Adwaith Korapati – SEC01 (NUID 002670980)

Big Data System Engineering with Scala  
Fall 2023   
Assignment No. Spark 2



**-List of Tasks Implemented**

Exploratory Data Analysis- Analysed the Titanic dataset to find out the following statistics,

* Calculate the survival rate by passenger class
* Calculate the survival rate by gender
* Calculate the average age of passengers by class
* Calculate the average fare by class
* Calculate the count of passengers embarked from each port
* Calculate the survival rate based on the "Embarked" column

Feature Engineering - Create new attributes that may be derived from the existing attributes. This may include removing certain columns in the dataset.

* Created a new column Family size which includes the count of kids and spouse.
* Changed the Embarked column data from Char codes to name of the places to help for better visualization

Prediction - Used the train.csv to train a Random Forest Machine Learning model & test it on the test.csv. Predicted if the records in test.csv survived or not.

**-Findings and analysis**

**Survival rate based on Class-**

First class people had the best chance of surviving the disaster with 62.9% followed by second class with 47.28% followed by the third class with the survival chance of 24.23%.

More money you spend, better are your survival chances.

**Survival rate based on Sex-**

Females survived the crash better with 74.20% of them got out followed by male with mere 18.8%.

**Survival rate based on Embarked-**

People from Cherbourg had the highest rate of survival with 55.35% followed by Queenstown and Southampton with 38.96% and 33.69% respectfully.

**Average fare based on class-**

First class – 84.15 (assuming it is in pounds)

Second class- 20.66

Third class – 13.67

**Average age based on Class-**

First class – 32.9

Second class- 28.09

Third class – 18.17

Predication Model had an Accuracy of 90.19138755980862%

**-Code**

**FE and EDA code**

A screenshot of a computer

Description automatically generatedA computer screen shot of a black screen

Description automatically generated

**ML Code**

A screenshot of a computer

Description automatically generatedA computer screen shot of a black screen

Description automatically generated

**- Result**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA computer screen shot of a black screen

Description automatically generated