

REPORT ON

TITANIC SURVIVAL PREDICTION

at

Faculty of Computing, Engineering and Media

**Name: Aathira Puthenpurayil**

**PNO: P2736418**

**Course: MSc Data Analytics**

|  |  |
| --- | --- |
| **Module name:** | **Big Data Analytics** |
| **Module code:** | **IMAT5322** |
| **Title of the Assignment:** | **Big Data Analytics – Assessment 1** |

Introduction

The Titanic dataset is a famous dataset that contains the demographic information of the passengers including age, gender, and class, as well as information about the passenger's tickets and cabins.

The main goal of this project is to predict whether the passenger survived or not considering various characteristics by taking the target variable as survived from the dataset. The dataset is split into a training and testing set, and machine learning algorithms are used to predict survival on the testing set based on the training set.

The relevance of the Titanic survival prediction problem lies in the insights it can provide into the factors that influenced survival rates during the disaster by analysing the dataset and building predictive models.

Literature Review

In 2014 authors J. Wijaya and J. T. Agee published a paper "Machine Learning Techniques for Predictive Maintenance of Shipboard Systems," for titanic survival prediction using decision trees, random forests, and support vector machines, and found random forests outperformed the other algorithms with an accuracy of 80.36%[1].In 2015 authors M. Manikandan and K. Balamurugan published a paper "Predicting Survival on the Titanic: A Comparison of Machine Learning Techniques,"[2] using decision trees, k-nearest neighbors, and logistic regression, in predicting survival and got an accuracy of 80.58%. In 2019 authors O. Nedelcu and A. Ionescu[4] published a paper "Titanic: Machine Learning from Disaster," using logistic regression, support vector machines, and neural networks and found that neural networks achieved the highest accuracy, with 79.9%.In 2020 authors S. Mishra and S. Singh a paper "Survival Prediction of Titanic Passengers Using Data Mining Techniques,"[5] using decision trees, random forests, and gradient boosting, found gradient boosting outperformed the other algorithms with an accuracy of 81.39%.

While different studies have found varying levels of accuracy in predicting the survival of passengers, most studies agree that factors such as gender and age class are important predictors of survival [3].

**Workflow**

Calendar

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Details of Approach**

# 1. Importing Main Libraries

Text

Description automatically generated

# 2. Data Pre-processing

Graphical user interface, text

Description automatically generated

The Train set consists of 891 rows and 12 columns; the test contains 418 rows and 11 columns.

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

train. columns is an attribute in Python that is used to return the column labels of a Data Frame in an array-like data structure.

Text

Description automatically generated

Text, application

Description automatically generated  
Shows the first five columns by default.

Text, application

Description automatically generated

Gives the last 3 columns.

Graphical user interface, text

Description automatically generated

train.info () is a Python method that shows a concise summary of a Data Frame, including the data types of each column, the number of non-null values, and the memory usage of the Data Frame.

Table

Description automatically generated

# To create a statistical summary of a data frame, use the train. describe () function in Python. It gives a brief summary of the dataset's numerical factors, such as the count, mean, standard deviation, minimum and maximum values, and quartile values (25%, 50%, and 75%).

# 3. Data Cleaning

Identifying the columns containing missing values in the train and test titanic set.

Table

Description automatically generated with medium confidence

There are 177 ages, 687 Cabin, and 2 Embarked fields missing in total.

A picture containing chart

Description automatically generated

There are 86 ages, 327 Cabin and 1 Fare fields missing in total.

# 3.1 Handling with missing columns.

Step 1: Dropping Cabin columns as this is not important for survival prediction.

Graphical user interface, text, application, email

Description automatically generated

Step 2: The mean value is used to fill in the "Age" column's blank entries.

Graphical user interface, text, application, website

Description automatically generated

Step 3: Locate the "Embarked" column's mode value, which will have happened the most frequently, and use it to replace the missing values in the column with the mode value.Graphical user interface, Word

Description automatically generated with medium confidence

Step 4: Replacing the Fare value using the Imputation technique Mean where we can use the mean fare value of the non-missing values to replace the missing fare values.

Graphical user interface, text, application, Word

Description automatically generated

Step 5: Check the null function again to see all null values removed and cleaned successfully.

Graphical user interface

Description automatically generated with medium confidence

# 3.2 Removing unnecessary columns.

# Can remove a few more columns such a passengerId,Name, and ticket are not considered for survival prediction

# Graphical user interface, text, application, email Description automatically generated

# 4. Data Normalization

# Converting the categorical features to numerical features.

# Graphical user interface, text, application Description automatically generated

# Experimental Result Analysis

# 5. Data Visualization

# 5.1 Survival rate based on Gender.

Graphical user interface, application, Word

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Males are much less likely to survive than females.

# 5.2 Survival rate based on Pclass

Text

Description automatically generated

Chart, pie chart

Description automatically generated

Higher socioeconomic class individuals had higher survival rates.

# 5.3 Survival Rate based on SibSp

A picture containing graphical user interface

Description automatically generated

Chart

Description automatically generated

People with no siblings or spouses were less to likely to survive than those with one or two.

# 5.4 Survival Rate based on Embarked

A picture containing company name

Description automatically generated

Chart, line chart

Description automatically generated

Passengers who embarked at Cherbourg(1) had a higher survival rate than those who embarked at Southampton(0) or Queenstown(2).

# 5.5 Histogram of Passenger Ages

Diagram

Description automatically generated with low confidence

Chart, histogram

Description automatically generated

Majority of passengers were between 20 and 40 years old. There were also a significant number of passengers under 20 years old and a smaller number of passengers over 60 years old.

# 5.6 Scatter plot for Age vs Fare.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Passengers who paid higher fares tend to be older. The plot also shows that there is no clear relationship between 'Age' and 'Survived', and that the survival rate is higher for passengers who paid higher fares.

# 5.7 Heat Map

Text

Description automatically generated with low confidence

Chart

Description automatically generated

It can be useful for identifying which features are most strongly correlated with survival. The features that are most strongly correlated with survival are "Sex", "Pclass", "Age", "Fare" and "Embarked".

# 5.8 Pair Plot



Calendar

Description automatically generated with medium confidence

Calendar

Description automatically generated

The resulting pair plot shows the pairwise relationships between the remaining columns in the dataset, and how they relate to survival.

6. Data Analyses

Analyzing the data by checking the total number of passengers, the median age of passengers, the survival rate, and the mean fare paid by passengers.

Text

Description automatically generated

Text

Description automatically generated

# 7. Feature Extraction

# 7.1 Age Group

Sort the ages into logical categories and created a new field in the train data set as Age Group for prediction which age group had survived the most.

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Babies are more likely to survive than any other age group.

Converting the new field of categorical values into numerical for model prediction.

Text

Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

**7.2 Family Size**

The term "FamilySize" in the Titanic train dataset refers to the entire number of family members (siblings, spouses, parents, and children) that a passenger has on the Titanic, including themselves.

For example, if a passenger has 1 sibling/spouse and 2 parents/children onboard, their 'FamilySize' would be 1+2+1=4.

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

The resulting plot shows that passengers with small family sizes (1-3) had a higher chance of survival compared to those with large family sizes (4-11). Passengers traveling alone had a survival rate that was between these two groups.

8. Choosing the best models

# 8.1 Splitting the training data.

Using Python's scikit-learn tool, we divided the dataset into two subsets: a training set and a testing set. The testing set is used to assess the model's success on fresh, untested data, while the training set is used to train the model.

Text

Description automatically generated

8.2 Testing with different clustering ModelsGraphical user interface, text, application

Description automatically generated

# K-means

# Graphical user interface, text, application, chat or text message Description automatically generated

# Chart, line chart Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

# Hierarchical clustering

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

# DBSCAN

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

8.3 Testing with different classification models

A picture containing text, person, screenshot

Description automatically generated

# Gaussian Naive Bayes

It's a probabilistic algorithm that produces predictions based on the probability of each input belonging to a given class.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, treemap chart

Description automatically generated

# Logistic Regression

Logistic Regression is a linear model that makes predictions based on the probability of each input belonging to a certain class.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

**Support Vector Machines**

SVM model comparison involves comparing multiple SVM models with different hyperparameters, feature selections, or pre-processing techniques to find the best model for predicting whether a passenger survived or not.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

# Linear SVC

LinearSVC is a popular choice for classification problems that involve linearly separable data.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

# Perceptron

A perceptron is a single-layer neural network that takes a set of input features and produces a binary output (0 or 1) based on a linear combination of the input features.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

# Decision Tree

Every internal node in a decision tree represents a test on a feature, each branch reflects the test's result, and each leaf node represents a decision or a class label.

Text

Description automatically generated

Text, application

Description automatically generated

Chart

Description automatically generated

# Random Forest

A random forest is based on constructing multiple decision trees and combining their outputs to make a final prediction.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

KNN or k-Nearest Neighbour’s

In KNN or k-Nearest Neighbors, the class of a new instance is predicted based on the class of its nearest neighbors in the training data.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

# Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a variation of the gradient descent algorithm that computes the gradient of the loss function with respect to the model parameters using a random subset of the training data.

Text

Description automatically generated

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

# Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) is a variant of the gradient boosting algorithm that combines multiple weak classifiers to form a strong classifier.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

## 9. Model Comparison

9.1 Clustering Models

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

9.2 Classification Models

Text, letter

Description automatically generated

Table

Description automatically generated

9.3 KFOLD Score

KFold cross Validation is also performed to check if gradient Boosting itself is the best model to be chosen.

Text

Description automatically generated

# 10. Model Prediction

Clustering does not make predictions about the survival of new passengers, as k-means, hierarchical and DBSCAN clustering is an unsupervised learning technique.

Based on the metrics provided in the table, the Gradient Boosting Classifier appears to be the best model for titanic survival data prediction with an accuracy score of 85.79% and a high precision score of 89.47%.

Also checked KFold Cross Score with 10 folds and got Gradient Boosting Classifier with the highest score(80.97%).

I decided to use the Gradient Boosting Classifier model for the testing data.

# Discussion and Conclusion

The key findings include the identification of key features that affect the survival rate, such as gender, age, and class. Various machine learning models, such as logistic regression or a decision tree etc, can be trained on these features to predict the survival rate with reasonable accuracy.

The limitation includes that the dataset may not be representative of the entire population on board the Titanic, which could limit the generalizability of the findings.

Future directions for improvement may include incorporating additional features, such as the location of the passenger on the ship or their occupation or exploring different other machine learning models not used in the current project to improve the accuracy of the predictions. Additionally, collecting and incorporating more data, if available, could also help to improve the accuracy of the model.

# References

# Wijaya, J. and Agee, J.T., 2014. Machine learning techniques for predictive maintenance of shipboard systems. In Proceedings of the ASNE Ship Maintenance and Modernization Symposium (pp. 1-12).

# Manikandan, M. and Balamurugan, K., 2015. Predicting survival on the Titanic: A comparison of machine learning techniques. International Journal of Applied Engineering Research, 10(67), pp.49-56.

# Alvarez, A., 2017. Exploring the Titanic dataset with R. arXiv preprint arXiv:1703.05921.

# Nedelcu, O. and Ionescu, A., 2019. Titanic: Machine learning from disaster. In 2019 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR) (pp. 1-5). IEEE.

# Mishra, S. and Singh, S., 2020. Survival prediction of Titanic passengers using data mining techniques. Journal of Big Data, 7(1), pp.1-15.

1. Butte, S., R, P.A. and Patil, S. (1970) Machine Learning Based Predictive Maintenance Strategy: A super learning approach with deep neural networks: Semantic scholar, 2018 IEEE Workshop on Microelectronics and Electron Devices (WMED).