The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 1, No. 1. Online ISSN: 2831-4794

Available at: https://www.isods.org/publication/

Titanic Survival Prediction Using Predictive Modeling Algorithms

Leakey O. Omolo*[a], Nguyet Nguyen[b]
[ab]Department of Mathematics and Statistics, Youngstown State University, Cafaro Hall, Youngstown, OH, USA 44555

ABSTRACT

The Titanic accident is a remarkable historical loss in the cruise industry with only 32% survival rate. In this project, a binary classification model is built to predict whether a Titanic accident passenger survived or not using python programming language. The paper analyzes 11 variables of each customer to find out which factors have the most significant effect on the customers' survival rate and then uses binary classifiers for prediction. The process involves identifying relevant features, highlighting and visualizing the interesting relationships between them, and that with the target variable (survived). We predict survival using Random Forest, Logistic Regression, Naive Bayes, k-nearest neighbors(KNN) and Decision Trees. We evaluate the performance by using a confusion matrix to calculate recall, accuracy, f1-score and precision. The results from the paper will help cruise companies have better planning to protect and rescue people from a similar accidents. It will also help customers to make a appropriate decision to increase their chances of survival in case of a similar incident.

Keywords: Binary Classification, Survived, Predictive Models, Algorithms, Logistic Regression, Decission Trees, Random Forests.

1.0 INTRODUCTION

The Titanic shipwreck occurred on April 1912, after the ship hit an iceberg and got submerged in the sea within 3 hours. The accident occurred during the ship's maiden voyage to the United States, with over 2000 passengers and crew aboard. Only 705 passengers were rescued, against the expectation of the builders of the ship. According to the builders, the ship should have stayed for at least 3 days in case of the worst accident.

Machine learning algorithms provide a foundation upon which the titanic data can be explored for classification, and be used to predict the survival of the Titanic passengers. According to the study by [2], predictions using Decission Tree and Naive Bayes algorithms depends greatly on size of the data, the choice of the features and data analysis applied. Despite the good prediction results, the model used was not robust enough as noted by [8], since only two algorithms were compared and the one with the better accuracy used for predictions. According to [4], a combination of the exploratory data analysis and machine learning was used to undertake predictions on the same data set, and the results showed that without feature selection, logistic regression could be used for the predictions with an accuracy of 83%, and use of advanced graphical presentations were recommended.

In this project, we perform exploratory data analysis on the Titanic dataset, and use machine learning algorithms to build a more robust predictive model. Data exploration helps identify the correlation and importance of various features for appropriate selection. The final model is obtained by comparing the performance of 5 models[5]. The three best performing models, are selected and improved using baseline cross-validation and voting classifier.

Purpose of the Project

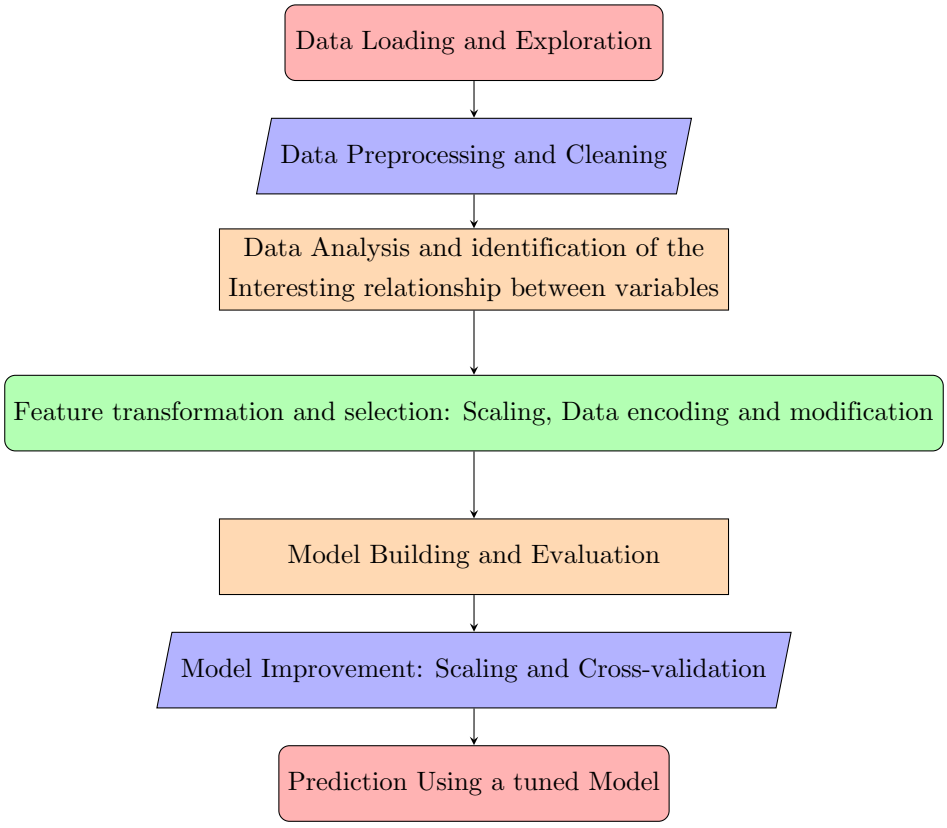
- To create an understanding of how predictive models can be used to identify high-risk factors in fatal accidents by illustrating the significance of scientific models in data analytic and visualization.
- To predict the survival outcomes of the victims given in the test dataset using data exploration and predictive models, thereby creating a means by which victims of accidents can be identified in situations where the identification is problematic.
- To create an understanding of how scientific algorithms can inform decision making, and risk mitigation.

1.1 Problem Description

Binary classification is a supervised learning algorithm that categorizes new observations in one of the two classes [6]. In the project, binary classification was applied to predict whether a passenger survived or not.

A. Steps Followed

The following steps were followed:



B. Model Evaluation - *The Confusion Matrix*

The evaluation for binary classifiers to predict whether a passenger survived or not involves analyzing the performance of the models based on four key metrics: true positives, true negatives, false positives, and false negatives[9].

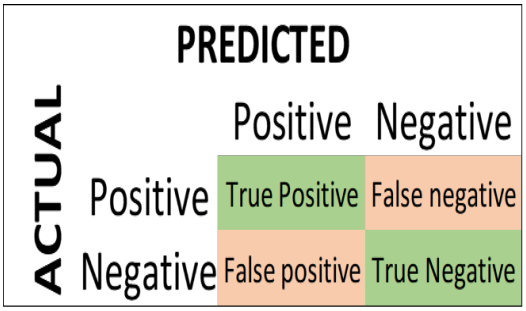


Figure 1: Confusion matrix

- True positives (TP) are the cases where the model correctly predicted that a passenger survived and the passenger actually survived.
- True negatives (TN)are cases where the model correctly predicted that a passenger did not survive and the model correctly identified passengers who actually did not survive.
- False positives (FP) are the cases where the model incorrectly predicted the passenger survived when they actually did not. This represents a type I error, where the model falsely identifies a non-survivor as a survivor.
- False negative (FN) are the cases where the model incorrectly predicted the passenger did not survive when they actually did. This represents a type II error, where the model falsely identifies a survivor as a non-survivor.

C. Performance Evaluation Metrics

The performance evaluation metrics are derived from the confusion matrix parameters [6]

$$\text{Accuracy, } a = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Recall, } r = \frac{TP}{TP+FN}$$

$$\text{Precission, } p = \frac{TP}{TP+FP}$$

$$\text{F1 score, } f = \frac{2rp}{r+p}$$

- Accuracy: The percentage of correct predictions (TP + TN) out of the total number of instances.
- Precision: The proportion of true positive predictions (TP) out of all positive predictions (TP + FP). It represents the model's ability to correctly identify the positive class.
- Recall (Sensitivity or True Positive Rate): The proportion of true positive predictions (TP) out of all instances that actually belong to the positive class (TP + FN). It represents the model's ability to find all positive instances.
- F1 Score: The harmonic mean of precision and recall. It provides a single metric that balances both precision and recall.

The performance of various algorithms was evaluated, and based on the metrics above, the following were used to build the final model:

- Random Forests.
- Decision Trees.
- k-nearest neighbors.

2.0 METHODOLOGY

2.1 Data

In this project, we performed a binary classification on the titanic passengers dataset[2]. The dataset (already split into train and test sets) was taken from the Kaggle website, where a competition had been published [3]. We used python to perform the analysis and build the prediction model. The train set was used to build the model, evaluate, and test its performance. On the other hand, the test dataset consisted of 418 rows, with fewer columns compared to the train set. The Kaggle's test set was used to illustrate the working of the model by predicting whether the passengers survived or not.

2.2 The Steps Followed

Step 1: Definition of Explanatory and Target Variables and Data Exploration

Upon loading the data onto python, we performed data exploration to understand its structure, and recorded the outcomes in figure 2.

Step 2: Data Pre-processing

Some columns like PassengerId, Name, Ticket and Cabin were not needed for the classification task, hence, they were dropped. The missing values for Age, fare and Embarked were filled using appropriate strategies: For Age and Fare median values were used and for Embarked, mode was applied, followed with creation of binary columns for each category. Sex attribute was converted to numerical values(0 for female and 1 for male) for easy manipulation.

Using box plots and histograms for the numerical variables, normality of the distribution and the presence of the extreme outliers were checked and reported in figure 5.

Step 3: Feature Selection

We used the correlation matrix to check on the strength of the association between the numerical explanatory variable and the target variable. From this correlation matrix, the Fare attribute is seen to have a stronger correlation with the target variable as compared to the other explanatory variables as seen in the figure below. Features with very low correlation co-efficient with the target variable were dropped. All the non-numerical features were used in model construction as they were found to be of significance.

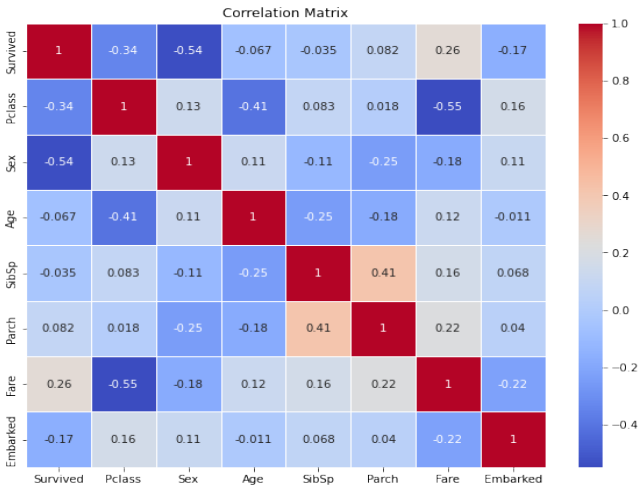


Figure 2: Correlation Matrix for the Numerical Features

Step 4: Model Building and Evaluation

Accuracy is one of the metrics used to asses a model's performance as it measures the proportion of correctly predicted instances in the test set. A higher accuracy indicates better performance. The train set was encoded, scaled and divided in the ratio 8:2 for the model's train and test subsets respectively. We compared accuracy of different models and identified those with the best performance for further tuning.

Step 5: Model Improvement

Improvement of a machine learning model aims at enhancing its performance and generalization capabilities. Various techniques can be used to improve a model, however, we settled on baseline cross-validation to achieve this since other improvement strategies like cleaning and feature engineering, had been applied in building the baseline models. Also, cross-validation was essential in assessing the performance on unseen data, while addressing the possibility of overfitting.

Step 6: Model Fitting and Prediction

After model evaluation, validation and improvement, we fitted the model and predicted using the test dataset. We recorded the results and visualizations after prediction.

3.0 RESULTS

The table below shows the outcome for a search of the various attributes of the train data set.

	PassengerId	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	Cumings, Mrs. John Bradley (Florence Briggs Th...	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, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Table 1: Data Structure

The table above represents data from the train dataset which has 891 rows with 12 columns. For the purposes of the model construction, we divided the train dataset into train and test subsets.

The test dataset provided by Kaggle has 418 rows with 11 columns. It does not have the survived column. The dataset was used as the prediction sample, and the results recorded.

From the above data the explanatory variables are the characteristics of the passengers that can be used to predict the target variable[7]. From our dataset, the explanatory variables include:

- PassengerId: unique identifier for each passenger.
- Pclass: Passenger's class(first, second or third)
- Name: Passenger's name
- Sex: Male or female
- Age: Passenger's age
- Sibsp: Number of siblings or spouses aboard the Titanic.
- Parch: Number of parents or children aboard the Titanic.
- Ticket: ticket number.
- Fare: Ticket price
- Cabin: cabin number.
- Embarked: Port of embarkation(C=Cherbourg, Q= Queenstown, S= Southampton).

The target variable is Survived: indicates whether a passenger survived or not(0=No, 1=Yes). We will use the explanatory variables to predict the survival outcome of each passenger, following the steps used by [1]. We also plotted visualizations of the actual to observe the data before prediction.

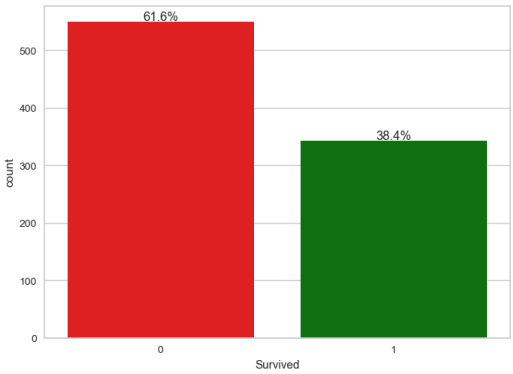
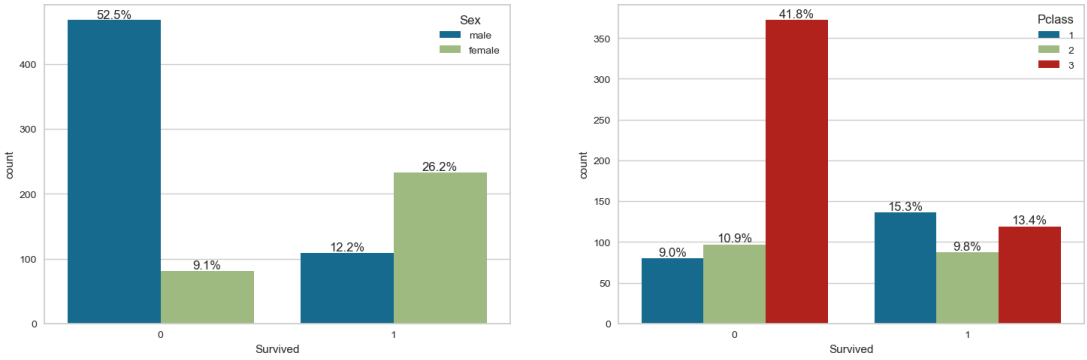


Figure 3: Bar plot of Survivors and Non-survivors from the train dataset



(a) Gender

(b) Passenger Class

Figure 4: Count plots for the Survival Rates based on Gender and Passenger Class

Both the histogram and box plots show that the age data is relatively symmetric, this is because the median is close to the center of the box and the whiskers seem to be of similar length though outliers are present. For the fare data, the data is skewed to the lower values and the whiskers are unequal in length and also has outliers. We also conducted the Shapiro-Wilk test to check for normality and both the Age and fare data are not normally distributed since the p-value is less than the level of significance. Finally we split the combine data back into the train and test data sets based on their original sizes.

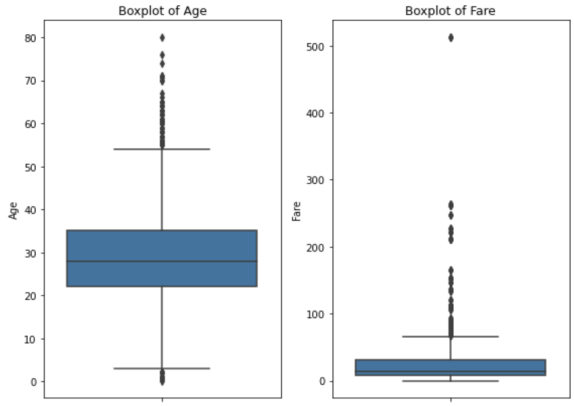


Figure 5: Age and Fare Boxplots

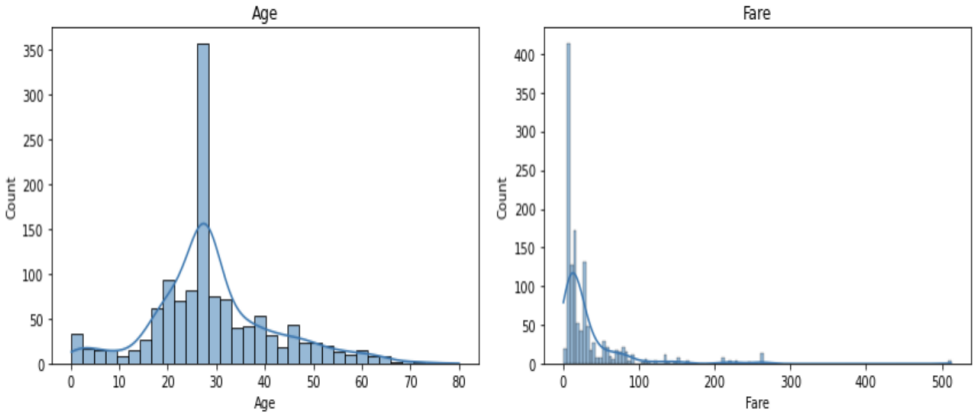


Figure 6: Histogram Plot for Age and Fare Distribution

An extreme outlier was present in the variable 'Fare'. It was intriguing to explore the details of the passenger who paid the outlier fare, and the following details were noted.

```
PassengerId        259
Survived             1
Pclass               1
Sex             female
Age               35.0
SibSp                0
Parch                0
Ticket        PC 17755
Fare          512.3292
Cabin              NaN
Embarked             C
Name: 258, dtype: object
```

Figure 7: Details of the passenger with an outlier 'Fare'

We compared the performance of the following classifiers: Logistic regression, Decision trees, Random forests, KNN, and Naive Bayes.

Based on performance and cross-validation results, Decission Trees, Random Forests and KNN, were selected for further cross-validation and tuning.

	precision	recall	f1-score	support
0	0.85	0.90	0.88	110
1	0.83	0.75	0.79	69
accuracy			0.84	179
macro avg	0.84	0.83	0.83	179
weighted avg	0.84	0.84	0.84	179

The Accuracy score for the RF is: 84%

Figure 8: Random Forests

	precision	recall	f1-score	support
0	0.85	0.79	0.82	110
1	0.70	0.78	0.74	69
accuracy			0.79	179
macro avg	0.78	0.79	0.78	179
weighted avg	0.79	0.79	0.79	179

The Accuracy score Logistic Regression is: 79%

(a) Logistic Regression

	precision	recall	f1-score	support
0	0.85	0.80	0.83	110
1	0.71	0.78	0.74	69
accuracy			0.79	179
macro avg	0.78	0.79	0.79	179
weighted avg	0.80	0.79	0.79	179

The Accuracy score for the NB is: 79%

(b) Naive Bayes

	precision	recall	f1-score	support
0	0.82	0.85	0.83	110
1	0.74	0.70	0.72	69
accuracy			0.79	179
macro avg	0.78	0.77	0.77	179
weighted avg	0.79	0.79	0.79	179

The Accuracy score for the DT is: 79%

(c) Decision Trees

	precision	recall	f1-score	support
0	0.81	0.83	0.82	110
1	0.71	0.68	0.70	69
accuracy			0.77	179
macro avg	0.76	0.75	0.76	179
weighted avg	0.77	0.77	0.77	179

The Accuracy score for the knn is: 77%

(d) KNN

Figure 9: Performance of various Predictive Models

After the model improvement through cross validation, we noted an improvement in performance of most of the models. The aim was to achieve a more robust model[6]. The three models were identified from the baseline performance results and cross-validation performed. The results before and after cross-validation are shown in the table below:

	Model	Baseline Performance	Tuned Performance
0	Random_Forest	84%	83%
1	KNN	77%	82%
2	Decission_Tree	79%	81%

Figure 10: Performance before and after cross-validation

After the cross-validation and improvement, voting classifier was used to build the new model, which was then applied on the test data to produce the predictions based on the passenger Id's. The results were recorded, and the figures below shows the summary.

	precision	recall	f1-score	support
0	0.84	0.91	0.87	110
1	0.83	0.72	0.78	69
accuracy			0.84	179
macro avg	0.84	0.82	0.82	179
weighted avg	0.84	0.84	0.84	179

(a) Performance of the improved model

	PassengerId	Prediction
0	892	perished
1	893	perished
2	894	perished
3	895	perished
4	896	Survived
...	...	...
413	1305	perished
414	1306	Survived
415	1307	perished
416	1308	perished
417	1309	Survived

418 rows × 2 columns

(b) Survival Predictions

Figure 11: Performance and Prediction Results

The following are the visualizations for the for the survival rates based on the overall counts, sex, and passenger class.

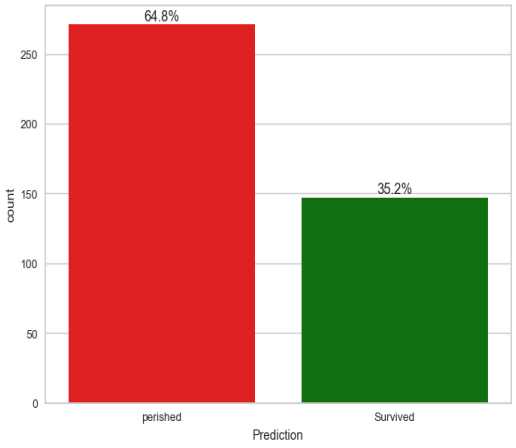
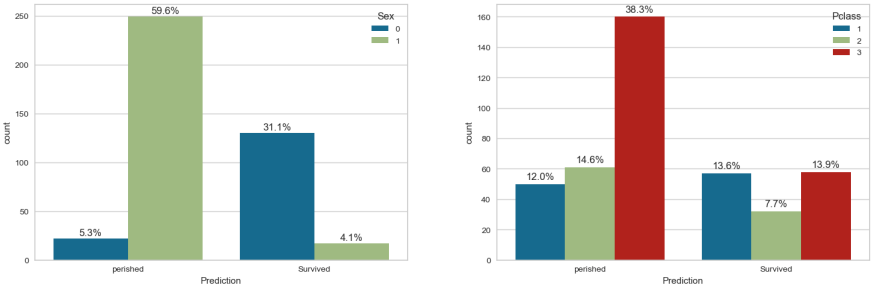


Figure 12: Count plots of Survivors and Non-survivors from the test dataset



(a) Gender (b) Passenger Class

Figure 13: Count plots for the Survival Rates based on Gender and Passenger Class

Comparing the predicted plots and the actual plots, we notice a close similarity in the number of the survived passengers. The result indicates that the voting classifier was able to capture and learn the underlying patterns in the data effectively, resulting in a relatively accurate predictions of survival outcomes (the predictions were 84% accurate).

4.0 CONCLUSION

The improved model (voting classifier) achieved an accuracy of 84%, indicating that it correctly predicted the survival outcome for a significant proportion of the passengers. The accuracy falls within the top 1% of the 2023 scores in the Kaggle leader board for the titanic accident data set.

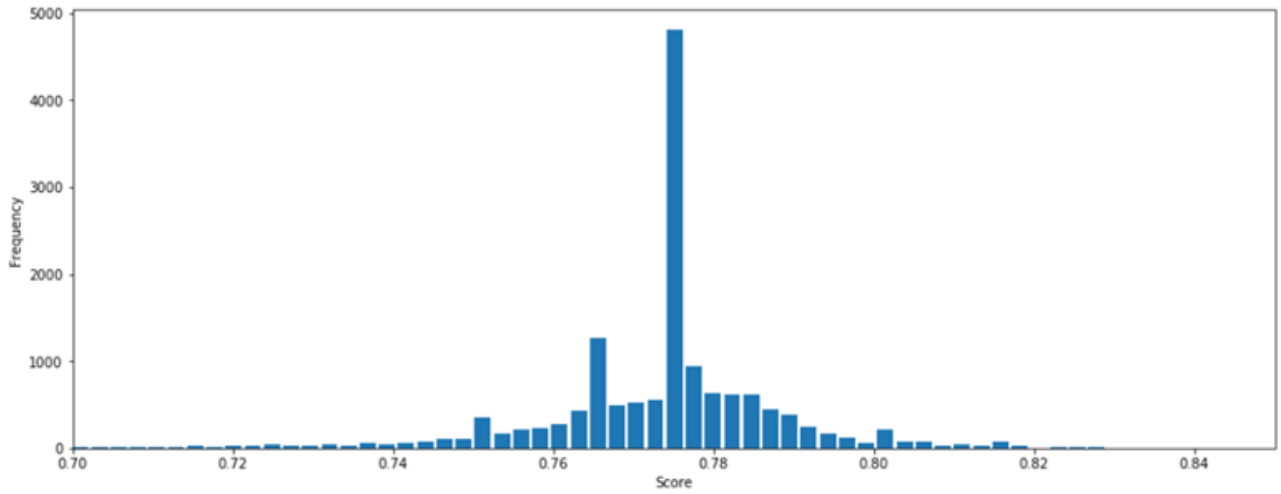


Figure 14: Kaggle Leaderboard scores Distribution (2023) for the Titanic dataset. Source: https://www.kaggle.com/code/carlmcbrideellis/titanic-leaderboard-a-score-0-8-is-great

To achieve this, we built a model based on an ensemble of four machine learning algorithms, followed by a baseline cross-validation process to improve performance. The result demonstrates the effectiveness of the voting classifier algorithm in improving the performance of models, and in capturing the underlying patterns and

relationships within a given dataset. Also, the outcome shows that techniques like scaling and cross-validation can improve the performance and tune the model for better prediction.
By analyzing the feature importance provided by various models, we identified that certain factors, like passenger class, sex, and age, played a crucial role in determining the survival chances of individuals aboard the Titanic.

The predictive analysis of the Titanic dataset has provided valuable insights into the factors that influenced the survival of passengers on the Titanic. Through exploratory data analysis and the application of various machine learning techniques, we have identified several key factors, including gender, age, socio status (Pclass), and ticket class, that strongly correlate with survival outcomes.

However, the Titanic dataset is relatively small, and its characteristics may not be representative of other real-world scenarios. Indeed, the ship had a total of over 2000 passengers, but Kaggle only provided a sample of about 1300 passengers (both train and test datasets), which reduced further due to cleaning. The small sample size might have had a negative effect on the performance of the models, hence low prediction accuracy. Additionally, the available data may be subject to biases and measurement errors that were not thoroughly addressed.

Overall, the predictive analysis of the Titanic dataset demonstrates the power of machine learning techniques to extract insights from data and inform decision-making. As the availability and volume of data continue to increase, we can expect these techniques to become even more valuable in a wide range of fields, from health care and finance to transportation and social services.

References

[1] Sunday Oluwafemi Akinwamide, Taiwo Fele, and Bamidele Ibitayo Faluyi. Exploratory data analysis and machine learning techniques on titanic training dataset. *Journal for Pure and Applied Sciences (JPAS)*, 1(1):18--31, 2022.

[2] Nadine Farag and Ghada Hassan. Predicting the survivors of the titanic kaggle, machine learning from disaster. In *Proceedings of the 7th international conference on software and information engineering*, pages 32--37, 2018.

[3] Shubham Gupta. Titanic dataset. `https://www.kaggle.com/datasets/shubhamgupta012/titanic-dataset`, 2017. Accessed: 20 April 2023.

[4] Yogesh Kakde and Shefali Agrawal. Predicting survival on titanic by applying exploratory data analytics and machine learning techniques. *International Journal of Computer Applications*, 179(44):32--38, 2018.

[5] Roshan Kumari and Saurabh Kr Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7), 2017.

[6] Mukrimah Nawir, Amiza Amir, Ong Bi Lynn, Naimah Yaakob, and R Badlishah Ahmad. Performances of machine learning algorithms for binary classification of network anomaly detection system. In *Journal of Physics: Conference Series*, volume 1018, page 012015. IOP Publishing, 2018.

[7] Shashank Shekhar, Deepak Arora, and Puneet Sharma. Classifying titanic passenger data and prediction of survival from disaster. In *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019*, pages 181--187. Springer, 2021.

[8] Aakriti Singh, Shipra Saraswat, and Neetu Faujdar. Analyzing titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 406--411. IEEE, 2017.

[9] Sofia Visa, Brian Ramsay, Anca L Ralescu, and Esther Van Der Knaap. Confusion matrix-based feature selection. *Maics*, 710(1):120--127, 2011.