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

**Raspoznavanje uzoraka i strojno učenje**

**Predviđanje dijabetesa**

**Projektni zadatak**

**Marko Cvitković**

**Osijek, 2023.**

Sadržaj

[**1.Uvod** 3](#_Toc128326748)

[**2. Opis podatkovnog skupa i eksplorativna analiza** 4](#_Toc128326749)

[**3. Modeli** 8](#_Toc128326750)

[**3.1.Metoda K-najbližih susjeda** 8](#_Toc128326751)

[**3.2. Logističku regresije** 9](#_Toc128326752)

[**3.3.SVM** 10](#_Toc128326753)

[**4. Usporedba algoritma** 11](#_Toc128326754)

[**5. Aplikacija** 12](#_Toc128326755)

[**6. Zaključak** 15](#_Toc128326756)

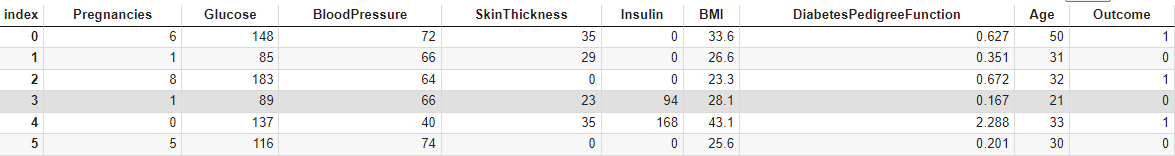
# **1.Uvod**

Dijabetes, također poznat i kao šećerna bolest, je kronično stanje koje se javlja kada tijelo ne može regulirati razinu šećera (glukoze) u krvi na normalan način. Postoji nekoliko tipova dijabetesa, ali najčešći su tip 1, tip 2 i gestacijski dijabetes. Simptomi dijabetesa mogu uključivati ​​povećanu žeđ, često mokrenje, umor, gubitak težine, zamućen vid i usporeno zacjeljivanje rana. Dijabetes može dovesti do komplikacija, uključujući probleme s bubrezima, očima, živcima, krvnim žilama i srcem. Dijagnoza dijabetesa uključuje testiranje razine glukoze u krvi. Liječenje dijabetesa uključuje prilagođavanje prehrane, vježbanje i uzimanje lijekova, uključujući inzulin. Važno je kontrolirati razinu šećera u krvi kako bi se spriječile komplikacije. Prevencija dijabetesa uključuje održavanje zdrave tjelesne težine, vježbanje i pravilnu prehranu bogatu cjelovitim žitaricama.

Glavni cilj projekta je na temelju određenih ulaznih vrijednosti odrediti ima li osoba dijabetes ili ne. Ovaj problem je klasifikacijski i uspoređivati će se 3 metode.

# **2.** **Opis podatkovnog skupa i eksplorativna analiza**

Funkcionalnost aplikacije se ostvaraju tako što se na temelju ulaznih vrijednosti za skup podataka koji sadrži 9 atributa predviđa ima li osoba dijabetes ili ne. Ovaj skup sadrži sljedeće atribute:1. trudnoća, 2. glukoza, 3. krvni tlak, 4.debljina kože, 5. inzulin, 6. indeks tjelesne mase, 7. dijabetes pedigre funkcija, 8.starost i 9.klasa. koje su prikazane u tablici 2.1.



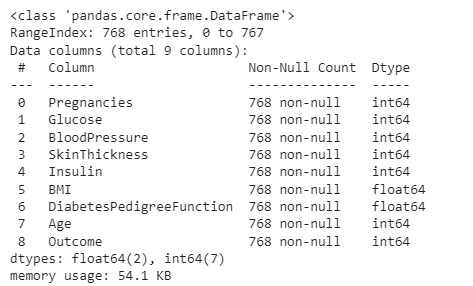
**Tablica 2.1** Atributi podatkovnog skupa

Kako bi se pristupilo podacima, korištena je funkcija *read\_csv()* kako je prikazano na *slici 2.1.*



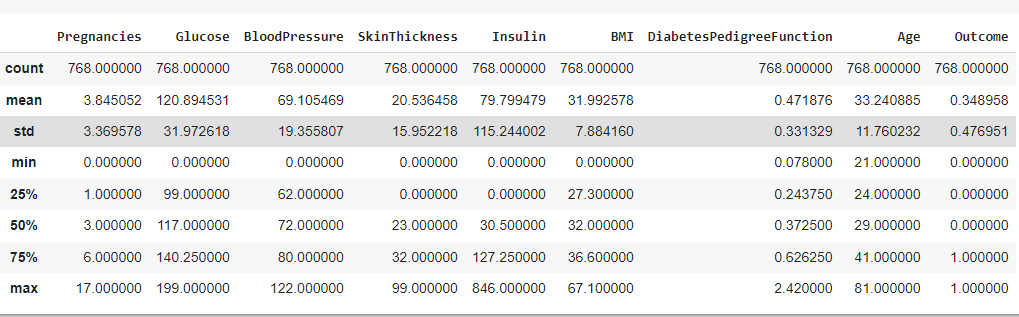
**Slika 2.1.** Učitavanje podataka iz CSV datoteke

Podatkovni skup sadržava 768 retka i 9 stupaca (atributi i ciljna varijabla). Informacije o varijablama prikazane su na *slici 2.2.*



**Slika 2.2.** Broj stupaca i redova, te informacije o varijablama

Pomoću funkcije describe() dobili smo tablicu 2.2. gdje su prikazane minimalne i maksimalne vrijednosti pojedinih atributa, srednja vrijednost, njihov broj…

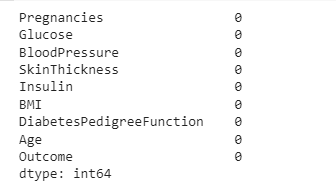


**Tablica 2.2.** Vrijednosti atributa

Uz pomoć funkcije *isnull().sum()* provjereno je ima li u neko stupcu

podatak koji fali. Kao sto je prikazano na *slici 2.3* nema vrijednosti koje

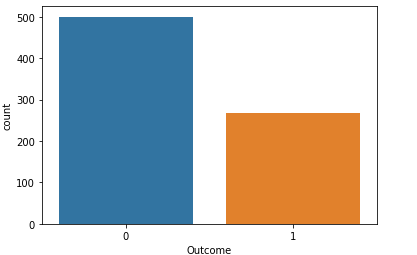
nedostaju.



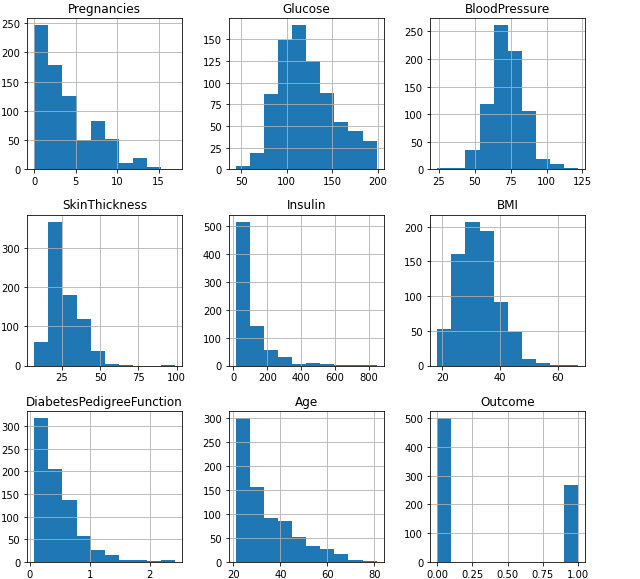
**Slika 2.3.** Vrijednosti koje nedostaju

U ovom modelu smo koristili tehniku standardizacija. Standardizacija je postupak skaliranja ulaznih varijabli tako da imaju srednju vrijednost od nula i standardnu devijaciju od jedan. To se radi kako bi se osigurala jednakost težina za svaku varijablu prilikom izgradnje modela strojnog učenja i kako bi se poboljšale performanse.

Od 768 slučajeva njih 500 je negativno dok je 268 pozitivno kao što je prikazano na grafu 2.1.,a to smo dobili pomoću funkcije *sns.countplot('Outcome',data=df)*.

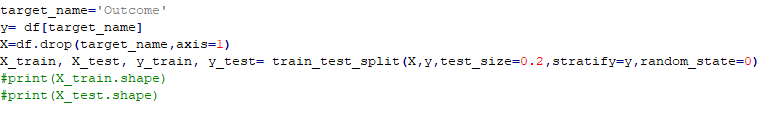


**Graf 2.1.** Prikaz pozitivnih i negativnih rezultat



**Slika 2.5**. Histogrami atributa

Podaci su podijeljeni na test i train u omjeru 0.2:0.8. Test podataka ima 154, a trening 614.



**Slika 2.6**. Korištenje *train\_test\_split*

# **3. Modeli**

## **3.1.Metoda K-najbližih susjeda**

Metoda k najbližih susjeda (engl. K-nearest neighbors, skraćeno KNN) je jedna od najjednostavnijih i najpopularnijih metoda nadziranog strojnog učenja. Ova metoda se često koristi za rješavanje problema klasifikacije ili regresije.

Osnovna ideja KNN metode je da se za novi ulazni primjerak (tj. novi skup značajki) pronađe K postojećih primjeraka iz skupa za učenje koji su mu najbliži prema nekoj mjeri udaljenosti, npr. euklidskoj udaljenosti. Kada se pronađu ti K najbližih susjeda, novi primjerak se klasificira (ili regresira) prema klasi (ili vrijednosti) koja je najzastupljenija među tim K susjeda.

KNN metoda je vrlo intuitivna i jednostavna za implementaciju, ali ima i nekoliko nedostataka. Na primjer, njezina izvedba se pogoršava kada postoji veliki broj dimenzija u prostoru značajki, kada skup za učenje ima puno primjeraka ili kada postoje šumovi i odstupanja u skupu za učenje. Također, odabir pravog K može biti izazovan problem i ovisi o specifičnosti problema koji se rješava.

Za odabir K u KNN metodi smo koristili križnu validaciju. Križna validacija (engl. cross-validation) je postupak procjene performansi modela strojnog učenja. Koristi se za procjenu kvalitete modela na način da se skup podataka dijeli na nekoliko dijelova, pri čemu se svaki dio koristi kao testni skup, dok se ostali dijelovi koriste kao skup za treniranje modela. Na taj način se model trenira i testira više puta, te se računa srednja vrijednost performansi modela na svim testnim skupovima.

Za K=5 dobiveno je 85,67% preciznost na trening podacima dok za test podatke je dobiveno 79,22% kao što se vidi na slici 3.1.



**Slika 3.1.** Rezultat KNN-metode

Slika 3.2. prikazuje matricu konfuzije za K metodu najbližih susjeda gdje 87 predstavlja točne pozitivne, 35 točne negativne, 13 netočne pozitivne i 19 netočne negativne.



**Slika 3.2.** Matrica konfuzije

## **3.2. Logističku regresije**

Logistička regresija je algoritam klasifikacije nadziranog učenja koji se koristi za predviđanje vjerojatnosti ciljne varijable. Priroda ciljne (zavisne) varijable je dihotomna, što znači da postoje dvije moguće klase. 1 označava uspjeh, a 0 označava neuspjeh. Ovaj model jedan je od najjednostavnijih ML algoritama koji se može koristiti za različite probleme klasifikacije, poput:, otkrivanja neželjene pošte, predviđanje dijabetesa, otkrivanje raka itd.

Za logističku regresiju dobiveno je 78,01% preciznost za trening podatke, a za test podatke 77,92%.



**Slika 3.3.** Rezultat logističke regresije

Slika 3.4. prikazuje matricu konfuzije za logističku regresiju gdje 89 predstavlja točne pozitivne, 31 točne negativne, 11 netočne pozitivne i 23 netočne negativne.



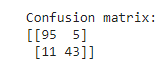
**Slika 3.4.** Matrica konfuzije

## **3.3.SVM**

Support Vector Machine (SVM) ili stroj potpornih vektora je popularan algoritam strojnog učenja koji se koristi za rješavanje problema klasifikacije i regresije. Osnovni cilj SVM-a je pronaći hiperravninu u visokodimenzionalnom prostoru koja najbolje razdvaja dvije ili više klase podataka.

Kada se koristi za klasifikaciju, SVM pokušava pronaći najbolju moguću hiperravninu koja dijeli podatke na dva različita razreda. Hiperravnina se nalazi tako da maksimizira razmak između dviju razreda. Razmak se definira kao udaljenost između hiperravnine i najbližih točaka iz svakog razreda, koje se nazivaju potpornim vektorima. Nedostatak SVM-a je da je osjetljiv na pretreniranje i zahtijeva temeljitu optimizaciju hiperparametara.

Za SVM dobiveno je 90,34% preciznost za trening podatke, a za test podatke 85,11%. Slika 3.5. prikazuje matricu konfuzije za SVM regresiju gdje 95 predstavlja točne pozitivne, 43 točne negativne, 5 netočne pozitivne i 11 netočne negativne.



**Slika 3.5.** Matrica konfuzije

# **4. Usporedba algoritma**

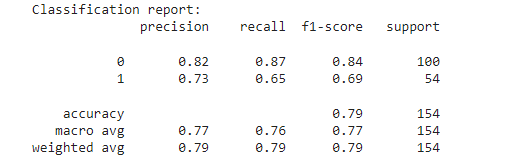
Izvještaj o klasifikaciji (engl. classification report) je statistički izvještaj koji se koristi za evaluaciju performansi klasifikacijskog modela. Sadrži informacije o točnosti (engl. accuracy) modela te preciznosti (engl. precision), odzivu (engl. recall) i F1-score-u (engl. F1-score) za svaku klasu modela.

Točnost modela je jednostavno razumijevanje - označava postotak točnih predviđanja klasa u odnosu na ukupan broj predviđanja.

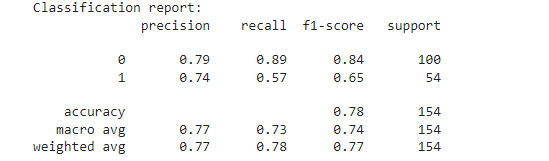
Preciznost se računa kao omjer broja stvarnih pozitivnih (TP) predviđanja i ukupnog broja pozitivnih predviđanja (TP + FP). Preciznost mjeri koliko su predviđanja pozitivne klase točna. U tom parametru najbolji se pokazao SVM kao što je prikazano na slici 4.3, a najgori logistička regresija.

Odziv se računa kao omjer broja stvarnih pozitivnih (TP) predviđanja i ukupnog broja pozitivnih primjera (TP + FN). Odziv mjeri koliko su stvarnih pozitivnih primjera identificirali model. U tom parametru najbolji se također pokazao SVM kao što je prikazano na slici 4.3, a najgori logistička regresija i isto je i za f1-score

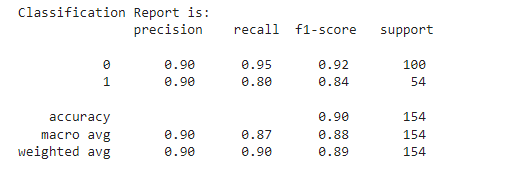
F1-score je harmonijska sredina između preciznosti i odziva te predstavlja njihovu ravnotežu. Računa se kao 2 \* (preciznost \* odziv) / (preciznost + odziv).



**Slika 4.1.** Izvještaj o klasifikaciji za KNN



**Slika 4.2.** Izvještaj o klasifikaciji za Logističku regresiju

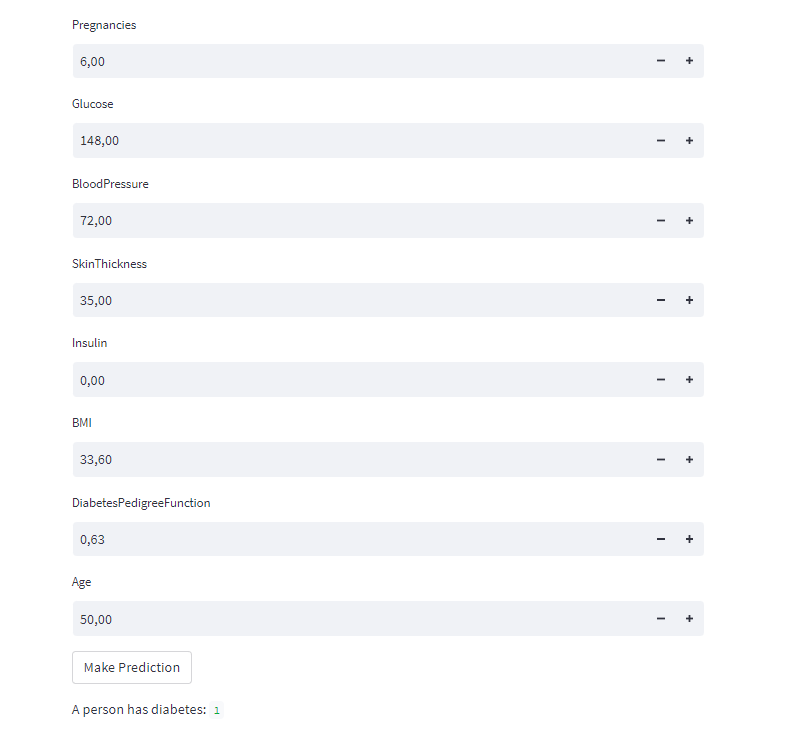


**Slika 4.3.** Izvještaj o klasifikaciji za SVM

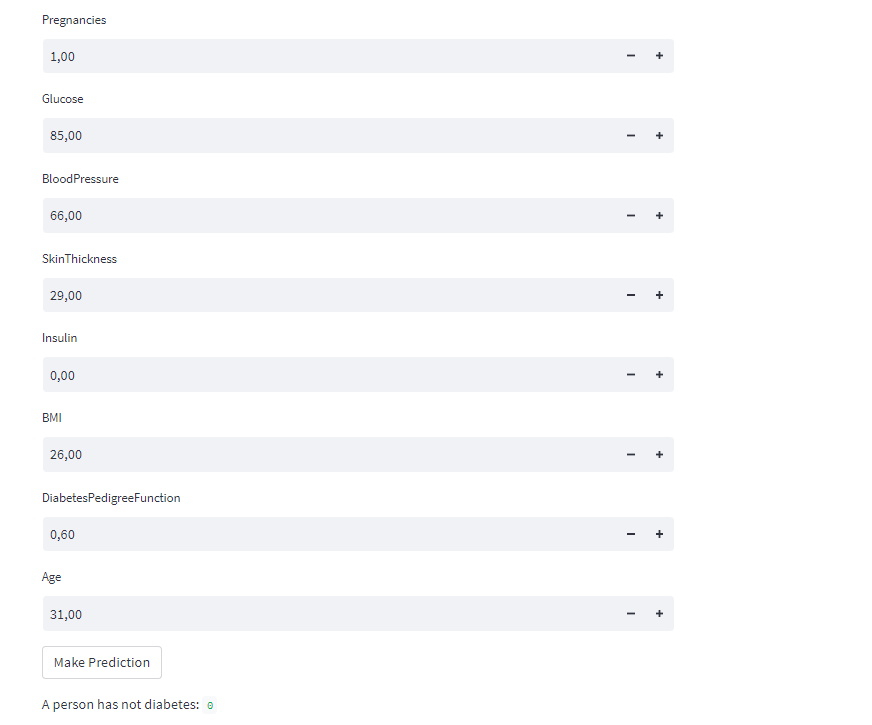
# 

# **5.Aplikacija**

Aplikacija je napravljen u Streamlitu koristeći metodu SVM zbog toga što se pokazala najveću preciznost u odnosu na ostale 2 metode. Streamlit je otvoreni okvir za izradu web aplikacija pomoću Pythona. Omogućuje brzu i jednostavnu izradu interaktivnih web aplikacija koje se mogu koristiti za vizualizaciju podataka, strojno učenje, analizu podataka…Korisnik ispunjava 8 podatak i nakon toga pritiskom na gumb “Make predicion“ dobiva konačan rezultat kao što je prikazano na slikama 4.1 i 4.2.



**Slika 4.1.** Prikaz rezultata predikcije



**Slika 4.2.** Prikaz rezultata predikcije

# **6. Zaključak**

Odabirom jednog od tri modela (SVM, linearne regresije ili metoda K najbližih susjeda) može se predvidjeti ima li osoba dijabetes ili ne. Model s najvećom točnošću je SVM, a model s najmanjom točnošću je metoda logistička regresija Korištenje aplikacije vrlo je jednostavno, potrebno je unijeti parametre u formu, a zatim kliknuti na gumb “Make predicition“ nakon čega se prikazuje rezultat predikcije.