



УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У
НОВОМ САДУ



Катарина Тополић

ПРИМЕНА МАШИНСКОГ УЧЕЊА ЗА ПРЕДИКЦИЈУ СРЧАНИХ БОЛЕСТИ

ДИПЛОМСКИ РАД
- Основне академске студије -

Нови Сад, 2024

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, РБР:	
Идентификациони број, ИБР:	
Тип документације, ТД:	Монографска публикација
Тип записа, ТЗ:	Текстуални штампани документ
Врста рада, ВР:	Дипломски рад
Аутор, АУ:	Катарина Тополић
Ментор, МН:	Проф. др Дарко Чапко
Наслов рада, НР:	Примена машинског учења за предикцију срчаних болести
Језик публикације, ЈП:	Српски/латиница
Језик извода, ЈИ:	Српски
Земља публикавања, ЗП:	Србија
Уже географско подручје, УГП:	Војводина
Година, ГО:	2024.
Издавач, ИЗ:	Ауторски репринт
Место и адреса, МА:	Факултет Техничких Наука (ФТН), Д. Обрадовића 6, 21000 Нови Сад
Физички опис рада, ФО: <small>(поглавља/страна/ цитата/табела/слика/графика/прилога)</small>	7/33/36/2/16/0/0
Научна област, НО:	Електротехничко и рачунарско инжењерство
Научна дисциплина, НД:	Рачунарски управљачки системи
Предметна одредница/Кључне речи, ПО:	Машинско учење
УДК	
Чува се, ЧУ:	Библиотека ФТН, Д. Обрадовића 6, 21000 Нови Сад
Важна напомена, ВН:	
Извод, ИЗ:	У оквиру рада имплементирани су алгоритми машинског учења са циљем предикције срчаних болести. Коришћене су четири методе – логистичка регресија, метод потпорних вектора, К најближих суседа и стабло одлуке.
Датум прихватања теме, ДП:	
Датум одбране, ДО:	12.09.2024.
Чланови комисије, КО:	Председник: Др Срђан Вукмировић, ред. проф.
	Члан: Доц. Немања Недић
	Члан, ментор: Др Дарко Чапко, ред. проф.
	Потпис ментора

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	Monographic publication
Type of record, TR :	Textual material, printed
Contents code, CC :	Bachelor thesis
Author, AU :	Katarina Topolić
Mentor, MN :	Darko Čapko, PhD
Title, TI :	Heart disease prediction using machine learning algorithms
Language of text, LT :	Serbian/Latin
Language of abstract, LA :	Serbian
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2024.
Publisher, PB :	Author reprint
Publication place, PP :	Faculty of Technical Sciences, D. Obradovića 6, 21000 Novi Sad
Physical description, PD : (chapters/pages/ref./tables/pictures/graphs/appendixes)	7/33/36/2/16/0/0
Scientific field, SF :	Electrical and computer engineering
Scientific discipline, SD :	Computing and control engineering
Subject/Key words, S/KW :	Machine learning
UC	
Holding data, HD :	Library of the Faculty of Technical Sciences, D. Obradovića 6, 21000 Novi Sad
Note, N :	
Abstract, AB :	The paper describes the implementation of machine learning algorithms with the aim of predicting heart diseases. Four methods were used – logistic regression, support vector machine, K nearest neighbors and decision tree.
Accepted by the Scientific Board on, ASB :	
Defended on, DE :	12.09.2024.
Defended Board, DB :	President: Srđan Vukmirović, PhD, full professor
	Member: Nemanja Nedić, doc.
	Member, Mentor: Darko Čapko, PhD, full professor

	Menthor's sign



УНИВЕРЗИТЕТ У НОВОМ САДУ • ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА
21000 НОВИ САД, Трг Доситеја Обрадовића 6

Број:

Датум:

ЗАДАТАК ЗА ЗАВРШНИ РАД

(Податке уноси предметни наставник - ментор)

СТУДИЈСКИ ПРОГРАМ:	РАЧУНАРСТВО И АУТОМАТИКА
РУКОВОДИЛАЦ СТУДИЈСКОГ ПРОГРАМА:	Др Милан Рапаић, р. проф.

Студент:	Катарина Тополић	Број индекса:	РА 164/2020
Област:	Електротехничко и рачунарско инжењерство		
Ментор:	Др Дарко Чапко		

НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ – МАСТЕР РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА:

- проблем – тема рада;
- начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна;

НАСЛОВ ДИПЛОМСКОГ РАДА:

Примена машинског учења за
предикцију срчаних болести

ТЕКСТ ЗАДАТКА:

- Представити скуп података коришћен у раду
- Описати поступак прављења модела машинског учења
- Описати коришћене алгоритме машинског учења за предвиђање срчаних болести
- Приказати и анализирати резултате коришћених метода

Руководилац студијског програма:	Ментор рада:

Примерак за: о - Студента; о - Ментора;

Spisak korišćenih skraćenica

ANOVA	Analysis of variance
SVM	Support vector machine
KNN	K-nearest neighbors
TP	True positive
TN	True negative
FP	False positive
FN	False negative

Sadržaj

1 Uvod.....	1
2 Teorijske osnove.....	2
2.1 Kardiovaskularne bolesti.....	2
2.2 Mašinsko učenje.....	2
2.3 Opis skupa podataka.....	4
3 Predobrada podataka.....	6
3.1 Tretiranje nedostajućih vrednosti.....	6
3.2 Otkrivanje i obrada anomalija.....	7
3.3 Konverzija kategorijskih podataka u numeričke vrednosti.....	10
4 Eksplorativna analiza podataka.....	11
4.1 Odabir najbitnijih atributa.....	11
5 Kreiranje klasifikacionog modela.....	17
5.1 Logistička regresija.....	17
5.2 Metod potpornih vektora.....	18
5.3 K najbližih suseda.....	21
5.4 Stablo odluke.....	22
6 Analiza rezultata.....	25
6.1 Mere kvaliteta modela.....	25
6.2 Rezultati klasifikacije.....	26
7 Zaključak.....	29
Literatura.....	30
Biografija.....	33

1 Uvod

Srčane bolesti predstavljaju jedan od vodećih uzroka smrti širom sveta, čineći pravovremenu dijagnostiku i prevenciju ključnim izazovima u medicini. Prema statistici Svetske zdravstvene organizacije (WHO), približno 17.9 miliona smrtnih slučajeva u svetu tokom jedne godine su posledica srčanih bolesti [1]. U poslednjih nekoliko godina, primena mašinskog učenja u medicini omogućila je značajan napredak u identifikaciji rizičnih faktora i predikciji bolesti na osnovu velikih skupova podataka. Mašinsko učenje, kao grana veštačke inteligencije, omogućava sistemima da uče iz podataka, identifikuju obrasce i donose odluke uz minimalnu ljudsku intervenciju. Ovi algoritmi mogu analizirati velike količine medicinskih podataka, prepoznati skrivena pravila i uzorke, te predvideti rizik od razvoja određenih bolesti sa visokim stepenom tačnosti. Cilj ovog rada je istraživanje različitih klasifikacionih modela mašinskog učenja u predikciji srčanih bolesti, uz fokus na evaluaciju njihovih performansi i identifikaciju najpouzdanijih modela

Zadatak ovog rada bio je napraviti modele mašinskog učenja za predikciju srčanih bolesti i analizom performansi različitih modela utvrditi koji od njih je najpouzdaniji. Drugo poglavlje detaljnije objašnjava šta su to kardiovaskularne bolesti, mašinsko učenje i skup podataka koji je korišćen. Treće poglavlje rada opisuje procese korišćene za predobradu podataka – tretiranje nedostajućih vrednosti, identifikacija i obrada anomalija i konverzija kategorijskih u numeričke vrednosti. Četvrto poglavlje je posvećeno eksplorativnoj analizi podataka gde se pomoću grafika i različitih statističkih metoda dolazi do informacije koji atributi iz skupa podataka su najbitniji. Takođe grafički prikaz doprinosi boljem razumevanju podataka i veza između njih. U petom poglavlju su opisani svi algoritmi klasifikacije koji su korišćeni za izradu modela, parametri koji se definišu prilikom pravljenja svakog i validacija koja je ključan korak u procesu razvoja mašinskog učenja. U šestom poglavlju su predstavljani i analizirani rezultati svakog modela i zaključeno je koji model ima najbolje performanse a koji najgore.

2 Teorijske osnove

2.1 Kardiovaskularne bolesti

Kardiovaskularne bolesti (KVB) obuhvataju širok spektar stanja koja utiču na srce i krvne sudove, čineći ih jednim od vodećih uzroka smrti širom sveta. Ova stanja su odgovorna za milione smrtnih slučajeva svake godine, pri čemu su glavni uzročnici srčani udari, moždani udari i druge vaskularne bolesti. Približno 640 miliona ljudi boluje od neke vrste srčanog oboljenja [2]. Najistaknutija među ovim bolestima je koronarna arterijska bolest, koja nastaje kada krvni sudovi koji snabdevaju srce postanu suženi ili blokirani, što dovodi do smanjenog protoka kiseonika do srčanog mišića. Ovo stanje često izaziva bol u grudima, poznat kao angina, ili, u ozbiljnijim slučajevima, srčani udar.

Faktori rizika za kardiovaskularne bolesti su višestruki, uključujući i životni stil i genetske komponente. Hipertenzija (visok krvni pritisak), visoki nivoi holesterola, pušenje, fizička neaktivnost, gojaznost i dijabetes su dobro poznati doprinosi razvoju kardiovaskularnih bolesti [3]. Genetski faktori takođe mogu predisponirati pojedince na ova stanja, posebno u slučajevima kada postoji porodična istorija ranih srčanih bolesti. Pored toga, godine i pol igraju značajnu ulogu, pri čemu rizik značajno raste kod osoba starijih od 50 godina i češće kod muškaraca, iako žene u postmenopauzi takođe doživljavaju nagli porast rizika [4][5].

U poslednjih nekoliko godina, napredak u medicinskoj nauci i tehnologiji poboljšao je razumevanje kardiovaskularnih bolesti i obezbedio alate za rano otkrivanje i lečenje. Međutim, prevencija je i dalje ključna, posebno u smislu promena životnog stila, poput održavanja zdrave ishrane, redovnog vežbanja, izbegavanja duvana i kontrole stanja poput dijabetesa i hipertenzije. Ipak, predviđanje nastanka i napredovanja kardiovaskularnih bolesti ostaje izazov, s obzirom na složenost faktora koji su uključeni. Ovde mašinsko učenje počinje da igra transformativnu ulogu, posebno u oblasti prediktivne medicine.

2.2 Mašinsko učenje

Mašinsko učenje je grana veštačke inteligencije koja se bavi razvojem algoritama i statističkih modela koji omogućavaju računarima da "uče" iz podataka i donose odluke ili predikcije bez eksplicitnog programiranja za svaki zadatak [6]. U osnovi, mašinsko učenje se fokusira na kreiranje modela koji identifikuju obrasce u podacima, uče iz tih obrazaca i koriste ovo učenje za donošenje odluka ili predikcija. Mašinsko učenje je revolucionisalo brojne industrije, uključujući zdravstvo, omogućavajući sistemima da uče iz podataka i donose odluke uz minimalnu ljudsku intervenciju.

Polje mašinskog učenja može se podeliti na nekoliko kategorija, od kojih svaka ima različite pristupe i primene. Nadgledano učenje (eng. supervised learning) je najčešći oblik, gde se model obučava na obeleženim podacima, što znači da su ulazni podaci upareni sa tačnim ishodima [6]. Odnosno, skup podataka nad kojima model uči se sastoji i od ulaza i od željenih izlaza. Na primer, u slučaju predikcije srčanih bolesti, model bi učio iz skupa podataka u kojem su karakteristike pacijenata (starost, krvni pritisak, holesterol, itd.) povezane sa poznatim ishodima (prisustvo ili odsustvo srčane bolesti). Kada se model obučuje, može predviđati verovatnoću srčanih bolesti kod novih, neviđenih pacijenata na osnovu njihovih karakteristika.

S druge strane, nenadgledano učenje (eng. unsupervised learning) podrazumeva obučavanje modela na podacima koji nisu obeleženi, što znači da skup podataka ne sadrži obeležene izlaze. Cilj ovde nije predviđanje ishoda, već prepoznavanje skrivenih šablona ili struktura, grupa u podacima [6]. Na primer, nenadgledano učenje može se koristiti za klasifikaciju pacijenata u različite kategorije rizika na osnovu njihovih zdravstvenih profila, bez potrebe za eksplicitnim etiketama kao što su "visok rizik" ili "nizak rizik". Uglavnom se koristi za identifikaciju anomalija. Polu-nadgledano učenje (eng. semi-supervised learning) predstavlja kombinaciju prethodne dve metode. Ova metoda koristi mali broj označenih izlaza zajedno sa velikim brojem neoznačenih podataka [6]. Cilj ove metode je da se što bolje obučuje model u slučaju kada su označeni podaci skupi ili teško dostupni, dok su neoznačeni podaci relativno lako dostupni. Još jedna grana, učenje sa podsticajem (eng. reinforcement learning), podrazumeva modele koji uče da donose niz odluka interakcijom sa okolinom i dobijanjem povratnih informacija u obliku nagrada ili kazni [6]. Iako se češće povezuje sa robotikom i igrama, učenje sa podsticajem ima obećavajuće primene u optimizaciji strategija lečenja u zdravstvu.

Mašinsko učenje u zdravstvu, a posebno u predikciji srčanih bolesti, ima ogroman potencijal zbog svoje sposobnosti da obradi velike količine podataka i otkrije složene šablone koji mogu izmaći tradicionalnim statističkim metodama. U oblasti predikcije srčanih bolesti, algoritmi mašinskog učenja se koriste za izradu modela koji mogu predvideti da li je pacijent u riziku od razvoja srčane bolesti na osnovu različitih faktora poput starosti, pola, nivoa holesterola i drugih kliničkih parametara. Analizom istorijskih podataka o pacijentima, ovi modeli mogu pomoći lekarima da identifikuju osobe sa visokim rizikom ranije nego što bi to omogućile konvencionalne metode.

Pored toga, modeli mašinskog učenja mogu poboljšati donošenje odluka u kliničkoj praksi kroz unapređenje stratifikacije rizika, što podrazumeva kategorizaciju pacijenata na osnovu verovatnoće razvoja srčane bolesti. To omogućava personalizovaniju negu, gde pacijenti sa visokim rizikom mogu biti pomno praćeni i dobijati preventivne tretmane, dok pacijenti sa niskim rizikom možda zahtevaju manje intenzivno praćenje. Takođe, mašinsko učenje se sve više koristi u medicinskom snimanju za otkrivanje abnormalnosti na srčanim snimcima, čime se dodatno poboljšava dijagnostička tačnost.

2.3 Opis skupa podataka

Skup podataka korišćen u ovom radu za predikciju srčanih bolesti preuzet je sa Kaggle sajta i sadrži podatke o 918 ispitanika [7]. Sadrži nekoliko ključnih varijabli koje su poznati faktori rizika za kardiovaskularne bolesti. Razumevanje ovih varijabli je ključno za izgradnju modela mašinskog učenja i tumačenje njegovih predikcija.

- **Starost (Age):** Starost je značajan faktor u srčanim bolestima, pri čemu su stariji pojedinci u većem riziku. Kako ljudi stare, njihove arterije prirodno postaju tvrđe, a srce kao i svi ostali mišići postaje manje efikasno, što čini starije pacijente sklonijim kardiovaskularnim problemima [4].
- **Pol (Sex):** Pol igra ulogu u riziku od srčanih bolesti, pri čemu su muškarci generalno u većem riziku nego žene, posebno pre nego što žene dostignu menopauzu. Postmenopauzalne žene, međutim, doživljavaju oštar porast rizika od srčanih bolesti zbog hormonskih promena [5].
- **Vrsta bolova u grudima (ChestPainType):** Vrsta bolova u grudima koju pacijent doživljava je ključna dijagnostička karakteristika. Tipična angina (TA) je bol u grudima koja se javlja pri naporu i popušta nakon odmora, što često ukazuje na koronarnu arterijsku bolest. Atypical angina (ATA) je manje specifična, dok je neangijalni bol (NAP) obično nepovezan sa srčanom bolešću. Asimptomatski (ASY) pacijenti, uprkos odsustvu simptoma, mogu imati srčanu bolest [8].
- **Krvni pritisak u mirovanju (RestingBP):** Krvni pritisak u mirovanju je ključni pokazatelj kardiovaskularnog zdravlja. Povišen krvni pritisak (hipertenzija) može oštetiti arterije tokom vremena, što vodi srčanim bolestima [9]. Ovaj skup podataka meri krvni pritisak u mirovanju u milimetrima žive (mm Hg).
- **Holesterol (Cholesterol):** Visoki nivoi holesterola, posebno lipoproteina niske gustine (LDL), doprinose razvoju ateroskleroze, nakupljanja plaka u arterijama [10]. Praćenje nivoa holesterola je stoga od esencijalnog značaja za procenu rizika od srčanih bolesti.
- **Šećer u krvi (FastingBS):** Nivoi šećera u krvi pružaju uvid u metabolizam glukoze pacijenta. Visok nivo šećera u krvi, posebno kada prelazi 120 mg/dL, ukazuje na dijabetes, što značajno povećava rizik od kardiovaskularnih bolesti [11].
- **EKG u mirovanju (RestingECG):** Ova varijabla beleži rezultate EKG-a pacijenta u mirovanju, koji meri električnu aktivnost srca. Abnormalnosti u EKG-u, kao što su ST-T talasne promene ili hipertrofija leve komore (LVH), mogu ukazivati na osnovnu srčanu bolest.
- **Maksimalna srčana frekvencija/brzina srca (MaxHR):** Maksimalna srčana frekvencija postignuta tokom fizičke aktivnosti je indikator kardiovaskularne kondicije. Niže

maksimalne srčane frekvencije mogu ukazivati na loše zdravlje srca i povećan rizik od srčanih bolesti.

- Angina pri vežbanju (ExerciseAngina): Ova varijabla beleži da li pacijent doživljava anginu (bol u grudima) tokom vežbanja. Angina izazvana vežbanjem je indikator koronarne arterijske bolesti [12].
- Stari maksimum (Oldpeak): Oldpeak se odnosi na depresiju ST segmenta na EKG-u, merenu u odnosu na odmor. Viša vrednost Oldpeak-a ukazuje na ozbiljniju ishemiju, gde srčani mišić ne dobija dovoljno krvi tokom napora [13].
- ST nagib (ST_Slope): Nagib ST segmenta na EKG-u tokom vežbanja je još jedna važna dijagnostička karakteristika. Uspon ST segmenta je generalno znak zdravog srca, dok ravni ili silazni segmenti ukazuju na povećan rizik od koronarne arterijske bolesti [14].
- Srčana bolest (HeartDisease): Ovo je ciljna varijabla u skupu podataka, gde vrednost 1 označava prisustvo srčane bolesti, a 0 označava odsustvo srčane bolesti.

Svaka od ovih varijabli pruža vredne informacije koje se mogu koristiti za izgradnju modela mašinskog učenja za predikciju srčanih bolesti. Kombinacija ovih karakteristika omogućava modelu da uhvati i tradicionalne faktore rizika (kao što su starost i holesterol) i složenije kliničke indikatore (kao što su rezultati EKG-a i vrsta bolova u grudima), poboljšavajući ukupnu prediktivnu moć.

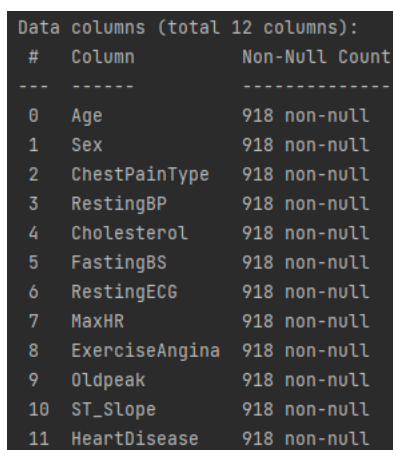
3 Predobrada podataka

Predobrada podataka je ključni korak u svakom projektu analize podataka i primene mašinskog učenja. Kvalitet predobrade može značajno uticati na performanse modela te je važno pažljivo sprovesti sve neophodne korake. Sastoji se od operacija koje su namenjene da transformišu podatke u struktuisan format pogodan za pravljenje modela mašinskog učenja. U ovom poglavlju detaljno će se opisati proces predobrade podataka za predikciju srčanih bolesti, uključujući tretiranje nedostajućih vrednosti, identifikaciju i obradu anomalija, konverziju kategorijskih podataka u numeričke vrednosti.

3.1 Tretiranje nedostajućih vrednosti

Nedostajuće vrednosti u skupu podataka mogu značajno uticati na performanse modela mašinskog učenja. Postoji nekoliko pristupa za tretiranje nedostajućih vrednosti. Jedan od pristupa je uklanjanje redova ili kolona koji sadrže nedostajuće vrednosti. Ovaj pristup se koristi kada nedostajuće vrednosti ne čine značajan deo podataka. Još jedan od načina je imputacija vrednosti. Popunjavanje nedostajućih vrednosti se vrši korišćenjem srednje vrednosti ostalih podataka. Ovaj pristup se koristi kada nedostajuće vrednosti čine značajan deo podataka [15].

Primenom metode `info()` dobijamo informacije prikazane na slici 1.



```
Data columns (total 12 columns):
#   Column      Non-Null Count
---  -
0   Age         918 non-null
1   Sex         918 non-null
2   ChestPainType 918 non-null
3   RestingBP   918 non-null
4   Cholesterol 918 non-null
5   FastingBS   918 non-null
6   RestingECG  918 non-null
7   MaxHR       918 non-null
8   ExerciseAngina 918 non-null
9   Oldpeak     918 non-null
10  ST_Slope    918 non-null
11  HeartDisease 918 non-null
```

Slika 1: Rezultat poziva metode `info()`

Skup podataka nema nedostajuće vrednosti. Moguće je još detaljnije ispitati podatke kako bi se možda uočila neka nepravilnost bez detaljnijih analiza. Na slici 2 prikazane su neke od specifičnih vrednosti kategoričkih varijabli.

	count	mean	std	min	25%	50%	75%	max
Age	918.0	53.510893	9.432617	28.0	47.00	54.0	60.0	77.0
RestingBP	918.0	132.396514	18.514154	0.0	120.00	130.0	140.0	200.0
Cholesterol	918.0	198.799564	109.384145	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	0.233115	0.423046	0.0	0.00	0.0	0.0	1.0
MaxHR	918.0	136.809368	25.460334	60.0	120.00	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.00	0.6	1.5	6.2

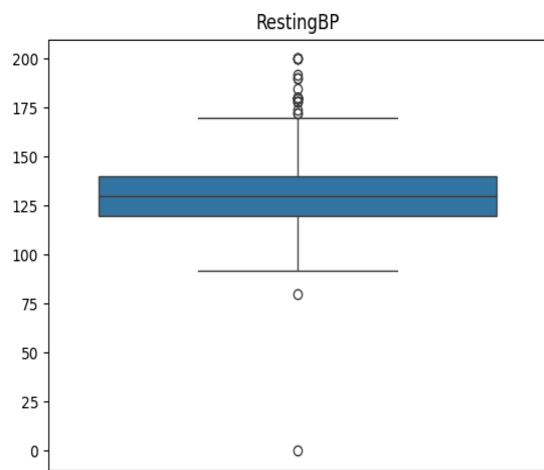
Slika 2: Specifične vrednosti kategoričkih varijabli

Jedna od nepravilnosti je ta što holesterol ima minimalnu vrednost 0, što nije realna vrednost. Ovo jedino odskake u odnosu na ostale varijable, tako da je korisno istražiti bolje. Prebrojavanjem kolona u kojima je vrednost holesterola jednaka 0, dobija se informacija da je ta vrednost zabeležena kod 171 pacijenta. Ovo može ukazati na grešku pri merenju ili prosto nedostatak informacije. Pošto je 171 relativno velik udeo od ukupno 918 podataka u skupu, preciznije 18.65% ukupnih uzoraka, uklanjanje vrednosti ne bi bio dobar izbor. Umesto toga se sve vrednosti jednake 0 zamenjuju srednjom vrednošću ostalih podataka za holesterol.

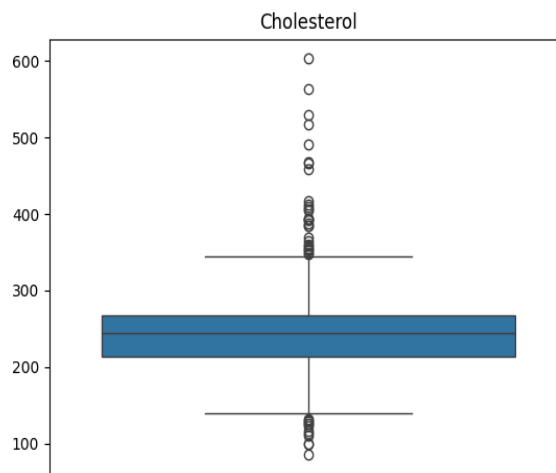
3.2 Otkrivanje i obrada anomalija

Anomalije (eng. outliers) su podaci koji značajno odstupaju od drugih vrednosti u skupu [16]. Važno je identifikovati ih i pravilno obraditi jer njihovo prisustvo može negativno uticati na performanse modela. Postoji nekoliko metoda za otkrivanje anomalija.

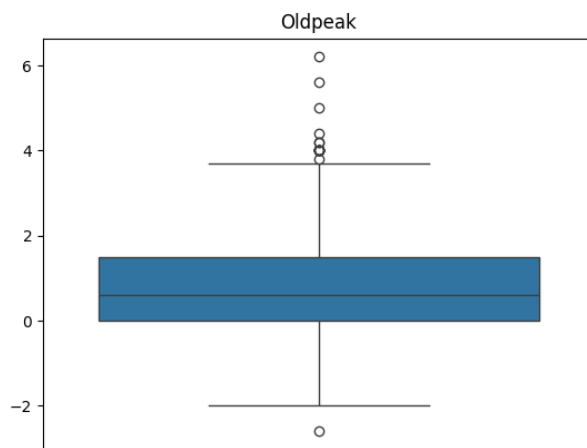
Jedna od metoda je vizuelizacija. Grafički prikazi pokazuju raspodelu podataka i omogućavaju lako uočavanje vrednosti koje odstupaju. Najčešće se koristi kvantilni graf (eng. boxplot). Kvantilni graf je standardizovani način prikaza distribucije podataka na osnovu pet brojeva odnosno veličina. Minimum predstavlja najmanju vrednost u podacima, ne računajući anomalije. Prvi kvartil ili prva četvrtina, označava se sa Q1, je tačka ispod koje se nalazi 25% podataka. Medijana, označava se sa Q2, je srednja vrednost podataka. Treći kvartil, označava se sa Q3, je tačka ispod koje se nalazi 75% podataka. Maksimum je najveća vrednost u podacima, ne računajući anomalije [17]. Kvantilni graf ove informacije vizuelno prikazuje i olakšava detekciju anomalija. Na slikama 3, 4 i 5 su prikazani kvantilni grafovi sa anomalijama za promenljive „RestingBP“, „Cholesterol“ i „Oldpeak“.



Slika 3: Kvantilni graf raspodele vrednosti promenljive "RestingBP"



Slika 4: Kvantilni graf raspodele vrednosti promenljive "Cholesterol"



Slika 5: Kvantilni graf raspodele vrednosti promenljive "Oldpeak"

Graf je strukturisan na sledeći način:

- kutija (eng. box) predstavlja međukvartilni raspon, koji se označava sa IQR, i predstavlja razliku između prvog kvartila (Q1) i trećeg kvartila (Q3),
- unutar kutije nalazi se linija koja označava medijanu
- „brkovi“ (eng. whiskers) su linije koje se prostiru od ivica kutije do minimuma i maksimuma

Veličina kutije pokazuje na raspon srednjih 50% podataka. Što je ona veća to znači da je veća varijabilnost među podacima. Ako je medijana bliža dnu ili vrhu kutije, to može sugerisati na asimetričnost podataka. Anomalije su prikazane kao pojedinačne tačke van „brkova“.

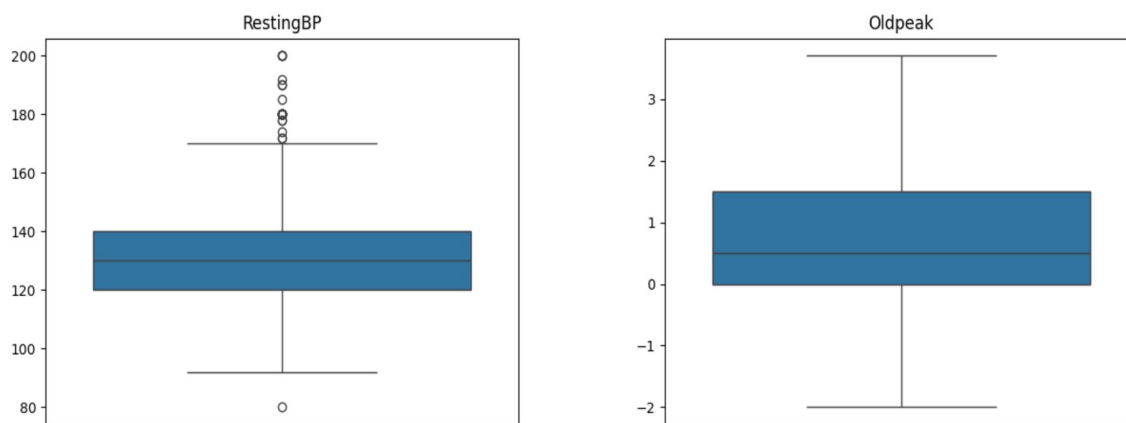
Nakon detekcije anomalija, postoji nekoliko načina da se tretiraju. U nekim slučajevima može biti prikladno ukloniti anomalije, posebno ako su rezultat grešaka u unosu podataka ili merenju. Posle analize anomalija, može se zaključiti da su one greške jer odstupaju od realnih slučajeva. Na slici 3 se vidi da postoji vrednost krvnog pritiska koja je jednaka 0, što u stvarnosti nije moguće, tako da ima smisla izbaciti tu vrednost. Vidi se i da ima ekstremno visokih vrednosti koje dosežu i do 200, međutim kako je to moguće te vrednosti neće biti izbačene. Holesterol takođe ima dosta anomalija, međutim kako u opisu skupa podataka nije specificirano koji je tip holesterola, izbacivanje podataka bi moglo biti pogrešno. Kada je u pitanju prethodni maksimum, odnosno „Oldpeak“, normalan opseg vrednosti za ovu veličinu je od 0 do 2mm [18]. Moguće je i da ima negativne vrednosti, zato su izbačene vrednosti koje se nalaze izvan kutije. To se radi tako što se izbace sve vrednosti koje su veće od maksimalne i manje od minimalne vrednosti. Ako se zna da je $IQR = Q_3 - Q_1$, gornja granica se računa na sledeći način:

$$\text{gornja granica} = Q_3 + 1.5 \cdot IQR,$$

a donja:

$$\text{donja granica} = Q_1 - 1.5 \cdot IQR,$$

Iako se postave uslovi izbacivanja vrednosti van izračunatih. Na slici 6 je prikazana raspodela vrednosti promenljivih „RestingBP“ i „Oldpeak“ nakon izbacivanja anomalija.



Slika 6: Kvantilni graf raspodele vrednosti promenljivih "RestingBP" i "Oldpeak" nakon izbacivanja anomalija

3.3 Konverzija kategorijskih podataka u numeričke vrednosti

Kategorijske promenljive ili atributi mogu se svrstati u dve grupe, nazivne (eng. nominal) ili redne (eng. ordinal). Nazivne promenljive imaju dve ili više kategorija koje nisu povezane specifičnim redosledom, kao na primer pol osobe. Sa druge strane redne promenljive imaju određene nivoe ili kategorije sa tačno određenim redosledom, kao na primer nivoi nizak, srednji i visok, važno je da idu baš tim redosledom [19]. Većina algoritama mašinskog učenja zahteva numeričke ulazne podatke, stoga kategorijske varijable (tipa string) treba konvertovati u numeričku vrednost. U ovom radu je korišćena Label Encoding metoda. Ova metoda konvertuje kategorijske promenljive u niz binarnih vrednosti [19]. Svakoј јединственој kategoriji u koloni se dodeljuje јединствен broj. Na primer, za kolonu „Pol“ koja ima vrednosti „M“ i „F“, metoda će dodeliti vrednost 0 za „M“ a 1 za „F“.

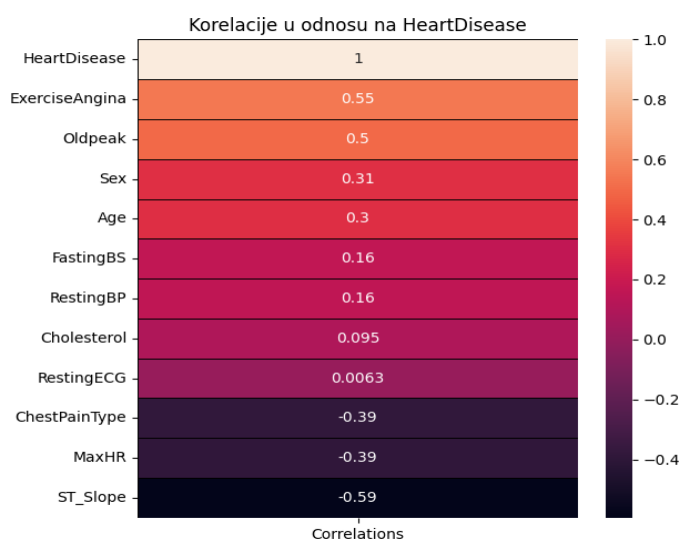
4 Eksplorativna analiza podataka

Eksplorativna analiza podataka je krucijalan korak koji za cilj ima razumevanje samog skupa podataka kao i veze između atributa. Fokusira se na analizu najvažnijih atributa pomoću statističkih metoda i vizuelizacije. Uobičajene tehnike uključuju generisanje sažetih statistika, kreiranje grafova poput histograma, dijagrama rasipanja, kao i ispitivanje korelacija među varijablama [20]. Iako je eksplorativna analiza podataka korak koji zahteva dosta vremena, veoma je bitan jer pravi selekciju najbitnijih promenljivih i time pravi temelj za pravljenje robustnog i tačnog modela.

4.1 Odabir najbitnijih atributa

Pre klasifikacije, potrebno je napraviti selekciju najbitnijih atributa. Ovo omogućava smanjenje dimenzionalnosti i time vremena izvršavanja programa. Takođe, uklanjanje nepotrebnih atributa rezultuje boljim performansama modela. Postoji više metoda koje se koriste u ove svrhe, konkretno u ovom radu su korišćene korelaciona matrica, Hi-kvadratni test i analiza varijanse.

Korelaciona matrica prikazuje korelacije odnosno veze između promenljivih u skupu podataka. Svako polje u matrici pokazuje korelaciju između dve promenljive i može imati vrednost od -1 do 1. Ako polje ima vrednost 1 to znači da su promenljive u savršenoj pozitivnoj korelaciji, ako se jedna varijabla povećava i druga će se takođe povećavati i to linearno. Ako polje ima vrednost -1 to znači da su promenljive u savršenoj negativnoj korelaciji, ako se jedna promenljiva povećava, druga će se smanjivati [21]. Na slici 7 prikazana je korelaciona matrica za korišćeni skup podataka.



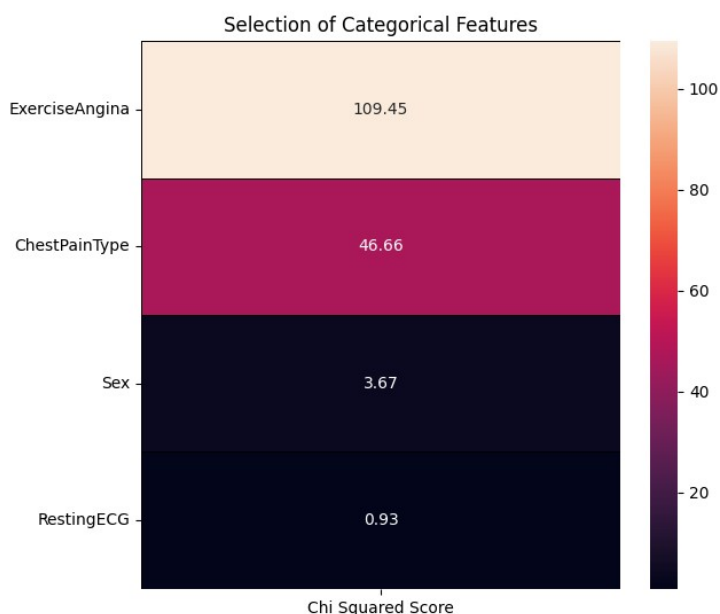
Slika 7: Korelaciona matrica

Na osnovu matrice se vidi da ciljna promenljiva „HeartDisease“ ima negativnu korelaciju sa promenljivom „MaxHR“ dok sa „ExerciseAngina“, „OldPeak“ i „Sex“ ima pozitivnu korelaciju. Korelaciona matrica pomaže da se olakša pregled najbitnijih podataka, to je posebno korisno kada podaci sadrže veliki broj atributa. Na taj način se i smanjuje vreme i složenost procesa.

Pored korelacione matrice, postoje i metode koje posebno ispituju korelacije kategorijskih i numeričkih vrednosti sa ciljnom promenljivom. Hi-kvadratni test (χ^2) ispituje vezu između kategorijskih nezavisnih promenljivih i zavisne promenljive. To je statistički test, zasniva se na Hi-kvadratnoj raspodeli. Hi-kvadratni test upoređuje stvarne (posmatrane) frekvencije u kategorijskim podacima sa očekivanim frekvencijama koje bi se javile da ne postoji povezanost između atributa i ciljne varijable [22]. Hi-kvadratni statistički podatak, koji se računa formulom:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

meri razliku između posmatrane (O_i) i očekivane (E_i) frekvencije za svaki atribut. Visoka vrednost Hi-kvadratnog statističkog podatka ukazuje na značajnu razliku između posmatranih i očekivanih frekvencija, što sugerise snažniju povezanost između karakteristike i ciljne klase. Ovo omogućava selekciju karakteristika koje imaju značajan uticaj na proces klasifikacije. Rezultat se može videti na slici 8.



Slika 8: Rezultat Hi-kvadratnog testa

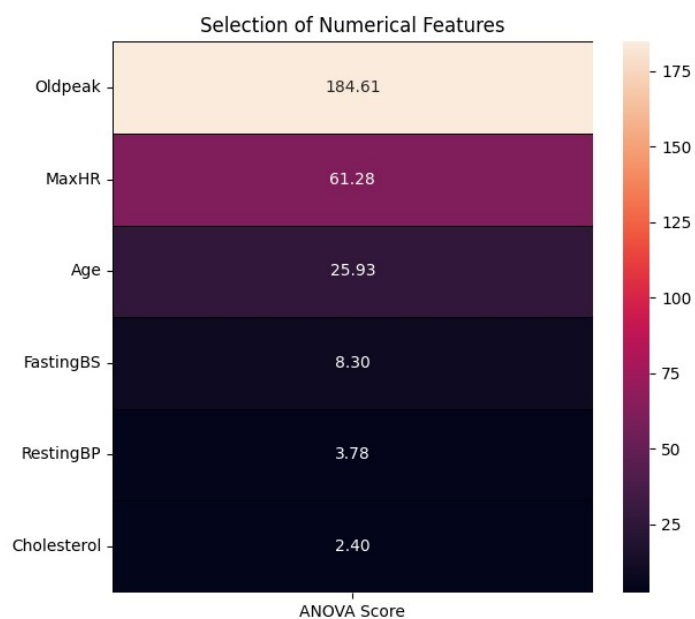
Jedna od ključnih prednosti Hi-kvadratnog testa u selekciji karakteristika je njegova jednostavnost i računarska efikasnost. On je posebno efektivan kada se radi sa velikim skupovima podataka koji sadrže kategorijske promenljive. Test pomaže u smanjenju dimenzionalnosti podataka, birajući samo one karakteristike koje imaju značajnu statističku vezu sa ciljnim promenljivama, čime se poboljšavaju performanse i interpretabilnost klasifikacionog modela [22]. Međutim, postoje i određena ograničenja pri korišćenju Hi-kvadratnog testa. On se može primeniti samo na kategorijske podatke, takođe, on podrazumeva da su karakteristike međusobno nezavisne, što nije uvek slučaj u realnim skupovima podataka.

ANOVA (Analysis of Variance), ili analiza varijanse, je statistička metoda koja se koristi za upoređivanje srednjih vrednosti više od dve grupe i utvrđivanje da li postoji značajna razlika između njih [23]. U kontekstu selekcije atributa za klasifikacione modele, ANOVA je posebno korisna kada se radi sa kontinuiranim podacima, jer pomaže u identifikaciji atributa koji imaju značajan uticaj na ciljnu varijablu. ANOVA funkcioniše tako što analizira varijansu unutar svake grupe i varijansu između grupa kako bi se testirala nulta hipoteza da su srednje vrednosti različitih grupa jednake. Ako je varijansa između grupa značajno veća od varijanse unutar grupa, nulta hipoteza može biti odbačena, što ukazuje da atribut koji se ispituje ima značajan uticaj na ciljnu varijablu. Statistički podatak koji se koristi u ANOVA testu je F-proporcija, koji se računa kao odnos između varijanse između grupa i varijanse unutar grupa:

$$F = \frac{\text{varijansa između grupa}}{\text{varijansa unutar grupa}}$$

Veći F-proporcija sugerise da je varijansa između srednjih vrednosti grupa veća u poređenju sa varijansom unutar grupa, što implicira da je atribut verovatno važan prediktor u klasifikacionom modelu.

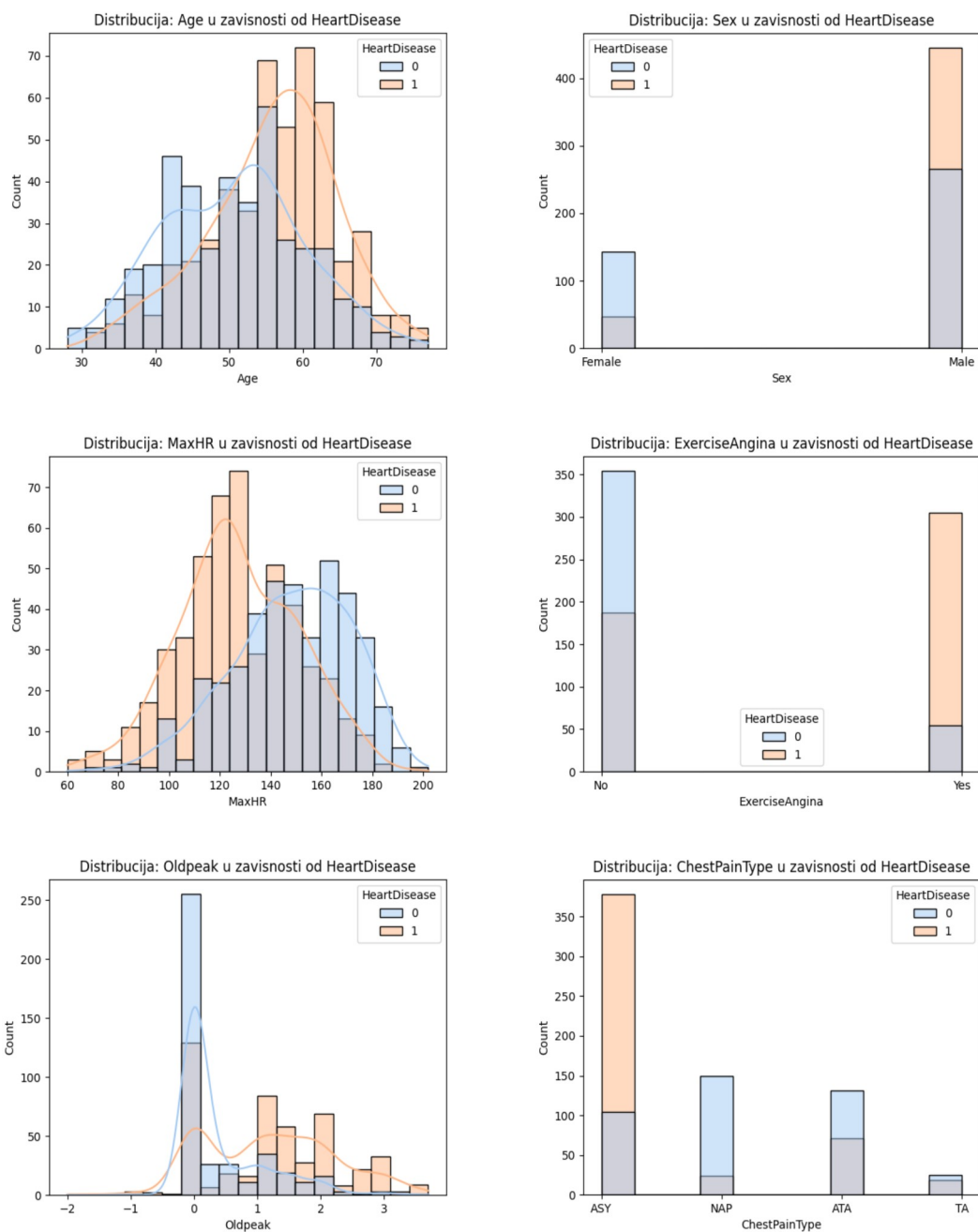
ANOVA je posebno vredna u selekciji atributa jer omogućava procenu više atributa istovremeno. Ona može identifikovati koji atributi značajno doprinose razlikama u ciljnoj varijabli, čime se smanjuje dimenzionalnost skupa podataka i poboljšava efikasnost klasifikacionog modela. Međutim, ANOVA ima i svoja ograničenja. Jedna od glavnih pretpostavki je da su podaci unutar svake grupe normalno distribuirani i da su varijanse među grupama jednake. Kada ove pretpostavke nisu ispunjene, rezultati ANOVA testa mogu biti pogrešni. Pored toga, ANOVA je osetljiva na ekstremne vrednosti, odnosno anomalije, koje mogu negativno uticati na rezultate testa [23]. Na slici 9 prikazani su rezultati ANOVA testa.



Slika 9: Rezultat ANOVA testa

Iz rezultata se vidi da ciljna promenljiva ima najveću korelaciju sa „Age“, „Sex“, „MaxHR“, „ExcerciseAngina“, „Oldpeak“ i „ChestPainType“. Na slici 10 je prikazana zavisnost ciljne varijable u odnosu na izabrane promenljive. Na taj način se jasnije vidi raspodela vrednosti promenljivih. Takođe, analizom se dolazi korisnih informacija koje od tih vrednosti su kritične odnosno koje imaju najveću verovatnoću da boluju od srčanih bolesti.

Takođe iz rezultata testova za najbitnije attribute se vidi da „RestingBP“, „RestingECG“, „Cholesterol“ imaju najmanje korelacije sa ciljnom promenljivom, što može značiti da nemaju velik uticaj na krajnji rezultat i potencijalno je moguće da se izostave prilikom pravljenja modela.



Slika 10: Zavisnost ciljne promenljive i najbitnijih atributa

Analizom grafika se može zaključiti da najveću sklonost srčanim oboljenjima imaju osobe:

- koje imaju između 55 i 63 godine
- koje su muškog pola
- kojima je maksimalan dostignut puls između 110 i 130
- koje su imale „ExerciseAngina“
- koje za „Oldpeak“ imaju vrednosti od 0 do 2
- koje su imali asimptomatski bol u grudima

5 Kreiranje klasifikacionog modela

Odabir algoritma za klasifikaciju je veoma bitan i utiče na performanse modela. Generalno ne postoje metode koje su bolje ili lošije, sve imaju različite performanse i biraju se u zavisnosti od problema. Zato je bitno definisati šta je cilj modela i koja performansa je za to najbitnija, da li tačnost, preciznost ili drugo. U ovom radu korišćene su četiri klasifikacione metode – logistička regresija, metod potpornih vektora, K najbližih suseda i stablo odluke.

5.1 Logistička regresija

Logistička regresija je jedan od fundamentalnih algoritama mašinskog učenja koji se koristi za klasifikacione probleme. Najčešće se koristi za binarnu klasifikaciju odnosno za odgovaranje na „da/ne“ pitanja. Takva vrsta logističke regresije naziva se binomijalna jer zavisna promenljiva ima samo dve moguće vrednosti. Pored nje postoji i multinomijalna gde zavisna promenljiva ima više od dve moguće vrednosti. Uprkos imenu ne koristi se za regresiju – opisivanje veze između dobijenih ulaza i očekivanog izlaza, već za predviđanje verovatnoće binarnog ishoda kao što je da li pacijent boluje od bolesti ili ne. Metoda je cenjena zbog svoje jednostavnosti, efikasnosti i brzine. Međutim nije primenjiva u svakom slučaju jer za neke probleme je suviše jednostavna, takođe podložna je preprilagođavanju i zahteva linearne odnose, što za većinu realnih problema nije slučaj.

Logistička regresija modeluje vezu između ulaznih atributa (odnosno nezavisnih promenljivih) i binarnog ishoda (zavisne promenljive) tako što primenjuje aktivacionu funkciju (ili drugim rečima funkciju prenosa) na linearnu kombinaciju ulaznih promenljivih [24]. Svrha aktivacione funkcije je da ulazne vrednosti mapira na izlaznu vrednost, koja se nalazi u opsegu $[0, 1]$ i na taj način odredi kojoj klasi pripadaju. Najčešće se koristi sigmoidna funkcija [25], koja se definiše kao:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

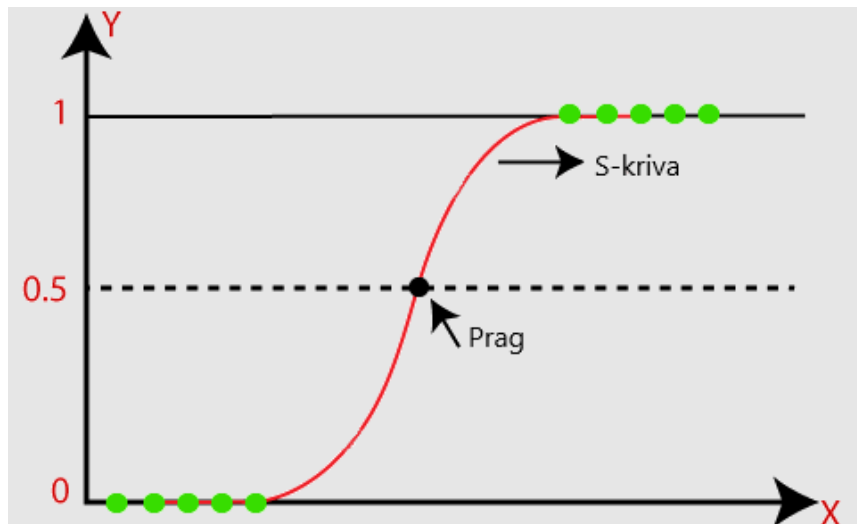
Gde je z linearna kombinacija nezavisnih promenljivih X i njihovih odgovarajućih težina β , odnosno:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Nakon primene sigmoidne funkcije, izlazna vrednost $P(y=1|X)$ predstavlja verovatnoću da ulaz X pripada pozitivnoj klasi ($y=1$). Ova verovatnoća se koristi za donošenje konačne odluke o klasifikaciji. Ako je verovatnoća veća od 0.5, model predviđa da ulaz pripada klasi 1, inače pripada klasi 0. Matematički, primena sigmoidne funkcije na dati skup podataka može se napisati kao:

$$P(y=1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

Grafički prikaz ove funkcije daje karakterističnu S-krivu, koja pokazuje kako se verovatnoće kreću od 0 do 1 sa promenom ulaznih karakteristika. Na slici 11 se vidi da vrednosti iznad praga pripadaju klasi 1, a ispod klasi 0.



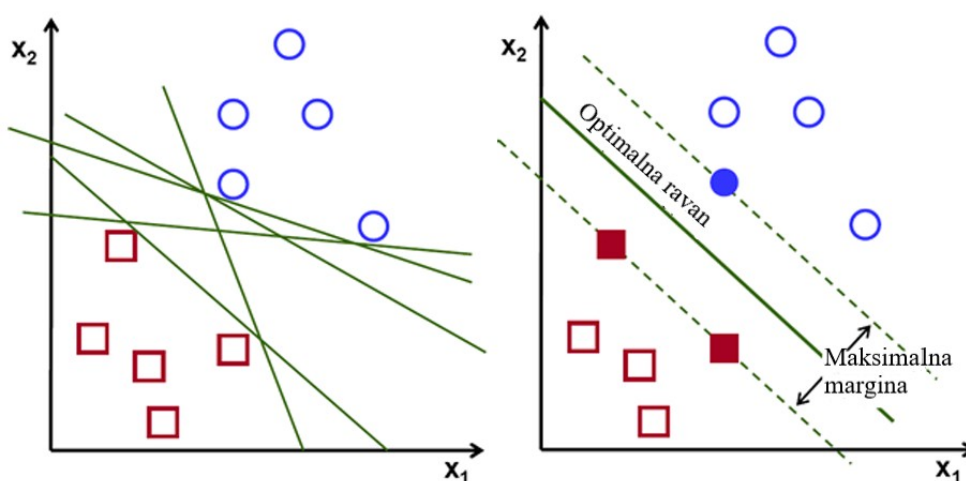
Slika 11: Grafik S-krive [26]

Parametri koji se definišu tokom pravljenja modela su „penalty“, „C“ i „solver“ [27]. „Penalty“ određuje tip regularizacije ili kazne koju treba upotrebiti da bi se izbeglo preprilagođavanje. Parametar „C“ predstavlja snagu regularizacije, manja vrednost ovog parametra ukazuje na veću regularizaciju. Poslednji parametar „solver“ određuje koji će se algoritam koristiti. Ako je skup podataka manji ili je problem binarna klasifikacija, pogodnije je koristiti „liblinear“ algoritam, dok je za veće skupove podataka ili multinomijalnu klasifikaciju bolji „saga“ algoritam.

5.2 Metod potpornih vektora

Metod potpornih vektora (eng. support vector machine) je jedan od važnijih metoda mašinskog učenja. Zasnovan je na jasnoj geometrijskoj intuiciji. Pretpostavlja se da postoje dve klase tačaka u ravni, pri čemu su klase takve da se između elemenata te dve klase može povući prava, tako da su svi elementi jedne klase sa jedne strane, a elementi druge klase sa druge strane [25]. Ovaj uslov linearne razdvajivosti nije realističan uslov, ali za sad se pretpostavlja kao validan. Primetno je, nakon crtanja različitih rasporeda takvih tačaka, da prava koja ih razdvaja retko kada bude jedinstvena, već je moguće povući više njih. Ovo je prikazano na slici 12. Ipak,

neke prave deluju bolje od ostalih. Na istoj slici je prikazana i optimalna prava, što je prava sa najvećim rastojanjem do najbliže joj tačke podataka, odnosno sa najširim pojasom praznog prostora oko nje – margine [25]. Intuitivno, posmatrajući sliku, prava koja bi bila pod drugačijim uglom i prolazila bliže nekoj od tačaka podataka bi nosila veći rizik da neka tačka koja nije u datim podacima završi sa pogrešne strane prave. Cilj ove metode je upravo da se nađe optimalna hiperravan koja će najbolje razdvojiti klase. Ideja je da se maksimizuje margina između hiperravni i najbližih tačaka iz svake klase, odnosno potpornih vektora. Time se poboljšava tačnost modela i njegova sposobnost generalizacije.



Slika 12: Prikaz rada metode potpornih vektora [25]

Jednačina hiperravni je

$$\omega \cdot x + \omega_0 = 0$$

gde je ω_0 slobodan član. Optimalna hiperravan, odnosno hiperravan najšireg pojasa ili margine je podjednako udaljena od najbližih predstavnika obe klase. Stoga, hiperravni paralelne optimalnoj se mogu opisati jednačinama

$$\omega \cdot x + \omega_0 = c$$

$$\omega \cdot x + \omega_0 = -c$$

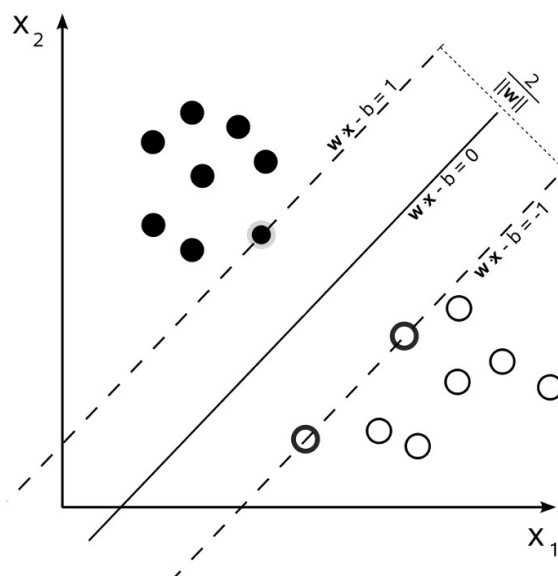
Deljenjem jednačina sa c , za neke nove koeficijente ω i ω_0 za koje zadržavamo iste oznake, dobijaju se jednačine sve tri hiperravni:

$$\omega \cdot x + \omega_0 = 0$$

$$\omega \cdot x + \omega_0 = 1$$

$$\omega \cdot x + \omega_0 = -1$$

Na slici 13 je grafički prikaz prethodno pomenutih jednačina.



Slika 13: Optimalna hiperravan i paralelne hiperravni koje leže na potpornim vektorima [25]

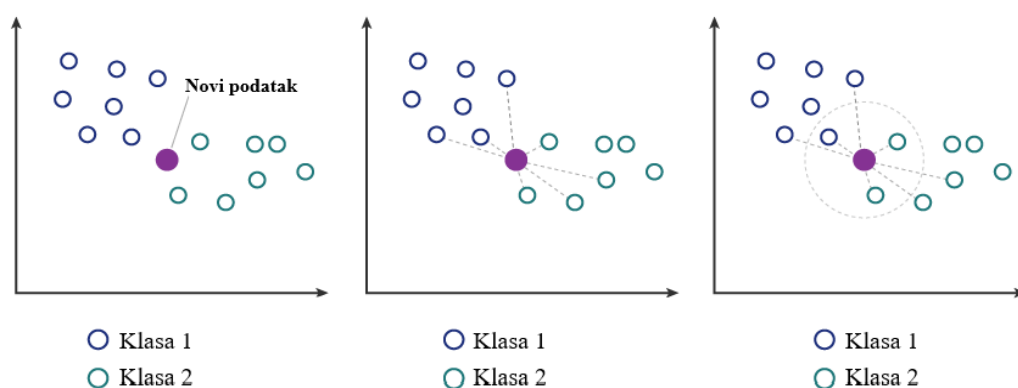
Ukoliko se zna da su podaci linarno razdvojivi, neophodno je u trening fazi pronaći optimalnu hiperravan, odnosno ravan sa maksimalnom marginom koja se računa:

$$m = \frac{2}{\|w\|}$$

Parametri koji se definišu za ovaj model su „C“ i „gamma“ [28]. „C“ ili parametar regularizacije kontroliše balans između maksimizacije margine i minimizacije greške klasifikacije. Ako je vrednost parametra velika model će se više fokusirati na to da tačno klasifikuje podatak, dok će se margina smanjivati. Ako je vrednost mala, model će se fokusirati na to da margina bude što veća što može da rezultira pogrešnom klasifikacijom. „Gamma“ ili parametar kernela određuje koliko će površina (hiperravan) koja razdvaja klase biti zakrivljena, odnosno koliko će model biti precizan u definisanju granice između različitih klasa. Ako je vrednost parametra velika to znači da će granica biti kriva što može izazvati preprilagođavanje, dok ako je vrednost mala granica će biti ravna što može dovesti do pogrešne klasifikacije.

5.3 K najbližih suseda

K najbližih suseda je jedan od jednostavnijih i najčešće korišćenim algoritama za klasifikacione probleme. Cilj ove metode je da uvrsti novi podatak u odgovarajuću klasu na osnovu K najbližih suseda odnosno tačaka. Ona se zasniva na pretpostavci da se podaci koji su slični, koji pripadaju istoj klasi, nalaze međusobno blizu u prostoru podataka [25]. Koraci algoritma su sledeći – prvo se zadaje ceo, prirodan broj K. Zatim se računa Euklidsko rastojanje između nove tačke i ostalih u prostoru, rangirajući ih od najmanjeg do najvećeg. Iz tog skupa uzima se K tačaka koje imaju najmanje rastojanje od neklasifikovane tačke. U zavisnosti od toga kojoj klasi pripada najviše tačaka iz tog podskupa, ta klasa se dodeljuje novom podatku [29]. Način rada ovog algoritma prikazan je na slici 14.



Slika 14: Prikaz rada algoritma K najbližih suseda [30]

Jedan od izazova ove metode je odabir parametra K. Ovaj korak je ključan jer značajno utiče na odnos pristrasnosti (eng. bias) i varijanse. Ove dve veličine uveliko utiču na sposobnost modela da generalizuje podatke i zbog toga je bitno dobro ih razumeti. Pristrasnost se odnosi na grešku uvedenu aproksimacijom problema iz stvarnog sveta, koji može biti veoma složen, sa pojednostavljenim modelom. U suštini, pristrasnost meri koliko predviđanja modela odstupaju od pravih vrednosti. Neke od posledica velike pristrasnosti su sledeće: model je suviše jednostavan da bi prepoznao obrasce u podacima, model se nedovoljno prilagođava podacima što rezultuje lošim performansama i nad obučavajućim skupom i nad skupom za testiranje. Varijansa se odnosi na osetljivost modela na male promene u podacima za obučavanje. Model sa visokom varijansom prevelik značaj daje podacima u skupu za obučavanje i previše se fokusira na šumove ili manje bitne detalje umesto na prepoznavanje obrazaca u podacima. Ovo dovodi do preprilagođavanja, model će imati malu grešku tokom obučavanja ali veliku tokom validacije jer nije naučio da dobro generalizuje nove, do sad neviđene, podatke. Potrebno je odabrati K tako da se postigne balans između pristrasnosti i varijanse. Mala vrednost parametra K rezultira malom pristrasnošću i visokom varijansom, što znači da se model dobro prilagođava podacima za obuku

ali je osetljiv na šumove i ne ponaša se dobro nad novim podacima. Velika vrednost parametra K rezultira visokom pristrasnošću i malom varijansom, što znači da model u ovom slučaju bolje generalizuje ali je moguće da je suviše jednostavan i moguće nedovoljno prilagođen [31]. Optimalno K bi obezbedilo malu pristrasnost i malu varijansu. Jedna od tehnika za dobijanje optimalne vrednosti parametra K je unakrsna validacija. Suštinski, procenjuju se performanse za različite vrednosti K i na kraju se uzima ona vrednost koja je dala najbolje performanse.

Parametri koji se definišu prilikom pravljenja modela su „ $n_neighbors$ “, „ $weights$ “ i „ $metric$ “ [32]. „ $N_neighbors$ “ određuje broj suseda, odnosno K . Za binarnu klasifikaciju za vrednost K obavezno se uzima neparan broj. „ $Weights$ “ predstavlja težinsku funkciju koja se koristi prilikom predviđanja. Postoji *uniformna* gde sve tačke imaju jednak značaj, dok se kod *daljinske* funkcije bliže tačke više vrednuju/cene. „ $Metric$ “ parametar određuje koja metrika će se koristiti za računanje udaljenosti tačaka. *Euklidska* je za većinu slučajeva dobar izbor. Kada su u pitanju višedimenzionalni ili rasejani podaci bolje je koristiti *menhetn* metriku.

5.4 Stablo odluke

Stablo odluke je vrsta grafa, zbog toga je potrebno uvesti nekoliko termina kako bi se jasnije opisala metoda. Graf je struktura podataka koja se koristi da prikaže veze između objekata. Sastoji se od dve glavne komponente – čvorova i ivica. Čvorovi predstavljaju objekte ili entitete unutar grafa dok ivice povezuju čvorove. Grafovi mogu biti usmereni ili neusmereni. Kod usmerenih ivice imaju određene smerove od jednog čvora ka drugom. Ako od čvora a postoji direktna putanja do čvora b , u tom slučaju čvor a predstavlja roditeljski čvor a čvor b njegovog potomka ili naslednika. U slučaju neusmerenih grafova, ivice nemaju određen smer i veza između čvorova je dvosmerna.

Stablo odluke je usmeren graf sa n čvorova i $n-1$ ivica, odnosno grana, gde svaki čvor ima tačno jednog roditelja, sem jednog. Koren je čvor koji nema roditelja i on predstavlja početnu tačku. Čvorovi unutar drveta nazivaju se čvorovi odluke. Svaki od njih odgovara atributu unutar skupa podataka. Drvo se gradi tako što se rekurzivno dele podaci, prateći različite putanje na osnovu odluke na svakom čvoru. Ivice ili grane povezuju čvorove i predstavljaju ishode odluka koje su donete na čvorovima. Ovaj postupak se ponavlja dok se svi podaci ne klasifikuju ili dok se ne ispune neki od uslova zaustavljanja. Listovi predstavljaju terminalne čvorove i oni nemaju naslednike. Nazivaju se i čvorovi odgovora jer predstavljaju rešenja dobijena iz stabla odluke, odnosno konačnu klasu koju je model predvideo.[31]

Jedna od ključnih prednosti stabla odluke je njegova transparentnost i interpretabilnost. Ovaj model je veoma intuitivan jer oponaša način na koji ljudi prirodno donose odluke – postavljanjem serije pitanja, a zatim grananjem u različitim smerovima na osnovu odgovora. Za razliku od nekih složenijih modela, stablo odluke pruža jasnu sliku o tome kako se dolazi do određene odluke. Lako je pratiti tok odluke od korena do lista, što omogućava korisnicima da vizualizuju proces donošenja odluka. Ovo je posebno korisno u aplikacijama gde korisnici žele

da razumeju na osnovu kojih faktora je dobijen rezultat, na primer, u medicinskoj dijagnostici ili pravnim odlukama. Međutim, iako je stablo odluke intuitivno i lako za razumevanje, ono može biti podložno preprilagođavanju. Zbog toga se često koristi tehnika obrezivanja stabla (eng. pruning), koja uklanja nepotrebne grane kako bi se smanjila složenost modela i poboljšala njegova sposobnost da generalizuje.

Parametri koji se definišu prilikom pravljenja modela su „criterion“, „gini“, „entropy“, „max_depth“ i „min_samples_split“ [34]. Parametar „criterion“ određuje funkciju koja se koristi za merenje kvaliteta podele u stablu. „Gini“ funkcija meri „nečistoću“ čvora odnosno verovatnoću da je došlo do pogrešne klasifikacije, dok „entropy“ meri količinu informacija koja se dobija podelom. „Max_depth“ odnosno maksimalna dubina određuje maksimalan broj podela od korena do lista. „Min_samples_split“ određuje minimalan broj uzoraka potreban za podelu unutrašnjeg čvora. Ako je količina uzoraka u unutrašnjem čvoru manja od definisane minimalne, onda će taj čvor postati terminalni odnosno list.

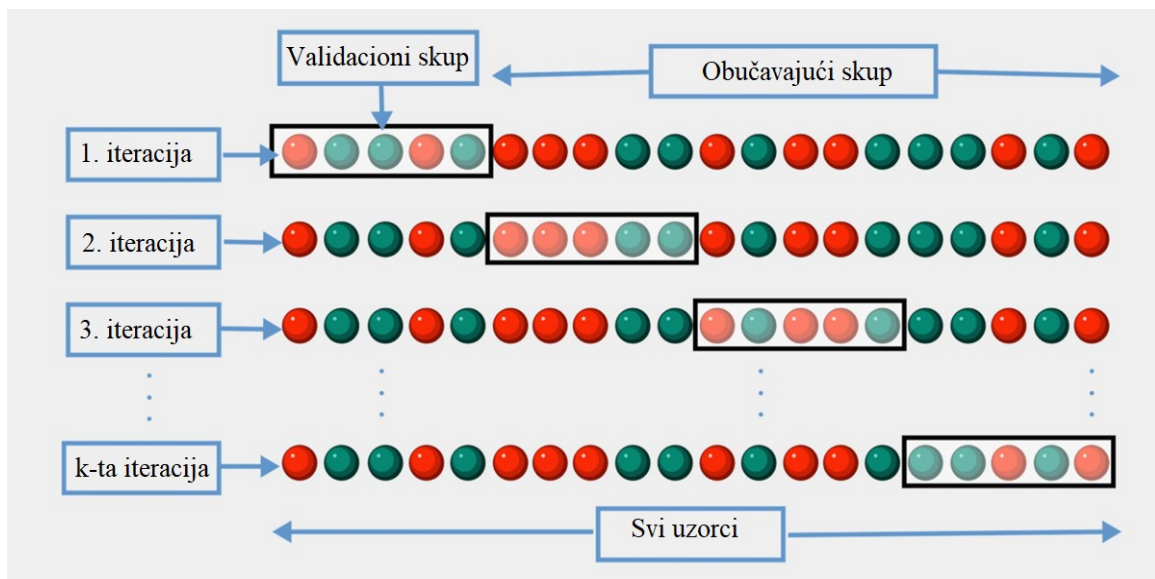
Nakon formiranja modela, sledeći korak je njegovo obučavanje. To se vrši pozivom funkcije *fit* kojoj se prosleđuju skup podataka za obučavanje odnosno trening i skup njihovih izlaza. Model pokušava da prepozna odnos između prosleđenih podataka i izlaza i zaključi od čega zavisi izlaz. Međutim, na ovaj način se izvršava samo jedna iteracija obučavanja, što nije dovoljno ako je cilj da se postignu što bolje performanse. Zato je poslednja faza obučavanja modela validacija.

Validacija je ključni korak u procesu razvoja modela mašinskog učenja i neophodna je kako bi se osiguralo da model može da pruži tačne i pouzdane rezultate u stvarnim situacijama. Omogućava procenu sposobnosti modela da postiže dobre rezultate nad nepoznatim, novim podacima, odnosno da generalizuje [35]. Njena osnovna svrha je da obezbedi da ne dođe do preprilagođavanja (eng. overfitting) modela. U tom slučaju model se suviše prilagodi uzorcima za obuku i nije u stanju da generalizuje nad novim, nepoznatim podacima jer je obučen da dobro prepoznaje isključivo podatke na kojima se obučavao. U mašinskom učenju, cilj je razviti model koji ne samo da postiže visoke performanse na podacima na kojima je treniran, već i na podacima koje nikada ranije nije video. Kada se model preprilagodi, to znači da je postao previše specifičan za podatke na kojima se obučavao, odnosno trenirao i naučio je čak i šum i slučajnosti, umesto da nauči osnovne obrasce koji su relevantni za zadatak. Kao rezultat toga, takav model obično postiže loše rezultate na novim podacima. Suprotno tome, model koji je premalo prilagođen (eng. underfitting) je previše jednostavan i nije u stanju da prepozna složene obrasce u podacima, što dovodi do loših performansi. Validacija omogućava identifikaciju ovih problema i omogućava optimizaciju modela kako bi se postigla bolja generalizacija.

Jedna od metoda validacije je unakrsna validacija (eng. cross validation). Jedna iteracija ove metode sastoji se od podele skupa uzoraka na komplementarne podskupove, zatim obuke korišćenjem jednog od podskupova (podskup za obuku ili trening) i na kraju validacije obuke korišćenjem drugog podskupa (podskupa za validaciju i testiranje). Kako bi se obezbedilo da

rezultati budu što bliže realnim, vrši se više iteracija unakrsne validacije, svaki put nad različito podeljenim skupovima uzoraka. Performanse se ocenjuju uzimanjem proseka po iteracijama. Unakrsna validacija sa K particija (eng. K-fold cross validation) deli skup uzoraka na K jednakih particija i zatim se vrši K iteracija [35]. U svakoj iteraciji koristi se K-1 particija za obuku dok se preostala particija koristi za testiranje. Način rada unakrsne validacije sa K particija prikazan je na slici 15. Glavna prednost ove metode je što se svi uzorci iz skupa podataka koriste u nekom trenutku, odnosno iteraciji, za obuku i testiranje. Zato je procenjena stopa greške približnija stvarnoj nego kod drugih tehnika. Ukupna estimacija stope greške određena je kao prosečna vrednost procena u svakoj od K iteracija:

$$E = \frac{1}{k} \sum_{i=1}^k E_i$$



Slika 15: Način rada unakrsne validacije sa K particija [36]

6 Analiza rezultata

6.1 Mere kvaliteta modela

Ukoliko na raspolaganju imamo konačan broj modela, od kojih je potrebno koristiti jedan, postavlja se pitanje izbora modela, koje se obično rešava tako što se na neki način evaluiraju svi raspoloživi modeli i izabere se najbolji. Evaluacija modela predstavlja kvantifikaciju njegove sposobnosti generalizacije. Ona počiva na merama kvaliteta modela odnosno njegovih performansi. Mere kvaliteta modela zavise od vrste problema koji se rešava, kao i od željenih ishoda.

Mere koje se najčešće koriste za klasifikaciju su tačnost, preciznost, odziv i F_1 mera. Sve ove mere počivaju na matrici konfuzije i pojmovima vezanim za nju. U slučaju binarne klasifikacije, matrica konfuzija ima specifičan oblik koji se može videti na slici 16.

Predviđeno/Stvarno	Pozitivno	Negativno
Pozitivno	stvarno pozitivno (TP)	lažno pozitivno (FP)
Negativno	lažno negativno (FN)	stvarno negativno (TN)

Slika 16: Matrica konfuzije

Stvarno pozitivne (eng. true positive) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne. Broj takvih instanci skraćeno se označava sa TP. Stvarno negativne (eng. true negative) instance su negativne instance koje su od strane modela prepoznate kao negativne, označavaju se sa TN. Lažno pozitivne (eng. false positive) instance su negativne instance koje su od strane modela prepoznate kao pozitivne, označavaju se sa FP. Ilustrativno ovaj slučaj bi se mogao opisati scenariom da model za pacijenta koji je zdrav predvidi da je bolestan. Lažno negativne (eng. false negative) instance su pozitivne instance koje su od strane modela proglašene negativnim, označavaju se sa FN. Ovaj slučaj bi se opisao primerom da model za bolesnog pacijenta predvidi da je zdrav. Klasifikacija je najbolja kada je ova matrica dijagonalna jer to znači da je ona potpuno ispravna i nema lažno predviđenih vrednosti [25].

Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci. Može se zapisati na sledeći način:

$$Tačnost = \frac{TP+TN}{TP+TN+FP+FN}$$

Iako je vrlo intuitivna, tačnost klasifikacije ne mora uvek biti pogodna mera kvaliteta. Jedan razlog je njena neinformativnost u slučaju da klase imaju vrlo različit broj instanci. Ukoliko jednoj klasi pripada 99% instanci, a drugoj 1%, naizgled impresivna tačnost od 0.99 može biti postignuta tako što će sve instance biti klasifikovane u prvu klasu. Ipak, takav klasifikator je beskoristan. Pritom, klase će često biti neizbalansirane, tako da i neki ekstremni slučajevi mogu biti realistični. Na primer, u slučaju detekcije prevara sa kreditnim karticama, detekcije retkih bolesti i slično.

Preciznost je udeo stvarno pozitivnih instanci u svim instancama koje su proglašene pozitivnim [25], odnosno:

$$\text{Preciznost} = \frac{TP}{TP + FP}$$

Visoka preciznost osigurava da kada model predvidi pozitivnu klasu, da je verovatno tačno. Ova veličina odgovara na pitanje – od svih predviđenih tačnih instanci, koliko je stvarno tačno?

Odziv je udeo stvarno pozitivnih instanci u svim pozitivnim instancama [25], odnosno:

$$\text{Odziv} = \frac{TP}{TP + FN}$$

Ova veličina odgovara na pitanje – od svih u stvarnosti pozitivnih instanci, koliko je model tačno predvideo? Odziv je ključan u situacijama kada propuštanje pozitivnih instanci ima velike posledice. Na primer, u medicinskoj dijagnostici odziv je ključan jer neuspeh u otkrivanju bolesti, odnosno u slučaju kad je predviđena lažno negativna instanca, može imati velike posledice po pacijenta. U nekim slučajevima, pojedinačno posmatrane ove dve mere daju samo delimičan uvid u performanse modela.. Zato ima smisla gledati ih zajedno, što se postiže pomoću F_1 mere. Ona predstavlja njihovu harmonijsku sredinu i računa se na sledeći način:

$$F_1 = 2 \frac{\text{Preciznost} \cdot \text{Odziv}}{\text{Preciznost} + \text{Odziv}}$$

Ova mera će biti visoka samo kad su i preciznost i odziv visoki.

6.2 Rezultati klasifikacije

Performanse su se računale u slučaju kada su modelu prosleđeni svi atributi iz skupa podataka i kada su modelu prosleđeni samo najbitniji atributi koji su određeni u poglavlju 4. Konkretno za zadatak predikcije srčanih bolesti, najvažnija mera kvaliteta je odziv jer će ona osigurati da je broj lažno negativnih predviđenih vrednosti minimalan. Cilj je da model što tačnije predvidi stvarno pozitivne (TP) i stvarno negativne (TN) vrednosti i da ima što manje

lažno negativnih (FN) predviđenih vrednosti. Performanse modela pre odabira najbitnijih atributa prikazane su u tabeli 1.

Model	Tačnost	Preciznost	Odziv	F1 mera	Matrica konfuzije
Logistička regresija	87.45%	88.51%	88.51%	88.51%	[[106 17] [17 131]]
SVM	86.35%	86.27%	89.19%	87.71%	[[102 21] [16 132]]
KNN	88.56%	88.74%	90.54%	89.63%	[[106 17] [14 134]]
Stablo odluke	84.13%	87.23%	83.11%	85.12%	[[105 18] [25 123]]

Tabela 1: Performanse modela pre odabira najbitnijih atributa

Logistička regresija pokazuje dobre ukupne performanse, sa uravnoteženim vrednostima preciznosti, odziva i F1 mere i velikom tačnošću od 87.45%. SVM pokazuje malo manju tačnost ali ima veći odziv. Međutim preciznost je manja i u matrici konfuzije se vidi da ima više predviđenih lažno pozitivnih slučajeva. KNN nadmašuje ostale modele po svim performansama. Predviđa najviše stvarno pozitivnih i stvarno negativnih slučajeva od svih modela i najvažnije od svega predviđa najmanje lažno negativnih slučajeva. Stablo odluke ima najlošije performanse od svih modela, sa najnižom tačnošću i F1 merom. Odziv je takođe nizak što ukazuje da predviđa dosta lažno negativnih slučajeva, što se vidi i u matrici.

U tabeli 2 prikazane su performanse modela nakon odabira najbitnijih atributa.

Model	Tačnost	Preciznost	Odziv	F1 mera	Matrica konfuzije
Logistička regresija	87.82%	89.12%	88.51%	88.81%	[[107 16] [17 131]]
SVM	87.45%	87.01%	90.54%	87.74%	[[103 20] [14 134]]
KNN	86.35%	84.91%	91.22%	87.95%	[[99 24] [13 135]]
Stablo odluke	84.13%	83.02%	89.19%	85.99%	[[96 27] [16 132]]

Tabela 2: Performanse modela nakon odabira najbitnijih atributa

Nakon odabira najbitnijih atributa, tačnost i preciznost logističke regresije su se poboljšale za manje od 1%. Stvarno pozitivni slučajevi su se povećali za 1, dok su se lažno negativni smanjili za 1. Kod SVM modela tačnost, preciznost i odziv su se povećali za oko 1% i smanjio se broj lažno predviđenih slučajeva. Kod KNN modela su sve performanse sem odziva pokazale lošije rezultate. To znači da model predviđa manje lažno negativnih slučajeva, međutim kako je preciznost pala za skoro 4%, sada predviđa više lažno pozitivnih vrednosti. Stablo odluke i nakon odabira najbitnijih atributa ima najlošije performanse u odnosu na ostale modele. Odziv se primetno poboljšao ali je preciznost smanjena za skoro 5%, odnosno iako predviđa manje lažno negativnih slučajeva, povećan je broj lažno pozitivnih i smanjen broj stvarno pozitivnih.

Zaključno, pre odabira najbitnijih atributa KNN model se izdvojio kao najpouzdaniji, zbog sposobnosti da efikasno balansira između preciznosti i odziva i tako minimizuje broj lažno predviđenih slučajeva. Nakon odabira najbitnijih atributa, logistička regresija je pokazala najstabilnije performanse. KNN još uvek ima najveći odziv ali pokazuje primetnu slabiju preciznost. U oba slučaja stablo odluke pokazuje najgore performanse zbog najmanje tačnosti i najvećeg broja lažno predviđenih slučajeva.

7 Zaključak

Ovaj rad je istražio primenu različitih metoda mašinskog učenja za predikciju srčanih bolesti, uključujući logističku regresiju, metode potpornih vektora (SVM), K najbližih suseda (KNN), i stablo odluke. Rezultati pokazuju da su svi korišćeni modeli ostvarili solidne performanse, s time da je logistička regresija nakon odabira najbitnijih atributa pokazala najstabilnije performanse. Evaluacija modela je sprovedena uz korišćenje standardnih mera kvaliteta, kao što su tačnost, preciznost, odziv i F_1 mera. Analizom dobijenih rezultata uočeno je da su neki modeli bolje generalizovali podatke u poređenju sa drugima, što je naglašeno kroz validaciju i testiranje modela. Konkretno, metoda K najbližih suseda se pokazala kao najpouzdaniji u minimizaciji lažno negativnih predviđanja, što je ključno za medicinske primene gde je od kritične važnosti precizno identifikovati pacijente sa visokim rizikom od srčanih bolesti. Iako su postignuti rezultati zadovoljavajući, rad je identifikovao i nekoliko izazova, uključujući potrebu za daljom optimizacijom modela i potencijalnim unapređenjima u oblasti obrade podataka i izbora značajnih atributa. Neki od ključnih izazova uključuju balansiranje između pristrasnosti i varijanse, kao i izbor adekvatnih hiperparametara za svaki model.

U pogledu budućih istraživanja, predlaže se dalja analiza i optimizacija algoritama kroz primenu naprednijih tehnika kao što su neuronske mreže ili kombinacija više modela (eng. ensemble methods). Takođe, istraživanje bi moglo biti prošireno na veće i raznovrsnije skupove podataka, što bi omogućilo bolju generalizaciju modela. Pored toga, važnost etičkih aspekata i zaštite podataka pacijenata ostaje ključno pitanje koje zahteva dodatno istraživanje, posebno u kontekstu primene mašinskog učenja u medicinskoj praksi.

Ovaj rad doprinosi boljem razumevanju potencijala mašinskog učenja u medicinskoj dijagnostici, pružajući osnovu za dalji napredak i integraciju ovih tehnologija u svakodnevnu kliničku praksu.

Literatura

- [1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1,
Datum pristupa: 01.09.2024.
- [2] <http://www.bhf.org.uk/what-we-do/our-research/heart-statistics>
Datum pristupa: 08.09.2024.
- [3] T. Gaziano, K. Srinath Reddy, F. Paccaud, S. Horton, V. Chatuverdi, *Disease Control Priorities in Developing Countries*, 2nd edition, Washington DC, Oxford University Press, 2006, pp. 9-10.
- [4] A. Sniderman, C. Furberg, "Age as a modifiable risk factor for cardiovascular disease", *The Lancet*, vol. 371, pp. 1547-1549, May 2008.
- [5] G. Rosano, C. Vitale, G. Marazzi, M. Volterrani, „Menopause and cardiovascular disease: the evidence“, *Climacteric*, vol. 10, pp. 19-24, 2007.
- [6] V. Nasteski, "An overview of the supervised machine learning methods", *Horizons*, vol. 4, pp. 51-62, December 2017.
- [7] <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data>
Datum pristupa: 08.09.2024.
- [8] W. Cayley, "Diagnosing the cause of chest pain", *Am. fam. phys. Eau Claire*, vol. 72, pp. 2012-2021, November 2005.
- [9] S. Howell, J. Sear, P. Foex, "Hypertension, hypertensive heart disease and perioperative cardiac risk", *Br. J. Anaesth.* Oxford, vol. 92, pp. 570-583, April 2004.
- [10] D. McNamara, "Dietary cholesterol and atherosclerosis", *Biochim. et biophys. acta* Amsterdam, vol. 1529, pp. 310-320, December 2000.
- [11] B. Leon, T. Maddox, "Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research", *World J. diabetes*, Pleasanton, vol. 6, pp. 1246-1258, October 2015.
- [12] N. Goldschlager, A. Selzer, K. Cohn, "Treadmill stress tests as indicators of presence and severity of coronary artery disease", *Ann Intern Med.* Filadelfija, vol. 85, pp. 277-286, September 1976.
- [13] G. Lanza, M. Mustilli, A. Sestito, F. Infusino, G. Sgueglia, F. Crea, "Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test", *Heart Rim*, vol. 90, pp. 1417-1421, December 2004.
- [14] M. Kardash, M. Elamin, D. Mary, M. Whitaker, D. Smith, R. Boyle, J. Stoker, R. Linden, "The slope of ST segment/heart rate relationship during exercise in the prediction of severity of coronary artery disease", *Eur. Heart J.* Oxford, vol. 3, pp. 449-458, October 1982.

-
- [15] <https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/>
Datum pristupa: 30.08.2024.
- [16] K. Singh, S. Upadhya, "Outlier detection: application and techniques", *IJCSI*, vol. 9, pp. 307-323, January 2012.
- [17] D. Williamson, R. Parker, J. Kendrick, "The box plot: a simple visual method to interpret data", *Ann. Of Int. Med.* Philadelphia, vol. 110, pp. 916-921, June 1989.
- [18] D. Apostolopoulos, P. Davlourous, N. Patsouras, T. Spyridonidis, P. Vassilakos, D. Alexopoulos, "ST-segment depression during vasodilator stress is of minor clinical importance in women with normal myocardial perfusion imaging and low or intermediate risk of coronary artery disease", *Eur. J. Nucl. Med. Mol. Imaging* Berlin, vol. 39, pp. 437-445, December 2011.
- [19] <https://medium.com/@sunnykumar1516/what-is-label-encoding-application-of-label-encoder-in-machine-learning-and-deep-learning-models-c593669483ed>
Datum pristupa: 27.08.2024.
- [20] C. Chatfield, "Exploratory data analysis", *Eur. J. Oper. Research*, vol. 23, pp. 5-13, January 1986.
- [21] <https://www.hackersrealm.net/post/feature-selection-using-correlation-matrix>
Datum pristupa: 27.08.2024.
- [22] X. Jin, A. Xu, R. Bie, P. Guo, "Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles", *LNBI*, vol. 3916, pp. 106-115, 2006.
- [23] A. Cuevas, M. Febrero, R. Fraiman, "An anova test for functional data", *CSDA*, vol. 47, pp. 111-122, August 2004.
- [24] D. Jurafsky, J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition, Stanford University, 2024, pp. 77-84
- [25] M. Nikolić, A. Zečević, *Mašinsko učenje*, Beograd, 2019, pp. 50-113
- [26] <https://pianalytix.com/logistic-regression-in-machine-learning/>
Datum pristupa: 24.08.2024.
- [27] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
Datum pristupa: 30.08.2024.
- [28] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>
Datum pristupa: 30.08.2024.
- [29] M. Ali, B. Paul, K. Ahmed, F. Bui, J. Quinn, M. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison", *Comp. in Biol. Med.*, vol. 136, September 2021.
- [30] <https://www.geeksforgeeks.org/k-nearest-neighbours/>
Datum pristupa: 26.08.2024.
-

-
- [31] A. Mucherino, P. Papajorgji, P. Pardalos, *Data mining in agriculture*, vol. 34, New York, 2009, pp. 83-106
- [32] <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
Datum pristupa: 30.08.2024.
- [33] F. Alotaibi, "Implementation of machine learning model to predict heart failure disease", *IJACSA*, vol. 10, pp. 261-268, November 2019.
- [34] <https://scikit-learn.org/stable/api/sklearn.tree.html>
Datum pristupa: 30.08.2024.
- [35] D. Berrar, *Cross-validation*, 2nd Edition of Encyclopedia of Bioinformatics and Computational Biology, vol. 1, Tokyo, 2024, pp. 2-6
- [36] <https://towardsdatascience.com/what-is-k-fold-cross-validation-5a7bb241d82f>
Datum pristupa: 26.08.2024.

Biografija



Katarina Topolić rođena je 14.08.2001. godine u Novom Sadu. Osnovnu školu „Đorđe Natošević“ u Novom Sadu završila je 2016. godine. Gimnaziju „Jovan Jovanović Zmaj“ u Novom Sadu završava 2020. godine. Iste godine upisuje Fakultet tehničkih nauka, smer Računarstvo i automatika. U trećoj godini upisuje smer Računarski upravljački sistemi. Ispunila je sve obaveze i položila sve ispite predviđene studijskim programom.