Proces otkrivanja znanja primjenom tehnika rudarenja podataka

Ivandić, Viktorija

Master's thesis / Diplomski rad

2016

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:235:748117

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

DIPLOMSKI RAD

VIKTORIJA IVANDIĆ

Zagreb, godina 2016.

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

Proces otkrivanja znanja primjenom tehnika rudarenja podataka

Mentor: Student:

Prof. dr. sc. Dragutin Lisjak Viktorija Ivandić

Zagreb, godina 2016.

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

Ovom prilikom želim zahvaliti mentoru prof. dr. sc. Dragutinu Lisjaku i asistentici Marini Tošić, mag. ing. mech., na strpljivosti, stručnoj pomoći i savjetima.

Također želim zahvaliti svojim roditeljima što su mi omogućili odlazak na studij i bili velika potpora tijekom studiranja. Želim se zahvaliti i sestri, dečku, svim prijateljima i kolegama koji su vjerovali u mene kada ja sama nisam, te mi pružali potporu i pomoć kada je trebalo.

Viktorija Ivandić

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

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

DIPLOMSKI ZADATAK

Student: **VIKTORIJA IVANDIĆ** Mat. br.: 0035181787

Naslov rada na hrvatskom jeziku: **OTKRIVANJA ZNANJA PRIMJENOM TEHNIKA RUDARENJA PODATAKA**

Naslov rada na engleskom jeziku: **KNOWLEDGE DISCOVERY USING DATA MINING TECHNIQUES**

Opis zadatka:

Razvojem informacijskih sustava omogućena je pohrana velike količine podataka. Pohranjeni podaci većinom se koriste za praćenje informacija u određenom procesu ili za izvještavanje o prošlim aktivnostima u procesu. Vremenom je uočena važnost analize prošlih podataka jer se na temelju njih u budućnosti mogu donositi bolje poslovne odluke. Zbog velikih količina podataka, analiza istih nije više moguća ručnom obradom te zahtijeva naprednije tehnike obrade. Zbog toga su razvijene tehnike rudarenja podataka kako bi se u cilju otkrivanja znanja pohranjeni podaci što bolje iskoristili. Rudarenje podataka podrazumijeva primjenu matematičkih metoda i modela te alata za otkrivanje skrivenog znanja unutar određenog skupa podataka. Primjena tehnika rudarenja podataka još uvijek nije dovoljno zastupljena unutar područja industrijskog inženjerstva te su potrebna daljnja istraživanja unutar tog područja.
U radu je potrebno obraditi sljedeće:

1. Dati literaturni pregled o procesu rudarenja podataka te opisati vrste strojnog učenja.
2. Opisati najčešće korištene metode strojnog učenja.
3. Detaljno opisati proces rudarenja podataka.
4. Za realni skup podataka predložiti metode koje će se koristiti za otkrivanje znanja.
5. Primjenom softverskog alata "RapidMiner" potrebno je prikazati način implementacije predloženih metoda za otkrivanje znanja.
6. Zaključak.

Zadatak zadan:	Rok predaje rada:	Predviđeni datum obrane:
29. rujna 2016.	1. prosinca 2016.	7., 8. i 9. prosinca 2016.

Zadatak zadao: Predsjednik Povjerenstva:

Prof. dr. sc. Dragutin Lisjak Prof. dr. sc. Franjo Cajner

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

POPIS JEDNADŽBI

POPIS OZNAKA

Oznaka	Jedinica	Opis
μ	-	Srednja vrijednost
x	-	Podatak
N	-	Ukupan broj podataka populacije
σ	-	Standardno odstupanje
σ^2	-	Varijanca

POPIS SKRAĆENICA

Skraćenica	Opis
PI	Poslovna inteligencija
BI	Business inteligence
FCM	Fuzzy C-means
VFR	Visual flight rules
IFR	Instrument flight rules

SAŽETAK

U radu su prikazane teorijske osnove poslovne inteligencije i rudarenja podataka. Detaljno je opisan proces rudarenja podataka koji je primijenjen na setu podataka o vojnim i civilnim zrakoplovnim nesrećama koje su se dogodile u razdoblju od 1908. do 2009. godine. Podaci su najprije deskriptivno analizirani te su modelirani procesi klasifikacije, klasterizacije i tekstualne analize. Procesi su rezultirali predikcijom na temu hoće li biti preživjelih putnika u nesreći, grupiranjem tipova zrakoplova, te asocijativnim pravilima koja daju informaciju o najčešćim uzrocima zrakoplovnih nesreća. Opisano je i kako se ti rezultati mogu iskoristiti u budućnosti.

Ključne riječi: poslovna inteligencija, rudarenje podataka, klasifikacija, klasterizacija, tekstualna analiza

SUMMARY

This paper presents the theoretical foundations of business intelligence and data mining. It describes in detail the process of data mining applied to a data set of military and civil aviation accidents that occured in the period from 1908 to 2009. The data was first analyzed descriptively. In addition to that the classification, clustering and text analysis processes were modeled. The processes have resulted in the prediction of the topic if there will be surviving passengers after occured accident, grouping similar aircraft types based on the overall accident occurance, and associative rules that provide information about the most common causes of these kind of accidents. In the end, a description how these results could be used in the future research has been given.

Key words: business intelligence, data mining, classification, clustering, text mining

1. UVOD

Korištenjem informacijskih sustava počele su se sakupljati velike količine podataka koje sadržavaju korisna znanja i informacije o prošlim događajima i procesima. Tek nedavno je otkriven potencijal analiziranja tih podataka i otkrivanje sakrivenih informacija. Ponekad te informacije nisu vidljive „golim okom“, već zahtijevaju stručna znanja i specijalne alate za otkrivanje. Tako je započeo razvoj poslovne inteligencije (PI) tj. kontinuiranog procesa koji se sastoji od različitih metoda i koncepata za obradu podataka s ciljem lakšeg i uspješnijeg donošenja poslovnih odluka.

Budući da iz dana u dan važnost PI raste, njen razvoj i nastanak će biti prikazan u drugom poglavlju.

Rudarenje podataka, statistička analiza i prediktivna analitika nisu novi pojmovi, no ono što ih je promijenilo jest način kako su integrirani u PI jer je menadžment prepoznao koliko se široko mogu primjenjivati ove analize. Različite tehnike rudarenja podataka su objašnjene u trećem poglavlju.

Analizom literature uočeno je da je proces rudarenja podataka zahtjevan i dugotrajan te uključuje određene faze i korake. Kako bi se analitičarima olakšao posao te proces rudarenja učinio kraćim i kvalitetnijim, razvijeni su razni softverski alati. Različite faze rudarenja podataka te načini odabira metoda i alata za otkrivanje znanja će biti prikazani u četvrtom poglavlju.

Nakon usvajanja svih faza rudarenja podataka, proces otkrivanja znanja će biti detaljno prikazan u petom poglavlju na setu koji sadrži informacije o civilnim i vojnim zrakoplovnim nesrećama. Temeljem različitih tehnika i metoda rudarenja podataka, otkrivena znanja će biti prikazana u šestom poglavlju.

Naposljetku, dan je zaključak o razvoju ovog područja.

2. POSLOVNA INTELIGENCIJA

Informatizacijom tvrtki došlo je do prikupljanja ogromnih količina podataka, tj. do tzv. eksplozije podataka. Gomilanjem podataka nastajale su nove baze podataka u kojima se s vremenom otkrio potencijal za poboljšanje poslovanja. No, za dolaženje do informacija iz tih podataka, a uz to i novih znanja, potrebno je bilo razviti alate koji bi taj proces omogućili i ubrzali. Tako je počeo razvoj poslovne inteligencije (PI).

Postoje različite verzije definicije poslovne inteligencije:

1. *Poslovna inteligencija predstavlja ranije prikriveno znanje koje se otkriva iz operativnih, rutinskih, prikupljanih poslovnih podataka primjenom odgovarajućih računsko-logičkih metoda, obično podržavanih informacijskom tehnologijom [6].*
2. *Poslovna inteligencija je skup metodologija i koncepata za prikupljanje, analizu i distribuciju informacija uz pomoć različitih softverskih alata. Ona je jedna od tehnika poslovnog izvještavanja, koja omogućuje pronalaženje informacija potrebnih za lakše i točnije donošenje poslovnih odluka [7].*
3. *Poslovna inteligencija je pristup obradi podataka koji želi transformirati podatke u informacije, a informacije u znanje te tako pomoći „inteligentnom" ponašanju poduzeća. Poslovna se inteligencija ostvaruje u organiziranom integriranom informacijskom sustavu s usklađenim transakcijskim, analitičkim i ostalim vrstama obrada podataka kao što su rudarenje podataka i obrada polustrukturiranih sadržaja [8].*

Iz navedenog se može zaključiti da je poslovna inteligencija zapravo kontinuirani proces koji se sastoji od različitih metodologija i koncepata koji služe za obradu podataka njihovim prikupljanjem, analizom i distribucijom u svrhu lakšeg i uspješnijeg donošenja poslovnih odluka.

Pojmovi potrebni za razumijevanje ovog područja te začeci PI nalaze se u nastavku.

2.1. Povijest PI i osnovni pojmovi

Pojam „poslovna inteligencija - PI" (engl. *Business Intelligence* – BI) prvi put je koristio H. P. Luhn u članku naslova „*A Business Intelligence System*" objavljenom u IBM

istraživačkom dnevniku 1958. Luhn je definirao PI kao „sposobnost razumijevanja međuveza prezentiranih činjenica na takav način koji bi usmjerio akcije prema željenom cilju."
Sljedećih 30 godina, originalni koncept se razvijao kroz različite faze: sustavi za potporu odlučivanju DSS (engl. *Decision Support Systems*) i EIS (engl. *Executive Information Systems*). Ali glavna prekretnica se dogodila kada je 1989. Howard Dresner, analitičar u Gartner Inc., opisao poslovnu inteligenciju kao „koncepte i metode za poboljšanje poslovnih odluka nastalih korištenjem sustava na bazi činjenica." Većina posla učinjenog u tom periodu bila je fokusirana ka tehnologijama, standardima, procesima i alatima za podršku prikupljanju, racionalizaciji skladišta i dohvata podataka te kreiranju izvještaja. Nakon toga sve se promijenilo, uvelike upravljano disciplinom starom 2500 godina – statistikom [1].

Za bolje razumijevanje definicije poslovne inteligencije najprije je potrebno objasniti osnovne pojmove:
Podatak – jednostavna, neobrađena, izolirana, misaona činjenica koja ima neko značenje. Podaci se pamte, zapisuju i bilježe na način koji im je primjeren i koji im odgovara. Struktura podatka je apstraktna i čine ju: značenje (naziv i opis značenja određenog svojstva), vrijednost (mjera i iznos) i vrijeme [2].
Informacija – rezultat analize i organizacije podataka na način da daje novo znanje primatelju. Ona postaje znanje kad je interpretirana, odnosno stavljena u kontekst ili kad joj je dodano značenje. Informaciju čine podaci kojima je dano značenje putem relacijskih veza, odnosno organizirani podaci koji su uređeni za bolje shvaćanje i razumijevanje [3].
Znanje – prikladna kolekcija informacija i to takva da se može smatrati korisnom. Znanje je deterministički proces. Definira se tako da se referira na informacije koje su na neki način organizirane, procesuirane ili strukturirane [4].
Inteligencija – pojam je nastao od latinskih riječi *inter* (hrv.-među) i *legere* (hrv-brati, skupljati). Kombinacija tih pojmova tvori značenje koje se odnosi na uviđanje međuveza ili međuodnosa pojmova. Prema Rječniku hrvatskog jezika (Anić, 2007., p. 149.), inteligencija se tumači kao: sposobnost shvaćanja i brzog snalaženja u novim prilikama, sposobnost otkrivanja zakonitosti u odnosima među činjenicama i rješavanju problema, oštroumnost te pamet [5].
Mudrost – ekstrapolacijski i nedeterministički proces koji se poziva na prethodne nivoe svijesti, posebice na kategorije kao što su moral, etički kodovi. Ona je esencija filozofskog promišljanja. Mudrost je proces kojim procjenjujemo što je dobro ili loše, ispravno ili krivo [4].

Informacija se definira preko podataka, znanje preko informacija, inteligencija preko znanja, a mudrost preko inteligencije. Slika 1. prikazuje model hijerarhijskih i funkcionalnih odnosa između navedenih pojmova. Nakon prikaza hijerarhije i razlike između osnovnih pojmova može se razumjeti proces poslovne inteligencija prikazan i sljedećem podpoglavlju.

2.2. Proces poslovne inteligencije

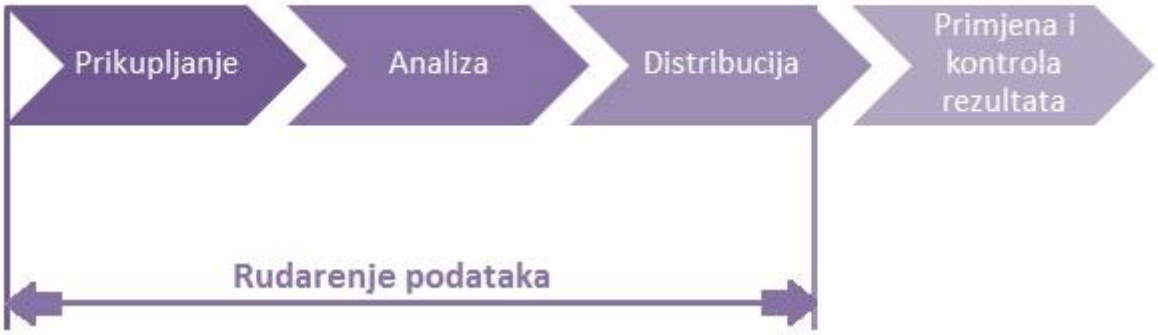
Proces PI je dugotrajan i složen, a sastoji od sljedeće četiri faze:

Prikupljanje podataka – prikupljanje dostupnih, sirovih podataka iz vanjskih i unutarnjih izvora

Analiza podataka – pregledavanje i ocjenjivanje prikupljenih podataka, davanje smisla informacijama i njihova nadogradnja u inteligenciju, pronalaženje uzoraka i međuodnosa među njima, i sve to uz znanstveni pristup, statistički softver i poznavanje tehnika modeliranja

Distribucija – završna faza procesa u kojoj se treba isporučiti gotove inteligentne proizvode donosiocima odluka

Primjena i kontrola rezultata – primjena rezultata istraživanja, osiguravanje povratnih veza i informacija te procjena novonastalog stanja i potreba



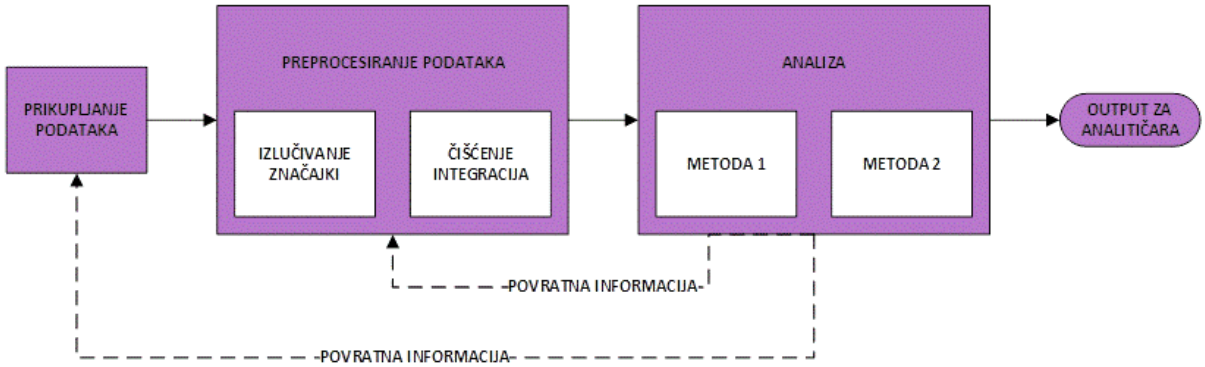
Slika 1. Faze procesa poslovne inteligencije [9]

Glavne faze procesa prikazane na slici 1. sastoje se od još nekoliko podprocesa i podfaza, ovisno o kompleksnosti zadatka i problema koji se pokušavaju riješiti. Postoje mnoge metode i koncepti te se svakog dana radi na razvijanju novih. Jedan od ključnih procesa PI je rudarenje podataka. Ono obuhvaća prikupljanje i analizu podataka bez čega PI ne bi ni postojala čime se bavi sljedeće podpoglavlje.

2.3. Rudarenje podataka

U ogromnim bazama podataka traže se upravo oni podaci koji čine ključne informacije za donošenje važnih poslovnih odluka koje garantiraju uspjeh.

Rudarenje podataka (engl. Data Mining) je prikupljanje, čišćenje, obrađivanje, analiziranje i dobivanje korisnih saznanja iz podataka [10].



Slika 2. Proces rudarenja podataka [10]

Na slici 2. prikazan je proces rudarenja podataka, tj. odnos između pojedinih faza procesa. Svaka od tih faza opisana je u nastavku.

Faze rudarenja podataka [10]:

1. **Prikupljanje podataka** – može zahtijevati specijalizirane hardvere kao što su mreža senzora, ručni rad kao što je prikupljanje anketa, ili softverske alate. Dobar odabir alata može značajno utjecati na cijeli proces. Nakon prikupljanja, podaci se najčešće pohranjuju u baze podataka (skladište podataka).
2. **Izlučivanje značajki i čišćenje podataka (preprocesiranje)** – Izlučivanje značajki se obično izvodi paralelno s čišćenjem podataka kojim se ispravljaju i procjenjuju podaci, neispravni i oni koji nedostaju. U mnogim slučajevima podaci se prikupljaju iz različitih izvora i moraju se integrirati u univerzalni format za rudarenje.
3. **Analitički procesi i algoritmi** – završna faza procesa rudarenja je konstruiranje učinkovite analitičke metode za obrađivanje podataka. U puno slučajeva neće biti moguća direktna uporaba standardnih metoda rudarenja podataka („superproblems“).

Tri vrste učenja koje će biti opisane u sljedećem poglavlju, pokrivaju puno slučajeva koji se razbijaju na manje komponente koje koriste ove različite metode.

3. TEHNIKE RUDARENJA PODATAKA

Tehnike rudarenja podataka dijelimo prema željenom ishodu učenja, odnosno želimo li nešto predvidjeti ili naći grupe podataka sa sličnim karakteristikama. Tehnike su podijeljene u tri glavne grupe od kojih je svaka detaljno opisana u nastavku.

3.1. Nadzirano učenje

Nadzirano učenje je tehnika strojnog učenja koja koristi poznati set podataka (tzv. trening set) za predviđanje. Trening set podataka uključuje ulazne varijable i odgovarajuće ciljne varijable. Na osnovu toga algoritam za nadzirano učenje gradi model koji može predvidjeti ciljne varijable kod novih setova podataka. Često se koristi test set podataka za validaciju modela. Korištenjem što većih trening setova podataka dobivaju se precizniji modeli s boljim prediktivnim rezultatima.

Nadzirano učenje uključuje dvije kategorije algoritama [14]:

- **Klasifikacija:** za kategoričke ciljne varijable, gdje podaci mogu biti razvrstani u specifične „klase“
- **Regresija:** za kontinuirane ciljne varijable

3.2. Nenadzirano učenje

Nenadzirano učenje je tehnika strojnog učenja kojom se pokušavaju utvrditi pravila u setu podataka koji se sastoji samo od ulaznih varijabli, bez ciljnih varijabli.

Najčešća metoda nenadziranog učenja je klaster analiza (grupiranje), koja se koristi za istraživačku analizu podataka, za pronalazak skrivenih uzoraka ili grupa u podacima.

Česte metode klasterizacije su [14]:

- **Hijerarhijsko grupiranje**: gradi višerazinsku hijerarhiju klastera kreirajući klaster drvo
- **k-Means klasterizacija**: gradi model u kojem svaka grupa ima svoju srednju vrijednost (centroid). Svaki primjer pripada grupi čiji mu je centroid najbliži (po euklidskoj udaljenosti)
- **Gaussov model**: modelira klastere kao mješavinu multivarijantnih komponenti normalne gustoće
- **Samoorganizirajuće mape**: koristi neuronske mreže koje uče topologiju i distribuciju podataka

3.3. Podržano učenje

Podržano učenje je tehnika strojnog učenja koja omogućava računalima i softverskim agentima automatsko određivanje idealnog ponašanja unutar specifičnog konteksta, s ciljem maksimalizacije performansi. Kako bi agent naučio svoje ponašanje potrebna mu je jednostavna povratna informacija u obliku nagrade.

Podržano učenje je definirano za specifičan tip problema, i sva njihova rješenja su klasificirana kao algoritmi podržanog učenja. Kod problema agent bi trebao odlučiti koja je najbolja radnja bazirano na trenutnom stanju. Kada se ovaj korak ponavlja, problem je poznat kao Markovljev proces [15].

Poznatiji algoritmi podržanog učenja su [16]:

- **Q-learning**: to je akcija koja maksimizira sumu trenutne i odgođene nagrade u slučaju da slijedimo optimalnu strategiju
- **Temporal difference learning**: kombinacija Monte Carlo ideje i dinamičkog programiranja

4. PROCES RUDARENJA PODATAKA

Već je spomenuto da se proces rudarenja podataka sastoji od tri faze koje su prikupljanje podataka, preprocesuiranje te analiza podataka. Faza preprocesuiranja koja slijedi nakon prikupljanja podataka, krucijalna je faza procesa rudarenja podataka. Rijetko joj se daje dovoljno pažnje jer se većina fokusa stavlja na samu analizu podataka. Analiza podataka je također jako bitna, ali bez pravilne pripreme i odabira podataka ni analiza neće dati zadovoljavajuće rezultate. Faza preprocesuiranja sastoji se od sljedećih koraka [10]:

1. ***Izlučivanje značajki***: Analitičar može biti suočen s ogromnim količinama sirovih dokumenata, logiranja u sustav, ili trgovačkih transakcija, bez uputa kako bi se ti sirovi podaci trebali transformirati u smislene značajke baze podataka za procesuiranje. Ova faza jako puno ovisi o mogućnosti analitičara da izluči najrelevantnije značajke za traženu primjenu. Ta mogućnost uvelike zahtijeva razumijevanje specifičnog područja primjene.
2. ***Čišćenje podataka***: Odabrani podaci mogu sadržavat nepravilne ili prazne unose. Točnije, neke zapise treba odbaciti, ili prazne unose aproksimirati. Nedosljednosti se trebaju ukloniti.
3. ***Odabir značajki i transformacija***: Kada podaci imaju puno dimenzija, mnogi algoritmi rudarenja podataka nisu efikasni. Mnogo više-dimenzioniranih značajki sadrži šumove i mogu izazvati greške u procesu rudarenja podataka. Postoje mnoge metode koje služe ili za uklanjanje irelevantnih značajki ili za transformaciju trenutnog seta značajki u novi podatkovni prostor koji je prikladniji za analizu. Postoji i transformacija seta podataka s određenim setom atributa u set podataka s drugim setom atributa istog ili drugačijeg tipa.

Kada završi ova faza počinje analitička faza procesa rudarenja podataka. Tu je najvažnije odabrati metodu koja najbolje obuhvaća problem i cilj procesa. Svaka primjena rudarenja podataka je jedinstvena i teško je kreirati model koji je općenit i primjenjiv u različitim područjima. Ali različite formulacije rudarenja podataka mogu se iskoristiti u kontekstu različitih primjena rudarenja podataka, ovisno o vještini i iskustvu analitičara. Faze su detaljnije objašnjene u nastavku poglavlja.

4.1. Prikupljanje i čišćenje podataka

Prva faza procesa rudarenja podataka je kreiranje seta podataka s kojima će analitičar moći raditi. U slučajevima gdje su podaci u sirovom i nestrukturiranom obliku (npr. sirovi tekst, signali senzora), moraju se izlučiti relevantne značajke za procesuiranje. U nekim slučajevima gdje su dostupne heterogene mješavine značajki u različitim formama, često nema već gotovog analitičkog postupka za procesuiranje takvih podataka. U takvim slučajevima potrebno je transformacija u oblik pogodan za proces.

Oblik i stanje podataka ovisi o domeni iz koje dolaze [10]:

1. **Senzorski podaci**: Senzorski podaci se često prikupljaju kao velike količine niskorazinskih signala, koji su masivni. Niskorazinski signali se ponekad konvertiraju u visokorazinske značajke koristeći wavelet ili Fourierove transformacije. U ostalim slučajevima vremenske serije se koriste direktno nakon čišćenja. Ove tehnologije su također korisne za prenošenje vremensko-serijskih podataka u multidimenzionalne podatke.
2. **Slikovni podaci**: U najprimitivnijoj formi, slikovni podaci su predstavljeni kao pikseli. Na malo višem nivou, histogrami u boji se mogu koristiti za predstavljanje značajki u različitim segmentima neke slike. U zadnje vrijeme je postalo popularnije korištenje *vizualnih riječi.* To je semantički bogat prikaz sličan dokumentima. Izazov kod procesuiranja slika je taj da su podaci općenito jako visoko dimenzionirani. Izlučivanje značajki može se izvoditi na različitim razinama. ovisno o primjeni.
3. **Web logovi**: Web logovi se obično prikazuju kao tekst stringovi u predodređenom formatu. Zbog toga što su ti logovi određeni i odvojeni, relativno je lako konvertirati Web pristupne logove u multidimenzionalni prikaz (relevantnih) kategoričkih i numeričkih atributa.
4. **Mrežni promet**: U mnogim aplikacijama za detekciju provala, karakteristike mrežnih paketa se koriste za analizu provala ili drugih zanimljivih aktivnosti. Ovisno o osnovnoj aplikaciji, iz tih paketa se mogu izlučiti razne značajke, kao što je broj transferanih bajtova, korišten mrežni protokol, itd.
5. **Dokumenti**: Dokumenti su često dostupni u sirovoj i nestrukturiranoj formi, i podaci mogu sadržavati bogate lingvističke relacije između različitih entiteta. Jedan pristup je

da se uklone stop riječi, zadrže podaci, i koriste bag-of-words prikaz. Druge metode koriste izlučivanja entiteta za određivanje lingvističkih veza.

Prikupljanje podataka i izlučivanje značajki je umjetnost koja jako puno ovisi o vještini analitičara da odabere značajke i njihovu prezentaciju koja najviše odgovara zadatku koji se rješava. Ako nisu odabrani pravi atributi, analiza može biti dobra samo onoliko koliko i dostupni podaci [10].

Prikupljeni podaci su često heterogeni i mogu sadržavati različite tipove podataka. To stvara izazov za analitičara koji mora izraditi algoritam s proizvoljnim tipovima podataka. Heterogeni tipovi podataka onemogućavaju analitičaru korištenje već gotovih algoritama, a traženje i korištenje algoritama specificiranih za određene kombinacije tipova podataka je nepraktično i zahtijeva puno vremena. Zbog toga postoji potreba za pretvaranjem različitih tipova podataka. Teži se korištenju numeričkih tipova podataka jer su oni najzastupljeniji u algoritmima rudarenja podataka. No to ne isključuje pretvorbe u drugačije tipove podataka. Tablica 1. prikazuje pretvorbe među tipovima podataka.

Tablica 1. Pretvorbe tipova podataka [10]

Izvorni tip podatka	Željeni tip podataka	Metoda
Numerički	Kategorički	Diskretizacija (klasterizacija, asocijativna pravila)
Kategorički	Numerički	Binarizacija (klasifikacija, regresija)
Tekstualni	Numerički	Latentna semantička analiza (LSA)

Neke od metoda detaljnije će biti objašnjene u poglavlju 5.1.

Proces čišćenja podataka je važan zbog grešaka povezanih s procesom prikupljanja podataka. Neki izvori sadržavaju prazne ulaze i greške koje se mogu pojaviti u podacima. Slijede neki primjeri [10]:

1. Neke tehnologije za prikupljanje, kao što su senzori, su svojstveno netočne zbog ograničenja hardvera povezanih s prikupljanjem i prijenosom. Ponekad senzori mogu preskočiti očitanje zbog greške u hardveru ili prazne baterije.

2. Podaci prikupljeni korištenjem tehnologija za skeniranje mogu sadržavati greške povezane s tehnologijom optičkog prepoznavanja karaktera su daleko od savršenog. Podaci koji nastaju pretvaranjem govora u tekst također su podložni greškama.
3. Korisnici možda ne žele dati tražene informacije iz privatnih razloga, ili namjerno upisuju netočne vrijednosti. Na primjer, primijećeno je da korisnici ponekad upisuju krivi datum rođenja na stranicama s automatskom registracijom kao što su socijalne mreže. U nekim slučajevima, korisnici mogu odabrati hoće li nekoliko polja ostaviti praznima.
4. Značajna količina podataka se upisuje ručno. U takvim slučajevima česte su greške kod upisivanja podataka.
5. Neki subjekti odgovorni za prikupljanje podataka neće prikupiti određena polja kod zapisa, ako su preskupi. Zbog toga zapisi možda neće biti potpuno specificirani.

Ovi problemi mogu biti značajni izvori nepravilnosti u rudarenju podataka. Potrebne su metode kojima se uklanjaju i ispravljaju podaci koji nedostaju ili su nepravilno uneseni. Ovo je nekoliko važnih aspekata čišćenja podataka [10]:

1. **Rukovanje nedostajućim ulazima**: Mnogi ulazi u podacima mogu ostati neodređeni zbog nepravilnosti pri prikupljanju podataka ili inherentnosti prirode podataka. Takvi nedostajući ulazi se možda mogu aproksimirati. Proces aproksimacije nedostajućih ulaza se također naziva imputacija.
2. **Rukovanje netočnim ulazima**: U slučajevima kada su iste informacije dostupne iz više izvora mogu se detektirati nedosljednosti. One se uklanjaju kao dio analitičkog procesa. Druga metoda za detektiranje netočnih ulaza je korištenje znanja određene domene o tome što se već zna o tim podacima. Općenitije, podaci koji su nedosljedni s distribucijom preostalih podataka često su šum. Takvi podaci su poznatiji kao iznimke. Ali opasno je pretpostaviti da su ti podaci uvijek uzrokom greške.
3. **Skaliranje i normalizacija**: Podaci se često mogu prikazati u različitim skalama (npr. godine i plaća). To može uzrokovati da su neke značajke nenamjerno precijenjene pa se druge značajke implicitno ignoriraju. Zbog toga je važno normalizirati različite značajke.

Priprema podataka je dugotrajan posao, ali i krucijalan. Bez dobrih i pripremljenih podataka nema ni dobre analize i odluka. Nakon što se prikupe svi podaci koji bi mogli biti relevantni za proces rudarenja podataka kreće njihova transformacija i redukcija za nastavak analize.

4.2. Redukcija i transformacija podataka

Cilj redukcije podataka je njihov kompaktniji prikaz. Kada je količina podataka mala , puno je lakše primijeniti sofisticirane i računski zahtjevne algoritme. Redukcija podataka se može odnositi na smanjenje broja redova (zapisa) ili broja kolona (dimenzija). Redukcija podataka uzrokuje određeni gubitak informacija. Korištenje sofisticiranijih algoritama može kompenzirati gubitak informacija nastalih redukcijom podataka. Različite redukcije podataka se koriste u različitim slučajevima [10]:

1. *Uzorkovanje podataka*: Zapisi iz osnovnih podataka se uzorkuju kako bi se kreirale manje baze podataka. Uzorkovanje je općenito znatno teže u slučajevima gdje se uzorci moraju dinamički održavati.
2. *Selekcija značajki*: Samo se podskup značajki iz osnovnih podataka koristi u analitičkom procesu. Taj podskup se bira na osnovi toga za što se primjenjuje. Na primjer, izbor značajki koja je pogodna za klasterizaciju možda neće biti dobra za klasifikaciju, i obrnuto.
3. *Redukcija podataka i osna rotacija*: Korelacije među podacima se mogu iskoristiti za njihov prikaz s manjim brojem dimenzija. Primjeri takvih metoda redukcije podataka uključuju analizu glavnih komponenti (eng. *principal component analysis* – PCA), dekompoziciju jedinstvenih vrijednosti (eng. *singular value decomposition* – SVD), ili latentnu semantičku analizu (eng. *latent semantic analysis* – LSA) za tekstualnu domenu.
4. *Redukcija podataka s transformacijom tipa*: Ovaj oblik redukcije podatka je strogo povezan s prenosivošću tipa podataka. Na primjer, vremenske serije se konvertiraju u multidimenzionalne podatke manje veličine i složenosti pomoću diskretne wavelet transformacije. Slično, grafovi se mogu konvertirati u multidimenzionalne prikaze korištenjem ugradbenim tehnikama.

4.3. Odabir metoda rudarenja podataka

Kod odabira metode koja će se koristiti za rudarenje podataka bitno je znati željeni cilj, odnosno koju vrstu rezultata se želi dobiti. Prema vrsti rezultata koji se želi dobiti mogu se koristi neke od metoda navedenih u nastavku:

Binomna varijabla (1 ili 0)

Predikcijom se pokušava predvidjeti ciljani atribut, odnosno hoće li njegov iznos biti 1 ili 0. Neki od najčešćih operatora koji to omogućavaju su:

a) *Rule induction* - operator radi s numeričkim, polinominalnim i binominalnim atributima, te također može predvidjeti i takve rezultate. Radi na temelju modificiranog RIPPER algoritma koji se kreće manje relevantnim klasama te zatim iterativno raste i obrezuje dobivena pravila sve dok ne ukloni pozitivne primjere ili greška algoritma ne bude veća od 50%. U fazi rasta, u svako pravilo dodaju se pohlepni uvjeti dok pravilo ne bude savršeno(100% točno). Procedura isprobava svaku moguću vrijednost za svaki atribut i selektira uvjet s najvećom informacijskom dobiti.

b) *Naive Bayes* – klasifikator Naive Bayes je jednostavni probabilistički klasifikator koji se temelji na primjeni Bayesovog teorema (iz Bayesove statistike) s jakim (naivnim) neovisnim pretpostavkama. Klasifikator pretpostavlja da prisutnost (odsutnost) određene značajke neke klase (ili atribut) je nepovezan s prisutnošću (odsutnošću) bilo koje druge značajke. Prednost ovog klasifikatora je ta da zahtijeva malu količinu trening podataka za procjenu sredstava i varijanci potrebnih za klasifikaciju.

c) *Decision Tree* – graf ili model u obliku stabla. Ono je više kao izokrenuto stablo jer mu se korijeni nalaze na vrhu i raste prema dole. U usporedbi s drugim pristupima, reprezentacija ovih podataka je simbolička i laka za interpretaciju. Cilj je kreirati klasifikacijski model koji predviđa vrijednost ciljanog atributa (često nazvanog klasa ili oznaka), temeljen na nekoliko ulaznih atributa u primjer setu.

Numerička varijabla

a) *Regresija* – klasifikacija pomoću operatora za regresiju je model koji sadrži podproces. Podproces mora sadržavati učenika regresije, odnosno operator koji generira model regresije. Za svaku *i* klasu danog primjer seta, model regresije je naučen da postavi oznaku na +1 ako je oznaka *i* te na -1 ako to nije. Tada se model udružuje u klasifikacijski model. Kako bi odredio predikciju za neoznačeni primjer, svi se

regresijski modeli primjenjuju i odabire se klasa pripadajućeg modela koji predviđa najveću vrijednost.

b) *Neuronske mreže* – ovaj operator služi za treniranje neuronske mreže. Radi na principu neuronske mreže s povratnim prostiranjem pogreške i ima mogućnost učenja. U parametrima neuronske mreže moguće je podesiti broj skrivenih slojeva mreže, momentum i koeficijent učenja.

Klasteri

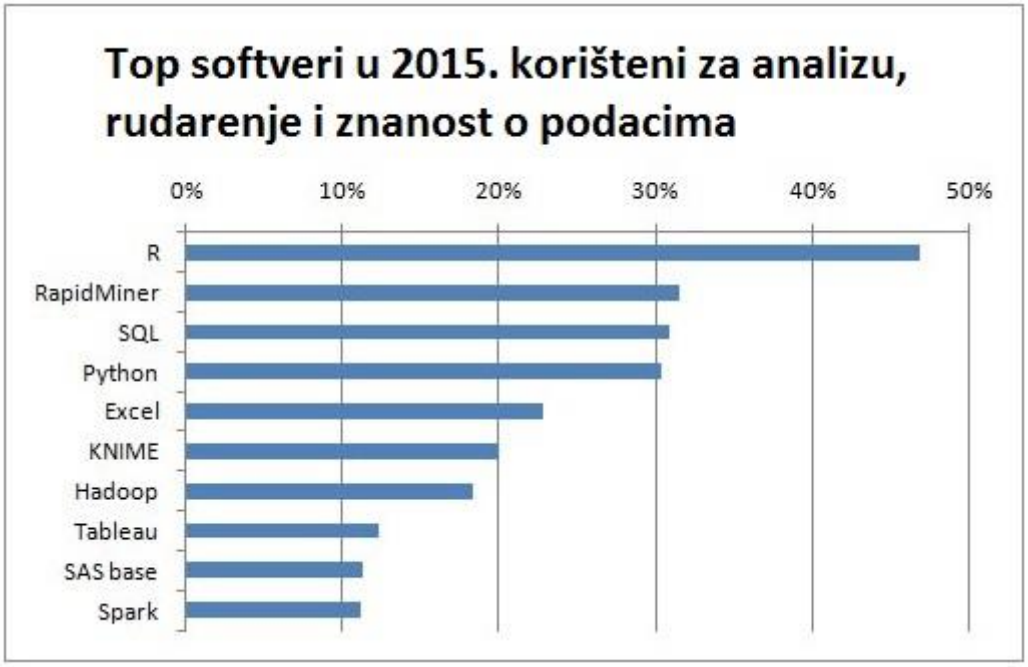
a) *K-means* – ovaj operator provodi klasterizaciju korištenjem k-means algoritma. Klasterizacija je postupak grupiranja objekata koji su slični jedan drugome i različiti od objekata koji pripadaju drugim klasterima. K-means klasteriranje je poseban algoritam, odnosno svaki objekt je dodijeljen točno jednom klasteru. Objekti u jednom klasteru su slični jedan drugome, a sličnost između objekata se temelji na mjerenju udaljenosti među njima.

b) *Fuzzy C-means (FCM)* – metoda klasteriranja koja omogućava jednom dijelu podataka da pripada u dva ili više klastera. Metoda se često koristi za prepoznavanje uzoraka. Slična je k-means metodi. Algoritam minimizira varijance u klasterima, ali sadrži problem zbog toga što su minimumi lokalni pa rezultat ovisi o inicijalnom izboru težina.

Kako je područje PI sve popularnije i raširenije, logično je da se razvijaju softveri koji već imaju u sebi navedene operatore i analitičarima puno pojednostavljuju proces rudarenja. Koji su to navedeno je u nastavku.

4.4. Softverski alati za rudarenje podataka

Tržište alata za rudarenje podataka zadnjih je godina u velikom porastu. Mnogi alati imaju u sebi integrirano više različitih postupaka strojnog učenja i pripreme podataka te tako omogućavaju kvalitetno otkrivanje znanja u podacima. Često je bitno da su ti alati i javno dostupni. Prema godišnjem istraživanju koje provodi *KDnuggets*, a temelji se na anketiranju od oko 3000 korisnika koji biraju između 93 različitih alata za rudarenje podataka (slika 3.), R je proglašen najpopularnijim u 2015. godini. Za njim slijedi RapidMiner koji je 2013. i 2014. zauzimao prvo mjesto.



Slika 3. *KDnuggets* istraživanje o alatima rudarenja podataka [19]

Tri takva alata slijede u nastavku [17]:

- **R-programiranje**

R je besplatni softverski jezik za programiranje i softversko kruženje za statističko računanje i grafiku. R jezik je u širokoj upotrebi među rudarima podataka za razvoj statističkih softvera i analizu podataka. Zbog lakoće upotrebe i proširivosti, njegova popularnost je bitno narasla zadnjih godina što se vidi iz spomenutog istraživanja (slika 3.). Uz rudarenje podataka omogućava i statistike te grafičke tehnike uključujući linearno i nelineareno programiranje, klasične statističke testove, analizu vremenskih serija, klasifikaciju, klasterizaciju i drugo.

- **RapidMiner (ranije poznat kao YALE)**

RapidMiner je suvremeni sustav za dubinsku analizu podataka koji se odlikuje kvalitetnim korisničkim sučeljem. Pisan je u Java programskom jeziku. Kao dodatak rudarenju podataka, RapidMiner također omogućava funkcije kao što je preprocesuiranje podataka i vizualizacija podataka, prediktivnu analizu i statističko modeliranje, evaluaciju, te razvoj. Moćnim ga čini i to što Za RapidMiner nije potrebna licenca i može biti skinut s SourceForge stranice gdje je ocijenjen kao broj 1 softver za poslovnu analizu. U petom poglavlju opisana je analiza provedena upravo u RapidMineru.

- **Excel (kodnog naziva Odyssey)**

Excel je Microsoftov softverski program koji je dio Microsoft Office paketa softverskih programa. Sposoban je za stvaranje i uređivanje proračunskih tablica koje se spremaju s ekstenzijama .xls ili .xlsx. Opća namjena Excela uključuje kalkulacije bazirane na ćelijama, pivot tablice i razne grafičke alate. Sastoji se od redova i stupaca, izrađenih od individualnih ćelija. Oni se mogu mijenjati na mnoge načine, uključujući boju pozadine, broj ili format datuma, font teksta i drugo. Također omogućuje rudarenje podataka. U petom poglavlju prikazano je uređivanje seta podataka te njegova statistička analiza izvedena pomoću Excela [24].

5. PRIMJENA TEHNIKA RUDARENJA PODATAKA NA SKUPU PODATAKA O ZRAKOPLOVNIM NESREĆAMA

Obrađenu teoriju u prethodnim poglavljima potrebno je prikazati na primjeru kako bi se bolje predočile prednosti PI i rudarenja podataka. Odabrani podaci i svi koraci procesa prikazani su u ovom poglavlju.

5.1. Opis seta podataka

Podaci odabrani za analizu sadrže podatke o civilnim i vojnim zrakoplovnim nesrećama te smrtnim slučajevima izazvanim zrakoplovima u vremenskom periodu od 1908. do 2009. godine skinuti su sa stranice *Kaggle* [20].

Zrakoplovne nesreće u izvještaju sadrže 13 atributa prikazanih u tablici 2.

Tablica 2. Opis atributa korištenih za analizu u danom setu podataka

Atribut	**Opis**	**Tip podataka**
Date	Datum odvijanja događaja	Datumski
Time	Vrijeme odvijanja događaja	Vremenski
Location	Mjesto odvijanja događaja (4 287)	Nominalni
Operator	Ime operatora (2 475)	Nominalni
Flight	Broj leta	Nominalni
Route	Ruta odvijanja leta	Nominalni
Type	Tip zrakoplova (2 440)	Nominalni
Registration	Jedinstveni, službeni, registracijski broj zrakoplova	Nominalni
cn/ln	Konstrukcijski broj koji daje proizvođač	Nominalni
Aboard	Putnici ukrcani u zrakoplov (144 551)	Numerički
Fatalities	Broj smrtno stradalih u zrakoplovu (105 358)	Numerički
Ground	Broj smrtno stradalih izvan zrakoplova kao posljedica zrakoplovne nesreće (8 440)	Numerički
Summary	Opis zrakoplovne nesreće	Nominalni

Baza podataka sadrži ukupno 5 246 zapisa o zrakoplovnim nesrećama, a brojevi u tablici 2. koji se nalaze u zagradama sadrže informaciju da su se nesreće dogodile na 4 287 različitih

lokacija u svijetu. Sudionici su 2 440 različitih tipova zrakoplova te 2 475 različitih operatera. Ukupno je ukrcano 144 551 osoba u te zrakoplove, od kojih je 105 358 poginulo. Zabilježeno je dodatnih 8 440 smrtnih slučajeva uzrokovanih zrakoplovnim nesrećama u koje nisu uključeni ukrcani putnici.

Za bolje razumijevanje područja potrebno je objasniti što podrazumijeva zrakoplovna nesreća i koje događaje isključuje [21]:

Zrakoplovna nesreća

Pojava povezana s događajem zrakoplova između trenutka kad se bilo koja osoba s namjerom leta ukrca na zrakoplov i trenutka do kojeg u kojem su se sve osobe iskrcale iz zrakoplova, u kojem

- zrakoplov trpi značajna oštećenja
- zrakoplov je nestao ili je potpuno nedostupan
    - Zrakoplov se smatra nestalim kada je službena potraga završena i olupina nije locirana
- smrt ili ozbiljne ozljede nastale od
    - bivanja u zrakoplovu
    - izravnog kontakta s zrakoplovom ili bilo čime vezanim za njega
    - izravnom izlaganju eksploziji

Isključeni događaji

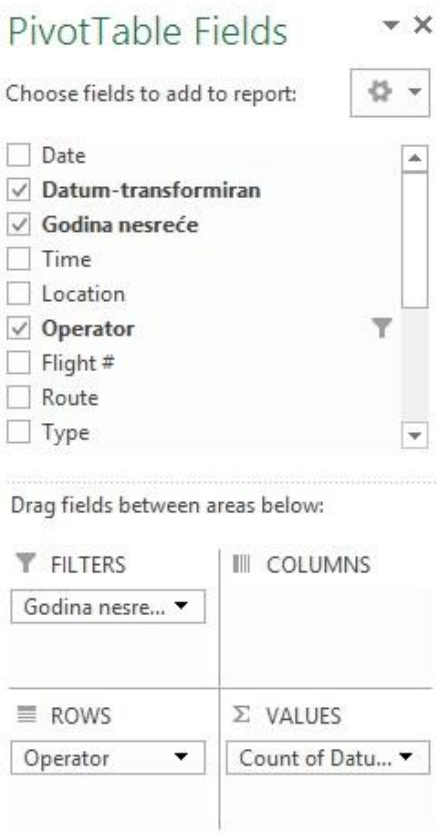
- Smrtne i lakše ozljede nastale prirodnim uzrokom
- Smrtne i lakše samonanešene ozljede ili ozljede nanesene od strane drugih osoba
- Smrtne i lakše ozljede slijepih putnika skrivenih izvan područja normalno dostupnih putnicima i posadi
- Lakše ozljede kao rezultat atmosferskih turbulencija, normalnih manevriranja, labavih objekata, lijetanja, iskrcavanja, evakuacije te održavanja i servisiranja
- Lakše ozljede osoba koje nisu ukrcane na zrakoplov

Sljedeći događaji se ne smatraju zrakoplovnim nesrećama: oni koji su rezultat eksperimentalnih testnih letova ili neprijateljskih akcija, uključujući sabotažu, otmicu, terorizam i vojnu akciju.
Na početku istraživanja potrebno je izvršiti početnu analizu koja služi za lakše definiranje i razjašnjavanje problema. Upravo tome služi eksplorativna analiza podataka prikazana u nastavku.

5.1.1. Eksplorativna analiza podataka

Za lakšu manipulaciju podataka upotrijebljen je alat Pivot Tablica (eng. *pivot table*) koji se nalazi unutar softverskog paketa Excel. Pivot tablice su dizajnirane za baratanje s velikim brojem podataka. Omogućavaju da se ogromna količina podataka pretvori u sumirani izvještaj. Osim navedenog, korištenjem Pivot Tablica je omogućeno da se iz „šume“ podatka izluče trendovi na osnovu kojih se donose poslovne odluke. Excel u memoriji stvara višedimenzionalnu sliku podataka, koji se zatim mogu transformirati i mogu se stvarati presjeci iz različitih perspektiva.

Pivot tablice omogućuju filtriranja pod različitim uvjetima što je prikazano na slici 4.

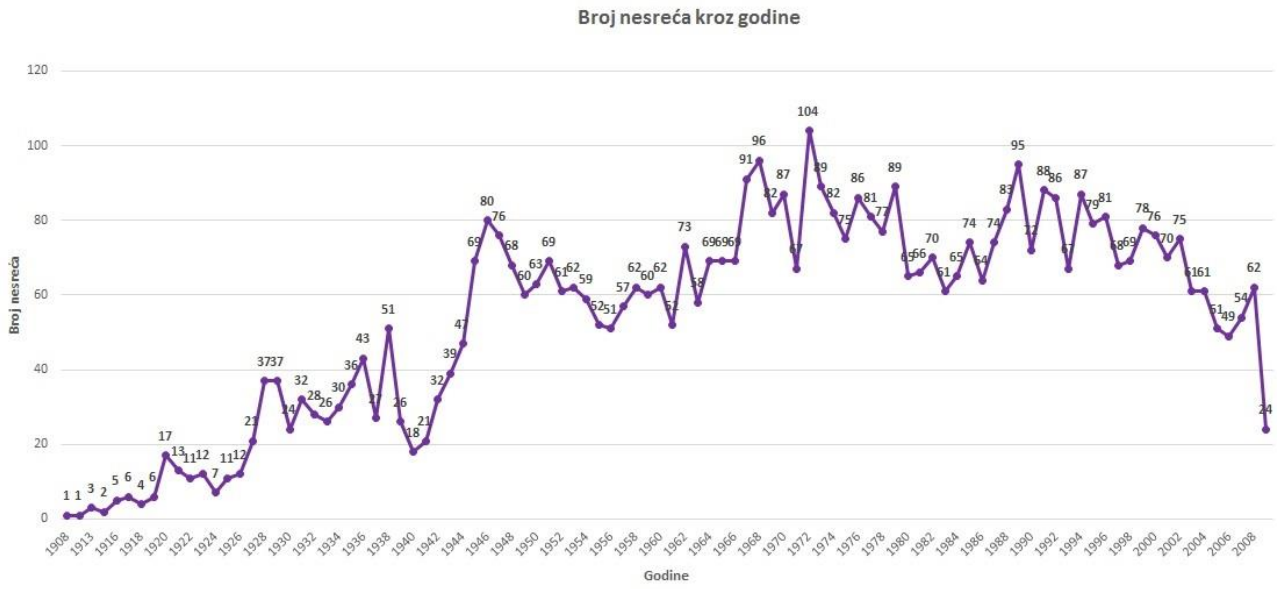


Slika 4. Prikaz stvaranja Pivot Tablice

Na slici 5. prikazana je Pivot Tablica koja sadržava informacije o ukupnom broju nesreća na godišnjoj razini te je na temelju te tablice prikazan graf koji prikazuje trend kretanja zrakoplovnih nesreća (slika 6.).

Godina	Broj nesreća
1908	1
1912	1
1913	3
1915	2
1916	5
1917	6
1918	4
1919	6
1920	17
1921	13
1922	11
1923	12
1924	7
1925	11

Slika 5. Prikaz ukupnog broja zrakoplovnih nesreća na godišnjoj razini



Slika 6. Broj zrakoplovnih nesreća kroz prikazan na godišnjoj razini

Iz slike 6. je vidljivo da je početkom razvitka zrakoplovne industrije bilo vrlo malo nesreća godišnje. Prva kobna zrakoplovna nesreća dogodila se 17. rujna 1908. godine kada se srušio zrakoplov braće Wright, u kojoj je poginuo mladi američki poručnik Thomas Selfridge. Što se više zrakoplovna industrija razvijala to je više rastao i broj letova. Broj nesreća nepravilno raste

i doseže maksimum 1972. godine. Nakon toga vidi se nepravilan, ali očiti pad iz čega se može zaključiti da s razvitkom avioindustrije raste i sigurnost zrakoplova, te veliki utjecaj ima završetak svjetskih ratova.

Koristeći isti alat, prikazan je ukupan broj nesreća za svakog operatera (slika 7.).

Godina nesreće	(Multiple Items)
Operator	**Broj nesreća**
Aeroflot	179
Air France	70
Air Taxi	48
American Airlines	36
China National Aviation Corporation	44
Deutsche Lufthansa	65
Military - Royal Air Force	36
Military - U.S. Air Force	176
Military - U.S. Army Air Forces	43
Military - U.S. Navy	36
Pan American World Airways	41
United Air Lines	44
US Aerial Mail Service	36
Ukupno	**854**

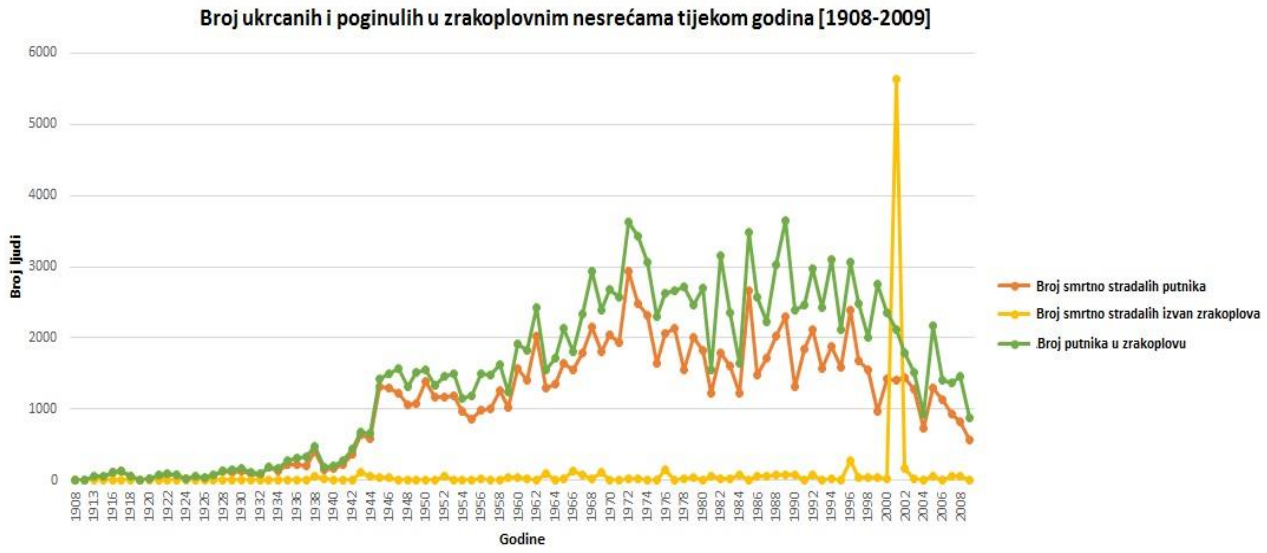
Slika 7. Prikaz 10 operatera s najviše zrakoplovnih nesreća

Osim prikaza trenda kretanja zrakoplovnih nesreća po godinama i operatera s najvećim brojem nesreća, prikazan je i broj ukupno stradalih (poginulih) putnika u nesrećama na godišnjoj razini (slika 8.).

Godina	Zbroj poginulih putnika	Zbroj poginulih izvan aviona	Zbroj putnika
1908	1	0	2
1912	5	0	5
1913	45	0	51
1915	40	0	60
1916	108	0	109
1917	124	0	124
1918	65	0	65
1919	5	0	5
1920	24	0	31
1921	68	1	69
1922	80	5	91
1923	77	0	80
1924	18	0	18
1925	39	0	68

Slika 8. Prikaz broja stradalih putnika u i izvan zrakoplova u odnosu na broj ukrcanih putnika

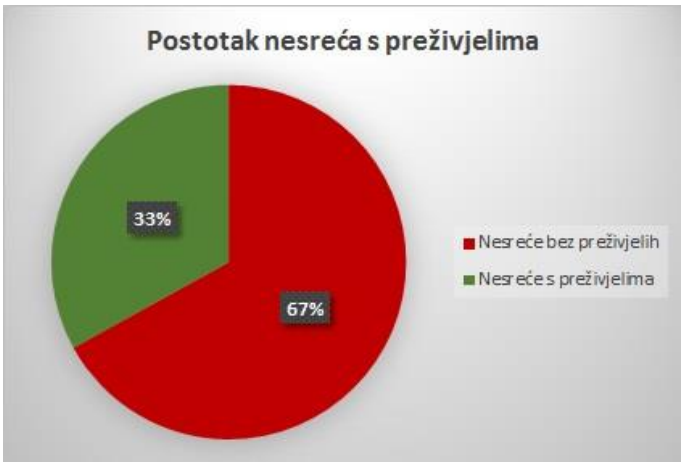
Slika 8. prikazuje broj ukrcanih i poginulih putnika, te poginulih izvan zrakoplova za svaku godinu. Time se dobiva mogućnost usporedbe kretanja tih vrijednosti prikazanih na slici 9.



Slika 9. Broj ukrcanih i poginulih osoba kroz godine

Kod krivulja ukrcanih i poginulih putnika vidimo sličan trend kao i kod krivulje zrakoplovnih nesreća kroz godine, što je bilo i očekivano. No velika nepodudarnost javlja se kod krivulje koja prikazuje osobe koje nisu bile putnici zrakoplova, a stradale su zbog zrakoplovne nesreće. Veliki skok se odnosi na dva od četiri teroristička napada 11. rujna 2001. godine u Sjedinjenim Američkim Državama, kada su oteta dva zrakoplova koja su se zabila u dva nebodera Svjetskog trgovačkog centra na Manhattanu u New Yorku. U tom događaju poginulo je 2 750 ljudi.

Analizirajući i uspoređujući podatke o broju ukrcanih na zrakoplov te smrtnim slučajevima u i izvan zrakoplova uzrokovanih zrakoplovnom nesrećom, dodatno su izvučena dva grafa koja slikovito prikazuju spomenute odnose (slike 10. i 11.).



Slika 10. Odnos zrakoplovnih nesreća u kojima ima stradalih izvan zrakoplova u usporedbi s onima u kojima ih nema

Iz slike 10. vidljivo je da upola manje zrakoplovnih nesreća ima preživjelih putnika. Neki zapisi nisu sadržavali sve podatke o ukrcanima i poginulima pa oni nisu uzeti u obzir. Također, 22 zapisa nisu sadržavala informaciju o ukrcanima na zrakoplov, a od toga 12 zapisa nije sadržavalo informaciju o poginulima u zrakoplovu. Budući da je izračunata statistika broja poginulih u odnosu na broj ukrcanih putnika izostavljeno je svih 22 zapisa.

Također je napravljena usporedba o nastradalima izvan zrakoplova (slika 11.).

Slika 11. Odnos zrakoplovnih nesreća s preživjelima u usporedbi s onima bez preživjelih

U 4% slučajeva su smrtno stradali ljudi koji nisu bili putnici zrakoplova. Ovdje je isti slučaj kao i kod prošlog grafa, odnosno neki zapisi nisu bili potpuni. Zapisi koji nisu sadržavali broj poginulih izvan zrakoplova nisu uzeti u obzir, točnije 22 od 5.268 zapisa što je prihvatljiv broj za točnost statistike.

Ono što bi moglo biti zanimljivo kada se promatraju operatori i tipovi zrakoplova, su upravo oni s najviše nesreća. Zato je izrađen graf koji prikazuje podatke za top 10 operatora i zrakoplova s najviše nesreća (slika 12.).

Slika 12. Prikaz 10 tipova zrakoplova s najviše nesreća (1908.-2009.)

Daleko najviše zrakoplovnih nesreća dogodilo se s zrakoplovom Douglas DC-3. On je američki propelerni zrakoplov, čija su brzina i obim razvoja napravili revoluciju u zračnom prometu 1930-ih i 1940-ih godina. Zbog trajnog učinka na zrakoplovnu industriju i Drugi svjetski rat, općenito je smatran jednim od najznačajnijih transportnih zrakoplova ikad napravljenih [23].

Slika 13. prikazuje top 10 operatora s najviše zrakoplovnih nesreća.

Slika 13. Prikaz 10 operatora s najviše zrakoplovnih nesreća (1908.-2009.)

Uvjerljivo najviše nesreća dogodilo se kod Aeroflota i Military – U.S. Air Force operatora. Aeroflota je najveća ruska zrakoplovna kompanija te jedna od najstarijih u svijetu. Military – U.S. Air Force je američka vojnozrakoplovna kompanija. Obje su sudjelovale u Drugom svjetskom ratu što daje smisao podacima s grafa.

5.1.2. Statistička analiza podataka

U nastavku su prikazani glavni statistički pokazatelji kako bi se bolje opisali podaci [22]:

1. Srednja vrijednost – predstavlja sumu svih podataka podijeljenu s ukupnim brojem podataka. Računanje središnje vrijednosti predstavlja jedan od najčešće primjenjivanih statističkih postupaka kojeg koristimo kako bismo sažeto i zorno prikazali određeni skup podatka. Računanje srednje vrijednosti cijeli skup podataka zamjenjujemo jednom vrijednošću za koju smatramo da ga dobro reprezentira, te stoga moramo biti jako pažljivi prilikom odabira prikladne mjere srednje vrijednosti [25].

$$\mu = \frac{x_1 + x_2 + .. + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \textbf{(1)}$$

Ukupan broj nesreća je 5268, a veličina populacije je 98 godina.

$$\mu = \frac{5268}{98} = 53{,}76$$

Dobiveni broj daje informaciju koliko se prosječno nesreća dogodilo u jednoj godini, odnosno u jednoj godini se prosječno dogodilo između 53 i 54 zrakoplovne nesreće.

Nakon što se izračuna srednja vrijednost potrebno je izračunati i mjere koje prikazuju raspršenost skupa podataka.

2. Standardna devijacija – je pozitivna vrijednost drugog korijena varijancije uzorka

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad \textbf{(2)}$$

$$\sigma = 27{,}35$$

Ona govori da je prosječno odstupanje od srednje vrijednosti 27,35, odnosno da broj nesreća po godini prosječno odstupa od broja 53,76 za 27,35 nesreća.

3. Varijanca – je suma kvadrata odstupanja svih podataka od njihove srednje vrijednosti podijeljene s N gdje N predstavlja ukupan broj podataka u uzorku.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \quad \textbf{(3)}$$

$$\sigma^2 = 747{,}90$$

Srednja vrijednost kvadrata odstupanja vrijednosti numeričke varijable od prosjeka (srednje vrijednosti) iznosi 747,90.

Standardna devijacija i varijanca su apsolutne mjere disperzije, a poznavanje disperzije je bitno da bi se mogla spoznati važnost srednjih vrijednosti kao mjera centralne tendencije. Prema dobivenim vrijednostima vidi se da je disperzija velika što znači da je niz vrijednosti nije homogen već varijabilan.

Nakon pripreme i statističke obrade podataka slijedi njihova transformacija i prilagodba za modeliranje procesa prikazana u podpoglavlju 5.2.

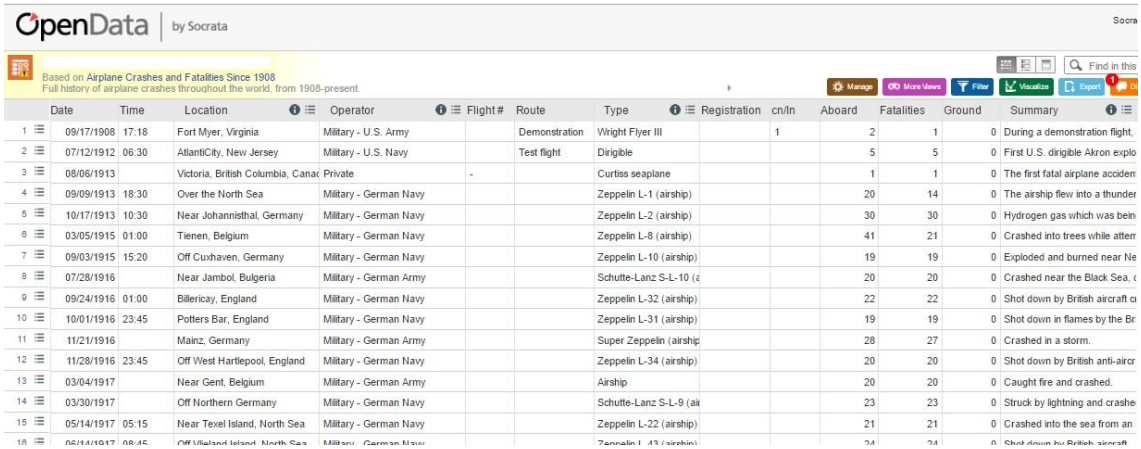
5.2. Prikupljanje i transformacija podataka

Nakon prikaza eksplorativne analize slijedi korak prikupljanja i transformacije podataka. Kako bi se maksimalno omogućila iskoristivost podataka, oni moraju biti dobro pripremljeni. Stoga je potrebno provesti sljedeće korake:

5.2.1. Prikupljanje i opća transformacija podataka

K 1: Prikupljanje podataka

Stvarni prikaz seta podataka preuzet s izvora [20] prikazan je na slici 14.



Slika 14. Podaci na stranici kaggle.com

Set podataka s interneta je skinut u „csv“ obliku, odnosno u obliku teksta u kojem zarez (,) predstavlja razdjelnik (slika 15.).

Slika 15. Prikaz CSV dokumenta u softverskom alatu Excel

K 2: Transformacija podataka- općenito

Kako bi se analiza mogla nastaviti potrebno je srediti tablicu u kojoj su podaci pregledni. Sređeni podaci prikazani su na slici 16.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Date	Datum-transformir	Godina nesr	Tim	Location	Operator	Fligh	Route	Type	Regis	cn/In	Aboa	Fatalitie	Groun	Summary
2	09/17/1908	17.9.1908	1908	17:18	Fort Myer, Virginia	Military - U.S. Army		Demonstratic	Wright Flyer III		1	2	1	0	During a demons
3	7.12.1912	7.12.1912	1912	6:30	AtlantiCity, New Jersey	Military - U.S. Navy		Test flight	Dirigible			5	5	0	First U.S. dirigible
4	8.6.1913	8.6.1913	1913		Victoria, British Columbia, Canada	Private	-		Curtiss seaplane			1	1	0	The first fatal airp
5	9.9.1913	9.9.1913	1913	18:30	Over the North Sea	Military - German Navy			Zeppelin L-1 (airship)			20	14	0	The airship flew i
6	10/17/1913	17.10.1913	1913	10:30	Near Johannisthal, Germany	Military - German Navy			Zeppelin L-2 (airship)			30	30	0	Hydrogen gas wh
7	9.3.1915	9.3.1915	1915	15:20	Off Cuxhaven, Germany	Military - German Navy			Zeppelin L-10 (airship)			19	19	0	Exploded and bu
8	3.5.1915	3.5.1915	1915	1:00	Tienen, Belgium	Military - German Navy			Zeppelin L-8 (airship)			41	21	0	Crashed into tree
9	10.1.1916	10.1.1916	1916	23:45	Potters Bar, England	Military - German Navy			Zeppelin L-31 (airship)			19	19	0	Shot down in flar
10	07/28/1916	28.7.1916	1916		Near Jambol, Bulgeria	Military - German Army			Schutte-Lanz S-L-10 (airship)			20	20	0	Crashed near the
11	09/24/1916	24.9.1916	1916	1:00	Billericay, England	Military - German Navy			Zeppelin L-32 (airship)			22	22	0	Shot down by Bri
12	11/21/1916	21.11.1916	1916		Mainz, Germany	Military - German Army			Super Zeppelin (airship)			28	27	0	Crashed in a stor
13	11/28/1916	28.11.1916	1916	23:45	Off West Hartlepool, England	Military - German Navy			Zeppelin L-34 (airship)			20	20	0	Shot down by Bri
14	03/30/1917	30.3.1917	1917		Off Northern Germany	Military - German Navy			Schutte-Lanz S-L-9 (airship)			23	23	0	Struck by lightnir
15	3.4.1917	3.4.1917	1917		Near Gent, Belgium	Military - German Army			Airship			20	20	0	Caught fire and c
16	05/14/1917	14.5.1917	1917	5:15	Near Texel Island, North Sea	Military - German Navy			Zeppelin L-22 (airship)			21	21	0	Crashed into the
17	06/14/1917	14.6.1917	1917	8:45	Off Vlieland Island, North Sea	Military - German Navy			Zeppelin L-43 (airship)			24	24	0	Shot down by Bri
18	08/21/1917	21.8.1917	1917	7:00	Off western Denmark	Military - German Navy			Zeppelin L-23 (airship)			18	18	0	Shot down by Bri
19	10/20/1917	20.10.1917	1917	7:45	Near Luneville, France	Military - German Navy			Zeppelin L-44 (airship)			18	18	0	Shot down by Fre
20	4.7.1918	4.7.1918	1918	21:30	Over the Mediterranean	Military - German Navy			Zeppelin L-59 (airship)			23	23	0	Exploded and cra
21	5.10.1918	5.10.1918	1918		Off Helgoland Island, Germany	Military - German Navy			Zeppelin L-70 (airship)			22	22	0	Shot down by Bri
22	8.11.1918	8.11.1918	1918	10:00	Ameland Island, North Sea	Military - German Navy			Zeppelin L-53 (airship)			19	19	0	Shot down by bri
23	12/16/1918	16.12.1918	1918		Elizabeth, New Jersey	US Aerial Mail Service			De Havilland DH-4	97		1	1	0	
24	10.2.1919	10.2.1919	1919		Newcastle, England	Aircraft Transport and Travel			De Havilland DH-4			1	1	0	
25	05/25/1919	25.5.1919	1919		Cleveland, Ohio	US Aerial Mail Service			De Havilland DH-4	61		1	1	0	Caught fire in mi
26	07/19/1919	19.7.1919	1919		Dix Run, Pennsylvania	US Aerial Mail Service			De Havilland DH-4	82		1	1	0	

Slika 16. Set podataka nakon uređivanja

Zbog problema s formatom datuma dodan je novi stupac u kojem se nalaze transformirani datumi u hrvatskom formatu. U C stupcu izlučena je samo godina događaja koja služi kako bi se lakše došlo do pojedinih statističkih podataka.

5.2.2. Transformacija podataka za klasifikaciju

Tablicu je potrebno transformirati za svaku metodu zasebno, ovisno o zahtjevima i potrebnim izlaznim vrijednostima (zavisnim varijablama) koji ovise o ulaznim podacima (nezavisnim varijablama). Za klasifikaciju je potrebno odabrati ciljani atribut koji se pokušava predvidjeti. Navedeni atribut (zavisna varijabla) treba sadržavati dvije vrijednosti (ili biti). U odabranom setu podataka, za zavisnu varijablu odabran je atribut *„Uvjet1"* koji prikazuje podatke o tome da li je nakon nesreće bilo preživjelih putnika ili nitko nije preživio. Ukoliko je bilo preživjelih ciljani atribut poprima vrijednost 1, dok u suprotnom poprima vrijednost 0. Na slici 17. prikazana je transformirana tablica korištena za metodu klasifikacije.

A	B	C	D	E	F	G	H	I	J	K	L
Location	City			Country	Type	Tip_zrakoplova_uređeno	Operator	Aboard	Fatalities	Survived	Uvjet1
Victoria, British Columbia, Canada	Victoria	British Columbia	Canada	Canada	Curtiss seaplane	Curtiss	Private	1	1	0	0
Elizabeth, New Jersey	Elizabeth	New Jersey		New Jersey	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Newcastle, England	Newcastle	England		England	De Havilland DH	De Havilland	Aircraft Transport and Travel	1	1	0	0
Cleveland, Ohio	Cleveland	Ohio		Ohio	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Dix Run, Pennsylvania	Dix Run	Pennsylvania		Pennsylvania	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Cantonsville, Maryland	Cantonsville	Maryland		Maryland	Curtiss R-4LM	Curtiss	US Aerial Mail Service	1	1	0	0
Long Valley, New Jersey	Long Valley	New Jersey		New Jersey	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Newark, New Jersey	Newark	New Jersey		New Jersey	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Cleveland, Ohio	Cleveland	Ohio		Ohio	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Tie Siding, Wyoming	Tie Siding	Wyoming		Wyoming	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
College Park, Maryland	College Park	Maryland		Maryland	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Hillersburg, Pennsylvania	Hillersburg	Pennsylvania		Pennsylvania	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
New Paris, Indiana	New Paris	Indiana		Indiana	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Newark, New Jersey	Newark	New Jersey		New Jersey	Curtiss R-4LM	Curtiss	US Aerial Mail Service	2	1	1	1
Batavia, Illinois	Batavia	Illinois		Illinois	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Heller Field, New Jersey	Heller Field	New Jersey		New Jersey	Curtiss JN-4H	Curtiss	US Aerial Mail Service	2	1	1	1
Oskaloosa, Iowa	Oskaloosa	Iowa		Iowa	De Havilland DH	De Havilland	US Aerial Mail Service	2	1	1	1
Elko, Nevada	Elko	Nevada		Nevada	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Mendotta, Minnisota	Mendotta	Minnisota		Minnisota	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Mitchel Field, NY	Mitchel Field	NY		NY	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
Cleveland, Ohio	Cleveland	Ohio		Ohio	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0
San Francisco, California	San Francisco	California		California	De Havilland DH	De Havilland	US Aerial Mail Service	1	1	0	0

Slika 17. Transformirana tablica za metodu klasifikacije

Pomoću stupca „*Survived"* provjereno je ima li preživjelih u avionu tako da je oduzet broj poginulih putnika iz stupca „*Fatalities*" od broja ukrcanih putnika iz stupca „*Aboard*". Uvjet, odnosno ciljani atribut, nalazi se u zadnjem stupcu „*Uvjet1*". Koristeći IF provjereno je da li je vrijednost u stupcu preživjelih veća od 0. Ukoliko je taj uvjet ispunjen određeni zapis će poprimiti vrijednost 1, tj. davat će informaciju o tome da je bilo preživjelih u toj zrakoplovnoj nesreći. U stupcu A , nalazi se atribut *„Location"* koji sadrži informaciju o mjestu nesreće. Iz njega su izlučene informacije o točnom nazivu grada gdje se nesreća odvila (stupac B „*City*") i države gdje se taj grad nalazi (stupac E „*Country*"). Stupci C i D su pomoćni stupci koji su služili za filtriranje informacija iz stupca A. Također je bilo potrebno izlučiti i proizvođače zrakoplova) budući da je eksplorativnom analizom uočeno da 10 tipova zrakoplova pokriva 70 % podataka. Slijedeći ove korake transformacije i filtriranja podataka, proces klasifikacije bi trebao davati bolje rezultate.

5.2.3. Transformacija podataka za klasterizaciju

Analizirajući dobiveni set podataka korištenjem eksplorativne analize uočeno je da određene grupe podataka (vezane uz proizvođača (tip) zrakoplova) sadrže slične vrijednosti određenih atributa. Zbog navedenog će se provesti analiza grupiranja (klasterizacija) kako bi se uočilo koji zapisi su slični.

Kako bi se olakšao proces klasterizacije transformirani su podaci na način prikazan na slici 18.

Tip zrakoplova	Broj nesreća
AAC-1	1
AEGK	1
Aermacchi	1
Aero Commander	12
Aerospatiale	32
Aerospeciale	1
Aerostar	4
Agusta	1
Airbus	35
Airship	1
Airspeed Ambassador	3
Antonov	247
Arado	1
Arava	2
Armstrong	2
Armstrong-Whitworth	2
AT	1
ATR	1
ATR-42-300	1
ATR-72-202	1
ATR-72-212	1
Avia	5
Aviation	1

Slika 18. Prikaz tablice za klasterizaciju tipova zrakoplova prema broju nesreća

Slika 18. prikazuje tablicu izrađenu pomoću alata Pivot tablice. Navedena tablica sadržava informacije o tome koliko je za pojedini tip zrakoplova zabilježeno nesreća u promatranom periodu (1908.-2009.).

5.2.4. Transformacija podataka za analizu tekstualnih zapisa

Dobiveni set podataka sadržava stupac *„Summary*“ koji sadržava informacije o opisu zrakoplovnih nesreća (slika 19.). Kako bi se ustanovilo koji su glavni uzročnici povezani s nesrećama, potrebno je provesti rudarenje teksta (analizu tekstualnih zapisa):

Summary
During a demonstration flight, a U.S. Army flyer flown by Orville Wright nose-dived into the ground from a heigh
First U.S. dirigible Akron exploded just offshore at an altitude of 1,000 ft. during a test flight.
The first fatal airplane accident in Canada occurred when American barnstormer, John M. Bryant, California aviat
The airship flew into a thunderstorm and encountered a severe downdraft crashing 20 miles north of Helgoland
Hydrogen gas which was being vented was sucked into the forward engine and ignited causing the airship to exp
Exploded and burned near Neuwerk Island, when hydrogen gas, being vented, was ignited by lightning.
Crashed into trees while attempting to land after being shot down by British and French aircraft.
Shot down in flames by the British 39th Home Defence Squadron.
Crashed near the Black Sea, cause unknown.
Shot down by British aircraft crashing in flames.
Crashed in a storm.
Shot down by British anti-aircraft fire and aircraft and crashed into the North Sea.
Struck by lightning and crashed into the BaltiSea.
Caught fire and crashed.
Crashed into the sea from an altitude of 3,000 ft. after being hit by British aircraft fire.
Shot down by British aircraft.
Shot down by British aircraft.
Shot down by French anti-aircraft fire.
Exploded and crashed into the sea off the southern coast of Italy.
Shot down by British aircraft crashing from a height of 17,000 ft.
Shot down by british aircraft.
Caught fire in midair. The pilot leaped from the plane to his death as the plane began to go into a dive.

Slika 19. Prikaz atributa koji sadržava informacije o zrakoplovnim nesrećama

Nakon što je izoliran stupac „*Summary*“ iz dobivenog seta podataka, potrebno je transformirati nominalne podatke u tom stupcu u tekstualne, kako bi se oni mogli koristiti u daljnjoj analizi teksta (slika 20).

Slika 20. Transformacija podataka za analizu teksta

Operator *Nominal to Text* pretvara nominalne atribute u tekst. Nakon što su podaci pretvoreni u tekstualni oblik, operator *Process Document from Data* generira vektore riječi iz atributa koji je u obliku nizova riječi. Podprocesi ovog operatora opisani su u sklopu prikazivanja metode analize teksta. Na kraju *Numerical to Binominal* operator pretvara numeričke atribute u binominalne.

5.3. Prikaz odabranih metoda

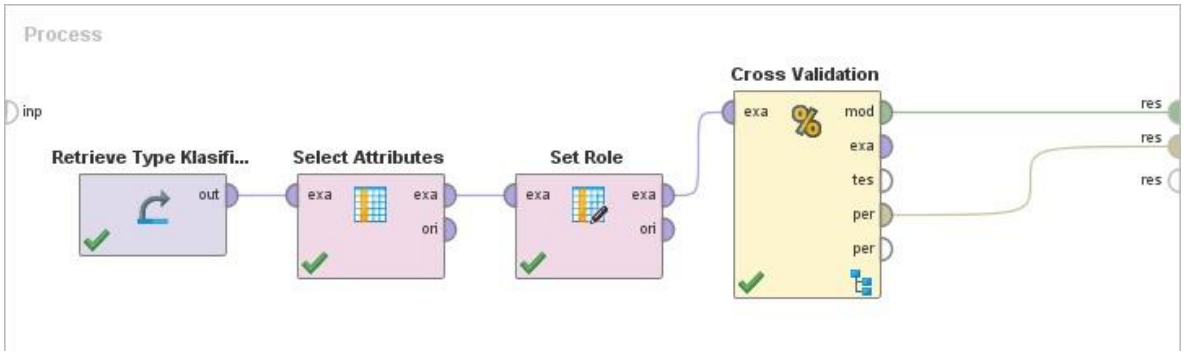
Kao što je prikazano u prethodnom poglavlju, metode rudarenja podataka kojima će se obraditi dani set podataka su klasifikacija, klasterizacija i analiza teksta. Nakon pripreme i transformacije podataka može započeti proces.

5.3.1. Klasifikacija

Već je spomenuto da klasifikacija služi za predviđanje vrijednosti ciljanog atributa (zavisne varijable) u odnosu na nezavisne atribute (varijable). Proces klasifikacije se sastoji od sljedećih operatera:

- Retrieve
- Select Attributes
- Set Role
- Cross Validation

Slika 21. prikazuje način povezivanja objašnjenih operatora.

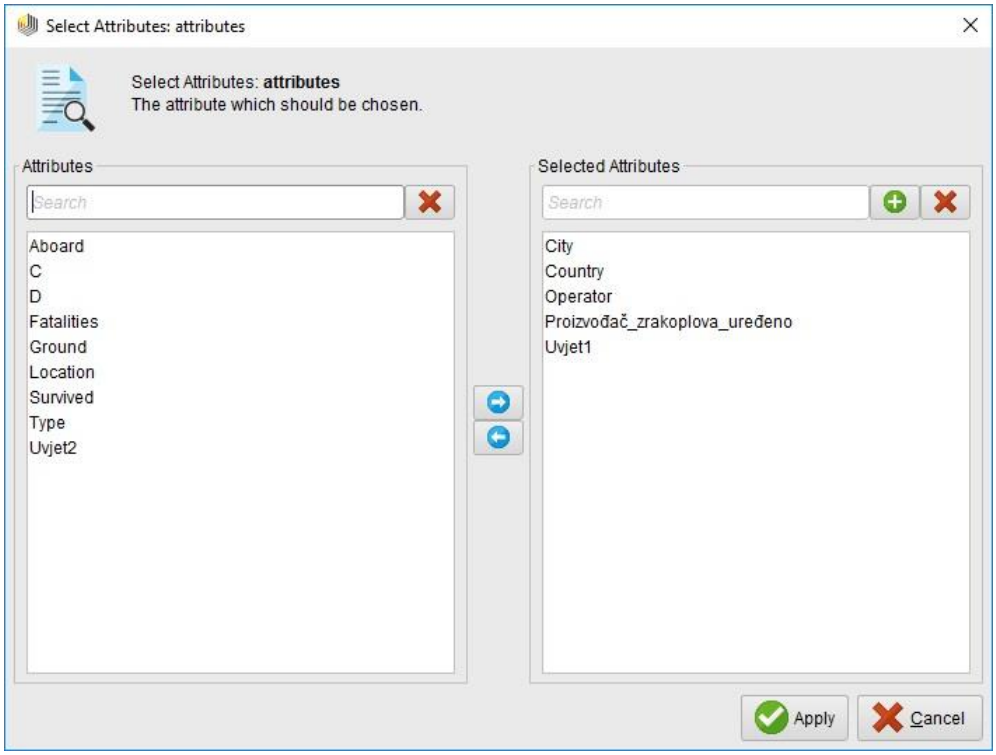


Slika 21. Glavni proces klasifikacije

Operator ***Retrieve*** dohvaća već pripremljene podatke koji su prethodno učitani u RapidMiner repozitorij kako bi se olakšao proces modeliranja.

Slijedi operator ***Select Attributes*** pomoću kojeg se odabiru atributi koji će se koristiti kao zavisne/nezavisne varijable. Koristeći ovaj operator, izostavljeni su pomoćni stupci korišteni u

fazi transformacije, kao i stupci koji ne sadržavaju strukturirane zapise i samim time nisu pogodni za klasifikaciju (npr. stupac „*Location*“). Slika 22. prikazuje odabrane atribute za proces klasifikacije.



Slika 22. Odabir atributa pomoću operatora *Select Attributes*

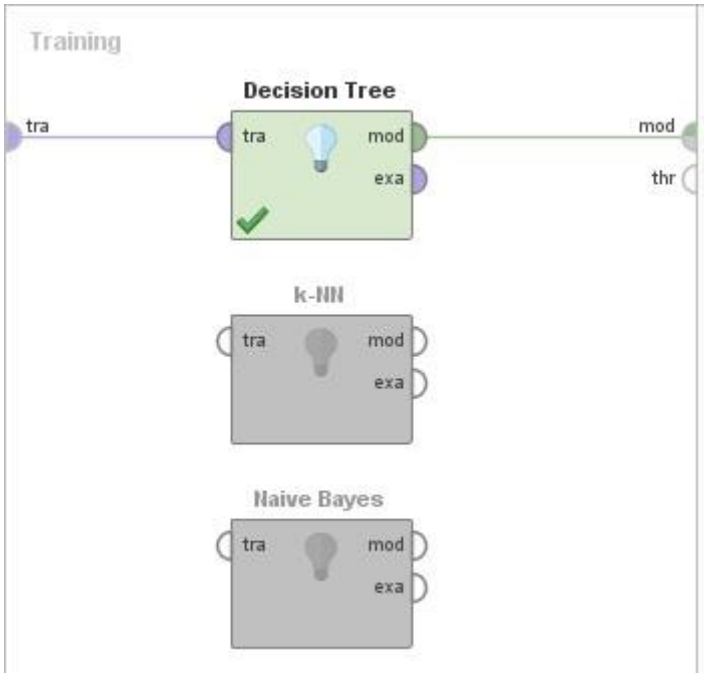
Iz slike 22. je vidljivo da se na lijevoj strani se nalaze atributi koji će biti izostavljeni iz procesa, dok su na desnoj oni koji će sudjelovati u procesu. To su City, *Country*, *Operator*, *Proizvođač_zrakoplova_uređeno* i *Uvjet1*.

Operatorom *Set Role* se određuje ciljani atribut, tj. zavisna varijabla (slika 23).

Slika 23. Parametri operatora *Set Role*

Operator ***Cross Validation*** je ključni operator koji izvodi unakrsnu validaciju kako bi se procijenile statističke performanse operatora za učenje koji se nalazi unutar njega. Točnije,

unutar njega se nalaze dva podprocesa, jedan za treniranje koji uči model (slika 24) i drugi za testiranje na kojem se primjenjuje naučeno te mjere performanse (slika 25.).



Slika 24. Podproces za treniranje

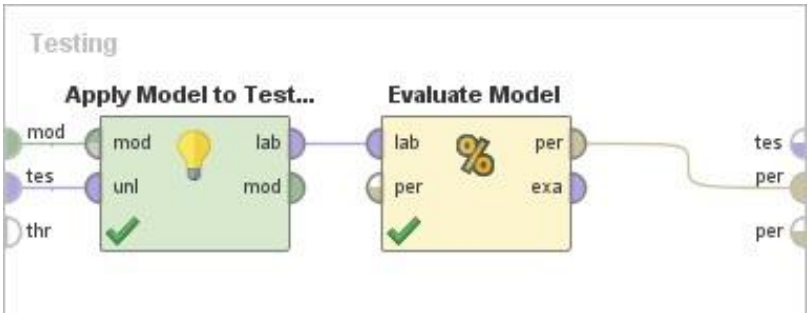
Podproces za trening sadrži operator za učenje. Na slici 24. su prikazana tri operatora. Po potrebi se omogućava rad određenom operatoru i uvijek je uključen samo jedan za vrijeme odvijanja procesa. Koriste se sva tri kako bi se moglo usporediti koji na zadanim podacima daje veću točnost.

Operator ***Decision Tree*** generira stablo odlučivanja, odnosno slikoviti model koji prikazuje cijelu strukturu odlučivanja. Ono klasificira primjere sortirajući ih od korijena (eng. *root*) do krajnjih čvorova (eng. *leaf*). Svaki čvor u stablu predstavlja neki atribut, a svaka grana koja izlazi iz čvora je određena s brojem mogućih vrijednosti za dati atribut.

Operator ***k-NN*** temelji se na algoritmu k najbližih susjeda, odnosno na uspoređivanju danog primjera za testiranje s primjerima za treniranje kojima su slični. Primjeri za testiranje su opisani sa n atributa. Svi primjer predstavlja točku u n-dimenzionalnom prostoru. Svi primjeri za treniranje se pohranjuju u tom prostoru i kada se dobije nepoznati primjer, ovaj algoritam traži u prostoru k primjera za treniranje koji su najbliži nepoznatom primjeru. Tih k primjera za treniranje čine k „najbližih susjeda“ nepoznatom primjeru. „Blizina“ se definira pojmom metričke udaljenost, kao npr. Euklidova udaljenost.

Operator ***Naive Bayes*** je jednostavni probabilistički klasifikator koji se temelji na primjeni Bayesovog teorema (iz Bayesove statistike) s jakim (naivnim) neovisnim pretpostavkama. Klasifikator pretpostavlja da prisutnost (odsutnost) određene značajke neke klase (ili atribut) je nepovezan s prisutnošću (odsutnošću) bilo koje druge značajke. Prednost ovog klasifikatora je ta da zahtijeva malu količinu trening podataka za procjenu sredstava i varijanci potrebnih za klasifikaciju.

Nakon podprocesa za trening slijedi podproces za testiranje (slika 25.)



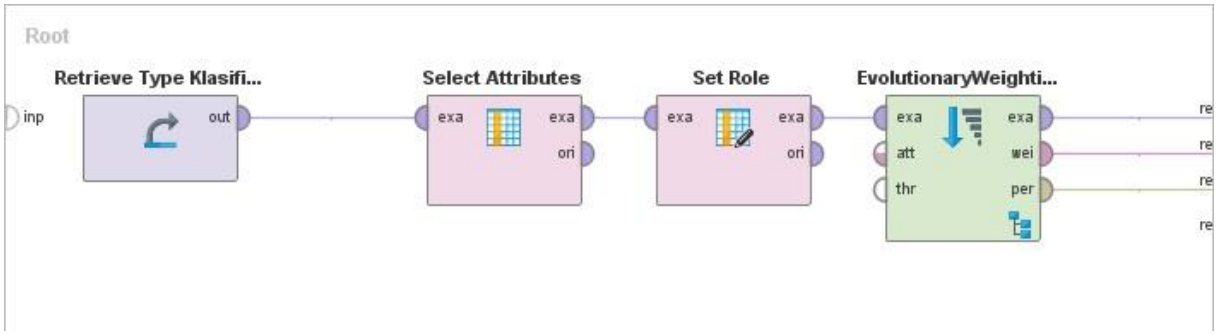
Slika 25. Podproces za testiranje

Operator ***Apply Model to TestSet*** primjenjuje već naučeni ili istrenirani model na primjere za testiranje.

Evaluate Model operator se koristi za procjenu statističkih performansi binomne klasifikacije, odnosno zadatka klasifikacije koji predviđa binomni atribut. Daje listu vrijednosti performansi klasifikacije.

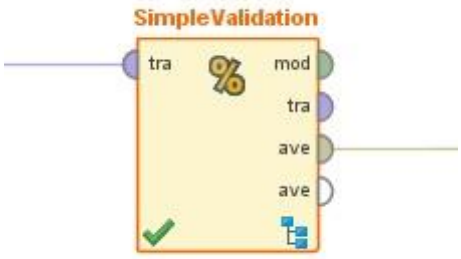
5.3.2. Klasifikacija s optimizacijom

Nakon klasifikacije napravljen je proces za optimizaciju kako bi se pokušali poboljšati rezultati predviđanja određivanjem težina za svaki atribut. Ovaj proces prikazan je na slici 26.



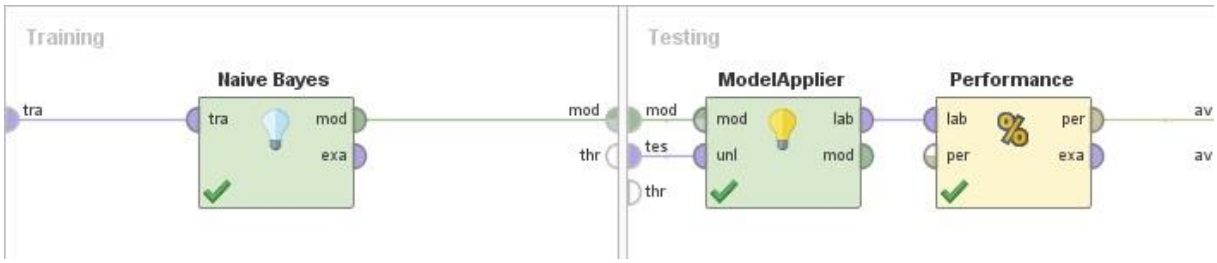
Slika 26. Proces klasifikacije s optimizacijom

Za određivanje težina atributa služi operator *Optimize Weights (Evolutionary)* koji računa relevantnost atributa danog seta primjera koristeći pristup evolucije. On u sebi sadrži podproces koji uvijek mora vraćati vektor performansi. Težine algoritama se računaju korištenjem genetičkog algoritma (GA). Što je veća težina atributa to je veća njegova relevantnost za proces klasifikacije. GA je heuristička potraga koja oponaša proces prirodne evolucije. Ovakva heuristika se rutinski koristi za generiranje korisnih rješenje za optimizaciju i pretraživanje problema. Unutar opisanog operatora nalazi se operator *Simple Validation* prikazan na slici 27.



Slika 27. Operator *Simple Validation*

Pomoću ovog operatora izvršava se jednostavna validacija, odnosno nasumično razdvajanje seta primjera na trening set i test set te procjenjuje model. Validacija razdvajanjem se provodi s ciljem da se ocijene performanse operatora za učenje. Slika 28. prikazuje operatore unutar ovog operatora.



Slika 28. Podprocesi operatora *Simple Validation*

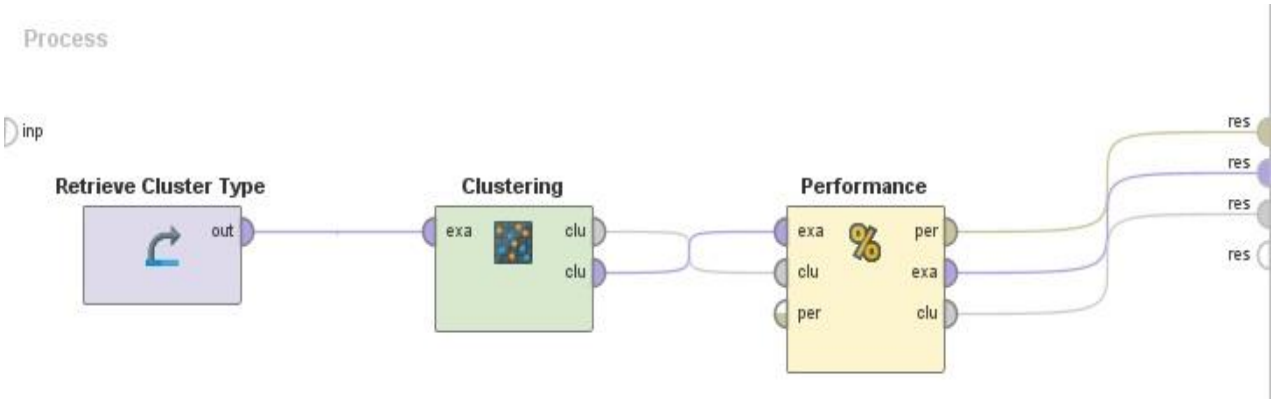
Izabrani operator za učenje je u ovom slučaju ***Naive Bayes*** i opisan je kod procesa klasifikacije. U podprocesu za testiranje naučenog modela su operatori ***Model Applier*** koji primjenjuje model na test skupu i ***Performance*** koji mjeri njegove performanse.

5.3.3. Klasterizacija

Na danom setu podataka korištena su dva algoritma za klasteriranje kako bi se mogli usporediti rezultati dobivenih grupa. Navedeni algoritmi su:

- k-means
- Fuzzy C-means (FCM)

Proces koji koristi k-means operator nalazi se na slici 29.

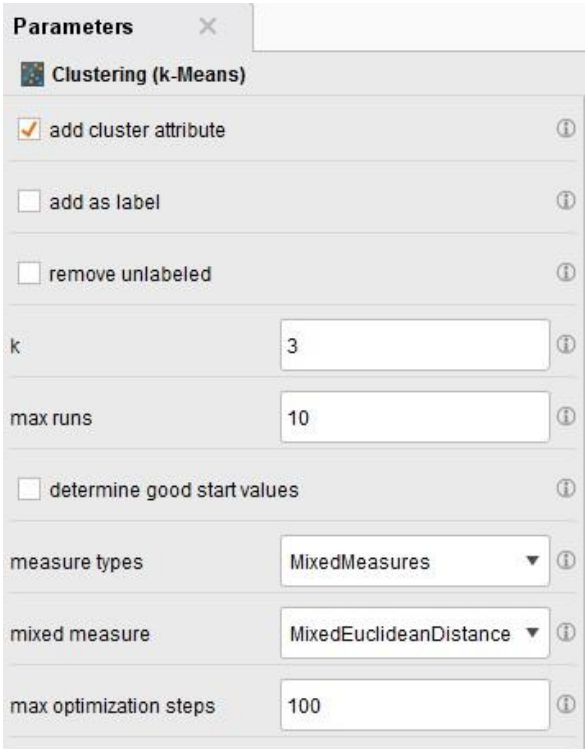


Slika 29. Glavni proces klasterizacije k-means metodom

Klasterizacija je izvršena za raspoređivanje tipova zrakoplova u klastere prema ukupnom broju nesreća za svaki tip. Operator ***Retrieve*** učitava pripremljenu tablicu za klasterizaciju.

Operator ***Clustering*** vrši klasterizaciju k-means metodom. K-means klasteriranje je poseban algoritam, odnosno svaki objekt je dodijeljen točno jednom klasteru. Objekti u jednom klasteru

su slični jedan drugome, a sličnost između objekata se temelji na mjerenju udaljenosti među njima. Na slici 30. prikazani su parametri ovog operatora.

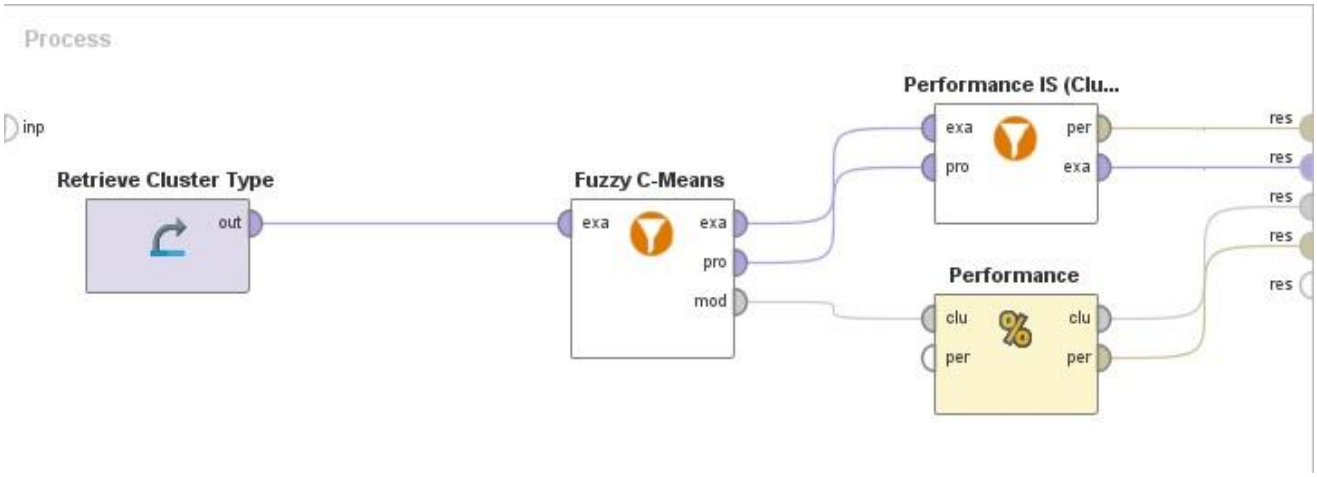


Slika 30. Prikaz parametara k-means operatora

Uključen je parametar *add cluster attribute* koji dodaje atribut s informacijom svakog tipa zrakoplova u koji je klaster smješten. Parametarom *k* se određuje željeni broj klastera. *Max runs* parametar određuje maksimalni broj izvođenja k-means algoritma. Još se mogu odrediti tipovi mjerenja i maksimalan broj koraka optimizacije.

Operator ***Performance***, točnije *Cluster Distance Performance,* koristi se za procjenu performansi metode klasteriranja temeljene na centroidima. Bilježi listu vrijednosti performansi na temelju centralnog klastera.

Glavni FCM proces prikazan je na slici 31.



Slika 31. Glavni proces klasterizacije FCM metodom

Prvi operator isti je kao i kod k-means metode.

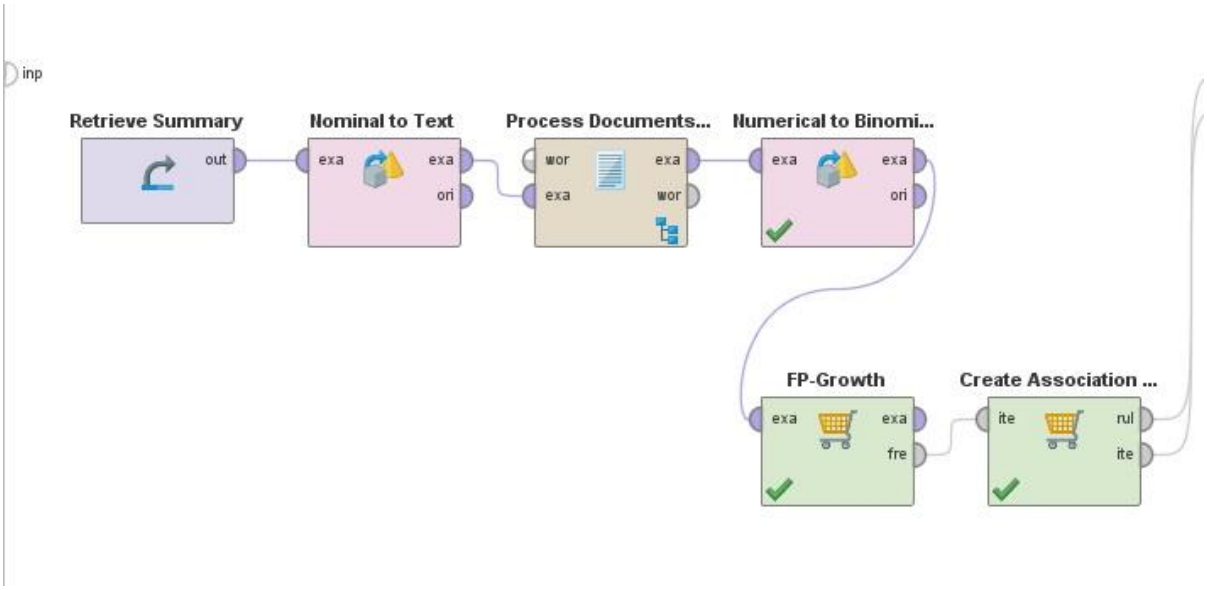
Operator ***Fuzzy C-Means*** izvodi metodu klasteriranja koja omogućava jednom dijelu podataka da pripada u dva ili više klastera. Metoda se često koristi za prepoznavanje uzoraka. Slična je k-means metodi. Algoritam minimizira varijance u klasterima, ali sadrži problem zbog toga što su minimumi lokalni pa rezultat ovisi o inicijalnom izboru težina.

Performance IS (Clustering) operator se koristi za analizu performansi klasterizacije na način da dobiva set prototipova klastera i set primjera kao input te računa varijance unutar klastera. Isti se može koristiti i za k-means algoritam.

Operator ***Performance*** odnosno *Item Distribution Performance* koristi se evaluaciju performansi metode klasteriranja baziranu na distribuciji primjera.

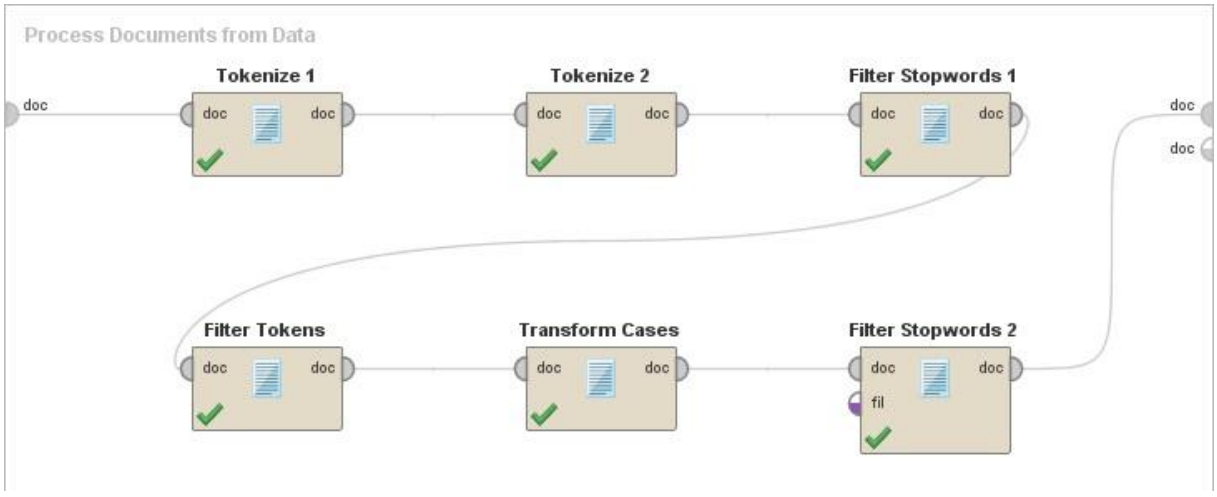
5.3.4. Analiza tekstualnih zapisa

Nakon transformacije nominalnih zapisa unutar atributa *„Summary“* u tekstualne zapise, provedena je analiza teksta korištenjem asocijativnih pravila. Cijeli proces prikazan je na slici 32.



Slika 32. Glavni proces analize tekstualnih zapisa

Operatori za dohvat i transformaciju podataka su navedeni kao operatori za transformaciju. Podproces koji se nalazi u operatoru *Process Documents from Data* prikazan je na slici 33.



Slika 33. Podproces operatora *Process Documents from Data*

Prvi operator ***Tokenize 1*** odvaja riječi na osnovu svako znaka koji nije slovo i postiže da se znak sastoji od jedne riječi.

Tokenize 2 operator odvaja jezične rečenice i podešen je na engleski jezik jer je i set podataka pisan na engleskom.

Operator ***Filter stopwords 1*** izbacuje engleske stopriječi iz dokumenta.

Operator ***Filter Tokens*** je podešen da izbacuje iz dokumenta riječi koja sadrže manje od 3 slova i više od 50 slova.

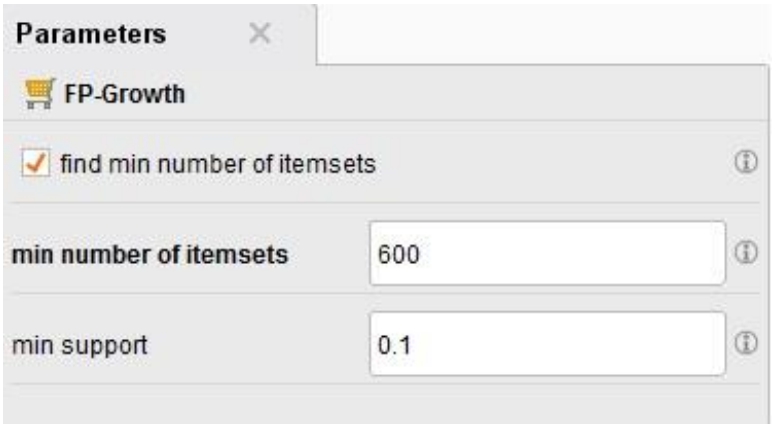
Operator ***Transform Cases*** transformira sva slova u mala slova.

Filter Stopwords 2 operator izbacuje riječi koje su određene od strane korisnika i u ovom slučaju su to sljedeće riječi:

aircraft	*plane*	*crashed*	*crash*	*flight*
flew	*killed*	*land*	*resulted*	*cause*
caused	*air*	*due*	*en*	

Navedene riječi se često pojavljuju u dokumentu, ali ne daju nikakve informacije o zrakoplovnoj nesreći budući da predstavljaju standardne informacije o nesrećama. Filtrirane su iz teksta zbog toga što loše utječu na rezultate interpretacije mogućih uzroka nesreća.

Nakon njih slijedi operator ***FP-Growth*** *(Frequent Pattern-Growth)* učinkovito izračunava sve frekventne skupove parova koristeći FP-stablo strukturu podataka. Svi atributi skupova parova moraju biti binomni zbog čega je potreban prethodno objašnjen operator *Numerical to Binominal*. Potrebno je namjestiti parametar *min support*. Podrška (engl. *support*) se definira kao odnos broja instanci u kojima postoje elementi jednog podskupa, parovi atribut-vrijednost, u odnosu na ukupan broj instanci analiziranog skupa. U frekventne skupove spadaju samo oni podskupovi za koje je podrška veća ili jednaka od definirane vrijednosti minimalne podrške, *min support* (slika 34).



Slika 34. Parametri *FP-Growth* operatora

Operator ***Create Association Rules*** generira set asocijativnih pravila iz danog seta frekventnih skupova parova koristeći kriterije *support* (hrv. podrška) i *confidence* (hrv. pouzdanost) za identifikaciju najvažnijih veza. Podrška je indikacija učestalosti pojavljivanja riječi u bazi podataka. Parametar *criterion* određuje da će se asocijativna pravila selektirati po kriteriju pouzdanosti. (slika 35.).

Slika 35. Parametri operatora *Create Association Rules*

Odabrani kriterij *confidence* kreće se u rasponu od 0 do 1 i pokazuje broj puta kada je if/then uvjet zadovoljen. Definiran je izrazom conf(X implies Y) = supp(X ∪Y)/supp(X), odnosno pouzdanost pojavljivanja riječi X i Y jednaka je podršci pojavljivanja riječi i X i Y podijeljenoj s podrškom pojavljivanja samo riječi X. Od svih generiranih frekventnih podkupova podataka za kreiranje asocijativnih pravila odabiru se samo oni za koje je vrijednost pouzdanosti veća od korisnički definiranog minimalnog praga pouzdanosti, *min confidence*.

6. INTERPRETACIJA REZULTATA I OTKRIVENIH ZNANJA NA SKUPU PODATAKA O ZRAKOPLOVNIM NESREĆAMA

Prije prikazivanja rezultata treba napomenuti da su prikazani samo rezultati s parametrima operatora koji su davali najbolje rezultate.

6.1. Rezultati klasifikacije

Na primjeru stabla odlučivanja, odnosno na rezultatima operatora *Decision Tree* (slika 36.) objašnjene su performanse klasifikacije.

accuracy: 66.70% +/- 1.11% (mikro: 66.69%)

	true 0	true 1	class precision
pred. 0	2269	1087	67.61%
pred. 1	95	98	50.78%
class recall	95.98%	8.27%	

Slika 36. Rezultati operatora *Decision Tree*

Red *pred. 0* govori o tome koliko je puta model predvidio da će vrijednost ciljanog atributa biti 0, a *pred. 1* pokazuje koliko puta je predvidio 1. Stupac *true* pokazuje kolika je bila stvarna vrijednost ciljanog atributa, odnosno kada je bila 0, a kada 1.

To znači da je model 3 356 puta predvidio da će vrijednost ciljanog atributa biti 0, odnosno 2 269 puta je pogodio i 1087 puta nije. Preciznost klase 0 (eng. *class precision*) jest 67,61%. Nadalje, model je 193 puta predvidio da će ciljani atribut biti 1, od čega je 95 puta krivo predvidio, 98 točno. Točnost klase 1 je 50,87%.

Od ukupno 2 364 zrakoplovnih nesreća (vrijednost 0) u kojima nije bilo preživjelih putnika model je pogodio 2 269 puta, a 95 nije. Točnost ove klase još se zove i odziv (eng. *class recall*) i iznosi 95,98%. Od 1185 zrakoplovnih nesreća u kojima je bilo preživjelih (vrijednost 1) model je pogodio samo 98 puta dok u 1087 slučajeva nije. Odziv klase 1 iznosi 8,27%. Ukupna točnost modela (eng. *accuracy*) je 66,70%.

Stablo odlučivanja je grafički prikazano na slici 37.

Slika 37. Stablo odlučivanja

Iz stabla se može vidjeti za svakog proizvođača zrakoplova je li veća vjerojatnost da će biti preživjelih ili da neće. Za proizvođače zrakoplova Convair i McDonnell je veća vjerojatnost da će biti preživjelih dok je za sve ostale izglednije da ih neće biti. Odnos plave i crvene linije prikazuje odnose između broja 0 i 1, točnije plava predstavlja vrijednost 0, a crvena vrijednost 1. Ovo stablo se zapravo sastoji samo od korijena i lista pa se stoga iz njega ne može izvući puno informacija. Zbog toga nije bilo potrebno ni „orezivanje“ stabla za povećanje točnosti.

Sljedeći operator koji se koristio jest *k-NN*, odnosno operator koji koristi algoritam najbližih susjeda (eng. *nearest neighbor*) i njegovi rezultati su prikazani na slici 38.

accuracy: 60.69% +/- 2.62% (mikro: 60.69%)

	true 0	true 1	class precision
pred. 0	1749	780	69.16%
pred. 1	615	405	39.71%
class recall	73.98%	34.18%	

Slika 38. Rezultati operatora *k-NN*

Prije pokretanja procesa bilo je potrebno odrediti parametar k, tj. koliko najbližih susjeda algoritam treba uzeti u obzir. Eksperimentiranjem s vrijednošću tog parametara dobiveno je da najbolje rezultate daje kada zaprima vrijednost k=1. Ovaj operator daje lošiju preciznost klase 1 te odziv klase 0 u odnosu na operator *Decision Tree*, ali zato bolju preciznost klase 0 i što je najvažnije puno veći odziv klase 1. Ukupna točnost iznosi 60,69%.

Zadnji operator koji je korišten je *Naive-Bayes* i njegovi rezultati su prikazani na slici 39.

accuracy: 54.75% +/- 3.39% (mikro: 54.75%)

	true 0	true 1	class precision
pred. 0	1253	495	71.68%
pred. 1	1111	690	38.31%
class recall	53.00%	58.23%	

Slika 39. Rezultati operatora *Naive-Bayes*

Sa slike 39. je vidljivo da ovaj operator za razliku od prošla dva, u više slučajeva predviđa točnu vrijednost ciljanog atributa 1. Time se znatno povećava odziv klase 1, ali to loše utječe na ukupnu točnost modela koja iznosi 54,75%.

Za bolju usporedbu operatora u tablici 3. prikazana je usporedba točnosti sva tri modela.

Tablica 3. Usporedba točnosti operatora klasifikacije

Operator	*Točnost [%]*	*Odziv klase 1 [%]*
Decision Tree	66,70	8,27
k-NN	60,69	34,18
Naive-Bayes	54,75	58,23

Iz tablice je vidljivo da operator *Decision Tree* daje najveću ukupnu točnost predviđanja, ali bez obzira na ukupnu točnost očito je da operator *Naive Bayes* daje bolju točnost kod predviđanja izlaza 1, odnosno odziv klase 1 je kod njega najveći. Upravo zbog toga je ovaj operator odabran za optimizaciju procesa prikazanu u nastavku.

6.2. Rezultati klasifikacije s optimizacijom

Nakon što je utvrđeno da *Naive Bayes* operator najbolje predviđa, izvršena je optimizacija cijelog procesa s istim operatorom. Pomoću operatora za evolucijsko optimiziranje težina dobivene su težine atributa prikazane na slici 40.

attribute	weight ↓
Proizvođač_zrakoplova_uređeno	1
City	0.884
Operator	0.655
Country	0

Slika 40. Težine atributa

Sa slike je vidljivo da atribut *Proizvođač_zrakoplova_uređeno* ima najveću težinu što znači da najviše utječe na predviđanje. Malo manju težinu ima atribut *City*. Zatim slijedi atribut *Operator*, a *Country* odnosno zemlja u kojoj se dogodila nesreća nema utjecaja na predviđanje.

Dobivena točnost ovim procesom prikazana je na slici 41.

accuracy: 57.56%

	true 0	true 1	class precision
pred. 0	383	149	71.99%
pred. 1	303	230	43.15%
class recall	55.83%	60.69%	

Slika 41. Točnost procesa klasifikacije s optimizacijom

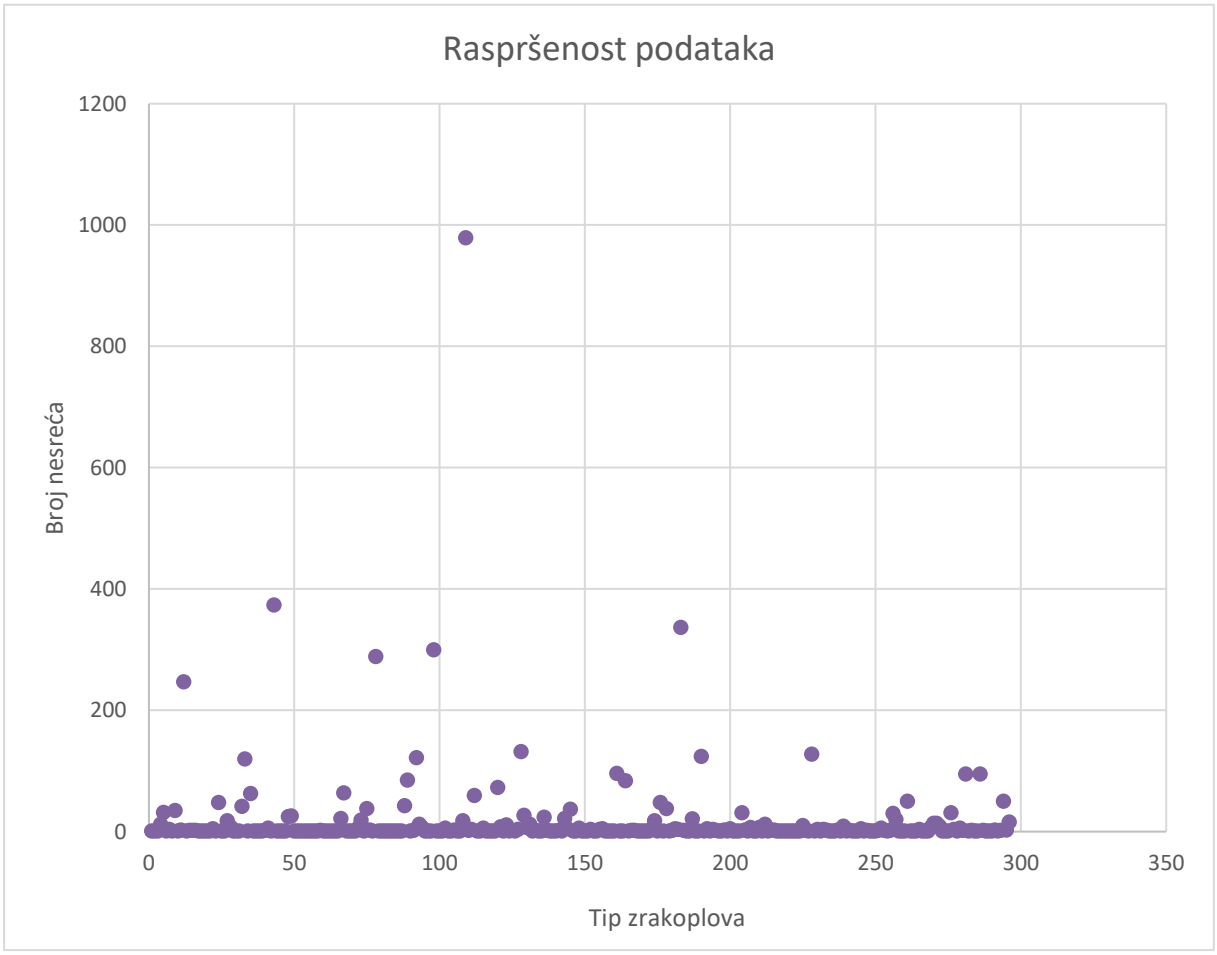
Operator *Simple Validation* dijeli set na trening set i test set u omjeru 70/30 (%). Zbog toga je u ovom slučaju broj pogađanja znatno manji, odnosno 30% ukupnog broja zapisa.

S dobivenom ukupnom točnošću od 57,56% i poboljšanjem odziva klase 1 na 60,69% proces je neznatno poboljšan.

Operator *Evolutionary Weighting* poboljšavao je proces kroz 20 generacija s veličinom populacije 5, a u PRILOGU 1 nalazi se tablica koja prikazuje mjerenje performansi kroz generacije i iz nje je vidljivo da su najbolje performanse dobivene već u 7. generaciji.

6.3. Rezultati klasterizacije

Prvi parametar koji treba odrediti da bi se uopće mogla provesti klasterizacija jest broj klastera „k“. To ujedno predstavlja i glavni problem jer ne postoji zadovoljavajuće rješenje, a iteratitvne metode zahtijevaju od korisnika da unaprijed odredi broj klastera. Postoje jedino mjere koje govore o povezanosti određenih klastera. Osim navedenih mjera, moguće je određivanje klastera i vizualnom metodom. što će biti primijenjeno i u ovom radu (slika 42.).



Slika 42. Raspršenost podataka broja nesreća za tipove zrakoplova

Za navedene podatke odabran je broj klastera k=3, nakon čega je izvršena klasterizacija tipova zrakoplova prema ukupnom broju zabilježenih nesreća. Kao što je prethodno spomenuto, korištene su dvije metode, k-means i FCM.

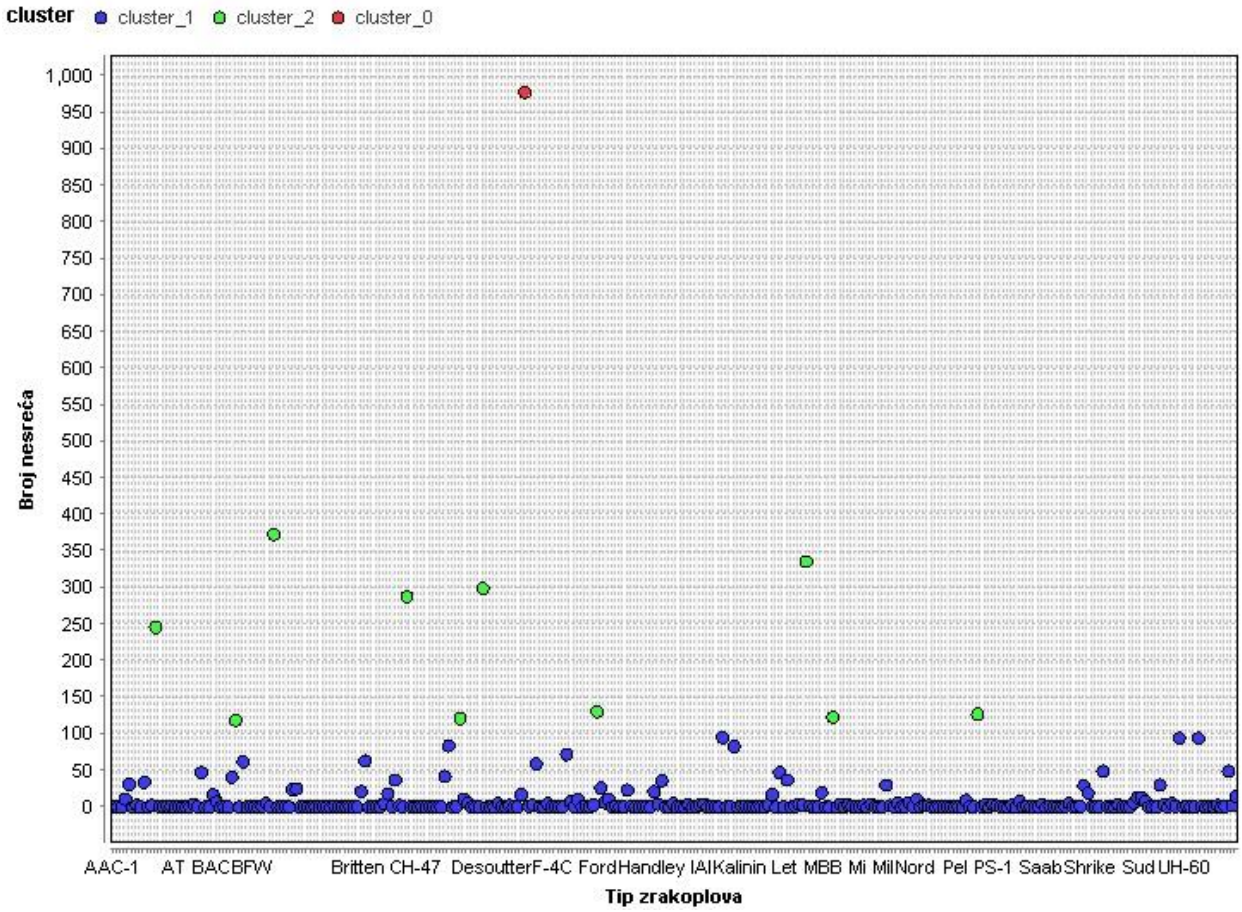
6.3.1. Rezultati k-means algoritma

Tablica 4. prikazuje raspoređenost u klasterima za k-means metodu.

Tablica 4. Raspoređenost zapisa po klasterima za k-means metodu

Nominalna vrijednost	Broj
klaster 0	1
klaster 1	285
klaster 2	10

Iz tablice je vidljivo da raspoređenost nije ravnomjerna. U klasteru 1 se nalazi većina tipova zrakoplova (PRILOG 2), dok je u klasteru 0 samo jedan tip (slike 43.).



Slika 43. Graf klastera k-means metode za tipove zrakoplova

Na grafu se vidi da je većina tipova zrakoplova imala manje od 100 nesreća u promatranom razdoblju (tamno plavi kružići) i oni pripadaju u klaster 1. U klasteru 2 nalazi se 10 tipova zrakoplova (zeleni kružići) i prikazani su na slici 44.

Tip zrakoplo...	cluster ↑	Broj nesreća
Antonov	cluster_2	247
Beechcraft	cluster_2	120
Boeing	cluster_2	374
Cessna	cluster_2	289
Curtiss	cluster_2	122
De Havilland	cluster_2	300
Fokker	cluster_2	132
Lockheed	cluster_2	337
McDonnell	cluster_2	124
Piper	cluster_2	128

Slika 44. Rezultati klasterizacije za klaster 2

Posljednji klaster 0 (označen crvenom bojom) sadrži samo jedan tip zrakoplova koji je sudjelovao u čak 979 nesreća. On po broju nesreća znakovito odskače od ostalih tipova i zbog toga je dobio zaseban klaster. Radi se o tipu zrakoplova Douglas u koji je ukrcano ukupno 20 422 putnika od kojih je 16 619 poginulo. Tim nesrećama zahvaćeno je još 107 osoba koje su smrtno stradale, a nisu sudjelovale u letu.

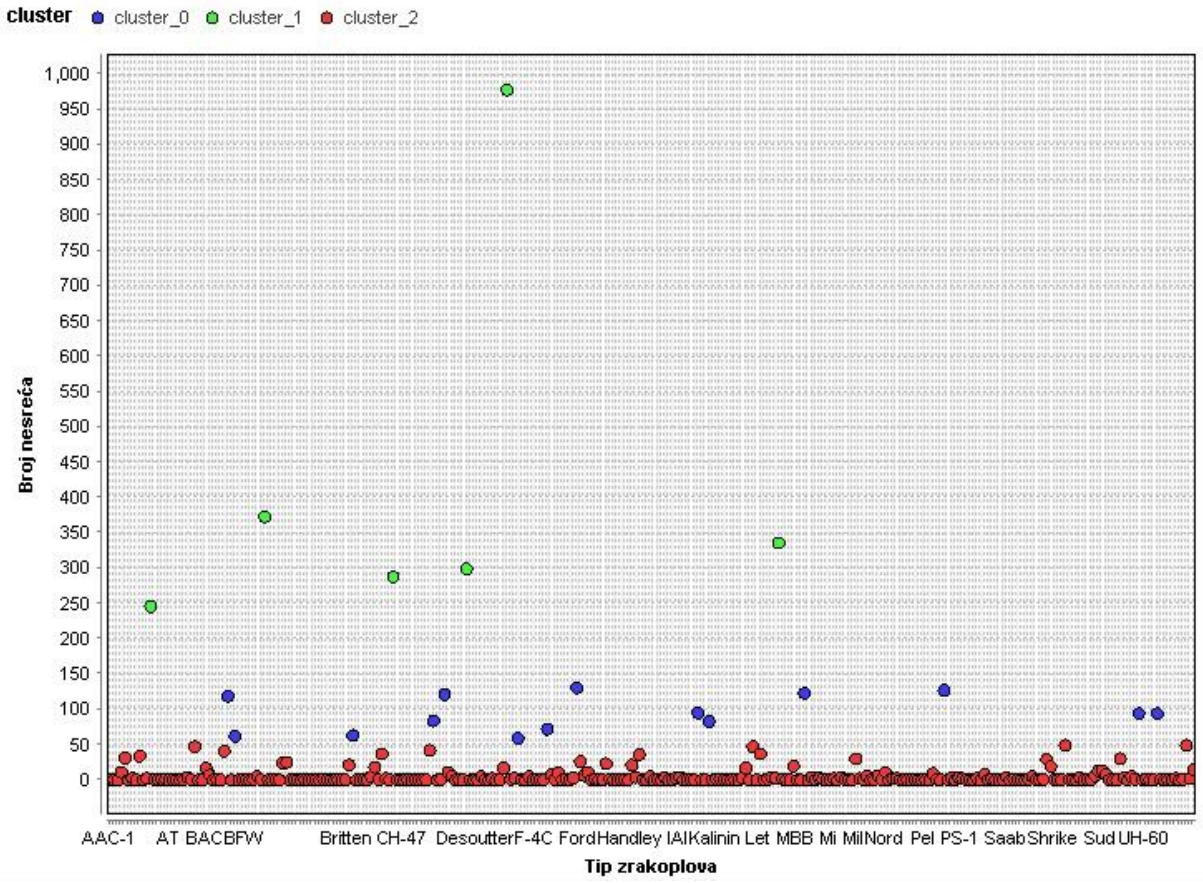
6.3.2. Rezultati Fuzzy C-means metode

Tačica 5. prikazuje raspored tipova zrakoplova po klasterima dobivenog FCM metodom.

Tablica 5. Raspoređenost zapisa po klasterima za FCM metodu

Nominalna vrijednost	Broj
klaster 0	14
klaster 1	6
klaster 2	276

Raspoređenost po klasterima kod ove metode razlikuje se nego kod KM metode što se vidi i na grafu sa slike 45.



Slika 45. Graf klastera FCM metode za tipove zrakoplova

U ovom slučaju, u najbrojnijem klasteru 2 nalaze se tipovi zrakoplova s do 50 nesreća u promatranom razdoblju (crveni kružići). U klasteru 0 nalaze se tipovi s više od 50 do 130 nesreća (plavi kružići). Klaster 1 s najmanje tipova, odnosno s najvećim odstupanjem u broju

nesreća sadrži 6 tipova prikazanih na slici 46. Detaljna distribucija tipova zrakoplova u ostalim klasterima je prikazana u PRILOGU 3.

Tip zrakoplo...	cluster ↑	Broj nesreća
Antonov	cluster_1	247
Boeing	cluster_1	374
Cessna	cluster_1	289
De Havilland	cluster_1	300
Douglas	cluster_1	979
Lockheed	cluster_1	337

Slika 46. Rezultati klasterizacije za klaster 1

U tablici sa slike vidljivo je da su to tipovi zrakoplova koji su doživjeli od 247 pa do 979 nesreća.

6.3.3. Usporedba rezultata

Budući da je za mjeru učinkovitosti metoda klasteriranja odabrana suma kvadrata odstupanja, rezultati pojedine metode su prikazani u tablici 6.

Tablica 6. Usporedba vrijednosti sume kvadrata odstupanja za obje metode

METODA	Suma kvadrata odstupanja
k-means	0,928
FCM	0,872

Analizirajući rezultate sume kvadrata odstupanja temeljem raspodijeljenosti zapisa dobivenih pomoću dvije metode klasteriranja , uočeno je sljedeće:

- ukoliko u zapisima postoje ekstremne vrijednosti (kao što je u ovom slučaju broj nesreća za tip zrakoplova Douglas), k-means metoda teži ekstremnom grupiranju zapisa, tj. generiranju jednog manjeg klastera i dva veća i time će suma kvadrata biti bolja na manjim klasterima (bliža vrijednosti 1), što u konačnici ne znači bolju distribuciju zapisa

- FCM metoda teži ravnomjernijem grupiranju zapisa unutar grupa klastera te su time dobiveni lošiji rezultati sume kvadratnih odstupanja, no bolja distribucija zapisa

Tablica 7. prikazuje podudarnost, tj. raspodjelu zapisa (tipova zrakoplova vidljiv iz PRILOGA 2 i PRILOGA 3) unutar uspoređenih metoda.

Tablica 7. Podudarnost dobivenih klastera k-means i FCM metode

	k-means vs. FCM [% podudarnosti zapisa]	
Usporedba 1	klaster 1 (k-means) = klaster 2 (FCM)	96,84%
Usporedba 2	klaster 2 (k-means) = klaster 0 (FCM)	35,71%
Usporedba 3	klaster 0 (k-means) = klaster 1 (FCM)	16,67%

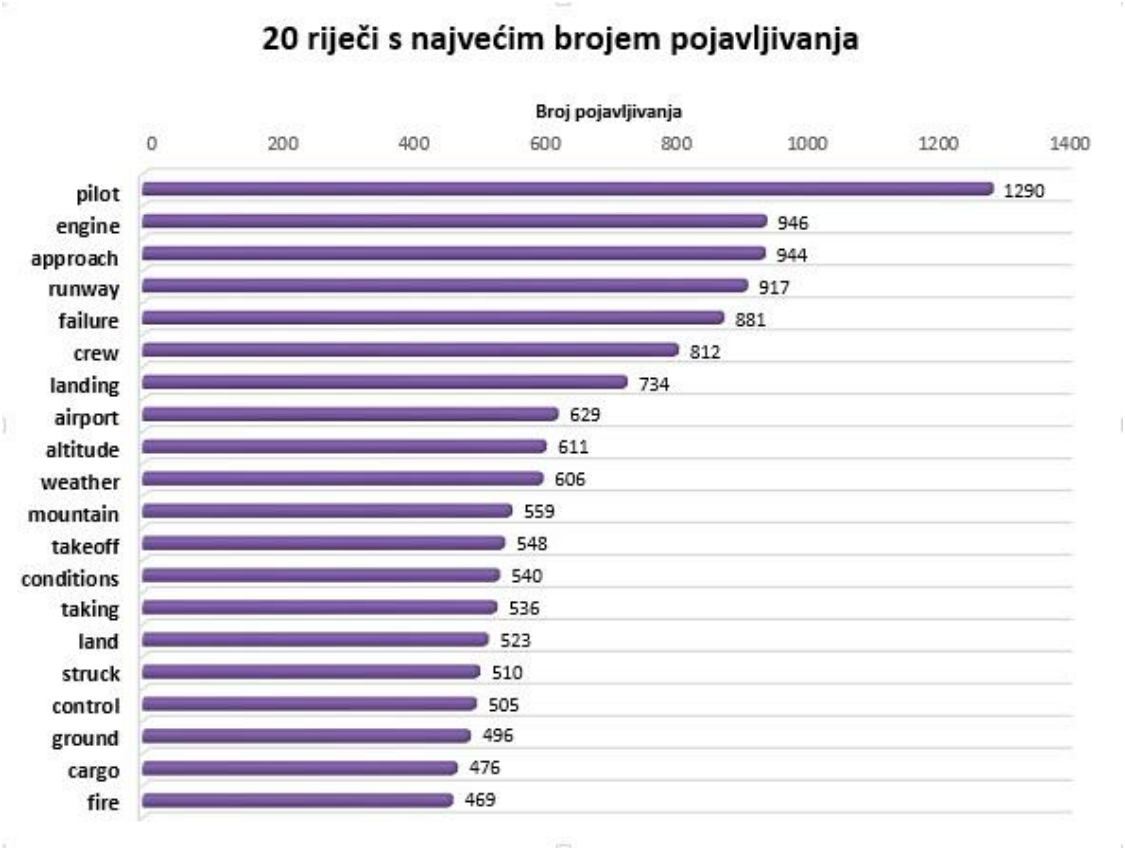
Iz tablice 7. je vidljivo da je podudarnost u rezultatima najveća u slučaju *usporedbe 1*, tj. kada su uspoređeni najveći klasteri obiju metoda, tj. klasteri koji kod k-means metode prikazuju broj grupe zrakoplova s brojem nesreća manjim od 50, a kod FCM metode manjim od 100.
Najmanja podudarnost je dobivena u manjim klasterima, što potvrđuje i prethodno navedene zaključke o ekstremnom grupiranju zapisa k-means metode.

6.4. Rezultati tekstualne analize

Za bolje razumijevanje dobivenih asocijativnih pravila najprije je potrebno navesti najfrekventnije riječi, a zatim usporediti za svaku riječ posebno povezanost s ostalim riječima.

6.4.1. Interpretacija pojave frekventnih riječi

Nakon što su filtrirane riječi s velikom frekvencijom koje nemaju utjecaj na otkrivanje novih znanja, istaknuto je 20 riječi s najvećom frekvencijom i prikazane su u grafu na slici 47.



Slika 47. Prikaz 20 riječi s najvećim brojem pojavljivanja

Za 20 riječi prikazanih na slici 47. je pretpostavljeno da su najčešći uzrok zrakoplovnih nesreća. Analizom prvih pet pojmova: *pilot* (hrv. pilot), *engine* (hrv. motor), *approach* (hrv. prilaz), *runway* (hrv. pista) i *failure* (hrv. kvar) stvaraju se sljedeće pretpostavke: Je li pilot kriv za pad zrakoplova ili se radi o općem terminu? Je li kvar motora najčešće uzrokom nesreće ili prilaz na pistu? Kako bi se riješile navedene nedoumice, potrebno je detaljno analizirati povezanost frekventnih pojmova s ostalim pojmovima u zapisima o nesrećama pronađenih asocijativnim pravilima.

6.4.2. Interpretacija asocijativnih pravila

Za točnije donošenje zaključaka potrebno je proučiti povezanost među dobivenim riječima. Za svaku promatranu riječ prikazana je tablica s riječima koje su povezane uz nju s najvećom pouzdanošću. Popis svih dobivenih asocijativnih pravila nalazi se u PRILOGU 4.

Tablica 8. prikazuje s kojim riječima je povezan pojam „pilot“.

Tablica 8. Povezanost s pojmom „PILOT“

"PILOT"					
command	*factors*	*contributing*	*minimum*	*clearance*	*instrument*
0,89	0,70	0,58	0,57	0,57	0,54
error	*reported*	*maintain*	*decision*	*accident*	*ifr*
0,53	0,52	0,51	0,50	0,43	0,43
attempted	*vfr*	*improper*	*continued*	*conditions*	*descend*
0,41	0,41	0,40	0,39	0,39	0,38
make	*adverse*	*terrain*	*low*	*procedures*	*turn*
0,37	0,36	0,36	0,36	0,34	0,34
control	*weather*				
0,31	0,30				

Neke riječi navedene u tablici su blisko povezane s pojmom „pilot“, ali proučavanjem njihove povezanosti ne može se detaljnije opisati uzrok nesreće. Zbog toga su niže navedene samo riječi koje tvore smisleno asocijativno pravilo:

- *command* (hrv. naredba, zapovjedništvo, komandant)

Uzrok zrakoplovne nesreće povezan s riječi *command* može biti kriva naredba zapovjedništva, nepoštivanje danih naredbi, loša procjena zapovjednog pilota, odnosno kapetana zrakoplova.

- *minimum* (hrv. minimum)

Minimum se može povezati s minimalnom visinom za slijetanje koju pilot nije poštovao ili nije imao uvjete za to. Također se može raditi o minimalnim vremenskim uvjetima pogodnim za letenje.

- *clearance* (hrv. čistina, dozvola), *terrain* (hrv. teren, zemljište)

Ako se gleda prijevod riječi čistina može se zaključiti da se radi o nedostatku čistine za slijetanje ili terena bez prepreka, tj. nemogućnost pilota za pronalazak iste. Prijevod dozvola govori o nedostatku dozvole, npr. za slijetanje koju pilot nije poštovao.

- *instrument* (hrv. instrument), *conditions* (hrv. uvjeti)

Prva asocijacija riječi instrument su instrumenti na ploči u pilotskoj kabini koje pilot prati tijekom letenja. Može doći do greške mjerenja stanja zrakoplova kao i njegove pozicije, a takve greške lako mogu dovesti do nesreće. Druga asocijacija je izraz *Instrument meteorological conditions* (IMC) (hrv. instrumentalni meteorološki uvjeti), a radi se o uvjetima manjim od minimuma utvrđenih za vizualne meteorološke uvjete (vidljivost, odstojanje od oblaka i baza oblaka). Znači česti uzrok zrakoplovnih nesreća je letenje u instrumentalnim meteorološkim uvjetima.

- *error* (hrv. greška)

Ovdje se bez sumnje može reći da je čest uzrok nesreća upravo greška pilota.

- *maintain* (hrv. održavati), *low* (hrv. nizak)

Ako pilot ne može održavati adekvatnu brzinu te održavati kontrolu nad zrakoplovom, to može biti čest uzrok nesreće. Ako ne može održavati sigurnu visinu letenja letjet će prenisko i postoji velika mogućnost da će zapeti za prepreku.

- *decision* (hrv. odluka)

Očiti uzrok nesreće može biti kriva odluka pilota.

- *ifr - instrument flight rules* (hrv. instrumentalna pravila letenja), *vfr - visual flight rules* (hrv. pravila vizualnog letenja)

Za letenje prema instrumentalnim i vizualnim pravilima treba slijediti propisane procedure kako ne bi došlo do nesreće. Često pilot ne slijedi zadane procedure ili uopće nema certifikat za letenje u posebnim uvjetima koji zahtijevaju spomenuta pravila. Može doći i do letenja po pravilima VFR-a umjesto IFR-a.

- *attempt* (hrv. pokušat), *descend* (hrv. spustiti), *turn* (hrv. okrenuti, zaokret)

Uzrok nesreće može biti neuspješan pokušaj pilota slijetanja ili pokušaj okreta zrakoplova. Neuspješno slijetanje može uzrokovati prerano spuštanje zrakoplova, neravan teren, premala visina. Također to može biti odluka pilota da okrene zrakoplov i vrati ga na polazište kako bi pokušao izbjeći neke druge faktore kao moguće uzroke nesreće.

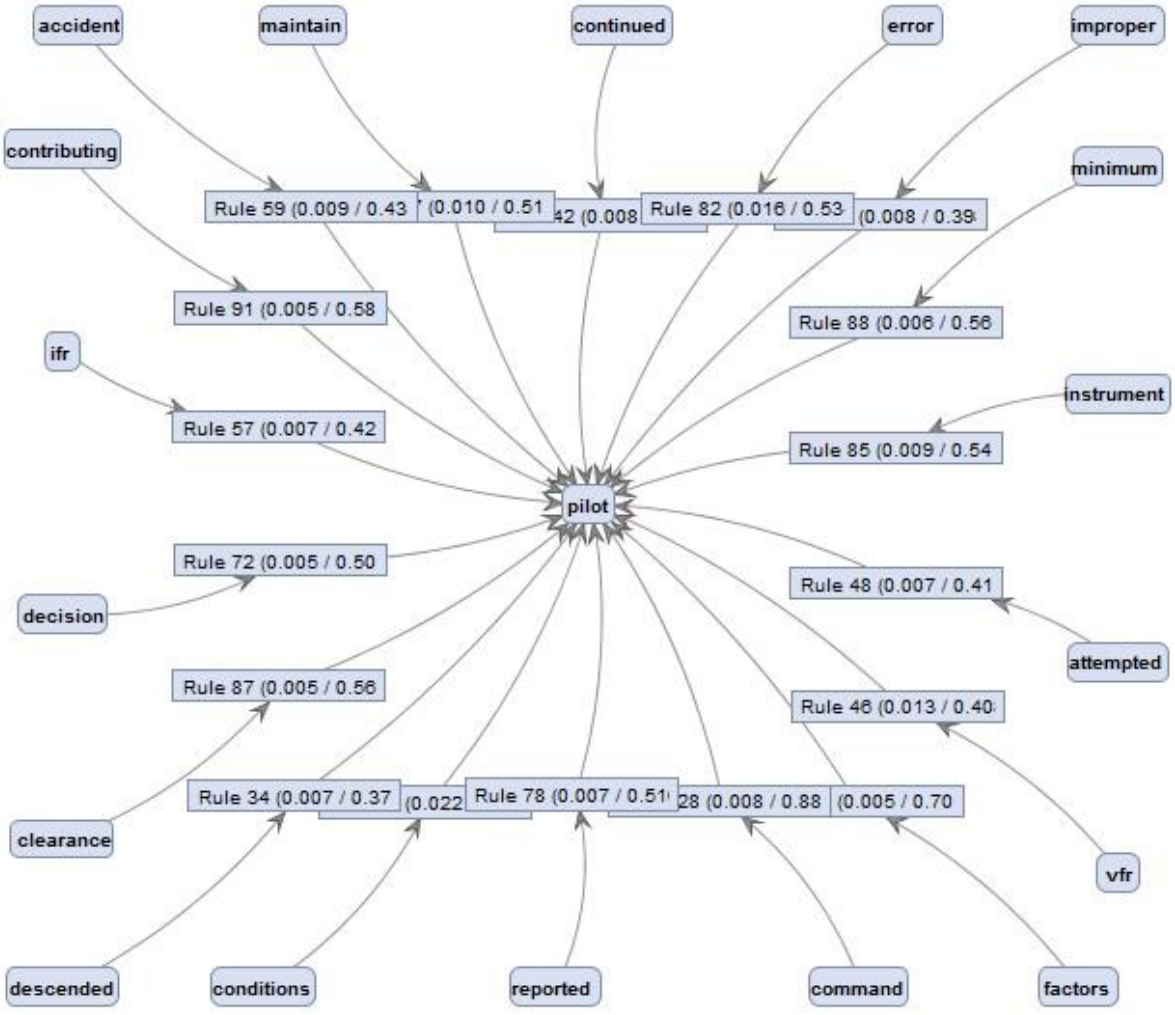
- adverse (hrv. nepovoljno), weather (hrv. vrijeme)

Većina dosada spomenutih uzroka nesreća mogu biti posljedica upravo nepovoljnih vremenskih uvjeta. Smanjenje vidljivosti, jak vjetar, udari gromova. Sve to spada u širok opseg nepovoljnih vremenskih uvjeta.

- *control* (hrv. kontrola)

Ukoliko pilot iz nekog razloga izgubi kontrolu nad zrakoplovom teško će se izbjeći nesreća.

Povezanost riječi „pilot“ sa spomenutim riječima također je prikazana grafički na slici 48.



Slika 48. Graf povezanosti za pojam „*pilot*“

Iz grafa sa slike 48. je vidljivo koje riječi imaju najveću podršku (prvi broj u zagradi) i pouzdanost (drugi broj u zagradi) s pojmom pilot.

Prema statističkim podacima PlaneCrashInfo.com baze podataka [27] koja sadrže točne informacije o 1 104 zrakoplovne nesreće(u vremenskom razdoblju od 1.1.1960. do 31.12.2015.) čak 58% nesreća uzrokovano je zbog greške pilota što opravdava i dobivene rezultate u ovom radu gdje je dobiveno uvjerljivo najviše riječi u korelaciji s pojmom.

Tablica 9. prikazuje s kojim je riječima najviše povezan pojam „engine“ .

Tablica 9. Povezanost s pojmom „ENGINE“

"ENGINE"					
trouble	*experiencing*	*losing*	*return*	*failure*	*emergency*
0,87	0,77	0,68	0,62	0,49	0,46
power	*failed*	*fire*	*caught*	*attempted*	*lost*
0,46	0,39	0,37	0,36	0,33	0,32
loss	*takeoff*				
0,31	0,31				

U ovom slučaju također će biti navedene i objašnjene samo riječi koje tvore smisleno asocijativno pravilo:

- *trouble* (hrv. nevolja, kvar, nezgoda), *failure* (hrv. kvar, nedostatak)

Uzrok nesreće je kvar motora. Ovo je dosta općenit uzrok, ali često se ne zna koji se specifični problem pojavio na motoru.

- *losing* (hrv. gubljenje), *lost* (hrv. izgubljen)

Uzrok nesreće je obično gubljenje jednog ili više motora. Ovdje se radi o gubljenju u kontekstu prestanka rada.

- *power* (hrv.snaga), *loss* (hrv. gubitak)

Jedan od većih problema koji se može dogoditi s motorom je i pad snage. To se može dogoditi zbog neispravnosti jednog od sustava, nedostatka goriva, ali nekad iz sasvim nepoznatih razloga.

- *fire* (hrv. vatra), *caught* (hrv. zahvaćen)

Zbog kvara na motoru dolazi do požara, odnosno motor se može zapalit što je razumljiv uzrok nesreće.

- *takeoff* (hrv. uzlijetanje)

Zatajenje motora jako često se događa upravo kod uzlijetanja jer vjerojatno motor nije dobro održavan i provjeren, što dovodi do velikih problema kada treba proizvesti dovoljno snage za uzlijetanje.

Tablica 10. prikazuje koje su riječi povezane s pojmom „approach“.

Tablica 10. Povezanost s pojmom „APPROACH“

"APPROACH"					
ils	*final*	*missed*	*visual*	*instrument*	*procedure*
0,93	0,92	0,79	0,63	0,59	0,56
minimum	*making*	*descent*	*procedures*	*improper*	*ifr*
0,47	0,47	0,39	0,38	0,38	0,35
captain	*descend*				
0,34	0,33				

- *ILS – Intrument Landing Sytem* (hrv. Sustav za instrumentalno slijetanje)

Ovaj sustav jest navigacijski sustav za precizno instrumentalno vođenje zrakoplova u završnoj fazi prilaženja i slijetanja na uzletno-sletnu stazu. Do nesreće može doći ako pilot nije dovoljno iskusan s ovim sustavom ili sustav daje krive upute.

- *final* (hrv. konačni, završni)

Velik dio nesreća događa se upravo pri završnom prilaženju iz čega se može zaključiti da je to kritična faza slijetanja.

- *missed* (hrv. promašeno)

Missed approach je engleski izraz za postupak neuspjelog prilaženja koji se mora započeti ako ne postoji potrebna vidljivost vizualnih orijentira za nastavak prilaženja. Tu može nastati problem ukoliko pilot nije dovoljno iskusan ili su uvjeti previše nepovoljni.

- *visual* (hrv. vizualno)

Visual approach (hrv. vizualno prilaženje) daje informaciju o tome da će nesreća nastati ukoliko se slijetanje zrakoplova provodi u uvjetima koji nisu za vizualno prilaženje, odnosno kada dođe do krive prosudbe uvjeta.

- *minimum* (hrv. minimum), *descent, descend* (hrv. slijetati)

Prilikom slijetanja, definirana je minimalna visina kod koje je potrebno započeti fazu slijetanja te ako pilot pokuša sletjeti ispod tog minimuma u navedenoj fazi postoji velika šansa za nesreću.

- *captain* (hrv. kapetan), *improper* (hrv. nepravilno, neispravno)

Čest uzrok nesreća su neispravne upute kapetana zrakoplova pri fazi prilaženja. Ovaj uzrok se može primijeniti na sve faze leta.

Tablica 11. prikazuje povezanosti s pojmom „runway“.

Tablica 11. Povezanost s pojmom „RUNWAY“

"RUNWAY"	
overran	*end*
0,97	0,79

Značenje riječi „runway“ na hrvatskom je pista, a niže su navedene riječi koje su s njom povezane:

- *overran* (hrv. pregaziti, preletjeti), *end* (hrv. kraj, završetak)

Nesreća se dogodi ako se zrakoplov ne uspije zaustaviti, već preleti pistu preko završetka gdje se obično nalaze prepreke koje mogu zazvati kobne posljedice.

Tablica 12. prikazuje povezanosti s pojmom „failure“.

Tablica 12. Povezanost s pojmom „FAILURE“

"FAILURE"					
structural	*maintain*	*experiencing*	*procedures*	*fatigue*	*engine*
0,93	0,57	0,44	0,43	0,4	0,39

Već je navedene da je značenje riječi „failure“ kvar, a zbog čega do njega dolazi nalazi se u nastavku:

- *structural* (hrv. strukturni)

Najčešća vrsta kvara je upravo strukturni ili mehanički kvar povezan sa strukturom zrakoplova.

- *maintain* (hrv. održavati)

Ovdje se obično radi o neuspjehu da se održe potrebne brzina i visina zrakoplova, te kontrola nad njim. Također se radi o kvaru motora do kojeg može doći zbog lošeg održavanja.

- *procedures* (hrv. procedure)

Ovdje kao uzrok možemo navesti neuspješno obavljanje posebnih procedura od strane članova posade, ali i tima za održavanje.

- fatigue (hrv. zamor, umor)

Ako se ova riječ odnosi na pilota i posadu, do kraha može doći zbog njihovog umora, a s time i manje sposobnosti razlučivanja. Podaci sa stranice Skybrary [28] koji istražuju utjecaj ljudskog faktora u zrakoplovnim nesrećama navodi listu „Dirty Dozen“ s najčešćim uzrocima

među kojima se nalazi upravo umor. Zamor materijala pak može napraviti štetu na bilo kojem dijelu zrakoplova o čemu treba voditi računa odjel za održavanje zrakoplova.
Tablica 13. prikazuje povezanost riječi s pojmom „landing“.

Tablica 13. Povezanost s pojmom „LANDING“

"LANDING"					
gear	*make*	*forced*	*emergency*	*attempt*	*made*
0,86	0,75	0,66	0,65	0,61	0,42
attempted					
0,37					

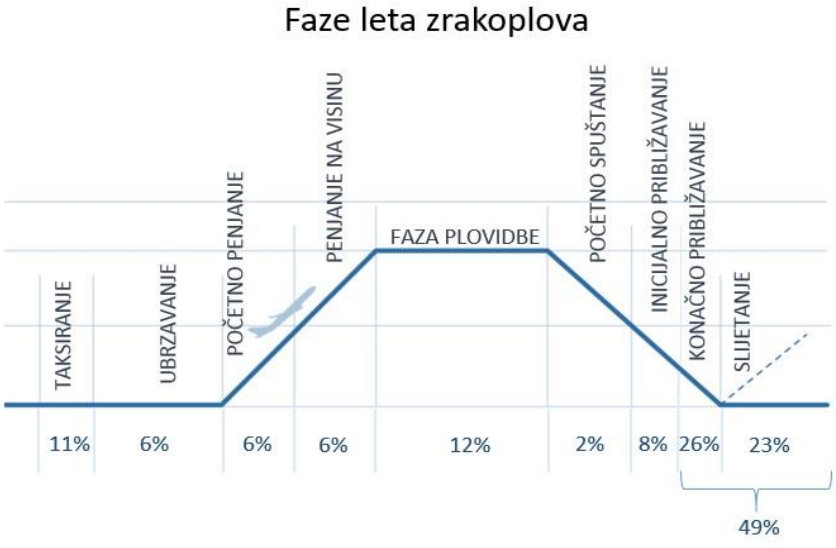
Prijevod riječi „landing“ na hrvatskom je slijetanje i najčešći uzroci fatalnog slijetanja su:

- *gear* (hrv. pogon)

Landing gear u kombinaciji daju opremu za slijetanje. Ako samo jedan kotač za slijetanje zapne ili na neki drugi način podbaci dolazi do nesreće. U fazi slijetanja to je jedan od najčešćih problema koji se javlja.

- *forced* (hrv. prisilno), *emergancy* (hrv. hitno)

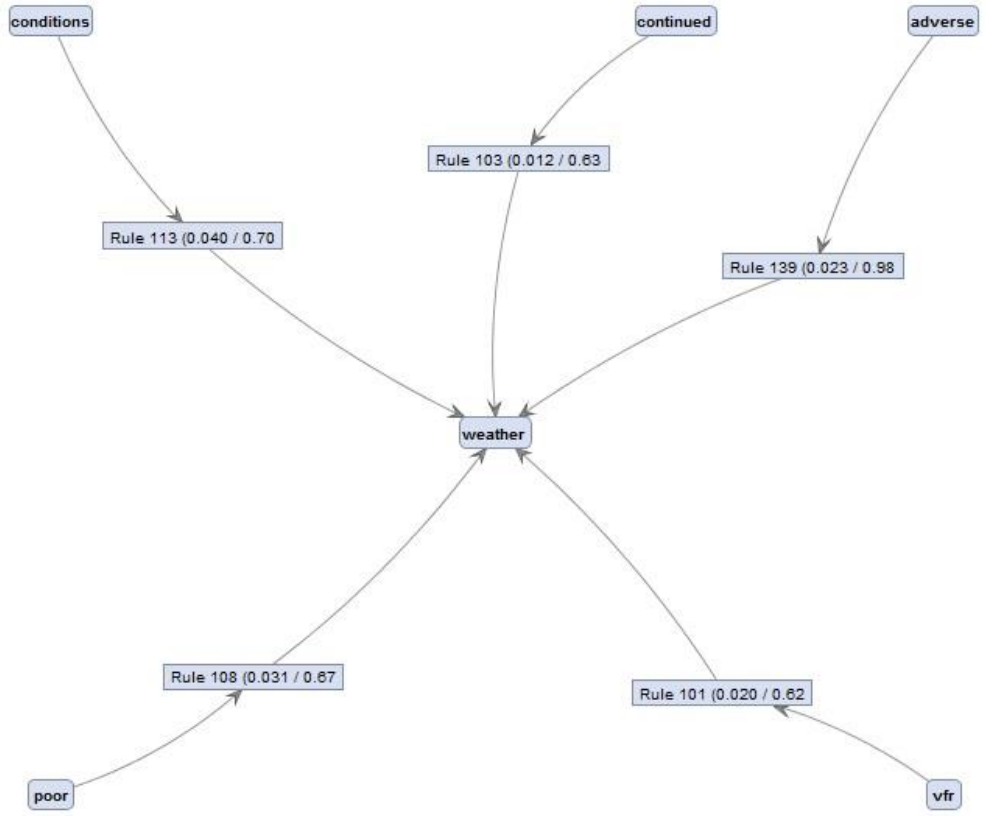
Do nesreće često dolazi kod pokušaja prisilnog slijetanja no takvo slijetanje je zapravo samo posljedica nekog drugog uzroka.
Iz analize rezultata se može zaključiti da je upravo završna faza leta, odnosno slijetanje, najopasnija faza. To potvrđuje i graf iz Boeingovog istraživanja [26] koji prikazuje postotak nesreća po fazama leta (slika 49.)



Slika 49. Postotak nesreća po fazama leta (2006.-2015.) [26]

Iz slike 49. je vidljivo da je najopasnije faze leta upravo faze konačnog približavanja i slijetanja jer se u njima događa čak 49% nesreća.

Na slici 50. je prikazan graf povezanosti riječi s pojmom „weather“.



Slika 50. Graf povezanosti za pojam „weather“

Weather conditions su pojmovi koji se mogu promatrati kao cjelina, a na hrvatskom znače vremenski uvjeti. Loši vremenski uvjeti znatno smanjuju vidljivost i zbog toga se ne mogu koristiti pravila vizualnog letenja (VFR). Posada se može osloniti samo na instrumente i tu dolazi do krive prosudbe, nepredvidivih faktora koji postaju uzrok zrakoplovne nesreće.

7. ZAKLJUČAK

Analizom literature o poslovnoj inteligenciji uočeno je da su osnovni pojmovi potrebni za razumijevanje ovog područja podaci, informacije, znanje, inteligencija i mudrost. Informacija se definira preko podataka, znanje preko informacija, inteligencija preko znanja, a mudrost preko inteligencije. Također je uočeno da različite tehnike rudarenja podataka, koje su opisane u trećem poglavlju, omogućavaju identifikaciju uzoraka i trendova između podataka kako bi se predvidjeli budući događaji.

Proces rudarenja podataka, od prikupljanja i čišćenja podataka, sve do odabira metoda, detaljno je opisan u četvrtom poglavlju. Slijedeći taj proces na setu podataka koji opisuje vojne i civilne zrakoplovne nesreće (u periodu od 1908. do 2009. godine) dobivena su korisna znanja i otkriveni slični zapisi. Deskriptivnom analizom podataka postignuto je bolje razumijevanje „sirovih" zapisa čime je olakšan nastavak procesa. Opisan je trend kretanja broja nesreća kroz godine, izdvojeni su određeni tipovi zrakoplova i zrakoplovni operatori koji se ističu po broju nesreća te je uspoređeno kretanje ukrcanih i poginulih osoba koje su sudjelovale u spomenutim nesrećama.

Nakon odabira različitih algoritama unutar metoda klasifikacije i klasterizacije stvoreni su različiti prediktivni modeli. Kod metoda klasifikacije uočeno je da najtočnije predikcije za vrijednost 1 ciljanog atributa daje algoritam Naive Bayes. Sukladno tome, navedeni proces je optimiziran pomoću operatora koji generira težine atributa po njihovu utjecaju na klasifikaciju, a na temelju genetičke evolucije. Klasterizacijom su dobivene grupe tipova zrakoplova prema ukupnom broju nesreća u promatranom razdoblju. Korištene su dvije metode, k-means i FCM. Uočeno je da k-means teži ekstremnom grupiranju zapisa , tj. generiranju jednog manjeg klastera i dva veća, dok FCM metoda teži ravnomjernijem grupiranju zapisa unutar grupa klastera.

Naposljetku, odrađena je analiza teksta, tj. izdvojeni su najfrekventniji pojmovi za koje je zaključeno da su i najčešći uzrok zrakoplovnih nesreća. Asocijativnim pravilima otkriveni su uzroci koji su doveli do nesreće. Najčešći uzroci zrakoplovnih nesreća su loši vremenski uvjeti i slaba vidljivost zbog koje je pilot izgubio kontrolu nad vozilom, mehanički kvarovi na zrakoplovu kao što su neispravnost motora i opreme za slijetanje, nemogućnost održavanja dovoljne brzine i visine zrakoplove, pad snage motora, nepoštivanje procedura letenja i održavanja, umor pilota. Prema statističkim podacima PlaneCrashInfo.com baze podataka čak

58% nesreća uzrokovano je zbog greške pilota, što je sukladno rezultatima u ovom radu, gdje je uočen velik broj riječi u korelaciji s pojmom „pilot“. Također, prema statistikama drugi najčešći razlog bio je mehanički kvar motora ili sustava na zrakoplovu uzrokovan pogreškama u održavanju, odnosno to je bio uzrok 17% zrakoplovnih nesreća.

Iz svega navedenog, vidljivo je da je tekstualna analiza rezultirala otkrivanjem najkorisnijih znanja na prikazanom setu podataka. Iako se analiza teksta najviše koristi u uslužnim djelatnostima, u budućnosti bi se trebalo raditi na tome da se poveća primjena na područje proizvodnih djelatnosti i proces održavanja.

8. LITERATURA

[1] SearchCIO, http://searchcio.techtarget.com/opinion/The-history-of-business-intelligence-and-analytics-and-what-comes-next, 20.1.2016.

[2] Sveznadar, http://razno.sveznadar.info/10-doc-PDF/01-04PodatakInformacijaDIKW.pdf, 29.1.2016.

[3] gorila.hr, http://gorila.hr/profile/digitalac/2011/06/03/koja-je-razlika-izmeu-podatka-i-informacije-to-je-podatak-a-to-informacija, 29.1.2016.

[4] Uvod u upravljanje znanjem, https://www.fer.unizg.hr/_download/repository/UVOD_U_UPRAVLJANJE_ZNANJEM1.pdf, 29.1.2016.

[5] Poslovna inteligencija i upravljanje opskrbnim lancem, http://www.skladistenje.com/wp-content/uploads/2013/07/Luetic_disertacija_BI_SCM.pdf, 15.1.2016.

[6] Hrvatska znanstvena BIBLIOGRAFIJA, https://bib.irb.hr/datoteka/481181.PISSHP_-_Glavnina_teksta.pdf, 1.2.2016.

[7] Wikipedija, https://hr.wikipedia.org/wiki/Poslovna_inteligencija, 1.2.2016.

[8] Hrvatska znanstvena BIBLIOGRAFIJA, https://bib.irb.hr/datoteka/166465.Poslovna_inteligencija_HrOUG_2004.pdf, 1.2.2016.

[9] EFOS, http://www.efos.unios.hr/arhiva/dokumenti/PRISTUPNI%20rad%20-%20PRIMJER.pdf, 3.2.2016.

[10] Aggarwal, Charu C. (2015). *Data Mining*. New York: Springer.

[11] Witten, Frank, Hall (2011). *Data Mining*. USA: Elsevier

[12] INFOTEH-JAHORINA, http://infoteh.etf.unssa.rs.ba/zbornik/2015/radovi/RSS-3/RSS-3-6.pdf, 23.3.2016.

[13] Gartner, http://www.gartner.com/newsroom/id/2848718, 22.4.2016.

[14] MathWork, https://www.mathworks.com/, 6.10.2016.

[15] Reinforcement learning, http://reinforcementlearning.ai-depot.com/, 7.10.2016.

[16] DeGiorgi, http://degiorgi.math.hr/~singer/ui/ui_1415/ch_18a.pdf, 7.10.2016.

[17] TheNewStack, http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/, 13.10.2016.

[18] RapidMiner, http://rapidminer.com/wp-content/uploads/2016/06/rapidminer-logo-retina.png, 13.10.2016.

[19] KDNuggets, http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html, 13.10.2016.

[20] OpenData, https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq, 19.10.2016.

[21] Skybrary, http://www.skybrary.aero/index.php/Accident, 22.10.2016.

[22] Osnove statistike, https://www.pmf.unizg.hr/_download/repository/PREDAVANJE7.pdf, 22.10.2016.

[23] Croinfo, http://croinfo.net/vijesti-hrvatska/7812-zrakoplov-legenda-douglas-dc-3-propada-u-otocu.html, 24.10.2016.

[24] Computer Hope, http://www.computerhope.com/jargon/e/excel.htm, 29.10.2016.

[25] Osnove statistike u društvenim i obrazovnim znanostima, http://marul.ffst.hr/~abubic/nastava/statistika/statistika_prirucnik_ucitelji.pdf, 25.11.2016.

[26] Statistical Summary of Commercial Jet Airplane Accidents, http://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf, 27.11.2016.

[27] PlaneCrashInfo.com, http://www.planecrashinfo.com/cause.htm, 27.11.2016.

[28] Skybrary, http://www.skybrary.aero/index.php/The_Human_Factors_%22Dirty_Dozen%22, 27.11.2016.

PRILOZI

1. Mjerenje performansi operatora Evolutionary Weighting
2. Klasteri k-means metode
3. Klasteri FCM metode
4. Asocijativna pravila
5. CD-R disc

PRILOG 1: *Mjerenje performansi operatora Evolutionary Weighting*

Generacija	Najbolje performanse	Performanse
0.0	?	?
0.0	0.528	0.528
0.0	0.531	0.531
0.0	0.531	0.531
0.0	0.531	0.531
1.0	0.548	0.548
1.0	0.548	0.537
1.0	0.548	0.537
1.0	0.557	0.557
1.0	0.561	0.561
1.0	0.561	0.561
1.0	0.561	0.561
2.0	0.561	0.561
2.0	0.561	0.550
2.0	0.561	0.550
2.0	0.561	0.560
2.0	0.561	0.560
2.0	0.561	0.560
2.0	0.561	0.560
2.0	0.561	0.560
2.0	0.561	0.560
3.0	0.561	0.560
3.0	0.561	0.551
3.0	0.561	0.551
3.0	0.561	0.551
3.0	0.561	0.553
3.0	0.561	0.553
3.0	0.572	0.572
4.0	0.572	0.572
4.0	0.572	0.546
4.0	0.572	0.548
4.0	0.572	0.548
4.0	0.572	0.548
5.0	0.572	0.569
5.0	0.572	0.549
5.0	0.572	0.549
11.0	0.576	0.562
11.0	0.576	0.562
11.0	0.576	0.562
11.0	0.576	0.562
12.0	0.576	0.562
12.0	0.576	0.532
12.0	0.576	0.535
12.0	0.576	0.565
12.0	0.576	0.565
12.0	0.576	0.565
12.0	0.576	0.565
12.0	0.576	0.565
12.0	0.576	0.565
13.0	0.576	0.565
13.0	0.576	0.538
13.0	0.576	0.570
13.0	0.576	0.570
13.0	0.576	0.570
13.0	0.576	0.570
13.0	0.576	0.570
13.0	0.576	0.570
13.0	0.576	0.570
14.0	0.576	0.570
14.0	0.576	0.534
14.0	0.576	0.541
14.0	0.576	0.541
14.0	0.576	0.541
14.0	0.576	0.541
14.0	0.576	0.557
14.0	0.576	0.557
14.0	0.576	0.557
15.0	0.576	0.557
15.0	0.576	0.526
15.0	0.576	0.540
15.0	0.576	0.550
15.0	0.576	0.557

5.0	0.572	0.549	16.0	0.576	0.557
5.0	0.572	0.564	16.0	0.576	0.547
5.0	0.572	0.564	16.0	0.576	0.547
5.0	0.572	0.564	16.0	0.576	0.547
5.0	0.572	0.564	16.0	0.576	0.566
6.0	0.572	0.564	17.0	0.576	0.573
6.0	0.572	0.538	17.0	0.576	0.527
6.0	0.572	0.545	17.0	0.576	0.553
6.0	0.572	0.545	17.0	0.576	0.553
6.0	0.572	0.545	17.0	0.576	0.561
6.0	0.572	0.545	17.0	0.576	0.561
6.0	0.572	0.545	18.0	0.576	0.561
6.0	0.572	0.545	18.0	0.576	0.549
6.0	0.572	0.545	18.0	0.576	0.549
7.0	0.572	0.545	18.0	0.576	0.549
7.0	0.572	0.564	18.0	0.576	0.554
7.0	0.572	0.564	18.0	0.576	0.554
7.0	0.576	0.576	18.0	0.576	0.569
7.0	0.576	0.576	18.0	0.576	0.569
7.0	0.576	0.576	18.0	0.576	0.569
7.0	0.576	0.576	19.0	0.576	0.574
8.0	0.576	0.576	19.0	0.576	0.526
8.0	0.576	0.527	19.0	0.576	0.557
8.0	0.576	0.527	19.0	0.576	0.557
8.0	0.576	0.542	19.0	0.576	0.557
8.0	0.576	0.544	19.0	0.576	0.557
8.0	0.576	0.553	19.0	0.576	0.557
8.0	0.576	0.553	20.0	0.576	0.557
9.0	0.576	0.553	20.0	0.576	0.534
9.0	0.576	0.541	20.0	0.576	0.564
9.0	0.576	0.552	20.0	0.576	0.564
9.0	0.576	0.552	20.0	0.576	0.564
9.0	0.576	0.552	20.0	0.576	0.564
9.0	0.576	0.552	20.0	0.576	0.564
9.0	0.576	0.552	20.0	0.576	0.564
10.0	0.576	0.552			
10.0	0.576	0.548			
10.0	0.576	0.549			
10.0	0.576	0.549			

10.0	0.576	0.549
10.0	0.576	0.549
10.0	0.576	0.549
10.0	0.576	0.554
10.0	0.576	0.554
11.0	0.576	0.554
11.0	0.576	0.545
11.0	0.576	0.547
11.0	0.576	0.562
11.0	0.576	0.562

PRILOG 2: *Klasteri k-means metode*

KLASTER 0		
Douglas		
KLASTER 1		
AAC-1	Black	British Aerospace
AEGK	Blackburn	Nimrod MR-2
Aermacchi	Bleriot	British Aerospace
Aero Commander	Bloch	Nimrod MR-2P
Aerospatiale	Boeing-Vertol	Britten
Aerospeciale	Bombardier	Britten-Norman
Aerostar	Boulton	Brittonnorman
Agusta	Brantly	Brittonorman
Airbus	Breguet	Burgess
Airship	Bristol	C-47
Airspeed Ambassador	British Aerospace BAe	Cams
Arado	Jetstream 32EP	Canadair
Arava	British Aerospace 146-300	Caravelle
Armstrong	British Aerospace 3101	CASA
Armstrong-Whitworth	Jetstream 31	Caudron
AT	British Aerospace 748 2A	Cesna
ATR	British Aerospace APT	Cessnea
ATR-42-300	British Aerospace ATP	CF-100
ATR-72-202	British Aerospace BAe-125-700A	CH-47
ATR-72-212	British Aerospace BAe-125-800A	CH-47D
Avia	British Aerospace BAe-146-100	CH-53D
Aviation	British Aerospace BAe-146-200A	CH53E
Avro	British Aerospace BAe-146-300	Chance
B-17C	British Aerospace Jetstream 3201	Channel
B17G	British Aerospace Jetstream 4101	CMASA
BAC	British Aerospace Jetstream BA-3100	Consolidated
BAe		Convair
BAe-748		ConvairCV-440
Bambardier		Curtis
Bandeirante		Curtiss-Wright
Beech		Dassault
BeechJet		Dassault-Breguet
Bell		DC-2-243
Bellanca		DC-3-65TP
Bernard		De Havilland
BFW		deHavilland

Desoutter
Dewoitine
DHC-5
DHC-6
Dirigible
Dornier
Domier
Dormier
EC-121H
EMB
Embraer
Embraer-110
Enstrom
Eurocopter
Evangel
F-4C
F-86
F-88
Fairchild
Fairchild-Hiller
Fairey
Farman
Faucett
Fiat
Five
Focke-Wulf
Ford
GAF
Gates
GD
General
Goodyear
Goodyear-Zeppelin
Grumman
Grummand
Gulfstream
H-21B
Hadley
HAL-748-224
Hamilton
Handley
Harbin
Hawker
Hawker-Siddeley
HBB
Heinkel
Helicopter,
Helo
Hiller
Hindustan
Howard
HS-125-700B
Hughes
IAI
Illyushin
Ilushin
Iluyshin
Ilysushin
Ilyushin
IPTN
Israel
Junkers
Kaiser-Fraser
Kalinin
Kawasaki
KB-50
KJ-2000
Koolhoven
L-100-20
Lasco
LatecoĂ¨re
Latecoere
Lear
Learjet
Learjet35A
Let
Let-410UVP-E
Liore
Liore-et-Olivier
Lisunov
Loening
LTVF-8J
Macchi
Martin
MBB
McDonnel
McDonnell-Douglas
MD
MD-87
MDonnell
Messerschmitt
MH-47
Mi
Mi-17
Mi-35
Mi-8
Mi-8MTV-1
MiG-15
MiG-23
Mil
Military
Mitsubish
Mitsubishi
NA
Nakajima
NAMC
NAMC-YS-11-111
Nord
Norman
North
Northrop
OFM
PA-
PA-23-250
PA-34-220T
Partenavia
PBY4-2
PBY-5A
Pel
Piaggio
Pilatus
Pilatus-Britten
Pilgrim
Pipper
Pitcairn
Pitcairns
Potez
PS-1
PT-LCN
PZL-MieleAN-2R
PZL-MieleM28
Robertson
Rochrbach
Rockwell
Rohrbach
Royal
Rutan
Ryan
S2F-1

Saab	Soloy	Waco,
Saab340B	SPCA	Westland
Sabca	Stearman	Wibault
Salmson	Stinson	Wright Flyer
Saro	Sud	Xian
Savbia-Marchetti	Sud-Aviation	Yakovlev
Savoia	Sukhoi	Yunshuji
Savoia-Marchetti	Super	Zeppelin
Schutte-Lanz	Swear.	
Sepecat	Swearingen	
Shaanxi	Transall	
Short	Transportes	
Shorts	Travel	
Shrike	Tupelov	
Siebel	Tupolev	
Sikorksky	UC-64A	
Sikorsky	UH-60	
Silver	V6	
Sinson	VEB	
Sirkorsky	Vickers	
SNCASE	Volpar	
SNIAS	Vultee	

KLASTER 2

Antonov	Curtiss	Lockheed
Beechcraft	De Havilland	McDonnell
Boeing	Fokker	Piper
Cessna		

PRILOG 3: *Klasteri FCM metode*

KLASTER 0		
Beechcraft	Embraer	McDonnell
Bell	Fairchild	Piper
Britten-Norman	Fokker	Tupolev
Convair	Ilyushin	Vickers
Curtiss	Junkers	
KLASTER 1		
Antonov	Cessna	Douglas
Boeing	De Havilland	Lockheed
KLASTER 2		
AAC-1	BFW	British Aerospace
AEGK	Black	Jetstream BA-3100
Aermacchi	Blackburn	British Aerospace
Aero Commander	Bleriot	Nimrod MR-2
Aerospatiale	Bloch	British Aerospace
Aerospeciale	Boeing-Vertol	Nimrod MR-2P
Aerostar	Bombardier	Britten
Agusta	Boulton	Brittonnorman
Airbus	Brantly	Brittonorman
Airship	Breguet	Burgess
Airspeed Ambassador	Bristol	C-47
Arado	British Aerospace BAe	Cams
Arava	Jetstream 32EP	Canadair
Armstrong	British Aerospace 146-300	Caravelle
Armstrong-Whitworth	British Aerospace 3101	CASA
AT	Jetstream 31	Caudron
ATR	British Aerospace 748 2A	Cesna
ATR-42-300	British Aerospace APT	Cessnea
ATR-72-202	British Aerospace ATP	CF-100
ATR-72-212	British Aerospace BAe-125-700A	CH-47
Avia	British Aerospace BAe-125-800A	CH-47D
Aviation	British Aerospace BAe-146-100	CH-53D
Avro	British Aerospace BAe-146-200A	CH53E
B-17C	British Aerospace BAe-146-300	Chance
B17G	British Aerospace Jetstream 3201	Channel
BAC	British Aerospace Jetstream 4101	CMASA
BAe	Hiller	Consolidated
BAe-748		ConvairCV-440
Bambardier		Curtis
Bandeirante		Curtiss-Wright
Beech		Dassault
BeechJet		Dassault-Breguet
Bellanca		DC-2-243
Bernard		DC-3-65TP

De Havilland
deHavilland
Desoutter
Dewoitine
DHC-5
DHC-6
Dirigible
Domier
Dormier
Dornier
EC-121H
EMB
Embraer-110
Enstrom
Eurocopter
Evangel
F-4C
F-86
F-88
Fairchild-Hiller
Fairey
Farman
Faucett
Fiat
Five
Focke-Wulf
Ford
GAF
Gates
GD
General
Goodyear
Goodyear-Zeppelin
Grumman
Grummand
Gulfstream
H-21B
Hadley
HAL-748-224
Hamilton
Handley
Harbin
Hawker
Hawker-Siddeley
HBB
Heinkel
Helicopter,
Helo
Hindustan
Howard
HS-125-700B
Hughes
IAI
Illyushin
Ilushin
Iluyshin
Ilysushin
IPTN
Israel
Kaiser-Fraser
Kalinin
Kawasaki
KB-50
KJ-2000
Koolhoven
L-100-20
Lasco
LatecoÄ¨re
Latecoere
Lear
Learjet
Learjet35A
Let
Let-410UVP-E
Liore
Liore-et-Olivier
Lisunov
Loening
LTVF-8J
Macchi
Martin
MBB
McDonnel
McDonnell-Douglas
MD
MD-87
MDonnell
Messerschmitt
MH-47
Mi
Mi-17
Mi-35
Mi-8
Mi-8MTV-1
MiG-15
MiG-23
Mil
Military
Mitsubish
Mitsubishi
NA
Nakajima
NAMC
NAMC-YS-11-111
Nord
Norman
North
Northrop
OFM
PA-
PA-23-250
PA-34-220T
Partenavia
PBY4-2
PBY-5A
Pel
Piaggio
Pilatus
Pilatus-Britten
Pilgrim
Pipper
Pitcairn
Pitcairns
Potez
PS-1
PT-LCN
PZL-MieleAN-2R
PZL-MieleM28
Robertson
Rochrbach
Rockwell
Rohrbach
Royal
Rutan
Ryan
S2F-1
Saab
Saab340B
Sabca
Salmson
Saro
Savbia-Marchetti
Savoia
Savoia-Marchetti

Schutte-Lanz
Sepecat
Shaanxi
Short
Shorts
Shrike
Siebel
Sikorksky
Sikorsky
Silver
Sinson
Sirkorsky
SNCASE
SNIAS
Soloy
SPCA
Stearman
Stinson
Sud
Sud-Aviation
Sukhoi
Super
Swear.
Swearingen
Transall
Transportes
Travel
Tupelov
UC-64A
UH-60
V6
VEB
Volpar
Vultee
Waco,
Westland
Wibault
Wright Flyer
Xian
Yakovlev
Yunshuji
Zeppelin

PRILOG 4: *Asocijativna pravila*

Riječ1	Riječ2	Pouzdanost
descent	altitude	0.302
weather	pilot	0.302
procedures	improper	0.302
trees	struck	0.303
takeoff	engine	0.305
weather	adverse	0.305
control	pilot	0.306
loss	engine	0.310
procedures	follow	0.315
lost	engine	0.320
attempted	engine	0.325
poor	conditions	0.325
error	crew	0.328
descended	approach	0.329
heavy	fog	0.335
turn	pilot	0.337
take	attempting	0.337
found	route	0.337
low	altitude	0.340
instrument	conditions	0.341
procedures	pilot	0.342
captain	approach	0.342
ifr	approach	0.346
control	loss	0.349
conditions	adverse	0.350
low	pilot	0.354
terrain	pilot	0.356
adverse	pilot	0.359
caught	engine	0.363
make	pilot	0.365
fire	engine	0.366
ifr	improper	0.373
attempted	landing	0.373
descended	pilot	0.376
improper	approach	0.377
procedures	approach	0.381
conditions	vfr	0.382

Riječ 1	Riječ 2	Pouzdanost
minutes	taking	0.503
adverse	continued	0.508
maintain	pilot	0.510
reported	pilot	0.516
disappeared	route	0.52
make	emergency	0.523
weather	conditions	0.531
error	pilot	0.534
found	wreckage	0.537
reasons	unknown	0.543
instrument	pilot	0.544
loss	control	0.553
clearance	pilot	0.565
minimum	pilot	0.566
maintain	failure	0.572
procedure	approach	0.577
contributing	pilot	0.581
visibility	poor	0.581
return	airport	0.583
follow	procedures	0.585
instrument	approach	0.594
continued	adverse	0.610
attempt	landing	0.612
return	engine	0.616
shortly	taking	0.621
flames	burst	0.621
vfr	weather	0.625
visual	approach	0.631
continued	weather	0.631
adverse	vfr	0.640
continued	conditions	0.642
emergency	landing	0.65
forced	landing	0.659
poor	weather	0.669
losing	engine	0.684
collision	midair	0.685
vfr	conditions	0.697

Riječ1	Riječ2	Pouzdanost
descent	approach	0.385
conditions	pilot	0.386
engine	failure	0.386
failed	engine	0.389
continued	pilot	0.389
fatigue	failure	0.396
improper	pilot	0.397
course	mountain	0.406
vfr	pilot	0.407
taking	shortly	0.408
attempted	pilot	0.409
undetermined	reasons	0.412
ifr	conditions	0.413
weather	poor	0.414
maintain	altitude	0.416
severe	turbulence	0.416
made	landing	0.423
heavy	rain	0.423
unknown	reasons	0.424
ifr	pilot	0.426
turbulence	severe	0.428
accident	pilot	0.43
procedures	failure	0.434
rain	heavy	0.439
experiencing	failure	0.442
vfr	continued	0.453
reasons	undetermined	0.456
power	engine	0.460
emergency	engine	0.462
ifr	vfr	0.466
making	approach	0.467
minimum	approach	0.471
vfr	adverse	0.480
failure	engine	0.488
decision	pilot	0.5
clearance	altitude	0.5
clearance	terrain	0.5

Riječ1	Riječ2	Pouzdanost
factors	pilot	0.702
conditions	weather	0.703
continued	vfr	0.726
minimum	altitude	0.735
mountainous	terrain	0.736
make	landing	0.746
experiencing	engine	0.769
missile	shot	0.774
missed	approach	0.790
end	runway	0.790
wreckage	found	0.811
midair	collision	0.818
rebels	shot	0.843
adverse	conditions	0.850
gear	landing	0.859
trouble	engine	0.866
command	pilot	0.886
caught	fire	0.893
wooded	area	0.897
final	approach	0.923
structural	failure	0.928
ils	approach	0.930
gain	altitude	0.939
international	airport	0.962
lines	power	0.965
overran	runway	0.973
burst	flames	0.978