**TITANIC DATASET SURVIVAL PREDICTION MODEL**

****

**UNIVERSITY OF ENGINEERING AND MANAGEMENT,**

**JAIPUR**

**TITANIC DATASET SURVIVAL PREDICTION MODEL**

Submitted in the partial fulfillment of the degree of

**MASTER OF COMPUTER APPLICATION**

In

**DATA SCIENCE AND AI**

Under

**UNIVERSITY OF ENGINEERING AND MANAGEMENT,**

**JAIPUR**

BY

**TARUN GUPTA**

**Enrollment No: 12022007017011**

UNDER THE GUIDANCE OF

**Prof. SUMAN ACHARAYA**

MASTER OF COMPUTER APPLICATION

****

UNIVERSITY OF ENGINEERING AND MANAGEMENT, JAIPUR

**Approval Certificate**

This is to certify that the project report entitled “**Titanic Dataset Survival Prediction Model**” submitted by **TarunGupta (12022007017011)**  in partial fulfilment of the requirements of the degree of **MASTER OF COMPUTER APPLICATION** from **University of Engineering and Management, Jaipur** was carried out in a systematic and procedural manner to the best of our knowledge. It is a bona fide work of the candidate and was carried out under our supervision and guidance during the academic session of 2023-2022.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Suman Acharya**

Project Guide, Assistant Professor (MCA)

UEM, JAIPUR

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Somen Nayak** **Prof. A Mukherjee**

HOD DEAN

UEM, JAIPUR UEM, JAIPUR

.

**ACKNOWLEDGEMENT**

The endless thanks goes to Lord Almighty for all the blessings he has showered onto me, which has enabled me to write this last note in my research work. During the period of my research, as in the rest of my life, I have been blessed by Almighty with some extraordinary people who have spun a web of support around me. Words can never be enough in expressing how grateful I am to those incredible people in my life who made this thesis possible. I would like an attempt to thank them for making my time during my research in the Institute a period I will treasure. I am deeply indebted to my project supervisor, Prof. Suman Acharya me such an interesting thesis topic. Each meeting with him added in valuable aspects to the implementation and broadened my perspective. He has guided me with his invaluable suggestions, lightened up the way in my darkest times and encouraged me a lot in the academic life.

Tarun Gupta

**ABSTRACT**

This project aims to analysis a dataset regarding Titanic Survival and predicts whether a person has survived or not in the titanic accident. It involves various steps such as data pre-processing, exploratory data analysis, feature selection, and model selection. The code visualizes the data, checks for missing values, replaces missing values with median values and replaces categorical variables with numerical values using encoding technique. Further, it uses a Logistic Regression algorithm to predict the no of people survived after the accident. The final score obtained from the model is used to evaluate the model's accuracy.

**Table of Contents**

Table of Contents

List of Figures

1. CHAPTER

INTRODUCTION

1.1 What is Titanic Disaster?

1.2 Why we used Titanic Disaster Survival analysis and prediction?

1.3 Elements of Titanic Disaster Survival analysis and prediction.

1. CHAPTER

2.1 About Python Libraries

2.2 Proposed Model / Flow Chart

3. RESULTS AND DISCUSSIONS

4. CONCLUSION

5. APPENDIX

6. BIBLIOGRAPHY

## **List of Figures**

Figure1. Linear algebra on dataset

Figure2. Accuracy

Figure3. Scatter plot

Figure4. Heat map

### CHAPTER

INTRODUCTION

This project is a data analysis and machine learning project that aims to predict whether a passenger survived or not after the titanic accident in the dataset that contains various details of passengers. The project involves data cleaning, exploration, visualization, feature selection, and model training using the Logistic Regression algorithm. The dataset is analysed for correlations .Feature selection is performed using variance threshold, and the final model is evaluated using stratified sampling and standardization techniques. The project uses Python libraries such as Pandas, Scikit-Learn, Seaborn, and Matplotlib.

1.1 **What is Titanic Disaster?**

Titanic sank in the early morning hours of 15 April 1912 in the North Atlantic Ocean, four days into her maiden voyage from Southampton to New York City. The largest ocean liner in service at the time, Titanic had an estimated 2,224 people on board when it struck an iceberg at around 23:40 (ship's time) on Sunday, 14 April 1912. Its sinking two hours and forty minutes later at 02:20 ship's time (05:18 GMT) on Monday, 15 April, resulted in the deaths of more than 1,500 people, making it one of the deadliest peacetime maritime disasters in history. Titanic received six warnings of sea ice on 14 April but were travelling at a speed of roughly 22 knots (41 km/h) when it’s lookouts sighted the iceberg. Unable to turn quickly enough, the ship suffered a glancing blow that buckled it’s starboard side and opened six of it’s sixteen compartments to the sea. Titanic had been designed to stay afloat with up to four of it’s forward compartments flooded, and the crew used distress flares and radio (wireless) messages to attract help as the passengers were put into lifeboats.

1.2 **Why we used Titanic Disaster Survival analysis and prediction?**

Titanic Disaster Survival analysis and prediction techniques can be used to analyse large datasets of passenger records travelling in Titanic Ship. This can help researchers to make informed decisions about the passengers and monitor their survival state after the accident closely, leading to improve outcomes and better quality for further research. Therefore, Titanic Disaster Survival analysis and prediction is an important area of research and application.

1.3 **Elements of Titanic Disaster Survival analysis and prediction.**

**1.** Import necessary libraries such as Pandas, Seaborn, Matplotlib and Scikit-Learn.

2. Load the kidney disease dataset using Pandas.

3. Drop the “Name”,” Ticket”, ”Cabin” columns from the dataset.

4. Merged the “Sibsp” and “Parch” columns into the “Family” column.

5. Dropped the “Sibsp” and “Parch” columns from the dataset.

6. Print dataset information such as column names, data types and non-null count.

7. Check for missing values in the dataset and fill them with median of the dataset.

8. Convert categorical class “Sex” and “Embarked” into numerical values.

9. Created a reduced dataset by removing columns.

10. Created a scatter plot to visualize the relationship between Age and Passenger class.

11. Create a Heatmap to visualize the correlation between numerical features.

12. Label encoded the categorical columns in the dataset.

13.Split the dataset into training and testing sets using stratified sampling to ensure a balanced class distribution.

14. Fitted a Logistic Regression model to predict the target variable class “Survived”.

15. Calculate the accuracy score, precision and F1 score of the classifier on the test set.

**2. CHAPTER**

* 1. **About Python Libraries**

The Following Libraries we Can Import:-

1. Import pandas as pd.
2. Import seaborn as sns.
3. Import matplotlib.pyplot as plt.
4. from sklearn.linear\_model import Logistic Regression.
5. from sklearn.preprocessing import Label Encoder.
6. from sklearn.model\_selection import train\_ test\_ split.
7. from sklearn. metrics import accuracy\_score, precision\_score, fbeta\_score.
8. **Pandas (imported as Pd):**- The Pandas library is used for data manipulation and analysis. It provides tools to read and write data in various formats, including CSV, Excel, and SQL databases. It is also used for cleaning, filtering, and transforming data. And also Pandas library is used for working with data sets.
9. **Matplotlib: -**Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error.
10. **Seaborn: -** Seaborn is a library for making statistical graphics in python. It builds on top of Matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.
11. **sklearn.model\_selection. Train\_test\_split**: It is a machine learning library for Python that provides a tool for splitting a data set into training and testing sets.
12. **Sklearn.preprocessing. LabelEncoder:** Used to Label Encode the categorical columns. It is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project.Through this type of encoding, we try to preserve the meaning of the element where higher weights are assigned to the elements having higher priority.
13. **Sklearn. linear\_model. LogisticRegression:** It is a machine learning library for Python that provides a tool to implement the algorithm Logistic Regression which portrays a relationship between the one dependent categorical variable with one or more nominal, ordinal, interval variables.

**7. Sklearn.metrics .accuracy\_score, precision\_score, fbeta\_score:** It is a machine learning library for Python that is imported to measure the accuracy, precision and the f1score of the Logistic Regression algorithm. Accuracy is the most common metric to be used in everyday talk. Accuracy answers the question **“Out of all the predictions we made, how many were true?”** .Precision is a metric that gives you the proportion of true positives to the amount of total positives that the model predicts. It answers the question **“Out of all the positive predictions we made, how many were true?”.F1 score** is an alternative machine learning evaluation metric that assesses the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by accuracy.

2.2**ProposedModel / Flow Chart**

**Data**

**Preprocessing**

**Read CSV file**

**Import**

**Libraries**

**Preform Data preprocessing**

**Handle Missing values**

**Drop unwanted columns Renaming and Cleaning**

**Perform Data Exploration**

**Perform Feature Selection**

**Perform data visualization**

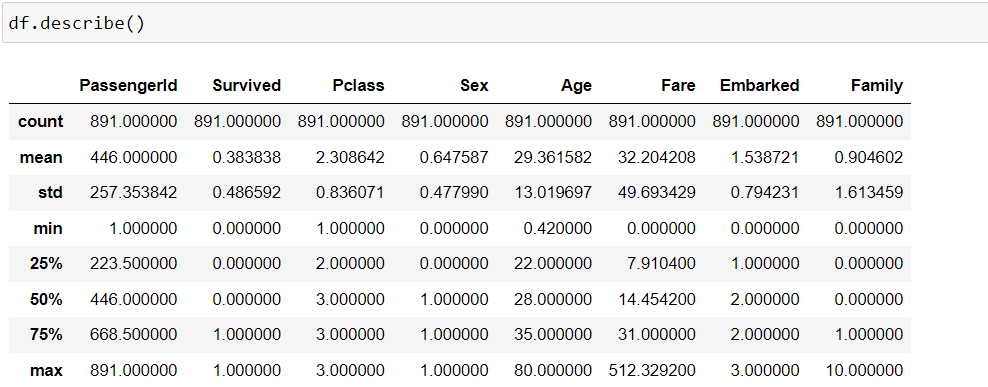
**Perform Model Evaluation and interpretation**

**Print Results**

**Perform Data Modelling and evaluation**

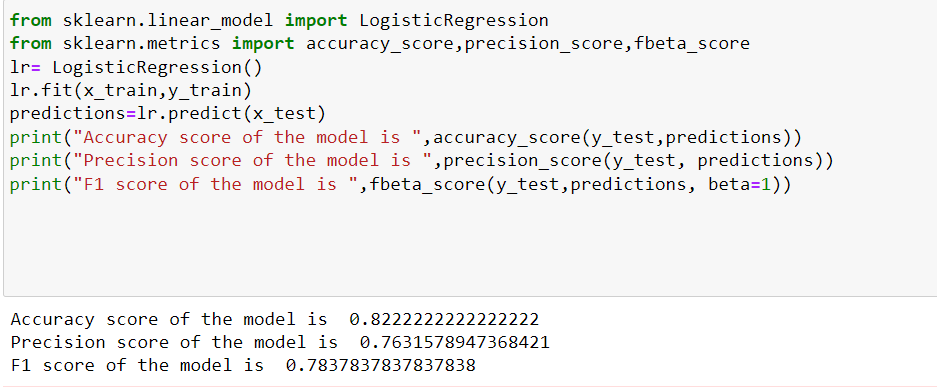
**RESULTS & DISCUSSION**

1. **Applying Linear Algebra on the dataset**

****

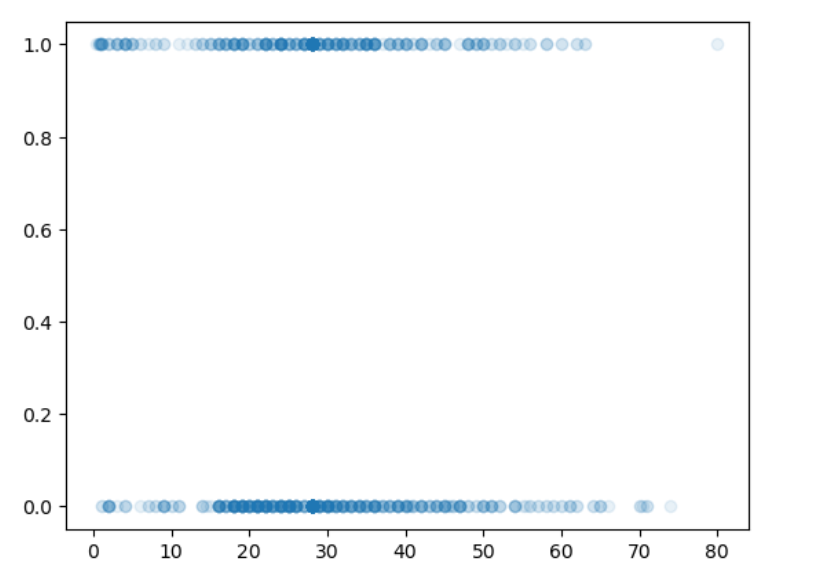
**Figure 1: Linear Algebra on dataset**

1. **The model which we train and make prediction using Logistic Regression algorithm we get accuracy of 82 percent and precision of 76 percent and f1 score of 78 percent.**



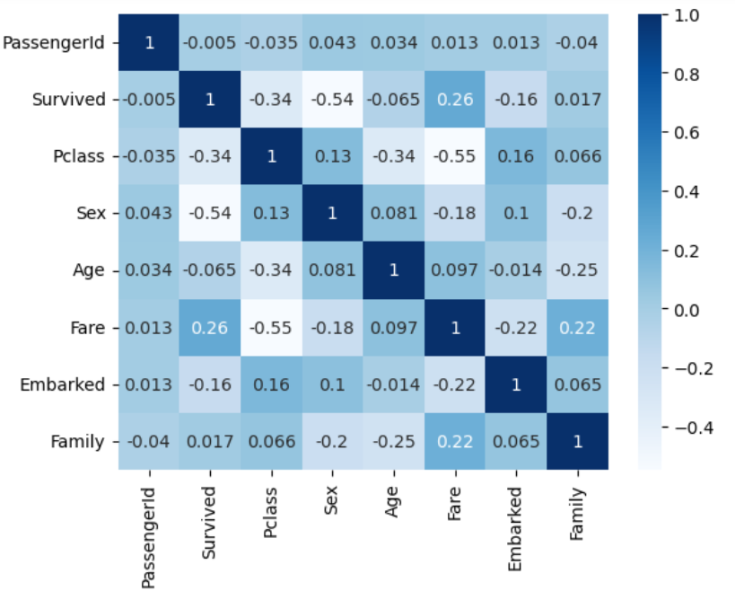
**Figure2: Accuracy, Precision and f1 score**

1. **A scatter plot is a type of graph used to display the relationship between two variables. In the context of dataset, a scatter plot can be used to show the relationship between age of the passengers and their Survival state after the titanic disaster.In this scatter plot graph we find out that which age group passengers survived more and which age group passengers survived less after the disaster.**



**Figure 3: Scatter plot**

1. **A heat map is a type of visual representation that is used to display data in a two-dimensional matrix format. Here we identify pattern and trend in the prevalence of the correlation of the numerical columns in the dataset.**



**Figure 4: Heatmap**

### 4. CONCLUSION

In conclusion, this code performs various data pre-processing and analysis tasks on a dataset related to Titanic Disaster. The code starts by importing necessary libraries such as Pandas, Seaborn, and Matplotlib and Skit-Learn. It then reads the dataset and performs data cleaning tasks such as dropping unnecessary columns, handling missing values, and converting categorical values to numerical values by label encoding technique. Data visualization is performed including statistical descriptive and feature selection. Finally, it prepares the data for machine learning by splitting the dataset into training and testing sets. A Logistic Regression model is trained on the data and evaluated using accuracy score, precision score and fbdeta score.

**APPENDIX**

In this appendix, we provide additional details on the use of the Logistic Regression algorithm for Survival prediction of passengers after the titanic disaster. Logistic Regression is a popular and simple machine learning algorithm that can be used for classification tasks, such as predicting whether a passenger has survived or not after the titanic disaster based on certain features.

Data Pre-processing:

1. Data Collection: A dataset containing relevant features such as Name, Age, Passenger id, embarked, etc., along with the corresponding labels indicating whether a passenger survived or not after the titanic disaster, is collected. The dataset is split into two parts: a training set and a testing set.
2. Data cleaning: The collected dataset is cleaned by handling any missing values, outliers, or inconsistencies in the data. This may involve imputing missing values and encoding categorical features.
3. Feature Selection: Relevant features are selected based on domain knowledge and feature importance analysis to reduce the dimensionality of the dataset and improve the performance of the Logistic Regression algorithm.

Logistic Regression Algorithm:

1. Logistic Regression Overview: Logistic regression, despite its name, is a [classification model](https://www.sciencedirect.com/topics/computer-science/classification-models) rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. Scikit-learn have a highly optimized version of logistic regression implementation, which supports multiclass [classification task](https://www.sciencedirect.com/topics/computer-science/classification-task) .
2. Prediction Step: Given a new data point from the testing set, the Logistic Regression algorithm predict the results based on the sigmoid line it draws on the training dataset. When the data is trained then the algorithm draws a sigmoid line which has a cut-off point 0f 0.5.When the test data is given to algorithm then it finds the sigmoid function value of the test data and if the sigmoid function value is greater than 0.5 then it is considered as True(1) and if it is below 0.5 then it is considered as false(0).

Model Evaluation:

Evaluation Metrics: To assess the performance of the KNN model, various evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix are calculated. These metrics provide insights into the performance of the model in terms of its ability to correctly classify kidney disease cases and non-cases.

**BIBLIOGRAPHY**

1. <https://www.kaggle.com/>
2. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>
3. <https://youtu.be/ehO0-6i9qD4?list=PLKnIA16_Rmvbr7zKYQuBfsVkjoLcJgxHH>