Metoda temeljena na stojnom učenju za predviĎanje rizika srčanih bolesti iz pregleda pacijent

Vinaj, Ema

Undergraduate thesis / Završni rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek / Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:200:534611

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-05-14**

Repository / Repozitorij:

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Sveučilišni studij

Metoda temeljena na strojnom učenju za predviđanje rizika srčanih bolesti iz pregleda pacijenta

Završni rad

Ema Vinaj

Osijek, 2023.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

Obrazac Z1P - Obrazac za ocjenu završnog rada na preddiplomskom sveučilišnom studiju

Osijek, 13.09.2023.

Odboru za završne i diplomske ispite

Prijedlog ocjene završnog rada na preddiplomskom sveučilišnom studiju

Ime i prezime Pristupnika:	Ema Vinaj
Studij, smjer:	Sveučilišni prijediplomski studij Elektrotehnika i informacijska tehnologija
Mat. br. Pristupnika, godina upisa:	4899, 30.07.2020.
OIB Pristupnika:	68159965252
Mentor:	prof. dr. sc. Irena Galić
Sumentor:	Marin Benčević, mag. ing. comp.
Sumentor iz tvrtke:	
Naslov završnog rada:	Metoda temeljena na stojnom učenju za predviđanje rizika srčanih bolesti iz pregleda pacijenta
Znanstvena grana rada:	**Umjetna inteligencija (zn. polje računarstvo)**
Zadatak završnog rad:	Istražiti i opisati trenutno stanje kardiovaskularnih bolesti u svijetu i proces njihove dijagnostike i detekcije. Istražiti koja mjerenja mogu biti indikatori rizika od kardiovaskularnih bolesti. Razviti algoritam koji će iz raznih mjerenja i podataka o nekoj osobi uključujući dob, spol, puls, razinu kolesterola, podaci o elektrokardiogramu i sl. predvidjeti rizik od kardiovaskularnih bolesti za tu osobu. Koristiti javno dostupan skup podataka. Napraviti statističku analizu razvijene metode na testnom skupu podataka, te usporedbu s trenutno postojećim metodama. Napraviti
Prijedlog ocjene završnog rada:	Izvrstan (5)
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: 3 bod/boda Postignuti rezultati u odnosu na složenost zadatka: 2 bod/boda Jasnoća pismenog izražavanja: 3 bod/boda Razina samostalnosti: 2 razina
Datum prijedloga ocjene od strane mentora:	13.09.2023.
Datum potvrde ocjene od strane Odbora:	24.09.2023.

Potvrda mentora o predaji konačne verzije rada:	*Mentor elektronički potpisao predaju konačne verzije.*
	Datum:

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA
I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O ORIGINALNOSTI RADA

Osijek, 25.09.2023.

Ime i prezime studenta:	Ema Vinaj
Studij:	Sveučilišni prijediplomski studij Elektrotehnika i informacijska tehnologija
Mat. br. studenta, godina upisa:	4899, 30.07.2020.
Turnitin podudaranje [%]:	7

Ovom izjavom izjavljujem da je rad pod nazivom**: Metoda temeljena na stojnom učenju za predviđanje rizika srčanih bolesti iz pregleda pacijenta**

izrađen pod vodstvom mentora prof. dr. sc. Irena Galić

i sumentora Marin Benčević, mag. ing. comp.

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

SADRŽAJ

.

1. UVOD

Kardiovaskularne bolesti – skupina bolesti i poremećaja srca i krvnih žila vodeći su uzrok smrti u svijetu prema statistikama iz posljednjih nekoliko godina. [1] [24] Procjenjuje se da godišnje odnose oko 17,9 milijuna života, što predstavlja 32% svih smrtnih slučajeva globalno. [27]

Razni su čimbenici koji utječu na rizik od srčanih bolesti, a neki od najznačajnijih su prehrana, tjelesna (ne)aktivnost, pušenje, prekomjerna uporaba alkohola, ali i spol, dob te obiteljska povijest kardiovaskularnih bolesti. [21]

Ipak, dobra je vijest to da se većina srčanih bolesti može spriječiti usvajanjem zdravog načina života, brigom o vlastitom tijelu te redovitim liječničkim pregledima. Identificiranjem osoba sa velikim rizikom od srčanih bolesti i osiguravanjem da se istima pruži odgovarajuće liječenje može se spriječiti veliki broj preuranjenih smrti.

Jedna od metoda predviđanja rizika od srčanih bolesti kod pacijenata jest korištenjem strojnog učenja. To uključuje razvijanje algoritma koji iz različitih podataka o osobi predviđa rizik koji ta osoba ima od razvoja srčanih bolesti. Upravo to je i zadatak ovog završnog rada.

U ovom će se radu detaljnije opisati trenutno stanje kardiovaskularnih bolesti u svijetu te čimbenika koji ih uzrokuju; opisat će se pojedine metode dijagnosticiranja rizika od srčanih bolesti, specifičnije, usporedit će se (na algoritamskoj i statističkoj razini) različite modele koje se može koristiti pri dijagnosticiranju i evaluaciji spomenutih rizika.

1.1. Zadatak završnog rada

Zadatak ovog završnog rada jest razvoj algoritma za procjenu rizika od razvoja srčanih bolesti ovisno o različitim utjecajnim čimbenicima kod osobe korištenjem javno dostupnog skupa podataka, zatim statistička analiza razvijene metode na testnom skupu podataka te prikaz pojedinih čimbenika koji utječu na rizik od srčanih bolesti. Naposljetku će se usporediti razvijena metoda s već postojećim metodama.

2. MEDICINSKA POZADINA I POSTOJEĆA RJEŠENJA

U ovom će se poglavlju navesti neki od najznačajnijih faktora rizika od srčanih bolesti te će se za iste dati opis (zašto utječu na povećanje rizika). Također će biti opisani modeli strojnog učenja koji se trenutno koriste pri prepoznavanju rizika od srčanih bolesti.

2.1. Utjecajni čimbenici

Utjecajne čimbenike koji povećavaju rizik od bolesti srca može se podijeliti na promjenjive i nepromjenjive: promjenjivi su oni na koje osoba može utjecati i time smanjiti rizik od obolijevanja, dok su nepromjenjivi oni koji su nasljedni ili neizbježni te se na njih ne može utjecati. [2] [3]

Promjenjivi čimbenici:

- visoki krvni tlak
- visoke razine kolesterola
- pušenje
- tjelesna neaktivnost
- nezdrava prehrana
- pretilost
- dijabetes
- prekomjerna konzumacija alkohola
- stres

Nepromjenjivi čimbenici:

- dob
- spol
- obiteljska povijest obolijevanja od bolesti srca

Hipertenzija, to jest visoki krvni tlak je vodeći uzrok bolesti srca. Definiran je kao tlak od 130/80 mm Hg i više. Sistolički krvni tlak, zvan još i „gornji tlak“, predstavlja onaj tlak koji stvara srce kada pumpa krv kroz arterije, za razliku od dijastoličkog ili „donjeg“ tlaka koji predstavlja tlak u arterijama između 2 otkucaja srca. Visoki krvni tlak oštećuje unutrašnjost arterija te ih čini osjetljivijima na nakupljanje tzv. plaka (naslaga masnih tvari, kolesterola i staničnih otpadnih proizvoda) što znači da lakše može doći do sužavanja arterija i smanjenja protoka korisnih tvari kroz žilu.

Visoke razine LDL kolesterola – lipida koji je u određenim dozama neophodan za zdravo funkcioniranje organizma, ali u pretjeranim dozama može biti vrlo štetan. Kolesterol se, ukoliko ga u tijelu ima previše, može nakupiti u stjenkama arterija i ograničiti dotok krvi u srce.

Pušenje je nakon visokog krvnog tlaka drugi vodeći uzročnik srčanih bolesti. Osim što uzrokuje zadebljanje i sužavanje krvnih žila, pušenjem se podižu razine triglicerida (vrste masnoće u krvi) u tijelu pa krv postaje „ljepljiva“ i veća je vjerojatnost da će se zgrušati, što može blokirati protok krvi u srce.

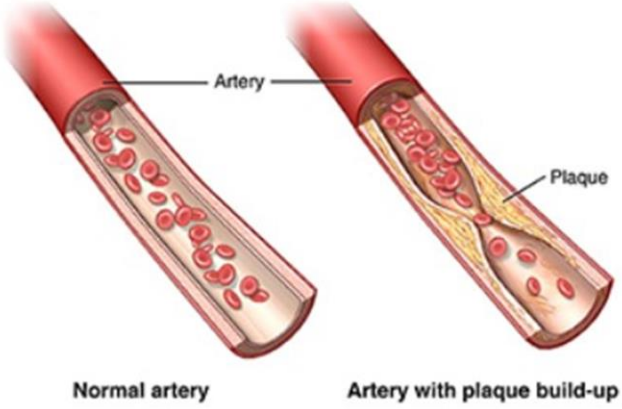
Prekomjerna tjelesna težina povećava rizik osobe da ima visoki krvni tlak, visoki kolesterol ili da oboli od dijabetesa, drugim riječima, može dovesti do nakupljanja masnoće u arterijama, a time i do njihovog začepljivanja. Isto vrijedi za nezdravu prehranu. Zdravom prehranom smanjuje se unos zasićenih i trans masti, dodatnih šećera te natrija – uzročnika gore navedenih čimbenika – u tijelo.

Tjelesna neaktivnost može dovesti do bolesti srca iz razloga što povećava mogućnost razvoja drugih opasnih čimbenika kao što su pretilost, visok kolesterol, visoki krvni tlak i dijabetes.

Dijabetes je također uzročnik srčanih bolesti iz razloga što visoka razina šećera u krvi uzrokovana dijabetesom može s vremenom oštetiti krvne žile i omogućiti nakupljanje tzv. „plaka“.

Pretjerano konzumiranje alkohola s vremenom može dovesti do istezanja i povećavanja srca. Kako se mišići postepeno rastežu, tako i slabe, što sprječava srce da pumpa krv kako bi trebalo te može dovesti do zatajenja srca. [28]

Visoke razine kortizola uzrokovane dugotrajnim stresom mogu povećati razinu kolesterola u krvi te količinu triglicerida i šećera u krvi. Kao što je navedeno iznad, ovo su uobičajeni čimbenici rizika za bolesti srca.



Sl. 2.1. Razlika u arterijama bez plaka i sa plakom. Izvor slike: Hopkins Medicine [37]

Kako čovjek stari, povećava se krutost velikih arterija u tijelu, tzv. „arterioskleroza“ ili otvrdnuće arterija, što uzrokuje povećanje krvnog tlaka te čini osobe starije životne dobi podložnijima obolijevanju od srčanih bolesti.[4]

Žene su sklonije obolijevanju od bolesti srca. Smatra se da je razlog tomu povećanje ostalih rizičnih čimbenika (kao što su krvni tlak i visoki kolesterol) za vrijeme menopauze. Također, kao rezultat dugogodišnjih uvjerenja da su žene „zaštićene“ od bolesti srca, i dalje se manje pažnje posvećuje simptomima bolesti srca kod žena te se iste pripisuje nekim drugim bolestima, što nažalost rezultira kasnijim otkrivanjem i slabijim liječenjem srčanih bolesti kod žena. [5]

Ukoliko jedan član obitelji ima srčanu bolest, ostali članovi s kojima je u krvnom srodstvu mogu naslijediti gene koji ih čine podložnijima obolijevanju od srčanih bolesti. Osim toga, osobe koje su u krvnom srodstvu često dijele i slične životne navike koje mogu uzrokovati bolesti srca.

Još neki od čimbenika koji utječu na rizik od razvoja srčanih bolesti, ali u znatno manjem postotku od prethodno navedenih:

Količina sna – manjak sna je povezan sa razvojem visokog krvnog tlaka te posljedično i sa mogućim razvojem srčanih bolesti

Zagađenost zraka – velik broj istraživanja je pokazao da nečist zrak može dodatno potaknuti razvoj srčanih bolesti uz postojanje drugih rizičnih čimbenika, razlog tomu je unos štetnih stvari udisanjem

Radijacija – dokazano je da terapija radijacijom, iako igra ključnu ulogu u tretiranju različitih oblika raka, može stvoriti i značajna oštećenja na srcu

Važno je napomenuti da, iako navedeni čimbenici mogu znatno povećati rizik od obolijevanja od srčanih bolesti, oni ne jamče razvoj istih. U svakom se slučaju ipak uvijek preporučuje zdraviji način života: prevencija je bolja nego liječenje. [6] [7]

2.2. Postojeća rješenja

U nastavku će biti opisani neki od najpoznatijih i najraširenijih postojećih modela strojnog učenja koje se koristi za predviđanje rizika od srčanih bolesti. To su modeli koji su dokazali svoju učinkovitost i korist u prepoznavanju rizika od srčanih bolesti na vrijeme.

Svi opisani modeli su tzv. *Cox Proportional Hazards* modeli odnosno Coxovi regresijski modeli. To je klasa statističkih modela koje se koristi za analizu podataka o preživljavanju. Takav model povezuje vrijeme proteklo prije no što se neki događaj dogodi s jednom ili više kovarijabli koje se mogu povezati s proteklom količinom vremena. Glavni cilj takvog modela je dakle procijeniti učinak kovarijabli na vrijeme koje će proteći do nekog događaja. [8] [9]

Coxov model ima eksponencijalni oblik:

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \qquad (1)$$

gdje:

- t predstavlja vrijeme kada se dogodi određeni događaj

- λ(t) je „*funkcija opasnosti"* (funkcija hazarda) za subjekt u vremenu t, određena skupom od m broja kovarijabli (X1, X2,…, Xk) – za funkciju opasnosti može se reći da predstavlja trenutnu stopu opasnosti u trenutku t

- β1, β2,…βk su regresijski koeficijenti koji mjere veličinu učinka pojedinačnih kovarijabli

- exp je eksponencijalna funkcija [$exp(X) = ex$]

- λ0(t) je osnovna stopa opasnosti - proizvoljna (nepoznata) funkcija, koja odgovara vrijednosti opasnosti kada su svi Xi jednaki nuli.

Coxov model se tada može napisati kao „*funkcija preživljavanja*":

$$S(t) = [S_0(t)]^{\exp\left(\sum_{i=1}^{k} \beta_i X_i\right)} \qquad (2)$$

Funkcija preživljavanja predstavlja vjerojatnost preživljavanja vremena t za pojedinog subjekta.

Opća formula za izračunavanje procjene rizika ima sljedeći oblik:

$$\widehat{H(t)} = 1 - [S_0(t)]^{\exp\left(\sum_{i=1}^{k} \beta_i X_i - \sum_{i=1}^{k} \beta_i \overline{X_i}\right)} \qquad (3)$$

gdje je:

- H(t) rizik od kardiovaskularne bolesti procijenjen za pojedinca

- S0(t) je osnovna stopa preživljenja u vremenu praćenja t, gdje je t = 10 godina

- βi je koeficijent regresije

- Xi je vrijednost i-tog čimbenika rizika (ako je kontinuiran to je log-transformirana vrijednost)

- $\overline{X}$i je odgovarajuća sredina,

- k je broj faktora rizika.

Takav oblik funkcije naziva se „*funkcija kumulativnog hazarda*“ te se lako može dobiti iz funkcije preživljavanja kao: $H(t) = -lnS(t)$

Funkcija rizika od kardiovaskularnih bolesti može se izvesti iz (3), korištenjem koeficijenata regresije i osnovnih stopa rizika.

U nastavku su navedeni najpoznatiji modeli temeljeni na Coxovom regresijskom modelu, a koji su korišteni u svrhu predviđanja rizika od srčanih bolesti:

1. Framingham Risk Score model

Ovo je jedan od najraširenijih modela predviđanja rizika od srčanih bolesti. Razvijen je na temelju podataka iz Framinghamske studije srca – dugogodišnjeg istraživanja srca provedenog u Framinghamu (savezna država Massachusetts, SAD). Model procjenjuje 10-godišnji rizik od razvoja srčane bolesti kod osobe temeljen na slijedećim ulaznim varijablama: dob, spol, ukupna količina kolesterola u krvi, kolesterol lipoproteina visoke gustoće (HDL), sistolički krvni tlak, status pušenja i status dijabetesa. Kao što se može vidjeti, sve ulazne varijable su značajni rizični čimbenici bolesti srca. Rezultat se izvodi kao postotak te se rizik smatra niskim ukoliko rezultat iznosi 10% ili manje, umjerenim ako je rezultat između 10% i 19%, a visokim ukoliko je rezultat 20% ili veći.

2. SCORE (*Systematic COronary Risk Evaluation)* model

Osim što je ovo najjednostavniji od opisanih modela, SCORE se razlikuje i po tome što se ne fokusira na rizik od obolijevanja bolestima srca, već na 10-godišnji rizik od smrtonosne bolesti srca. Rizični čimbenici koje se koristi u procjeni rizika su: dob, spol, status pušenja, sistolički krvni tlak te ukupna razina kolesterola. Također se koriste i tzv. tablice rizika specifične za različite geografske regije i dobne skupine. Kao rezultat, SCORE kategorizira osobu u jednu od četiri kategorije: „niski rizik“, „umjereni rizik“, „visoki rizik“ i „vrlo visoki rizik“. SCORE model je prevladavajući model u Europi uz Farmingham Risk Score model. Razvilo ga je Europsko kardiološko društvo (ESC). [10]

3. ACC/AHA Pooled Cohort Equations model

Ovaj model predviđa 10-godišnji rizik pojedinca od razvijanja aterosklerotskog kardiovaskularnog poremećaja (Atherosclerotic Cardiovascular Disease). Čimbenici koji se koriste za njegov rad su: dob, spol, rasa, ukupna količina kolesterola, količina HDL kolesterola, sistolički krvni tlak, upotreba antihipertenzivnih lijekova, status pušenja i status dijabetesa. ACC/AHA model razvili su, kao što samo ime sugestira, u suradnji American College of Cardiology (ACC) i American Heart Association (AHA) te se on većinom koristi u Sjedinjenim Američkim Državama. U praksi se koristi online kalkulator zvan ASCVD Risk Estimator temeljen na ACC/AHA modelu za izračun rizika od spomenutog poremećaja.

4. QRISK model

Kao i Farmingham model, QRISK model predviđa 10-godišnji rizik od razvoja bolesti srca. Razlika je u ulaznim varijablama, to jest rizičnim čimbenicima koji se koriste, a kod ovog modela to su: dob, spol, rasa, status pušenja, sistolički krvni tlak, ukupna količina kolesterola, količina HDL kolesterola, BMI (Body Mass Index), obiteljska anamneza (povijest obolijevanja u obitelji) i socioekonomski status. Razvijen je od strane liječnika i akademika Nacionalne Zdravstvene Službe Ujedinjenog Kraljevstva, a temelji se na podacima prikupljenim od preko tisuću liječnika opće prakse diljem zemlje.

Unatoč kontinuiranom radu na modelima i poboljšanjima na području strojnog učenja, svaki model ima određena ograničenja te ne može dati apsolutnu procjenu rizika, stoga njihovu primjenu uvijek valja kombinirati sa kliničkom prosudbom. [11] [17]

3. PREDVIĐANJE RIZIKA OD SRČANIH BOLESTI

Za ovaj je zadatak odabran Python kao programski jezik. Zbog svoje čiste i čitljive sintakse, širokog spektra različitih biblioteka i alata korisnih pri strojnom učenju, velike aktivne zajednice, fleksibilnosti i svestranosti u različitim fazama strojnog učenja te kombatibilnosti sa više platformi, Python je izvrstan izbor jezika za razvoj i implementaciju algoritama, modela i sustava strojnog učenja.

Kao IDE (Integrated Development Environment), to jest programsko okruženje korišten je Jupyter, specifičnije Jupyter Notebook (bilježnica) koja kombinira kod, tekst objašnjenja, jednadžbe, vizualizacije i druge korisne elemente. Jupyter je općenito izrazito popularan u Python zajednici te je koristan kod strojnog učenja jer omogućava razvoj, treniranje i evaluaciju raznih modela.

Za instalaciju Pythona i potrebnih Python biblioteka korištena je Anaconda kao popularna distribucijska platforma za Python. Anaconda pojednostavljuje upravljanje podacima i implementaciju istih te pruža sveobuhvatan sustav upravljanja alatima i bibliotekama posebno prilagođenim za analizu podataka i strojno učenje. [35]

Dataset, odnosno skup podataka na kojem se u ovom završnom radu radilo, preuzet je sa Kaggle-a. Kaggle je online platforma koja je, između ostalog, jedna od najistaknutijih platformi koje pružaju skupove podataka za analizu i strojno učenje.

Biblioteke koje su korištene pri radu sa podacima su:

1. Pandas

Korišten za manipulaciju i analizu podataka. Nudi funkcionalnosti za čišćenje podataka, filtriranje, grupiranje, spajanje i slične operacije, što ga čini nezamjenjivom bibliotekom za predprocesiranje i istraživanje podataka u strojnom učenju.

2. NumPy

NumPy pruža podršku za velike, višedimenzionalne nizove i matrice zajedno sa zbirkom matematičkih funkcija za različite operacije. Ova je biblioteka temelj za numeričko računanje u Pythonu.

3. Scikit-learn

Sveobuhvatna biblioteka koja pruža širok raspon alata za zadatke kao što su klasifikacija, regresija, grupiranje, smanjenje dimenzionalnosti, odabir modela i evaluacija. Koristi se za obradu podataka, izgradnju i obuku modela strojnog učenja ili za procjenu izvedbe modela.

4. Matplotlib

Biblioteka korištena za vizualizaciju podataka. Omogućuje kreiranje dijagrama te se dobro integrira sa Pandas i NumPy bibliotekama što ju čini dobrim alatom za vizualizaciju rezultata.

5. Seaborn

Također biblioteka za vizualizaciju, Seaborn omogućuje prikaz statističkih podataka te nudi ugrađenu podršku za statističku procjenu i palete boja.

6. Missingno

Pomoću ove biblioteke je moguće vizualizirati i analizirati podatke koji nedostaju u određenom datasetu (skupu podataka).

3.1. Algoritmi predviđanja rizika od srčanih bolesti

Razni su algoritmi koje se može koristiti u svrhu predviđanja rizika od srčanih bolesti. Neki od najčešće korištenih su: logistička regresija (*Logistic Regression*), stabla odlučivanja (*Decision trees*), neuronske mreže (*Neural networks*) i slučajne šume (*Random forests*).

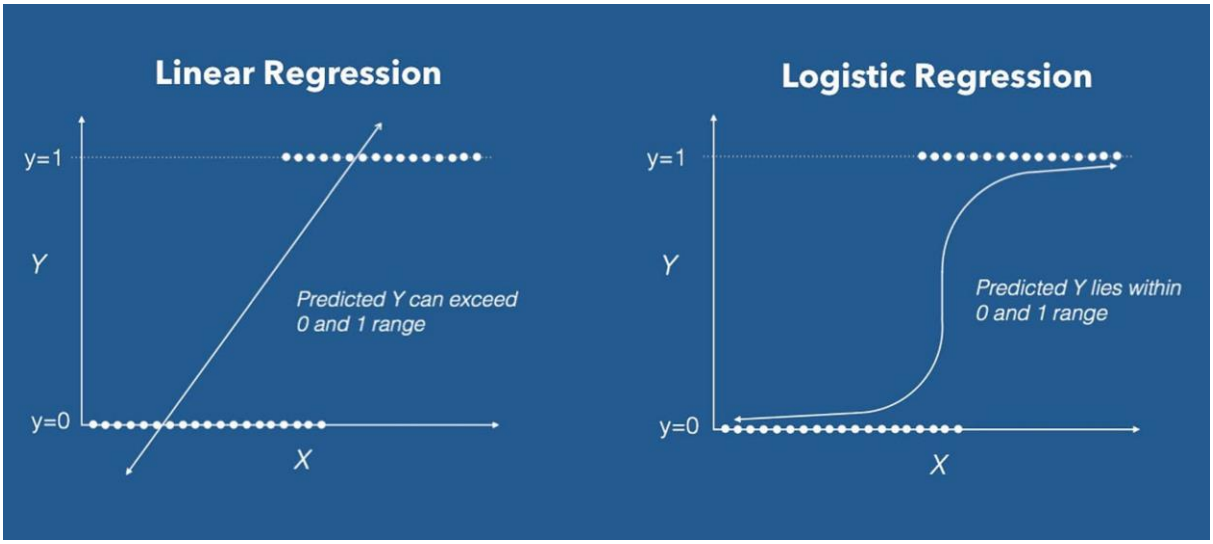
Logistička regresija

Logistička regresija je statistički algoritam i algoritam strojnog učenja koji se koristi za zadatke binarne klasifikacije. Ova metoda modelira vjerojatnost binarnog ishoda, odnosno ishoda tipa da/ne ili 1/0 na temelju jedne ili više prediktorskih varijabli. To ju čini pogodnom za zadatke gdje se predviđa ima li osoba rizik od razvijanja bolesti ili ne, ali i za zadatke kao što su predviđanje hoće li kupac kupiti proizvod ili ne, je li e-pošta spam ili nije i slično. [26]

Tri su osnovne vrste logističke regresije: binarna logistička regresija (odgovor ima samo dva moguća ishoda, npr. e-mail je spam ili e-mail nije spam); multinomna logistička regresija (tri ili više kategorija bez poretka, npr. kategorizacija osoba na one koji preferiraju kavu, one koji preferiraju čaj i one koji preferiraju neku treću opciju); ordinalna logistička regresija (tri ili više

kategorija s redoslijedom, npr. ocjena filma od 1 do 5). Bitno je ne miješati logističku regresiju s linearnom regresijom koja je zasebni algoritam i koristi se u potpuno različite svrhe. Dok se logistička regresija koristi za predviđanje binarnog ishoda (dakle predviđa se vrijednost kategoričkih varijabli), linearna se regresija pak koristi kod predviđanja kontinuiranih numeričkih vrijednosti (kontinuiranih varijabli). Linearna regresija koristi se u slučajevima kada je potrebno predvidjeti cijene nekretnina, temperaturu, utjecaj dobi i spola na visinu osobe i slično. Logistička regresija se, kao što je već spomenuto, koristi za potpuno različite zadatke pa je jasno zašto je ona i prikladnija za zadatak predviđanja rizika od bolesti srca. [12] [22]

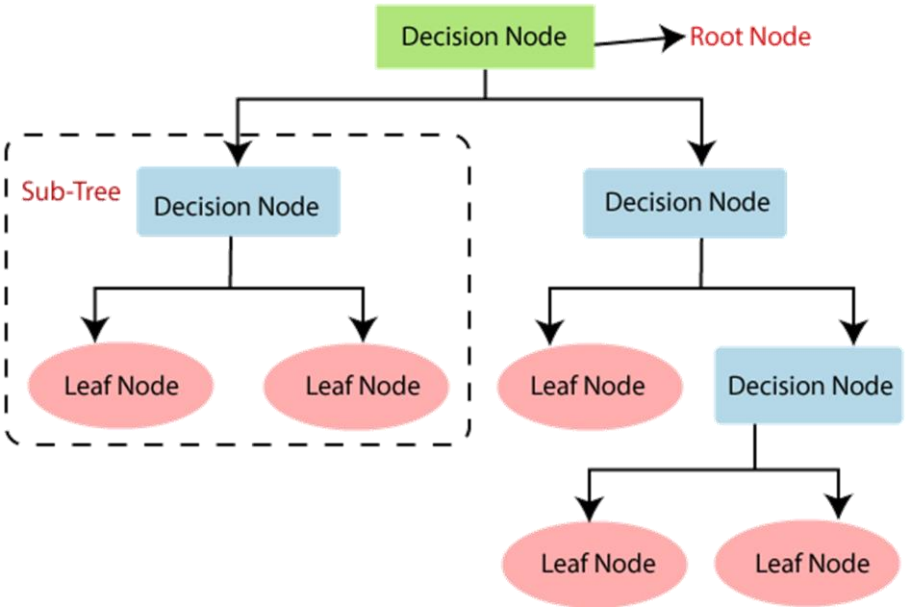
Prednosti algoritma logističke regresije su njegova jednostavnost, robusnost te mogućnost rukovanja i numeričkim i kategoričkim varijablama. [18]



Sl. 2.1. Razlika između linearne i logističke regresija. Izvor slike: Pant, Ayush [38]

Stabla odlučivanja

Još jedan algoritam koji se može koristiti kod ovakvog zadatka je algoritam stabla odlučivanja (*decision tree*). To je neparametarski algoritam koji se koristi kod zadataka klasifikacije i regresije. Ima hijerarhijsku strukturu stabla koja se sastoji od korijenskog čvora, grana,



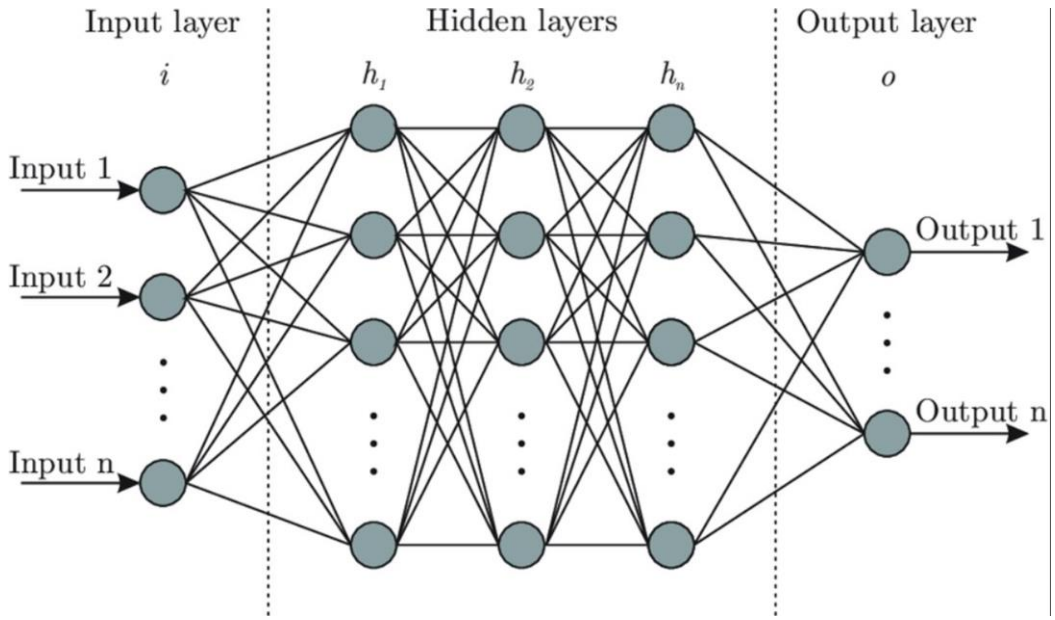
Sl. 2.2. Prikaz algoritma stabla odlučivanja. Izvor slike: Javatpoint [39]

Stablo odlučivanja počinje s korijenskim čvorom koji nema nikakve ulazne grane. Izlazne grane iz korijenskog čvora zatim ulaze u unutarnje čvorove, koje se zovu još i „čvorovi odlučivanja“. Na temelju dostupnih značajki podataka, svaki unutarnji čvor provodi procjene kako bi formirao homogene podskupove, koji su označeni završnim čvorovima (na dijagramu označeni kao *leaf node*). Proces izgradnje stabla je rekurzivan što znači da nakon što se podjela izvrši na unutarnjem čvoru, isti postupak podjele primjenjuje se na rezultirajuće podskupove, stvarajući podređene čvorove. Ovaj se proces nastavlja sve dok se ne ispuni kriterij zaustavljanja, kao što je maksimalna dubina stabla, minimalni broj uzoraka u čvoru ili prag homogenosti. Završni čvorovi predstavljaju sve moguće ishode unutar skupa podataka. Cilj ovakvog algoritma je stvoriti model koji predviđa vrijednost ciljne varijable učenjem jednostavnih pravila odlučivanja izvedenih iz značajki podataka. [14] [25]

Prednosti ovog algoritma su to što može rukovati i kategoričkim i numeričkim varijablama, njegova otpornost na odstupanja i šumove među podacima te jednostavnost pri tumačenju i razumijevanju. Ipak, valja znati da je ovaj algoritam sklon tzv. *overfittingu* (nepoželjno ponašanje strojnog učenja koje se događa kada model strojnog učenja daje točna predviđanja za istrenirane podatke, ali ne i za nove podatke [23]), posebice kod dubokih stabala te da mu je izražajnost ograničena u usporedbi s nekim drugim, složenijim algoritmima. [33] Često se stoga koristi algoritam kao što su Slučajne šume koje kombiniraju višestruka stabla odlučivanja. Detaljnije o istima opisat će se u nastavku.

Umjetne neuronske mreže

Korisni alat u bilo kakvom zadatku predviđanja jesu umjetne neuronske mreže (*Artificial Neural Networks*) - klasa modela strojnog učenja inspirirana strukturom i funkcijom ljudskog mozga. Sastoje se od međusobno povezanih čvorova, zvanih umjetni neuroni, organiziranih u slojeve (ulazni sloj, jedan ili više skrivenih slojeva i izlazni sloj). Svaki čvor, ili umjetni neuron, povezuje se s drugim i ima pridruženu težinu i prag. Ako je izlaz bilo kojeg pojedinačnog čvora iznad navedene vrijednosti praga, taj se čvor aktivira, šaljući podatke sljedećem sloju mreže. Inače se podaci ne prosljeđuju na sljedeći sloj mreže. [32] Prikaz strukture neuronske mreže dan je slikom 2.3.



Sl. 2.3. Prikaz algoritma umjetnih neuronskih mreža. Izvor slike: ResearchGate [40]

Postoje različite vrste neuronskih mreža (ovisno o njihovoj arhitekturi i načinu učenja), no bez obzira na oblik i broj veza unutar njih, princip rada je isti - uvježbavanje kroz niz ponavljajućih postupaka analize. Od cijeloga skupa podataka veći je dio upotrijebljen za učenje, a manji za ponovno predviđanje poznatih vrijednosti. Na taj je način moguće izračunati pogrešku predviđanja, koja bi s većim brojem pokušaja trebala biti manja. [13]

Neuronske mreže se naširoko primjenjuju u istraživanju i predviđanju jer mogu modelirati izrazito nelinearne sustave u kojima je odnos među varijablama nepoznat ili vrlo složen. Primjenjuju se u zadaćama klasifikacije predmeta, prepoznavanju oblika, govora ili rukom pisanog teksta, a neke su namijenjene modeliranju određenih pojava (npr. pojave srčanih bolesti).

Osim kod predviđanja rizika od srčanih bolesti, neuronske mreže mogu se koristiti i u predviđanju rizika od razvoja dijabetesa, otkrivanju i dijagnosticiranju stanja poput tumora, prijeloma, upale pluća i kožnih bolesti iz analize slika, modeliranju napredovanja bolesti te predviđanju interakcija lijekova.

Potrebno je ipak napomenuti da, iako neuronske mreže nude značajne prednosti i korist u iznad navedenim zadacima, njihova primjena u stvarnoj medicinskoj dijagnostici zahtijeva velike skupove podataka visoke kvalitete i pažljivu provjeru valjanosti modela kako bi se osigurali pouzdani rezultati i, još važnije, sigurnost pacijenata. U praksi se za izgradnju kvalitetnog sustava za predviđanje rizika od srčanih bolesti koriste neuronske mreže u kombinaciji drugih tehnika strojnog učenja.

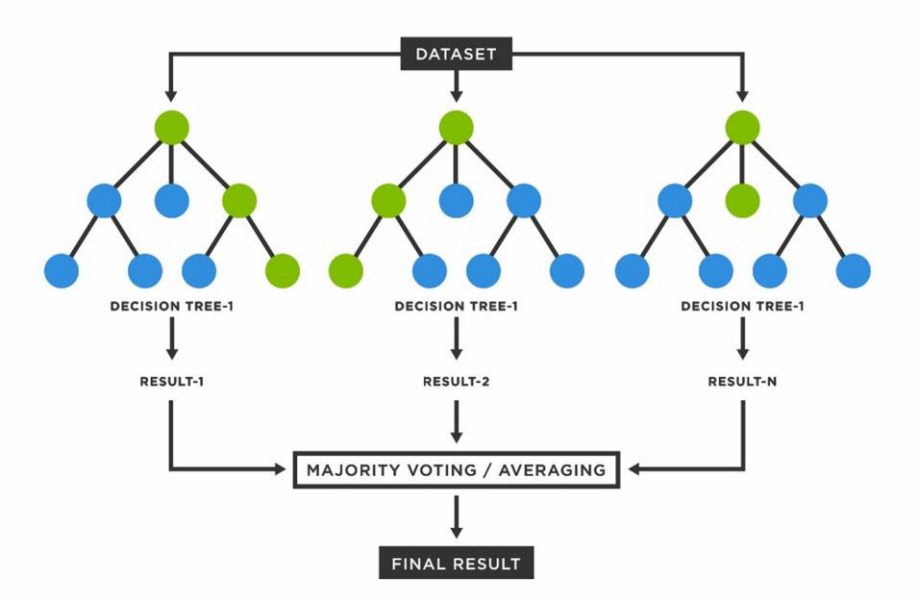
Slučajne šume

Slučajne ili nasumične šume predstavljaju algoritam strojnog učenja koji kombinira izlaz više stabala odlučivanja kako bi se postigao jedan rezultat. Ime „slučajne šume“ dolazi od činjenice da se ovaj algoritam sastoji od kombinacije više stabala odlučivanja. Iako su stabla odlučivanja široko primjenjivi algoritmi za strojno učenje, mogu biti sklona problemima, kao što su pristranost određenim rezultatima i *overfitting*. Međutim, kada više stabala odlučivanja formira ansambl u algoritmu slučajne šume, oni predviđaju točnije rezultate, osobito kada pojedinačna stabla nisu u međusobnoj korelaciji. [34]

Ova se metoda koristi za klasifikaciju (npr. otkrivanje neželjene pošte, dijagnozu bolesti i analizu raspoloženja), regresiju (predviđanje cijena kuća, cijena dionica i dugotrajne vrijednosti kupaca)

te detekciju anomalija (prepoznavanje neobičnih uzoraka u podacima). Upravo je to i velika prednost ove metode: može rukovati skupovima podataka s kontinuiranim varijablama, kao u regresiji, ali i sa kategoričkim varijablama, kao u klasifikaciji.

Princip rada algoritma slučajnih šuma prikazan je slikom 2.4.:



Sl. 2.4. Prikaz algoritma slučajnih šuma. Izvor slike: TIBCO Software [41]

Odabir odgovarajuće metode koja će se koristiti u rješavanju problema ovisi o karakteristikama korištenog skupa podataka, prirodi zadatka te o dostupnim resursima. Dobra je praksa isprobati više različitih metoda te na taj način usporediti njihovu izvedbu kako bi se odabralo najprikladniju za specifični zadatak. [15]

4. ANALIZA PODATAKA

4.1. Exploratory Data Analysis (EDA) i predobrada podataka

Prije upuštanja u bilo kakvo stvaranje modela, potrebno je upoznati, razumjeti i prirediti podatke koje se planira koristiti. Prilikom izrade modela nije uvijek slučaj da se na raspolaganju ima čiste i ispravno formatirane podatke. Budući da se često radi o podacima iz stvarnog svijeta, oni mogu sadržavati šumove, poneke vrijednosti mogu nedostajati ili biti u obliku koji nije pogodan za korištenje pri izradi modela. Stoga je prvi i ključni korak analiza i obrada podataka za korištenje. [18]

Exploratory Data Analysis (EDA), odnosno Istraživačka analiza podataka predstavlja proces istraživanja podataka kako bi se otkrili uzorci među podacima, uočile potencijalne anomalije te provjerile pretpostavke vezane za određeni skup podataka. Svrha ovakve analize jest da se podacima da smisao prije nego što ih se krene koristiti. [16] [31]

Najprije se uvozi potrebne biblioteke te dataset kojeg će se koristiti:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
import missingno as msno
from scipy import stats

data = pd.read_csv('heart_2020_cleaned.csv')
```

Ime koje je dodijeljeno datasetu jest „*data*“ te se taj dataset otvara pomoću *read_csv* naredbe iz **Pandas** biblioteke.

Pomoću *dtypes* naredbe dobija se rezultat koji pokazuje tipove varijabli u datasetu:

```
data.dtypes
```

```
HeartDisease         object
BMI                 float64
Smoking              object
AlcoholDrinking      object
Stroke               object
PhysicalHealth      float64
MentalHealth        float64
DiffWalking          object
Sex                  object
AgeCategory          object
Race                 object
Diabetic             object
PhysicalActivity     object
GenHealth            object
SleepTime           float64
Asthma               object
KidneyDisease        object
SkinCancer           object
dtype: object
```

Sl. 3.1. Tipovi varijabli koje se nalaze u korištenom datasetu

Kratki opis svake od varijabli:

- HeartDisease je target varijabla koja predstavlja srčanu bolest

- BMI (Body Mass Index) je broj koji predstavlja stupanj korespondencije između mase i visine osobe te omogućava procjenu je li masa nedovoljna, normalna ili prekomjerna

- Pušenje (Smoking) kao jedan od glavnih uzročnika bolesti srca

- Konzumacija alkohola (AlcoholDrinking) – alkohol može uzrokovati trajne smetnje u radu srca

- Moždani udar (Stroke) je često uzrokovan poremećajem krvotoka u arterijama i smanjenjem opskrbe mozga krvlju

- parametar Fizičko zdravlje (PhysicalHealth) predstavlja broj dana u mjesecu koliko osoba procjenjuje da se osjećala fizički slabo/loše

- parametar Mentalno zdravlje (MentalHealth) predstavlja broj dana u mjesecu koliko osoba procjenjuje da se osjećala mentalno slabo/loše

- Poteškoće u hodanju/penjanju na stepenice (DiffWalking)

- parametri Dobna skupina (AgeCategory), Rasa (Race) i Spol (Sex) kao osnovni opisni parametri pojedine osobe i utjecajni parametri na rizik od bolesti srca

- Dijabetes (Diabetes) jer osobe s dijabetesom imaju veću šansu razvoja bolesti srca

- Fizička aktivnost (PhysicalActivity) – bavljenje tjelesnom aktivnošću (koja ne uključuje redoviti posao osobe) unazad mjesec dana

- Zdravlje općenito (GenHealth) izuzetak ili postojanje nekih drugih zdravstvenih poteškoća koje nisu spomenute u ranijim varijablama

- Broj sati sna (SleepTime)

- Astma (Asthma) jer osobe s astmom imaju veću šansu razvoja bolesti srca

- Bolesti bubrega (KidneyDisease) i Rak kože (SkinCancer) kategorije su uključene jer ti slučajevi povećavaju šansu za razvoj bolesti srca

(Za detaljniji opis rizika pojedinih varijabli pogledati poglavlje 2.1. Utjecajni čimbenici)

Naredba *shape* vraća kao rezultat oblik dataseta:

```
data.shape
```

```
(319795, 18)
```

Sl. 3.2. Rezultat *Shape* naredbe

Vidi se da dataset ima 18 redova i 319795 stupaca.

Naredbom *isnull().values.any()* provjerava se fali li ijedna vrijednost varijable u cijelom datasetu:

```
data.isnull().values.any()
```

```
False
```

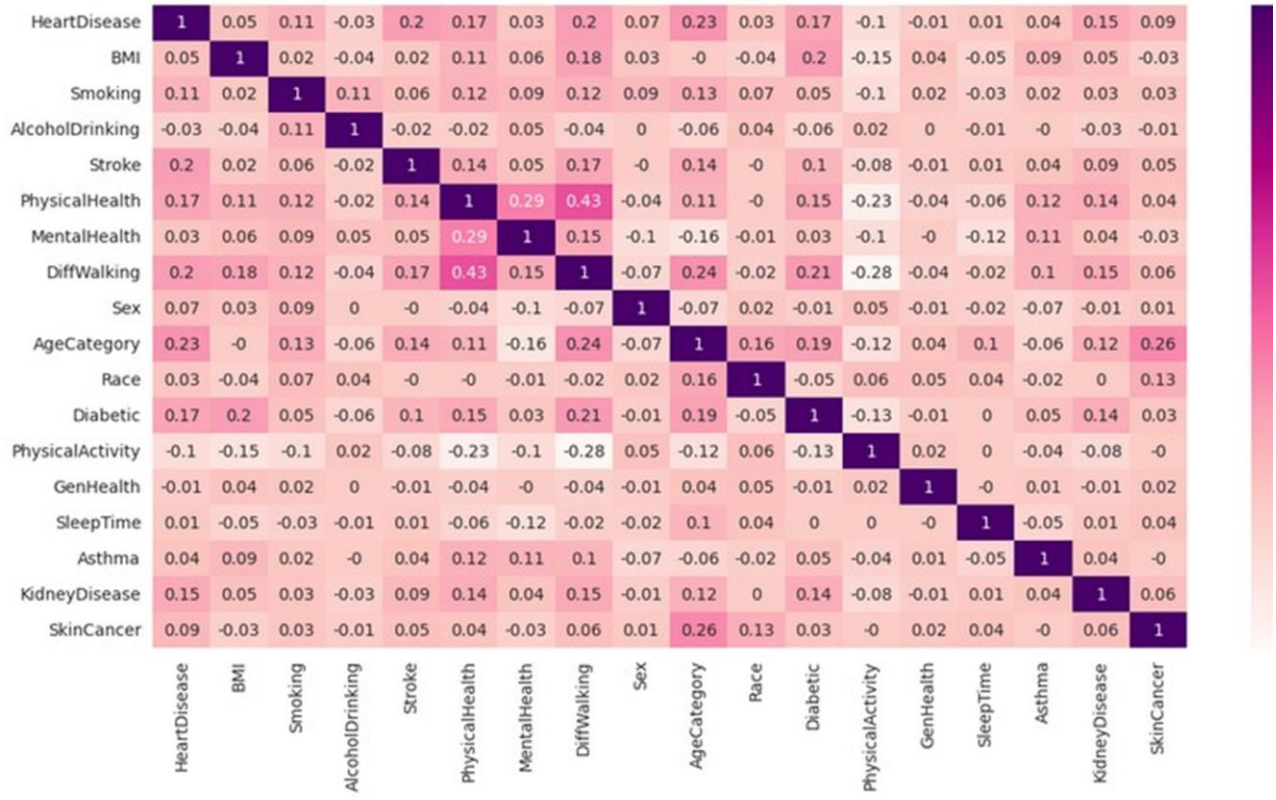
Sl. 3.3. Rezultat *isnull().values.any()* naredbe je False dakle nema vrijednosti koje fale

U ovom slučaju ne fali niti jedna vrijednost.

Kako bi se vidjelo međusobnu korelaciju između podataka u datasetu, koristi se **Matplotlib** biblioteku te naredbe za stvaranje korelacijske matrice:

```
correlation = data.corr().round(2)
plt.figure(figsize = (14, 7))
sns.heatmap(correlation, annot = True, cmap = "RdPu")
```

```
<Axes: >
```



Sl. 3.5. Korelacijska matrica

Što je boja polja tamnija, to je korelacija između 2 podatka veća.

Iz korelacijske matrice se može vidjeti da GenHealth, MentalHealth i SleepTime varijable nemaju prevelik utjecaj na ishod (na HeartDisease varijablu kao target varijablu), stoga ih se može izbaciti. Izbacivanje se vrši pomoću *drop* naredbe:

```
data.drop(['MentalHealth', 'GenHealth', 'SleepTime'], axis=1, inplace=True)
```

5. IZGRADNJA MODELA

Korak koji slijedi nakon EDA je tzv. feature engineering, to jest, inženjering značajki podataka koje se koristi. Dok je glavni cilj EDA razumijevanje podataka, vizualizacija istih te identifikacija uzoraka, anomalija i odnosa među podacima, inženjering značajki je specifičniji proces, usmjeren na poboljšanje performansi modela, a koji uključuje transformaciju podataka u oblike pogodne za korištenje, stvaranje značajki koje potencijalno nedostaju te odabir značajki koje će biti od vrijednosti za korištenje u zadatku. EDA dakle omogućava stjecanje uvida u strukturu, kvalitetu, ali i potencijalne izazove podataka te pomaže usmjeriti slijedeće korake, kao što je inženjering značajki, u procesu izrade modela. [20]

U ovom će poglavlju biti detaljno opisani koraci inženjeringa značajki te koraci izrade modela strojnog učenja za predviđanje rizika od srčanih bolesti te će biti priložen odgovarajući kod.

5.1. Enkodiranje kategoričkih vrijednosti i normaliziranje brojčanih vrijednosti

Prvi korak je enkodiranje kategoričkih vrijednosti. Kao što se može vidjeti iz slike 3.1. u poglavlju 4.1. (Exploratory Data Analysis (EDA) i predobrada podataka), postoje varijable koje su tipa „object“, to jest kategoričke varijable u ovom datasetu. Kada se radi sa kategoričkim varijablama, često ih je potrebno pretvoriti u numerički oblik kako bi ih algoritmi mogli obraditi. Taj se proces naziva „Enkodiranje kategoričkih varijabli“. Postoje različiti načini enkodiranja, a u ovom projektu koristi se *OrdinalEncoder* klasa iz **Scikit-learn** biblioteke:

```
from sklearn.preprocessing import OrdinalEncoder
enc = OrdinalEncoder()
enc.fit(data[categorical_features])
data[categorical_features] = enc.transform(data[categorical_features])
```

Svakoj kategoričkoj varijabli dodjeljuje se jedinstvena cjelobrojna vrijednost na temelju njenog redoslijeda ili ranga te ju se tako pretvara u numeričku vrijednost.

Osim što je potrebno enkodirati kategoričke vrijednosti, valja i normalizirati brojčane vrijednosti. Normalizacija služi za skaliranje numeričkih vrijednosti na određeni standardni raspon. Normalizacija je u ovom projektu odrađena pomoću *normalize* naredbe iz **Scikit-learn** biblioteke:

```
from sklearn import preprocessing
normalized_data = preprocessing.normalize(data)
scaled_data = pd.DataFrame(normalized_data, columns = data.columns)
print(scaled_data)
```

Nakon normaliziranja, sve vrijednosti će se nalaziti u rasponu između 0 i 1.

5.2. Treniranje modela

Posljednji korak jest treniranje samog modela. Budući da je cilj klasificirati podatke u 1 od 2 moguće klase (osoba ima bolest srca/osoba nema bolest srca), koristit će se *binarna klasifikacija*. Binarna klasifikacija je vrsta strojnog učenja koju se koristi kako bi se model naučilo točno predvidjeti klase novih instanci na temelju ulaznih varijabli. Ciljna varijabla (*target* varijabla) poprima 2 različite vrijednosti: 0 (negativno) i 1 (pozitivno). Upotreba binarne klasifikacije je česta u modelima predviđanja i dijagnoze bolesti gdje se predviđa ima li osoba bolest (1, pozitivno) ili nema (0, negativno).

Postoje razni algoritmi koje možemo koristiti za binarnu klasifikaciju, a najprikladniji se odabire, kao što je već objašnjeno u poglavlju 3.2., ovisno o potrebama modela. Algoritam odabran za ovaj model je Logistička Regresija (*Logistic Regression*). To je metoda koja se koristi za klasifikaciju i modeliranje vjerojatnosti pojave određenog događaja. Ishod je vjerojatnost, stoga je zavisna varijabla ograničena između 0 i 1. Za izvođenje Logističke regresije u Pythonu može se upotrijebiti **Scikit-learn** biblioteka:

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

X_train, X_test, y_train, y_test = train_test_split(features, target, shuffle
= True, train_size = 0.8, test_size = 0.2, random_state = 102)

model = LogisticRegression(class_weight='balanced')
model.fit(X_train, y_train)
```

Funkcija *train_test_split* koristi se za podjelu podataka u skup za treniranje (80% podataka) i skup za testiranje (20% podataka). Parametar „random_state“ je proizvoljno odabrana cjelobrojna vrijednost koja služi za kontrolu nasumičnog miješanja i dijeljenja skupa podataka. Instanci model dodjeljuje se *LogisticRegression* klasa te se koristi *fit* metoda za treniranje modela.

6. REZULTATI I RASPRAVA

Nakon što je model istreniran, može ga se koristiti za predviđanje na temelju novih podataka koristeći *predict* metodu:

```
y_pred = model.predict(X_test)
print('Accuracy score: ', accuracy_score(y_test, y_pred))
```

Naposlijetku, korisno je dodati metrike koje klasificiraju izvedbu modela. U ovom slučaju to su točnost modela, matrica grešaka, ROC-AuC i ROC krivulja, osjetljivost, specifičnost i F1.

1. *Accuracy_score* funkcija vraća točnost modela. U ovom slučaju iznosi: 0.7256 odnosno 72.56%.

```
y_pred = model.predict(X_test)
print('Accuracy score: ', accuracy_score(y_test, y_pred))
```

Iako je takva točnost na prvi pogled zadovoljavajuća, mora se uzeti u obzir to da je ovaj parametar u stvarnosti relevantan samo kod slučajeva gdje je podatkovni skup balansiran (ovdje bi to značilo da je 50% ljudi bolesno, a drugih 50% zdravo, što nije slučaj). Razlog tomu je svojstvo parametra točnosti da sve klase tretira kao jednako važne pa ne uzima u obzir lažno pozitivne (*false positive* – osoba nema rizik od bolesti srca, ali model predvidi da ima) i lažno negativne (*false negative* – osoba ima rizik od bolesti srca, ali model to ne predvidi) rezultate, već ih sve tretira kao točna predviđanja. [36] Jasno je da u medicinskoj primjeni, lažno negativan rezultat može biti opasan pa čak i fatalan. Iz tog se razloga uz točnost koriste i drugi parametri kako bi se procijenila stvarna izvedba modela.

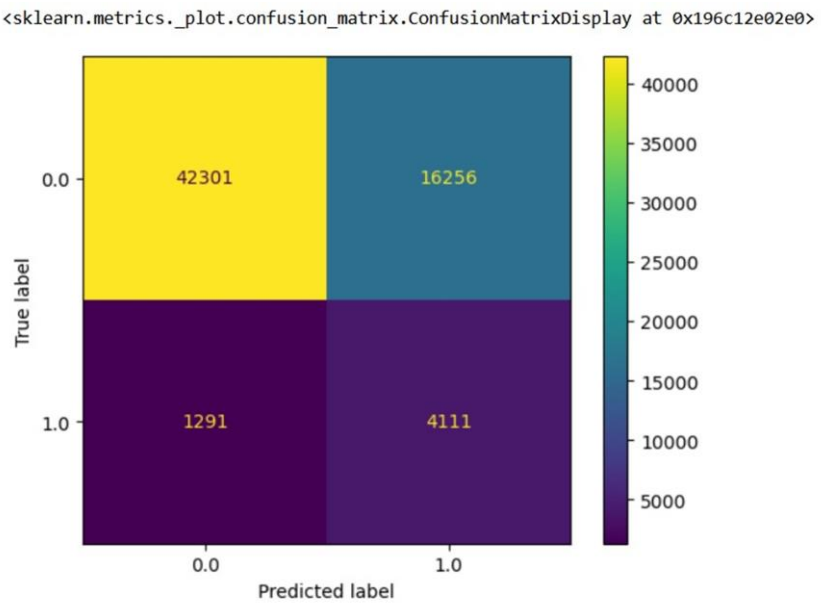
2. Confusion matrix predstavlja sažetak predviđanja modela u obliku matrice – prikazuje koliko je točnih/netočnih predviđanja po klasi:

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
[[42301 16256]

 [ 1291  4111]]
```

Model je odradio 42301 točnih negativnih i 16256 točnih pozitivnih predviđanja te 1291 netočnih negativnih i 4111 netočnih pozitivnih previđanja.

```
from sklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(model, X_test, y_test)
```



Sl. 6.1. Vizualizacija Confusion matrice

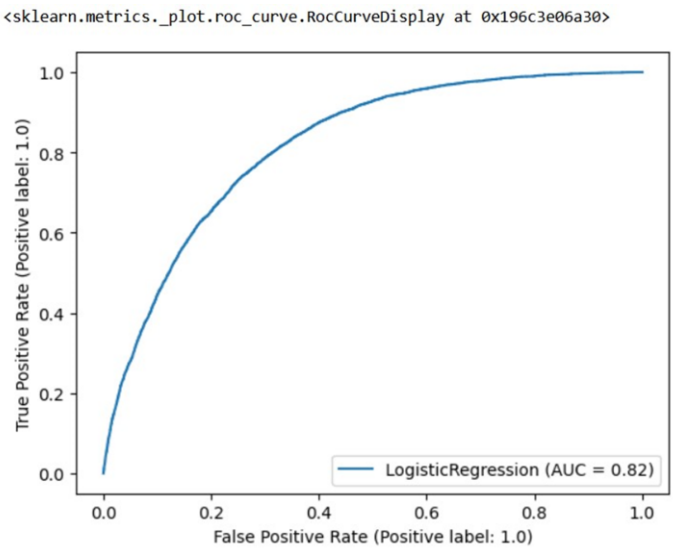
3. „*Receiver Operating Characteistic – Area under the Curve*“ odnosno ROC-AuC rezultat (score) daje podatak o tome koliko je model učinkovit – što je veći AuC to je izvedba modela bolja u razlikovanju pozitivnih i negativnih klasa, npr. AuC rezultat '1' bi značio da model može savršeno razlikovati pozitivne i negativne klase:

```
from sklearn.metrics import roc_auc_score

auc_score = roc_auc_score(y_test, y_pred)
print("ROC AuC score:", auc_score)


ROC AuC score: 0.7417022944326163
```

4. ROC krivulja je grafička reprezentacija iznad opisanog parametra: `plot_roc_curve(model, X_test, y_test)`



Sl. 6.2. ROC curve

5. Osjetljivost, specifičnost i F1 metrike se koriste u slučajevima kada točnost nije relevantna.

```
model_eval = evaluate_model(model, X_test, y_test)

print('Precision:', model_eval['prec'])
print('Recall:', model_eval['rec'])
print('F1 Score:', model_eval['f1'])
```

Osjetljivost ili *recall* mjeri sposobnost modela da točno identificira stvarno pozitivne klase, to jest daje odgovor na pitanje „Od svih stvarnih pozitivnih slučajeva, koliko ih je model točno predvidio kao pozitivne slučajeve?“. Na ovom primjeru to bi bio broj slučajeva gdje osoba zaista ima bolest srca te je model to uspješno predvidio. Formula za osjetljivost je slijedeća:

$$\text{Recall (Sensitivity/True Positive Rate)} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

Sl. 6.3. Formula za izračun osjetljivosti. Izvor slike: ResearchGate [42]

A model ima osjetljivost od 0.7464 odnosno 74.6% što znači da je model ispravno identificirao 75% slučajeva gdje osoba zaista ima srčanu bolest od ukupnog broja takvih slučajeva.

Specifičnost ili precision, s druge strane mjeri koliko je, od svih slučajeva koje je model predvidio kao pozitivne, stvarno pozitivnih slučajeva. Specifičnost koja iznosi 1 bi značila da su sve instance koje je model predvidio kao pozitivne, zaista pozitivne. Formula za specifičnost je:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Sl. 6.4. Formula za izračun specifičnosti Izvor slike: ResearchGate [42]

Često se između osjetljivosti i specifičnosti mora postići kompromis budući da povećanje jednog od ta dva parametra može dovesti do smanjenja drugog. Npr. povećanjem specifičnosti, model postaje selektivniji u pozitivnim predviđanjima te se time smanjuje osjetljivost i obrnuto.

U ovom slučaju model ima specifičnost od 0.1999 odnosno 19.9%. Kao što se može vidjeti, specifičnost je znatno manja od osjetljivosti, što je u ovome slučaju bolje nego obrnuti ishod. Naime, ukoliko model za nekog pacijenta predvidi da ima rizik od srčane bolesti, no taj pacijent zapravo nema rizik, šteta nije velika. S druge strane, ako model ne uspije predvidjeti rizik od srčane bolesti kod pacijenta, a pacijent je u stvarnosti rizičan, to može stvoriti znatan problem. Iz tog je razloga bolje imati veliku osjetljivost (model je uspio predvidjeti što više osoba s rizikom), nego veliku specifičnost (model je predvidio velik broj osoba s rizikom od kojih nisu sve zaista rizične) .

Naposljetku, F1 parametar kombinira osjetljivost i specifičnost kako bi se dobila uravnotežena mjera izvedbe modela. Cilj ovog parametra jest postizanje ravnoteže između osjetljivosti i specifičnosti. Posebno je koristan u slučajevima kada postoji neravnoteža između pozitivnih i negativnih klasa te kada se želi procijeniti sposobnost modela da daje točna pozitivna, a izbjegava lažno pozitivna predviđanja. Formula je:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Sl. 6.5. Formula za izračun F1. Izvor slike: ResearchGate [42]

F1 ocjena modela iznosi 0.3154 odnosno 31.5%. To znači da ravnoteža između osjetljivosti i specifičnosti nije baš velika, no kao što je iznad objašnjeno, u ovom slučaju to i nije od velike važnosti, s obzirom da je kod predviđanja rizika od bolesti važnije imati veliku osjetljivost.

U nastavku su prikazane pojedine ulazne vrijednosti i izlazna vrijednost (1 – ima rizik od bolesti srca/0 – nema rizik od bolesti srca) za nasumično odabrane pojedince iz skupa:

```
        Row Index  HeartDisease    BMI  Smoking  AlcoholDrinking  DiffWalking  \
133506     133507           0.0  42.51      0.0              0.0          0.0
1162         1163           1.0  41.60      1.0              0.0          1.0
173415     173416           1.0  50.36      1.0              0.0          1.0
216472     216473           1.0  28.13      1.0              0.0          0.0
23184       23185           0.0  17.81      0.0              0.0          0.0
85038       85039           0.0  30.54      0.0              0.0          0.0
85681       85682           0.0  41.71      1.0              1.0          0.0
153047     153048           1.0  32.92      0.0              0.0          0.0

        Diabetic  Asthma  KidneyDisease  SkinCancer
133506       0.0     1.0            0.0         0.0
1162         2.0     1.0            1.0         0.0
173415       2.0     0.0            0.0         1.0
216472       2.0     0.0            0.0         1.0
23184        0.0     0.0            0.0         0.0
85038        0.0     0.0            0.0         0.0
85681        0.0     0.0            0.0         0.0
153047       2.0     1.0            0.0         0.0
```

Sl. 6.6. Ulazne vrijednosti i pripadajuća izlazna vrijednost za nasumično odabrane pojedince

Za lakše razumijevanje vrijednosti su prikazane i u slijedećoj tablici:

	Indeks osobe	Rizik od bolesti srca	BMI	Pušenje	Konzumacija alkohola	Poteškoće u hodanju	Dijabetes	Astma	Bolest bubrega	Rak kože
1	133506	0	42.51	0	0	0	0	1	0	0
2	1162	1	41.60	1	0	1	2	1	1	0
3	173415	1	50.36	1	0	1	2	0	0	1
4	216472	1	28.13	1	0	0	2	0	0	1
5	23184	0	17.81	0	0	0	0	0	0	0
6	85038	0	30.54	0	0	0	0	0	0	0
7	85681	0	41.71	1	1	0	0	0	0	0
8	153047	1	32.92	0	0	0	2	1	0	0

Tablica 6.1. Ulazne vrijednosti i pripadajuća izlazna vrijednost za nasumično odabrane pojedince

Iz tablice se da zaključiti da su pojedinci kojima je predviđen rizik od bolesti srca u pravilu dijabetičari i pušači, osim kod 8. osobe u tablici kojoj je predviđen rizik iako ista nije pušač. BMI kod pojedinaca s predviđenim rizikom je također u pravilu veći, ali to nije slučaj kod 1. i 7. osobe u tablici koje, iako im je BMI vrlo visok, nemaju predviđen rizik od bolesti srca. Bitno je primijetiti i kako, unatoč tomu što je konzumacija alkohola dokazan rizični čimbenik, u ovom slučaju ni jedan od pojedinaca kojima je predviđen rizik nisu redoviti u konzumiranju alkohola. Poteškoće u hodanju, astma i rak kože su čimbenici od kojih prema rezultatima tablice svaki ima 50% utjecajnosti na rizik – od 4 osobe s predviđenim rizikom, samo po 2 osobe imaju ili su imale poteškoće u hodanju/astmu/rak kože. Uz konzumaciju alkohola, bolesti bubrega imaju najmanji utjecaj na rizik.

Jasno je dakle kako je previđanje rizika od bolesti srca vrlo delikatan proces te da pojedinci kod kojih čimbenici gotovo sigurno ukazuju na mogućnost razvoja bolesti srca, u stvarnosti mogu biti i ostati zdravi. Valja ipak imati na umu i to da su neki pojedinci koji su u tablici (pa tako i u čitavom modelu) prikazani bez rizika, mogu pripadati u onih 25% koje ovaj model ne predviđa ispravno.

7. ZAKLJUČAK

Kako bi se zadatak predviđanja rizika od srčanih bolesti pomoću strojnog učenja uspješno izveo, potrebna su najprije znanja iz područja umjetne inteligencije i strojnog učenja, znanja o Pythonu kao odabranom jeziku za korištenje u zadatku, upoznavanje sa medicinskom pozadinom srca, srčanih bolesti i rizika koji ih uzrokuju te temeljito poznavanje podataka koje se pri strojnom učenju u ovome zadatku koristi. Za početak je bilo potrebno razumjeti kako radi Jupyter Notebook koji se koristio za realizaciju ovog zadatka, upoznati se sa Kaggleom kao platformom sa koje su preuzeti korišteni podaci i naravno, naučiti dobro baratati Pythonom. U tim su koracima korišteni ponajviše online i video tečajevi, ali i stručna literatura osigurana od strane mentora.

Sljedeći je korak bio upoznati se sa širokim spektrom čimbenika koji mogu dovesti do razvoja srčane bolesti, razumjeti zašto do istih dolazi, to jest, kako i koliko koji rizični čimbenik utječe na razvoj bolesti. Zaključeno je da, npr. pušenje, pretilost i stres znatno utječu na razvoj srčane bolesti jer povisuju razinu krvnog tlaka i razinu kolesterola u krvi, dok su čimbenici kao radijacija i količina sna dokazano povezani sa razvojem bolesti, ali u znatno manjim količinama. Zatim je bilo potrebno upoznati, očistiti i urediti skup podataka koji se u zadatku koristi, što je ujedno i najveći izazov te odnosi najviše vremena s obzirom na to da je potrebno dobro shvatiti što koja varijabla predstavlja u skupu podataka, kako se varijable odnose jedna na drugu te za svaku odlučiti igra li značajnu ulogu u određivanju rizika od bolesti. U ovom je zadatku od značajne pomoći bilo prethodno izvedeno istraživanje i shvaćanje rizičnih čimbenika.

Sa matematičke, statističke strane istražen je pojam „Coxov regresijski model“ te su istraženi pojedini, već postojeći modeli temeljeni na Coxovom i njihove prednosti/nedostaci. Osim modela, trebalo se upoznati i sa statističkim algoritmima koje je moguće koristiti u svrhu predviđanja, a za ovaj zadatak je nakon detaljnije procjene odabrana logistička regresija kao algoritam. Za kraj su ostali izgradnja i treniranje modela te procjena rezultata istoga. Pri tim su zadacima kao referenca korišteni slični postupci, to jest, primjeri logističke regresije na nekim drugim skupovima. Dobiva se Accuracy score odnosno ocjena točnosti modela od otprilike 0.7257, odnosno oko 72.57%, no budući da korišteni podatkovni skup nije balansiran, ta metrika nije od prevelike koristi. Iz tog se razloga mjere dodatni parametri, a to su ROC AuC, osjetljivost, specifičnost i F1. Rezultati mjerenja istih pokazuju da model ima ROC AuC ocjenu od 74.2%, specifičnost od 19.9%, F1 ocjenu od 31,5% te najvažnije, osjetljivost koja iznosi 74,6% . Da se zaključiti iz danih metrika kako model zaista vrši previđanje, a ne samo obično

„pogađanje“ rezultata. Iako postoje mnogi modeli koji bi imali puno bolje ocjenu točnosti izvedbe i ovaj model izvršava svoj zadatak. Dodatna poboljšanja modela su moguća na boljom filtracijom i obradom podataka, potencijalnim isprobavanjem ili korištenjem drugih algoritama, boljim rukovanjem neuravnoteženog skupa (npr. podjelom skupa na više manjih skupa koji su bolje uravnoteženi), dodatnim skaliranjem i normalizacijom varijabli, prikupljanjem više podataka te općenito boljim upoznavanjem sa domenom strojnog učenja.

8. SAŽETAK

U ovom je radu u Pythonu razvijen model logističke regresije namijenjen za korištenje pri predviđanju rizika od razvoja srčanih bolesti. Nakon upoznavanja podataka te detaljne obrade istih, izgradnje i treniranja modela, dobivena je ocjena osjetljivosti od otprilike 0.746 što znači da model od 100% rizičnih slučajeva točno predviđa prisutnost rizika u 74.7% slučajeva. To ukazuje na to da, iako ne savršeno, model zaista predviđa rizik te ga se može koristiti u medicinske svrhe. Model je dakle ispunio početni zadatak predviđanja rizika, a sa daljnjim usavršavanjem i proširenjem s dodatnim podacima, tehnikama i znanjem potencijalno bi se mogla i poboljšati njegova prediktivna točnost i izvedba.

Motivirano činjenicom da srčane bolesti u današnje vrijeme zauzimaju broj 1 kao glavni uzrok smrti među ljudima, ovakav model predviđanja može itekako biti od koristi jer je prevencija najbolji lijek.

Ključne riječi

strojno učenje, Python, predviđanje rizika, srce, srčane bolesti

8.1. Summary

In this assignment, a logistic regression model was developed in Python to be used in predicting the risk of developing heart disease. After getting familiar with the data and detailed processing of it, building and training the model, a recall score of approximately 0.726 was given for the model, meaning that the model accurately predicts the presence of risk in 72.7% out of a 100% of positive cases, which indicates that, although not perfectly, the model really predicts risk. The model therefore fulfilled the initial task of risk prediction, and with further refinement and extension of the model with additional data, techniques and knowledge, its predictive accuracy and performance could potentially be improved.

Motivated by the fact that heart disease is currently the number 1 cause of death among humans, this kind of prediction model can be very useful since prevention really is the best medicine.

Keywords

machine learning, Python, risk prediction, heart, heart disease

LITERATURA

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.[21] „CVD Statistics“. European Heart Network. *European Cardiovascular Disease Statistics 2017*. 2017. https://ehnheart.org/cvd-statistics.html (pristupljeno 07. rujan 2023.).

[22] „What is Logistic regression? | IBM“. https://www.ibm.com/topics/logistic-regression (pristupljeno 07. rujan 2023.).

[23] „What is Overfitting? | IBM“. https://www.ibm.com/topics/overfitting (pristupljeno 07. rujan 2023.).

[24] „FastStats“, 18. siječanj 2023. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm (pristupljeno 07. rujan 2023.).

[25] „1.10. Decision Trees“, *scikit-learn*. https://scikit-learn/stable/modules/tree.html (pristupljeno 07. rujan 2023.).

[26] Swaminathan, Saishruthi. *Logistic Regression — Detailed Overview*. 15. March 2018. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc (pokušaj pristupa 2023).

[27] „The top 10 causes of death“. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (pristupljeno 07. rujan 2023.).

[28] „The impact of alcohol consumption on cardiovascular health“, *World Heart Federation*. https://world-heart-federation.org/news/no-amount-of-alcohol-is-good-for-the-heart-says-world-heart-federation/ (pristupljeno 07. rujan 2023.).

[29] L. M. Tierney Jr., S. Saint, M. A. Whooley, „Essentials of Diagnosis & Treatment“, Lange Medical Books/McGraw – Hill, 2002.

[30] R. Jurilj, I. Božić, „Ehokardiografija – drugo, dopunjeno i obnovljeno izdanje“, Medicinska naklada, Zagreb, 2013., https://issuu.com/medicinskanaklada/docs/ehokardiografija

[31] A. D’Agostino, „Exploratory Data Analysis in Python — A Step-by-Step Process“, *Medium*, 22. kolovoz 2023. https://towardsdatascience.com/exploratory-data-analysis-in-python-a-step-by-step-process-d0dfa6bf94ee (pristupljeno 07. rujan 2023.).

[32] „What are Neural Networks? | IBM“. https://www.ibm.com/topics/neural-networks (pristupljeno 07. rujan 2023.).

[33] „What is a Decision Tree | IBM“. https://www.ibm.com/topics/decision-trees (pristupljeno 07. rujan 2023.).

[34] „What is Random Forest? | IBM“. https://www.ibm.com/topics/random-forest (pristupljeno 07. rujan 2023.).

[35] „List of Top 10 Libraries in Python (2023)“, *InterviewBit*, 04. travanj 2023. https://www.interviewbit.com/blog/python-libraries/ (pristupljeno 07. rujan 2023.).

[36] „Accuracy vs. precision vs. recall in machine learning: what’s the difference?“ https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall (pristupljeno 07. rujan 2023.).

[37] „Atherosclerosis“. https://www.hopkinsmedicine.org/health/conditions-and-diseases/atherosclerosis (pristupljeno 07. rujan 2023.).

[38] A. Pant, „Introduction to Logistic Regression“, Medium, 22. siječanj 2019. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 (pristupljeno 07. rujan 2023.).

[39] „Decision Tree Algorithm in Machine Learning - Javatpoint“. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (pristupljeno 07. rujan 2023.).

[40] „Figure 4: A hypothetical example of Multilayer Perceptron Network.“, ResearchGate. https://www.researchgate.net/figure/A-hypothetical-example-of-Multilayer-Perceptron-Network_fig4_303875065 (pristupljeno 07. rujan 2023.).

[41] „What is a Random Forest?“, TIBCO Software. https://www.tibco.com/reference-center/what-is-a-random-forest (pristupljeno 07. rujan 2023.).

[42] „What is the best metric (precision, recall, f1, and accuracy) to evaluate the machine learning model for imbalanced data?“, ResearchGate. https://www.researchgate.net/post/What_is_the_best_metric_precision_recall_f1_and_accuracy_to_evaluate_the_machine_learning_model_for_imbalanced_data (pristupljeno 07. rujan 2023.).