

Titanic Survival Classification and Modeling using Machine Learning

A Technical Report on IMAT5322 Coursework Submitted to The Faculty of Computing, Engineering and Media, De Montfort University

**damilare jinadu student**

**March 12, 2023**

**student id: p2733425**

**Introduction**:

The Titanic was a ship disaster that got sunk and got some passengers killed. There exist conclusions regarding the cause of the sinking and the analysis of the available data on what impacted the survival of passengers hence, continues. (Haque, Shivaprasad and Guruprasad, 2021)

This coursework presents the Titanic dataset of 891 passengers aboard and requires the prediction of survival or no survival of these passengers based on the available raw data through machine learning and feature analysis.

The available raw data contains missing and uncorrelated features which could impact the performance of prediction. In other to enhance the accuracy of prediction, it’s important to carry out some specific tasks as itemised below;

* The effect of the features would be investigated for a detailed data analysis.
* Some new features added to the dataset while some also removed from existing features to enhance the accuracy of prediction.
* Specific model selection for the classification of the Titanic test data to check the passengers who survived or not would be made (i.e., Logistic Regression, SVC, KNN, Random Forest, etc.).

Hitherto, several classifications and predictions had been made on the Titanic Dataset using different models. According to (Chatterjee, 2018) during his application of multiple logistic regression and logistic regression to check whether passengers survived, reported performance metrics across different cases and determined that the maximum accuracy achieved from Multiple Linear Regression is 78.426%; for Logistic Regression is 80.756%. (Basu, 2021) compared the results of Decision tree and Random Forests algorithms for Titanic dataset. He reported that the Decision tree classified result was 0.84%, while Random Forests resulted 0.81%.

(Meyer et al., 2003) compared SVM implementation for Titanic dataset with 16 classification algorithms and they achieved %20.81 and %21.27 error rates with neural networks and SVM respectively as minimum errors. (Rätsch, Onoda and Müller, 2001) also compared Adaboost classifiers to SVM and RBF classifiers and achieved %22.4 error rate from SVM as the minimum error rate. (Li, Wang and Sung, 2008) also contributed to the body of knowledge by using SVM as a component classifier for Adaboost.

This study intends to classify and predict the Titanic dataset with Logistic regression, Random Forest, SVC, KNN, CatBoost, Lightgbm, Xgboost and Extra Trees.

To achieve the above tasks, the following outlined processes shall be implemented.

1. Data Loading/Checking – Checking the available features of the data and its feature correlation.
2. Data pre-processing (i.e., Data cleaning and Data normalisation); This entails the cleaning and transformation of the raw data in preparation for analysis.
3. Data visualisation: Representation of the data in diagram, charts etc to draw insight about the data.
4. Data analysis: The extraction of meaningful information and conclusion on the data with different analytic tools.
5. Feature extraction: This is construction of a combination of variables to predict the target variable while still describing the data with sufficient accuracy.
6. Model Selection: This involves the application of clustering and classification models to predict the survival rate on the **Titanic Dataset**
7. Experimental results analysis: Result of analysis conducted i.e.: codes ran to generate different results.

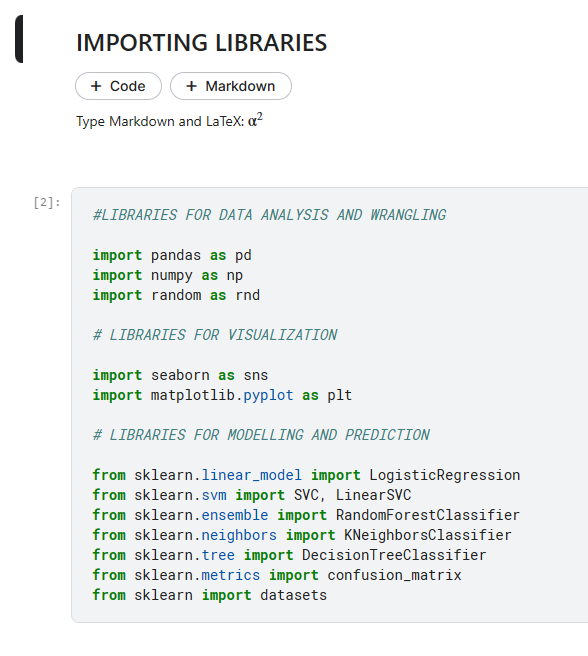
The dataset variable features for this task are as shown below:

|  |  |  |
| --- | --- | --- |
| Feature Type | Feature | Feature Values |
| Target Variable | Survived | If survived or not (0=No, 1=Yes) |
| Numerical Variable | PassengerId | Passengers’ Unique identity |
|  | Age | Age in years |
|  | SibSp | Number of Siblings/Spouses Aboard |
|  | Parch | Number of Parents/Children Aboard |
|  | Fare | Passenger Fare (British Pound) |
| Strings | Name | Name of Passenger |
|  | Cabin | Cabin Number |
|  | Ticket | Ticket Number |
| Categorical Variables | Pclass | Passenger Ticket Class (1 = 1st, 2 = 2nd, 3 = 3rd) |
|  | Sex | Sex (Male or Female) |
|  | Embarked | Port of Embarkation (C= Cherbourg, Q = Queenstown, S = Southampton) | |

*Table 1: Titanic Dataset Variables*

**DETAILED APPROACH**

1. **DATA LOADING/MANIPULATION**



The modules needed to execute the codes are first loaded i.e., libraries for data analysis, data visualization and libraries for modelling and prediction.

**Fig.1: Importing Libraries for Data Manipulation**

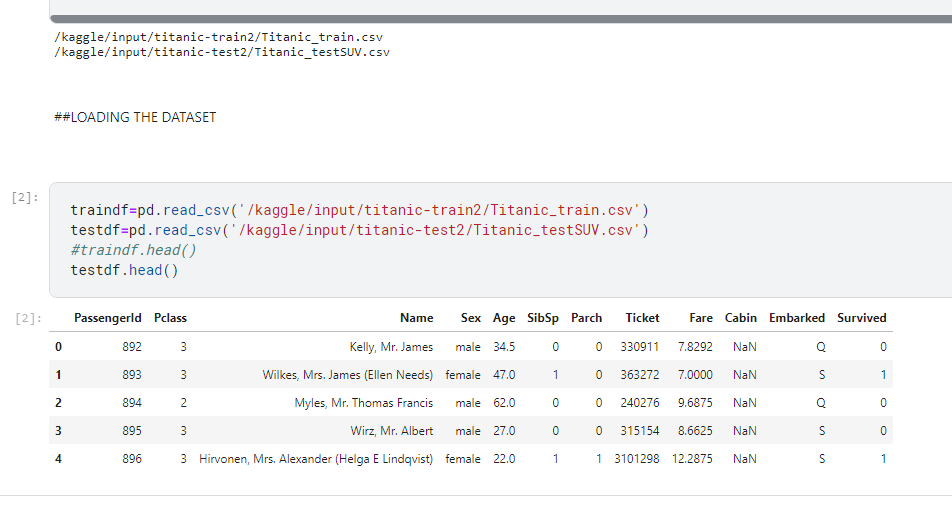
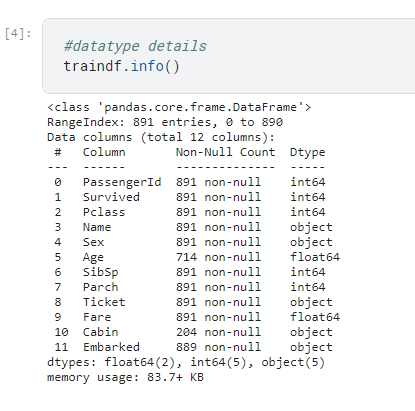


Figure 3: Checking the variables for missing values

Figure 2: Loading of Dataset

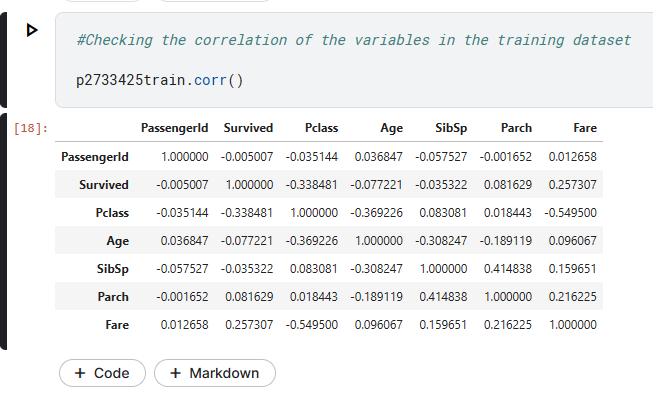
The titanic train dataset contains 891 rows and 12 columns while the titanic test dataset has 418 rows and 11 columns, hence a total of 1309 observations which makes the training dataset a 68.1% of the entire data. The survived is the target/dependent variable to be used to model and predict the available dataset.

Figure 4: Checking the variable correlation.

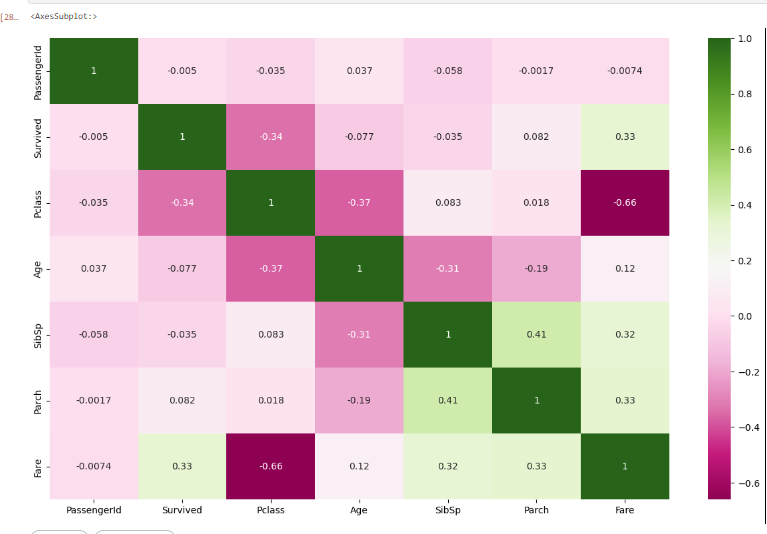


Figure 5: Correlation Matrix for PassengerId, Pclass, Age, SibSp, Parch, Fare against Survived,

From the correlation matrix above, the features that have some correlation with the survived feature will be determinants in modeling and prediction (i.e., Pclass, Age are somewhat correlated (negatively) and the Fare is positively correlated with the Survived and this shows that these set of features will be significant to the prediction model.

1. **DATA CLEANING**

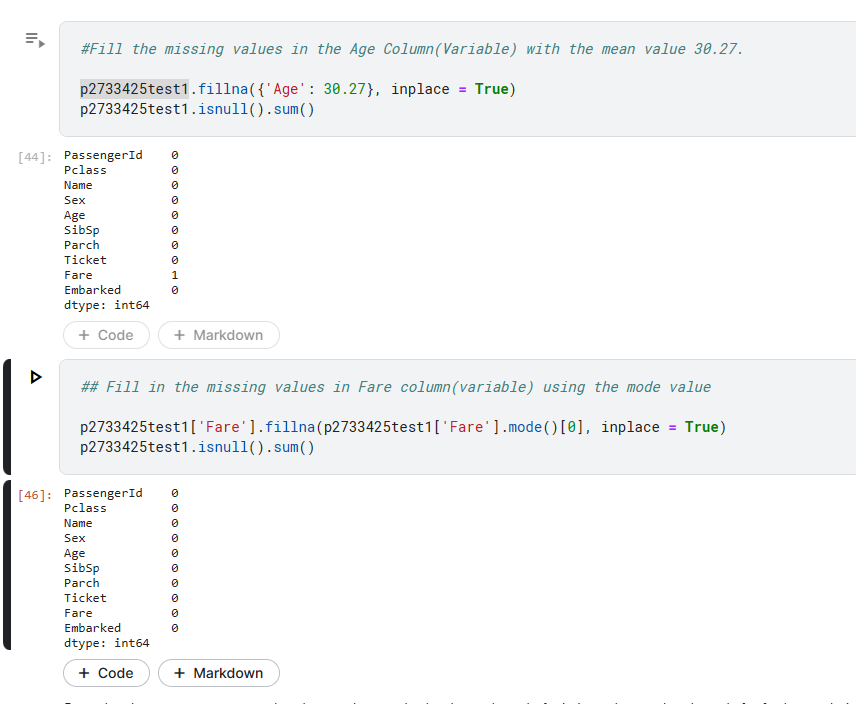
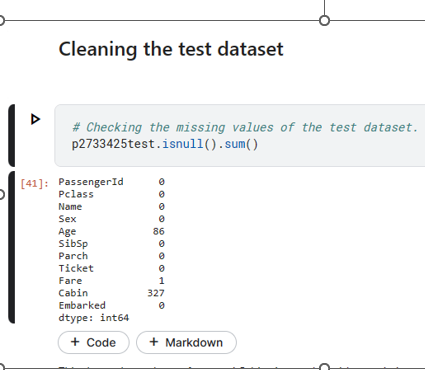
Missing values in the dataset were resolved with respect to their various features and importance to the prediction model.

Figure 6: Checking the variables for missing values

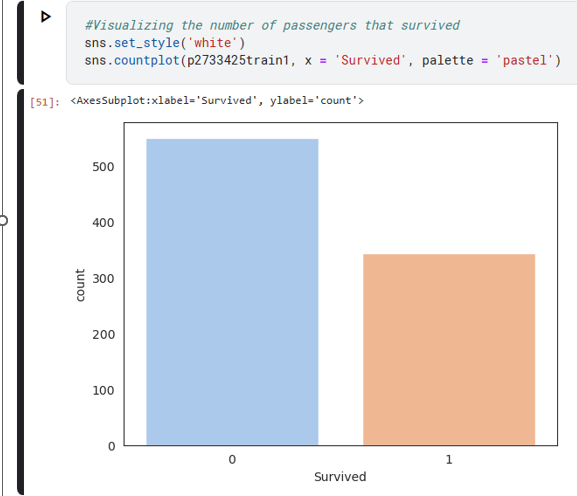
Figure 7: Filling the missing values of Age and Fare Variable

The above diagram shows that (Age, Cabin, Fare and Embarked) consist of missing values with Cabin having the highest volume of missing values. We can remove or replace the missing values in the dataset.

**Decision on the features with missing values**

The Cabin will be dropped due to the large volume of missing values in the feature, and it is not significantly important in our model prediction.

The Age missing values will be replaced with the mean value of the Age features. Age is a significant feature to the prediction model.

The Embarked features is also perceived to be significantly important to the prediction model and as such, the missing value in the features will be filled, replaced by the most frequent value (i.e., Mode).

1. **DATA VISUALIZATION**

This shows that in the train dataset containing a maximum of 891 variable counts (i.e., passengers), more than 500 passengers did not survive while less than 400 passengers survived.

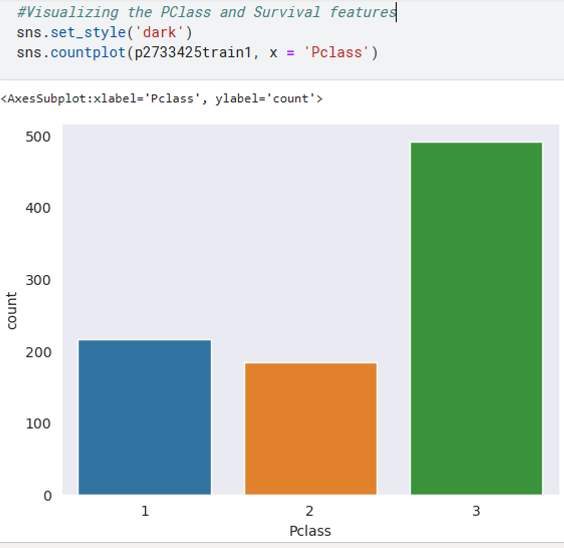


Figure 8: Survived Feature(Target Variable)

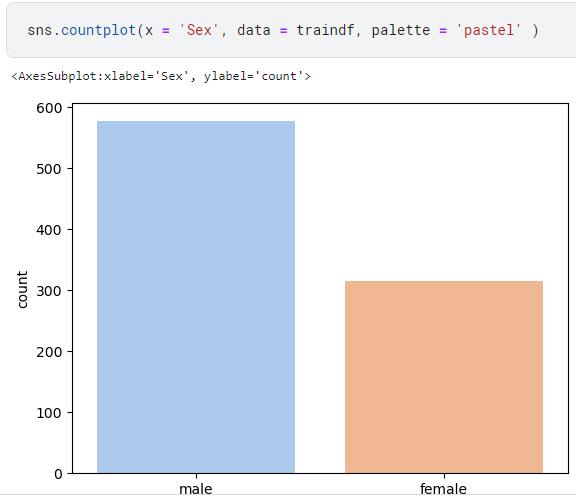


Figure 10: Sex features in relation to survival

Figure 9: Pclass features relative to Survival.

*Fig 9* shows there’s relationship between the Pclass and Survived features. The 3rd Class passengers are less likely to survive while the 1st Class passengers are more likely to survive.

*Fig 10* shows that Sex is an important feature as male passengers are less likely to survive, and the females are more likely to survive. More than 500 Male did not survive while less than 100 Females did not survive.

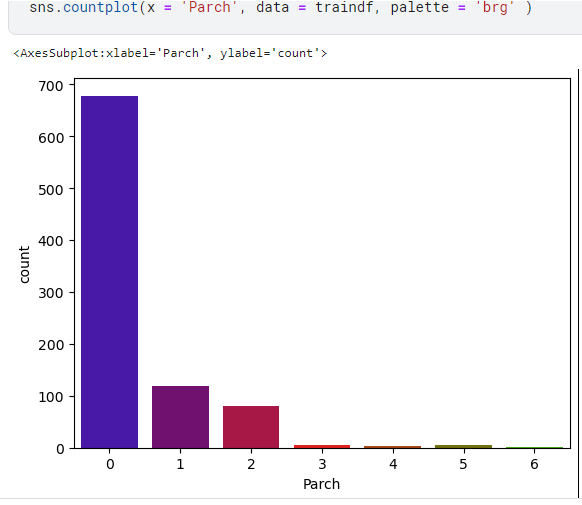


Figure 2: Parch bar chart showing rate of survival in Titanic Dataset

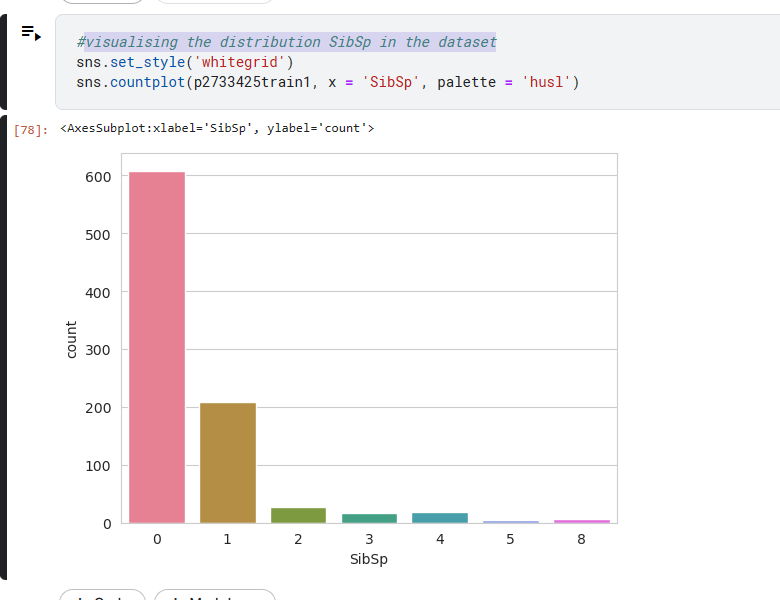
**

Figure 11: Sibsp bar chart showing rate of survival in Titanic Dataset

The SibSp figure above shows the count of passengers with Siblings and Spouse, which shows that more passengers travelled alone or with at least one dependant.

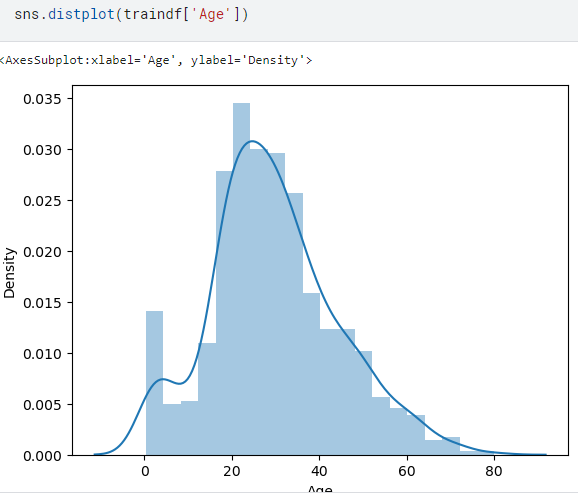
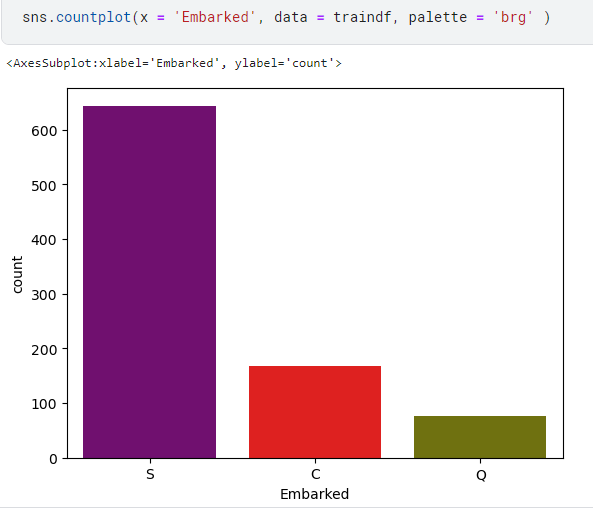


Figure 14: Age distribution

Figure 13: Embarked bar chart showing the effect of embarkation on survival

The distribution of the Age which shows a normal distribution and Average Age is between 20 and 30 with more younger people in the dataset.

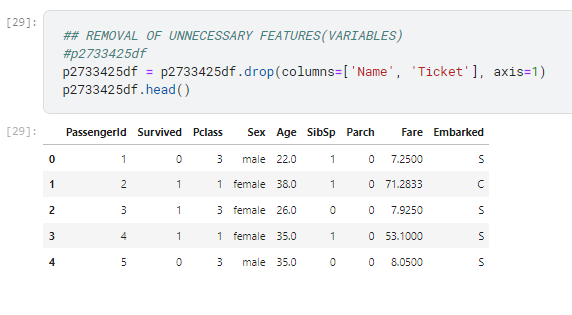
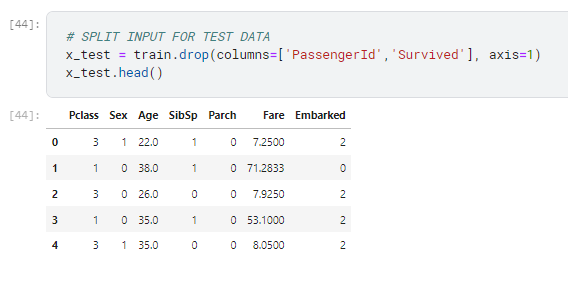
 Figure 25: Dropping of PassengerId and Survived

Figure 15: Dropping of Name and Ticket Variable

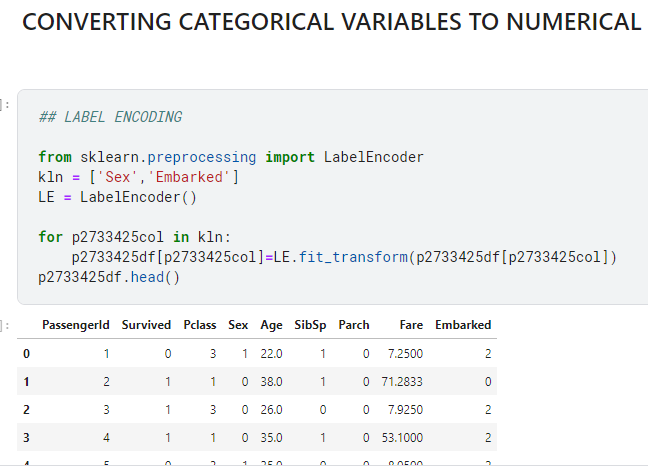


Figure 16: Converting Categorical Variable to Numerical

Converting the Sex and Embarked Categorical variables into numerical variable.

Sex has only 2 categories (Male and Female)

Embarked has 3 categories (S, C, Q)

**RESULTS ANALYSIS**:

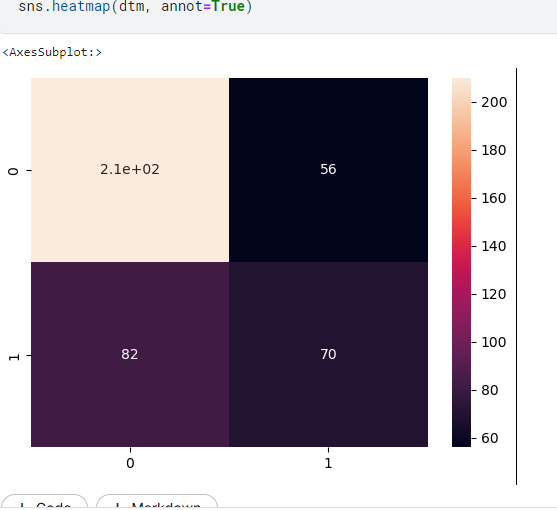
Confusion Matrix

Figure 18: KNN Classification report analysis

Figure 17: Confusion matrix of Linear KNN

**KNN Confusion Matrix Analysis:**

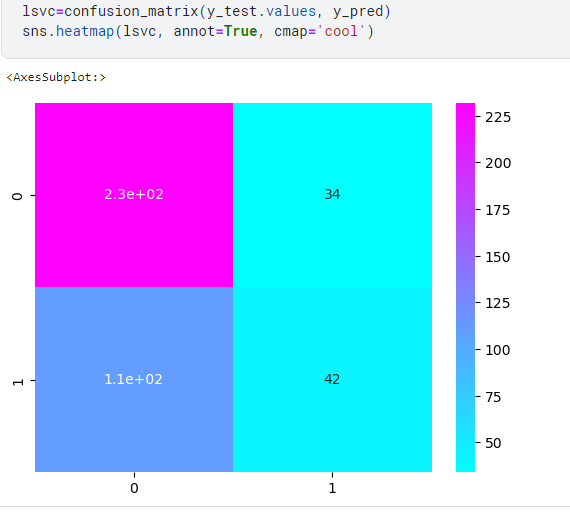
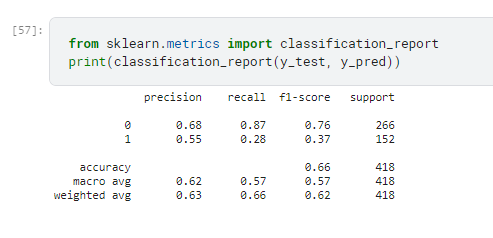
210 passengers are confirmed not survived (False Positive), 70 are confirmed Survived (True Positive), 56 Passengers reported not Survive but survived (False Negative), and 82 Passengers reported Survived but did not survive. (True Negative)

Figure 0: Classification Report on SVC

Figure 19: confusion matrix on SVC

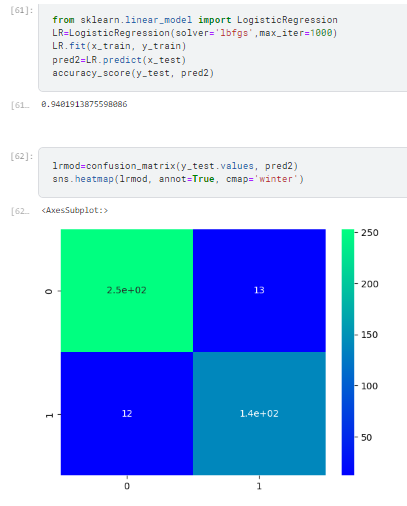
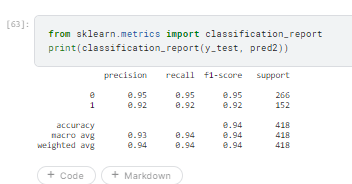
 **Logistic Regression Confusion Matrix**

Figure 22: Classification Report on Logistic Regression

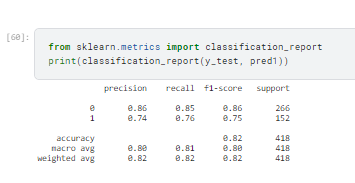
250 passengers didn’t survive (False +ve),140 confirmed Survived (True +ve), 13 Passengers reported not Survive but survived (False \_ve), and 12 Passengers reported Survived but did not survive (True \_ve).

Figure 21: Confusion Matrix on Logistic Regression

Figure 24: Classification Report on Random Forest

**Random Forest Confusion Matrix**

230 passengers didn’t survive (False +ve),120 confirmed Survived (True +ve), 40 Passengers reported not Survive but survived (False \_ve), and 36 Passengers reported Survived but did not survive (True \_ve).

Figure 23: Confusion Matrix on Random Forest Classifier



Figure 26: Model Training

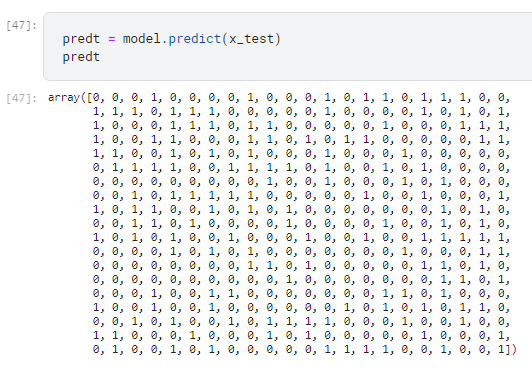
Figure 28: XGBoost, LightGBM, catBoost Classifiers



Figure 27: Decision Tree, Random Forest, Extra Tree Classifiers



Figure 29: KNN Classifier



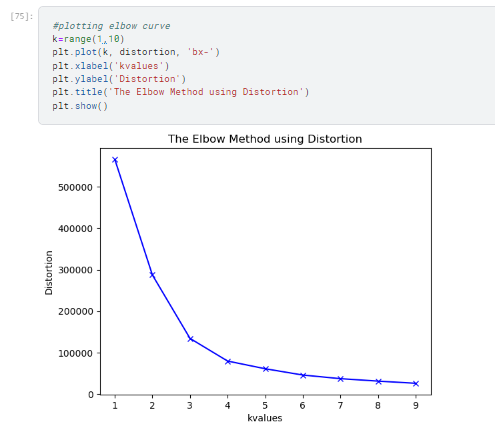
Figure 30: Model Cluster on Titanic Test Dataset

Figure 2: K-Mean Distortion (K=3)

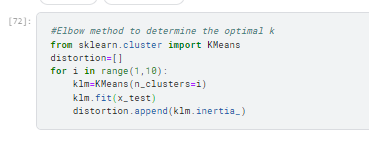


Figure 31: Determining K value of KMean clustering technique.

**Experimental Result**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ALGORITHM | SURVIVED | PRECISION | RECALL | F-SCORE | SUPPORT | ACCURACY |
| KNN | 0 | 0.72 | 0.79 | 0.75 | 266 | 0.67 |
|  | 1 | 0.56 | 0.46 | 0.50 | 152 |  |
| LINEAR SVC | 0 | 0.68 | 0.87 | 0.76 | 266 | 0.66 |
|  | 1 | 0.55 | 0.28 | 0.37 | 152 |  |
| RANDOM FOREST | 0 | 0.87 | 0.85 | 0.86 | 266 | 0.82 |
|  | 1 | 0.75 | 0.77 | 0.76 | 152 |  |
| LOGISTIC REGRESSION | 1 | 0.95 | 0.95 | 0.95 | 266 | 0.94 |
|  | 0 | 0.92 | 0.92 | 0.92 | 152 |  |

*Table 2: Model performances on Titanic Test Dataset*

The above Algorithms are evaluated based on accuracy and F-Score to analyse the likelihood of survival and deduce the features in correlation with survival of passengers. Some reasonable adjustments were made during modelling (e.g., the K factor in KMeans as seen from the Elbow method gave a value of 3, for KNN, number of neighbors selected is 10 and for Random Forest Tree estimator, 100 was selected)

In summary Logistic Regression classifier gave the highest F1 score for both classes (i.e., survived or not survived) indicating best balance between precision and recall, while the KNN and SVM classifiers have room for improvement in predicting one of the two classes.

**DISCUSSION AND CONCLUSIONS**

Obtaining valuable results from raw data with missing values could be quite daunting but machine learning and feature engineering methods have reduced the gap in analysis of raw data to a reasonable accuracy. I have proposed models for predicting the survival of passengers in the Titanic disaster.

Firstly, I conducted a detailed data analysis and investigated the features and correlation of variables to the target (i.e., Survived). During data preprocessing, missing values were handled by scaling numerical features and encoding categorical variables and some features were excluded e.g., Cabin, Name, Ticket, which were deemed less relevant and/or no correlation with Survived feature.

Secondly, training of some machine learning algorithms (i.e., KNN, SVM, Logistic regression and Random Forest) were conducted and performance evaluation was carried out via confusion matrix and classification report analysis.

The proposed model i.e., Logistic Regression predicted the survival of passengers with 0.95 F-score and 94% accuracy on the test dataset.

In conclusion, comparison of four different machine learning techniques were used to analyze Titanic dataset and improved some features (i.e., normalization and replacement of missing values) that impacted the classification results and was able to identify a robust technique among the four compared.

**References**

Basu, A. (2021). *Titanic dataset Analysis (~80% accuracy)*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/titanic-dataset-analysis-80-accuracy-9480cf3db538.

Chatterjee, T. (2018). Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms. *International Journal of Emerging Research in Management and Technology*, 6(6), p.1.

Haque, M.A., Shivaprasad, G. and Guruprasad, G. (2021). Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. *IOP Conference Series: Materials Science and Engineering*, 1065(1), p.012042.

Li, X., Wang, L. and Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21(5), pp.785–795.

Meyer, D., Karatzoglou, A., Leisch, F., Buchta, C. and Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, [online] 55(1-2), pp.169–186. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0925231203004314 [Accessed 15 Jun. 2021].

Rätsch, G., Onoda, T. and Müller, K.-R. . (2001). Soft Margins for ‘AdaBoost,’ Machine Learning. *Machine Learning*, 42(3), pp.287–320.