Titanic survival Analysis

Nutheti Nikhila Priya
Computer Science & Engineering
PES University
Bangalore, India
nikhila.nutheti@gmail.com

Divya Gannanmaneni
Computer Science & Engineering
PES University
Bangalore, India
g13divya@gmail.com

Yasaswini Madineni
Computer Science & Engineering
PES University
Bangalore, India
yasaswinim17@gmail.com

***Abstract*— The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.. This research is aimed at achieving an exploratory data analysis and understanding the effect or parameters key to the survival of a person had they been on the ship. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After this the results of applying machine learning models are compared and analyzed on the basis of accuracy.The survival prediction has been done by applying various algorithms like Logistic Regression and Support vector machines.**

Keywords—data analytics, titanic, prediction, logistic regression, kaggle dataset , model, xgboost, extra tree classifier

I. Introduction

The Titanic disaster is one of the most famous shipwrecks in world history. Titanic is a British cruise liner that sank in the North Atlantic Ocean a few hours after colliding with an iceberg .Many categories of people of all ages and genders where he was on that fateful night, but luckily that was only a few lifeboats helped. The dead include a large number of men who were given their place by the majority women and children riding.

The inevitable development of technology has both facilitated our life and brought some difficulties with it. One of the benefits brought by the technology is that a wide range of data can be obtained easily when requested. However, it is not always possible to acquire the right information. Raw data that is easily accessed from internet sources alone does not make sense and it should be processed to serve an information retrieval system. In this regard, machine learning algorithms play an important role in this process.

The aim of this study is to get as reliable results as possible from the raw and missing data by using machine learning. Therefore one of the most popular datasets. Titanic is used. This dataset records various features of passengers on the Titanic, including who survived and who didn't. It is realized that some missing and uncorrelated features decreased the performance of

prediction. For a detailed data analysis, the effect of the features has been investigated. Thus some new features are added to the dataset and some existing features are removed from the dataset.Over the years, data of survived as well as deceased passengers has been collected. This dataset has been studied and analyzed using various machine learning algorithms like logistic regression and decision tree.

Machine learning algorithms are used to make predictions on passengers who survived drowning The Titanic. Factors such as ticket boarding, age, gender, category will be used to make predictions. Predictive analysis is the process that including the use of calculation methods to find patterns are important and useful for large data.Using the machine learning algorithms, survival is predicted on different combinations of features.

The paper has been organised as per the following sections: II. Related Works/Literature Review, III Solution Approach; Dataset description, preprocessing, descriptive analysis, and proposed solution approach.

II. Related Works/Literature Review

The observations made by this analysis is that females and children were given more priority than men and the loss of lives was mostly caused due to shortage of lifeboats. In this research paper[1] Logistic Regression has been used for the classification. Before data classification, data pre- processing is done by replacing the missing values with the average of that column and transforming all the raw data into an understandable format.

Analysis of accuracy of the model and performance of the algorithm is done by the confusion matrix. Two attributes are used at a time for the confusion matrix plotting. The values shown in the confusion matrix are the probability of survival of the individual considering only those parameters. Based on the confusion matrix, the logistic regression gave the accuracy of 95%. That’s a good percentage. So, it's good to use logistic regression for classification of data. Finally it was concluded that it works better with the binary dependent variables. Here, ROC curve is also used to depict the performance of different algorithms and helps to decide which algorithm is best with their dataset.

In this research paper[2] the objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After applying they are compared and analysed on the basis of accuracy. At first the data cleaning was done by replacing the missing values with a random sample in the age column and using the median method to find out missing values in the fare column. Also performed exploratory data analysis to figure out the features which would influence the survival rate and use those features for analysis and exclude the remaining. Finding the relationship between each attribute and the survival rate. Feature engineering is also done to this dataset which helps in selecting correct features which are used in making predictions. Bad feature selection may lead to less accurate or poor predictive models.

Initially, it is realized that some missing and uncorrelated features decreased the performance of prediction.They applied multiple logistic regression and logistic regression to check whether a passenger is survived.Here they have used 14 different machine learning techniques, including Logistic Regression (LR), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Support Vector Machines, Decision Tree, Bagging, AdaBoost, Extra Trees, Random Forest (RF), Gradient Boosting (GB), Calibrated GB, Artificial Neural Networks (ANN), Voting (GB, ANN, kNN) and, Voting (GB, RF, NB, LR, kNN).

The reported performance metrics across different cases comparison and concluded that, the maximum accuracy obtained from Multiple Linear Regression is 78.426%; the maximum accuracy obtained from Logistic Regression is 80.756%, then compared the results of Decision tree and Random Forests algorithms for Titanic dataset. Decision tree is resulted 0.84% correctly classified instances, while Random Forests resulted 0.81%, and then compared SVM implementation to 16 classification algorithms and for titanic dataset they achieved 20.81% and 21.27% error rates with neural networks and SVM respectively as minimum errors, then compared Adaboost classifiers to SVM and RBF classifiers. For the titanic dataset, 22.4% error rate is obtained from SVM as the minimum error rate. Also used SVM as a component classifier for Adaboost. They used a titanic dataset as one of the experimental data and the minimum error rate they obtained is 21.8%.

For comparing these 14 algorithms accuracy, F-measure are used. Both the accuracy and F-measure are highest for Voting (GB, ANN, kNN). (Whereas accuracy was same for both Voting and Gradient boosting). So it is concluded that the proposed model can predict the survival of passengers and crew with 0.82 F-measure score with Voting (GB, ANN, kNN).

Here, four machine models are used for the analysis. They are 1) Logistic Regression 2) Decision Tree 3) Random Forest and 4) Support Vector Machine. After analysing the data in each model, the model evaluation is done using Confusion Matrix drawn and the accuracy is calculated for each machine learning model. The Logistic Regression machine model got the highest accuracy among all four. So, it is concluded that this model is best suitable for this dataset. The accuracy of the models may vary when the choice of feature modelling is different. Ideally logistic regression and support vector machines are the models which give a good level of accuracy in case of classification problems.

[3]Before building a model, data exploration is done to determine the factors that can be beneficial while creating the classifier for the prediction. Few plots are plotted to get an overall idea on the relations between the attributes. Here, they have used four machine learning algorithms to predict the model accuracy. They are 1) Logistic Regression 2) Decision Tree 3) Random Forest. All of these algorithms are compared to one another on the basis of accuracy. The observed accuracies of Logistic Regression, Decision Tree and Random Forest are 94.26% ,93.06%, 91.86% respectively.

For comparing the four algorithms two measures are used. Those are accuracy and false discovery rate. The observed false discovery rates of Logistic Regression, Decision Tree and Random Forest are 7.90%, 9.26%, 10.66% respectively. To consider an algorithm as the best, the accuracy score should be high and false discovery rate should be low. By these observations, it has been proven that Logistic Regression is the best algorithm with high accuracy and low false discovery rate.

III. Solution Approach

We performed the data analysis and visualisation on our dataset titanic_train.csv to understand and analyse the relationships that were existing among the attributes.

A. Dataset Description

The dataset used in this analysis is titanic_train.csv from Kaggle. The dataset contained the information about the passengers on the ship including name, gender, ticket, age, passenger class etc..along with the target variable ‘survived’, which tells if the particular passenger has survived or not.

B. Pre- processing and Data Cleaning

The dataset contained 7 numerical attributes and 5 categorical attributes.We checked for the missing data across the entire dataset and found that two attributes- ‘age’ and ‘cabin’ had many NaN values.

Roughly 20 percent of the Age data is missing.

The proportion of Age missing is likely small enough for reasonable replacement with some form of imputation. Looking at the Cabin column, it looks like we are just missing too much of that data to do something useful with at a basic level.So we decide to drop the 'Cabin' column.

For 'age' we decide to check for the mean values for all the passenger classes and fill the missing data with the mean of the passenger classes accordingly. Because the age of the passengers follows normal distribution and many of them belonged to the age group 17-34.

I.CORRECTING

Reviewing the data, there does not appear to be any aberrant or non-acceptable data inputs. In addition, we see we may have potential outliers in age and fare. However, since they are reasonable values, we will wait until after we complete our exploratory analysis to determine if we should include or exclude from the dataset. It should be noted, that if they were unreasonable values, for example age = 800 instead of 80, then it's probably a safe decision to fix now. However, we want to use caution when we modify data from its original value, because it may be necessary to create an accurate model.

II.COMPLETING

There are null values or missing data in the age, cabin, and embarked field. Missing values can be bad, because some algorithms don't know how-to handle null values and will fail. While others, like decision trees, can handle null values. Thus, it's important to fix before we start modeling, because we will compare and contrast several models. There are two common methods, either delete the record or populate the missing value using a reasonable input. It is not recommended to delete the record, especially a large percentage of records, unless it truly represents an incomplete record. Instead, it's best to impute missing values. A basic methodology for qualitative data is impute using mode. A basic methodology for quantitative data is impute using mean, median, or mean + randomized standard deviation. An intermediate methodology is to use the basic methodology based on specific criteria; like the average age by class or embark port by fare and SES. There are more complex methodologies, however before deploying, it should be compared to the base model to determine if complexity truly adds value. For this dataset, age will be imputed with the median, the cabin attribute will be dropped, and embark will be imputed with mode. Subsequent model iterations may modify this decision to determine if it improves the model's accuracy.

III.CREATING

Feature engineering is when we use existing features to create new features to determine if they provide new signals to predict our outcome. For this dataset, we will create a title feature to determine if it played a role in survival.

IV.CONVERTING

Last, but certainly not least, we'll deal with formatting. There are no date or currency formats, but data type formats. Our categorical data is imported as objects, which makes it difficult for mathematical calculations. For this dataset, we will convert object data types to categorical dummy variables.

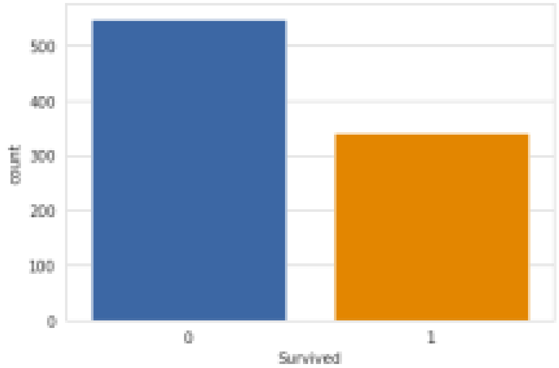


Fig. 1.countplot showing the count of survived and non survived

Survivors are less when compared to the survival ratio. This Tells many of them that they could not survive the Titanic.

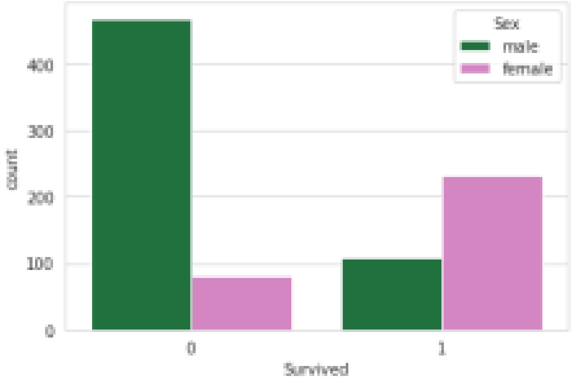


Fig. 2.countplot showing the deaths with hue varibale 'sex'

Here we can obtain the insight that many of the non-survivors are males. Amongst the survived females are more than the males.Most of men could not survive, this can be due to the reason that male were trying to save females and children and lost lives.

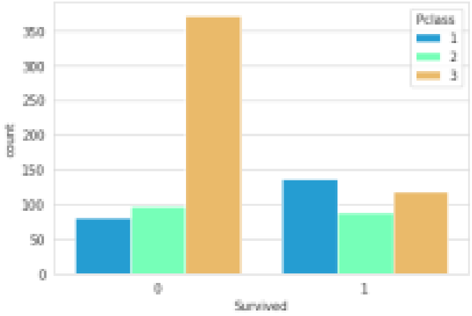


Fig. 3.countplot showing the deaths with hue variable as 'Pclass'

The passengers from the passenger class 3 are more in number when compared to passenger class 1 and 2. It tells that many passengers on the ship belonged to poor and below middle class families. The death rate of the passenger class 3 people is enormously higher than 2 and 1.Most of the class 3 passengers could not survive and rich people from class 1 somehow managed to survive using their power or authority in getting the lifeboats. Class 3 passengers were probably given least priority to get life boats and hence many deaths.

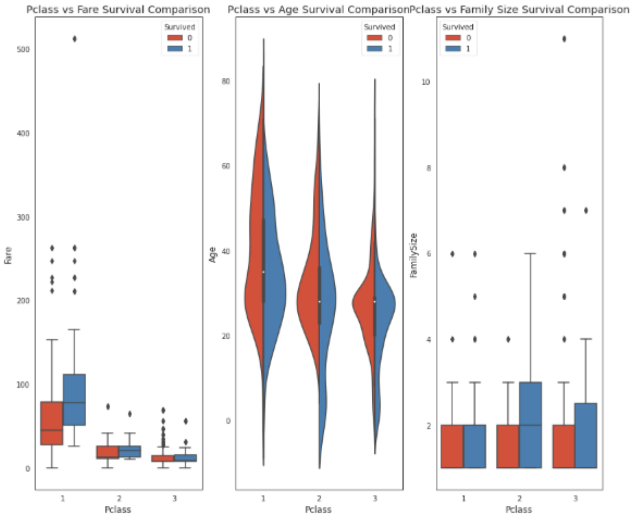


Fig. 4.violinplot and boxplot showing the survived wrt Pclass with Age, Fare and Familysize

We observe that there are many deaths in pclass 3 when compared to the survival rate. The non survivors mostly belonged to the age class 20 to 30 as it can be observed in violinplot.

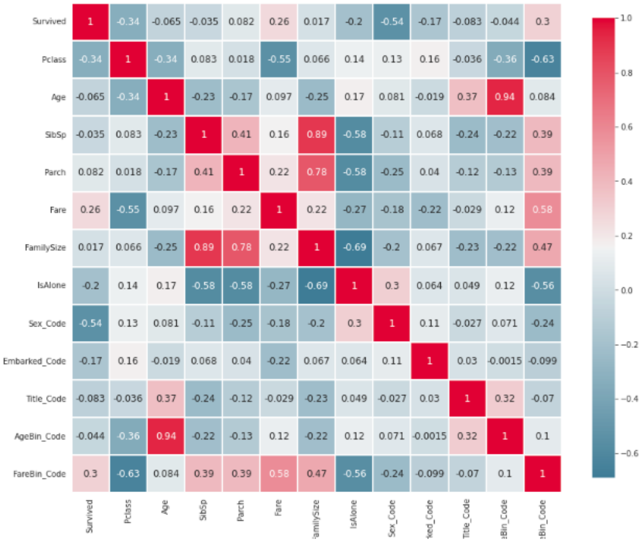


Fig. 5.Pearson correlation of all features

We see there are quite enough strongly correlated and few badly correlated and we can work on those strongly correlated features to get more accuracy and get more insights.

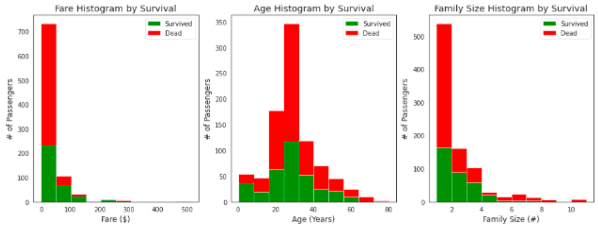


Fig. 6.Fare, Age, Familysize histplot by Fare

It is histplot of fare vs Age we see that follows a normal distribution. Also there is a passenger who had a fare of around 500 which seems to be an outlier.

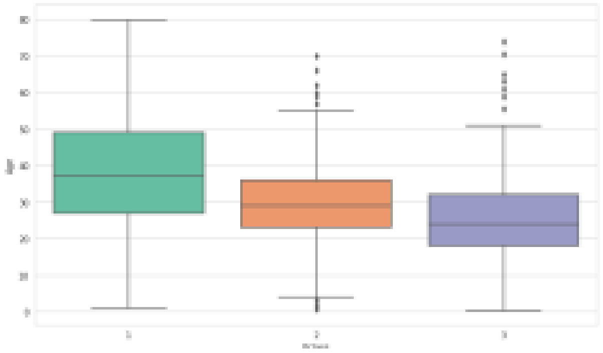
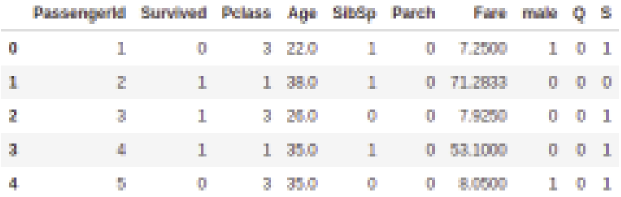


Fig. 7.boxplot of all three passenger classes

We want to fill in missing age data instead of just dropping the missing age data rows. One way to do this is by filling in the mean age of all the passengers (imputation). However we can be smarter about this and check the average age by passenger class.We used a function to replace the NaN values.For passenger class 1 age=37, class 2=29 and class 3=24.

Here in class 1 passengers most of the people belonged to all the age groups, wherein class 2 had age groups below 57 and class 3 passengers were younger with few older people.



	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	1	0	3	22.0	1	0	7.2500	1	0	1
1	2	1	1	38.0	1	0	71.2833	0	0	0
2	3	1	3	26.0	0	0	7.9250	0	0	1
3	4	1	1	35.0	1	0	53.1000	0	0	1
4	5	0	3	35.0	0	0	8.0500	1	0	1

Fig. 8.final dataset given to the model for training

We converted the categorical variable like age and embark into dummy variables so that it gets easier for the model to analyse.We also dropped the attributes 'Sex', 'Embarked', 'Name', 'Ticket' and appended the dummy variables S, Q and male. So this is our final dataset with the target variable as ‘survived’.

4.MODEL DATA

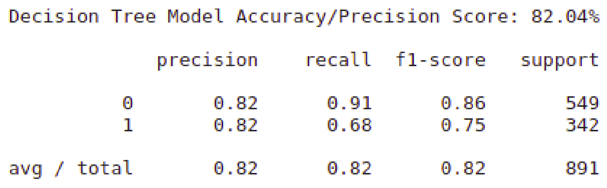
There are three main fields in the data science,they are mathematics,computer science and business management.These three have equal importance in data science.Machine learning can be classified as supervised,unsupervised and reinforced learning.In this, we are doing supervised learning.Because we are training our algorithm by presenting it with a set of features and their corresponding target. We then hope to present a new subset from the same dataset and have similar results in prediction accuracy. Since our problem is predicting if a passenger survived or did not survive, this is a discrete target variable. We will use a classification algorithm from the *sklearn* library to begin our analysis. We will use cross validation and scoring metrics, discussed in later sections, to rank and compare our algorithms' performance.

5.EVALUATE PERFORMANCE

We have to decide how to make our model better on every level. We can consider this as a binary problem because there are only two possible outcomes.Those are whether the passenger survived or died. So we have a 50-50 chance of surviving.So without any information about the dataset,we can always get 50% with a binary problem.But we need to do it better because we have total information about our dataset. According to the information there, 67.5% of people died. So, we can consider 68% as bad model performance.To show this, we're going to build our own decision tree model, because it is the easiest to conceptualize and requires simple addition and multiplication calculations. When creating a decision tree, you want to ask questions that segment your target response, placing the survived as1 and dead as 0 into homogeneous subgroups.

When we used sklearn Decision Tree (DT) Classifier, we accepted all the function defaults. This leaves the opportunity to see how various hyper-parameter settings will change the model accuracy. In order to apply this to our model, we need to actually understand it. Some advantages of Decision trees are that they are Simple to understand and to interpret. Trees can be visualized,requiring little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values, the cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree, easy to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable. See algorithms for more informat.This is also able to handle multi-output problems.

Decision Tree can handle both continuous and categorical variables.Handles nonlinear parameters efficiently**:** Non linear parameters don't affect the performance of a Decision Tree unlike curve based algorithms. So, if there is high nonlinearity between the independent variables, Decision Trees may outperform as compared to other curve based algorithms. This uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.It is possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model. This Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.



```
Decision Tree Model Accuracy/Precision Score: 82.04%

             precision    recall  f1-score   support

          0       0.82      0.91      0.86       549
          1       0.82      0.68      0.75       342

avg / total       0.82      0.82      0.82       891
```

This includes some of the disadvantages.Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble. They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret.

In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.Little bit of noise can make it unstable which leads to wrong predictions.If data size is large, then one single tree may grow complex and lead to overfitting. So in this case, we should use Random Forest instead of a single Decision Tree.There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best in present times.It can be used to solve regression, classification, ranking, and user-defined prediction problems.Advantages of xgboost includes it has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why, XGBoost is also called a regularized form of GBM (Gradient Boosting Machine).

XGBoost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model. XGBoost has an in-built capability to handle missing values. When XGBoost encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node. It then does the same when working on the testing data.XGBoost allows users to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we have to run a grid-search and only a limited value can be tested.

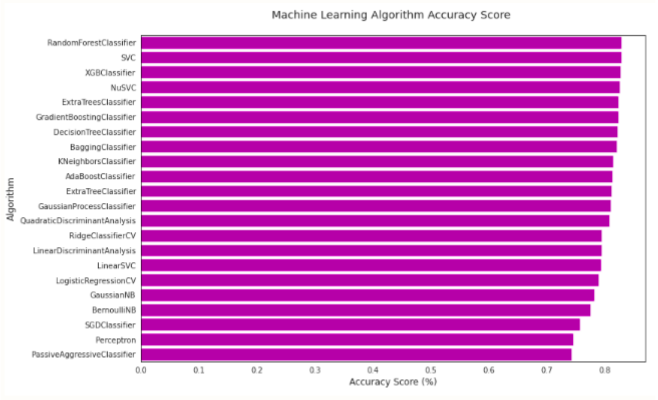


Fig. 9. Plot showing how accurate different models perform on the training and test dataset.

Some observations from the dataset to improve the accuracy: Our sample survival is different from our population of 68%. If we assumed everybody died, our sample accuracy is 62%.

Majority (81%) male passengers died. Majority (74%) female passengers survived. This gives us an accuracy of 79%.

If we consider classes,the majority (97%) of class 1 survived and the majority (92%) of class 2 survived. Class 3, is even at a 50-50 split. No new information is gained to make our model better.

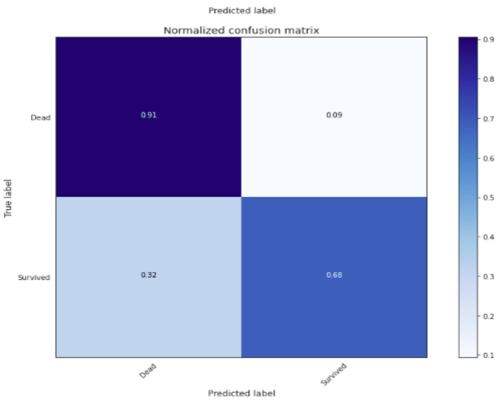


Fig. 10. Confusion matrix with normalisation

If we consider embark from ports C,Q and S.The passengers embarked from ports C and Q, the majority still survived, so no change. Also, the dead subgroup is less than 10, so we will stop there. The passengers embarked from port S, the majority (63%) died. So, we will change passengers who come under females, class 3, embarked from port S from assuming they survived, to assuming they died. Our model accuracy increases to 81%.

we get to 82% accuracy. On a worst, bad, good, better, and best scale, we'll set 82% to good, since it's a simple model that yields us decent results.

Now, let's code what we just wrote above. Please note, this is a manual process created by hand. It's important to understand it before we start working with MLA.

6.CROSS VALIDATION OF MODEL PERFORMANCE

Here, we use a different subset for train data to build our model and test data to evaluate our model. Otherwise, our model will be overfitted. Cross Validation is basically a shortcut to split and score our model multiple times, so we can get an idea of how well it will perform on new data. It’s a little more expensive in computer processing, but it's very important to do that. In addition to Cross Validation, we used a sklearn train test splitter, to allow a little more randomness in our test scoring.

7.CONCLUSION

Using the same dataset and different implementation of a decision tree (adaboost, random forest, gradient boost, xgboost, etc.) without tuning we got an accuracy of 0.7655 and with tuning does not exceed the submission accuracy of 0.77990. Interesting for this dataset, the simple decision tree algorithm had the best default submission score and even with tuning achieved the same best accuracy score. Some additional steps that may be taken to improve one's score/accuracy could be:

i.Implementing a good cross-validation strategy in training the models to find optimal parameter values

ii.Introduce a greater variety of base models for learning. The more uncorrelated the results, the better the final score

References

[1] https://bit.ly/3vmzViN

[2] https://www.researchgate.net/profile/Yogesh-Kakde/publication/325228831_Predicting_Survival_on_Titanic_by_Applying_Exploratory_Data_Analytics_and_Machine_Learning_Techniques/links/5c068f63a6fdcc315f9c0bb9/Predicting-Survival-on-Titanic-by-Applying-Exploratory-Data-Analytics-a nd-Machine-Learning-Techniques.pdf

[3]]http://sajrest.com/Archives/vol4issue4_2019/v4i4p2.pdf [4] https://www.researchgate.net/profile/Neytullah-Acun/publication/32490954_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset/links/607533bc299bf1f56d51db20/A-Comparative-Study-on-M achine-Learning-Techniques-Using-Titanic-Dataset.pdf