Final Project: Titanic Classification

Programmer: Pragya Saxena

What is titanic data?

Titanic data, is a popular dataset, that includes information like, survived: whether the passenger lived (1) or died (0), passenger class, age, sex, fare, parch, sibsp. The goal in most Titanic projects is to predict whether or not the passenger survived or not based on those conditions you make. By building machine learning models, you can see which models have the best predictability.

survived	pclass	name	sex	age	fare	sibsp	pa
0	3	Braund, Mr. Owen Harris	male	22.0	7.25	1	0
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	71.2833	1	0
1	3	Heikkinen, Miss. Laina	female	26.0	7.925	0	0
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1	1	0
0	3	Allen, Mr. William Henry	male	35.0	8.05	0	0
0	1	McCarthy, Mr. Timothy J	male	54.0	51.8625	0	0
0	3	Palsson, Master. Gosta Leonard	male	2.0	21.075	3	1

Why do this project?

This project is a part of stempeers final summer project. This allows me to gain the skills I need to explore data preprocessing, model training, and evaluation while predicting survival outcomes.

Model Accuracy

I have built 3 models, that allow me to classify whether a passenger on the titanic would survive, and those are, logistic regression: 75%, decision tree: 74%, and random forest: 78%. Although all the models were pretty accurate, all of their accuracy scores were around the same, so I can improve it by adding feature engineering which is selecting, creating, and transforming features to better the models accuracy. Instead of dropping the rows that had missing data, I could've filled it with the median age. I could've used feature scaling, which improves performance in logistic regression

What I learned

I learned how to preprocess data by handling missing values and converting categorical variables into numeric form. I also gained experience in training and evaluating classification models, interpreting accuracy scores, and understanding how different algorithms approach the same prediction problem.