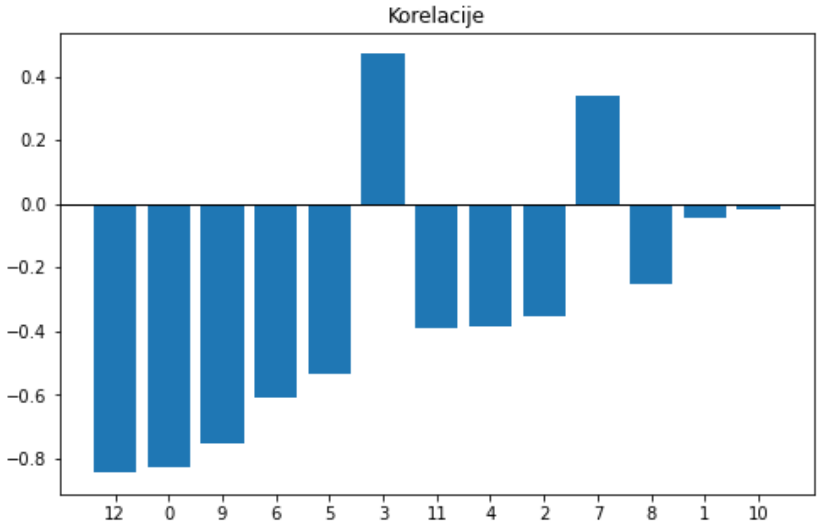
Mašinsko učenje – Domaći 4
Selekcija obeležja i ansambli

Nemanja Saveski 2023/3163

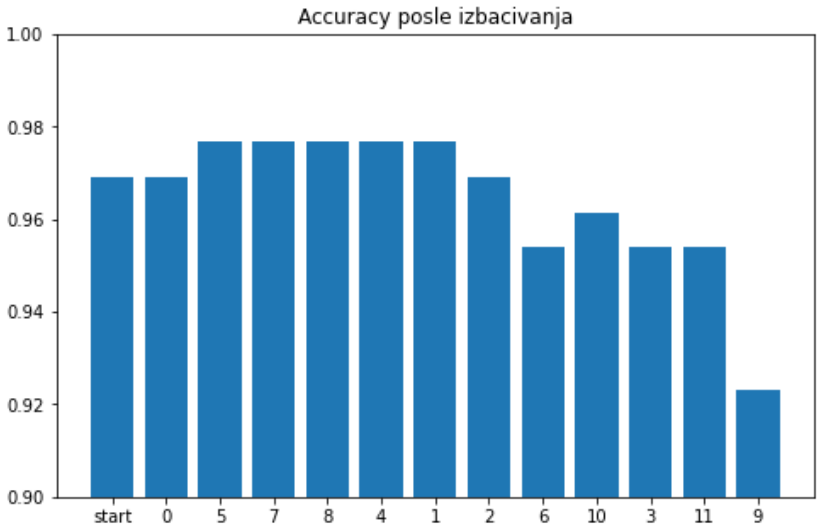
16. decembar 2023.

1 Selekcija obeležja

Prvo ćemo sortirati prediktore na osnovu koeficijenata Pearsonovog korelacije sa ciljnom promenljivom. Sva obeležja su sortirana po apsolutnoj vrednosti (nama odgovara da koeficijent korelacije bude bliži 1 ili -1, a dalji od 0):



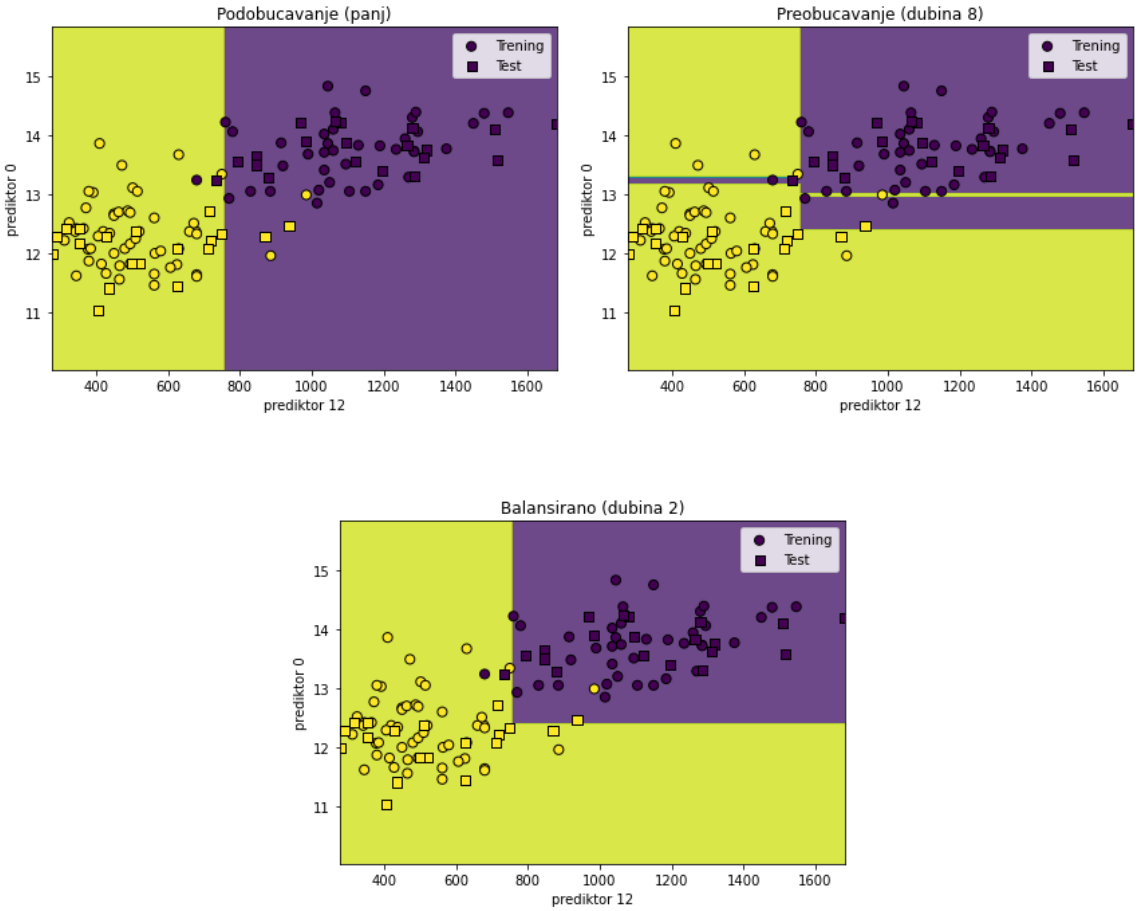
Drugi postupak selekcije obeležja podrazumeva da se iskoristi wrapper algoritam. U ovom slučaju ćemo iskoristiti pretragu unazad, tako što ćemo prvo obučiti model logističke regresije sa svim mogućim prediktorima, a zatim iz tog skupa izbaciti onaj prediktor koji, kada se izbaci iz skupa prediktora kojim obučavamo model, daje najmanju razliku u krosvalidacionom skoru. Grafik menjanja tačnosti nakon izbacivanja najslabijek prediktora izgleda ovako:



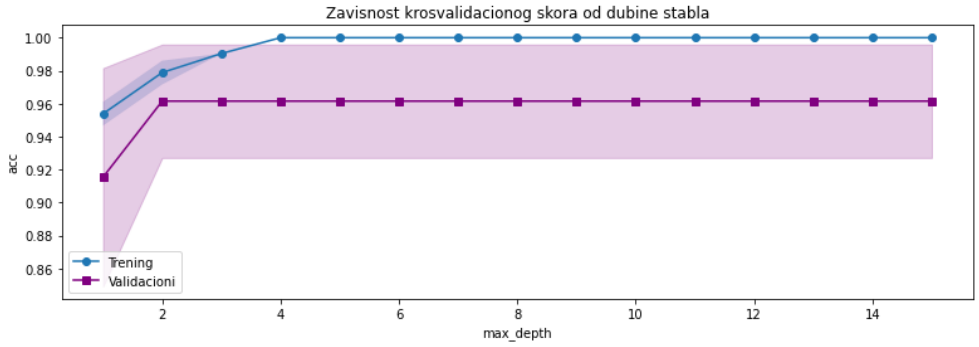
Moze se primetiti da fali jedan prediktor (poslednji, 12.), ali on je najbolji i kada bismo ga izbacili ne bi imali čime da obučavamo model.

2 Stabla odlučivanja

Za različite slučajeve (tj. za različite dubine), granice u ravni prediktora izgledaju ovako:

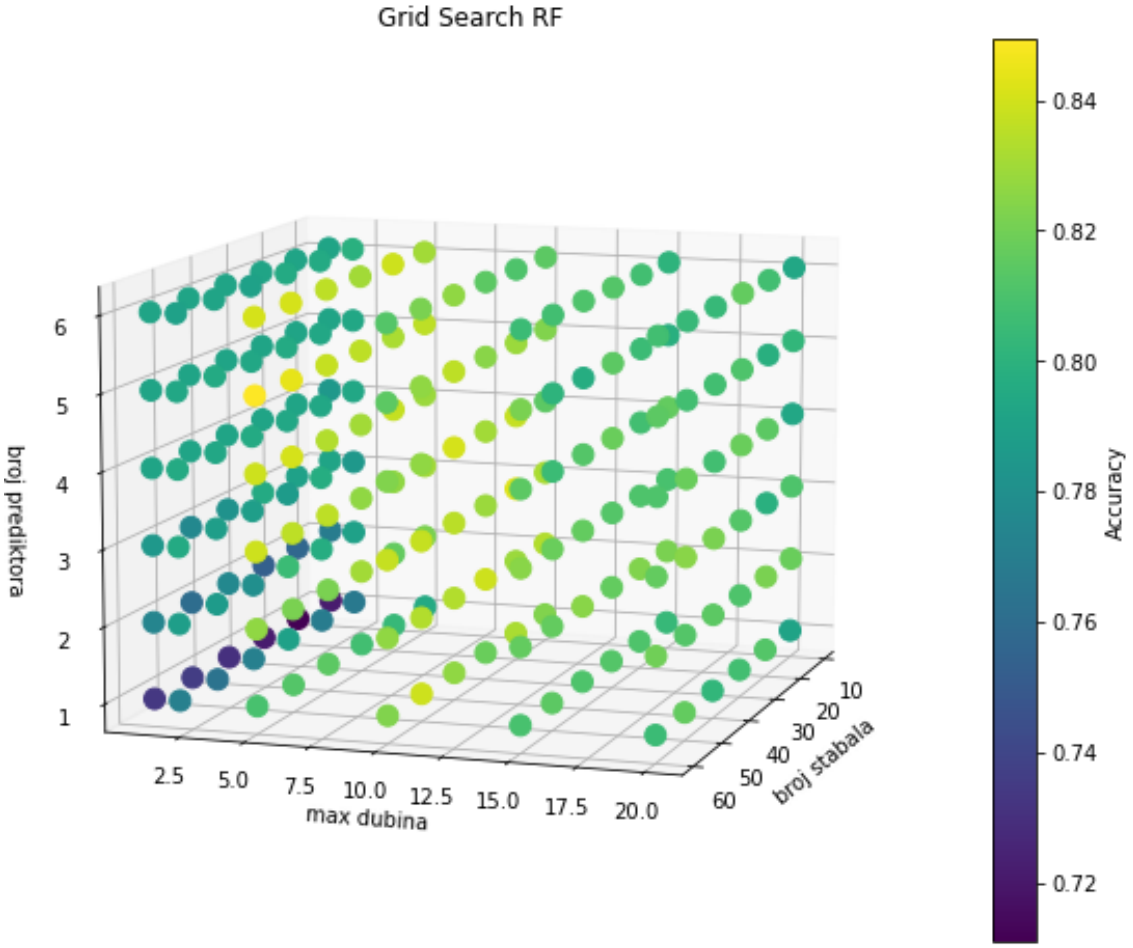


Dubine stabala su birane na osnovu krosvalidacionog skora na različitim dubinama. Vidimo da je najbolji slučaj kada je dubina 2, a već posle dubine 4 se ništa ne dešava i stablo je preobučeno:



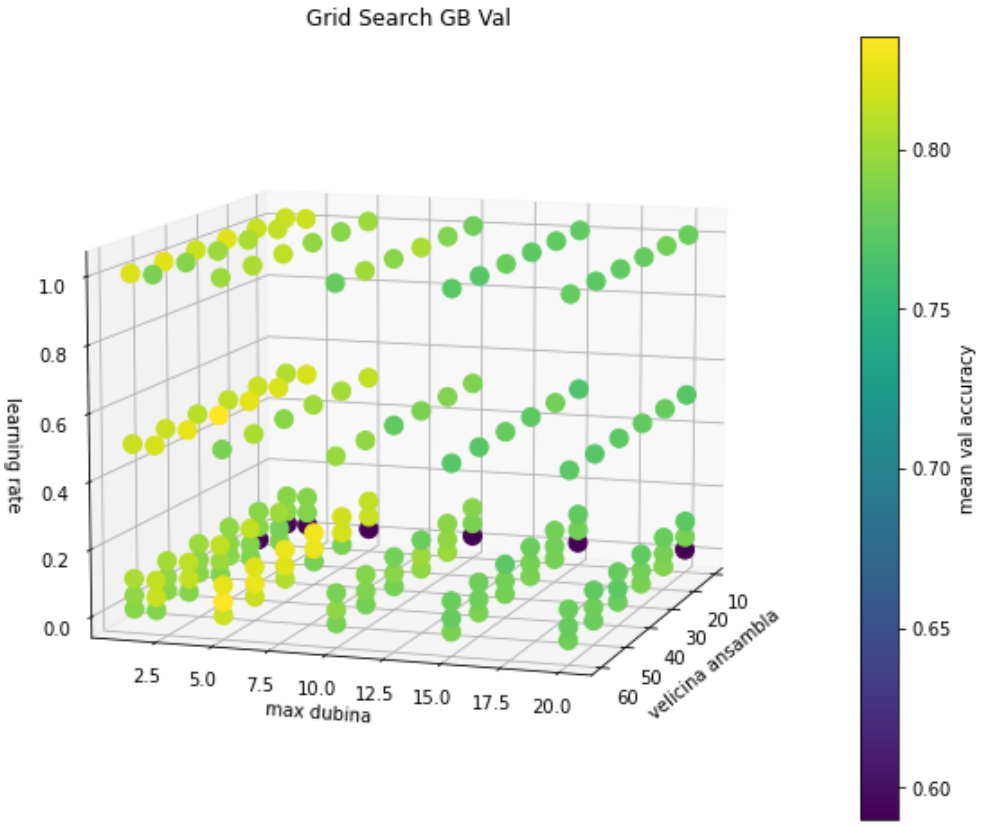
3 Ansambli

Za slučajne šume, isprobano je 6 vrednosti različitih vrednosti za svaki parametar (veličina ansambla, maksimalna dubina i broj prediktora koji se razmatraju kada se dodaje novi čvor u stablo). Metrika koja je korišćena da oceni kvalitet klasifikatora je srednja tačnost na validacionom skupu (za svaku kombinaciju parametara, sa 5-tostrukom unakrsnom validacijom). Sa grafika možemo primetiti najbolju kombinaciju parametara:



gde je optimalna dubina stabla 5, optimalan broj prediktora isto 5, a veličina ansambla 60. Može se primetiti i da je najbolji parametar maksimalna dubina, jer nam daje osećaj maksimuma blizu optimalne maksimalne dubine (levo i desno od nje se smanjuje tačnost), dok se za druga dva parametra mogu dati neke pretpostavke (vidimo da je tačnost veća tamo gde zajedno rastu broj stabala i broj prediktora, kao da su u nekoj korelaciji).

U slučaju gradient boostinga, parametri koji su korišćeni su veličina ansambla, maksimalna dubina i stopa obučavanja. Ovde možemo videti da je veći region kombinacije parametara koji daje rezultate slične najboljim:



Ovde se dobija najbolja kombinacija parametara: optimalna stopa obučavanja 0.5, optimalna maksimalna dubina 2 i optimalna veličina ansambla 40.

Značajnosti prediktora naći ćemo pomoću atributa *feature_importances_* nakon što se fituje slučajna šuma. Ovaj atribut se računa na osnovu Gini gaina (razlika između Gini nečistoće ciljne promenljive i Gini nečistoće ciljne promenljive da je dataset podeljen prema određenom prediktoru):

$$\Delta \mathrm{Gini}(x_i) = \mathrm{Gini}(y) - \mathrm{Gini}_{x_i}(y)$$

Najveći značaj imaju oni prediktori koji imaju najveću očekivanu vrednost Gini gaina po svim učesnicima u ansamblu: