Prva domača naloga

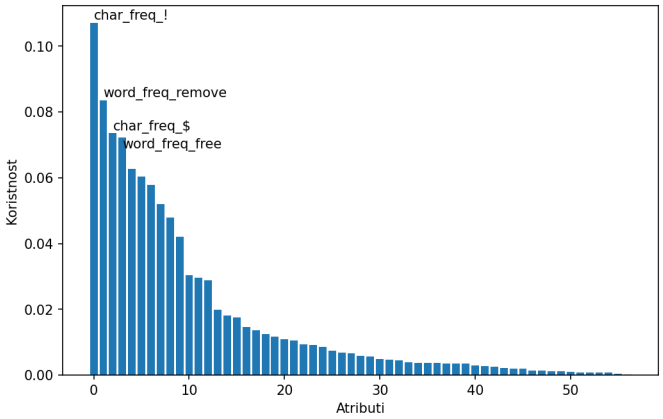
Tine Markočič

Maj 2023

1 Model

Nalogo sem reševal v Pythonu, s pomočjo knjižnice scikit-learn. Podatke sem razdelil na učno in testno množico v razmerju $4 : 1$ in na njih preizkusil različne algoritme za klasifikacijo. Za dobrega se je izkazal algoritem naključnih gozdov z natančnostjo $94,90\%$.

Hkrati me je zanimalo tudi, kateri atributi so pri tej klasifikaciji najpomembnejši. To sem izvedel iz atributa *feature_importances_*, urejene koristnosti pa sem predstavil v spodnjem grafu.

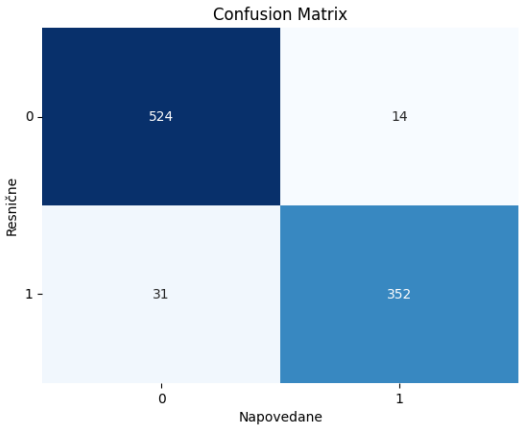


Slika 1: Koristnost atributov

Najpomembnejši štirje atributi so pogostost znakov ! in $ ter pogostost besed free in remove. Gre pričakovati, da se vsi štirje atributi veliko pogosteje pojavijo v spam pošti, kar pomeni, da bi model moral v splošnem znati dobro ločiti med vsiljeno in želeno pošto.

2 Izboljšave

Model sem želel še dodatno izboljšati. Poleg natančnosti je pomembno da model ne klasificira navadne pošte pod vsiljeno. Z drugimi besedami, poskusil sem zmanjšati število lažno pozitivnih primerov. Rezultati napovedi izhodiščnega modela so predstavljene spodaj v tako imenovani matriki zmede. Lažno pozitivnih je bilo 14 primerov.



Slika 2: Matrika zmede

Z metodo GridSearchCV sem poskusil najti vrednosti hiperparametrov *max_depth*, *min_samples_leaf* in *class_weight*, pri katerih bi bil model najbolj precizen. Žal pa pri tem nisem bil uspešen, saj je optimalen model na testni množici priredil 15 lažno pozitivnih primerov.

Z isto metodo sem želel izboljšati še natančnost modela. Tokrat sem izbral hiperparametra *n_estimators* in *criterion*. Natančnost optimalnega modela pa se je na testni množici izboljšala na $95,11\%$.