# ML18-Z3-code10: Izveštaj

## Pristup zadatku

Zadatak je urađen u Python-u uz biblioteke Pandas, Matplotlib i Scikit-Learn.
Najpre smo plotovali zadate podatke. S obzirom na veliku dimenzionalnost problema, plotovanjem nismo dobili posebno korisne informacije.
Podaci su pre obučavanja modela normalizovani, s tim što nismo normalizovali kolonu koja označava klasu (*NoYes*) s obzirom na to da su njene jedine vrednosti 0 i 1.
Tačnost modela utvrđivali smo korišćenjem unakrsne validacije, a za splitovanje trening podataka koristili smo *RepeatedStratifiedKFold* iz *scikit-learn* bibilioteke.

## Isprobani algoritmi

### Nedostajuće vrednosti

Nedostajuće vrednosti trening skupa smo pokušali da rešimo na više načina:

1. Izbacivanje 32 uzorka sa nedostajućim vrednostima
2. Zamena nedostajućih vrednosti sa *mean*, *median* ili *most_frequent* vrednostima (*scikit-learn* implementacija)
3. Predikcija nedostajućih vrednosti korišćenjem KNN (preuzeto sa scikit-learn imputation by knn)

### Odabir obeležja

Selekciju obeležja smo radili korišćenjem linearnog SVM-a sa L1 regularizacijom i izbačena je *Trees* kolona.

### SVM

Isprobali smo linearni, polinomijalni, Sigmoidni i Gausov kernel. Ručnim isprobavanjem nam se kao najpreciznji pokazao Gausov kernel. Kako bi smo utvrdili najbolje parametre ovog modela, radili smo *grid search* pretragu. Početne parametre za ovu pretragu smo preuzeli iz ovog rada. Najbolje rezultate ove pretrage smo posmatrali kao lokalne maksimume, pa smo zatim korigovali opseg i korak pretrage da traži bolja rešenja u okolini tih lokalnih maksimuma.

# Konačno rešenje

Kao najbolje rešenje pokazao se SVM sa Gausovim kernelom sa parametrima C=72 i gamma=9.3 uz izbacivanje uzoraka sa nedostajućim vrednostima i izbacivanjem *Trees* kolone.

# Ostvareni rezultati

Konačna tačnost na test skupu iznosi 0.940520446096654.