# Predikcija sklonosti ka kardiovaskularnim poremećajima

Anđela Ristić, SV67-2022

### **1. Pun naziv teme**

“Predikcija sklonosti ka kardiovaskularnim poremećajima na osnovu hematoloških i biohemijskih parametara krvi primenom mašinskog učenja”

### **2. Definicija problema**

Kardiovaskularne bolesti (KVB), uključujući srčani udar, moždani udar, hipertenziju i koronarnu bolest srca, vodeći su uzrok smrtnosti globalno. Rana identifikacija osoba sa povećanim rizikom može značajno smanjiti smrtnost i komplikacije. Laboratorijski podaci krvi i biohemijski parametri omogućavaju rano prepoznavanje rizika, ali su često kompleksni i međusobno povezani, što zahteva primenu mašinskog učenja za preciznu klasifikaciju pacijenata u nisku, srednju i visoku verovatnoću oboljenja.

Ključni aspekti problema:

* Podaci iz laboratorije (krv, biohemija) su lako dostupni i rutinski prikupljeni;
* Predikcija pre pojave kliničkih simptoma;
* Omogućavanje preventivne medicine i personalizovanog praćenja pacijenata.

### **3. Motivacija**

Kardiovaskularne bolesti predstavljaju vodeći uzrok smrti i invaliditeta širom sveta, što ih čini jednim od najvažnijih zdravstvenih problema današnjice. Projekat predviđa značajnu praktičnu i društvenu vrednost jer omogućava rano otkrivanje pacijenata sa visokim rizikom, čime se smanjuje mogućnost nastanka ozbiljnih komplikacija poput infarkta ili moždanog udara.

Pored toga, pravovremena identifikacija rizičnih osoba može doprineti značajnom smanjenju zdravstvenih troškova, dok lekarima pruža dodatni alat za donošenje informisanih odluka i efikasno praćenje stanja pacijenata. Projekat takođe omogućava razvoj veština u primeni mašinskog učenja na medicinske podatke, čime doprinosi akademskom i profesionalnom usavršavanju.

### **4. Skup podataka**

Za projekat koristim javni dataset: [Cardiovascular Disease Dataset](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset), preuzet sa Kaggle. Karakteristike ds:

~70.000 pacijenata iz Rusije

Raspodela klasa: ~65–70% zdravih, ~30–35% sa KVB

Prednost: veliki broj instanci, pogodan za neuronske mreže i klasifikaciju

Atributi

1. Demografski:

age – starost (u danima)

gender – pol (1 = muškarac, 2 = žena)

2. Faktori rizika i vitalni parametri:

height – visina

weight – telesna masa

BMI – indeks telesne mase

ap\_hi – sistolni krvni pritisak

ap\_lo – dijastolni krvni pritisak

smoke – pušač (0 = ne, 1 = da)

alco – konzumiranje alkohola (0 = ne, 1 = da)

active – fizička aktivnost (0 = ne, 1 = da)

3. Biohemijski parametri:

cholesterol – nivo holesterola (1 = normalno, 2 = povišeno, 3 = visoko)

glucose – nivo glukoze u krvi (CDD – gluc)

4. Ciljni atribut:

cardio – prisustvo kardiovaskularne bolesti (0 = nema, 1 = ima)

### **5. Način pretprocesiranja podataka**

Planiram sledeće korake:

1. Standardizacija i konverzija atributa

* Kategorijski atributi (gender, smoke, alco, active, cholesterol, glucose) biće uniformno kodirani.

1. Rukovanje nedostajućim vrednostima

* Za numeričke parametre – imputacija medijanom ili prosekom.
* Za kategorijske parametre – imputacija najčešćom vrednošću.
* Nedostajući ciljni atribut (cardio) – instanca se odbacuje.

1. Skaliranje numeričkih atributa

* Parametri kao što su height, weight, ap\_hi, ap\_lo, glucose će biti standardizovani (z-score) ili min-max skalirani, zavisno od modela.

1. Balansiranje klasa

* cilj je omogućiti modelu da ne bude pristrasan prema većinskoj klasi.

### **6. Metodologija**

Za projekat ću koristiti modele Logistic Regression, Random Forest, (X)GBoost i neuronske mreže. Ulazni podaci su demografski, faktori rizika i biohemijski parametri, dok je izlaz klasifikacija pacijenata u nisku, srednju i visoku verovatnoću KVB. Pretprocesiranje uključuje standardizaciju, imputaciju nedostajućih vrednosti, kodiranje kategorijskih atributa i balansiranje klasa. Hiperparametri će se optimizovati korišćenjem GridSearchCV, a interpretacija modela će biti obavljena preko feature importance ili SHAP vrednosti.

### **7. Način evaluacije**

Dataset će biti podeljen na train (70%), validation (15%) i test (15%) skupove. Trening skup služi za obučavanje modela, validacioni skup za optimizaciju hiperparametara, test skup za finalnu evaluaciju performansi modela. Evaluacija će se vršiti korišćenjem tačnosti, F1-score, precision, recall i ROC-AUC, kako bi se pratila sposobnost modela da razlikuje pacijente sa i bez KVB.

### **8. Tehnologije**

Projekat će se realizovati u Pythonu, koristeći Pandas i NumPy za obradu podataka, Scikit-learn za Logistic Regression i Random Forest, XGBoost za boostovane modele, te TensorFlow/PyTorch za neuronske mreže. Imbalanced-learn će se koristiti za balansiranje klasa, dok će Matplotlib i Seaborn služiti za vizualizaciju rezultata.

### **9. Relevantna literatura**

D. R. Mookadam et al., “Prediction of Cardiometabolic Risk Using Machine Learning,” Frontiers in Cardiovascular Medicine, 2022.

[M. K. Alaa et al., “Machine learning for risk prediction in cardiometabolic disease,” PLoS One, 2020.](https://www.researchgate.net/publication/378277461_Machine_Learning_for_Risk_Prediction_of_Cardiovascular_Disease_Current_Advances_and_Future_Prospects)

[https://www.stetoskop.info/bolesti-srca-i-krvnih-sudova-kardiologija/kardiolog-o-prevenciji-kardiovaskularnih-bolesti#:~:text=Promenjivi%20faktori%20rizika%20su%20pu%C5%A1enje,faktor%20i%20pripadnost%20mu%C5%A1kom%20polu.](https://www.stetoskop.info/bolesti-srca-i-krvnih-sudova-kardiologija/kardiolog-o-prevenciji-kardiovaskularnih-bolesti#:~:text=Promenjivi%20faktori%20rizika%20su%20pu%C5%A1enje,faktor%20i%20pripadnost%20mu%C5%A1kom%20polu)