Titanic Survival Classification – Projekat

# **Opis problema**

Pretpostavlja se strateško rešavanje zadatka klasifikacije u kojem će **biti predviđeno preživljavanje putnika** (*survived / not survived*) na osnovu demografskih, porodičnih i putnih karakteristika. Dataset sadrži informacije o putnicima na Titaniku, a očekuje se da model omogući identifikaciju faktora koji su imali najznačajniji uticaj na preživljavanje. Poseban naglasak se stavlja na sposobnost analize, interpretacije i vizualnog predstavaljanja podataka uz minimalni skup resursa — isključivo *train.csv*.

# **Pregled kolona i njihovo značenje**

Dataset (*train.csv*) obuhvata **sledeće kolone**, svaka sa specifičnim značenjem:

* **PassengerId** — jedinstveni identifikator svakog putnika, korišćen za praćenje.
* **Survived** — **ciljana promenljiva** označena vrednostima 0 (nije preživeo) i 1 (preživeo).
* **Pclass** — klasa karte (1 = prva, 2 = druga, 3 = treća), funkcioniše kao pokazatelj socio-ekonomske pozicije.
* **Name** — puno ime putnika, iz kojeg se može izvesti titula (npr. “Mr.”, “Mrs.”).
* **Sex** — pol putnika (male / female), značajan demografski faktor.
* **Age** — starost u godinama; ponekad izražena kao decimalna vrednost (npr. 0.5 za bebe).
* **SibSp** — broj braće/supružnika koji su bili na brodu (uz striktno definisane relacije).
* **Parch** — broj roditelja/dece na brodu (uz definisane kategorije kao što su majka, otac, dete).
* **Ticket** — broj karte, koji može sadržati informacije o tipu ili grupi putnika.
* **Fare** — cena karte, izražena u britanskim funtama, ukazuje na platežnu moć i položaj.
* **Cabin** — broj kabine, često nepotpun i korišćen kao potencijalni indikator dodate vrednosti (npr. blizina palube).
* **Embarked** — luka ukrcavanja (C = Cherbourg, Q = Queenstown, S = Southampton), može ukazivati na geografske i društvene obrasce.

# **Očekivani izlazi (output):**

* **Evaluacija modela klasifikacije** obuhvata diskretan prikaz performansi kroz metrike: *tačnost, preciznost, odziv* i *F1-skor*.
* **Vizualna analiza** najuticajnijih atributa (feature importance) koja će se jasno predstaviti grafički.
* Grafički prikazi **odnosa između atributa i rezultata** (npr. raspodela starosti, preživljavanje po klasi, polu, luku ukrcavanja).
* **Uključeno tumačenje rezultata**: koje karakteristike se najjače povezuju sa preživljavanjem i koje su preporuke za eventualno dalje unapređenje modela.

# **Zahtevi za procesiranje podataka:**

* **Nedostajući podaci** u kolonama poput *Age*, *Cabin* i *Embarked* treba da budu obrađeni na način koji minimizira njihov negativan uticaj na performanse modela.
* **Kategorijski atributi** (*Sex*, *Embarked*, *Pclass*) trebaju da budu prikazani u formatu prikladnom za model, a razlog primene odabrane obrade treba da bude dat.
* **Odluke o uklanjanju ili transformaciji** kolona kao što su *Name*, *Ticket*, *Cabin*, ili eventualna ekstrakcija dodatnih informacija (npr. titule iz *Name*) moraju biti opravdane.

# **Dokumentovanje rezultata:**

* **Rezultati modela i vizuelni elementi** (grafici, tabele) treba da budu prikazani jasno, u formatu pogodno za integraciju u Word dokument.
* **Interpretacije i zaključci** moraju biti jasni i logično utemeljeni, uz preporuke za potencijalna poboljšanja.