Državni univerzitet u Novom Pazaru
Departman za Tehničko-tehnološke nauke
Softversko inženjerstvo

Seminarski rad

Predmet: Mašinsko učenje

Tema:
Klasifikacija podataka o dijabetesu

Mentor:
Doc. dr Ulfeta Marovac

Student:
Ibrahimović Tarik
Hamzić Ensar

Novi Pazar, 2024.godina

Contents

Uvod i opis problema

Problem koji ćemo rešiti primenom različitih klasifikacionih metoda jeste klasifikacija podataka iz istraživanja za dijabetes. Klasifikacija će se bazirati na tome da li ispitanik ima dijabetes ili ne.

Svaki red podataka sadrzi 22 unosa:

- HighBP
- HighChol
- CholCheck
- BMI
- Smoker
- Stroke
- HeartDiseaseorAttack
- PhysActivity
- Fruits
- Veggies
- HvyAlcoholConsump
- AnyHealthcare
- NoDocbcCost
- GenHlth
- MentHlth
- PhysHlth
- DiffWalk
- Sex

- Age
- Education
- Income
- Diabetes_binary

Opis skupa podataka

Ovaj dataset sadrži informacije prikupljene iz ankete o dijabetesu, fokusirajući se na različite zdravstvene indikatore, faktore rizika i demografske podatke koji mogu biti povezani sa dijabetesom. Podaci omogućavaju analizu i identifikaciju ključnih faktora koji doprinose razvoju dijabetesa, kao i drugih hroničnih bolesti.

Opis atributa

- HighBP: Da li osoba ima visok krvni pritisak; 0 = nema visok krvni pritisak, 1 = ima visok krvni pritisak.
- HighChol: Da li osoba ima visok holesterol; 0 = nema visok holesterol, 1 = ima visok holesterol.
- CholCheck: Da li je osoba proverila holesterol u poslednjih 5 godina; 0 = nije proverila, 1 = jeste proverila.
- BMI: Indeks telesne mase (Body Mass Index).
- Smoker: Da li je osoba pušila bar 100 cigareta u životu; 0 = nije pušila, 1 = jeste pušila.

- Stroke: Da li je osoba ikada imala moždani udar; 0 = nije imala, 1 = jeste imala.
- HeartDiseaseorAttack: Da li je osoba imala koronarnu srčanu bolest ili infarkt miokarda; 0 = nije imala, 1 = jeste imala.
- PhysActivity: Da li je osoba bila fizički aktivna u poslednjih 30 dana, ne uključujući posao; 0 = nije bila aktivna, 1 = jeste bila aktivna.
- Fruits: Da li osoba konzumira voće jednom ili više puta dnevno; 0 = ne konzumira, 1 = konzumira.
- Veggies: Da li osoba konzumira povrće jednom ili više puta dnevno; 0 = ne konzumira, 1 = konzumira.
- HvyAlcoholConsump: Da li osoba prekomerno pije alkohol (muškarci više od 14 pića nedeljno, žene više od 7 pića nedeljno); 0 = ne pije prekomerno, 1 = pije prekomerno.
- AnyHealthcare: Da li osoba ima neku vrstu zdravstvenog osiguranja ili pokrića; 0 = nema, 1 = ima.
- NoDocbcCost: Da li je osoba imala potrebu da vidi doktora u poslednjih 12 meseci, ali nije mogla zbog troškova; 0 = nije imala taj problem, 1 = jeste imala taj problem.
- GenHlth: Opšta ocena zdravlja osobe; skala od 1 do 5: 1 = odlično, 2 = vrlo dobro, 3 = dobro, 4 = zadovoljavajuće, 5 = loše.
- MentHlth: Broj dana u poslednjih 30 dana kada je mentalno zdravlje osobe bilo loše; skala od 0 do 30.
- PhysHlth: Broj dana u poslednjih 30 dana kada je fizičko zdravlje osobe bilo loše; skala od 0 do 30.
- DiffWalk: Da li osoba ima ozbiljnih problema sa hodanjem ili penjanjem uz stepenice; 0 = nema problema, 1 = ima problema.

- Sex: Pol osobe; 0 = žensko, 1 = muško.
- Age: Starosna kategorija osobe; skala od 1 do 13: 1 = 18-24, 8 = 55-59, 13 = 80 ili stariji.
- Education: Nivo obrazovanja osobe; skala od 1 do 6: 1 = nikada nije išao u školu ili samo vrtić, 2 = 1. do 8. razreda.
- Income: Skala prihoda; skala od 1 do 8: 1 = manje od $10,000, 5 = manje od $35,000, 8 = $75,000 ili više.
- Diabetes_binary: Da li osoba ima dijabetes; 0 = nema dijabetes, 1 = ima dijabetes.

Ovi podaci obuhvataju širok spektar informacija, od demografskih (kao što su pol i starost), preko zdravstvenih ponašanja (kao što su pušenje i fizička aktivnost), do medicinskih indikatora (kao što su visok krvni pritisak, visok holesterol i indeks telesne mase). Analizom ovih podataka, istraživači mogu bolje razumeti kako različiti faktori utiču na prisustvo dijabetesa i razviti strategije za prevenciju i upravljanje ovim stanjem.

Opis metoda koje se koristi i metrike za evaluaciju

Naš odabir metoda za rešavanje ovog problema se svodi na sledećih pet:

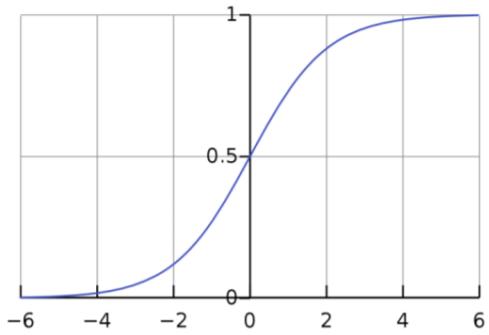
- **Logistička regresija**

  Podvrsta logističke regresije koju ćemo u ovom primeru iskoristiti jeste binarna logistička regresija, kod koje zavisna promenljiva uzima vrednost iz binarnog skupa. Cilj logističke regresije je modelovanje verovatnoće da neka instanca iz skupa podataka pripada određenoj kategoriji.

  - $$p(y|x) = \begin{cases} \mu, & y = 1 \\ 1 - \mu, & y = 0 \end{cases}$$

  S' obzirom da logistička regresija radi sa verovatnoćama, granične vrednosti su zadate intervalom [0,1], a ovaj model se dobio transformacijom linearnog modela sa intervala [-∞,∞] pomoću neprekidne diferencijabilne funkcije, koja se još naziva sigmoidna funkcija.

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



- **<u>Naivni Bajesov algoritam</u>**

  Zasniva se na modelovanju raspodele ciljne promenljive y pri datim vrednostima promenljive x, korišćenjem Bajesove formule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

  Karakteristika ovog algoritma je ta da se na samom početku uvodi pretpostavka da su atributi nezavisni, a zatim se pretpostavlja da su svi atributi podjednako važni. Iako ove pretpostavke u realnosti nikada nisu tačne, ovaj algoritam se pokazao kao dosta pouzdan za problem klasifikacije.

- **<u>KNN</u>**

  Osnovna pretpostavka ovog algoritma je postojanje rastojanja nad prostorom atributa (eng. feature space).

  Algoritam k najbližih suseda klasifikuje nepoznatu instancu tako što pronalazi k instanci iz skupa za obučavanje koje su joj najbliže u smislu neke izabrane metrike i pridružuje joj klasu koja se najčešće javlja među tih k instanci. Funkcija rastojanja se bira nezavisno od podataka.

Dobra matrica rastojanja bi bila ona za koju su tačke iz iste klase blizu, dok su sve tačke iz različitih klasa međusobno daleko.

- **Decision Tree (stablo odlučivanja)**

  Ovo je algoritam za klasifikaciju i regresiju koji koristi drvo odluka kao model predikcije. Drvo odluka predstavlja niz pravila baziranih na atributima podataka koji vode do predikcije ciljne promenljive.
  Svaki čvor u stablu predstavlja atribut na kojem se vrši podela, svaki krak predstavlja ishod te podele, a svaki list predstavlja finalnu klasifikaciju ili regresiju.

Prednost odlučujućih stabala je njihova jednostavnost i interpretabilnost, ali su sklona prekomernom prilagođavanju (overfitting) posebno kada su stabla duboka.

- **Random Forest (slučajna šuma)**

  Ovo je ansambl algoritam koji koristi više odlučujućih stabala za klasifikaciju ili regresiju. Svako stablo u šumi trenira se na nasumičnom uzorku podataka i koristi nasumičan podskup atributa za svaku podelu.

Predikcija Random Forest modela se dobija agregacijom predikcija svih stabala u šumi, obično putem glasanja (za klasifikaciju) ili proseka (za regresiju).
Prednost Random Forest algoritma je njegova otpornost na prekomerno prilagođavanje i bolja generalizacija u odnosu na pojedinačna odlučujuća stabla. Takođe, može da se nosi sa velikim brojem atributa i ima dobru performansu na visokodimenzionalnim podacima.

- **GAN (Generative Adversarial Networks)**

  Ovo su vrsta generativnih modela mašinskog učenja koji se sastoje od dva konkurentna neuronska modela - generatora i diskriminatora. Ova metoda se

razlikuje od KNN i Decision Tree algoritama koje ste opisali, jer se ne koristi za klasifikaciju ili regresiju, već za generisanje novih, veštačkih podataka koji liče na podatke iz originalnog skupa za obučavanje.

Generator nastoji da proizvede lažne uzorke koji izgledaju realistično, poput slika, audio snimaka ili teksta. Diskriminator pokušava da razlikuje generisane lažne uzorke od pravih uzoraka iz obučavajućeg skupa. Tokom obučavanja, generator i diskriminator se takmičarima i iterativno poboljšavaju - generator nastoji da "prevari" diskriminatora produciranjem sve realističnijih lažnih uzoraka, dok diskriminator postaje bolji u detekciji lažnih uzoraka od generatora.

GAN modeli se široko koriste u raznim primjenama kao što su generisanje slika visokog kvaliteta, povećanje rezolucije slika, prevođenje stilova između domena (npr. pretvaranje skica u fotografije), sinteza govora i muzike, itd. Za razliku od KNN i odlučujućih stabala koji se bave predikcijom na osnovu obučavajućih podataka, GAN modeli kreativno generišu nove podatke koristeći obučavajuće podatke kao referencu.

- **<u>Conditional Generative Adversarial Networks (CGAN)</u>**

  Ovo je proširenje generativnih modela poznatih kao Generative Adversarial Networks (GAN), koji uključuju dodatne informacije kao uslov (condition) za generisanje novih podataka. Dok klasični GAN sadrži generator i diskriminator koji funkcionišu nezavisno od specifičnih oznaka ili ulaznih podataka, CGAN dodaje dodatnu komponentu - uslovne informacije koje mogu biti u obliku oznaka, klasa ili drugih atributa. Kod CGAN-a, generator prima ne samo nasumični šum kao ulaz, već i dodatnu informaciju o željenoj klasi ili atributima podataka koje treba generisati. Na primer, ako generišemo slike rukom pisanih brojeva, generator može primiti oznaku broja koji treba da generiše (npr. broj "5"). Diskriminator takođe dobija ovu dodatnu informaciju i

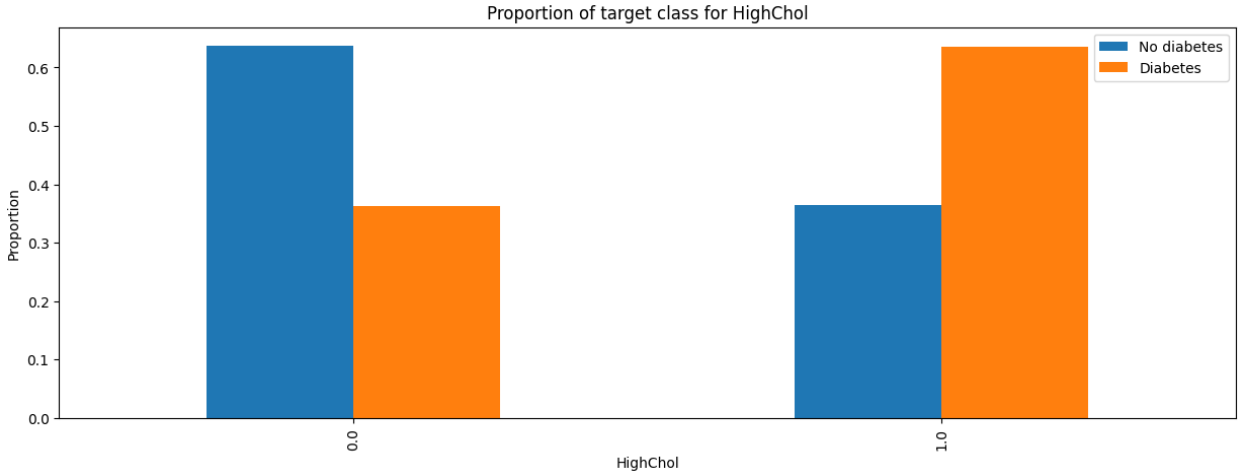
koristi je za razlikovanje stvarnih podataka sa oznakama od generisanih podataka sa istim oznakama.

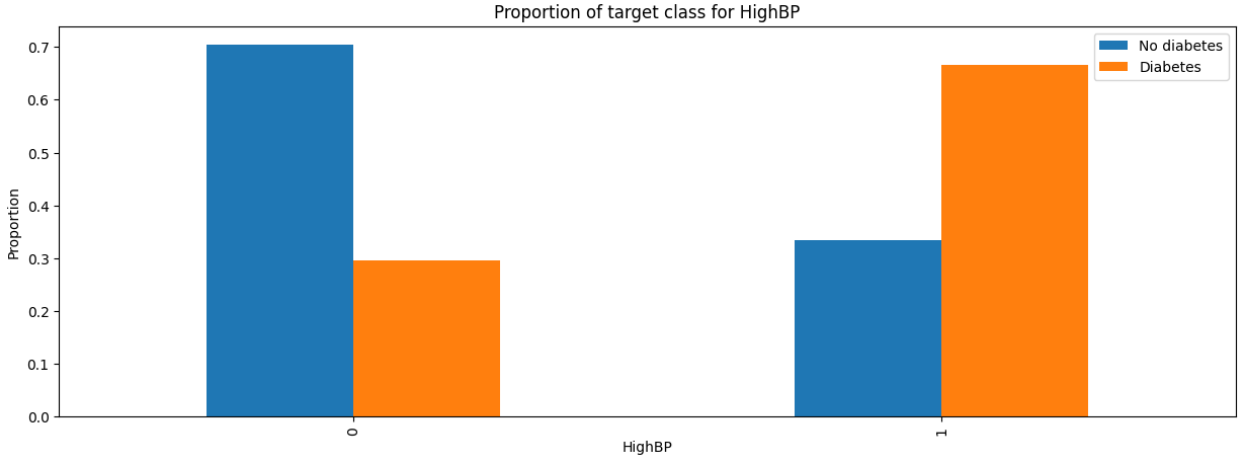
Priprema podataka

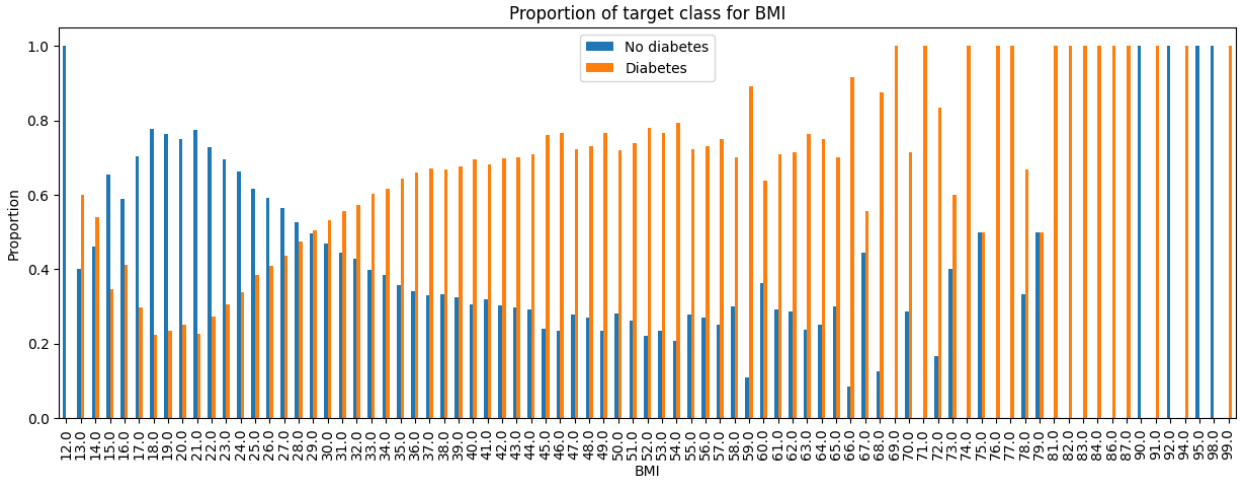
Priprema podataka je ključni korak u analizi i modelovanju, posebno kada se radi o klasifikaciji dijabetesa. U ovom procesu, podaci su učitani, očišćeni i pripremljeni za dalje korake analize. Prvo, balansirane su klase kako bi se osiguralo da je broj instanci sa i bez dijabetesa približno jednak, što pomaže u izbegavanju pristrasnosti modela. Svi podaci su od samog početka bili kompletni, nije imalo nikakvih nepoznatih vrednosti niti vrednosti koje nedostaju, tako da nije bilo potrebno nikakvo dodatno preprocesiranje. Dakle, nakon spomenutog, podaci su podeljeni na trening i test skupove kako bi se omogućila validacija modela i osigurala njihova generalizacija na nove podatke. Ovim koracima pripremljen je čvrst temelj za primenu različitih klasifikacionih algoritama.

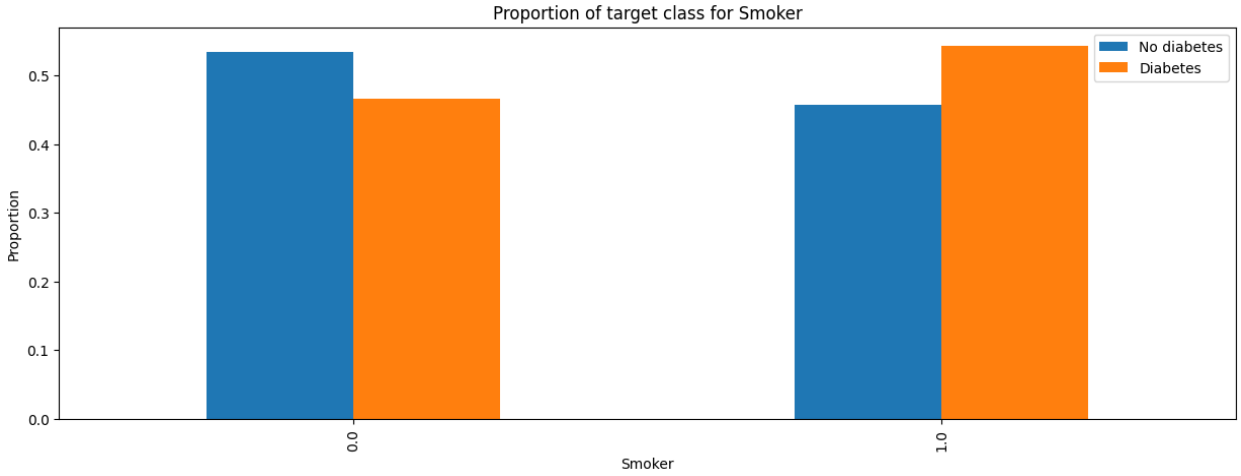
```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, recall_score, precision_score, f1_score, roc_curve, auc
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

df = pd.read_csv('dataset/diabetes.csv')

class_0 = df[df.Diabetes_binary == 0]
class_1 = df[df.Diabetes_binary == 1]

class_0_downsampled = class_0.sample(len(class_1))

balanced_df = pd.concat([class_0_downsampled, class_1], axis=0)

X = balanced_df.drop("Diabetes_binary", axis=1)
y = balanced_df["Diabetes_binary"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

balanced_df.head()
```

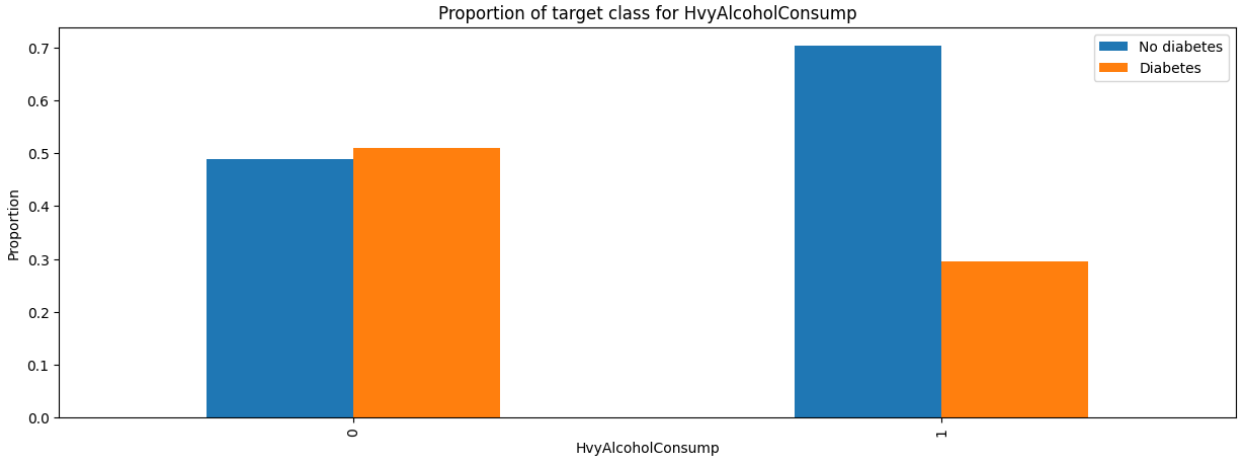
Odnosi između podataka

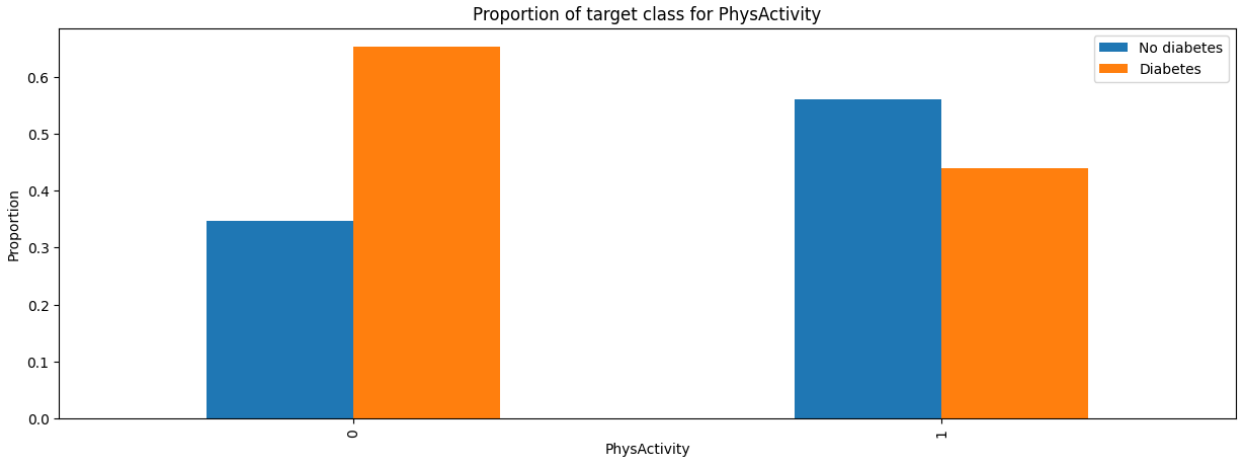


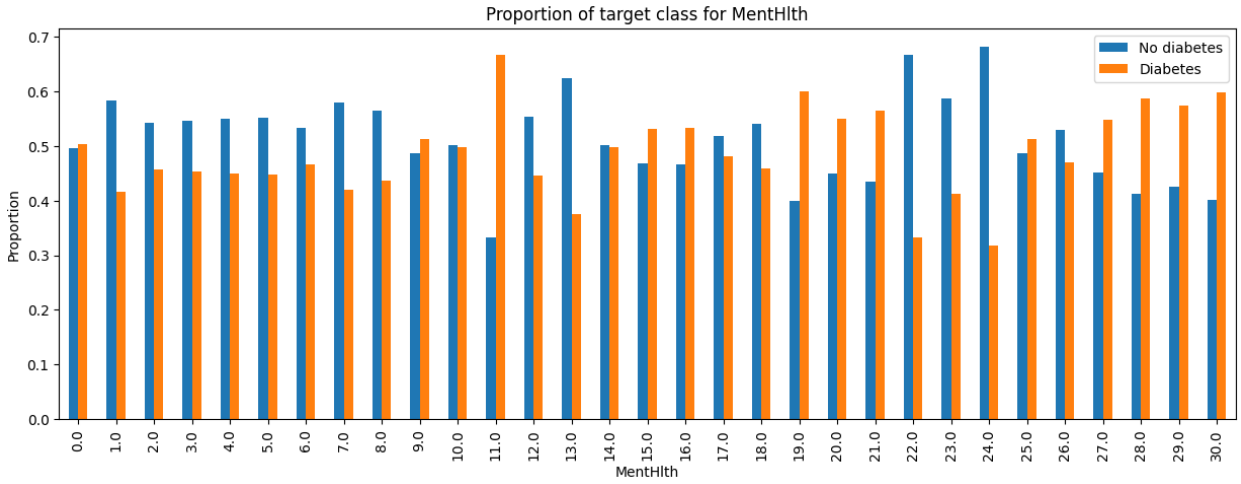





Proportion of target class for Smoker
No diabetes
Diabetes
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.0
1.0
Smoker

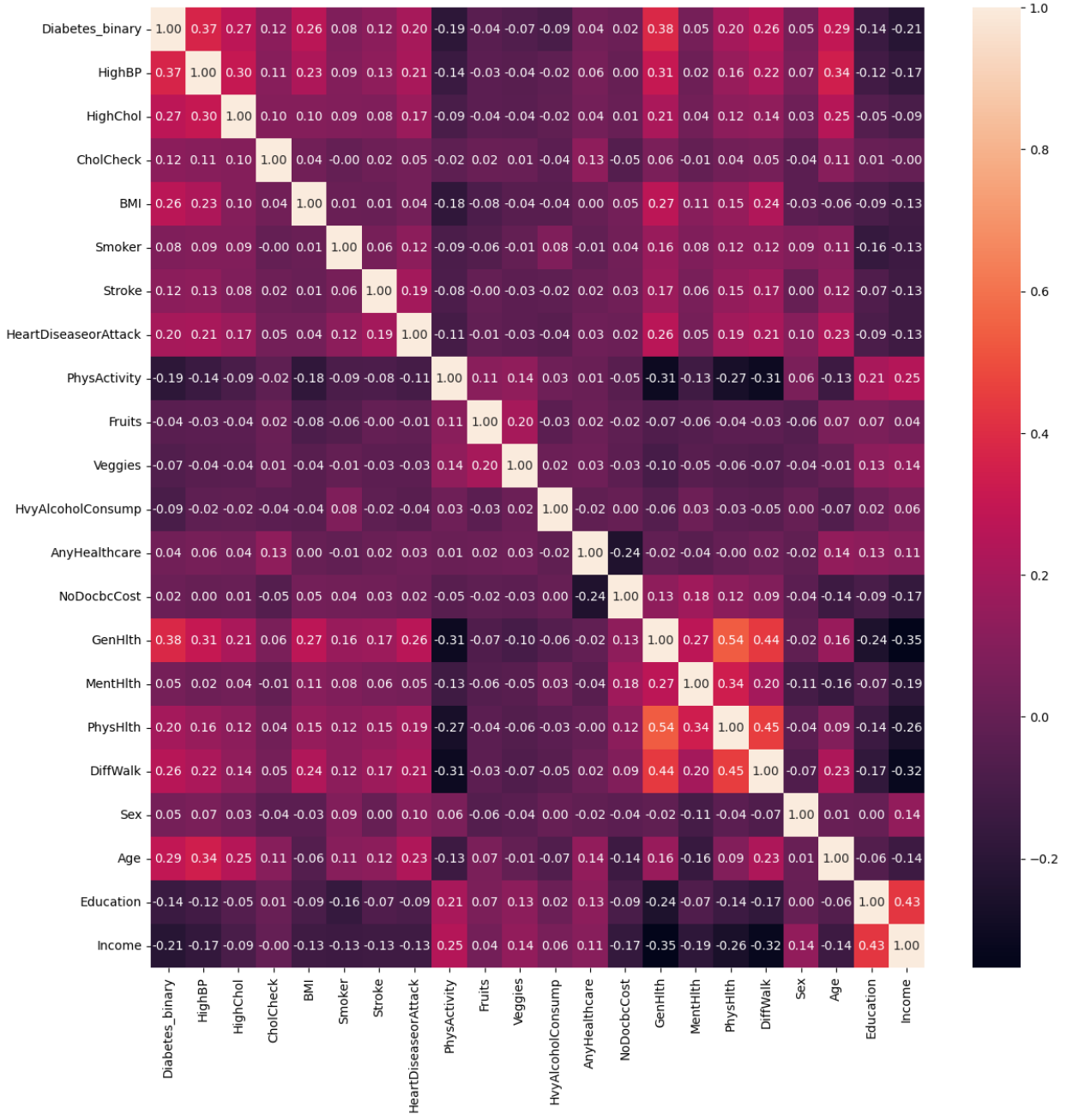

Proportion of target class for HvyAlcoholConsump
No diabetes
Diabetes
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
HvyAlcoholConsump


Proportion of target class for PhysActivity
No diabetes
Diabetes
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
PhysActivity


Proportion of target class for MentHlth
No diabetes
Diabetes
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
MentHlth

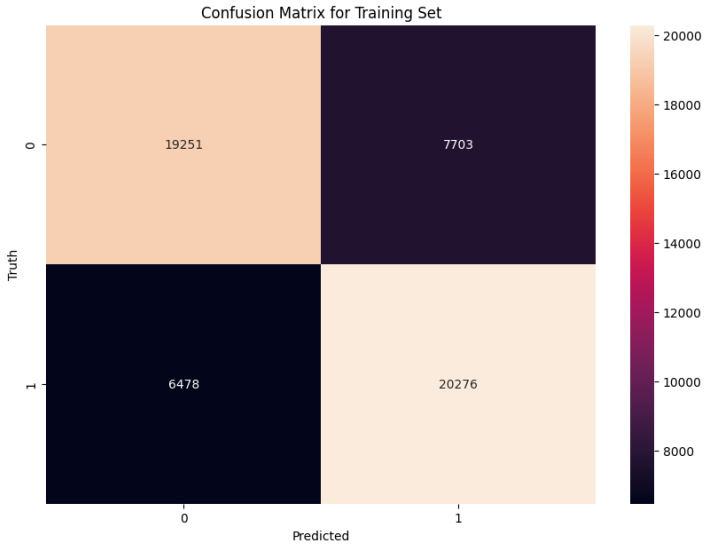
Rezultati

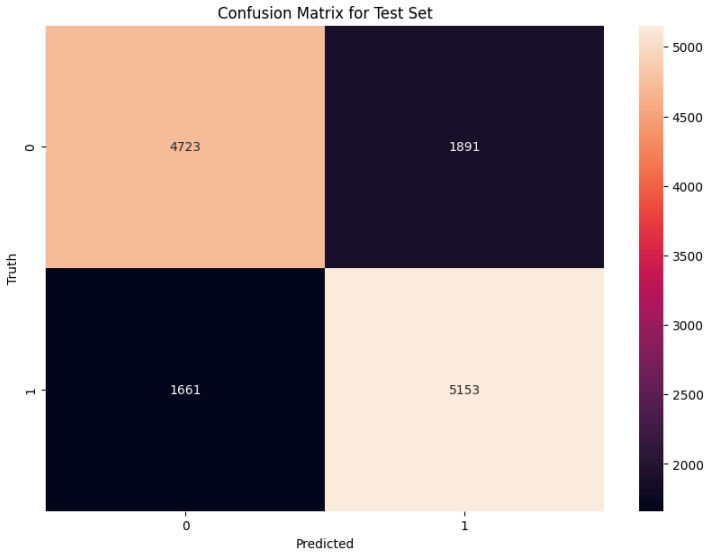
Heat-map dataset-a

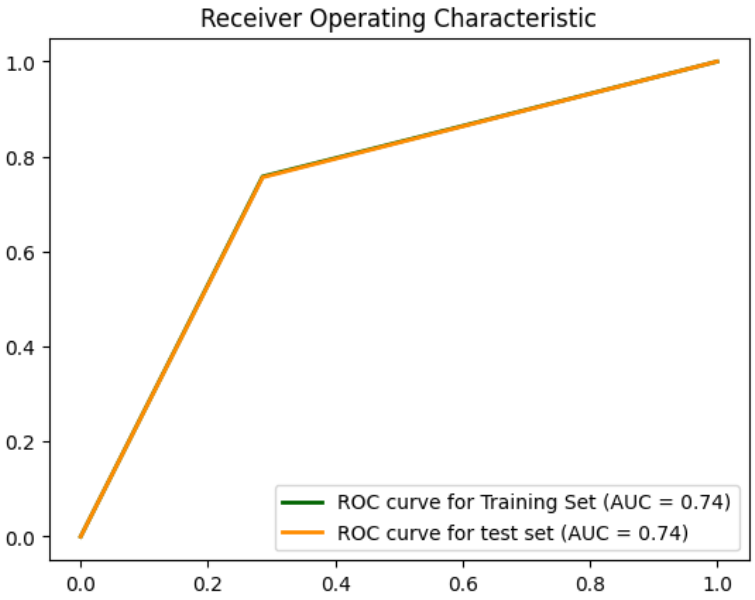


Naive-Bayes

	Accuracy	Precision	Recall	F1 Score
Training set	0.74	0.72	0.76	0.74
Test set	0.74	0.73	0.76	0.74

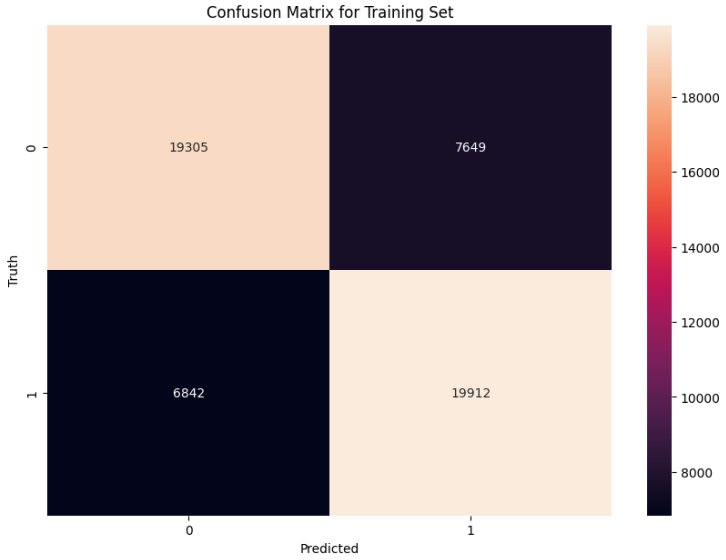


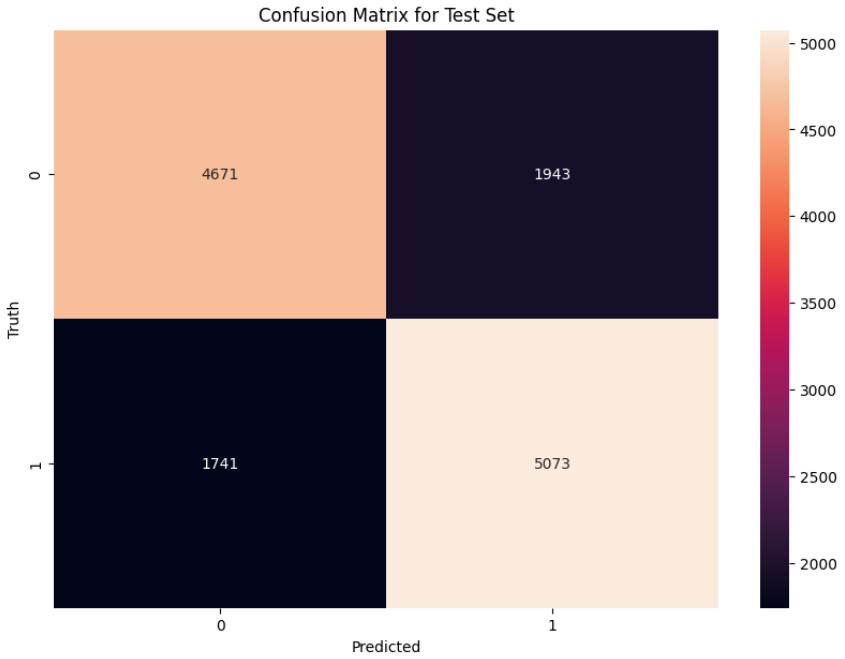


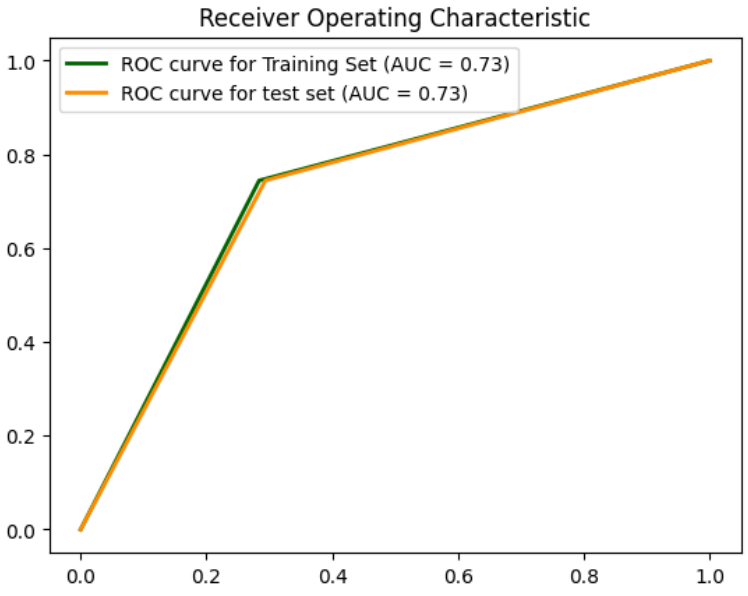


Decision Tree Classification

	Accuracy	Precision	Recall	F1 Score
Training set	0.73	0.72	0.74	0.73
Test set	0.73	0.72	0.74	0.73

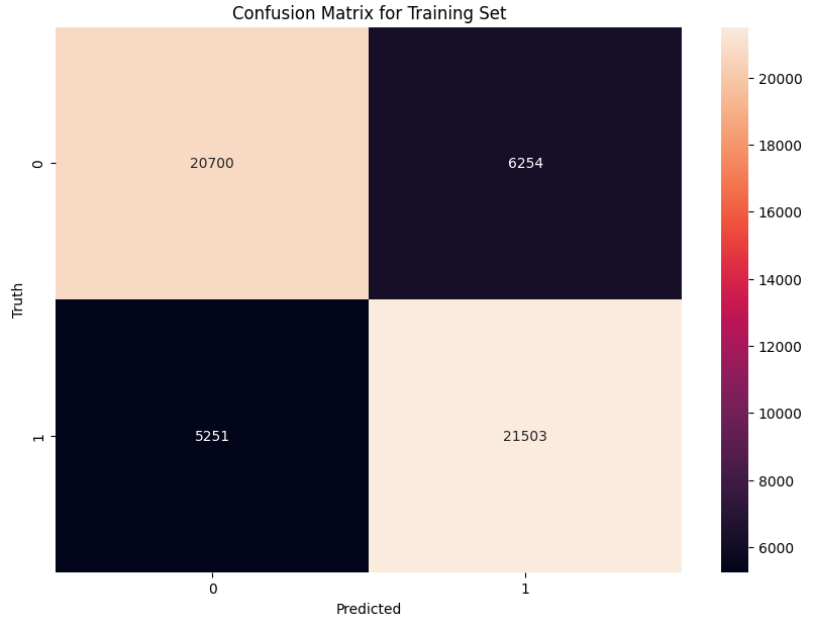


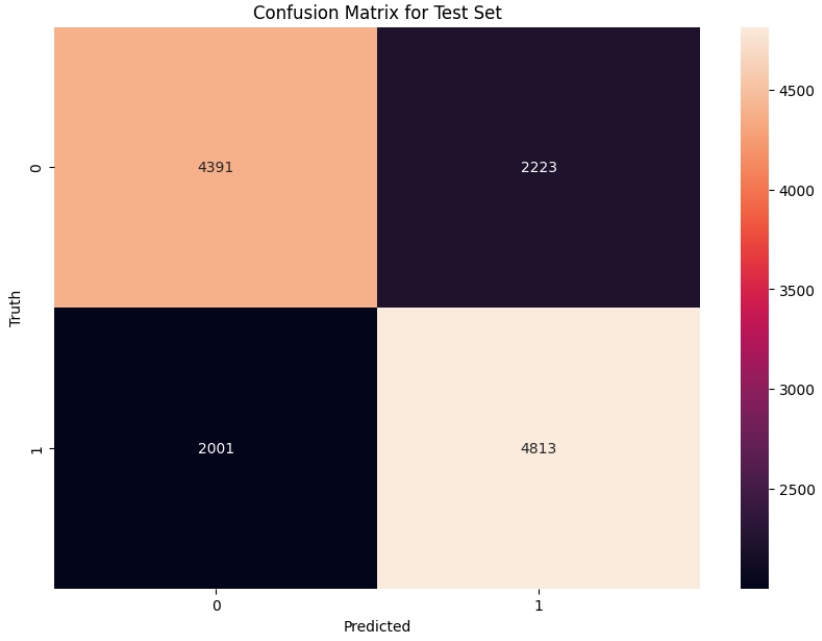


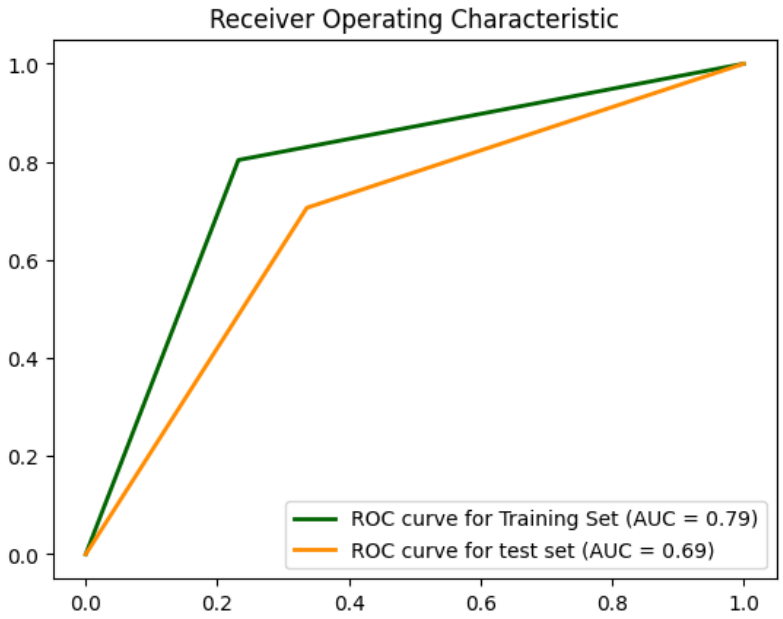


KNN

	Accuracy	Precision	Recall	F1 Score
Training set	0.79	0.77	0.8	0.79
Test set	0.69	0.68	0.71	0.7

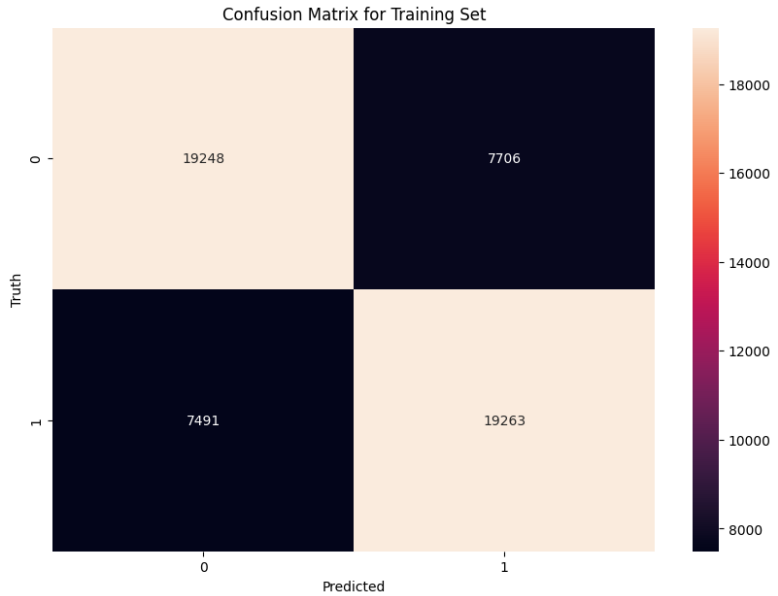

Confusion Matrix for Training Set
20700
6254
5251
21503
Truth
Predicted
0
1
20000
18000
16000
14000
12000
10000
8000
6000

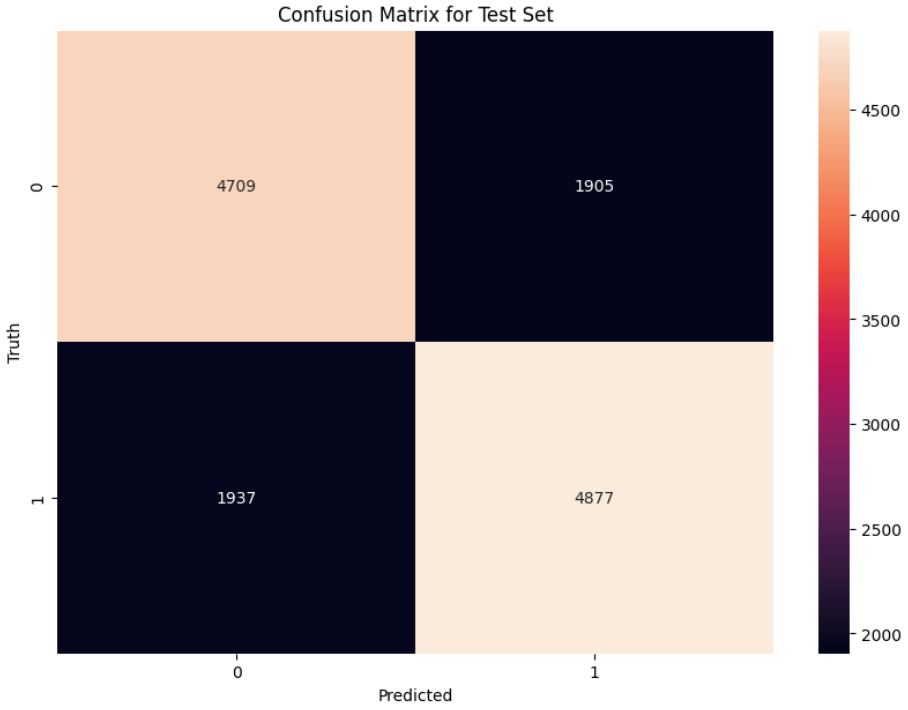

Confusion Matrix for Test Set
4391
2223
2001
4813
Truth
Predicted
0
1
4500
4000
3500
3000
2500

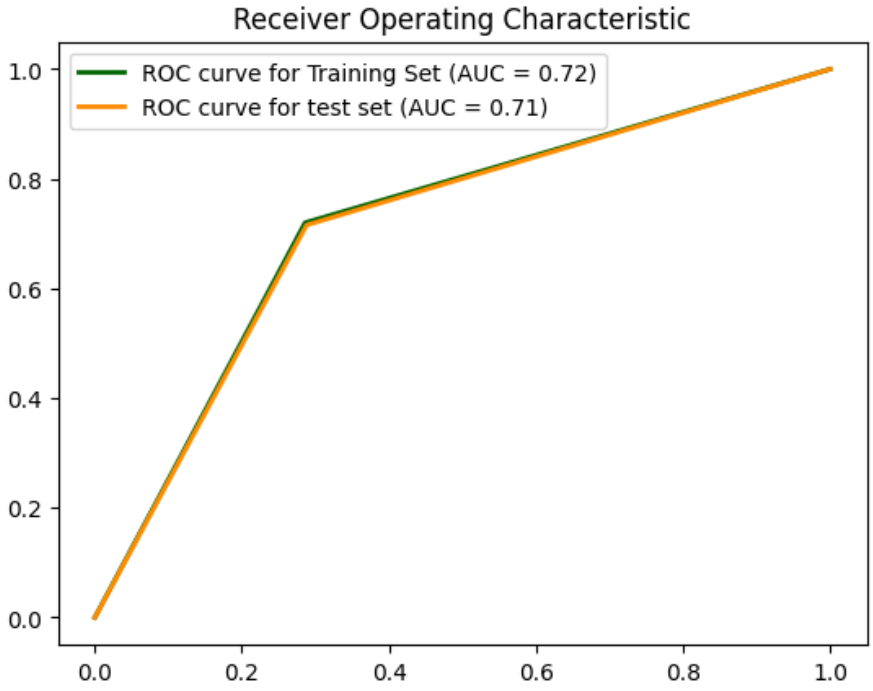


Logistička regresija

	Accuracy	Precision	Recall	F1 Score
Training set	0.72	0.71	0.72	0.72
Test set	0.71	0.72	0.72	0.72

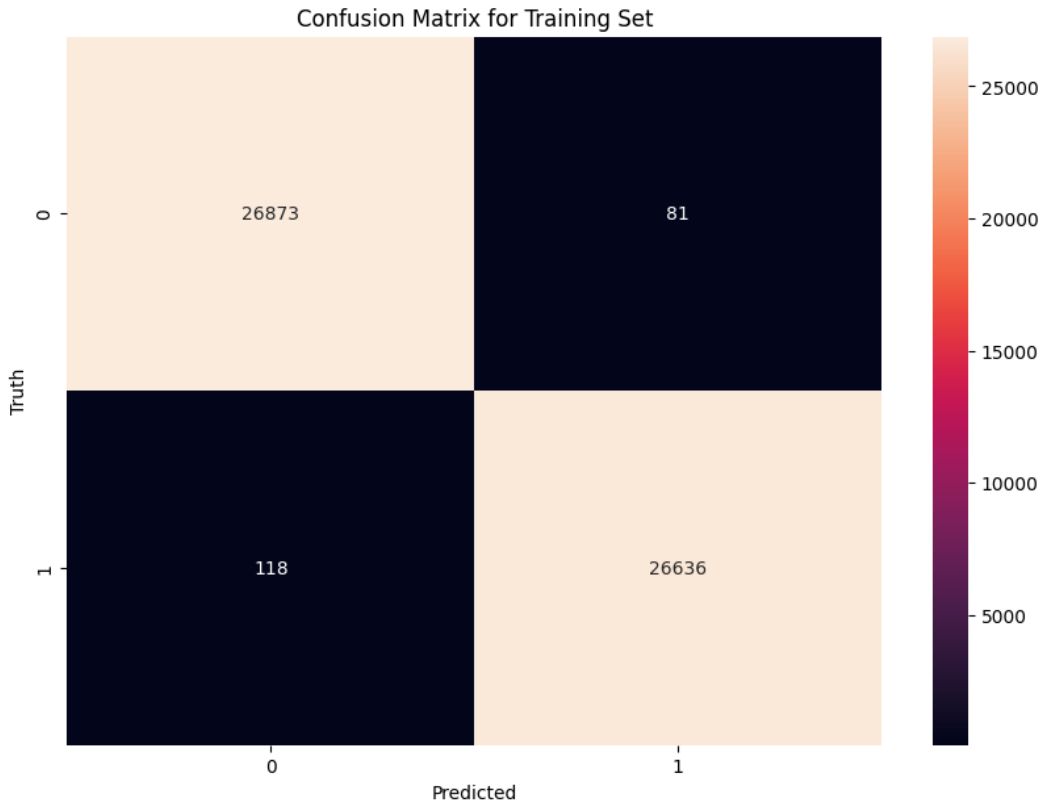


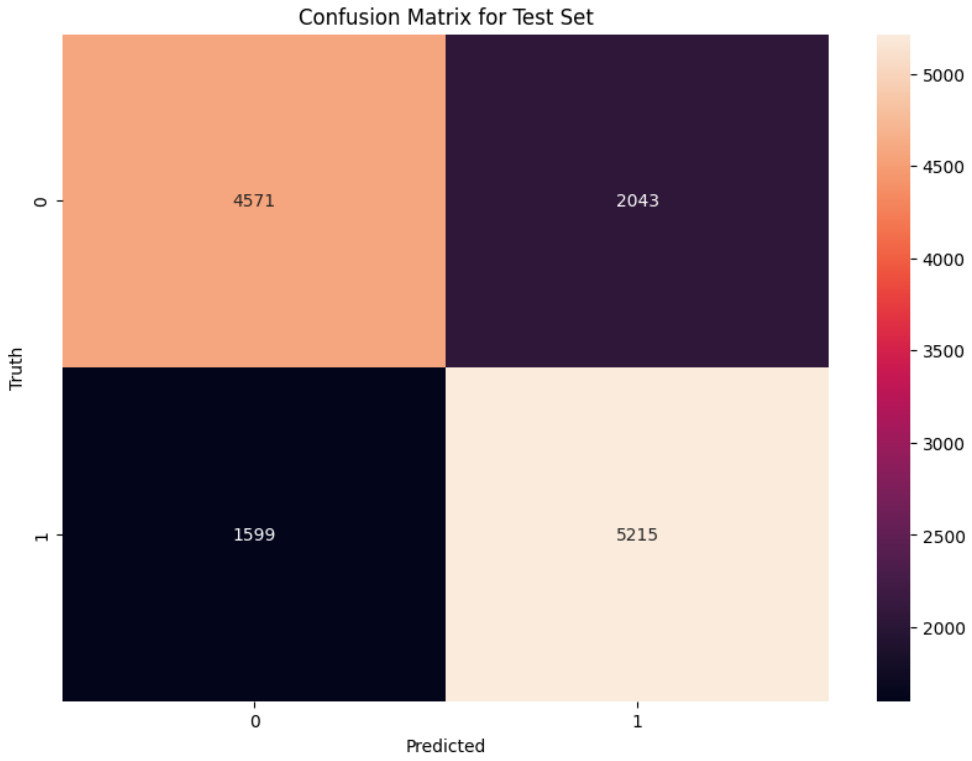

Confusion Matrix for Test Set
4709
1905
1937
4877
0
1
Truth
Predicted
4500
4000
3500
3000
2500
2000

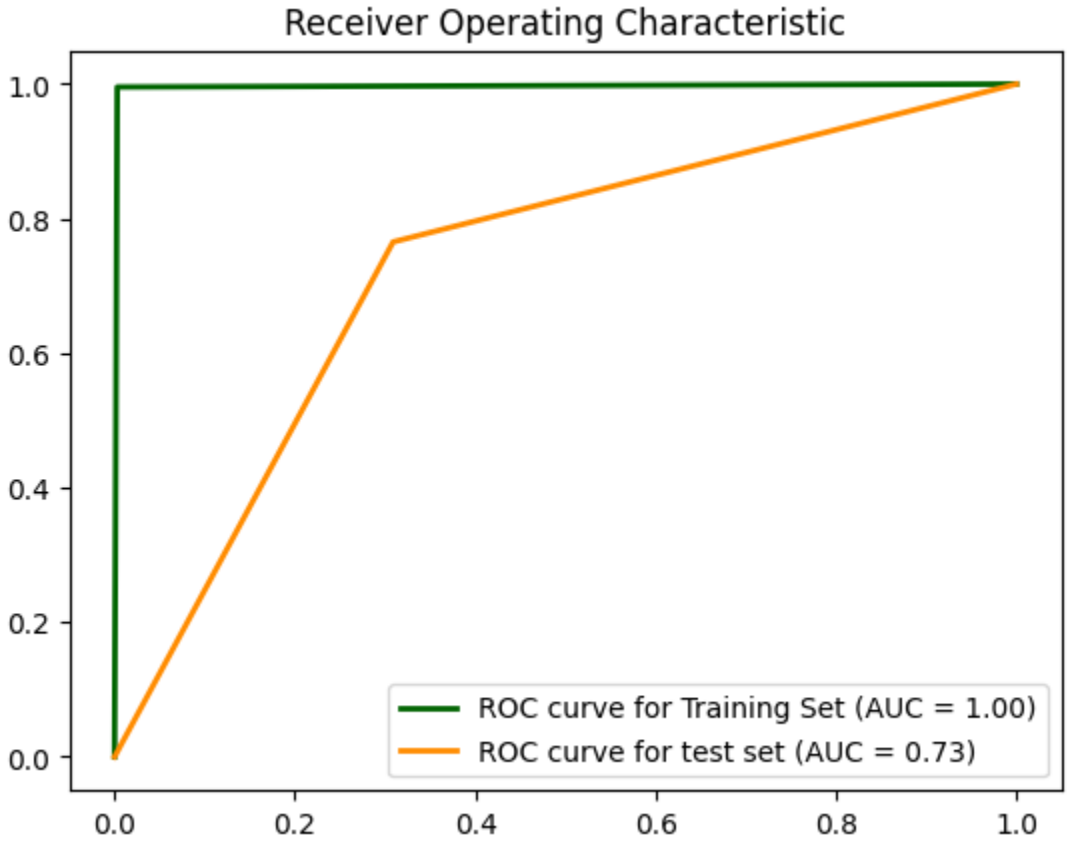

Receiver Operating Characteristic
ROC curve for Training Set (AUC = 0.72)
ROC curve for test set (AUC = 0.71)
0.0
0.2
0.4
0.6
0.8
1.0

Random Forest

	Accuracy	Precision	Recall	F1 Score
Training set	0.9	0.9	0.9	0.9
Test set	0.72	0.71	0.72	0.72




Confusion Matrix for Test Set
4571
2043
1599
5215
Truth
Predicted
0
1
5000
4500
4000
3500
3000
2500
2000


Receiver Operating Characteristic
ROC curve for Training Set (AUC = 1.00)
ROC curve for test set (AUC = 0.73)
0.0
0.2
0.4
0.6
0.8
1.0

GAN

GENERISANI PODACI:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	1	1.0	1	24.0	1.0	0.0	0.0	1	1	0	1	1	0.0	3.0	0.0	12.0	0.0	1	12	3.0	7.0
1	1.0	1	1.0	1	29.0	1.0	0.0	1.0	1	0	1	1	1	0.0	4.0	0.0	2.0	0.0	1	11	4.0	11.0
2	1.0	1	1.0	1	20.0	1.0	0.0	0.0	0	0	0	1	1	0.0	3.0	4.0	0.0	0.0	0	6	6.0	2.0
3	0.0	0	1.0	1	23.0	1.0	0.0	0.0	0	0	1	1	1	0.0	2.0	0.0	0.0	0.0	1	4	5.0	7.0
4	0.0	0	0.0	0	32.0	0.0	0.0	0.0	1	1	1	1	1	0.0	2.0	0.0	0.0	0.0	1	9	6.0	4.0
5	1.0	1	1.0	1	30.0	1.0	0.0	0.0	1	0	0	1	1	0.0	5.0	2.0	30.0	0.0	0	10	4.0	1.0
6	1.0	1	1.0	1	30.0	0.0	0.0	1.0	0	0	1	1	1	0.0	4.0	10.0	30.0	0.0	0	11	3.0	6.0
7	0.0	0	0.0	1	32.0	0.0	0.0	0.0	1	0	1	1	1	0.0	3.0	5.0	0.0	0.0	1	6	6.0	5.0
8	1.0	0	1.0	1	30.0	0.0	0.0	0.0	0	0	1	1	1	0.0	4.0	3.0	30.0	0.0	0	11	5.0	6.0
9	0.0	0	0.0	1	20.0	0.0	0.0	0.0	1	0	1	1	1	0.0	2.0	0.0	0.0	0.0	1	3	6.0	11.0
10	0.0	0	0.0	1	24.0	1.0	0.0	0.0	1	0	0	1	1	1.0	2.0	12.0	0.0	0.0	1	5	4.0	5.0
11	0.0	1	0.0	1	21.0	0.0	0.0	0.0	1	0	0	1	1	0.0	5.0	30.0	5.0	0.0	0	5	6.0	2.0
12	0.0	0	0.0	0	29.0	0.0	0.0	0.0	1	1	0	1	1	0.0	2.0	0.0	3.0	0.0	0	3	5.0	11.0
13	1.0	1	1.0	1	34.0	0.0	0.0	0.0	0	0	0	1	1	0.0	3.0	0.0	0.0	0.0	1	12	6.0	7.0
14	0.0	1	1.0	1	36.0	1.0	0.0	0.0	0	0	1	1	1	0.0	2.0	0.0	30.0	0.0	1	11	6.0	6.0
15	0.0	0	0.0	1	33.0	0.0	0.0	0.0	1	1	0	1	1	0.0	4.0	30.0	0.0	0.0	0	7	5.0	6.0
16	0.0	1	1.0	1	24.0	0.0	0.0	0.0	0	0	0	1	1	0.0	3.0	3.0	3.0	0.0	0	12	6.0	2.0
17	1.0	1	1.0	1	28.0	1.0	0.0	0.0	0	1	1	1	1	0.0	3.0	12.0	0.0	0.0	0	10	5.0	6.0
18	1.0	1	1.0	1	31.0	0.0	0.0	0.0	1	1	0	1	1	0.0	2.0	0.0	0.0	0.0	1	11	6.0	6.0
19	1.0	1	1.0	1	28.0	0.0	0.0	0.0	1	0	1	1	1	0.0	4.0	9.0	18.0	1.0	0	5	5.0	7.0
20	1.0	1	1.0	1	29.0	0.0	0.0	1.0	1	0	1	1	1	0.0	4.0	0.0	19.0	0.0	0	13	5.0	3.0

PRECIZNOST KLASIFIKACIJA NA GENERISANE PODATKE I NA PRAVI DATA SET:

Logistic regression:

	Accuracy	Precision	Recall	F1 Score
Real data	0.71	0.72	0.72	0.72
Synthetic data	0.7	0.86	0.5	0.63

Naive bayes:

	Accuracy	Precision	Recall	F1 Score
Real data	0.74	0.73	0.76	0.74
Synthetic data	0.58	0.91	0.15	0.26

Random forest:

	Accuracy	Precision	Recall	F1 Score
Real data	0.72	0.72	0.72	0.72
Synthetic data	0.74	0.8	0.61	0.7

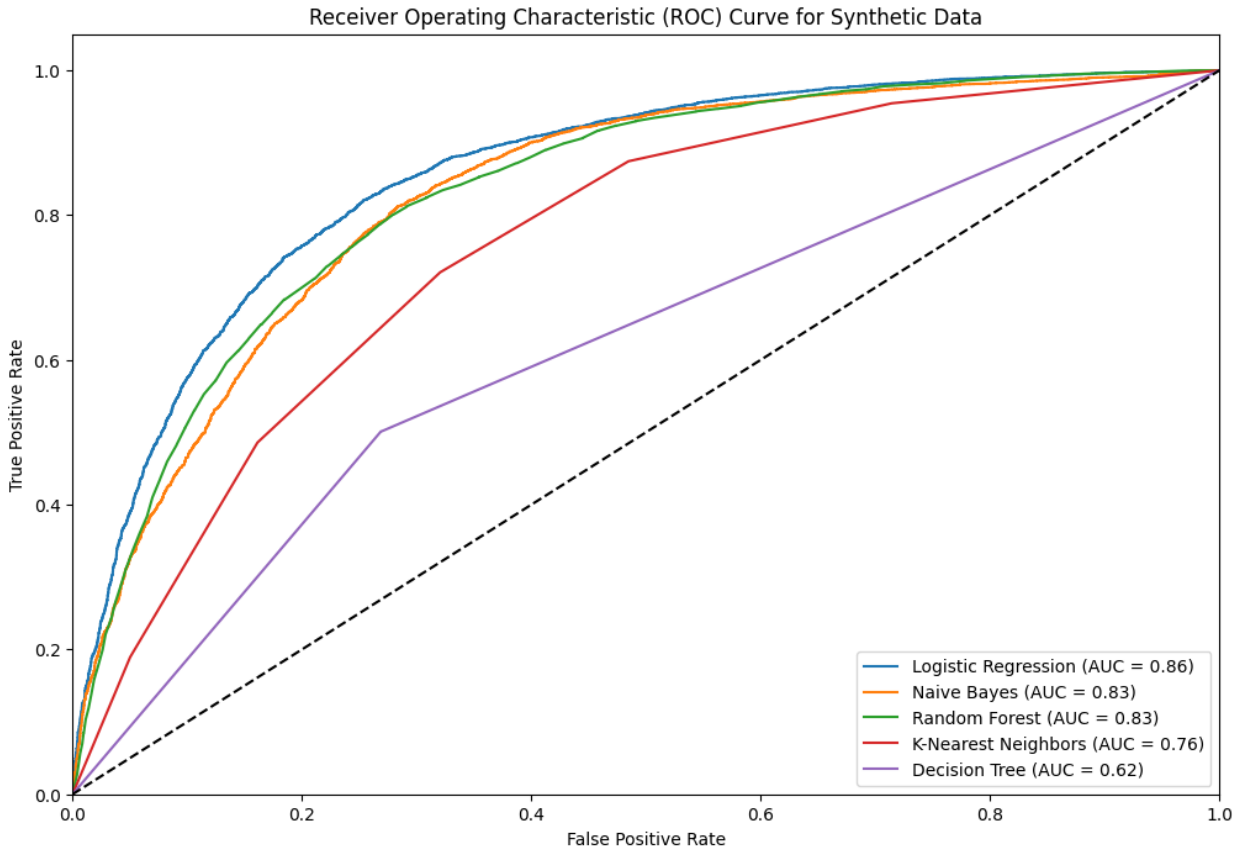
KNN:

	Accuracy	Precision	Recall	F1 Score
Real data	0.69	0.68	0.71	0.7
Synthetic data	0.7	0.69	0.72	0.7

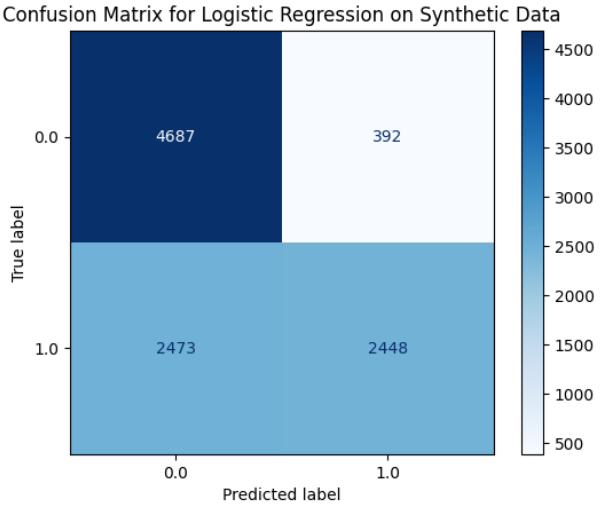
Decision tree:

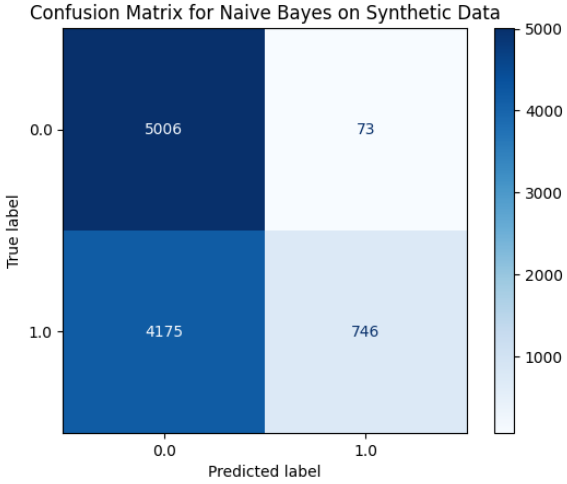
	Accuracy	Precision	Recall	F1 Score
Real data	0.73	0.72	0.74	0.73
Synthetic data	0.62	0.64	0.5	0.56

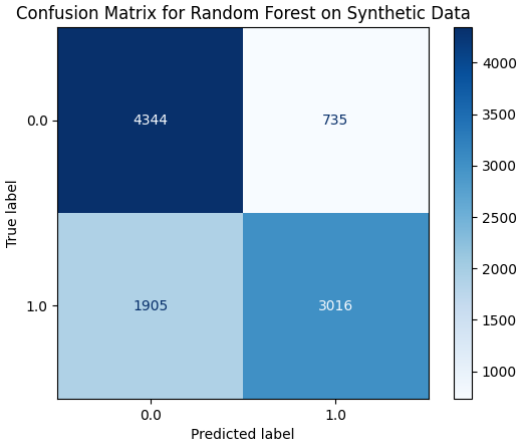
ROC-Kriva

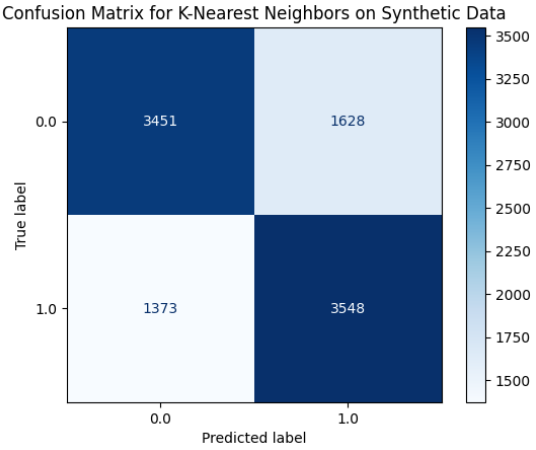


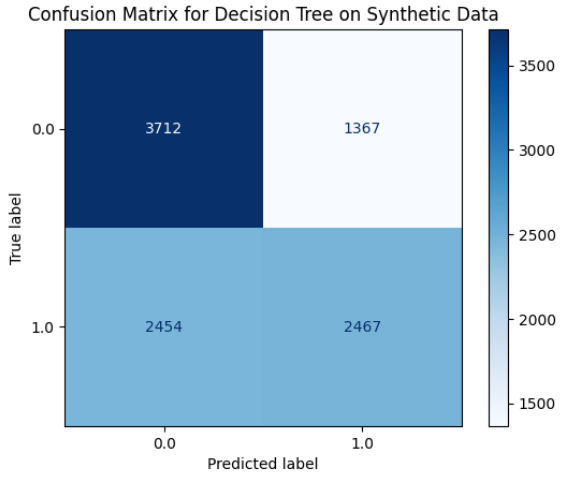
Matrice konfuzije za generisane podatke











CGAN

GENERISANI PODACI:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0	0.0	1	30.0	0.0	0.0	0.0	1	1	0	0	1	0.0	3.0	0.0	2.0	0.0	1	12	4.0	1.0
1	0.0	0	1.0	0	24.0	0.0	0.0	0.0	0	1	0	0	1	0.0	2.0	1.0	2.0	0.0	0	7	4.0	7.0
2	0.0	1	0.0	1	22.0	1.0	0.0	0.0	1	1	0	0	0	0.0	2.0	0.0	0.0	0.0	0	9	5.0	10.0
3	1.0	0	0.0	1	26.0	0.0	0.0	0.0	1	1	1	0	1	0.0	3.0	2.0	0.0	0.0	1	12	6.0	10.0
4	1.0	1	1.0	1	38.0	1.0	0.0	1.0	0	1	1	0	1	0.0	4.0	0.0	30.0	1.0	0	10	4.0	2.0
5	0.0	1	0.0	1	27.0	1.0	0.0	0.0	1	0	1	0	1	0.0	4.0	0.0	14.0	1.0	1	13	5.0	4.0
6	0.0	0	1.0	1	31.0	0.0	0.0	0.0	1	1	1	0	1	0.0	2.0	1.0	0.0	0.0	1	7	4.0	7.0
7	0.0	0	1.0	1	28.0	0.0	0.0	0.0	0	1	1	0	1	0.0	2.0	2.0	0.0	0.0	1	8	5.0	4.0
8	1.0	1	1.0	1	43.0	1.0	1.0	0.0	0	0	1	0	1	0.0	1.0	0.0	0.0	0.0	1	10	5.0	7.0
9	1.0	1	1.0	1	33.0	1.0	0.0	0.0	0	0	0	0	1	0.0	3.0	0.0	20.0	1.0	1	9	4.0	5.0
10	1.0	1	1.0	1	45.0	1.0	0.0	0.0	0	1	0	0	1	0.0	3.0	0.0	10.0	0.0	1	7	6.0	4.0
11	1.0	0	1.0	1	43.0	0.0	0.0	0.0	1	1	1	0	1	0.0	3.0	21.0	9.0	0.0	1	10	5.0	11.0
12	0.0	1	1.0	1	47.0	1.0	0.0	0.0	1	1	1	0	1	0.0	3.0	0.0	0.0	0.0	1	7	5.0	6.0
13	1.0	1	1.0	1	47.0	1.0	0.0	0.0	1	1	1	0	1	0.0	3.0	0.0	0.0	0.0	1	10	3.0	1.0
14	0.0	1	0.0	1	35.0	0.0	0.0	0.0	1	0	1	0	1	0.0	3.0	5.0	0.0	0.0	1	5	4.0	10.0
15	1.0	1	1.0	0	32.0	1.0	0.0	0.0	0	1	0	0	1	0.0	4.0	24.0	0.0	1.0	0	11	6.0	3.0
16	1.0	0	0.0	1	47.0	1.0	0.0	0.0	0	0	0	0	1	0.0	4.0	15.0	0.0	0.0	1	2	5.0	7.0
17	0.0	1	0.0	1	28.0	1.0	0.0	0.0	1	1	0	0	1	0.0	4.0	4.0	0.0	0.0	1	8	4.0	8.0
18	0.0	0	0.0	1	27.0	1.0	0.0	0.0	1	1	1	0	1	0.0	2.0	0.0	0.0	0.0	1	10	6.0	9.0
19	1.0	1	1.0	1	32.0	0.0	0.0	0.0	0	1	1	0	1	0.0	4.0	30.0	10.0	0.0	1	13	5.0	8.0
20	1.0	1	1.0	1	27.0	0.0	0.0	0.0	0	1	0	0	1	0.0	4.0	7.0	30.0	0.0	0	8	5.0	7.0

PRECIZNOST KLASIFIKACIJA NA GENERISANE PODATKE I NA PRAVI DATA SET:

Logistic regression:

	Accuracy	Precision	Recall	F1 Score
Real data	0.71	0.72	0.72	0.72
Synthetic data	0.7	0.75	0.64	0.69

Naive bayes:

	Accuracy	Precision	Recall	F1 Score
Real data	0.74	0.73	0.76	0.74
Synthetic data	0.7	0.76	0.65	0.7

Random forest:

	Accuracy	Precision	Recall	F1 Score
Real data	0.72	0.72	0.72	0.72
Synthetic data	0.69	0.72	0.67	0.7

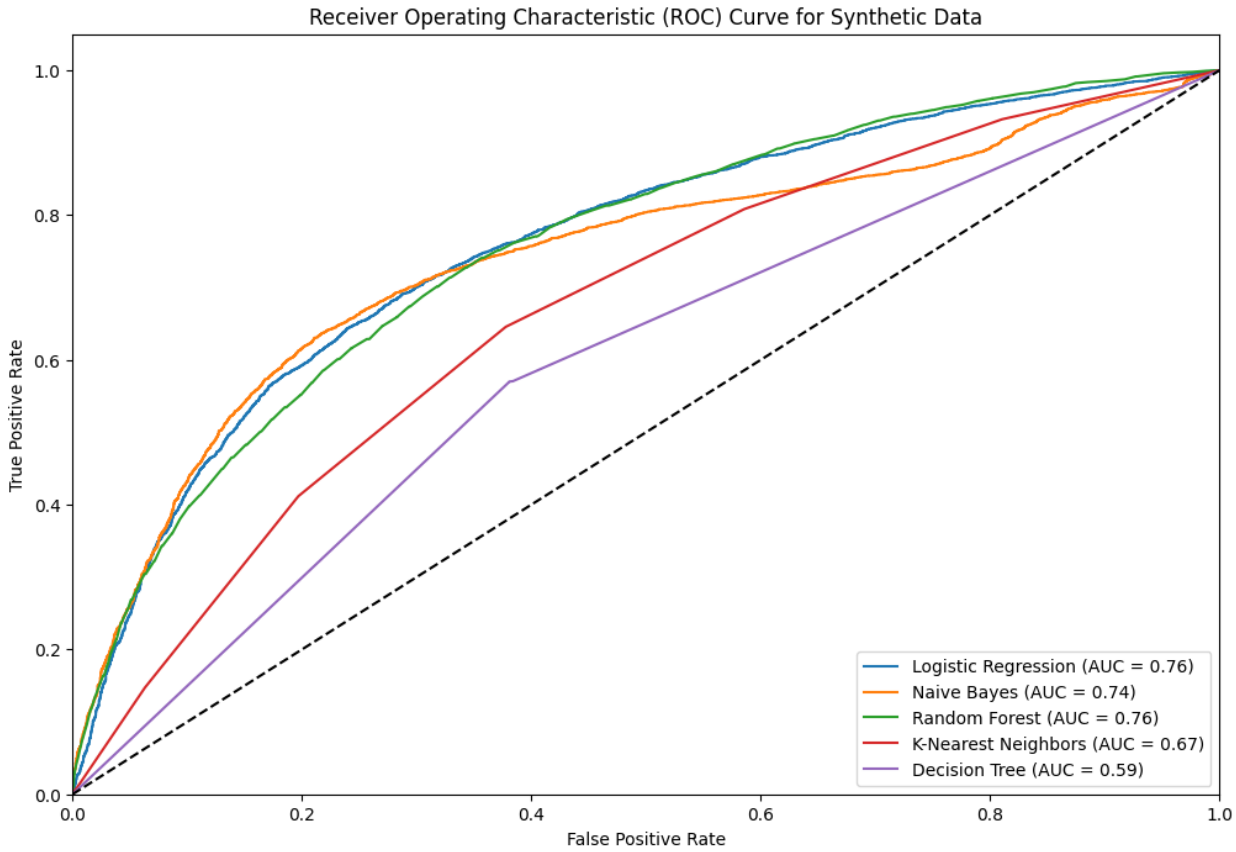
KNN:

	Accuracy	Precision	Recall	F1 Score
Real data	0.69	0.68	0.71	0.7
Synthetic data	0.63	0.66	0.65	0.65

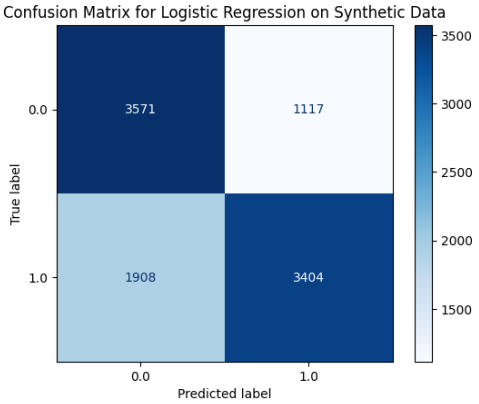
Decision tree:

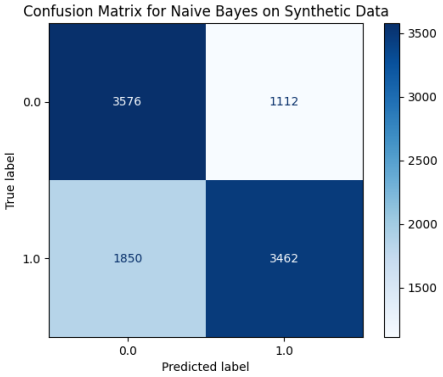
	Accuracy	Precision	Recall	F1 Score
Real data	0.73	0.72	0.74	0.73
Synthetic data	0.59	0.63	0.57	0.6

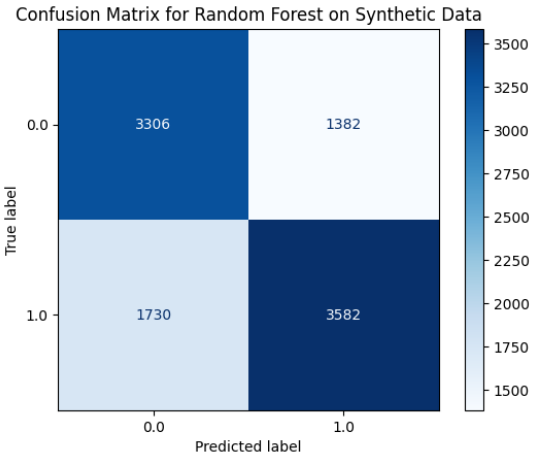
ROC-Kriva

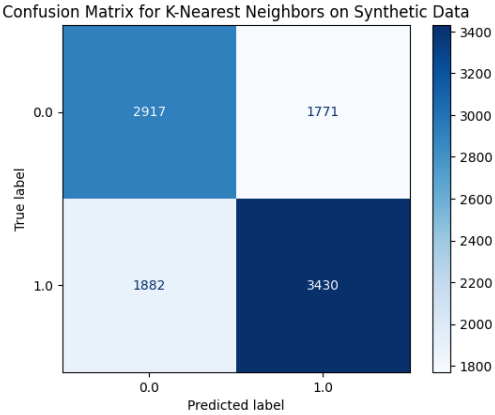


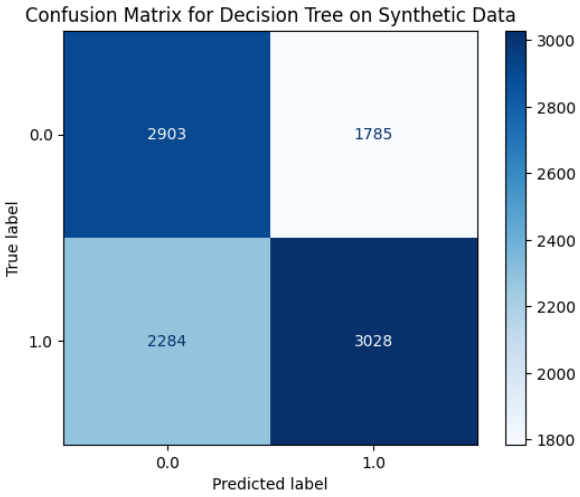
Matrice konfuzije za generisane podatke











Analiza rezultata

Svi klasifikacioni algoritmi su postigli skoro jednake performanse na pravim podacima, sa prosekom od oko 72%

Naive-Bayes model pokazuje konzistentne performanse na oba skupa podataka (trening i test). Tačnost, preciznost, odziv i F1 skor su gotovo isti za oba skupa (oko 0.74), što ukazuje da model nije previše prilagođen trening podacima (nema overfittinga) i da dobro generalizuje na neviđene podatke.

Decision Tree model takođe pokazuje slične performanse na trening i test skupu, sa tačnošću, preciznošću, odzivom i F1 skorom oko 0.73. Ovaj model takođe ne pati od overfittinga i pruža konzistentne rezultate na neviđenim podacima, slično Naive-Bayes modelu.

KNN model pokazuje značajno bolje performanse na trening skupu nego na test skupu. Tačnost na trening skupu je 0.79, dok je na test skupu samo 0.69, što ukazuje na overfitting. Model dobro funkcioniše na trening podacima, ali ne uspeva da generalizuje na nove podatke, što ga čini manje pouzdanim u stvarnom okruženju.

Logistička regresija pokazuje veoma konzistentne performanse na oba skupa podataka, sa tačnošću, preciznošću, odzivom i F1 skorom oko 0.72. Ovaj model dobro generalizuje i nema problema sa overfittingom, pružajući slične performanse kao Decision Tree model.

Random Forest model pokazuje izuzetne performanse na trening skupu (tačnost 0.90), ali značajan pad performansi na test skupu (tačnost 0.72). Ovo je jasan znak overfittinga, gde model dobro uči specifične karakteristike trening podataka, ali ne uspeva da se dobro ponaša na neviđenim podacima, što ga čini manje poželjnim za upotrebu na stvarnim podacima uprkos visokim performansama na trening skupu.

Za većinu klasifikacionih algoritama (logistička regresija, slučajna šuma), tačnost klasifikacije na sintetičkim podacima generisanim pomoću CGAN metode je bliska tačnosti na pravim podacima. Ovo ukazuje da CGAN model uspešno generiše sintetičke podatke koji su slični pravim podacima i mogu se koristiti za obučavanje ovih algoritama.

Model Logistic Regression na realnim podacima pokazuje stabilne performanse. Koristeći GAN generisane podatke, preciznost je vrlo visoka, ali je odziv veoma nizak, što rezultira nižim F1 skorom. CGAN generisani podaci daju bolje uravnotežene rezultate sa boljim F1 skorom, što pokazuje da CGAN generisani podaci bolje simuliraju realne podatke.

Naive Bayes model na realnim podacima pokazuje visoku tačnost i uravnotežene metrike. GAN generisani podaci daju vrlo visoku preciznost, ali ekstremno nizak odziv, što rezultira veoma niskim F1 skorom. CGAN generisani podaci pokazuju mnogo bolje uravnotežene rezultate sa većom tačnošću i F1 skorom, što znači da su CGAN podaci bolji za treniranje ovog modela.

Random Forest model na realnim podacima pokazuje uravnotežene performanse. GAN generisani podaci daju višu tačnost i preciznost, ali niži odziv, što rezultira sličnim F1 skorom kao kod realnih podataka. CGAN generisani podaci daju uravnoteženije performanse sa sličnim F1 skorom kao GAN, ali bolju raspodelu između preciznosti i odziva.

KNN model na realnim podacima pokazuje stabilne performanse. GAN generisani podaci daju slične rezultate kao i realni podaci sa sličnom tačnošću i F1 skorom. CGAN generisani podaci daju nešto nižu tačnost i F1 skor, što ukazuje na nešto slabiju generalizaciju u poređenju sa GAN metodom, ali i dalje pružaju solidne performanse.

Decision Tree model na realnim podacima pokazuje uravnotežene performanse. GAN generisani podaci daju nižu tačnost i F1 skor zbog niskog odziva. CGAN generisani podaci pokazuju blago poboljšanje sa višim F1 skorom i boljim balansom između preciznosti i odziva u poređenju sa GAN podacima.

Generalno, svi algoritmi su pokazali bolju tačnost na pravim podacima nego na sintetičkim, što je i očekivano s obzirom na to da sintetički podaci nisu savršeni. Međutim, rezultati ukazuju na to da CGAN metoda može biti korisna za generisanje sintetičkih podataka koji se mogu koristiti za obučavanje klasifikacionih modela, posebno kada su pravi podaci ograničeni ili nedostupni. Važno je napomenuti da kvalitet sintetičkih podataka može varirati u zavisnosti od primene i korišćenog GAN modela.

U poređenju sa GAN-om, CGAN pokazuje potencijal za generisanje sintetičkih podataka koji su više prilagođeni specifičnim uslovima ili atributima, što može biti korisno u određenim primenama.

Zaključak

Ovaj proces pripreme podataka za klasifikaciju dijabetesa bio je korak ka stvaranju pouzdanih modela koji mogu pomoći u identifikaciji rizika od ove bolesti. Iako su korišćeni algoritmi dali zadovoljavajuće rezultate, važno je imati na umu da tačnost modela zavisi od mnogih faktora, uključujući prirodu problema i veličinu skupa podataka. Dalji koraci u analizi mogu obuhvatiti proširivanje trening skupa ili dodavanje novih atributa kako bi se poboljšala prediktivna moć modela. Kroz ovaj rad, stekli smo dragoceno iskustvo u obradi podataka i primeni algoritama mašinskog učenja na konkretan problem zdravstvene analitike. Ovo iskustvo će biti od neprocenjive važnosti za buduće radove u oblasti mašinskog učenja i analize podataka.

Literatura

https://www.kaggle.com/code/samanemami/gan-on-tabular-data

https://aws.amazon.com/what-is/gan/

https://www.learndatasci.com/glossary/binary-classification/

https://www.freecodecamp.org/news/binary-classification-made-simple-with-tensorflow/