Titanic Machine Learning from Disaster

Sandeep Naraynasetty
Computer Science Department
Kent State University
Kent, OH 44240
snaraya5@kent.edu

Naveena Kanderi
Computer Science Department
Kent State University
Kent, OH 44240
nkanderi@kent.edu

Jashwanth Reddy Baggari
Computer Science Department
Kent State University
Kent, OH 44240
jbaggari@kent.edu

Vaishnavi Valluri
Computer Science Department
Kent State University
Kent, OH 44240
vvalluri@kent.edu

***Abstract*—One of the most tragic events in history was the Titanic shipwreck, which claimed the lives of thousands of passengers and crew members. The lack of life boats was mostly to blame for the deaths. The tragedy revealed a startling fact: certain people were more resilient to suffering than others, such as children and women, who received higher priority for rescue. By using exploratory data analytics on the given training data, the algorithm's main goal is to first identify predictable or previously undiscovered data. It then applies several machine learning models and classifiers to finish the study. Finally the accuracy of different methods used is compared.**

I. Introduction

The main aim of this project is to build a predictive model that helps us to get the probability of survival for the passengers that boarded the Legendary Titanic Ship.This is to be done using the data of the passengers and crew on the ship at the time of disaster such as Age, Passenger ID, Gender, Passenger class, Fare, Embarked place etc... We could implement this predictive model for any disaster that might happen in future. Get insights on how various algorithms are used to train the given data set to get the probability of survival using Machine learning techniques. To get familiar working with large datasets, processing the data, data cleaning (prioritizing given data and working with missing data), training of datasets, visualization of the data and predicting the possible outcomes.

Many models such as Linear Regression, k-Nearest Neighbor, Naïve Bayes, Decision Tree, AdaBoost, Random Forest, Gradient Boosting, Artificial Neural Networks (ANN) are used by other people who have worked on this Problem. Each model has its own benefits and drawbacks depending on the approach they have chosen to obtain the results. We have chosen to use the Decision trees model to train our dataset and predict the results as we can use the independent variables to predict the survivability of the passengers.

Tools such as Numpy, Pandas, Seaborn, Matplot, Sklearn (Scikit -learn), Jupyter notebook are utilized in this building this model to predict the survival rate of the passesngers.

II. Literature Review

Neetu Faujdar and Et al.[1] have utilized the ML algorithms such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest. to fill in the missing data they considered the title such as Mr., Mrs. ,Miss., etc. For analyzing these methods they considered accuracy and false discovery date as the performance metrics using a confusion matrix using the formula (FP/FP+TP) *100, implying, lower the false discovery rate the better the algorithm works. Logistic regression has the most accurate results compared to others.

Nahush Phalke and Et al.[2] used logistic regression for predicting the survival rate.To fill in the missing data,average values of the column are used.They generated a regression line between the dependent variable and independent variables to predict the value of the dependent variable. For testing the data, they have used the k-fold cross validation technique. Confusion matrix technique is used to calculate the accuracy of the algorithm using (TN + TP)/Total number of rows* 100 and achieved an accuracy of 95%.

Neytullah Acun and Et al.[3] have applied fourteen different algorithms to the data set and F-measure score from all the techniques are compared with one another and the F-measure score which is taken from Kaggle. It's seen that best results are obtained by using Voting (GB,ANN,kNN) algorithm with a F-measure score of 0.82. calibration can be done to improve the results but theirs don't yield much in F-score whereas Kaggle does.

Ghada Hassan and Et al.[4] has explained to fill in the missing values average median and a logic equation from the general age difference between that of a woman and her child is used to fill in the age column. The finding demonstrates that Decision tree was the most accurate and Naïve Bayes has the least. Findings implied that if we use sex as the major feature, Decision Tree provided the accuracy of 81%.

Chun Yu Cheung [5] has explained Random Forest(RF) technique is implemented using Weka tool, where data is classified into either survived and not survived. Using RF model, 5 different models are computed using alternative datasets to predict the accuracy of survival rate. The author concludes that the predicted accuracy is 82% and by comparing the weka and leader board, the prediction error does not differ more than 5% with each other.

III. Potential Customers and End Users

Any Government organizations, travel agencies or any institutions that organize large events involving any potential risks to the lives of people that use their services. Authorized personnel that handle Disaster Management in fields such as Police, Fire, Medical etc.

A. Constraints Imposed by the Customer

Obtained results shall be displayed to end users in the form of graphs so they can be easily understood. Use the Box plots and density plots to visualize the probability of survival for various constraints. Choose the most suitable patterns of correlation to predict your results.

B. Assumptions and Risks

When features that have less correlation are selected to train the data over the ones that might provide higher correlation, it could possibly predict the result with less accuracy. Additional data such as differentiating the crew from passengers might help us to train the data with more accuracy. Care should be taken while replacing missing values in the dataset as considering improper values might lead to less accurate results. We should be cautious while adding more data to the dataset, as it might predict incorrect results. While neglecting data from a given dataset, care should be taken not to omit data that might be relevant in producing results of higher accuracy.

C. Stakeholder

Srinadh B
Application Development Manager
Email: bssrinadh@gmail.com

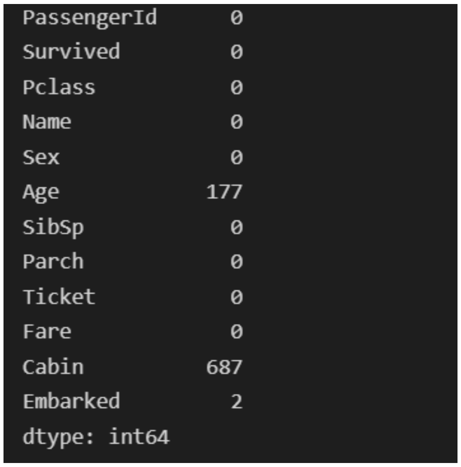
IV. Dataset

The data we used for our project was provided on the Kaggle website. We are given with two datasets one is the training dataset to train the algorithm. In training dataset, we are given with 891 samples of data with 12 unique features. The test file does not have the survived feature, but it has the remaining 11 features and in test data we have 418 passengers' samples which are different from the samples given from the training dataset. We have to predict the survival rate of the passengers in the test dataset by using the training dataset. The features of the training dataset, value of the feature, Characteristic of the feature and Description are shown in the Table 1.

V. Data Cleaning and Analysis

In the above dataset, the passenger ID and Cabin features are not considered for predicting the survival chances. And in our model, we have considered the Survived feature as a target feature to get the accurate prediction of survived passengers.

PassengerID: It is a numerical feature which represents the number of passengers and it is used to just sort the dataset. So, we are removing this feature from the dataset as it does not add any significance to the predictive model.



```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Fig. 1. Number of null values for each feature

PClass: The feature PClass is given as a number but we can use it as a category. By using this feature, we can categorize the passenger from the wealthiest classes to other classes.
Name: From the feature Name we can extract the titles of the passenger like Mr., Mrs., Miss, Master. We can use the extracted title for filling the missing values in another feature.
Sex: This is given as string feature which represents the gender of passengers in the dataset. According to it, this feature considers 65% passengers as male and 35% passengers as female. So, we can use the label male to 0 and female to 1 while training the data.
Age: It is a numeric feature which can be highly relevant, but this feature has 177 null values (fig 1 for ref) for replacing these null values we can use the average median of the age, or we can use the extracted titles from the name and taking the average of each title and the fill the missing value.
SibSp and Parch: These two features are numerical features which can be used to group the families together.
Ticket: In the dataset, ticket feature is considered as a string which can be dropped from the table because this feature has a lot of uncertainty, and we cannot parse the string as we do not have a single pattern for this feature.
Cabin: It is a string feature which shows us which cabin is booked by the passenger, if we use this feature, we can get some good results as we by this we can see classify cabins are closer to lifeboats. But unfortunately, we have 687 null values (fig1 for reference) from 891 sample data we have so we can drop this feature.
Embarked: This is considered as a string feature. We have three ports where people have boarded the titanic, we have only 2 missing values in this feature. We can either drop those two rows from the dataset or we can fill it with S as in the dataset the most number of people embarked at Southampton.

The below figure2 shows the relation between age attribute and the survival rate. This shows that age is a key factor in determining the survival rate. We have filled the 177 null values in Age column by using the title of the passenger obtained form the passengers Name. each title is grouped and the median of the all the passengers with the same title, sex, PClass is replaced in the null values.

When the "age" feature is taken into account, it can be seen that passengers' ages range from 0 to 80. The majority of the passengers in the 0-13 age group survived, whereas the majority of those in the 61-80 age group perished, if we divide the passengers into age groups such as 0-13, 14-60, and 61-80. These statistical findings demonstrate that the first kids were saved as the ship began to sink.

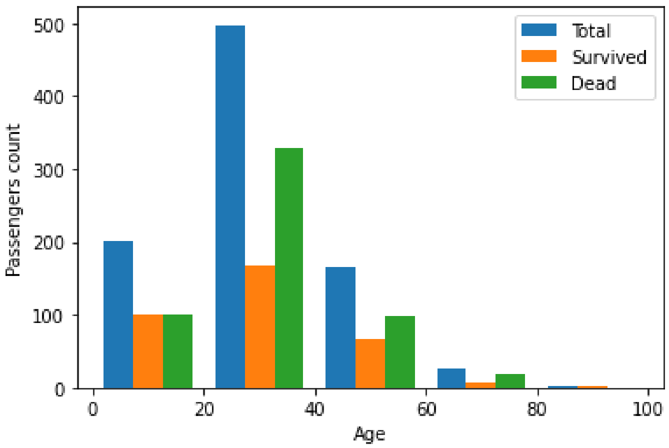


Fig. 2. Survival Rate depending on age

According to the distribution of the "Embarked" feature, the ship's ports "S," "C," and "Q" are each home to 644, 168, and 77 passengers, respectively. Fig. 2 provides the survival rates of passengers boarding from various ports . When this graph is examined, C is the port with the greatest rate of 55% survival. So, this could be regarded as "embarked" feature provides crucial information about getting by.

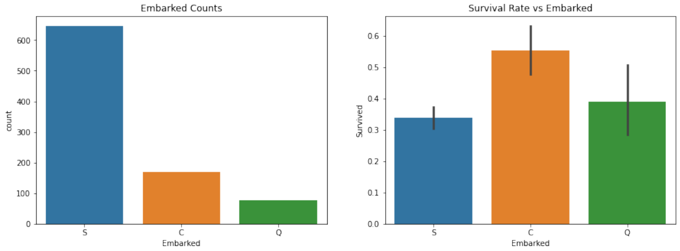


Fig. 3. Survival Rate depending on Embarked Port

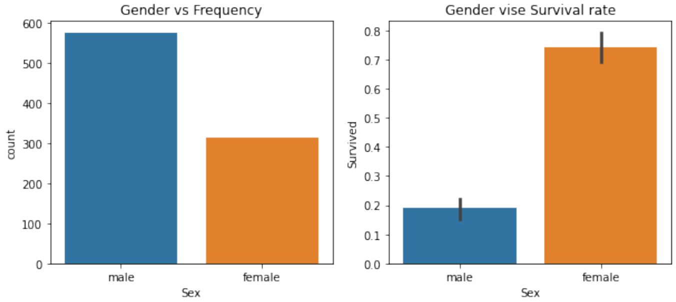


Fig. 4. Survival Rate depending on Gender

Figure4 analyses the relation between sex and survival rate. It shows that there are 314 female travelers and 577 male passengers. While some passengers have died, 233 female travelers have been saved. On the other hand, 109 of the male passengers were saved, while others perished. When we examine these distributions, it becomes clear that women have a higher survival rate than men. It has been determined that this attribute has a considerable impact on predicting the class label.

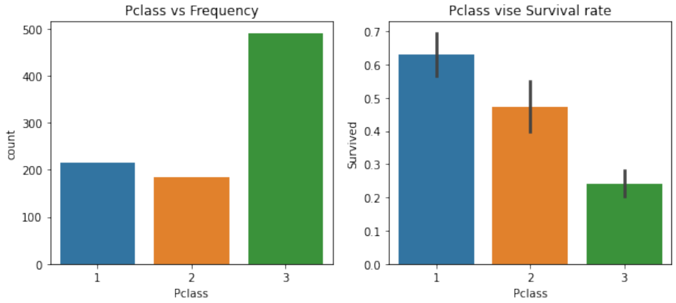


Fig. 5. Survival Rate depending on Pclass

Three different passenger classes are described by the "Pclass" characteristic. There are 491 passengers in class 3, 216 in class 1, 184 in class 2, and a total of 216 in class 1. Fig. 5 displays the passenger survival rates as a result of the "Pclass" feature. First-class travelers had the best survival statistics, with a 63% success rate. This ratio also demonstrates the existence of affluent people.

In addition to the already existing features, a new feature called "Family size" is established in this study. This feature is calculated by multiplying the Sibsp feature's value by the Parch feature's value. Following that, we've divided this feature into two groups. There are passengers in the first group whose family size is 0, 4, 5, 6, 7, or 10, and passengers in the second group whose family size is 1, 2, or 3. It has been noted that the bulk of the first group perished while most of the second group survived. These findings demonstrate that having more family members increases one's chance of surviving.

VI. Algorithms Used

Random Forest

The random forest is made up of a set of bootstrap samples that are produced from the original data set and a collection of decision trees. The entropy (or Gini index) of a chosen subset of the characteristics determines how the nodes are divided. The subsets that are generated using bootstrapping from the original data set are the same size as the original data set. It is usually produced by creating a large number of decision trees and averaging their behavior using the well-known wisdom of the crowd. We could get superior outcomes as a result of its randomness, as would be predicted.

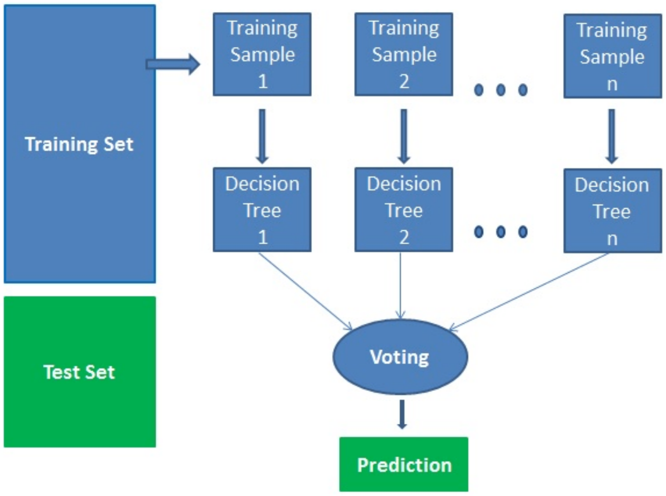


Fig. 6. Random Forest

Gaussian Naive Bayes

This is based on Bayes theorem assuming all features are independent given the value of the class variable. This is a conditional independence assumption and true in real world applications. When characteristic values are continuous in nature, the values associated with each class are presumed to be distributed according to the Gaussian distribution, or normal distribution.Due to this assumption NB performs well on high dimensional and complex datasets.

We employed the Naive Bayesian classifier to be thorough. Its advantages include accurate handling of missing values and resistance to noise in the data as well as irrelevant features. These strengths, that there are no missing values, that the noise has been minimized with the outlier's detection, and that unnecessary attributes were previously discarded at the beginning, are beneficial because of the pre-processing on our created data.

In comparison to more complex techniques, naive Bayes learners and classifiers can operate at lightning speeds. Each distribution can be individually estimated as a one-dimensional distribution due to the decoupling of the class conditional feature distributions. This in turn aids in resolving issues brought on by the dimensionality curse.

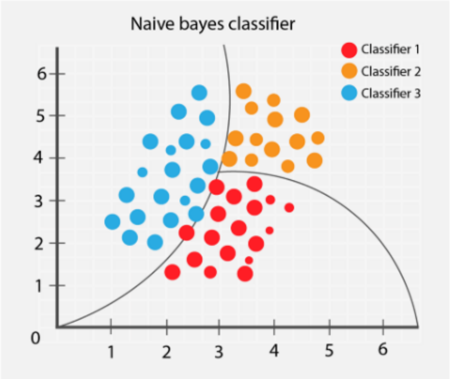


Fig. 7. Gaussian Naive Bayes

Decision Trees

The Decision Tree algorithm is used to identify the patterns through predictive models, by making use of simple probability calculations. It is a supervised machine learning algorithm; it uses the tree structures to predict the outcome. This algorithm deals with both numerical and categorical data. Conceptually, decision tree starts with ROOT node and data is divided into levels where the last level is called as leaf node where decision has been taken.

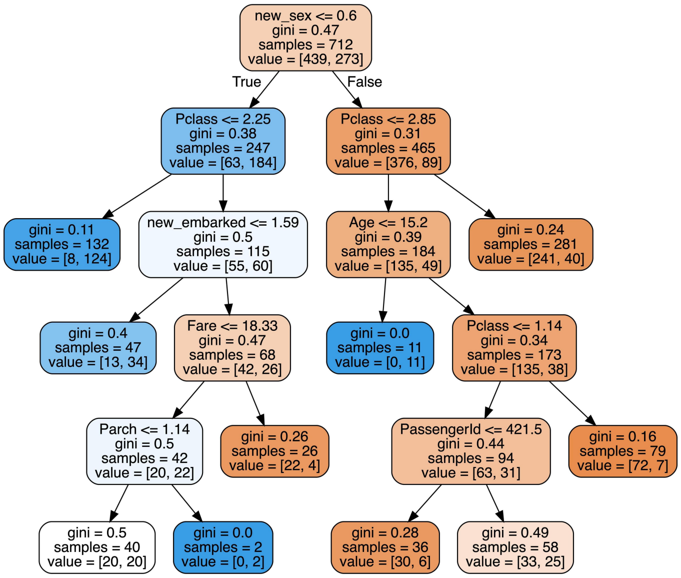


Fig. 8. Decision Tree

SVM Kernel Linear

Support vector machines (SVMs) are a group of supervised learning techniques for classifying data, performing regression analysis, and identifying outliers.The key concept is that the algorithm searches for the best hyperplane that may be used to categorize new data points based on the labeled data

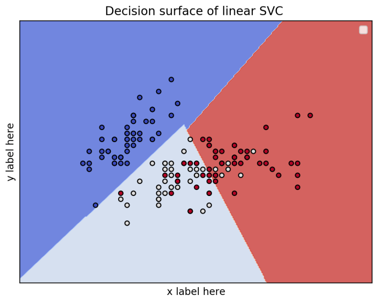


Fig. 9. Support Vector Machine Linear Kernel

(training data). The hyperplane is a straightforward line in two dimensions. The categorization of a class is typically based on the representative qualities that a learning algorithm learns to reflect the most prevalent traits (what distinguishes one class from another) (so classification is based on differences between classes). The SVM operates in the reverse direction. It locates the class samples that are most comparable. The support vectors will be those.

VII. Experimental Results

The objective of running all algorithms is to determine which features are related to the survival of passengers and to analyze the likelihood of survival. To implement the algorithms to the Titanic dataset additional modifications to specific model parameters are needed in order to improve algorithm accuracy which is performed in data cleaning.
We have implemented Decision Tree for the train data and the test data and we got a accuracy of 88.6% and 96.08% respectively. We also implemented Random Forest as this method is produced by creating different decision trees and got an accuracy of87.07% for train data and 95.5% for test data which is similar to the accuracy of decision trees. We wanted to compare the accuracy with some other classifying techniques so we implemented Support Vector Machine (kernel linear) and Gaussian Naive Bayes for the dataset which provided an accuracy of 85.11% & 88.07% for train data respectively and 94.41% & 96.08% for test data respectively. Below is the table that the accuracy results for both train and test data to the different algorithms we used.

Experimental Results		
Method	**Train**	**Test**
Decision Trees	88.6%	96.08%
Random forest	87.07%	95.5%
SVM Kernel linear	85.11%	94.41%
Decision Trees	88.07%	96.08%

VIII. Conclusion

In a knowledge-based society, it is crucial to get useful results from incomplete and unprocessed data utilizing machine learning and feature engineering techniques. In this essay, we present models for determining if a Titanic tragedy survivor will be identified. In order to study features that have correlation or are uninformative, a thorough data analysis is first carried out. Additionally, as part of the preprocessing process, certain new features are added to the dataset, like family size, while others are removed, like name, ticket, and cabin. Second, four distinct machine learning methods are employed in the classification stage to categorize the dataset created in the preprocessing step. As a conclusion, this paper presents a comparative study on machine learning techniques to analyze Titanic dataset to learn what features effect the classification results and which techniques are robust

IX. References

1) Aakriti Singh, Shipra Saraswat, Neetu Faujdar, "Analyzing Titanic Disaster using Machine Learning Algorithms" International Conference on Computing, Communication and Automation (ICCCA2017).
2) Vaishnav Kshirsagar, Nahush Phalke, Titanic Survival Analysis using Logistic Regression, (IRJET-2019).
3) Ekin Ekinci* , Sevinç İlhan Omurca* , Neytullah Acun*, "A Comparative Study on Machine Learning Techniques using Titanic Dataset" 7th International Conference on Advanced Technologies (ICAT'18)
4) Nadine Farag, Ghada Hassan, "Predicting the Survivors of the Titanic-Kaggle Machine Learning from Disaster-" (ACM-2018)
5) Chun Yu Cheung, Machine Learning from Disaster: Predicting the Titanic Survival Rate, a Random Forest approach (Research gate- 2015)
6) L. Breiman, "Random Forests," Machine Learning, vol. 45, pp.5-32,2001
7) T. Chatterjee, "Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms," International Journal of Emerging Research in Management & Technology, vol. 6, pp. 1-5, June 2017