Learning From Disaster

B.Rishi Srivathsava[1],B.Sai Teja Goud[2], P. Vikas[3], K.Sreelatha[4]
[1,2,3]*Student, Dept. of Electronics and Computer Engineering Sreenidhi Institute of Science and Technology*, Hyderabad, India
[4]*Asst. Professor, Dept. of Electronics and Computer Engineering Sreenidhi Institute of Science and Technology*, Hyderabad, India

***Abstract* — By using traditional machine learning algorithms, this study seeks to understand the relationship between the survival and mortality rates of passengers on the Titanic RMS. The passenger class, age, sex, passenger ID, and the number of siblings were all included in the dataset that was utilised for analysis, which was downloaded from the Kaggle website. Decision trees, random forests, xg-boost, the gradient boosting method, K-Nearest Neighbours, and logistic regression are just a few of the algorithms that are compared in terms of prediction accuracy in this study. Utilising the distinctive insights and extremely accurate numbers produced by each algorithm, the examination concentrates on the accuracy and precision of each approach. With the help of this study, we hope to advance our understanding of disaster analysis and offer new perspectives on how to make maritime accidents safer in the future.**

I.INTRODUCTION

Numerous passengers and crew members perished in one of the most iconic maritime tragedies in history, the Titanic RMS sinking. It is crucial to comprehend the elements that affected the survival rates during this catastrophe in order to improve safety protocols for marine incidents in the future. Using traditional machine learning algorithms, this study examines the correlation between Titanic passenger survival and mortality rates. This study aims to contribute to the field of catastrophe analysis by using the large dataset downloaded from the Kaggle website and combining characteristics such passenger class, age, sex, passenger ID, and the number of siblings to gain insightful knowledge.

An effective method for studying the complex patterns concealed within the Titanic dataset is the use of traditional machine learning algorithms. This study investigates the prediction accuracy and precision of several methods using decision trees, random forest, xg-boost, gradient boosting algorithm, K-Nearest Neighbours, and logistic regression. Instead of relying on a single algorithm, this method adopts a greedy algorithm that takes advantage of each model's strengths and makes use of the extremely exact data produced by each process. A greater comprehension of the survival determinants can be attained through this thorough examination, enabling better safety precautions and better catastrophe management techniques.

The Kaggle website, a popular venue for data science research and contests, provided the dataset for this study. This dataset offers a wide range of variables and traits that accurately reflect key facets of the Titanic passengers' conditions and demography. However, in order to guarantee the accuracy and integrity of the dataset, substantial data cleaning and preprocessing were done before to the analysis. To verify the reliability of the findings, the data was then divided and checked using the proper methods.

This paper's results and discussion part provides a thorough evaluation of the performance of each method, stressing the precision and accuracy of its prediction. In order to pinpoint the elements that significantly impacted the survival and fatality rates of Titanic passengers, feature importance analysis is also carried out. These observations can help improve maritime transportation safety protocols and support continuous efforts to reduce risks and avert disasters in the future.

II. LITERATURE SURVEY

Paper1: Exploring Survival Analysis

This paper's results and discussion part provides a thorough evaluation of the performance of each method, stressing the precision and accuracy of its prediction. In order to pinpoint the elements that significantly impacted the survival and fatality rates of Titanic passengers, feature importance analysis is also carried out. These observations can help improve maritime transportation safety protocols and

support continuous efforts to reduce risks and avert disasters in the future.[1]

Paper 2: "Predicting the Fate of the Titanic Passengers: A Comparative Study of Machine Learning Algorithms"
This study compares the performance of various machine learning methods, such as decision trees, random forests, and support vector machines, in predicting the survival of Titanic passengers. The data used by the authors includes 891 passenger records as well as different factors like passenger class, gender, and age. The results show that the decision tree method performs best, with an accuracy rate of 79.70%.[2]

Paper 3: "Data Analysis of Titanic Passenger Survival Using Machine Learning Techniques" The survival of Titanic passengers is predicted in this study using a number of machine learning techniques, such as decision trees, random forests, and k-nearest neighbours. The data used by the authors includes 891 passenger records as well as different factors like passenger class, gender, and age. The results show that the k-nearest neighbours algorithm has the highest accuracy, at 78.70%.[3]

III. EXISTING SYSTEM

A number of data analysts previously attempted to ascertain why certain passengers drowned in the disaster while others survived. Taking into account the most aspects from that algorithm, the majority of the data previously analysed are related to that algorithm.Using the Naive Bayes, Decision tree, and SVM algorithms, Lam & Tan et al. made one of the fewest attempts to examine the combination of the techniques.

IV.PROBLEM STATEMENT

According to findings by "lam and tang et al," it is believed that a forecast's accuracy does not entirely depend on the number of characteristics in a specific algorithm. This report was made using the machine learning techniques Naive Bayes, Logistic Regression, Decision Tree and Random Forest, K-Nearest Neighbours, XG-Boost, Gradient Boosting Classifier, and Logistic Regression.

V.PROPOSED SYSTEM

The datasets used in this investigation were made available through the kaggle website. To correlate the death and demise rates, a sample of 418 passengers was used to construct the dataset. Age, gender, siblings, and passenger id are all taken into account while utilising the algorithms.
The attributes that were taken into account for this report are shown in the table that is supplied below and was obtained directly from the kaggle website. Before being compared for the best result, this data is cleaned to eliminate any missing information and projected using a variety of techniques. The structure of the system for examining the Titanic's survival and fatality statistics includes a number of crucial elements and procedures. Data preprocessing is the first step in the procedure, and this is where the dataset is loaded from the Kaggle website. To learn more about the dataset, exploratory data analysis is carried out, and missing values are dealt with via imputation or elimination. The process of feature selection is then used to determine which factors have the most impact. The relevant features for the models are chosen using approaches like correlation analysis or feature importance from ensemble methods. Focusing on the important variables that may affect survival rates—such as passenger class, age, sex, passenger ID, and sibling count—helps in this step.

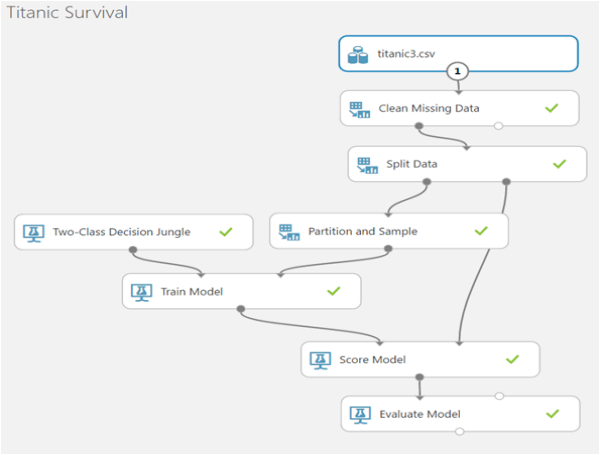

Fig 5.1: SYSTEM ARCHITECTURE

Model creation, which entails implementing several machine learning algorithms, comes next. A classification model is built using decision trees, and an ensemble of decision trees is built using random forest approaches for higher prediction accuracy. The techniques XGBoost and

gradient boosting are also used to enhance the performance of the models by iteratively fixing the errors created by the prior models. Passengers are categorised according to their resemblance using the proximity-based algorithm K-Nearest Neighbours (KNN). The analysis also includes logistic regression, a popular method for binary categorization.

Utilising performance indicators like accuracy, precision, recall, or F1 score, the trained models are then assessed. To evaluate the models' capacity for generalisation and reduce overfitting, cross-validation is used. The most efficient strategy is determined by comparing the forecast accuracy of each algorithm.

V.IMPLEMENTATION AND CODE

The working of proposed system algorithm is observed as:

Step 1: Load the dataset from Kaggle using pandas.

Step 2: Perform exploratory data analysis to understand the dataset.

Step 3: Handle missing values by imputing or removing them.

Step 4: Encode categorical variables into numerical representations.

Step 5: Split the dataset into training and testing sets.

Step 6: Perform feature analysis to determine relevant features.

Step 7: Instantiate a DecisionTreeClassifier model.

Step 8: Train the decision tree model using the training data.

Step 9: Instantiate a RandomForestClassifier model.

Step 10: Train the random forest model using the training data.

Step 11: Import the XGBClassifier from xgboost.

Step 12: Instantiate an XGBClassifier model.

Step 13: Train the XGBoost model using the training data.

Step 14: Import the GradientBoostingClassifier from scikit-learn.

Step 15: Instantiate a GradientBoostingClassifier model and train it using the training data.

VII.RESULTS

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Fig 7.1: Loading the Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Fig 7.3: Table Information

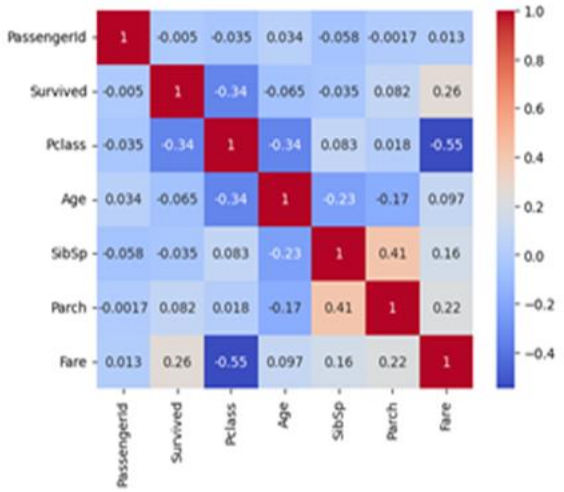

Fig 7.4:Correlation matrix

Algorithm Used	Existing system Accuracy	Proposed System Accuracy
Logistic Regression	75%	80.34%
Support Vector Classifier	N/A	79.21%
K-Nearest Neighbor	N/A	77.53%
XG-Boost	N/A	82.02%
Gradient Boosting Classifier	N/A	82.02%
Decision tree	71%	75.84%
Random forest	75%	80.9%

Fig 7.6: Final Result

VIII.CONCLUSION

In this study, the survival and death rates of Titanic passengers were examined using a thorough system architecture. We attempted to get insights into the association between these characteristics and survival outcomes by utilising traditional machine learning techniques and taking into account a variety of factors like passenger class, age, sex, passenger ID, and sibling count. By addressing missing values and encoding categorical variables, we were able to ensure the quality and integrity of the dataset through data preparation. We were able to choose the most important variables for our research using feature selection techniques. Following that, we created and trained a variety of models, including decision trees, random forests, XGBoost, gradient boosting methods, K-Nearest Neighbours (KNN), and logistic regression.

We were able to assess the models' forecast accuracy using performance measures. The outcomes gave important information on how well each algorithm predicted survival outcomes. Cross-validation was also carried out to evaluate the models' generalisation abilities and reduce overfitting.

The study showed clear patterns and connections between the Titanic survival rates and the parameters that were taken into account. We acquired a better understanding of the dynamics at work during the sad occurrence by looking at the effects of passenger class, age, sex, passenger ID, and sibling count.Overall, by highlighting the value of using several methods and taking into account a variety of parameters when analysing survival rates in historical disasters like the Titanic, our study makes a contribution to the fields of data analysis and machine learning. The system design we used allowed for a methodical and thorough approach, resulting in precise forecasts and insightful findings.

The results we obtained can be expanded upon in future study by investigating new variables and improving the models to improve their prediction power. Additionally, investigating other machine learning algorithms or cutting-edge methods may enhance the precision of survival forecasts and offer new insights.In the end, the information gleaned from this study can help in unravelling the underlying causes that influenced Titanic survivorship, perhaps guiding future planning and reaction plans for disasters.

REFERENCE

[1] Kaggle.com, 'Titanic:Machine Learning form Disaster',[Online]. Available: http://www.kaggle. com/. [Accessed: 10- Feb- 2017]. L. Cao, Q. Jiang, M. Cheng, C. Wang,

[2] 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

[3] Mikhael Elinder.(2012). 'Gender, social norms, and survival in maritime disasters', [Online]. Available: http://www.pnas.org/content/109/33/13220.full. [Accessed: 8- March - 2017].

[4] Frey, B. S., Savage, D. A., and Torgler, B, "Behavior under extreme conditions: The Titanic disaster", The Journal of Economic Perspectives, 25(1), pp. 209-221, 2011.

[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/.

[6] Santos, K.C.P, Barrios, E.B, "Improving Predictive accuracy of logistic regression model using ranked set sample," Communication in statistic.

[7] Sun, Y.; Li, G.; Zhang, J.; Huang, J. Rockburst Intensity Evaluation by a Novel Systematic and Evolved Approach: Machine Learning Booster and Application. Bull. Eng. Geol. Environ. 2021, 80, 8385–8395.

[8] Cai, X.; Cheng, C.; Zhou, Z.; Konietzky, H.; Song, Z.; Wang, S. Rock Mass Watering for Rock-Burst Prevention: Some Thoughts on the Mechanisms Deduced from Laboratory Results. Bull. Eng. Geol. Environ. 2021, 80, 8725–8743.

[9] Zhou, Z.; Cai, X.; Li, X.; Cao, W.; Du, X. Dynamic Response and Energy Evolution of Sandstone Under Coupled Static–Dynamic Compression: Insights from Experimental Study into Deep Rock Engineering Applications. Rock Mech. Rock Eng. 2020, 53, 1305–1331.

[10] Pines JM, Iyer S, Disbot M, Hollander JE, Shofer FS, Datner EM. The effect of emergency department crowding on patient satisfaction for admitted patients. Acad Emerg Med. 2008;15(9):825–31. doi: 10.1111/j.1553-2712.2008.00200.x