Primjena klaster analize na otvorenim podacima

Blažek, Luka

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Luka Blažek

PRIMJENA KLASTER ANALIZE NA OTVORENIM PODACIMA

ZAVRŠNI RAD

Varaždin, 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Luka Blažek

Matični broj: 0016131726

Studij: Ekonomika poduzetništva

Primjena klaster analize na otvorenim podacima

ZAVRŠNI RAD

Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, rujan 2022.

Luka Blažek

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Znanost o podacima ima široko područje djelovanja. Također, u proteklih tridesetak godina joj se pridodaje veliko značenje te ona napreduje svakodnevno. Koristimo ju kako bi izvukli korisne uzorke iz velikih podataka, ali i za poboljšanje kod donošenja odluka u poslovanju koja su temeljna na analizi podataka. Područja rudarenja podataka i strojnog učenja su povezana sa znanosti o podacima, ali sama znanost o podacima ima širi opseg djelovanja. Dok je strojno učenje bazirano na dizajnu te procjeni algoritama da se izvuku uzorci iz podataka, rudarenje podataka je bazirano na analizi podataka koji su strukturirani.

Cilj ovog završnog rada je prikazati i provesti klaster analizu na otvorenim podacima koji se odnose na automobilsku industriju. Alati koji su korišteni u izradi rada su: Microsoft Excel i BigML. Klaster analiza je metoda nenadziranog učenja i koristi nam kako bi smo opisali podatke. Svrha klaster analize je da pronađe slične grupe unutar podataka. Omogućuje nam da vidimo sličnosti, ali i razlike. Grupe se međusobno razlikuju, dok su podaci unutar grupa slični po određenim karakteristikama. Klaster analiza ima široko područje djelovanja pa nam tako može pomoći u: ekonomiji, marketingu, medicini, radu s dokumentima, analizi zločina itd. Zbog svoje jednostavnosti i mogućnost vizualizacije klaster analizu uglavnom koristimo s drugim metodama rudarenja podataka da bi bili bolji rezultati. Na kraju prikazan je odabir najboljeg rješenja te napravljen zaključak.

Ključne riječi: klaster analiza; automobilska industrija; znanost o podacima; rudarenje podataka; strojno učenje; podaci; metodologije

Sadržaj

Sadržaj iii
1. Uvod 1
2. Znanost o podacima 2
3. Rudarenje podataka 5
3.1. CRISP metodologija 7
3.2. SEMMA metodologija 8
4. Primjena rudarenja podataka u analizi cijena automobila 9
5. Prethodna istraživanja 10
5.1. Korištenje neuronske mreže i Naive Bayesovog klasifikatora u rudarenju podataka za procjenu automobila 10
5.2. Predviđanje cijene automobila pomoću tehnika strojnog učenja 11
5.3. Predviđanje cijena starih automobila pomoću tehnika strojnog učenja 12
6. Razumijevanje podataka 13
7. Klaster analiza 18
8. Zaključak 34
9. Popis literature 35
Popis slika 36
Popis tablica 37

1. Uvod

Na samom početku rada biti će riječ o znanosti o podacima, te njezinoj povijesti. Nadalje, biti će navedene definicije rudarenja podataka te primjena koncepata. U praktičnom djelu rada provesti će se klaster analiza na javnom dostupnim podacima. Ti podaci odnose se na automobilsku industriju.

Za izradu praktičnog djela rada, korišteni su alat Microsoft Excel kako bi se sredili podaci te je nakon toga za razumijevanje podataka i provedbu klaster analize korišten alat BigML. BigML alat je osnovan 2011. s misijom da učini strojno učenje jednostavnijim i ljepšim za sve korisnike. Također, cilj je da sam rad u alatu i doživljaj bude prijatan. Velik broj organizacija svih veličina i iz različitih industrija se koristi ovim alatom te im on pomaže da izgrade rješenja temeljena na strojnom učenju. BigML je potrošna, programska i skalarna platforma strojnog učenja koja olakšava rješavanje i automatiziranje klasifikacije, regresije, predviđanje događaja u određenom vremenu, analize klastera, otkrivanje anomalija i asocijacija i modeliranja zadataka [1].

Cilj ovog završnog rada je objasniti pojmove koji su povezani uz klaster analizu te prikazati primjere na praktičnom djelu. Klaster analiza će nam omogućiti da vidimo sličnosti i povezanosti između različitih atributa, ali i razlike.

2. Znanost o podacima

Kelleher i Tierney (2021, str.1) navode da znanost o podacima obuhvaća skup načela, definicija, algoritama i procesa kako bi izvukli nejasne, ali također i korisne uzorke iz velikih podataka. Velik broj elemenata znanosti o podacima možemo naći u srodnim područjima, u koja možemo svrstati strojno učenje, te rudarenje podataka. U današnje vrijeme, pojmovi znanost o podacima, strojno učenje i rudarenje podataka se koriste kao sinonimi. Cilj ovih disciplina je da se usredotoče na unapređenje donošenja odluka koja su temeljena na samoj analizi podataka (Kelleher i Tierney, 2021, str. 1).

Strojno učenje temelji se na dizajnu i procjeni algoritama kako bi se izvukli uzorci iz samih podataka, dok se rudarenje podataka temelji na analizi strukturiranih podataka i uz to ističe komercijalnu primjenu. Iako postoje sličnosti između ovih područja, potrebno je naglasiti da je znanost o podacima šireg opsega od strojnog učenja i rudarenja podataka. Pa tako kod same znanosti o podacima dolaze u obzir i sva ova razmatranja, ali također ulaze i druga. U ta razmatranja ubrajamo dohvaćanje, čišćenje i oblikovanje podataka koji nisu strukturirani iz samog područja društvenih mreža i Interneta. Uz sve to koristi i rješenja koja dolaze iz područja velikih podataka za spremanje i obradu drugačijih tipova podataka, te koristi i sama pitanja koja su povezana uz etiku i pravnu regulativu (Kelleher i Tierney, 2021, str. 1-2).

Kada govorimo o samoj znanosti o podacima ona nam omogućava da ekstrahiramo (izlučimo) drugačije tipove uzoraka podataka iz nekih određenih velikih skupova. Uzmimo na primjer auto kuću. Možda ćemo htjeti podatke o kupcima koji bi se mogli slično ponašati i imati izražene slične karakteristike, ali korisno će nam biti ako primijetimo i razlike. Ovaj primjer u znanosti o podacima nazivamo klasteriranje o kojem će posebno riječ biti kasnije (Kelleher i Tierney, 2021, str. 1-2).

Kada na raspolaganju imamo veliki broj pojedinačnih podataka, te kada su sami obrasci i njihova ponašanja presložena za nas ljude da bi ih otkrili i izvadili rukom, tada znanost o podacima postaje najkorisnija (Kelleher i Tierney, 2021, str. 2-3).

Ovisno o samoj složenosti uzoraka podataka, možemo i definirati ljudsku sposobnost nošenja sa složenosti. Pa smo tako mi ljudi dobri kada određujemo pravila koja imaju jedan, dva ili tri atributa, ali dolazimo do problema kad se susretnemo s više od tri atributa, zato što tada dolazi do većeg međudjelovanja između samih atributa. Kada je riječ o znanosti o podacima onda govorimo često i o vezi između više desetaka, stotina, tisuća i milijuna atributa. Uzorci podataka koji su izdvojeni korištenjem znanosti o podacima, korisni su nam samo ako pružaju uvid u postojeći problem koji nam pomaže u njegovom rješavanju. Sama riječ uvid

naglašava da uzorak treba pružiti važne i vjerodostojne informacije o problemu koje nam nisu očite (Kelleher i Tierney, 2021, str. 3-5).

Sam pojam znanost o podacima je relativno novi pojam, te se on javio u devedesetim godinama prošlog stoljeća. Područja s kojima je pojam povezan imaju dosta dulju povijest od samog pojma znanosti o podacima. Znanstvena područja sa kojima je znanost o podacima povezana jest povijest prikupljanja samih podataka i povijest analiziranja podataka. Prvi načini (metode) na koji su se bilježili podaci su bili urezi na štapovima kako bi se zabilježilo izmjenjivanje dana, ali također i zabijanje stupova u tlo, za označavanje zimske kratkodnevice. Kada je došlo do pojave pisma, povećala se sposobnost prikupljanja i bilježenja podataka. U prvim pismima bilježila su se iskustva i događaji. Prvi oblik pisma pojavio se u Mezopotamiji oko 3200. godine prije Krista. Taj oblik pisma koristio se kako bi se vodile poslovne knjige, a nama je to danas poznatije kao transakcijski podaci. U tim podacima zabilježeni su različiti događaji u koje ubrajamo prodaju robe, plaćanje karticama, izdavanje računa itd. Naravno osim tih transakcijskih podataka postojali su i ne transakcijski. Prvi takvi podaci su bili demografski, te je najraniji poznati bio popis stanovništva iz Egipta 3000.godina prije Krista. Izum računala, razvoj elektroničkog senzora te digitalizacija podataka su u proteklih 150 godina pomogli u ubrzanom i velikom povećanju količine podataka koji se na dnevnoj bazi prikupe i pohrane. Prijelomna točka je bila objava rada relacijski model podataka, kojeg je napisao Edgar F. Codd 1970.godine. Taj model pruža korisnicima da mogu izdvojiti podatke iz baze podataka kroz jednostavne upite, pomoću kojih se utvrdi koje podatke korisnici žele izdvojiti od drugih podataka. Korisnik je oslobođen brige o unutarnjoj strukturi podataka, te ne treba znati gdje su oni fizički pohranjeni. Coddov rad je upravo pridonio razvoju strukturiranog upitnog jezika (SQL), te je postavio temelje za suvremene baze podataka. SQL je međunarodni standard pomoću kojeg se formuliraju upiti prema bazama podataka. Relacijska baza podataka ima savršenu strukturu za pohranjivanje podataka jer može raščlaniti u prirodne atribute. Rane verzije algoritama klasterizacije, k – sredina, te stabla odlučivanja javila su se oko 1966. godine od velikih broja istraživača (Kelleher i Tierney, 2021, str. 6-8, 14).

Područje strojnog učenja je u „srcu“ suvremene znanosti o podacima jer ima algoritme pomoću kojih se mogu automatski provesti analize velikog skupa podataka, da bi iz njih dobili bitne i korisne stvari. Iz dana u dan strojno se učenje razvija i raste. 1997. godine na predavanju C. F. J. Wua koje se zvalo „Statistics = Data Science?“ govorilo se o sve većoj upotrebi algoritama i modela, ali također naglasio se niz za statistiku trendova koji obećavaju u budućnosti. C. F. J. Wu je nakon predavanja je predložio da se termin statistika preimenuje u termin „znanost o podacima“. Od 2001. godine pojam znanosti o podacima proširio je svoje područje djelovanja od samog redefiniranja statistike. Nakon 2010. godine došlo je do masovnog povećanja količine podataka koji su generirani online aktivnostima kao što su:

zabava, društvene mreže, te maloprodaja. Podatkovni znanstvenici danas bi trebali imati određeni skup vještina. U te vještine ubrajamo: vještine komuniciranja, statistika i vjerojatnost, vizualizacija podataka, strojno učenje, poznavanja problemskog područja, računalna znanost i RVP, podatkovna etika i zakonska regulativa, priprema podataka i baze podataka. Navedene osobine važne su kako bi oni mogli izlučiti, obuhvatiti i pročistiti podatke koji nam dolaze iz vanjskih mrežnih izvora. Veoma je jasno da svaki pojedinac (podatkovni znanstvenik) ne može posjedovati sve vještine koje su prethodno navedene, pa je uobičajeno da posjeduje mnogo znanja i struke u nekima. Nadalje, podatkovni znanstvenici trebaju poznavati srž problema, kako se upotrebom metoda znanosti o podacima taj može problem riješiti te nakon toga pomoći poslovanju određene organizacije. Centar svakog projekta znanosti o podacima su podaci (Kelleher i Tierney, 2021, str. 15 - 21).

Za pokretanje donošenja odluka u svim suvremenim društvima bitna je znanost o podacima. Korištenje znanosti o podacima ima širok aspekt, pa tako nam pomaže u prodaji i marketingu, u državnoj upravi, sportu i slično. U prodaji i marketingu možemo prikupiti velik broj podataka o željama i potrebama kupaca i možemo pratiti ponašanje kupaca na Internetu. Također, uvođenje novih proizvoda nam može biti olakšano ako imamo prikupljene podatke s tržišta o preferencijama kupaca te možemo dobiti povratne informacije. Platforma Netflix također prikuplja naše podatke kada gledamo neki film ili seriju, te nam može predložiti slične. Države u današnje vrijeme također prikupljaju podatke o okolišu, načinu života, transportne sustave, medicinu i sl. Znanost o podacima pomaže i u samom sportu, kada se prikupljaju informacije o igračima, drugim timovima, statistika, te samim pratiteljima sporta koji prate određeni tim i sl. (Kelleher i Tierney, 2021, str. 25 – 30).

3. Rudarenje podataka

U današnje vrijeme informacijski sustavi svakog dana prikupe velik broj različitih podataka, koji dolaze iz različitih sadržaja, domena i izvora. Kako bi se pretražile velike količine podataka svakodnevno se pojavljuje potreba za različitim metodama, alatima i tehnikama. Također, te metode, tehnike i alati trebaju prepoznati zakonitosti između njih i prikazati ih na razini konkretnih izvještaja. Kada imamo složene zahtjeve klasični analitički pristup nam ne može biti dovoljan. Rudarenje podataka je sastavni dio poslovne inteligencije, koja uključuje alate za prikupljanje, pohranu, analizu i vizualizaciju podataka (Garača i Jadrić, 2011, str. 14).

Kada govorimo o pojmu rudarenja podataka on se može poimati na dva načina: uže i šire poimanje. Uže se odnosi na određenu fazu obrade podataka, dok se šire odnosi na cijeli proces otkrivanja znanja iz dostupnih podataka (Garača, 2008, str. 14).

Faze od kojih se sastoji cijeli proces rudarenja podataka su: određivanje cilja prema potrebama korisnika, priprema te izrada skladišta podataka, rudarenje podataka i na kraju ocjena i uporaba znanja koje je otkriveno. U nastavku će biti nabrojane metode koje se koriste u određenoj fazi obrade podataka pri rudarenju podacima. Određivanje najbližeg susjeda, grupiranje, stabla odlučivanja, genetski algoritmi, umjetne neuronske mreže, asocijativna pravila itd. (Garača i Jadrić, 2011, str. 14-15).

Rudarenje podataka se fokusira na matematiku, informatiku i statistiku, kada govorimo o pronalaženju struktura i odnosa između varijabla te kreiranju novog znanja (Berry i Linoff, 2004, str. 15). Kada je riječ o velikim nizovima podataka na njih se primjenjuju metode umjetne inteligencije i statističke analize (Olson i Delen, 2008, str. 15).

U modeliranju sustava, procesa i organizacija metode i tehnike umjetne inteligencije nam pružaju niz prednosti. Omogućuju nam učenje iz podataka, dinamičnosti, neizrazitosti i nelinearnosti. Prednosti koje su prethodno navedene dolaze do izražaja prilikom izrade modela rudarenja podataka koji nam uvelike pomažu kod rješavanja konkretnih problema te u različitim područjima znanosti (Garača i Jadrić, 2011, str. 15 – 16).

Prije otprilike dvadesetak godina rudarenje podataka je dobilo strateški značaj u organizacijama. Primjena tehnika rudarenja podataka i koncepata raširena je u mnogim područjima kao što su: transport, proizvodnja, farmaceutska industrija, planiranje infrastrukture, komunikacije, medicina itd. (Cox, 2005, str. 16). Garača (2008, str. 16) navodi da se tipične primjene rudarenja podataka odnose na područje odnosa organizacije sa svojim postojećim ili budućim klijentima, a to su: upravljanje odnosima s kupcima, segmentacija koja se odnosi na to da utvrdimo grupe klijenata koji imaju isti ili sličan način ponašanja, direktni marketing u kojem ponudu šaljemo kupcu za kojeg vjerujemo da će se na nju odazvati,

predviđanje rizika u financijama, tržištu kapitala i bankarstvu, za detektiranje mogućih zlouporaba kod kartičnog poslovanja i sl., za izradu profila kupca u kojem utvrđujemo kako se pojedinac ponaša kako bi mu poslali adekvatnu ponudu te za istraživanje povezanosti prodaje između različitih proizvoda da bi određenim mjerama potaknuli prodaju određenog proizvoda.

Uglavnom su tehnike rudarenja podataka izvedene iz statistike i računalnih znanosti. Odabir tehnike koju ćemo koristiti u postojećoj situaciji ovisi o samoj prirodi problema, vještinama i znanju osobe te dostupnosti podataka. Za izradu modela uvelike se koristi rudarenje podataka. „Model predstavlja algoritam ili skup pravila koja povezuje ulaze s određenom ciljnom varijablom ili izlazom.“ (Garača i Jadrić, 2011, str. 17). Kada govorimo o velikoj količini podataka tada ima smisla da govorimo i o rudarenju podataka. Metode koje se danas najčešće koriste u istraživanjima su: klasteriranje , stabla odlučivanja i regresija, ali postoje i ostale metode kao što su neuronske mreže i asocijativni algoritmi. Važan segment istraživanja u društvenim znanostima je analiza podataka koji su dobiveni kroz ankete i materijala iz studija slučaja koje je potrebno ubaciti u kombinaciju s podacima prikupljenih iz baza podataka. „Prema samoj definiciji rudarenja podataka, smisao je primijeniti niz tehnika i metoda koje predstavljaju ponešto alternativni, suplementni pristup u procesu otkrivanja znanja, zakonitosti, pojava.“ (Garača i Jadrić, 2011, str. 17).

3.1. CRISP metodologija

Jedna od dvije metodologije koje danas dominiraju. „CRISP (Cross Industry Standard. Process for data mining) je nastao kao inicijativa konzorcija proizvođača i korisnika softvera za rudarenje podataka kako bi se proces standardizirao.“ (Garača i Jadrić, 2011, str. 18). CRISP metodologija se sastoji od šest faza: razumijevanje problema, razumijevanje podataka, priprema podataka, modeliranje, evaluacija modela i razvijanje modela. Te faze bi se morale odvijati ciklično. U prvoj fazi, razumijevanju problema podrazumijeva se da odredimo ciljeve određenog istraživanja, procijenimo postojeću situaciju i odredimo razvoj plana samog projekta. „Rudarenje podataka je redovito predstavljeno kao tehnički problem pronalaska modela koji objašnjavaju odnos grupne ulaznih varijabli i ciljne varijable.“ (Garača i Jadrić, 20011, str. 19). Druga faza je razumijevanje podataka. Ona uključuje početno sakupljanje podataka, opisivanje samih podataka, istraživanje podataka te provjeru kvalitete samih podataka. Jasno je da se danas podaci razlikuju od izvora do izvora, od industrije do industrije te od problema do problema. Korisni podaci se danas mogu prikupiti anketiranjem ljudi. Uz to podaci koji mogu biti korisni su podaci o načinu života i podaci o web logovima. Deskriptivna statistika nam omogućuje da na kraju ove faze istražimo i opišemo vlastite podatke te nam ona pruža i vizualni prikaz danih varijabli. Iduća je faza priprema podataka. U ovoj fazi podaci se moraju odabrati, očistiti te transformirati u odgovarajući oblik. Postoji mogućnost da u ovoj fazi dođe do problema kad se provode prethodno navedeni postupci. Tip nastalog problema ovisi o tehnici rudarenja koja se koristila. Za metodu rudarenja podataka kao što je neuronska mreža to može izazvati velik broj problema, dok za stabla odlučivanja, nedostajuće i ekstremne vrijednosti ne bi trebalo stvoriti neke probleme. Nakon faze pripreme podataka, slijedi faza modeliranje. Kada je riječ o početnoj analizi prikladne su nam metode rudarenja podataka vizualizacijom i klaster analizom. Kada su nam podaci jasniji, postoji mogućnost da se razvije detaljniji model prilagodbe podacima. Također, ako dođe do određenih situacija potrebno je da se podaci podijele u one koji nam služe za treniranje i one koji nam služe za testiranje samih modela. Slijedeća je faza evaluacija modela. U ovoj fazi rezultati koji su dobiveni modelom trebaju se vrednovati s obzirom na to kako su postavljeni ciljevi u fazi razumijevanja problema (prva faza). Pomoću toga ćemo identificirati druge potrebe, te ako je potrebno vratiti se na neke od prethodnih faza. Posljednja faza (šesta faza) je razvijanje modela. Modeli koje razvijemo mogu nam biti korisni da bi potvrdili hipoteze koje su prethodno postavljene ili za otkrivanje znanja i predikciju. Znanjem kojeg smo dobili kroz prethodne faze možemo dobiti izrazito korisne modele (Garača i Jadrić, 2011, str. 18-20).

3.2. SEMMA metodologija

Druga metoda koja dominira u današnje vrijeme je SEMMA. SEMMA je zapravo jedan engleski akronim koji redom označava S – sample (uzorkovanje), E – Explore (istraživanje), M – Modify (modificiranje), M – Model (modeliranje), A – Assess (procjena). „Razvio ju je SAS institut koji je i proizvođač platforme za rudarenje podataka koji koristi tu metodologiju – SAS Enterprise Miner.“ (Garača i Jadrić, 2011, str. 18). „Počevši sa statistički reprezentativnim uzorkom podataka, SEMMA olakšava primjenu statističkih i vizualizacijskih tehnika, odabir i transformaciju najznačajnijih prediktivnih varijabli, modeliranje varijabli za predviđanje izlaza i završno, potvrđivanje vjerodostojnosti modela.“ (Garača i Jadrić, 2011, str. 20).

Sam proces počinje s fazom uzorkovanje. U ovoj fazi podrazumijeva se da nakon što se prikupe podaci slijedi kreiranje jedne ili više tablica s podacima. Uzorak bi trebao biti malen da bi se obradio u što kraćem roku, ali u drugu ruku i velik kako bi sadržavao bitne informacije. Nakon faze uzorkovanja slijedi faza istraživanje. Kako bi smo što bolje razumjeli podatke i lakše pronašli ideje o mogućim rješenjima problema, u ovoj fazi se ,kao što i samo ime govori, istražuju očekivane veze, neočekivanih trendova ili anomalije. Slijedeća je faza modificiranje. Ona obuhvaća modificiranje podataka tako da se kreiraju, odabiru i transformiraju varijable koje su nam potrebne tijekom procesa modeliranja. Četvrta faza je modeliranje. Modeliranje je aktivnost koja uključuje kombiniranje različitih istraživačkih alata da bi se dobila ona kombinacija koja predviđa određenu varijablu ili je rezultat drugačiji cilj. Posljednja faza je procjena. Ona uključuje procjenu podataka evaluacijom njihove upotrebljivosti i pouzdanosti rezultata koji su dobiveni pomoću rudarenja podataka (Garača i Jadrić, 2011, str. 18 i 20).

4. Primjena rudarenja podataka u analizi cijena automobila

Kineska automobilska tvrtka Geely Auto teži ulasku na američko tržište automobila. Cilj je postaviti svoje proizvodne jedinice na američkom području i lokalna proizvodnja automobila kako bi bila konkurenta američkom i europskom tržištu. Angažirali su konzultantsku tvrtku za automobile kako bi razumjeli čimbenike o kojima ovisi cijena automobila na tržištu. Konkretno, žele razumjeti čimbenike koji utječu na cijene automobila na američkom tržištu, budući da se one mogu jako razlikovati od kineskog tržišta. Između ostaloga, tvrtka želi znati: koje su varijable značajne u predviđanju cijena automobila, koliko dobro te varijable opisuju cijenu automobila, te posljednje na temelju različitih istraživanja tržišta, konzultantska tvrtka prikupila je veliki skup podataka o različitim tipovima automobila na američkom tržištu (Kumar, 2019.).

Analiza će omogućiti tvrtki uvid u cijene različitih automobila, koji imaju različite specifikacije te određene razlike i sličnosti, ali će im također pomoći oko odluke postavljanja vlastitih cijena na tržištu.

5. Prethodna istraživanja

U sljedećem dijelu rada biti će prikazana prethodna istraživanja koja su koristila razne algoritme i metode rudarenja podataka u domeni automobila.

5.1. Korištenje neuronske mreže i Naive Bayesovog klasifikatora u rudarenju podataka za procjenu automobila

U rudarenju podataka, klasifikacija je oblik podatkovne analize koja se može koristiti da izvuče modele koji opisuju važne klase podataka. Dva od vrlo poznata algoritma koja se koriste u klasifikaciji rudarenja podataka su neuronske mreže (Backpropagation Neural Network – BNN) i Naïve Bayes metoda (NB). Ovaj rad istražuje svojstva te dvije klasifikacijske metode koristeći skup podataka o procjeni automobila. Napravljena su dva modela za oba algoritma te su uspoređeni rezultati. Eksperimentalni rezultati pokazali su da BNN klasifikator daje veću točnost u odnosu na NB klasifikator, ali je manje učinkovit zato što oduzima puno vremena i teško ga je analizirati zbog implementacije crne kutije (black-box). Također, sama provedba BNN metode je sporija, može imati nesigurne i dvosmislene rezultate. Atributi koji su korišteni u ovom radu su: buying, maint, doors, persons, lug_boot i safety (Makki, Mustapha, Kassim, Gharayebeh, Alhazmi, 2011).

5.2. Predviđanje cijene automobila pomoću tehnika strojnog učenja

Predviđanje cijena automobila je veoma zanimljivo područje za istraživanje, zato što zahtijeva značajan napor i znanje stručnjaka. Značajan broj različitih atributa je ispitan kako bi se dobilo točno i pouzdano predviđanje. Kako bi se izgradio model za predviđanje cijena korištenih automobila u Bosni i Hercegovini, primijenjene su tri tehnike strojnog učenja: umjetna neuronska mreža (Artifical Neural Network), metode potpornih vektora (Support Vector Machine – SVM) i slučajne šume (Random Forest). Sve spomenute tehnike su u radu primijenjene kao cjelina, a ne zasebno. Podaci korišteni za predviđanje u ovom radu prikupljeni su sa web portala autopijaca.ba korištenjem web ekstrakcije podataka (web scraper) koji je napisan u PHP programskom jeziku. Odnosna svojstva različitih algoritama su zatim bila uspoređena, kako bi se pronašao onaj algoritam koji najbolje odgovara dostupnom skupu podataka. Konačni model predviđanja integriran je u aplikaciju Java. Model je ocijenjen korištenjem testnih podataka i dobivena je točnost od 87,38%. Za razliku od toga, kada je primijenjen jedan algoritam strojnog učenja na skup podataka točnost je bila manja od 50%. Predviđanje cijena automobila može biti veoma zahtijevan zadatak zbog velikog broja atributa koji bi se trebali uzeti u obzir za samo predviđanje. Neki od atributa koji su korišteni u ovom radu su: brand, model, fuel, power in kilowats, year of man, miles, leather, cruise control, price (Gegic, Isakovic, Keco, Masetic, Kevric, 2019).

5.3. Predviđanje cijena starih automobila pomoću tehnika strojnog učenja

Svijet raste iz dana u dan, a isto tako očekivanja svakog čovjeka također rastu. Od svih očekivanja jedno od njih je kupnja automobila. Nisu svi u mogućnosti uvijek kupiti novi automobil, pa će onda neki kupiti rabljeni. Osobe uobičajeno ne znaju tržišnu cijenu njegovog ili njezinog automobila iz snova za rabljeni ili stari. U ovom radu dolazi se do platforme koja je napravljena pomoću tehnologije strojnog učenja, a ta platforma će pomoći ljudima u predviđanju cijena automobila. Korištenjem nadziranih metoda strojnog učenja kao što su linearna regresija, KNN (K – nearest neighbours algorithm), slučajne šume (Random Forest), XG boost i stablo odlučivanja (Decision Tree) pokušalo se izgraditi statistički model koji će moći predvidjeti cijenu rabljenog automobila. U tome su pomogli prethodni podaci o potrošačima te zadani skup značajki. Na kraju su uspoređeni rezultati modela, tj. točnost predviđanja, kako bi se odredio optimalan model. Atributi koji su korišteni u ovom radu su kilometers traveled, year of registration, fuel type, car model, fiscal power, car brand i gear type. Najbolje rezultate dao je model slučajnih šuma (Gajera, Gondaliya, Kavathiya, 2021).

6. Razumijevanje podataka

Podaci na kojim se izvršava praktični dio završnog rada prikupljeni su s Interneta i odnose se na automobilsku industriju. Za početak podaci će se srediti u alatu Microsoft Excel kako bi se mogli ubaciti u alat BigML, u kojem će se provoditi daljnja analiza. Biti će napravljen prikaz samih podataka koji su podijeljeni u kategorijske i kontinuirane atribute, uz to biti će i prikaz nedostajućih vrijednosti. Grafički će biti prikazana distribucija podataka i prikaz stršila. Nakon toga slijedi klaster analiza za koju su odabrana šest klastera, te će se nakon njihovih prikaza napraviti analiza kako bi se odabrao najoptimalniji klaster.

U nastavku će biti prikazani podaci razvrstani u dvije tablice. Prva tablica će prikazati kontinuirane atribute, dok će druga tablica prikazati kategorijske atribute. Uz sve to prikazana će biti i tablica distribucije samih podataka za koje će biti naveden tip distribucije. U tablicama možemo vidjeti prikaz svih podataka koji su korišteni u samom završnom radu. Sam skup podataka se sastoji od 26 različitih podataka koji su podijeljeni u prethodno navedene skupine. Na njima će se kasnije provesti klaster analiza.

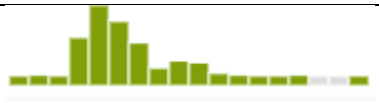
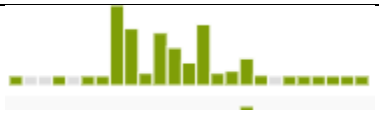
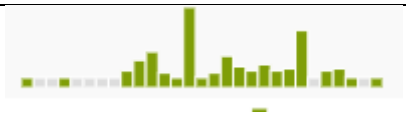
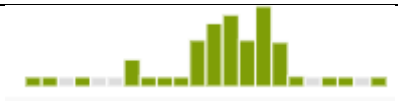
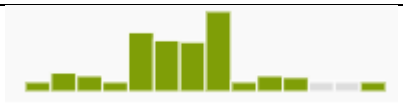
Naziv atributa	*Broj vrijednosti atributa*	*Udio nedostajućih vrijednosti*	*Minimalna vrijednost*	*Aritmetička sredina*	*Medijan*	*Maksimalna vrijednost*	*Standardna devijacija*
Car_ID	205	0	1.00	103.00	103.00	205.00	59.32
Symboling	205	0	-2.00	0.83	1.00	3.00	1.25
wheelbase	205	0	86.60	98.76	97.00	120.90	6.02
carlength	205	0	141.10	174.05	173.20	208.10	12.34
carwidth	205	0	60.30	65.91	65.50	72.30	2.15
carheight	205	0	47.80	53.72	54.10	59.80	2.44
curbweight	205	0	1488.00	2555.57	2414.00	4066.00	520.68
enginesize	205	0	61.00	126.91	120.00	326.00	41.64
boreratio	205	0	2.54	3.33	3.31	3.94	0.27
stroke	205	0	2.07	3.26	3.29	4.17	0.31
compressionratio	205	0	7.00	10.14	9.00	23.00	3.97
horsepower	205	0	48.00	104.12	95.00	288.00	39.54
peakrpm	205	0	4150.00	5125.12	5200.00	6600.00	476.99
citympg	205	0	13.00	25.22	24.00	49.00	6.54
highwaympg	205	0	16.00	30.75	30.00	54.00	6.89

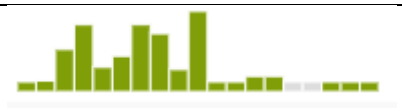
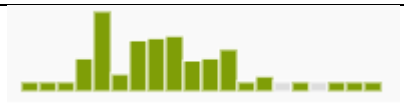
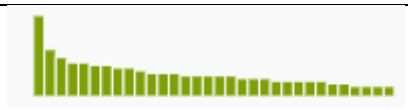
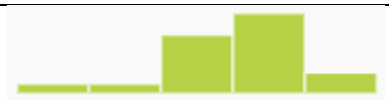
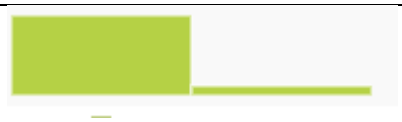
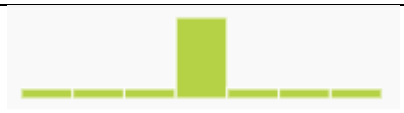
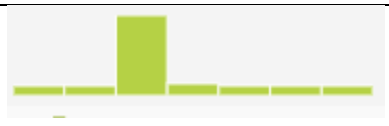
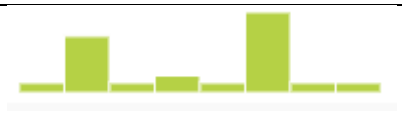
price	205	0	5118.00	13276.71	10295.00	45400.00	7988.85

Tablica 1 Opis kontinuiranih atributa

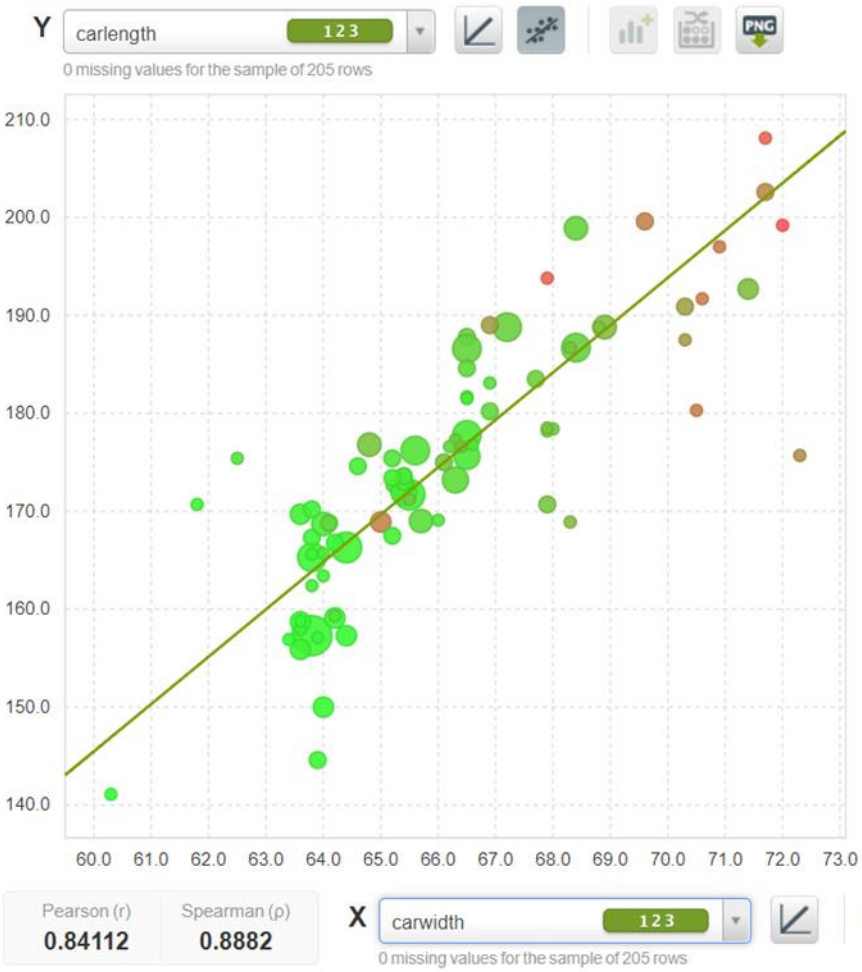
Naziv atributa	*Broj vrijednosti atributa*	*Udio nedostajućih vrijednosti*	*Mod*
CarName	205	0	toyota
fueltype	205	0	gas
aspiration	205	0	std
doornumber	205	0	four
carbody	205	0	sedan
drivewheel	205	0	fwd
enginelocation	205	0	front
enginetype	205	0	Ohc
cylindernumber	205	0	Four
fuelsystem	205	0	Mpfi

Tablica 2 Opis kategorijskih atributa

Naziv atributa	*Grafički prikaz distribucije*	*Tip distribucije*
Car_ID		Uniformna
Symboling		Normalna/normalna pomaknuta udesno
wheelbase		Normalna pomaknuta udesno
carlength		Multimodalna
carwidth		Multimodalna
carheight		Multimodalna
curbweight		Multimodalna
enginesize		Normalna pomaknuta udesno
boreratio		Multimodalna
stroke		Multimodalna
compressionratio		Normalna pomaknuta udesno
horsepower		Normalna pomaknuta udesno
peakrpm		Multimodalna

citympg		Multimodalna
highwaympg		Multimodalna
price		Normalna pomaknuta udesno
CarName		Eksponencijalna
fueltype		Normalna pomaknuta ulijevo
aspiration		Normalna pomaknuta udesno
doornumber		Normalna pomaknuta udesno
carbody		Normalna pomaknuta ulijevo
drivewheel		Multimodalna
enginelocation		Normalna pomaknuta udesno
enginetype		Normalna
cylindernumber		Normalna pomaknuta udesno
fuelsystem		Multimodalna

Tablica 3 Distribucija vrijednosti atributa



Slika 1 Scatterplot prikaz

Gornja slika nam prikazuje odnos carwidth (os x) i carlength (os y). Ona nam govori o stupnju povezanosti između dva atributa numeričkog tipa (carwidth i calength). Spearmanov koeficijent korelacije iznosi 0,8882, dok je Pearsonov koeficijent korelacije 0,84112. Što je koeficijent korelacije bliže broju 1 to je i povezanost atributa jača. Ovdje imamo veliku povezanost između atributa. Iz slike vidimo da su instance manje-više grupirane te da prate određenu liniju. Kako ovdje imamo pozitivan predznak koeficijenta, to znači da ako raste vrijednost jednog atributa tada raste i vrijednost drugog atributa. Vrijedi i obrnuto.

7. Klaster analiza

Kada govorimo o najčešćoj vrsti učenja bez nadzora tada je riječ o klaster analizi (Kelleher i Tierney, 2021, str. 102). Klaster analiza je metoda deskriptivnog modeliranja. Njezin je cilj da pronađe grupe koje su različite jedna od druge, ali da su atributi unutar grupa slični jedni drugima po karakteristikama. Kada započinje sam proces nije poznato po kojim atributima će klasteriranje biti izvršeno. Najčešća uporaba metode klasteriranja je u svrhu segmentacije tržišta, ali također koristi se i za istraživanje ostalih područja (Garača i Jadrić, 2011, str. 38). Možemo reći da postoje mnoge koristi od grupiranja sličnih objekata. Na primjer, u ekonomskoj primjeni može nas zanimati da pronađemo zemlje koja imaju slična gospodarstva. Nadalje, u medicinskoj primjeni bi htjeli pronaći klastere pacijenata koji imaju slične simptome. Kada govorimo o marketingu cilj nam je pronaći klastere kupaca s sličnim kupovnim ponašanjem (Bramer, 2007, str. 221). Još jedan primjer je izrada preporuka kupcima za kupovinu određenog proizvoda. Ako se kupcu svidjela određena majica i hlače, onda postoji i vjerojatnost da će mu se svidjeti i majica i hlače iz istog klastera (Kelleher i Tierney, 2021, str. 104). Kada su nam svi atributi u određenom skupu podataka numerički i ako im je sličan raspon vrijednosti, tada možemo izračunati euklidsku udaljenost između podataka. Formula za izračun euklidske udaljenosti glasi: $d(A,B) = [(p_1 - p_2)^2 + (q_1 - q_2)^2]^{0,5}$. Postoje mogućnosti da nam numerički atributi imaju različit raspon vrijednosti, pa je u tim situacijama potrebno normalizirati atribute. U obzir se mora uzeti i to da su neki atributi važniji od drugih, tako da im se prilikom izračuna udaljenosti mora dodati određena težina (Kelleher i Tierney, 2021, str. 102-103).

Kao prednosti klaster analize možemo navesti da je ona veoma jednostavna, lako razumna i ima mogućnost vizualizacije grafičkih prikaza. Pod nedostatke također možemo navesti tu jednostavnost, zbog koje nije preporučeno da se previše oslanjamo na nju. Uz to možemo nadodati da su klasteri osjetljivi na odstupajuće vrijednosti. Zbog toga se klaster analiza najčešće kombinira s ostalim metodama rudarenja podataka kako bi došli do boljih rezultata i interpretacija (Garača i Jadrić, 2011, str-39).

Kako bi nam sama analiza bila što jasnija u samom radu isproban je različiti broj klastera. Postoje dvije mogućnosti odabira algoritma u klasteriranju, a to su K – means i G – means. Algoritam klasteriranja koji se koristi u ovom radu je K – means koji je ujedno i najčešće korišteni algoritam klasteriranja. Taj algoritam unaprijed određuje broj klastera i on je temeljen na aritmetičkim sredinama, dakle računaju se aritmetičke sredine grupa od klastera i u grupi se nalaze one instance koje su najbliže dobivenoj aritmetičkoj sredini. Središte klastera

nazivamo centroid. Postoje i druge metode klasteriranja, a jedna od češće korištenih je aglomerativno hijerarhijsko grupiranje. Ono je također jednostavno. Započnemo s svakim objektom u posebnom klasteru, a zatim uzastopno spajamo najbliži par klastera dok ne završimo sa samo jednim klasterom koji obuhvaća sve (Bramer, 2007, str. 231). U samom radu korištena su tri uvjeta kako bi odabrali optimalan broj klastera:

- ratio_ss nam treba biti što bliži vrijednosti broja 1,
- broj instanci u klasteru je manji od 35% ukupnog broja instanci, a veći od 5% ukupnog broja instanci,
- klastere je moguće objasniti, a postoje i razlike u srednjim vrijednostima varijabli između različitih klastera.

U tablici slijedi prikaz različitih brojeva klastera iz skupa podataka. Uz to navedeni je ratio_ss za sve klastere, te je prikazana distribucija instanci.

Broj klastera	**Ratio_ss**	**Distribucija instanci**
3	0.398140	Global: 100% (205 instances) • Cluster 0: 23.41% (48 instances) • Cluster 1: 54.15% (111 instances) • Cluster 2: 22.44% (46 instances)
4	0.441790	Global: 100% (205 instances) • Cluster 0: 44.39% (91 instances) • Cluster 1: 5.37% (11 instances) • Cluster 2: 30.73% (63 instances) • Cluster 3: 19.51% (40 instances)
6	0.555020	Global: 100% (205 instances) • Cluster 0: 5.37% (11 instances) • Cluster 1: 17.56% (36 instances) • Cluster 2: 27.32% (56 instances) • Cluster 3: 5.37% (11 instances)

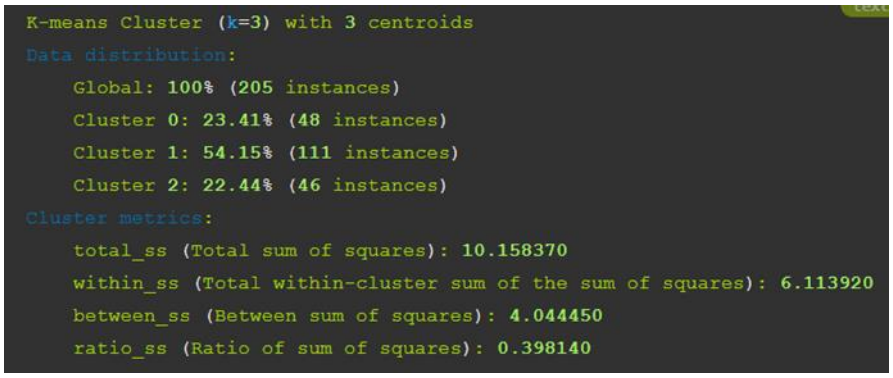
		• Cluster 4: 25.85% (53 instances) • Cluster 5: 18.54% (38 instances)
8	0.615290	Global: 100% (205 instances) • Cluster 0: 21.46% (44 instances) • Cluster 1: 4.88% (10 instances) • Cluster 2: 4.88% (10 instances) • Cluster 3: 20.00% (41 instances) • Cluster 4: 16.10% (33 instances) • Cluster 5: 11.22% (23 instances) • Cluster 6: 4.88% (10 instances) • Cluster 7: 16.59% (34 instances)
10	0.646340	Global: 100% (205 instances) • Cluster 0: 22.44% (46 instances) • Cluster 1: 9.76% (20 instances) • Cluster 2: 12.68% (26 instances) • Cluster 3: 5.37% (11 instances) • Cluster 4: 4.88% (10 instances) • Cluster 5: 16.59% (34 instances) • Cluster 6: 6.34% (13 instances) • Cluster 7: 8.78% (18 instances) • Cluster 8: 4.88% (10 instances) • Cluster 9: 8.29% (17 instances)
12	0.680050	Global: 100% (205 instances) • Cluster 00: 8.78% (18 instances) • Cluster 01: 10.24% (21 instances) • Cluster 02: 2.44% (5 instances)

		• Cluster 03: 2.93% (6 instances) • Cluster 04: 26.83% (55 instances) • Cluster 05: 16.10% (33 instances) • Cluster 06: 7.32% (15 instances) • Cluster 07: 9.76% (20 instances) • Cluster 08: 4.88% (10 instances) • Cluster 09: 2.93% (6 instances) • Cluster 10: 2.93% (6 instances) • Cluster 11: 4.88% (10 instances)

Tablica 4 Prikaz klastera

U nastavku slijedi vizualizacija klastera, te njihove karakteristike.

Broj klastera (k=3)



Slika 2 Distribucija i metrika k=3

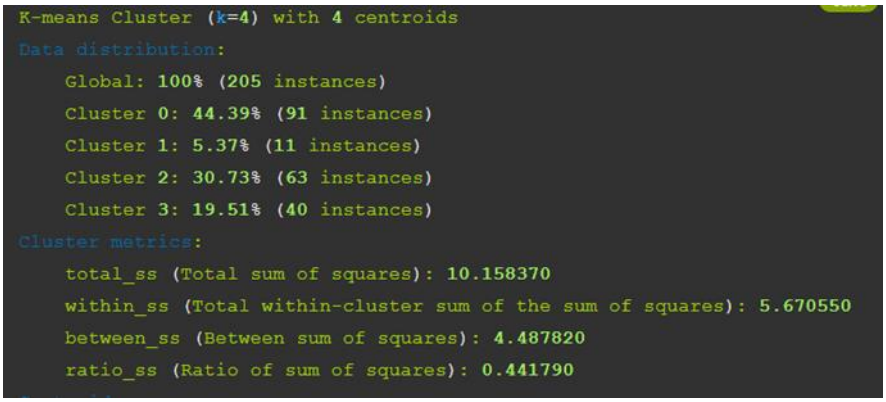
Iz slike možemo vidjeti da ratio_ss iznosi 0.398140, što nije najbliže broju 1 kako bi se zadovoljio uvjet. Klaster 1 (54,15%) ne zadovoljava uvjet da je broj instanci u klasteru manji od 35% ukupnog broja instanci u klasteru. Ostali klasteri (klaster 0 i klaster 2) zadovoljavaju gore navedene uvjete.

Klaster 0 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 21707.35

Klaster 1 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 8369.79

Klaster 2 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 16635.22

Broj klastera (k=4)



Slika 3 Distribucija i metrika k=4

U ovom primjeru možemo vidjeti da klaster 0 (44,39%) ne zadovoljava uvjete za postotke koji su prethodno navedeni, tj. 44,39% premašuje 35%. Ostali klasteri su unutar zadanog uvjeta. Ratio_ss za navedeni primjer iznosi 0.441790 što je s obzirom na tri klastera veća vrijednost.

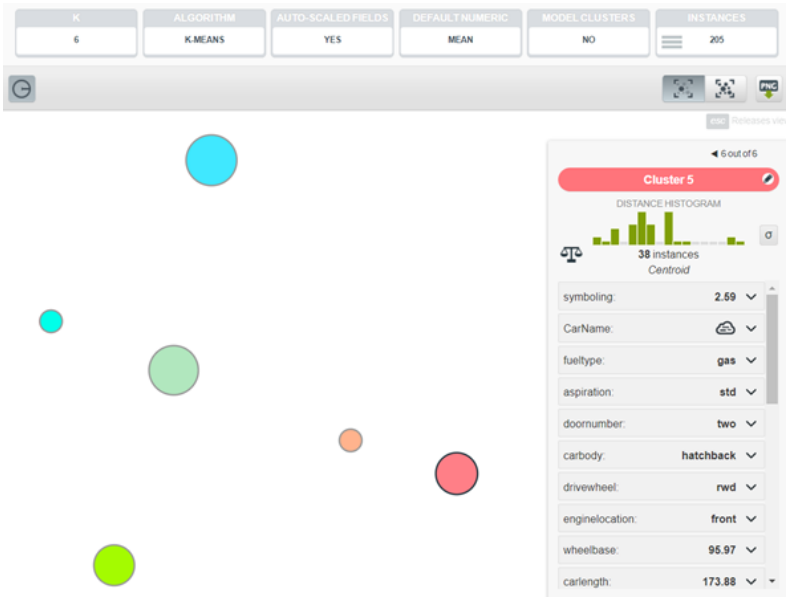
Klaster 0 → fueltype: gas, aspiration: std, doornumber: four, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 7793.24

Klaster 1 → fueltype: gas, aspiration: std, doornumber: two, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 36341.41

Klaster 2 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 13719.00

Klaster 3 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 18394.08

Broj klastera (k=6)



Slika 4 Vizualni prikaz k=6

```
Data distribution:
    Global: 100% (205 instances)
    Cluster 0: 5.37% (11 instances)
    Cluster 1: 17.56% (36 instances)
    Cluster 2: 27.32% (56 instances)
    Cluster 3: 5.37% (11 instances)
    Cluster 4: 25.85% (53 instances)
    Cluster 5: 18.54% (38 instances)
Cluster metrics:
    total_ss (Total sum of squares): 10.158370
    within_ss (Total within-cluster sum of the sum of squares): 4.520280
    between_ss (Between sum of squares): 5.638090
    ratio_ss (Ratio of sum of squares): 0.555020
```

Slika 5 Distribucija i metrika k=6

U primjeru sa šest klastera svi klasteri zadovoljavaju uvjete postotka koji su u uvodu navedeni, tako da što se tiče toga nema odstupanja. Ratio_ss za šest klastera iznosi 0.555020 što je s obzirom na prethodne primjere ipak nešto veća vrijednost iako i dalje nije blizu broja 1.

Klaster 0 → fueltype: gas, aspiration: std, doornumber: two, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 36263.88

Klaster 1 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 16849.03

Klaster 2 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 7165.64

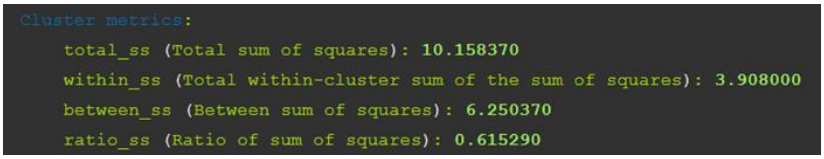
Klaster 3 → fueltype: diesel, aspiration: turbo, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 21446.60

Klaster 4 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 9220.47

Klaster 5 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 15183.77

Broj klastera (k=8)

Slika 6 Distribucija k=8



Slika 7 Metrika k=8

Kod ovog primjera možemo vidjeti da su minimalna odstupanja kod klastera 1, klastera 2 i klastera 6, postotak instanci iznosi 4,88%. Ostali klasteri zadovoljavaju uvjete koji su navedeni. Ratio_ss iznosi 0,615290.

Klaster 0 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 6756.82

Klaster 1 → fueltype: diesel, aspiration: turbo, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 22124.73

Klaster 2 → fueltype: diesel, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10875.24

Klaster 3 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 9286.34

Klaster 4 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 11227.99

Klaster 5 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 18566.46

Klaster 6 → fueltype: gas, aspiration: std, doornumber: two, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 36899.55

Klaster 7 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 17547.91

Broj klastera (k=10)

```
Data distribution:
    Global: 100% (205 instances)
    Cluster 0: 22.44% (46 instances)
    Cluster 1: 9.76% (20 instances)
    Cluster 2: 12.68% (26 instances)
    Cluster 3: 5.37% (11 instances)
    Cluster 4: 4.88% (10 instances)
    Cluster 5: 16.59% (34 instances)
    Cluster 6: 6.34% (13 instances)
    Cluster 7: 8.78% (18 instances)
    Cluster 8: 4.88% (10 instances)
    Cluster 9: 8.29% (17 instances)
```

Slika 8 Distribucija k=10

```
Cluster metrics:
    total_ss (Total sum of squares): 10.158370
    within_ss (Total within-cluster sum of the sum of squares): 3.592640
    between_ss (Between sum of squares): 6.565730
    ratio_ss (Ratio of sum of squares): 0.646340
```

Slika 9 Metrika k=10

Klaster 4 i klaster 8 imaju minimalna odstupanja, tj. ne zadovoljavaju uvjet jer im je postotak instanci manji od 5% (iznosi 4,88%). Ratio_ss iznosi 0,646340.

Klaster 0 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 6882.08

Klaster 1 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10560.54

Klaster 2 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 9278.88

Klaster 3 → fueltype: gas, aspiration: std, doornumber: two, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 36341.41

Klaster 4 → fueltype: diesel, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10173.20

Klaster 5 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 15699.86

Klaster 6 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 17265.57

Klaster 7 → fueltype: gas, aspiration: turbo, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 16842.18

Klaster 8 → fueltype: diesel, aspiration: turbo, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 21503.10

Klaster 9 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10483.00

Broj klastera (k=12)

```
Cluster 00: 8.78% (18 instances)
Cluster 01: 10.24% (21 instances)
Cluster 02: 2.44% (5 instances)
Cluster 03: 2.93% (6 instances)
Cluster 04: 26.83% (55 instances)
Cluster 05: 16.10% (33 instances)
Cluster 06: 7.32% (15 instances)
Cluster 07: 9.76% (20 instances)
Cluster 08: 4.88% (10 instances)
Cluster 09: 2.93% (6 instances)
Cluster 10: 2.93% (6 instances)
Cluster 11: 4.88% (10 instances)
```

Slika 10 Distribucija k=12

```
Cluster metrics:
    total_ss (Total sum of squares): 10.158370
    within_ss (Total within-cluster sum of the sum of squares): 3.250150
    between_ss (Between sum of squares): 6.908220
    ratio_ss (Ratio of sum of squares): 0.680050
```

Slika 11 Metrika k=12

U ovom primjeru sa dvanaest klastera možemo vidjeti najveća odstupanja od uvjeta te on sigurno neće ulaziti u obzir prilikom odabira najoptimalnijeg broja klastera. Klasteri koji ne zadovoljavaju uvjete su: klaster 2, klaster 3, klaster 8, klaster 9, klaster 10, klaster 11. Ratio_ss iz ovog primjera ima najveću vrijednost s obzirom na prethodne primjere te iznosi 0,680050.

Klaster 0 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 16745.66

Klaster 1 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 17005.99

Klaster 2 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 37668.80

Klaster 3 → fueltype: gas, aspiration: std, doornumber: two, carbody: convertible, drivewheel: rwd, enginelocation: front, price: 25932.83

Klaster 4 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: fwd, enginelocation: front, price: 7180.27

Klaster 5 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 9631.62

Klaster 6 → fueltype: gas, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10066.54

Klaster 7 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 13294.61

Klaster 8 → fueltype: diesel, aspiration: turbo, doornumber: four, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 21503.10

Klaster 9 → fueltype: gas, aspiration: std, doornumber: two, carbody: hatchback, drivewheel: rwd, enginelocation: front, price: 11819.33

Klaster 10 → fueltype: gas, aspiration: std, doornumber: two, carbody: sedan, drivewheel: rwd, enginelocation: front, price: 35235.25

Klaster 11→ fueltype: diesel, aspiration: std, doornumber: four, carbody: sedan, drivewheel: fwd, enginelocation: front, price: 10173.20

Odabir najbolje opcije

Najoptimalniji broj klastera je k=6 zato što je on najbliže gore postavljenim uvjetima. Što se tiče samih postotaka tamo uopće nema odstupanja. Odstupanje se javlja kod ratio_ss zato što njegova vrijednost iznosi 0.555020 što nije najbliže broju 1. Također, isprobani su i ostali klasteri koji ovdje nisu prikazani, ali su imali veća odstupanja. Što se tiče samih sličnosti možemo vidjeti da kod goriva prevladava benzin. Kod same aspiracije najčešća je std (standard transmission). Najčešći broj vrata je četiri, dok se ponekad jave i automobili sa dvoja vrata. Svi automobili imaju motor u prednjem dijelu automobila. Kod četiri klastera možemo primijetiti da im je pogon na stražnje kotače (rwd), dok je kod dva klastera pogon na prednje kotače. Što se tiče samog tijela automobila, klasteri 0,1,3,4, imaju sedan, dok preostala dva hatchback.

U nastavku slijedi prikaz tablice najoptimalnijeg klastera.

Klasteri	**Aritmetička sredina**	**Instance**
Klaster 0	0.17999	Horsepower – 194.37435 Compressionratio – 8.63225 Citympg – 15.28886 Peakrpm – 5052.33161 Fuelsystem - mpfi Enginesize – 246.73575 Stroke – 3.3592 Boreratio – 3.65312 Price – 36263.8763 Highwaympg – 19.88731 Cylindernumber - six Curbweight – 3635.23187 Enginetype - ohcv Carwidth – 70.0158 Carheight – 53.40738 Carlength – 192.64145 Carbody - sedan Drivewheel - rwd Aspiration - std Doornumber - two Carname – bmw, bmw x3, buick, cayenne, century, coupe, deluxe, isuzu, jaguar, porsche, porsche cayenne, turbo, x3 Fueltype - gas Symboling – 0.5272 Enginelocation - front Wheelbase – 106.85609
Klaster 1	0.14419	Horsepower – 126.06588 Compressionratio – 8.68303 Citympg – 19.77333 Peakrpm – 5205.23529 Fuelsystem - mpfi Enginesize – 139.14941 Stroke – 3.09099

		Boreratio – 3.47493 Price – 16849.0306 Highwaympg – 25.12549 Cylindernumber - four Curbweight – 2948.15804 Enginetype - ohc Carwidth – 67.43392 Carheight – 55.67686 Carlength – 185.96784 Carbody - sedan Drivewheel - rwd Aspiration - std Doornumber - four Carname – 504, peugeot, volvo Fueltype – gas Symboling – 0.04706 Enginelocation - front Wheelbase – 104.15302
Klaster 2	0.1229	Horsepower – 73.18708 Compressionratio – 9.6504 Citympg – 31.92214 Peakrpm – 5375.23253 Fuelsystem – 2bbl Enginesize – 94.13154 Stroke – 3.23844 Boreratio – 3.05383 Price – 7165.63593 Highwaympg – 37.63752 Cylindernumber - four Curbweight – 2004.76641 Enginetype - ohc Carwidth – 63.94337 Carheight – 52.74095 Carlength – 160.30369 Carbody - hatchback Drivewheel - fwd Aspiration - std Doornumber - two Carname – honda, nissan Fueltype - gas Symboling – 1.07414 Enginelocation - front Wheelbase – 94.01576
Klaster 3	0.12401	Horsepower – 105.33018 Compressionratio – 21.26861 Citympg – 25.16373 Peakrpm – 4292.55751 Fuelsystem - idi Enginesize – 162.67794 Stroke – 3.56189 Boreratio – 3.57119 Price – 21446.60217 Highwaympg – 28.37618 Cylindernumber - four Curbweight – 3383.59811

		Enginetype - ohc Carwidth – 69.1138 Carheight – 56.71705 Carlength – 190.66942 Carbody - sedan Drivewheel - rwd Aspiration - turbo Doornumber - four Carname – 504, 604sl, buick, century, custom, diesel, gs, mazda, mazda rx-7 gs, peugeot, peugeot 504, peugeot 604sl, rx, sw, turbo, volvo Fueltype - diesel Symboling - -0.6806 Enginelocation - front Wheelbase – 109.99689
Klaster 4	0.14632	Horsepower – 82.66958 Compressionratio – 10.66132 Citympg – 27.46986 Peakrpm – 4898.57746 Fuelsystem – 2bbl Enginesize – 110.69634 Stroke – 3.23428 Boreratio – 3.31132 Price – 9220.46845 Highwaympg – 33.24761 Cylindernumber - four Curbweight – 2357.51155 Enginetype - ohc Carwidth – 65.26155 Carheight – 54.53651 Carlength – 172.66527 Carbody – sedan Drivewheel - fwd Aspiration - std Doornumber - four Carname – subaru, toyota, volkswagen Fueltype - gas Symboling – 0.29662 Enginelocation - front Wheelbase – 97.65634
Klaster 5	0.15569	Horsepower – 132.06438 Compressionratio – 8.73454 Citympg – 20.29987 Peakrpm – 5264.99365 Fuelsystem – mpfi Enginesize – 139.9208 Stroke – 3.36197 Boreratio – 3.45914 Price – 15183.77347 Highwaympg – 26.3939 Cylindernumber - four Curbweight – 2695.00805 Enginetype - ohc Carwidth – 66.01546 Carheight – 51.13172

		Carlength – 173.88407 Carbody - hatchback Drivewheel - rwd Aspiration - std Doornumber - two Carname – mitsubishi, toyota Fueltype - gas Symboling – 2.58662 Enginelocation - front Wheelbase – 95.91788

Tablica 5 Prikaz najoptimalnijeg klastera

Klaster 0 ima srednju vrijednost 0.17999 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 194.37435, Compressionratio – 8.63225, Citympg – 15.28886, Peakrpm – 5052.33161, Fuelsystem - mpfi, Enginesize – 246.73575, Stroke – 3.3592, Boreatio – 3.65312, Price – 36263.8763, Highwaympg – 19.88731, Cylindernumber - six, Curbweight – 3635.23187, Enginetype - ohcv, Carwidth – 70.0158, Carheight – 53.40738, Carlength – 192.64145, Carbody - sedan, Drivewheel - rwd, Aspiration - std, Doornumber - two, Carname – bmw, bmw x3, buick, cayenne, century, coupe, deluxe, isuzu, jaguar, porsche, porsche cayenne, turbo, x3, Fueltype - gas, Symboling – 0.5272, Enginelocation - front, Wheelbase – 106.85609.

Klaster 1 ima srednju vrijednost 0.14419 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 126.06588, Compressionratio – 8.68303, Citympg – 19.77333, Peakrpm – 5205.23529, Fuelsystem - mpfi, Enginesize – 139.14941, Stroke – 3.09099, Boreatio – 3.47493, Price – 16849.0306, Highwaympg – 25.12549, Cylindernumber - four, Curbweight – 2948.15804, Enginetype - ohc, Carwidth – 67.43392, Carheight – 55.67686, Carlength – 185.96784, Carbody - sedan, Drivewheel - rwd, Aspiration - std, Doornumber - four, Carname – 504, peugeot, volvo, Fueltype - gas, Symboling – 0.04706, Enginelocation - front, Wheelbase – 104.15302.

Klaster 2 ima srednju vrijednost 0.1229 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 73.18708, Compressionratio – 9.6504, Citympg – 31.92214, Peakrpm – 5375.23253, Fuelsystem – 2bbl, Enginesize – 94.13154, Stroke – 3.23844, Boreatio – 3.05383, Price – 7165.63593, Highwaympg – 37.63752, Cylindernumber - four, Curbweight – 2004.76641, Enginetype - ohc, Carwidth – 63.94337, Carheight – 52.74095, Carlength – 160.30369, Carbody - hatchback, Drivewheel - fwd, Aspiration - std, Doornumber - two, Carname – honda, nissan, Fueltype - gas, Symboling – 1.07414, Enginelocation - front, Wheelbase – 94.01576.

Klaster 3 ima srednu vrijednost 0.12401 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 105.33018, Compressionratio – 21.26861, Citympg – 25.16373, Peakrpm – 4292.55751, Fuelsystem - idi, Enginesize – 162.67794, Stroke – 3.56189, Boreatio

– 3.57119, Price – 21446.60217, Highwaympg – 28.37618, Cylindernumber - four, Curbweight – 3383.59811, Enginetype - ohc, Carwidth – 69.1138, Carheight – 56.71705, Carlength – 190.66942, Carbody - sedan, Drivewheel - rwd, Aspiration - turbo, Doornumber - four, Carname – 504, 604sl, buick, century, custom, diesel, gs, mazda, mazda rx-7 gs, peugeot, peugeot 504, peugeot 604sl, rx, sw, turbo, volvo, Fueltype - diesel, Symboling - -0.6806, Enginelocation - front, Wheelbase – 109.99689.

Klaster 4 ima srednju vrijednost 0.14632 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 82.66958, Compressionratio – 10.66132, Citympg – 27.46986, Peakrpm – 4898.57746, Fuelsystem – 2bbl, Enginesize – 110.69634, Stroke – 3.23428, Boreatio – 3.31132, Price – 9220.46845, Highwaympg – 33.24761, Cylindernumber - four, Curbweight – 2357.51155, Enginetype - ohc, Carwidth – 65.26155, Carheight – 54.53651, Carlength – 172.66527, Carbody - sedan, Drivewheel - fwd, Aspiration - std, Doornumber - four, Carname – subaru, toyota, volkswagen, Fueltype - gas, Symboling – 0.29662, Enginelocation – front, Wheelbase – 97.65634.

Klaster 5 ima srednju vrijednost 0.15569 te u njega spadaju instance kojima je srednja vrijednost Horsepower – 132.06438, Compressionratio – 8.73454, Citympg – 20.29987, Peakrpm – 5264.99365, Fuelsystem - mpfi, Enginesize – 139.9208, Stroke – 3.36197, Boreatio – 3.45914, Price – 15183.77347, Highwaympg – 26.3939, Cylindernumber - four, Curbweight – 2695.00805, Enginetype - ohc, Carwidth – 66.01546, Carheight – 51.13172, Carlength – 173.88407, Carbody - hatchback, Drivewheel - rwd, Aspiration - std, Doornumber - two, Carname – mitsubishi, toyota, Fueltype - gas, Symboling – 2.58662, Enginelocation - front, Wheelbase – 95.97188.

Klaster 0 ima najveću vrijednost za varijable Horsepower, Enginesize, Boreratio, Price, Curbweight, Carwidth, Carlength, a najmanju vrijednost za varijable Compressionratio, Citympg, Highwaympg.

Klaster 1 ima najmanju vrijednost za varijablu Stroke.

Klaster 2 ima najveću vrijednost za varijable Citympg, Peakrpm, Highwaympg, a najmanju vrijednost za varijablu Horsepower, Enginesize, Boreratio, Price, Curbweight, Carwidth, Carlength, Wheelbase.

Klaster 3 ima najveću vrijednost za varijable Compressionratio, Stroke, Carheight, Wheelbase, a najmanju vrijednost za varijable Peakrpm, Symboling.

Klaster 5 ima najveću vrijednost za varijablu Symboling, a najmanju vrijednost za varijablu Carheight.

Klaster 4 nema niti jednu najveću niti najmanju varijablu. Klaster 0 ima najviše najvećih vrijednosti, dok klaster 2 ima najviše najmanjih vrijednosti.

8. Zaključak

U nastavku slijedi zaključak o samom radu te provedenoj klaster analizi. Znanost o podacima je dobila na značaju u proteklih tridesetak godina te nam uvelike pomaže u radu sa podacima, ali i na donošenju odluka u organizacijama. Rudarenje podataka bazira se na analizi podataka koji su strukturirani te se sam proces odvija kroz nekoliko faza. Najpoznatije i najčešće korištene metode rudarenje podataka u današnje vrijeme su: klasteriranje, stabla odlučivanja, neuronske mreže i regresija.

Klaster analiza nam je u ovom radu omogućila da na vizualni način prikažemo rezultate koji su dobiveni samom analizom. Rezultati koji su dobiveni mogu olakšati tvrtki uvid u grupe automobila koji imaju određene sličnosti i razlike. Kao najoptimalniji klaster odabran je klaster k=6 koji je najbolje zadovoljio postavljene uvjete. Također, možemo vidjeti kako različite specifikacije automobila utječu na cijenu automobila. Npr. iz analize možemo vidjeti kako najveća snaga automobila izražena u konjima (Horsepower), najveća veličina motora (Enginesize) i najveća duljina auta (Carlength) utječu na to da je i cijena najveća. Dok u drugu ruku klaster koji je sadržavao aute koji su najslabiji, s najmanjom veličinom motora i najmanjom duljinom i širinom je imao i najniže cijene.

Kako bi klaster analiza bila što bolje iskorištena bilo bi dobro da se za daljnja istraživanja kombinira s još nekim metodama rudarenja podataka da bi smo dobili preciznije i bolje rezultate.

9. Popis literature

1. About BigML, preuzeto 20.08.2022. s https://bigml.com/about/
2. Bramer, M. (2007). *Principles of Data Mining*. London: Springer.
3. Cox, E. *Fuzzy Modeling and Genetic Algorithms for Dana Mining and Exploration*. Elsevier Inc., 2005.
4. Gajera, P., Gondaliya, A., i Kavathiya, J. (2021). *Old Car Price Prediction with Machine Learning.* Preuzeto 07.08.2022. s www.irjmets.com https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf
5. Garača, Ž. (2008). *Poslovni informacijski sustavi,* Ekonomski fakultet, Split
6. Garača, Ž., i Jadrić, M. (2011). *Rudarenje podataka: Različiti aspekti informacijskog društva.* Split: Ekonomski fakultet u Splitu
7. Gegic, E., Isakovic B., Kečo, D., Kevric, J., i Mašetić, Z. (2019). *Car Price Prediction using Machine Learning Techniques.* Preuzeto 07.08.2022. s https://www.researchgate.net/publication/331994496_Car_price_prediction_using_machine_learning_techniques
8. Kelleher, J. D., i Tierney, B. (2021). *Znanost o podacima.* Zagreb: MATE d.o.o.
9. Kelleher, J. D., Namee B. M., i D' Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics.* Cambridge. MA: The MIT Press.
10. Kumar, M. (2019). *Car Price Prediction Multiple Linear Regression.* Preuzeto 05.08.2022. s https://www.kaggle.com/datasets/hellbuoy/car-price-prediction
11. Makki, S., Mustapha A., Kassim J. M., Gharayebeh, E. H., i Alhazmi M. (2011). *Employing Neural Network and Naive Bayesian Classifier in Mining Dana for Car Evaluation.* Preuzeto 07.08.2022. s https://www.researchgate.net/publication/233831571_Employing_Neural_Network_and_Naive_Bayesian_Classifier_in_Mining_Data_for_Car_Evaluation
12. Olson, D. L., i Delen, D. (2008). *Advanced Data Mining Techniques.* Berlin Heidelberg: Springer – Verlag

Popis slika

Slika 1 Scatterplot prikaz 17
Slika 2 Distribucija i metrika k=3 22
Slika 3 Distribucija i metrika k=4 23
Slika 4 Vizualni prikaz k=6 24
Slika 5 Distribucija i metrika k=6 24
Slika 6 Distribucija k=8 25
Slika 7 Metrika k=8 25
Slika 8 Distribucija k=10 26
Slika 9 Metrika k=10 26
Slika 10 Distribucija k=12 27
Slika 11 Metrika k=12 27

Popis tablica

Tablica 1 Opis kontinuiranih atributa.....14
Tablica 2 Opis kategorijskih atributa 14
Tablica 3 Distribucija vrijednosti atributa 16
Tablica 4 Prikaz klastera.....21
Tablica 5 Prikaz najoptimalnijeg klastera.....32