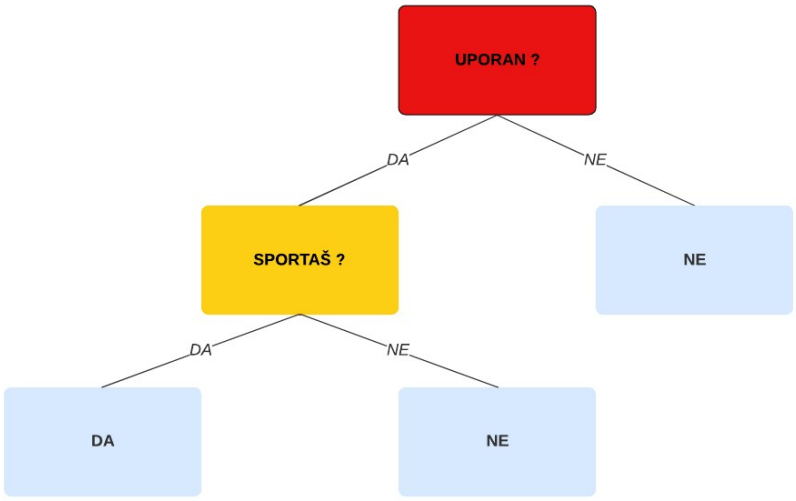
Izračun Gini indeksa za značajku „uporan“ i to za oba skupa podataka te na kraju skupno, slijedi:

$$Gini(D_1) = 1 - \left(\frac{5}{1+5}\right)^2 - \left(\frac{1}{1+5}\right)^2 = 1 - 0.69 - 0{,}03 = 0{,}28$$

$$Gini(D_2) = 1 - \left(\frac{0}{0+4}\right)^2 - \left(\frac{4}{0+4}\right)^2 = 1 - 0 - 1 = 0$$

$$Gini\ (D) = \frac{6}{10} 0{,}28 + \frac{4}{10} 0 = 0{,}17$$

Na osnovu manjeg Gini indeksa za značajku „uporan“ odlučuje se postaviti podjelu na osnovu značajke „uporan“ u korijenski čvor koji je na vrhu. S obzirom kako postoji samo još jedna značajka, ona preostaje kao jedini izbor za drugi podređeni čvor. Na ovom jednostavnom primjeru kategorizacije opisan je način rada, odnosno treniranja modela baziranog na stablu odlučivanja. Prije nego što se prijeđe na kompleksniji primjer u programu KNIME, treba naglasiti da tehnika stabla odlučivanja može koristiti i s kontinuiranim značajkama.



Slika 228. Stablo odlučivanja generirano iz primjera

Na kraju treba spomenuti prednosti ove tehnike, a to su niska zahtjevnost što se tiče računalnih resursa za treniranje, laka interpretacija, nema potrebe za normalizacijom i mogućnost rada s kategorijalnim i numeričkim značajkama. Nedostaci su ograničena točnost i nedostatak efikasnijih mogućnosti upravljanja problemima prenaučenosti i podnaučenosti (Egger, 2022).

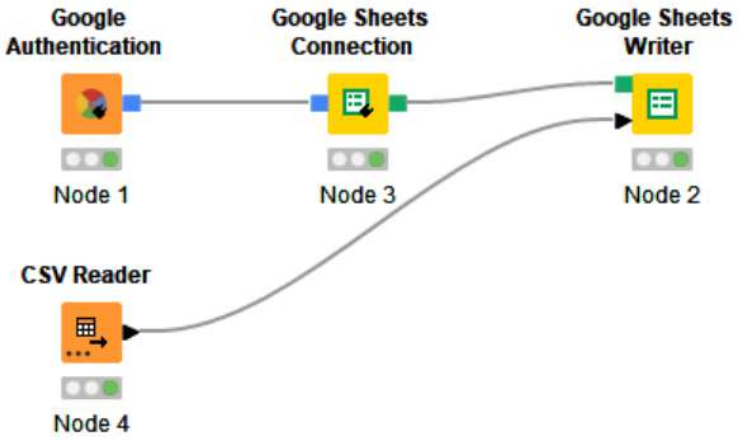
9.1. Priprema podataka

Jedna od čestih metoda prikupljanja podataka na području društvenih znanosti je anketiranje. U novije vrijeme često se provode online ankete i pri tom je dosta popularan alat Google Obrasci (*Google Forms*). Radi se o jednostavnom i besplatnom online alatu s kojim se vrlo brzo mogu kreirati ankete, a podaci su trenutno dostupni online u tabličnom obliku u formatu Google Tablice (*Google Sheets*). S obzirom da je alat Google Obrasci prilično jednostavan za korištenje i nije tema ovog priručnika neće se posebno opisivati, ali će se učitati podaci iz datoteke u formatu Google Tablica i na osnovu njih kreirati model stabla odlučivanja. Kratki uvod u izradu Google Obrazaca dostupan je na poslužitelju www.youtube.com pri čemu treba tražiti video pod naslovom „Kako koristiti Google obrasce?“ čiji autori su djelatnici Googlea. U primjeru će se pretpostaviti da se radi o podacima prikupljenim anketiranjem, a bit će predstavljen način dohvaćanja podataka pohranjenih na serverima Googlea u formatu Google Tablica.

Problem koji će se rješavati korištenjem tehnike stabla odlučivanja vezan je uz izostajanje s posla. Radi se o problemu s kojim se suočavaju tvrtke i odjeli, a strojno učenje omogućuje da se s tim problemom može bolje upoznati. Podaci koji će se pri tom koristiti su dostupni na adresi https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset. Treba naglasiti kako su podaci izmišljeni i to je autor naglasio na stranici. Podaci su u CSV formatu i potrebno je preuzeti jednu datoteku te ju spremiti tako da joj se može pristupiti iz programa KNIME. Na stranici KAGGLE gdje se nalaze izvorni podaci, navedeni su nazivi stupaca u datoteci, a to su: identifikator zaposlenika, prezime, ime, spol, grad, naziv radnog mjesta, naziv odjela, lokacija trgovine, podjela, dob, radni staž, odsutni sati i poslovna jedinica.

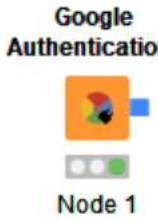
Nakon preuzimanja komprimirane datoteke istu je potrebno dekomprimirati da bi se dobila jedna CSV datoteka. Da bi se kroz ove primjere steklo što više znanja i vještina, za rješavanje ovog problema kreirat će se dodatni hodogram koji će služiti za konverziju podataka iz CSV datoteke u format Google

Tablica i spremanje na prostor koji tvrtka Google besplatno ustupa uz kreiran Google/Gmail račun. U glavnom hodogramu u kojem se trenira model na osnovu podataka te iste podatke preuzet će iz datoteke formata Google Tablice sa servera tvrtke Google na internetu. Slika 229 prikazuje taj dodatni hodogram, a prije opisa novih čvorova dan je pregled osnovne funkcionalnosti hodograma. Čvor **CSV Reader** služi za učitavanje podataka iz CSV datoteke koja se može preuzeti s poslužitelja www.kaggle.com, da bi te podatke dostavio čvoru **Google Sheets Writer**. Kako mu samo ime kaže, taj čvor zapisuje podatke u formatu Google Tablica. Osim toga na hodogramu su vidljiva još dva čvora koji služe za prijavu na Google i kreiranje veze prema Google Tablicama. U nastavku će biti opisani novi čvorovi i njihove postavke.



Slika 229. Dodatni hodogram za konverziju iz CSV formata u format Google Tablice

Slika 230 prikazuje čvor **Google Authentication** ili Google autentikacija koji služi za prijavu kako bi se mogli koristiti servisi tvrtke Google. Poznato je da se za korištenje različitih usluga koje ta tvrtka nudi potrebno prijaviti putem web preglednika i to samo jednom. U dodatnom hodogramu je potrebna prijava i ona se ostvaruje korištenjem ovog čvora.

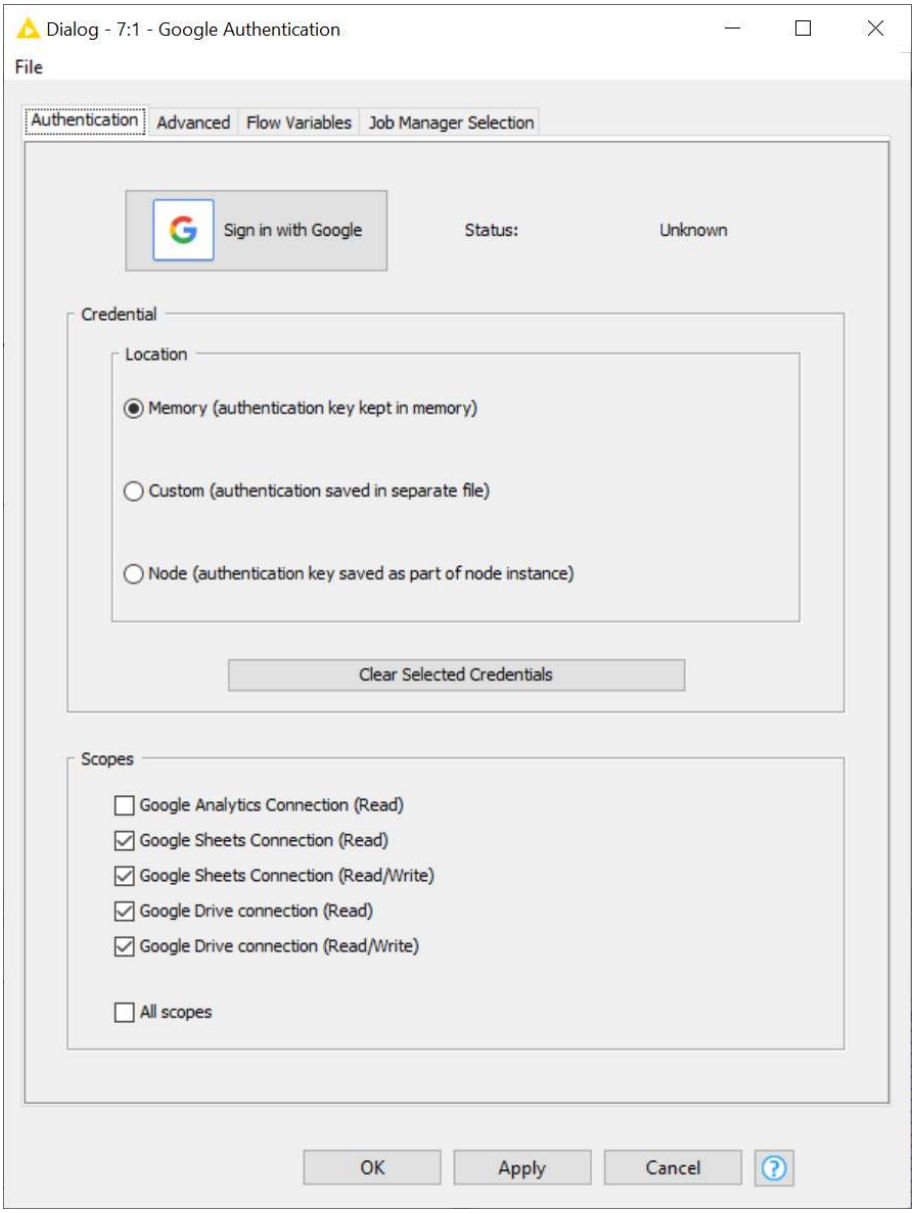


Slika 230. Čvor Google Authentication

Slika 231 prikazuje postavke čvora **Google Authentication**. U gornjem dijelu dijaloškog okvira nalazi se dugme *Sign in with Google* koje pokreće proces prijave na način da se automatski otvara web preglednik u kojem se prijavljuje. Ako je u web pregledniku osoba već prijavljena s Google računom dovoljno je izabrati račun s kojim je prijavljena i dati određene ovlasti KNIME aplikaciji nad podacima osobe na Google poslužitelju. U slučaju da osoba nije prijavljena, potrebno je odraditi prijavu unošenjem korisničkog imena i lozinke. Nakon tog postupka web preglednik ispisuje sljedeću rečenicu: *Received verification code. You may now close this window.* Tada se možete vratiti u KNIME aplikaciju jer je postupak prijave završen.

U postavkama čvora **Google Authentication** osim prijave može se izabrati i gdje će biti pohranjene vjerodajnice. Moguće ih je pohraniti u memoriji što je zadana mogućnost pri čemu se vjerodajnice brišu

izlaskom iz aplikacije KNIME. U slučaju da vjerodajnice želite pohraniti u obliku datoteke, možete izabrati jednu od druge dvije mogućnosti (*Custom* i *Node*). U donjem dijelu postavki definiraju se dozvole s obzirom na aplikacije (*Analytics, Sheets i Drive*) i ovlasti čitanja/pisanja (*Read* ili *Read/Write*).

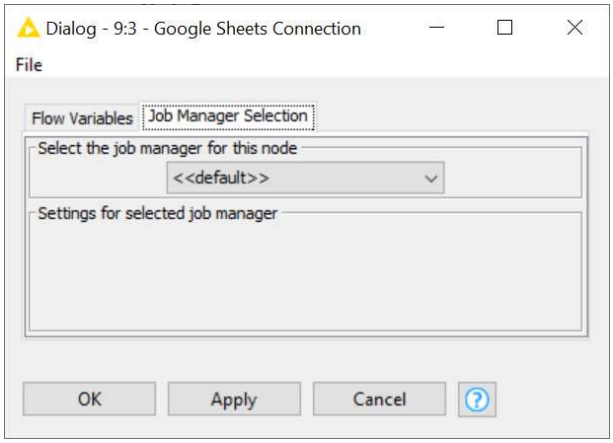


Slika 231. Postavke čvora Google Authentication

Slika 232 prikazuje čvor **Google Sheets Connection** ili Spojnica Google tablice. Njegova funkcija je povezivanje s Google Tablice uslugom tvrtke Google nakon uspješne prijave.

Slika 232. Čvor Google Sheets Connection

Slika 233 prikazuje dijaloški okvir postavki čvora **Google Sheets Connection**. U pravilu postavke ovog čvora se ne mijenjaju.

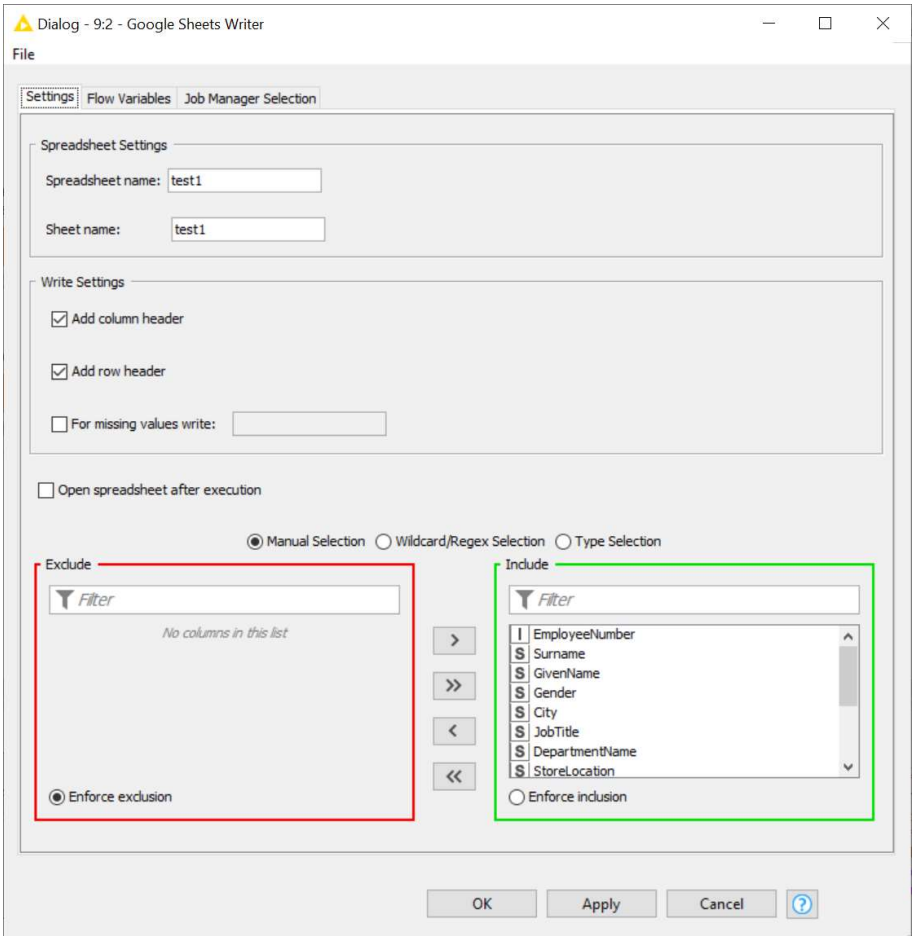


Slika 233. Postavke čvora Google Sheets Connection

Slika 234 prikazuje čvor **Google Sheets Writer** ili Pisač Google tablice. Njegova funkcija je zapisivanje podataka u izabran radni list izabrane radne knjige, odnosno datoteke u formatu Google Tablice.

Slika 234. Čvor Google Sheets Writer

Slika 235 prikazuje postavke čvora **Google Sheets Writer**. U gornjem dijelu dijaloškog okvira bira se naziv datoteke te naziv radnog lista u koji se želi upisivati podatke. U srednjem dijelu dijaloškog okvira bira se željeno zaglavlja na stupce i redove, kao i vrijednost koja će biti unesena u ćeliju ako je u izvornoj tablici prazna. U donjem dijelu dijaloškog okvira biraju se stupci iz izvorne tablice koji će biti zapisani u novu tablicu u formatu Google Tablice.

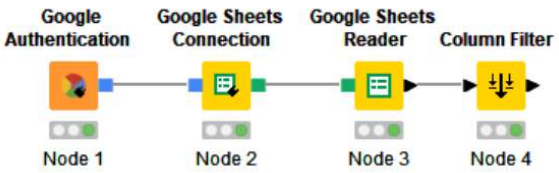


Slika 235. Postavke čvora Google Sheets Writer

Nakon što se uspješno izvrše svi čvorovi na dodatnom hodogramu, kao rezultat trebali bi se dobiti podaci iz CSV datoteke u formatu Google Tablice na poslužitelju tvrtke Google. Treba još jednom naglasiti kako novonastala datoteka neće biti pohranjena lokalno na računalu, nego na poslužitelju tvrtke Google. Ključna prednost je da joj se može pristupiti s bilo kojeg računala spojenog na internet.

9.2. Izrada modela temeljenog na klasifikatoru stabla odlučivanja

Hodogram za model baziran na klasifikatoru stabla odlučivanja ne razlikuje se bitno od prethodno kreiranih hodograma. Na početku se učitavaju podaci, a s obzirom da se za ovaj primjer podaci nalaze u datoteci na poslužiteljima tvrtke Google, pristup će biti drugačiji nego u prethodnim primjerima.

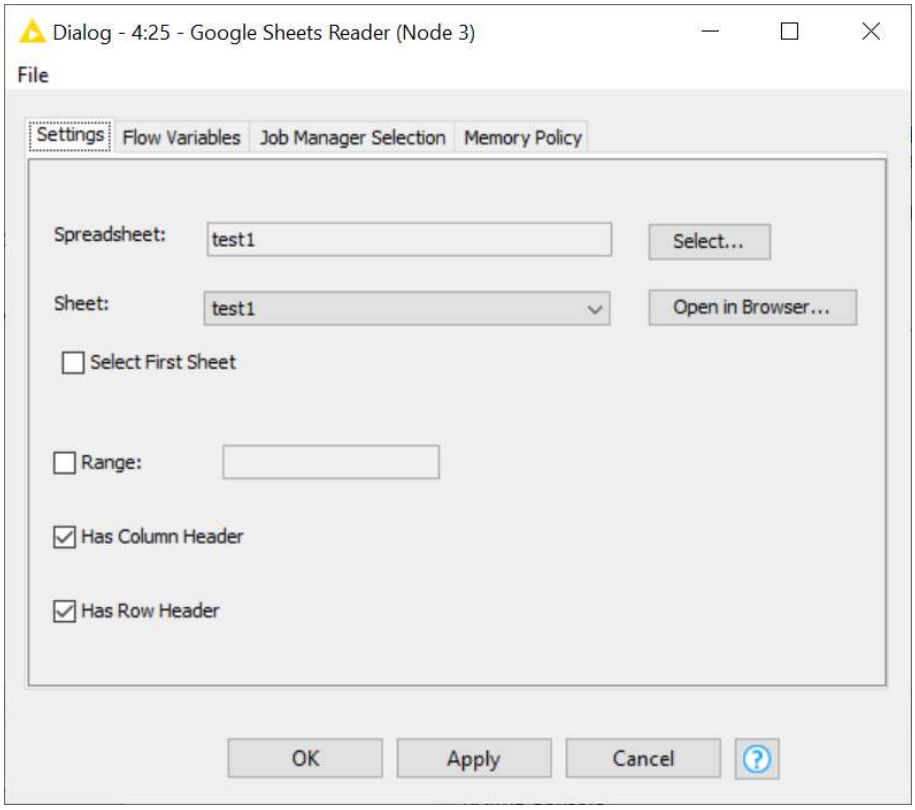


Slika 236. Prvi dio hodograma za učitavanje i filtraciju podataka iz Google Tablice

Slika 236 prikazuje prvi dio hodograma u koji se prijavljuje s Google računom te se učitavaju podaci koristeći čvor **Google Sheets Reader** ili Čitač Google tablice. Radi se o novom čvoru čija funkcionalnost je jasna u samom nazivu. Slika 237 prikazuje čvor.

Slika 237. Čvor Google Sheets Reader

Slika 238 prikazuje postavke čvora **Google Sheets Reader**. Postavke su slične čvoru **Google Sheets Writer**, pri čemu je omogućen izbor radne knjige i radnog lista. Osim toga može se definirati područje u radnom listu s kojeg se preuzimaju podaci. Ako radni list ima zaglavlja redova i stupaca, to se također može izabrati u postavkama.



Slika 238. Postavke čvora Google Sheets Reader

Zadnji čvor u prvom dijelu prikazanog hodograma služi za filtraciju stupaca, odnosno značajki koje su nepotrebne u modelu. Radi se o imenima i prezimenima djelatnika, kao i o jedinstvenim brojevima svakog djelatnika. S obzirom da te značajke nemaju nikakve veze s izostajanjem s posla, ovim čvorom se uklanjaju. Nakon uklanjanja viška stupaca, može se pogledati sadržaj iz kontekstnog izbornika klikom na izlazni priključak čvora **Column Filter** i izborom *Filtered Table*. Slika 239 prikazuje filtrirane podatke.

Filtered table - 4:2 - Column Filter (Node 4)

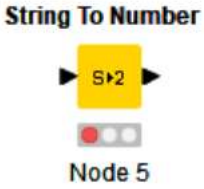
File Edit Hilite Navigation View

Table "default" - Rows: 8149 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	Gender	City	JobTitle	Depart...	StoreLo...	Division	Age	Length...	Absent...	Busines...
Row0	F	Burnaby	Baker	Bakery	Burnaby	Stores	33.02881569	7.018478474	36.57730606	Stores
Row1	M	Courtenay	Baker	Bakery	Nanaimo	Stores	41.32090167	6.532444578	30.16507231	Stores
Row2	M	Richmond	Baker	Bakery	Richmond	Stores	49.82204661	5.389973118	83.80779766	Stores
Row3	F	Victoria	Baker	Bakery	Victoria	Stores	45.59935722	4.081735738	70.02016505	Stores
Row4	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	36.69787561	4.619091448	0.0	Stores
Row5	M	Richmond	Baker	Bakery	Richmond	Stores	49.44031059	3.717692452	81.83007916	Stores
Row6	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	51.75273	11.157918	60.49507152	HeadOffice
Row7	M	Sechelt	Baker	Bakery	West Vanco...	Stores	37.2160312	5.432122862	30.07290192	Stores
Row8	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	59.42738025	7.940120524	181.630819	Stores
Row9	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	40.85398	14.848321	30.66440832	HeadOffice
Row10	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	47.54758052	5.872038321	28.01835332	Stores
Row11	M	North Vanco...	Baker	Bakery	North Vanco...	Stores	38.72801116	4.621141838	0.0	Stores
Row12	F	Vananda	Baker	Bakery	Nanaimo	Stores	31.78519091	5.583328091	34.33444296	Stores
Row13	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	50.92338	5.883225	0.0	HeadOffice
Row14	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	43.79789	20.107198	21.65982312	HeadOffice

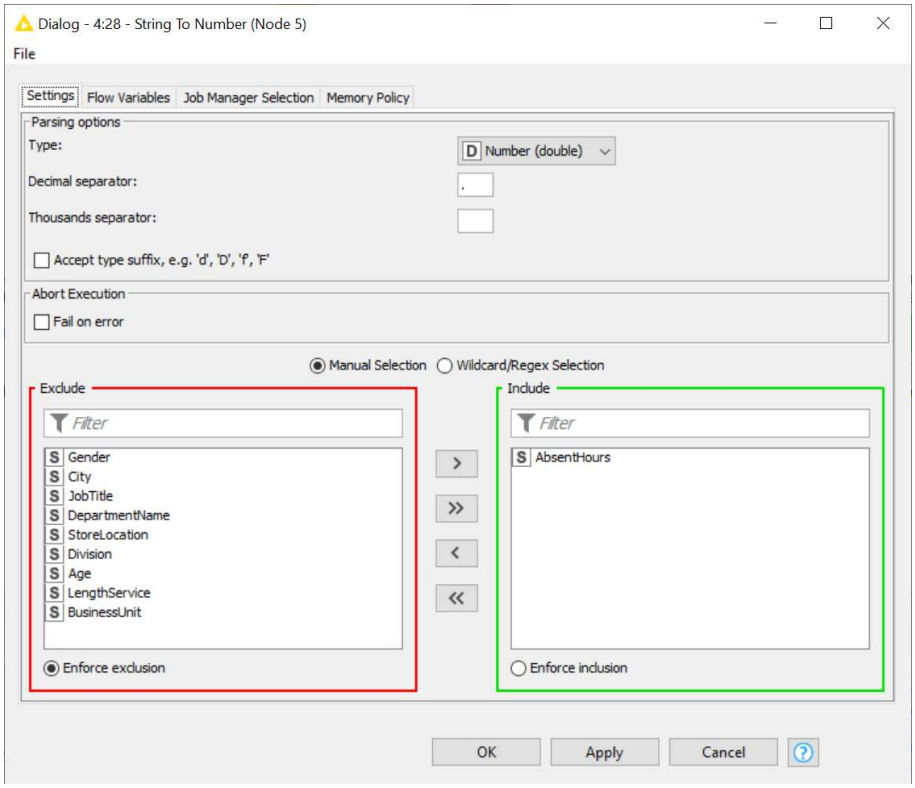
Slika 239. Podaci nakon filtriranja

U tablici s podacima uočljivo je kako se radi isključivo o tekstualnim podacima, jer je uz naziv svakog stupca prisutno slovo „S" što označava *String*. S obzirom da se iz ovih podataka koristeći klasifikator stabla odlučivanja želi predvidjeti tko će od djelatnika izostajati s posla, ciljna varijabla je kategorija koja označava beznačajno ili značajno izostajanje s posla. U podacima postoji samo stupac *AbsentHours* u kojem su navedeni sati i to u tekstualnom formatu. Prvi korak u pripremi podataka je konverzija tekstualnih podataka u brojčane. Za to postoji čvor i naziva se **String To Number**. Do sada je obrađen čvor **Number To String**, a ovaj čvor ima suprotnu funkciju. Slika 240 prikazuje čvor.



Slika 240 . Čvor String To Number

Slika 241 prikazuje postavke čvora **String To Number** ili Tekstualni zapis u broj. U padajućem izborniku na vrhu dijaloškog okvira može se izabrati jedan od tri vrste brojčanih zapisa: *double*, *integer* i *long*. Opcija *integer* koristi se kada su podaci cjelobrojne vrijednosti bez decimalnih brojeva, dok su preostale dvije opcije namijenjene za decimalne brojeve. *Decimal separator* služi za izbor točke ili zareza između cijelih i decimalnih brojeva u zapisu. U donjoj polovini dijaloškog okvira biraju se stupci čije vrijednosti se želi konvertirati iz tekstualnih u brojčane. U ovom primjeru za sada se konvertira samo stupac *AbsentHours*, za koji je navedeno da sadrži broj sati izostanka s posla za svakog djelatnika.



Slika 241. Postavke čvora String To Number

Slika 242 prikazuje tablicu s podacima nakon konverzije pri čemu se vidi kako je u predzadnjem stupcu pri konverziji dio decimalnih brojeva odbačen. S obzirom da se radi o četvrtoj decimali, za ovaj primjer to nema nikakvog utjecaja.

Transformed input - 4:28 - String To Number (Node 5)

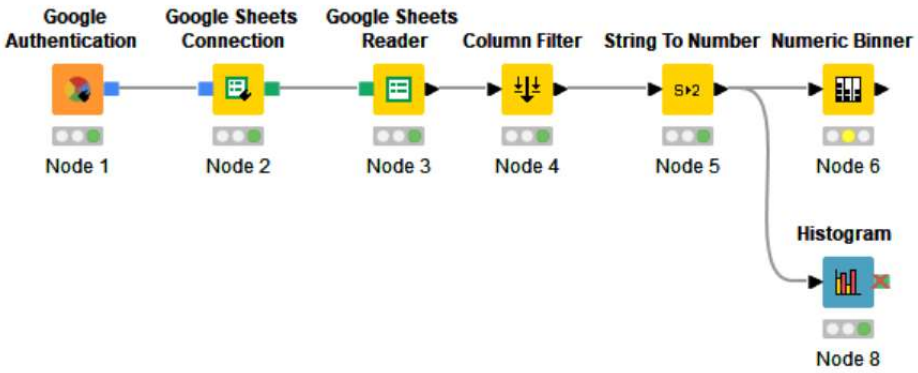
File Edit Hilite Navigation View

Table "default" - Rows: 8149 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	Gender	City	JobTitle	Depart...	StoreLo...	Division	Age	Length...	Absent...	Busines...
Row0	F	Burnaby	Baker	Bakery	Burnaby	Stores	33.02881569	7.018478474	36.577	Stores
Row1	M	Courtenay	Baker	Bakery	Nanaimo	Stores	41.32090167	6.532444578	30.165	Stores
Row2	M	Richmond	Baker	Bakery	Richmond	Stores	49.82204661	5.389973118	83.808	Stores
Row3	F	Victoria	Baker	Bakery	Victoria	Stores	45.59935722	4.081735738	70.02	Stores
Row4	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	36.69787561	4.619091448	0	Stores
Row5	M	Richmond	Baker	Bakery	Richmond	Stores	49.44031059	3.717692452	81.83	Stores
Row6	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	51.75273	11.157918	60.495	HeadOffice
Row7	M	Sechelt	Baker	Bakery	West Vanco...	Stores	37.2160312	5.432122862	30.073	Stores
Row8	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	59.42738025	7.940120524	181.631	Stores
Row9	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	40.85398	14.848321	30.664	HeadOffice
Row10	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	47.54758052	5.872038321	28.018	Stores
Row11	M	North Vanco...	Baker	Bakery	North Vanco...	Stores	38.72801116	4.621141838	0	Stores
Row12	F	Vananda	Baker	Bakery	Nanaimo	Stores	31.78519091	5.583328091	34.334	Stores
Row13	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	50.92338	5.883225	0	HeadOffice
Row14	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	43.79789	20.107198	21.66	HeadOffice

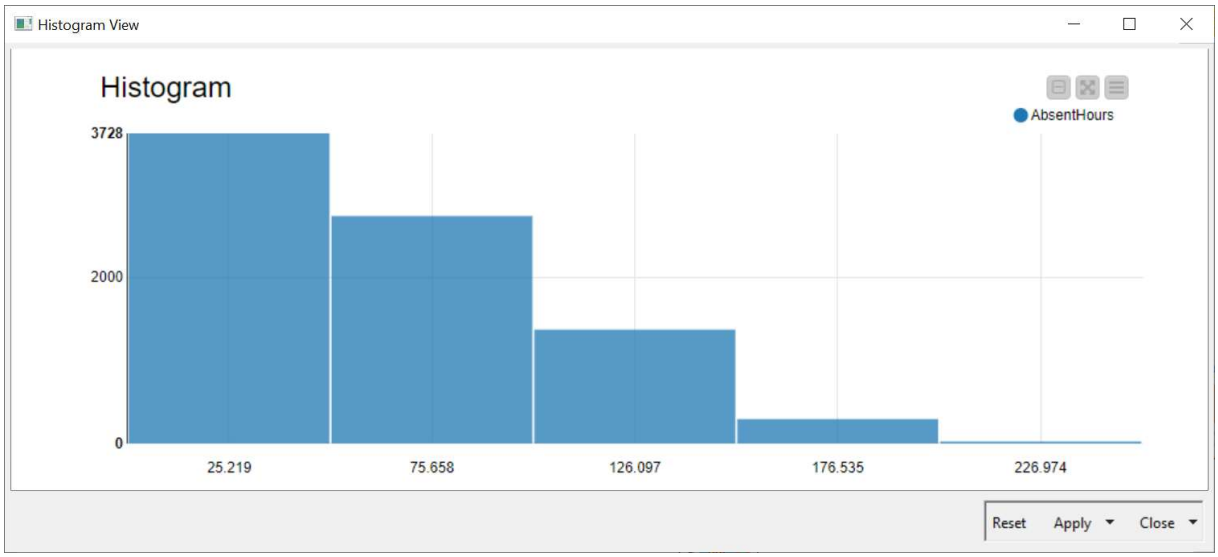
Slika 242. Podaci nakon konverzije iz tekstualnih u brojčane

S obzirom da je cilj kategorizirati djelatnike na osnovu izostanaka u dvije kategorije, beznačajno i značajno izostajanje s posla, potrebno je definirati što znači beznačajno izostajanje, a što značajno. Drugim riječima, potrebno je značajku *AbsentHours* koja je trenutno kontinuirana i brojčana pretvoriti u diskretnu koja uključuje samo dvije kategorije: beznačajno i značajno. Da bi se napravila konverzija koristi se čvor **Numeric Binner**, ali prije upotrebe čvora preporučljivo je pogledati histogram izostanaka s posla da bi lakše bilo donijeti odluka gdje postaviti granicu dvaju navedenih kategorija. Slika 243 prikazuje hodogram nadograđen na opisan način.



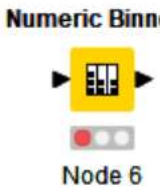
Slika 243. Hodogram nadograđen čvorovima String To Number, Numeric Binner i Histogram

Kao što je navedeno, čvor **Histogram** služi za analizu izostanaka s posla od strane djelatnika. Slika 244 prikazuje histogram. Broj sati koji će podijeliti djelatnike u dvije kategorije može se u ovom primjeru postaviti na 100. To bi u stvarnoj situaciji trebala biti odluka odjela upravljanja ljudskim resursima koji bi tu vrijednost trebali dogovoriti s upravom, ali svakako bi trebali sagledati i dosadašnje vrijednosti, koristeći grafičke prikaze kao što je ovaj histogram.



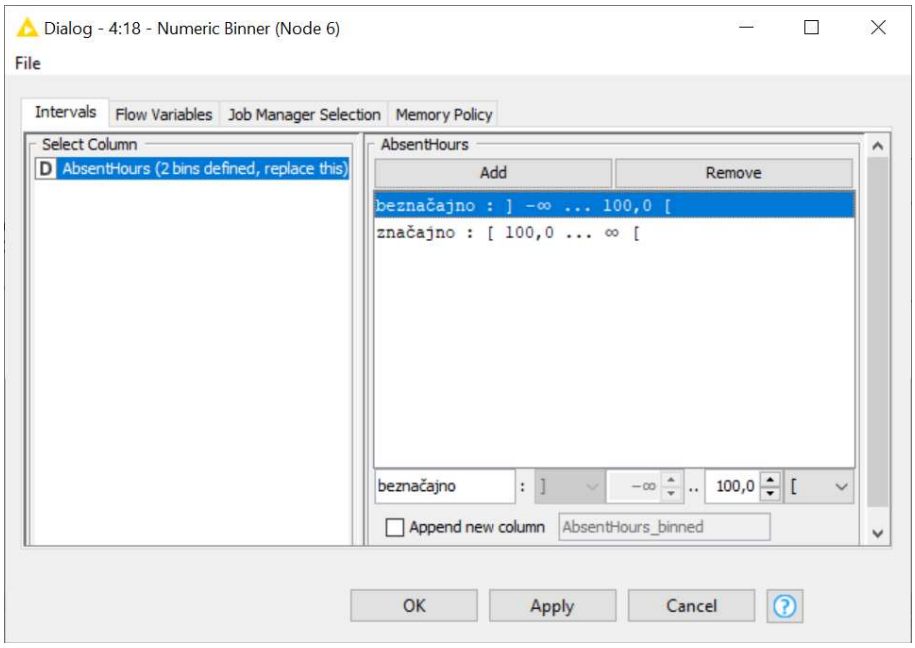
Slika 244. Prikaz podataka o izostancima djelatnika histogramom

U nadograđenom hodogramu pojavljuje se još jedan novi čvor, a to je **Numeric Binner** ili Numerički spremnik. Kao što je prethodno spomenuto, taj čvor će podijeliti djelatnike u dvije skupine na osnovu broja sati izostanaka s posla. Funkcija čvora je dijeljenje slučajeva u kategorije na osnovu kontinuirane varijable, odnosno značajke. Slika 245 prikazuje taj čvor.



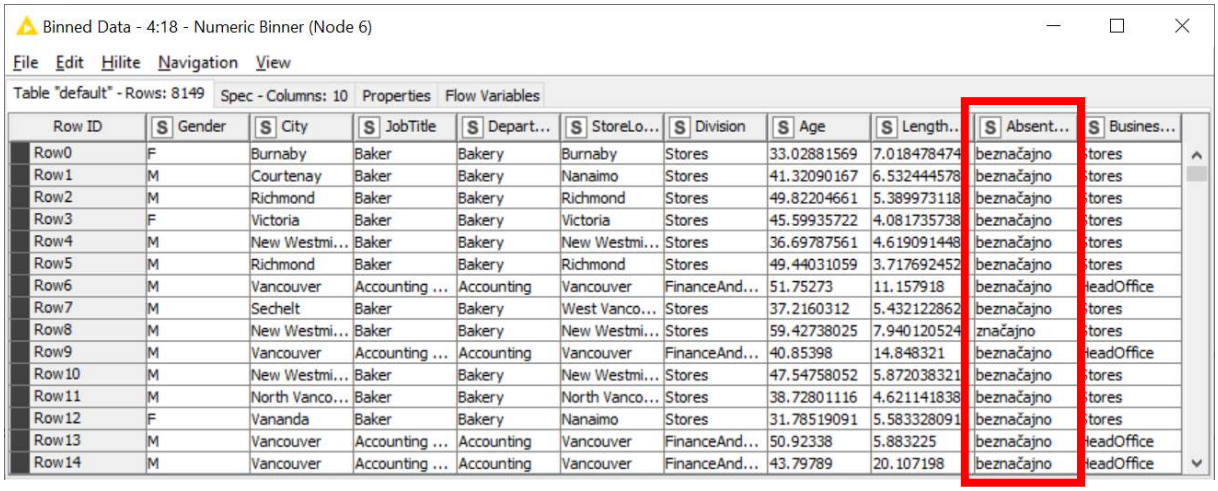
Slika 245. Čvor Numeric Binner

Slika 246 prikazuje postavke čvora **Numeric Binner**. S lijeve strane dijaloškog okvira može se izabrati stupac na osnovu čijih podataka se želi podijeliti slučajeve u kategorije. U ovom slučaju dostupan je samo jedan stupac i to *AbsentHours*. Čvor će u postavkama prikazati samo kontinuirane varijable, dok će kategorijalne izostaviti. Iz tog razloga je ovdje vidljiv samo jedan stupac i to onaj koji je prethodno konvertiran iz tekstualnog u brojčani. S desne strane se definiraju granice vrijednosti pojedinih kategorija i to na način da se klikne na dugme Add ovisno o tome koliko kategorija se želi. S obzirom da se u ovom primjeru želi slučajeve podijeliti u samo dvije kategorije, dovoljno je kliknuti dva puta. Time se dobivaju dva reda u desnom dijelu postavki. Klikom na svaki red definiraju se granice kategorija, a u ovom slučaju dovoljno je definirati granice za prvu kategoriju. To se radi na način da se za označenu kategoriju mijenjaju brojevi u dnu dijaloškog okvira. U ovom primjeru ostavljena je granica „minus beskonačno", dok je definirana gornja granica na 100. U donjem dijelu dijaloškog okvira mogu se mijenjati i nazivi kategorija. Zadani nazivi su Bin1, Bin2 itd, a izmijenjeni su u „beznačajno" i „značajno".



Slika 246. Postavke čvora Numeric Binner

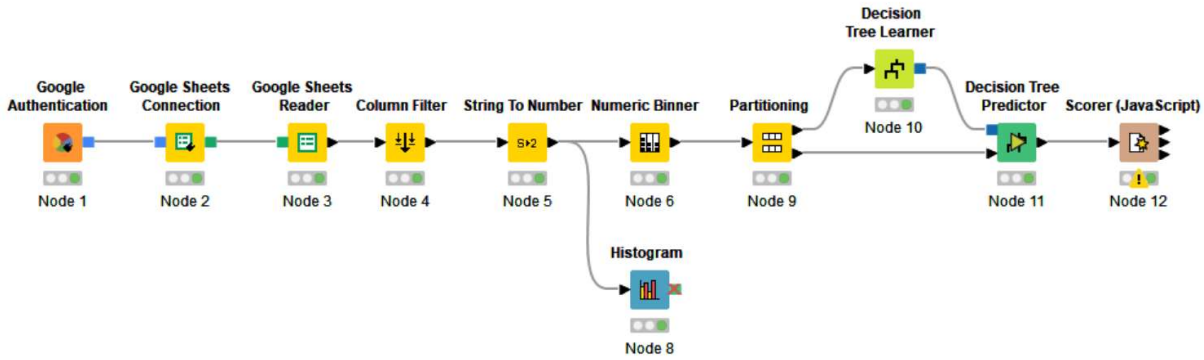
Nakon izvršavanja čvora **Numeric Binner**, koristeći kontekstualni izbornik i izborom *Binned Data* može se dobiti ispis podataka nakon konverzije u kategorijalne podatke za stupac *AbsentHours*. Slika 247 prikazuje podatke nakon konverzije, a uokviren je stupac *AbsentHours*.



Row ID	Gender	City	JobTitle	Depart...	StoreLo...	Division	Age	Length..	Absent...	Busines...
Row0	F	Burnaby	Baker	Bakery	Burnaby	Stores	33.02881569	7.018478474	beznačajno	Stores
Row1	M	Courtenay	Baker	Bakery	Nanaimo	Stores	41.32090167	6.532444578	beznačajno	Stores
Row2	M	Richmond	Baker	Bakery	Richmond	Stores	49.82204661	5.389973118	beznačajno	Stores
Row3	F	Victoria	Baker	Bakery	Victoria	Stores	45.59935722	4.081735738	beznačajno	Stores
Row4	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	36.69787561	4.619091448	beznačajno	Stores
Row5	M	Richmond	Baker	Bakery	Richmond	Stores	49.44031059	3.717692452	beznačajno	Stores
Row6	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	51.75273	11.157918	beznačajno	HeadOffice
Row7	M	Sechelt	Baker	Bakery	West Vanco...	Stores	37.2160312	5.432122862	beznačajno	Stores
Row8	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	59.42738025	7.940120524	značajno	Stores
Row9	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	40.85398	14.848321	beznačajno	HeadOffice
Row10	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	47.54758052	5.872038321	beznačajno	Stores
Row11	M	North Vanco...	Baker	Bakery	North Vanco...	Stores	38.72801116	4.621141838	beznačajno	Stores
Row12	F	Vananda	Baker	Bakery	Nanaimo	Stores	31.78519091	5.583328091	beznačajno	Stores
Row13	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	50.92338	5.883225	beznačajno	HeadOffice
Row14	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	43.79789	20.107198	beznačajno	HeadOffice

Slika 247. Podaci nakon konverzije stupca AbsentHours

Nakon konverzije preostaju čvorovi **Partitioning**, **Decision Tree Learner, Decision Tree Predictor** i **Scorer (JavaScript).** Očigledno je kako se model izgrađuje od praktički istih čvorova kao i u prethodnim primjerima, a jedina razlika je što se ovdje koriste čvorovi **Decision Tree Learner** i **Decision Tree Predictor** zato jer se ta tehnika koristi u primjeru.



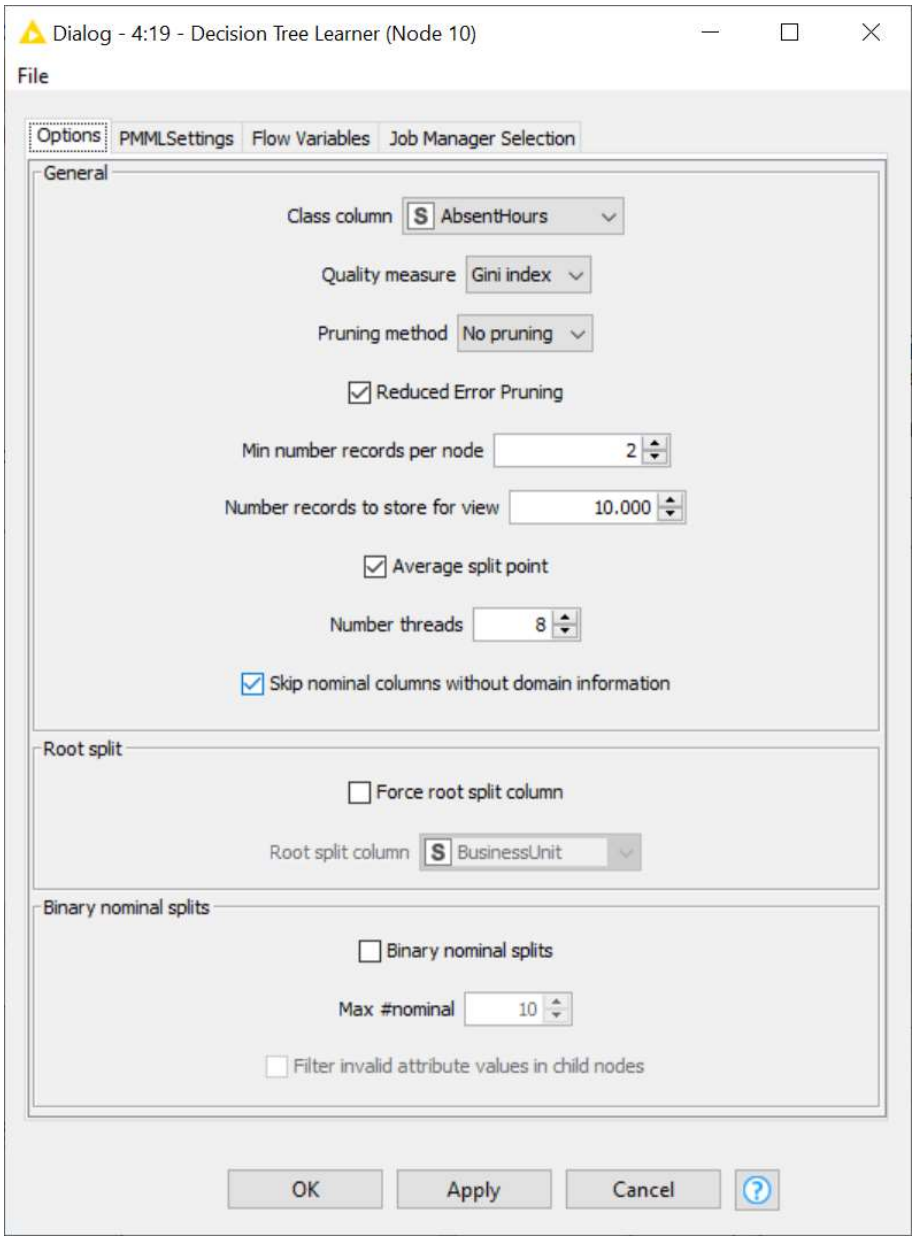
Slika 248. Kompletan hodogram

Prije analize matrice konfuzije i točnosti, slijedi osvrt na dva nova čvora i njihove postavke. Slika 249 prikazuje čvor **Decision Tree Learner** ili pomalo nespretan naziv Učenik stabla odlučivanja.

Slika 249. Čvor Decision Tree Learner

Slika 250 prikazuje postavke čvora **Decision Tree Learner**. U postavkama u vrhu definira se ciljna varijabla, odnosno naziv stupca koji mora biti tekstualni, odnosno kategorijalni (*Class column*). Ispod toga se bira način izbora značajke za čvor (*Quality measure*), a tu je ponuđen *Gini index* koji je opisan u uvodu ovog poglavlja. U nastavku se može izabrati MDL (eng *Minimum Description Length*) metodu orezivanja stabla. Radi se o metodi optimizacije nakon generiranja stabla. Počevši od listova, svaki se

čvor zamjenjuje svojom najpopularnijom kategorijom, ali samo ako se točnost predviđanja ne smanjuje. Opcija *Min number records per node* omogućuje izbor najmanjeg broja zapisa u čvoru. Ako je broj zapisa manji ili jednak, stablo se ne širi dalje. Nadalje je dostupna opcija definiranja maksimalnog broja zapisa za prikaz stabla (*Number records to store for view*). Postavljena opcija *Average split point* omogućuje da se vrijednost dijeljenja za numeričke atribute određuje prema srednjoj vrijednosti dviju vrijednosti atributa koje odvajaju dvije particije. Postavka *Number treads* omogućuje korištenje više jezgri procesora pri izračunu, a *Skip normal columns without domain information* omogućuje preskakanje stupaca, odnosno značajki, s previše jedinstvenih vrijednosti. Ako se zna da za prvo dijeljenje treba izabrati određenu značajku, to se može postaviti u oviru *Root split*.

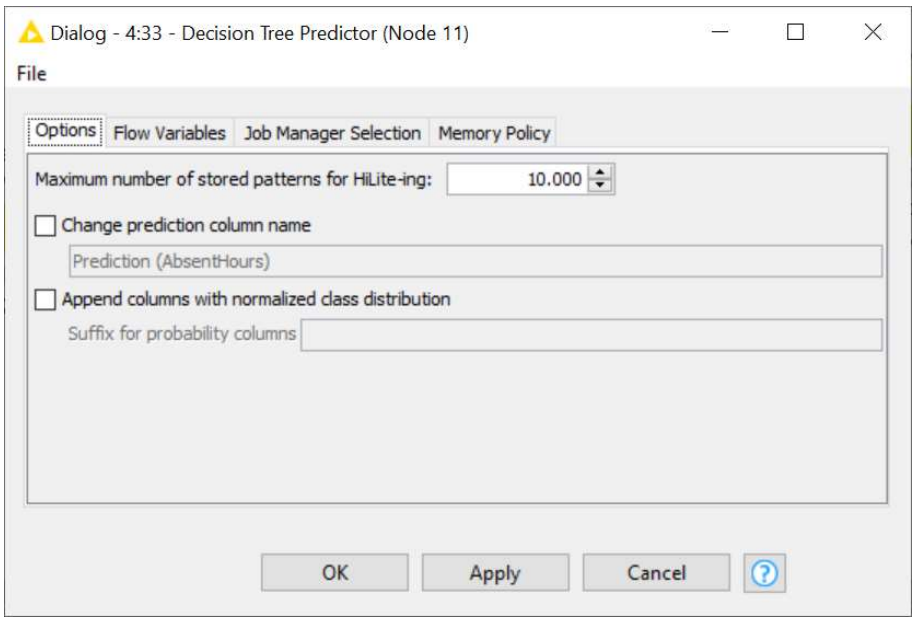


Slika 250. Postavke čvora Decision Tree Learner

Slika 251 prikazuje čvor **Decision Tree Predictor** ili Prediktor stabla odlučivanja.

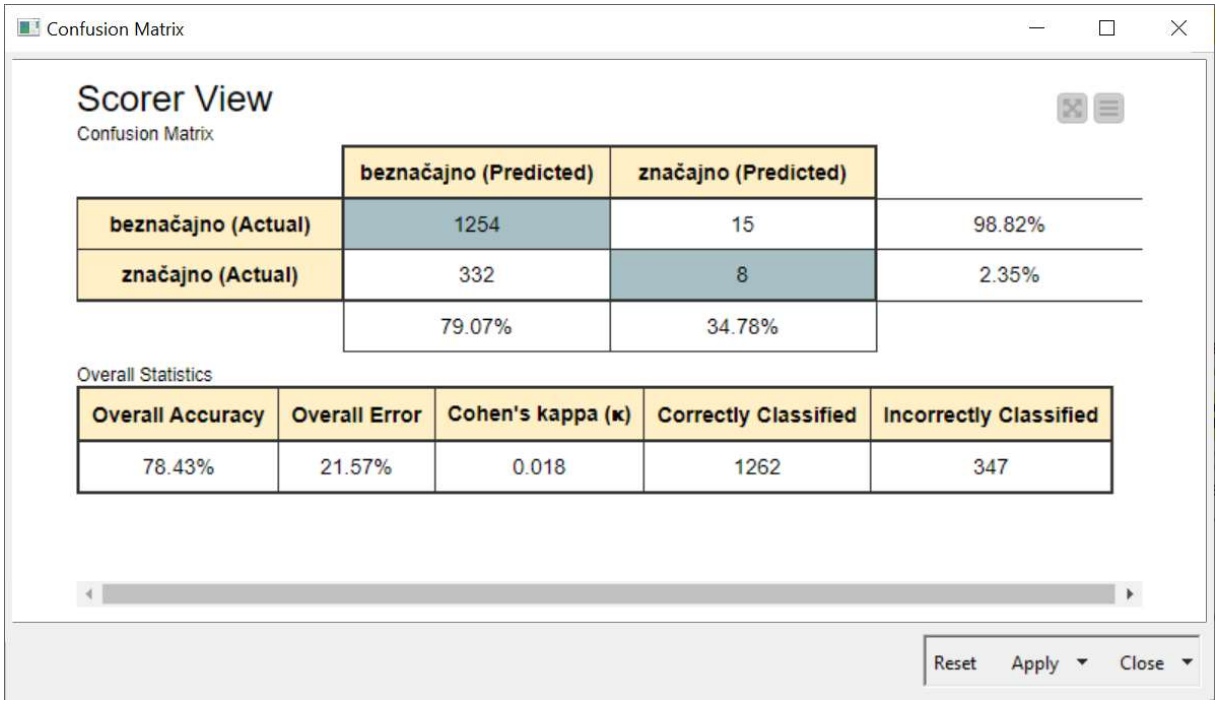
Slika 251. Čvor Decision Tree Predictor

Slika 252 prikazuje postavke čvora **Decision Tree Predictor**. Kao i u prethodnim modelima ove postavke u pravilu se ne mijenjaju.



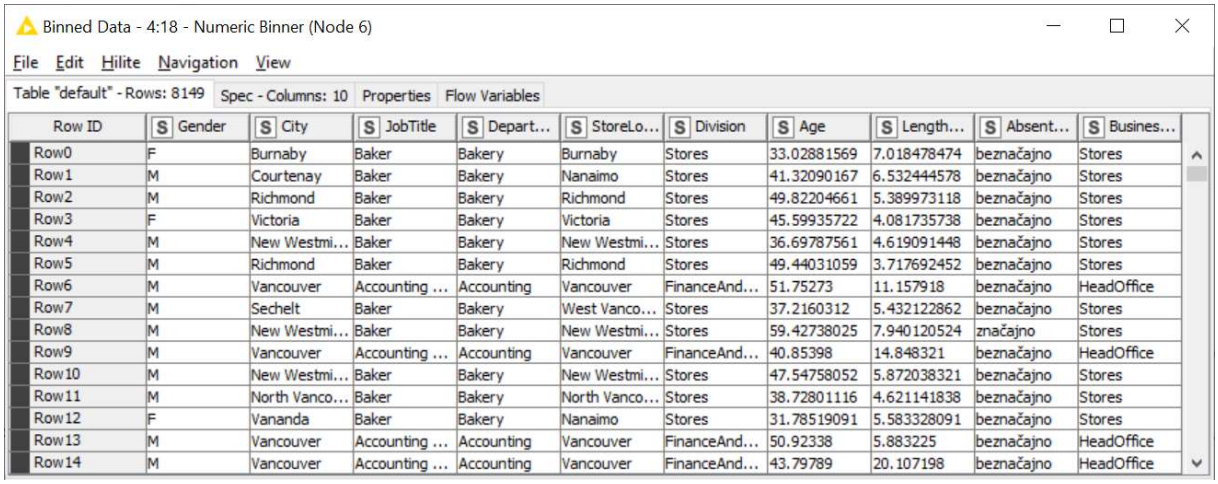
Slika 252. Postavke čvora Decision Tree Predictor

Slika 253 prikazuje matricu konfuzije i točnost modela koristeći čvor **Scorer (JavaScript**). Sveukupna točnost je skoro 80 % što bi se moglo činiti zadovoljavajuće. Ipak u matrici konfuzije postoji jedna neprihvatljiva vrijednost. Od 340 slučajeva, odnosno djelatnika koji su značajno izostajali s posla, model je točno prepoznao njih 8, odnosno 2,35 %. S obzirom da je većina djelatnika u grupi koja beznačajno izostaje s posla, model ima ukupnu točnost blizu 80 %, ali je izuzetno loš kod predikcije značajnog izostajanja. Može li se tu što učiniti ?



Slika 253. Matrica konfuzije

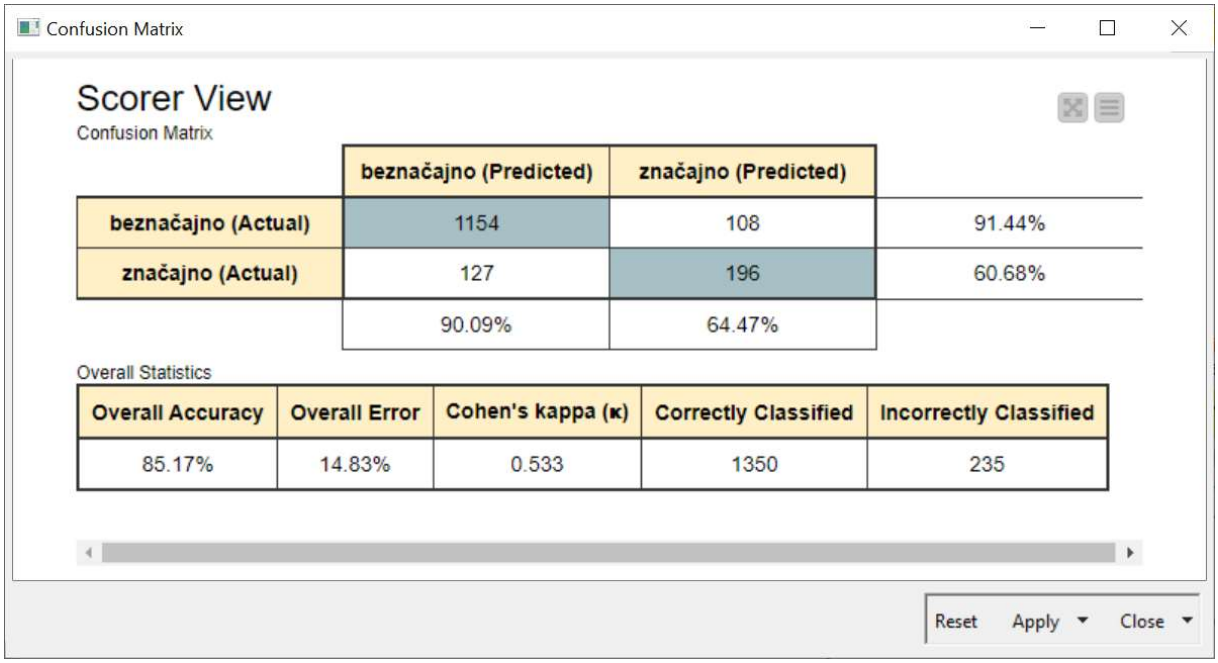
Za početak se mogu proanalizirati značajke koje se koriste za izradu modela. Slika 254 prikazuje tablicu s podacima.



Row ID	Gender	City	JobTitle	Depart...	StoreLo...	Division	Age	Length...	Absent...	Busines...
Row0	F	Burnaby	Baker	Bakery	Burnaby	Stores	33.02881569	7.018478474	beznačajno	Stores
Row1	M	Courtenay	Baker	Bakery	Nanaimo	Stores	41.32090167	6.532444578	beznačajno	Stores
Row2	M	Richmond	Baker	Bakery	Richmond	Stores	49.82204661	5.389973118	beznačajno	Stores
Row3	F	Victoria	Baker	Bakery	Victoria	Stores	45.59935722	4.081735738	beznačajno	Stores
Row4	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	36.69787561	4.619091448	beznačajno	Stores
Row5	M	Richmond	Baker	Bakery	Richmond	Stores	49.44031059	3.717692452	beznačajno	Stores
Row6	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	51.75273	11.157918	beznačajno	HeadOffice
Row7	M	Sechelt	Baker	Bakery	West Vanco...	Stores	37.2160312	5.432122862	beznačajno	Stores
Row8	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	59.42738025	7.940120524	značajno	Stores
Row9	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	40.85398	14.848321	beznačajno	HeadOffice
Row10	M	New Westmi...	Baker	Bakery	New Westmi...	Stores	47.54758052	5.872038321	beznačajno	Stores
Row11	M	North Vanco...	Baker	Bakery	North Vanco...	Stores	38.72801116	4.621141838	beznačajno	Stores
Row12	F	Vananda	Baker	Bakery	Nanaimo	Stores	31.78519091	5.583328091	beznačajno	Stores
Row13	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	50.92338	5.883225	beznačajno	HeadOffice
Row14	M	Vancouver	Accounting ...	Accounting	Vancouver	FinanceAnd...	43.79789	20.107198	beznačajno	HeadOffice

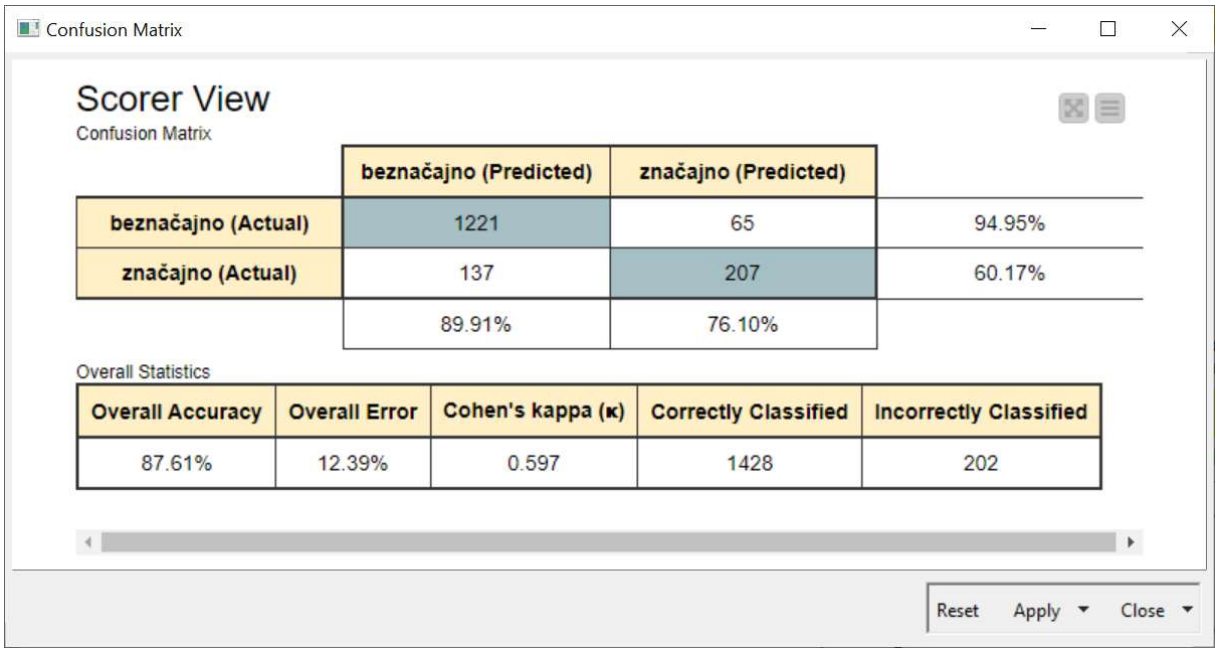
Slika 254. Podaci koji se koriste za izradu modela

Vidi se kako se za izradu modela koriste isključivo kategorijalne, odnosno tekstualne vrijednosti. To se vidi tako što je ispred naziva svake značajke u tablici slovo „S" koje označava da se radi o tekstualnoj varijabli (*String*). Spomenuto je u uvodnom dijelu kako tehnika stabla odlučivanja može koristiti i kontinuirane varijable, a očigledno je da postoje među podacima. Radi se o značajkama starost (*Age*) i radni staž (*LengthService*). S obzirom da se radi o decimalnim brojevima koji su prikazani kao kategorijalne tekstualne varijable, njihov doprinos točnosti modela uz ovakav prikaz nije optimalan. Razlog je što praktički svaka vrijednost u ta dva stupca pripada jednoj jedinstvenoj kategoriji. Bolji pristup bi bio da ih se konvertira u kontinuirane varijable, odnosno u realne brojeve. To se radi tako da se samo u čvoru **String To Number** dodaju ta dva stupca u konverziju. Nakon promjene opet se pokreću svi čvorovi i analiziraju rezultati. Slika 255 prikazuje matricu konfuzije.



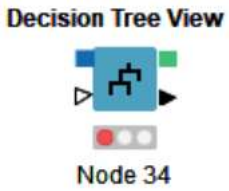
Slika 255. Matrica konfuzije nakon transformacije dvije tekstualne varijable u brojčane

Ukupna točnost porasla je na preko 85 %, a točnost kod predikcije značajnog izostajanja s posla porasla je s 2,35 % na preko 60 %. Za kraj se može pokušati povećati točnost modela uključenjem MDL metode orezivanja stabla odlučivanja. Radi se o metodi koja uklanja pojedine čvorove na krajevima stabla ako time ne smanjuje točnosti. Ta postavka se uključuje u čvoru **Decision Tree Learner**. Nakon uključenja rezultati su malo bolji.



Slika 256. Matrica konfuzije nakon uključenja MDL metode orezivanja

Za kraj poglavlja o klasifikacijskoj metodi stabla odlučivanja spomenut će se još dva čvora koji se koriste za prikaz rezultata modela u obliku stabla i u obliku niza pravila. Slika 257 prikazuje čvor **Decision Tree View** ili Prikaz stabla odlučivanja koji ima fleksibilniji prikaz stabla odlučivanja od čvora **Decision Tree Predictor**.

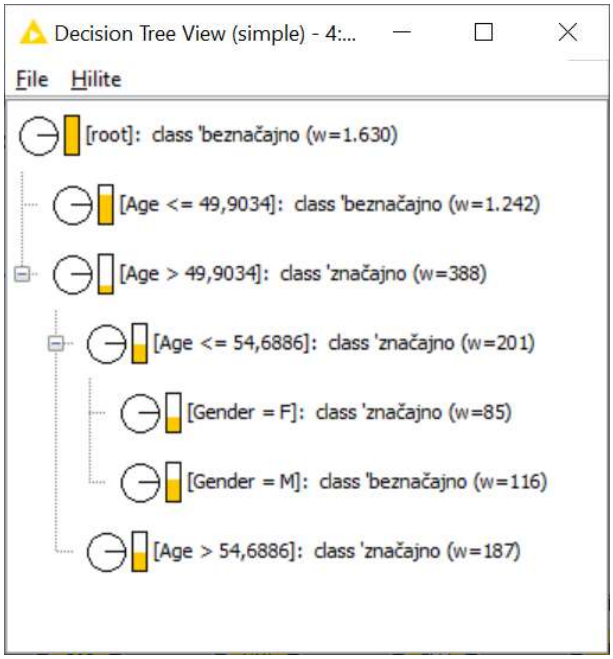


Slika 257. Čvor Decision Tree View

Slika 258 prikazuje čvor **Decision Tree to Ruleset** ili Stablo odlučivanja u skup pravila, omogućuje ispis pravila koja su definirana modelom.

Slika 258. Čvor Decision Tree to Ruleset

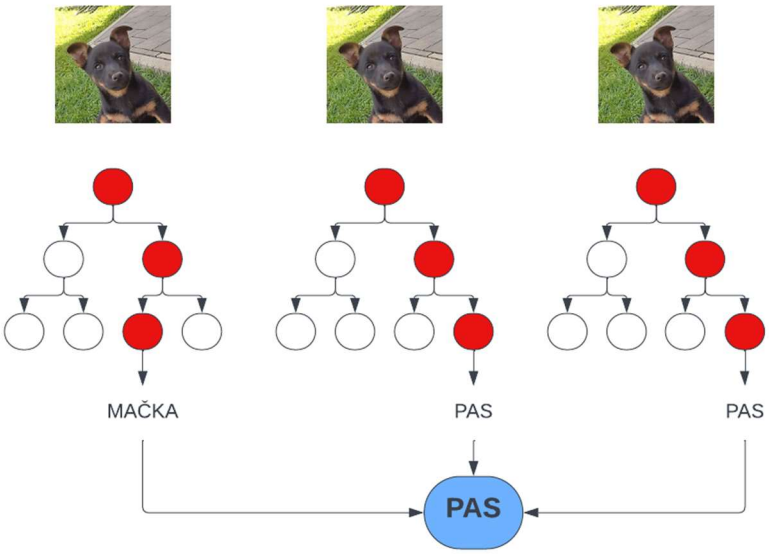
Zadnja dva opisana čvora neće biti postavljena na hodogram, ali za kraj poglavlja će se iz čvora **Decision Tree Predictor** izvesti pojednostavljena slika modela stabla odlučivanja koji je nastao kroz rješavanje ovog primjera. Slika 259 prikazuje taj model.



Slika 259. Pojednostavljena slika modela stabla odlučivanja

10. Tehnika slučajnih šuma

Nedostatak koji je opisan u literaturi i spomenut u opisu prethodno obrađenoj tehnici stabla odlučivanja je prenaučenost u slučaju da mu se postavi velik broj grana, odnosno podnaučenost ako se značajno smanji broj grana. Da bi se taj problem riješio uvodi se jednostavno rješenje, a to je za primjer klasifikacije izrada više stabala odlučivanja gdje se konačna odluka o kategoriji donosi glasanjem. Jedna od tehnika koja koristi takav pristup glasanja je tehnika slučajne šume i predlaže je Breiman početkom ovog stoljeća (Breiman, 2001). Ako se tehnika slučajnih šuma koristi kod regresijskih problema, tada se računa prosjek rezultata pojedinih stabala. Slika 260 prikazuje način rada tehnike slučajnih šuma za klasifikaciju.



Slika 260. Način rada tehnike slučajnih šuma

Pri stvarnom korištenju tehnike slučajnih šuma broj stabla odlučivanja može biti i nekoliko stotina, a pojedino stablo se generira iz uzorka kompletnog skupa podataka. Osim toga svako pojedino stablo ne uključuje sve značajke koje su dostupne u skupu podataka nego samo dio. Za kreiranje pojedinog stabla kod klasifikacije je preporuka da se za broj značajki uzme samo kvadratni korijen ukupnog broja, dok je kod regresije preporuka da se uzme trećina ukupnog broja značajki. Neće se ulaziti u detaljniju analizu tehnike, a jedan od razloga je i taj što sam model nije jednostavno interpretirati kao što je slučaj kod drugih tehnika (Brownlee, 2016; Hastie, et al., 2009).

Prednosti tehnike su neosjetljivost na šum, nije potrebna normalizacija, rezultati modela su u pravilu dobri i bez podešavanja parametara modela. Nedostatak je zahtjevnost za računalnim resursima (Egger, 2022).

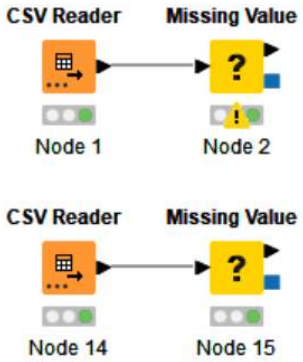
10.1. Priprema podataka

Deseti dan mjeseca listopada je svjetski dan mentalnog zdravlja. Tog dana se širom svijeta organiziraju događaji koji imaju za cilj osvijestiti važnost mentalnog zdravlja u populaciji. Zbog lošeg upravljanja COVID19 krizom i neopravdanog zaključavanja koje je uzrokovalo i masovan rad od kuće, došlo je do globalnog porasta problema vezanih uz mentalno zdravlje. Da bi se održala razina produktivnosti, u tvrtkama bi se odjeli za upravljanje ljudskim resursima trebali baviti tim područjem i na osnovu nekih pokazatelja, moći prepoznati problem i poduzeti korake da se on ublaži ili riješi. Skup

podataka koji se koristi u primjeru klasifikacije korištenjem tehnike slučajnih šuma pokriva značajke vezane uz mentalno zdravlje, a dostupan je na adresi:
https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out .
Nakon preuzimanja podataka s poslužitelja i dekomprimiranja preuzete datoteke, dobivaju se tri datoteke od kojih će se koristiti dvije: *train.csv* i *test.csv*. U datoteci train.csv nalazi se preko 22 tisuće redova s podacima. Stupci koji se nalaze u datoteci train.csv su sljedeći:

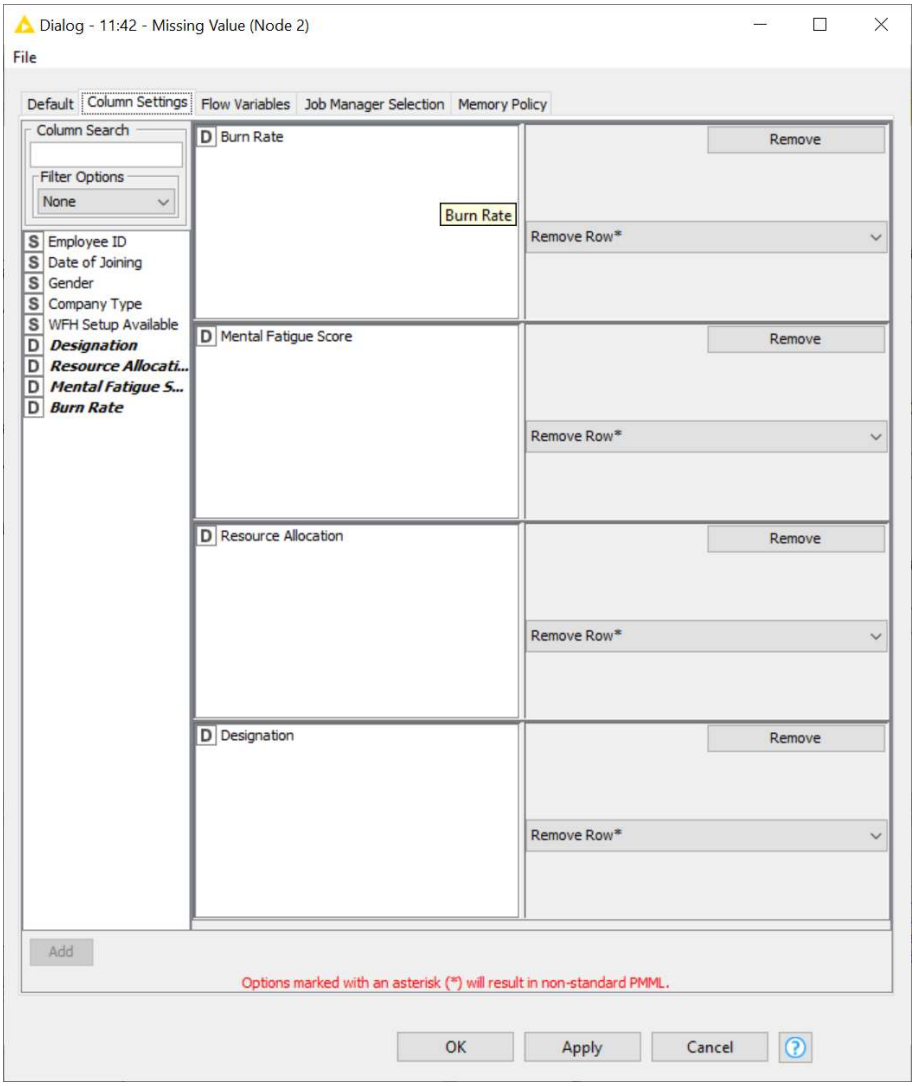
a) *Employee ID* - jedinstveni identifikator dodijeljen svakom zaposleniku
b) *Date of Joining* - datum kada se zaposlenik zaposlio u tvrtki
c) *Gender* – spol [*Male/Female*]
d) *Company Type* - vrsta tvrtke u kojoj zaposlenik radi [*Service/Product*]
e) *WFH Setup Available* - je li rad od kuće dostupan zaposleniku [*Yes/No*]
f) *Designation* - imenovanje zaposlenika za rad u organizaciji na skali od 0 do 5
g) *Resource Allocation* - količina resursa dodijeljena zaposleniku za rad, tj. broj radnih sati na skali od 1 do 10
h) *Mental Fatigue Score* - razina psihičkog umora s kojom se zaposlenik suočava na skali od 1 do 10
i) *Burn Rate* - vrijednost koju se treba predvidjeti i koja govori o stopi „izgaranja“ tijekom rada na skali od 0 do 1, pretvorena u kategorijalnu varijablu niska i visoka s granicom na 0,5.

U datoteci *test.csv* nalazi se nešto preko 12 tisuća redova podataka koji se sastoje samo od značajki, ali nedostaje ciljna varijabla. Za početak će se učitati podaci u dva **CSV Reader** čvora i izbaciti vrijednosti koje nedostaju. Radi se o prethodno korištenim i opisanim čvorovima koje nije potrebno posebno opisivati. Slika 261 prikazuje taj hodogram.



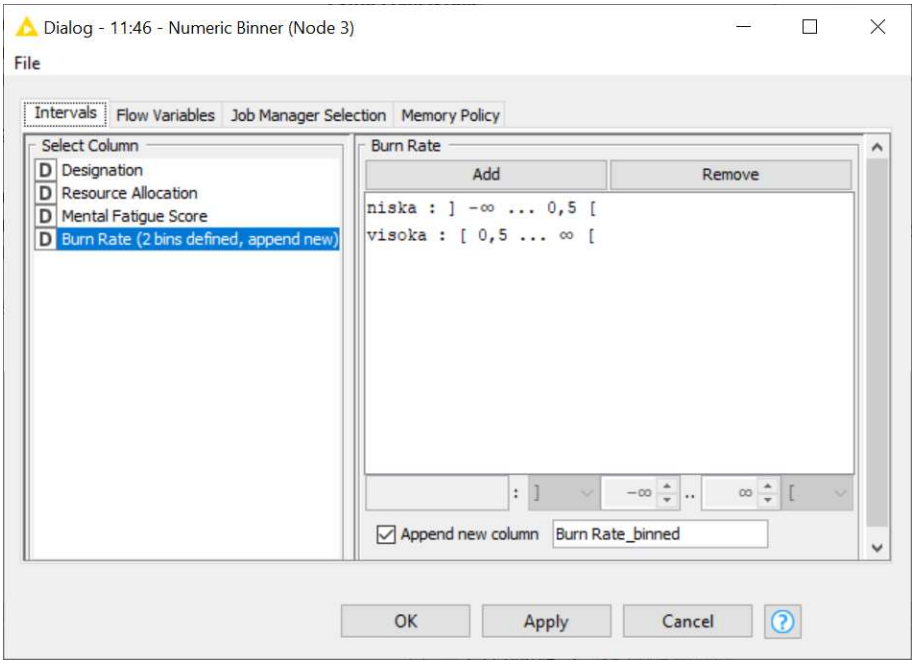
Slika 261. Hodogram s čvorovima za učitavanje i upravljanje s praznim ćelijama

U gornjem čvoru **Missing Value** uklonjeni su redovi gdje nedostaje neka od numeričkih značajki (*Designation, Resource Allocation, Mental Fatigue Score i Burn Rate*). Time se smanjio broj redova na nešto preko 18 tisuća. Slika 262 prikazuje postavke čvora za uklanjanje redova.



Slika 262. Postavke čvora Missing Value za brisanje redova

Sljedeći korak je pretvaranje numeričke vrijednosti u nominalnu i to ciljne varijable „stopa izgaranja" (*Burn rate)*. Ako se pogledaju vrijednosti koje se pojavljuju u tom stupcu vidi se kako su to vrijednosti između 0 i 1. Koristeći čvor **Numeric Binner** pretvorit će se brojevi manji od 0,5 u kategoriju „niska", a brojevi veći od 0,5 u kategoriju „visoka". Navedenu aktivnost pretvaranja nema potrebe raditi s podacima koji su učitani iz datoteke *test.csv* jer u tim podacima nema stupca s ciljnom varijablom „stopa izgaranja" (*Burn Rate)*. Slika 263 prikazuje postavke čvora **Numeric Binner**. Treba naglasiti kako je ta vrijednost uzeta proizvoljno, a da bi u ozbiljnoj primjeni sličnog modela netko stručan trebao definirati granicu između dvije kategorije ciljne varijable „stopa izgaranja" (*Burn Rate)*.



Slika 263. Postavke čvora Numeric Binner

Slika. 264 prikazuje podatke nakon transformacije zadnjeg stupca u dvije kategorije. Ono što je uočljivo, svi djelatnici u oba skupa podataka zaposleni su tijekom 2008. godine, a datum je zapisan kao tekstualna varijabla. Kao takav, datum nije od neke koristi, ali ako se pretvori u dan, mjesec i godinu, onda bi mogao doprinijeti točnosti modela jer je veća vjerojatnost da netko ima problema sa stresom na radnom mjestu ako je zaposlen duže. Trajanje zaposlenja je povezano s mjesecom zaposlenja pod pretpostavkom da se podaci analiziraju krajem 2008. ili početkom 2009. godine te taj stupac može doprinijeti točnosti modela. Tu se uvode novi čvorovi koji do sada nisu korišteni.

Binned Data - 11:46 - Numeric Binner (Node 3)

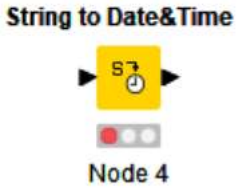
File Edit Hilite Navigation View

Table "default" - Rows: 18590 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	Employee ID	Date of...	Gender	Compa...	WFH S...	Design...	Resour...	Mental ...	Burn Rate	Burn R...
Row0	fffe32003000360033003200	2008-09-30	Female	Service	No	2	3	3.8	0.16	niska
Row1	fffe3700360033003500	2008-11-30	Male	Service	Yes	1	2	5	0.36	niska
Row3	fffe32003400380032003900	2008-11-03	Male	Service	Yes	1	1	2.6	0.2	niska
Row4	fffe31003900340031003600	2008-07-24	Female	Service	No	3	7	6.9	0.52	visoka
Row5	fffe3300350037003500	2008-11-26	Male	Product	Yes	2	4	3.6	0.29	niska
Row6	fffe33003300340039003100	2008-01-02	Female	Service	No	3	6	7.9	0.62	visoka
Row7	fffe32003600320037003400	2008-10-31	Female	Service	Yes	2	4	4.4	0.33	niska
Row10	fffe33003100330032003700	2008-03-16	Male	Product	Yes	2	3	5.3	0.5	visoka
Row11	fffe3400310035003800	2008-05-12	Male	Service	Yes	0	1	1.8	0.12	niska
Row14	fffe33003100330036003300	2008-05-14	Male	Product	Yes	1	3	5.8	0.51	visoka
Row15	fffe31003700350033003100	2008-02-03	Female	Service	Yes	3	5	4.7	0.32	niska
Row16	fffe33003200360037003000	2008-03-17	Male	Service	Yes	1	2	5.9	0.39	niska
Row17	fffe31003500350030003400	2008-03-28	Male	Service	No	3	6	6.7	0.59	visoka
Row18	fffe31003000380035003800	2008-05-29	Female	Product	Yes	2	4	4	0.22	niska
Row20	fffe31003300360039003000	2008-08-31	Male	Product	No	2	4	7.6	0.57	visoka

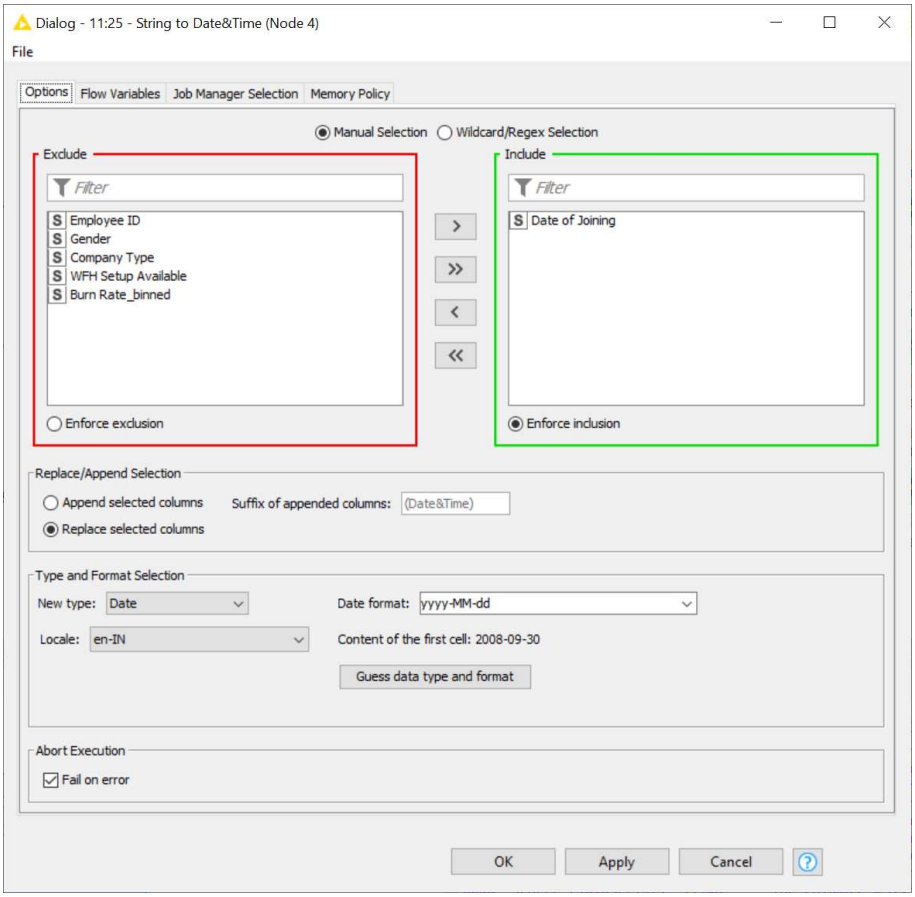
Slika. 264. Podaci nakon klasifikacije stupca Burn Rate

Slika 265 prikazuje čvor **String to Date&Time** ili Tekstualni zais u datumsko/vremenski zapis. Kao što mu samo ime kaže, taj čvor služi za pretvaranje iz tekstualnog oblika zapisa datuma u oblik u kojem je definirana godina, mjesec i dan u mjesecu.



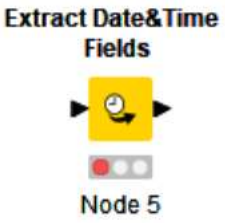
Slika 265. Čvor String to Date&Time

Slika 266 prikazuje postavke čvora **String to Date&Time**. U vrhu postavki biraju se stupci koji se pretvaraju u datumski oblik prebacivanjem u okvir omeđen zelenim pravokutnikom. U dijelu *Replace/Append Selection* bira se mogućnost zamjene stupca s novim ili dodavanje još jednog stupca. U donjem dijelu postavki bira se format datuma pri čemu je dostupno automatsko prepoznavanje formata klikom na gumb *Guess date type and format*.



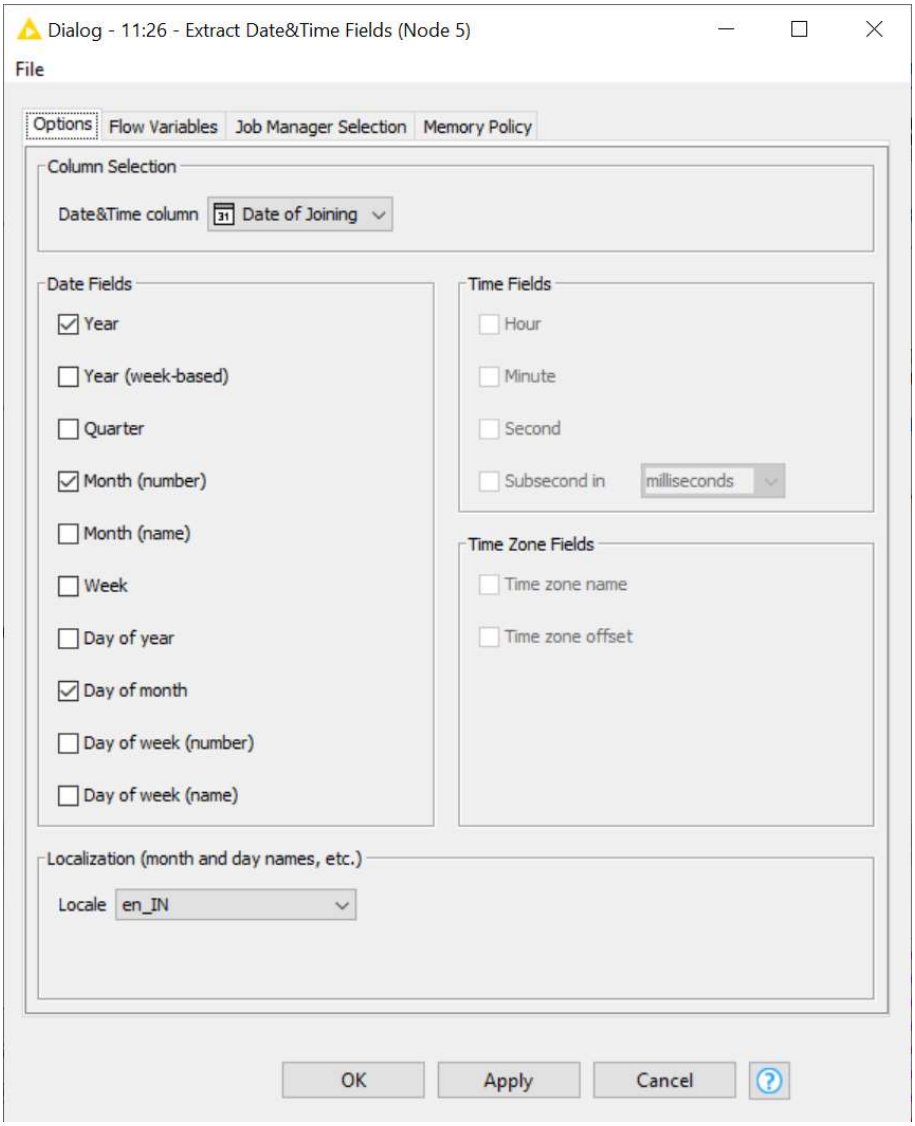
Slika 266. Postavke čvora String to Date&Time

Da bi se iz datumskog formata dobilo tri stupca s izlučenom godinom, mjesecom i danom, koristi se čvor **Extract Date&Time Fields** ili Ekstrakcija datumsko/vremenskih polja. Slika 267 prikazuje taj čvor.



Slika 267. Čvor Extract Date&Time Fields

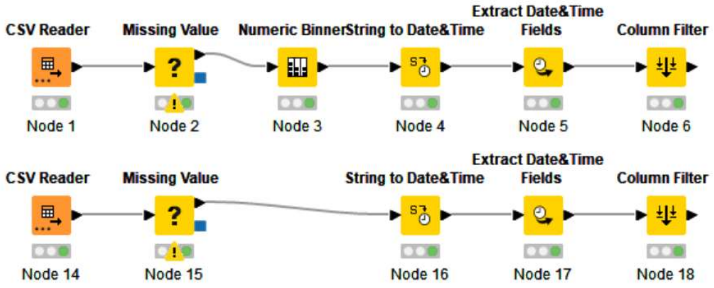
Slika 268 prikazuje postavke tog čvora na kojemu se u vrhu bira koji stupac se želi raščlaniti, dok se u okvirima *Date Fields i Time Fields* biraju novi stupci koji će biti generirani iz datumskog zapisa. U ovom slučaju je zapisan samo datum, ali ovaj čvor može ekstrahirati i podatke o vremenu, odnosno sate, minute, sekunde i milisekunde, kao i vremenski zonu.



Slika 268. Postavke čvora Extract Time&Date Fields

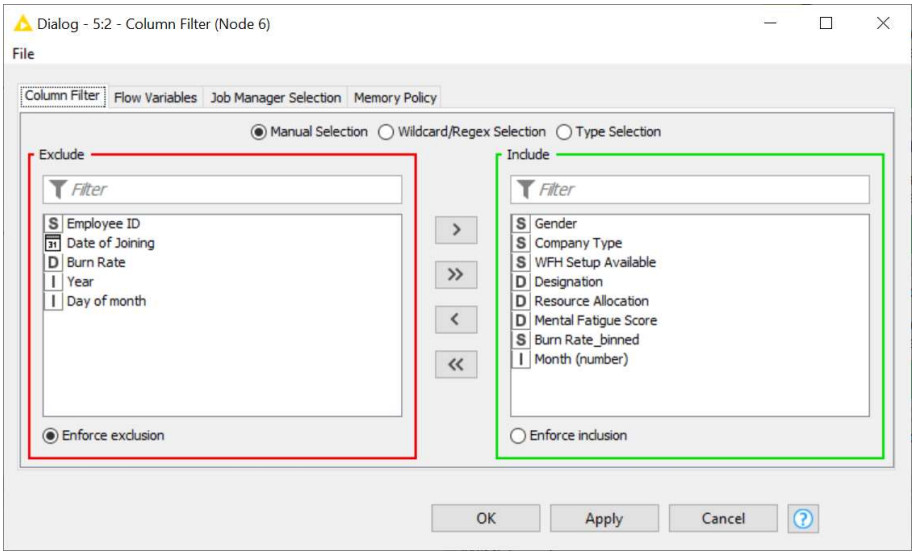
Kao što je vidljivo iz postavki, iz datumskog zapisa generirani su stupci s godinom, mjesecom i danom.

Slika 269 prikazuje prvi dio hodograma na kojem su čvorovi za učitavanje i pripremu podataka. S obzirom da ovaj skup podataka ima razdvojene podatke za treniranje modela od podataka za testiranje, kreirana su dva niza čvorova.



Slika 269. Dio hodograma za pripremu podataka

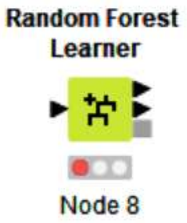
Na kraju svakog niza su dodani čvorovi **Column Filter**, koji služe za uklanjanje stupaca. Uklonjeni su stupci s danom i godinom zaposlenja, jedinstvenim brojem zaposlenika, tekstualnim zapisom dana zaposlenja i konačno stupac *Burn Rate* koji je iz numeričke pretvoren u kategorijalnu varijablu i više nije potreban. Važno je naglasiti kako se stupac *Burn Rate* mora ukloniti, jer ako se ostavi tehnika će prepoznati vezu između te numeričke značajke i kategorijalne ciljne varijable koja je nastala iz značajke i pojavit će se točnost od 100 %. Takva točnost najčešće ukazuje da je među značajkama za treniranje ostala varijabla koja s ciljanom varijablom ima faktor povezanosti 1 te tu značajku treba ukloniti.



Slika 270. Postavke čvora Column Filter

10.2. Izrada modela temeljenog na klasifikatoru slučajnih šuma

Konačno slijedi izrada modela temeljenog na klasifikatoru slučajnih šuma. Očekivano, model se sastoji od dva čvora, kao što je bio slučaj i u prethodnim primjerima. Slika 271 prikazuje prvi od ta dva čvora i to **Random Forest Learner** ili opet pomalo nespretan Učenik slučajne šume.



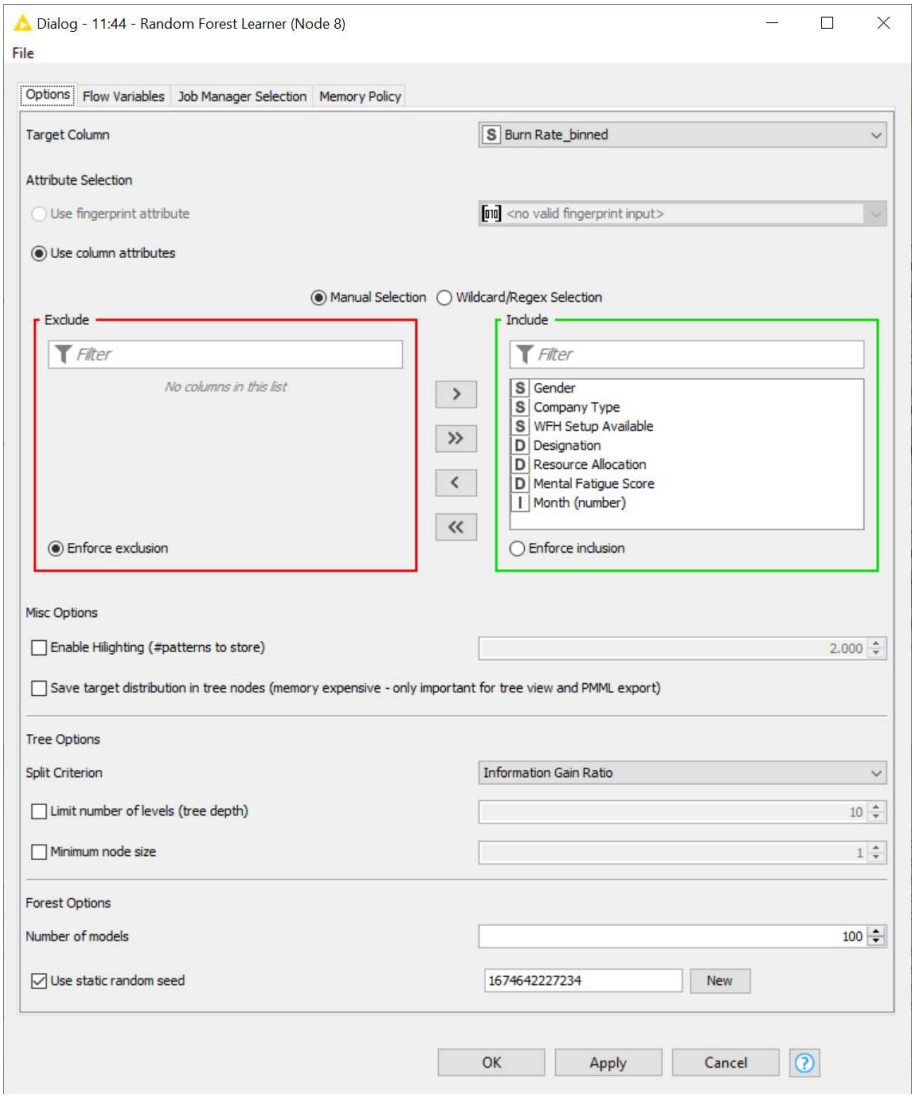
Slika 271. Čvor Random Forest Learner

Slika 272 prikazuje postavke čvora **Random Forest Learner**. Prvi padajući izbornik je očekivano ciljna varijabla. Nakon toga je moguće birati način selekcije značajki pri čemu postoji mogućnost izbora *Use fingerprint attribute* koji omogućuje unos vektora kao posebne varijable, pri čemu vektori moraju biti iste dužine. Osim toga je na raspolaganju standardni izbor značajki korištenjem zelenog i crvenog pravokutnika oko polja s unosom.

Sljedeća opcija (*Enable Hilighting*) omogućuje pohranu informacije o odabranom broju redaka u svakom čvoru, iduća opcija (*Save target distribution in tree nodes*) omogućuje pohranu distribucije vrijednosti ciljne kategorije u svakom čvoru. Obje opcije znatno doprinose potrošnji radne memorije te je preporuka da ih se ne koristi ako nije nužno.

U skupini opcija vezanih za kreiranje stabla odlučivanja može se izabrati kriterij podjele i na raspolaganju su tri mogućnosti (*Information Gain, Information Gain Ratio i Gini Index*). Osim toga može se ograničiti dubina kreiranja čvorova, kao i minimalni broj zapisa u podređenim čvorovima.

U skupini opcija vezanih za kreiranje modela je broj stabala u modelu. Zadan broj je 100, a po potrebi može se postaviti i veći broj.

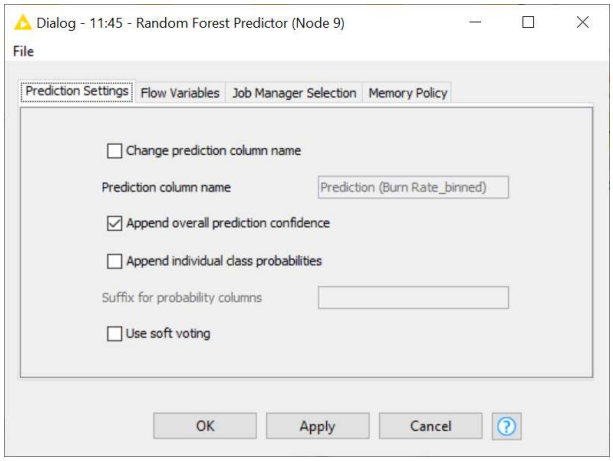


Slika 272. Postavke čvora Random Forest Learner

Slika 273 prikazuje čvor **Random Forest Predictor** ili Prediktor slučajnih šuma koji služi za predikciju kategorije na osnovu modela kreiranog čvorom **Random Forest Learner** i podataka.

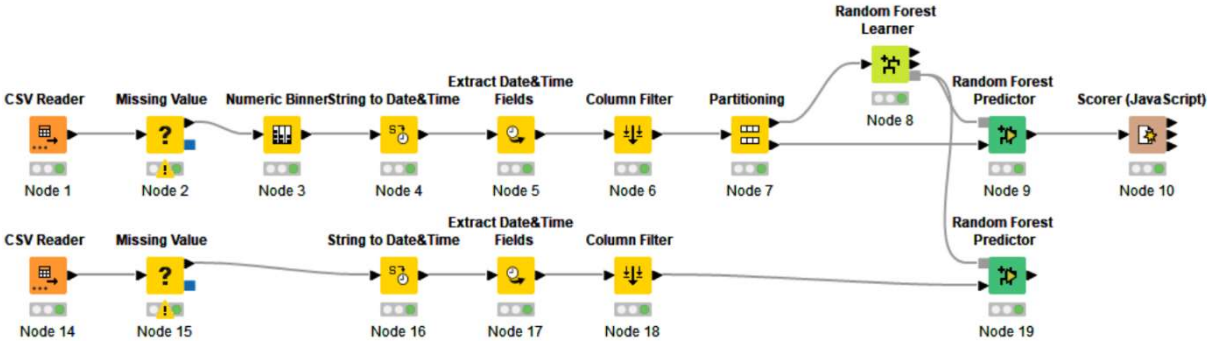
Slika 273. Čvor Random Forest Predictor

Slika 274 prikazuje postavke čvora **Random Forest Predictor** u kojima se mogu postaviti dodatne opcije prikaza pouzdanosti predviđanja. Omogućavanjem mekog glasovanja (*Use soft voting*) tehnika uzima u obzir i spremljene vjerojatnosti, dok kod tvrdog glasovanja prevladava najveći broj glasova.



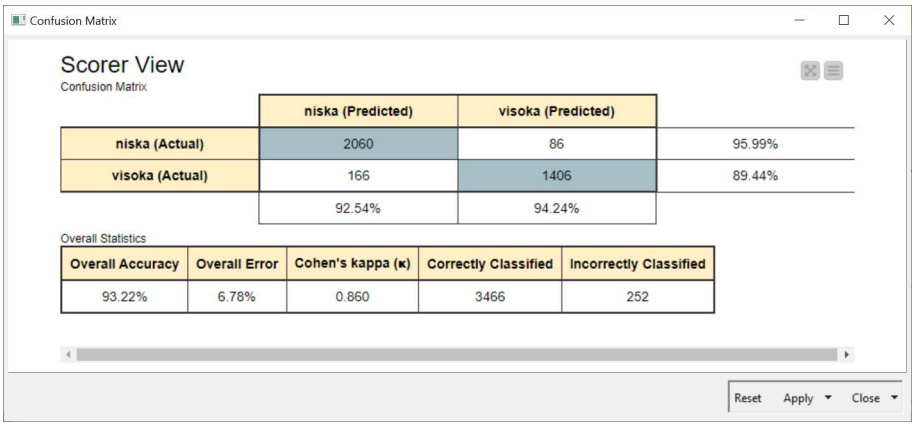
Slika 274. Postavke čvora Random Forest Predictor

Slika 275 prikazuje kompletan hodogram modela slučajnih šuma.



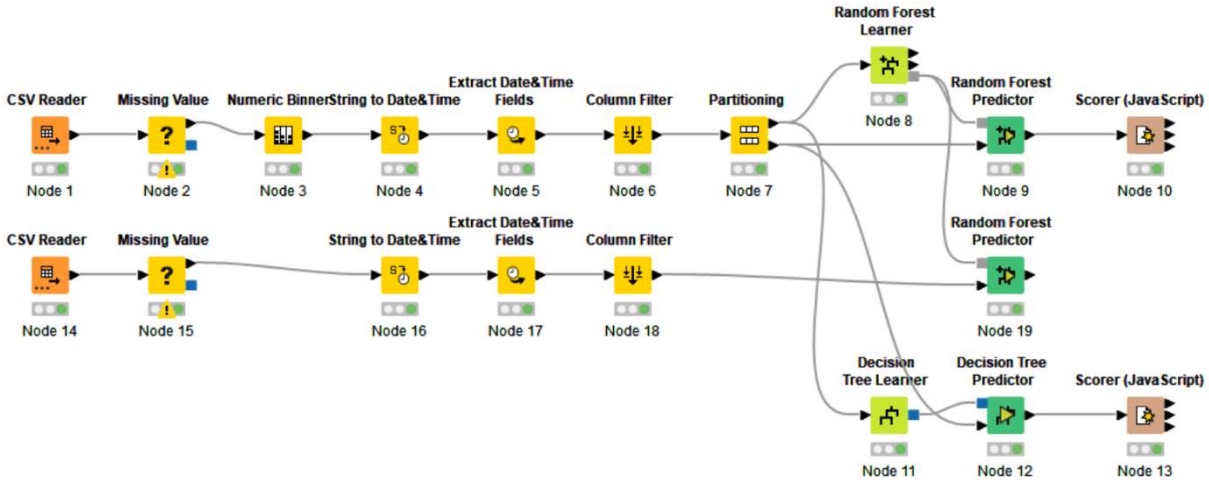
Slika 275 – Kompletan hodogram modela slučajnih šuma

Slika 276 prikazuje matricu konfuzije i ukupnu točnost modela.



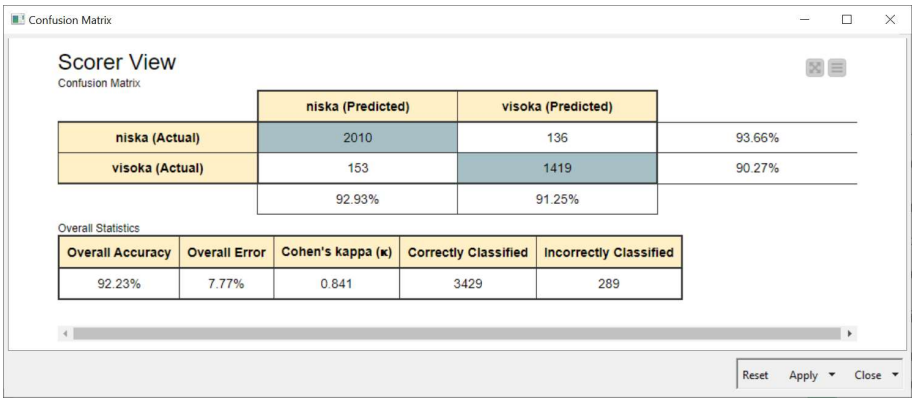
Slika 276. Matrica konfuzije i ukupna točnost modela

S obzirom da je tehnika slučajnih šuma na neki način nadogradnja tehnike stabla odlučivanja, u hodogram se mogu dodati i čvorovi koji čine model stabla odlučivanja da bi se utvrdilo koliko je model slučajnih šuma bolji od modela stabla odlučivanja. Slika 277 prikazuje nadograđeni hodogram s dodatna tri čvora koji čine model stabla odlučivanja.



Slika 277. Hodogram s uključenim modelom stabla odlučivanja za usporedbu

Slika 278 prikazuje matricu konfuzije za model stabla odlučivanja, kao i ukupnu točnost. Razlika je nešto manje od 1 %.

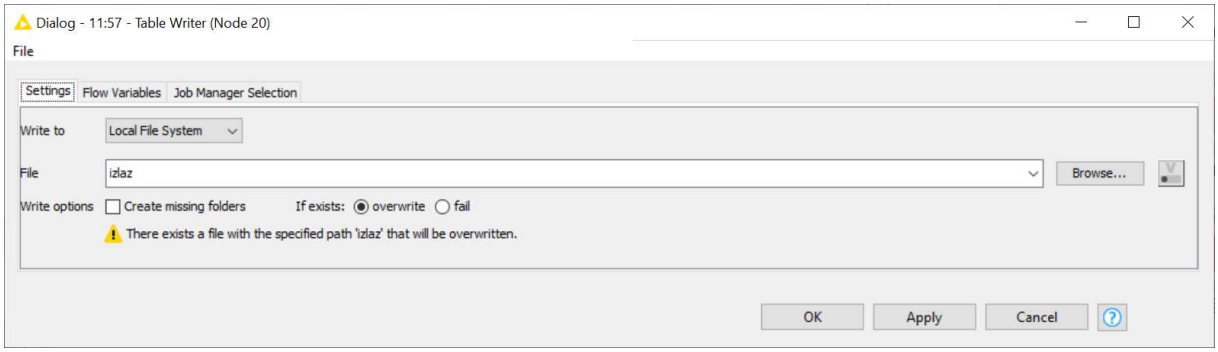


Slika 278. Matrica konfuzije za model stabla odlučivanja s istim podacima

Specifičnost ovog hodograma je da su podaci za treniranje i testiranje bili odvojeni, a u podacima za testiranje nije bilo stupca s ciljnom varijablom. Na taj način nije se mogla utvrditi točnost modela na tim podacima. Nakon klasifikacije slučajeva kojima pripadaju testni podaci, tablica treba biti spremljena za kasniju analizu. U ovom slučaju bit će spremljena u format s nastavkom *table*. Radi se o formatu koji program KNIME koristi za pohranu tabličnih podataka. Slika 279 prikazuje čvor **Table Writer** ili Zapisivač tablice koji služi za spremanje tabličnih podataka u taj format datoteke.

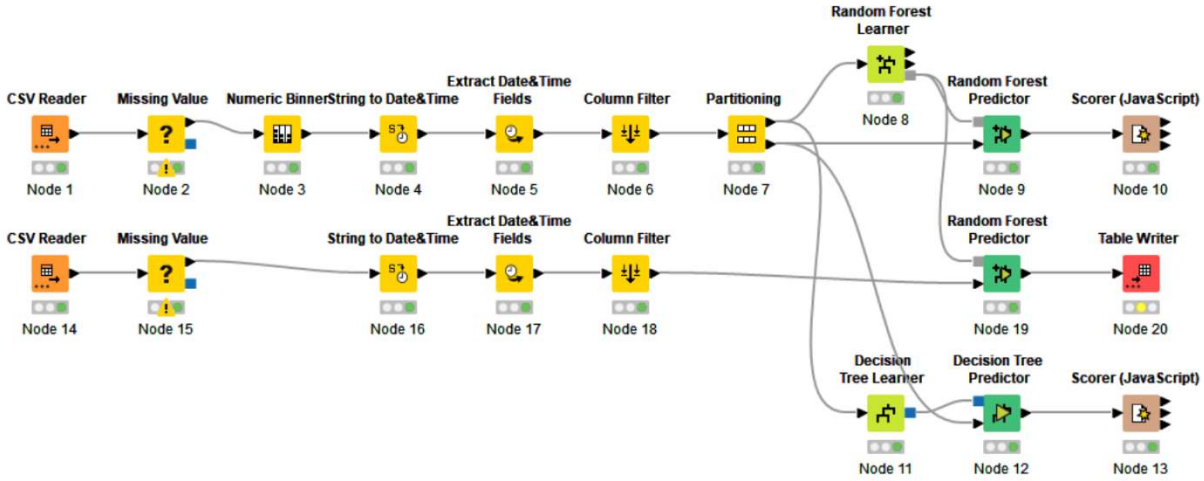
Slika 279. Čvor Table Writer

Slika 280 prikazuje postavke čvora **Table Writer** koji od opcija zahtijeva ime datoteke kojem će se dodati nastavak *.table* i spremiti u datotečni sustav računala ili na neku drugu lokaciju ovisno o odabiru.



Slika 280. Postavke čvora Table Writer

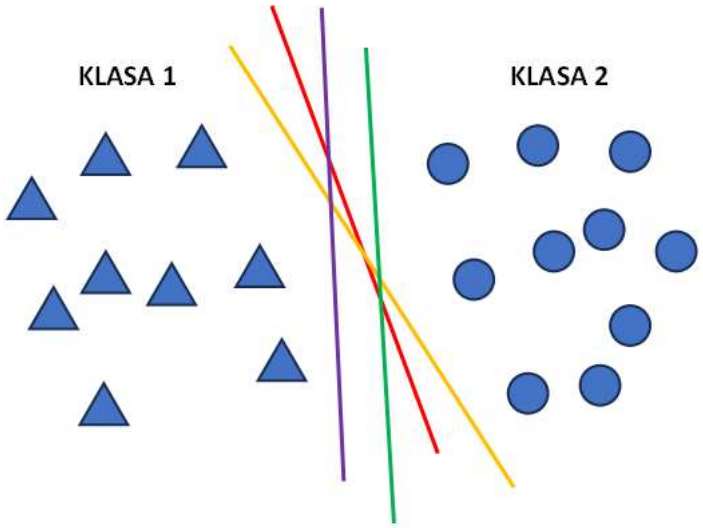
Slika 281 prikazuje kompletan hodogram koji uključuje i spremanje podataka u tabličnom formatu.



Slika 281. Kompletan hodogram

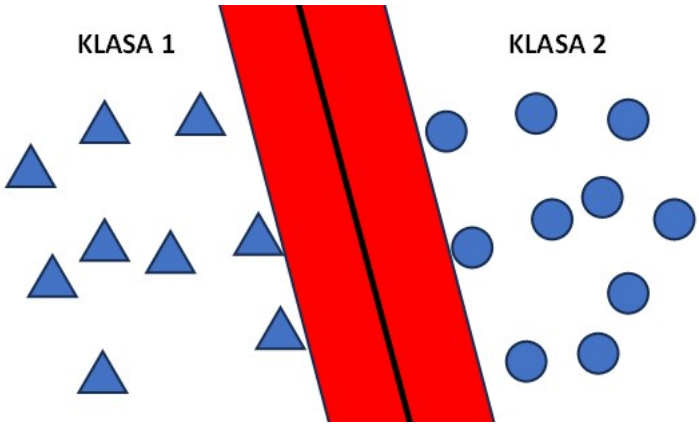
11. Metoda potpornih vektora

Metoda potpornih vektora pripada metodama nadziranog strojnog učenja i koristi se za rješavanje klasifikacijskih i regresijskih problema. Bez korištenja matematičkih izraza u nastavku će se pokušati objasniti specifičnosti ove metode. Slika 282 prikazuje slučajeve koji pripadaju dvjema kategorijama i može ih se podijeliti pravcem. Očigledno je kako postoji beskonačno mnogo pravaca koji mogu podijeliti te dvije kategorije slučajeva, odnosno pripadajuću ravninu u dvije poluravnine. Pojam pravca koristi se kada se ima samo dvije dimenzije, a opći naziv za element koji dijeli višedimenzionalni prostor na dva dijela je hiperravnina. Ako se radi o jednodimenzionalnom prostoru, njega dijeli točka. Dvodimenzionalni prostor dijeli pravac, a trodimenzionalni dijeli ravnina. Broj dimenzija hiperravnine jednak je broju dimenzija prostora koji dijeli umanjen za 1 (Brownlee, 2016).



Slika 282 . Pravci dijele ravninu u dva dijela

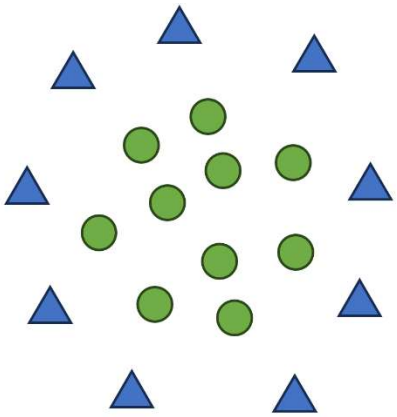
Za optimalno rješenje uzima se ona hiperravnina koja maksimizira marginu između dvije kategorije skupa podataka. Pojam margina označava udaljenost hiperravnine od najbližeg slučaja iz skupa podataka koji je potrebno klasificirati. Slika 283 prikazuje hiperravninu u sredini crvene površine koja dijeli dvije kategorije podataka s tim da je margina maksimalna.



Slika 283. Podjela ravnine s maksimalnom marginom

Metoda potpornih vektora poznaje pojam tvrde i meke margine. Kod tvrde margine nije dopušteno narušavanje uvjeta margine, dok je kod meke to dopušteno i toleriraju se greške ulaska u područje

margine ili čak na krivu stranu hiperravnine (Géron, 2022). Ipak, neki problemi uključuju slučajeve koji nisu linearno djeljivi. Slika 284 prikazuje dvije kategorije podataka koji nisu djeljivi pravcem u dvodimenzionalnom prostoru.

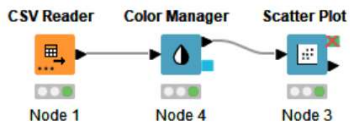


Slika 284 – Problem klasifikacije koji nije rješiv pravcem

Da bi se zorno prikazalo kako taj problem rješava metoda potpornih vektora i da bi se čitatelji upoznali s još jednim čvorom, potrebno je preuzeti datoteku s adrese
https://github.com/kristian1971/knimeprirucnik_v1/blob/main/svm3.csv
klikom na dugme sa strjelicom prema dolje uz desni rub prozora web preglednika. Za kontrolu ta datoteku se može otvoriti u programu Blok za pisanje (*Notepad*). Slika 285 prikazuje dio podataka iz datoteke u Bloku za pisanje. Radi se o četiri stupca podataka, tri stupca su koordinate na 3 osi, a četvrti stupac je kategorija kojoj slučaj pripada.

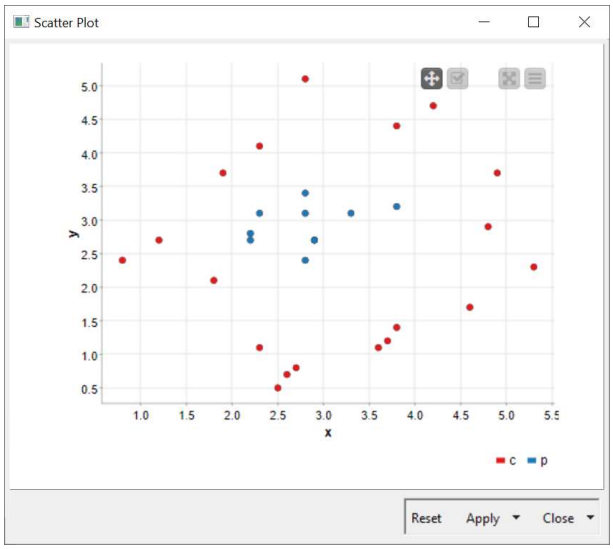
Slika 285. Gornji dio prozora Bloka za pisanje sa zalijepljenim podacima

Sljedeći korak je kreiranje jednostavnog hodograma koji će prikazati grafički slučajeve iz CSV datoteke uzevši u obzir samo X i Y os, a čvor **Color Manager** ili Upravljač bojama će svakoj kategoriji dodijeliti jedinstvenu boju. S obzirom kako su svi čvorovi u hodogramu korišteni u prethodnim poglavljima ili se spominju kasnije, nema potrebe posebno ih opisivati. Slika 286 prikazuje hodogram.



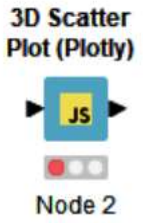
Slika 286. Hodogram za 2D grafički prikaz

Slika 287 prikazuje dvodimenzionalni prikaz slučajeva koje treba pravcem podijeliti u dvije kategorije. Očigledno je kako je to nemoguće u dvije dimenzije, ali je moguće ako se uvede još jedna dimenzija. Datoteka koja se koristi ima u trećem stupcu podatke za Z os pa slijedi prikaz slučajeva u tri osi. Pri tom će se koristiti novi čvor po imenu **3D Scatter Plot (Plotly)** ili 3D raspršeni dijagram (Plotly).



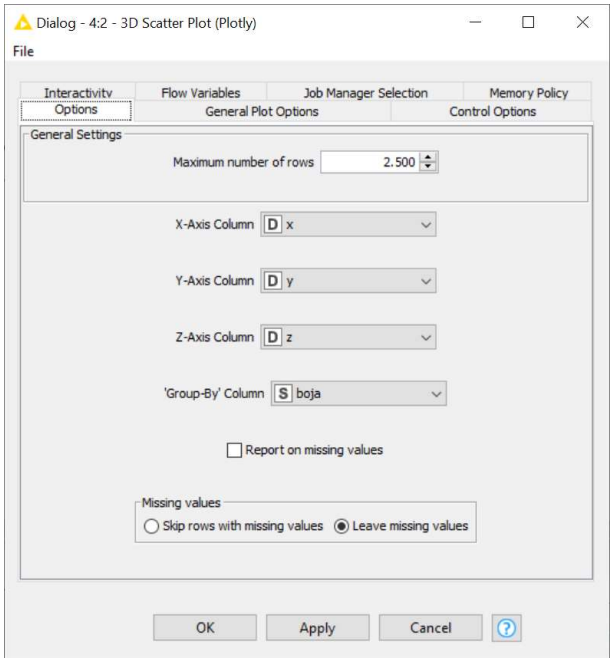
Slika 287. 2D grafički prikaz slučajeva

Slika 288 prikazuje čvor **3D Scatter Plot (Plotly)** koji omogućuje trodimenzionalne prikaze podataka. Čvor ne dolazi sa zadanom instalacijom programa KNIME te ga je potrebno naći u repozitoriju KNIME Hub i prevući na uređivač hodograma kako bi se pokrenula instalacija čvora. Nakon instalacije može se koristiti.



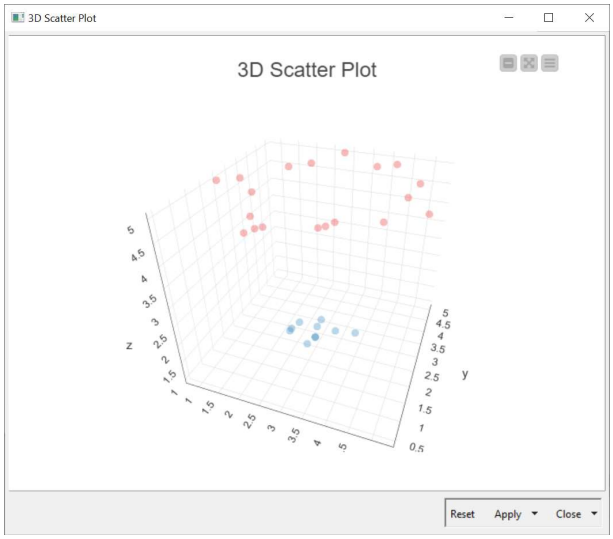
Slika 288. Čvor 3D Scatter Plot (Plotly)

Slika 289 prikazuje postavke čvora **3D Scatter Plot (Plotly)**. Postavke su intuitivne i potrebno je iz padajućih izbornika izabrati numeričke varijable za svaku od tri osi. Postoji i mogućnost grupiranja pri čemu se bira neka kategorijalna varijabla. Osim toga postoji još nekoliko opcija vezanih uz prazne ćelije što se susretalo i prije u postavkama čvorova.



Slika 289. Postavke čvora 3D Scatter Plot (Plotly)

Slika 290 prikazuje i treću dimenziju za učitan skup podataka. I dalje se vidi jedna kategorija slučajeva u obliku prstena, dok je druga nakupina u sredini tog prstena, ali se to vidi iz perspektive koju omogućuje treća Z os. Svakako treba obratiti pažnju na vrijednosti trećeg stupca za dvije kategorije „c" i „p" u tablici s podacima. Za kategoriju „c" vrijednost je 5, dok je za kategoriju „p" vrijednost trećeg stupca, odnosno Z osi jednaka 1. Ta razlika je vidljiva u grafikonu ako se pažljivije pogleda. Grafikon omogućuje rotaciju po svim osima tako da je moguće mijenjati kut gledanja.



Slika 290. Trodimenzionalni točkasti grafikon

Iz ovog primjera je očito kako metoda potpornih vektora koristi „trik" koji omogućuje razdvajanje linearno nedjeljivih podataka dodavanjem nove dimenzije. Očito je kako je sada moguće „provući" ravninu između prikazana dva skupa podataka. Sljedeći korak je rješavanje primjera koristeći metodu potpornih vektora, a prije toga treba naglasiti da se ta metoda na engleskom jeziku naziva *Support Vector Machines* dok je kratica za istu SVM. Ta tri slova koriste se i u nazivu čvorova.

Na kraju treba navesti i prednosti ove tehnike. Ona dobro radi i s malim skupom podataka za treniranje te ima mali broj parametara modela koji se mogu podešavati. Od nedostataka treba spomenuti nužnost normalizacije podataka, osjetljivost na šum, zahtjevnost za računalnim resursima i ovisnost izbora vrste treniranja (*HyperTangent, Polynomial and RBF*) o primjeni.

11.1. Priprema podataka

Jedan od problema s kojim se susreću hoteli je otkazivanje rezervacija. Ta pojava utječe na alociranje resursa, kako ljudskih tako i materijalnih, a tehnike strojnog učenja mogu donekle pomoći da se smanji neizvjesnost u procesu rezerviranja i otkazivanja rezervacija. S obzirom na karakteristike rezervacije, moguće je predvidjeti hoće li i za koliko smještajnih kapaciteta rezervacije biti otkazane. Kako bi se testirala metoda potpornih vektora, kreirat će se model koji će predviđati hoće li pojedina rezervacija biti otkazana ili neće.

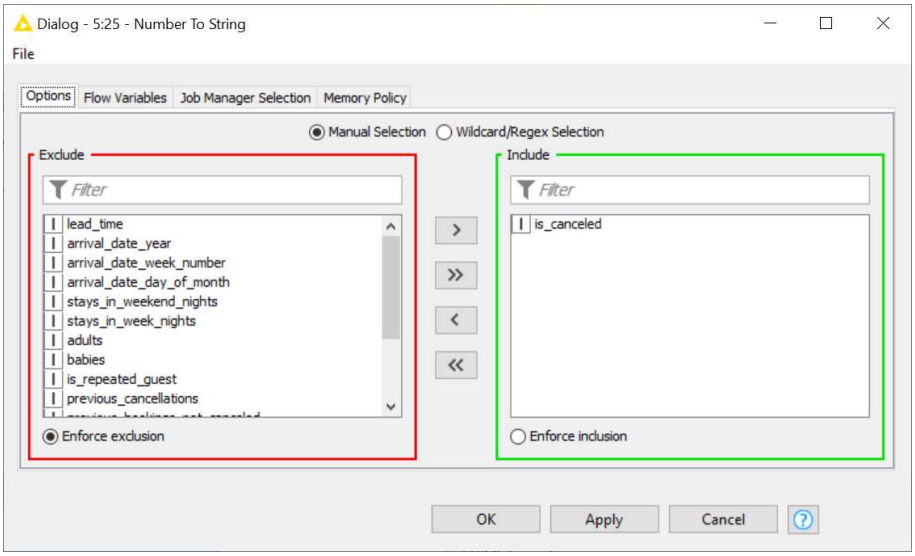
Skup podataka koji će se koristiti za treniranje i testiranje modela sastoji se od preko 120 000 rezervacija hotelskog smještaja s cijelim nizom značajki koje se mogu koristiti za treniranje modela. Skup podataka dostupan je na adresi: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand . Značajke koje se nalaze u skupu su sljedeće:

- *hotel* – Vrsta hotela (H1 = Resort Hotel ili H2 = City Hotel).
- *is_canceled* - Vrijednost koja pokazuje je li rezervacija otkazana (1) ili ne (0).
- *lead_time* - Broj dana koji je prošao između datuma unosa rezervacije u PMS i datuma dolaska.
- *arrival_date_year* - Godina datuma dolaska.
- *arrival_date_month* - Mjesec datuma dolaska.
- *arrival_date_week_number* - Broj tjedna godine za datum dolaska.
- *arrival_date_day_of_month* - Datum dolaska.
- *stays_in_weekend_nights* - Broj noćenja vikendom (subota ili nedjelja) u kojima je gost boravio ili rezervirao boravak u hotelu.
- *stays_in_week_nights* - Broj noćenja u tjednu (od ponedjeljka do petka) u kojima je gost boravio ili rezervirao boravak u hotelu.
- *adults* - Broj odraslih osoba.
- *children* - Broj djece.
- *babies* - Broj beba.
- *meal* - Vrsta rezerviranog obroka. Kategorije su u standardnim ugostiteljskim paketima obroka: BB, HB, FB.
- *country* - Zemlja podrijetla. Kategorije su predstavljene u formatu ISO 3155–3:2013.
- *market_segment* - Označavanje tržišnog segmenta. U kategorijama, izraz „TA" znači „putnički agenti", a „TO" znači „turoperateri".
- *distribution_channel* - Kanal distribucije rezervacije. Izraz „TA" znači „putnički agenti", a „TO" znači „turoperatori".
- *is_repeated_guest* - Vrijednost koja pokazuje je li ime rezervacije bilo od ponovljenog gosta (1) ili ne (0).
- *previous_cancellation* - Broj prethodnih rezervacija koje je kupac otkazao prije trenutne rezervacije.
- *previous_bookings* - Broj prethodnih rezervacija koje kupac nije otkazao prije trenutne rezervacije.
- *reserved_room_type* - Šifra vrste rezervirane sobe. Šifra je prikazana umjesto oznake iz razloga anonimnosti.

- *assigned_room_type* - Šifra za vrstu sobe dodijeljene rezervaciji. Ponekad se dodijeljena vrsta sobe razlikuje od rezervirane vrste sobe.
- *booking_changes* - Broj promjena/dopuna napravljenih na rezervaciji od trenutka kada je rezervacija upisana na PMS.
- *deposit_type* - Naznaka je li gost uplatio depozit za jamstvo rezervacije. Ova varijabla može pretpostaviti tri kategorije.
- *agent* - ID turističke agencije koja je izvršila rezervaciju.
- *company* - ID tvrtke/subjekta koji je izvršio rezervaciju ili je odgovoran za plaćanje rezervacije.
- *days_in_waiting_list* - Broj dana koliko je rezervacija bila na listi čekanja prije nego što je potvrđena kupcu.
- *customer_type* - Vrsta rezervacije.
- *adr* - Prosječna dnevna stopa definirana dijeljenjem zbroja svih transakcija smještaja s ukupnim brojem noćenja.
- *required_car_parking_spaces* - Broj parkirnih mjesta za automobile prema zahtjevu kupca.
- *total_of_special_requests* - Broj posebnih zahtjeva kupca (npr. bračni krevet ili visoki kat).
- *reservation_status* - Zadnji status rezervacije.
- *reservation_status_date* - Datum kada je zadnji status postavljen.

Prvi čvor će naravno biti **CSV Reader** kao i u većini slučajeva do sada. Na postavkama čvora i to na trećoj kartici postavki (*Advanced settings*) potrebno je postaviti pregled većeg broja redova prije nego što KNIME predloži kakvi se podaci nalaze u pojedinom stupcu. Potrebno je navesti 50000 da ne bi došlo do grešaka zbog neispravnog zapisa u ćeliji.

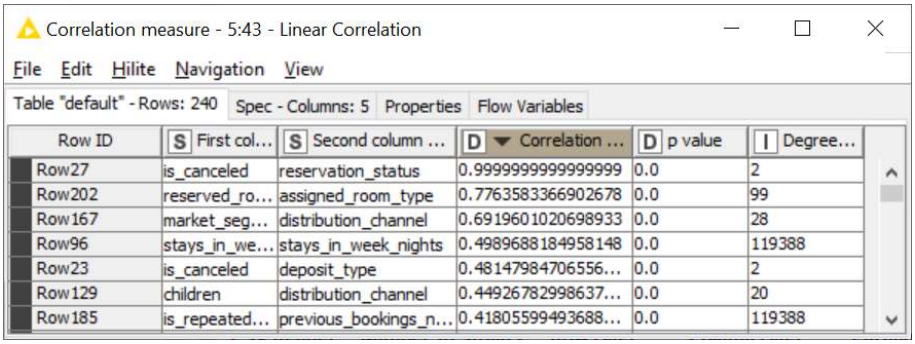
Kako je zadatak istrenirati klasifikator koji će moći odrediti hoće li rezervacija biti otkazana ili neće, treba pogledati koji stupac je ciljna varijabla. Radi se o stupcu *is_canceled* koji govori je li rezervacija otkazana i to tako da je za otkazane rezervacije vrijednost 1, a za ostale vrijednost 0. S obzirom da je KNIME taj stupac pretvorio u brojčanu vrijednost, potrebno ju je konvertirati u tekstualnu, odnosno kategorijalnu varijablu. Za to se koristi čvor **Number to String** koji je obrađen u prethodnom tekstu. Taj čvor koristi se samo za konverziju varijable *is_canceled* iz numeričkog u tekstualni oblik.



Slika 291. Postavke za konverziju stupca is_canceled

Nakon tog čvora s obzirom da ima puno značajki, preporuka je da se testira ima li stupaca koji imaju jaku korelaciju sa stupcem *is_canceled*. U tom slučaju dobio bi se model sa skoro 100 % točnosti, ali

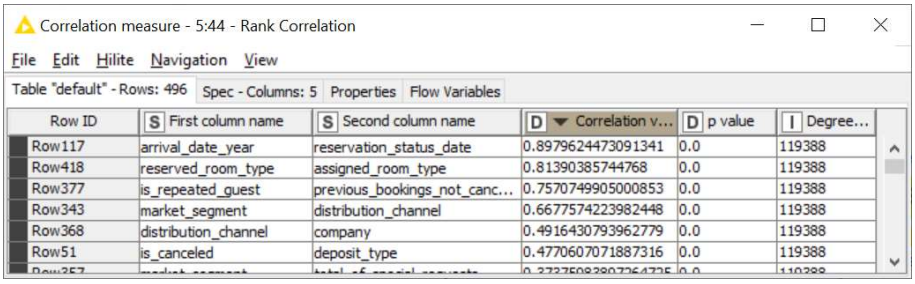
kad bi se pokušalo unijeti podatke o nekoj rezervaciji s ciljem klasifikacije vidjelo bi se da podatak s kojim je velika povezanost, vjerojatno nedostaje. Iza prethodnog čvora konverzije ubačena su dva čvora i to **Linear Correlation** i **Rank Correlation**. Najbolje je nakon pokretanja čvorova pogledati tablice koje se dobivaju iz kontekstnog izbornika čvora izborom *Correlation Measure*. Tablice dobivene iz oba čvora treba sortirati po vrijednosti stupca *Correlation value*. Slika 292 prikazuje korelaciju od približno 1 između stupaca *is_canceled* i *reservation_status*. Stupac *is_canceled* mora se ostaviti jer je to ciljna varijabla, ali stupac *reservation_status* mora se ukloniti u sljedećem čvoru.



Row ID	First col...	Second column ...	Correlation ...	p value	Degree...
Row27	is_canceled	reservation_status	0.9999999999999999	0.0	2
Row202	reserved_ro...	assigned_room_type	0.7763583366902678	0.0	99
Row167	market_seg...	distribution_channel	0.6919601020698933	0.0	28
Row96	stays_in_we...	stays_in_week_nights	0.4989688184958148	0.0	119388
Row23	is_canceled	deposit_type	0.48147984706556...	0.0	2
Row129	children	distribution_channel	0.44926782998637...	0.0	20
Row185	is_repeated...	previous_bookings_n...	0.41805599493688...	0.0	119388

Slika 292. Povezanost među varijablama

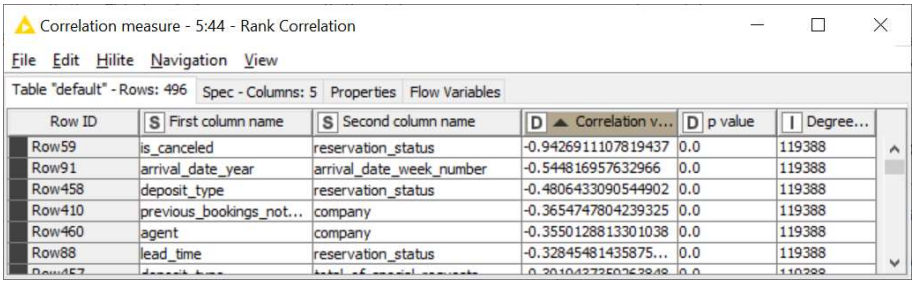
Slika 293 prikazuje sortirane vrijednosti povezanosti pri čemu se vidi da postoji prilično jaka povezanost između varijable *arrival_date_year* i *reservation_status_date*. S obzirom da stupac *reservation_status_date* govori kada je zadnji puta status rezervacije promijenjen, očigledno je da se taj stupac mora ukloniti u slijedećem čvoru. Razlog je što ta informacija nije dostupna za nove rezervacije za koje se želi raditi predikcije otkazivanja.



Row ID	First column name	Second column name	Correlation v...	p value	Degree...
Row117	arrival_date_year	reservation_status_date	0.8976924473091341	0.0	119388
Row418	reserved_room_type	assigned_room_type	0.81390385744768	0.0	119388
Row377	is_repeated_guest	previous_bookings_not_canc...	0.7570749905000853	0.0	119388
Row343	market_segment	distribution_channel	0.6677574223982448	0.0	119388
Row368	distribution_channel	company	0.4916430793962779	0.0	119388
Row51	is_canceled	deposit_type	0.4770607071887316	0.0	119388

Slika 293 – Povezanosti među varijablama – pozitivna

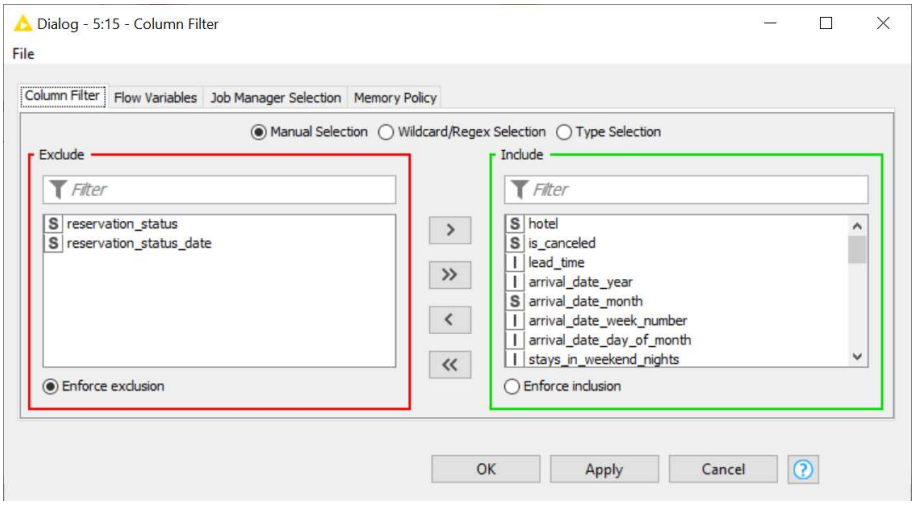
Slika 294 prikazuje sortirane negativne vrijednosti povezanosti pri čemu se vidi kako postoji prilično jaka povezanost između varijable *is_canceled* i *reservation_status.* Već je uočena neprihvatljiva povezanost s varijablom *reservation_status* u čvoru **Linear Correlation** tako da će i ta varijabla biti uklonjena u sljedećem čvoru.



Row ID	First column name	Second column name	Correlation v...	p value	Degree...
Row59	is_canceled	reservation_status	-0.9426911107819437	0.0	119388
Row91	arrival_date_year	arrival_date_week_number	-0.544816957632966	0.0	119388
Row458	deposit_type	reservation_status	-0.4806433090544902	0.0	119388
Row410	previous_bookings_not...	company	-0.3654747804239325	0.0	119388
Row460	agent	company	-0.3550128813301038	0.0	119388
Row88	lead_time	reservation_status	-0.32845481435875...	0.0	119388

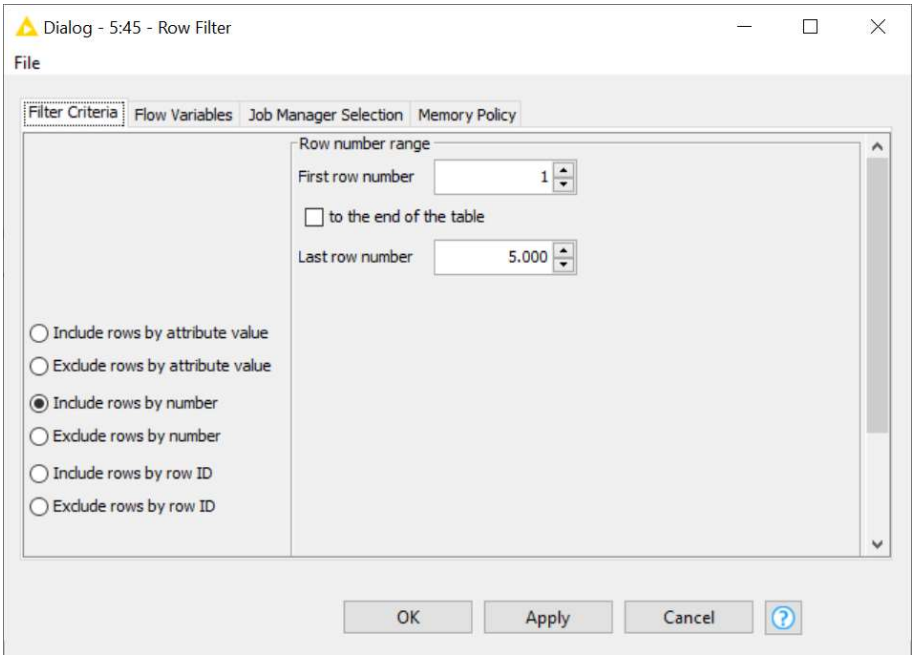
Slika 294. Povezanost među varijablama – negativna

Navedene dvije varijable se uklanjaju čvorom **Column Filter**. Slika 295 prikazuje postavke za uklanjanje navedenih varijabli.



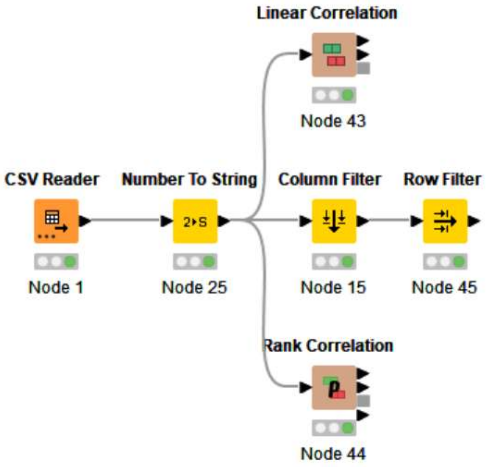
Slika 295. Postavke čvora Column Filter za uklanjanje varijabli

Nakon uklanjanja nepotrebnih stupaca umetne se čvor **Row Filter** koji će omogućiti treniranje i testiranje modela na dijelu skupa podataka. S obzirom da je treniranje SVM tehnike u programu KNIME zna potrajati, što je navedeno kao nedostatak tehnike, ograničavanjem skupa podataka skratit će se trajanje treniranja. Činjenica je da se tim postupkom u pravilu smanjuje točnost modela, ali za potrebe analize tehnike može se koristiti i dio skupa podataka. Za produkcijske modele koristit će se kompletan skup podataka.



Slika 296. Filtriranje samo prvih 5000 redova

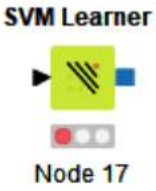
Slika 297 prikazuje prvi dio hodograma za učitavanje i pripremu podataka. Sljedeći korak je izrada modela.



Slika 297. Dio hodograma za učitavanje i pripremu podataka

11.2. Izrada modela temeljenog na metodi potpornih vektora

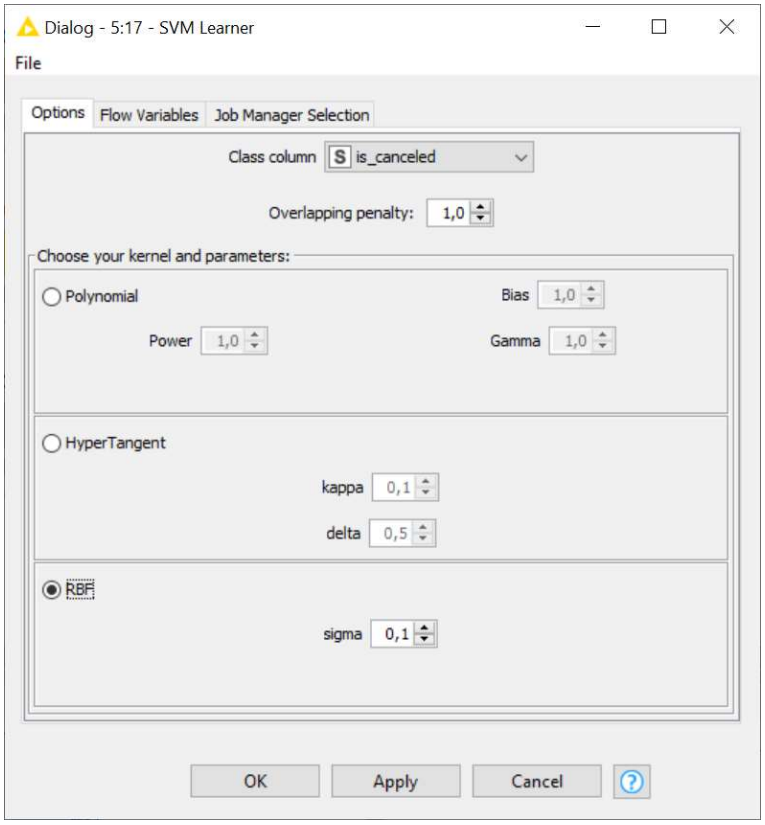
Model temeljen na metodi potpornih vektora očekivano se sastoji od dva ključna čvora. Slika 298 prikazuje čvor **SVM Learner** ili SVM učenik koji služi za treniranje modela.



Slika 298. Čvor SVM Learner

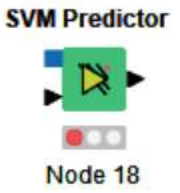
Slika 299 prikazuje postavke čvora **SVM Learner**. Prvi padajući izbornik služi za izbor kategorijalne varijable za čiju klasifikaciju se trenira model. Ta varijabla ne može biti numerička, odnosno brojčana. Ispod toga se može definirati vrijednost za postavku *Overlapping Penalty*. Vrijednost te postavke se koristi kada ulazne podatke nije moguće razdvojiti u kategorije pa se navedena vrijednost koristi kao kazna za netočnu klasifikaciju. Predložena vrijednost za to je 1.

Najveću površinu dijaloškog okvira postavki zauzima izbor vrste treniranja. Dostupne su tri vrste (*HyperTangent, Polynomial and RBF*) i svaka od njih ima drugačije parametre. Neki autori preporučuju RBF kao početnu funkciju, osim ako se model trenira s velikim brojem značajki. U nastavku će se koristiti RBF.



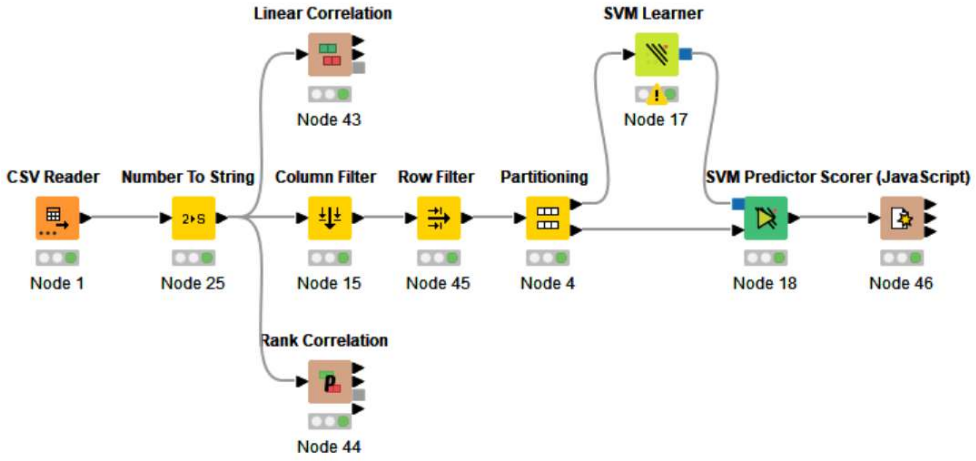
Slika 299. Postavke čvora SVM Learner

Slika 300 prikazuje čvor **SVM Predictor** ili SVM prediktor. S obzirom da čvor nema neke bitne karakteristike u postavkama, taj dijaloški okvir nije prikazan.



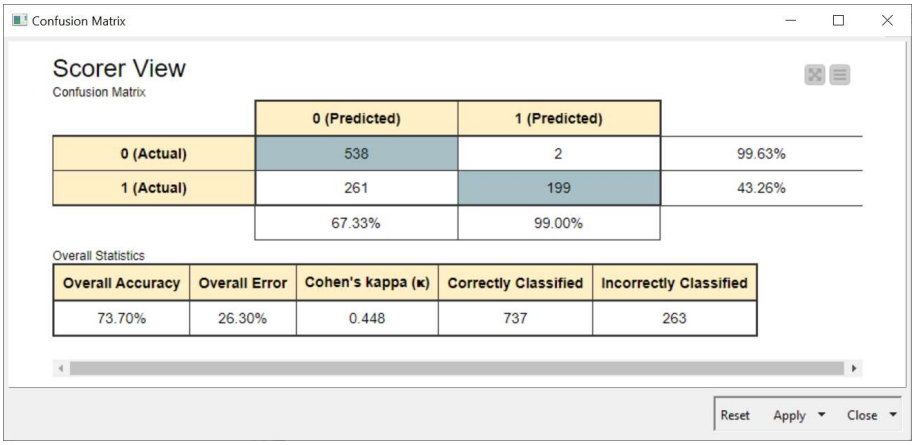
Slika 300. Čvor SVM Predictor

Slika 301 prikazuje hodogram koji učitava i priprema podatke, a spreman je i za treniranje modela te ispis matrice konfuzije koristeći čvor **Scorer (JavaScript)**.



Slika 301. Hodogram koji uključuje sve osnovne čvorove modela potpornih vektora

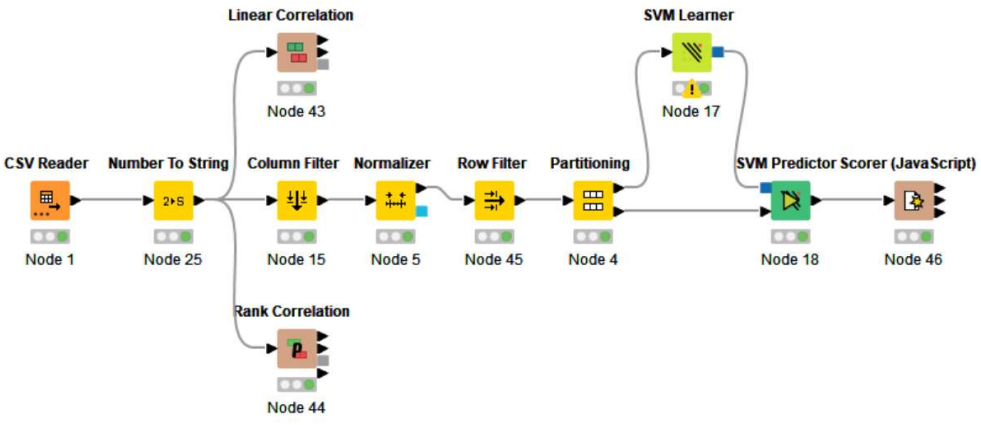
Slika 302 prikazuje ukupnu točnost i matricu konfuzije. Ukupna točnost je 73,70 %, a točnost predikcije prekida rezervacije iznosi samo 43,25 %.



Slika 302. Matrica konfuzije i ukupna točnost modela potpornih vektora

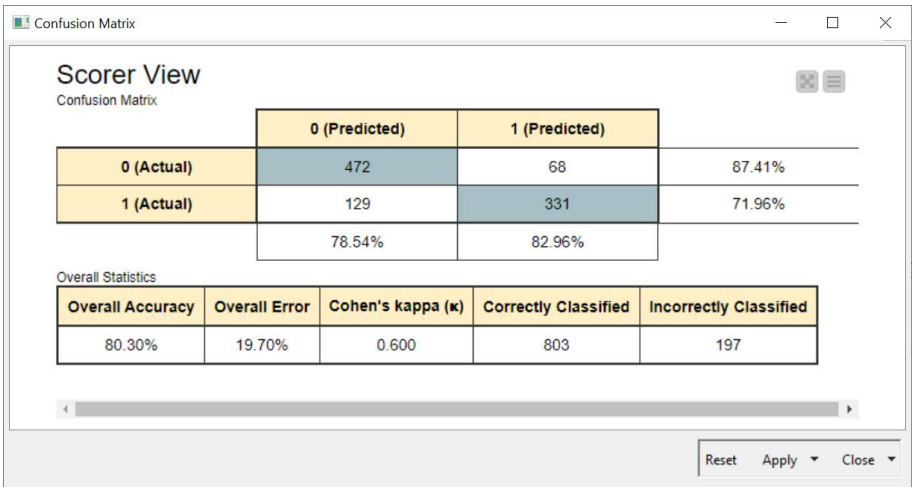
11.3. Optimizacija modela

Jedna od metoda koja se već koristila za optimizaciju u priručniku je normalizacija podataka koja će se primijeniti i u ovom primjeru. Slika 303 prikazuje hodogram s dodanim čvorom za normalizaciju.



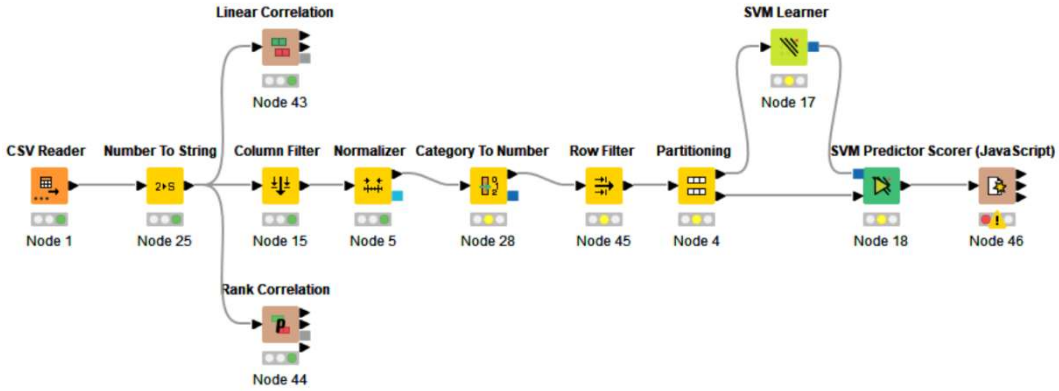
Slika 303. Hodogram s dodanim čvorom za normalizaciju

Nakon pokretanja čvorova dobivena ukupna točnost je znatno veća i iznosi preko 80 %. Radi se o značajnom skoku koji ukazuje na osjetljivost modela na normalizaciju podataka što je navedeno kao nedostatak tehnike u uvodnom opisu istog. Čvor za normalizaciju nije posebno konfiguriran prije pokretanja, tako da je ostala zadana *Min-max* normalizacija. Hodogram je testiran i nakon promjene vrste normalizacije u *Z-score normalization* i *Normalisation by decimal scaling*, ali rezultati su bili lošiji i to oko 73 % i 76 %.



Slika 304. Matrica konfuzije i ukupna točnost modela potpornih vektora nakon normalizacije

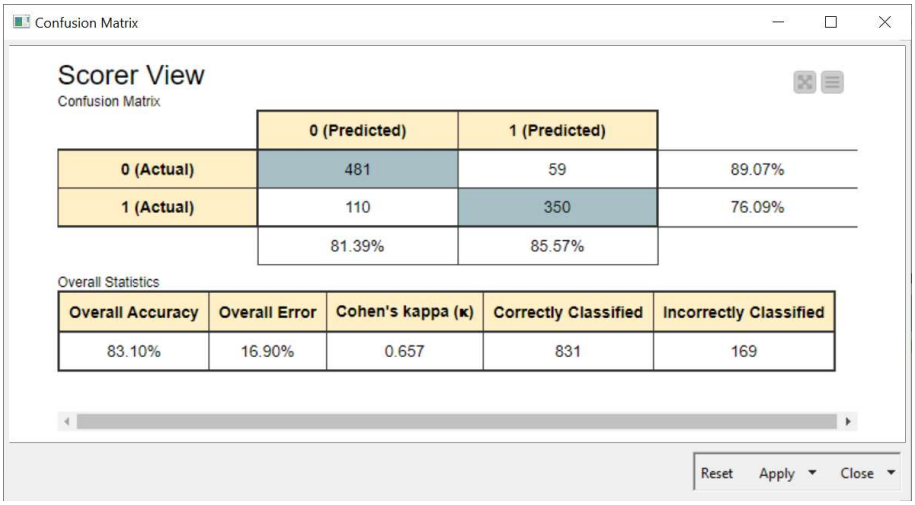
Sljedeći korak u povećanju točnosti modela je rezultat upozorenja koje je KNIME ispisao u prozoru konzole. Ono glasi: `Rejecting 13 column(s) due to incompatible type: [hotel, arrival_date_month, children, meal, ... (remainder truncated)].` Očigledno je čvor koji je zadužen za treniranje odbacio pojedine značajke zbog nekompatibilnosti. Radi se o kategorijalnim značajkama čiji sadržaj su tekstualni podaci. U odbačenim stupcima sigurno postoji korisnih informacija koje mogu poslužiti za povećanje točnosti modela. Kako bi se iskoristile odbačene značajke potrebno ih je pretvoriti u numeričke varijable. Za to će poslužiti čvor **Category to Number** koji je već opisan u prethodnom tekstu. Slika 305 prikazuje hodogram s dodanim čvorom.



Slika 305. Hodogram s čvorom za konverziju kategorijalnih vrijednosti u numeričke

Nakon pokretanja svih čvorova u hodogramu dobivaju se upozorenja o greškama jer neki stupci koje čvor pokušava konvertirati imaju više od zadanog maksimuma od 100 različitih tekstualnih vrijednosti.

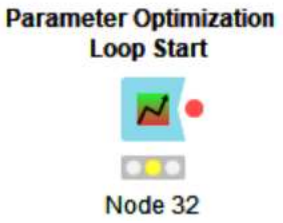
Da bi se otklonile greške jednostavno se te stupce uklanja s prethodnim čvorom **Column Filter**. Radi se o stupcima *country, agent* i *company*. Slika 306 prikazuje rezultate nakon normalizacije i transformacije kategorijalnih varijabli. Ukupna točnost prešla je 83 %.



Slika 306. Matrica konfuzije nakon normalizacije i transformacije kategorijalnih varijabli

Treća mogućnost optimizacije modela je izmjena parametara samog modela. U postavkama čvora **SVM Laerner** izabrana je funkcija RBF kod koje se može mijenjati parametar « sigma ». Zadano taj parametar ima vrijednost 0,1. Postoji mogućnost da ta vrijednost ne daje optimalan model, ali ručna izmjena tog parametra i višestruko treniranje s različitim vrijednostima je dosta nespretno rješenje za optimizaciju. Srećom KNIME ima rješenje za automatsko testiranje parametara modela.

Slika 307 prikazuje čvor **Parameter Optimization Loop Start** ili Početak petlje optimizacije parametara. Radi se o čvoru koji se u hodogramu primjenjuje u paru s čvorom **Parameter Optimization Loop End** koji će biti opisan kasnije. Čvor **Parameter Optimization Loop Start** pokreće petlju optimizacije parametara.

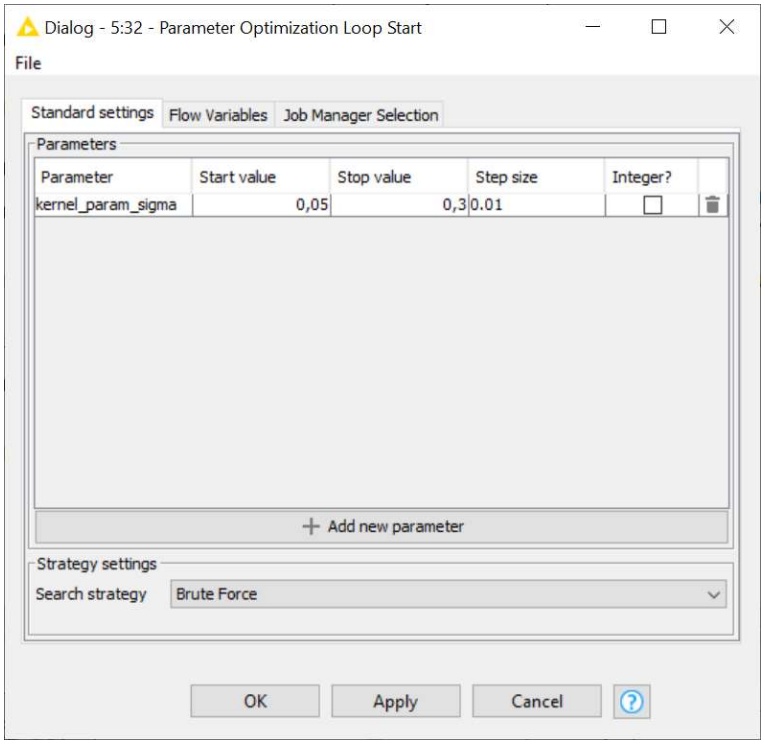


Slika 307. Čvor Parameter Optimization Loop Start

Slika 308 prikazuje postavke čvora u kojem se može definirati koji parametri se mijenjaju, u kojim granicama i s kolikim korakom. Dostupne su četiri strategije optimizacije i to:

a) *Brute Force* - provjeravaju se sve kombinacije parametara i vraća se najbolja.
b) *Hillclimbing* - generira se slučajna početna kombinacija i procjenjuju se izravni susjedi. Najbolja kombinacija među susjedima je početna točka za sljedeću iteraciju. Ako niti jedan susjed ne poboljša funkciju cilja, petlja se prekida.
c) *Random Search* - kombinacije parametara se nasumično biraju i procjenjuju. Navedene početne i zaustavne vrijednosti definiraju prostor parametara iz kojeg se nasumično izvlači kombinacija parametara. Petlja završava nakon određenog broja ponavljanja ili, ako je aktivirano rano zaustavljanje, kada se za određeni broj pokušaja vrijednost nije poboljšala.

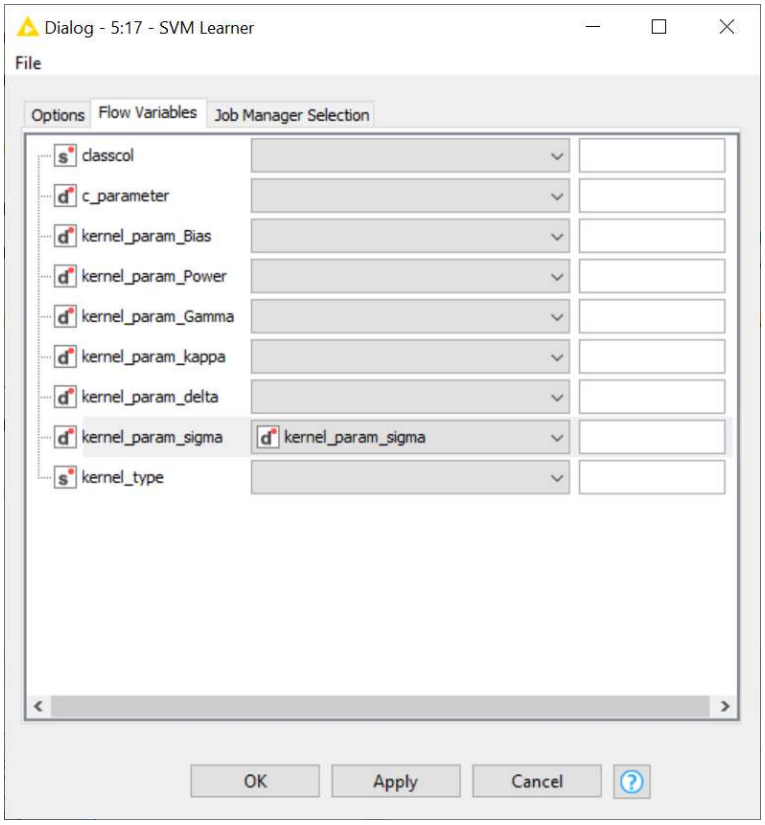
d) *Bayesian Optimization (TPE)* - Ova strategija sastoji se od dvije faze. Prva je zagrijavanje u kojem se kombinacije parametara nasumično biraju i ocjenjuju. Na temelju rezultata zagrijavanja, druga faza pokušava pronaći najbolje kombinacije parametara.



Slika 308. Postavke čvora Parameter Optimization Loop Start

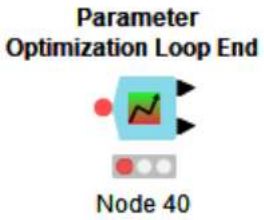
Slika 308 prikazuje kako je zadan parametar *kernel_param_sigma* te se njegova vrijednost mijenja od vrijednosti 0,05 do 0,30 s korakom od 0,01. Izabrana je strategija *Brute Force* što znači da će se sa svakom mogućom vrijednošću parametra trenirati model i na taj način doći do vrijednosti parametra. Ta strategija je ponekad vremenski zahtjevna ako se ima puno parametara jer se isprobavaju sve moguće kombinacije.

Kada se koristi optimizacija parametara modela, pri tom se u čvoru u kojem se trenira model ,treba postaviti da parametar koji se optimizira nije naveden u samom čvoru nego dolazi kao varijabla protoka ili protočna varijabla (eng. *Flow Variable*). Radi se o varijablama koje čvorovi mogu izmjenjivati, kako bi to bilo moguće, potrebno je učiniti vidljivim priključke za varijable protoka. To se radi tako da se desnim klikom dobije kontekstni izbornik čvora i zatim izabere *Show Flow Variable Ports*. Nakon toga se u gornjim kutovima čvora prikazuju crveni krugovi koji služe kao ulazni i izlazni priključci za varijable protoka. Tu radnju potrebno je izvršiti na svim čvorovima između kojih se želi razmjenjivati varijable protoka , konačno ih i povezati koristeći te priključke. Nakon povezivanja, kao što je navedeno, potrebno je u čvoru za treniranje definirati izvor varijable koja se optimizira. Slika 309 prikazuje drugu karticu postavki čvora **SVM Learner** gdje se postavlja izvor varijable koja se optimizira.



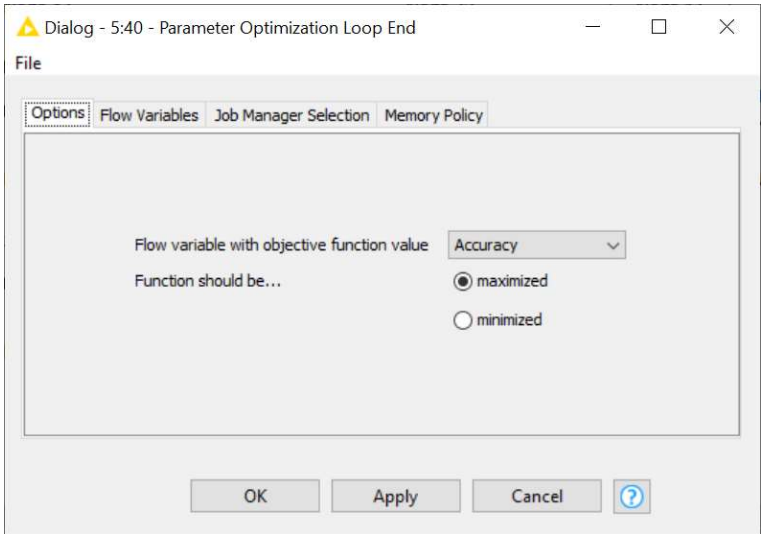
Slika 309. Postavke čvora SVM Learner gdje treba izmijeniti varijablu

Slika 310 prikazan je čvor **Parameter Optimization Loop End** ili Kraj petlje optimizacije parametara koji služi kao kraj petlje optimizacije i omogućuje očitavanje optimalne vrijednosti parametra.



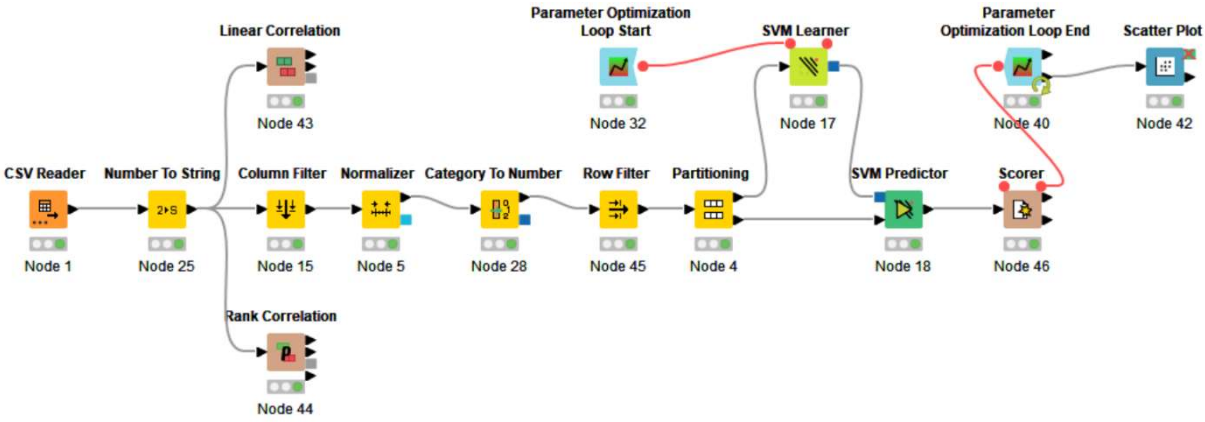
Slika 310. Čvor Parameter Optimization Loop End

Slika 311 prikazuje postavke čvora u kojem iz padajućeg izbornika treba izabrati koju vrijednost treba pratiti i pri tom zabilježiti vrijednost parametra (*sigma!*) koji se optimizira kad je vrijednost u padajućem izborniku bila najveća ili najmanja. Prati li se maksimum ili minimum, također se definira u postavkama čvora.



Slika 311. Postavke čvora Parameter Optimization Loop End

Slika 312 prikazuje konačan histogram modela u kojem su uključene i veze između priključaka čvorova kojima se izmjenjuju varijable protoka. Te veze se razlikuju zato jer su crvene boje.



Slika 312. Konačan hodogram modela

Na Slika 313 prikazana je tablica koja se dobiva na način da se iz kontekstnog izbornika čvora **Parameter Optimization Loop End** izabere *All parameters.* U drugom stupcu je vidljivo kako je čvor **SVM Learner** kreirao 26 modela koristeći vrijednosti parametra *sigma* od 0,05 do 0,30 u koracima od 0,01. Pri tom je svaki puta zabilježena vrijednost ukupne točnosti, a to je vidljivo u trećem stupcu tablice. Da bi se vidjelo koja vrijednost parametra *sigma* je optimalna za model, iz kontekstnog izbornika se bira *Best parameters*. Dobiva se da je optimalna vrijednost 0,09, što je vrlo blizu zadanoj vrijednosti 0,1. U ovom slučaju izmjenom parametra *sigma* iz 0,1 u 0,09 poboljšava se točnost modela za 0,1 % što nije značajno poboljšanje, mada bi ova metoda optimizacije u nekom drugom slučaju mogla biti puno korisnija.

All parameters - 5:40 - Parameter Opti...

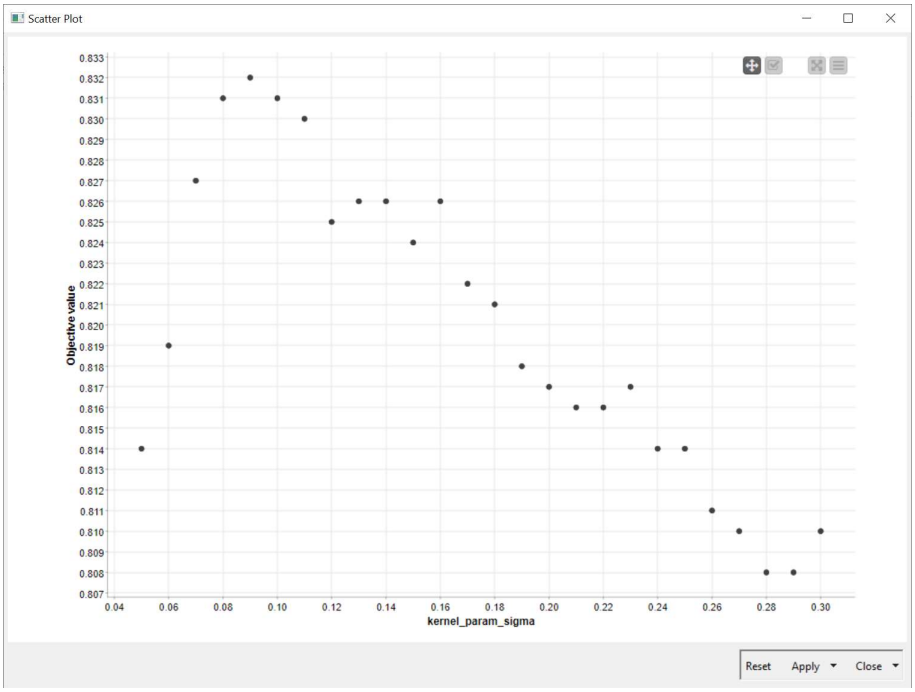
File Edit Hilite Navigation View

Table "default" - Rows: 26 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	kernel_param_sigma	Objective value
Row0	0.05	0.814
Row1	0.06	0.819
Row2	0.07	0.827
Row3	0.08	0.831
Row4	0.09	0.832
Row5	0.1	0.831
Row6	0.11	0.83
Row7	0.12	0.825
Row8	0.13	0.826
Row9	0.14	0.826
Row10	0.15	0.824
Row11	0.16	0.826
Row12	0.17	0.822
Row13	0.18	0.821
Row14	0.19	0.818
Row15	0.2	0.817
Row16	0.21	0.816
Row17	0.22	0.816
Row18	0.23	0.817
Row19	0.24	0.814
Row20	0.25	0.814
Row21	0.26	0.811
Row22	0.27	0.81
Row23	0.28	0.808
Row24	0.29	0.808
Row25	0.3	0.81

Slika 313. Tablica ovisnosti točnosti o vrijednosti parametra sigma

Slika 314 prikazuje graf ovisnosti točnosti SVM modela o vrijednosti parametra *sigma*. Taj dijagram raspršenosti dobiven je uz pomoć čvora **Scatter Plot** koji je postavljen na kraju hodograma.



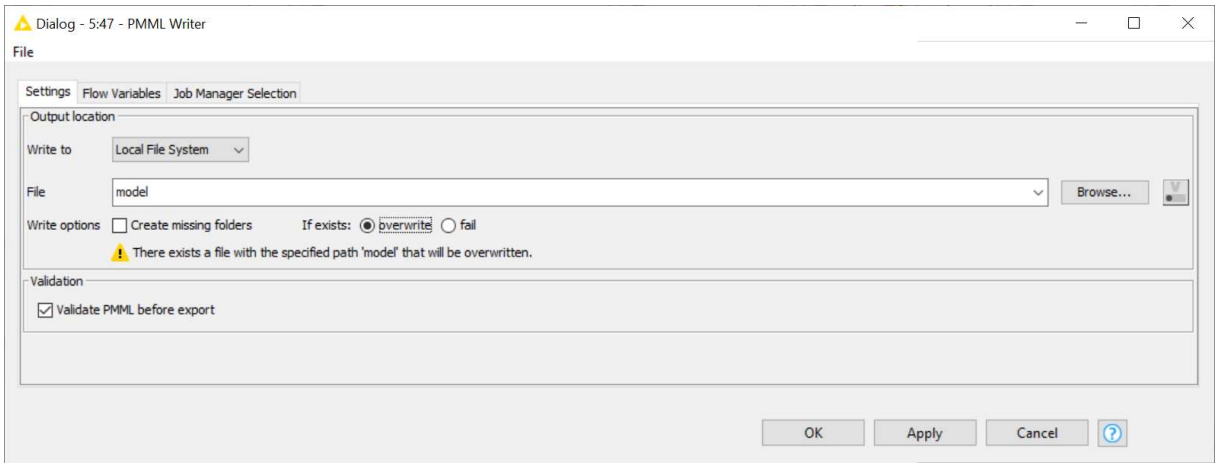
Slika 314. Graf ovisnosti točnosti SVM modela o vrijednosti parametra sigma

11.4. Spremanje modela

Nakon što je model optimiziran, isti se može spremiti za naknadno korištenje. Za to se koristi čvor **PMML Writer** ili PMML pisač.

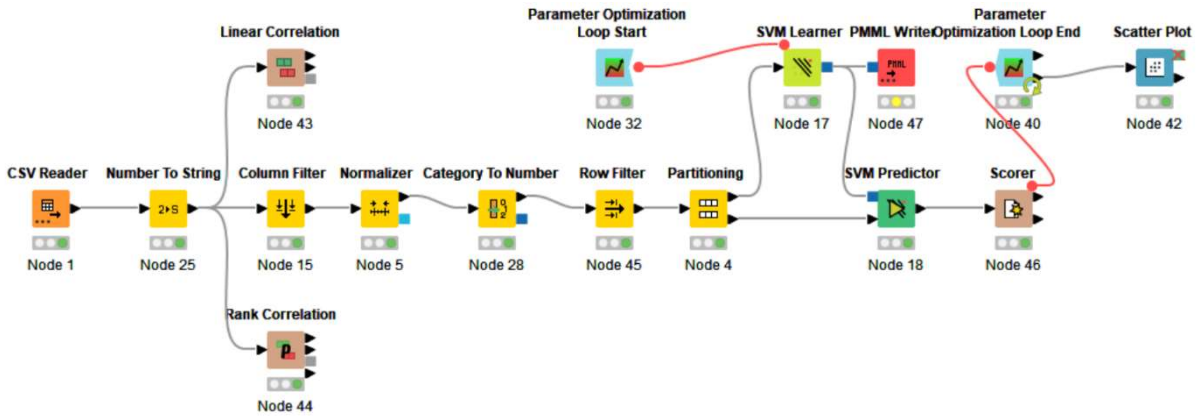
Slika 315. Čvor PMML Writer

Slika 316 prikazuje postavke čvora **PMML Writer** koji je vrlo sličan drugim čvorovima koji služe za zapisivanje datoteka. Potrebno je izabrati naziv datoteke, a model se sprema u standardiziranom PMML formatu.



Slika 316. Postavke čvora PMML Writer

Slika 317 prikazuje konačan hodogram koji uključuje i zadnji umetnuti čvor za spremanje modela. Taj čvor može se dodati na bilo koji čvor koji služi za treniranje modela, a u sljedećem poglavlju bit će opisan čvor koji služi za učitavanje modela iz PMML datoteke.



Slika 317. Konačan hodogram SVM modela

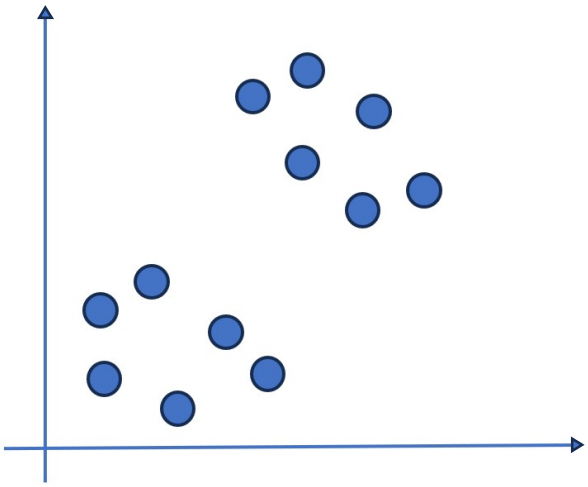
12. Tehnika k-srednjih vrijednosti

Tehnika k-srednjih vrijednosti pripada grupi tehnika nenadziranog učenja i služi za klasteriranje. Osnovni cilj klasteriranja je automatsko grupiranje podataka prema kriterijima sličnosti uz postizanje velike sličnosti između podataka unutar iste grupe, a male sličnosti između podataka koji pripadaju različitim grupama. Sve prethodno obrađene tehnike pripadale su skupini tehnika nadziranog učenja kod kojih se koristi niz značajki i ciljna varijabla za izradu modela. Tehnikama nenadziranog učenja dostupne su samo značajke i oni na osnovu njih kreiraju grupe ili klastere, na osnovu sličnosti dostupnih slučajeva. Bilo bi jednostavnije kada bi bila dostupna ciljna varijabla, ali u nekim situacijama je preskupo kategorizirati podatke ili uopće nije poznato koje kategorije postoje (Géron, 2022).

Sam naziv tehnike je prvi put spomenut u radu 1967. mada je sama tehnika predložena i korištena u tvrtki *Bell Labs* još pedesetih godina prošlog stoljeća (MacQueen & others, 1967). Smatra se da ju je osmislio Lloyd te se ponekad naziva i *Lloyd-ova* tehnika (Lloyd, 1982).

Prije pokretanja treniranja tehnike je nužno definirati na koliko grupa se želi podijeliti podatke. Svaka grupa ima svoju srednju vrijednost koja se naziva centroid. Svaki slučaj u skupu podataka pripada jednoj grupi i to onoj čiji centroid mu je najbliži. Kada se pokrene, prije prve iteracije se slučajnim izborom postavlja zadani broj centroida. Svaki slučaj se pridružuje najbližem centroidu, a nakon toga se centroid postavlja u centar slučajeva koji mu pripadaju. Pri tom dolazi do pomaka centroida i ponovnog dodjeljivanja slučajeva najbližem centroidu. Navedena dva koraka se ponavljaju (pridruživanje centroidu i računanje nove lokacije) sve dok više nema promjene u centroidima grupa. U nastavku će se s 12 slučajeva i dvije grupe prikazati kako funkcionira tehnika (Géron, 2022).

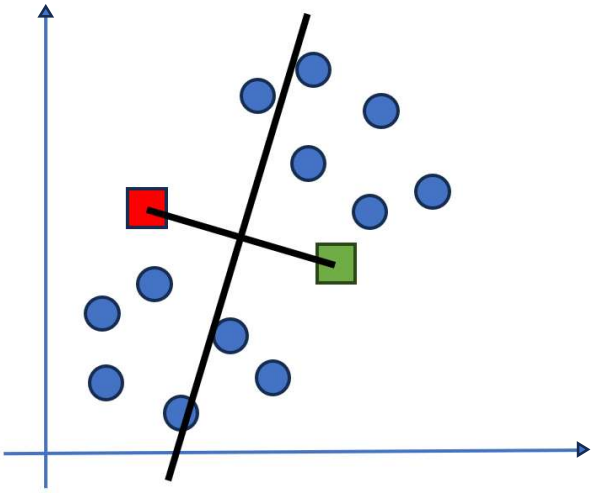
Primjer na kojem će se prikazati klasteriranje uz pomoć tehnike k-srednjih vrijednosti, ali bez korištenja matematičkih izraza, može biti klasteriranje turista po primanjima i potrošnji. Nekako je za očekivati da će turisti s većim primanjima trošiti više novca i obratno. Slika 318 prikazuje 12 turista, na osi X su njihova primanja dok je na osi Y prikazana njihova potrošnja na lokaciji koju posjećuju. Očigledno je kako postoje dvije grupe turista, ali u nastavku je objašnjeno kako će te dvije grupe otkriti tehnika.



Slika 318. Grafički prikazana potrošnja i prihodi 12 turista

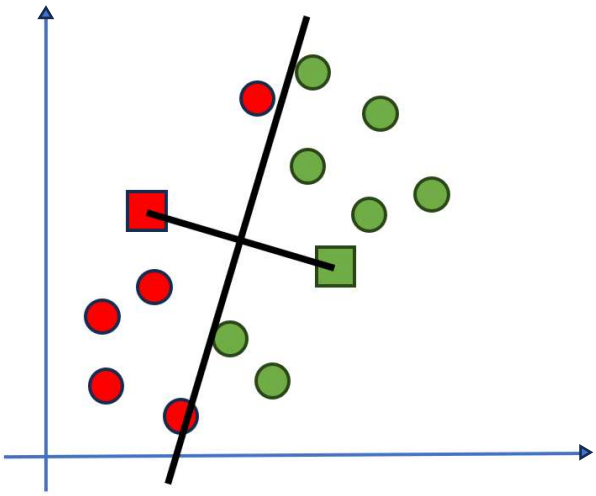
Tehnika k-srednjih vrijednosti kao što je navedeno definira centroide slučajnim izborom u dvodimenzionalnom prostoru jer se radi sa samo dvije značajke. Jedan centroid je crveni kvadrat, a drugi je zeleni kvadrat. Nakon toga se može zamisliti kako se povlači dužina između centroida te da

okomiti pravac na sredini te dužine dijeli slučajeve u grupe. Slika 319 prikazuje postavljanje centroida i pravac koji dijeli slučajeve.



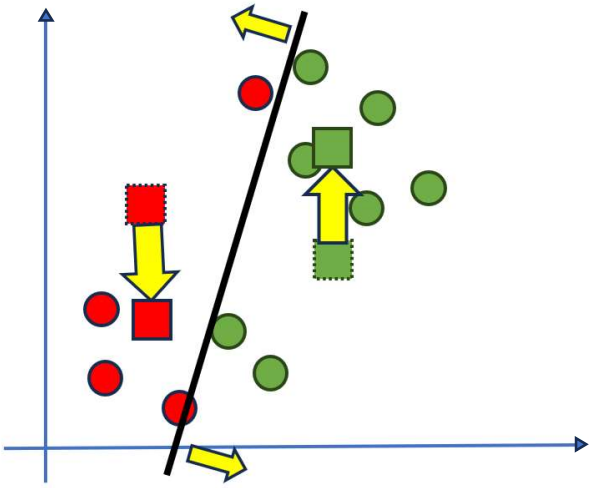
Slika 319. Slučajno postavljeni centroidi

Slika 320 prikazuje slučajeve koji su podijeljeni u grupe.



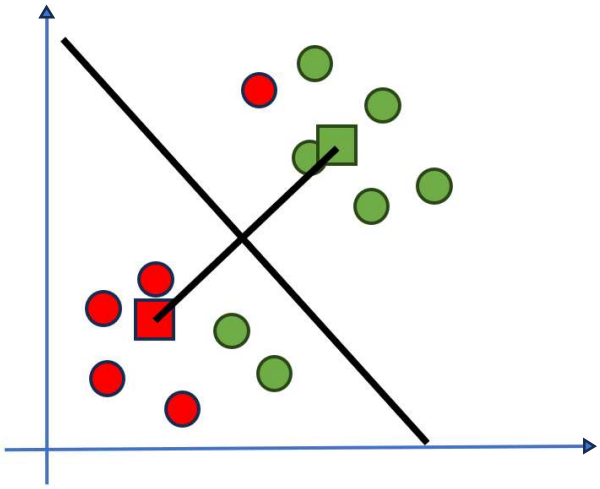
Slika 320. Slučajevi podijeljeni centroidima

Slika 321 prikazuje postupak računanja novih lokacija centroida s obzirom na slučajeve koji su pripali pojedinom centroidu. Treba uočiti kako je nova lokacija crvenog centroida niže od stare jer je većina crvenih slučajeva u donjem dijelu grafičkog prikaza. Nova lokacija zelenog centroida se pomiče gore prema većini zelenih slučajeva. Postavljanjem centroida na nove lokacije ponovno se povlači zamišljena dužina među centroidima te se definira pravac koji razdvaja dvije grupe.



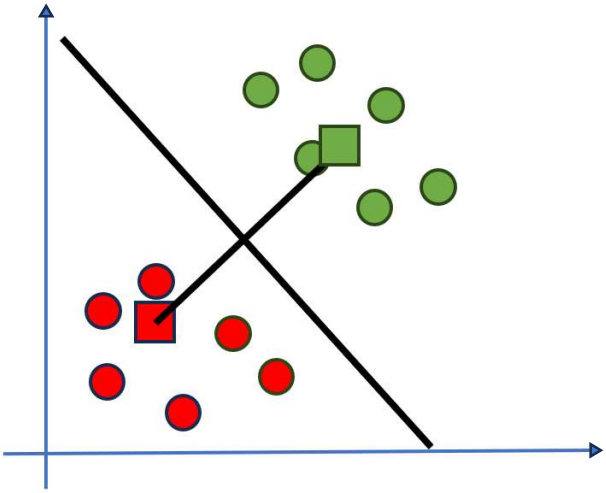
Slika 321. Postupak pomicanja centroida na nove lokacije

Slika 322 prikazuje novonastalu situaciju i novi pravac koji dijeli slučajeve u grupe.



Slika 322. Postavljanje pravca

Slika 323 prikazuje novu podjelu slučajeva s obzirom da se pripadnost nekih slučajeva promijenila. Nakon ove faze očekuju se još jedan ili dva koraka i tehnika bi završila jer ne bi više bilo promjene.



Slika 323. Ponovljena podjela slučajeva

Na ovom jednostavnom primjeru sa samo 12 slučajeva i dvije značajke čije vrijednosti nisu bile zadane kvantitativno je objašnjeno kako radi tehnika. Za razliku od ovog primjera u kojem je prikazano samo jedno pomicanje centroida, u stvarnim situacijama dogodi se i na desetke i stotine iteracija što se zadaje u čvoru za treniranje modela.

U slučaju kad se ima četiri značajke, više nije moguće taj proces ni prikazati grafički jer se sve odvija u prostoru od četiri dimenzije koji se ne može zamisliti. Ipak, bez obzira koliko značajki ima tehnika nalazi centroide za zadani broj grupa. Nedostatak tehnike je što uvijek dolazi do rješenja koje je stabilno, ali ne mora biti optimalno. Drugi nedostatak je što samo može grupirati slučajeve koji su odvojivi hiperravninama. Prednost je jednostavnost modela koji je razumljiv praktički svakom.

12.1. Izrada modela baziranog na tehnici k-srednjih vrijednosti

Jedna od čestih zadaća strojnog učenja je podjela korisnika u klastere, odnosno grupe. Jasno je da su tvrtkama neki korisnici važniji od drugih, a bitno je prepoznati koje vrste korisnika tvrtka ima i oko kojih se treba više potruditi da bi ih se zadržalo. Jasno je kako taj problem ne postoji kad tvrtka ima nekoliko korisnika, ali kad su u pitanju tisuće ili milijuni korisnika, tada jedino strojno učenje i moćna računala mogu pomoći u podjeli korisnika u klastere.

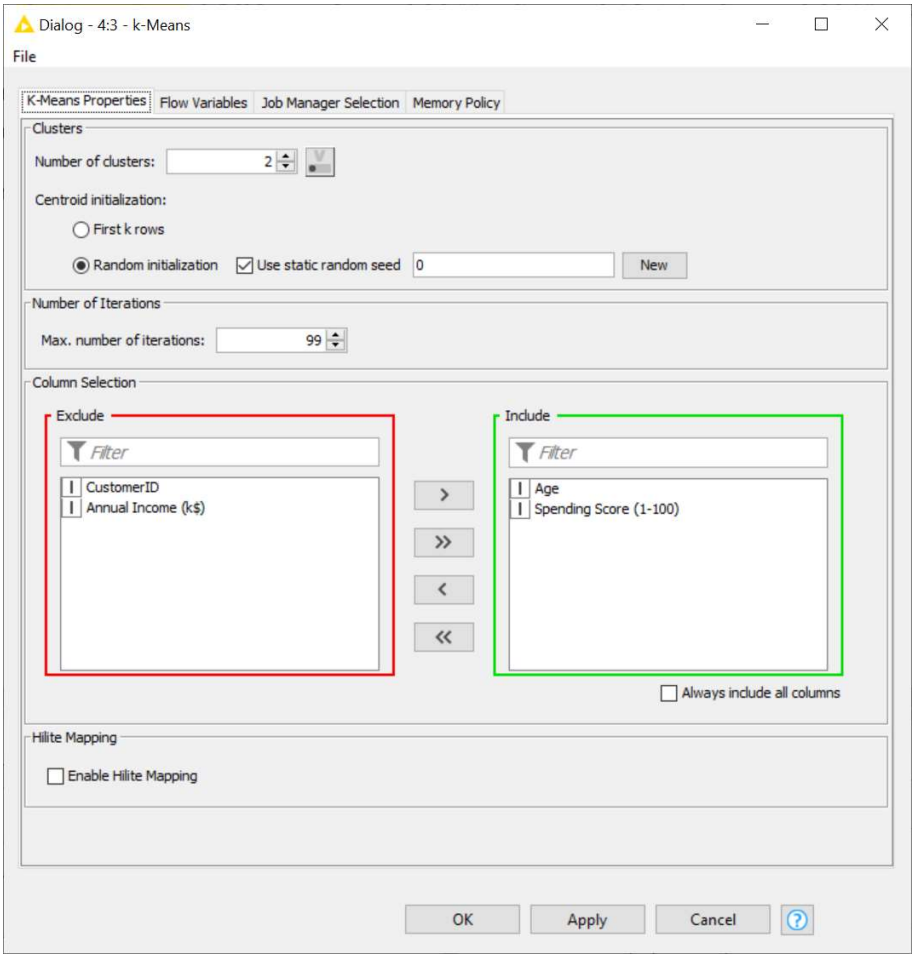
U primjeru za tehniku k-srednjih vrijednosti anketirani posjetitelji trgovačkog centra će se dijeliti u dva klastera. Podaci su dostupni na adresi: https://www.kaggle.com/datasets/kandij/mall-customers . Radi se o relativno malom skupu podataka od sto slučajeva, odnosno posjetitelja. Značajke koje su navedene u stupcima, kojih ima pet, su sljedeće:

a) *CustomerID* – identifikator posjetitelja
b) *Genre* – spol posjetitelja
c) *Age* – starost posjetitelja
d) *Annual Income* – godišnji prihodi u tisućama dolara
e) *Spending Score* – pokazatelj potrošnje.

Učitavanje podataka rješava čvor **CSV Reader**, nakon što se s prethodno navedene adrese preuzme skup podataka i dekomprimira ga se na mjesto gdje je dostupan programu KNIME. Nakon učitavanja podaci su dostupni i izlazni priključak čvora **CSV Reader** povezuje se s čvorom **k-Means**. Slika 324 prikazuje čvor.

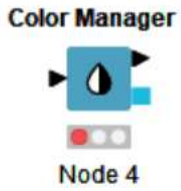
Slika 324. Čvor k-Means

Slika 325 prikazuje postavke čvora **k-Means** ili tehnika k-srednjih vrijednosti. U postavkama se zadaje broj grupa (*Number od clusters*) i način inicijalizacije centroida. Zadan je slučajni izbor lokacije, a mogu biti i lokacije prvih nekoliko slučajeva ovisno koliko je zadano grupa. Osim toga definira se broj iteracija koji je zadano 99, a u zeleni pravokutnik uključuju se značajke na osnovu kojih se generiraju grupe. U primjeru su izabrani starost (*Age*) i pokazatelj potrošnje (*Spending Score*) koji je numerička vrijednost između 0 i 100.



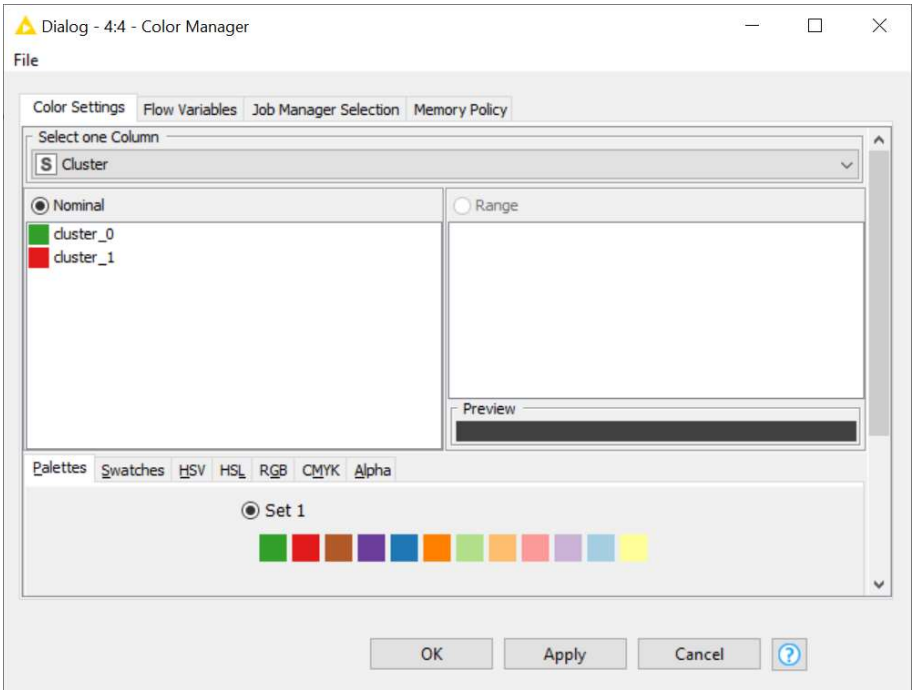
Slika 325. Postavke čvora k-Means

Slika 326 prikazuje čvor **Color Manager** ili Upravljač bojama koji slijedi iza čvora **k-Means**. Njegova funkcija je da za pojedine grupe koje pripadaju nekoj kategorijalnoj varijabli definira boje koje će se koristiti u grafičkim prikazima u nekom od sljedećih čvorova.



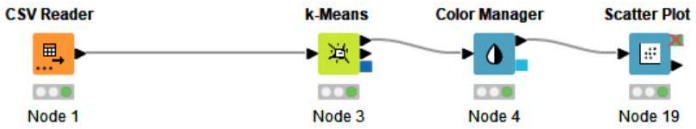
Slika 326. Čvor Color Manager

Slika 327 prikazuje postavke čvora **Color Manager** u kojem se bira kategorijalna varijabla za koju se definiraju boje te se za svaku jedinstvenu vrijednost definira boja. Na raspolaganju je više različitih paleta, a i više različitih načina zadavanja boje.



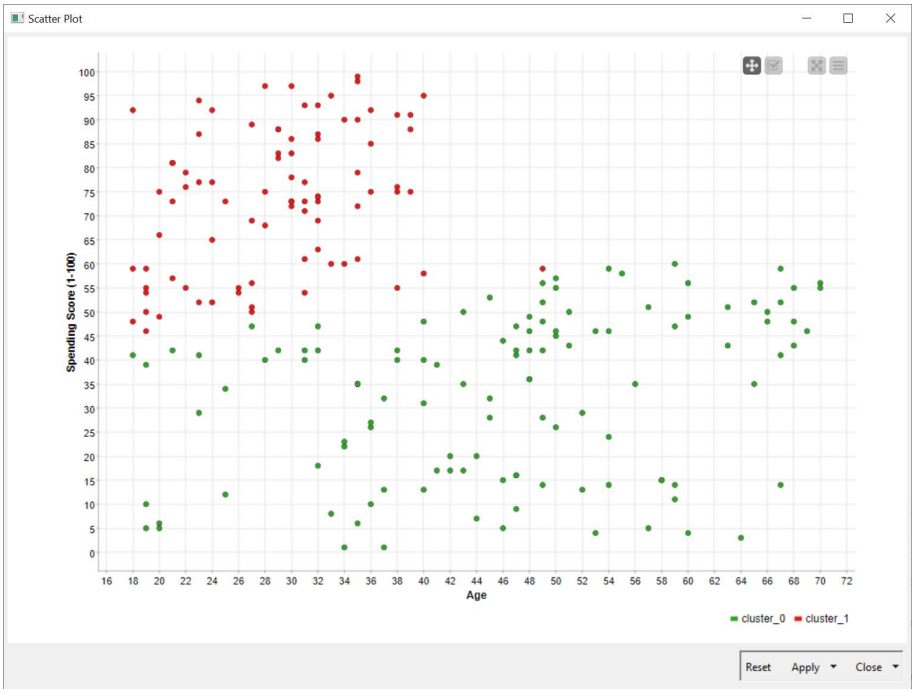
Slika 327. Postavke čvora Color Manager

Slika 328 prikazuje osnovni hodogram modela k-srednjih vrijednosti.



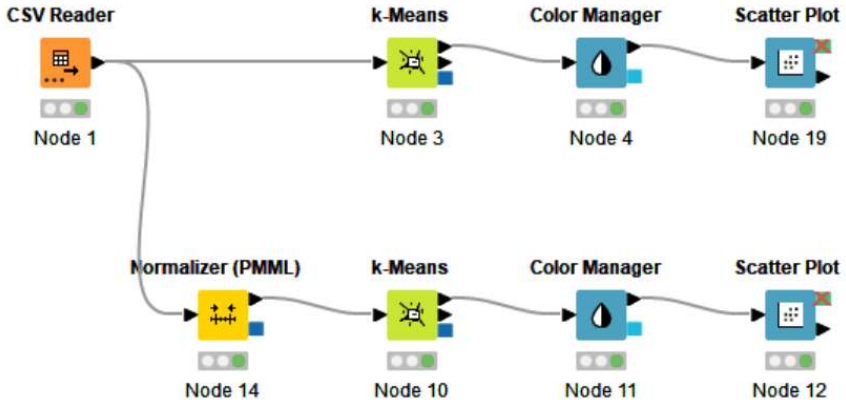
Slika 328. Hodogram osnovnog modela k-srednjih vrijednosti

Na slici 329 prikazane su grupe u različitim bojama na dijagramu raspršenosti. Podjela u grupe je lako uočljiva.



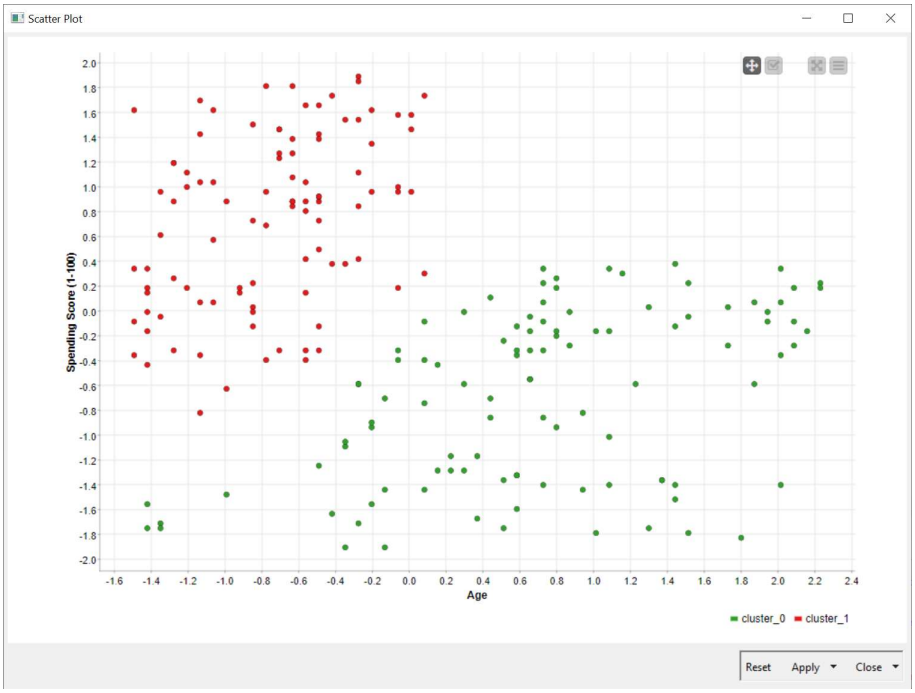
Slika 329. Grafički prikaz klastera bez normalizacije

Jedna od preporuka prije korištenja tehnike k-srednjih vrijednosti je da se podaci normaliziraju. Da bi se vidjelo razlikuju li se otkrivene grupe, na postojeći hodogram dodat će se još jedan niz čvorova uz prethodnu normalizaciju podataka. Izabrana je *Z-Score Normalization*. Svi korišteni čvorovi su već opisani prethodno, tako da ne bi trebao biti problem kreirati hodogram i postaviti potrebne opcije. Slika 330 prikazuje postojeći hodogram na koji je dodan još jedan niz čvorova uz prethodnu normalizaciju.



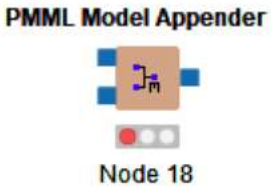
Slika 330. Hodogram modela k-srednjih vrijednosti uz normalizaciju

Slika 331 prikazuje novu podjelu u grupe, pri čemu se vidi da je zamišljeni pravac među grupama postavljen na drugom mjestu. Ako se pažljivo pogleda podjela na oba grafikona, može se uočiti da podjela nakon normalizacije izgleda za nijansu uvjerljivije.



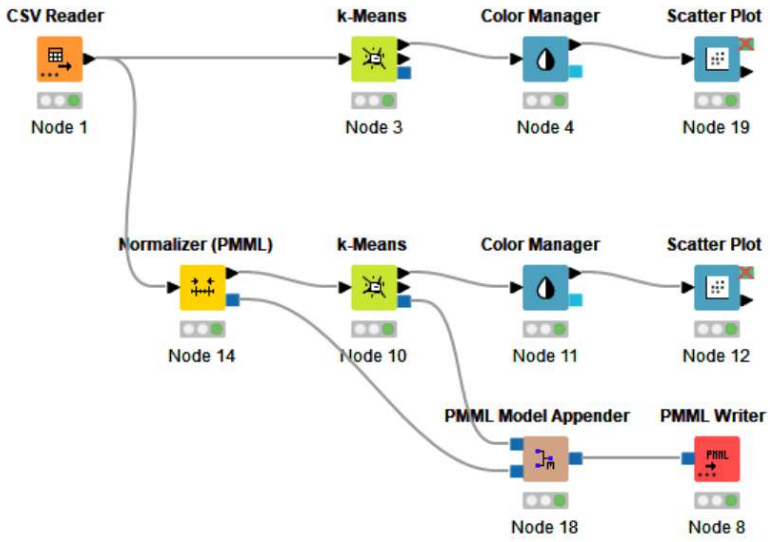
Slika 331. Grafički prikaz klastera s normalizacijom

Za kraj preostaje spremanje modela koji će biti korišten u drugom hodogramu. Ako se pogleda hodogram koji se koristio vidi se da postoje dva čvora koji među izlaznim priključcima imaju plavi kvadrat. To su čvorovi **Normalizer (PMML)** i **k-Means**. Model koji bi se pohranio bez prethodne normalizacije bio bi neupotrebljiv sa stvarnim podacima, tako da je nužno pohraniti i model koji je služio za normalizaciju podataka. Za to će se koristiti čvor **PMML Model Appender** ili Dodavanje PMML modelu. Slika 332 prikazuje taj čvor.



Slika 332. Čvor PMML Model Appender

Navedeni čvor nema značajnijih postavki, ali bitno je na gornji ulazni priključak povezati PMML treniranog modela dok se na donji povezuje PMML transformacija podataka.

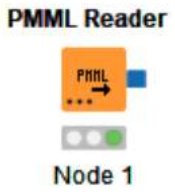


Slika 333. Hodogram kompletnog modela uz čvor za spremanje u PMML formatu

Slika 333 prikazuje kompletan hodogram. Nakon treniranja i spremanja može se prijeći i na učitavanje i primjenu modela.

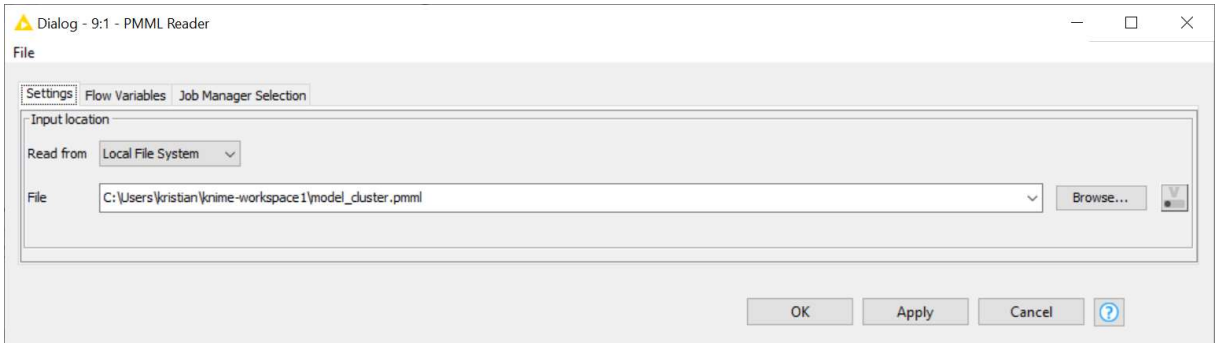
12.2. Korištenje spremljenog modela baziranog na tehnici k-srednjih vrijednosti

Konačno se dolazi do primjene modela, ali u posebnom hodogramu. U praksi hodogram može biti dosta kompleksan i nekome tko ga želi samo koristiti, velik broj čvorova može djelovati zbunjujuće. Iz tog razloga gotov model može se lako distribuirati i omogućiti njegovo jednostavno korištenje. U novom hodogramu prvi čvor će biti **PMML Reader** ili PMML čitač koji ima funkciju učitavanja modela. Slika 334 prikazuje njegov izgled.



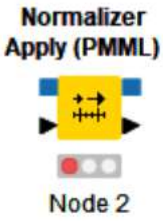
Slika 334. Čvor PMML Reader

Slika 335 prikazuje postavke čvora **PMML Reader** koje su jednostavne i omogućuju definiranje putanje do modela.



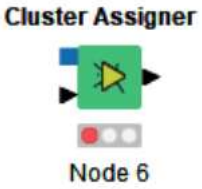
Slika 335. Postavke čvora PMML Reader

Nakon učitavanja PMML datoteke koja u sebi uključuje trenirani model k-srednjih vrijednosti i model normalizacije, potrebno je iste modele primijeniti, odnosno postaviti čvorove koji će ih primijeniti. Čvor koji primjenjuje normalizaciju pohranjenu u PMML datoteci naziva se **Normalizer Apply (PMML)** ili Primjena normalizacije (PMML). Slika 336 prikazuje taj čvor.



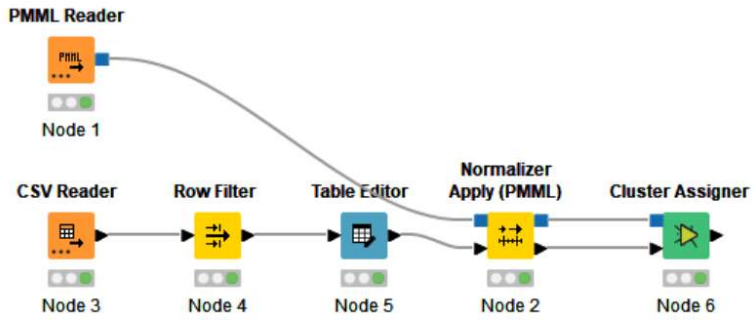
Slika 336. Čvor Normalizer Apply (PMML)

Čvor **Normalizer Apply (PMML)** nema značajnijih postavki, a iza njega slijedi čvor **Cluster Assigner** ili Dodjelitelj klastera koji služi za primjenu modela k-srednjih vrijednosti na novim slučajevima. Taj čvor prikazan je na sliciSlika 337.



Slika 337. Čvor Cluster Assigner

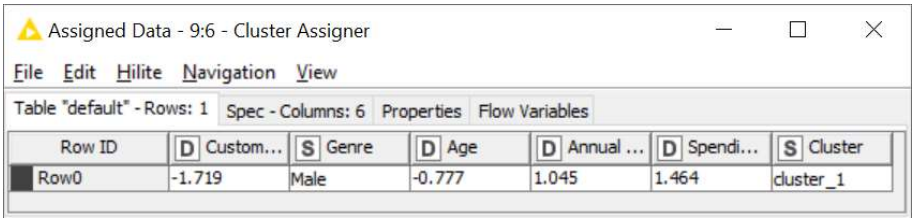
Slika 338 prikazan je konačan hodogram za primjenu modela na kojem se vidi na koji način treba povezati tokove podataka i informacije iz PMML datoteke u kojoj je definirana normalizacija i sam model k-srednjih vrijednosti. Tu su i čvorovi **CSV Reader**, **Row Filter** i **Table Editor** koji služe za učitavanje i mogućnost uređivanja jednog reda podataka.



Slika 338. Hodogram za učitavanje i primjenu modela k-srednjih vrijednosti

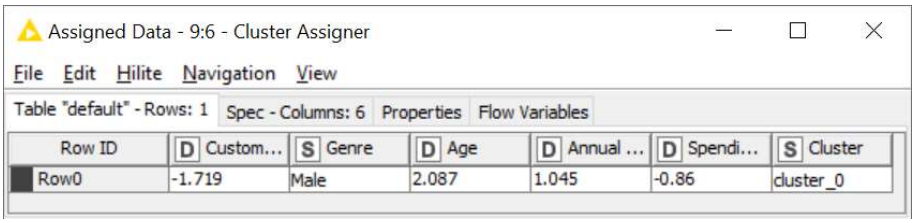
Konačno se dolazi i do testiranja modela. Slika 329 prikazuje kako je tehnika podijelila posjetitelje trgovačkog centra. Na osi X je starost posjetitelja, a na osi Y je pokazatelj potrošnje. U grupu *cluster_0* pripadaju osobe svih starosti čiji pokazatelj potrošnje je manji od 50. U grupu *cluster_1* pripadaju osobe koje su mlađe i pokazatelj potrošnje im je preko 50.

Slika 339 prikazuje rezultat odnosno pripadnost grupi *cluster_1* za osobu koja je stara 28 godina i pokazatelj potrošnje joj je 88. Navedeni podaci uneseni su u čvor **Table Editor**. Nakon normalizacije vidi se kako je starost od 28 godina manja 0,777 standardnih devijacija od prosječne vrijednosti populacije, a pokazatelj potrošnje je veći 1,464 standardnih devijacija od pokazatelja potrošnje populacije.



Slika 339. Rezultat za mlađu osobu s visokim pokazateljem potrošnje

Slika 340 prikazuje rezultat odnosno pripadnost grupi *cluster_0* za osobu koja je stara 68 godina i pokazatelj potrošnje joj je 28. Navedeni podaci su uneseni u čvor **Table Editor**. Nakon normalizacije se vidi kako je starost od 68 godina veća 2,087 standardnih devijacija od prosječne vrijednosti za populaciju, a pokazatelj potrošnje je manji 0,86 standardnih devijacija od prosječne vrijednosti pokazatelja potrošnje populacije.



Slika 340. Rezultat za stariju osobu s niskim pokazateljem potrošnje

Može se zaključiti da je model kategorizirao testne primjere u odgovarajuće grupe.

13. Zaključak

Samo dva desetljeća smo udaljeni od stogodišnjice uključenja prvog računala i na tom putu je čovječanstvo svjedočilo izrazito brzom razvoju računala. U početcima prije osamdeset godina računalo su mogli i znali programirati samo vrhunski znanstvenici, a danas se programiranje računala uči u osnovnoj školi. Prva računala zauzimala su stotine kvadratnih metara prostora, a danas u zadnjem džepu hlača nosimo pametne telefone koji su i tisuće puta moćniji od prvih računala. Slična priča se zbiva i sa strojnim učenjem i umjetnom inteligencijom. Područje koje je desetljećima bilo rezervirano samo za znanstvenike, odjednom je postalo dostupno i običnim ljudima. Za početak kroz primjenu tehnika u proizvodima i uslugama koje se svakodnevno koriste, ali polako se pojavljuju alati koji omogućuju prilagođavanje i izbor tehnika strojnog učenja i umjetne inteligencije nestručnjacima i to bez poznavanja programiranja. Jedan od tih alata je i KNIME koji je opisan u ovom priručniku.

Priručnik koji pokriva područje strojnog učenja je prilično nezahvalno pisati, jer se radi o području koje se intenzivno razvija i gdje morate biti svjesni da će se koristiti nekoliko godina te će nakon toga vrlo brzo zastarjeti. Baš iz tog razloga u ovom priručniku su opisane tehnike koje nisu nove i koje su poznate desetljećima, a priručnik služi da se bez poznavanja programiranja i matematičke pozadine samih tehnika čitatelji upuste u izradu modela s vlastitim podacima. Time je priručniku ipak malo produženo vrijeme korištenja. Autor ne očekuje kako će čitateljima ovaj priručnik biti dovoljan, jer on služi za upoznavanje sučelja i osnovnih tehnika strojnog učenja. Nakon toga na čitateljima ostaje da sami nastave proučavati mogućnosti i tehnika koje nudi KNIME.

Malo bolji poznavatelji strojnog učenja zamijetit će da u priručniku nisu opisane neuronske mreže, a nema ni jezičnih modela koji su u posljednje vrijeme izrazito popularni. Autor je odlučio te teme ostaviti za nastavak serije o korištenju KNIME-a pri čemu će biti obrađene napredne tehnike kroz primjere u turizmu i ekonomiji. U nastavku će se vjerojatno koristiti verzija KNIME-a 5.x ili čak 6.x, a ovaj priručnik pokriva stariju verziju s kompleksnijim sučeljem.

Jedan od najvećih problema pri izradi priručnika autor je imao s terminologijom. U želji da pojmovi budu lakše shvatljivi i osobama koje ne koriste engleski jezik, pri prevođenju je napravljen niz kompromisa. Neki prijevodi ne zvuče dobro na hrvatskom jeziku i moguće je kako postoje bolji, ali nekim čitateljima će pomoći da bolje razumiju način na koji KNIME radi.

Ovaj priručnik bit će objavljen u elektroničkom obliku i u otvorenom pristupu što znači da ga možete slobodno kopirati i dijeliti u izvornom obliku bez ograničenja. Time je izbjegnut proces tiska koji bi dodatno odgodio trenutak izdavanja, a osim toga tisak bi učinio sam priručnik teže dostupnim jer bi se tiskana inačica plaćala. Za sve sugestije i prijedloge vezano uz ovaj priručnik ili inačicu s naprednim metodama koja je u planu, sam otvoren. Slobodno me možete pronaći na internetu i javiti mi se.

Kristian

Popis slika

Popis tablica

Popis čvorova

Popis korištenih skupova podataka po poglavljima

POGLAVLJE	NAZIV	LINK
4.	Data Expo 2009: Airline On Time Data	https://www.kaggle.com/datasets/wenxingdi/data-expo-2009-airline-on-time-data?select=2008.csv
5.	E-commerce - Users of a French C2C fashion store	https://www.kaggle.com/datasets/jmmvutu/ecommerce-users-of-a-french-c2c-fashion-store
6.	Apartment data	https://www.kaggle.com/datasets/gunhee/koreahousedata
7.	Adult	http://archive.ics.uci.edu/dataset/2/adult
8.	Churn in Telecom's dataset	https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset
9.	Absenteeism Dataset	https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset
10.	Are Your Employees Burning Out?	https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out
11.	Hotel booking demand	https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
12.	Mall customers	https://www.kaggle.com/datasets/kandij/mall-customers

Bibliografija

Artasanchez, A. & Joshi, P., 2020. *Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3. x.* s.l.:Packt Publishing Ltd.

Berrar, D., 2018. Bayes' theorem and naive Bayes classifier. *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics,* Svezak 403, p. 412.

Breiman, L., 2001. Random forests. *Machine learning,* Svezak 45, p. 5–32.

Brownlee, J., 2016. *Machine learning algorithms from scratch with Python.* s.l.:Machine Learning Mastery.

Brownlee, J., 2016. *Master Machine Learning Algorithms.* 1.1 ur. s.l.:Jason Brownlee.

Buglear, J., 2010. *Stats means business.* s.l.:Routledge.

Bussmann, N., Giudici, P., Marinelli, D. & Papenbrock, J., 2021. Explainable machine learning in credit risk management. *Computational Economics,* Svezak 57, p. 203–216.

Chen, J. & Jenkins, W. K., 2017. *Facial recognition with PCA and machine learning methods.* s.l., an., p. 973–976.

Choy, G. i dr., 2018. Current applications and future impact of machine learning in radiology. *Radiology,* Svezak 288, p. 318–328.

Coad, J., 2021. *Why You (Probably) Shouldn't Use Reinforcement Learning.* [Mrežno] Available at: https://towardsdatascience.com/why-you-shouldnt-use-reinforcement-learning-163bae193da8
[Pokušaj pristupa 17 kolovoz 2023].

Copeland, B., 2023. *artificial intelligence.* [Mrežno] Available at: https://www.britannica.com/technology/artificial-intelligence/Methods-and-goals-in-AI
[Pokušaj pristupa 17 kolovoz 2023].

Dakić, B. & Elezović, N., 2019. *Matematika 1, udžbenik i zbirka zadataka za 1. razred gimnazija i strukovnih škola, 2. dio.* s.l.:Element.

Damale, R. C. & Pathak, B. V., 2018. *Face recognition based attendance system using machine learning algorithms.* s.l., an., p. 414–419.

Eberly, L. E., 2007. Multiple linear regression. *Topics in Biostatistics,* p. 165–187.

Egger, R., 2022. *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications.* s.l.:Springer.

Ganapathiraju, A., Hamaker, J. E. & Picone, J., 2004. Applications of support vector machines to speech recognition. *IEEE transactions on signal processing,* Svezak 52, p. 2348–2355.

Ganapathiraju, A., Hamaker, J. & Picone, J., 2000. *Hybrid SVM/HMM architectures for speech recognition..* s.l., an., p. 504–507.

Géron, A., 2022. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* s.l.:" O'Reilly Media, Inc.".

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* New York, NY: Springer.

Horvat, J. & Mijoč, J., 2019. *Istraživački SPaSS.* s.l.:Naklada Ljevak.

Hosch, W. L., 2023. *machine learning.* [Mrežno] Available at: https://www.britannica.com/technology/machine-learning [Pokušaj pristupa 24 srpanj 2023].

Kelley, K. & Lai, K., 2011. Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research,* Svezak 46, p. 1–32.

Khairi, M. & Darmawan, D., 2021. The Relationship Between Destination Attractiveness, Location, Tourism Facilities, And Revisit Intentions. *Journal of Marketing and Business Research (MARK),* Svezak 1, p. 39–50.

Khandani, A. E., Kim, A. J. & Lo, A. W., 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance,* Svezak 34, p. 2767–2787.

KNIME AG, 2023. *About KNIME.* [Mrežno] Available at: https://www.knime.com/about [Pokušaj pristupa srpanj 2023].

Kohli, M., Prevedello, L. M., Filice, R. W. & Geis, J. R., 2017. Implementing machine learning in radiology practice and research. *American journal of roentgenology,* Svezak 208, p. 754–760.

Kovač, A., Dunđer, I. & Seljan, S., 2022. *An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services.* s.l., an., p. 954–961.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C. & Lu, C.-J., 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis,* Svezak 50, p. 1113–1130.

Liu, Y. H., 2017. *Python machine learning by example.* s.l.:Packt Publishing Ltd.

Lloyd, S., 1982. Least squares quantization in PCM. *IEEE transactions on information theory,* Svezak 28, p. 129–137.

MacQueen, J. & others, 1967. *Some methods for classification and analysis of multivariate observations.* s.l., an., p. 281–297.

Maulana, Y., Ulinnuha, H. & Chandra, D. L. T., 2021. *The effect of tourism attractions on tourists' visiting interest to Penglipuran village, Bangli district.* s.l., an., p. 012035.

McCulloch, W. S. & Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics,* Svezak 5, p. 115–133.

Mehra, S. & Hasanuzzaman, M., 2020. *Detection of Offensive Language in Social Media Posts,* s.l.: an.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2018. *Foundations of machine learning.* s.l.:MIT press.

Nagy, Z., 2018. *Regex quick syntax reference: understanding and using regular expressions.* s.l.:Apress.

Novarlia, I., 2022. Tourist Attraction, Motivation, and Prices Influence on Visitors' Decision to Visit the Cikandung Water Sources Tourism Object. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal),* Svezak 5.

Ostertagová, E., 2012. Modelling using polynomial regression. *Procedia Engineering,* Svezak 48, p. 500–506.

Powers, D. M. W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061.*

Probyto Data Science and Consulting Pvt. Ltd., 2020. *Data Science for Business Professionals: A Practical Guide for Beginners.* New Delhi: BPB Publications.

Sahay, A., 2021. *Essentials of Data Science and Analytics: Statistical Tools, Machine Learning, and R-Statistical Software Overview.* s.l.:Business Expert Press.

Seemakurthi, P., Zhang, S. & Qi, Y., 2015. *Detection of fraudulent financial reports with machine learning techniques.* s.l., an., p. 358–361.

Silipo, R., 2011. *KNIME Beginner's Luck.* s.l.:KNIME Press, Zurich, Switzerland.

Sterne, J., 2017. *Artificial intelligence for marketing: practical applications.* s.l.:John Wiley & Sons.

Su, X., Yan, X. & Tsai, C.-L., 2012. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics,* Svezak 4, p. 275–294.

Tatsat, H., Puri, S. & Lookabaugh, B., 2020. *Machine Learning and Data Science Blueprints for Finance.* s.l.:O'Reilly Media.

Wang, S. & Summers, R. M., 2012. Machine learning and radiology. *Medical image analysis,* Svezak 16, p. 933–951.

Watt, J., Borhani, R. & Katsaggelos, A. K., 2020. *Machine learning refined: Foundations, algorithms, and applications.* s.l.:Cambridge University Press.

Wei, J., Jian-Qi, Z. & Xiang, Z., 2011. Face recognition method based on support vector machine and particle swarm optimization. *Expert Systems with Applications,* Svezak 38, p. 4390–4393.

William, P. i dr., 2022. *Machine learning based automatic hate speech recognition system.* s.l., an., p. 315–318.

Xu, H., Fan, G., Song, Y. & others, 2022. Novel key indicators selection method of financial fraud prediction model based on machine learning hybrid mode. *Mobile Information Systems,* Svezak 2022.