Primene tehnika klasifikacije na primeru ocene kvaliteta različitih vina

Seminarski rad u okviru kursa Mašinsko učenje
Univerzitet u Novom Sadu
Prirodno-matematički fakultet

Nikola Pujaz - 21m/19

14. septembar 2020.

Sažetak

Sa ciljem uspešne analize kvaliteta hrane i pića, stručnjaci iz različitih oblasti primenjuju specifične metode iz svojih domena. Analiza velikih podataka, zajedno sa mašinskim učenjem omogućava pristup automatskoj proceni kvaliteta hrane i pića sa određenom tačnošću. U ovom radu je prikazan proces analize i pripreme podataka za učenje i evaluaciju više modela mašinskog učenja na primeru klasifikacije crvenog vina na osnovu kvaliteta, kao i sami rezultati evaluacije dobijeni nakon različitih ciklusa pripreme podataka i kreiranja modela.

Sadržaj

1 Uvod

Svest o moći modernih koncepata mašinskog učenja kvalitetno opisuju autori u [7], navodeći ekstremne primere korišćenja tehnika mašinskog učenja kao što su izbor kvalitetnih embriona sa ciljem uspešnog začeća IVF metodom, ili specifičniji izbor jedinki iz stočnog krda koje se klasifikuju za eutanaziju, pri čemu se može primetiti da koncepti mašinskog učenja optimizuju procese u različitim sferama života do razmera uticaja na pitanja života i smrti.

Iako prethodni koncept deluje zastrašujuće, optimistična strana posmatranja na ovu tehnologiju omogućava veliki napredak u poboljšavanju kvaliteta poslovnog procesa, naučnog istraživanja, razumevanja nepoznatog, a pre svega poboljšanja kvaliteta života.

Kao jednu od široko rasprostranjenih metoda mašinskog učenja, koja ima primenu u različitim poljima i oblastima, predstavlja metoda klasifikacije. Klasifikacija ima za cilj automatsku identifikaciju pripadajuće klase sa većom preciznošću nego što je to slučaj u situacijama kada klasifikaciju vrši čovek.

Jedan takav primer klasifikacije se odnosi na posebnu vrstu napitka, čiju proizvodnju ljudi usavršavaju od početaka civilizacije pa sve do danas. Nijedno piće nije bilo element čovekovog realnog i imaginarnog sveta kao što je to vino. Ljubav prema ovom poznatom napitku od grožđa se najbolje ogleda kroz kulturu i nasleđe, posmatrajući lokalne primere narodnih epskih pesama, pa sve do istorije Stare Grčke, koja je vino uzdizala do božanskog pojma.

Danas, vinski stručnjaci poznati kao somelijeri (franc. *Sommelier*), i poznavaoci vina (enofili) vrše kvalitativnu ocenu vina na osnovu boje, mirisa i ukusa, a njihova ocena se bazira na osnovu njihovog iskustva i sposobnosti njihovih fizičkih čula, konkretno čula vida, mirisa i ukusa. Intuitivno se može naslutiti da ukoliko vino ima visoku ocenu od više različitih vinskih stručnjaka, može se smatrati vinom visokog kvaliteta. Postavlja se pitanje, da li postoji precizniji način određivanja kvaliteta vina na osnovu preciznijih parametara koje je moguće kvantifikovati, kao i koji od tih parametara utiču na kvalitet vina? Na osnovu fizičko-hemijskih osobina supstance kao parametara koji opisuju vino i utiču na kvalitet vina, otvara se mogućnost kreiranja modela mašinskog učenja koji uspešno klasifikuje vina prema kvalitetu.

Postoji mnogo različitih pristupa rešavanju ovog problema klasifikacije. Autori u [5] smatraju da je sistem senzora jedan od načina sa najvećim potencijalom razvoja brzih metoda niske cene namenjenih kontroli kvaliteta hrane i pića, te u svom radu ispituju rezultate primene "*elektronskog jezika*"u klasifikaciji određenih sorti italijanskih vina na skupu podataka sličnih atributa kao onog koji će biti korišćen u ovom radu. Sa druge strane, koristeći tehnike mašinskog učenja kao što su linearna regresija, neuronske mreže i SVM(*Support Vector Machine*) autor u [4] pokazuje da se bolje performanse klasifikacije postižu izborom značajnih atributa - atributa koji imaju statistički značaj u poboljšanju rezultata klasifikacije.

Kroz cikluse unapređivanja faze pretprocesiranja podataka i treniranja različitih modela mašinskog učenja, u ovom radu su predstavljeni rezultati uticaja pripreme ulaznih podataka na rezultate klasifikacije sa određenim modelom mašinskog učenja.

2 Analiza i pretprocesiranje podataka

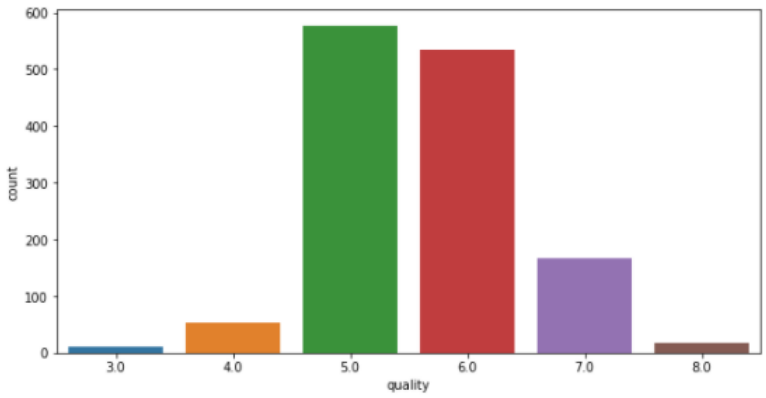
Kako bi cilj ovog rada bio uspešno ispunjen, korišćeno je *Google Colab* okruženje, kao i *scikit-learn* i *pandas* biblioteka, programski jezik *Python* i njegove konstrukcije i druge biblioteke. Implementacija ovog rešenja se nalazi u *Jupiter Notebook* datoteci u prilogu ovog rada.

Skup podataka sadrži podatke o fizičko-hemijskim svojstvima crvenog vina. Skup podataka sadrži ukupno 1599 instanci, koje određuje 12 atributa sledećeg naziva: *'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality'*, pri čemu se poslednji atribut smatra klasnim - određuje klasu instance.

Pristrasnost (eng. *BIAS*) prema određenoj klasi dovodi do povećanja greške klasifikacije, stoga iz skupa podataka uklanjamo sve duplikate instanci kako bi izbegli pristrasnost prema određenoj klasi. Nakon uklanjanja duplikata skup podataka se smanjuje na 1359 instanci.

Različite metode mašinskog učenja na različite načine tretiraju instance kojima nedostaju vrednosti atributa, međutim, iako je zbog čistoće podataka inicijalna namera bila da se instance sa nedostajućim vrednostima izuzmu iz skupa podataka, ustanovljeno je da ovaj skup podataka nema instance sa nedostajućim vrednostima, pa bi korak uklanjanja instanci sa nedostjućim vrednostima bio redudantan.

Pregledom tipova atributa ustanovljeno je da su sve vrednosti atributa u *DataFrame* strukturi tipa *float64*, nakon čega je izvršen pregled broja instanci prema pripadajućim klasama prikazan na slici 1.

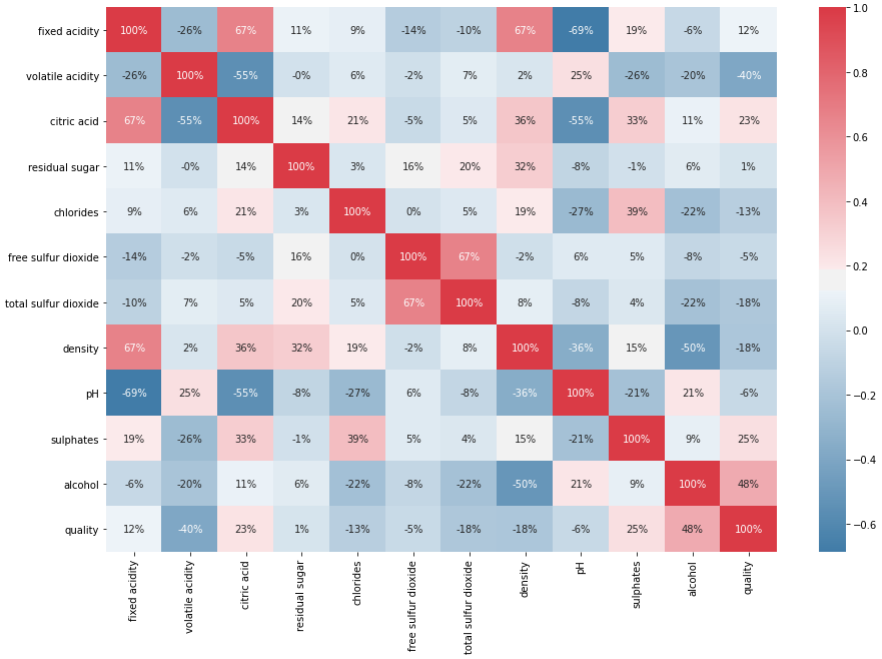


Slika 1: Pregled broja instanci po pripadajućim klasama

Rudimentalno rečeno, modeli mašinskog učenja koji se odnose na klasifikaciju instanci imaju za cilj prepoznavanje obrazaca između podataka koji predstavljaju svojevrsni tip znanja. Veliki broj promenljivih utiče na prepoznavanje ovih obrazaca, stoga je neophodno podatke adekvatno pripremiti pre pristupanju učenju i evauciji modela mašinskog učenja.

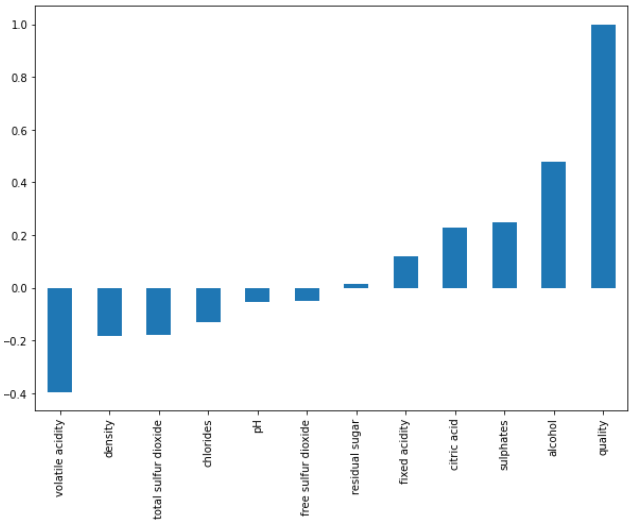
Pojedini algoritmi mašinskog učenja ne pružaju zadovoljavajuće rezultate na podacima veće dimenzionalnosti, odnosno podacima sa velikim brojem atributa. Sa druge strane, manja dimenzionalnost skupa podataka omogućava bolju identifikaciju obrazaca među podacima, pogotovo ukoliko postoji statistička korelacija između pomenutih atributa. U sledećoj

fazi pretprocesiranja napravljen je pregled statističkih korelacija između atributa, korišćenjem Pirsonovog koeficijenta korelacije (eng. *Pearson correlation coefficient*) koji se kreće u intervalu između -1 i +1 i zahteva linearnu povezanost i neprekidnu normalnu distribuciju između promenljivih [1]. Korelacija je prikazana toplotnom kartom na slici 2



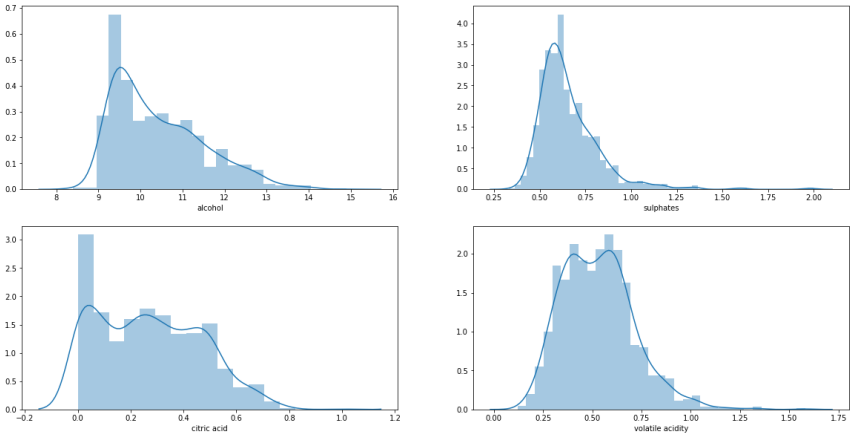
Slika 2: Pregled međusobnih zavisnosti između atributa

Sa toplotne karte se primećuje da postoje atributi koji imaju visoku međusobnu korelaciju. Kako je cilj ovog koraka pretprocesiranja smanjenje dimenzionalnosti skupa podataka, iskoristićemo informaciju o korelaciji između klasnog atributa *'quality'* i ostalih atributa (slika 3) kako bi izdvojili atribute čije vrednosti imaju veću korelaciju sa vrednostim klasnog atributa, i to veću od definisane *threshold* vrednosti koja u ovom slučaju iznosi 0.2. Izdvojena su četiri atributa, *'alcohol','sulphates','citric acid','volatile acidity','quality'*, koji zajedno sa klasnim atributom čine novi skup podataka koji će biti korišćen za učenje i evaluaciju modela.

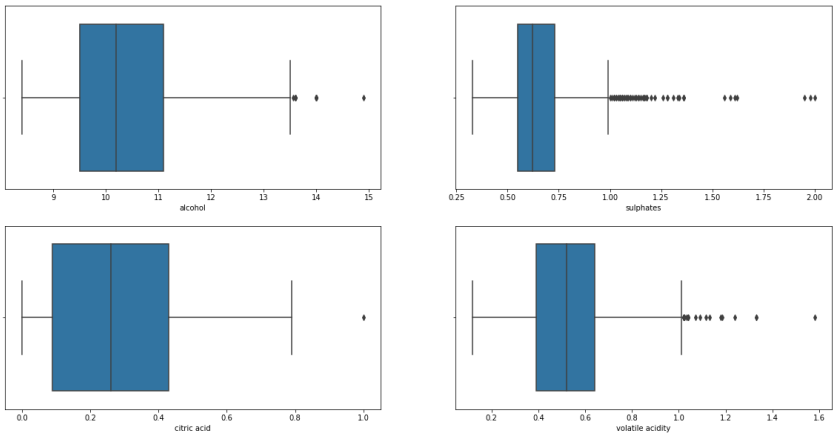


Slika 3: Korelacija između atributa *quality* i ostalih atributa

U narednom koraku pretprocesiranja napravljen je pregled distribucija vrednosti po atributima (slika 4), pri čemu su uočene ekstremne vrednosti na krajevima skupova vrednosti atributa koje potencijalno označavaju prisustvo *outlier* vrednosti, tj. vrednosti koje odstupaju od ostalih podataka. Ove vrednosti mogu značajno da utiču na rezultate učenja i evaluacije modela i postoji više različitih metoda rešavanja problema prisustva ovih vrednosti. Radi bolje vizualizacije *outlier* vrednosti korišćen je *box-and-whisker* dijagram (slika 5), koji prikazuje podatke od donjeg do gornjeg kvartila. Iako se gornja i donja granica mogu različito definisati, najčešće predstavljaju najmanju i najveću vrednost koja se nalazi unutar 1.5 umnoženog interkvartilnog raspona gledajući od donjeg, odnosno gornjeg kvartila. Sve tačke izvan tih granica se smatraju *outlier* vrednostima.[3]



Slika 4: Pregled distribucija vrednosti po atributima



Slika 5: *Box-and-whisker* dijagram

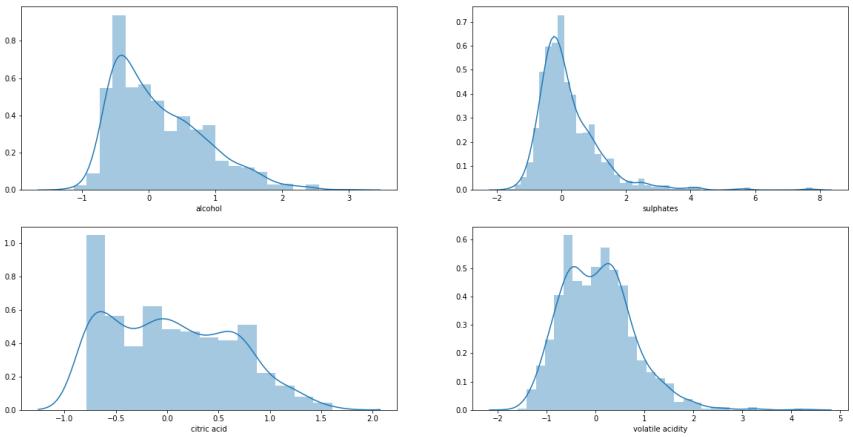
Na *box-and-whisker* dijagramu se može jednostavno uočiti velika količina *outlier* vrednosti kada se posmatraju atributi ,*'sulphates'* i *'volatile acidity'*. Veličina *outlier* vrednosti može biti u direktnoj vezi sa značenjem podataka koje jedino može da tumači domenski ekspert, pružajući informaciju o tome da li su primećene vrednosti zapravo posledica loše čistoće i čuvanja podataka ili ljudske greške. Sa druge strane, količina *outlier* vrednosti može da ide u prilog značaju ovih podataka, pa ih ne treba uklanjati iz skupa. U slučaju kada je količina *outlier* vrednosti mala, kao što je to slučaj sa vrednostima atributa *'citric acid'*, praksa nalaže uklanjanje ovih vrednosti, međutim, kada skup podataka ne sadrži veliku količinu podataka, ove vrednosti se ne uklanjaju već se pristupa drugim tehnikama rešavanja problema prisustva ovih vrednosti.

Uz prisustvo *outlier* vrednosti, kao drugi problem se javlja i različit opseg vrednosti u svakom atributu koji može znatno da utiče na kvalitet učenja i evaluacije modela mašinskog učenja, stoga je neophodno primeniti i određenu tehniku transformacije(normalizacije, sklairanja) podataka.

Biblioteka *scikit-learn* sadrži različite tipove klasa za skaliranje podataka, od kojih se najčešće koriste *StandardScaler i MinMaxScaler*, od kojih prvi ne može da garantuje balansiranu skalu atributa u prisustvu *outlier* vrednosti, dok drugi skalira vrednosti na interval [0,1], pri čemu se zbog prisustva *outlier* vrednosti sve *inlier* vrednosti nalaze u uskom intervalu oko nule, što ga čini osetljivim na prisustvo *outlier* vrednosti. Zbog ovih nedostatka izabran je *RobustScaler* koji svoja izračunavanja bazira na kvantilima, što smanjuje uticaj *outlier* vrednosti.[6][2] Vizualizacija primene *RobustScaler* skaliranja je prikazana na slici 6

Pripremljeni podaci su podeljeni u dva skupa, trening i test skup, od kojih trening skup sadrži 80% podataka, a test skup 20% podataka.

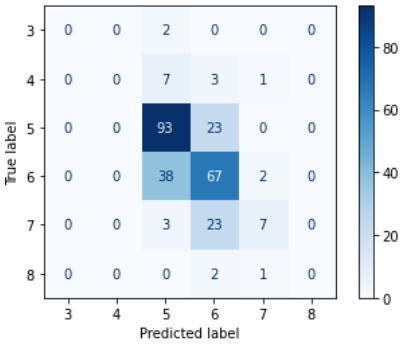
Sa ovim korakom je završen prvi ciklus pretprocesiranja koraka i u sledećoj sekciji su prikazani rezultati učenja i evaluacije regresionog modela.



Slika 6: Pregled distribucija vrednosti nakon skaliranja

3 Logistička regresija

Rezultat učenja i evaluacije modela logističke regresije, za koji su podešeni odgovarajući *multi_class* i *solver* parametri, prikazan je merom tačnosti klasifikacije koja iznosi 61.4%. Matrica konfuzije prikazana je na slici 7, sa koje se broj ispravno klasifikovanih instanci po klasama može pročitati sa glavne dijagonale matrice, dok se broj pogrešno klasifikovanih instanci može pročitati iznad ili ispod glavne dijagonale matrice. Ovaj princip čitanja broja ispravno i pogrešno klasifikovanih instanci važi i za rezultate ostalih klasifikatora u nastavku.



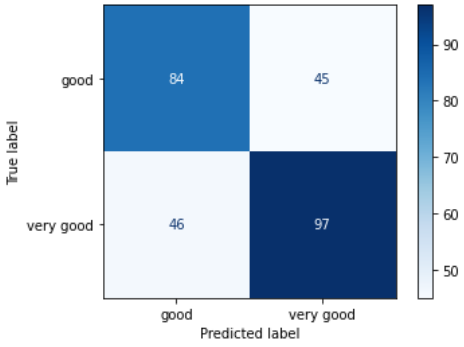
Slika 7: Matrica konfuzije: Logistička regresija

4 Pretvaranje numeričkog atributa u nominalni

Sa ciljem uspešnog kreiranja i evaluacije modela klasifikatora drugog tipa, vrednosti klasnog atributa *'quality'* su pretvorene u nominalne vrednosti po sledećoj šemi: *0-3: 'bad', 3-6: 'good', 6-9: 'very good', 9-10: 'excellent'*. Nakon ovog koraka, kreirana su dva modela, model stabla odlučivanja i probabilistički model Naïve Bayes, za koje su dobijeni rezultati evaluacije prikazani u nastavku.

4.1 Stablo odlučivanja

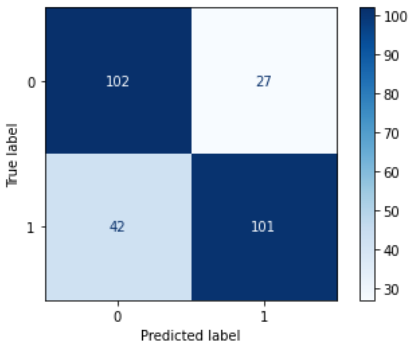
Mera tačnosti klasifikacije stabla odlučivanja iznosi 66.54%. Matrica konfuzije prikazana je na slici 8.



Slika 8: Matrica konfuzije: Stablo odlučivanja

4.2 Naïve Bayes

Pre pristupa treniranju klasifikatora na osnovu preporuka iz [6] i [2], nominalni klasni atribut je enkodiran u odgovarajuće numeričke vrednosti, što u slučaju *DecisionTree* klasifikatora nije bilo potrebno, jer on ima mogućnost rada sa nominalnim klasnim atributom. Mera tačnosti klasifikacije Naïve Bayes algoritma iznosi 74.63%. Matrica konfuzije prikazana je na slici 9.



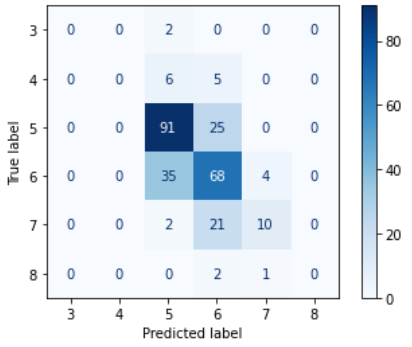
Slika 9: Matrica konfuzije: Naïve Bayes

5 *PCA - (Principal Component Analysis)*

U prethodnom ciklusu pretprocesiranja urađen je izbor značajnih atributa na osnovu korelacije njihovih vrednosti sa vrednostima klasnog atributa. Ovaj ciklus pretprocesiranja pristupa drugoj metodi redukcije dimenzionalnosti skupa podataka koja se naziva Analiza glavnih komponenti (eng. *PCA - (Principal Component Analysis)*). Ova nenadgledana metoda transformacije podataka funkcioniše po principu identifikacije smera varijanse podataka, tj. smera disperzije podataka. Primenom PCA dobijeni su novi atributi, koji se u kontekstu PCA nazivaju glavne komponente, dobijene linearnom kombinacijom prvobitnog skupa podataka. Prva nova komponenta obuhvata najveći deo disperzije skupa podataka, a svaka sledeća obuhvata manji deo disperzije, koji nije pokriven prethodnom komponentom.[7] U ovom slučaju izdvojeno je 7 komponenti koje obuhvataju 95% kumulativnog zbira varijansi ovih komponenti, pri čemu su dobijeni rezultati prikazani u narednim sekcijama.

5.1 Logistička regresija nakon PCA

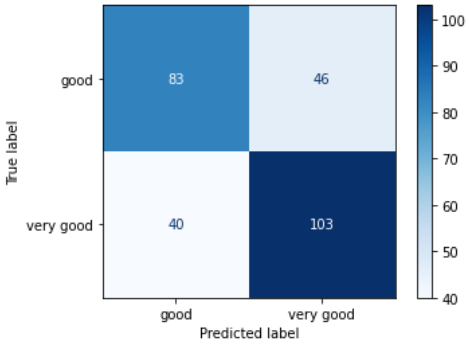
Mera tačnosti klasifikacije stabla odlučivanja nakon PCA iznosi 62.13%. Matrica konfuzije prikazana je na slici 10.



Slika 10: Matrica konfuzije: Logistička regresija nakon PCA

5.2 Stablo odlučivanja nakon PCA

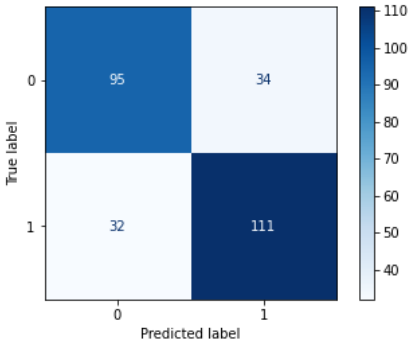
Mera tačnosti klasifikacije stabla odlučivanja nakon PCA iznosi 68.38%. Matrica konfuzije prikazana je na slici 11.



Slika 11: Matrica konfuzije: Stablo odlučivanja nakon PCA

5.3 Naïve Bayes nakon PCA

Mera tačnosti klasifikacije Naïve Bayes algoritma nakon PCA iznosi 75.74%. Matrica konfuzije prikazana je na slici 12.



Slika 12: Matrica konfuzije: Naïve Bayes nakon PCA

6 Zaključak

Posmatrajući rezultate prikazane u tabeli 1, može se primetiti da klasifikacija nakon PCA pruža za nijansu bolje rezultate nego što je to u slučaju izbora atributa na osnovu međusobne korelacije, pri čemu se kao najbolji klasifikator pokazao Naïve Bayes nakon analize glavnih komponenti, čija evaulacija preciznosti iznosi 75.74%. Sa druge strane, mala razlika između tačnosti klasifikacije primenom prvog i drugog ciklusa pretprocesiranja u kojima se na različite načine izvršila redukcija dimenzionalnosti skupa svedoči o uspešnom izboru atributa u prvom ciklusu pretprocesiranja, analizom korelacija klasnog atributa sa ostalim atributima.

Klasifikator	Manuelni izbor atributa	Nakon PCA
Logistička regresija	61.40%	62.13%
Stablo odlučivanja	66.54%	68.38%
Naïve Bayes	74.63%	75.74%

Tabela 1: Pregled tačnosti klasifikacije

U skladu sa prethodnim rezultatima, poboljšanje tačnosti klasifikacije prikazanih klasifikatora se može postići dodatnim podešavanjem klasifikatora, ispitujući tačnost klasifikacije za različite vrednosti parametara. Uz ovo treba napomenuti da je pri izboru parametara potrebna obazrivost, jer izbor pogrešnih parametara može da rezultuje pogoršanju tačnosti klasifikatora, kao i povećanju cene samog procesa klasifikacije.

Dodatno, u fazi pretprocesiranja bi trebalo obratiti pažnju na veliki broj *outlier* vrednosti atributa i izabrati neku drugu metodu rešavanja pristupa istih, kao što je brisanje ili neki drugi tip transformacije, zatim nakon toga proveriti da li dolazi do poboljšanja performansi klasifikatora.

Literatura

[1] C.C. Aggarwal. *Data Mining: The Textbook*. Springer International Publishing, 2015.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[3] Krstić Dijana. Istraživačka analiza podataka (EDA) uz upotrebu statističkog softvera R. Master rad, Univerzitet u Novom Sadu, Prirodno-matematički fakultet, Trg Dositeja Obradovića 3, Novi Sad, Srbija, 6 2016.

[4] Yogesh Gupta. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125:305–312, 2018.

[5] A. Legin, A. Rudnitskaya, L. Lvova, Yu. Vlasov, C. Di Natale, and A. D'Amico. Evaluation of italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta*, 484(1):33–44, May 2003.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] I.H. Witten, E. Frank, M.A. Hall, and C.J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2016.