# Detekcija prijevara u osiguranju motornih vozila

**Hajdarović, Dijana**

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

**Dijana Hajdarović**

# DETEKCIJA PRIJEVARA U OSIGURANJU MOTORNIH VOZILA

SPECIJALISTIČKI RAD

Zagreb, 2023.

UNIVERSITY OF ZAGREB
**FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING**
SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

**Dijana Hajdarović**

# MOTOR INSURANCE FRAUD DETECTION
# DETEKCIJA PRIJEVARA U OSIGURANJU MOTORNIH VOZILA

SPECIALIST THESIS
SPECIJALISTIČKI RAD

Zagreb, 2023.

*Specijalistički rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva u sklopu poslijediplomskog specijalističkog studija Proizvodi, digitalne inovacije i tehnologije u osiguranju – INSURTECH.*

*Mentor: Prof. dr. sc. Zvonko Kostanjčar*

*Specijalistički rad ima: 65 stranica*
*Specijalistički rad br.:_____*

Povjerenstvo za ocjenu u sastavu:

1. doc. dr. sc. Stjepan Begušić – predsjednik

2. prof. dr. sc. Zvonko Kostanjčar – mentor

3. doc. dr. sc. Azra Tafro, Sveučilište u Zagrebu Fakultet šumarstva i drvne tehnologije – članica

Povjerenstvo za obranu u sastavu:

1. doc. dr. sc. Stjepan Begušić – predsjednik

2. prof. dr. sc. Zvonko Kostanjčar – mentor

3. doc. dr. sc. Azra Tafro, Sveučilište u Zagrebu Fakultet šumarstva i drvne tehnologije – članica

Datum obrane: 28. rujna 2023.

# SAŽETAK

U radu se analiziraju metode za detekciju prijevarnih šteta na podacima o nastalim štetama na motornim vozilima u periodu od dvije godine u jednom osiguravajućem društvu u Republici Hrvatskoj.

Cilj rada je ispitati dvije metode za nadzirano i jedne za nenadzirano učenje te na temelju dobivenih rezultata preložiti najbolju metodu i korake za poboljšanje dobivenih rezultata.

Ispitivanje metoda napravljeno je na podacima koji se prikupljaju u procesu prijave i obrade odštetnih zahtjeva. Prije ispitivanja je napravljena analiza dostupnih varijabli i priprema za korištenje u prediktivnom modelu. Ispitivani su modeli logističke regresije i slučajnih šuma na 21 varijabli koja opisuje štete. Dodatno je implementirana *k-means* metoda kako bi se korištenjem metode lakta i metode siluete ispitalo koji bi bio optimalan broj klastera na korištenom uzorku podataka.

**Ključne riječi:**

metode za detekciju prijevarnih šteta, štete, nadzirano i nenadzirano učenje, logistička regresija, slučajne šume, k-means, metoda lakta, metoda siluete

# ABSTRACT

This paper analyzes methods for detecting fraudulent claims on data for motor claims in a period of two years in one insurance company in the Republic of Croatia.

The aim of this paper is to test two methods for supervised and two methods for unsupervised learning and based on the obtained results to detect the best method and propose steps to improve it.

Data used in this paper were collected during the first notice of loss and claim handling process. Before implementing predictive methods, data were analyzed and prepared for use. In this paper were tested logistic regression and random forest utilizing 21 variables as inputs for the models. Additionally, the *k-means* method was used to find an optimal number of clusters using the elbow and silhouette method.



**Key words:**

methods for detecting fraudulent claims, claims, supervised and unsupervised learning, logistic regression, random forest, k-means, elbow method, silhouette method

# Sadržaj

# 1 UVOD

Prijevara u osiguranju je namjerna prijevara počinjena protiv ili od strane osiguravajućeg društva ili agenta u svrhu financijske dobiti. Počinitelj prijevare može biti bilo koji sudionik u procesu nastanaka, prijave i obrade štete – ugovaratelj police[1], osiguranik[2], korisnik osiguranja[3], oštećenik[4], djelatnik zadužen za obradu zahtjeva, procjenitelj[5].

Najčešće prijevare su lažno prikazivanje činjenica u prijavi štete, preuveličavanje veličine i visine štete, prijava štete koja se nije dogodila i insceniranje nesreće, a

najčešći počinitelji prijevara u osiguranju su:

- organizirane kriminalne skupine koje kradu velike svote novaca kroz prijevarne aktivnosti u poslovnim odnosima,
- stručni djelatnici koji povećavaju troškove usluga ili naplaćuju neizvršene usluge i
- obični ljudi koji žele pokriti iznos uplaćene premije ili prijavu štete vide kao mogućnost zarade.

Desetljećima su procjene godišnjih troškova prijevare u osiguranju bile preniske te su nedostajale ažurirane informacije o inflaciji, međutim istraživanje koje je proveo *The Coalition Against Insurance Fraud* (CAIF) u 2022. godini pokazalo je da prijevare šteta osiguranike u SAD-u koštaju 308.6 milijardi dolara godišnje [1].

[1] Ugovaratelj osiguranja pravna je ili fizička osoba koja s društvom za osiguranje sklapa ugovor o osiguranju. Plaćanjem premije osiguranja može imati pravo na osigurninu ili može odrediti drugu osobu koja time stječe pravo na osigurninu. [1]

[2] Osiguranik je ona strana ugovora o osiguranju čija su imovina, život ili zdravlje osigurani, a koja ima pravo i ovlaštena je potraživati osigurninu od društva za osiguranje u slučaju nastanka osiguranog slučaja. Uglavnom se uloga ugovaratelja osiguranja i osiguranika poklapa, ali je moguće i osiguranje za tuđi račun, kao što je to kod osiguranja osoba (npr. moguće je da poslodavac kao ugovaratelj osiguranja sklopi ugovor o zdravstvenom osiguranju za svoje djelatnike – osiguranike). No kada se osiguranje života ne odnosi na život ugovaratelja osiguranja, nego na život neke druge osobe (osiguranika), za određivanje korisnika, odnosno osobe kojoj će se isplatiti osigurnina, potrebna je pisana suglasnost i samog osiguranika. Kod osiguranja imovine, ugovaratelj osiguranja najčešće je ujedno i osiguranik, no moguće je i osiguranje kod kojeg je osiguranik i jedna ili više trećih osoba. [1]

[3] Korisnik osiguranja pravna je ili fizička osoba koja ima pravo na naknadu temeljem ugovora o osiguranju ako nastupi osigurani slučaj. Primjerice, kod osiguranja za slučaj smrti i doživljenja, u slučaju smrti osiguranika korisniku osiguranja (obično je to član obitelji) isplaćuje se osigurnina za slučaj smrti. U slučaju doživljenja osigurnina se isplaćuje osiguraniku ili korisniku kojeg odredi ugovaratelj osiguranja, dakle to može biti i on sam. [1]

[4] Oštećenik je pravna je ili fizička osoba koja podnosi zahtjev za naknadu štete za koju je odgovoran osiguranik koji je sklopio osiguranje od odgovornosti prema trećim osobama. [3]

[5] Procjenitelj je predstavnik osiguratelja koji nastoji utvrditi opseg odgovornosti osiguratelja za prijavljeni zahtjev za nastalu štetu. [4]

Procjenjuje se da u Republici Hrvatskoj godišnja šteta od prijevara u osiguranju se kreće od 250 do 300 milijuna kuna [5], čije su izravne posljedice više premije[6] za sve osiguranike. Osiguravajuće kuće pokušavaju detektirati i dokazati potencijalne prijevare stručnim znanjem (npr. oštećenja na vozilu ne odgovaraju uvjetima u kojima se navodno dogodila nezgoda) te korištenjem različitih *software*-a koji detektiraju nepravilnosti u podacima.

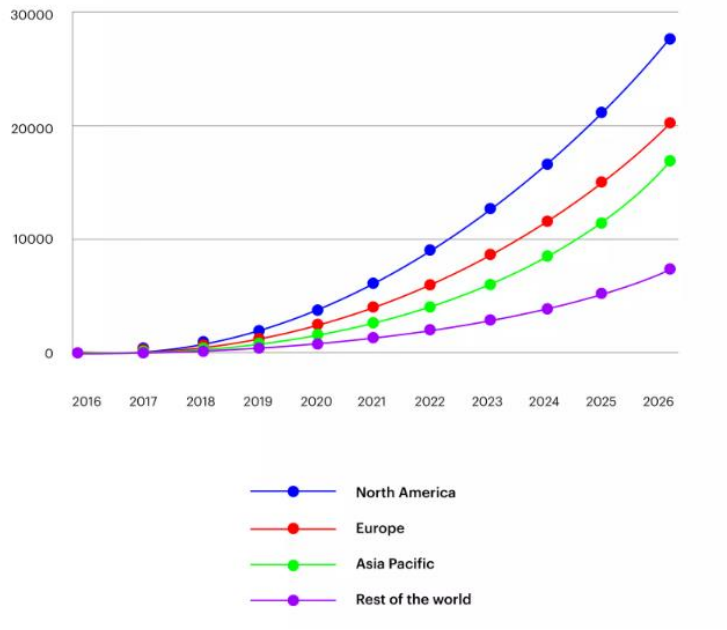
Kroz ovaj rad će se napraviti analiza dostupnih podataka o štetama iz jednog osiguravajućeg društva u Hrvatskoj te će se na njima ispitati nekoliko metoda za detekciju potencijalnih prijevara.

Prvo će se napraviti analiza radova u kojima su već ispitane različite metode za detekciju prijevara, te će objasniti nekoliko metoda za detekciju prijevarnih šteta koje će se kasnije i ispitati na dostupnim podacima o štetama. Zatim slijedi centralni dio ovog rada u kojem će se opisati podaci i njihov način pripreme za korištenje u ispitivanju metoda te na kraju rezultati i zaključak kao i preporuke za kalibraciju ispitanih metoda.

[6] Premija je cijena osiguranja odnosno novčani iznos koji je ugovaratelj osiguranja obvezan platiti osiguratelju temeljem zaključenog ugovora o osiguranju. [6]

# 2 PREGLED POVEZANIH RADOVA

R. Devi Burri, R. Burii, R. Reddy Bojja i S. Rao Buruga u radu „***Insurance Claim Analysis Using Machine Learning Algorithms***“ [7] spominju tri načina na koja se strojno učenje može primjenjivati u industriji osiguranja, a to su: automatizirane i personalizirane ponude, bolja procjena rizika i bolja detekcija prijevara. Na *Slika 2.1* prikazana su potencijalna tržišta za korištenje strojnog učenja u industriji osiguranja za period 2017.- 2026. (njihova veličina iskazna je u američkim dolarima).




***Slika 2.1*** *Primjena strojnog učenja u industriji osiguranja*
Izvor: R. Devi Burri, R. Burii, R. Reddy Bojja i S. Rao Buruga u radu „Insurance Claim Analysis Using Machne Learning Algorithms“ [7]

G. Kowshalya i Dr. M. Nandhini u radu „***Predicting Fraudulent Claims in Automobile Insurance***“ [8] ispitivali su klasifikacijske algoritme na podacima vezanim za štete nastale na motornim vozilima. Razlikuju dva tipa šteta – krađu vozila i prometne nezgode. U izradi modela korištena je 31 varijabla, pri čemu je 20 varijabli bilo zajedničko svim štetama, 5 je primjenjivo samo za štete u kojima je ukradeno vozilo i 6 je primjenjivo samo za ostale štete. Popis svih varijabli koji su korišteni u modelu nalazi se u *Tablica 2.1.*

***Tablica 2.1*** *Varijable korištene u radu 'Predicting Fraudulent Claims in Automobile Insurance'*

| Rbr | Varijabla | Moguće vrijednosti | Za koji uzrok nastanka štete je primjenjiv atribut |
|---|---|---|---|
| 1 | Oštećenik učestalo mijenja adresu/broj telefona | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 2 | Oštećenik mijenja učestalo posao | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 3 | Teško je stupiti u kontakt s oštećenikom | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 4 | Kontakt mjesto oštećenika je hotel | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 5 | Oštećenik je nezaposlen | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 6 | Oštećeniku je nedavno oduzeta vozačka dozvola | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 7 | Oštećenik je agresivan te radi pritisak za brzu nagodbu | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 8 | Oštećenik je predobro upoznat sa procesom brade šteta i osigurateljnom terminologijom | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 9 | Oštećenik plaća kredit za vozilo ili ima financijske probleme | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 10 | Oštećenik izbjegava susret sa istražiteljima | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 11 | Oštećenik je nedavno kupio novo pokriće na polici | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 12 | Između datuma nastanka štete i datuma prijave štete je više od 7 dana | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 13 | Između početka premije na polici i datuma štete je manje od 5 dana | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 14 | Oštećenik je vlasnik vozila tijekom svih osigurateljnih razdoblja | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 15 | Iznos premije na polici je više od 5000 (?) i kreditni rejting je slabi ili loš | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 16 | Oštećenik je priložio uz štetu račun pisan rukom | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 17 | Uz prijavu štete nije priložena dokumentacija | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 18 | Slaba komunikacija | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 19 | Ne postoji policijski zapisnik | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 20 | Uzrok nastanka štete je krađa ili nesreća (1 - krađa, 0- prometna nezgoda) | Da (1)<br>Ne (0) | Krađa i Prometna nezgoda |
| 21 | Je li vozilo skupo u trenutku nezgode | Da (1)<br>Ne (0) | Krađa |

| | | | |
|---|---|---|---|
| 22 | Prije krađe, u vozilu se nalazio natpis da je vozilo na prodaju | Da (1)<br>Ne (0) | Krađa |
| 23 | Vozilo je parkirano na ulici iako je dostupna garaža | Da (1)<br>Ne (0) | Krađa |
| 24 | Vozilo je ukradeno unutar mjesec dana od izdavanja police | Da (1)<br>Ne (0) | Krađa |
| 25 | Susjedi/prijatelji/obitelj ne znaju da je vozilo ukradeno | Da (1)<br>Ne (0) | Krađa |
| 26 | Ne postoje svjedoci za tipove šteta oštećenje ili ozljeda | Da (1)<br>Ne (0) | Prometna nezgoda |
| 27 | Vrijeme nastanka štete je između 23:00h i 3:00 i tip štete je oštećenje | Da (1)<br>Ne (0) | Prometna nezgoda |
| 28 | Velika je udaljenost između bolnice i mjesta nezgode | Da (1)<br>Ne (0) | Prometna nezgoda |
| 29 | Broj fotografija oštećenja na vozilu je manji od 5 | Da (1)<br>Ne (0) | Prometna nezgoda |
| 30 | Započeo je popravak na oštećenom vozilu | Da (1)<br>Ne (0) | Prometna nezgoda |
| 31 | Oštećenik je već prevario osiguranje | Da (1)<br>Ne (0) | Prometna nezgoda |

Izvor: Prilagodba autorice prema tablici iz članka Kowshalya G., Dr. Nandhini M., „Predicting Fraudulent Claims in Automobile Insurance“ [8]

Autori su ispitivali model slučajnih šuma[7], *Naive Bayes* i *J48* te su pri tome koristili tri različita omjera podjele uzorka podataka na set za treniranje i set za testiranje (50:50, 66:34, 10CV[8]). Mjere uspješnosti korištenih modela (točnost, preciznost i odziv) prikazane su u *Tablica 2.2*.

U procesu treniranja modela koristili su dva različita scenarija:

1) U prvom scenariju kod izrade prediktivnog modela korištene su sve varijable navedene u *Tablica 2.1*. U ovom scenariju najveću točnost[9] imao je model slučajnih šuma 99,41% u slučaju kad je uzorak podataka za treniranje i testiranje podijeljen u omjeru 66:34. Model slučajnih šuma je imao najveću točnost i u slučaju podjele uzorka podataka na set za učenje i testiranje u omjeru 50:50, dok je u omjeru podjele podataka *10CV* najveću točnost 99,39% imao *J48*.

[7] engl. *random forest*

[8] *10CV* ili *10-Fold Cross Validation* je metoda u kojoj se uzorak podataka na slučajan način podijeli u 10 grupa. Devet grupa koristi se za učenje, dok se jedna zatim koristi za testiranje i taj se postupak ponavlja 10 puta (svaka grupa koristi se kao set za testiranje). Ukupna greška je prosjek grešaka koje su se pojavile kod svakog treniranja, odnosno testiranja podataka. [9]

[9] engl. *accuracy*

***Tablica 2.2*** *Rezultati prediktivnog modela za osiguranjeve podatke o štetama bez predprocesiranja*

| Metoda | 50:50 | | | 66:34 | | | 10CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Točnost (%) | Preciznost | Odaziv | Točnost (%) | Preciznost | Odaziv | Točnost (%) | Preciznost | Odaziv |
| J48 | 97.59 | 0.977 | 0.976 | 97.05 | 0.972 | 0.971 | **99.39** | 0.994 | 0.994 |
| Slučajna šuma | 98.79 | 0.988 | 0.988 | **99.41** | 0.994 | 0.994 | 99.19 | 0.992 | 0.992 |
| Naïve Bayes | 97.59 | 0.938 | 0.968 | **98.23** | 0.983 | 0.982 | 97.79 | 0.978 | 0.978 |

Izvor: Prilagodba autorice prema tablici iz članka Kowshalya G., Dr. Nandhini M., „Predicting Fraudulent Claims in Automobile Insurance“ [8]

U drugom scenariju u implementaciji modela korištene su samo varijable koje doprinose modelu, dok su varijable „*Vrijeme nastanka štete je između 23:00h i 3:00* , „*Tip štete je oštećenje*“, „*Oštećenik učestalo mijenja adresu/broj telefon*“ i „*Oštećenik je nezaposlen*“ maknute. Mjere uspješnosti izrade modela nad ovakvim setom podataka prikazane su u *Tablica 2.3.*

2) I u ovom scenariju, model slučajne šume imao je najveću točnost, međutim za razliku od prethodnog scenarija to je postignuto kod podjele uzorka podataka u omjeru 10CV. *J48* i *Naive Bayes* imaju najveću točnost za podjelu podataka u omjeru 10CV.

   U slučaju podjele uzorka podataka na set za učenje i set za testiranje u omjeru 66:34, slučajna šuma ima istu točnost i prvom i u drugom scenariju kad je korišteno manji broj značajki, *J48* ima za 2,36% veću točnost, dok *Naive Bayes* ima za 2,97% manju točnost.

   U omjeru 50:50 podjele uzorka podataka na set za učenje i set za testiranje, također *Naive Bayes* ima manju točnost kao i u drugom scenariju, dok slučajna šuma i *J48* i u ovom slučaju imaju veću preciznost.

***Tablica 2.3*** *Rezultati prediktivnog modela za osiguranjeve podatke o štetama sa predprocesiranjem*

| Metoda | 50:50 | | | 66:34 | | | 10CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Točnost (%) | Preciznost | Odaziv | Točnost (%) | Preciznost | Odaziv | Točnost (%) | Preciznost | Odaziv |
| J48 | 98.39 | 0.985 | 0.984 | 99.41 | 0.994 | 0.994 | **99.39** | 0.994 | 0.994 |
| Slučajna šuma | 99.59 | 0.996 | 0.996 | 99.41 | 0.994 | 0.994 | **99.79** | 0.998 | 0.998 |
| Naïve Bayes | 95.98 | 0.964 | 0.960 | 95.29 | 0.957 | 0.953 | **98.19** | 0.983 | 0.982 |

*Izvor: Prilagodba autorice prema tablici iz članka Kowshalya G., Dr. Nandhini M., „Predicting Fraudulent Claims in Automobile Insurance“ [8]*

Shah S., Koli S.P.P., Sharma S. u radu „***Insurance Fraud Detection using Machine Learning***“ [10] ispitivali su četiri algoritma za detekciju prijevarnih šteta – logistička regresija, slučajna šuma, SVM (stroj potpornih vektora) i *XGBoost*. Za sve algoritme izračunali su najbolje parametre korištenjem „*GridSearch*“ funkcionalnosti, a rezultati korištenja različitih modela opisani su u *Tablica 2.4*.

***Tablica 2.4*** *Rezultati korištenja različitih modela u radu „Insurance Fraud Detecting using Machine Learning“*

| Metoda | Točnost | Preciznost | Odziv | F- mjera |
|---|---|---|---|---|
| Logistička regresija | 0,864 | 0,750 | 0,706 | 0,727 |
| Slučajna šuma | 0,879 | 0,800 | 0,706 | 0,750 |
| SVM | 0,886 | 0,842 | 0,600 | 0,750 |
| XGBoost | 0,932 | 0,903 | 0,824 | 0,862 |

Izvor: Shah S., Koli S.P.P., Sharma S: „Insurance Fraud Detecting using Machine Learning“ [10]

Na setu podataka na kojem je napravljeno testiranje (u radu nije navedeno koje značajke su korištene, ni na koji način je odabran set podataka za učenje i testiranje) *XGBoost* algoritam ima najbolje rezultate.

Arun Kumar Rai i Rajendra Kumar Dwivedi u članku „***Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme***“ [11] predlažu modele nenadziranog učenja za detekciju prijevara u transakcijama kreditnih kartica. Modele koje su predložili su neuronske mreže, *Auto Encoder*, *Local Outlier Factor*, *Isolation Forest* i *k-means* klastering.
U *Tablica 2.5* prikazana je matrica zabune za modele, a u *Tablica 2.6* su prikazani rezultati mjera korištenih za validaciju modela te se iz njih vidi da neuronske mreže imaju najbolje rezultate.

***Tablica 2.5*** *Matrica zabune za modele testirane u članku " Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme"*

| Metoda | Stvarno pozitivna (engl. *True Positive*) | Lažno pozitivna (engl. F*alse Positive*) | Lažno negativna (engl. *False Negative*) | Stvarno negativna (engl. *True Negative*) |
|---|---|---|---|---|
| Neuronske mreže | 56851 | 13 | 18 | 80 |
| Auto encoder | 55491 | 1373 | 20 | 78 |
| Isolation Forest | 55789 | 1075 | 48 | 50 |
| Local Outlier Factor | 55091 | 773 | 52 | 46 |
| K-Means | 55734 | 130 | 13 | 85 |

Izvor: Rai A.K., Dwivedi R.K., „Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme“ [11]

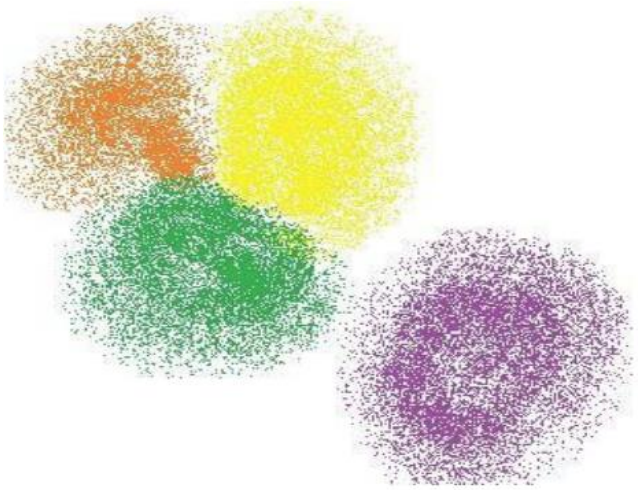
***Tablica 2.6*** *Rezultati modela iz članka " Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme“*

| Metoda | Točnost | Preciznost | Odziv | F- mjera |
|---|---|---|---|---|
| Neuronske mreže | **0,990** | **0,860** | 0,810 | 0,830 |
| Auto encoder | 0,970 | 0,053 | 0,790 | 0,100 |
| Isolation Forest | 0,980 | 0,510 | 0,044 | 0,081 |
| Local Outlier Factor | 0,980 | 0,056 | 0,460 | 0,100 |
| K-Means | **0,990** | 0,390 | 0,860 | 0,540 |

Izvor: Rai A.K., Dwivedi R.K., „Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based [11]

Vaishali, M.Tech u radu „***Fraud detection in Credit Card by Clustering Approach***“ [12] opisuju korištenje *k-means* algoritma u problemu detekcije prijevara na kreditnim karticama. U implementaciji su korišteni podaci o transakcijama te su podaci sa sličnim karakteristikama grupirani u iste klastere. Krajnji rezultat je bio grupiranje transakcija u četiri klastera (prikaz na *Slika 2.2* ).

Narančasti klaster predstavlja slabo rizične transakcije, žuti rizične transakcije, zeleni visoko rizične transakcije, a ljubičasti izrazito visoko rizične transakcije.



***Slika 2.2*** *Prikaz transakcija kreditnih kartica u klasterima*
Izvor: Vaishali, M.Tech u radu „Fraud detection in Credit Card by Clustering Approach“ [12]

Nakon implementacije algoritma, napravljena je usporedba rezultata dobivenih klasifikacijom transakcija korištenjem algoritma i podatka radi li se doista o prijevarnoj transakciji ili ne. Rezultati usporedbe su prikazani u *Tablica 2.7*.

***Tablica 2.7*** *Usporedba rezultata algoritma i stvarnih transakcija*

| Broj transakcije | Oznaka koju je algoritam dodijelio transakciji | Oznaka je dosta bila prijevarna transakcija |
|---|---|---|
| **1** | Slabo rizična (narančasto) | Ne |
| **2** | Slabo rizična (narančasto) | Da |
| **3** | Rizična (žuto) | Da |
| **4** | Visoko rizična (zeleno) | Ne |
| **5** | Izrazito visoko rizična (ljubičasta) | Da |

Izvor: Izrada autorice prema iskazanim rezultatima u radu Vaishali, M.Tech u radu „Fraud detection in Credit Card by Clustering Approach“ [12]

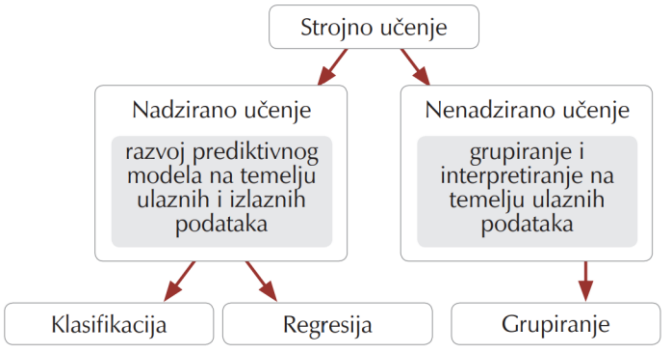
Autori rada iznijeli su zaključak da korištenje *k-means* algoritma pokazuje značajne rezultate u detekciji prijevara, iako se na temelju klasifikacije ne može sa 100% sigurnošću tvrditi da je neka transakcija prijevarna usprkos činjenici da je označena kao jako rizična transakcija.

# 3 METODE ZA OTKRIVANJE PRIJEVARNIH ŠTETA

U prethodnom poglavlju opisani su radovi u kojima su autori ispitivali različite metode strojnog učenja za otkrivanje prevara u osiguranju i u bankovnim transakcijama, a neke od predloženih metoda ispitat će se i u ovom završnom radu.

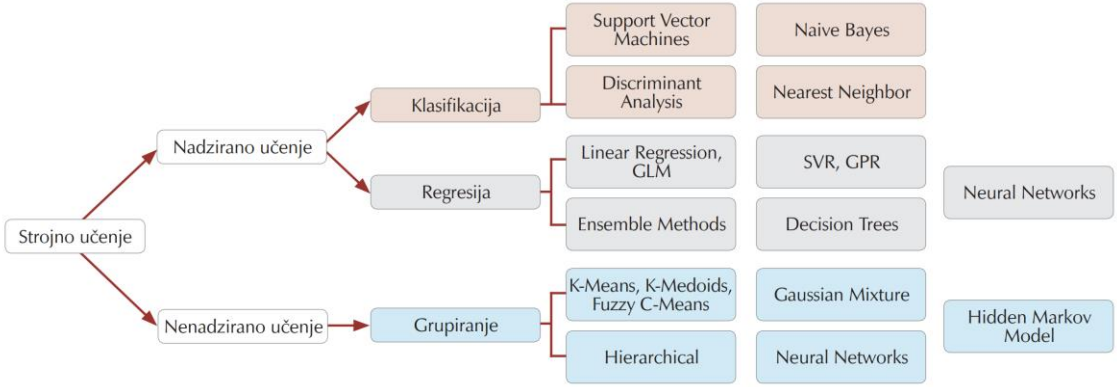
Zadatak algoritama strojnog učenja je pronaći prirodne uzorke i poveznice u podacima te na temelju toga steći uvid i zatim odlučiti i predviđati[10].

Osnovna podjela strojnog učenja je na metode za nadzirano i nenadzirano učenje, a prikaz podjele se nalazi na *Slika 3.1*.




***Slika 3.1*** *Osnovna podjela strojnog učenja*
Izvor: Bolf N., „Strojno učenje“ [13]

Metode za nadzirano učenje grade model sa skupom ulaznih i poznatim skupom izlaznih podataka te uvježbavaju model za predikciju izlaznih podataka, a razlikuju se metode za klasifikaciju i regresiju. Za razliku od njih, metode za nenadzirano učenje pronalaze skrivene uzorke u ulaznim podacima bez poznavanja izlaza.




***Slika 3.2*** *Podjela metoda strojnog učenja*
Izvor: Bolf N., „Strojno učenje“ [13]

[10] Bolf N. , „Strojno učenje“ [13]

U ovom radu će se ispitivati metode za nadzirano učenje (logistička regresija i slučajne šume) na klasifikacijskom problemu - je li prijavljena šteta osiguravajućem društvu prijevarna ili ne. Dodatno će se ispitati metoda za nenadzirano učenje *k-means* kako bi se provjerio optimalan broj klastera u koje se štete mogu grupirati prema nekim sličnostima.

## 3.1 Metode za nadzirano učenje

Kao što je spomenuto u uvodu, u nastavku su opisane dvije metode za nadzirano učenje - logistička regresija i slučajne šume[11], obzirom da će se one ispitivati na podacima za otkrivanje prijevarnih šteta.

Teorijski dio temeljen je na definiciji u [14] i [15].

Prva metoda koja će se opisati malo detaljnije je **logistička regresija**. Algoritam logističke regresije proizlazi iz ideje za modeliranje aposteriornih vjerojatnosti K razreda koristeći linearne funkcije za varijablu „X“ istodobno osiguravajući da njihov zbroj bude u intervalu [0,1]. Model ima oblik:

$$log\frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$

$$log\frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$log\frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_1^T x \tag{3.1}$$

Može se pokazati da je suma:

$$\Pr(G=k|X=x) = \frac{exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} exp(\beta_{k0} + \beta_k^T x)}, k = 1, \ldots., K-1$$

$$\Pr(G=K|X=x) = \frac{1}{1 + \sum_{l=1}^{K-1} exp(\beta_{k0} + \beta_k^T x)} \tag{3.2}$$

jednaka 1.

Uvođenjem nove varijable $\theta = \{\beta_{10}, \beta_1^T, \ldots\ldots, \beta_{(K-1)0}, \beta_{(K-1)}^T\}$, posteriorna vjerojatnost definirana u (3.2) može se zapisati:

$$\Pr(G=k|X=x) = p_k(x;\theta) \tag{3.3}$$

[11] engl. *random forest*

Model logističke regresije maksimizira uvjetnu vjerojatnost od G za dani X. Log vjerojatnost (*engl. log-likelihood*) za N opservacija može se izračunati kao:

$$l(\theta) = \sum_{k=0}^{N} log p_{g_i}(x_i; \theta) \tag{3.4}$$

gdje je $p_k(x; \theta) = Pr(G = x|X = x_i; \theta)$

Klasifikacija u samo dvije klase značajno pojednostavljuje algoritam pa će se u nastavku objasniti algoritam za klasifikaciju u dvije klase. Označimo klase sa 0 i 1, pri čemu je $y_i = 1$ kada je $g_i = 1$ i $y_i = 0$ kada je $g_i = 2$. Dodatno, neka je $p_1(x; \theta) = p(x; \theta)$ i $p_2(x; \theta) = 1 - p(x; \theta)$.

Sada se log vjerojatnost (engl. *log-likelihood)* može zapisati kao

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \{y_i \log p(x_i; \beta) + (1 - y_i) \log(1 - p(x_i; \beta))\} \\ &= \sum_{i=1}^{N} \{y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})\} \end{aligned} \tag{3.5}$$

gdje je $\beta = \{\beta_{10}, \beta_1\}$, uz pretpostavku da vektor opservacija $x_i$ sadrži konstantu 1.

Da bi se maksimizirala log vjerojatnost potrebno je derivaciju matrice izjednačiti sa 0:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta)) = 0 \tag{3.6}$$

Rješenje dobivenih jednadžbi nije moguće dobiti u zatvorenoj formi već se rješavaju numerički, najčešće korištenjem *Newton–Raphson* iterativnog algoritma.

Druga metoda za nadzirano učenje koja će se koristiti u ovom radu je algoritam **slučajnih šuma**[12].

Algoritam slučajnih šuma je općenit naziv za skupinu metoda koje se koriste kolekcijom stablastih klasifikatora pri čemu je skup nezavisnih slučajnih vektora jednake distribucije, a x ulazni vektorski uzorak. Svako stablo daje svoj glas za ulazni uzorak, a šuma određuje klasu uzorka na temelju većine glasova.

Obzirom da svako stablo generirano *bagging* algoritmom ima istu distribuciju, prosječno očekivanje B jednako distribuiranih stabla je isto kao i očekivanje bilo kojeg stabla. Što znači da je pristranost *bagging* stabla ista kao i individualnog stabla, odnosno zaključuje se da je jedina mogućnost poboljšanja algoritma smanjenjem varijance.

[12] engl. *random forest*

Cilj slučajnih šuma u odnosu na ostale *bagging* algoritme je unaprijediti izračun za smanjenje varijance na način da se smanjuje korelacija između stabala. To je postignuto tako da se u procesu izgradnje stabla uzorci odabiru na slučajan način.

Za p-dimenzionalni vektor $X = (X_1, \dots, X_P)^T$ koji predstavlja slučajne varijable koje se koriste za predikciju i slučajnu varijablu $Y$ koja predstavlja stvarne vrijednosti, pretpostavimo da ja $P_{XY}(X, Y)$ njihova zajednička nepoznata distribucija te je potrebno pronaći funkciju $f(X)$ za predikciju $Y$. Funkcija predikcije je određena funkcijom gubitka $L\big(Y, f(X)\big)$ koja je definirana tako da minimizira očekivani gubitak:

$$E_{XY}[L(Y, f(X)] \tag{3.7}$$

Intuitivno gledajući, $L\big(Y, f(X)\big)$ govori koliko je predikcija $f(X)$ blizu stvarne vrijednosti Y.

Kod regresije, $L$ je definirana kao :

$$L\big(Y, f(X)\big) = (Y - f(X))^2 \tag{3.8}$$

Dok je kod klasifikacije definirana kao :

$$L\big(Y, f(X)\big) = I\big(Y \neq f(X)\big) = \begin{cases} 0 \;\; ako\; je\; y = f(X) \\ 1 \;\; inače \end{cases} . \tag{3.9}$$

Može se pokazati da minimalizacija $E_{XY}[L(Y, f(X)]$ za regresiju daje uvjetno očekivanje koje se može definirati kao:

$$f(x) = E[Y|X = x] \tag{3.10}$$

odnosno regresijsku funkciju (*engl. regression function*).

U slučajevima klasifikacije, označimo li sa $\mathcal{Y}$ skup svih vrijednosti koje može poprimiti varijabla Y, minimizacija $E_{XY}[L(Y, f(X)]$ daje:

$$f(x) = \arg max\, P(Y = y|X = x) \tag{3.11}$$

Ansambli konstruiraju $f$ u smislu kolekcije osnovnih modela za učenje (*engl. base-learners)* $h_1(x), \dots, h_J(x)$, te su oni kombinirani kako bi dali najbolji prediktor ansambla *f(x).*

Kod linearne regresiji to se može zapisati kao:

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \tag{3.12}$$

Dok kod problema klasifikacije, $f(x)$ je klasa sa najvećim brojem glasova

$$f(x) = \arg\max_{y \in \mathcal{Y}} \sum_{j=1}^{J} I\left(y = h_j(x)\right) \tag{3.13}$$

*J*-ta osnova za učenje (engl. *base learner*) je stablo označeno $h_j(X, \Theta_j)$ , gdje je $\Theta_j, j = 1, \dots, J$ skup slučajnih varijabli koje su međusobno nezavisne.

Neka je $\mathcal{D} = \{(x_1, y_1), \dots., (x_N, y_N)\}$ označeni set podataka, gdje je $x_i = (x_{i,1}, \dots, x_{i,p})^T$. Za generiranje svakog slučajnog stabla koristi se idući **algoritam**:

Za *j=1 do J:*

i) Iz seta podataka za učenje $\mathcal{D}$ odaberi uzorak $\mathcal{D}_j$ veličine N

ii) Za odabrani podskup $\mathcal{D}_j$, kreiraj stablo koristeći binarno rekurzivno particioniranje:

(1) Na slučajan način odaberi *m* od *p* ulaznih slučajnih varijabli

(2) Od odabranih *m* slučajnih varijabli, pronađi najbolju varijablu koja će biti mjesto podjele

(3) Podijeli čvor u dva čvora kćeri

Izlaz je ansambl stabala $\{T_j\}_1^J$.

Za predikciju vrijednosti u novoj točki x, koriste se funkcije:

- $\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J} \hat{h}_j(x)$ za regresiju
- $\hat{f}(x) = arg\max_{y} \sum_{j=1}^{J} I(\hat{h}_j(x) = y)$ za klasifikaciju

gdje je $\hat{h}_j(x)$ predikcija za varijablu *x* koristeći *j*-to stablo.

Iz opisa algoritma jasno je da se točnost algoritma može povećati definiranjem iduća tri parametra:

- *m* koji označava broj slučajno odabranih prediktora u svakom čvoru
- *J*, odnosno broj stabala u šumi
- *veličina stabla,* koja se mjeri kao najmanja veličina čvora za podjelu ili kao maksimalan broj završnih čvorova

U problemu klasifikacije kao inicijalna vrijednost za *m* koristi se $m = \sqrt{M}$, gdje je *M* ukupan broj prediktora. U problemu regresije inicijalna vrijednost je *m=N/3* pri čemu je *N* veličina uzorka. Obično algoritmi slučajnih šuma nisu jako osjetljivi na veličinu *m,* te postoji jako mala mogućnost prenaučenosti modela uzrokovanog odabirom vrijednosti *m*.

## 3.2 Metode za nenadzirano učenje

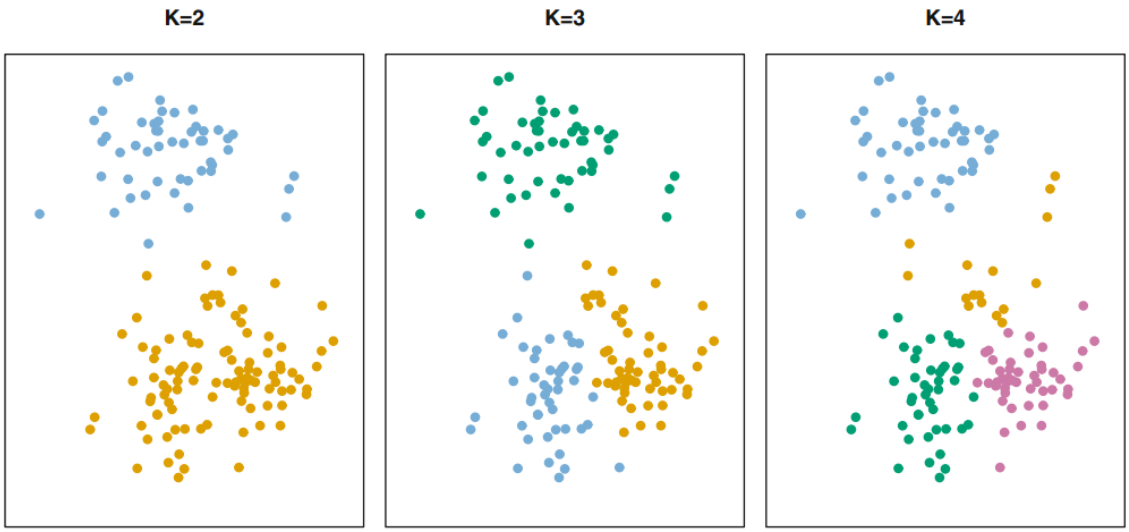
U nastavku će se objasniti teorijski *k-means* metoda za nenadzirano učenje koja je ispitana u ovom radu.

Teorijski dio temeljen je na definiciji metode u [16].

*K-means* klasteriranje je metoda za grupiranje podataka u $K$ različitih klastera. Neka $C_1, \dots, C_K$ označavaju opservacije u svakoj klasi, pri čemu kase zadovoljavaju iduće uvjete:

1. $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots., n\}$. Svaka opservacija pripada jednom od klastera.
2. $C_k \cap C_{k'} = \emptyset$ za sve $k \neq k'$

Ukoliko $i$ -ta opservacija pripada $k$-tom klasteru, tada vrijedi $i \in C_k$. *K-means* algoritam leži nad idejom da je unutar klastera što manja varijanca.




***Slika 3.3*** *K-means klastering*
Izvor: „An Introduction to Statistical Learning with Applications in R“, stranica: 387 [16]

Za klaster $C_k$ označimo sa $W(C_k)$ mjeru za varijancu unutar jednog klastera i tada se problem koji se rješava *k-means* algoritmom može opisati:

$$\min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{3.14}$$

odnosno, želi se grupirati opservacije u $K$ klastera, tako da suma varijanci svih klastera bude minimalna.

Varijanca se definira kao:

$$\mathrm{W}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{x'j})^2 \tag{3.15}$$

gdje je $|C_k|$ broj opservacija u k-tom klasteru, odnosno varijanca u *k*-tom klasteru je suma Euklidskih udaljenosti između svih parova opservacija koje pripadaju k-tom klasteru podijeljena sa brojem opservacija koje pripadaju k-tom klasteru.

Iz (3.14) i (3.15), problem optimizacije *k-means* klasteriranja može se zapisati kao

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{3.16}$$

Problem podjele *n* opservacija u *K* klastera može se riješiti na $K^n$ načina, a predložen algoritam za *k-means* klasteriranje glasi:

1. Na slučajan način svakoj opservaciji dodijeli brojeve od 1 do K. Na ovaj način napravljena je inicijalna dodjela klastera opservacijama.
2. Ponavljaj dok se za određene opservacije dodjela klastera neće mijenjati:
   a) Za svaki od *K* klastera izračunaj centroid. *k*-ti centroid je vektor od *p* značajki za opservacije u k-tom centroidu
   b) Svaku opservaciju dodaj u klaster čiji centroid mu je najbliži (izračun najbližeg centroida računa se koristeći Euklidsku udaljenost).

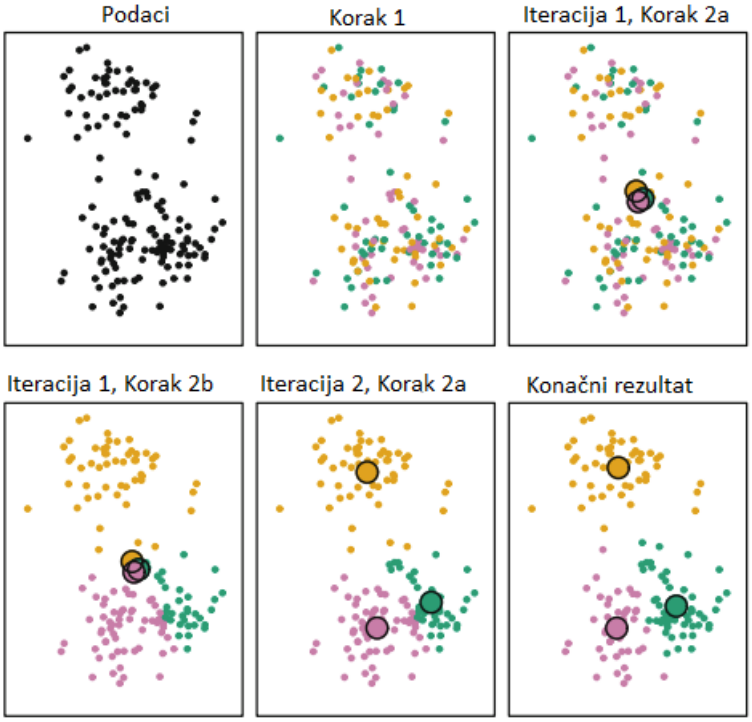
Predloženim algoritmom u svakom koraku smanjuje se vrijednost definirana u (3.16), a to se može opisati i kao:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{x'j})^2 = 2 \sum_{i, \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \tag{3.17}$$

gdje je $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ aritmetička sredina za značajku *j* u klasteru $C_k$.

U algoritmu, u koraku 2.a aritmetičke sredine za svaku značajku su konstante koje minimaliziraju sumu kvadrata razlika između značajki, a u koraku 2.b preraspodjela značajki može samo poboljšati algoritam.

Na *Slika 3.4* grafički su prikazani koraci *k-means* algoritama za K=3.




***Slika 3.4*** *Napredak po koracima K-Means algoritma za K=3*
Izvor: „An Introduction to Statistical Learning with Applications in R“, stranica: 389 [16]

Odabir broja klastera u praksi najčešće ovisi o problemu koji se rješava i koji je cilj, a neke od metoda koje se koriste za određivanje broja klastera prema [17] su:

- **Metoda lakta**[13] u kojoj se za različite brojeve klastera ($K$) za sve podatke izračuna prosječna udaljenost od centroida i nacrtaju se dobiveni rezultati. Optimalan $K$ onaj gdje se prosječna udaljenost od centroida počinje smanjivati. Kad se udaljenost počinje smanjivati, na grafu to izgleda kao kut pa se zbog toga zove i metoda lakta.
- **Analiza siluete**[14] koja mjeri koliko su slični podaci unutar klastera u odnosu na druge klastere. Formula za izračun sličnosti je:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.18}$$

Gdje je

- *S(i)* koeficijent siluete za točku *i*
- *a(i)* prosječna udaljenost između podatka *i* i svih drugih podataka koji pripadaju istom klasteru kao i podatak *i*
- *b(i)* prosječna udaljenost i od svih klastera kojemu podatak *i* ne pripada

[13] engl. *elbow method*
[14] engl. *silhouette analysis*

Tada je

$$prosječna\ vrijednost\ siluete = prosjek\ \{S(i)\} \tag{3.18}$$

Koeficijenti siluete zadovoljavaju iduće uvjete:

- Vrijednosti koeficijenata su u intervalu [-1,1].
- Vrijednost siluete 1 označava da je podatkovna točka kompaktna unutar klastera kojemu pripada, te da je daleko od ostalih klastera.
- Vrijednosti blizu 0 označavaju da se klasteri preklapaju.

Broj klastera je optimalan ukoliko prosječna vrijednost siluete za svaki klaster je veća od ukupne prosječne vrijednosti siluete.

# 4 OPIS PODATAKA

Za implementaciju modela korišteni su podaci o prijavljenim štetama na motornim vozilima u periodu od dvije godine kod jednog osiguravajućeg društva iz Republike Hrvatske i zbog tajnosti podataka u radu neće biti spomenuti ni opisani osjetljivi podaci.
U uzorku podataka promatrane su varijable koje se prikupljaju u procesu prijave i obrade šteta. Analiza podataka napravljena je u programskom alatu *Python*.

## 4.1 Definicija varijabli

U *Tablica 4.1* nalazi se popis i značenje varijabli koje su analizirane i koriste se u ispitanim modelima za detekciju prijevara.

***Tablica 4.1*** *Popis i opis varijabli iz uzorka kojima se opisuju štete*

| Varijabla | Opis |
|---|---|
| STETA_ID | Jedinstveni identifikator štete. |
| VRSTA_STETE | Varijabla kojom se označava vrsta štete, a ovisi o riziku po kojem je prijavljena šteta. Kod spremanja štete dodjeljuje se automatski ovisno o riziku po kojem je prijavljena šteta.<br>Moguće vrijednosti:<br>AK - štete prijavljene po riziku automobilskog kaska<br>AO - štete prijavljene po riziku auto odgovornosti<br>MD - međunarodne štete prijavljene po riziku auto odgovornosti |
| RIZIK_SIFRA, RIZIK NAZIV | Varijabla koja označava rizik po kojem se prijavljuje šteta.<br>RIZIK_*SIFRA* je deveteroznamenkasti broj i njegovo značenje opisano je u varijabli RIZIK_*NAZIV*. |
| BROJ_POLICE | Broj police po kojoj je prijavljena šteta. |

| | |
|---|---|
| BROJ_STETA_PO_ISTOJ_POLICI | Broj šteta koje su prijavljene po istoj polici kao i promatrana šteta. |
| DATUM_PRIJAVE | Datum prijave štete. |
| DATUM_NASTANKA | Datum nastanka štete. |
| VRIJEME_NASTANKA_DOGADJAJA | Vrijeme nastanka štetnog događaja. |
| NAZIV_DRZAVA_DOGADJAJA | Naziv države u kojoj se dogodila prometna nezgoda. Vrijednosti se popunjavaju iz šifrarnika. |
| ZUPANIJA_DOGADJAJA | Naziv županije u kojoj se dogodila prometna nezgoda. Podatak se evidentira samo za štetne događaje koji su nastali u Hrvatskoj. |
| UZROK | Uzrok nastanka štete. Popunjava se kod prijave te se vrijednosti odabiru iz šifrarnika koji je mapiran uz vrste rizika. |
| TIP_OSTECENIKA | Tip oštećenika. Moguće vrijednosti su F (fizička osoba), P (pravna osoba) i O (obrtnik). |
| SPOL_OSTECENE_OSOBE | Spol oštećenika. Moguće vrijednosti 1 (Ženski), 2 (Muški), 3 (Nepotreban). |
| GODINA_ROĐENJA_OSTECENIKA | Godina rođenja oštećenika. |
| RAC | Oznaka je li oštećenik ujedno i *Rent a car*. |
| POLICIJA | Oznaka je li na mjesto štetnog događaja bila policija. Moguće vrijednosti su D i N. |
| VATROGASCI | Oznaka jesu li na mjesto štetnog događaja bili vatrogasci. Moguće vrijednosti su D i N. |
| HITNA_POMOC | Oznaka je li na mjestu štetnog događaja bila policija. Moguće vrijednosti su D i N. |
| VRSTA_OSTECENOG_VOZILA | Vrsta oštećenog vozila. |
| MARKA_OSTECENOG_VOZILA | Marka oštećenog vozila. |
| GOD_PRO_OST_VOZILA | Godina proizvodnje oštećenog vozila. |
| PREMIJA_OD | Datum od kad vrijedi premija police po kojoj se prijavljuje šteta. |

| | |
|---|---|
| PREMIJA_DO | Datum do kad vrijedi premija police po kojoj se prijavljuje šteta. |
| OSTECENA_OSOBA_U_SDT | Indikator je li u istom štetnom događaju uz vozilo oštećena i osoba. Moguće vrijednosti su:<br>• 0 - ukoliko nije oštećena osoba<br>• 1 - ukoliko je oštećena i osoba |
| OSTECENA_IMOVINA_U_SDT | Indikator je li u istom štetnom događaju uz vozilo oštećena i neka druga imovina. Moguće vrijednosti su:<br>• 0 - ukoliko nije oštećena druga imovina<br>• 1 - ukoliko je oštećena i druga imovina |
| BRZA_NAGODBA | Oznaka je li nakon prijave štete sa strankom dogovorena isplata odštete brzom nagodbom. |
| TOTALNA_STETA | Oznaka je li vozilo totalno oštećeno. Moguće vrijednosti su D i N. |
| ODVJETNIK_OSTECENOG | Jedinstveni identifikator odvjetnika ukoliko je zastupao oštećenika u procesu prijave štete osiguranjnom društvu. |
| VISINA_STETE | Varijabla koja sadrži iznos (visinu) štete. |
| STRANKA_VRSI_PRITISAK | Varijabla u kojoj je označeno radi li u procesu obrade štete oštećenik pritisak ili predobro poznaje procese u osiguranju. |
| PRIJEVARA | Varijabla kojom označavamo je li za određenu štetu dokazana prijevara. Moguće vrijednosti su 0 (nije dokazana prijevara) i 1 (dokazana je prijevara). |

## 4.2 Deskriptivna analiza podataka

Uzorak čini 160 tisuća šteta nastalih na motornim vozilima prijavljene po policama automobilske odgovornosti[15] i auto kaska[16].

Svaka šteta prijavljuje se po polici koju je ugovaratelj sklopio sa društvom za osiguranje[17]. U uzorku se nalaze štete prijavljene na 130 tisuća različitih polica (sve police su jednogodišnje). Tri četvrtine podataka u uzorku su auto kasko štete, dok ostatak čine štete prijavljene po riziku automobilske odgovornosti.

U nastavku se nalazi deskriptivna analiza varijabli[18]:

- RIZIK (RIZIK SIFRA i RIZIK NAZIV)

  Rizik po kojem osnovu je prijavljena šteta nalazi se u varijablama RIZIK_SIFRA i RIZIK NAZIV. U uzorku podataka nalazi se 125 različitih vrijednosti, međutim više od 2/3 šteta iz uzorka prijavljeno je po svega tri rizika koji imaju najveću frekvenciju i odnose se na osiguranje osobnih vozila (puni kasko i osiguranje od automobilske odgovornosti).

  U uzorku podataka je varijabla RIZIK_SIFRA bila deveteroznamenkasti broj koji se koristi da bi se detaljnije opisao rizik koji se osigurava, međutim sva izvještavanja prema regulatoru (Hanfa[19]) rade se na razini peteroznamenkastog broja, te je zbog toga uvedena nova varijabla CJENIK_SIFRA_5 i na taj način dobiveno je 6 različitih vrijednosti u varijabli CJENIK_SIFRA_5 i ta varijabla će se koristiti u ispitivanju metoda. I u novoj

[15] U ovoj vrsti osiguranja društvo za osiguranje umjesto osiguranika koji je prouzročio štetu trećoj osobi plaća nastalu štetu do ugovorenog limita ili limita propisanog zakonom. [2]

[16] U slučaju kasko osiguranja radi se o dobrovoljnom osiguranju kojim se osigurava djelomična ili potpuna šteta na vozilima, zrakoplovima, jahtama i sl., ovisno o tome što je ugovoreno. Za razliku od osiguranja od automobilske odgovornosti, kod kasko osiguranja od štete osiguravate svoje vozilo. [2]

[17] Društvo za osiguranje je pravna osoba sa sjedištem u Republici Hrvatskoj koja obavlja poslove životnog ili neživotnog osiguranja, koja ima odobrenje Agencije za obavljanje poslova osiguranja i upisana je u sudski registar nadležnog trgovačkog suda. [18]

[18] Zbog osjetljivosti podataka u deskriptivnoj analizi nisu navođeni apsolutni iznosi ili kao ni postoci koji su povjerljivi.

[19] Hrvatska agencija za nadzor financijskih usluga (Hanfa) je nadzorno tijelo u čiji djelokrug i nadležnost spada nadzor financijskih tržišta, financijskih usluga te pravnih i fizičkih osoba koje te usluge pružaju.

Hanfa provodi nadzor nad poslovanjem burzi i uređenih javnih tržišta, investicijskih društava i izdavatelja vrijednosnih papira, brokera i investicijskih savjetnika, vezanih zastupnika, središnjeg klirinškog depozitarnog društva, društava za osiguranje i reosiguranje, zastupnika i posrednika u osiguranju i reosiguranju, društava za upravljanje investicijskim i mirovinskim fondovima, mirovinskih osiguravajućih društava, investicijskih i mirovinskih fondova, Središnjeg registra osiguranika, Fonda hrvatskih branitelja iz Domovinskog rata i članova njihovih obitelji i pravnih osoba koje se bave poslovima leasinga i faktoringa, osim ako ih banke obavljaju unutar svoje registrirane djelatnosti. [19]

kreiranoj varijabli CJENIK_SIFRA_5 ne postoji jednaka distribucija broja šteta po vrijednostima, već je po tri vrijednosti prijavljeno 99% šteta.
Nadalje, ako se promatra udio prijevara u odnosu na varijablu CJENIK_SIFRA_5 - na riziku po kojem je prijavljeno najmanje šteta je najveći udio prijevarnih šteta[20].

- UZROK

Za auto kasko štete, osim rizika po kojem se prijavljuje šteta, evidentira se i uzrok nastanka štete (za štete prijavljene po auto odgovornosti taj podatak se ne evidentira). U uzorku podataka se nalazi 47 različitih uzroka. Uzroci sa najvećim frekvencijama su *'Nepoznata osoba oštetila vozilo na parkiralištu* ' i '*Udar vozila u objekt'* (pojavljuju se na više od 50% auto kasko šteta).
U uzorku podataka, neki uzroci imaju isto ili slično značenje (kao na primjer 'Nepoznati uzroci' i 'Ostalo') te je iz tog razloga napravljeno čišćenje podataka na način da ukoliko se pojavljuje više uzroka sa istim ili sličnim nazivima, svi se zamijene sa istom vrijednosti.
Također, uzroci koji imaju frekvenciju manju od 30, zamijenjeni su sa uzrokom 'Ostalo'.
Nakon ovakvih transformacija, dobiveno je 26 različitih vrijednosti.
Kad se gledaju uzroci i broj šteta na kojima je dokazana prijevara, a imaju evidentiran uzrok, samo na dva uzroka nema ni jedne prijevarne štete.

- VATROGASCI, POLICIJA i HITNA POMOĆ

Varijable VATROGASCI, POLICIJA i HITNA POMOĆ mogu poprimiti vrijednosti D ili N te označavaju je li na mjestu štetnog događaja bili određena žurna služba. Od spomenutih žurnih službi najčešće je na mjestu štetnog događaja bila policija i to na 13% šteta u uzorku, dok su najmanje puta bili VATROGASCI i to na manje od 0,02% šteta. Kao priprema za korištenje podataka u modelu vrijednosti 'D' zamijenjene su sa 1, a 'N' sa 0.

[20] Zbog osjetljivosti podataka u radu nije naveden točan udio prijevarnih šteta.

- BRZA_NAGODBA

Varijabla BRZA_NAGODBA ima popunjenu vrijednost samo u slučaju kad je oštećeniku isplaćena šteta odmah nakon prijave brzom nagodbom, inače je ta vrijednost prazna. U pripremi podataka za korištenje u prediktivnom modelu prazne vrijednosti su popunjene sa 0. Ukoliko gledamo u kontekstu prijevara koje su kasnije potvrđene, na štetama koje su isplaćene brzom nagodbom ta brojka je beznačajna te iznosi manje od 0,01%.

- TOTALNA_STETA

Varijabla TOTALNA_STETA popunjena je na 6% podataka u uzorku sa vrijednostima D i N, dok je na ostalim štetama ta vrijednost prazna. Vrijednost D je zamijenjena sa 1, a vrijednost N i nepopunjene vrijednosti zamijenjene su sa vrijednosti 0. U uzorku podataka veći je udio prijevara na totalnim štetama nego na ostalim štetama i to gotovo za 2%.

- DATUM_NASTANKA

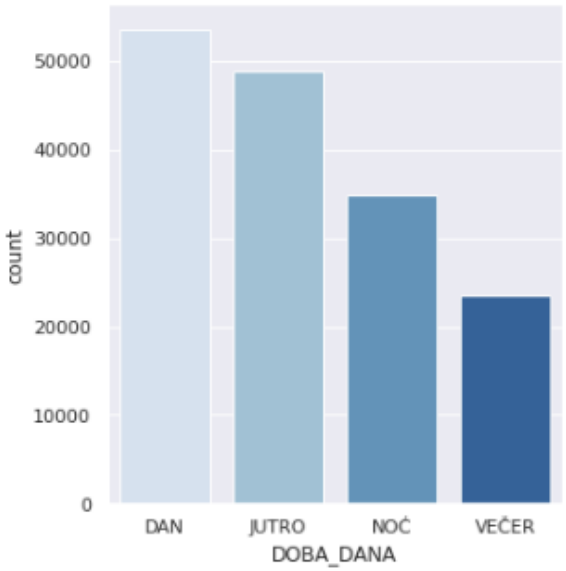
99% šteta u uzorku nastale su nakon 01.01.2018., međutim postoje datumi koji odstupaju, pa je najmanji datum nastanka štete u uzorku 17.06.1993. Kad se promatraju štete količinski i financijski u odnosu na datum nastanka, značajna odstupanja su se pojavila na dane elementarnih nepogoda (potres, tuča,…).

- VRIJEME_NASTANKA_DOGADJAJA

Vrijeme nastanka štetnog događaja nalazi se u formatu hh:mm (sati:minute), ali je za potrebe ove analize uvedena nova značajka DOBA_DANA u kojoj je vrijeme podijeljene u četiri grupe:

- noć (00:00-05:59)
- jutro (06:00-11:59)
- dan (12:00-17:59) i
- večer (18:00-23:59).

Ukoliko se promatra udio prijavljenih šteta u odnosu na doba dana, najviše šteta je nastalo u periodu 12-17:59 (*dan*), zatim slijedi *jutro* (06:00-11:59), pa *noć* (00:00-05:59) i najmanje na *večer* (čak duplo manje nego preko dana). Prikaz broja nastalih šteta u odnosu na doba dana nalazi se na *Slika 4.1.*

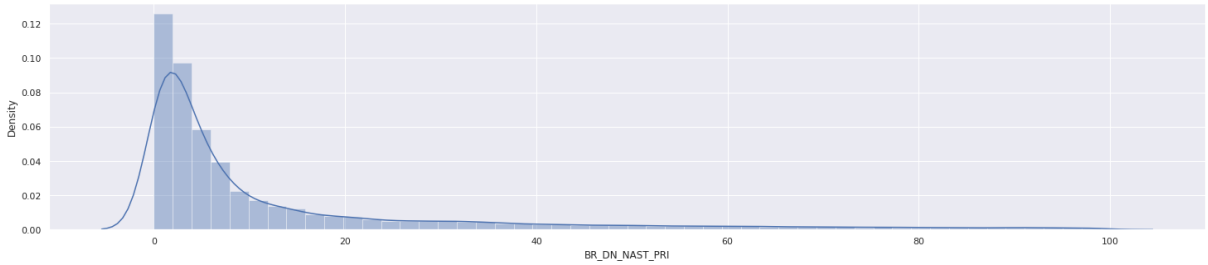



***Slika 4.1*** *Frekvencija broja šteta u odnosu na doba dana*
Izvor: Izrada autorice

Uspoređujemo li udio prijevarnih šteta u odnosu na doba dana, najmanje prijevara ima na štetama koje su nastale ujutro, dok najviše ima na štetama nastalim navečer, iako je je udio prijevarnih šteta kroz dan jako blizu onima nastalim kroz večer.

- DATUM_PRIJAVE

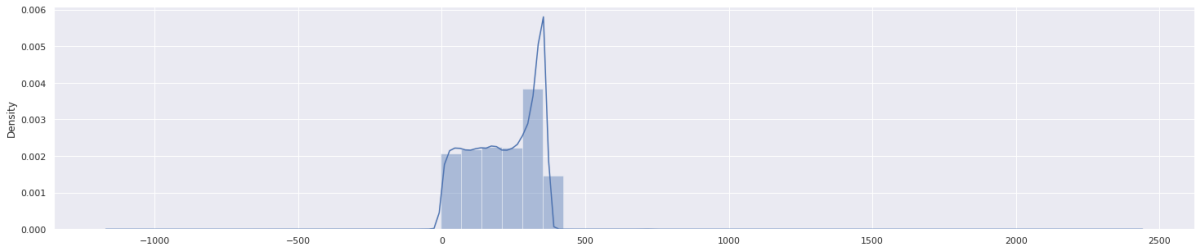
Osim podataka o datumu i vremenu nastanka, postoji i podatak kad je oštećenik prijavio štetu osiguranjeljnom društvu i taj podatak je evidentiran u varijabli *DATUM_PRIJAVE*. Podatak kao takav nije posebno zanimljiv ni koristan za analizu i prediktivan model, ali koristeći taj datum i datum nastanka štete kreirana je nova varijabla *BR_DN_NAST_PRI* koja označava koliko je dana prošlo od nastanka do prijave štete i ta će se varijabla dalje koristiti u modelu. Iz nove kreirane varijable se vidi da je 50% šteta prijavljeno unutar 5 dana od nastanka, dok je 75% šteta u uzroku prijavljeno unutar 25 dana od nastanka (distribucija je prikazana na *Slika 4.2*).




***Slika 4.2*** *Distribucija broja dana od nastanka do prijave štete (za štete koje su prijavljene unutar 100 dana)*
Izvor: Izrada autorice

- PREMIJA_OD, PREMIJA_DO

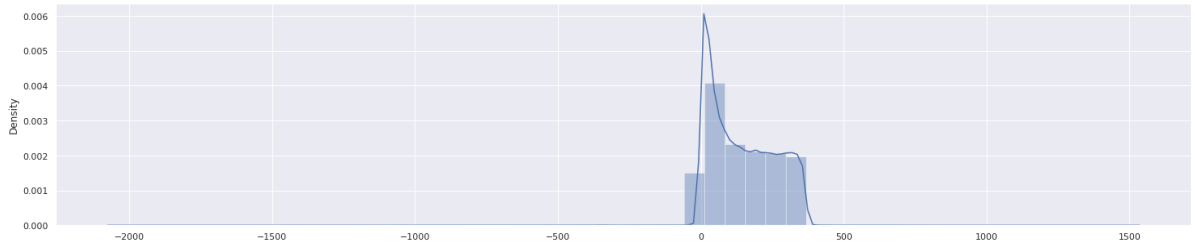
Dodatno, na polici se nalazi podatak od kad do kad vrijedi premija osiguranja te taj podatak sam za sebe nije posebno zanimljiv, ali je korišten da se izvedu još dvije nove varijable u kojima se evidentira broj dana koliko je prošao od početka premije do nastanka štete (varijabla: *BR_DN_NAST_POC_PREM*) te koliko je dana prošlo od nastanka štete do kraja premije (varijabla: *BR_DN_NAST_KRAJ_PREM*). Na *Slika 4.3* prikazana je distribucija varijable *BR_DN_NAST_POC_PREM*.




***Slika 4.3*** *Distribucija broja dana od početka premije do nastanka šteta*
Izvor: Izrada autorice

U uzorku se nalaze primjeri šteta kojima je datum nastanka štete prije datuma početka premije na polici po kojoj je prijavljena šteta. U nekim primjerima radi se o slučajevima kad je krivo evidentiran datum premije na polici, međutim postoje i primjeri u kojima je evidencija ispravna, te se kroz status štete može dalje analizirati jesu li bili zadovoljeni svi uvjeti za isplatu odštete, te radi li se o prijevarnim štetama. Zbog drugog navedenog primjera, na štetama nisu korigirani datumi početka premije ili datumi nastanka šteta jer se ne može utvrditi koji primjeri su ispravni, a gdje je došlo do pogreške pri unosu podataka.

Druga nova varijabla je BR_DN_NAST_KRAJ_PREM i ona označava broj dana od nastanka štete do kraja premije te je na *Slika 4.4* je prikazana distribucija njezinih vrijednosti. Negativne vrijednosti ove varijable mogu se objasniti na isti način kao i za varijablu BR_DN_NAST_POC_PREM.




***Slika 4.4*** *Distribucija broja dana od dana nastanka štete do kraja premije*
Izvor: Izrada autorice

- BROJ_STETA_PO_ISTOJ_POLICI

Za policu po kojoj je prijavljena šteta, u ovoj varijabli evidentira se broj šteta koje su prijavljene po toj istoj polici. Napravljena je analiza ugovaratelja polica po kojima je prijavljeno više od 10 šteta.

Ugovaratelji takvih polica su pravne osobe koje su ujedno servisne radionice ili leasing, a to su primjeri polica gdje je na jednoj polici osigurano nekoliko vozila. Preko 90% šteta na takvim policama je prijavljeno po riziku puni kasko.

- TIP_OSTECENE_OSOBE

Za gotovo svakog oštećenika u uzorku postoji evidencija radi li se o fizičkoj ili pravnoj osobi ili o obrtniku, dok samo nekoliko primjera u uzorku taj podatak poprima vrijednost *Razno* i *Nedefinirano*. Obzirom da u uzorku najveći udio čine fizičke osobe, vrijednosti *Razno* i *Nedefinirano* označene su kao fizičke osobe.

- SPOL_OSTECENE_OSOBE

Varijabla SPOL_OSTECENE_OSOBE poprima tri različite vrijednosti – 1 (žensko), 2 (muško) i 0 (nepotreban). Vrijednost nepotreban je u pravilu popunjen kod pravih subjekata i obrtnika. Kada se promatra udio prijevarnih šteta u odnosu na spol oštećenika, značajno manje ih je na štetama gdje je spol štetnika upisan kao nepotreban, dok između muškaraca i žena nema značajne razlike u udjelima ni ukoliko se uspoređuje količinski ni financijski.

- RAC

Varijabla u kojoj je označeno je li oštećenik ujedno i *Rent a car.* Varijabla je pripremljena za korištenje u prediktivnom modelu tako da je upisano 1 tamo gdje je evidentiran podatak, a gdje nije bila popunjena nikakva vrijednost, popunjeno je 0. Na štetama na kojima je oštećenik *Rent a car* značajno je manje detektiranih prijevara u odnosu na prosjek.

- LEASING

Za oštećenike se evidentira koja je vrsta osobe (leasing, radionica, zdravstvena ustanova, odvjetnik, banka, ljekarna,….), međutim kvaliteta podatka je poprilično loša. Od oštećenika u uzorku kojima je upisana vrsta osobe, najviše ima leasing kuća, pa je uvedena nova varijabla *LEASING* sa mogućim vrijednostima 1 (ukoliko je oštećenik leasing) i 0 (ukoliko nema navedenu vrstu osobe ili ima neki drugu vrstu).

- GODINA_ROĐENJA_OSTECENIKA

Za oštećenike evidentira se i datum rođenja. Iako podatak ima smisla samo za fizičke osobe naveden je i na nekim obrtnicima i pravnim osobama. Da bi se ova varijabla mogla dalje koristiti u prediktivnom modelu, izračunata je starost oštećenika u trenutku nastanka štetnog događaja (godina nastanka štetnog događaja minus godina rođenja oštećenika), te je za pravne osobe i obrtnike upisano 0. Nova varijabla gdje se evidentira starost oštećenika zove se *STAROST_OSTECENIKA*.

- NAZIV_DRZAVA_DOGADJAJA

Štetni događaji su nastali u 49 različitih država, iako ipak značajno najveći udio imaju štete koje su nastale u Hrvatskoj. Na manjem dijelu šteta nije bila navedena država gdje se dogodio nesretni slučaj, pa su takve vrijednosti i vrijednosti država u kojima je bilo manje od 5 nesretnih slučajeva zamijenjene sa vrijednosti NEPOZNATO.
Na taj način dobiveno je u uzorku 33 različitih država, te je u 16 od njih detektirana barem jedna prijevara.

- ZUPANIJA_DOGADJAJA

Za štetne događaje koji su nastali u Hrvatskoj, evidentira se i županija u kojoj se dogodio štetni događaj. Jedna trećina šteta koje su nastale u Hrvatskoj nastale su u gradu Zagrebu. Ukoliko se promatra udio prijevarnih šteta u pojedinoj županiji u odnosu na ukupni broj šteta u određenoj županiji, devet županija ima udio veći od prosjeka u ukupnom uzorku. Uz udio prijevarnih šteta u pojedinoj županiji, promatran je i udio šteta po županijama u ukupnom broju šteta te udio prijevarnih šteta po županijama u ukupnom broju prijevarnih šteta, te je zbog značajne razlike u distribuciji broja šteta po županijama, umjesto županija uvedena nova značajka REGIJA. Županije su grupirane u šest regija: Grad Zagreb, Sjeverozapadna Hrvatska, Istočna Hrvatska, Sjeverni Jadran i Lika te Srednji i Južni Jadran, a prikaz na koji način je napravljeno grupiranje nalazi se u *Tablica 4.2.*

***Tablica 4.2*** *Grupiranje županija u regije*

| Regija | Županija |
|---|---|
| Grad Zagreb | GRAD ZAGREB |
| Središnja Hrvatska | BJELOVARSKO - BILOGORSKA |
| | KARLOVAČKA |
| | SISAČKO - MOSLAVAČKA |
| | ZAGREBAČKA |
| Sjeverozapadna Hrvatska | MEĐIMURSKA |
| | KOPRIVNIČKO - KRIŽEVAČKA |
| | VARAŽDINSKA |
| | KRAPINSKO - ZAGORSKA |
| Istočna Hrvatska | VIROVITIČKO - PODRAVSKA |
| | POŽEŠKO - SLAVONSKA |
| | VUKOVARSKO - SRIJEMSKA |
| | BRODSKO - POSAVSKA |
| | OSJEČKO - BARANJSKA |
| Sjeverni Jadran i Lika | LIČKO - SENJSKA |
| | ISTARSKA |
| | PRIMORSKO - GORANSKA |
| Srednji i Južni Jadran | ŠIBENSKO - KNINSKA |
| | DUBROVAČKO - NERETVANSKA |
| | ZADARSKA |
| | SPLITSKO - DALMATINSKA |

Izvor: Izrada autorice

- VRSTA_OSTECENOG_VOZILA

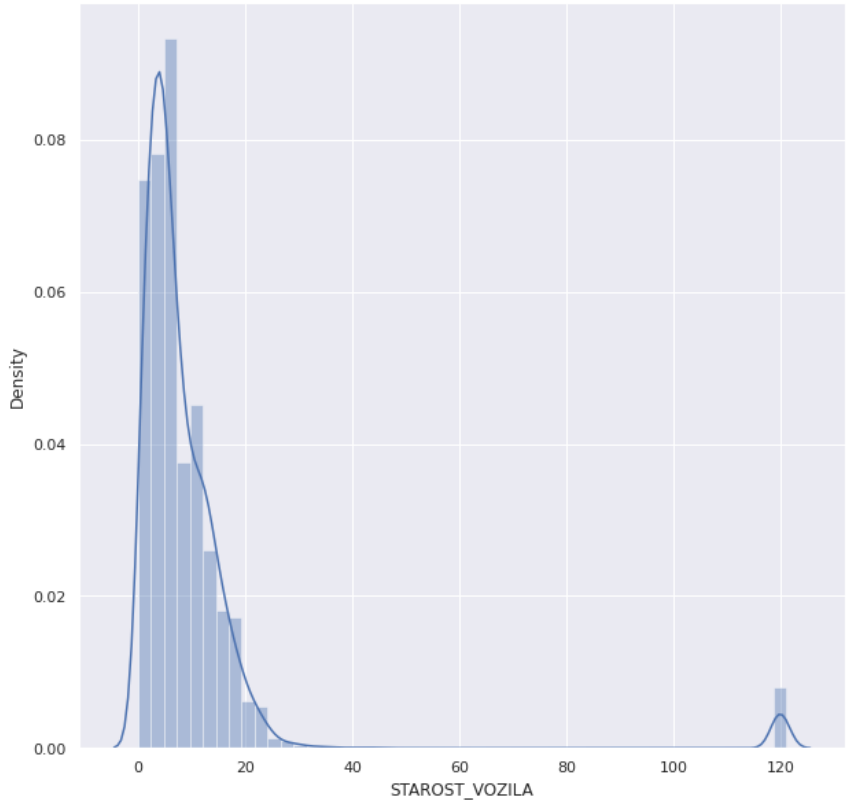
Analiza vrste oštećenih vozila (varijabla VRSTA_OSTECENOG_VOZILA) u štetama iz uzorka pokazuje da najveći udio imaju osobna vozila (81%), dok teretna vozila, traktori, motocikli, autobusi, priključna i specijalna vozila te ona za koje nije navedena vrsta imaju znatno manji udio. Međutim, obzirom da velikom udjelu vozila nije navedena vrsta, varijabla se neće koristiti u prediktivnom modelu.

- MARKA_OSTECENOG_VOZILA

Osim vrste vozila, evidentira se i marka vozila - u uzorku nalazi 425 različitih marki, međutim nakon čišćenja podataka (npr. ista marka je upisana 2 puta sa različitim nazivima) te označavanjem svih marki vozila koje se u uzorku pojavljuju na manje od 100 šteta sa 'OSTALO VOZILO', dobivene su 52 različite marke vozila. Najveći udio prijevarnih šteta ima marka BMW.

- GOD_PRO_OST_VOZILA

U varijabli GOD_PRO_OST_VOZILA evidentirana je za oštećena vozila godina njihove proizvodnje. Varijabla je pripremljena na način da su prazne vrijednosti ili vrijednosti koje su manje od 1900 popunjene sa 1900, a zatim je izvedena nova varijabla STAROST_VOZILA u kojoj je izračunata starost vozila u trenutnu nastanka štete. Prosječna starost vozila u trenutku nastanka štete je 9,6 godina, kolika je i prosječna starost vozila u štetama na kojima nije detektirana prijevara, dok u štetama na kojima je detektirana prijevara je prosječna starost vozila 10 godina. Na *Slika 4.5* prikazana je distribucija broja šteta u odnosu na starost oštećenog vozila.




***Slika 4.5*** *Distribucija broja šteta u odnosu na starost oštećenog vozila*
Izvor: Izrada autorice

- OSTECENA_OSOBA_U_SDT

U štetnom događaju osim vozila može biti oštećena/ozlijeđena osoba ili neka imovina. Obzirom da se takve štete ne nalaze u uzorku koji je predmet ove analize, na šteti se nalaze varijable koje označavaju postoji li u štetnom događaju još neko oštećenje.

OSTECENA_OSOBA_U_SDT poprima vrijednost 1 ukoliko je u štetnom događaju oštećena osoba i takvih je u uzorku svega 2% šteta. Ukoliko se gleda udio prijevarnih šteta u odnosu na to je li oštećena i osoba ili ne, neznatno je više prijevarnih šteta nastalih u

štetnom događaju u kojem je ozlijeđena i osoba. Ukoliko se to još stavi u kontekst sa štetama kod kojih na mjestu štetnog događaja nije bila hitna pomoć, može se zaključiti da je upravo to jedan od indikatora prijevarnih šteta.

- OSTECENA_IMOVINA_U_SDT

Na isti način kao i OSTECENA_OSOBA_U_SDT pripremljena je i varijabla OSTECENA_IMOVINA_U_SDT koja označava je li u štetnom događaju prijavljena i šteta na nekoj imovini. Udio takvih šteta u uzorku je još manji (0,6%), te je udio prijevarnih šteta manji na štetama gdje je u štetnom događaju oštećena i imovina od ukupnog udjela prijevarnih šteta, ali nije zanemariv.

- ODVJETNIK_OSTECENOG

Još jedna zanimljiva varijabla je ODVJETNIK_OSTECENOG u kojoj se nalazi podatak o odvjetniku oštećenika ukoliko ga je zastupao u procesu prijave štete. U uzorku takvih šteta ima manje od 500, a najveće frekvencija pojavljivanja određenih odvjetnika u štetama je 25 puta. Iz navedene varijable kreirana je nova varijabla ODVJETNIK_OSTECENOG_IND u kojoj je sa 1 i 0 označeno postoji li na šteti odvjetnik ili ne. Zanimljivo je da je udio prijevarnih šteta na štetama u kojima se nalazi odvjetnik gotovo 1% veći od prosječnog udjela prijevarnih šteta u uzorku.

- STRANKA_VRSI_PRITISAK

U procesu prijave i obrade štete, osobe zadužene za obradu štete evidentiraju na šteti radi li stranka pritisak na njih ili predobro poznaje sve procese u osiguranju (subjektivni dojam). Ukoliko je označeno da stranka vrši pritisak, vrijednost u ovoj varijabli je 1, inače je 0. Na više od 50% šteta na kojima je u procesu obrade označeno da stranka vrši pritisak, dokazana je prijevara.

- VISINA_STETE

Svaka šteta ima numeričku vrijednost koja označava njezinu vrijednost, međutim uvedena je nova varijabla RANG_STETE u kojoj je u ovisnosti o visini štete šteti dodijeljen jedan od 5 rangova:

  - Rang 1: štete čija je visina u rasponu 0-2000 kn
  - Rang 2: štete čija je visina u rasponu 2001-5000kn

- Rang 3: štete čija je visina u rasponu 5001-10000kn
- Rang 4: štete čija je visina u rasponu 10001-20000kn
- Rang 5: štete čija je visina iznad 20001kn

Prvi i drugi rang čine najveći udio svih šteta.

Na temelju deskriptivne analize podataka, definirano je **26** varijabli koje se mogu koristiti u prediktivnom modelu te je popis je prikazan u *Tablica 4.3*.

***Tablica 4.3*** *Popis varijabli koje će se koristiti u prediktivnom modelu*

| KVALITATIVNE VARIJABLE | KVANTITATIVNE VARIJABLE |
|---|---|
| CJENIK_SIFRA_5 | BR_DN_NAST_KRAJ_PREM |
| DOBA_DANA | BR_DN_NAST_POC_PREM |
| NAZIV_DRZAVA_DOGADJAJA | BR_DN_NAST_PRI |
| ZUPANIJA_DOGADJAJA | BROJ_STETA_PO_ISTOJ_POLICI |
| SPOL_OSTECENE_OSOBE | STAROST_OSTECENIKA |
| TIP_OSTECENE_OSOBE | STAROST_VOZILA |
| UZROK | |
| MARKA_OSTECENOG_VOZILA | |
| BRZA_NAGODBA | |
| HITNA_POMOC | |
| POLICIJA | |
| VATROGASCI | |
| ODVJETNIK_OSTECENOG_IND | |
| OSTECENA_IMOVINA_U_SDT | |
| OSTECENA_OSOBA_U_SDT | |
| RAC | |
| TOTALNA_STETA | |
| LEASING | |
| STRANKA_VRSI_PRITISAK | |
| RANG_STETE | |

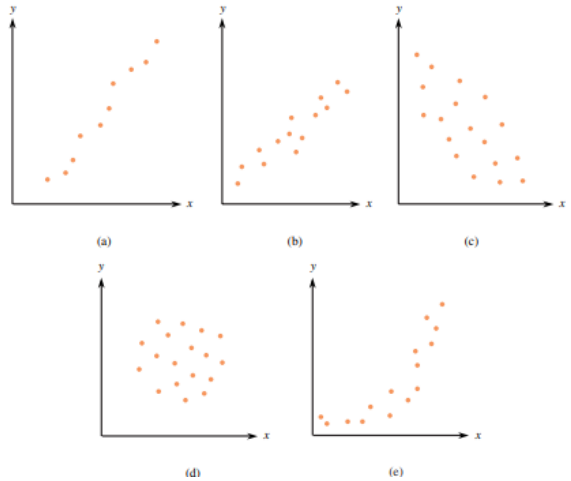
Izvor: Izrada autorice

Međutim, prije samog ispitivanja modela na ovim podacima, napravit će se još dodatne analize. Kako bi se u prediktivnom modelu izbjegle moguće pogreške u procjeni parametara modela, ispitano je postoji li korelacija između varijabli.

## KORELACIJA

**Koeficijent korelacije** je numerička mjera kojom se iskazuje povezanost između dvije varijable. Primjer korelacije između dvije varijable prikazan je na *Slika 4.6*:

a) i b) Prikaz pozitivne korelacije između varijabli x i y
c) Prikaz negativne korelacije između varijabli x i y
d) Ne postoji korelacija između varijabli x i y
e) Postoji korelacija između varijabli x i y, međutim ona nije linearna




***Slika 4.6*** *Prikaz korelacije između dvije varijable*
Izvor: Introduction to Statistics and Data Analysis, stranica:300 [20]

Za dvije slučajne varijable X i Y korelacija se definira [21]:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.1}$$

Gdje je *Cov(X,Y)* kovarijanca između slučajnih varijabli X i y i definira se kao:

$$Cov(X,Y) = E\big[\big(x - E(X)\big)\big(Y - E(Y)\big)\big] \tag{4.2}$$

Za korelaciju vrijedi:

$$-1 \leq Corr(X,Y) \leq 1 \tag{4.3}$$

Ukoliko su X i Y nezavisne slučajne varijable, tada je

$$E(X \cdot Y) = E(X) \cdot E(Y), \tag{4.4}$$

Pa tada vrijedi:

$$Cov(X,Y) = E\big[\big(x - E(X)\big)\big(Y - E(Y)\big)\big] = E[(X - E(X)] \cdot E[(Y - E(Y)] = 0 \cdot 0 = 0 \tag{4.5}$$

Isto vrijedi i za korelaciju:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0 \tag{4.6}$$

Najpoznatiji koeficijent korelacije je ***Pearsonov* koeficijent** korelacije i on mjeri linearnu povezanost dvije varijable.

Ukoliko promatramo dva obilježja X i Y u populaciji veličine N, *Pearsonov* koeficijent korelacije obilježja X i Y definiran je kao [21]:

$$\rho = \frac{\frac{1}{N}\sum(x_i - \mu_x)(y_i - \mu_y)}{\sigma_X \sigma_Y} \tag{4.7}$$

Gdje je:

- $\sigma_X$- standardna devijacija obilježja X
- $\sigma_y$ – standardna devijacija obilježja Y
- $\mu_x$ – aritmetička sredina obilježja X
- $\mu_y$ – aritmetička sredina obilježja Y

Za testiranje *Pearsonovog* koeficijenta korelacije korištena je gotova metoda definirana u Pythonu *pandas.DataFrame.corr* za testiranje koeficijenta korelacije između dvije varijable [22].

Prije ispitivanja korelacije, kategorijskim varijablama koje su imale vrijednosti u tekstualnom obliku (*DOBA_DANA*, *CJENIK_SIFRA_5*, *NAZIV_DRZAVA_DOGADJAJA*, *ZUPANIJA_DOGADJAJA*, *SPOL_OSTECENE_OSOBE*, *TIP_OSTECENE_OSOBE*, *UZROK*, *MARKA_OSTECENOG_VOZILA* i *RANG_STETE)* vrijednosti su zamijenjene prirodnim brojevima sa vrijednostima u rangu [0, broj klasa-1] koristeći gotovu metodu iz Pythona *sklearn.preprocessing.LabelEncoder* [23].

***Tablica 4.4*** *Ispitivanje korelacije između varijabli*

| | CJENIK_SIFRA_5 | DOBA_DANA | NAZIV_DRZAVA_DOGADJAJA | SPOL_OSTECENE_OSOBE | TIP_OSTECENE_OSOBE | UZROK | MARKA_OSTECENOG_VOZILA | RANG_STETE | REGIJA | BR_DN_NAST_POC_PREM | BR_DN_NAST_KRAJ_PREM | BR_DN_NAST_PRI | BROJ_STETA_PO_ISTOJ_POLICI | STAROST_OSTECENIKA | STAROST_VOZILA | BRZA_NAGODBA | HITNA_POMOC | POLICIJA | VATROGASCI | ODVJETNIK_OSTECENOG_IND | OSTECENA_OSOBA_U_SDT | OSTECENA_IMOVINA_U_SDT | RAC | TOTALNA_STETA | LEASING | STRANKA_VRSI_PRITISAK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CJENIK_SIFRA_5 | 1,00 | -0,10 | 0,00 | 0,32 | -0,34 | 0,61 | -0,06 | 0,15 | 0,07 | -0,21 | 0,24 | -0,08 | -0,02 | 0,23 | 0,33 | 0,04 | 0,23 | 0,28 | 0,00 | 0,07 | 0,18 | 0,05 | -0,03 | 0,25 | -0,28 | 0,01 |
| DOBA_DANA | -0,10 | 1,00 | 0,02 | -0,01 | 0,01 | -0,12 | 0,01 | 0,00 | 0,00 | 0,02 | -0,03 | 0,03 | 0,01 | -0,02 | 0,01 | -0,01 | 0,02 | 0,01 | 0,00 | 0,00 | -0,02 | 0,00 | 0,02 | -0,01 | 0,02 | 0,01 |
| NAZIV_DRZAVA_DOGADJAJA | 0,00 | 0,02 | 1,00 | -0,04 | 0,04 | 0,00 | 0,00 | 0,05 | 0,14 | -0,02 | 0,01 | 0,03 | 0,00 | -0,06 | 0,10 | -0,01 | 0,11 | 0,01 | 0,00 | -0,01 | 0,00 | 0,02 | 0,00 | -0,01 | 0,00 | 0,00 |
| SPOL_OSTECENE_OSOBE | 0,32 | -0,01 | -0,04 | 1,00 | -0,91 | 0,13 | -0,07 | 0,02 | 0,05 | 0,03 | 0,00 | -0,11 | -0,13 | 0,81 | 0,18 | 0,20 | 0,07 | 0,09 | 0,00 | 0,03 | 0,07 | 0,00 | -0,06 | 0,14 | -0,58 | 0,02 |
| TIP_OSTECENE_OSOBE | -0,34 | 0,01 | 0,04 | -0,91 | 1,00 | -0,14 | 0,07 | -0,02 | -0,05 | -0,05 | 0,00 | 0,12 | 0,15 | -0,86 | -0,19 | -0,21 | -0,07 | -0,09 | 0,00 | -0,03 | -0,07 | 0,00 | 0,06 | -0,15 | 0,66 | -0,02 |
| UZROK | 0,61 | -0,12 | 0,00 | 0,13 | -0,14 | 1,00 | -0,05 | 0,24 | 0,04 | -0,14 | 0,16 | -0,04 | -0,03 | 0,09 | 0,21 | 0,05 | 0,19 | 0,13 | 0,00 | 0,05 | 0,15 | 0,05 | -0,01 | 0,21 | -0,14 | 0,01 |
| MARKA_OSTECENOG_VOZILA | -0,06 | 0,01 | 0,00 | -0,07 | 0,07 | -0,05 | 1,00 | -0,06 | -0,01 | 0,01 | -0,02 | 0,03 | 0,04 | -0,05 | -0,02 | -0,01 | -0,02 | -0,02 | 0,00 | 0,00 | 0,00 | -0,01 | 0,00 | -0,01 | 0,07 | -0,01 |
| RANG_STETE | 0,15 | 0,00 | 0,05 | 0,02 | -0,02 | 0,24 | -0,06 | 1,00 | 0,06 | -0,12 | 0,14 | -0,07 | -0,06 | -0,02 | 0,12 | -0,14 | 0,19 | 0,31 | 0,01 | 0,07 | 0,18 | 0,09 | 0,01 | 0,28 | -0,02 | 0,02 |
| REGIJA | 0,07 | 0,00 | 0,14 | 0,05 | -0,05 | 0,04 | -0,01 | 0,06 | 1,00 | -0,02 | 0,02 | 0,00 | -0,10 | 0,01 | 0,12 | 0,04 | 0,15 | 0,03 | 0,00 | 0,01 | 0,00 | 0,02 | 0,02 | 0,03 | -0,06 | -0,01 |
| BR_DN_NAST_POC_PREM | -0,21 | 0,02 | -0,02 | 0,03 | -0,05 | -0,14 | 0,01 | -0,12 | -0,02 | 1,00 | -0,97 | -0,07 | -0,03 | 0,06 | -0,06 | 0,07 | -0,05 | -0,11 | 0,00 | -0,02 | -0,05 | -0,03 | -0,02 | -0,07 | 0,01 | 0,01 |
| BR_DN_NAST_KRAJ_PREM | 0,24 | -0,03 | 0,01 | 0,00 | 0,00 | 0,16 | -0,02 | 0,14 | 0,02 | -0,97 | 1,00 | 0,07 | 0,00 | -0,02 | 0,07 | -0,06 | 0,06 | 0,13 | 0,00 | 0,02 | 0,06 | 0,03 | 0,02 | 0,08 | -0,03 | 0,00 |
| BR_DN_NAST_PRI | -0,08 | 0,03 | 0,03 | -0,11 | 0,12 | -0,04 | 0,03 | -0,07 | 0,00 | -0,07 | 0,07 | 1,00 | 0,12 | -0,10 | 0,02 | -0,03 | 0,07 | 0,00 | 0,02 | 0,17 | 0,02 | 0,01 | 0,03 | -0,06 | 0,10 | 0,00 |
| BROJ_STETA_PO_ISTOJ_POLICI | -0,02 | 0,01 | 0,00 | -0,13 | 0,15 | -0,03 | 0,04 | -0,06 | -0,10 | -0,03 | 0,00 | 0,12 | 1,00 | -0,12 | -0,01 | -0,05 | -0,02 | 0,01 | 0,00 | 0,02 | 0,06 | 0,02 | 0,00 | 0,00 | 0,09 | 0,00 |
| STAROST_OSTECENIKA | 0,23 | -0,02 | -0,06 | 0,81 | -0,86 | 0,09 | -0,05 | -0,02 | 0,01 | 0,06 | -0,02 | -0,10 | -0,12 | 1,00 | 0,09 | 0,19 | -0,06 | 0,06 | 0,00 | 0,02 | 0,05 | 0,00 | -0,05 | 0,13 | -0,55 | 0,01 |
| STAROST_VOZILA | 0,33 | 0,01 | 0,10 | 0,18 | -0,19 | 0,21 | -0,02 | 0,12 | 0,12 | -0,06 | 0,07 | 0,02 | -0,01 | 0,09 | 1,00 | 0,02 | 0,53 | 0,10 | 0,01 | 0,05 | 0,08 | 0,04 | -0,02 | 0,14 | -0,22 | 0,00 |
| BRZA_NAGODBA | 0,04 | -0,01 | -0,01 | 0,20 | -0,21 | 0,05 | -0,01 | -0,14 | 0,04 | 0,07 | -0,06 | -0,03 | -0,05 | 0,19 | 0,02 | 1,00 | -0,04 | -0,10 | 0,00 | -0,01 | -0,03 | -0,02 | -0,01 | -0,03 | -0,15 | -0,02 |
| HITNA_POMOC | 0,23 | 0,02 | 0,11 | 0,07 | -0,07 | 0,19 | -0,02 | 0,19 | 0,15 | -0,05 | 0,06 | 0,07 | -0,02 | -0,06 | 0,53 | -0,04 | 1,00 | 0,09 | 0,01 | 0,05 | 0,04 | 0,04 | -0,01 | 0,00 | -0,09 | -0,01 |
| POLICIJA | 0,28 | 0,01 | 0,01 | 0,09 | -0,09 | 0,13 | -0,02 | 0,31 | 0,03 | -0,11 | 0,13 | 0,00 | 0,01 | 0,06 | 0,10 | -0,10 | 0,09 | 1,00 | 0,02 | 0,08 | 0,29 | 0,14 | 0,00 | 0,31 | -0,08 | -0,01 |
| VATROGASCI | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,01 | 0,00 | 0,00 | 0,00 | 0,02 | 0,00 | 0,00 | 0,01 | 0,00 | 0,01 | 0,02 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,02 | 0,00 | 0,00 |
| ODVJETNIK_OSTECENOG_IND | 0,07 | 0,00 | -0,01 | 0,03 | -0,03 | 0,05 | 0,00 | 0,07 | 0,01 | -0,02 | 0,02 | 0,17 | 0,02 | 0,02 | 0,05 | -0,01 | 0,05 | 0,08 | 0,00 | 1,00 | 0,15 | 0,04 | 0,00 | 0,06 | -0,03 | 0,00 |
| OSTECENA_OSOBA_U_SDT | 0,18 | -0,02 | 0,00 | 0,07 | -0,07 | 0,15 | 0,00 | 0,18 | 0,00 | -0,05 | 0,06 | 0,02 | 0,06 | 0,05 | 0,08 | -0,03 | 0,04 | 0,29 | 0,00 | 0,15 | 1,00 | 0,12 | 0,00 | 0,30 | -0,05 | 0,00 |
| OSTECENA_IMOVINA_U_SDT | 0,05 | 0,00 | 0,02 | 0,00 | 0,00 | 0,05 | -0,01 | 0,09 | 0,02 | -0,03 | 0,03 | 0,01 | 0,02 | 0,00 | 0,04 | -0,02 | 0,04 | 0,14 | 0,00 | 0,04 | 0,12 | 1,00 | 0,00 | 0,11 | -0,01 | 0,00 |
| RAC | -0,03 | 0,02 | 0,00 | -0,06 | 0,06 | -0,01 | 0,00 | 0,01 | 0,02 | -0,02 | 0,02 | 0,03 | 0,00 | -0,05 | -0,02 | -0,01 | -0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | -0,01 | -0,03 | 0,00 |
| TOTALNA_STETA | 0,25 | -0,01 | -0,01 | 0,14 | -0,15 | 0,21 | -0,01 | 0,28 | 0,03 | -0,07 | 0,08 | -0,06 | 0,00 | 0,13 | 0,14 | -0,03 | 0,00 | 0,31 | 0,02 | 0,06 | 0,30 | 0,11 | -0,01 | 1,00 | -0,11 | 0,01 |
| LEASING | -0,28 | 0,02 | 0,00 | -0,58 | 0,66 | -0,14 | 0,07 | -0,02 | -0,06 | 0,01 | -0,03 | 0,10 | 0,09 | -0,55 | -0,22 | -0,15 | -0,09 | -0,08 | 0,00 | -0,03 | -0,05 | -0,01 | -0,03 | -0,11 | 1,00 | -0,01 |
| STRANKA_VRSI_PRITISAK | 0,01 | 0,01 | 0,00 | 0,02 | -0,02 | 0,01 | -0,01 | 0,02 | -0,01 | 0,01 | 0,00 | 0,00 | 0,00 | 0,01 | 0,00 | -0,02 | -0,01 | -0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,01 | -0,01 | 1,00 |

Izvor: Izrada autorice

Korelacijska matrica varijabli prikazana je u *Tablica 4.4*, a iz rezultata je vidljivo da postoji korelacija između idućih varijabli:

- **BR_DN_NAST_POC_PREM** i **BR_DN_NAST_KRAJ_PREM** – koeficijent korelacije je -0,97, što je očekivano obzirom da se u uzorku nalaze samo štete prijavljene po jednogodišnjim policama pa broj dana od kad je počela premija do nastanka štete i broj dana od nastanka štete do kraja premije nose istu informaciju, te će se dalje u modelu koristiti samo varijabla BR_DN_NAST_POC_PREM.
- **TIP_OSTECENE_OSOBE** i **SPOL_OSTECENE_OSOBE** – koeficijent korelacije je -0,91. Obzirom da je spol definiran samo za fizičke osobe, dok za pravne i obrtnike nije očekivan je veći koeficijent korelacije. U ovom slučaju definirana je nova varijabla *SPOL_TIP* koja je kombinacija varijabli TIP_OSTECENE_OSOBE i

SPOL_OSTECENE_OSOBE. Vrijednosti koje može poprimiti nova varijabla definirane su u *Tablica 4.5.*

***Tablica 4.5*** *Moguće vrijednosti varijable SPOL_TIP*

| Vrijednost | Opis |
|---|---|
| P | Pravna osoba |
| F2 | Fizička osoba, Muško |
| F1 | Fizička osoba, Žensko |
| O | Obrtnik |
| F0 | Fizička osoba, Nedefinirano |

Izvor: Izrada autorice

- **STAROST_OSTECENIKA** i **SPOL_OSTECENE_OSOBE** – koeficijent korelacije je 0,81, što se isto može objasniti na način da za pravne osobe i obrtnike ne postoji podatak o spolu, a u analizi podataka pravnim osobama i obrtnicima stavljena je starost 0 obzirom da podatak o rođenju nema smisla, a u uzorku ne postoji podatak o osnivanju pravnog subjekta. Varijabla SPOL_OSTECENE_OSOBE bit će zamijenjena sa varijablom *SPOL_TIP* kao što je navedeno u prethodnom koraku pa se neće poduzimati dodatne akcije zbog velikog koeficijenta korelacije ove dvije varijable.
- **STAROST_OSTECENIKA** i **TIP_OSTECENE_OSOBE** – koeficijent korelacije iznosi -0,86, te je obrazloženje analogno onom za korelaciju STAROST_OSTECENIKA i SPOL_OSTECENE_OSOBE
- **CJENIK_SIFRA_5** i **UZROK** – koeficijent korelacije je 0,61. Kako je bilo spomenuto kod zasebne analize svake varijable, uzrok se evidentira samo za auto kasko štete, dok je za štete automobilske odgovornosti ova vrijednost prazna.
- **LEASING i STAROST_OSTECENIKA** – koeficijent korelacije je -0,55. Slično kao i u prethodnim primjerima - leasing je pravna osoba, a za njih je definirano da je starost 0. Obzirom da će varijabla SPOL_OSTECENE_OSOBE biti zamijenjena s varijablom SPOL_TIP neće se poduzimati dodatne akcije zbog velike korelacije između ove dvije varijable.
- **LEASING i TIP_OSTECENE_OSOBE** – koeficijent korelacije je 0,66. Obzorom da su leasing kuće u pravilu pravne osobe, očekivan je ovaj koeficijent korelacije. Obzirom da će varijabla TIP_OSTECENE_OSOBE biti zamijenjena s

varijablom SPOL_TIP neće se poduzimati dodatne akcije zbog velike korelacije ove dvije varijable.

Nakon predloženih dorada, ponovno je ispitana korelacija između varijabli te je matrica rezultata prikazana u *Tablica 4.6.* I dalje postoji veća korelacija između varijabli STAROST_OSTECENIKA i SPOL_TIP (0,82), LEASING i SPOL_TIP (0,63) te UZROK i CJENIK_SIFRA_5 (0,61) međutim neće se raditi dodatne akcije zbog spomenutih korelacija.

***Tablica 4.6*** *Ispitivanje korelacije između varijabli nakon određenih transformacija*

| | CJENIK_SIFRA_5 | DOBA_DANA | NAZIV_DRZAVA_DOGADJAJA | SPOL_TIP | UZROK | MARKA_OSTECENOG_VOZILA | RANG_STETE | REGIJA | BR_DN_NAST_POC_PREM | BR_DN_NAST_PRI | BROJ_STETA_PO_ISTOJ_POLICI | STAROST_OSTECENIKA | STAROST_VOZILA | BRZA_NAGODBA | HITNA_POMOC | POLICIJA | VATROGASCI | ODVJETNIK_OSTECENOG_IND | OSTECENA_OSOBA_U_SDT | OSTECENA_IMOVINA_U_SDT | RAC | TOTALNA_STETA | LEASING | STRANKA_VRSI_PRITISAK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CJENIK_SIFRA_5 | 1,00 | -0,10 | 0,00 | -0,32 | 0,61 | -0,06 | 0,15 | 0,07 | -0,21 | -0,08 | -0,02 | 0,23 | 0,33 | 0,04 | 0,23 | 0,28 | 0,00 | 0,07 | 0,18 | 0,05 | -0,03 | 0,25 | -0,28 | 0,01 |
| DOBA_DANA | -0,10 | 1,00 | 0,02 | 0,01 | -0,12 | 0,01 | 0,00 | 0,00 | 0,02 | 0,03 | 0,01 | -0,02 | 0,01 | -0,01 | 0,02 | 0,01 | 0,00 | 0,00 | -0,02 | 0,00 | 0,02 | -0,01 | 0,02 | 0,01 |
| NAZIV_DRZAVA_DOGADJAJA | 0,00 | 0,02 | 1,00 | 0,04 | 0,00 | 0,00 | 0,05 | 0,14 | -0,02 | 0,03 | 0,00 | -0,06 | 0,10 | -0,01 | 0,11 | 0,01 | 0,00 | -0,01 | 0,00 | 0,02 | 0,00 | -0,01 | 0,00 | 0,00 |
| SPOL_TIP | -0,32 | 0,01 | 0,04 | 1,00 | -0,14 | 0,06 | -0,01 | -0,04 | -0,05 | 0,11 | 0,14 | -0,82 | -0,17 | -0,20 | -0,06 | -0,08 | 0,00 | -0,02 | -0,07 | 0,00 | 0,06 | -0,13 | 0,63 | -0,02 |
| UZROK | 0,61 | -0,12 | 0,00 | -0,14 | 1,00 | -0,05 | 0,24 | 0,04 | -0,14 | -0,04 | -0,03 | 0,09 | 0,21 | 0,05 | 0,19 | 0,13 | 0,00 | 0,05 | 0,15 | 0,05 | -0,01 | 0,21 | -0,14 | 0,01 |
| MARKA_OSTECENOG_VOZILA | -0,06 | 0,01 | 0,00 | 0,06 | -0,05 | 1,00 | -0,06 | -0,01 | 0,01 | 0,03 | 0,04 | -0,05 | -0,02 | -0,01 | -0,02 | -0,02 | 0,00 | 0,00 | 0,00 | -0,01 | 0,00 | -0,01 | 0,07 | -0,01 |
| RANG_STETE | 0,15 | 0,00 | 0,05 | -0,01 | 0,24 | -0,06 | 1,00 | 0,06 | -0,12 | -0,07 | -0,06 | -0,02 | 0,12 | -0,14 | 0,19 | 0,31 | 0,01 | 0,07 | 0,18 | 0,09 | 0,01 | 0,28 | -0,02 | 0,02 |
| REGIJA | 0,07 | 0,00 | 0,14 | -0,04 | 0,04 | -0,01 | 0,06 | 1,00 | -0,02 | 0,00 | -0,10 | 0,01 | 0,12 | 0,04 | 0,15 | 0,03 | 0,00 | 0,01 | 0,00 | 0,02 | 0,02 | 0,03 | -0,06 | -0,01 |
| BR_DN_NAST_POC_PREM | -0,21 | 0,02 | -0,02 | -0,05 | -0,14 | 0,01 | -0,12 | -0,02 | 1,00 | -0,07 | -0,03 | 0,06 | -0,06 | 0,07 | -0,05 | -0,11 | 0,00 | -0,02 | -0,05 | -0,03 | -0,02 | -0,07 | 0,01 | 0,01 |
| BR_DN_NAST_PRI | -0,08 | 0,03 | 0,03 | 0,11 | -0,04 | 0,03 | -0,07 | 0,00 | -0,07 | 1,00 | 0,12 | -0,10 | 0,02 | -0,03 | 0,07 | 0,00 | 0,02 | 0,17 | 0,02 | 0,01 | 0,03 | -0,06 | 0,10 | 0,00 |
| BROJ_STETA_PO_ISTOJ_POLICI | -0,02 | 0,01 | 0,00 | 0,14 | -0,03 | 0,04 | -0,06 | -0,10 | -0,03 | 0,12 | 1,00 | -0,12 | -0,01 | -0,05 | -0,02 | 0,01 | 0,00 | 0,02 | 0,06 | 0,02 | 0,00 | 0,00 | 0,09 | 0,00 |
| STAROST_OSTECENIKA | 0,23 | -0,02 | -0,06 | -0,82 | 0,09 | -0,05 | -0,02 | 0,01 | 0,06 | -0,10 | -0,12 | 1,00 | 0,09 | 0,19 | -0,06 | 0,06 | 0,00 | 0,02 | 0,05 | 0,00 | -0,05 | 0,13 | -0,55 | 0,01 |
| STAROST_VOZILA | 0,33 | 0,01 | 0,10 | -0,17 | 0,21 | -0,02 | 0,12 | 0,12 | -0,06 | 0,02 | -0,01 | 0,09 | 1,00 | 0,02 | 0,53 | 0,10 | 0,01 | 0,05 | 0,08 | 0,04 | -0,02 | 0,14 | -0,22 | 0,00 |
| BRZA_NAGODBA | 0,04 | -0,01 | -0,01 | -0,20 | 0,05 | -0,01 | -0,14 | 0,04 | 0,07 | -0,03 | -0,05 | 0,19 | 0,02 | 1,00 | -0,04 | -0,10 | 0,00 | -0,01 | -0,03 | -0,02 | -0,01 | -0,03 | -0,15 | -0,02 |
| HITNA_POMOC | 0,23 | 0,02 | 0,11 | -0,06 | 0,19 | -0,02 | 0,19 | 0,15 | -0,05 | 0,07 | -0,02 | -0,06 | 0,53 | -0,04 | 1,00 | 0,09 | 0,01 | 0,05 | 0,04 | 0,04 | -0,01 | 0,00 | -0,09 | -0,01 |
| POLICIJA | 0,28 | 0,01 | 0,01 | -0,08 | 0,13 | -0,02 | 0,31 | 0,03 | -0,11 | 0,00 | 0,01 | 0,06 | 0,10 | -0,10 | 0,09 | 1,00 | 0,02 | 0,08 | 0,29 | 0,14 | 0,00 | 0,31 | -0,08 | -0,01 |
| VATROGASCI | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,01 | 0,00 | 0,00 | 0,02 | 0,00 | 0,00 | 0,01 | 0,00 | 0,01 | 0,02 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,02 | 0,00 | 0,00 |
| ODVJETNIK_OSTECENOG_IND | 0,07 | 0,00 | -0,01 | -0,02 | 0,05 | 0,00 | 0,07 | 0,01 | -0,02 | 0,17 | 0,02 | 0,02 | 0,05 | -0,01 | 0,05 | 0,08 | 0,00 | 1,00 | 0,15 | 0,04 | 0,00 | 0,06 | -0,03 | 0,00 |
| OSTECENA_OSOBA_U_SDT | 0,18 | -0,02 | 0,00 | -0,07 | 0,15 | 0,00 | 0,18 | 0,00 | -0,05 | 0,02 | 0,06 | 0,05 | 0,08 | -0,03 | 0,04 | 0,29 | 0,00 | 0,15 | 1,00 | 0,12 | 0,00 | 0,30 | -0,05 | 0,00 |
| OSTECENA_IMOVINA_U_SDT | 0,05 | 0,00 | 0,02 | 0,00 | 0,05 | -0,01 | 0,09 | 0,02 | -0,03 | 0,01 | 0,02 | 0,00 | 0,04 | -0,02 | 0,04 | 0,14 | 0,00 | 0,04 | 0,12 | 1,00 | 0,00 | 0,11 | -0,01 | 0,00 |
| RAC | -0,03 | 0,02 | 0,00 | 0,06 | -0,01 | 0,00 | 0,01 | 0,02 | -0,02 | 0,03 | 0,00 | -0,05 | -0,02 | -0,01 | -0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | -0,01 | -0,03 | 0,00 |
| TOTALNA_STETA | 0,25 | -0,01 | -0,01 | -0,13 | 0,21 | -0,01 | 0,28 | 0,03 | -0,07 | -0,06 | 0,00 | 0,13 | 0,14 | -0,03 | 0,00 | 0,31 | 0,02 | 0,06 | 0,30 | 0,11 | -0,01 | 1,00 | -0,11 | 0,01 |
| LEASING | -0,28 | 0,02 | 0,00 | 0,63 | -0,14 | 0,07 | -0,02 | -0,06 | 0,01 | 0,10 | 0,09 | -0,55 | -0,22 | -0,15 | -0,09 | -0,08 | 0,00 | -0,03 | -0,05 | -0,01 | -0,03 | -0,11 | 1,00 | -0,01 |
| STRANKA_VRSI_PRITISAK | 0,01 | 0,01 | 0,00 | -0,02 | 0,01 | -0,01 | 0,02 | -0,01 | 0,01 | 0,00 | 0,00 | 0,01 | 0,00 | -0,02 | -0,01 | -0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,01 | -0,01 | 1,00 |

Izvor: Izrada autorice

Korištenje velikog broja značajki u modelu doprinosi njegovoj složenosti i može dovesti do prevelikog prilagođavanja modela podacima za učenje (engl.*overfitting)*. Neke od tih značajki mogu stvoriti šum (engl. *noise*) te potencijalno uzrokovati lošiji model, zbog toga micanje takvih značajki bolje generalizira model [25].

U ovom radu je za odabir **kvalitativnih varijabli** korištena **univarijatna analiza varijabli**[21]. Metoda univarijatne analize značajki kod odabira varijabli svaku varijablu uspoređuje sa zavisnom varijablom kako bi se ispitalo postoji li statistički značajna veza između njih. Kod ispitivanja veza između nezavisnih i zavisnih varijabli uvijek se ispituje veza između jedne nezavisne varijable i zavisne varijable i pri tome svaka varijabla ostvari određene bodove (engl. *score*) te se vrijednosti međusobno uspoređuju i one sa najviše ostvarenih bodova se koriste u modelu.
Za ispitivanje univarijatne analize kvalitativnih varijabli korišten je Hi-kvadrat test [24].

**Definicija Hi-kvadrat**

Nulta hipoteza $H_0$: Dvije varijable su nezavisne

Alternativna hipoteza $H_a$: Varijable su zavisne

Testna statistika:

$$X^2 = \sum \frac{(f_o - f_t)^2}{f_t} \tag{4.8}$$

Gdje su :

$f_o$- opažene frekvencije

$f_t$- očekivane (teoretske) frekvencija

Pretpostavke za korištenje Hi-kvadrat testa:

- Opservacija su iz slučajnog uzorka
- Testiranje se provodi na velikom uzorku podataka

U Pythonu postoji već gotova metoda *sklearn.feature_selection.chi2(X, y*) koja se može koristiti za odabir *n* varijabli koje imaju najveće vrijednosti $X^2$ statistike[26]. Pretpostavka je da su u X sve vrijednosti pozitivne (na primjer 1- 0 vrijednosti ili frekvencije) .
Rezultati univarijatne analize kvalitativnih varijabli nalaze se u *Tablica 4.7.*

[21] engl. *Univariate feature selection*

***Tablica 4.7*** *Rezultati univarijatne analize kvalitativnih varijabli*

| R.Br. | Varijabla | Dodijeljen broj bodova[22] |
|---|---|---|
| 1 | STRANKA_VRSI_PRITISAK | 13.703,86 |
| 2 | CJENIK_SIFRA_5 | 638,19 |
| 3 | BRZA_NAGODBA | 326,89 |
| 4 | UZROK | 175,25 |
| 5 | TOTALNA_STETA | 92,39 |
| 6 | HITNA_POMOC | 74,28 |
| 7 | LEASING | 54,83 |
| 8 | REGIJA | 54,29 |
| 9 | RANG_STETE | 31,28 |
| 10 | SPOL_TIP | 23,72 |
| 11 | MARKA_OSTECENOG_VOZILA | 17,47 |
| 12 | VATROGASCI | 8,79 |
| 13 | RAC | 5,53 |
| 14 | POLICIJA | 2,29 |
| 15 | NAZIV_DRZAVA_DOGADJAJA | 2,80 |
| 16 | ODVJETNIK_OSTECENOG_IND | 2,69 |
| 17 | OSTECENA_OSOBA_U_SDT | 0,14 |
| 18 | OSTECENA_IMOVINA_U_SDT | 0,08 |
| 19 | DOBA_DANA | 0,01 |

Izvor: Izrada autorice

Varijable OSTECENA_OSOBA_U_SDT, OSTECENA_IMOVINA_U_SDT i DOBA_DANA neće se koristiti u prediktivnom modelu obzirom da mu doprinose manje od 1, dok će se kvantitativne varijable koristiti sve (ima ih samo pet).

U analizi korelacije i univarijatne analize varijabli kvalitativnim varijablama koje nisu binomne korištenjem tehnike *sklearn.preprocessing.LabelEncoder* [23] kategorije su zamijenjene cijelim brojevima u rangu [0,broj kategorija -1]. Međutim, obzirom da unutar kategorija se ne može odrediti poredak niti je jednaka udaljenost vrijednosti među svim kategorijama, u izradi modela za takve kategorijske varijable stvorit će se tzv. „*dummy*“ ili „*one hot*“ varijable, odnosno jedna varijabla za svaku kategoriju. Svaka varijabla je binarna numerička varijabla koja poprima vrijednost 1 za štete za koje vrijedi ta kategorija, odnosno vrijednost 0 za štete za koje ne vrijedi ta kategorija [27].

[22] engl. *score*

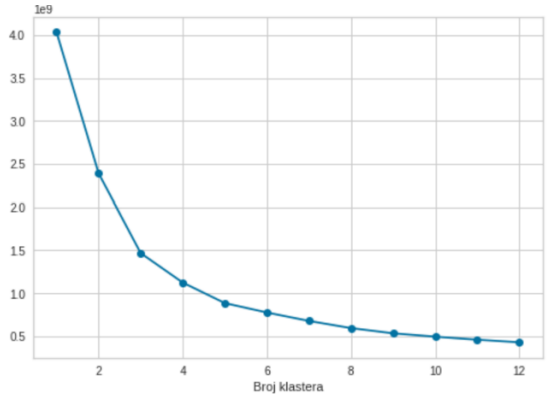
U dostupnom uzorku podataka, štete su klasificirane u dvije kategorije – prijevarna ili nije prijevarna. U praksi, ukoliko neki algoritam štetu automatizmom označi da je prijevarna, stručna osoba treba provjeriti sve dostupne podatke vezane uz nastanak štetnog događaja te prikupiti dokaze koji pokazuju da je šteta doista prijevarna. Obzirom da je samo dokazivanje prijevare izuzetno kompleksan posao koji zahtjeva značaji angažman, najoptimalnije bi bilo da:

- algoritam kao prijevarne označi samo one štete koje to doista jesu jer se time smanjuje trošak resursa na dodatno istraživanje šteta koje nisu prijevarne
- sve prijevarne štete budu označene kao prijevarne i da osiguravajuća društva doista uspiju detektirati sve njih.

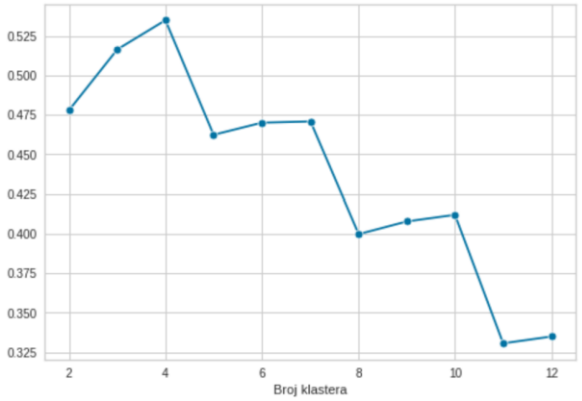
Ovo je idealan scenarij, koji je poprilično nerealan. Također, ukoliko algoritam označi veliki broj šteta prijevarnima, ponekad je teško napraviti prioritizaciju obrade. Možda bi klasifikacija šteta u 3 ili 4 klase (npr. velika vjerojatnost prijevare, potencijalna prijevara, mala vjerojatnost prijevare, nije prijevara) olakšala određivanje prioriteta obrade u dokazivanju da je neka šteta prijevarna. Upravo zbog tog razloga, na dostupnom uzorku podataka ispitana je metoda nenadziranog učenja *k-means* [28] kako bi se odredio optimalan broja klastera u koje se mogu podijeliti dostupni podaci iz uzorka. Za određivanje optimalnog broja klastera korištene su metoda lakta i metoda siluete.

Prikaz rezultata dobivenih metodom lakta prikazan je na *Slika 4.7,* te se na temelju njih ne može se odrediti je li optimalan broj klastera 3, 4 ili 5 pa je zbog toga implementirana i metoda siluete čiji rezultati su prikazani na *Slika 4.8.* i tu se vidi da je najveći koeficijent siluete (0,53) postignut je u slučaju podjele podataka u četiri klastera.

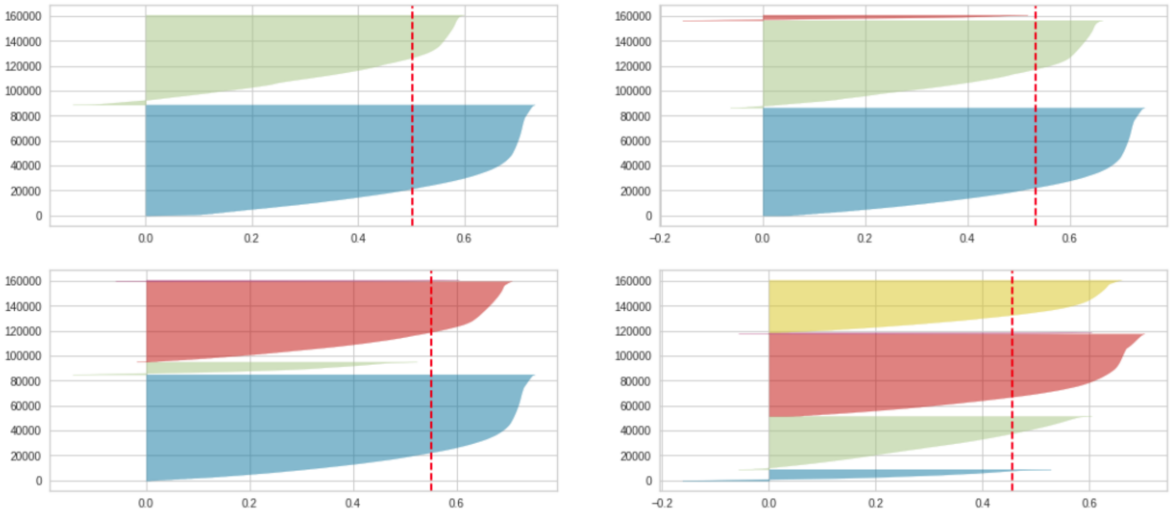
Dodatno je napravljena analiza koeficijenata siluete za podjelu podataka u 2, 3, 4 i 5 klastera kako bi se usporedili koeficijenti siluete svakog klastera sa prosječnim koeficijentom siluete čiji rezultati su prikazani na *Slika 4.9.*




***Slika 4.7*** *Metoda lakta za k-means*
Izvor: Izrada autorice




***Slika 4.8*** *Metoda za k-means*
Izvor: Izrada autorice



***Slika 4.9*** *Analiza metode siluete za 2,3,4,5 klastera*
Izvor: Izrada autorice

Iz analize siluete (*Slika 4.9*) može se zaključiti da podjela podataka u tri i četiri klastera nije dobar odabir obzirom da koeficijent siluete nije za sve klastere iznad ukupne prosječne vrijednosti koeficijenta siluete, odnosno metoda siluete pokazuje da je podjela podataka u 2 ili 5 klastera puno bolji odabir. Koeficijent siluete za podjelu u 2 klastera iznosi 0,48, dok za podjelu u 5 iznosi 0,43, te se prema tome može zaključiti da su dva klastera optimalnija za grupiranje podataka iz uzorka. Obzirom da se u klasteru nalaze podaci koji su slični, provjereno je nalaze li se u jednom klasteru prijevarne štete, a u drugom sve ostale štete, te je isto provjereno za slučaj grupiranja šteta u pet klastera.

Za slučaj grupiranja šteta u dva klastera, u klaster 1 grupirano je 45% šteta, dok je u klaster 2 grupirano 55% šteta, te je unutar klastera gotovo identičan udio prijevarnih i redovnih šteta kao što je ukupan udio šteta u klasteru pa se zbog toga može zaključiti da grupiranje podataka u dva

klastera nije napravljeno zbog sličnih obilježja koja pokazuju da je šteta prijevarna, već zbog nekih drugih elemenata sličnosti.

Na isti način ispitana je klasifikacija šteta u pet klastera (*Tablica 4.9*), te je udio prijevarnih i redovnih šteta unutar klastera jednak udjelu broja šteta koje su raspoređene u isti klaster, dakle ne dobiva se dodatna informacija o sličnosti podataka, a koja je vezana za prijevarne štete, pa u implementaciji prediktivnih metoda neće koristiti dobiveni rezultati.

***Tablica 4.8*** *Usporedba stvarne klasifikacije štete i k-means klasifikacije (2 klastera)*

| | | STVARNA VRIJEDNOST | | |
|---|---|---|---|---|
| | | Prijevara | Nije Prijevara | |
| PREDVIĐENA VRIJEDNOST | Klaster 1 | 43,63% | 44,82% | 44,78% |
| | Klaster 2 | 56,37% | 55,18% | 55,22% |
| | | 100,00% | 100,00% | |

Izvor: Izrada autorice

***Tablica 4.9*** *Usporedba stvarne klasifikacije štete i k-means klasifikacije (5 klastera)*

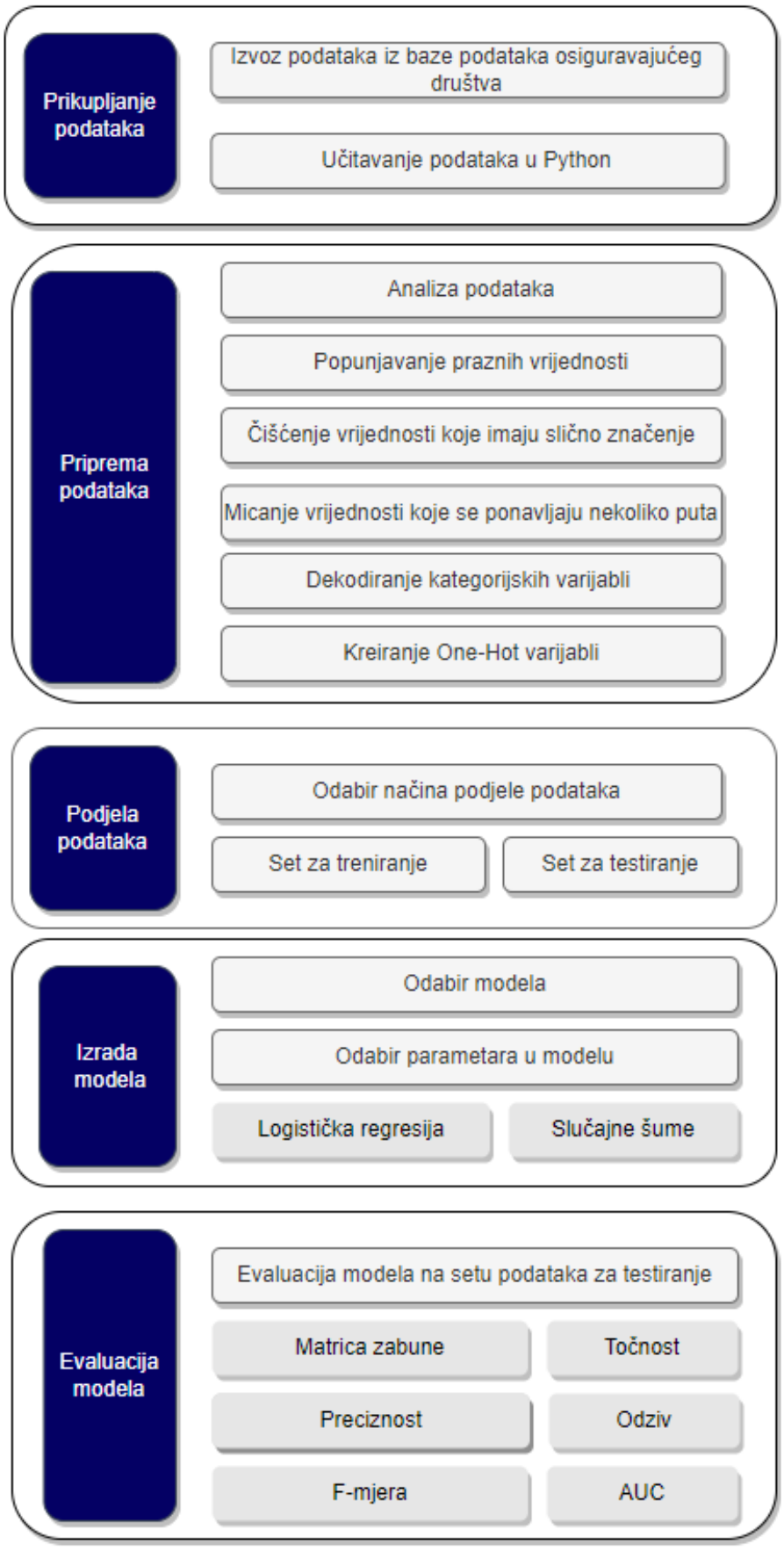
| | | STVARNA VRIJENDOST | | |
|---|---|---|---|---|
| | | Prijevara | Nije Prijevara | |
| PREDVIĐENA VRIJENDOST | Klaster 1 | 41,01% | 42,36% | 41,05% |
| | Klaster 2 | 26,18% | 25,54% | 26,16% |
| | Klaster 3 | 0,62% | 0,63% | 0,62% |
| | Klaster 4 | 26,58% | 24,96% | 26,53% |
| | Klaster 5 | 5,61% | 6,51% | 5,63% |
| | | 100,00% | 100,00% | |

Izvor: Izrada autorice

# 5 REZULTATI

U prethodnim poglavljima opisane su varijable, njihove vrijednosti  i način na koji su one pripremljene za izradu prediktivnog modela.

Prije samog testiranja modela, uzorak podataka podijeljen je na skup podataka za učenje i testiranje u omjeru 70:30 i to na način da su stariji podaci korišteni za učenje modela, a noviji podaci za testiranje.
Implementirane su metode za nadzirano učenje logistička regresija i slučajne šume, a prikaz svih koraka pripreme podataka za izradu modela nalazi se na *Slika 5.1*.




***Slika 5.1*** *Koraci u izradi modela za predikciju*
Izvor: Izrada autorice

## 5.1 Primjena odabranih metoda

U ovom poglavlju prikazani su rezultati primjena odabranih modela na uzorku podataka.
Za ispitivanje su korištene gotove metode implementirane u *Pythonu*.
Za implementaciju **logističke regresiju** korištena je metoda *linear_model.LogisticRegression* unutar paketa *sklearn* [29]. Model je ispitan na šest grupa podataka, a grupe podataka su kreirane na temelju rezultata dobivenih korištenjem testa univarijatne analize kvalitativnih varijabli i to prema dobivenom broju bodova[23] *:*

- prvi skup čine kvalitativne varijable kojima je broj bodova veći od 1,
- drugi skup čine kvalitativne varijable kojima je broj bodova veći od 5,
- treći i peti skup čine kvalitativne varijable kojima je broj bodova veći od 20,
- četvrti i šesti skup čine kvalitativne varijable kojima je broj bodova veći od 70,

s time da su u prva četiri skupa uključene i sve dostupne kvantitativne varijable. Prikaz svih varijabli po skupovima za ispitivanje nalazi se *u Tablica 5.1.*

Za implementaciju **slučajne šume** (engl. *random forest*) korišten je model *ensemble.RandomForestClassifier* koji se nalazi u paketu *sklearn* [30]. U modelu se mogu definirati različiti parametri, a u procesu treniranja tražene su najbolje vrijednosti za iduće parametre:

- *n_estimators*: Označava broj stabala u šumi.
- *criterion*: Funkcija kojom se mjeri kvaliteta podjele varijabli u čvoru.
- *max_depth*: Označava maksimalnu dubinu stabla.
- *max_features*: Označava broj varijabli koje se trebaju uzeti u skup za odabir kod podjele u čvoru. Moguće vrijednosti:
    - Ukoliko je *pozitivan cijeli broj (int),* on označava broj varijabli koje se moraju uzeti u obzir.
    - Ukoliko je decimalan broj *(float), max_features* maksimum od 1 i najvećeg cijelog umnoška upisanog broja i ukupnog broja varijabli.
    - Ukoliko je „*auto*“, *max_features* je drugi korijen od ukupnog broja varijabli.
    - Ukoliko je „sqrt“, *max_features* je drugi korijen od ukupnog broja varijabli.
    - Ukoliko je „log2“, *max_features* je logaritam po bazi 2 od ukupnog broja varijabli.

[23] engl. *score*

- Ako nije ništa navedeno, *max_features* je jednak ukupnom broju varijabli.

***Tablica 5.1*** *Oznaka podskupa varijabli korištenih za ispitivanje Logističke regresije*

| **Oznaka podskupa varijabli** | **X** | **X2** | **X3** | **X4** | **X5** | **X6** |
|---|---|---|---|---|---|---|
| **KVALITATIVNE VARIJABLE** | | | | | | |
| CJENIK_SIFRA_5 | D | D | D | D | D | D |
| UZROK | D | D | D | D | D | D |
| NAZIV_DRZAVA_DOGADJAJA | D | | | | | |
| REGIJA | D | D | D | | D | |
| SPOL_TIP | D | D | D | | D | |
| MARKA_OSTECENOG_VOZILA | D | D | | | | |
| BRZA_NAGODBA | D | D | D | D | D | D |
| HITNA_POMOC | D | D | D | D | D | D |
| POLICIJA | D | | | | | |
| VATROGASCI | D | D | | | | |
| ODVJETNIK_OSTECENOG_IND | D | | | | | |
| RAC | D | D | | | | |
| LEASING | D | D | D | | D | |
| STRANKA_VRSI_PRITISAK | D | D | D | D | D | D |
| TOTALNA_STETA | D | D | D | D | D | D |
| RANG_STETE | D | D | D | | D | |
| **KVANTITATIVNE VARIJABLE** | | | | | | |
| BR_DN_NAST_POC_PREM | D | D | D | D | | |
| BR_DN_NAST_PRI | D | D | D | D | | |
| BROJ_STETA_PO_ISTOJ_POLICI | D | D | D | D | | |
| STAROST_OSTECENIKA | D | D | D | D | | |
| STAROST_VOZILA | D | D | D | D | | |
| **Broj korištenih varijabli** | **21** | **18** | **15** | **11** | **10** | **6** |

Izvor: Izrada autorice

Za validaciju modela korištene su mjere [31]:

1. **Matrica zabune** (engl. *confusion matrix*) uspoređuje klase dobivene kao predikcije modela sa stvarnim klasama primjera. Razlikujemo četiri ishoda klasifikacije:
   - **a. TP** (engl. *true positive*) – ispravna klasifikacija u pozitivnu klasu
   - b. **TN** (engl. *true negative*) – ispravna klasifikacija u negativnu klasu
   - **c. FP** (engl. *false positive*) – neispravna klasifikacija u pozitivnu klasu
   - **d. FN** (engl. *false negative*) – neispravna klasifikacija u negativnu klasu

***Tablica 5.2*** *Matrica zabune (engl. confusion matrix)*

| | | **STVARNA VRIJEDNOST** | |
|---|---|---|---|
| | | **Pozitivna (P)** | **Negativna (N)** |
| **PREDVIĐENA VRIJEDNOST** | **Pozitivna (P)** | Stvarno pozitivna (engl. *true positive*, **TP**) | Lažno pozitivna (engl. *false positive,* **FP**) |
| | **Negativna (N)** | Lažno negativna (engl. *false negative*, **FN**) | Stvarno negativna (engl. *true negative*, **TN**) |

Izvor: Prilagodba autorice prema članku Narkhede S., „Understanding Confusion Matrix“ [31]

2. **Točnost** (engl. *accuracy*) se definira kao omjer ispravno klasificiranih primjera i ukupnog broja primjera:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Kod izuzetno disbalansiranih klasa može se dobiti visoka vrijednost točnosti iako se radi o lošem modelu. To je slučaj kad je udio primjera negativne klase izuzetno visok pa klasifikator svaki novi primjer klasificira kao negativan.

3. **Preciznost** (engl. *precision*) je definiran kao omjer broja primjera koje je klasifikator ispravno označio kao pozitivne i ukupnog broja primjera koje je klasifikator označio kao pozitivne:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5.2}$$

4. **Odziv** (engl. *recall*) je definiran mjerom broja primjera koje je klasifikator označio kao pozitivne i ukupnog broja pozitivnih primjera:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.3}$$

5. **F-mjera** se izračunava kao harmonijska sredina između preciznosti i odziva:

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R} \tag{5.4}$$

6. **ROC** (engl. *Receiver Operatig Characteristic*) krivulja je grafički prikaz sposobnosti klasifikatora u slučaju da se mijenja njegov prag diskriminacije.

7. **AUC** (engl. *Area under the ROC Curve)* odgovara površini ispod ROC krivulje. Što je veća površina, to je bolji model.

Obzirom da se u radu ispituju modeli čiji je zadatak klasificirati štete jesu li prijevarne ili ne, cilj je da implementirani algoritam što manje prijevarnih šteta pogrešno označi da nisu prijevarne, odnosno da udio točno označenih šteta kao „prijevara“ u ukupnom udjelu prijevarnih šteta bude što veći. Taj podatak izražava se mjerom odziv i zbog toga će se ona koristiti za usporedbu implementiranih metoda.

## 5.2 Diskusija rezultata

Predloženi modeli su sukladno njihovim specifičnostima ispitani na nekoliko pristupa. Kod ispitivanja modela **logističke regresije** podaci su podijeljeni na skup za treniranje i skup za testiranje u omjeru **70:30** na način da su stariji podaci iz uzorka korišteni za treniranje, a noviji za testiranje. Prvo je ispitan na skupu varijabli koje su u *Tablica 5.1* označene sa „X“ i to sa četiri različita pristupa:

1. Korištene su varijable na način kao što su pripremljene i opisane u prethodnim poglavljima bez dodatne obrade.
2. Varijable su standardizirane korištenjem implementiranih naredbi u Pythonu *StandardScaler(), RobustScaler(), MinMax(),* a prema [32], [33] i [34]:
   - **StandardScaler** - standardizira vrijednosti tako da od svake vrijednosti oduzme aritmetičku sredinu vrijednosti, te podijeli razliku sa standardnom devijacijom,
   - **RobustScaler** - standardizira vrijednosti tako da od svake vrijednosti oduzme medijan, te podijeli razliku sa IQR[24]-om ,
   - **MinMax** – standardizira vrijednost tako da od svake vrijednosti oduzme minimalnu vrijednost u uzorku i dobiveni rezultat podijeli sa razlikom između minimalne i maksimalne vrijednosti u uzorku.
3. Varijable nisu standardizirane, međutim obzirom da je uzorku značajno veći udio šteta koje nisu prijevarne, odnosno postoji veliki disbalans između klase prijevarnih i

[24] engl. *interquartile range*, računa se kao razlika između trećeg i prvog kvartila

regularnih šteta, korištena je opcija iz Pythona *class_weight='balanced'* koja daje veću težinu klasi koja ima manji udio [35]. Težina razreda[25] računa se na način:

$$w(j) = \frac{n}{K \cdot n(j)} \tag{5.5}$$

Gdje je:

w(j) – težina klase,

n - broj opservacija,

K - ukupan broj klasa,

n(j) - broj opservacija u svakoj klasi

4. Varijable su standardizirane korištenjem gotovih naredbi iz Pythona navedenih u točki 2, te je dodatno korištena opcija *class_weight='balanced'* opisana u točki 3.

Rezultati implementiranih metoda na podacima za testiranje prikazani su u *Tablica 5.3*, te je iz njih vidljivo da model ima veći odziv ukoliko se dodaje veća težina klasi koja ima manji udio i kad je korištena tehnika za standardizaciju varijabli, s time da tip tehnike ne utječe na odziv kao ni na ostale mjere koje su korištene za validaciju modela.

[25] engl. *class weight*

***Tablica 5.3*** *Rezultati logističke regresije na skupu podataka za testiranje*

| **Logistička regresija** | | | | | | | |
|---|---|---|---|---|---|---|---|
| R Br | TEHNIKE KORIŠTENE U MODELU | **Skup podataka za testiranje** | | | | | AUC - na setu za treniranje |
| | | Točnost | Preciznost | Odziv | F mjera | AUC | |
| 1 | - bez korištenja dodatnih tehnika | 0,968 | 0,618 | 0,089 | 0,156 | 0,715 | 0,704 |
| 2 | - Standard Scaler() | 0,969 | 0,627 | 0,111 | 0,189 | 0,724 | 0,727 |
| 3 | - RobustScaler() | 0,969 | 0,627 | 0,111 | 0,189 | 0,724 | 0,727 |
| 4 | - MinMax() | 0,969 | 0,627 | 0,111 | 0,189 | 0,724 | 0,727 |
| 5 | - class weight= 'balanced' | 0,699 | 0,061 | 0,573 | 0,111 | 0,724 | 0,727 |
| 6 | - Standard Scaler()<br>- class weight= 'balanced' | 0,689 | 0,063 | 0,612 | 0,114 | 0,725 | 0,729 |
| 7 | - RobustScaler()<br>- class weight= 'balanced' | 0,689 | 0,063 | 0,612 | 0,114 | 0,725 | 0,729 |
| 8 | - MinMax()<br>- class weight= 'balanced' | 0,689 | 0,063 | 0,612 | 0,114 | 0,725 | 0,729 |

Izvor: Izrada autorice

Dodatno je model logističke regresije ispitan za šest podskupova varijabli koji su definirani obzirom na rezultate univarijatne analize kvalitativnih varijabli *(*u *Tablica 5.1* nalazi se popis varijabli označen po podskupima):

- u prvoj grupu koja je označna sa „X“ nalazi se 21 varijabla,
- drugu grupu označenom kao „*X2*“ čini 18 varijabli,
- treću grupu označenu kao „*X3*“ čini 15 varijabli,
- četvrtu grupu označeno kao „*X4*“ čini 11 varijabli,
- petu grupu označenu kao „*X5*“ čini 10 varijabli,
- šestu grupu označenu kao „*X6*“ čini 6 varijabli.

Uspoređujući rezultate dobivene implementacijom modela logističke regresije za svaku grupu varijabli, najveću vrijednost mjere odziv ima modelu u kojem je korišteno 10 varijabli (označen „X5“), a najmanji odziv u modelu u kojem je korištena 21 varijabla (označen „X“). Temeljem tih rezultata, možemo zaključiti da varijable koje su korištene u prvom modelu, a nisu korištene u petom modelu, ne donose modelu značajne informacije za klasifikaciju šteta je li prijevara ili nije prijevara. Prikaz rezultata implementacije logističke regresije na navedene grupe varijabli se nalazi u *Tablica 5.4*.

***Tablica 5.4*** *Logistička regresija na različitim setovima podataka (korišten je Standard Scaler () i class_weight='balanced' za predprocesiranje podataka)*

| **Logistička regresija** | | | | | | |
|---|---|---|---|---|---|---|
| Skup podataka | **Skup podataka za testiranje** | | | | | AUC - na setu za treniranje |
| | Točnost | Preciznost | Odziv | F mjera | AUC | |
| X | 0,689 | 0,063 | 0,612 | 0,114 | 0,725 | 0,729 |
| X2 | 0,680 | 0,062 | 0,626 | 0,113 | 0,728 | 0,720 |
| X3 | 0,680 | 0,062 | 0,625 | 0,113 | 0,728 | 0,720 |
| X4 | 0,677 | 0,061 | 0,621 | 0,112 | 0,717 | 0,712 |
| X5 | 0,675 | 0,062 | 0,636 | 0,113 | 0,722 | 0,715 |
| X6 | 0,682 | 0,062 | 0,619 | 0,113 | 0,712 | 0,704 |

Izvor: Izrada autorice

Za ispitivanje modela **slučajnih šuma** uzorak podataka podijeljen je u omjeru 70:30 na način da 30% čine noviji podaci i oni će se koristiti za testiranje modela. Preostalih 70% čine stariji podaci i taj skup je još dodatno podijeljen na skup za treniranje i validaciju na način da 4/5 podataka čini skup za treniranje, a 1/5 skup za validaciju modela.

Model je treniran na različitim dubinama stabla (2, 3, 4, 5, 7, 10, 12, 15, 20) i što se više povećavala dubina stabla to je bila veća točnost modela, međutim odabir modela sa najvećom točnosti dobivenom povećavanjem dubine stabla na skupu podataka za treniranje mogao bi uzrokovati prenaučenost[26] modela.

U Pythonu unutar paketa *sklearn* postoji gotova metoda *model_selection.GridSearchCV* [36] koja pronalazi najbolje parametre modela na način da maksimizira zadanu mjeru za validaciju modela, te je navedena metoda korištena kako bi pronašli najbolje parametre za model slučajnih šuma tako da odziv bude što veći. Metoda *GridSearchCV* ima u sebi implementiranu metodu

[26] engl. *overfitting*

za *cross-validaciju* modela, na način da podijeli skup podataka na pet grupa, te četiri koristi za treniranje i jednu za validaciju modela.
Ulazni parametri u *GridSearchCV* su bili:

- broj stabala[27]: 50,100
- broj varijabli koje se algoritam uzima u obzir kod podjele u čvorove: auto, sqrt, log2, 0.3, 0.5, 0.7
- dubina stabla: 2, 3, 4, 5, 7, 10, 12, 15, 20
- funkcija koja će se koristiti kao kriterij podjele podataka u čvoru: gini, entropy

a najveći odziv je imao model sa parametrima:
*{'criterion': 'gini', 'max_depth': 3, 'max_features': 'auto','n_estimators': 50}.*

Točnost na implementiranom modelu sa navedenim parametrima na skupu za testiranje iznosi 0,689, dok je odziv 0,601. Prikaz svih dobivenih mjera na skupu za testiranje nalazi se u *Tablica 5.5*.

***Tablica 5.5*** *Rezultati ispitivanja modela slučajne šume sa najboljim parametrima*

| **Slučajna šuma - model dubine 3** | | | | | | |
|---|---|---|---|---|---|---|
| Rezultati mjera | **Skup podataka ta testiranje** | | | | | AUC - na setu za treniranje |
| | Točnost | Preciznost | Odziv | F mjera | AUC | |
| | 0,689 | 0,062 | 0,601 | 0,112 | 0,717 | 0,720 |

Izvor: Izrada autorice

Za implementirani model slučajne šume ispitano je koje varijable su bile najznačajnije u kreiranju stabla (korištena je gotova naredba *feature_importances_* implementirana u Pythonu [37] koja izračunava značajnost varijable u slučajnoj šumi tako da za svaku značajku računa prosječnu mjeru *gini impurity* u svim kreiranim stablima u slučajnoj šumi). Dobiveni rezultati su prikazani u *Tablica 5.6,* a na *Slika 5.2* i *Slika 5.3*. prikazana su dva primjera stabla koja su kreirana u modelu.
Dobiveni rezultati pokazuju da varijabla vrši li stranka pritisak na proces obrade štete je daleko najznačajnija u modelu, a ispod nje sa većom značajnosti su varijable koje sadrže informaciju je li šteta isplaćena brzom nagodbom, je li na mjestu nezgode bila intervencija hitne pomoći te starost oštećenog vozila.

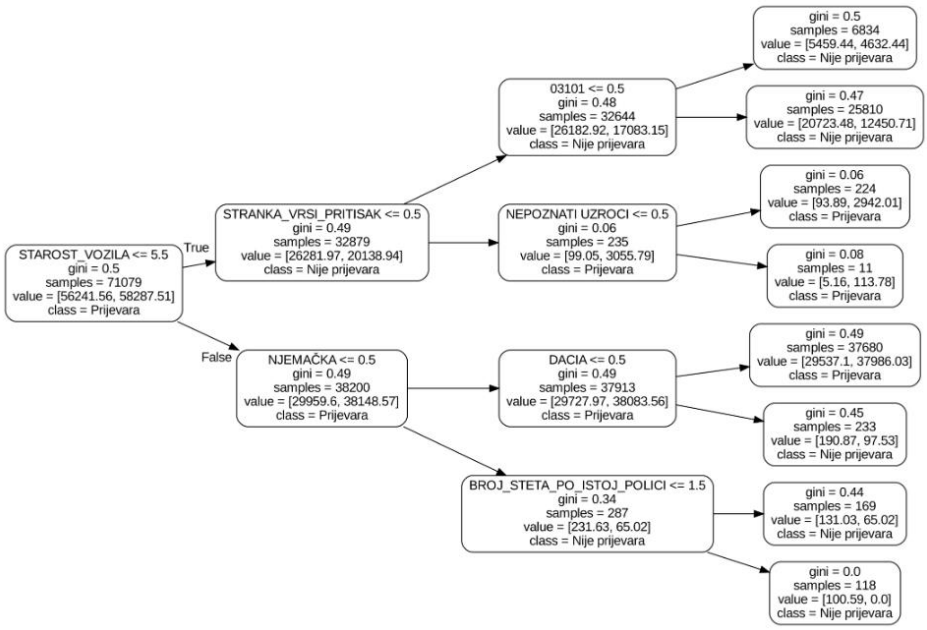
[27] engl. *number of estimators*

**Tablica 5.6** Važnost značajki[28] za model slučajne šume dubine 3

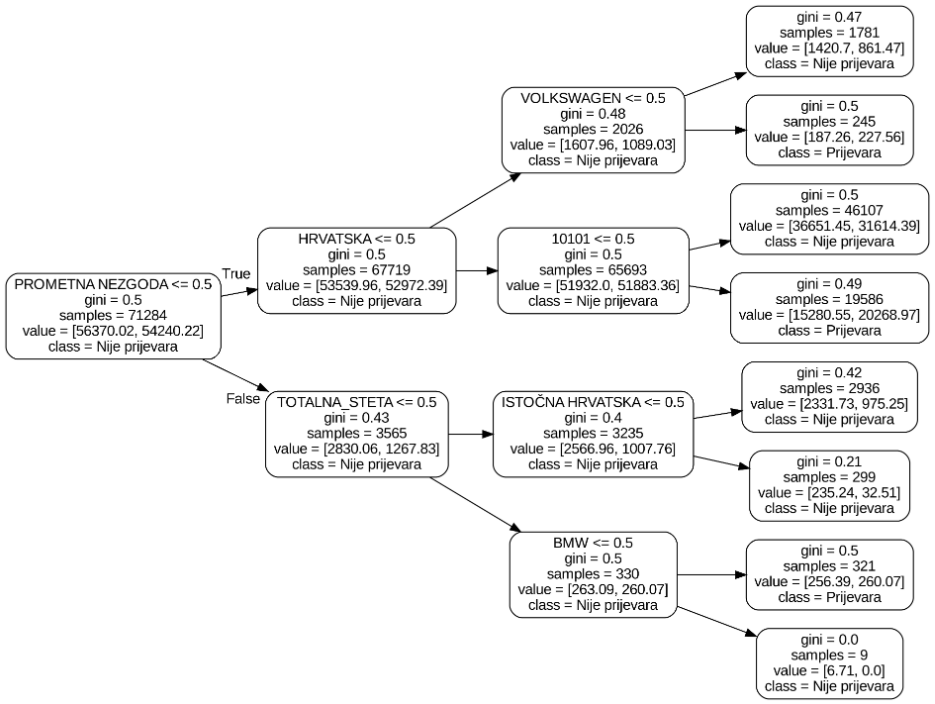
| R.Br. | Varijabla | Važnost značajki |
|---|---|---|
| 1 | STRANKA_VRSI_PRITISAK | 0,308 |
| 2 | BRZA_NAGODBA | 0,225 |
| 3 | HITNA_POMOC | 0,065 |
| 4 | STAROST_VOZILA | 0,056 |
| 5 | 10101 | 0,039 |
| 6 | 03101 | 0,039 |
| 7 | PAD ILI UDARAC NEKOG PREDMETA | 0,030 |
| 8 | STAROST_OSTECENIKA | 0,027 |
| 9 | LEASING | 0,025 |
| 10 | BR_DN_NAST_POC_PREM | 0,023 |
| 11 | BR_DN_NAST_PRI | 0,023 |
| 12 | PROMETNA NEZGODA | 0,020 |
| 13 | GRAD ZAGREB | 0,015 |
| 14 | BROJ_STETA_PO_ISTOJ_POLICI | 0,011 |
| 15 | TOTALNA_STETA | 0,010 |
| 16 | BMW | 0,009 |
| 17 | UDAR | 0,008 |
| 18 | MERCEDES | 0,006 |
| 19 | Udar vozila u objekt | 0,005 |
| 20 | F2 | 0,004 |
| 21 | OSTALO_VOZILO | 0,004 |
| 22 | Udar vozila u vozilo u kretanju | 0,004 |
| 23 | SJEVERNI JADRAN I LIKA | 0,004 |
| 24 | HRVATSKA | 0,003 |
| 25 | Slijetanje vozila s ceste | 0,003 |
| 26 | TUČA (GRAD) | 0,003 |
| 27 | RENAULT | 0,003 |

Izvor: Izrada autorice

[28] engl. *feature importance*




***Slika 5.2*** *Primjer 1 generiranog stabla algoritmom slučajnih stabala*
Izvor: Izrada autorice




***Slika 5.3*** *Primjer 2 generiranog stabla algoritmom slučajnih stabala*
Izvor: Izrada autorice

# 6 ZAKLJUČAK

Osiguravajuća društva učestalo se susreću sa lažnim odštetnim zahtjevima te koriste različita softverska rješenja za detekciju, a kasnije i za dokazivanje prijevarnih radnji. Dokazivanje prijevarnih radnji je složen proces u kojem se koriste znanja i vještine stručnjaka iz različitih područja i to je svakako najteži korak u cijelom procesu, ali korak prije je detekcija potencijalno prijevarnih šteta.

Sama detekcija se odvija na različite načine – djelatnik koji obrađuje štete iz ponašanja oštećenika prepoznaje potencijalne sumnjive radnje te predloži odštetni zahtjev za detaljniju analizu, softversko rješenje na temelju nekih predefiniranih indikatora i prediktivnih modela identificira potencijalne prijevarne štete, anonimna dojava prijevarne radnje,…

Ručna detekcija potencijalnih prijevara na temelju sumnjivog ponašanja oštećenika nije dovoljna pogotovo u vrijeme sve veće digitalizacije, obzirom da se štete prijavljuju putem web-a ili mobilnih aplikacija te nema uopće kontakta sa fizičkom osobom.

Softversko rješenje na temelju predefiniranih indikatora je odličan alat za detekciju prijevara, međutim radi se o predefiniranim indikatorima, a osobe ili organizacije koje se bave prijevarama osiguranjima pronalaze svakodnevno nove kreativne metode prijevara i upravo zbog toga se predlaže korištenje prediktivnih metoda koje pronalaze nepravilnosti u uzorku podataka.

U ovom radu su na podacima o štetama iz jednog osiguravajućeg društva u Hrvatskoj ispitane metode logistička regresija i slučajne šume kao prediktivni alati za detekciju prijevara. Temeljem dobivenih rezultata kao prediktivni model za detekciju prijevara predlaže se model **logističke regresije** sa 10 varijabli, u kojem su varijable standardizirane i koristi funkcija class_weight='balanced' koja daje veću težinu klasi koja ima manji udio. Razlozi zbog kojih se predlaže upravo taj model su:

- ima najveću mjeru **odziv** – model logističke regresije u kojem se koristi 21 varijabla[29] ima manji odziv od modela logističke regresije sa 10 varijabli
- logistička regresija je **jednostavniji** model od modela slučajnih šuma

Predloženi model mogao bi se dodatno poboljšati ukoliko se kod analize podataka i izrade modela uključe još neke varijable koje se evidentiraju u procesu obrade šteta, a neki od primjera su podaci o :

- štetniku za štete prijavljene po auto odgovornosti,

[29] broj varijabli prije korištenja One Hot Encoding tehnike

- procjenitelju koji radi procjenu,
- servisu koji radi popravak vozila,
- načinu prijave štete,
- osobama koje su već počinile prijevaru,
- osobe koje su u bilo kakvoj ulozi sudjelovale u štetama u kojima je detektirana prijevara,
- adresa osoba koje su sudjelovale u štetama.

A isto tako preporučuje se čišćenje podataka na izvoru za varijable

- marka i vrsta vozila, te godina proizvodnje vozila,
- vrsta osobe

kako bi se mogle koristiti u implementaciji modela.

# 7 LITERATURA


[1] Insurance Information Institute, „Background on: Insurance fraud“, dostupno na: https://www.iii.org/article/background-on-insurance-fraud (27.kolovoza 2022.)

[2] HANFA, "Osiguranje", dostupno: https://www.szp.hr/UserDocsImages/dokumenti/publikacije/za-web-osiguranje-2019.pdf (16. ožujka 2022.)

[3] Generali, „Osnovni pojmovi u osiguranju“, dostupno: https://www.generali.hr/rjecnik-osiguruteljnih-termina (27.kolovoza 2022.)

[4] National Insurance Brokers Association, „Glossary of Insurance Terms“, dostupno na: https://statewideinsurance.com.au/wp-content/uploads/2018/01/Insurance-Glossary.pdf (27.kolovoza 2022.)

[5] Hrvatski ured za osiguranje (HUO), „Prijevare u osiguranju“, dostupno: https://huo.hr/hr/ostale-korisne-informacije/prijevare-u-osiguranju (28.kolovoza 2022.)

[6] GRAWE, „Pojmovi u osiguranju“, dostupno: https://www.grawe.hr/pitanja-i-pojmovi/ (27. kolovoza 2022.)

[7] R.Devi Burri, R. Burii, R. Reddy Bojja i S.Rao Buruga u radu „Insurance Claim Analysis Using Machine Learning Algorithms“, dostupno na: https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf (5.kolovoz 2022.)

[8] Kowshalya G., Dr. Nandhini M., „Predicting Fraudulent Claims in Automobile Insurance“, dostupno na: https://ieeexplore.ieee.org/document/8473034 (5.kolovoz 2022.)

[9] ISTA 321 - Data Mining, „Cross Validation“ Dostupno na: https://bookdown.org/ndirienzo/ista_321_data_mining/cross-validation.html (05.kolovoz 2022.)

[10] Shah S., Koli S.P.P., Sharma S., „Insurance Fraud Detection using Machine Laerning“, dostupno na: https://www.irjet.net/archives/V8/i4/IRJET-V8I4577.pdf (5.kolovoz 2022.)

[11] Rai A.K., Dwivedi R.K., „Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme“, dostupno na: https://ieeexplore.ieee.org/document/9155615 (22.kolovoz 2022.)

[12] Vaishali, M.Tech, „Fraud detection in Credit Card by Clustering Approach“, dostupno na:

https://www.researchgate.net/publication/272863425_Fraud_Detection_in_Credit_Card_by_Clustering_Approach (5. kolovoz 2022.)

[13] Bolf N. , „Strojno učenje“, dostupno na: https://hrcak.srce.hr/file/382926 (9. kolovoz 2022.)

[14] Hastie T., Tibshirani R., Friedman J., „The Elements of Statistical Learning Dana Mining, Inference, and Predicting“, Springer, 2009.

[15] Cutler A., Cutler D.R., Stevens J.R., „Random Forests“, dostupno na: https://www.researchgate.net/publication/236952762_Random_Forests (13. kolovoz 2022.)

[16] Jamer G., Witten D., Tibshirani R., Hastie T., „An Introduction to Statistical Learning with Applications in R“, dostupno na: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf (13.kolovoz 2022.)

[17] Congming Shi, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, Jialei Liu „A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm , dostupno na: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-021-01910-w (10. ožujka 2023)

[18] Zakon o osiguranju, dostupno: https://www.zakon.hr/z/369/Zakon-o-osiguranju (12. travanj 2022.)

[19] HANFA, „O nama“, dostupno na: https://www.hanfa.hr/o-nama/ (9. kolovoza 2022.)

[20] Peck R., Olsen C., Devore J., „Introduction to Statistics and Data Analysis“, Duxbury, Third edition

[21] Marušić M., „Korelacija“, dostupno na: https://web.math.pmf.unizg.hr/~rus/nastava/stat/SLAJD/7-korelacija.pdf (9. kolovoz 2022.)

[22] Padas, „pandas.DataFrame-corr“, https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html (10.kolovoz 2022.)

[23] Scikit-learn, „Label Encoder“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (10.kolovoz 2022.)

[24] Grubišić, A., „Hi-kvadrat test i njegove primjene“, seminarski rad, Sveučilište u Zagrebu, Fakultet elektronike i računarstva Zagreb, Split, 2004, dostupno na: https://bib.irb.hr/datoteka/145851.Ani_Grubisic_hi_kvadrat.pdf (10.kolovoz 2022.)

[25] Liang R., „ Feature selection using Python for classification problems“, Towards Data Science, dostupno na: https://towardsdatascience.com/feature-selection-using-python-for-classification-problem-b5f00a1c7028#:~:text=Univariate%20feature%20selection,analysis%20of%20variance%20(ANOVA) (26.kolovoz 2022.)

[26] Scikit-learn, „Chi2“, dostupno na: sklearn.feature_selection.chi2 — scikit-learn 1.1.2 documentation (10.kolovoz 2022.)

[27] Kaggle, „Using Categorical Data with One Hot Encoding“, dostupno na: https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook (16.siječanj 2022.)

[28] Scikit-learn, „KMeans“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (16.siječanj 2022.)

[29] Scikit-learn, „Logistic Regression“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (16.siječanj 2022.)

[30] Scikit-learn, „Random Forest Classifier“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (16.siječanj 2022.)

[31] Narkhede S. , „Understanding Confusion Matrix“, Toward Data Science, dostupno na: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (26.kolovoz 2022.)

[32] Scikit-learn, „Standard Scaler“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (16.siječanj 2022.)

[33] Scikit-learn, „Robust Scaler“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler (16.siječanj 2022.)

[34] Scikit-learn, „Min Max Scaler“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html#sklearn.preprocessing.MinMaxScaler (16.siječanj 2022.)

[35] Darshan M , „Handling imbalanced data with class weights in logistic regression“, dostupno na: https://analyticsindiamag.com/handling-imbalanced-data-with-class-weights-in-logistic-regression/ (04. veljača 2023.)

[36] Scikit-learn, „Grid Search CV“, dostupno na: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (16.siječnja 2022.)

[37] Scikit-learn, „Feature importances with a forest of trees“, dostupno na:https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (04. veljača 2023.)

[38] Merčep A., Đuričić T., „Strojno učenje – bilješke sa predavanja“, dostupno na: https://repozitorij.pmf.unizg.hr/islandora/object/pmf%3A9099/datastream/PDF/view (26.kolovoz 2022.)

[39] Vondraček Z., Sandrić N, „Vjerojatnost – predavanja“, dostupno na: https://www.pmf.unizg.hr/images/50023697/vjer_predavanja.pdf (28.kolovoz 2022.)

[40] Čular M., „Modeli slučajnih šuma i primjene“, diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, 2020, dostupno na: https://repozitorij.pmf.unizg.hr/islandora/object/pmf%3A9099/datastream/PDF/view (27.kolovoz 2022.)

[41] Tadić T., „Najveće cijelo $\lfloor x \rfloor$ i njegovi prijatelji“, dostupno na: https://hrcak.srce.hr/file/3143 (27.kolovoz 2022.)

[42] Johnson K., Kuhn M., „Applied Predictive Modeling“, Springer, New York, 2013.

# 8 KAZALO POJMOVA

1) Euklidska udaljenost ili $L^2$ norma između vektora $v = (v_1, v_2, \dots, v_n)$ i $u = (u_1, u_2, \dots, u_n)$ definira se kao $d(u, v) = \sqrt{\sum_{i=1}^{n}(v_i - u_i)^2}$. [38]
2) Neka je $X$ slučajna varijabla s funkcijom gustoće $f$ i očekivanjem $\mathbb{E}[X]$. **Varijanca** od $X$ definira se kao $Var(X) := \mathbb{E}[(x - \mathbb{E}[X])^2]$.

   **Standardna devijacija** od $X$ definirana je kao $\sigma(X) := \sqrt{Var(X)}$ . [39]
3) (*Bayesova formula*) Neka je $(H_i)_{i \in I}$ potpun sustav događaja na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Tada za svaki $A \in \mathcal{F}$ takav da je $P(A) > 0$ vrijedi

$$P(H_j|A) = \frac{P(H_j)P(A|H_j)}{\sum_{i \in I} P(H_i)\, P(A|H_i)}.$$

   Dobivene vjerojatnosti $P(H_j|A)$ zovu se ***aposteriorne vjerojatnosti***. [39]
4) **Gini indeks** čvora je vjerojatnost da će nasumice odabran uzorak u čvoru biti netočno klasificiran, a računa se kao $I_G(n) = 1 - \sum_{i=1}^{J}(p_i)^2$, gdje je $n$ trenutni čvor, $p_i$ vjerojatnost klase $i$ u čvoru $n$, a $J$ je broj klasa u modelu. [40]
5) **Entropija** je mjera nečistoće u skupu podataka, te se za slučajnu varijablu $V$ sa pripadnim vrijednostima $v_k$ računa kao $H(V) = -\sum_k P(v_k) \cdot log_2 P(v_k)$, gdje je $P(v_k)$ vjerojatnost ishoda te vrijedi $\sum_k P(v_k) = 1$. [40]
6) Funkcija **najveće cijelo** od $x$ ili $\lfloor x \rfloor$ svakom realnom broju $x$ pridružuje najveći cijeli broj ne veći od x. [41]

# 9 LISTA TABLICA I SLIKA

# 10 ŽIVOTOPIS

Dijana Hajdarović rođena je 19. srpnja 1987. godine u Čakovcu. Diplomirala je na Matematičkom odsjeku na Prirodoslovno-matematičkom fakultetu u Zagrebu te je stekla akademski naziv magistra matematike. Poslijediplomski specijalistički studij „Proizvodi, digitalne inovacije i tehnologije u osiguranju – INSURTECH“ upisala je u listopadu 2020. godine.

Radno iskustvo započela je u Vindiji d.d. na radnom mjestu Analitičara transporta i distribucije, a zatim karijeru nastavlja na radnim mjestima poslovnog analitičara u tvrtki Comping d.o.o. , društvu za osiguranje Croatia Osiguranju d.d., te tvrtkama Atos IT Solutions and Services d.o.o i Bilog d.o.o .

Uspješno je sudjelovala u više domaćih i inozemnih projekata koji se odnose na razvoj i implementaciju informacijskih rješenja u osiguravajućim društvima, bankama, javnim ustanovama, prehrambenoj industriji i transportu, s time da je najznačajnije projekte odradila u društvima za osiguranje, od čega se ističu: implementacija sustava za detekciju prijevarnih šteta, implementacija web prijave štete kao i sudjelovanje u implementaciji digitalnog brenda Laqo by Croatia Osiguranje. Trenutno sudjeluje u razvoju core rješenja za društva za osiguranje u ulozi vodećeg poslovnog analitičara za područje obrade šteta.

# 11 BIOGRAPHY

Dijana Hajdarović was born on July 19th in Čakovec. She graduated from the Department of Mathematics in the Faculty of Science, University of Zagreb, and obtained a master's degree Master of Science in Mathematics. She enrolled in the postgraduate specialist study "Products, Digital Innovations and Technologies in Insurance – INSURTECH" in October 2020.

She started her business path in Vindija d.d. as a Transportation and Distribution Analyst and continued her career as a Business Analyst in Comping d.o.o., Croatia osiguranje d.d., Atos IT Solutions and Services d.o.o. and Bilog d.o.o.

She has successfully participated in multiple domestic and international projects related to the development and implementation of information solutions in insurance companies, banks, public institutions, the food industry, and transportation. Her most significant projects were in insurance companies, including the implementation of a fraud detection system, the implementation of a web-based claims reporting solution, and participation in the implementation of the digital brand "Laqo" by Croatia Osiguranje. Currently, she is working in the development of core solutions for insurance companies as a lead business analyst for claims processes.