ECS784 Coursework-1: Titanic Case Study

I. INTRODUCTION

The sinking of the Titanic in 1912, caused by colliding with an iceberg, resulted in the deaths of approximately 1502 passengers and crew, making it one of the deadliest commercial maritime disasters in history. The tragedy led to the fact that some individuals had a better chance of survival than others was an eye-opening observation. The objective of this study is to apply data analytics (EDA) to find new insights in the available dataset. Furthermore, machine learning models (Decision Tree and Logistic Regression) are used to determine which types of individuals were more likely to survive. The accuracy of the different models is further compared and analyzed to draw conclusions.

1.1 PROBLEM STATEMENT

The problem statement of the Titanic dataset is to determine which factors influenced the survival of passengers on the Titanic. The issue might be presented as a classification problem, where the task is to predict whether a passenger survived or not based on their attributes.

1.2 HYPOTHESIS

Hypothesis suggests that being female, being a child, having a higher social status, and paying a larger fare may have been factors that increased a passenger's chance of survival. However, the hypothesis may also be challenged by other factors that were not considered, such as the passenger's location on the ship or the timing of their evacuation. Ultimately, the hypothesis serves as a starting point for analyzing the data and testing various machine learning models to identify which factors were most predictive of survival.

II. LITERATURE REVIEW

Akriti Singh et.al [1] proposed the survival rate of passengers on the Titanic using four major approaches: logistic regression, Random Forest, Naive Bayes, and decision tree. It was concluded that logistic regression had the highest accuracy and lowest false rate among all the algorithms tested. However, they noted that the analysis is still ongoing to determine the survival of all passengers.

Hitesh Pujari et.al [2] performed prediction on dataset by using logistic regression, SVC and Random Forest and concluded that Random Forest model performs the best as it achieves the highest accuracy.

Shawn Cicoria et.al [3] performed Decision tree classification and Cluster analysis suggested that sex is the dominant feature as compared to other features.

Kunal Vyas [4] suggested that dimensionality reduction and playing more with the dataset could improve the accuracy of the algorithms. The most important conclusion provided by them is that more features utilized in the models do not necessarily make results better.

Bruno S. et. al [5] concluded that people in their prime age died less often than older people. Passengers with high financial stability, traveling in first class, are better able to save themselves as are passengers in second class as compared to third class.

III. EXPLORING DATA

3.1 Dataset: The Dataset is collected from Kaggle [6]. The data is present in csv file.

3.2 Exploring Data Size and Type:

Variable	
PassengerId	Unique identifier fo
Survived	A binary variable indicating whether the passenger survived or no
Pclass	The class of the ticket the passenger purchased. 1 = 1s
Name	The name
Sex	The gende
Age	The age
SibSp	The number of siblings/spouses the passenger ha
Parch	The number of parents/children the passenger ha
Ticket	
Fare	The fare
Cabin	
Embarked	The port of embarkation for the passenger. C = Cherbourg, Q = Queenstown,

Table 1: Dataset variable and its description

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

Table 2: Data types

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

Table 3: First 5 columns of Dataset

Analysis:

- 11 feature variables and one target variable “Survived”
- PassengerID : Primary key that is unique (no use)
- Continuous Data : Age, Fare
- String Columns : 5
- Numerical Columns : 5
- String Data : Cabin, Name, Ticket
- Categorical String Data : Sex, Embarked
- Categorical Data : Sibsp, Pclass, Parch, Embarked

3.3 Missing Data

```
PassengerId      0.000000
Survived         0.000000
Pclass           0.000000
Name             0.000000
Sex              0.000000
Age             19.865320
SibSp            0.000000
Parch            0.000000
Ticket           0.000000
Fare             0.000000
Cabin           77.104377
Embarked         0.224467
dtype: float64
```

Table 4: Missing Data

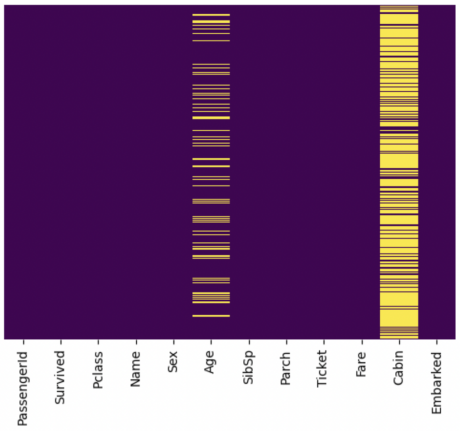


Fig 1: Missing Data Heatmap

Analysis:

- Cabin and Age columns have the most missing values.
- Embarked Column also has a few missing values.

3.4 Distribution of continuous variable (Age)

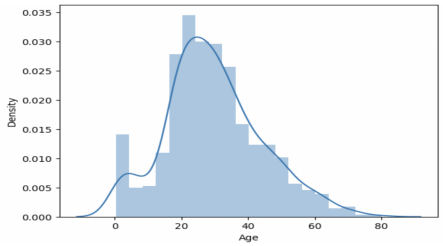


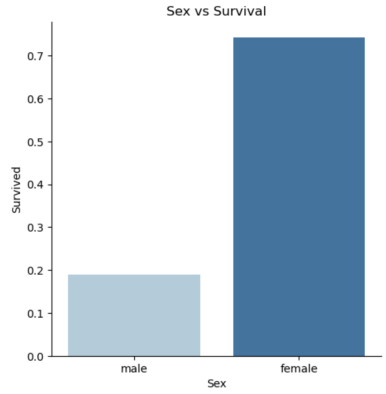
Fig 2: Age Distribution

Analysis:

- Age has good distribution (normal distribution)
- Majority age: 20 between 40
- Peak : 20-30 yrs
- Includes very young (less than 1 year) and some elderly passengers (over 60 years), but they are rare
- Slight Dip around 5-15 years old indicating fewer children in this age range
- Data Addition: For better visualization of the dataset, we have added features like Title and Family.

3.6 Visualization

1. Sex v/s Survival:



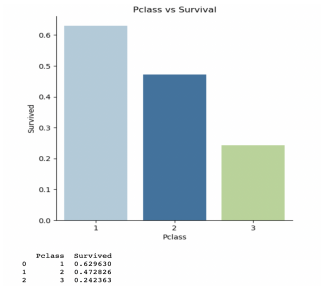
	Sex	Survived
0	female	0.742038
1	male	0.188908

Fig 3: Sex v/s Survival

Analysis:

- Female survival rate is ~75% and Men's survival rate is ~19%.
- Female has higher chance of survival than male

2. Pclass v/s Survival:



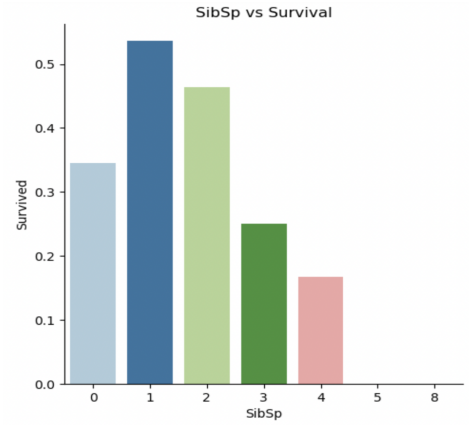
	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

Fig 4: Pclass v/s Survival

Analysis:

- Pclass 1 survival rate: 63%
- Pclass 2 survival rate: 47%
- Pclass 3 survival rate: 24%
- Pclass 1 has the highest survival rate; Pclass3 has the lowest Survival rate.

3. SibSp v/s Survival:



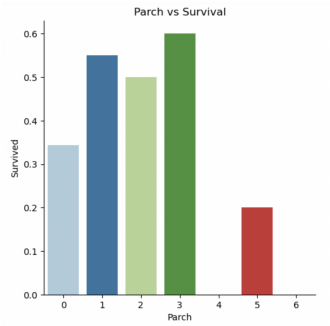
	SibSp	Survived
0	0	0.345395
1	1	0.535885
2	2	0.464286
3	3	0.250000
4	4	0.166667
5	5	0.000000
6	8	0.000000

Fig 5: Sibsp v/s Survival

Analysis:

- SibSp 1 has highest chance of survival (53%) , followed by SibSp 2 (46%)
- SibSp 5 and 8 has 0 chances of survival

4. Parch v/s Survival



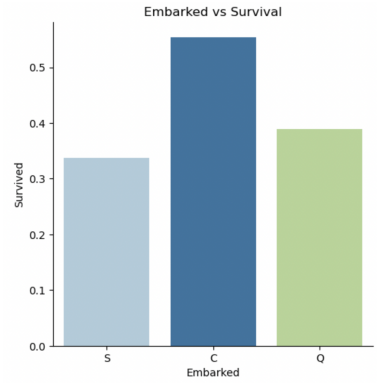
	Parch	Survived
0	0	0.343658
1	1	0.550847
2	2	0.500000
3	3	0.600000
4	4	0.000000
5	5	0.200000
6	6	0.000000

Fig 6: Parch v/s Survival

Analysis:

- Parch 3 has highest survival rate (60%), followed by Parch 1 (55%) and Parch 2 (50%)
- Parch 4 and Parch 6 has 0% chances of survival

5. Embarked v/s Survival



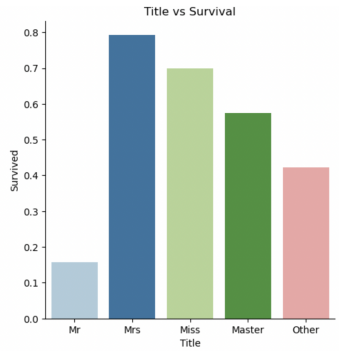
	Embarked	Survived
0	C	0.553571
1	Q	0.389610
2	S	0.336957

Fig 7: Embarked v/s Survival

Analysis:

- Embarked C has highest chance of survival 55% among others
- Embarked Q chance of survival: 38%
- Emarked S chance of survival: 33% (lowest among others)

6. Title v/s Survival



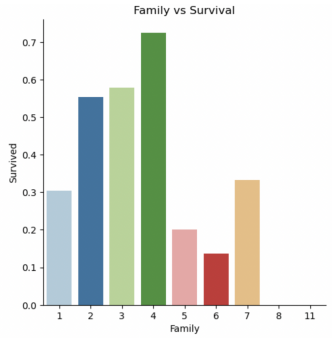
	Title	Survived
0	Master	0.575000
1	Miss	0.699454
2	Mr	0.156673
3	Mrs	0.792000
4	Other	0.423077

Fig 8: Title v/s Survival

Analysis:

- Mrs. has the highest chance of survival, followed by Miss
- Mr. has lowest chance of survival (15.66%)

7. Family v/s Survival



	Family	Survived
0	1	0.303538
1	2	0.552795
2	3	0.578431
3	4	0.724138
4	5	0.200000
5	6	0.136364
6	7	0.333333
7	8	0.000000
8	11	0.000000

Fig 9: Family v/s Survival

Analysis:

- Family size 8 and 11 have 0% chance of survival
- Family size 4 has maximum chance of survival (72%) followed by family size 3 (57%)

- Family size 6,5, 1 and 7 also has very chances of survival.

IV. DATA PREPROCESSING

Column “Sex” changed from categorical to integer (0 - male , 1 - female). Missing values of “Age” column filled up using Python function, which fills the mean value of column.

V. FEATURE SELECTION

In Feature Selection, we looked at correlation v/s Survival Analysis, Correlation matrix and then dropping unnecessary columns.

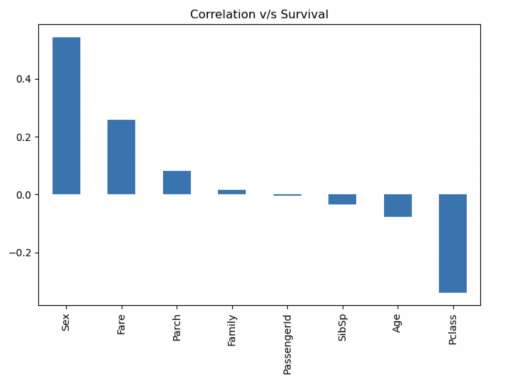


Fig 10: Correlation v/s Survival Analysis:

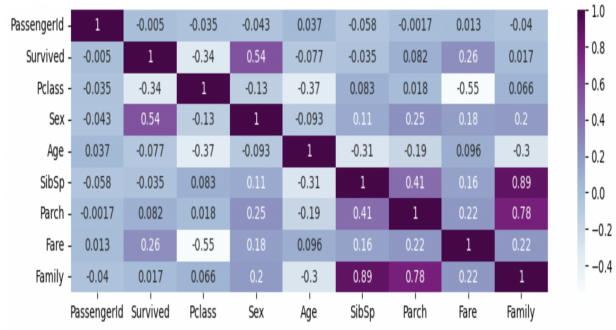


Fig 11: Correlation matrix

Analysis: Sex and Pclass has high correlation with survival. This is useful information and we may use while designing a decision tree.Fare and Age also has a significant role. Passenger Id doesn’t play any significant role. (least value in matrix) We further dropped few columns (PassengerId, Name, SibSp, Parch, Ticket, Fare, Cabin, Embarked) that were unnecessary and redundant.

VI. DATA ANALYTIC METHODS:

For our dataset, we have used Decision trees and logistic regression methods as they are relatively simple algorithms and are used for classification tasks which we are trying to achieve with Titanic dataset. They also perform well with a small number of features (like the Titanic Dataset) and are less prone to overfitting.

6.1 Decision Trees

Decision tree learning is a type of Supervised Learning method where the data is split continuously according to a certain parameter.The tree can be explained by two entities, namely decision nodes and leaves.We have selected Pclass, Sex and Age as features and Survived / Not Survived as class.

In the research the generated decision tree gave some useful insights. Some of the insights are that if a passenger is female and she belongs to a passenger class of either 1 or 2, then the probability of survival is more and if a passenger is male and greater than or equal to an age of 13, then the probability of his survival is less.

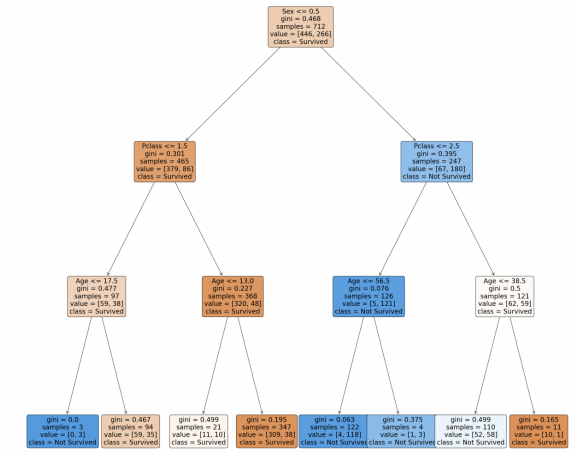


Fig 12 : Decision Tree

Performance Management:

1. Accuracy: 80.4%
2. Precision: 80.5%
3. Recall: 71%
4. AUC-ROC: 79.21%
5. Confusion Matrix: array([[90, 13], [22, 54]])

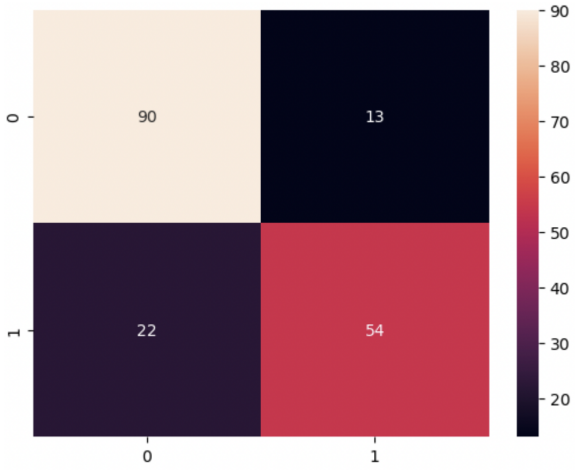


Fig 13: Decision Tree - Confusion Matrix

6.2 Logistic Regression

Logistic Regression is used to calculate or predict the probability of a binary (yes/no) event occurring. In our case, we use to calculate if passenger survives or not.

Performance Management:

1. Accuracy: 79.3%
2. Precision: 78.2%
3. Recall: 71%
4. AUC-ROC: 78.2%
5. Confusion matrix: array([[88, 15], [22, 54]])

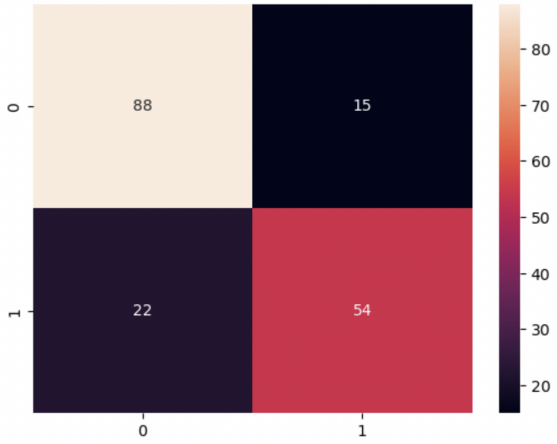


Fig 14: Logistic Regression - Confusion Matrix

VII. CONCLUSION

7.1 Valuation Analysis

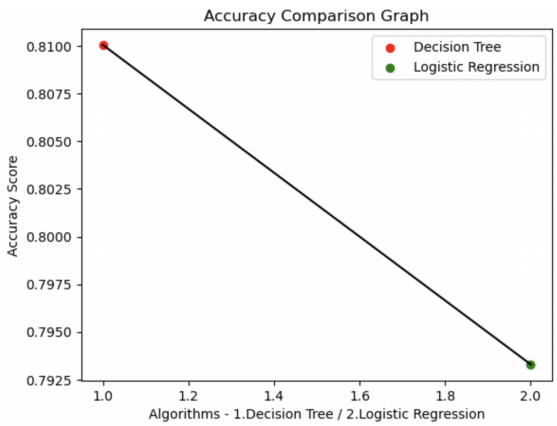


Fig 15: Valuation Analysis

In our analysis of the dataset, we found that decision trees outperformed logistic regression models in terms of accuracy. However, the difference in accuracy between the two models was not significant.

7.2 Challenges and Success

The Titanic dataset presented several challenges such as incomplete data, unbalanced classes with a much smaller number of survivors than non-survivors, biased data, and the need for feature selection. However, these challenges were successfully tackled by using visualization techniques such as charts, applying machine learning algorithms, and focusing on predictive accuracy. Despite the limitations, the dataset was still able to provide valuable insights into the factors that contributed to the survival of passengers on the Titanic.

7.3 Key Findings

Sex, age, and passenger class are the most crucial characteristics for predicting survival in the decision tree and logistic regression models.The decision tree model has a higher accuracy rate compared to the logistic regression model. Both models show that being a female or having a higher passenger class increases the likelihood of survival, while being older decreases the likelihood of survival.

7.4 Possible Extensions

To enhance the model performance and provide more accurate predictions for the Titanic dataset, possible extensions include ensemble methods, feature engineering, model tuning, other machine learning models, and cross-validation techniques. These approaches can improve the accuracy and generalization performance of the models, explore additional variables and features, optimize the models' hyperparameters, and evaluate the models' performance on different subsets of the data, resulting in more reliable and accurate predictions.

VIII. REFERENCES

[1] A. Singh, S. Saraswat and N. Faujdar, "Analyzing Titanic disaster using machine learning algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 406-411, doi: 10.1109/CCAA.2017.8229835.

[2] Hitesh Pujari, Rohit Singh Adhikari, Plaban Patra," Predicting the Survival of Titanic Passengers using Different Machine" , 2020 International Research Journal of Engineering and Technology, India, 2020

[3] Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, “Classification of Titanic Passenger Data and Chances of Surviving the Disaster”, Proceedings of Student-Faculty Research Day, CSIS, pp. 1-6, May 2014.

[4] Vyas, Kunal, Zeshi Zheng, and Lin Li, “Titanic-Machine Learning From Disaster”, Machine Learning Final Project, UMass Lowell, pp. 1-7, 2015.

[5] Trevor Stephens. (2014), ‘Titanic: Getting Started With R - Part 3: Decision Trees’, [Online]. Available: http://trevorstephens.com/kaggletitanic-tutorial/r-part-3-decision-trees/. [Accessed: 11-March- 2017].

[6] https://www.kaggle.com/datasets/brendan45774/test-file