Introduction

This report presents my findings from participating in the renowned Kaggle machine learning competition titled "***Titanic - Machine Learning from Disaster***". The competition focused on predicting survival outcomes based on passenger data from the tragic sinking of the RMS Titanic on April 15, 1912. Despite its reputation as an "unsinkable" ship, the Titanic succumbed to an iceberg collision during its maiden voyage, leading to the loss of 1502 lives out of 2224 passengers and crew members. Due to a shortage of lifeboats, survival was not guaranteed for everyone on board, and certain groups appeared to have higher chances of survival. The challenge tasked participants with constructing a predictive model to answer the question: "Which demographics were more likely to survive?" using the available passenger data.

Methodology

1. Data Collection and Exploratory data analysis

The Kaggle competition titled "***Titanic - Machine Learning from Disaster***" offers a dataset featuring information about Titanic passengers. The challenge involves developing a predictive model to determine which passengers survived the disaster. The dataset comprises 891 entries and 12 columns, encompassing details such as passengers' names, ages, genders, ticket classes, and their survival status. The data is divided into two files: train.csv, containing both passenger information and survival outcomes, and test.csv, which includes only passenger details without survival.
The variables that were provided had the following descriptors:

- Survived: Survived (1) or died (0)
- Pclass: Passenger's class
- Name: Passenger's name
- Sex: Passenger's sex
- Age: Passenger's age
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Fare
- Cabin: Cabin
- Embarked: Port of embarkation

Exploratory data analysis was performed on the training set to identify any missing variables or anomalies requiring cleaning. Furthermore, an investigation into survival and mortality rates was carried out, considering various factors like passenger seating, gender, age, and more.

2. Data pre-processing

There was missing values that were found in the column that consisted age values and I filled it using the average value of age, in the embarked column the missing value were replaced with the symbol that symbolise unclassified category. In addition, the type of data that were given as objects were converted to float and this conversion allowed me to perform some numerical operations when normalizing the data.

3. Model development and evaluation

Different popular machine learning models were used such as logistic regression, K nearest neighbour, Decision tree, support vector machine and even random forest. Their comparison was done based on their precision and accuracy on data that a model is familiar with (train set) as well as the new unseen data.

Results

As shown in figure 1 below passengers who were seating in the first class survived at higher rate compared to other class while those who were seating in the third class died at higher rate compared to others. First-class cabins were typically located closer to the lifeboats, making it easier for passengers in the first class to reach the lifeboats quickly. Third-class cabins, on the other hand, were often located in the lower decks, farther away from the lifeboats. This made it more challenging for passengers in the third class to access the lifeboats during the limited time available for evacuation.

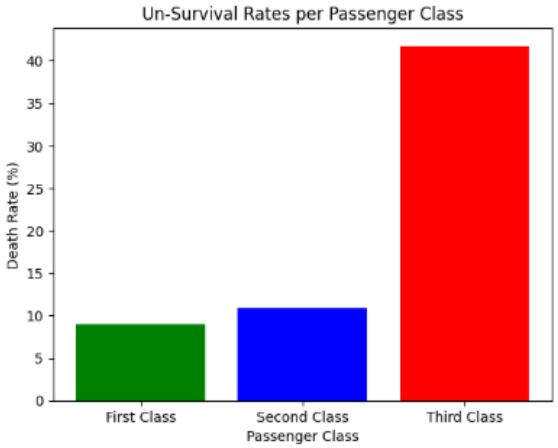

Figure 1:The rate of death based on passenger seats

The figure 2 below shows that female survived that incident at higher rate compared to male. The observation that a higher percentage of females survived the sinking of the Titanic compared to males is primarily due to the implementation of the "women and children first" protocol during the evacuation of the ship. This protocol was a maritime tradition that prioritized the evacuation of women and children over adult males in life-threatening situations

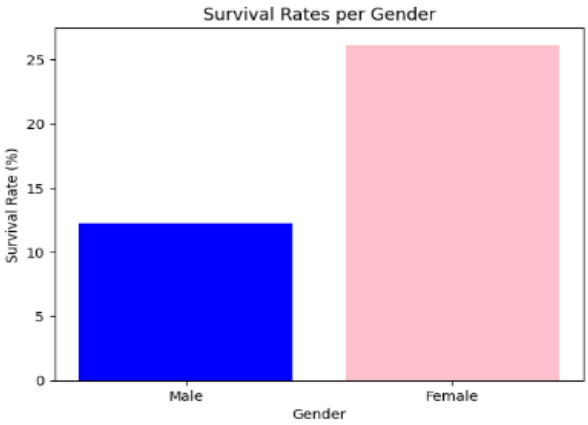

Figure 2:The survival rate based on passenger gender

Passengers who with the age less that 5 years were classified as babies, the ones with the age between 5 and 17 years were classified as children, the ones with age between 18 and 35 years were classified as young adults, the ones with the age between 36 and 50 years were classified as middle-aged adults, the ones with the age above 50 years were classified as old adults. The category with which the age is not known were not classified.

As shown in figure 3 below young adults survived at higher rate compared to other age categories and old adults died at higher rate compared to other categories. Young adults may have been more physically fit and agile, making it easier for them to navigate the ship, access lifeboats, and respond to the emergency situation.

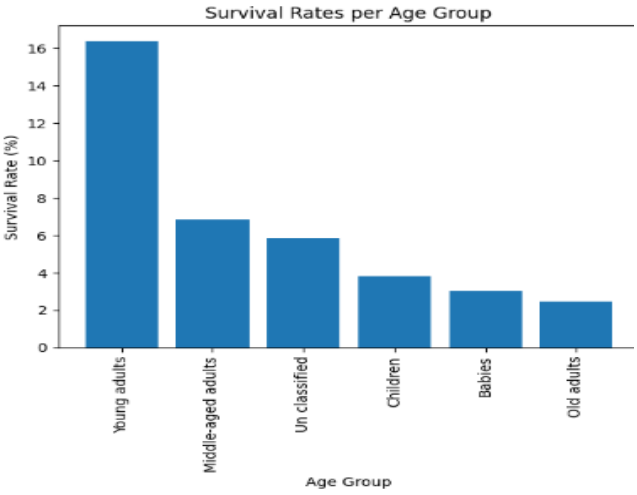

Figure 3: The survival rate of passengers based on their age

Before submitting my predictions on Kaggle I divided the given dataset into two sets such as training and testing set. The testing set consisted 20% of the given training data and I did this to determine a best model to predict unseen data before submitting my model, Figure 4 below shows the performance of different models that were trained, and Random Forest was the one that was performing better on unseen data with the accuracy of 84.35%.

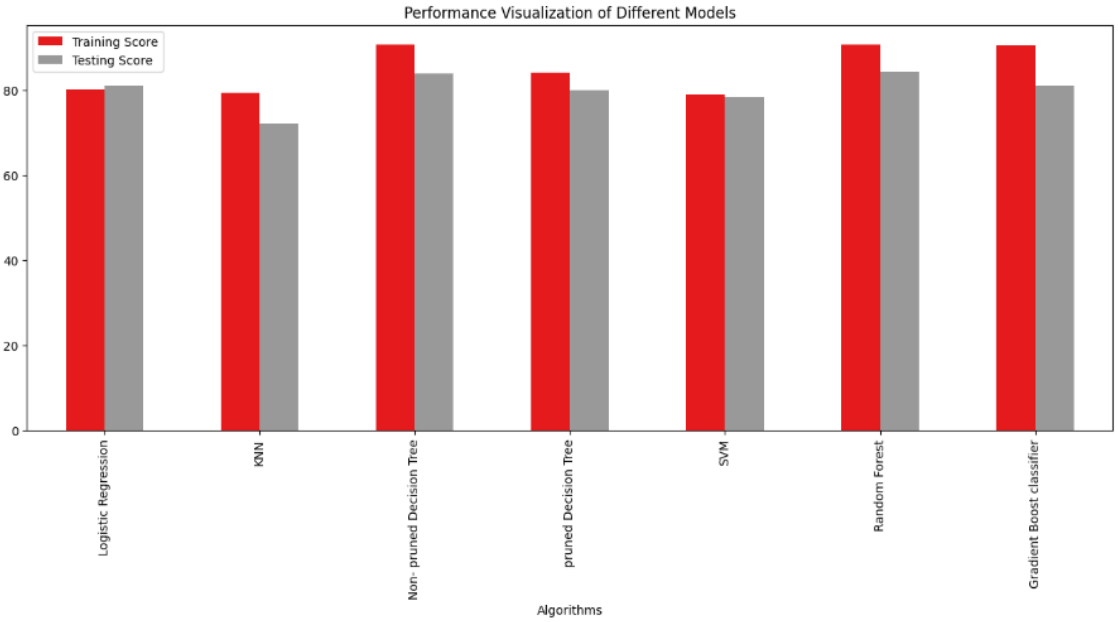

Figure 4: Model performance and their difference in predicting whether a passenger will survive

Figure 5 below shows the performance of my chosen model after on kaggle's testing dataset and it was 77.511%. Here are some of the best parameters used for the Random Forest Classifier to predict the survival of passengers on the Titanic :

- **criterion**: 'gini '
- **n_estimators**: 700
- **min_samples_split**: 10
- **min_samples_leaf:** 1
- **oob_score**: True
- **random_state**: 1
- **n_jobs**:. -1
- **bootstrap**: True

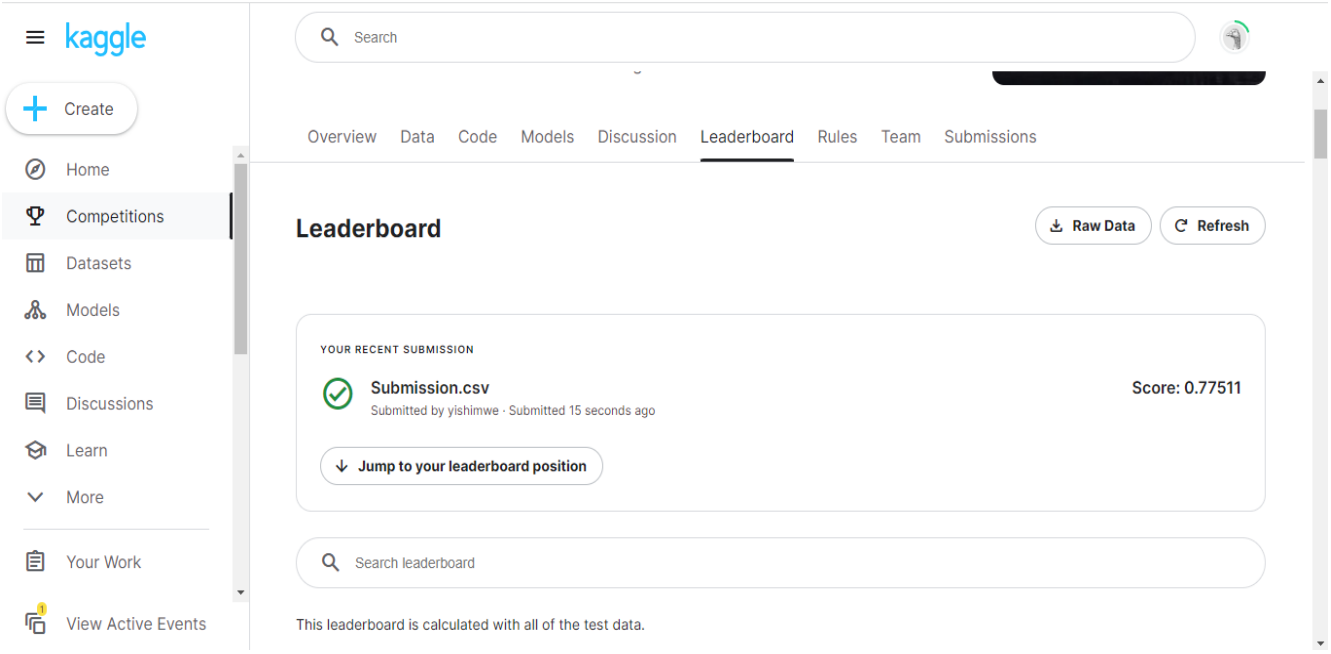

Figure 5: The score my chosen model on Kaggle competition