Primjena mašinskog učenja u predikciji dijabetesa

1. Definicija problema

Dijabetes je hronično oboljenje koje često ostaje neprepoznato u ranoj fazi, što može dovesti do ozbiljnih zdravstvenih komplikacija. Trenutno, rani skrining zahtijeva medicinske preglede i laboratorijske testove, što može biti skupo ili nepristupačno za neke korisnike. Problem koji ovaj projekat rješava je **predviđanje prisustva dijabetesa kod osobe koristeći dostupne zdravstvene i demografske podatke**, bez potrebe za neposrednim medicinskim testovima. Projekat omogućava razvoj sistema koji može pomoći u ranom otkrivanju dijabetesa, podršci ljekarima i samoprocjeni korisnika.

Cilj projekta je napraviti sistem koji predviđa da li osoba ima dijabetes na osnovu zdravstvenih i demografskih podataka. Koriste se Random Forest, KNN i Gradient Boosting modeli, trenirani na dva javno dostupna skupa podataka. Rezultati i performanse modela prikazani su putem web aplikacije.

2. Motivacija

Rano otkrivanje dijabetesa je ključno, a mašinsko učenje može pomoći u identifikaciji rizičnih slučajeva. Povećava svijest o važnosti prevencije i može biti temelj za buduće zdravstvene aplikacije.

3. Skup podataka

Baze podataka su preuzete sa HuggingFace, korištene su dvije baze CDC Health Indicators i Diabetes Prediction Dataset.

- CDC Health Indicators - 243,532 instanci
  - **Ciljno obilježje**: *Diabetes_binary* - da li osoba ima dijabetes(“Diabetic”: 13.9%, “Non-diabetic”: 86.1%)
  - *BMI* (numeric) - indeks tjelesne mase.
  - *Age* (ordinal): starosna grupa (npr 18-24, 25-29...)
  - *GenHlth* (ordinal) - samoprocjena generalnog zdravstvenog stanja (Fair, Good, Poor, Very Good, Excellent).
  - *MenthHlth* (numeric) - broj dana u zadnjih 30 u kojima je mentalno zdravlje bilo loše.
  - *PhysActivity* (ordinal) - ocjena nivoa fizičke aktivnosti.
  - *Income* (ordinal) - kategorija prihoda.
  - *HighBP* (nominal) - da li osoba ima povišen krvni pritisak (da/ne).
  - *HighChol* (nominal) -da li osoba ima povišen holesterol (da/ne).
  - *Smoker* (nominal) - da li je osoba pušač.
  - *Stroke* (nominal) - da li je osoba ikada imala moždani udar.
  - *Sex* (nominal) - Pol
  - *Fruits* (nominal)- da li osoba konzumira voće.
  - *Veggies* (nominal) - da li osoba konzumira povrće.
  - *AnyHealthcare* (nominal) - da li osoba ima bilo kakav oblik zdravstvenog osiguranja.
  - *Education* (ordinal) - najviši stepen obrazovanja.

- *HeartDiseaseorAttack* (nominal) - da li imaju istoriju sa srčanim bolestima ili udarima.
- *NoDocbcCost* (nominal) - da li osoba nije mogla da posjeti ljekara zbog troškova.

- Diabetes Prediction Dataset - 100,00 instanci

  - **Ciljno obeležje**: *diabetes*- da li osoba ima dijabetes ("1": 3.9%, "0": 96.1%)
  - *Blood_glucose_leve*l (numeric) - nivo glukoze u krvi.
  - *Age* (ordinal) - starosna grupa.
  - *bmi* (numeric) - indeks tjelesne mase.
  - *HbA1c_level* (numeric) - nivo HbA1c (glikoziliranog hemoglobina), indikator dugoročne glikemije.
  - *Gender* (nominal) - pol
  - *Hypertension* (nominal) - da li osoba ima hipertenziju
  - *Heart_disease* (nominal) - da li osoba ima srčanu bolest
  - *Smoking_history* (nominal) - status pušenja (npr. current, never, former...).

4. Način pretprocesiranja podataka

4.1 Konverzija ordinalnih atributa:

- U *CDC Health Indicators* datasetu atributi poput *Age*, *GenHlth* i *Education* pretvoreni su u numeričke vrijednosti prema unaprijed definisanim mapiranjima.
- U *Diabetes Prediction Datasetu* starost (*age*) je mapirana u starosne grupe.

4.2 Rukovanje nedostajućim vrijednostima

- Za numeričke podatke korišten je *SimpleImputer** sa strategijom popunjavanja srednjom vrijednošću (*mean*).
- Za kategorijske podatke korišten je *SimpleImputer** sa strategijom popunjavanja najčešćom vrijednošću (*most frequent*).

Nakon testiranja, strategije sa najboljim perfomsama su upisane u fajl sa konstantama, te se lako mogu naknadno promjeniti.

4.3 Standardizacija numeričkih vrijednosti

- Numerički atributi (*BMI*, *MentHlth*, *HbA1c_level*, *blood_glucose_level*, itd.) standardizovani su pomoću *StandardScaler** kako bi se uklonile razlike u mjerilima.

4.4 Kodiranje nominalnih vrijednosti

- Nominalni atributi (npr. *Smoker*, *HighBP*, *Gender*, *Smoking_history*) kodirani su pomoću *OneHotEncoder** sa opcijom *drop="if_binary"* radi optimizacije broja kolona.

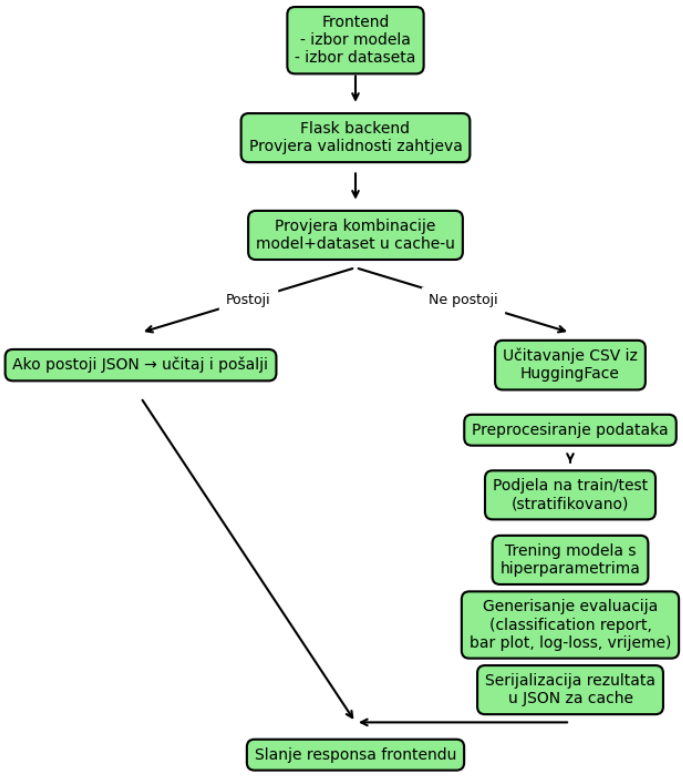
4.5 Podjela na karakteristike i ciljnu varijablu

- Iz skupa podataka je uklonjena ciljna kolona (*Diabetes_binary* ili *diabetes*) za formiranje ulazne matrice X.

*SimpleImputer, StandardScaler i OneHotEncoder pripadaju biblioteci scikit_learn.

5. Metodologija

Trenutni modeli koji su integrisani u projekat su KNN, Random Forest i Gradient Boosting. Projekat je kreiran tako da dodavanje novog modela zahtjeva minimalne promjene. Za određivanje hiperparametara je korišten *GridSearchCV,* koji je mjerio performanse na osnovu tačnosti. Nakon ovog testiranja hiperapametri koji su se pokazali kao optimalni u odnosu vremena i rezultata su postavljeni kao stalni pri treniranju modela.



6. Način evaulacije

Podaci su podeljeni na 80% train i 20% test (stratifikovano).
Evaluacija se vrši pomoću tačnosti i metrika iz *classification report*-a (*precision*, *recall*, *F1-score*). Prikazuju se i bar graf predikcija te log-loss i train/val loss grafikoni.

7. Tehnologije

Frontend: Angular, Backend: Python, Flask

Biblioteke: scikit-learn, numpy, matplotlib, huggingface_hub, pandas

8. Relevantna literatura

https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/

https://ai.plainenglish.io/diabetes-prediction-using-machine-learning-classification-approaches-a-capstone-project-by-team-cbd0b784a30b