Primjena metoda strojnog učenja u proizvodnim sustavima

De Marco, Marjam

Master's thesis / Diplomski rad

2018

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture / Sveučilište u Zagrebu, Fakultet strojarstva i brodogradnje**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:235:657226

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of Faculty of Mechanical Engineering and Naval Architecture University of Zagreb







SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

DIPLOMSKI RAD

MARJAM DE MARCO

Zagreb, 2018.

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

PRIMJENA METODA STROJNOG UČENJA U PROIZVODNIM SUSTAVIMA

Mentor: Student:

Prof. dr. sc. Dragutin Lisjak Marjam De Marco

Zagreb, 2018.

Izjavljujem da sam ovaj rad izradila samostalno koristeći znanja stečena tijekom studija i navedenu literaturu.

Zahvaljujem se mentoru prof. dr. sc. Dragutinu Lisjaku i asistentu Davoru Kolaru, na ukazanom povjerenju, stručnim savjetima, pomoći kao i na izdvojenom vremenom prilikom izrade ovog diplomskog rada.

Isto tako, zahvaljujem se braći, prijateljima i kolegama na podršci tokom studiranja. Ponajviše se zahvaljujem majci na potpori, strpljenju i riječima ohrabrenja te što je imala vjere u mene kad mi je to bilo najpotrebnije.

Marjam De Marco

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE
Središnje povjerenstvo za završne i diplomske ispite
Povjerenstvo za diplomske radove studija strojarstva za smjerove:
proizvodno inženjerstvo, računalno inženjerstvo, industrijsko inženjerstvo i menadžment, inženjerstvo materijala te mehatronika i robotika

Sveučilište u Zagrebu Fakultet strojarstva i brodogradnje	
Datum:	Prilog:
Klasa:	
Ur. broj:	

DIPLOMSKI ZADATAK

Student: **MARJAM DE MARCO** Mat. br.: 0035188697

Naslov rada na hrvatskom jeziku: **Primjena metoda strojnog učenja u proizvodnim sustavima**

Naslov rada na engleskom jeziku: **Application of machine learning methods in production systems**

Opis zadatka:

Razvojem informacijskih sustava za upravljanje proizvodnjom (eng. Manufacturing execution systems, MES) i njihovom integracijom sa sustavima za planiranje i upravljanje resursima poduzeća (eng. Enterprise resource planning, ERP), omogućeno je planiranje i praćenje pretvorbe ulaznih resursa u proizvod. Pohranjeni podaci većinom se koriste za praćenje informacija o određenom procesu ili za osnovno izvještavanje o prošlim aktivnostima. Primjenom metoda strojnog učenja, moguće je iz pohranjenih podataka generirati novih skup implicitnih znanja, stvarajući pritom dodanu vrijednost kroz mogućnost dodatne optimizacije proizvodnih procesa. U skladu s navedenim, u radu je potrebno:

1. Opisati metodologiju rudarenja podataka .
2. Analizirati dostupne podatke primjenom CRISP-DM procesnog modela te predložiti prikladnu metodu za otkrivanje znanja u dostupnom skupu podataka.
3. Prikazati način implementacije predložene metode za otkrivanje znanja primjenom softverskog alata uz detaljni prikaz i objašnjenje dobivenih rezultata.
4. Analizom dobivenih rezultata izvesti zaključke.

U radu je potrebno navesti korištenu literaturu i eventualno dobivenu pomoć.

Zadatak zadan:
27. rujna 2018.

Rok predaje rada:
29. studenog 2018.

Predviđeni datum obrane:
05. prosinca 2018.
06. prosinca 2018.
07. prosinca 2018.

Zadatak zadao:

prof. dr. sc. Dragutin Lisjak

Predsjednica Povjerenstva:

prof. dr. sc. Biserka Runje

SADRŽAJ

SADRŽAJ I
POPIS SLIKA III
POPIS TABLICA V
SAŽETAK VI
SUMMARY VII
1. UVOD 1
2. STROJNO UČENJE 2
2.1. Primjene strojnog učenja 2
2.2. Tehnike strojnog učenja 3
3. TEHNIKE ZA DUBINSKU ANALIZU PODATAKA 6
3.1. Statistika 6
3.2. Baze podataka 7
3.3. Vizualizacija 8
4. CRISP – DM METODOLOGIJA 9
4.1. Razumijevanje područja i podataka 9
4.2. Priprema podataka 10
4.3. Modeliranje 10
4.4. Evaluacija i implementacija 10
5. KLASTER ANALIZA 13
5.1. Tehnike klaster analize 16
5.2. Metoda k-srednjih vrijednosti 19
5.3. Mjere udaljenosti između uzoraka 21
6. REGRESIJA I KLASIFIKACIJA 24
6.1. Regresija 24
6.1.1. Linearna regresija 24

6.1.2. Višestruka regresija 25
6.1.3. Logistička regresija 25
6.1.4. Nelinearna regresija 25
6.2. Klasifikacija 25
6.3. Tumačenje rezultata 27
7. PRIMJENA METODA STROJNOG UČENJA U PROIZVODNOM POGONU 28
7.1. PROCES PRIPREME PODATAKA 28
7.1.1. Dokumentiranje dostupnog skupa podataka 28
7.1.2. Statistika podataka 33
7.1.3. Izrada novih atribut 35
7.1.4. Transformacija podataka 37
7.1.5. Priprema podataka za proces dubinske analize 38
7.2. EKSPLORATIVNA ANALIZA PODATAKA 42
7.3. REGRESIJSKI MODEL 50
7.3.1. Prvi regresijski model 51
7.3.2. Drugi regresijski model 54
7.3.3. Analiza rezultata i prijedlog za daljnju analizu 55
7.4. KLASIFIKACIJSKI MODEL 60
7.4.1. Prvi klasifikacijski model 60
7.4.2. Analiza rezultata i prijedlog za daljnje istraživanje 65
7.4.3. Klaster analiza s ciljem optimizacije rezultata 66
7.4.4. Drugi klasifikacijski model 68
7.4.5. Optimizacija rezultata proširivanjem ulaznog skupa podataka 70
8. ZAKLJUČAK 73
LITERATURA 75

POPIS SLIKA

Slika 1. Tehnike strojnog učenja [7] 3
Slika 2. Tehnike dubinske analize podatakan [7] 6
Slika 3.CRISP DM [18] 9
Slika 4. CRISP DM faze i opis [18] 12
Slika 5. Klaster analiza [4] 13
Slika 6. Raspored podataka u koordinatnom sustavu [14] 14
Slika 7. Podjela klaster postupaka [23] 15
Slika 8.Hijerarhijsko grupiranje [24] 16
Slika 9. Aglomerativni i divizijski postupak [22] 17
Slika 10 k-means algoritam [23] 20
Slika 11. Princip rada k-means algoritma [9] 20
Slika 12. Euklidska udaljenost [23] 21
Slika 13. Blokovska udaljenost 22
Slika 14. Razlika između regresije i klasifikacije [30] 26
Slika 15. Izlazne varijable klasifikacije i regresije 26
Slika 16. Vizualni prikaz atributa 29
Slika 17. Prikaz dijela podataka 33
Slika 18. Prikaz atributa evid_trajanje_procesa_sekundi 36
Slika 19. Prikaz atributa proizvod_motor 36
Slika 20. Prikaz vremenskih atributa 37
Slika 21. Prikaz atributa i tipa podatka 38
Slika 22. Prikaz procesa pripreme podataka 38
Slika 23. Operator Filter Examples 39
Slika 24. Mapiranje atributa 40
Slika 25. Operator Replace Missing Values 40
Slika 26. Operator Select Attributes 41
Slika 27. Dijagram trajanja procesa i broja djelatnika na op. liniji 42
Slika 28. Dijagram završetka proizvodnje i broja djelatnika na op. liniji 43
Slika 29. Histogram raspodjele duljine trajanja procesa 44
Slika 30. Histogram raspodjele duljine trajanja procesa prije pripreme podataka 45
Slika 31. Odnos promjera proizvoda i vremena trajanja procesa 46

Slika 32. Odnos širine proizvoda i vremena trajanja procesa 47
Slika 33. Odnos visine proizvoda i vremena trajanja procesa 48
Slika 34. Omjer pravokutnih i cilindričnih proizvoda 49
Slika 35. Podjela podataka na cilindrične i pravokutne 50
Slika 36. Odabir atributa za cilindrične proizvode 50
Slika 37. Odabir atributa za pravokutne proizvode 51
Slika 38. Prvi regresijski model 51
Slika 39. Podjela podataka na trening i test skup prvog regresijskog modela 52
Slika 40. Grafikon koeficijenta korelacije za prvi regresijski model 56
Slika 41. Graf prosječne vrijednosti koeficijenta korelacije za prvi regresijski model 57
Slika 42. Grafikon koeficijenta korelacije za drugi regresijski model 57
Slika 43. Graf prosječne vrijednosti koeficijenta korelacije za drugi regresijski model 58
Slika 44. Usporedne vrijednosti za prvi i drugi regresijski model 59
Slika 45. Prvi dio prvog klasifikacijskog modela 60
Slika 46. Priprema podataka za prvi klasifikacijski model 60
Slika 47. Odabir ciljanog atributa 61
Slika 48. Prikaz operatora Binning 61
Slika 49. Raspodjela podataka po spremnicima 61
Slika 50. Operator Replace Missing Values 62
Slika 51. Prikaz ulaznih i izlaznih signala [35] 63
Slika 52. Slojevi neuronske mreže [35] 63
Slika 53. Parametri za duboko učenje 64
Slika 54. Rectifier aktivacijska funkcija [38] 65
Slika 55. Matrica konfuzije prvog klasifikacijskog modela 65
Slika 56. Prikaz modela za klaster analizu 67
Slika 57. Prikaz atributa koji nisu ušli u klaster analizu 67
Slika 58. Prikaz grupiranja podataka 68
Slika 59. Raspodjela podataka po klasterima 68
Slika 60. Matrica konfuzije drugog klasifikacijskog modela 69
Slika 61. Tlocrt operacijske linije 70
Slika 62. Raspodjela podataka po spremnicima 71
Slika 63. Matrica konfuzije optimiranog klasifikacijskog modela 71

POPIS TABLICA

Tablica 1. Prikaz matrice konfuzije [20] 11
Tablica 2. Prikaz evaluacijskih mjera temeljenih na analizi matrice konfuzije [20] 11
Tablica 3. Skup podataka [14] 13
Tablica 4. Popis atributa i njihov opis, tip podatka 29
Tablica 5. Maksimalna, minimalna i prosječna vrijednost podataka 33
Tablica 6. Različitost tekstualnih podataka 34
Tablica 7. Atributi sa nedostajućim vrijednostima 34
Tablica 8. Tip i opis podatka [34] 37
Tablica 9. Rezultati prvog regresijskog modela za cilindrične proizvode 53
Tablica 10. Rezultati prvog regresijskog modela za pravokutne proizvode 54
Tablica 11. Rezultati drugog regresijskog modela za cilindrične proizvode 54
Tablica 12. Rezultati drugog regresijskog modela za pravokutne proizvode 55

SAŽETAK

Tema ovog diplomskog rada je primjena metoda strojnog učenja u linijskoj proizvodnji protupožarnih zaklopki i regulatora varijabilnog protoka za operaciju montaže. Primjena strojnog učenja i dubinske analize podataka postaje standard u svim aspektima proizvodnje, kako bi se otkrile skrivene informacije i znanje utkano u podacima, a uvelike pridonoseći procesu donošenja odluka i poslovanja.

Cilj ovog rada je pronaći matematički model koji će prema kriteriju točnosti analizirati podatke te doprinijeti razumijevanju procesa montaže linijske proizvodnje. U prvom djelu rada opisana je teorija strojnog učenja, tehnike za dubinsku analizu podataka – klaster analiza, regresija i klasifikacija.

U praktičnom dijelu rada, na podacima iz poslovnog sustava napravljena je sustavna analiza kao i priprema podataka. U cilju rješavanja navedenog problema, u radu su kreirani regresijski i klasifikacijski modeli te je napravljena klaster analiza sa svrhom optimizacije rezultata klasifikacijskog modela. U zaključnom dijelu rada dane su i smjernice za daljnja istraživanja.

Ključne riječi: strojno učenje, dubinska analiza podataka, klasifikacija, regresija, klaster analiza.

SUMMARY

The subject of this thesis is application of machine learning methods in line production of fire dampers and variable flow regulators for the assembly line. The use of machine learning and deep data analysis has become a standard in all aspects of production. Its use is to reveal hidden information and knowledge in the data which greatly contributes to the decision making process.

The aim of this thesis is to find a mathematical model that will analyze the data according to the criterion of accuracy and contribute to the understanding of the line production assembly process. The first part of the paper describes the theory behind machine learning, the techniques for deep data analysis – cluster methods, regression and classification.

In the practical part of the thesis, systematic analysis and data preparation were performed on the data from the business system In order to address this problem, regression and classification models were created. To optimize the classification model results a cluster analysis was used. Further research guidelines are noted in the conclusion.

Key words: machine learning, deep data analysis, classification, regression, cluster analysis.

1. UVOD

Društvene, znanstvene i različite industrijske djelatnosti preplavljene su velikom količinom podataka koje se svakodnevno pohranjuju u bazama podataka. Pojavom tehnika i metoda strojnog učenja (eng**.** *Machine Learning – ML)* odgovara se na problem kako iz raspoloživih skupova podataka proizvesti novo znanje.

Nove tehnologije, poput bežičnog prijenosa senzorskih podataka sa strojeva u baze podataka, omogućuju da se relevantne informacije pohranjuju u digitalnom obliku. Prikupljeni i pohranjeni podaci sami po sebi ne predstavljaju dodanu vrijednost, već se mogu definirati kao potencijal u kojem poduzeća mogu, kroz njihovu analizu, doći do prednosti nad drugim poduzećima. Analiza prikupljenih podataka vrlo brzo postaje ograničena mogućnostima analitičkih alata, a može se povećati primjenom algoritama strojnog učenja, koji spadaju u domenu rudarenja podataka tj. dubinsku analizu podataka (eng**.** *Data Mining-DM*).

Dubinska analiza podataka predstavlja niz tehnika, koje omogućuju korisnicima izdvajanje korisnih informacija, odnosno otkrivanje implicitnog znanja, iz brzo rastućih količina podataka.

Ovaj rad opisuje mogućnost primjene dubinske analize podataka prikupljenih u proizvodnom procesu, s ciljem osiguravanja dodatne potpore odlučivanju u procesu planiranja proizvodnje.

U drugom poglavlju objašnjene su primjene i tehnike strojnog učenja, dok se treće poglavlje bavi tehnikama za dubinsku analizu podataka. Najraširenija metodologija za dubinsku analizu podataka prikazana je u četvrtom poglavlju. U petom i šestom poglavlju prikazane su i objašnjene metode za dubinsku analizu podataka koje će se koristiti nad stvarnim podacima, a to su analiza klastera, klasifikacija i regresija. Proces pripreme podataka iz proizvodnog sustava, analiza podataka, izrada modela za dubinsku analizu učenja, njihova optimizacija i analiza rezultata prikazani su u sedmom poglavlju. Na kraju rada dan je zaključak o rezultatima istraživanja te su dane smjernice za daljnji razvoj.

2. STROJNO UČENJE

Postoji mnogo definicija za dubinsku analizu podataka. Dubinska analiza podataka je interdisciplinarna znanost koja povezuje više grana kao što su npr. statistika, strojno učenje, baza podataka i vizualizacija. Dubinsku analizu podataka možemo definirati kao proces pronalaženja novog i potencijalnog korisnog znanja iz podataka, odnosno kao otkrivanje ili 'rudarenje' znanja iz velike količine podataka [1].

Kao što je navedeno, strojno učenje je jedna od temeljnih tehnika za dubinsku analizu podataka.

Tehnike strojnog učenja razlikuju dubinsku analizu podataka od tradicionalne analize. Strojno učenje je osnovno svojstvo po kojem se dubinska analiza podataka razlikuje od tradicionalne metoda analize podataka (deskriptivna statistika, diferencijalna statistika koja se bavi parcijalnim skupovima podataka). Strojno učenje je dio područja računarstva poznatog kao umjetna inteligencija, koja se bavi razvojem računarskih postupaka podložnih računalnim simulacijama, kao npr. čovjekovog ponašanja [2], [3].

Izgradnja modela dubinske analize podataka omogućuje lakše razumijevanje i upoznavanje s dostupnim podacima te otkrivanje logičkih veza između podataka. Rezultat strojnog učenja je modela, a točnost izgrađenog modela testira se na skupu podataka koji nisu korišteni za izgradnju modela (skup podataka za učenje) [2].

S velikim skupom podataka, očekuje se da će model strojnog učenja točnije procjenjivati prema kriteriju točnosti ponašanje budućih uzoraka. Ideja strojnog učenja jest stvoriti sustav koji će olakšati donošenje odluka (eng**.** *Decision support system*). Na ovaj način, omogućava se korisnicima računalna potpora u procesu odlučivanja. Iz navedenog proizlazi da se strojno učenje bavi izgradnjom sustava za podršku pri odlučivanju, čiji je rezultat donošenje boljih odluka i planova djelovanja koji omogućuju učinkovitije postizanje ciljeva poslovnog sustava [4].

2.1. Primjene strojnog učenja

Kao što je navedeno u prethodnom odjeljku, strojno učenje je temelj dubinske analize podataka, odnosno procesa otkrivanja znanja u velikim skupovima podataka. Samim time, strojno učenje je široko primijenjeno u različitim granama znanosti. Široko područje rada rezultira primjenom metoda strojnog učenja na rješavanje složenih problema na gotovo sva područja ljudske djelatnosti.

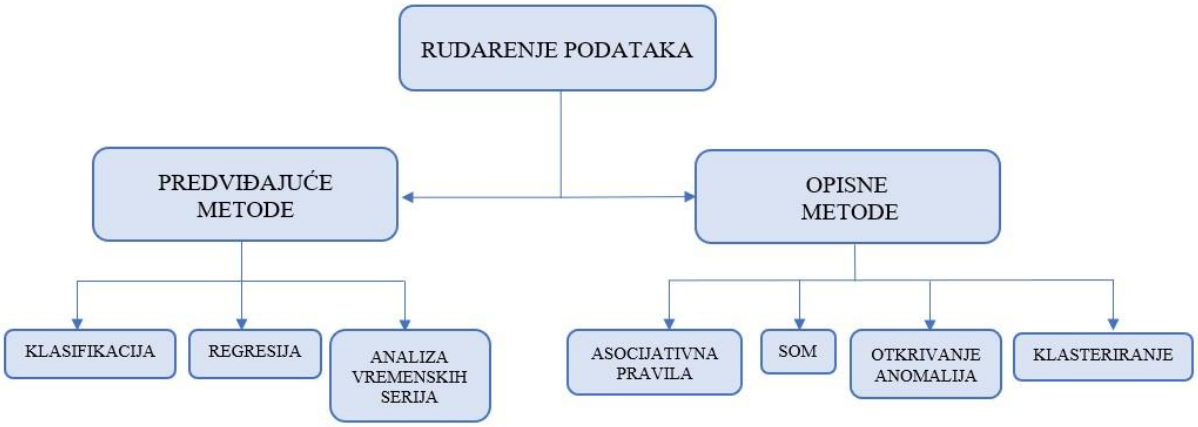
Neki od primjera gdje se koristi strojno učenje su [5]:

- Programske implementacije koje nisu rješive klasičnim programiranjem
- Prilagodljivi programski sustavi
- Bio informatika
- Obrada prirodnog jezika
- Raspoznavanje govora
- Raspoznavanje uzoraka
- Inteligentno upravljanje
- Predviđanje trendova.

Tehnologija svaki dan ide korak dalje, a strojno učenje prati napredak tehnologije. Tako se strojno učenje primjenjuje kod osobnih virtualnih pomoćnika kao što su Siri, Alexa i Bixby [4]. Strojno učenje u ovom primjeru prikuplja i pročišćuje podatke na temelju prethodnih pretraživanja. Nadalje, strojno učenje se primjenjuje prilikom GPS navigacije, gdje procjenjuje na kojim lokacijama su moguće gužve na temelju svakodnevnih iskustava. Još jedan od svakodnevnih primjera strojnog učenja može se naći na široko primijenjenim društvenim mrežama kao što je npr. *Facebook*. Prilikom učitavanja slike na navedenu mrežu, strojno učenje omogućuje prepoznavanje lica (eng**.** *Face recognition)* gdje analizira sliku te je povezuje s drugim korisnicima [6].

2.2. Tehnike strojnog učenja

Tehnike strojnog učenja mogu se podijeliti u dvije glavne kategorije; nadzirano učenje (eng**.** *Supervised learning),* nenadzirano učenje (eng**.** *Unsupervised learning)* i nekoliko potkategorija, prikazanih na slici 1.



Slika 1. Tehnike strojnog učenja [7]

Nadzirano učenje podrazumijeva da u skupu podataka postoji ciljna (zavisna) varijabla na temelju koje će model nadziranog učenja predviđati buduće ishode (npr. hoće li nastati kvar na stroju temeljem prikupljenih senzorskih podataka). Cilj ove vrste strojnog učenja je dobiti prediktivni model na osnovu vrijednosti prikupljenih podataka [8].

Nadzirano učenje za izgradnju modela koristi podatke (npr. senzorske zapise) za koje su unaprijed poznati razredi (npr. kvar / nije kvar) . Na temelju izgrađenog modela, predviđa se razred kojemu će pripadati nepoznati podaci (npr. novo prikupljeni senzorski podaci) [9].

Podaci su u obliku *(x,y),* gdje *x* označava ulaznu vrijednost, a *y* ciljnu vrijednost. Prilikom nadziranog učenja, uči se funkcija *f(x)=y* koja za nove ulazne varijable *x*, predviđa izlazne varijable *y.* Ukoliko je *y* diskretna varijabla, radi se o klasifikaciji (eng**.** *Classification),* a ako je *y* kontinuirana varijabla (cjeloviti broj) koristi se metoda regresije (eng**.** *Regression)* [9], [10].

Nadzirano učenje primjenjuje se za[10]:

- Predviđanje – na temelju ulaznih (nezavisnih) značajki, predviđaju se zavisne značajke
- Ekstrakciju znanja – učenje lako tumačenih modela (npr. stablo odluke)
- Sažimanje – model koji jezgrovito objašnjava podatke
- Upravljanje – upravljački ulazi dobiveni kao izlaz regresije.

Nenadzirano učenje sinonim je za grupiranje i suprotnost je nadziranoj metodi strojnog učenja. Ukoliko je definicija problema dubinske analize podataka nepoznata, tada se problem naziva nenadziran [11].

Tehnike nenadziranog učenja su:

- Otkrivanje asocijativnih pravila **–** otkrivanje relacijskih veza koje ukazuju na učestali odnos (relaciju) između vrijednosti značajke i pojedine značajke u skupu podataka [12]
- Samo – organizirajuće mape **–** na ulaznom skupu podataka, mapa po dobro definiranom algoritmu samoorganizira varijable podešavanjem svojih parametara[13]
- Otkrivanje anomalija **–** model prema ulaznim podacima nauči očekivane uzorke i ponašanja, te lako otkriva i izlučuje nepoznato i neviđeno tj. anomalije [14]
- Klaster analiza **–** metoda kojom se skup uzoraka svrstava u skupove međusobno što sličnijih podataka [9].

Za razliku od nadziranog učenja, gdje su poznate varijable *x* i *y,* kod nenadziranog učenja poznata je samo ulazna varijabla *x,* ali ne i odgovarajuća izlazna varijabla [8].

Drugim riječima, na temelju zadanih ulaznih podataka nenadzirano učenje formira grupe podataka, ali bez prethodnog saznanja o tome kojem razredu bi podaci mogli pripadati [9].

Nenadzirano učenje primjenjuje se za [10]:

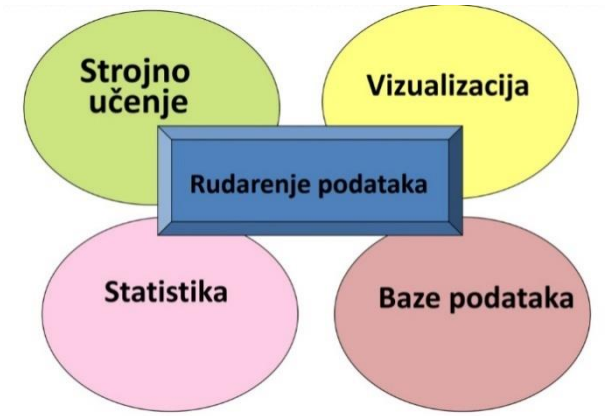
- Marketing – grupiranje korisnika prema potrošačkim navikama
- Biologija – grupiranje organizama prema njihovim značajkama
- Rudarenje po dokumentima – grupiranje sličnih dokumenata
- Pretraživanje informacija – grupiranje sličnih rezultata
- Obrada slike - sažimanje slike grupiranjem istobojnih piksela.

3. TEHNIKE ZA DUBINSKU ANALIZU PODATAKA

Dubinska analiza podataka objedinjuje više vrsta tehnika iz drugih domena istraživanja. Prema[1], tehnike koje posebno utječu na razvoj metoda DM su:

- Statistika
- Strojno učenje
- Baze podataka
- Vizualizacija
- Prepoznavanje uzoraka
- Dohvat informacija
- Algoritmi itd.

Ovaj rad baviti će se sa prvih četiri kao što je prikazano na slici 2.



Slika 2. Tehnike dubinske analize podatakan [7]

3.1. Statistika

Statistika odgovara na mnoga pitanja čiji su odgovori bitni prilikom dubinske analize podataka, kao što su kolika je vjerojatnost da će se stroj pokvariti, koliki je postotak ispravnih proizvoda, koliko često neki stroj treba održavati.

Dubinska analiza podataka je u svojim osnovama usko vezana na statističke pojmove i postupke. U većini postupaka dubinske analize u određenom obliku postoji prebrojavanje skupa podataka i uspoređivanje tako dobivenih veličina. Za razliku od statistike, koja je matematička disciplina usmjerena na pronalaženje i vjerodostojno vrednovanje odnosa koji postoje u

skupovima podataka, DM je tehnička disciplina koja teži otkrivanju bilo kojih potencijalno korisnih informacija sadržanih u podacima [2].

Prediktivna metoda dubinske analize podataka temelji se na prediktivnoj statistici koja modelira podatke na način da izdvaja slučajna ili nepouzdana zapažanja i oslikava zaključke procesa ili skupova koji se istražuju.

Kako bi obrada podataka prije dubinske analize bila uspješna, potrebno je steći cjelokupnu sliku baze podataka. To se obično radi općim statističkim opisima koji uvelike pomažu pri detektiranju nedostajućih podataka [11].

Statistika se u dubinskoj analizi koristi za potvrđivanje postavljene hipoteze. Ona se primjenjuje na malim skupovima podataka.

Uloga statističara je velika, dok je računalo samo pomoćni alat [7]

3.2. Baze podataka

Sustav za upravljanje bazom podataka sastoji se od povezanih podataka, skupa programa koji omogućuju pristup bazi podataka, stvaranje datoteka te unos i organizaciju podataka [11].

Sustav za upravljanje bazom podataka odgovoran je za:

- Pronalaženje i izdvajanje potrebnih informacija
- Sigurnosnim sustavom baze podataka
- Kreiranjem novih baza podataka
- Definiranje baze podataka
- Pisanje upita nad bazom
- Spremanje velikih količina podataka itd. [15]

Istraživanje sustava baza podataka fokusirano je na kreiranje, održavanje i korištenje baze podataka za krajnje korisnike. Većina zadaća dubinske analize podataka zahtjeva podršku velikim skupovima podataka i obradu u realnom vremenu te moraju stvoriti zadovoljstvo kod korisnika u naprednim analizama podataka [11].

Jasno je vidjeti zašto dubinska analiza podataka počiva na bazama podataka. Baze podatka omogućuju spremanje velikog broja podataka, grupiranjem prema atributima te pronalaženje i izdvajanje potrebnih informacija. Pomoću relacijskih baza podataka moguće je promatrati

trendove ponašanja ili uzorke podataka, na temelju koje dubinska analiza podataka dolazi do traženih zaključaka.

3.3. Vizualizacija

Vizualizacija (eng**.** *Visual Data Mining*) su postupci transformacije podataka u oblik prikladan za ljudsku interpretaciju. Rezultati se u završnoj fazi verificiraju statističkim postupcima, vizualno ilustriraju te eventualno integriraju sa znanjem iz drugih izvora.

Vizualizacija podataka pomaže korisniku tijekom obrade podataka i dubinske analize podataka. Kroz vizualizaciju izvornih podataka, korisnik može pregledavati podatke kako bi dobio uvid u njihovo kretanje. Konkretno, vizualizacija se može koristiti za otkrivanje anomalija tj. za otkrivanje podataka koji se ne podudaraju s ponašanjem modela. Osim toga, vizualizacija korisniku pomaže pri odabiru odgovarajućih podataka. Podaci koji ne dodaju vrijednost modelu ili podaci koji postižu jednake vrijednosti izbačeni su iz analize [16].

Vizualizacija podataka, može se koristiti i prilikom odabira modela učenja. Velike količine podataka se na ovakav način sažeto prikazuju i omogućuju korisnicima uvid u podatke koji nije moguć samim proučavanjem podataka. Pretvorbom numeričkih podataka u statističke grafikone, omogućuje se korisnicima jednostavno analiziranje velikog skupa složenih podataka [16].

Vizualizacija rezultata je bitna jer je ilustracija samih podataka dobar način prenošenja otkrivenog znanja. Na taj način je moguće prikazati i kvalitetu otkrivenih modela i novo otkrivene odnose u podacima [2].

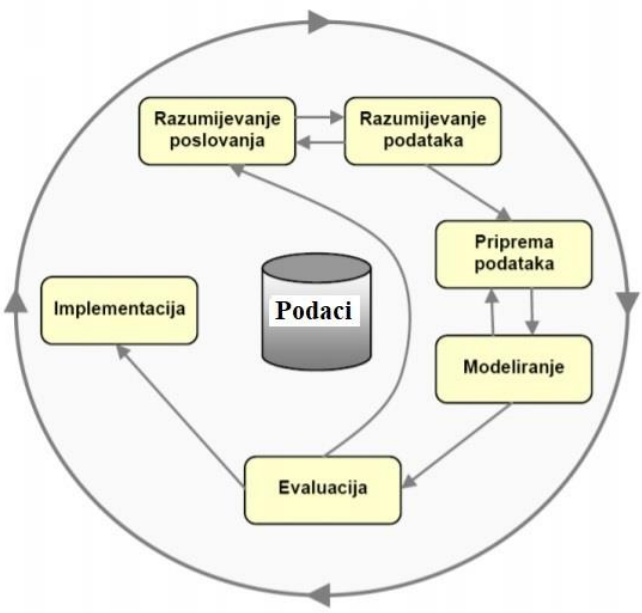
Cilj vizualizacije u rudarenju podataka jest dobiti početno razumijevanje podataka i procjenu njihove kvalitete. Stvarna točna procjena podataka i otkrivanje novih znanja su zadaci dubinske analize podataka. Stoga, vizualni prikaz treba biti razumljiv, jednostavan i sažet [16].

4. CRISP – DM METODOLOGIJA

CRISP DM (eng**.** *Cross Industry Standard Process for Data Mining*) je jedna od raširenijih metodologija za dubinsku analizu podataka pri rješavanju određenih problema [17].

Prema [18], CRISP metodologija sastoji se od šest osnovnih koraka:

1. Razumijevanje područja
2. Razumijevanje podataka
3. Priprema podataka
4. Modeliranje
5. Evaluacija
6. Implementacija



Slika 3.CRISP DM [18]

Kao što je prikazano na slici 3, razumijevanje područja i podataka povezano je povratnom vezom. Način na koji se definira poslovni cilj, imati će posljedice na vrstu i tip podataka koje će se primijeniti u modeliranju

4.1. Razumijevanje područja i podataka

U prvoj fazi CRISP metodologije definira se poslovni cilj. Ova faza se fokusira na razumijevanje poslovnih zahtjeva i ciljeve dubinske analize podataka. Oni se dalje pretvaraju u definiciju problema i preliminarni plan za ispunjavanje ciljeva.

Nakon definiranja poslovnih ciljeva u prvoj fazi, slijedi inicijalno prikupljanje i proučavanje podataka. U ovoj fazi se dolazi do prvog uvida u podatke koji daju odgovore na poslovna pitanja [7], [17], [18].

4.2. Priprema podataka

Faza pripreme podataka uobičajeno oduzima 90% od ukupnog vremena provedenog za dubinsku analizu podataka [7]. Kvaliteta podataka jedna je od najbitnijih stavki u dubinskoj analizi, jer baze podataka koje nisu kvalitetne mogu poremetiti analizu i dati krive rezultate što je ujedno jedan od razloga zašto se najviše vremena provodi na ovoj fazi.

Cilj ove faze je da se podaci transformiraju na takav način da se mogu prikazati nekim od modela za učenje, u namjeri da se kao konačni output dobe skupovi podataka za učenje, testiranje i validaciju. To se čini kroz odabir, čišćenje, spajanje, transformaciju i formatiranje podataka tj. kreiranje novih varijabli [19].

4.3. Modeliranje

Modeliranje je faza u kojoj se nakon definiranja uzorka za učenje i uzorka za testiranje, odabranom metodom dubinske analize podataka konstruiraju modeli za dubinsku analizu. Ova faza se sastoji od odabira metode modeliranja, kreiranje procedure za testiranje, izgradnje tog modela i na koncu procjene [18].

Odabir tehnike modeliranja ovisi o podacima i cilju dubinske analize. Modeliranje je iterativni proces te je različit za nadgledano i nenadgledano učenje.

Modelirati se može s ciljem deskripcije (opisa) ili predikcije (procjene).

4.4. Evaluacija i implementacija

Nakon što se model prihvati, potrebno je procijeniti koliko je on dobar. Upravo evaluacija modela predstavlja peti korak ove metodologije. Prije nego što se model primijeni, potrebno je proći niz koraka koji pored uspješnosti samog modela, evaluiraju i doprinos modela u ostvarivanju specificiranih ciljeva analize. Evaluacija modela je procjena performansi modela na testnim podacima. Ukoliko rezultati evaluacije nisu zadovoljavajući, potrebno je vratiti se na početnu fazu tj. razumijevanje područja i revidiranje poslovnog cilja te primijeniti neki drugi algoritam, kao što to prikazuje slika 3. Ova faza je iterativna i nema konačnog rezultata iz prvog puta. Praksa je pokazala da je iste podatke potrebno rudariti na različite načine kako bi se dobilo više rezultata nad kojima će se provesti analiza. Težina interpretacije ovisi o modelu [7], [17],[19].

Jedan od načina evaluacije je pomoću matrice konfuzije. Klasifikacijski algoritmi za predviđanje kategoričke klase se ocjenjuju matricom konfuzije koja se sastoji od četiri kategorije, kao što to prikazuje tablica 1.

Tablica 1. Prikaz matrice konfuzije [20]

		Stvarne vrijednosti	
		Nema kvara (0)	Kvar (1)
Predviđene vrijednosti	Nema kvara (0)	TN	FN
	Kvar (1)	FP	TP

Izrazi TN (eng**.** *True Negative)* predstavlja broj negativnih primjeraka koji su ispravno kvalificirani kao negativni.

Izraz FN (eng**.** *False Negative)* predstavlja broj pozitivnih primjeraka koji su pogrešno kvalificirani kao negativni.

Izraz FP (eng**.** *False Positive)* predstavlja broj negativnih primjeraka koji su pogrešno kvalificirano kao pozitivni.

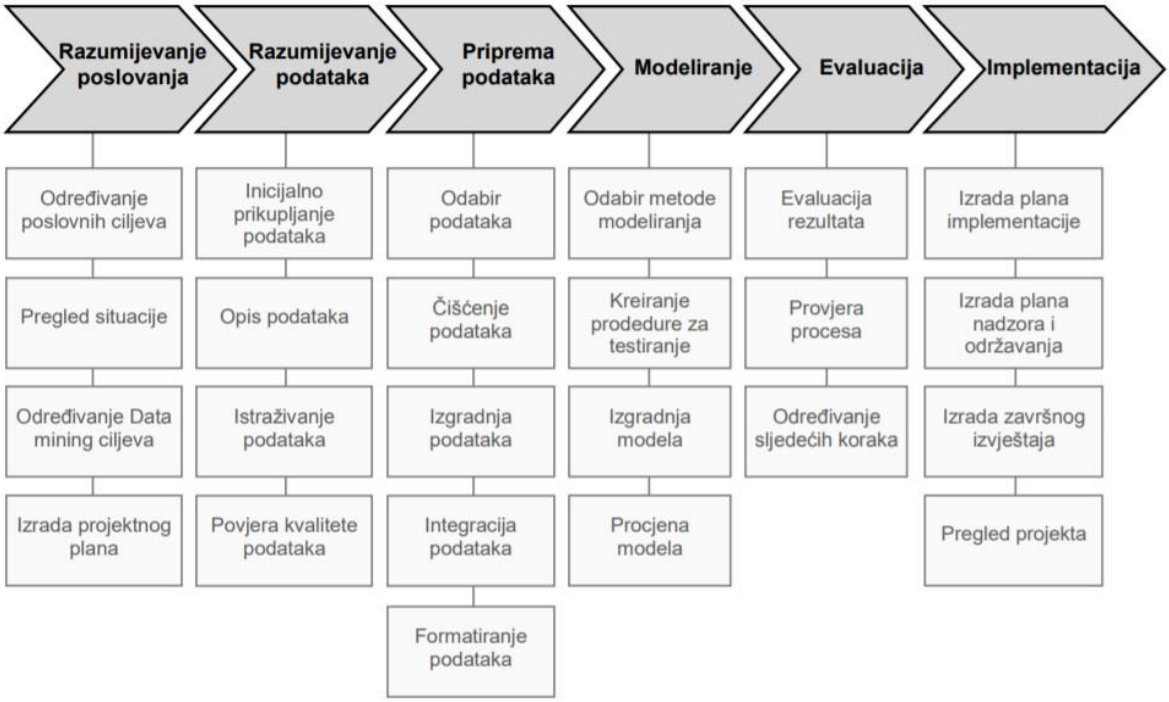
Izraz TP (eng**.** *True Positive)* predstavlja broj pozitivnih primjeraka koji su ispravno kvalificirani kao pozitivni.

Tablica 2. Prikaz evaluacijskih mjera temeljenih na analizi matrice konfuzije [20]

Mjera	**Formula**	**Interpretacija**
Točnost	$\frac{TP + TN}{TP + TN + FP + FN}$	Ukupna točnost Klasifikacijskog modela
Odziv	$\frac{TP}{TP + FN}$	Točnost pozitivnih primjeraka
Preciznost	$\frac{TP}{TP + FP}$	Ispravnost klasificiranja pozitivnih primjeraka

Posljednja faza CRISP metodologije predstavlja primjenu razvijenog modela.

Prilikom implementacije treba donijeti odluke o tome kako će rezultati biti iskorišteni, tko ih treba koristiti, koliko često se trebaju koristiti/ provoditi [17].



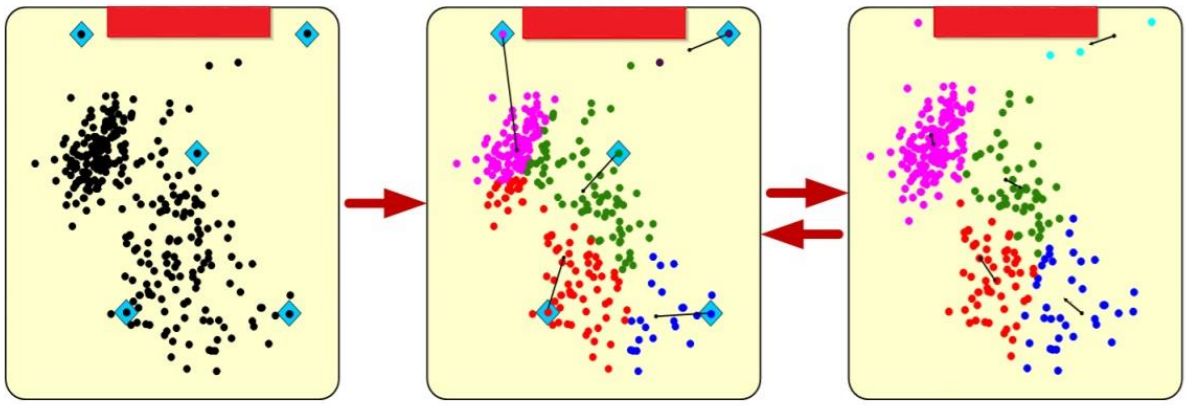
Slika 4. CRISP DM faze i opis [18]

Sažeti prikaz CRISP DM metodologije vidljiv je na slici 4. gdje su prikazane sve faze i što one uključuju

.

5. KLASTER ANALIZA

Klaster analiza je metoda kojom se skup uzoraka svrstava u skupove međusobno što sličnijih podataka. Objekti unutar određene grupe (eng**.** *Clusters*) moraju biti međusobno što sličniji i moraju se što više razlikovati od objekata koji se nalaze unutar druge grupe. Kao što je već napomenuto, skup podataka koji će se koristiti u klaster analizi nema ciljne značajke. Iz tog razloga podaci su grupirani prema sličnim vrijednostima koje se nalaze unutar ulaznih značajki [9], [20].



Slika 5. Klaster analiza [4]

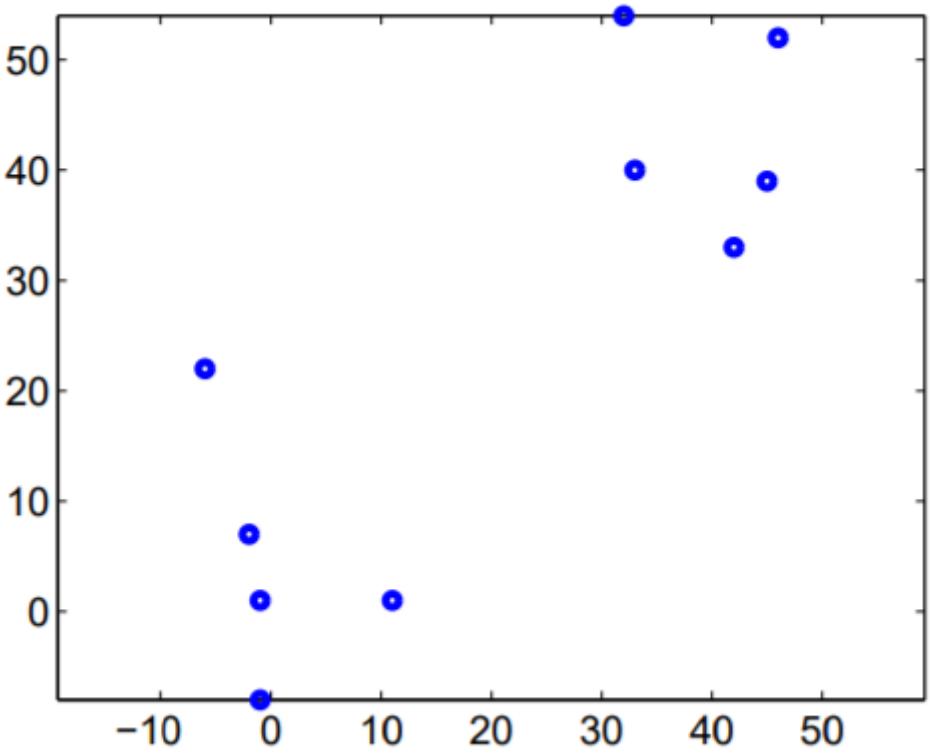
Kako bi se utvrdila sličnost ili različitost između pojedinih objekata, potrebno je definirati mjere kojima će se mjeriti sličnost, odnosno različitost. Kao mjeru sličnosti uzima se udaljenost između pojedinih objekata. [7].

Primjer [14]**:**

Tablica 3. prikazuje deset dvodimenzionalnih točaka. Samo promatrajući podatke, mogu se uočiti dvije grupe podataka.

Tablica 3. Skup podataka [14]

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
x_1	-2	-6	-1	11	-1	-46	33	42	32	45
x_2	7	22	1	1	-8	52	40	33	54	39
...	…	…	…	…	…	…	…	…	…	…



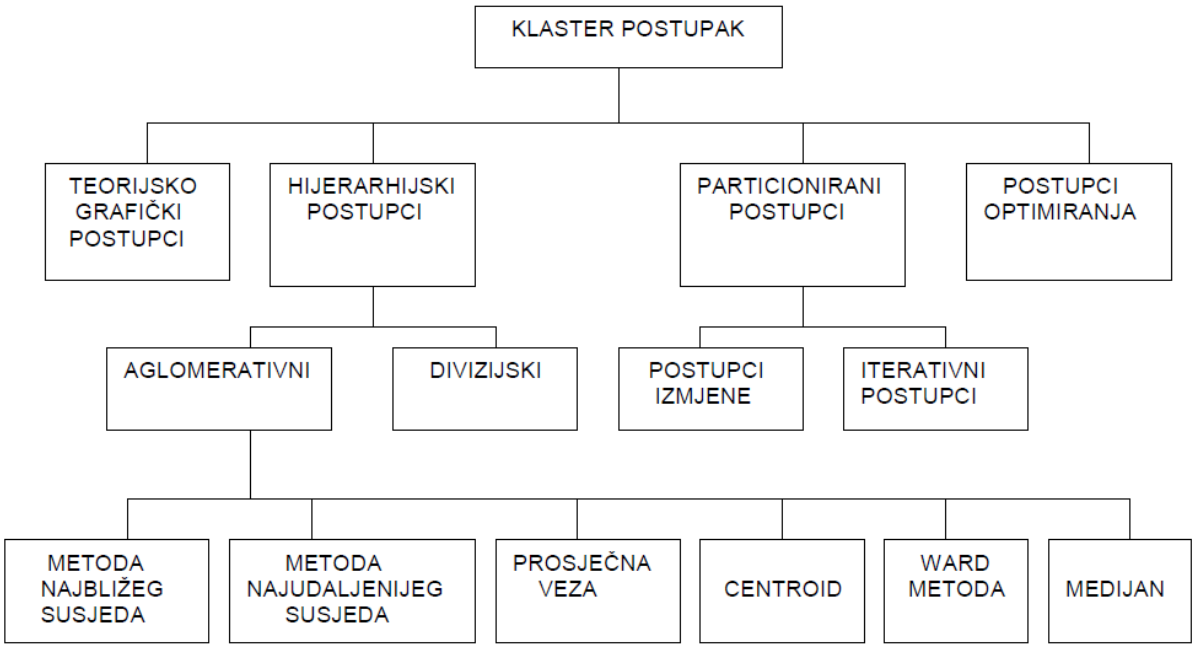
Slika 6. Raspored podataka u koordinatnom sustavu [14]

Jedna grupa podataka je oko točke (0;5) dok je druga grupa podataka oko točke (35;45). Jednostavnim pregledom podataka, dolazi se do zaključka da su podaci podijeljeni u dvije grupe: jedna oko (0;0), a druga oko (35;45) dok standardna devijacija (prosječno odstupanje od srednje vrijednosti) iznosi 10.

Klaster analiza bavi se problemom: za dani skup podatak *U,* odrediti pod skupove (klastere) *C* koji su homogeni i/ili dobro separirani u odnosu na mjerne varijable [21].

Kao što je navedeno u prethodnom odjeljku, konačan rezultat klaster analize je podjela objekata u klastere u skladu s definiranim ciljevima. Nadalje, klaster analiza odgovara na tri temeljna pitanja: kako mjeriti sličnost između objekata, kako formirati klastere i kako utvrditi konačan broj klastera [22].

Prikaz strategija grupiranja tj. postupaka klasteriranja, dan je na slici 7.



Slika 7. Podjela klaster postupaka [23]

Svi postupci sa slike 7 imaju jednake ciljeve koji se odnose na grupiranje ili segmentiranje skupa podataka u pod skupove ili klastere, tako da su podaci unutar svakog klastera usko povezani međusobno u odnosu na podatke u drugim klasterima. Objekt ili podatak može opisan je različitim značajkama ili pak prema njegovim odnosom s drugim objektima tj. podacima. Nadalje, cilj grupiranja može biti organiziranje grupa prema određenoj hijerarhiji. To podrazumijeva uspješno grananje klastera tako da na svakoj razini hijerarhije klasteri unutar jedne hijerarhije su više slični jedni drugima, nego drugim klasterima [21].

Koraci u rješavanju analizom grupiranja su [21]:

1. Odabrati skup podataka
2. Pripremiti varijable za dani problem, ukoliko su varijable numeričke, trebalo bi ih se normalizirati (npr. u rasponu od 0-100)
3. Izabrati prikladnu mjeru sličnosti među podacima za dani problem i tipove varijabli
4. Izabrati prikladnu metodu grupiranja (npr. hijerarhijsko ili particijsko)
5. Izabrati broj klastera za procjenu odabranog tipa grupiranja
6. Izabrati algoritam grupiranja (npr. k-means, k-median)
7. Ocijeniti dobivena rješenja kako bi se vidjela da li imaju neku temeljnu strukturu (evaluirati sa stručnjacima)

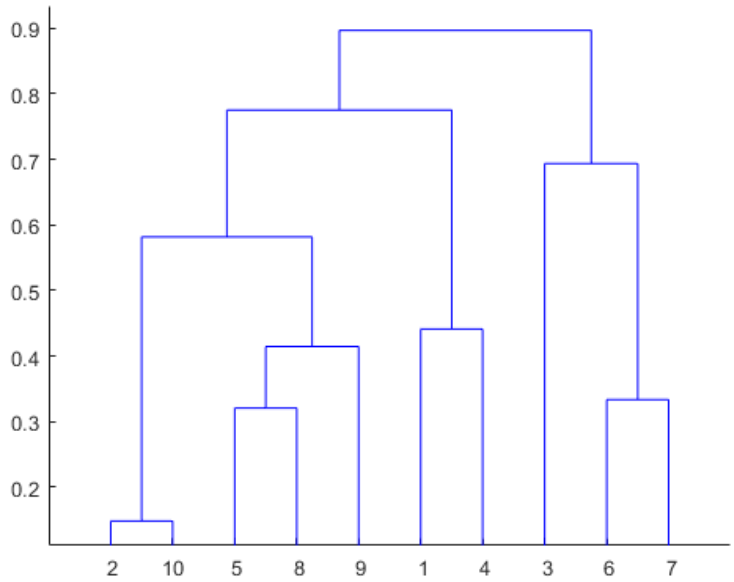
U okviru analize grupiranjem postoji veliki broj različitih algoritama koji odgovaraju na iste probleme.

5.1. Tehnike klaster analize

Iako postoje različite strategije grupiranja, strategije se mogu ugrubo podijeliti na hijerarhijske i ne hijerarhijske postupke. Hijerarhijski pristup kao krajnji rezultat ima dendogram. Dendogram je grafički prikaz klastera u obliku stabla povezivanja koja može nastati na gomilajući (aglomerativan) ili dijeleći (divizijski) način [22].

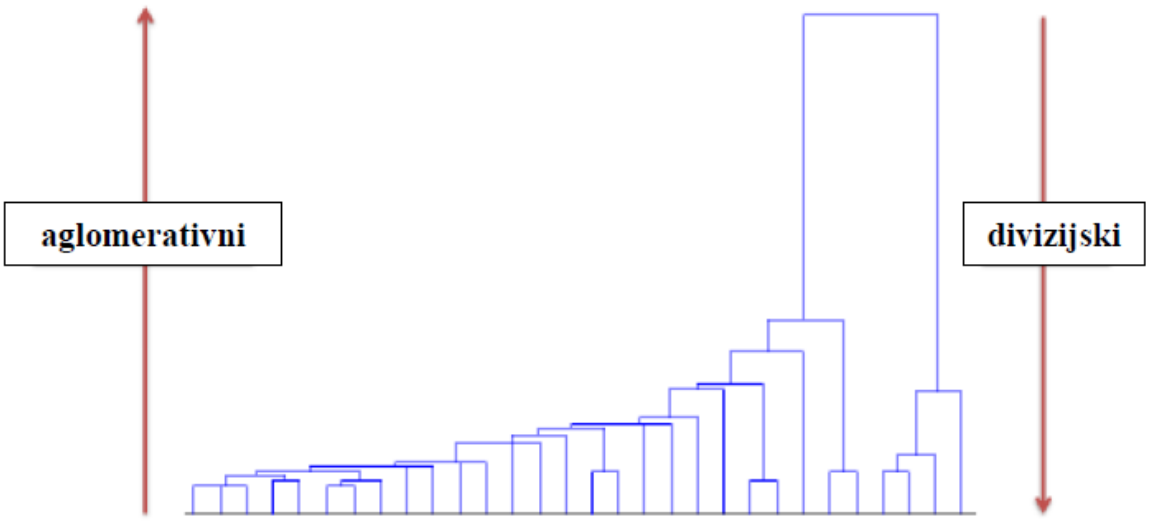
Nehijerarhijski pristup podrazumijeva raščlanjivanje tako da se uzorci mogu kretati iz jedne u drugu grupu u različitim fazama analize. Postupak je takav da se prvo nađe točka grupiranja (centroid) oko koje se nalaze uzorci, a potom se izračunavaju nove točke grupiranja na osnovu prosječne vrijednosti uzoraka [22].

Hijerarhijski pristup temeljen je na ideji da su objekti povezani s objektima koji su im bliži nego s onima koji su im udaljeniji. Objekti formiraju klastere pomoću mjere sličnosti između svakog novog elementa i svih ostalih već prije određenih klastera. Dendogram je stablo koji prikazuje raspored klastera nastalih hijerarhijskim grupiranjem. Sličnost između dva objekta u dendogramu predstavljena je visinom tj. duljinom najnižeg unutarnjeg čvora koji dijele. Ukoliko se dendogram stavi u koordinatni sustav, na *x* osi su raspoređeni objekti, dok su na *y* osi visine na kojoj se pojedini klasteri spajaju [21].



Slika 8.Hijerarhijsko grupiranje [24]

Nadalje, hijerarhijski postupci mogu se podijeliti na aglomerativne i divizijske. Aglomerativni postupci pojedine objekte povezuju u sve veće klastere, dok dijeleći, kao što i samo ime sugerira, polaze od svih objekata koji su udruženi u jedan klaster te ih dijele do pojedinih dijelova [22].



Slika 9. Aglomerativni i divizijski postupak [22]

Aglomerativni hijerarhijski postupci grupiranja su [25]:

a) Potpuna veza
b) Prosječna udaljenosti
c) Centroid
d) Medijan
e) Wardova metoda.

Potpuna veza je postupak gdje se udaljenost između dva klastera A i B definira kao maksimalna udaljenost dvaju pojedinačnih vrijednosti u A i B.

$$D(A,B) = \max\{d(y_i, y_j)\}, \tag{1}$$

za $y_i \in A$ i $y_j \in B$. U svakom se koraku određuju udaljenosti za svaki par vrijednosti te se par s najmanjom udaljenošću spaja u klaster [25]

Prosječna udaljenost (eng**.** *unweighted pair – group method using the average approach - UPGMA)* metoda je gdje se udaljenost dva klastera definira kao prosjek $n_A n_B$ udaljenosti između n_A točaka u A i n_B točaka u B prema formuli:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=!}^{n_A} \sum_{i=1}^{n_B} d(y_i, y_j), \tag{2}$$

U svakom se koraku udružuju dva klastera s najmanjom udaljenosti izračunatoj prema navedenoj formuli [25].

Centroid metoda udaljenosti između dva klastera A i B definira udaljenost kao Euklidsku udaljenost između dva vektora sredina (centroid) [25]:

$$D(A,B) = d(\overline{y_A}, \overline{y_B}) \tag{3}$$

Pri čemu su $\overline{y_A}$i $\overline{y_B}$ vektori sredina za opažanja iz A, odnosno opažanja iz B, a $d(\overline{y_A}, \overline{y_B}) = \sqrt{(\overline{y_A} - \overline{y_B})'(\overline{y_A} - \overline{y_B})}$. U svakom se koraku spajaju dva centroida s najmanjom udaljenošću. Nakon što su dva klastera A i B združena, centroid novog klastera se računa kao vagana aritmetička sredina:

$$\overline{y_{AB}} = \frac{n_A \overline{Y_A} + n_B \overline{Y_B}}{n_A + n_B}. \tag{4}$$

Medijan metoda radi na principu da za polovište pravca koji spaja dva podataka, A i B, određuje točku za računanje novih udaljenosti klastera A, B u odnosu na druge klastere:

$$m_{AB} = \frac{1}{2}(\overline{y_A} + \overline{y_B}). \tag{5}$$

U svakom se koraku klasteri s najmanjom medijalnom udaljenosti spajaju u novi klaster [25].

Wardova metoda, poznata kao inkrementalna suma kvadrata, upotrebljava (kvadrirane) udaljenosti unutar klase i (kvadrirane) udaljenosti između klastera. Ako je AB klaster dobiven kombiniranjem klastera A i B, tada je zbroj udaljenosti unutar klastera (elemenata od centroida):

$$W_A = \sum_{i=1}^{n_A} \sqrt{(y_i - \overline{y_A})'(y_i - \overline{y_A})}, \tag{6}$$

$$W_B = \sum_{i=1}^{n_B} \sqrt{(y_i - \overline{y_B})'(y_i - \overline{y_B})}, \tag{7}$$

$$W_{AB} = \sum_{i=1}^{n_{AB}} \sqrt{(y_i - \overline{y_{AB}})'(y_i - \overline{y_{AB}})}, \quad (8)$$

pri čemu je $\overline{y_{AB}} = \frac{n_A\overline{Y_A}+n_B\overline{Y_B}}{n_A+n_B}$, a n_A, n_B i $n_{AB} = n_A + n_B$, su brojevi točaka u A, B i AB. Budući da su zbrojevi udaljenosti ekvivalentni sumama kvadrata odstupanja točaka klastera od njihovih centroida, označene su s W_A, W_B, W_{AB}. Warodva metoda združuje dva klastera A i B koji minimiziraju prirast u W [25]:

$$I_{AB} = W_{AB} - (W_A + W_B). \quad (9)$$

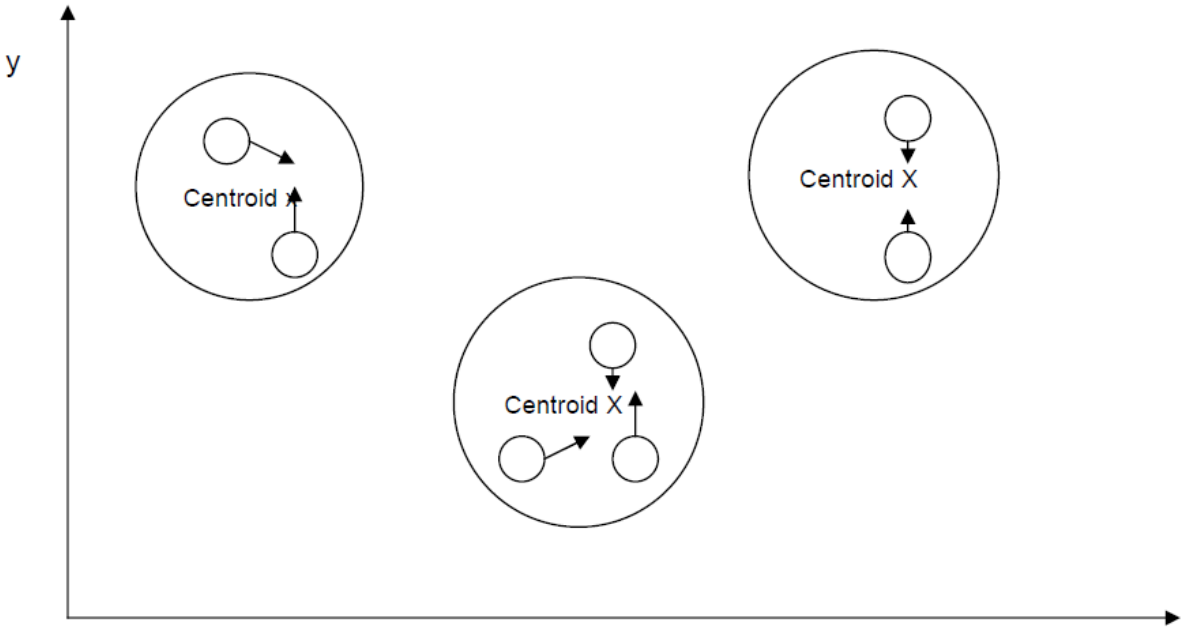
5.2. Metoda k-srednjih vrijednosti

Metoda k-srednjih vrijednosti (eng**.** *k-means)* je iterativna, optimizacijska metoda dijeljenja. Temelji se na podjeli osnovnog uzorka u *k* grupa koristeći koncept središta (centroida). Centroid klastera je srednja vrijednost točaka unutar klastera. Na temelju funkcije udaljenosti, algoritam procjenjuje sličnost elemenata [25],[26]. Metoda *k* – means dozvoljava pomicanje članova iz jednog klastera u drugi, što nije dozvoljeno u hijerarhijskim metodama. Ova metoda je osjetljiva na izbor početnih točaka. Preporučljivo je da se postupak počne ponovno s drugačijim izborom početnih točaka. Ukoliko takav izbor rezultira sasvim drugačijim konačnim klasterima ili ako je konvergencija iznimno spora, može se zaključiti da nema prirodnih klastera podataka [25].

Metoda se može koristiti kao moguća potvrda hijerarhijskog postupka. Članovi se prvo klasteriraju hijerarhijskom metodom, a zatim se centroidi klastera koriste kao početne točke za *k – means* pristup, koji dozvoljava premještaj točaka iz jednog klastera u drugi [25].

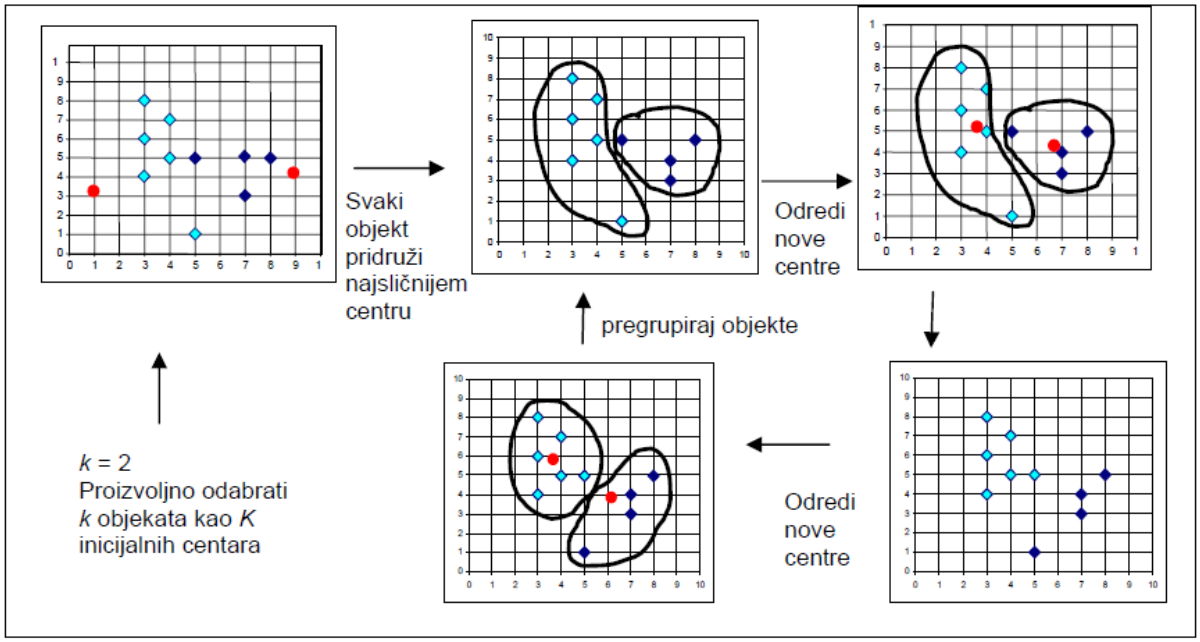
Algoritam *k-means* odvija se na idući način [23]:

1. Proizvoljan odabir *k*-segmenata
2. Određivanje središta (centroida) za svaki od k-segmenata
3. Pridruživanje pomoću funkcije udaljenosti svih elemenata populacije najbližim klasterima
4. Proračun nove vrijednosti središta klastera za svaki klaster pojedinačno kao prosječne vrijednosti objekata sadržanih unutar svakog klastera
5. Ponavljanje sve dok se vrijednosti središta klastera ne prestanu mijenjati tj. sve dok ne postigne konvergencija.



Slika 10 k-means algoritam [23]

Ova metoda se može prikazati slikom 10 koja prikazuje težnju vezivanja elemenata u iterativnim procesima uz centralne vrijednosti klastera.



Slika 11. Princip rada k-means algoritma [9]

Princip rada može se prikazati i slikom 11 gdje je prikazano kako se objekti ili uzorci grupiraju oko *k* objekata koji predstavljaju srednju vrijednost unutar pojedinog klastera [9].

5.3. Mjere udaljenosti između uzoraka

Tri su metode utvrđivanja sličnosti u klaster analizi: mjere korelacije, mjere udaljenosti i mjere udruživanja. Mjere udaljenosti i mjere korelacije primjenjuju se kod numeričkih vrijednosti značajki. Mjere udaljenosti su najčešće upotrebljavanje mjere sličnosti u analizi grupiranja [16].

Kako bi se utvrdila sličnosti ili različitost između pojedinih objekata, potrebno je definirati mjere kojima će se mjeriti sličnost tj. različitost [9].

Metoda grupiranja omogućuje tri različite matematička pristupa za mjerenje udaljenosti između veličina x i y.

1) **Euklidska udaljenost** [23]:

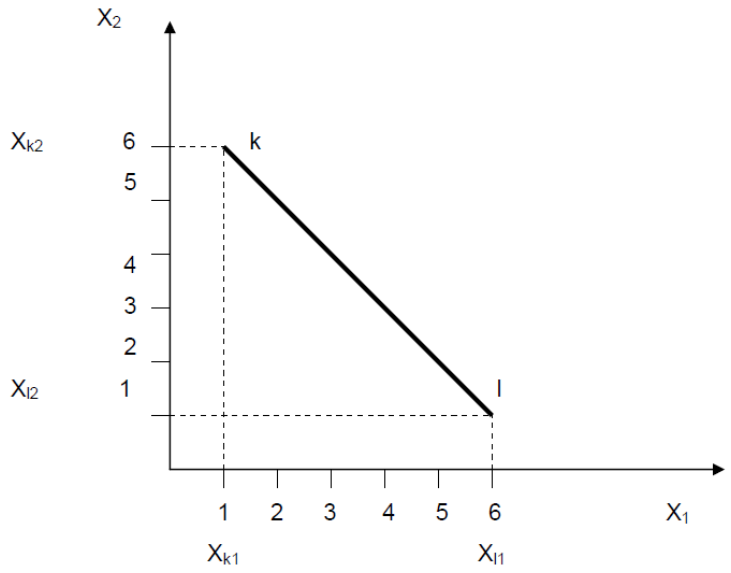
Kod Euklidske metrike, udaljenost između dvije točke prema njihovoj najkraćoj udaljenosti izračunava se formulom:

$$d(i,j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2} \tag{10}$$

Gdje je:

$d(i,j)$ – udaljenost točaka i, j

x_{im}, x_{jm} – koordinate točaka k, I



Slika 12. Euklidska udaljenost [23]

Slika 12 prikazuje euklidsku udaljenost između točaka k i I.

2) Euklidska težinska udaljenost [9]:

$$d(i,j) = \sqrt{\sum_{m=1}^{p} w_m(x_{im} - x_{jm})^2} \tag{11}$$

Gdje je w_m težina m-te varijable (veća težina je pridijeljena važnijim varijablama)

3) Manhattan ili blokovska udaljenost [23]:

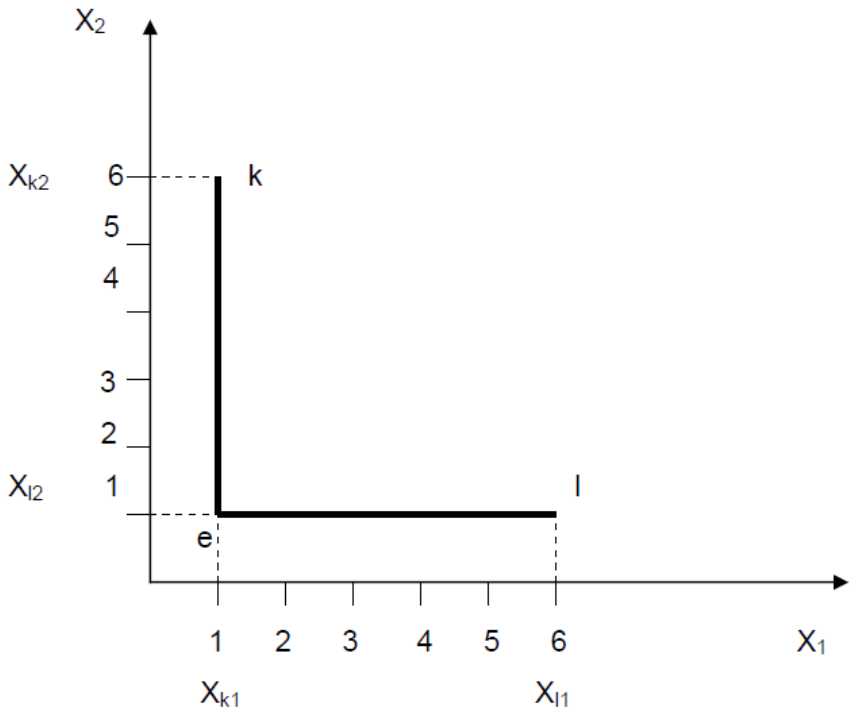
Kod blokovske metrike, udaljenost između dvije točke se izračunava kao suma apsolutnih udaljenosti između točaka pomoću formule:

$$d(i,j) = \sum_{m=1}^{p} |x_{im} - x_{jm}| \tag{12}$$

Gdje je:

$d(i,j)$ – udaljenost točaka i, j

x_{im}, x_{jm} – koordinate točaka k, I



Slika 13. Blokovska udaljenost

Služi za rješavanje problema najkraćeg puta između dviju točaka u grafu koji simulira mrežu ulica koje su međusobno ili paralelne ili okomite i spajaju se u čvorovima, po čemu je i ova metoda dobila svoj naziv [27]

.

6. REGRESIJA I KLASIFIKACIJA

Strojno učenje, kao što je navedeno u odjeljku 2.2, može se podijeliti na nadzirano i nenadzirano učenje, ali postoje i inherentne razlike u tim učenjima na temelju oblika njihovog izlaza. Promatrajući ih na ovaj način, dvije najčešće vrste metoda strojnog učenja su regresija i klasifikacija.

6.1. Regresija

Zadatak modela predviđanja regresijskom funkcijom je aproksimiranje nepoznate funkcije *f* na temelju ulazne varijable *x* do kontinuirane izlazne varijable *y*. Ono je metoda modeliranja ciljne vrijednosti na temelju nezavisnih atributa. Na temelju poznatih vrijednosti atributa zadanih podataka određuju se parametri modela, a zatim se na temelju parametara modela određuju nepoznate vrijednosti atributa novih podataka. Koristi se uglavnom za predviđanje i pronalaženje uzročno – posljedičnog odnosa između nezavisnih varijabli [9] [28].

Osnovne vrste regresije su [9]:

1. Linearna regresija
2. Višestruka regresija
3. Logistička regresija
4. Nelinearna regresija

6.1.1. Linearna regresija

Linearna regresija je najjednostavniji oblik regresije u kojoj se vrijednost zavisne slučajne varijable *Y* određuje kao linearna funkcija prediktivne varijable *X* [9].

$$Y = \alpha X + \beta$$

Gdje se regresijski koeficijenti α i β određuju metodom najmanjih kvadrata prema formuli [9]:

$$\beta = \frac{\sum_{i=1}^{u}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{u}(x_i - \bar{x})^2}, \alpha = \bar{y} - \beta\bar{x}.$$

Vrijednost predikcije predstavlja os X, dok Y os predstavlja vrijednost prediktora. Zadaća regresije je pronaći liniju između svake vrijednosti prediktora i predikcije, a linija gdje je udaljenost između vrijednosti osi X i Y najmanja, izabere se model predikcije. Linearna funkcija se može prikazati i na ovaj način [11]:

$$Predikcija = \alpha * Prediktor + \beta$$

6.1.2. Višestruka regresija

Višestruka regresija je proširenje linearne regresije. Za razliku od linearne regresije, uključuje više od jedne prediktivne varijable. Zavisna varijabla *Y* određuje se kao linearna funkcija višedimenzionalnog vektora prediktivnih varijabli. Općeniti oblik višestruke regresije je[9]:

$$Y = \alpha_1 X_1 + \cdots + \alpha_n X_n + \beta$$

Gdje se parametri α_i i β određuju metodom najmanjih kvadrata.

6.1.3. Logistička regresija

Logistička regresija je proširenje linearne regresije uz ograničenje da interval vrijednosti koje zavisna varijabla *Y* može poprimiti je [0,1]. Logistička funkcija time modelira vjerojatnost kao linearnu funkciju prediktivne varijable, a općeniti oblik je [9]:

$$Y = \frac{1}{1 + e^{-X}}.$$

6.1.4. Nelinearna regresija

Ukoliko varijable u modelu ne pokazuju linearnu ovisnost, već polinomnu ovisnost, tada je riječ o nelinearnoj regresiji. Potrebno je polinomnu ovisnost svesti na linearan oblik te je dalje tretirati kao linearnu regresiju. Neka je zadana funkcija međuovisnosti varijabli:

$Y = \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3 + \beta$

Funkcija se prvo treba svesti na linearni oblik uvođenjem novih varijabli:

$X_1 = X^1, X_2 = X^2, X_3 = X^3$, a zatim se rješava jednadžba

$Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \beta$ koja ima oblik višestruke regresije. Nepoznati parametri α i β se određuju metodom najmanjih kvadrata.

6.2. Klasifikacija

Dok je kod regresije izlazna varijabla kontinuirana vrijednost, zadatak klasifikacije je procjena klasa. Klasifikacijski prediktivni model ima za zadatak aproksimirati mapiranu funkciju *f* na temelju ulazne varijable *x,* kako bi se diskretizirala izlazna varijabla *y*. Izlazne varijable se najčešće nazivaju kategorije ili klase. Funkcija mapiranja procjenjuje klasu ili kategoriju za zadani zadatak. Jednostavan primjer klasifikacije je e mail poruka koja može biti klasificirana u skupine '*spam'* ili '*nije spam'* [29].

Slika 14. prikazuje razliku između klasifikacije i regresije s obzirom na ciljnu varijablu *y*.

Regresija

x1	x2	y
0.5	3.4	2.3
1.0	2.3	1.5
0.8	0.2	3.3
5.1	4.1	4.0

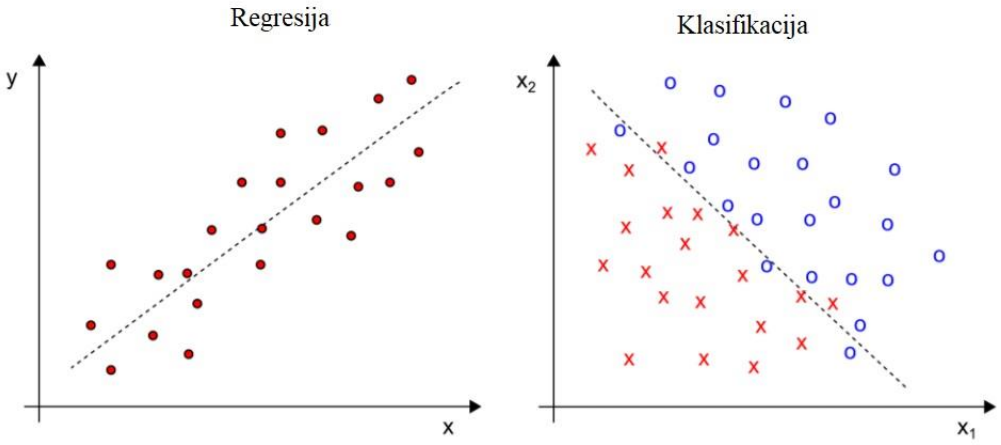
Klasifikacija

x1	x2	y
0.5	3.4	x
1.0	2.3	o
0.8	0.2	o
5.1	4.1	x

Slika 14. Razlika između regresije i klasifikacije [30]

Izlazna varijabla regresije je kontinuirana varijabla koja može poprimiti vrijednost realnog skupa brojeva$y \in \mathbb{R}$. Nasuprot tome, izlazna varijabla klasifikacije je klasa, u ovom slučaju $y \in \{x, o\}$ [30].

Slika 15. prikazuje izlazne varijable klasifikacije i regresije.



Slika 15. Izlazne varijable klasifikacije i regresije

Model učenja regresijom ima jednu zavisnu varijablu te postoji odnos između zavisne i nezavisne varijable, te se nastoji pogoditi vrijednost svake zavisne varijable, dok model učenja klasifikacijom nastoji klasificirati zavisnu varijable u dvije ili više klasa, ovisno o zadanim postavkama [28].

Kod klasifikacije i regresije potrebno je podijeliti ulazni skup podataka na dvije skupine trening i test skup. Trening skup podataka se koristi kako bi se izgradio model, dok se nad testnom skupu podataka provjerava točnost modela. Testni skup podataka uobičajeno iznosi oko 20% - 30% ukupnog skupa podataka. Testni skup podataka model nikad neće vidjeti tj. on se neće koristiti za izgradnju modela. Ovakav skup je nužno izdvojiti kako bi se nad njime ispitala točnost izgrađenog modela [9].

Neki od algoritama koji se mogu koristiti za klasifikaciju podataka su[9]:

- Stabla odlučivanja (eng. *Decision Tree)*
- Naivni Bayes-ov klasifikator (eng. *Naive Bayes)*
- Neuronske mreže – prenošenje informacija unatrag (eng. *Backpropagation)*
- Metoda *k* – najbližih susjeda
- Algoritam SVM (eng. *Support Vector Machines).*

Navedene algoritme moguće je podijeliti u dvije skupine[9], [31]:

1. Lijeni algoritmi učenja (eng. *Lazy Learner)* – klasifikator koji samo pohrani trening skup bez da uči na njemu, sve dok ne dobije testni skup. Kada dobije testni skup, pokreće proces klasifikacije. Potrebno mu je manje vremena da nauči, ali više vremena da klasificira podatke. Primjer za lijene algoritme učenja je metoda *k* – najbližih susjeda.
2. Željni algoritmi učenja (eng. *Eager Learner)* – za razliku od lijenih, željni algoritam ne čeka testni skup da bi započeo s učenjem, već čim primi trening skup započinje s učenjem. Primjer su stabla odlučivanja i naivni Bayes-ov klasifikator.

Lijeni algoritmi učenja su brži za stvaranje, ali sporiji prilikom klasifikacije novih uzoraka, dok su željni algoritmi brži prilikom susreta s novim uzorkom jer već imaju izrađen klasifikacijski model.

Metode klasifikacije mogu se usporediti prema sljedećim kriterijima [9]:

- Točnost predviđanja – sposobnost modela da točno predvidi klasu za testni uzorak
- Brzina – potrebno vrijeme za izgradnju i testiranje modela
- Razumljivost – razumljivost izgrađenog modela
- Robusnost – sposobnost modela da točno predvidi razred za nepoznate uzorke koji sadrže šum ili atribute koji imaju nepoznate vrijednosti
- Skalabilnost – sposobnost efikasne izgradnje modela koji sadrže velike količine podataka.

6.3. Tumačenje rezultata

Uspješnost modela učenja regresijom mjeri se koeficijentom korelacije. Koeficijent korelacije je broj između -1 i +1 koji mjeri stupanj povezanosti između zavisnog atributa i atributa za procjenu. Kad je njegova vrijednost jednaka nuli, zavisna i nezavisna varijabla su potpuno neovisne tj. između njih ne postoji nikakva zavisnost, a ako je jednaka 1, tada između varijabli postoji funkcionalna zavisnost i jedna se varijabla može izračunati preko druge. Kada je vrijednost koeficijenta +1 veza između varijabli je upravo razmjerna, a ukoliko je -1 tada je zavisnost u obliku obrnute razmjernosti [32] [33].

7. PRIMJENA METODA STROJNOG UČENJA U PROIZVODNOM POGONU

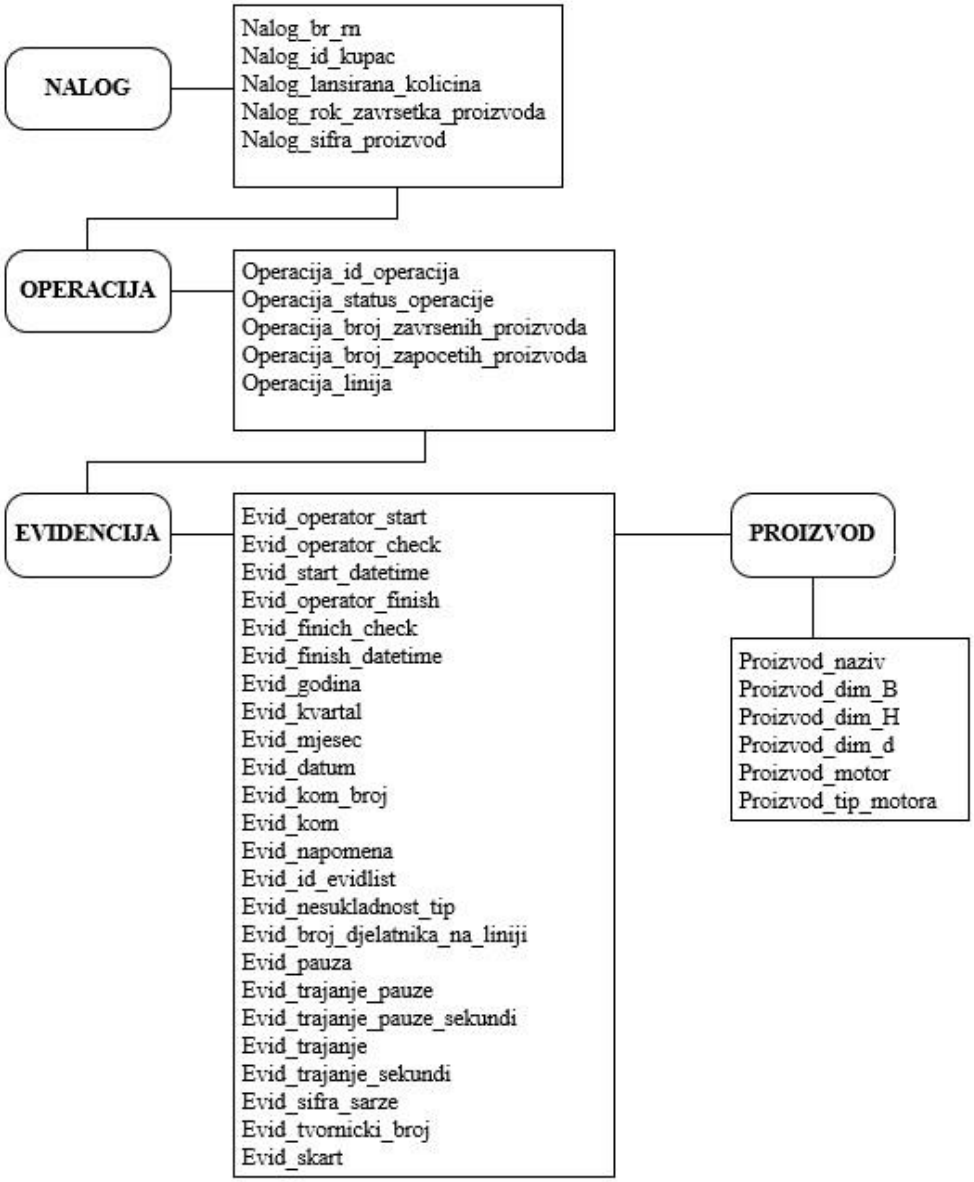
Podaci odabrani za analizu sadrže podatke o evidenciji proizvodnje u linijskoj proizvodnji protupožarnih zaklopki i regulatora varijabilnog protoka za operaciju montaže. Ulazni set podataka prikupljen je u periodu od 2016. do 2018. godine te sadrži 42 atributa i 87 541 podataka.

7.1. PROCES PRIPREME PODATAKA

U ovom poglavlju prikazani su svi procesi pripreme podataka, tj. dokumentiranje i statistika ulaznog skupa podataka, izrada novih atributa i transformacija podataka.

7.1.1. Dokumentiranje dostupnog skupa podataka

Vizualni prikaz atributa dan je na slici 16.



Slika 16. Vizualni prikaz atributa

Tablica 4. Popis atributa i njihov opis, tip podatka

NAZIV ATRIBUTA	**TIP PODATKA**	**OPIS**
Id	*Integer*	identifikacijski broj
Nalog_br_rn	*Integer*	broj radnog naloga za izradu proizvoda
Nalog_id_kupac	*Nominal*	identifikacijski broj kupca za koga je proizvod

Nalog_lansirana_kolicina	*Real*	ukupna količina proizvoda koju je potrebno izraditi
Nalog_rok_zavrsetka_proizvodnje	*Date time*	datum i vrijeme do kada je potrebno izvršiti proizvodnju
Nalog_sifra_proizvod	*Nominal*	šifra proizvoda kojeg je potrebno izraditi
Operacija_id_operacije	*Integer*	identifikacijski broj operacije koji se izvodi nad proizvodom
Operacija_status_operacije	*Integer*	atribut može poprimiti 3 različite vrijednosti koje opisuju u koji je status operacije nad proizvodom: 1 - čeka se da proizvod krene na operaciju 2 - proizvod je lansiran na operaciju 3 – proizvod je završio s operacijom.
Operacija_broj_zapocetih_proizvoda	*Real*	atribut označava koliko je ukupno proizvoda s radnog naloga započelo proizvodnju
Operacija_broj_zavrsenih_proizvoda	*Real*	atribut označava koliko je ukupno proizvoda s radnog naloga završilo proizvodnju
Operacija_linija	*Nominal*	atribut može poprimiti dvije vrijednosti: NGLIN7 – označava operacijsku liniju za pravokutne proizvode NGLIN7A – označava operacijsku liniju za cilindrične proizvode
Evid_operator_start	*Integer*	identifikacijski broj operatora koji je započeo operaciju nad proizvodom
Evid_start_check	*Integer*	može poprimiti vrijednosti 0 i 1, gdje 0 označava da nije ni započeta operacija nad proizvodom, a 1

		označava da je proizvod prošao '*check point'* i započeo sa operacijom
Evid_start_datetime	*Date time*	označava datum i vrijeme početka proizvodnje
Evid_operator_finish	*Integer*	identifikacijski broj djelatnika koji je završio i preuzeo proizvod kada se završila operacija nad proizvodom
Evid_ finish_check	*Integer*	može poprimiti vrijednosti 0 i 1, gdje 0 označava da proizvod nije dovršen, a 1 označava da je proizvod prošao '*finish point'* i završena je proizvodnja nad njime
Evid_finish_datetime	*Date time*	označava datum i vrijeme završetka proizvodnje
Evid_godina	*Integer*	godina proizvodnje
Evid_kvartal	*Integer*	kvartal proizvodnje
Evid_mjesec	*Integer*	mjesec proizvodnje
Evid_datum	*Date*	datum proizvodnje
Evid_kom_broj	*Integer*	označava broj komada koji se izradio, uvijek 1
Evid_kom	*Integer*	označava ukupan broj izrađenih proizvoda
Evid_napomena	*Nominal*	moguće napomene vezane za izradu proizvoda
Evid_id_evidlist	*Integer*	identifikacijski broj liste radnika na dan proizvodnje
Evid_nesukladnost_tip	*Integer*	klasificirana nesukladnost koja se pojavila tokom proizvodnje, a može biti: 1 – Škart proizvod 2 – Smanjena kvaliteta proizvoda 3 – Kašnjenje proizvodnje

		4 – Ostalo 5 – Zastoj stroja
Evid_broj_djelatnika_na_liniji	*Integer*	broj djelatnika na liniji tokom proizvodnje
Evid_pauza	*Integer*	poprima vrijednosti 0 i 1, gdje 1 označava da je bila pauza za vrijeme proizvodnje proizvoda, a 0 da nije bilo pauze
Evid_trajanje_pauza	*Time*	trajanje pauze u vremenskom obliku.
Evid_trajanje_pauza_sekundi	*Real*	trajanje pauze prikazano u sekundama
Evid_trajanje	*Time*	ukupno trajanje proizvodnje s uključenom pauzom u vremenskom obliku
Evid_trajanje_sekundi	*Real*	ukupno trajanje proizvodnje s uključenom pauzom u sekundama
Evid_sifra_sarze	*Nominal*	šifra šarže iz koje je izrađen proizvod.
Evid_tvornicki_broj	*Nominal*	tvornički identifikacijski broj
Evid_skart	*Integer*	proizvod koji nije imao potrebne karakteristike i kvalitetu, može poprimiti vrijednosti 0 i 1, gdje 1 označava da proizvod nije zadovoljio te se odnosi u otpad, a 0 da je zadovoljio sve potrebne tehničke karakteristike i kvalitetu.
Proizvod_naziv	*Nominal*	naziv proizvoda
Proizvod_dim_B	*Real*	širina proizvoda
Proizvod_dim_H	*Real*	visina proizvoda
Proizvod_dim_L	*Real*	duljina proizvoda
Proizvod_dim_d	*Real*	promjer proizvoda

Proizvod_motor	*Integer*	poprima vrijednosti 0 i 1, gdje 1 označava da proizvod sadrži motor, a 0 da ne sadrži
Proizvod_tip_motora	*nominal*	naziv ugrađenog motora

Dio podataka prikazan je na slici 17.

	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
1	evid_trajanje_sekundi	evid_sifra_sarze	evid_tvornicki_broj	evid_skart	proizvod_naziv	proizvod_dim_B	proizvod_dim_H	proizvod_dim_L	proizvod_dim_d	motor	tip_motora
2	1945,0		4.08202E+20		FD-40-1000x750x350-M24-S	1000,0	750,0	350,0		1	Belimo BF - T
3	1837,0		4.08202E+20		FD-25-800x450x350-M230-S	800,0	450,0	350,0		1	Belimo BFL - T
4	3873,0		4.08202E+20		FD-40-600x300x350-M230-S	600,0	300,0	350,0		1	Belimo BFL - T
5	1247,0		4.08202E+20		FD-40-700x400x350-M230-S	700,0	400,0	350,0		1	Belimo BFL - T
6	1087,0		4.08202E+20		FD-25-200x200x350-M230-S	200,0	200,0	350,0		1	Belimo BFL - T
7			4.08202E+20	1	FD-25-250x200x350-M230-S	250,0	200,0	350,0		1	Belimo BFL - T
8			4.08202E+20	1	FD-25-250x200x350-M230-S	250,0	200,0	350,0		1	Belimo BFL - T
9			4.08202E+20	1	FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
10			4.08202E+20	1	FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
11			4.08202E+20	1	FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
12			4.08202E+20	1	FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
13			4.08202E+20	1	FD-25-300x200x350-M230-S	300,0	200,0	350,0		1	Belimo BFL - T
14			4.08202E+20	1	FD-25-650x250x350-M230-S	650,0	250,0	350,0		1	Belimo BFL - T
15			4.08202E+20	1	FD-25-250x250x350-M230-S	250,0	250,0	350,0		1	Belimo BFL - T
16	1192,0		4.08202E+20		FD-25-250x200x350-M230-S	250,0	200,0	350,0		1	Belimo BFL - T
17	1124,0		4.08202E+20		FD-25-250x200x350-M230-S	250,0	200,0	350,0		1	Belimo BFL - T
18	1134,0		4.08202E+20		FD-25-250x250x350-M230-S	250,0	250,0	350,0		1	Belimo BFL - T
19	1236,0		4.08202E+20		FD-25-300x200x350-M230-S	300,0	200,0	350,0		1	Belimo BFL - T
20	1174,0		4.08202E+20		FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
21	1174,0		4.08202E+20		FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
22	1184,0		4.08202E+20		FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
23	1060,0		4.08202E+20		FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
24	1135,0		4.08202E+20		FD-25-350x250x350-M230-S	350,0	250,0	350,0		1	Belimo BFL - T
25	1232,0		4.08202E+20	1	FD-25-650x250x350-M230-S	650,0	250,0	350,0		1	Belimo BFL - T
26	1284,0		4.08202E+20		FD-40-1000x650x350-M24-S	1000,0	650,0	350,0		1	Belimo BFN - T
27	1071,0		4.08202E+20		FD-40-1000x700x350-M24-S	1000,0	700,0	350,0		1	Belimo BFN - T

Slika 17. Prikaz dijela podataka

7.1.2. Statistika podataka

Tablica 5. prikazuje statističke podatke ulaznih podataka prije pred procesuiranja podataka.

Tablica 5. Maksimalna, minimalna i prosječna vrijednost podataka

NAZIV ATRIBUTA	**MINIMALNA VRIJEDNOSTI**	**MAKSIMALNA VRIJEDNOST**	**PROSJEČNA VRIJEDNOST**
Nalog_lansirana_kolicina	1	575	46,9
Nalog_rok_zavrsetka_proizvodnje	21.04.2004.	17.11.2020.	-
Operacija_status_operacije	2	4	2,99
Operacija_broj_zapocetih_proizvoda	1	575	47
Operacija_broj_zavrsenih_proizvoda	0	575	46,89
Operacija_linija	NGLIN7 (26989)	NGLIN7A (60552)	-
Evid_start_datetime	04.08.2016.	16.10.2018.	-
Evid_finish_datetime	04.08.2016.	16.10.2018.	-
Evid_godina	2016	2018	2017,5
Evid_kvartal	1	4	2,52

Evid_mjesec	1	12	6,55
Evid_datum	04.08.2016.	16.10.2018.	-
Evid_kom	1	575	23,99
Evid_broj_djelatnika_na_liniji	0	332	64,83
Evid_trajanje_pauza_sekundi	2	2809735	67479,57
Evid_trajanje_sekundi	-1069219	1120879	13772,36
Proizvod_dim_B	0	1500	232,65
Proizvod_dim_H	0	800	150,15
Proizvod_dim_L	0	600	124,11
Proizvod_dim_d	100	800	224,48
Proizvod_motor	0	1	0,993

Različitost podataka kod atributa sa tekstualnim tipom podataka prikazani su u tablici 6.

Tablica 6. Različitost tekstualnih podataka

NAZIV ATRIBUTA	**OPIS**
Nalog_id_kupac	216 različitih kupaca
Nalog_sifra_proizvoda	3194 različitih šifra proizvoda
Evid_napomena	28 različitih napomena
Evid_šifra_šarže	110 različitih šifra šarže
Evid_tvornički_broj	86 808 različitih zapisa
Proizvod_naziv	3211 različitih proizvoda
Proizvod_tip_motora	55 različitih tipova motora

Atributi sa nedostajućim vrijednostima prikazani su u tablici 7.

Tablica 7. Atributi sa nedostajućim vrijednostima

NAZIV ATRIBUTA	**BROJ NEDOSTAJUĆIH VRIJEDNOSTI**
Nalog_sifra_proizvod	543
Evid_operator_finish	838
Evid_finish_check	795
Evid_finish_datetime	795
Evid_godina	795
Evid_kvartal	795

Evid_mjesec	795
Evid_napomena	87 118
Evid_nesukladnost_tip	85 576
Evid_pauza	82 237
Evid_trajanje_pauza_sekundi	82 237
Evid_trajanje	795
Evid_trajanje_sekundi	795
Evid_sifra_sarze	5 460
Evid_skart	87 399
Proizvod_dim_B	24 061
Proizvod_dim_H	24 061
Proizvod_dim_L	9 841
Proizvod_motor	607
Proizvod_tip_motora	607

7.1.3. Izrada novih atribut

Kao što je navedeno u odjeljku 4. priprema podataka oduzima 90 % vremena dubinske analize podataka. Potrebno je odrediti potrebne podatke, transformirati ih i vrednovati.

Prvoj verziji podataka dodane su novi stupci kako bi se poboljšala analiza i rezultati. Dodan je novi stupac *'Evid_trajanje_sekundi'* i *'Evid_trajanje_pauza_sekundi'* koji umjesto u vremenskom obliku, podatke prikazuje u sekundama, što je bitno budući da je trajanje varijabla koja se nastoji predvidjeti. (slika 18).

evid_trajanje_pauza	evid_trajanje_pauza_sekundi	evid_trajanje	evid_trajanje_sekundi
12:07:44 AM CET	464	12:13:53 AM ...	833
12:07:45 AM CET	465	10:06:22 PM ...	252382
12:07:52 AM CET	472	3:38:33 AM C...	13113
12:07:52 AM CET	472	12:36:39 AM ...	2199
12:07:55 AM CET	475	12:33:15 AM ...	1995
12:07:56 AM CET	476	12:56:13 AM ...	3373
12:07:58 AM CET	478	12:14:17 AM ...	857
12:07:58 AM CET	478	12:15:10 AM ...	910
12:08:02 AM CET	482	12:29:43 AM ...	1783
12:08:06 AM CET	486	12:34:03 AM ...	2043
12:08:07 AM CET	487	12:43:09 AM ...	2589

Slika 18. Prikaz atributa evid_trajanje_procesa_sekundi

Nadalje, dodan je stupac *'Proizvod_motor'* koji može poprimiti vrijednosti 0 i 1, gdje 1 označava da proizvod ima ugrađeni motor, a 0 da nema. Vrijednosti u novom stupcu su povezani sa stupcem *'Proizvod_tip_motora'*, kao što je prikazano na slici 19.

proizvod_motor	proizvod_tip_motora
1	M230
1	Motor LMV-D3-MF
1	M230
1	M230
1	M230
1	M230
1	M230
1	M230

Slika 19. Prikaz atributa proizvod_motor

Dodani su i stupci *'Evid_godina', 'Evid_kvartal', 'Evid_mjesec'*, koji se popunjavaju na temelju stupca *Evid_datum'*. Ovi atributi su dodani kako bi se dodatno mogla napraviti analiza proizvodnje s obzirom na vrijeme u godini (godišnji odmori, praznici itd.) (slika 20).

evid_godina	evid_kvartal	evid_mjesec	evid_datum
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016
2016	3	8	Aug 4, 2016

Slika 20. Prikaz vremenskih atributa

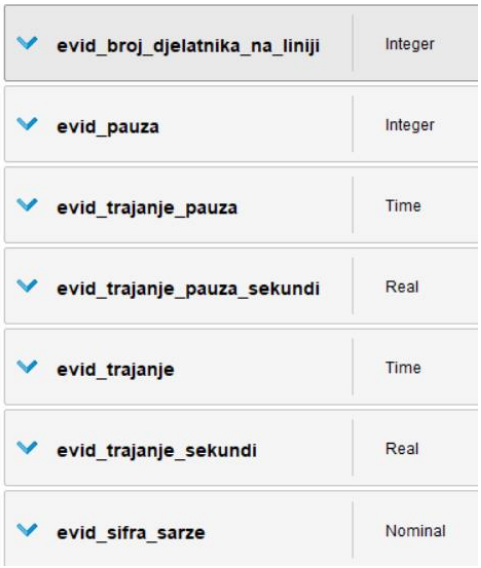
7.1.4. Transformacija podataka

Podatke je potrebno transformirati u prikladan oblik. U odjeljku 7.1.1. prikazani su atributi i tipovi podataka u koje su transformirani. Tablica 8. prikazuje tipove podataka i opis istih. Definiranjem tipa podatka, specificira se tip vrijednosti koji je dopušten za atribut.

Tablica 8. Tip i opis podatka [34]

Tip podatka	Opis
Binominal	Točno dvije vrijednosti (npr. da/ne, točno/netočno itd.)
Date	Datum bez vremena (npr. 09.11.2018.)
Date_time	Datum i vrijeme (npr. 09.11.2018. 21:37)
Integer	Cijeli broj (npr. -5, 23, 100 000 000)
Nominal	Različite tekstualne vrijednosti, uključujući polinominalne i binominalne
Numeric	Različite numeričke vrijednosti, uključujući datum, vrijeme, integer i brojeve
Polynominal	Različite povezane (eng. *string*) vrijednosti (npr. crvena, zelena, plava, žuta)
Real	Decimalni broj (npr. 22.14, -0.05)
Text	Tekstualne vrijednosti
Time	Vrijeme bez datuma (npr. 21:41)

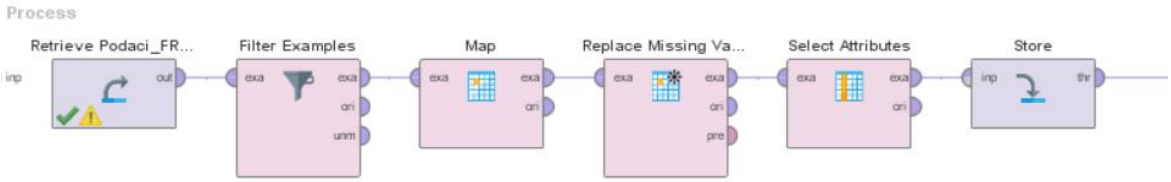
Prikaz djela atributa i njihovih tipova podataka prikazani su na slici 21.



Slika 21. Prikaz atributa i tipa podatka

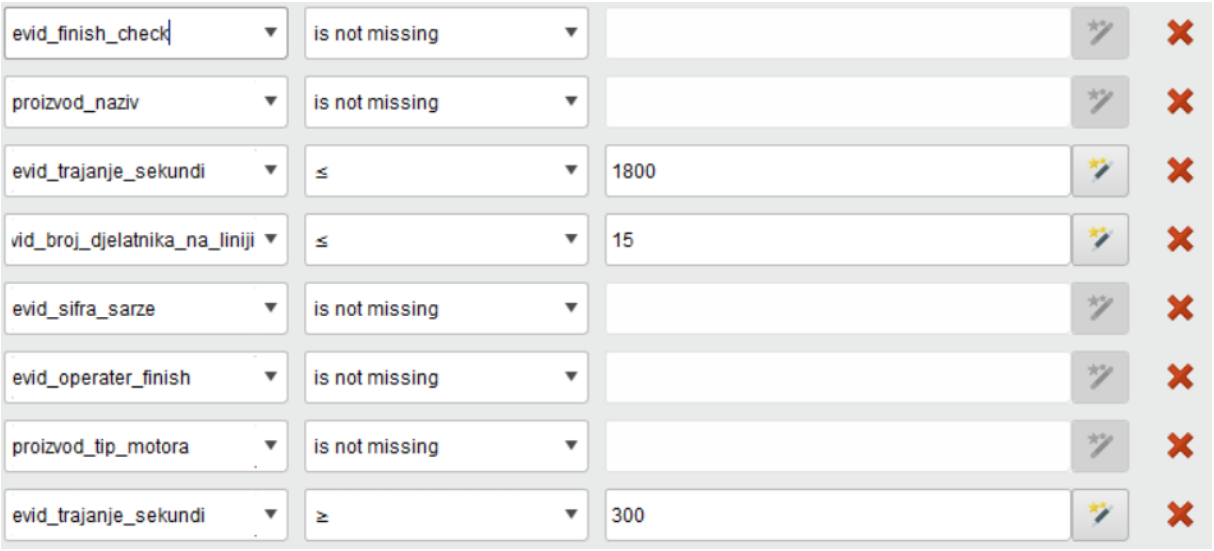
7.1.5. Priprema podataka za proces dubinske analize

Slika 22. prikazuje proces pripreme podataka. Nakon što su dodani novi atributi i nakon što je obavljena transformacija podataka, pomoću operatora *Retrieve* dohvaćaju se ulazni podaci. Ovaj operator ima mogućnost pristupanja podacima koji se nalaze u bazama podataka i unošenja istih u proces.



Slika 22. Prikaz procesa pripreme podataka

Operator *Filter examples* omogućuje filtriranje podataka prema zahtjevu korisnika. Filtrirani podaci prikazani su na slici 23.



Slika 23. Operator Filter Examples

Pomoću ovog operatora , iz procesa su uklonjeni:

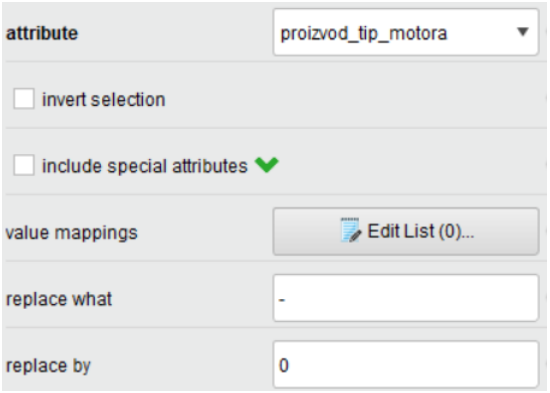
1. Podaci koji nisu dovršeni (*'evid_finish_check' is not missing*) – 795 podataka
2. Podaci kojima nedostaje naziv proizvoda *(proizvod_naziv is not missing)*
3. Podaci kojima nedostaje šifra šarže *(evid_sifra_sarze is not missing)* – 5 460 podataka
4. Podaci kojima nedostaje zapis o identifikacijskom broju djelatnika koji je preuzeo i završio proizvod (*evid_operator_finish)* – 838 podataka
5. Podaci kojima nedostaje zapis o tipu ugrađenog motora (*proizvod_tip_motora)* – 607 podataka

Na ovaj način, uklonjeni su svi podaci s kojima daljnja analiza ne bi bila moguća, budući da nisu dovršeni proizvodi, nemaju motor, ne zna se iz koje šarže su napravljeni ili o kojem se proizvodu radi.

Atribut *'evid_trajanje_sekundi'* koji daje podatak koliko je dugo trajao proces izrade postavljen je u granice od 300 do 1 800 sekundi, tj. analizirani su proizvodi čiji je proces trajao između 5 minuta i 30 minuta.

Atribut *'Evid_broj_djelatnika_na_liniji'* daje informaciju koliko je djelatnika bilo na liniji prilikom izrade proizvoda, a pomoću ovog operatora uzeti su u obzir samo podaci koji imaju manje od 15 djelatnika na proizvodnoj liniji.

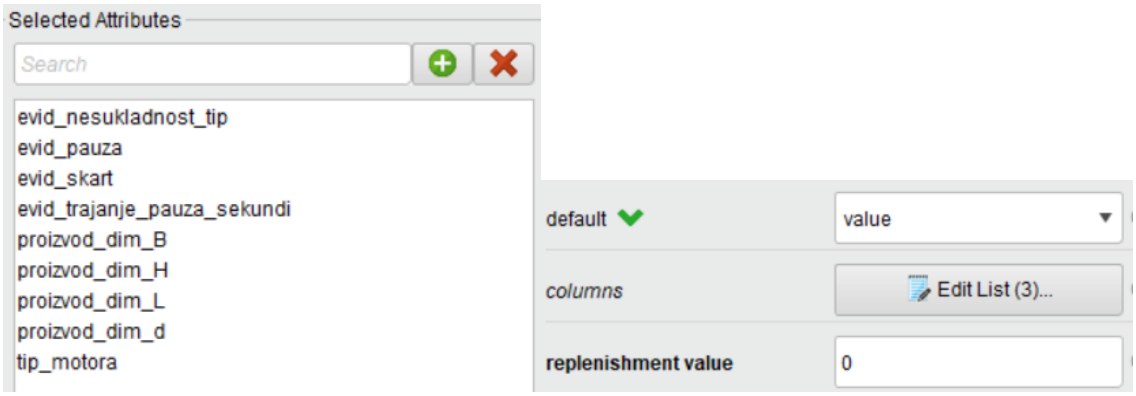
Operator *Map* omogućava mapiranje specificiranih vrijednosti određenih atributa u novu vrijednost. Može se koristiti i za numeričke i nominalne vrijednosti.



Slika 24. Mapiranje atributa

Kao što prikazuje slika 24. pomoću ovog operatora, atribut '*proizvod_tip_motora'* koji je imao nedostajale vrijednosti koje su označene s oznakom '-', mapirane su tj. transformirane u oznaku '0' kako bi se i nad tim podacima mogla provoditi analiza. U slučaju da nije zamijenjeno, program bi prethodnu oznaku čitao kao određeni tip motor. Na ovaj način, oznaka ’0' će se čitati kao da nedostaje motor.

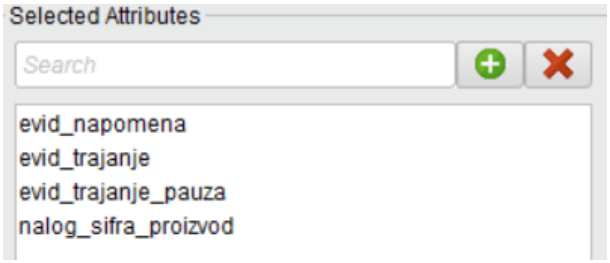
Operator *Replace missing values* zamjenjuje nedostajuće vrijednosti određenog atributa s točno određenom vrijednošću. Ta vrijednost može biti minimum, maksimum, prosjek, nula ili neka druga vrijednost. Budući da se radi o podacima koji imaju točno određene vrijednosti i na temelju tih vrijednosti određuje se zavisna varijabla trajanja procesa, nije bilo moguće nedostajuće vrijednosti zamijeniti s bilo kojom vrijednosti osim nula.



Slika 25. Operator Replace Missing Values

Pomoću ovog operatora, atributi prikazani na slici 25. kojima su nedostajale vrijednosti, poprimili su vrijednost 0.

Zadnji korak pripreme podataka je odabir atributa koji ulaze u analizu. Na slici 26. prikazani su atributi koji su isključeni iz analize pomoću operatora '*Select attributes'*.



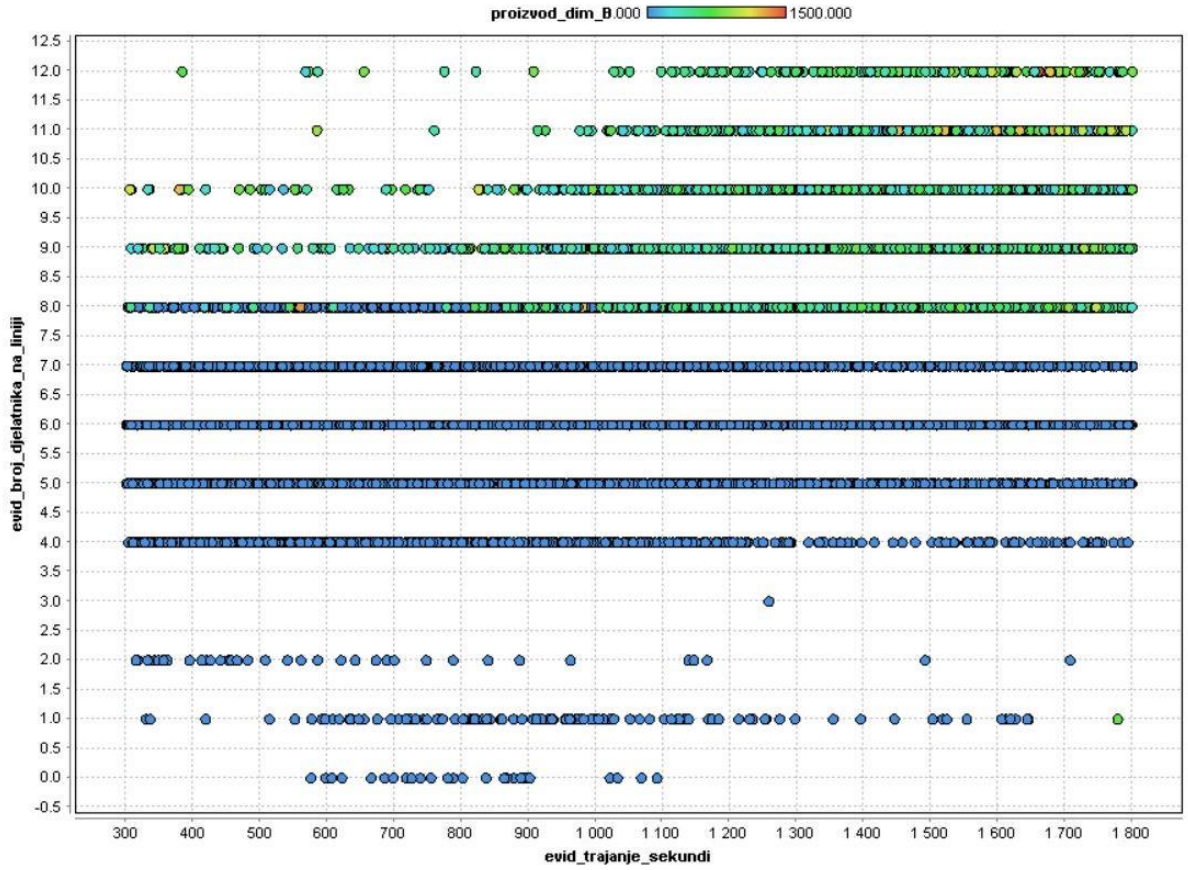
Slika 26. Operator Select Attributes

Ovi atributi ne donose vrijednost pri analizi podataka, već ju prigušuju stoga ne ulaze u daljnju analizu.

Rezultat ovog procesa je da nema proizvoda s nedostajućim vrijednostima, proizvodi koji su imali neprikladne vrijednosti su isključeni, a atributi koji nisu bili prikladni za analizu su također isključeni. Od početnih 44 atributa i 87 542 podatka, nakon pripreme podataka, model će se graditi na 40 atributa i 50 629 podataka.

7.2. EKSPLORATIVNA ANALIZA PODATAKA

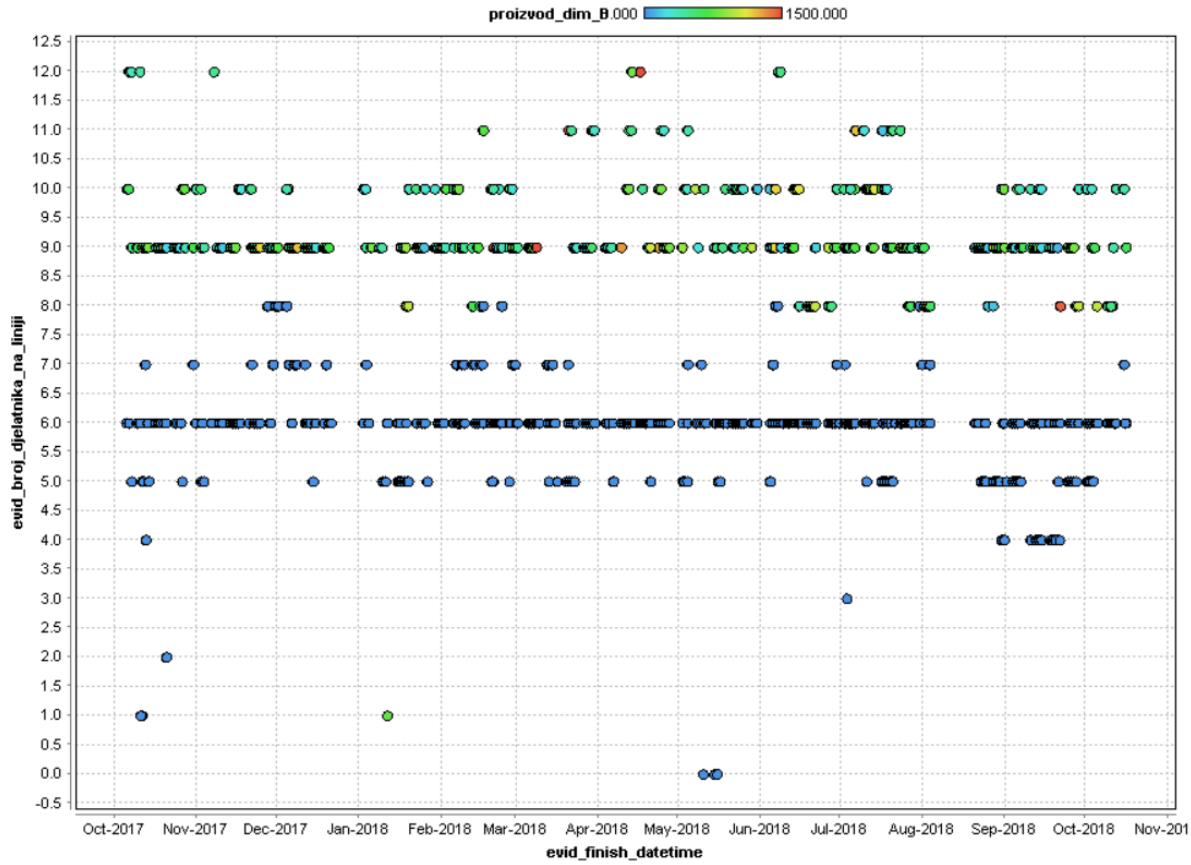
Kako bi se lakše razumjeli ulazni podaci, potrebno je provesti eksplorativnu analizu podataka. Analiza je napravljena pomoću analitičke platforme *RapidMiner* koji radi s već prethodno pripremljenim podacima. Pomoću ovakve analize, moguće je predvidjeti rezultate učenja.



Slika 27. Dijagram trajanja procesa i broja djelatnika na op. liniji

Slika 27. prikazuje dijagram koji na *x* osi ima vrijednost trajanja procesa u sekundama, dok je na *y* osi prikazana vrijednosti broja djelatnika na liniji. Različite boje prikazuju različite dimenzije *B*. Iz dijagrama je vidljivo da je malo procesa u kojima je broj djelatnika manji od 4 i da su procesi tada kraće trajali. Vidljivo je i da kad je broj radnika bio od 4 do 7, da su uglavnom radili sa malim dimenzijama *B* pravokutnih proizvoda. Trajanje procesa u tom slučaju je ravnomjerno raspoređeno, odnosno veći broj radnika ne utječe na kraće trajanje procesa. Nadalje, vidljivo je da kada je broj radnika bio od 8 do 12, da se radilo sa srednjim veličinama pravokutnih proizvoda. Kada je broj radnika bio najveći, tj. 10, 11 ili 12 očekivano je da je duljina trajanja procesa kraća. Iz prikaza je vidljivo da to nije slučaj, već da je proizvodnja tada trajala najduže. Iz ovakvog prikaza lako je zaključiti da će model imati poteškoća prilikom učenja i dobivanja kvalitetnih rezultata predviđanja.

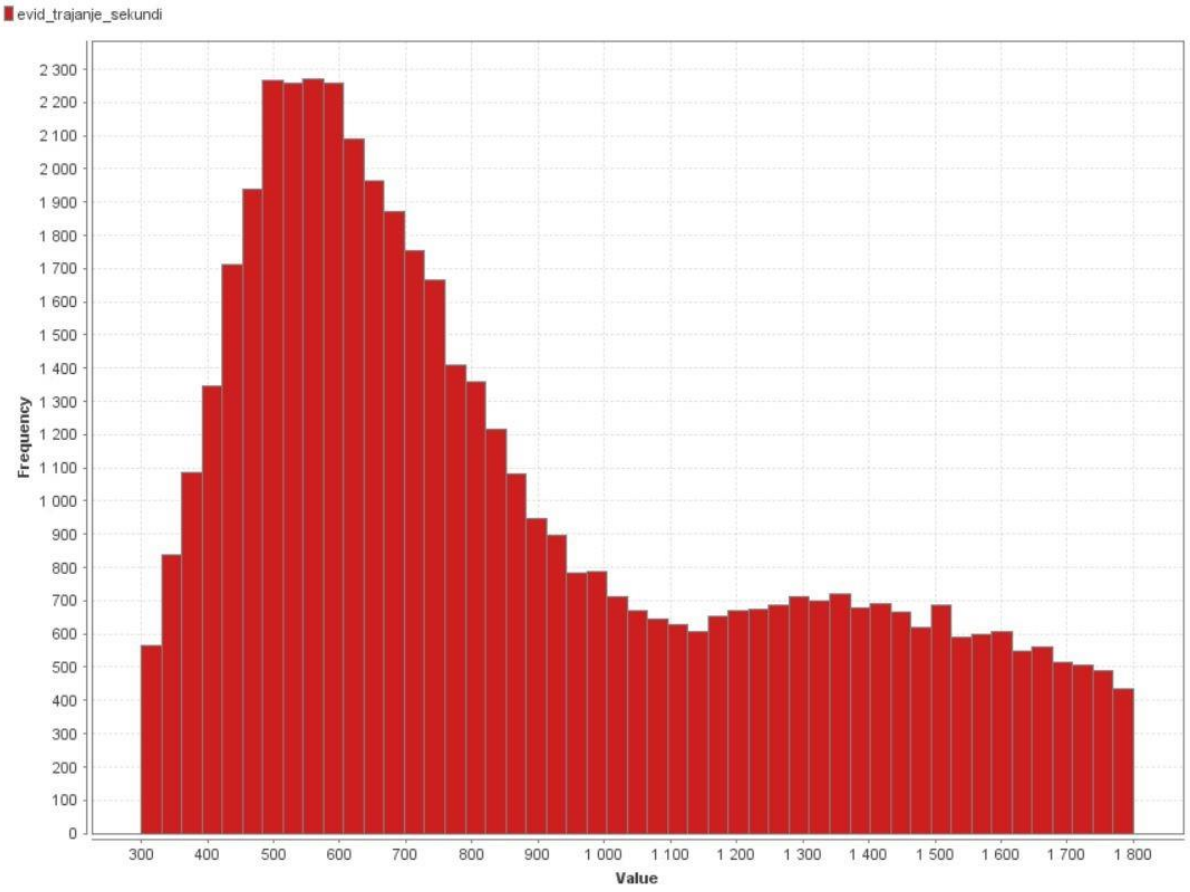
Slika 28. prikazuje broj djelatnika na operacijskoj liniji od listopada 2017. godine do listopada 2018. godine.



Slika 28. Dijagram završetka proizvodnje i broja djelatnika na op. liniji

Iz dijagrama je lako vidljivo da je najčešći broj djelatnika na operacijskoj liniji bio 6 i da su se tada proizvodili pravokutnih proizvodi manjih širina. Pravokutni proizvodi srednjih širina proizvodili su se najčešće kada je bilo 9 djelatnika na operacijskoj liniji. Iz dijagrama se da iščitati da je proizvodnja obustavljena tokom Novogodišnjih blagdana, kao i sredinom kolovoza za vrijeme kolektivnog godišnjeg odmora.

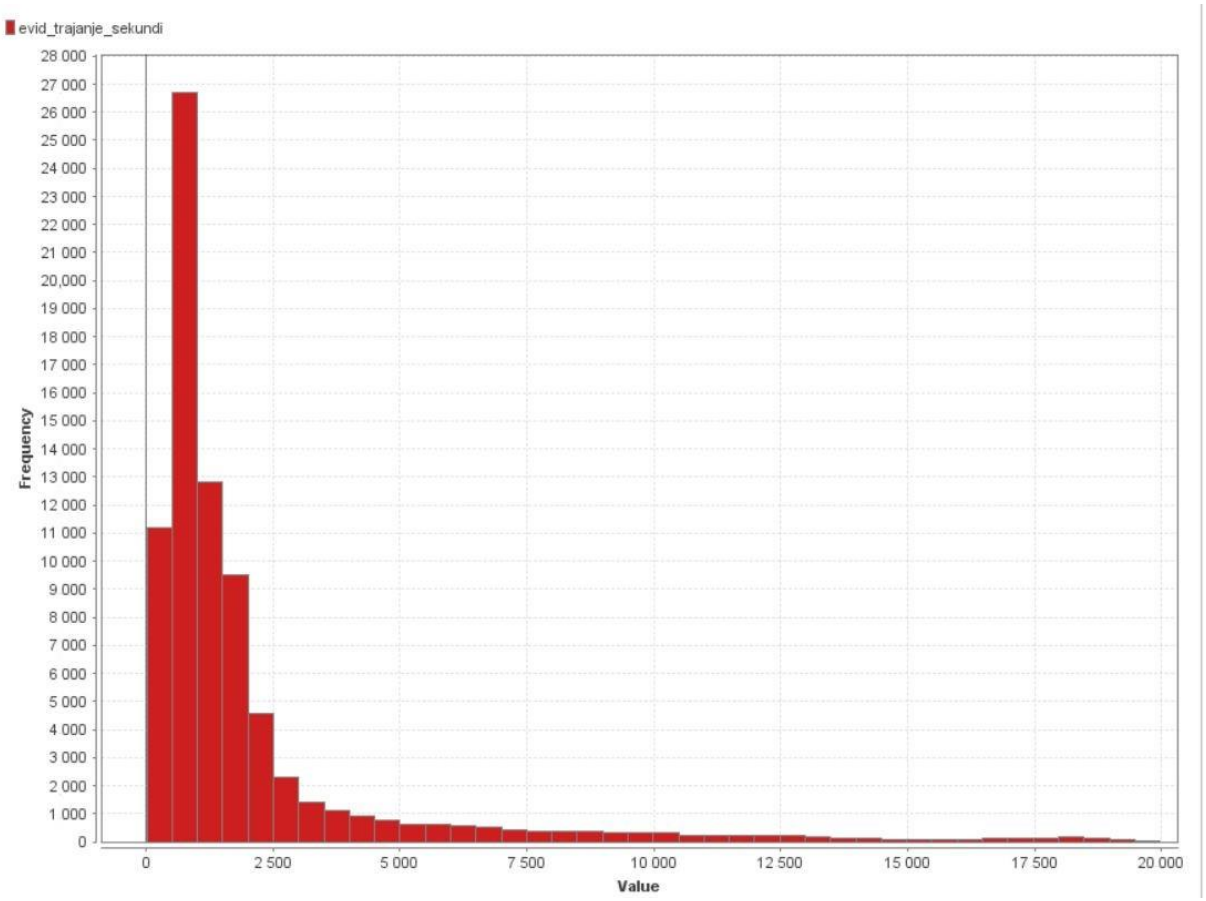
Slika 29. prikazuje histogram rasipanja vrijednosti trajanja procesa proizvodnje.



Slika 29. Histogram raspodjele duljine trajanja procesa

Najveća frekvencija pojavljivanja prisutna je kod procesa koji traju između 500 i 600 sekundi, tj. 8 i 10 minuta. Nadalje, linearna distribucija trajanja procesa je i kod procesa koji traju između 1000 i 1800 sekundi, tj. 16 i 30 minuta.

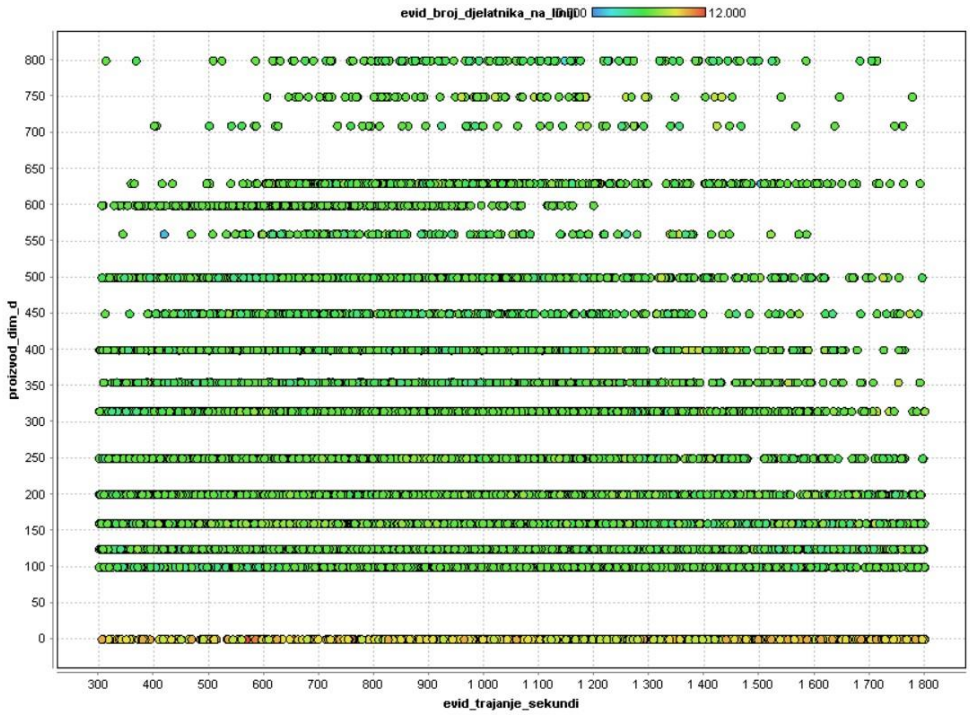
Na slici 30. prikazan je histogram raspodjele trajanja procesa u sekundama prije pripreme podataka.



Slika 30. Histogram raspodjele duljine trajanja procesa prije pripreme podataka

Jasno je vidljivo zašto je bilo potrebno ograničiti vrijeme trajanja na 30 minuta (1 800 sekundi). Najveća frekvencija trajanja procesa je između 0 i 30 minuta. Ukoliko su bili uključeni i ostali podaci, model za učenje bi lošije predviđao, budući da bi uzimao u obzir procese gdje je vrijeme trajanja bilo više od sat vremena što nije slučaju u stvarnoj proizvodnji. Na ovaj način se je pokušalo poboljšati učenje modela na primjerima koji su stvarni za poduzeće i proizvodnju. Iz histograma je jasno vidljivo da je ¼ ukupnih vremena trajanja proizvodnje bilo između 0 i 30 minuta. Budući da nije moguće da je proces proizvodnje trajao 0 minuta, kao minimalno vrijeme trajanja procesa odabrana je vrijednost od 5 minuta, kao što je i vidljivo na slici 23.

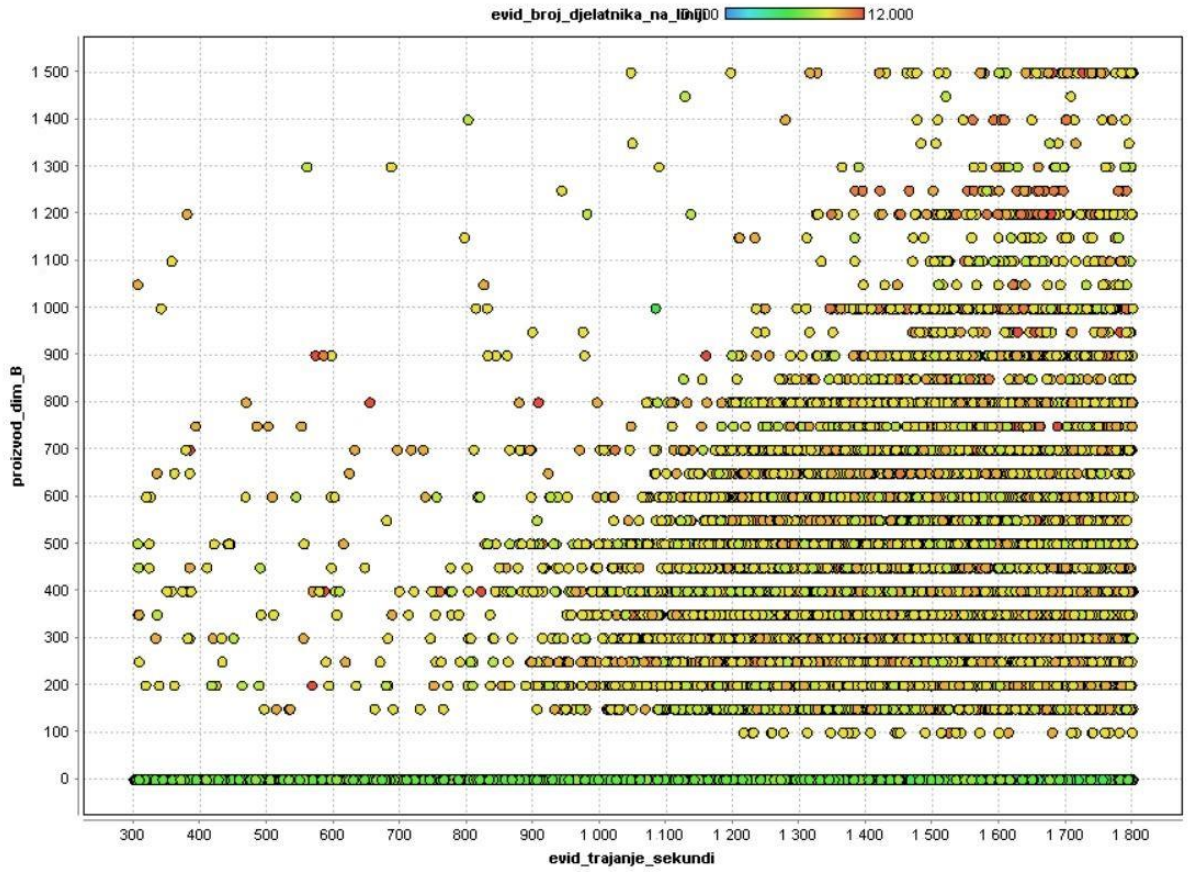
Slika 31. prikazuje dijagram raspodjele pojavljivanja različitih veličina promjera cilindričnih proizvoda s obzirom na vrijeme trajanja procesa.



Slika 31. Odnos promjera proizvoda i vremena trajanja procesa

Iz dijagrama je vidljivo kako nema jasne povezanosti između vremena trajanja procesa i veličine proizvoda. Vrijeme trajanja procesa za proizvode čije su dimenzije bile između 100 mm i 650 mm trajalo je između 5 i 30 minuta te se ne može iščitati pravilo vezano za duljinu trajanja procesa i veličinu proizvoda. Vidljivo je da je za proizvode čija je dimenzija *d* između 650 mm i 800 mm, vrijeme trajanja procesa je između 10 i 25 minuta, ali tih je proizvoda jako malo te je teško za model učenja napraviti pravilo.

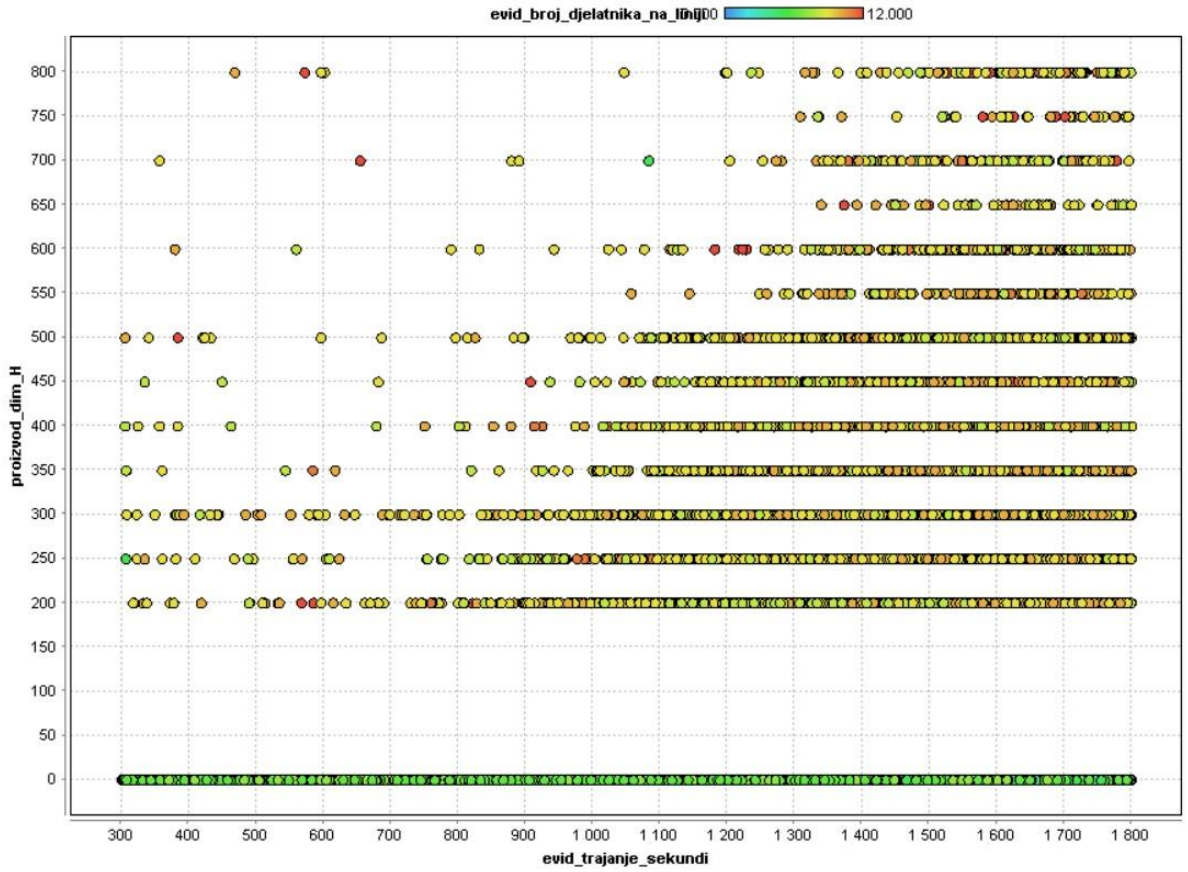
Slika 32. prikazuje dijagram vremena trajanja procesa u odnosu na veličinu dimenzije *B* pravokutnog proizvoda.



Slika 32. Odnos širine proizvoda i vremena trajanja procesa

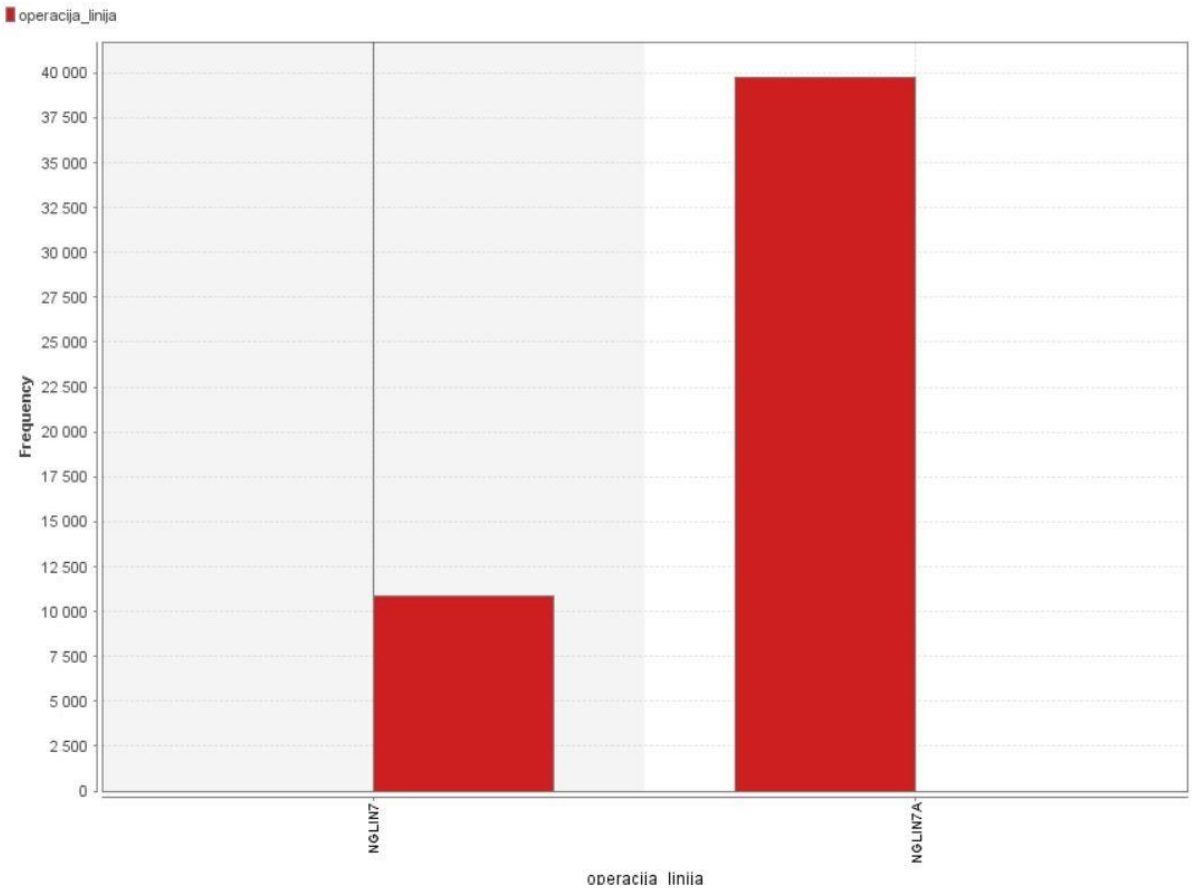
Ukoliko se usporede dijagrami sa slike 30. i ovaj, vidljivo je da kod dimenzije *B* postoji ponovljivost u učestalosti pojavljivanja, dok to kod dimenzije d nije bio slučaj. Vidljivo je da što je dimenzija *B* bila veća, trajanje procesa je bilo dulje. Također je vidljivo da je najčešća minimalna vrijednost trajanja procesa bila 15 minuta te se proporcionalno povećavala s veličinom dimenzije *B*. Iz dijagrama se može zaključiti da će model bolje predviđati pravokutne proizvode od cilindričnih proizvoda.

Slika 33. prikazuje dijagram vremena trajanja procesa u odnosu na veličinu dimenzije *H* pravokutnog proizvoda.



Slika 33. Odnos visine proizvoda i vremena trajanja procesa

Dimenzija *H* se slično ponaša kao i dimenzija *B* ukoliko se gleda odnos vremena trajanja procesa i veličine proizvoda. Ovakav prikaz je bio očekivan s obzirom da su dimenzije *B* i *H* povezane. Ovaj prikaz pokazuje manji broj kategorija dimenzija *H*, te da se dimenzije kreću od 200 mm pa do 800 mm. Očekivano, izrada proizvoda većih dimenzija će trajati dulje, dok će manje dimenzije trajati kraće. Ovaj prikaz je potvrdio da su podaci vezani za pravokutne proizvode pravilno prikupljeni. Treba također uzeti u obzir činjenicu da je pravokutnih proizvoda manje u odnosu na cilindrične proizvode što prikazuje graf na slici 34.



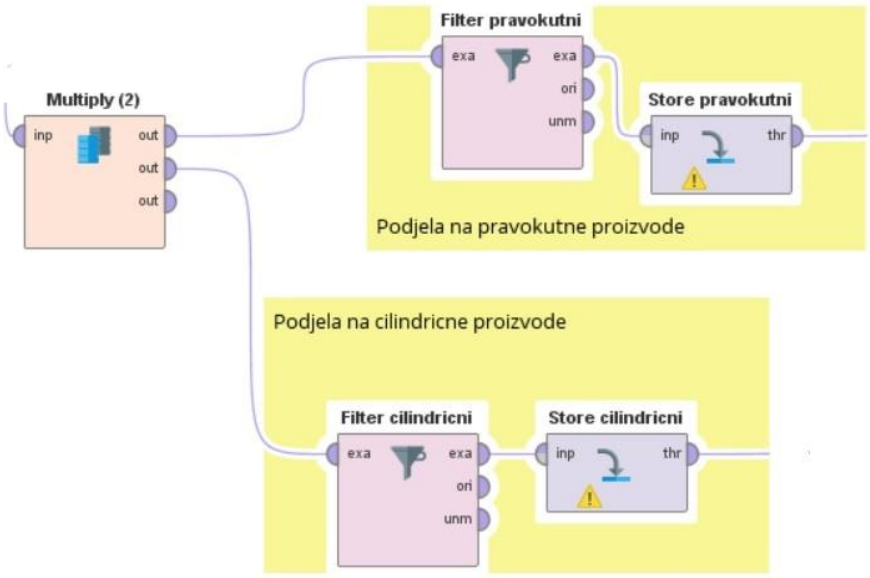
Slika 34. Omjer pravokutnih i cilindričnih proizvoda

Naziv NGLIN7 označava operaciju izrade pravokutnih proizvoda i izrađeno ih je 10 884, dok je cilindričnih proizvoda izrađeno 39 745. Vidljivo je da ¼ proizvoda čine pravokutni proizvodi dok je ostalih ¾ proizvoda cilindrični proizvodi. Ovaj podatak može uvelike (negativno) utjecati na kvalitetu dubinske analize podataka s obzirom da su ulazni podaci cilindričnih proizvoda lošije kvalitete, a njih je čak ¾ ukupne količine proizvoda.

Eksplorativna analiza podataka pridonijela je u razumijevanju podataka, tj. dobio se uvid u ponašanje podataka, njihovu kvalitetu i njihovo kretanje. Ono što vizualizacija ne može, a dubinska analiza podataka to uspješno radi, jest identificirati pravila ponašanja podataka kao i višedimenzijsku povezanost. Pomoću dubinske analize podataka moguće je otkriti skrivene obrasce i pravilnosti koje nije moguće vidjeti proučavanjem podataka, a ni njihovom vizualizacijom. Idući korak primjene strojnog učenja je izgradnja regresijskog modela pomoću kojeg će se pokušati dobiti nove informacije, koje nisu bile moguće dobiti eksplorativnom analizom podataka.

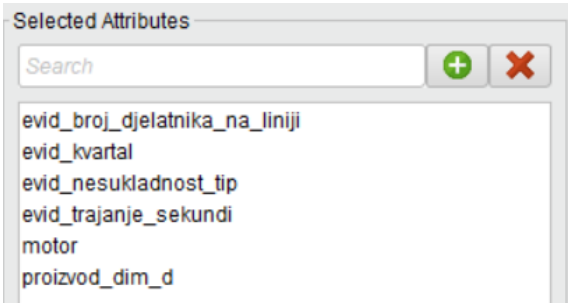
7.3. REGRESIJSKI MODEL

U nastavku je prikazan modela učenja regresijom. Kako bi se dobili što bolji rezultati i kako se model učenja ne bi gušio s različitim podacima, proizvodi su podijeljeni u dvije skupine; cilindrični i pravokutni proizvodi (slika 35).



Slika 35. Podjela podataka na cilindrične i pravokutne

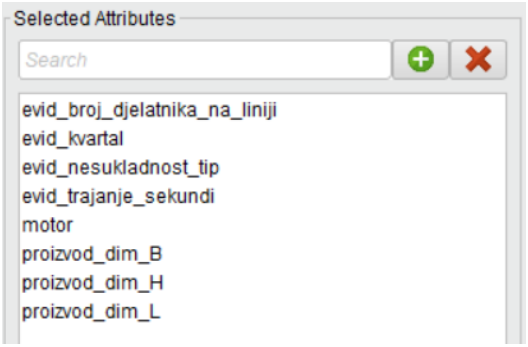
Ovaj postupak rezultirao je smanjenjem broja nezavisnih varijabli kako bi model učenja radio na povezanim podataka što bi trebalo rezultirati što većim koeficijentom korelacije. Slika 36. prikazuje odabrane atribute za cilindrične, a slika 37. za pravokutne proizvode.



Slika 36. Odabir atributa za cilindrične proizvode

Cilj izrade regresijskog modela jest definirati vrijeme trajanja operacije kao funkciju preostalih numeričkih atributa tj. vremena trajanja (*evid_trajanje_sekundi),* veličine proizvoda (*proizvod_dim_d)* i broja djelatnika na operacijskoj liniji (*evid_broj_djelatnika_na_liniji),* ujedno uzimajući u obzir ima li proizvod ugrađeni motor (*motor)*, kada je izrađen (*evid_kvartal)*

i da li je postojao nekakav tip nesukladnosti na proizvodu (*evid_nesukladnost_tip)*. Ukupni broj podataka za pravokutne proizvode nakon podjele je 10 884, dok je cilindričnih 39 745.

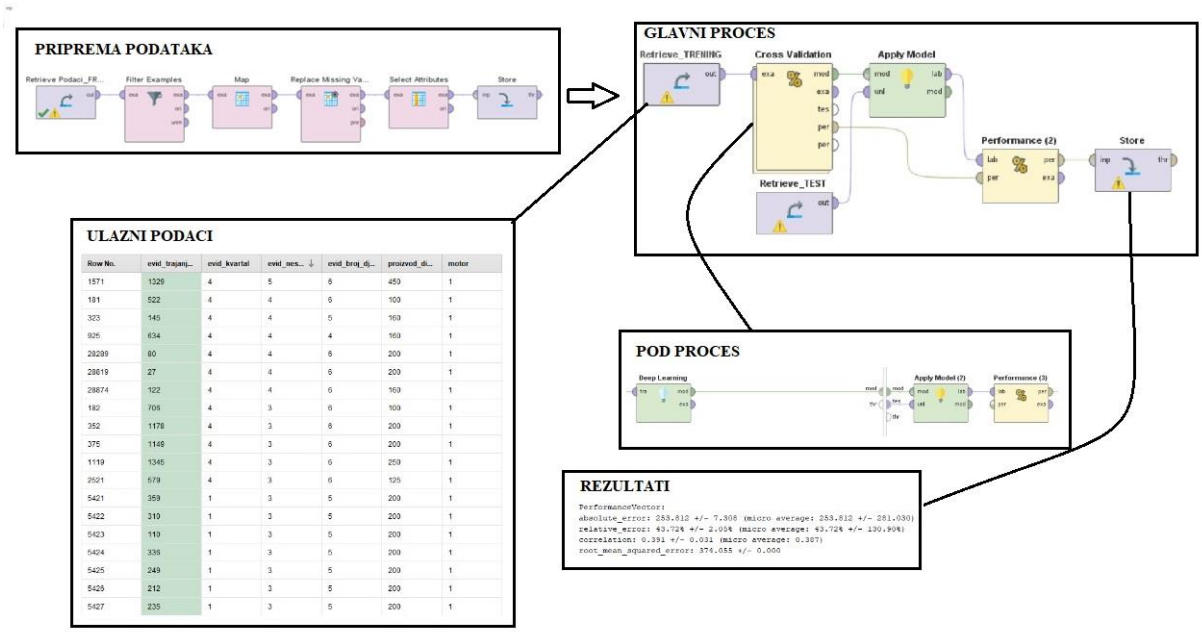


Slika 37. Odabir atributa za pravokutne proizvode

Jednako je napravljeno i za pravokutne proizvode, gdje su osim već navedenih atributa, bili uključeni i atributi dimenzije visine (*proizvod_dim_B),* dimenzije širine (*proizvod_dim_H)* i dimenzije duljine pravokutnog proizvoda (*proizvod_dim_L).*

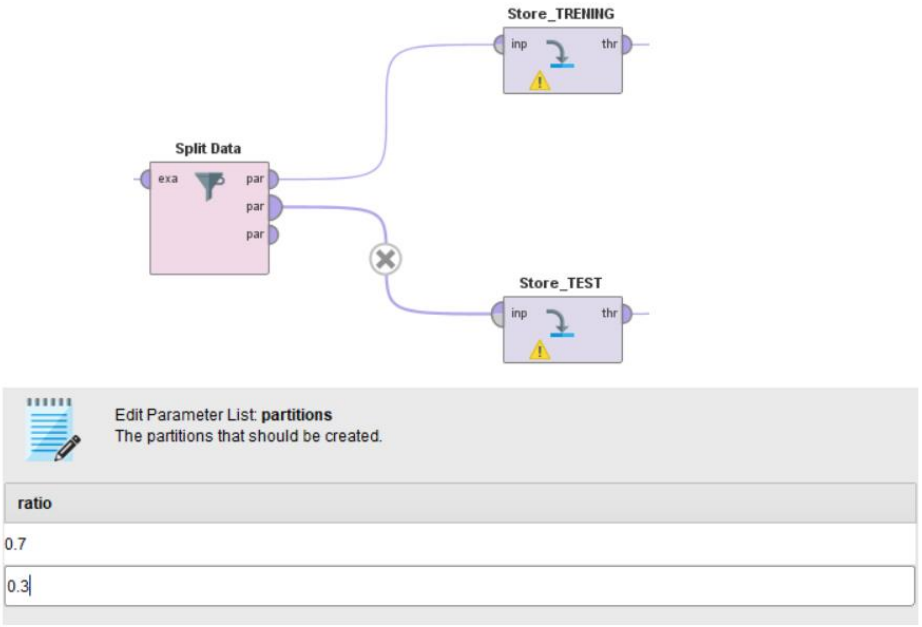
7.3.1. Prvi regresijski model

Slika 38. prikazuje sažete korake proces za prediktivno učenje, gdje je korišten operator *Deep Learning.*



Slika 38. Prvi regresijski model

Nakon pripreme podataka, potrebno je podijeliti ulazni skup podataka na trening i test skup. Podjela je napravljena pomoću operatora podjele podataka (eng. *Split data)* u omjeru 70% za trening skup i 30% za test skup, kao što je prikazano na slici 39.



Slika 39. Podjela podataka na trening i test skup prvog regresijskog modela

Regresijski model izgrađen je na trening skupu. Model nije vidio niti će koristiti podatke test skupa za učenje te će se uspješnost modela provjeravati na testnom skupu. Najčešće se koristi podjela u omjeru 70:30, no konačni omjer podjele ovisi o veličini skupa te raspodjeli podataka unutar skupa podataka. Često se, u svrhu provjere stabilnosti modela i procjene prenaučnenosti (eng. *Overfitting)* i pondnaučenosti (eng. *Underfitting*), koristi više različitih omjera.

Pomoću operatora ***Retrieve*** učitali su se već pripremljeni podaci čiji je proces prikazan na slici 22.

Prilikom izgradnje modela korišten je operator k – struka unakrsna validacija (eng**.** *Cross Validation – CV)* koji je ugniježđeni operator. Ulazni (ogledni) set podataka podijeljen je u *k* podskupova jednake veličine. Od *k* podskupova, jedan podskup se zadržava kao skup za testiranje izgrađenog modela tj. za unos za testiranje. Preostali *k*-1 podskupova koriste se kao skupovi podataka za izgradnju modela, tj. za trening. Proces unakrsne validacije se zatim ponavlja *k* puta, pri čemu se svaki od *k* podskupova podataka koristi samo jednom za testiranje. Naposljetku, iz *k* iteracija je dobivena prosječna točnost modela.

Unutar operatora ***Cross Validation*** još su dva pod procesa; pod proces za učenje i testiranje modela. U procesu za učenje korišteni su različiti operatori s ciljem dobivanja što boljeg koeficijenta korelacije. Proces za učenje se sastoji od dva operatora: ***Apply model*** i ***Performance***. Operator *Apply model* operator koristi se nakon što je model već izgrađen na *k*-1 podskupova te služi za testiranje *k*-tog podskupa. Operator *Performance* koristi se za procjenu statistički performansi evaluacije regresijskog zadatka tj. koeficijent korelacije.

Korišteni su različiti parametri za svaki od operatora, a u tablicama 9 i 10 su prikazani samo najbolji rezultati za svaki operator. Tablica 10. prikazuje rezultate za pravokutne, a tablica 9. za cilindrične proizvode.

Koeficijent korelacije (eng**.** *Correlation Coefficient)* – mjera korelacije između dviju matematičkih varijabli

Koeficijent determinacije (eng**.** *Squared Correlation)* – kvadrirana mjera korelacije između dviju matematičkih varijabli. Model je reprezentativniji što je koeficijent determinacije bliži vrijednosti 1.

RMSE (eng**.** *Root Mean Squared Error)* – korijen srednjeg kvadrata pogreške, mjera za razliku između vrijednosti predviđenih modelom i stvarne vrijednosti koja se procjenjuje

NRMSE (eng**.** *Normalized Root Mean Squared Error)* – normalizirani korijen srednjeg kvadrata pogreške. Računa se prema formuli:

$$NRMSE = \frac{RMSE}{\max(y) - \min(y)} \tag{13}$$

Tablica 9. Rezultati prvog regresijskog modela za cilindrične proizvode

CILINDRIČNI PROIZVODI				
Operator	**Koeficijent korelacije**	**Koef. determinacije**	**RMSE**	**NRMSE**
Neural Net	0,216	0,047	314,56	0,209
Deep Learning	0,323	0,106	258,13	0,172
Gradient Boosted Trees	0,363	0,132	253,07	0,168
Random Forest	0,364	0,133	252,86	0,168

Support Vector Machine	0,279	0,078	266,49	0,177

Tablica 10. Rezultati prvog regresijskog modela za pravokutne proizvode

PRAVOKUTNI PROIZVODI				
Operator	**Koeficijent korelacije**	**Koeficijent determinacije**	**RMSE**	**NRMSE**
Neural Net	0,373	0,139	209,95	0,139
Deep Learning	0,401	0,161	205,97	0,137
Gradient Boosted Trees	0,520	0,270	195,44	0,130
Random Forest	0,481	0,231	197,16	0,131
Support Vector Machine	0,361	0,130	213,66	0,142

7.3.2. Drugi regresijski model

Napravljen je novi regresijski model sa svrhom procjene modela. Procjena modela se vršila na promjeni podjele omjera trening i test skupa. U prvom regresijskom modelu podaci su podijeljeni na 70% za trening i 30% za test skup. U drugom regresijskom modelu podaci su podijeljeni u omjeru 80:20 tj. 80% podataka bilo je korišteno za trening, dok je ostalih 20% korišteno za treniranje naučenog modela.

Novi model napravljen je u cilju treninga nad više podacima što rezultira bolje naučenim modelom, dok je istovremeno skup za testiranje dovoljno velik za što precizniju mjeru pogreške. Rezultati drugog regresijskog modela prikazani su u tablicama 11 i 12.

Tablica 11. Rezultati drugog regresijskog modela za cilindrične proizvode

CILINDRIČNI PROIZVODI				
Operator	**Koeficijent korelacije**	**Koeficijent determinacije**	**RMSE**	**NRMSE**
Neural Net	0,163	0,080	291,54	0,194
Deep Learning	0,322	0,104	256,47	0,170
Gradient Boosted Trees	0,361	0,130	252,78	0,168
Random Forest	0,360	0,130	194,84	0,129

Support Vector Machine	0,283	0,080	266,01	0,177

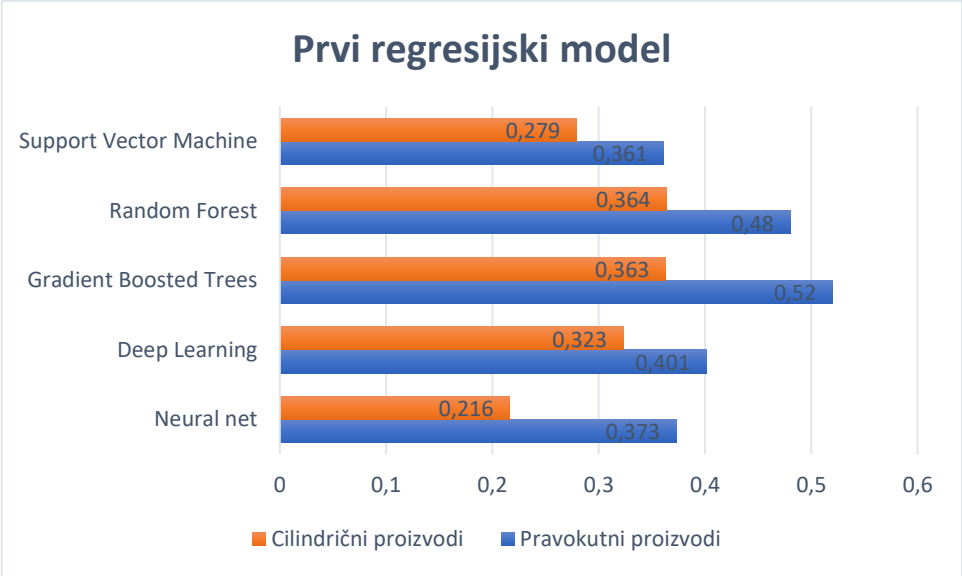
Tablica 12. Rezultati drugog regresijskog modela za pravokutne proizvode

PRAVOKUTNI PROIZVODI				
Operator	**Koeficijent korelacije**	**Koeficijent determinacije**	**RMSE**	**NRMSE**
Neural Net	0,364	0,132	213,81	0,142
Deep Learning	0,411	0,169	206,87	0,137
Gradient Boosted Trees	0,533	0,284	192,04	0,128
Random Forest	0,498	0,248	193,65	0,129
Support Vector Machine	0,371	0,138	211,45	0140

7.3.3. Analiza rezultata i prijedlog za daljnju analizu

Nezadovoljavajuće rezultate se moglo i očekivati s obzirom na ulazne podatke. Kao što se može vidjeti na dijagramima ulaznih podataka, teško je naći pravilnost koji bi rezultirala zadovoljavajućim modelom za učenje. Pravokutni proizvodi imaju nešto bolje rezultate u odnosu na cilindrične. Ukoliko se promotre grafovi gdje su prikazani odnosi vremena trajanja procesa i dimenzija *B, H* i *d*, primjećuje se da parametar dimenzije *B* ima pravilniju raspršenost s obzirom na vrijeme trajanja procesa. Pravokutni proizvodi većih dimenzija zahtijevati će dulje vrijeme izrade, a manji proizvodi kraće vrijeme. Sa slike 31. je vidljivo da su vrijednosti dimenzije *d* linearno raspoređeni neovisno o vremenu trajanja procesa, iz čega se može zaključiti da će jednako trajati proces izrade proizvoda većih i manjih dimenzija. Velika raspršenost podataka te slaba povezanost između ciljne varijable (vrijeme trajanja procesa) i nezavisnih varijabla kao rezultat dali su slabe koeficijente korelacije tj. modele za učenje koji neće dobro predviđati buduća vremena trajanja procesa. Veliki utjecaj na nezadovoljavajuće rezultate ima i činjenica da su većina atributa koji ulaze u analizu kategorički podaci.

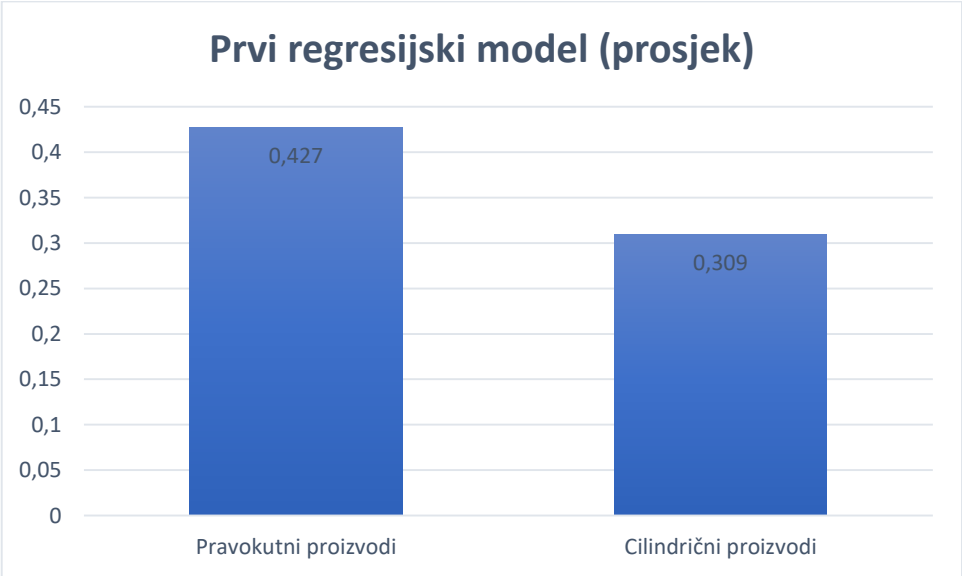
Slika 40. prikazuje grafikon koeficijenta korelacije prvog regresijskog modela, za cilindrične i pravokutne proizvode istovremeno.



Slika 40. Grafikon koeficijenta korelacije za prvi regresijski model

Iz grafikona je jasno vidljivo da će u prvom regresijskom modelu, model bolje predviđati pravokutne proizvode od cilindričnih proizvoda. Takvi rezultati su posljedica vrste i tipa podataka. Između promjera i vremena trajanja procesa, teško je razaznati uzročno – posljedične veze, budući da je iz podataka vidljivo da veličina proizvoda ne utječe na duljinu trajanja procesa. Takvi ulazni podaci predstavljaju prepreku regresijskom modelu pri stvaranju veza, što je uzrokovalo niski koeficijent korelacije i mali stupanj povezanosti. Kod pravokutnih proizvoda, uočljivije su povezanosti između vremena trajanja procesa i dimenzija *B* i *H*. Regresijski model je u tom slučaju bolje učio i postigao bolji koeficijent korelacije i veći stupanj povezanosti.

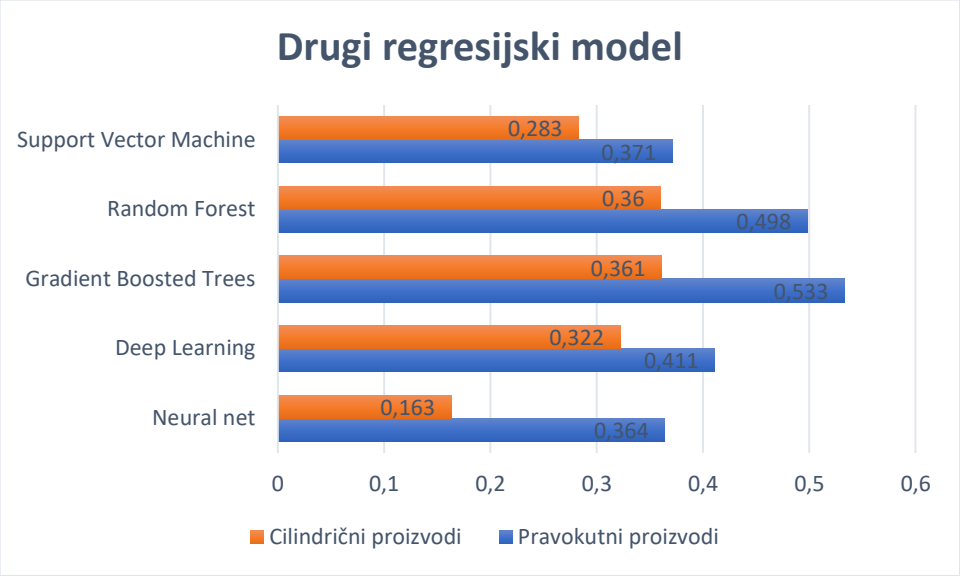
Slika 41. prikazuje graf prosječne vrijednosti koeficijenta korelacije za prvi regresijski model.



Slika 41. Graf prosječne vrijednosti koeficijenta korelacije za prvi regresijski model

Prosječni koeficijent korelacije prvog regresijskog modela je 0,427%, a cilindričnog modela 0,309% (slika 41).

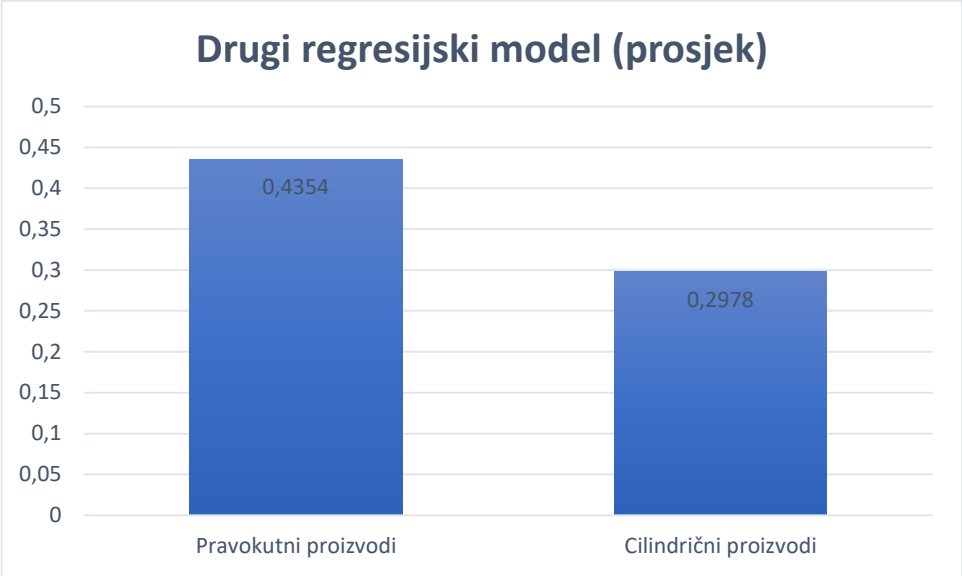
Slika 42. prikazuje grafikon koeficijente korelacije za različite operatore, drugog regresijskog modela.



Slika 42. Grafikon koeficijenta korelacije za drugi regresijski model

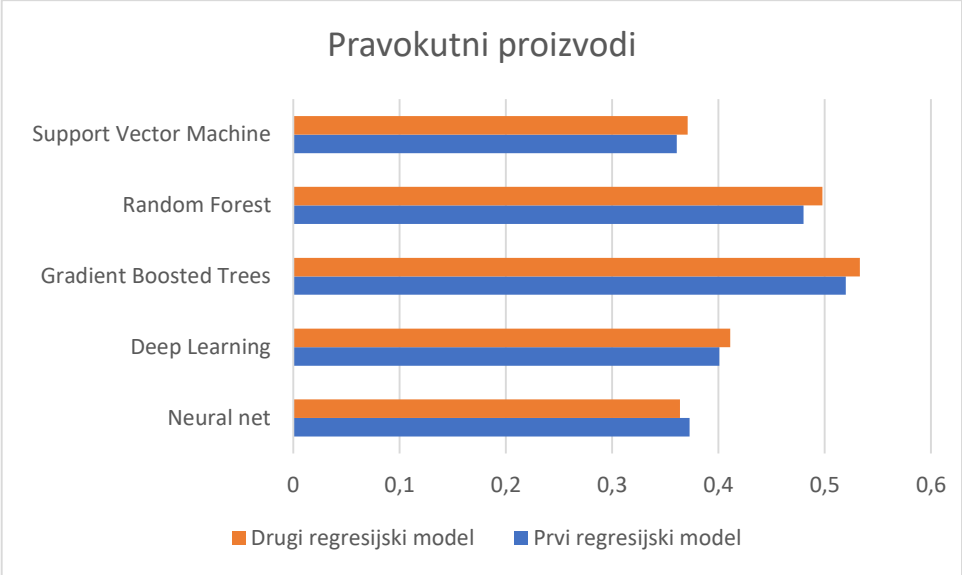
Drugi regresijski model koristio je 80% ulaznih podataka za trening, dok je ostalih 20% koristio za testiranje naučenog modela. Rezultati drugog modela daju slične rezultate kao i prvi. Povećanje skupa za trening rezultiralo je slabim povećanjem koeficijenta korelacije.

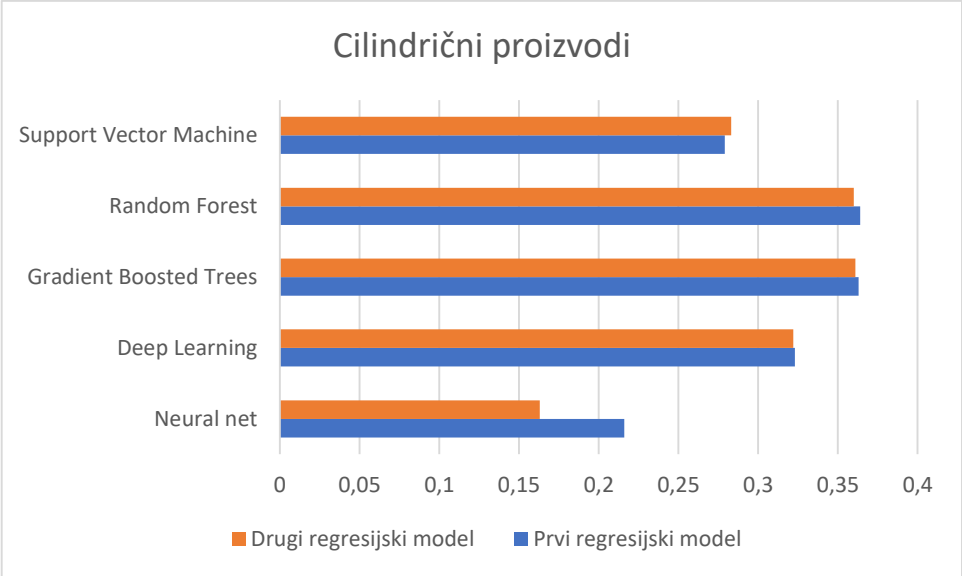
Slika 43. prikazuje prosječni iznos koeficijenta korelacije za drugi model.



Slika 43. Graf prosječne vrijednosti koeficijenta korelacije za drugi regresijski model

Ukupni porast prosječnog koeficijenta korelacije za pravokutne proizvode iznosi približno 3%, dok se za cilindrične proizvode smanjio za približno 4%.





Slika 44. Usporedne vrijednosti za prvi i drugi regresijski model

Slika 44. prikazuje usporedne vrijednosti koeficijenta korelacije za prvi i drugi regresijski model za pravokutne i cilindrične proizvode zasebno. Grafikon pravokutnih proizvoda vizualno prikazuje nešto bolje rezultate kada je korišten veći skup za trening. Grafikon cilindričnih proizvoda prikazuje da je model učenja davao bolje performanse kada se koristio manji skup za trening tj. 70% podataka.

Idući korak u procesu izrade modela za učenje koji će što bolje procjenjivati vremena trajanje izrade proizvoda jest izrada klasifikacijskog modela za procjenjivanje koji će izlaznu nezavisnu varijablu *vrijeme_trajanje_sekundi* podijeliti u klase. Regresijskim modelom nastojala se pronaći povezanost između podataka, odnosno linearna funkcija ovisnosti zavisne varijable vremena trajanja o preostalim numeričkim atributima. Regresijski model limitiran je kategoričkom naravi numeričkih atributa u skupu podataka. Usprkos tome što su se u modeli koristili atributi numeričkog tipa podatka, oni mogu biti smatrani i polinominalnim (kategoričkim) budući da nemaju pravi kontinuirani obrazac pojave. Samim time, teško je bilo očekivati bolje performanse regresijskog modela te se iz tog razloga gradi klasifikacijski model. Klasifikacijskim modelom, koji će podijeliti vrijeme trajanja procesa u klase, nastojati će se pronaći pravilo u učestalosti pojavljivanja podataka u klasama, tj. u ovom slučaju koliko najčešće traje proces. U ovom modelu, uključeni su atributi koji do sada nisu korišteni, kako bi model pokušao naći pravilo ili obrazac ponašanja skriven u podacima, koji korisniku nisu vidljivi samim pregledom ili vizualizacijom podataka.

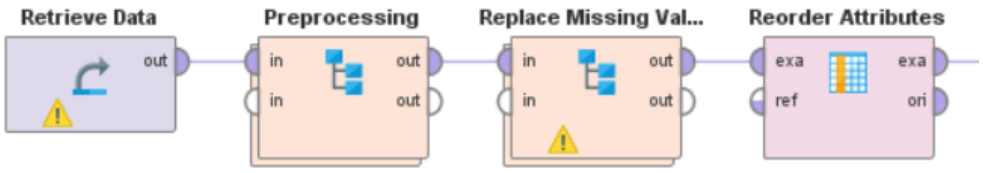
7.4. KLASIFIKACIJSKI MODEL

U ovom poglavlju prikazan je klasifikacijski model i analizirani su rezultati.

7.4.1. Prvi klasifikacijski model

Model za klasifikaciju koristiti će već pripremljene podatke koji ima 39 nezavisnih varijabla i 1 zavisnu varijablu koja se nastoji procijeniti '*vrijeme_trajanje_sekundi*'. Ukupan broj podataka koji će se koristiti u modelu za učenje je 50 629.

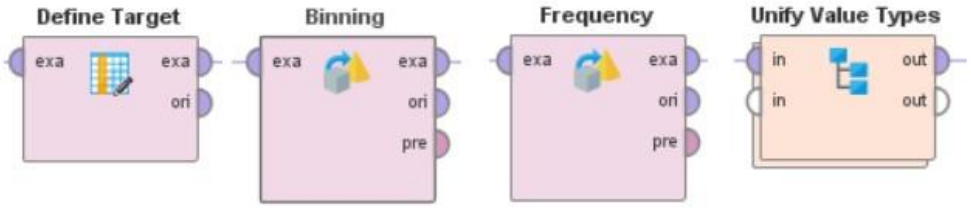
Prvi korak klasifikacijskog modela za učenje prikazan je na slici 45.



Slika 45. Prvi dio prvog klasifikacijskog modela

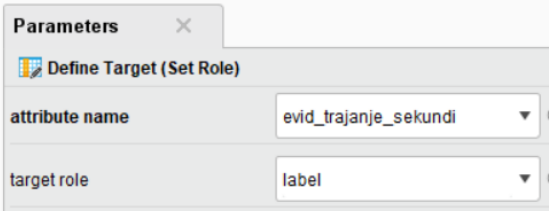
Pomoću operatora ***Retrieve Data*** učitavaju se podaci. Operator ***Preprocessing*** je ugniježđeni operator koji u sebi ima sve potrebne operatore za dodatnu pripremu podataka za klasifikaciju.

Slika 46. prikazuje operatore za pripremu podataka unutar operatora *Preprocessing*.



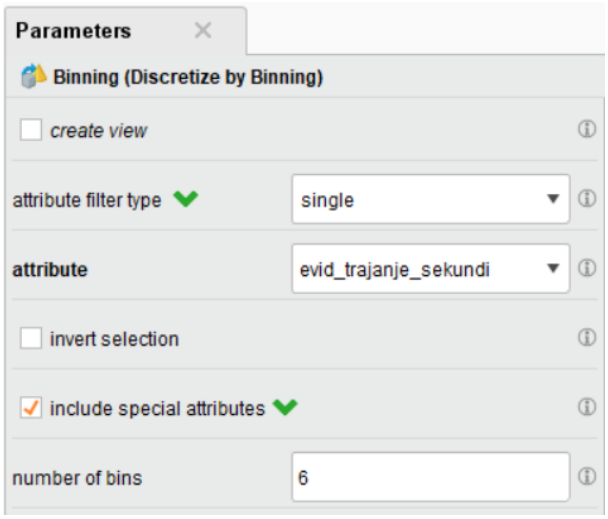
Slika 46. Priprema podataka za prvi klasifikacijski model

Operator *Define target* koristi se za određivanje zavisne varijable koju se nastoji procijeniti kao što je prikazano na slici 47.



Slika 47. Odabir ciljanog atributa

Operator *Binning* diskretizira odabrane numeričke atribute u broj klasa (spremnika) koje je korisnik odabrao. Pomoću ovog operatora, automatski se generiraju spremnici jednakog raspona, dok broj vrijednosti u različitim spremnicima može varirati. Odabran broj različitih klasa tj. spremnika je 6 kao što prikazuje slika 48.



Slika 48. Prikaz operatora Binning

Nakon što su podaci podijeljeni u klase jednakih veličina (intervala), operator *Frequency* je dodatno podijelio podatke tako da su u svakoj klasi podaci jednakih frekvencija.

Operator *Unify all value types* dodatno pred procesuira podatke u prikladni oblik za klasifikaciju i operator korišten za model učenja.

Kao što je navedeno, broj vrijednosti u spremnicima može varirati kao što je vidljivo na slici 49.

Index	Nominal value	Absolute count	Fraction
1	[550.0 - 800.0]	15278	0.302
2	[-∞ - 550.0]	12486	0.247
3	[800.0 - 1050.0]	7661	0.151
4	[1300.0 - 1550.0]	5494	0.109
5	[1050.0 - 1300.0]	5343	0.106
6	[1550.0 - ∞]	4367	0.086

Slika 49. Raspodjela podataka po spremnicima

Najviše podataka ima spremnik pod indeksom 1, iza kojeg slijedi spremnik indeksa 2, dok najmanje podataka ima spremnik 6.

Operator ***Replace Missing Values*** također je ugniježđeni proces koji unutar sebe može sadržavati više pod procesa. Pod procesi unutar ovog operatora prikazani su na slici 50.



Slika 50. Operator Replace Missing Values

Prvi operator *Replace Pos Infinite Values* zamjenjuje pozitivne beskonačne vrijednosti odabranih atributa s odabranom zamjenom. Operator je programiran tako da sve atribute s numeričkim tipom podatka koji imaju beskonačne vrijednosti zamjeni s nedostajućom vrijednosti. Drugi operator (*Replaces Neg Infinite Values)* jednako funkcionira, samo što se ovdje radi o podacima s negativnim beskonačnim vrijednostima. Budući da su se s ova dva operatora generirali podaci s nedostajućim vrijednostima, zadnji operator *Replace Numerical Missing* zamjenjuje nedostajuće vrijednosti s prosječnom (eng**.** *Average)* vrijednosti tog atributa.

Operator *Reorder Attributes* služi za određivanje redoslijeda atributa prema odabiru korisnika. Odabrano je da atributi budu poredani prema abecednom redu.

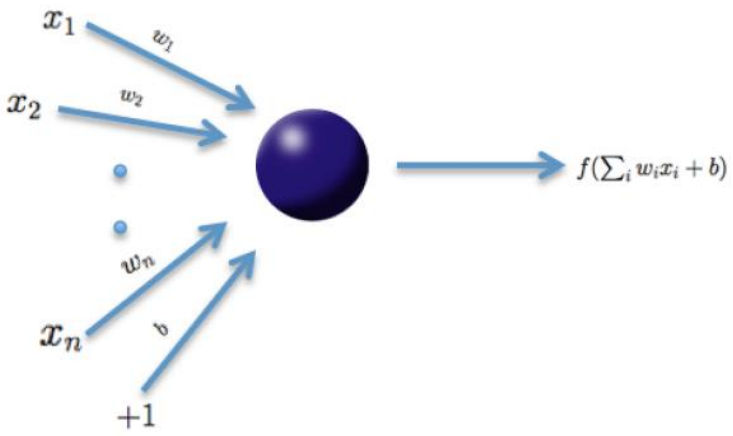
Deep learning operator

U klasifikacijskom modelu koristiti će se operator dubokog učenja (eng**.** *Deep Learning.*

Za razliku od drugih neuronskih mreža, operator *Deep Learning* pruža stabilnost, generalizaciju i skalabilnost s velikim broja podataka. Budući da pruža dobre performanse u različitim vrstama zadataka, duboko učenje brzo postaje prvi izbor algoritma za dobivanje najviše prediktivne točnosti. Duboko učenje radi na principu unaprijedne umjetne neuronske mreže (eng. *Feedforward).* Ovakva mreža nema povratnih veza između neurona pa signali koji krenu od ulaznih neurona nakon određenog broja prijelaza dolaze do izlaza iz mreže, tj. propagacija signala je jednosmjerna. Osnovna jedinica u modelu je neuron, biološki inspirirani model ljudskog neurona [35][36].

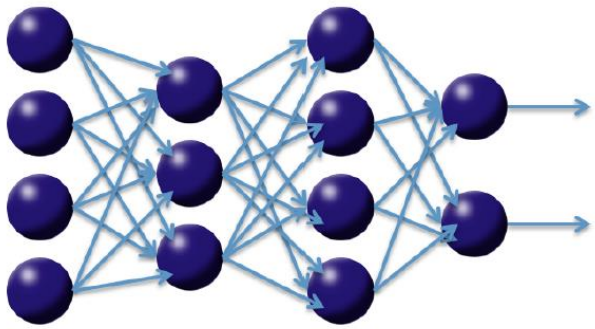
Neuroni su organizirani u slojeve, a izlazi iz svih neurona u jednom sloju postaju ulazi za svaki neuron u idućem sloju. U modelu dubokog učenja, ponderirana kombinacija $\alpha = \sum_{i=1}^{n} w_i x_i + b$

ulaznih signala je agregirana, a izlazni signal $f(\alpha)$ je odašiljan preko povezanog neurona. Funkcija predstavlja nelinearnu aktivnost funkcije upotrijebljenu kroz cijelu mrežu, a pristranost *b* (eng**.** *Bias)* predstavlja neuronov aktivacijski prag [37].



Slika 51. Prikaz ulaznih i izlaznih signala [35]

Višeslojne, unaprijedne umjetne neruonske mreže sastoje se od mnogih slojeva međusobno povezanih neuronskih jedinica (slika 52), počevši od ulaznog sloja, nakon čega slijedi više slojeva nelinearnosti, a završava s linearnom regresijom ili klasifikacijskim slojem koji odgovara izlaznom prostoru. Ulazi i izlazi jedinica modela slijede osnovnu logiku pojedinačnog neurona opisanog u prethodnom tekstu [37].



Slika 52. Slojevi neuronske mreže [35]

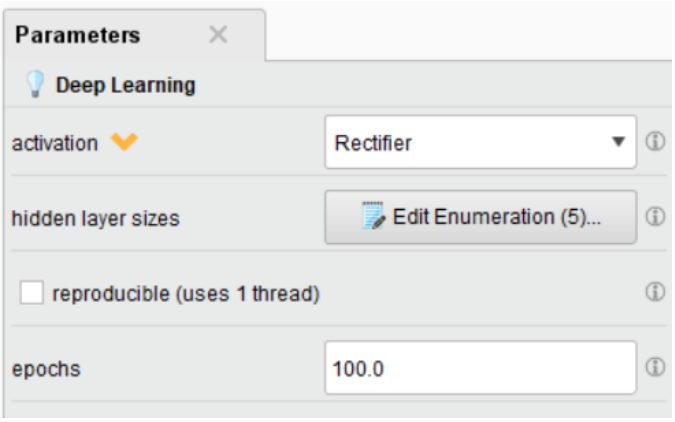
Prvi sloj čine ulazni podaci, trening ili test skup. Posljednji sloj je odgovor, tj. u ovom slučaju predikcija. Ukoliko se radi o regresiji gdje se uči jedna vrijednost, sloj će imati jedan neuron. Ukoliko je model klasifikacijski, utoliko će izlazni sloj imati jedan neuron za svaku moguću klasu. Slojevi između ulaznog i izlaznog sloja nazivaju se skriveni slojevi (eng**.** *Hidden Layers*) [35].

Svaki neuron u svakom skrivenom sloju ima težinu za svaki od njegovih ulaza, a mreža se uči tako da modificira te težine[35].

Drugi korak klasifikacijskog modela je treniranje i testiranje modela za učenje pomoću operatora dubokog učenja (eng**.** *Deep Learning***).**

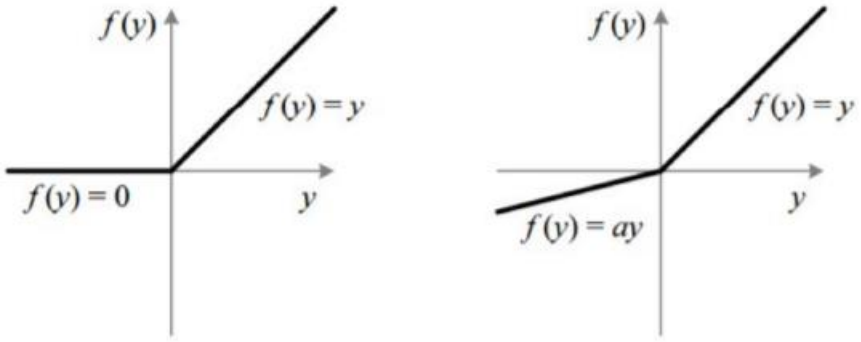
Prije nego što je moguće započeti proces treninga, potrebno je podijeliti ulazni set podataka na dvije skupine: trening i test. Pomoću operatora za podjelu podataka, ulazni set podataka podijeljen je u omjeru 80:20.

Parametri za *Deep Learning* operator prikazani su na slici 53, koji su optimizacijom različitih parametara dali najbolji rezultat za učenje.



Slika 53. Parametri za duboko učenje

Najčešće korištena aktivacijska funkcija je ispravljačka (eng**.** *Rectifier)* aktivacijska funkcija tzv. ReLU (eng**.** *Rectified Linear Unit)* definirana pozitivnim dijelom svog argumenta: $f(x) = x^+ = \max(0, x)$. Ova funkcija je pola ravna, a zatim raste linearno. Može se promatrati kao kombinacija stepenaste i linearne funkcije. Negativna strana ove funkcije je što loše radi s negativnim ulazima, stoga postoje ispravljene verzije ove funkcije kao što su propustljivi ReLU (eng**.** *Leaky ReLU)* i slučajni ReLU (eng**.** *Leaky ReLU).* ReLU funkcija i ispravljena tj. propustljiva funkcija, prikazani su na slici 54 [38].



Slika 54. Rectifier aktivacijska funkcija [38]

Odabrani broj skrivenih slojeva(eng**.** *Hidden Layers*) je 5, gdje je težina svakog sloja 50. Broj epoha je postavljen na 100.

7.4.2. Analiza rezultata i prijedlog za daljnje istraživanje

Slika 55. prikazuje matricu konfuzije za klasifikacijski model i parametre prikazane u prethodnom poglavlju.

accuracy: 58.57%

	true [-∞ - 550.0]	true [550.0 - 800...	true [800.0 - 105...	true [1050.0 - 13...	true [1300.0 - 15...	true [1550.0 - ∞]	class precision
pred. [-∞ - 550.0]	1707	713	105	25	10	4	66.58%
pred. [550.0 - 80...	652	1867	462	105	36	23	59.36%
pred. [800.0 - 10...	105	419	811	246	73	18	48.50%
pred. [1050.0 - 1...	16	37	123	447	175	53	52.53%
pred. [1300.0 - 1...	9	11	22	211	663	339	52.83%
pred. [1550.0 - ∞]	8	9	9	35	142	436	68.23%
class recall	68.36%	61.09%	52.94%	41.81%	60.33%	49.94%	

Slika 55. Matrica konfuzije prvog klasifikacijskog modela

Retci prikazuju koliko je puta model procijenio da je rezultat u tom intervalu, dok stupci prikazuju koliko je puta zaista procijenjeni rezultat bio u tom intervalu. Ukoliko analiziramo prvi red, model je od ukupno 2 564 puta procijenio da je vrijednost ciljanog atributa u intervalu $[-\infty - 550{,}0]$. Od 2 564 podataka, točno je pogodio 1 707 puta, dok je za ostalih 857 puta pogriješio. Preciznost klase $[-\infty - 550{,}0]$ (eng**.** *Class precision)* jest 66,58 %. Najlošija preciznost klase je u trećem intervalu $[800 - 1050{,}0]$ gdje iznosi 48,50%, što znači da je model procijenio da je vrijednost ciljane varijable 1 672 puta bila u ovom intervalu, od čega je samo 811 puta točno procijenio, dok je 861 puta krivo procijenio. Drugim riječima, u više od 50% slučajeva je pogrešno procijenio da je ciljna varijabla u tom intervalu.

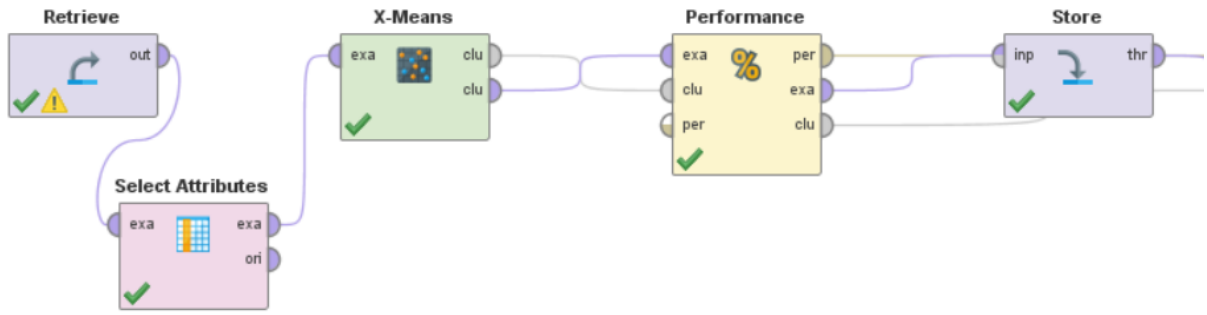
Analizirajući prvi stupac, vidljivo je da je ukupno 2 497 puta vrijeme trajanja procesa trajalo u intervalu $[-\infty - 550{,}0]$, a model je 1 707 puta procijenio točno, dok 790 puta nije. Ovaj podatak naziva se odziv klase (eng**.** *Class recall).* Odziv je najlošiji bio u slučaju intervala $[1050{,}0 - 1300{,}0]$, gdje je vrijeme trajanja procesa bilo u tom intervalu 1 069 puta, od čega je model točno procijenio tu klasu samo 622 puta.

Promatrajući matricu konfuzije, mogu se donijeti određeni zaključci. Za prvi interval tj. vrijednosti vremena trajanja procesa od 0 minuta do 9 minuta, odziv klase upućuje na to da će model u 68% slučajeva točno klasificirati pozitivne primjerke tj. kada je vrijeme trajanja procesa zaista u tom intervalu. Za intervale kada je vrijeme trajanja procesa između 13 i 17 minuta ($[105,.0 - 1300{,}0]$) i 25 i 30 minuta ($[1050{,}0 - 1800{,}0]$) model u gotovo 50% slučajeva ne će točno procjenjivati vrijeme trajanja procesa.

Ukupna točnost modela iznosit 58,57%, što nije zadovoljavajuće za predviđajući model. Istovremeno, analiza klasifikacijskim modelom daje mnogo prostora za daljnja istraživanja. Model je podijelio vremena trajanja procesa u klase koje se mogu koristiti za daljnja procjenjivanja u realnom procesu. Također, rezultati odziva i preciznosti klasa mogu biti smjernice za najčešće frekvencije vremena trajanja procesa koja mogu pomoći u proizvodnji. Klasifikacijski model je dao više informacija od regresijskog modela. Regresijski model je pružio podatke o povezanosti tj. nepovezanosti varijabla, koliko vrijednosti jedne utječu na druge varijable, dok je klasifikacijski model dao opipljive rezultate na temelju kojih se mogu donositi odluke.

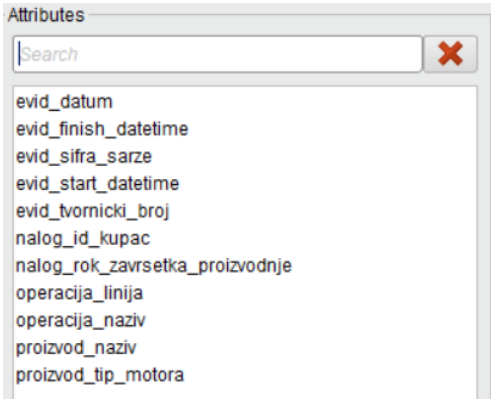
7.4.3. Klaster analiza s ciljem optimizacije rezultata

Kao što je navedeno u odjeljku 4, metodom klaster analize se nastoji skup podataka razvrstati u skupove međusobno što sličnijih podataka. Objekti unutar jednog klastera moraju međusobno biti što sličniji, a istovremeno se što više razlikovati od objekata u drugim klasterima. Model klaster analize prikazan je na slici 57.



Slika 56. Prikaz modela za klaster analizu

Klaster analiza primijenjena je na ulaznom skupu podataka koji se sastoji samo od numeričkih podataka budući da je ova metoda primjenjiva samo na takvim podacima. U analizu je ušlo 28 atributa, a na slici 57. prikazani su atributi koji nisu ušli u analizu.



Slika 57. Prikaz atributa koji nisu ušli u klaster analizu

Korišten je operator *x-means* koji sam određuje broj klastera, nakon što se postavi minimalan i maksimalan broj klastera. Operator *x-means* je klasterski algoritam koji određuje točan broj centroida baziranih na heurističkoj metodi. Započinje s minimalnim setom centroida, a zatim iterativno provjerava da li korištenje više centroida poboljšava rezultate grupiranja.

Operator ***Cluster Distance Performance*** koristi se za procjenu performansi metoda klasteriranja temeljene na centroidima. Indeks korišten za analizu klaster analize je ***David Bouldin*** indeks. Indeks se temelji na približnoj procjeni udaljenosti između klastera i njihovoj disperziji, kako bi se dobila konačna vrijednost koja predstavlja kvalitetu particije. tradicionalna mjera ovog indeksa za mjerenje disperzije je prosječna udaljenost podataka od centra do klastera. Indeks može biti u intervalu od 0 do 1, a teži se vrijednosti što bližoj 0.

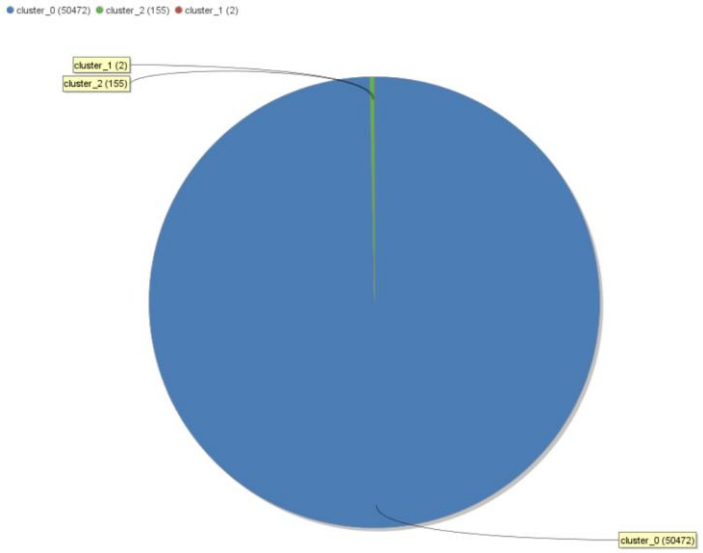
Rezultati

Iako je *David Bouldin* – ov indeks nizak, tj 0,255, podaci nisu zadovoljavajuće razmješteni po klasterima. Operator *x-means* odabrao je kao optimalan broj klastera minimalnih 3. Slika 58 prikazuje pripadnost podataka pojedinim klasterima.

Index	Nominal value	Absolute count	Fraction
1	cluster_0	50472	0.997
2	cluster_2	155	0.003
3	cluster_1	2	0.000

Slika 58. Prikaz grupiranja podataka

Slika 59. prikazuje neravnomjernu raspodjelu po klasterima.



Slika 59. Raspodjela podataka po klasterima

Iz slike je vidljivo da je čak 99,7 % svih klastera grupirano u jedan klaster, dok je ostalih 0,3% pripalo drugom klasteru. Ovakvi rezultati se mogu tumačiti malom disperzijom podataka i kategoričkim vrijednostima. Budući da je većina podataka u obliku kategoričkih vrijednosti, a kategorija nema puno, model je pretpostavio da je veličina podataka slična. Nadalje, atributi poput *tip_motora, evid_i_kvaliteta, evid_kom, evid_godina, evid_operator finish/start,* poprimaju jednaku vrijednost za većinu podataka, što je dodatno utjecalo na grupiranje podataka u jedan klaster. Također treba uzeti u obzir da je atribut *vrijeme_trajanje_sekundi* poprimao vrijednosti od 0 pa sve do čak 13 dana, što nije prikazivalo stvarno stanje procesa te je taj atribut reducirani na od 5 do 30 minuta. Takva redukcija je isto pridonijela ovakvim klasterima.

7.4.4. Drugi klasifikacijski model

Drugi klasifikacijski model za ulazne podatke koristi podatke koji su izlaz klaster analize. Nakon pred procesuiranja podataka, broj atributa koji su ušli u analizu je 14 nezavisnih atributa i dva zavisna atributa (*evid_trajanje_sekundi, cluster).* Ukupan broj podataka je 40 503.

Korišten je isti model kao u poglavlju 7.4.1. ali su ulazni podaci drugačiji. Rezultati su prikazani na slici 60.

accuracy: 41.06%

	true [-∞ - 550.0]	true [550.0 - 800...	true [800.0 - 105...	true [1050.0 - 13...	true [1300.0 - 15...	true [1550.0 - ∞]	class precision
pred. [-∞ - 550.0]	503	372	110	38	7	4	48.65%
pred. [550.0 - 80...	1734	1911	723	268	97	66	39.82%
pred. [800.0 - 10...	223	658	539	184	68	35	31.58%
pred. [1050.0 - 1...	14	58	94	220	112	51	40.07%
pred. [1300.0 - 1...	16	23	36	299	614	346	46.03%
pred. [1550.0 - ∞]	7	34	30	60	201	371	52.77%
class recall	20.14%	62.53%	35.18%	20.58%	55.87%	42.50%	

Slika 60. Matrica konfuzije drugog klasifikacijskog modela

Ukupna točnost modela je vidljivo lošija. Dok je u prvom klasifikacijskom modelu točnost iznosila 58,57%, u drugom klasifikacijskom modelu se spustila za 17% te iznosi 41,06%

Model će znatno lošije procjenjivati odziv klase prvog intervala. Dok je ta točnost u prvom modelu iznosila 68%, u drugom modelu iznosi 20%. Odzive klasa drugog i petog intervala oba modela procjenjuju jednako, dok sve ostale odzive klasa prvi model procjenjuje bolje.

Isto se ne može reći za preciznost klasa. Sve intervale će bolje procjenjivati prvi model.

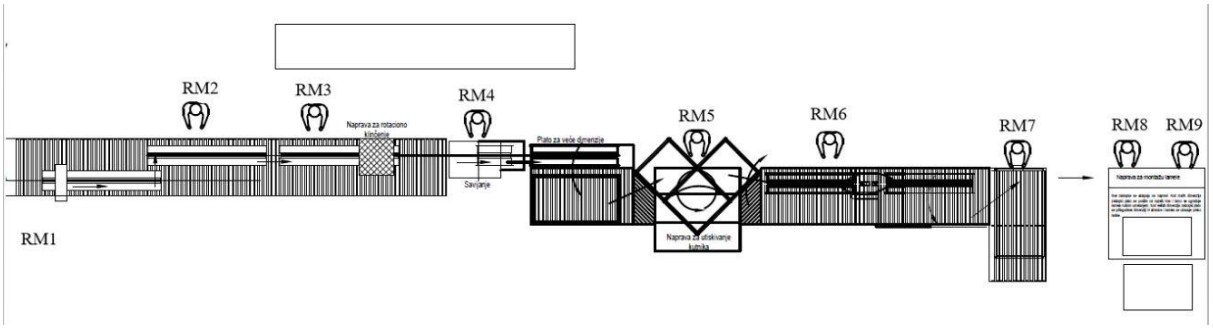
Dodavanjem klastera u klasifikacijsku analizu, nastojala se pospješiti ukupna točnost modela, što nije bilo uspješno. Težilo se boljem povezivanju podataka na temelju kojih će model bolje učiti. Ovakav rezultat je bio očekivan s obzirom raspodjele podataka po klasterima. Budući da je klaster analiza grupirala gotovo sve podatke u jedan klaster, takav rezultat je mogao samo negativno utjecati na rezultate klasifikacije. Model klasifikacije nastoji pronaći dublje, indirektne veze među atributima, a podaci koji su grupirani u jedan klaster ne pridonose pronalasku takvih veza. Smjer daljnje analize ovog klasifikacijskog modela bi bio izbacivanje dijela atributa iz ulaznih podataka za klaster analizu, koja bi rezultirala boljom raspodjelom podataka po klasterima, pa tako i boljim rezultatom klasifikacije.

Kvalitetni podaci su srž svake dubinske analize podataka. Raspršenost podataka o vremenu trajanja procesa i teško utvrdljiva povezanost s ostalim atributima ima za posljedicu nemogućnost kreiranja reprezentativnijih modela procesa. Dostupan skup podataka, prema istraženom, ne nosi dovoljnu količinu utjecajnih parametara procesa proizvodnje. Razlozi tome mogu se tražiti u samoj aktivnosti prikupljanja, odnosno načinu prikupljanja. Naime, podaci se u procesu prikupljaju pritiskom operatora na aktivacijske tipke, čime se označava početak i završetak procesa. U takvom procesu moguće su devijacije u proceduri, što posljedično uzrokuje netočan zapis o vremenu trajanja. S ciljem definiranja kvalitetnijeg procesa

prikupljanja, u sljedećem poglavlju dan je prijedlog optimizacije procesa prikupljanja podataka o vremenu trajanja linijske proizvodnje. Primijenjeni modeli su iskoristivi u slučaju postojanja kvalitetnog skupa ulaznih podataka. Cilj ovog rada je ostvaren kroz prikaz mogućnosti dubinske analize, ali isto tako su i prikazana ograničenja koja su uzrokovana nedovoljnom količinom atributa utjecajnih na ciljnu varijablu, u ovom slučaju vrijeme trajanja procesa.

7.4.5. Optimizacija rezultata proširivanjem ulaznog skupa podataka

Slika 61. prikazuje tlocrt operacijske linije od trenutka kada materijal uđe na proizvodnu liniju, sve do gotovog proizvoda.



Slika 61. Tlocrt operacijske linije

Kao što je vidljivo na slici, na operacijskoj liniji postoji devet radnih mjesta. Početak procesa bilježi se na radnom mjestu jedan (RM1), dok se kraj proizvodnje bilježi na radnom mjestu 9 (RM9). Trajanje izrade proizvoda između ova dva radna mjesta su ona koja se bilježe u bazi podataka kao *vrijeme_trajanja_sekundi* te se to vrijeme nastoji procijeniti. Nedostatak ovakvog prikupljanja vremena trajanja procesa je u tome što nedostaje informacija o tome što se događa između prvog i zadnjeg radnog mjesta. To se posredno može riješiti definiranjem broja jedinica u procesu (eng. *Work in progress – WIP)* od početka do kraja operacijske linije, tj. od ulaska sirovine u proizvodnju do izlaska proizvoda iz proizvodnje [39]. Više proizvoda na operacijskoj liniji, može rezultirati duljim trajanjem procesa. Procesu je potrebno dodati još jedan atribut, koji će računati trenutni broj proizvoda na operacijskoj liniji. Iskustva iz proizvodnje ukazuju na važnost tog podatka te je idući korak njegovo definiranje iz dostupnog skupa podataka.

Novi atribut definirati će se idućim ograničenjima:

$$vrijeme\ početka\ \leq vrijeme\ završetka\ (promatranog)$$

$$vrijeme\ početka\ \geq vrijeme\ početka\ (promatranog)$$

Na ovaj način obuhvatili su se svi proizvodi koji su na operacijskoj liniji u trenutku početka proizvodnje proizvoda, čiji se zapis promatra.

Novi, prošireni skup podataka uzima u obzir procese koji traju do sat vremena (3600 sekundi). Ulazni skup podataka prošao je isti proces pripreme podataka prikazan u poglavlju 7.1.5. Klasifikacijski model (prikazan u poglavlju 7.4.1.) izgrađen je na trening skupu koji iznosi 80% ulaznog skupa podataka, dok test skup iznosi ostalih 20%. Nakon pripreme podataka, model će se graditi na 45 539 podataka i 21 atributa. Raspodjela po spremnicima prikazana je na slici 62.

Index	Nominal value	Absolute count	Fraction
1	[-∞ - 849.7]	23832	0.523
2	[849.7 - 1399.3]	10601	0.233
3	[1399.3 - 1949.0]	7457	0.164
4	[1949.0 - 2498.7]	2558	0.056
5	[2498.7 - 3048.3]	836	0.018
6	[3048.3 - ∞]	255	0.006

Slika 62. Raspodjela podataka po spremnicima

Najviše podataka sadrži spremnik pod indeksom 1, dok najmanje sadrži zadnji spremnik.

Slika 63. prikazuje matricu konfuzije za optimirani klasifikacijski model.

accuracy: 79.26%

	true [-∞ - 849.7]	true [849.7 - 139...	true [1399.3 - 19...	true [1949.0 - 24...	true [2498.7 - 30...	true [3048.3 - ∞]	class precision
pred. [-∞ - 849.7]	5485	530	35	16	9	0	90.29%
pred. [849.7 - 13...	432	1747	273	17	4	5	70.50%
pred. [1399.3 - 1...	33	352	1404	283	39	8	66.26%
pred. [1949.0 - 2...	2	18	137	273	49	13	55.49%
pred. [2498.7 - 3...	6	3	15	48	105	29	50.97%
pred. [3048.3 - ∞]	0	0	0	2	3	9	64.29%
class recall	92.06%	65.92%	75.32%	42.72%	50.24%	14.06%	

Slika 63. Matrica konfuzije optimiranog klasifikacijskog modela

Točnost novog, optimiranog modela je 79.26%. Odziv klase je najlošiji za zadnji spremnik koji sadrži samo 255 podatka od ukupnih 45 539. Najbolja preciznost klase je u slučaju kada je trajanje procesa između 5 i 14 minuta i iznosi 90,29%, što znači da je model procijenio da je trajanje procesa u tom intervalu 6040 puta, od čega je točno procijenio 5485 puta. Odziv klase ovog intervala prati preciznost klase te je visokih 92,06%. Ovaj model posjeduje najbolju mogućnost procjene prema kriteriju točnosti.

Za daljnju analizu, preporuča se koristiti ovaj model. Nadalje, preporuča se dodavanje kontrolnih točaka (eng. *Check point)* na određenim radnim mjestima. Dodavanjem ovakvih

točaka, omogućila bi se mjerenja vremena trajanja procesa unutar operacijske linije (između dva ili više radnih mjesta) te bi se na taj način pomoglo u definiranju procesa koji uzrokuju usko grlo i usporavaju proizvodnju. Ovako redefiniran proces može dovesti do povećanja učinkovitosti prikupljanja podataka o trajanju proizvodnje.

Optimizacija procesa, a s time i smanjenje ukupnog vremena trajanja proizvodnje mogla bi se izvršiti kroz primjenu sistema povlačenja (eng. *Pull system)*. Ovakav tip proizvodnje značio bi da tek kad je jedan proizvod završen, može započeti proizvodnja idućeg proizvoda. Na ovaj način bio bi osiguran konstantan broj proizvoda na operacijskoj liniji te bi se spriječila pojava uskih grla, tj. podoperacija koji dovode do zastoja u proizvodnji. Za ovakav tip proizvodnje, potrebno je osigurati jaku dvosmjernu komunikaciju između zadnjeg i prvog radnog mjesta, kao i radnih mjesta unutar operacijske linije na kojima su kontrolne točke.

Zaključno, optimizacija procesa kojim bi se poboljšao proces prikupljanja podataka leži u mjerenju broja istovremenih proizvoda na operacijskoj liniji (*WIP*) koji dovode do zastoja u proizvodnji, zajedno s uvođenjem kontrolnih točaka koje bi pomogle u identificiranju radnih mjesta koji dovode do zastoja. Optimizacija cjelokupnog procesa postigla bi se uvođenjem tzv. *Pull* tipa proizvodnje kojim bi se osigurao konstantan broj proizvoda na operacijskoj liniji.

8. ZAKLJUČAK

U današnje konkurentno vrijeme u kojem su zahtjevi na tržištu sve veći, a konkurencija sve jača, neizbježna je upotreba metoda strojnog učenja kako bi se došlo do novih saznanja koja će poduprijeti proces proizvodnje i plasman na tržištu.

Razumijevanje područja i poslovnog cilja elementarni je dio svakog procesa strojnog učenja. Potrebno je razumjeti ulazne podatke i odrediti smjer kojim će se analiza kretati. Kvalitetni podaci srž su svakog kvalitetnog modela za učenje. Dobiveni ulazni set podataka zahtijeva pripremu u smislu njihova razumijevanja. U prvom redu, potrebno je analizirati raspršenost podataka jer o tome ovisi kvaliteta matematičkog modela. Eksplorativna analiza podataka pomaže pri boljem razumijevanju podataka, čime se u radu olakšala analiza i odredio smjer kretnje analize podataka. Pripremom podataka uklonila se skoro ¼ podataka koji su imali nedostajuće, nepravilne ili nerealne vrijednosti. Izrada novih atributa koji su povezani s već postojećim, kao i promjena tipa podataka potpomogli su dobivanju boljeg modela te otkrivanju novih znanja.

Regresijski model rezultirao je otkrivanjem povezanosti između atributa. Rezultati su upućivali na manjak povezanosti između nezavisnih varijabli i ciljne varijable. Podaci koji su većinom kategorički, kao i velika raspršenost podataka te slaba povezanost između ciljne varijable i ostalih nezavisnih varijabli dali su kao rezultat niže vrijednosti korelacije i model čija procjena ne zadovoljava potrebama procesa. Regresijski model koji radi s podacima koji predstavljaju stvarno stanje u poduzeću, uspješnije će procjenjivati buduće događaje. Rezultati takvog modela doprinijeti će donošenju kvalitetnih odluka, čiji će profit nadmašiti troškove ulaganja u dubinsku analizu podataka.

Klasifikacijski model, za razliku od regresijskog, koristio je veći broj atributa. Takvim pristupom nastojalo se pronaći ponašanje ili pravilo u podacima koje je nevidljivo samim pregledom podataka. Podjelom ciljane varijable po klasama, moguće je iščitati učestalosti ponavljanja trajanja procesa za pojedine intervale. Iako takav model nije rezultirao dobrim modelom za procjenjivanje, podjela po klasama otvorila je prostor za daljnja istraživanja s reprezentativnim klasama, koja će u velikoj mjeri doprinijeti razvoju procesa. Analizom klastera, koja spada u tehnike nenadziranog učenja, pokušalo se utjecati na kreiranje dodatnih atributa koji bi kvalitetnije opisivali proces.

Mogućnost provedbe temeljite analize nad podacima zasigurno je korak koji vodi razumijevanju događaja u procesu, a s time i donošenju kvalitetnijih odluka u budućnosti. Metode strojnog učenja omogućile su takve analize, a ovim diplomskim radom pokazana je mogućnost provedbe jedne takve analize i približavanju konceptu industrije 4.0. Industrija 4.0. bi u ovom poduzeću mogla napraviti značajne pomake, sa npr. ugradbom senzora koji će automatski pratiti vrijeme trajanja procesa, dimenzije i ostale vrijednosti bitne za poslovni cilj. Ti podaci bi u konačnici pridonijeli kvalitetnom i brzom predviđanju trajanja procesa koji otvaraju prostor za smanjene troškova, kvalitetno planiranje proizvodnje, smanjenje broja radnika i smanjenje nezadovoljavajućih proizvoda.

Optimizacija prikupljanja podataka u kombinaciji s tzv. *Pull* tipom proizvodnje može rezultirati poboljšanim procesom koji bi osigurao konstantan broj proizvoda na operacijskoj liniji, smanjenje vremena trajanja proizvodnje kao i pravilnije prikupljanje podataka o trajanju proizvodnje. Optimirani klasifikacijski model s uključenim atributom o zbroju proizvoda koji su istovremeno na operacijskoj liniji, dao je najbolje rezultate procjene vremena trajanja procesa.

Iz prethodnog navedenog, vidljivo je da postoji određena količina implicitnog znanja u prikupljenim podacima, koje nije moguće vidjeti samim pregledom podataka. Dubinska analiza podataka postaje ključna za izučavanje i ekstrakciju takvog znanja. Ovim radom prikazane su neke od mogućnosti koje ona nudi. Razvijeni procesi mogu poslužiti kao osnova prilikom definiranja procesa dubinske analize na novim skupovima podataka.

LITERATURA

[1] Han, J., Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001

[2] Institut Ruđer Bošković, Otkrivanje znanja dubinskom analizom podataka, Priručnik za istraživače i studente, http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf

[3] Statističke metode. Statistika-analiza. (Datum pristupa:03.06.2018.) http://statistika-analiza.com/index.php/vaznost-statistike

[4] D. Lisjak. Informatički menadžment. Podloge za predavanja. Fakultet strojarstva i brodogradnje. 2018.

[5] Dalbeo Bašić, Bojana, Umjetna inteligencija, uvod u strojno učenje, Fer, (Datum pristupa: 03.06.2018.) (http://degiorgi.math.hr/~singer/ui/ui_1415/UI_10_UvodUStrojnoUcenje.pdf

[6] A medium Corporation,9 Applications of Machine Learning from Day-to-Day life, 2017. Datum pristupa: 03.06.2018.) https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

[7] Lisjak, Dragutin. Rudarenje podataka. Menadžment u održavanju. 2018. Fakultet strojarstva i brodogradnje. Zagreb

[8] Brownlee, Jason. Supervised and Unsupervised Machine Learning Algorithms. 2016. Pristupljeno (03.06.2018.) https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[9] Domazet-Lošo, Mirjana. 2006. Usporedba postupaka dubinske analize primijenjenih nad biološkim podacima. Magistarski rad. Fakultet elektrotehnike i računarstva. Zagreb.

[10] Dalbeo Bašić, Bojana; Šnajder, Jan. Uvod u strojno učenje. 2015. (Datum pristupa:03.06.2018.)http://www.ieee.hr/_download/repository/SU-1-Uvod%5B1%5D.pdf

[11] Markić Ivan, Dubinsko pretraživanje, dohvaćanje i analiza digitalnih podataka, Kvalifikacijski doktorski ispit, Split, 05.03.2015.

[12] Jovanović, Milan. Dana mining alat. Centar za poslovno odlučivanje. Fakultet organizacionih nauka. Beograd. (Datum pristupa: 16.11.2018.) http://odlucivanje.fon.bg.ac.rs/wp-content/uploads/Orange-skripta.pdf

[13] Ceraj, Luka. 2015. Unaprijedna neuronska mreža s jednim ili dva skrivena sloja u predikciji ponašanja kaotičnog dinamičkog sustava. Završni rad. Fakultet strojarstva i brodogradnje. Zagreb

[14] Barber, David. Bayesian Reasoning and Machine Learning. 2016.

[15] H. Garcia-Molina, J. D. Ullman and J. Widom, Database Systems: The Complete Book, New Jersey 07458: Prentice Hal, 2002

[16] L Viktor Henra; Paquet Eric. Visualization Techniques for Data Minig. School of Information Technology and Engineering. Ontario, Canada. 2014.

[17] Gabelica Hrvoje, Rudarenje podataka i CRISP metodologija, 06.08.2013, http://www.skladistenje.com/rudarenje-podataka-i-crisp-metodologija/,

[18] Kraljević Goran, Rudarenje podataka, Sveučilište u Mostaru.

[19] Grljević Olivera, Bošnjak Zita, Primena CRISP-DM metodologije u analizi podatak o malim i srednjim poduzećima, http://www.academia.edu/1903275/Primena_CRISP-DM_metodologije_u_analizi_podataka_o_malim_i_srednjim_preduze%C4%87ima, (datum pristupa: 27.11.2017)

[20] Tošić Marina, Modeli za potporu pri odlučivanju o raspoloživosti zrakoplova temeljem dubinske analize podataka, Doktorski rad, Zagreb, 2017.

[21] Ungaro, Tea. Klasterska analiza. Diplomski rad. Prirodoslovno – matematički fakultet. Zagreb.2016

[22] Kapetanović, Lea. 2015.Primjena *cluster* analize u projektiranju proizvodnih sustava. Diplomski rad. Fakultet strojarstva i brodogradnje. Zagreb.

[23] Miliković, Vili. 2008. Analiza kvarova uređaja i opreme korištenjem metoda poslovnog izvješćivanja. Magistarski rad. Fakultet strojarstva i brodogradnje. Zagreb.

[24] Hyvarinen, Aapo. Unsupervised Machine Learning. 2015. University in Helsinki.

[25] Uremović, Kristina. 2016. Statističke metode grupiranja u analizi podataka. Diplomski rad. Fakultet strojarstva i brodogradnje. 2018.

[26] Al-Anazi, Sumayia; AlMahmoud, Hind; Al-Turaiki, Isra. Finding similar documents using different clustering techniques. Simposium on Dana Mining Applications. Saudi Arabia.

[27] Ćustić, Anja; Graf, Dajana; Petric Maretić, Grgur. A* algoritam - Ulice Manhattana. Umjetna inteligencija. Prirodoslovno – matematički fakultet. 2008.

[28] Gandhi, Rohith. Introduction to Machine Learning Algorithms: Linear Regression. 27.05.2018. https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a(Datum pristupa: 16.11.2018.)

[29] Brownlee, Jason. Difference Between Classification and Regression in Machine Learning. 11.12.2017.https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/(Datum pristupa: 16.11.2018.)

[30]Šmuc, Tomislav. Strojno učenje, Struktura metoda/ algoritama strojnog učenja. Prirodoslovno – Matematički fakultet. 2017.

[31] Yaghini, Masoud. Data mining, Nearest – Neighbor Classification. 2009. http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_881/DM_04_06_Nearest-Neighbor%20Classification.pdf. (Datum pristupa: 19.11.2018.)

[32] Hrvatski izričaj, Geomatematički pojmovnik. http://e.math.hr/category/klju-ne-rije-i/hrvatski-izri-aj. (Datum pristupa: 16.11.2018.)

[33]Correlation Matrix (Concurrency), RapidMiner. https://docs.rapidminer.com/latest/studio/operators/modeling/correlations/correlation_matrix.html (Datum pristupa: 16.11.2018.)

[34] Importatnt terms, RapidMiner. https://docs.rapidminer.com/latest/studio/getting-started/important-terms.html . (Datum pristupa: 16.11.2018.)

[35] Cook Darren. Practical Machine Learning with H2O. O'Reilly Median. Sebastopol 2016.

[36] Dalbeo Bašić Bojana; Čupić Marko; Šnajder Jan. Umjetne neuronske mreže. Fakultet elektrotehnike i računarstva. 2008.

[37] Candel Arno; LeDell, Erin. Deep learning booklet. H2O.ai Inc. Mountain View. 2018.

[38] Ljubić Hrvoje. Predviđanje ishoda učenja pomoću neuronskih mreža u okruženju konceptualnih mapa. Diplomski rad. Fakultet prirodoslovno matematičkih i odgojnih znanosti. Mostar. 2018.

[39] Hegedić Miro, Gudlin Mihael. Proizvodni menadžment. Podloge za predavanja. Fakultet strojarstva i brodogradnje. Zagreb. 2017.