Katedra za Signale i sisteme
Elektrotehnički fakultet
Univerzitet u Beogradu

13M051MU - Mašinsko učenje

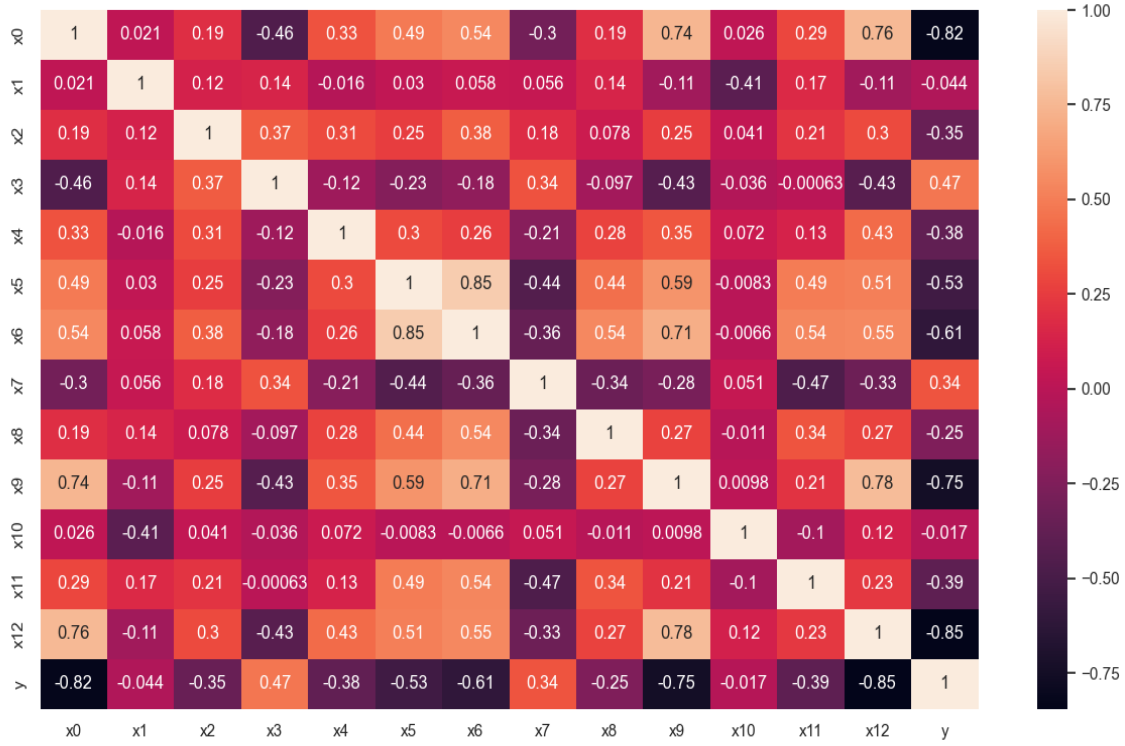
Domaći zadatak 4 - Izbor odlika. Stabla. Ansambli.

Aleksa Janjić 2023/3085

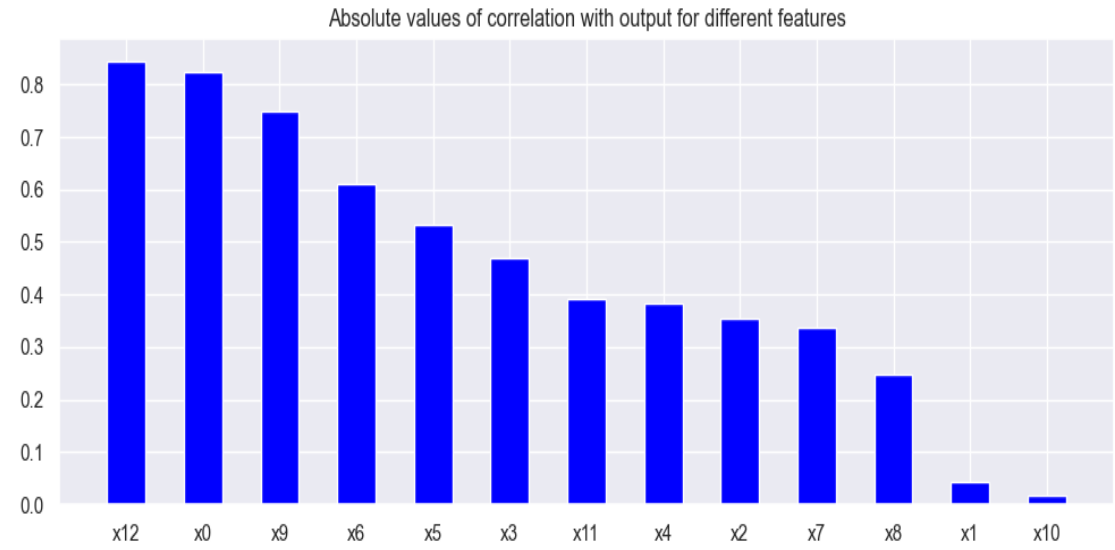
1 Zadatak 1 - Izbor prediktora

Na raspolaganju je skup podataka sa 13 obeležja i jednim izlazom. Cilj prvog dela zadatka je da odredimo koje obeležje je najinformativnije kada je u pitanju njegova povezanost sa izlazom, što se određuje nalaženjem koeficijenta korelacije između svakog prediktora i izlaza gde se za najinformativniji prediktor bira onaj čiji je koeficijent korelacije po apsolutnoj vrednosti najveći. Na slikama 1 i 2 ispod prikazani su matrica korelacije, kao i dijagram apsolutne vrednosti korelacije prediktora sa izlazom. Na osnovu priloženog, možemo zaključiti da su dva najinformativnija obeležja x_{12} i x_0.

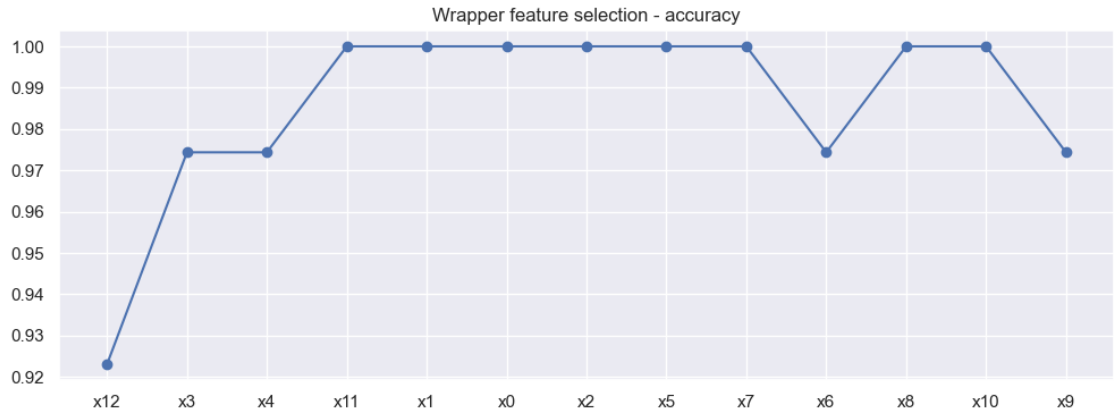
Informativnost obeležja moze se odrediti i na drugi način pomoću logističke regresije i "omotač" algoritma. Ulazne podatke delimo na trening i validacioni skup gde ćemo trenirati logističku regresiju na svakom obeležju pojedinačno i uzeti ono obeležje koje ima najveću tačnost na validacionom skupu. Nakon određivanja najinformativnijeg obeležja, nastavlja se pronalaženje drugog obeležja posmatranjem svake moguće kombinacije prvo izabranog obeležja i ostalih obeležja i uzima se kombinacija koja ima najveću tačnost na validacionom skupu i tako dalje. Dva najinformativnija obeležja, na osnovu slike 3 možemo zaključiti da su to obeležja x_{12} i x_3.



Slika 1: Matrica korelacije prediktora i izlaza



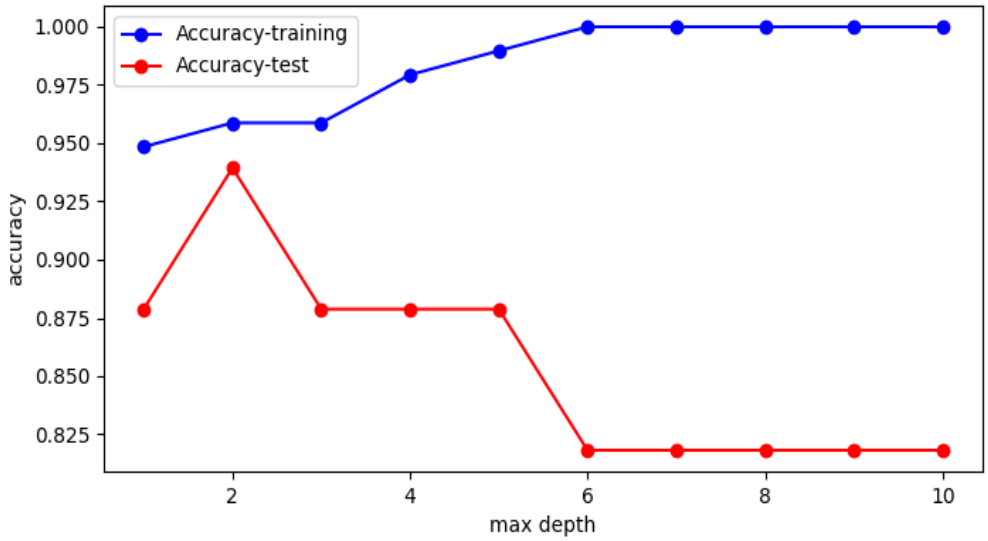
Slika 2: Dijagram apsolutnih vrednosti korelacije prediktora sa izlaznom promenljivom



Slika 3: Zavisnost tačnosti klasifikatora logističke regresije - ”omotač” algoritam selekcije obeležja

2 Zadatak 2 - Obučavanje stabla

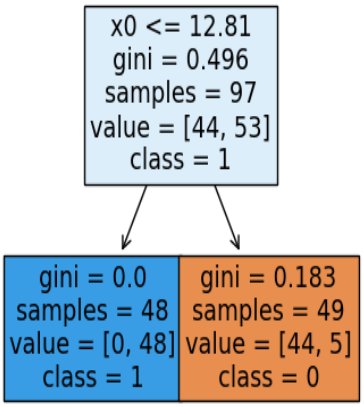
Drugi zadatak predstavlja obučavanje stabala. Za obučavanje stabala koristi se ugrađena funkcija *sklearn.tree.DecisionTreeClassifier*, a cilj zadatka je da se izvrši obučavanje stabala različitih dubina i da vidimo kako se takva stabla ponašaju na test skupu i kolike su njihove tačnosti, kao i da li se ispoljavaju efekti podobučavanja i preobučavanja.



Slika 4: Zavisnost tačnosti stabla odlučivanja za različite vrednosti dubine stabla

Kao što se može primetiti, posmatrajući grafik zavisnosti tačnosti stabla odlučivanja za različite vrednosti dubine stabla može se zaključiti da sa povećanjem dubine stabla tačnost na trening skupu raste, ali se tačnost na test skupu smanjuje iza vrednosti dubine stabla 2, što znači da je došlo do preobučavanja, tako da bi se za idealnu dubinu stabla mogla izabrati veličina 2, jer za veće dubine model loše generalizuje zbog efekta preobučavanja.

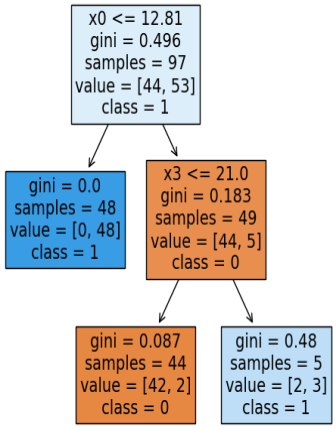
2.1 Podobučavanje



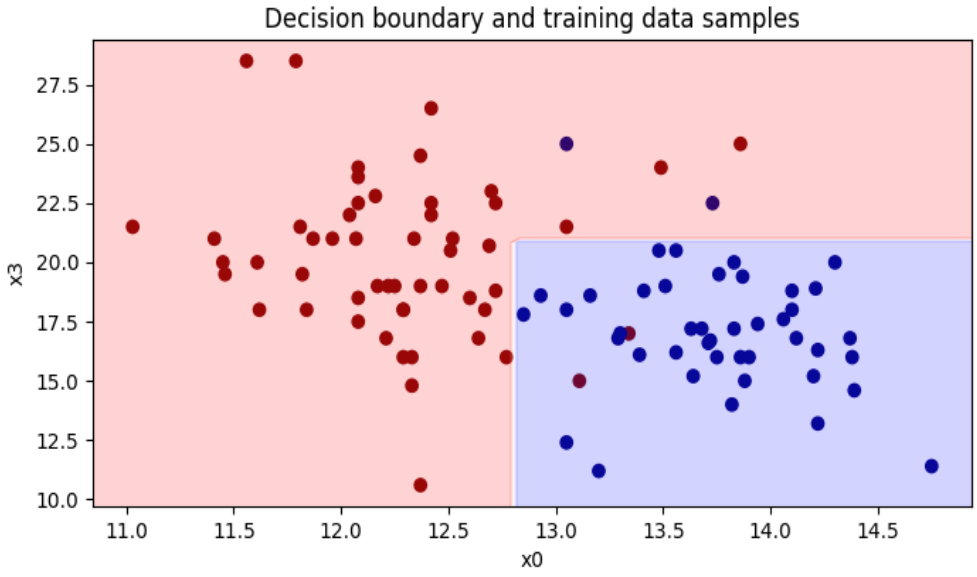
Slika 5: Stablo odlučivanja - dubina 1

Slika 6: Granica odluke i trening podaci - dubina 1

2.2 Optimalno obučavanje

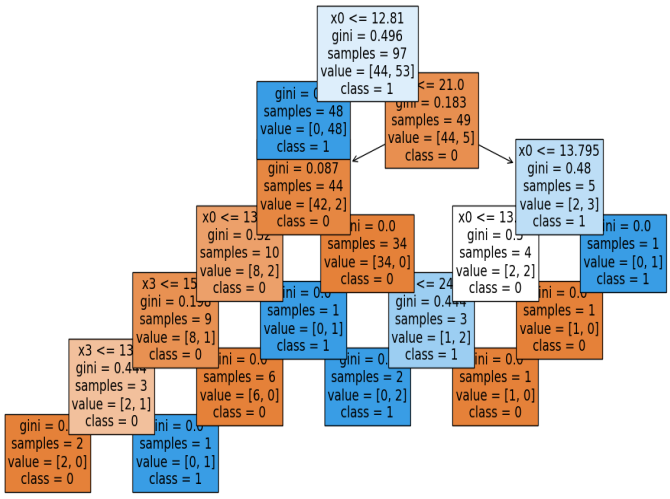


Slika 7: Stablo odlučivanja - dubina 2

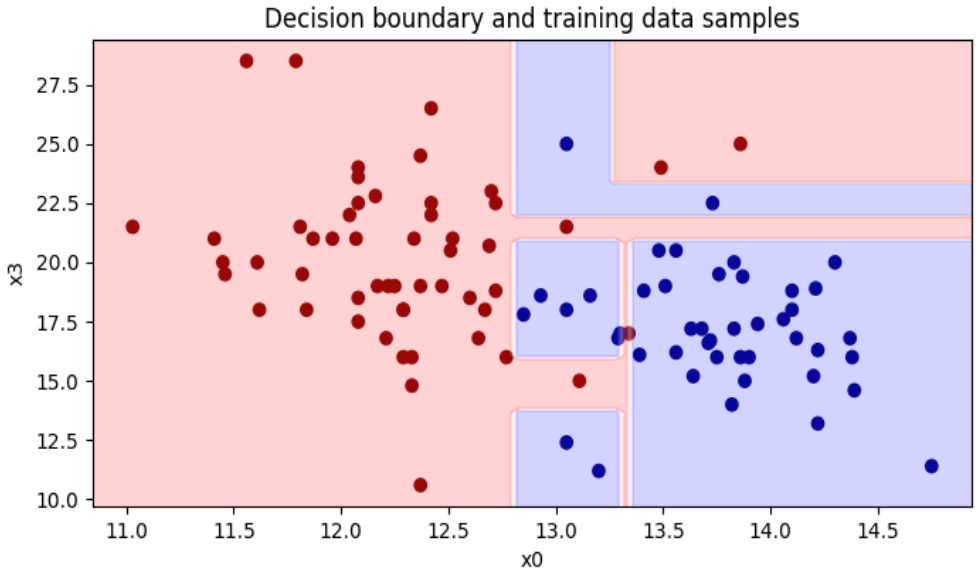


Slika 8: Granica odluke i trening podaci - dubina 2

2.3 Preobučavanje



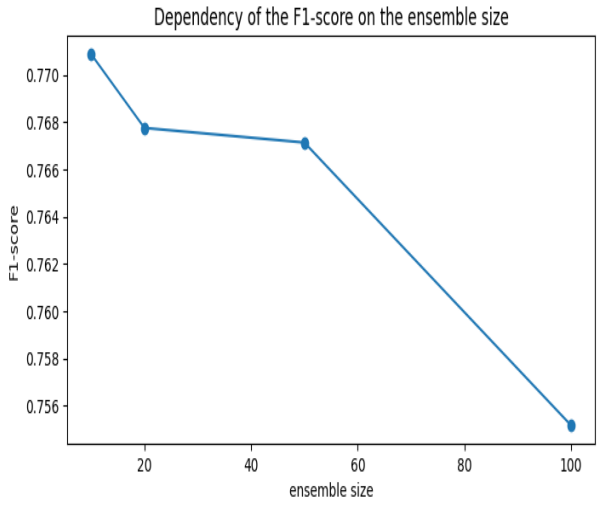
Slika 9: Stablo odlučivanja - dubina 6



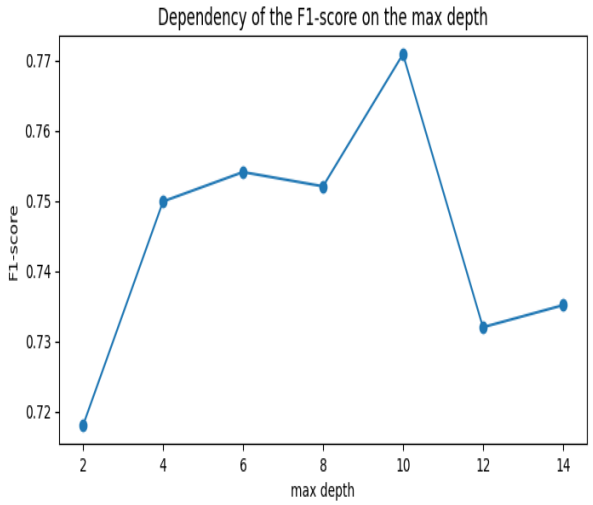
Slika 10: Granica odluke i trening podaci - dubina 6

3 Zadatak 3 - Ansambli

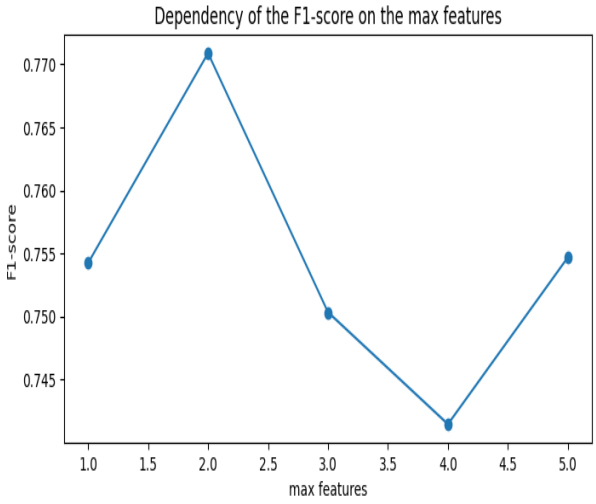
3.1 Random Forest



(a) Zavisnost F_1-skora od veličine ansambla



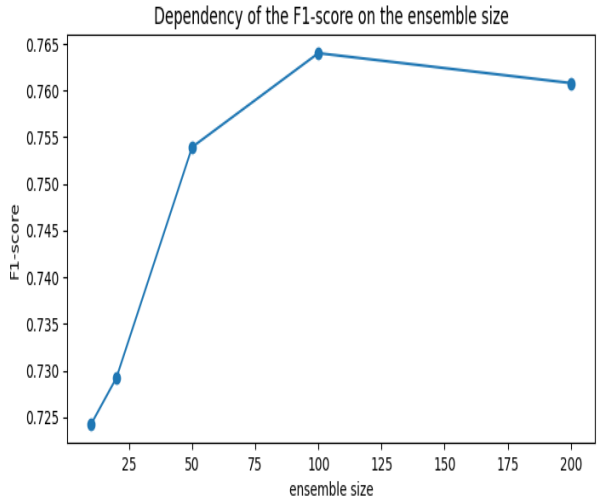
(b) Zavisnost F_1-skora od maksimalne dubine



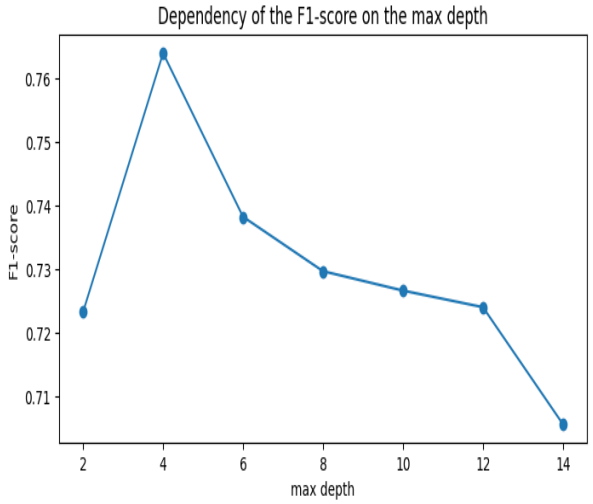
(c) Zavisnost F_1-skora od maksimalnog broja odlika

Slika 11: Performanse **Random Forest** algoritma izražene preko F_1-skora u zavisnosti od veličine ansambla, maksimalne dubine i maksimalnog broja odlika

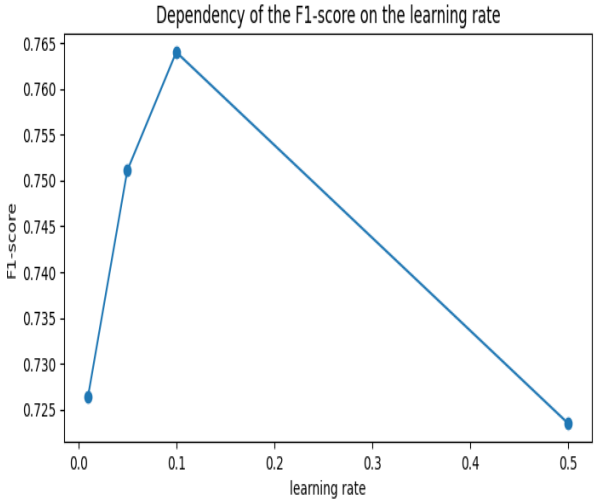
3.2 Gradient Boosting



(a) Zavisnost F_1-skora od veličine ansambla



(b) Zavisnost F_1-skora od maksimalne dubine



(c) Zavisnost F_1-skora od konstante učenja

Slika 12: Performanse **Gradient Boosting** algoritma izražene preko F_1-skora u zavisnosti od veličine ansambla, maksimalne dubine i konstante učenja