Predikcija dijabetesa korišćenjem metoda mašinskog učenja

Definicija problema

Problem koji se rešava jeste klasifikacija pacijenata na dve klase: oboleli od dijabetesa i zdravi. Na osnovu medicinskih parametara (glukoza, pritisak, BMI itd.) algoritam predviđa da li osoba ima povećan rizik od dijabetesa.

Motivacija

Dijabetes je jedno od najrasprostranjenijih hroničnih oboljenja, a rano otkrivanje je ključno za sprečavanje komplikacija. Automatizovana predikcija može pomoći lekarima i pacijentima u donošenju odluka, omogućiti pravovremene mere i smanjiti troškove zdravstvene zaštite.

Skup podataka

Koristi se PIMA Indians Diabetes Dataset (UCI Machine Learning Repository).

- Broj instanci: 768 pacijenata
- Broj atributa: 8 ulaznih + 1 izlazni
- Atributi: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
- Ciljni atribut: Outcome (0 = nema dijabetes, 1 = ima dijabetes)
- Link: PIMA dataset
    - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Raspodela klasa: broj uzoraka sa Outcome=0 i Outcome=1, i procenti (navesti iz EDA).

Exploratorna analiza podataka (EDA)

Ciljevi: razumeti raspodele, kvalitet podataka, veze sa ciljem, outliere.
Istraživačka pitanja:

1. Koji atributi najviše razlikuju klase (top 3)?
2. Kako se stope pozitivnih menjaju po kvartilima Glucose/BMI/Age?
3. Da li nule u Insulin/SkinThickness predstavljaju “missing” i kako utiču na metrike?
4. Da li standardizacija poboljšava SVM/MLP na validaciji?

Provera kvaliteta: procenat nula/NaN po atributu; tretiraj 0 u Insulin/SkinThickness kao missing.
Univarijantno: histogram + box-plot za svih 8 atributa.
Bivarijantno: box-plotovi po klasama za Glucose, BMI, Age; bar-plot pozitivnih po kvartilima.
Korelacije: Pearson/Spearman heatmap; zabeleži parove $|r|>0.7$.
Outlieri: IQR pravilo; strategija: winsorizacija ili RobustScaler gde je opravdano.
Sažetak EDA (3–5 tačaka): kratko navedi nalaze i posledice po pretprocesiranje.

Isporuka EDA: 6–8 grafika (histogrami/box-plotovi, korelacijska mapa, 2–3 bivarijantna prikaza) + sažetak nalaza (3–5 tačaka).

Povezivanje EDA i treniranja

- Zaključci iz EDA se koriste kao ulaz u eksperimente.
- Ako se otkrije kolinearnost između atributa, testira se model i bez jednog od tih atributa.
- Ako se potvrdi da nule u *Insulin* i *SkinThickness* predstavljaju nedostajuće vrednosti, porede se različite strategije imputacije (median, KNN, ili model-bazirana imputacija).
- Ako se identifikuju značajni outlier-i, upoređuju se rezultati sa i bez njihovog uklanjanja ili winsorizacije.

Pretprocesiranje podataka

- Imputacija: median ili KNN za Insulin i SkinThickness (0 tretirati kao missing).
- Standardizacija: StandardScaler nad svim ulazima.
- Podela podataka: stratified 80/20 (fiksiran random_state).
- Napomena: ako je klasa neuravnotežena, za LR/SVM koristiti class_weight='balanced'.

Metodologija

Modeli (baseline): Logistic Regression, Random Forest, SVM.
Kompleksniji model: MLP (scikit-learn)

• MLPClassifier(hidden_layer_sizes=(32,16), activation='relu', early_stopping=True, max_iter=200)

• Mali grid: hidden_layer_sizes ∈ {(32,16), (64,32)}, alpha ∈ {1e-4, 1e-3}, learning_rate_init ∈ {0.001, 0.01}

• Ulaz obavezno standardizovan (StandardScaler). Early stopping na validacionom skupu.

Kompleksniji model: Transformer

- Pored MLP-a, koristi se i Transformer-bazirani model za tabularne podatke (npr. TabTransformer iz biblioteke HuggingFace).
- Moguće primene:
    1. **Glavni klasifikacioni model** – poređenje sa LR, RF, SVM i MLP.
    2. **Detekcija outlier-a** – treniranje varijante modela za identifikaciju atipičnih pacijenata.
    3. **Imputacija podataka** – predikcija i popunjavanje nedostajućih vrednosti (posebno *Insulin* i *SkinThickness*).
- Evaluira se istim protokolom (5-fold CV, ROC-AUC kao primarna metrika).

Ulaz: medicinski parametri jedne osobe (8 vrednosti)
Izlaz: klasifikacija (0 = nema dijabetes, 1 = ima dijabetes)

Dijagram procesa:

```
Podaci → Pretprocesiranje → Trening modela → Evaluacija → Predikcija novog unosa
```

Način evaluacije

- **Podela podataka:** Stratified train/test = 80/20, fiksiran random_state.
- **Validacija:** Na train delu sprovodim **stratified 5-fold CV** za sve modele (LR, RF, SVM, MLP).
- **Transformer** model se evaluira istim protokolom (5-fold CV, ROC-AUC kao primarna metrika) radi poređenja sa ostalim algoritmima.
- **Metrike (na CV i na testu):**

    - Accuracy, Precision, Recall, F1
    - **ROC-AUC (obavezno)**
    - **PR-AUC (preporučeno kod neuravnoteženih klasa)**
    - Confusion matrix

- **Izbor modela:** prema **prosečnom ROC-AUC** iz 5-fold CV; test metrika se prijavljuje samo jednom na kraju.
- **Prag odlučivanja:** podešavam prag na validacionom skupu radi ciljane osetljivosti (npr. Recall ≥ 0.80), pa taj prag primenjujem na testu.
- **Izveštaj:** tabela metrika za sve modele + ROC kriva za najbolji model; opcionalno PR kriva.
- Primarna metrika za izbor modela: ROC-AUC iz 5-fold CV na train-u.
- Reproduktivnost: fiksiran random_state/seed za sve procedure.

Tehnologije

- Python
- scikit-learn (klasifikacioni algoritmi, evaluacija)
- Pandas, NumPy (obrada podataka)
- Matplotlib/Seaborn (vizualizacija)
- scikit-learn MLP
- Streamlit (jednostavan UI za unos podataka i prikaz predikcije) ili alternativno Django/Flask (za izradu web forme i prikaz rezultata predikcije)

Relevantna literatura

- UCI Repository: PIMA dataset
- Slični projekti i tutorijali:
    - GeeksForGeeks: Diabetes Prediction ML Project
    - https://www.geeksforgeeks.org/python/diabetes-prediction-machine-learning-project-using-python-streamlit/
    - Medium članci o predikciji dijabetesa ML metodama
- Dokumentacija biblioteka: scikit-learn, Pandas, NumPy