Predikcija dijabetesa kod pacijenata pomoću mašinskog učenja

Student: Milica Bosančić SV60/2022

Predmet: Računarska inteligencija

1. Naziv teme

Predikcija dijabetesa primenom metoda mašinskog učenja

2. Definicija problema

Cilj projekta je izgradnja modela mašinskog učenja koji predviđa da li pacijent ima povećan rizik od obolevanja od dijabetesa na osnovu medicinskih parametara.

Radi se o problemu **binarne klasifikacije** gde su ulaz podaci o pacijentu (godine, indeks telesne mase, krvni pritisak, nivo glukoze, nivo insulina itd.), a izlaz je odluka da li pacijent spada u grupu sa rizikom od dijabetesa (1) ili ne (0).

3. Motivacija

Dijabetes je jedno od najčešćih hroničnih oboljenja i predstavlja ozbiljan zdravstveni problem širom sveta. Pravovremeno otkrivanje osoba koje imaju povećan rizik može omogućiti preventivne mere i smanjiti broj obolelih, a samim tim i smanjiti opterećenje zdravstvenih sistema.

Model za predikciju dijabetesa može se koristiti u preventivnim pregledima, kao podrška lekarima pri dijagnostici i kao deo aplikacija za praćenje zdravlja. Time bi se omogućilo:

- rano otkrivanje rizičnih pacijenata,
- ušteda resursa zdravstvenih sistema,
- podizanje svesti kod populacije koja je u riziku.

4. Skup podataka

Za rešavanje problema koristi se **Pima Indians Diabetes Database**, dostupan na Kaggle-u i UCI repozitorijumu.

- Link: Kaggle – Pima Indians Diabetes Database
- Broj instanci: 768
- Broj atributa: 8 numeričkih atributa + ciljni atribut
- Atributi:

    - **Pregnancies** – broj trudnoća
    - **Glucose** – nivo glukoze u krvi
    - **BloodPressure** – krvni pritisak
    - **SkinThickness** – debljina kože
    - **Insulin** – nivo insulina
    - **BMI** – indeks telesne mase
    - **DiabetesPedigreeFunction** – porodična predispozicija za dijabetes
    - **Age** – starost pacijenta
    - **Outcome** (ciljno obeležje): 0 = nema dijabetesa, 1 = dijabetes
- Raspodela klasa: oko 65% instanci bez dijabetesa, oko 35% sa dijabetesom.

5. Način pretprocesiranja podataka

- Provera i tretiranje nelogičnih vrednosti (npr. nula za BMI ili krvni pritisak)
- Skaliranje numeričkih atributa (standardizacija ili min-max scaling)
- Analiza korelacije i eventualno uklanjanje redundantnih atributa
- Podela podataka na **train/validation/test** (70/15/15)

6. Metodologija

Modeli:

- Naivni Bajes (Naive Bayes)
- K-Nearest Neighbors (KNN)
- Logistic Regression
- Nauronska mreza MLP

Koraci:

1. Učitavanje i analiza dataset-a
2. Pretprocesiranje podataka (čišćenje, skaliranje)
3. Treniranje modela na **train skupu**
4. Evaluacija na **validation/test skupu**
5. Poređenje modela po metriki (accuracy, F1-score, ROC-AUC)

Ulaz modela: numerički atributi pacijenata

Izlaz modela: predikcija da li pacijent ima dijabetes (0/1)

7. Način evaluacije

- Podela podataka na **train / validation / test** skupove.
- Metrike koje se koriste:
    - Accuracy (tačnost)
    - F1-score (harmonična sredina preciznosti i osetljivosti)
    - ROC-AUC kriva za poređenje modela

8. Tehnologije

- **Python** (glavni jezik)
- **scikit-learn** (klasifikacioni modeli, evaluacija)
- **Pandas, NumPy** (obrada podataka)
- **Matplotlib, Seaborn** (vizualizacija podataka i rezultata)

9. Relevantna literatura

- Skup podataka:
    - UCI Machine Learning Repository – Pima Indians Diabetes Database
- Naučni radovi i slični projekti:
    - "Diabetes Prediction using Machine Learning" – IEEE paper: link
- Slični radovi:
    - R. S. Bharti, et al., *Machine Learning Techniques for Diabetes Prediction*, 2020
    - J. Smith, *Comparative Analysis of Diabetes Classification Models*, 2019