Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek

Računarstvo usluga i analiza podataka

SEMINARSKI RAD

Predikcija srčanog udara

Dominik Škrinjar

Osijek, 2023

Sadržaj

[1.Uvod 3](#_Toc145060885)

[2.Opis problema 4](#_Toc145060886)

[2.1 Korišteni podaci 4](#_Toc145060887)

[2.2 K najbližih susjeda 5](#_Toc145060888)

[2.3 Logistička regresija 5](#_Toc145060889)

[2.4 Klasifikator slučajnih šuma 6](#_Toc145060890)

[2.5 Support Vector Machine (SVM) 6](#_Toc145060891)

[2.5 XGBoost (Extreme Gradient Boosting) 6](#_Toc145060892)

[3. Opis programskog rješenja 6](#_Toc145060893)

[3.1 Model strojnog učenja. 7](#_Toc145060894)

[3.2 Web aplikacija 8](#_Toc145060895)

[Zaključak: 11](#_Toc145060896)

[Literatura: 11](#_Toc145060897)

[Prolozi 12](#_Toc145060898)

# 1.Uvod

Ovaj seminarski rad istražuje predviđanje srčanog udara na temelju podataka preuzetih s Kaggle stranice. Cilj je izraditi model koji će na temelju dostupnih podataka predvidjeti jel li osoba u opasnosti od srčanog udara. Za izradu modela koristi se platforma Colab i programski jezik Python i njegove različite biblioteke.

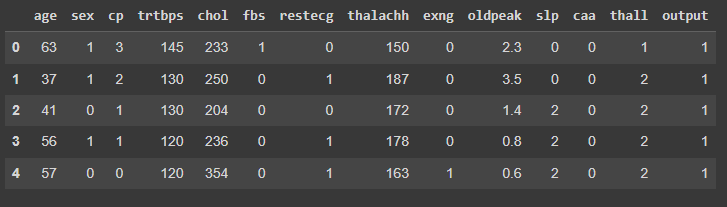
U rada će biti prikazan postupak pripreme podataka, izbor modela te evolucija točnosti predviđanja.

# 2.Opis problema

Problem predviđanja srčanog udara se temelji na zdravstvenim podacima pacijenta. Cilj nam je razviti model koji će predvidjeti srčani udar na temelju podataka o dobi pacijenta, spolu , angina izazvana vježbanjem, tip boli u prsima, krvni tlak, kolesterol , šećer u krvi, elektrokardiografski rezultati,broj otkucaja srca. Postoji nekoliko sličnih pristupa za predviđanje srčanog udara i korištenje različitih tehnika strojnog učenja poput regresije, neuronskih mreža i modela zasnovanim na stablima odlučivanja.Sve ove metode i pristupi mogu se kombinirati kako bi se stvorio učinkovit model za predviđanje srčanog udara.

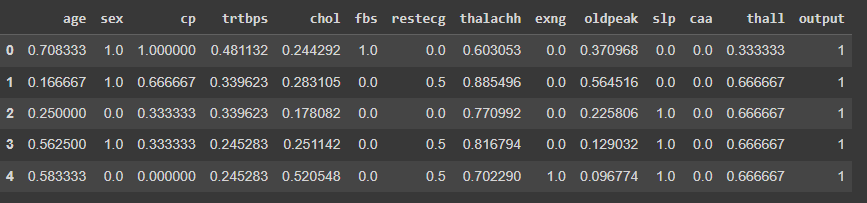
## 2.1 Korišteni podaci

Podaci korišteni za izradu seminara preuzeti su s Kaggle stranice.. Kaggle predstavlja online zajednicu gdje korisnici mogu pronaći i objaviti *datasets* odnosno skupove podataka te istraživati o izrađenim modelima na pojedinim skupovima. Odabrani skup za izradu ovog projekta, 'Predikcija srčanog udara ', sadrži sljedeće podatke: Dob, spol, angina uzrokovana tjelesnom aktivnošću, broj glavnih krvnih žila,tip boli u prsima, krvni tlak, kolesterol, šećer u krvi, rezultati elektrokardiografije,brzina otkucaja srca , depresija ST segmenta izazvana tjelesnom aktivnošću u odnosu na mirovanje, nagib ST segmenta tijekom vršne tjelesne aktivnosti i jel pacijent ima više ili manje od 50% šanse za srčanim udarom. Svi ti podaci nalaze u *heart.csv* datoteci. U prilozima se nalazi detaljnije objašnjenje podataka.



*Slika.1 Prikaz podataka dataseta Heart.csv*

Prije korištenja podataka moraramo skalirati podatke pomoću biblioteke skrelean koristimo funkcije MinMaxScaler() ili StandradScaler(), jer različita dimenzija ulaznih veličina može uzrokovati probleme kod izgradnje modela.



*Slika.2 Prikaz skaliranih podataka*

## 2.2 K najbližih susjeda

K-najbližih susjeda je algoritam za klasifikaciju u strojnom učenju koji se koristi za nadzirano učenje. Ovaj algoritam se često koristi za rješavanje problema klasifikacije, gdje je cilj odrediti kojoj klasi pripada određeni ulazni podatak na temelju sličnosti s drugim podacima u skupu za učenje.

Osnovna ideja K-najbližih susjeda je da novi podatak pripada istoj klasi kao i većina njegovih "k-najbližih" susjeda. "k" u ovom slučaju predstavlja broj najbližih susjeda koje algoritam uzima u obzir pri donošenju odluke o klasi novog podatka. Ova vrijednost "k" se obično odabire unaprijed i može utjecati na performanse algoritma.

Ovaj algoritam je popularan zbog svoje jednostavnosti i relativne efikasnosti, ali može biti manje učinkovit u slučajevima s velikim skupovima podataka ili visokim dimenzionalnim prostorima.

## 2.3 Logistička regresija

Logistička regresija (Logistic Regression) je statistički model koji se često koristi u strojnom učenju za rješavanje problema binarne klasifikacije. Ovaj model se također može prilagoditi za višeklasnu klasifikaciju. Glavni cilj logističke regresije je predviđanje vjerojatnosti da neki ulazni podatak pripada određenoj klasi.

Ključna prednost logističke regresije je njezina jednostavnost i interpretabilnost. Može se koristiti za razumijevanje kako svaka značajka doprinosi predikciji klasifikacije. Osim toga, logistička regresija može se koristiti za rješavanje problema s vjerojatnostima, kao što su procjene vjerojatnosti događaja.

## 2.4 Klasifikator slučajnih šuma

Random Forest Classifier (Klasifikator slučajnih šuma) je popularan algoritam u strojnom učenju koji se koristi za različite zadatke, uključujući klasifikaciju i regresiju. Ovaj algoritam temelji se na asembliranju, što znači da kombinira više stabala odlučivanja (decision trees) kako bi donio konačne odluke.

Ovaj algoritam često se koristi za raznolike primjene, uključujući medicinu, financije, analizu teksta i slično. Važno je napomenuti da se Random Forest Classifier može koristiti za klasifikaciju i regresiju, ovisno o postavkama modela.

## 2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM), ili na hrvatskom podržani vektor stroj, je algoritam za nadzirano strojno učenje koji se koristi za klasifikaciju i regresiju. Ovaj algoritam je posebno moćan u rješavanju problema klasifikacije i može se primijeniti na različite domene, uključujući medicinu, računalni vid, financije i mnoge druge.

Osnovna ideja SVM-a je pronalaženje hiperplohe (u dvodimenzionalnom prostoru to je prava linija, a u višedimenzionalnom prostoru je ravan) koja najbolje razdvaja podatke različitih klasa. Ova ploha se naziva "granična ploha" ili "graničnik", a njezina posebnost je da je maksimalno udaljena od najbližih točaka iz oba razreda. Te najbliže točke su podržani vektori (support vectors).

## 2.5 XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting), ili na hrvatskom ekstremno poticanje gradijentom, je popularan algoritam za strojno učenje koji se koristi za rješavanje problema klasifikacije, regresije i rangiranja. Ovaj algoritam spada u kategoriju ansambl metoda i temelji se na ideji poticanja gradijentom, što ga čini izuzetno efikasnim i preciznim.

Prednosti XGBoost-a uključuju:

* Izuzetno visoku preciznost i performanse u mnogim zadacima strojnog učenja.
* Mogućnost rada s različitim tipovima značajki i podacima.
* Učinkovitost u radu s velikim skupovima podatak

# 

# 3. Opis programskog rješenja

Za rješavanje ovog projektnog zadatka testirat ćemonekoliko varijatni rješavanja koja su nabrojana u drugom poglavlju.

## 3.1 Model strojnog učenja.

Usporedili smo rezultate 5 različitih modela gdje smo gledali:

* Točnost- Ova mjera opisuje postotak ukupnog broja predikcija koje su točne.
* Odziv- Ova mjera opisuje postotak pravih pozitivnih predikcija u odnosu na ukupan broj stvarnih pozitivnih primjera.
* F1 mjera - Predstavlja težinu između preciznosti i odziva, te se koristi kao mjera ukupne učinkovitosti modela

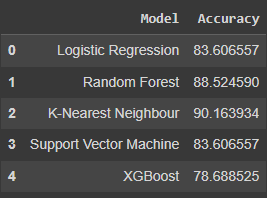
Rezultate koje smo dobili :

|  |  |
| --- | --- |
| K najbližih susjeda |  |
| LogisticRegression |  |
| RandomForestClassifier |  |
| Support Vector Machine |  |
| XGBoost |  |

*Tablica 1. Prikaz dobivenih rezultata modela*

Iz dobivenih rezultata možemo vidjeti da najbolju točnost ima Kneighbors i RandomForest (0.869),najbolji odziv ima Support Vector Machine (0.941), a F1 odziv isto Kneighbors.

Nakon dobivenih rezultata pokušali smo poboljšati modelima točnost te dobijemo nove rezultate.:



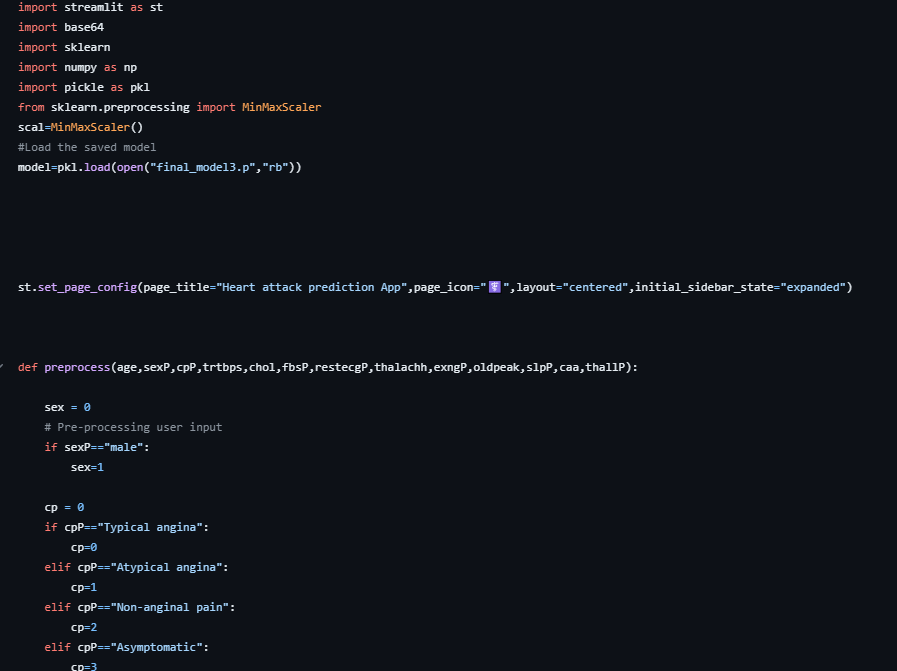
*Slika 3. Prikaz rezultata točnosti modela*

Možemo zaključiti da najbolji model Kneighbors jer ima najveći odziv (90.163) i F1 mjeru (0.886), što znači da ima može prepoznati više stvarnih pozitivnih primjera i uravnotežiti preciznost i odziv učinkovitije od ostalih modela.

## 

## 3.2 Web aplikacija

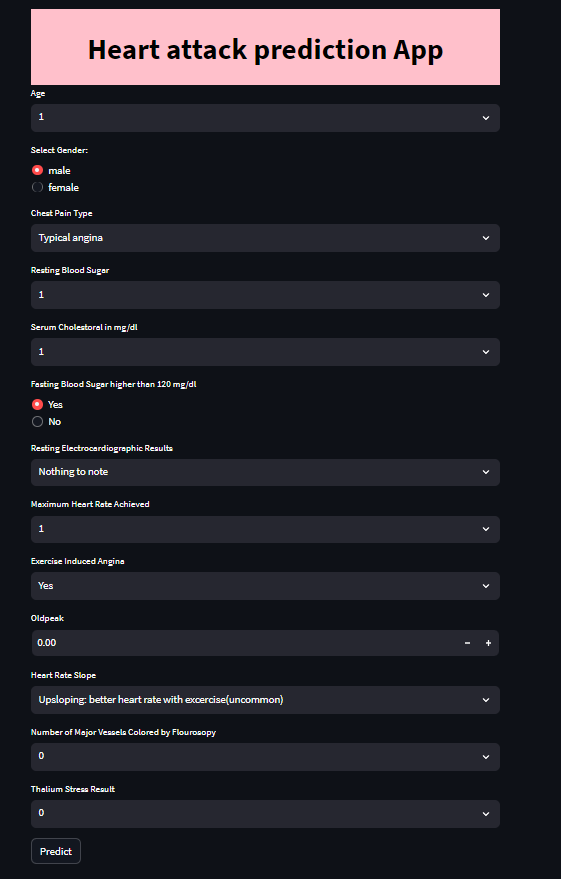
Nakon pronalaska modela s najboljim performansama, uključujemo biblioteku pickle uz koju pomoć možemo spremiti naš model.Nakon toga trebamo napisati programski kod za našu aplikaciju koristeći biblioteku streamlit i napisati Python skriptu.Streamlit služi da vrlo jednostavno napravimo web aplikaciju iz područja podataka i strojnog učenja



*Slika 4. Prikaz healthy-heart-app.py skripte*

Skripta sadrži korištene biblioteke koje nam koriste prilikom rješavanja projekta.Sastoji se od funkcije kojoj su parametri podaci koje korisnik unosi prilikom korištenja.Funkcija prevodi podatke tako da budu jednaki podacima iz skupa podataka.Nakon toga podatke skaliramo i računamo predikciju jel pacijent u opasnosti od srčanog udara. Osim funkcije sadrži i izgled stranice.

Za podizanje stranice na internet koristimo host streamlit.io te dobivamo izgled stranice. Kada korisnik ispuni obrazac klikne dugme predicit iskočiše prozor koji govori jel korisnik ugrožen od srčanog udara.



*Slika 5. Prikaz izgled web aplikacije*

# Zaključak:

U opisanom programu za predviđanje srčanog udara korišten je model strojnog učenja za predviđanje potencijalne opasnosti od srčanog udara koristeći različite značajke kao što su: Dob, spol, angina uzrokovana tjelesnom aktivnošću, broj glavnih krvnih žila itd.

Zaključak je da takvi modeli bi mogli pomoći doktorima o otkrivanju potencijalnog srčanog udara.Potrebno je imati pouzdane podatke i provjeriti ispravnost i pouzdanost modela prije nego se krene s korištenjem u praksi.

Također treba naglasiti da ovaj model nije savršen te da postoje ograničenja u korištenju strojnog učenja u ovom kontekstu.

# Literatura:

Github: <https://github.com/Snupix031/ruap1/tree/main>

WebApp: <https://ea2r8nrujumo3juiynhgp8.streamlit.app>

K najbližih susjeda: <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>

Logistička regresija: <https://en.wikipedia.org/wiki/Logistic_regression>

Support Vector Machine : <https://en.wikipedia.org/wiki/Support_vector_machine>

XGBoost: https://en.wikipedia.org/wiki/XGBoost

# Prolozi:

Podaci:

**Dob**: Dob pacijenta (numerčki).

**Spol**: Spol pacijenta (kategorijalno: 0 za žensko, 1 za muško).

**Cp**: Tip boli u prsima (kategorijalno):

Vrijednost 1: Tipična angina

Vrijednost 2: Atipična angina

Vrijednost 3: Neanginalna bol

Vrijednost 4: Asimptomatska

**Trtbps:** Mirovni krvni tlak u mm Hg

**Chol**: Razina kolesterola u mg/dL izmjerena putem BMI senzora (numerčki).

**Fbs**: Post način šećera u krvi > 120 mg/dL (1 za istina, 0 za laž)

**Rest\_ecg**: Rezultati elektrokardiografije u mirovanju

Vrijednost 0: Normalno

Vrijednost 1: Prisutnost abnormalnosti na ST-T valovima (inverzija T vala i/ili ST elevacija ili depresija veća od 0,05 mV)

Vrijednost 2: Pokazivanje vjerojatnog ili definitivnog hipertrofiranja lijevog srčanog pretklijetka prema Estesovim kriterijima

**Thalach**: Najviša postignuta brzina otkucaja srca

**Exang**: Angina uzrokovana tjelesnom aktivnošću (1 za da, 0 za ne).

**Oldpeak**: Depresija ST segmenta izazvana tjelesnom aktivnošću u odnosu na mirovanje.

**Slp:** Nagib ST segmenta tijekom vršne tjelesne aktivnosti

**Ca**: Broj glavnih krvnih žila (numerčki: 0-3).

**Output:** 0 = Manja vjerojatnost srčanog udara, 1 = Veća vjerojatnost srčanog udara