Titanic – Klasifikacija

Autor: Ognjen Ikrašev RA238/2022

### **1. Uvod**

Cilj ovog rada je analiza podataka sa *Titanica*, kao i izgradnja prediktivnih modela koji klasifikuju da li je putnik preživeo brodolom. Analiza obuhvata pripremu podataka, inženjering atributa, poređenje različitih algoritama mašinskog učenja i evaluaciju performansi.

### **2. Priprema i inženjering podataka**

Polazni skup sadrži osnovne informacije o putnicima (ime, pol, godine, cena karte, klasa i sl.). Kako bi povećao informativnost podataka, uveo sam nove atribute:

* **Title** – titula iz imena (npr. Mr, Mrs, Miss), koja odražava društveni status i često implicira pol i starost.
* **FamilySize** – ukupan broj članova porodice na brodu.
* **IsAlone** – indikator da li je putnik bio sam.
* **TicketPrefix** – prefiks broja karte (ponekad povezan sa klasom ili grupom).
* **CabinDeck** – paluba na kojoj se nalazila kabina.
* **AgeBin** – grupisanje godina u intervale (dete, tinejdžer, odrasli...).

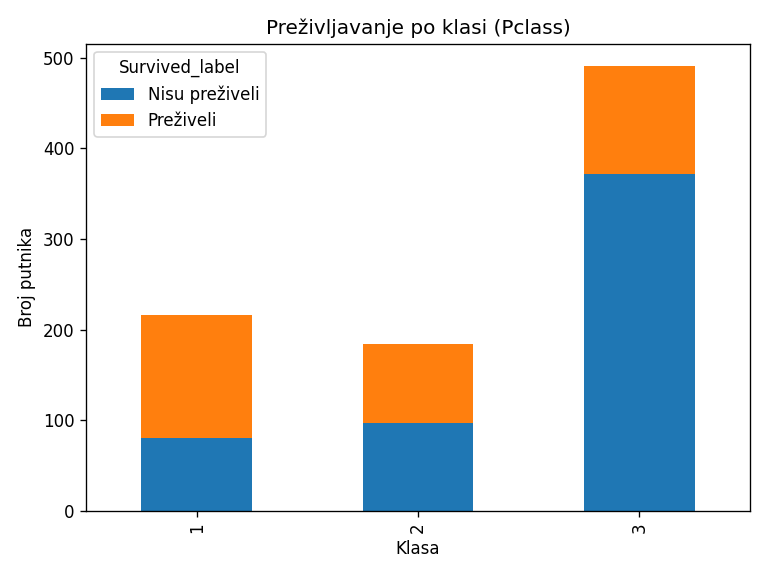
Ovi atributi su uvedeni jer pomažu modelima da lakše otkriju obrasce koji nisu očigledni u sirovim numeričkim vrednostima.

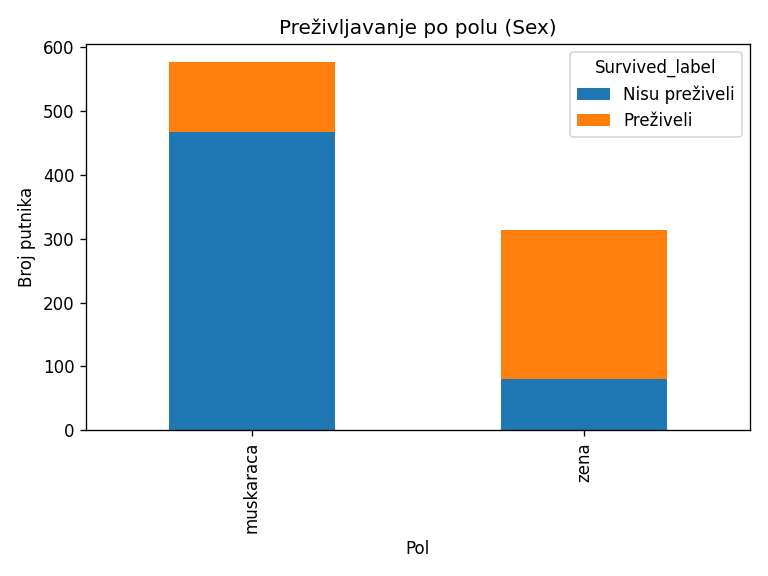
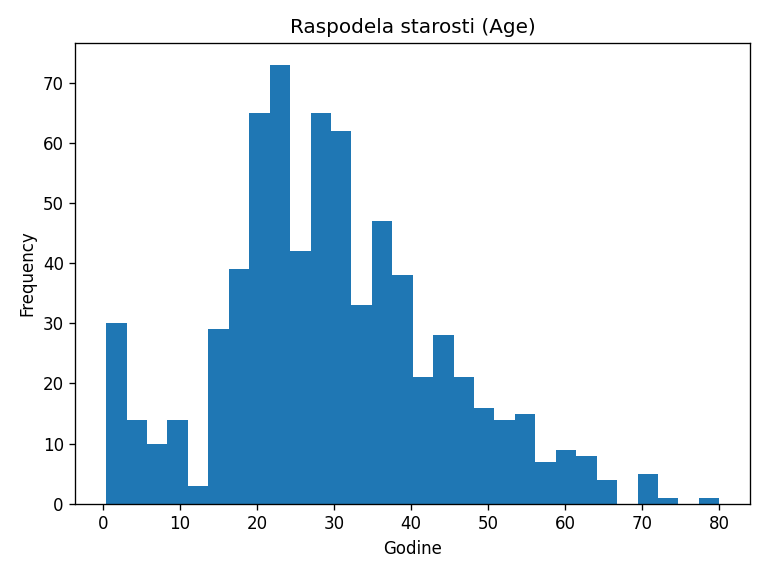
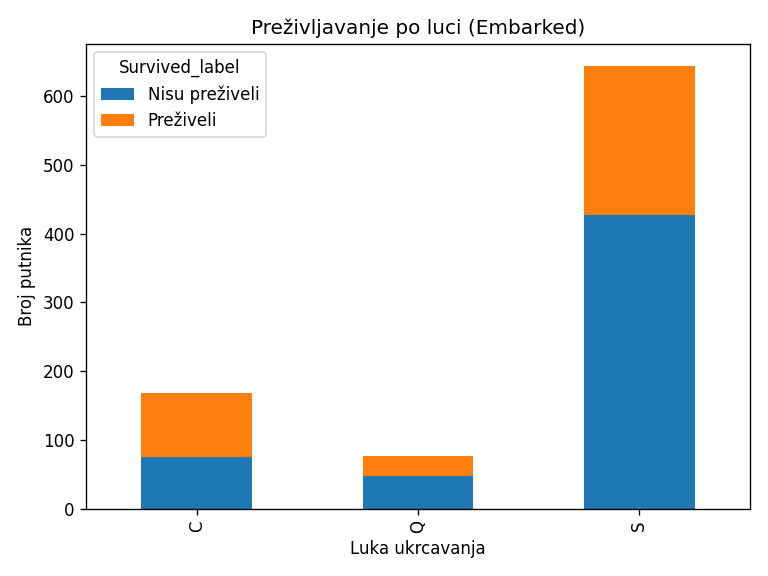
### **3. Eksplorativna analiza podataka (EDA)**

Vizuelizacije su pokazale:

* Žene i deca imali su znatno veću stopu preživljavanja.
* Putnici prve klase preživljavali su češće od ostalih.
* Cena karte i luka ukrcavanja imali su određeni uticaj na verovatnoću preživljavanja.

Ove uvide kasnije su potvrdili i modeli.





### **4. Izabrani algoritmi i obrazloženje**

Testirano je više algoritama mašinskog učenja:

1. **Logistička regresija**
   1. Koristi se kao osnovni model za binarnu klasifikaciju.
   2. Lako interpretabilna i daje dobru baznu tačnost.
2. **Random Forest**
   1. Ansambl metoda bazirana na velikom broju stabala odlučivanja.
   2. Dobro hvata nelinearne odnose i interakcije između atributa.
   3. Robustan na šum i dobro se nosi sa nedostajućim vrednostima.
3. **Gradient Boosting**
   1. Složeniji ansambl koji gradi stabla sekvencijalno, svako sledeće pokušava da ispravi greške prethodnog.
   2. Obično daje visoku tačnost na strukturisanim podacima kao što je Titanic skup.

Ovi algoritmi su izabrani jer pokrivaju različite pristupe: od jednostavnog linearnog (logistička regresija), preko bagging metoda (Random Forest), do boosting metoda (Gradient Boosting). To omogućava fer poređenje i izbor optimalnog modela.

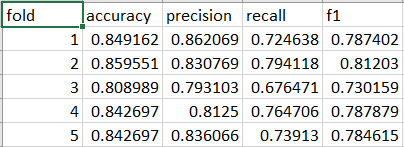
### **5. Evaluacija**

* Korišćena je **stratifikovana 5-fold cross-validation** metoda da bi se dobile stabilne procene performansi.
* Pored tačnosti (accuracy), korišćene su i **precision, recall i F1-score** metrike, jer je važno ravnotežno procenjivati i preživljavanje i smrt.

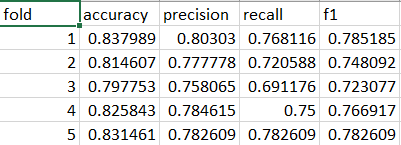
Rezultati su pokazali:

* Logistička regresija daje solidne rezultate, ali lošije od kompleksnijih metoda.
* Random Forest i Gradient Boosting postižu bolju ravnotežu između preciznosti i osetljivosti.
* Najbolji rezultati dobijeni su sa boosting modelom.

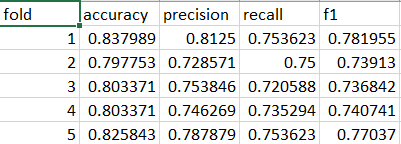
Gradient Boosting



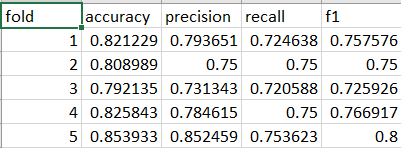
Logistička regresija



Random Forest



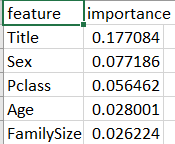
Random Forest top 7



### **6. Važnost atributa**

Analiza važnosti atributa pokazala je da najveći uticaj imaju:

* Pol (Sex),
* Klasa (Pclass),
* Godine (Age),
* Veličina porodice (FamilySize),
* Titula (Title).



Permutation importance je potvrdio da modeli ne zavise previše od manje relevantnih atributa, što povećava pouzdanost rezultata.

### **7. Selekcija atributa**

Kada su modeli trenirani samo na ograničenom broju najvažnijih osobina, performanse su ostale vrlo slične kao i sa celim skupom. To pokazuje da se zadatak može rešavati efikasno sa manjim brojem značajnih atributa.

### **8. Zaključak**

Postupak je obuhvatio celokupan tok rada u mašinskom učenju: od pripreme podataka i inženjeringa atributa, preko EDA, do poređenja više algoritama. Najbolje performanse postigli su modeli bazirani na boosting tehnikama, dok je logistička regresija poslužila kao dobra polazna tačka.

Kombinacija dobrog inženjeringa atributa i moćnih algoritama pokazala se kao ključ za uspešno rešavanje problema predviđanja preživljavanja na Titanicu.

# Metrike po modelima (5-fold CV)

**LogReg:** accuracy=0.824, precision=0.788, recall=0.740, f1=0.763

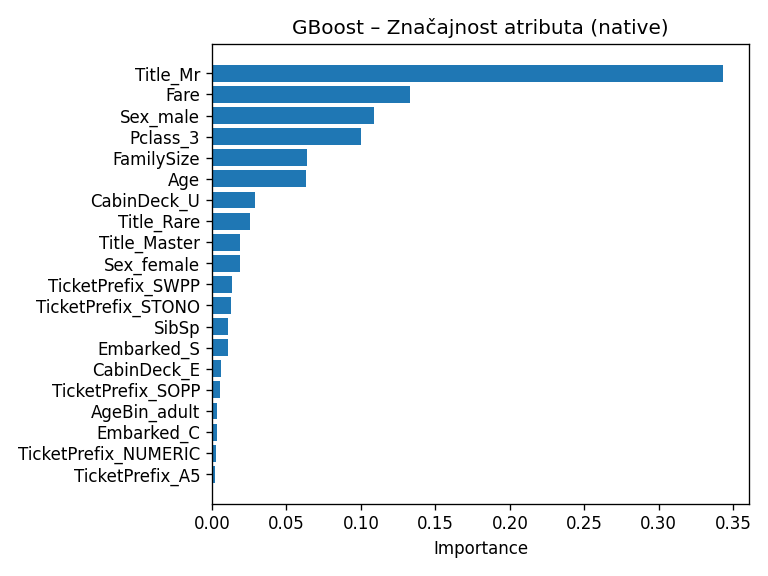
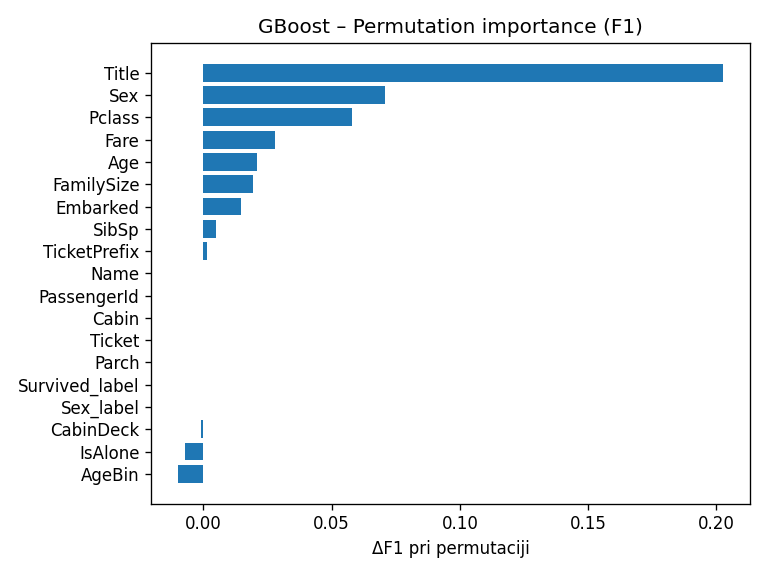
**RF:** accuracy=0.813, precision=0.766, recall=0.743, f1=0.753

**GBoost:** accuracy=0.839, precision=0.826, recall=0.737, f1=0.779

**RF\_top7:** accuracy=0.820, precision=0.782, recall=0.740, f1=0.760

# Grafici – EDA

# Značajnost atributa



# Tumačenje i preporuke

Najveći doprinos preživljavanju, u skladu sa literaturom i analizom, pokazuju pol (female), klasa karte (Pclass), kao i starost. Takođe, izvedeni atributi poput FamilySize/IsAlone i titula (Title) doprinose razdvajanju klasa. Preporuke: dodatno optimizovati hiperparametre (npr. pretraga mrežom), probati kalibraciju verovatnoća i balansiranje klasa.