2021/2022

Inteligentna analiza podatkov

MODELI IN ODLOČITVENI SISTEMI

SAŠA VINČIĆ

1 Uvod

Kot projektno nalogo sem implementiral Naive Bayes model odločanja v Python-u in ga ovrednotil z k-kratno križno validacijo nad podatkovno zbirko s temo »Rak na Dojki«. Uporabil sem tudi orodje Orange3, da sem na isti podatkovni zbirki uporabil še nekaj drugih klasifikacijskih algoritom. Na koncu pa sem pridobljene podatke primerjal.

2 Naive Bayes algoritem

Algoritem, ki sem ga implementiral deluje tako, da se najprej prebere csv datoteka, ki se oblikuje in s pomočjo zbirke »Pandas« spravi v strukturirano podatkovno zbirko. Nato se ta podatkovna zbirka razbije v n-število foldov (sam sem algoritem izvajal z 5-imi foldi). Eden izmed foldov se uporabi kot testna zbirka, ostali pa kot učna zbirka. Nad učno zbirko se izvaja učenje, nato pa nad testno zbirko algoritem izvaja predikcije. Rezultati se nato obdelajo in vrnejo se ocenitvene metrike modela, ter matrika zmede. Omenjen postopek se izvede za število foldov, tako, da se vsak fold enkrat uporabi za testiranje.

Algoritem je z eno majhno izjemo dinamičen in bi z majhnimi prilagoditvami deloval za vsako podatkovno zbirko, če bi ta ustrezala naslednjim pogojem:

1. Podana v csv formatu,
2. klasifikacijski atribut je binaren in je na prvem mestu
3. atributi so kategorični in ne numerični.

3 Podatkovna zbirka

Podatkovna zbirka, ki sem jo izbral vsebuje 286 instanc in 9 atributov. 201 instanca pripada enemu razredu, preostalih 85 pa drugemu. Tema podatkovne zbirke je medicinska in vsebuje kr nekaj znanstvenih terminov. Gre pa za klasifikacijo žensk, ki bolehajo za rakom na dojki.

Zbirka ne vsebuje manjkajočih podatkov.

4 Ovrednotenje algoritmov

4.1 Ovrednotenje implementiranega algoritma

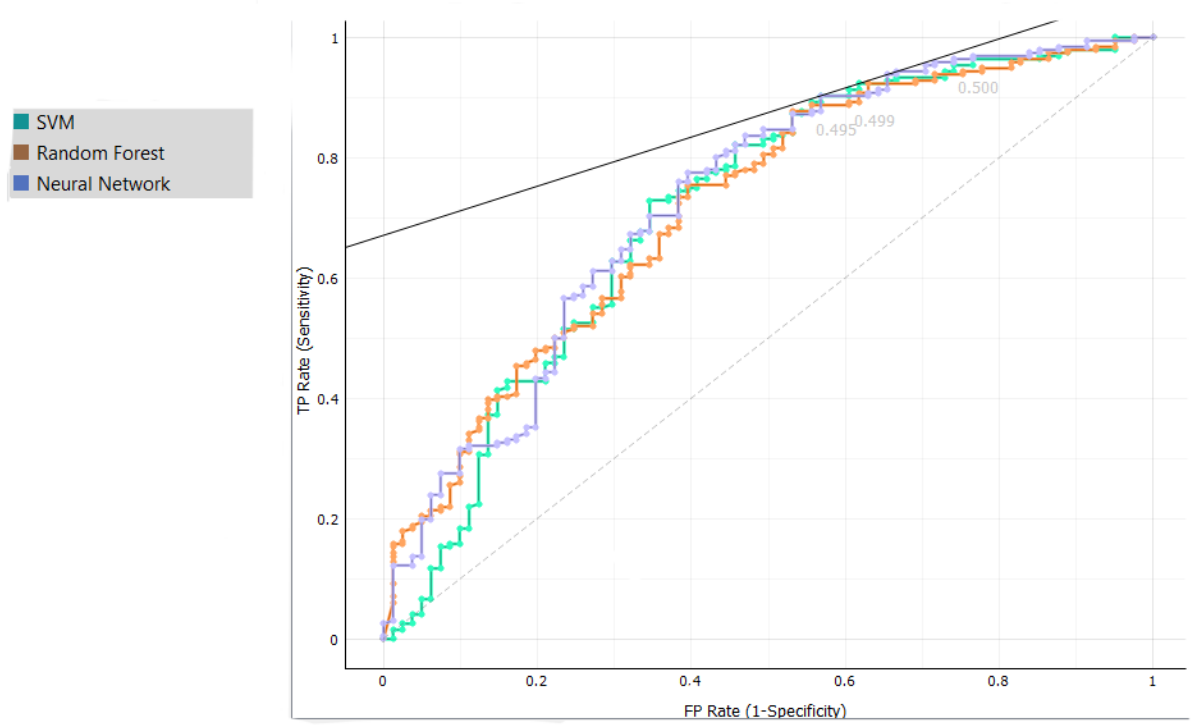
Povprečje 5-ih foldov:

- Točnost: 0.73
- Senzitivnost: 0.85
- Specifičnost: 0.58
- AUC: 0.66
- Preciznost: 0.56
- Recall: 0.5
- F-mera: 0.55

4.2 Ovrednotenje Orange3 algoritmov

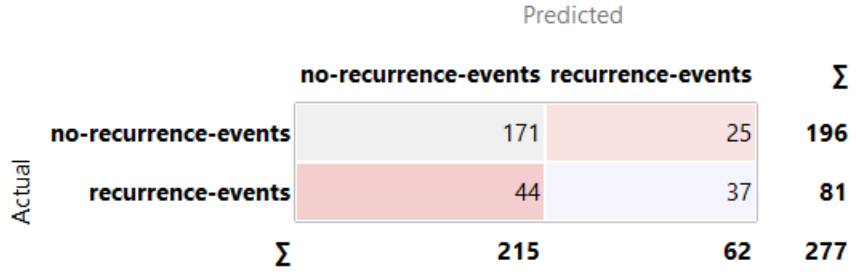
Model	AUC	CA	F1	Precision	Recall	Specificity
Neural Network	0.729	0.751	0.740	0.737	0.751	0.578
Random Forest	0.720	0.747	0.731	0.730	0.747	0.548
SVM	0.714	0.740	0.688	0.727	0.740	0.429

Slika 1: Metrike algoritmov iz orodja Orange3

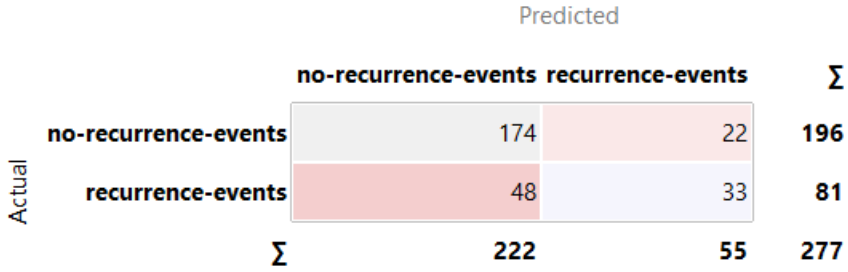


Slika 2: Krivulja ROC, za algoritme iz orodja Orange3

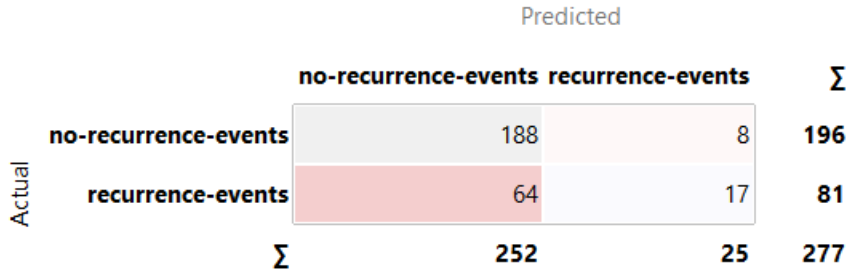
4.2.1 Matrike zmede Orange3 algoritmov



Slika 3: Matrika zmede algoritma Neural Network

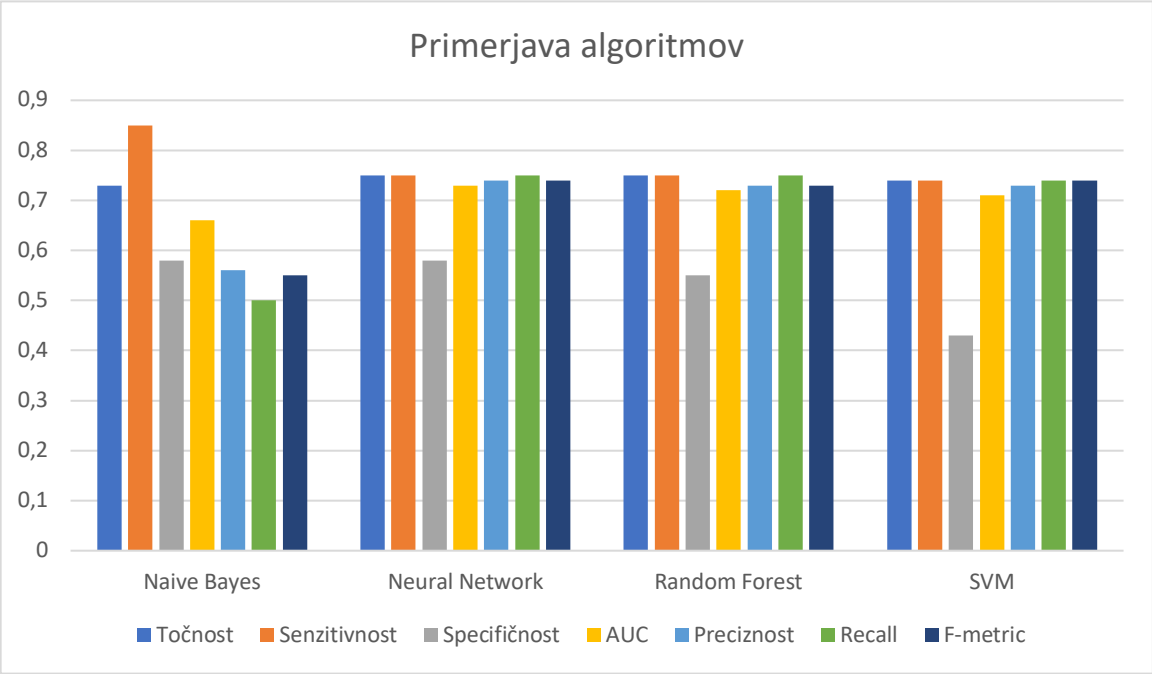


Slika 4: Matrika zmede algoritma Random Forest



Slika 5: Matrika zmede algoritma SVM

4.3 Primerjava algoritmov



Algoritmi so si precej blizu, posebej po točnosti. Očitno je da je orodje Orange3 v vseh treh algoritmih izvedlo bolj natančne »predikcije«. Vidna je superiornost že implementiranih algoritmov v večini matrik. Zaradi neuravnovesene zastopanosti instanc iz obeh razredov, je vidna tudi precej nizka specifičnost in visoka senzitivnost. Mislim, da bi za boljšo ovrednoteneje algoritmov lahko izbral kakšno bolj uravnoveseno podatkovno zbirko, saj je ta imela več kot 75% instanc enega razreda. To neravnovesje je vidno v količini »False-positive« predikcij, saj se algoritmi niso uspeli dovolj dobro naučiti značilnosti za razred, ki je bil v manjšini.

5 Zaključek

Algoritem, ki sem ga implementiral se je odrezal slabše kot tisti, ki so na voljo v orodju Orange3, kar je bilo za pričakovati. Sem mnenja, da bi lahko modele boljše ovrednotili nad kakšno drugo podatkovno zbirko. Nevem pa, če bi to povečalo razlike med mojim ter ostalimi modeli ali bi se te razlike zmanjšale. Kljub vsemu je bila projektna naloga koristna, teorijo, ki sem se jo naučil v sklopu predmeta sem uspel tudi praktično uporabiti.