Predicting Survival on Titanic by Applying ExploratoryData Analytics and Machine Learning Techniques

By

Mohaned Farg Seif
211014738

Omar Ayman Tawfik
211009801

Supervised by

DR Ahmed Abdelhafeez

Abstract:

The sinking of the RMS Titanic, among history's deadliest maritime disasters, claimed thousands of lives due to a shortage of lifeboats. Survival odds favored women and children, highlighting disparities in rescue efforts. Through exploratory data analytics and machine learning, we dissect demographic trends to understand survival factors. Comparative analysis of machine learning results provides insights into the Titanic tragedy's human dynamics.

Introduction:

On April 15, 1912, the sinking of the Titanic etched itself into history as a haunting catastrophe. Its collision with an iceberg tore through the vessel, leaving insufficient lifeboats for the diverse array of passengers aboard, with many men making the ultimate sacrifice for women and children.

Today, leveraging machine learning algorithms, researchers endeavor to predict which passengers survived this fateful night. Analyzing factors like ticket fare, age, sex, and class, these algorithms sift through vast datasets to unveil pivotal patterns, shedding light on the dynamics of survival.

Exploratory data analytics delve into the dataset, probing each field's influence on passenger survival, illuminating crucial insights. Through meticulous comparison of algorithmic accuracy, researchers aim to pinpoint the most effective models for predictive analysis, advancing our understanding of this historic tragedy.

Methodology

1- Data

There are two groups in the historical data, one is the training set, and the other is the testing set. For the training set, using the testing dataset, which Kaggle provides, can determine whether the passenger survived to build the model for generating prediction patterns.

Table 1 shows the training dataset

Table 1. Training dataset

	Passengers	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund,Mr.Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cummings,Mrs.JohnBradiey(Florence Briggs)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen,Miss.Laina	female	26.0	0	0	STON/ O2.3101282	7.9250	NaN	S
3	4	1	1	Futrelle,Mrs.Jacques Heath(Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen,Mrs.William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran,Mr.James	male	NaN	0	0	330877	8.4583	NaN	Q

2- Data cleaning

The purpose of data cleaning is to prevent similar errors in data collection. Data cleanup can be divided into these steps identifying the mistakes and either correcting or deleting the data that needs to be modified or manually processing the data, which is an essential part of machine learning to build models. Correct data deletion can save us time and cost and improve data efficiency, and the previous dataset after cleaning is summarized in Table.2.

Table 2. dataset after cleaning

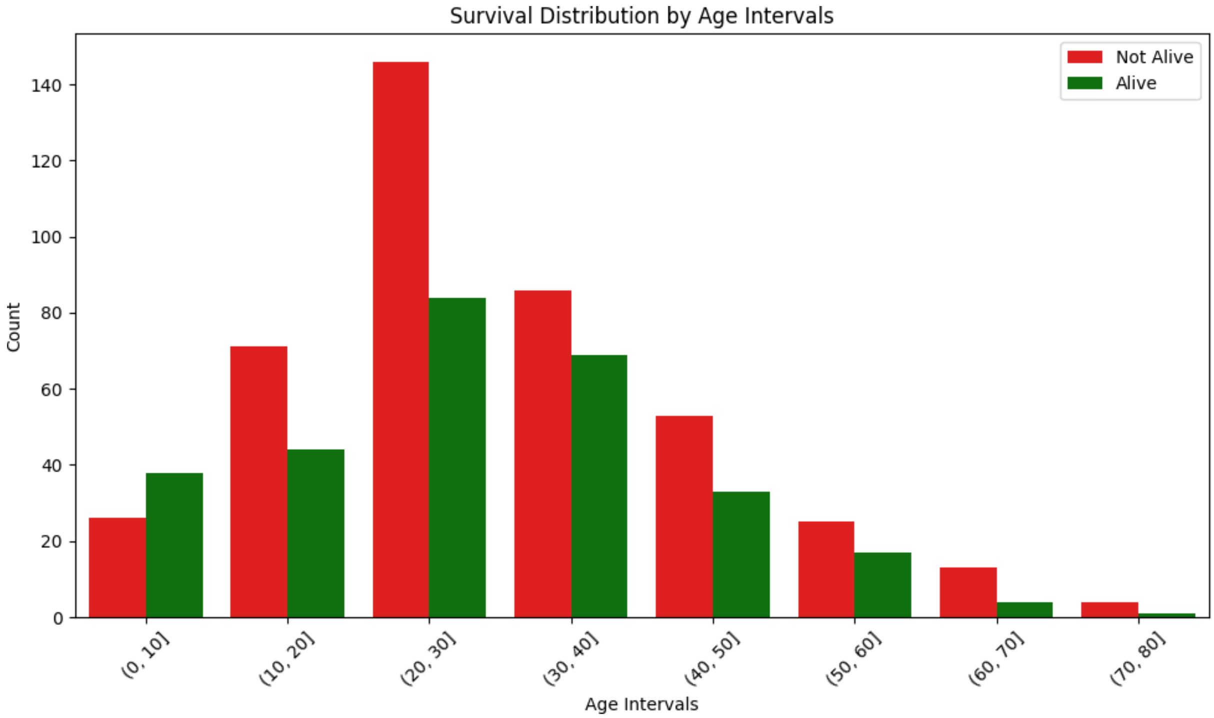
	Survived	Pclass	Sex	Age	Fare	Embarked	IsAlone	Title
0	0	3	1	1	0.0	0	0	1
1	1	1	0	2	3.0	1	0	3
2	1	3	0	1	1.0	0	1	2
3	1	1	0	2	3.0	0	0	3
4	0	3	1	2	1.0	0	1	1
5	0	3	1	1	1.0	2	1	1

- Exploratory Data Analysis

We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the survival rate. The data is deeply analysed by finding a relationship between each attribute and survival.

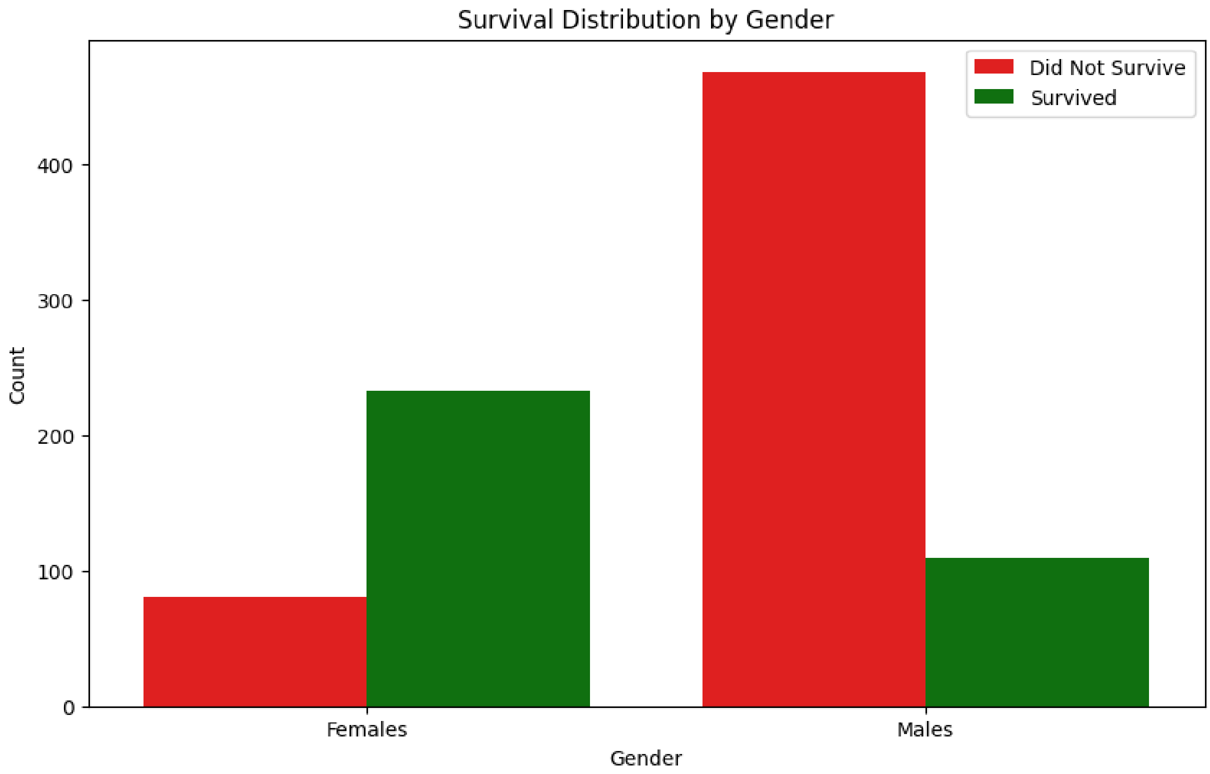
1- Age versus Survival

Here fig. 1 shows how survival rate will be affected by age. If the value of age is less then chances of survival are more and vice versa

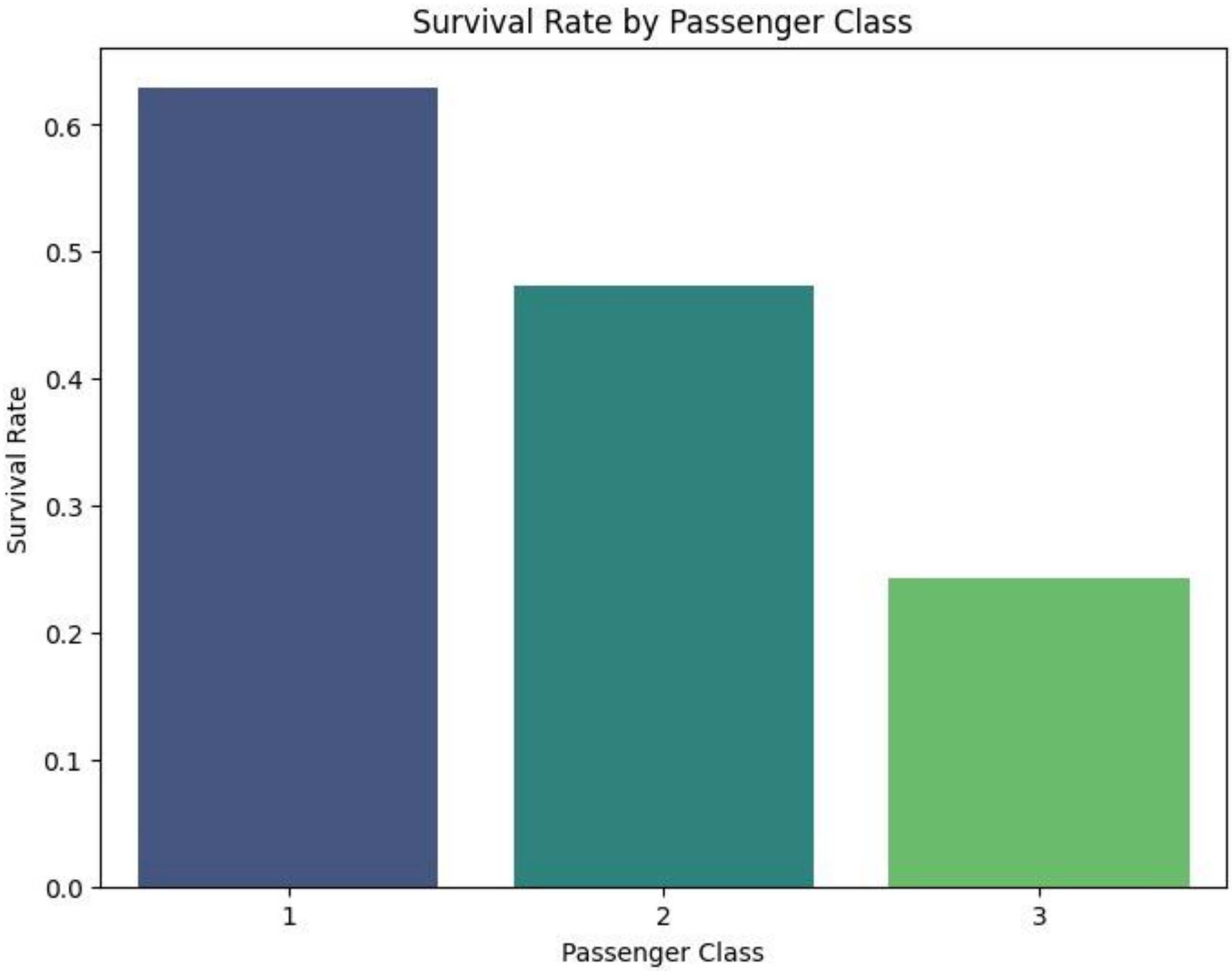


2-Sex versus Survival

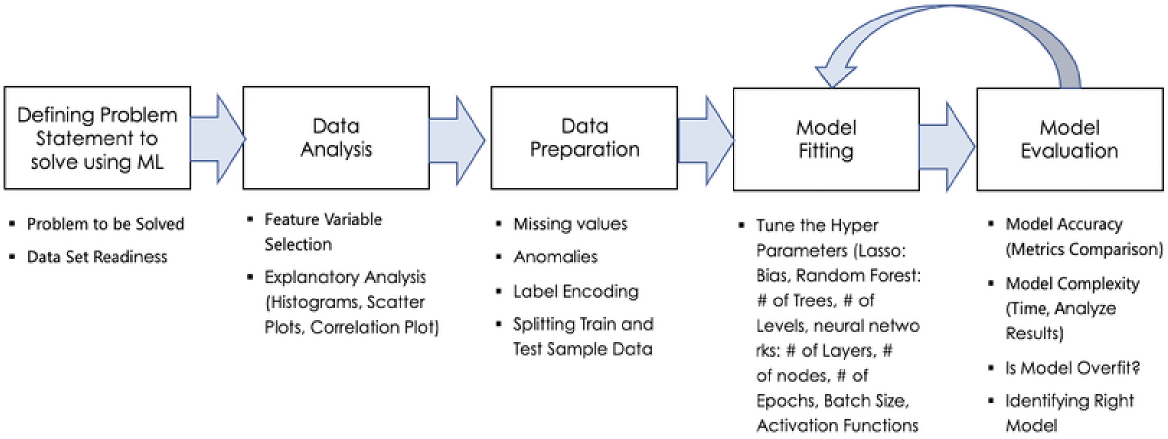
From Fig. 2 it is clear that females are more likely to survive than males. We calculated that survival rate of female and male are 74.20382% and 18.89081% respectively.



3-Passenger class versus Survival



- Machine Learning Models


Defining Problem Statement to solve using ML
Data Analysis
Data Preparation
Model Fitting
Model Evaluation
Problem to be Solved
Data Set Readiness
Feature Variable Selection
Explanatory Analysis (Histograms, Scatter Plots, Correlation Plot)
Missing values
Anomalies
Label Encoding
Splitting Train and Test Sample Data
Tune the Hyper Parameters (Lasso: Bias, Random Forest: # of Trees, # of Levels, neural netwo rks: # of Layers, # of nodes, # of Epochs, Batch Size, Activation Functions
Model Accuracy (Metrics Comparison)
Model Complexity (Time, Analyze Results)
Is Model Overfit?
Identifying Right Model

The process of ML

1-Support Vector machine

A support vector machine (SVM) is the basic model of the system, which is a linear classifier with the most considerable interval defined in the feature space. It is a generalized linear machine that supervises binary classification. It solves the classification problem of two groups of categories by using classification algorithms. After providing a set of SVM model sets marked with training data for each class, they can classify the new text. From the results of EDA, the SVM method of this project is still a problem to be improved. Before EDA, it was concluded that the survival rate of women and the upper class was higher than other classes. Therefore, eliminate unnecessary data to retrain the SVM model. Finally, 82.82% of the results can be determined by this improvement.

2-Decision tree

Decision tree is a supervised learning algorithm. This is generally used in problems based on classification. It is suitable for both categorical and continuous input and output variables. Each root node represents a single input variable (x) and a split point on that variable. The dependent variable (y) is present at leaf nodes. For example: Suppose there are two independent variables, i.e. input variables (x) which are height in centimeter and weight in kilograms and the task to find gender of person based on the given data. (Hypothetical example, for demonstration purpose only).

Model Evaluation:

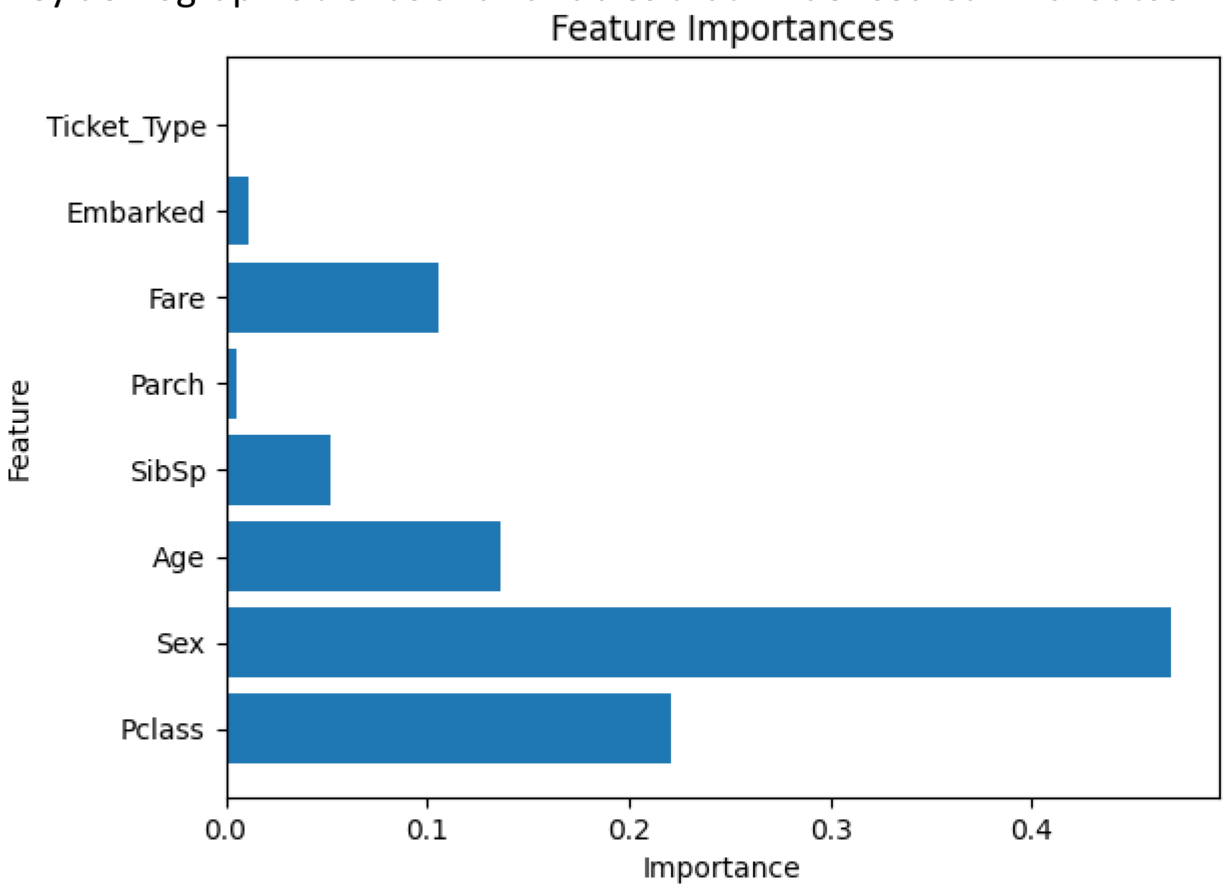
The accuracy of the model is evaluated using “confusion matrix”. A confusion matrix is a table layout that allows to visualize the correctness and the performance of an algorithm.

Confusion Matrix: A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N*N, here N is the number of values. Performance of such models is commonly evaluated using the data in the matrix.

- Our Contributions

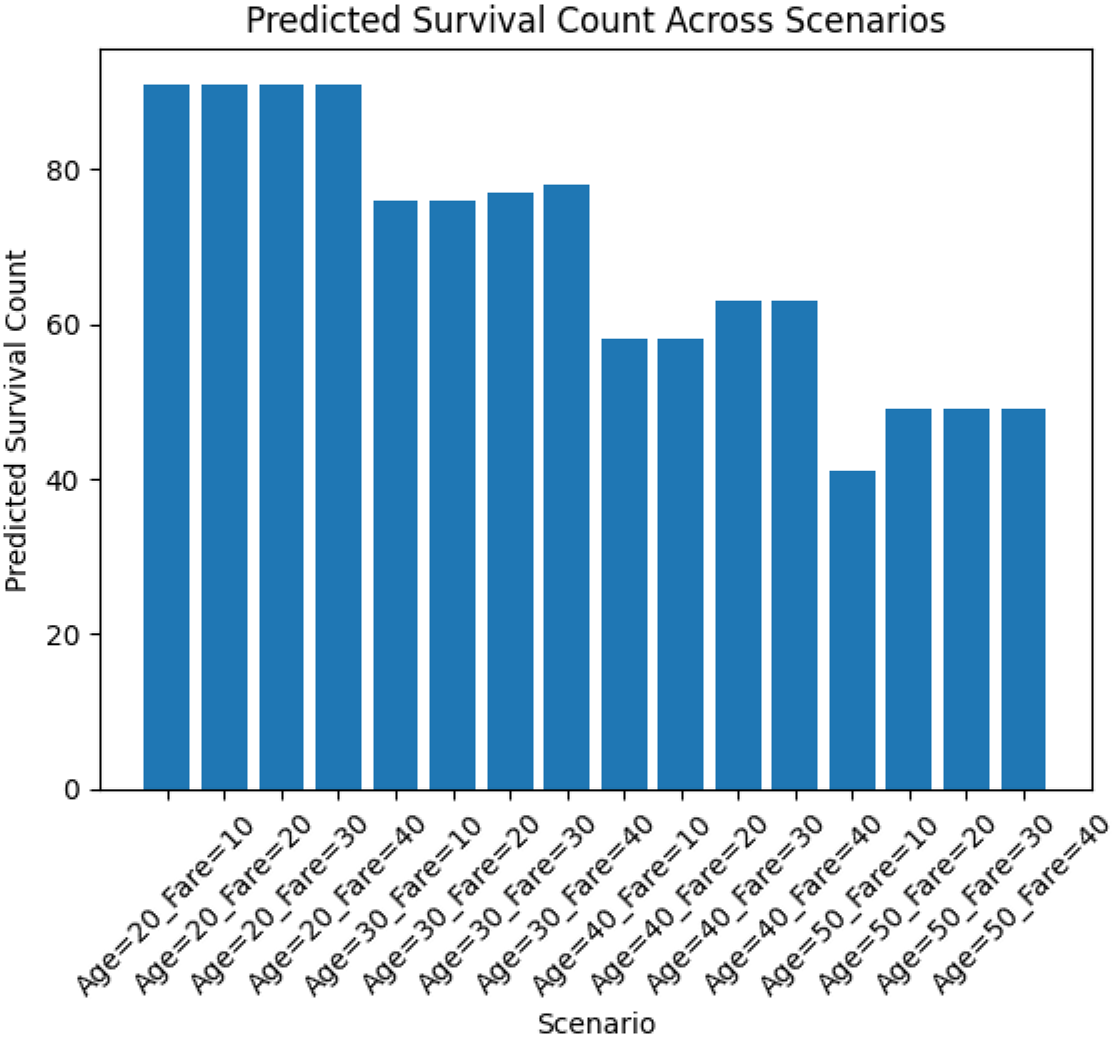
We made significant contributions to the paper are by Feature Importance and sensitivity analysis. Our primary contributions are outlined below:

1-Feature Importance: the feature importance analysis was performed to understand which factors had the most significant impact on predicting whether a passenger survived or not. By examining the importance of each input feature, the researchers aimed to uncover the key demographic trends and variables that influenced survival outcomes

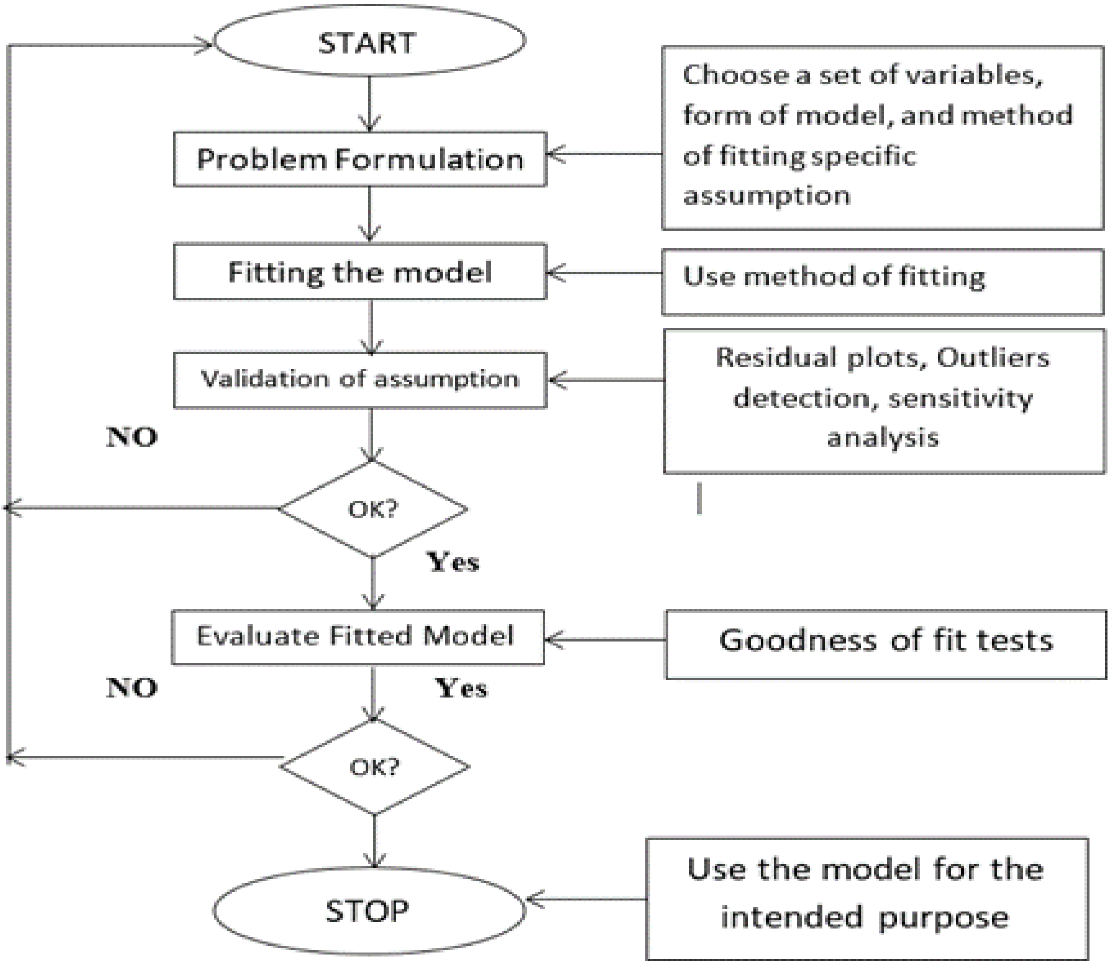


2-Sensitivity analysis: we performed sensitivity analysis to examine the robustness of our model to changes in input variables and parameters. This analysis not only provided us with valuable information about the stability of our predictions but also highlighted the factors that have the greatest impact on survival predictions. the model evaluation process played a crucial role in validating the effectiveness of our predictive model and ensuring its utility in real-world applications.

By rigorously assessing its performance and conducting sensitivity analysis, we were able to build a robust and reliable model for predicting Titanic survivors.



- Flowchart


START
Problem Formulation
Choose a set of variables, form of model, and method of fitting specific assumption
Fitting the model
Use method of fitting
Validation of assumption
Residual plots, Outliers detection, sensitivity analysis
NO
OK?
Yes
Evaluate Fitted Model
Goodness of fit tests
NO
Yes
OK?
STOP
Use the model for the intended purpose

- Conclusion

Applying exploratory data analytics and machine learning techniques to predict survival on the Titanic yields insightful conclusions. Initial data exploration and feature analysis unveil key insights into passenger characteristics. Through preprocessing and model training, various machine learning algorithms are employed to predict survival probabilities accurately. Evaluation metrics such as accuracy, precision, and recall rates validate model performance. Leveraging EDA enables a deeper understanding of the dataset's nuances, enhancing predictive capabilities. Ultimately, this integrated approach offers a robust framework for predicting survival outcomes on the Titanic.

- ## References

- The National Archives, J. C. 2012. Titanic, 100 years on. https://blog.nationalarchives.gov.uk/titanic-100-years-on/
- Kaggle. 2020, Aug 1. Titanic - machine learning from disaster. https://www.kaggle.com/c/titanic/data
- Cronan, J. 2012. *National Archives*.https://blog.nationalarchives.gov.uk/titanic-100-years-on/
- Lord, W., &amp; Philbrick, N. 2017. I Believe She's Gone, Hardy. In A night to remember (pp. 47–50).
- Mishra, V. P., Singh, B., Shukla, V. K., &amp; Dasgupta, A. 2021. Predicting the likelihoodof survival of Titanic's passengers by Machine Learning. https://www.researchgate.net/publication/351155499_Predicting_the_Likelihood_of_Survival_of_Titani c%27s_Passengers_by_Machine_Learning
- Brownlee, J. 2020. Evaluating Machine Learning Algorithms.https://machinelearningmastery.com/train- test-split-for-evaluating-machine-learning-algorithms/
- Cook, A. 2019. Titanic - Machine Learning from Disaster. https://www.kaggle.com/code/alexisbcook/titanic-tutorial
- Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.
- Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-Titanic Machine Learning From Disaster, 2012.
- S. Cicoria, J. Sherlock, M. Muniswamaiah, L. Clarke, "Classification of Titanic Passenger Data and Chances ofSurviving the Disaster", Proceedings of Student-Faculty Research Day CSIS, pp. 1-6, May 2014.
- Corinna Cortes, Vlasdimir Vapnik, "Support-vector networks", Machine Learning, Volume 20, Issue 3,pp273-297.
- L Breman- "random forests", Machine Learning, 2001 Ng. CS229 Notes, Standford University, 2012.

SJ Russsel P Norvig-"Artificial intelligence: A modern approach"-2016.