

**Državni univerzitet u Novom Pazaru**

**Departman za Tehničke nauke**

**Smer: Softversko inženjerstvo**

**Seminarski rad**

**Predmet: Mašinsko učenje**

**Tema:**

**Klasifikacija vesti uz pomoć text mining-a**

**Mentor: Student:**

**Doc.dr. Ulfeta Marovac Ibrahimović Tarik**

**Hamzić Ensar**

**Novi Pazar, 2024.godina**

Contents

[**Uvod I opis problema** 1](#_Toc166710283)

[**Opis skupa podataka** 1](#_Toc166710284)

[**Opis metoda koje se koristi i metrike za evaluaciju** 2](#_Toc166710285)

[ **logistička regresija** 2](#_Toc166710286)

[ **Naivni Bajesov algoritam** 3](#_Toc166710287)

[ **KNN** 4](#_Toc166710288)

[ **Decision Tree (odlučujuće drvo)** 4](#_Toc166710289)

[ **Random Forest (slučajna šuma)** 4](#_Toc166710290)

[**Priprema podataka** 5](#_Toc166710291)

[**Rezultati** 6](#_Toc166710292)

[**Heat-map dataset-a** 7](#_Toc166710293)

[**Naïve-Bayes** 8](#_Toc166710294)

[**Decision Tree Classification** 8](#_Toc166710295)

[**KNN** 10](#_Toc166710296)

[**Logistička regresija** 11](#_Toc166710297)

[**Random Forest** 11](#_Toc166710298)

[**GAN** 12](#_Toc166710299)

[**Analiza rezultata** 14](#_Toc166710300)

[**Zakljucak** 14](#_Toc166710301)

[**Literatura** 15](#_Toc166710302)

# **Uvod I opis problema**

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

Svaki red podataka sadrzi 21 podatak od kojih su neki bitniji:

* Diabetes\_binary
* HighBP
* HighChol

# **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

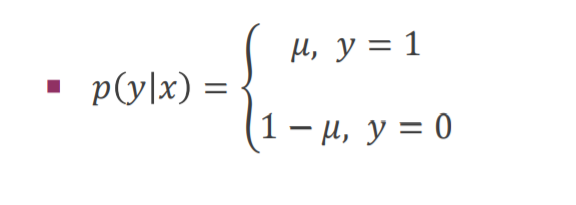
* Diabetes\_binary: Indikator da li osoba ima dijabetes (1) ili ne (0). Ovaj atribut je ključan za identifikaciju populacije sa dijabetesom.
* BMI (Indeks telesne mase): Mera telesne masti na osnovu visine i težine osobe. Visok BMI može ukazivati na gojaznost, što je faktor rizika za dijabetes.
* Smoker: Indikator da li je osoba trenutno pušač (1) ili ne (0). Pušenje je značajan faktor rizika za mnoge bolesti, uključujući bolesti srca i dijabetes.
* PhysActivity (Fizička aktivnost): Indikator da li se osoba bavi fizičkim aktivnostima (1) ili ne (0). Fizička aktivnost je važna za održavanje zdravlja i smanjenje rizika od dijabetesa.
* GenHlth (Opšte zdravstveno stanje): Subjektivna procena opšteg zdravstvenog stanja osobe, obično na skali od 1 do 5, gde niža vrednost označava bolje zdravlje.

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.

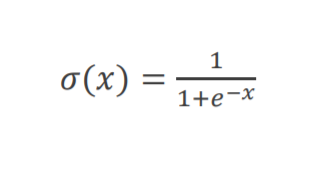
# **Opis metoda koje se koristi i metrike za evaluaciju**

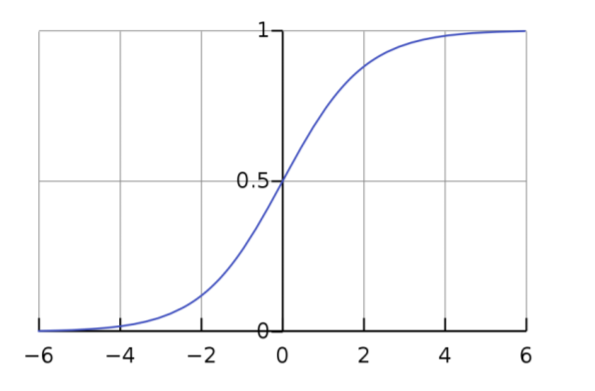
Naš odabir metoda za rešavanje ovog problema se svodi na sledećih pet:

* + **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.

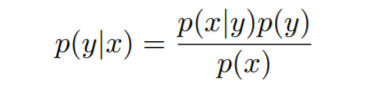


S' obzirom da logistička regresija radi sa verovatnoćama, granične vrednosti su zadate intervalom [0,1], a ovaj model se dobio transformacijom linearnog modela sa intervala [-∞,∞] pomoću neprekidne diferencijabilne funkcije, koja se još naziva sigmoidna funkcija.





* **Naivni Bajesov algoritam –** se zasniva na modelovanju raspodele ciljne promenljive y pri datim vrednostima promenljive x, korišćenjem Bajesove formule:

****

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.

* **KNN -** 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 (odlučujuće drvo)** - 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) -** 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.

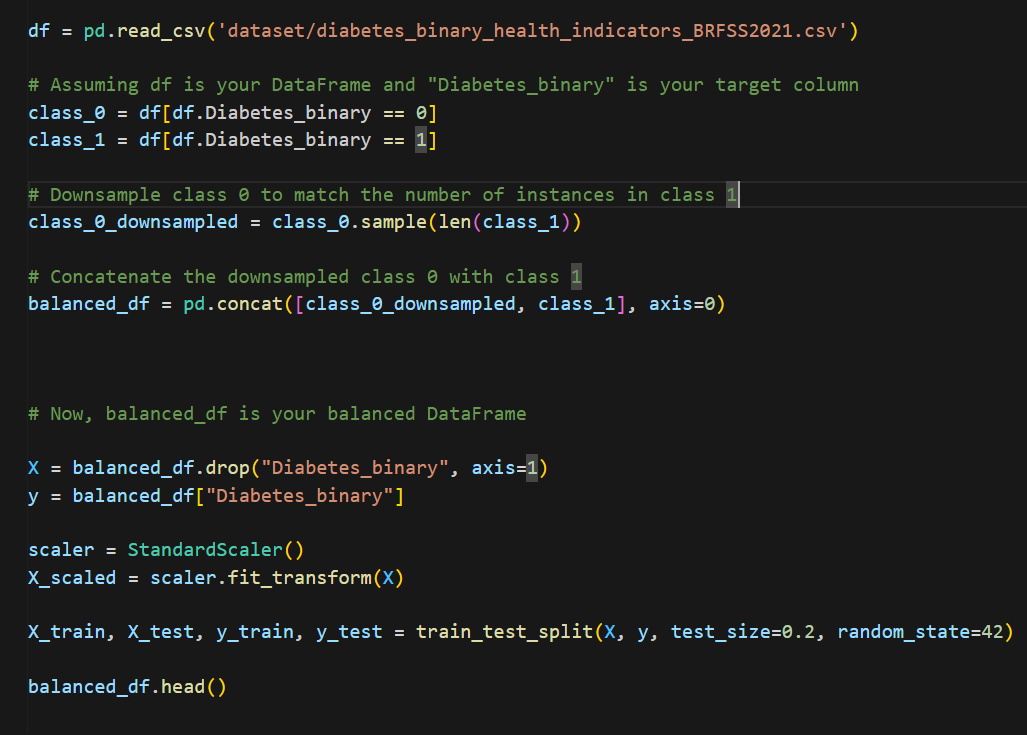
* **GAN (Generative Adversarial Networks)** su vrsta generativnih modela mašinskog učenja koji se sastoje od dva konkurentna neuronska modela - generatora i diskriminatora. Ova metoda se razlikuje od KNN i Decision Tree algoritama koje ste opisali, jer se ne koristi za klasifikaciju ili regresiju, već 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.

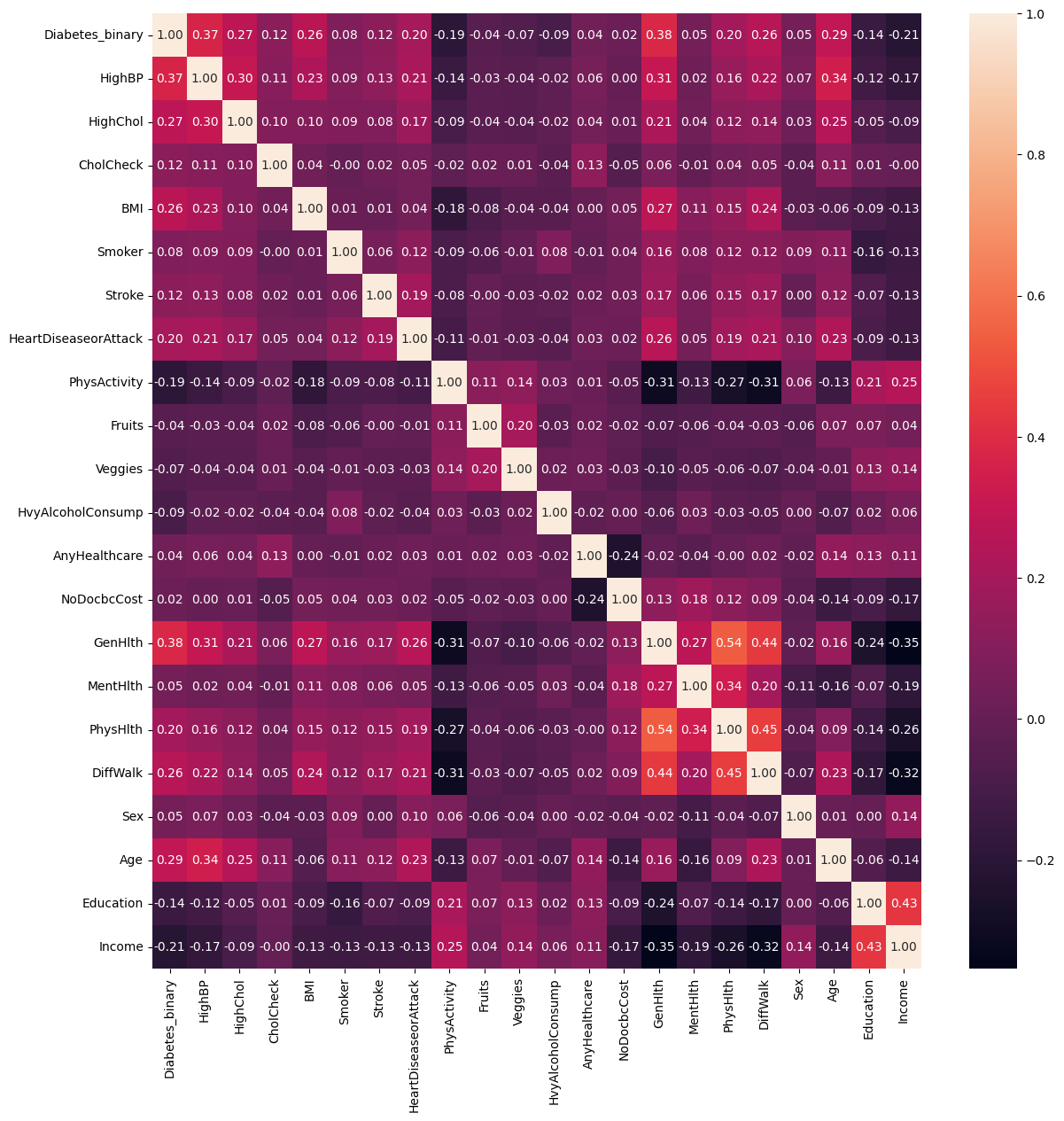
# **Priprema podataka**

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. Zatim su atributi standardizovani kako bi se svi podaci nalazili na istom rasponu vrednosti, što poboljšava performanse mnogih algoritama mašinskog učenja. Konačno, 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.

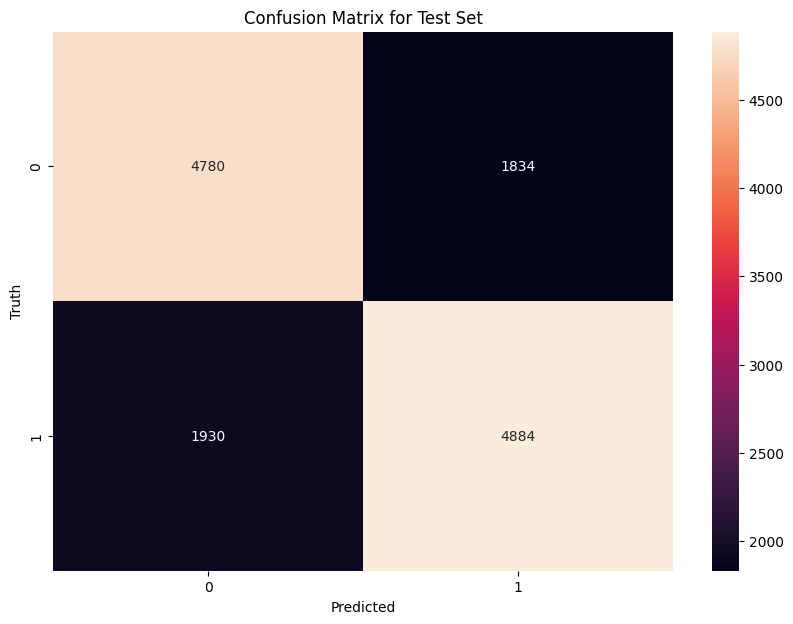


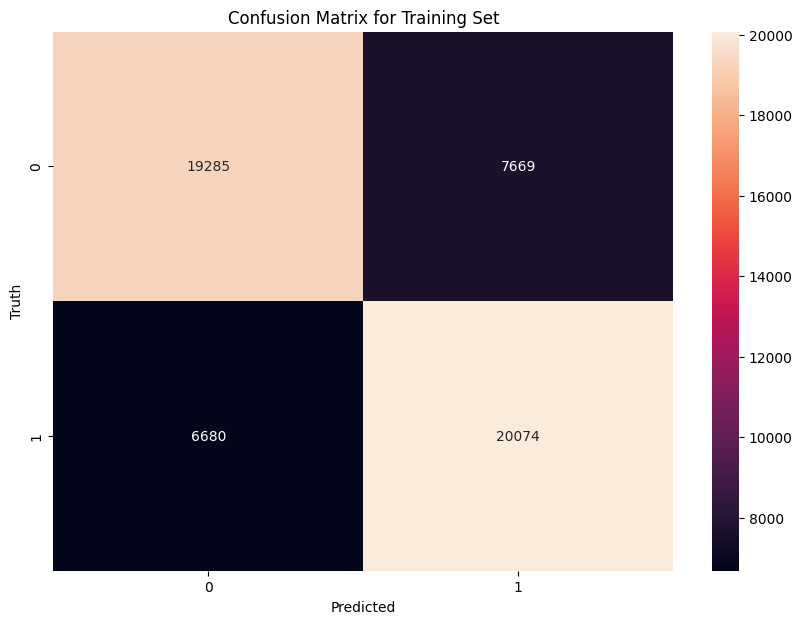
# **Rezultati**

### **Heat-map dataset-a**

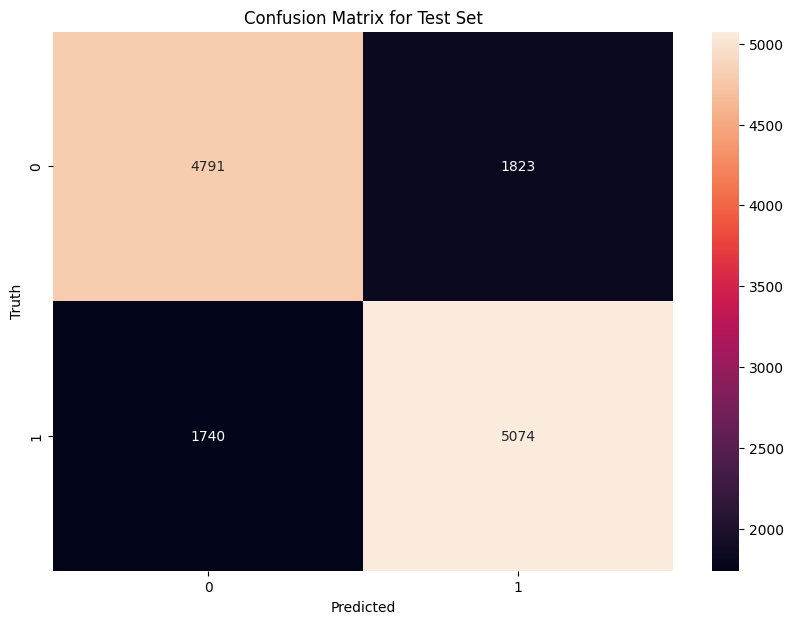


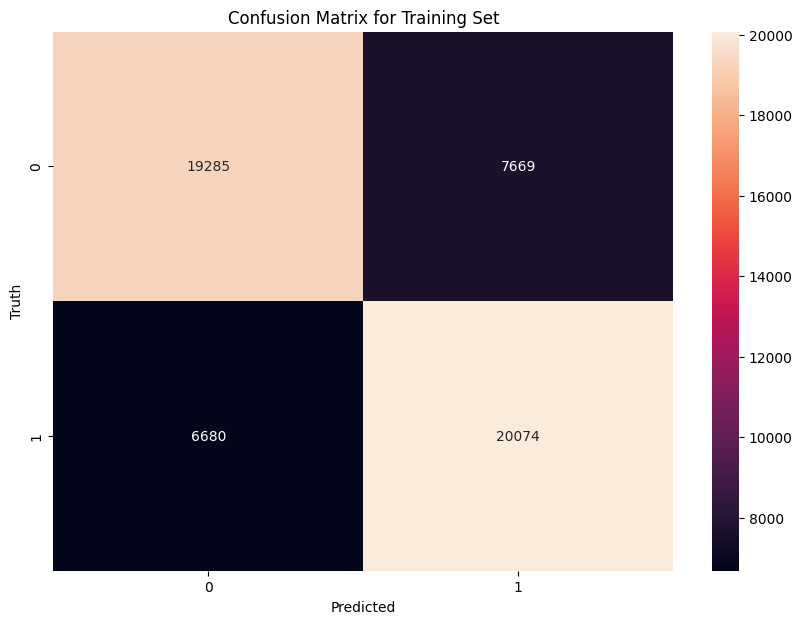
### **Naïve-Bayes**



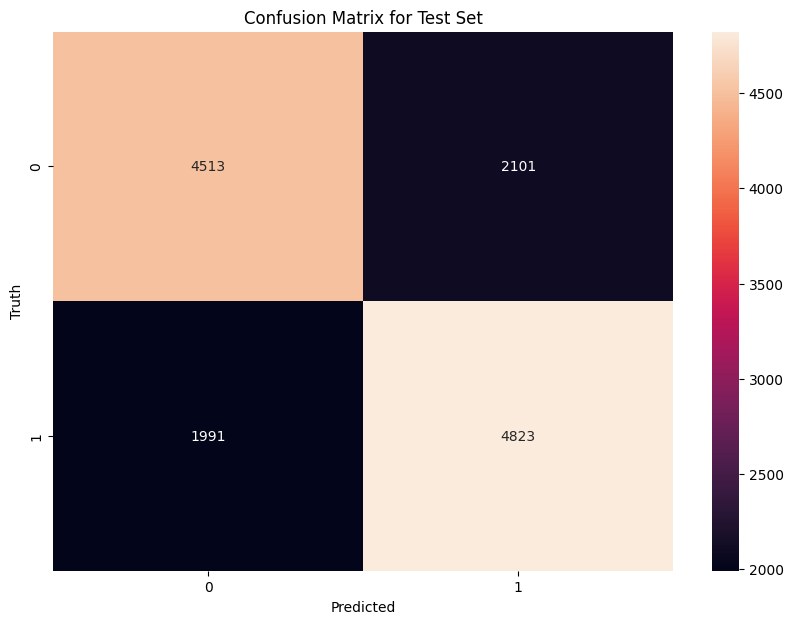


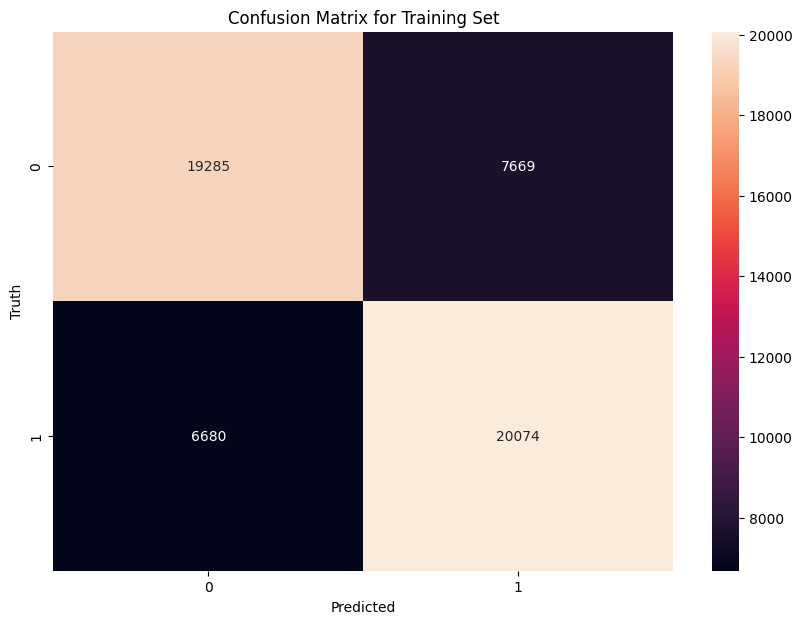
### **Decision Tree Classification**



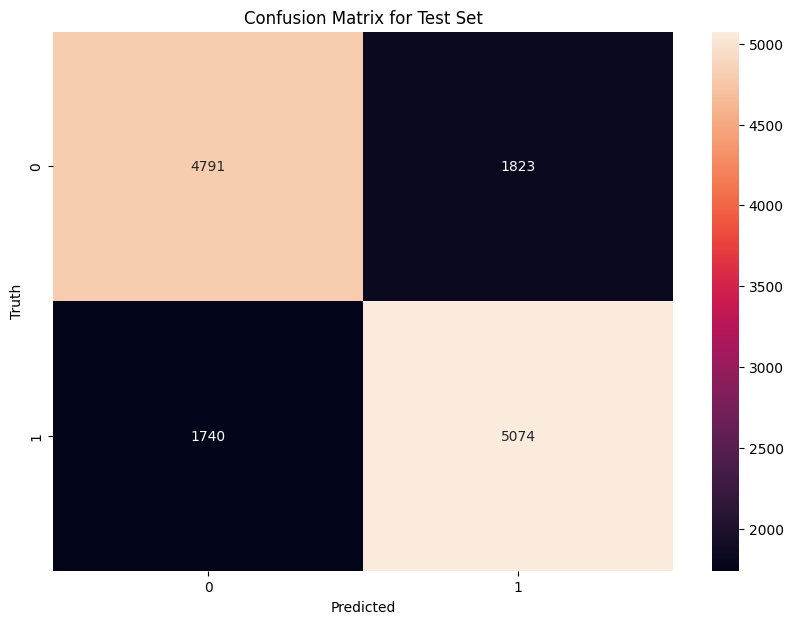


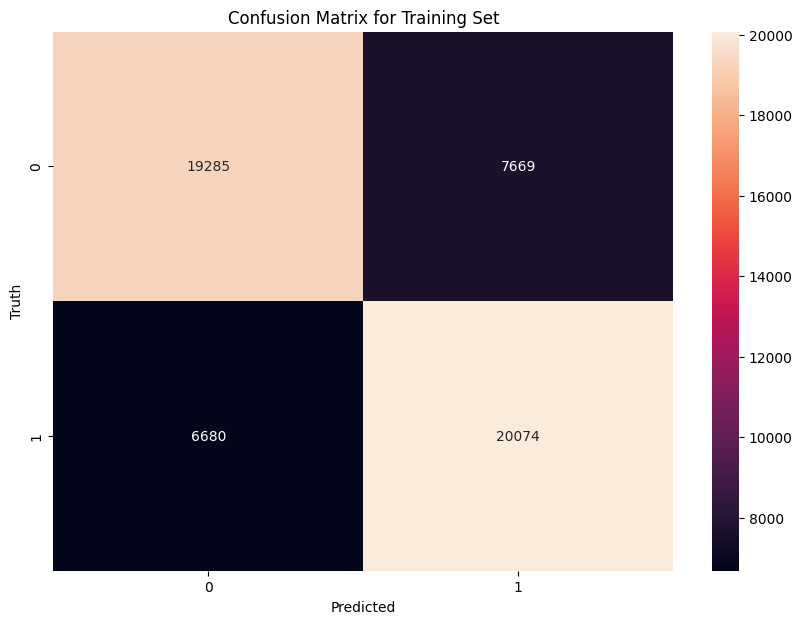
### **KNN**



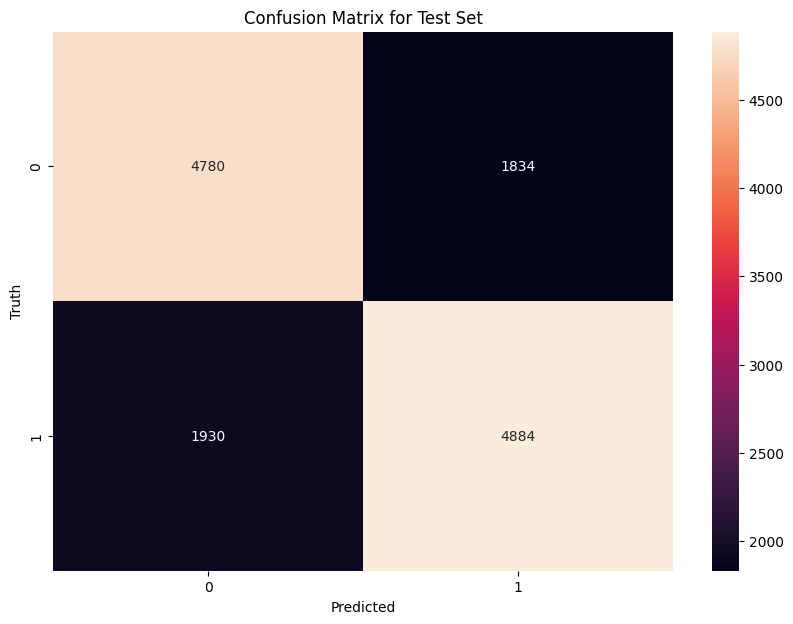


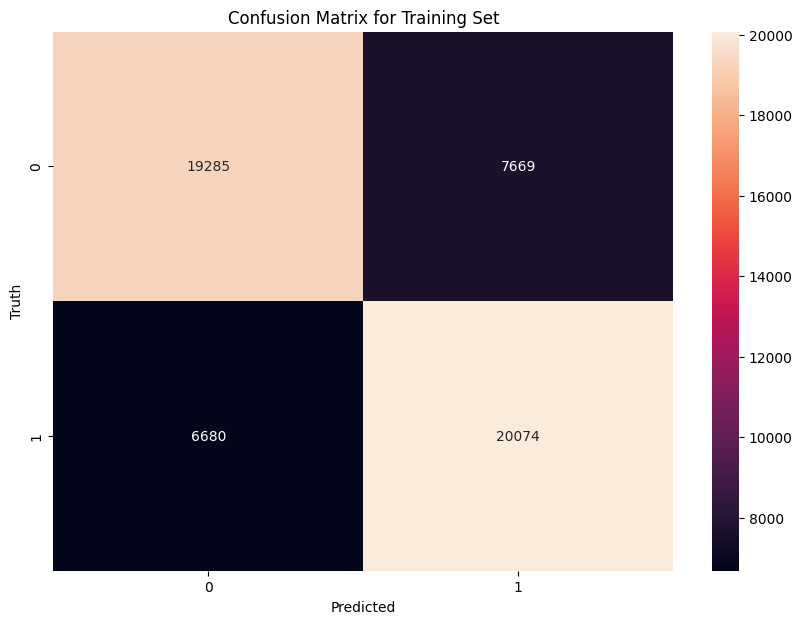
### **Logistička regresija**





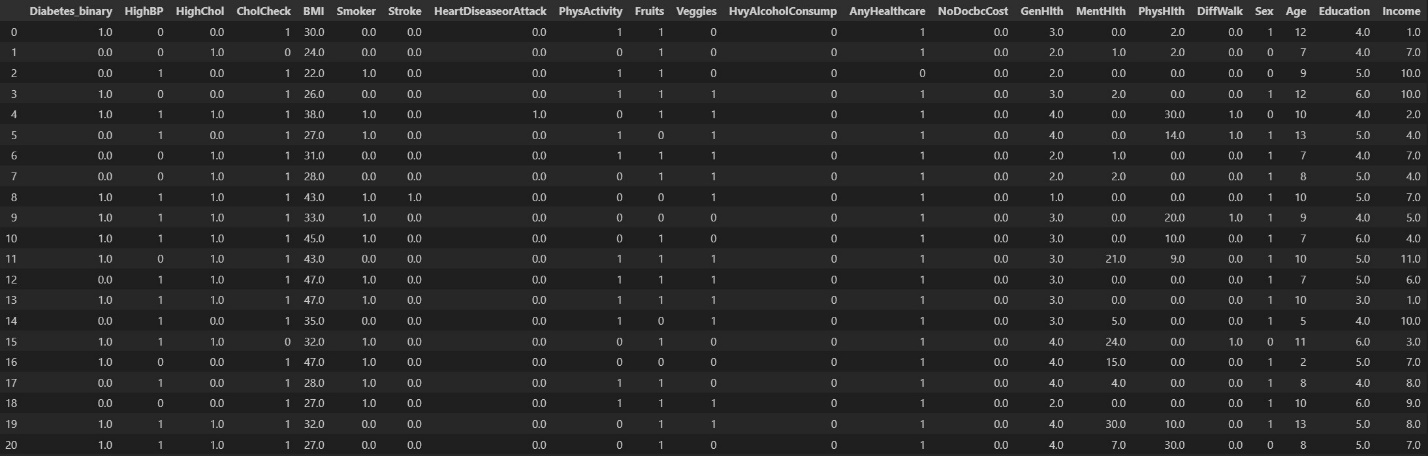
### **Random Forest**





### **GAN**

#### **GENERISANI PODACI:**



**PRECIZNOST KLASIFIKACIJA NA GENERISANE PODATKE I NA PRAVI DATA SET:**

**Logistic Regression:**

Accuracy on real data: 0.7357759904676795

Accuracy on synthetic data: 0.74

**Naive Bayes:**

Accuracy on real data: 0.7167858206732202

Accuracy on synthetic data: 0.7

**Random Forest:**

Accuracy on real data: 0.7349568066726244

Accuracy on synthetic data: 0.705

**K-Nearest Neighbors:**

Accuracy on real data: 0.6925826630920465

Accuracy on synthetic data: 0.67

**Decision Tree:**

Accuracy on real data: 0.649836163240989

Accuracy on synthetic data: 0.615

# **Analiza rezultata**

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

 Logistička regresija je postigla najbolje performanse kako na pravim, tako i na sintetičkim podacima, sa tačnošću od oko 73-74%.

 K-najbliži sused i stablo odlučivanja su postigli nešto lošije rezultate na obe vrste podataka u poređenju sa ostalim algoritmima. Ovo može ukazivati da su manje pogodni za ovu vrstu problema klasifikacije ili da su osetljiviji na kvalitet podataka.

 Generalno, svi algoritmi su pokazali bolju tačnost na pravim podacima nego na sintetičkim, što je i očekivano s obzirom da sintetički podaci nisu savršeni.

 Dobijeni rezultati sugerišu da GAN metoda može biti korisna za generisanje sintetičkih podataka koji se mogu koristiti za obučavanje klasifikacionih modela, posebno kada su pravi podaci ograničeni ili nedostupni. Međutim, važno je imati na umu da kvalitet sintetičkih podataka i dalje može varirati u zavisnosti od primene i korišćenog GAN modela.

# **Zakljucak**

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/>