www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.03
TPI 2019; SP-8(4): 30-37
© 2019 TPI
www.thepharmajournal.com
Received: 02-02-2019
Accepted: 12-03-2019

Navneet KR Pandey
Department of CSE, Accurate Institute of Management & Technology, Greater Noida, Uttar Pradesh, India

Ashish Jain
Department of CSE, Accurate Institute of Management & Technology, Greater Noida, Uttar Pradesh, India

Correspondence
Navneet KR Pandey
Department of CSE, Accurate Institute of Management & Technology, Greater Noida, Uttar Pradesh, India

Predicting survival on titanic through exploratory data analysis and logistic regression: Unraveling historical patterns and insights

Navneet KR Pandey and Ashish Jain

DOI: https://doi.org/10.22271/tpi.2019.v8.i4Sa.25259

Abstract
The research paper on the survival of passengers on the Titanic employs statistical techniques to investigate the factors that played a role in their survival. The study reveals that gender, age, and social class significantly impacted the likelihood of survival. The paper also acknowledges the limitations in the data and potential biases. The findings underscore the significance of access to lifeboats and the actions of crew members during emergencies and can potentially inform future disaster preparedness strategies. This study contributes valuable insights into the socio-economic factors that shaped the disaster and establishes the significance of thorough analysis and interpretation of historical data.

Keywords: GGPLOT, confusion matrix, feature engineering, random forest, model evaluation, logistic regression, data mining

Introduction
The sinking of "The Titanic," which took place on April 15, 1912, is history's most notorious catastrophe. Many Titanic components were destroyed in the collision with the iceberg. On that fatal night, there were many different classes of individuals of every age and gender present, but it was unfortunate that there weren't many lifeboats available to save them. The numerous women and children on board took the place of the many males who were among the deceased. The second-class male passengers were already dead.

It is predicted using machine learning algorithms which passengers were still alive when the Titanic sank. The projections will be based on factors such as ticket price, age, sex, and class. The predictive analytics process uses computational techniques to identify prominent and useful patterns in large quantity of data. Based on various feature combinations, survival is predicted using logistic machine learning techniques.

The goal is using exploratory data analytics to explore and expand knowledge on different information from the dataset given and using it to understand how each field affects passengers survival by applying analytics between dataset field and the "Survival" field. By using a machine learning algorithm, predictions are made for more recent data sets. The most accurate model is recommended for predictions after algorithms have been compared for accuracy. The correctness of the data analysis will be evaluated using the implemented algorithms. The most accurate model for making predictions is suggested after many algorithms have been compared for accuracy.

The importance of this problem lies in its potential to inform future safety protocols for the maritime industry, as well as to provide insights into the factors that affected passenger’s survival on the Titanic. By using modern data science and machine learning techniques to analyze historical data, we can gain a deeper understanding of the events that led to most tragic maritime disasters in history. Furthermore, the application of machine learning algorithms to survival prediction tasks has important implications for various other domains, such as healthcare, finance, and social science.

The problem is significant because it provides an opportunity to gain insights into the factors that affected passenger’s survival on the Titanic, and to inform future safety protocols for the maritime industry [1]. By analyzing historical data using modern data science and machine learning techniques, we can gain a deeper understanding of the events that led to one of the most tragic maritime disasters in history.

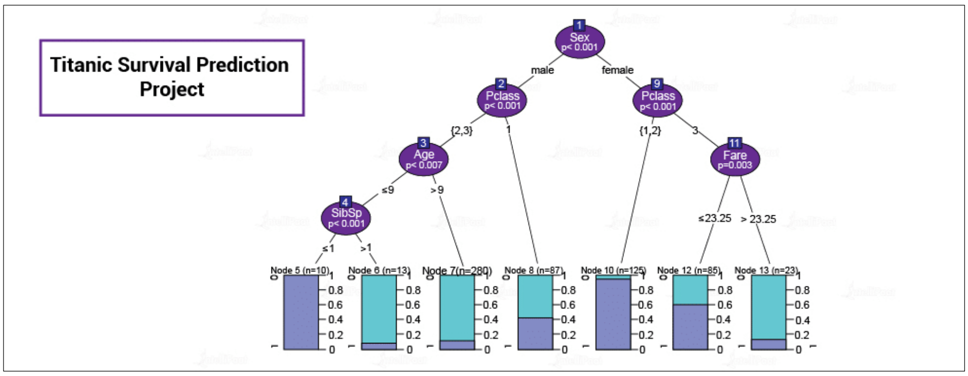


Fig 1: Decision Tree

Literature Survey

The Titanic was a British passenger liner that sank in the North Atlantic Ocean on 15 April, in 1912 following a harmful collision with a huge iceberg on its journey starting from Southampton to New York. Estimated of 1,500+ people were lost in the disaster, making it one of the fatal maritime disasters in history.

Since this disaster, there have been numerous studies and analyses of the factors that influenced survival rates among the passengers and crew. Below are some examples of literature related to the Titanic survival project

“**The Titanic and the Indifferent Stranger**”**:** A Study of the Lifeboat by John J. Pershing – This book examines the lifeboats on the Titanic and the decision-making process behind their deployment. The author argues that the lack of urgency and concern from both passengers and crew played a major role in the high number of casualties.

“**The Titanic Disaster**”**:** An Enduring Example of Money Management vs. Risk Management by H. Paul Jeffers – This book explores the financial decisions that contributed to the Titanic disaster, as well as the implications for risk management in modern business.

“**Women and Children First**”**:** The Myth of Titanic by Judith B.

Geller - This book challenges the notion that “women and children first” was the policy followed during the Titanic disaster, arguing that the survival rates of men and women were influenced by factors such as social class and access to lifeboats.

“**Titanic: The Story Lives On**!” By David G. Brown – This book provides an overview of the Titanic disaster and the subsequent investigations and inquiries that followed. It includes accounts from survivors and discusses the ongoing fascination with the Titanic in popular culture.

“**The Titanic**”**:** Enduring Legacy of a Disaster by John Maxton- Graham – This book offers a comprehensive look at the Titanic disaster, including the ship’s construction, the events leading up to the collision, and the aftermath. It also explores the impact of the disaster on maritime safety regulations and popular culture.

“**Titanic: Women and Children First**” by Judith B. Geller – This book focuses specifically on the experiences of women and children on the Titanic, including their survival rates and the challenges they faced in escaping the sinking ship.

“**Titanic: Voices from the Disaster”** by Deborah Hopkinson – This book draws on first-hand accounts and archival material to provide a detailed and immersive narrative of the Titanic disaster, including the experiences of individual passengers and crew members.

“**Titanic: A Survivor’s Story by Colonel Archibald Gracie**” – This memoir by a survivor of the Titanic disaster gives a detailed account of the events which led to the sinking including author’s own experiences during and after the disaster.

"Titanic: Machine Learning from Disaster" by Kaggle, a popular data science competition platform. This study provides a detailed analysis of the Titanic dataset and evaluates various machine learning algorithms for survival prediction, including logistic regression, decision trees, and random forests.

"Predicting the Fate of Titanic Passengers" by Dimitris Zervakis and Theophilos Papadopoulos. This study explores the role of feature selection and imputation techniques in improving survival prediction accuracy and compares the performance of several machine learning algorithms, including support vector machines and neural networks.

"Titanic”**:** Survival Prediction with Machine Learning" by Shi tong Luo and Feng Chen. This study uses a stacked ensemble model to combine multiple machine learning algorithms for survival prediction, and evaluates the impact of different feature sets on model performance.

"A Comparative Study of Machine Learning Algorithms for Titanic Survival Prediction" by Rana Attallah, Hatem A. Fayed, and Ahmed M. Khedr. This study evaluates the performance of several machine learning algorithms, including decision trees, support vector machines, and naive Bayes, for survival prediction on the Titanic dataset.

"Data Mining and Machine Learning in Disaster Survival Prediction: A Systematic Review" by Arif Sari and Ismail Rakip Karas. This study provides a comprehensive review of machine learning applications in disaster survival prediction, including studies on Titanic survival prediction.

“**Titanic Survival Analysis using Logistic Regression**” by Vaishnav Kshirsagar, Nahush Phalke: This paper provides a quick insight in the methodology used in our paper namely logistic regression and confusion matrix to helps us analyze the relationship between the factors we have considered in out research paper.

“Developing a Titanic survival scorecard: Risk analysis of populations through statistical scoring methods” by Dominic Vincent Light, CirroLytix Research Services: In this paper the authors employ the use of great mathematical model know as statistical scoring method in short is works like a scorecard to keep track of the association by determining a benchmark which to should be met before any further analysis can be drawn from the pair of similar effects on the event.

Overall, these studies provide valuable insights into the application of machine learning algorithms for survival prediction on the Titanic dataset and highlight the importance of feature selection, model selection, and imputation techniques in improving prediction accuracy.

Proposed System

The proposed system for the Titanic Survival Project is to develop advanced data analysis and machine learning models that predict the likelihood of passenger survival based on a wide range of factors. The project will involve several stages, which includes data pre- processing and cleaning, EDA, feature engineering, model selection and training, and optimization and model evaluation.

The proposed system will use a combination of both supervised and unsupervised machine learning algorithms to create a predictive models of passenger survival. Supervised learning algorithms, such as logistic regression and random forest, will predict passenger survival based on labelled training data. Unsupervised learning algorithms, such as K-means and Apriori Algorithm, will be formulated to identify patterns and correlation in the data that can be used to improve model performance.

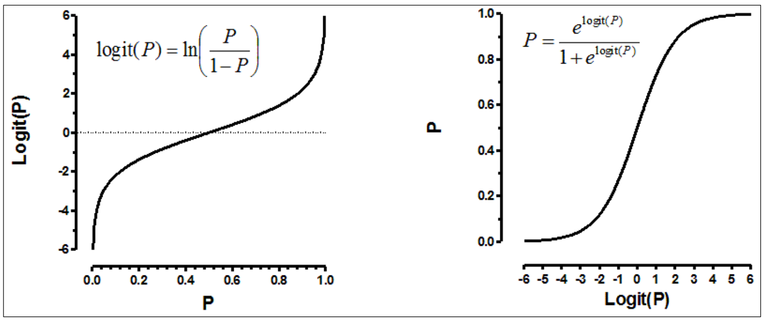
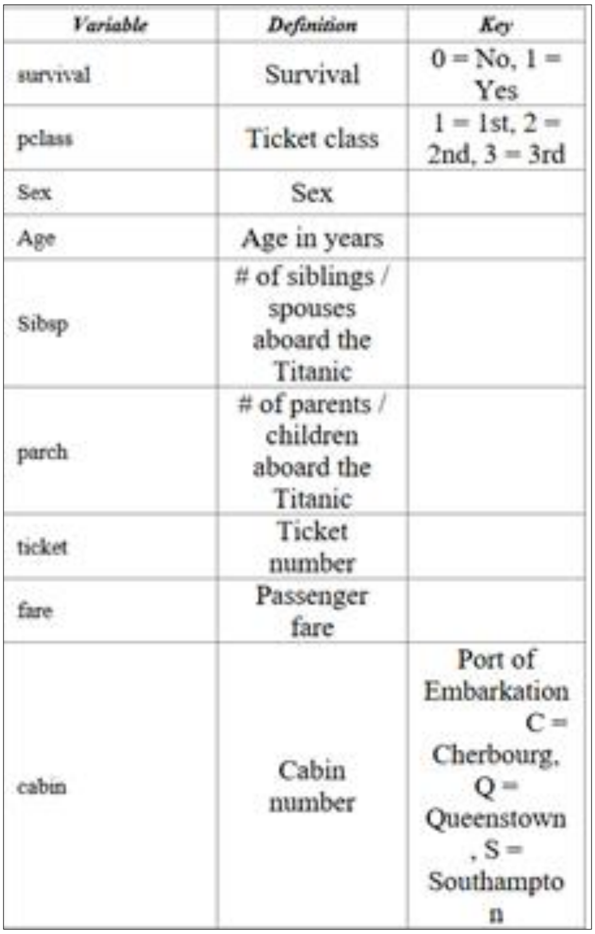


Fig 2: Sigmoid Curve for Logistic Regression

The data we use for our proposed test system is the Titanic. Data set for text data preprocessing and graph analysis.

A dataset with the following data dictionary

Table 1: Variable in the Data set



Variable	*Definition*	*Key*
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	Port of Embarkation C = Cherbourg, Q = Queenstown , S = Southampton

Methedology

The methodology used in the Titanic Survival Project involves techniques such as data exploration and preparation, feature engineering, model selection, hyperparameter tuning, and performance evaluation. Many machine learning algorithms namely logistic regression, XGBoost, also random forests were employed, the project was carried out using Python programming language and popular data analysis libraries such as pandas, NumPy, and scikit- learn.

1. Data collection: Start by collecting data related to the Titanic disaster. This could include passenger data, such as age, sex, class, and cabin location, as well as information about the circumstances of the sinking.

2. Data preprocessing: Clean and preprocess the data to make it accurate and ready for analysis. This might involve compensating for NULL values, normalizing the data, and converting the variables into appropriate formats.

3. Feature engineering: Extract relevant features from the data that may be useful in predicting survival. This involve generating new variables based on existing variables, such as the size of family or the fare per person.

4. Model selection: Choose a suitable machine learning model that can accurately predict survival based on the available data. This could include logistic regression, XGBoost, and random forests.

5. Model training: Training the selected model using the cleaned preprocessed data ensure the desired results and the engineered features further elevate the specificity of the required output.

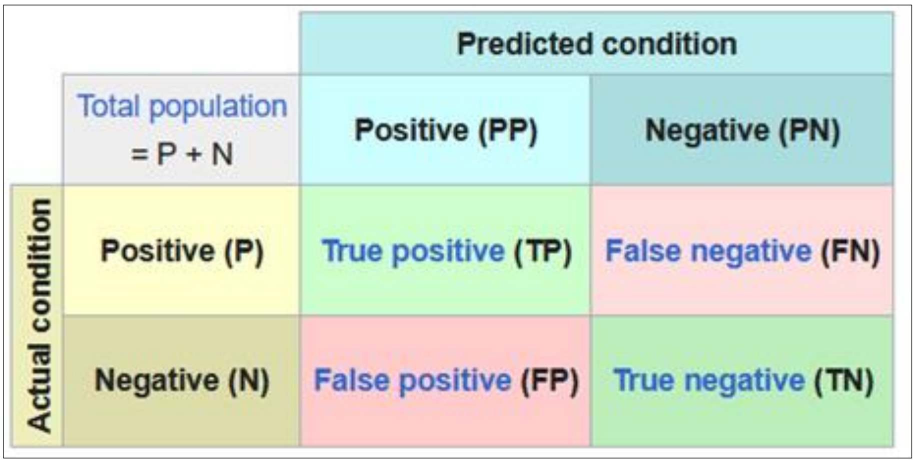
6. Model evaluation: Further evaluating the performance of the trained model using suitable metrics such as accuracy of output set, testing data accuracy, precision, recall, or Z score.

7. Interpretation: Interpret the results of the model to understand which variables are the most important in predicting survival. This could involve visualizations or feature importance scores.

Because logistic regression predicts probabilities, not just categories, we can fit it using probabilities. For each training data point, we have a feature vector xi and a class of observations yi. The class probability is p if yi= 1 and 1 − p if yi = 0. Then the probability is [1].

LL (*ββ*0, *ββ*) = ⎬ *pp* (*xxii*) *yyii* (1 – *pp* (*xxii*)1−*yyii*
ii=1

Fig 3: Algorithm of the Probability



		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Fig 4: Hypothesis Testing

A. Feature Engineering in Input dataset

Crucial phase of data analytics is feature engineering. It deals with formulating predictions and choosing the attributes that are given in training. In this, the dataset is searched for particular features that can be utilized to develop machine learning models. It aids in modelling the dataset and comprehending it. An inaccurate or subpar prediction model may result from inadequate feature selection. The selection of the appropriate characteristics affects both accuracy and predictive power. It eliminates all the extraneous or pointless elements [4].

The following variables are employed based on the exploratory study mentioned above: cabin, sex, age, Pclass, title, family size (parch plus sibsp columns), embarked, and fare. The result column is the survival column.

The selected features were opted for their ability to significantly influence survival rates and will be used as the independent variable in the bar-plots. Selecting incorrect features can result in inaccurate predictions, even if a good algorithm is used. Thus, feature engineering plays a crucial role in developing an accurate predictive model by serving as its backbone.

B. Logistic Regression

Logistic regression is a useful method for analyzing data that involves a binary or categorical dependent variable. It allows for the examination of the relationship between one dependent binary variable and one or more independent variables of varying levels, such as nominal, ordinal, interval, or ratio. In real-world applications, logistic regression is frequently used to address binary classification problems. For instance, it can be used in spam detection to determine whether an email is spam or not, in the medical field to determine whether a tissue lump is benign or cancerous, and in marketing to predict whether a user will purchase an insurance product.

C. Decision Tree

A set of options and their results are graphically represented by a decision tree. It may be applied to simulate different situations and assess the optimal course of action. There are nodes, branches, and leaves in a decision tree. Nodes are places in a graph where decisions are made, or questions are posed. The possibilities or solutions are depicted as branches. The ultimate findings or outcomes are represented by leaves. Decision trees may be used to visualize trade- offs in each choice and simplify difficult situations [7].

As shown in Fig 5.

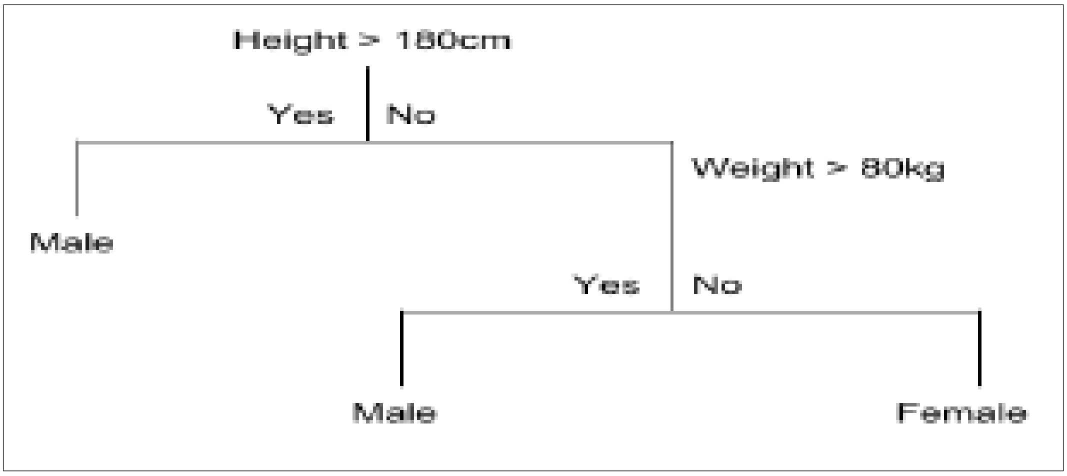


Fig 5: Sample Decision Tree of (Height and Weight)

D. Random Forest

It is a machine learning algorithm which can be used as a guardrail for classification and regression tasks. It works by forming many decision trees from randomly selected subsets of the training dataset, and then merging their predictions using a voting or averaging scheme.

Random Forest has several advantages over other algorithms, such as high accuracy, robustness to noise and outliers, and low risk of overfitting. However, it also has some drawbacks, such as high computational cost, difficulty in interpreting the results, and lack of control over the tree structure.

Model Evaluation

Model evaluation is a technique of assessing how well a machine learning model performs on a given task. It involves many measuring metrics, namely accuracy judgment, precision, Z-score, etc., these reflect the quality of the model's predictions and consistency. Model evaluation is crucial for selecting the best model among different candidates, tuning the model's hyperparameters, and identifying the model's strengths and weaknesses.

A. Confusion Matrix

It is a useful tool for the purpose of evaluation of classification model’s accuracy. It helps in comparing the predicted results of the model with the actual data, providing a clear indication of the number of correct and incorrect predictions. The matrix is of size N by N, with N values, making it an effective way to assess the performance of a model. As a result, the confusion matrix is commonly used for evaluating the effectiveness of classification models.

Sensitivity: Sensitivity is a term that refers to the accurate identification of true positives, and it complements the false negative rate. It can be calculated by dividing the true positive by the sum of true negative and false positive. A sensitivity score of "1.0" is considered optimal, while "0.0" is the lowest possible score.

Specificity: It measures the percentage of accurately identified negatives and is a complement to the false positive rate. True negatives/ (true negatives plus false positives) equals specificity. Specificity has a range from "0.0" to "1.0," with "1.0" being the ideal value.

Positive Predictive Value: The metric is a measure of the test's effectiveness and is determined by dividing the true positives (where the prediction and outcome are both correct) by the sum of true positives and false positives (where the prediction is wrong despite the outcome being correct).

Negative Predicted Value: The negative predictive value (NPV) is a measure that helps in assessing the effectiveness of a diagnostic test. It represents the probability that a negative test result is indeed accurate. The NPV is calculated by dividing the number of true negative results (where the test predicted negative and the actual result was also negative) by the sum of true negative and false negative results (where the test predicted positive despite the actual result being negative).

Accuracy: It provides a proportion of correctly predicted events based on the model or algorithm. In terms of values, "1.0" is the best and "0.0" is the worst.

Implementaion in Python

Here we have implemented the model building and cross verification process in Jupyter Notebook using Python Libraries. The Python contains libraries such as Pandas, numpy and matplotlib.

The notebook contains multiple tabs that present the analysis of research data. These tabs include Survival by Age, Gender Survival, Cabin Survival, P Class Survival, Survival vs. Tickets, Graph Between Name Survival, Survival v/s Family, Departure v/s Survival. In addition, one of the tabs showcases predictive analysis information, categorized under Logistic Regression, Decision Trees, and Random Forests [7].

Results and Output

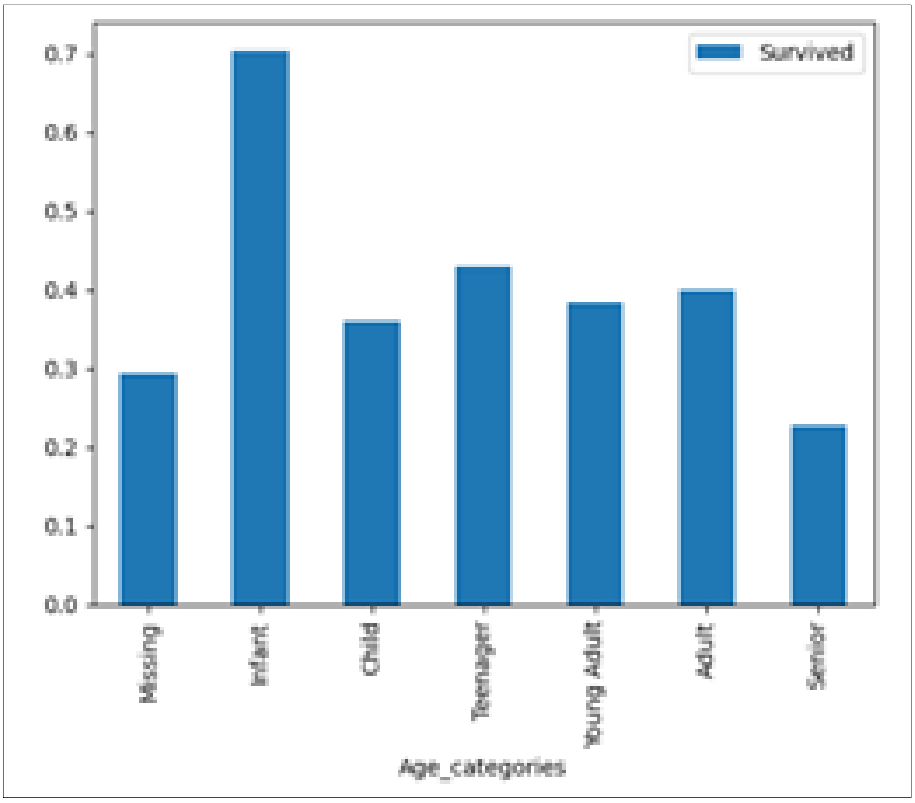


Fig 6: Demographics of Survivors

The graph shown in fig 6 shows the bar chart of the Representative Demographics of Survivors in Titanic Mishap. This Chart show the trend that Age was one of Deciding factor determining the survival of the passenger boarded the Ship.

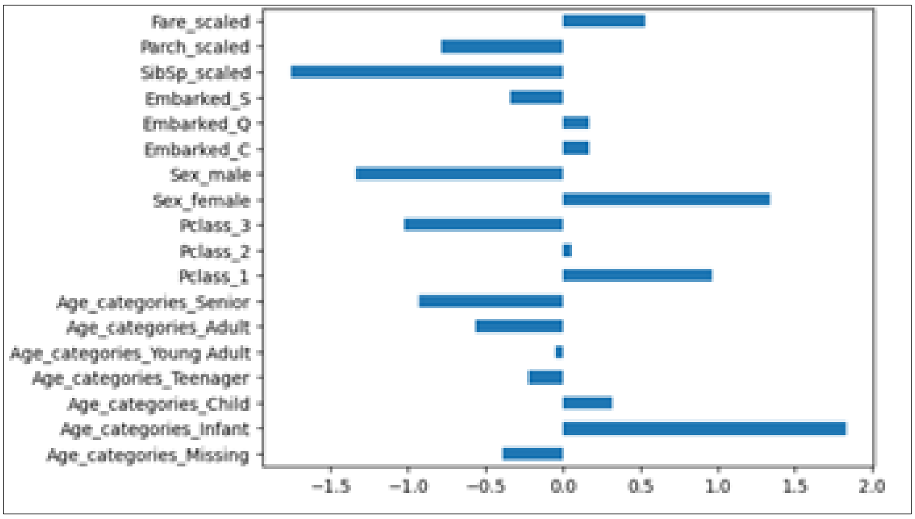


Fig 7: Accessing feature Relevance of the categories

The graph shown in fig 7 depicts the correlation between the factor in the dataset and how much influence it had in the survival of individual.

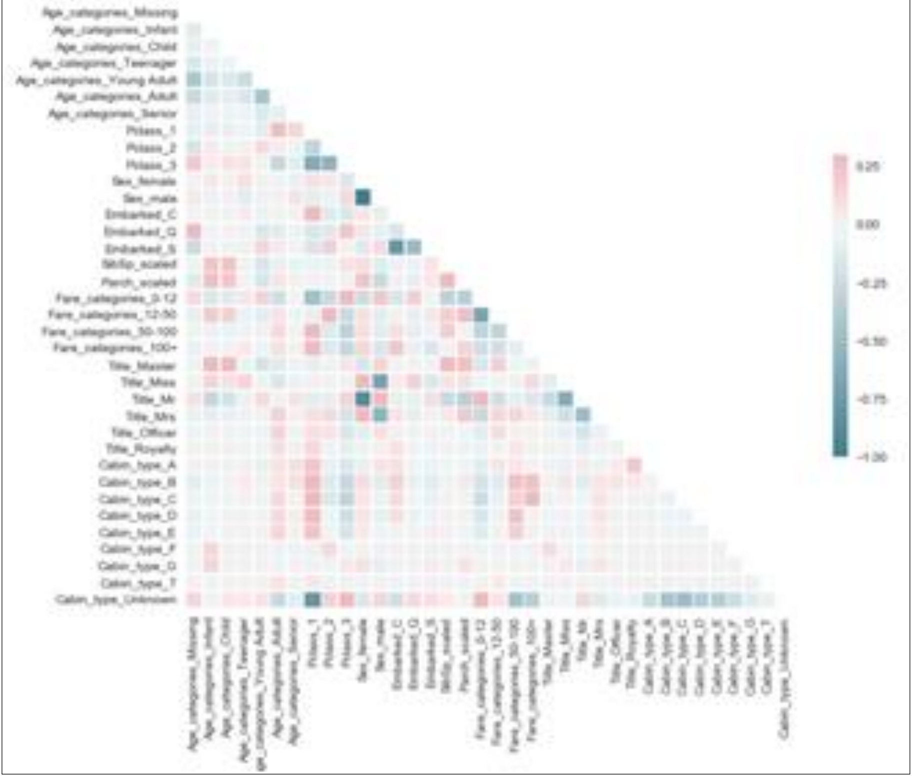


Fig 8: Heat Map of Correlating Variables of Survival

The heat map shows the cross-table relationship in an Aggregated way, the darker portions convey the fact that they are prominently involved in the affecting the survival probability of the passenger.

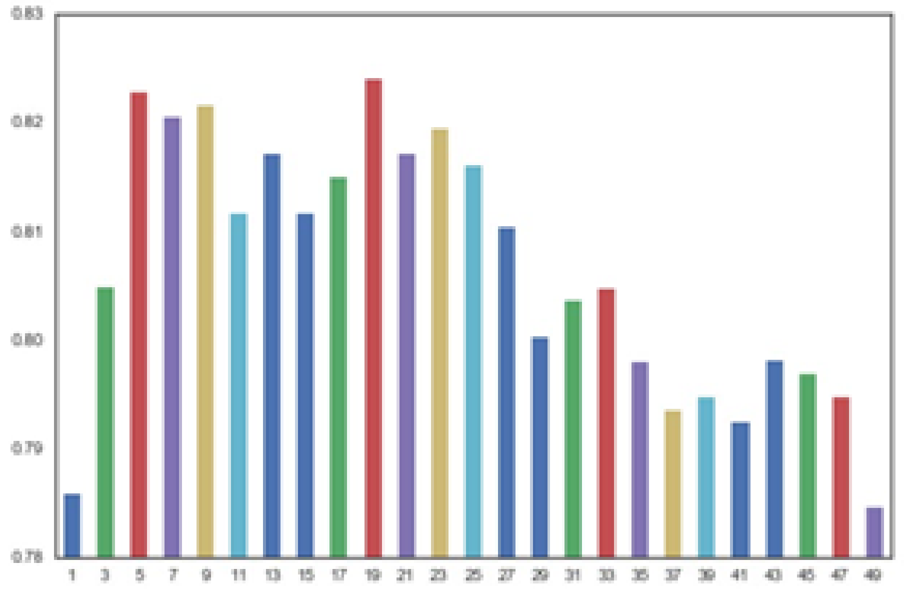


Fig 9: Hyperparameter optimization on k-nearest neighbors

Using the Hyperparameter optimization before the learning models begin it helps us to reduce the steps involved in fine-tuning and refining result in later stage which may be time intensive task and involve pruning ,tree structing , etc.

Conclusion

The initial step in data analysis is data cleaning, followed by exploratory data analysis (EDA) to comprehend the dataset and the interrelationships among attributes. EDA utilizes graphical techniques like ggplot and histograms to determine the association between features. From the results of EDA, conclusions and facts can be derived, such as the substantial impact of age on survival. According to Table-2, as age increases, survival rates decrease. Moreover, the analysis indicates that females had a high survival rate of approximately 74%, while males had a low survival rate. This fact is further supported by examining the titles in the name column, where the survival rate for Mr. was approximately 16% compared to 79% for Mrs.

To determine the family size of a specific passenger, we merged the parch and sibsp columns. We discovered that there is direct relation between survival rate increase and family size growth. However, when family size exceeds three, the survival rate decreases. Similarly, passengers with more cabins have a higher survival rate. As provided in Table 2.

Table 2: Survival Rate Vs. Class Type

Passenger class	Survival Rate (%)
1	61.34545
2	46.34534
3	23.45645

With these figures, we can conclude that the survival rate will increase with fare.

The exact parameters being used for constructing the models for Training and prediction are determined in feature engineering based on the exploratory data analytics approach.

Models created using machine learning anticipate the data values of travelers who survived. To make accurate predictions in classification problems, the logistic regression technique is used.

With an accuracy of 0.81504 according to the confusion matrix, the technique of logistic regression emerges as the one of the most accurate model.

This indicates that logistic regression has a very high level of prediction ability in this dataset using the selected features.

It is made obvious that when using a different feature modelling approach, the models' accuracy may change. The ideal models for classification problems are logistic regression and support vector machines since they provide a high level of accuracy.

Future Scope

This project combines the use and deployment of data analysis and machine learning and should be employed like a learning resource for EDA and machine learning at a basic level. Additionally, the project can be further developed using newer libraries such as shiny in R to create advanced graphical user interfaces. An interactive page can be created that allows you to proportionally change property values to change the corresponding graphical values (ggplot or histogram). By combining the obtained results, more specific conclusions can be drawn.

Due to the current work, there are many directions for future research. First, the current study considered only a small number of factors that affect a passenger's chances of survival. Other aspects could be examined to test whether other factors, such as family size or hut location, are equally important in predicting survival.

Although the machine learning models utilized in this study have good accuracy, there is still potential for development. The prediction performance of the models could be further improved by using more intricate models or ensemble techniques.

Third, applying the same technique to other historical shipwreck datasets to determine if the same characteristics hold across different contexts could be an intriguing path for future research.

Finally, while this study provides insights into the factors that influenced Titanic survival, it would be interesting to investigate how these factors have changed over time and how they might impact survival rates in modern shipwrecks.

References

1. Allison PD. Logistic regression using SAS®: Theory and application. SAS Institute; c1999.
2. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
3. Davison AC, Hinkley DV. Bootstrap methods and their application (Vol. 1). Cambridge University Press; c1997.
4. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics, 2001, 1.
5. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. CRC press, 2014, 2.
6. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer; c2009.
7. Hosmer Jr DW, Lemeshow S. Applied logistic regression. John Wiley & Sons; c2004.
8. Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions. John Wiley & Sons; c1994.
9. Kruschke JK. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press; c2015.
10. McKinney W. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference; c2010.
11. Kaushik P, Yadav R. Reliability design protocol and blockchain locating technique for mobile agent. J Adv Sci Technol (JAST). 2017;14(1):136-141. https://doi.org/10.29070/JAST.
12. Kaushik P, Yadav R. Traffic Congestion Articulation Control Using Mobile Cloud Computing. J Adv Scholar Res Allied Educ (JASRAE). 2018;15(1):1439-1442. https://doi.org/10.29070/JASRAE.
13. Kaushik P, Yadav R. Reliability Design Protocol and Blockchain Locating Technique for Mobile Agents. J Adv Scholar Res Allied Educ (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE.
14. Kaushik P, Yadav R. Deployment of Location Management Protocol and Fault Tolerant Technique for Mobile Agents. J Adv Scholar Res Allied Educ (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE.
15. Kaushik P, Yadav R. Mobile Image Vision and Image Processing Reliability Design for Fault-Free Tolerance in Traffic Jam. J Adv Scholar Res Allied Educ (JASRAE). 2018;15(6):606-611. https://doi.org/10.29070/JASRAE.
16. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; c2018.
17. Tidyverse. R package version 1.3.0; c2019. https://www.tidyverse.org.
18. Wickham H. ggplot2: Elegant graphics for data analysis. Springer; c2016.
19. Wickham H, Francois R. dplyr: A grammar of data manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr; c2016.