Who Can Survive the Titanic Disaster? A Decision Tree Analysis (July 2023)

Rijan Ghimire[1] and Sujan Bhattrai[1]

[1]Department of Electronics and Computer Engineering, IOE, Thapathali Campus, Kathmandu 44600, Nepal
Corresponding author: Rijan Ghimire (rijanghimire96@gmail.com)

ABSTRACT Decision trees are widely used and highly interpretable machine learning algorithms that can solve classification and regression problems. They consist of internal nodes representing features and leaf nodes indicating class labels or predictions. The construction process involves recursive partitioning based on the most informative features, using criteria like Gini Index or Information Gain. Decision trees possess qualities such as interpretability, ease of use, and the ability to handle categorical and numerical features, making them robust against outliers and capable of capturing non-linear relationships. This paper explores the capabilities of Decision tree by applying it on the titanic dataset and making inference about survival. we constructed a decision tree model to uncover the underlying patterns and predictors of survival. Through prepossessing, model construction, and evaluation, we extract valuable insights and validate the model's effectiveness in accurately predicting survival outcomes. The findings contribute to a better understanding of the factors that influenced survival on the Titanic, offering valuable knowledge for historians, researchers, and enthusiasts, and potentially aiding in the development of predictive models for similar scenarios.

INDEX TERMS Decision tree analysis, Titanic dataset, survival outcomes, patterns, interpretability

I INTRODUCTION

A decision tree is a powerful and interpretable machine learning model that constructs a tree-like structure to represent the relationships between features and the target variable. In the context of the Titanic dataset, a decision tree can capture the underlying patterns and rules that determined the survival outcomes of passengers. By examining the features and their values, the decision tree algorithm selects the most informative feature to split the data, creating branches based on different feature values. This recursive splitting process continues until a stopping criterion is met, resulting in a tree that can be used to make predictions for new instances. Decision trees are known for their ability to handle both numerical and categorical features, capture complex relationships, and provide interpretability by visualizing the learned rules.

The Titanic on its voyage in April 1912 resulted in the loss of approximately 1,502 lives out of the 2,224 passengers and crew on board, has remained a subject of fascination and study. Researchers and data scientists have diligently analyzed the available data to gain insights into the factors that influenced the survival of passengers. This ongoing analysis seeks to uncover patterns and correlations among various variables, such as age, gender, social class, and cabin location, to understand the dynamics that determined who survived and who perished in this historic maritime tragedy. By examining the Titanic dataset [1], researchers aim to shed light on the human stories and the socio-economic factors that played a role in shaping the outcomes of this catastrophic event. This paper aims to utilize a decision tree algorithm to conduct a comprehensive analysis of the Titanic dataset. By employing a decision tree model, the we intend to uncover and interpret the underlying patterns and rules that influenced the survival outcomes of passengers in the Titanic disaster. The decision tree algorithm will enable the identification of the most informative features and their thresholds for splitting the data, allowing for the creation of a tree-like structure that represents the relationships between the features and the target variable. This approach promises to provide valuable insights into the factors that played a significant role in determining passenger survival, contributing to a better understanding of this historic event. The Titanic dataset underwent some preprocessing steps, including handling missing values, outliers, and categorical variables, to ensure the quality of the data. Once the dataset was cleaned and prepared, a decision tree was constructed using a recursive splitting approach. The decision tree algorithm selected the most informative features and used a suitable criterion, such as entropy or information gain, to create decision nodes that effectively separated survivors from non-survivors. By iteratively partitioning the data based on these learned rules, the decision tree uncovered valuable insights into the patterns and factors that could have determined the survival outcomes of passengers in the Titanic disaster.

The resulting decision tree provided a visual representation of the learned rules, with each node representing a feature and each branch representing a decision. By analyzing the decision tree, we could infer about factors that could have played a crucial role in determining passengers survival. The decision tree structure allows us to interpret the importance and hierarchy of features in influencing the outcome.

To evaluate the performance of the decision tree model, we employ appropriate metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques are utilized to assess the model's generalization ability and ensure its reliability in predicting survival outcomes accurately. By utilizing a decision tree algorithm for analyzing the Titanic dataset, we not only extract valuable insights into the factors influencing survival but also provide a transparent and interpretable framework for understanding these relationships. The decision tree model offers a logical and intuitive representation of the learned rules, enabling historians, researchers, and enthusiasts to comprehend the intricate dynamics of the Titanic disaster. Additionally, the findings from this analysis may have implications for decision-making processes, safety measures, and disaster management strategies in related scenarios.

II RELATED WORK

Decision tree algorithms have been widely utilized for analyzing various datasets to uncover patterns and predict outcomes. These algorithms provide a structured and interpretable approach for understanding the relationships between features and the target variable. The hierarchical nature of decision trees allows for a visual representation of the decision-making process, enabling researchers to derive meaningful rules and insights.

Durmuş et al. applied decision trees to analyze the Titanic dataset and found that using data that contributes significantly to the model yields more successful results, with a classification accuracy of 81.57% [2].Singh et al. compared various machine learning methods including decision trees on the Titanic dataset and obtained the accuracy of 93.6% with decision trees [3].Kakde et al. conducted a comprehensive analysis of the Titanic dataset, employing various machine learning algorithms including logistic regression, decision tree, random forest, and support vector machines. Their study aimed to compare the performance of these algorithms in terms of accuracy for classification problems. The results indicated that both logistic regression and support vector machines exhibited good accuracy in predicting survival outcomes [4].Barhoom et al.used JustNN Tool for analysis and examined the impact of input variables based on existing literature. The ANN model yielded a remarkably high prediction accuracy of 99.28%, surpassing the performance of the original datasets . The study identified that variables such as sex, passenger class (Pclass), and cabin had significant effects on the likelihood of survival for the given problem. These findings emphasize the importance of these variables in determining the survival outcomes of Titanic passengers and contribute to the understanding of the dynamics of the disaster [5].

Decision tree models have demonstrated accuracy in predictions and offer interpretability, making them valuable tools for analysis and predictions. This paper contributes to the existing body of knowledge by applying a decision tree algorithm to the Titanic dataset, aiming to uncover significant predictors and provide valuable insights into the dynamics of passenger survival.

III METHODOLOGY

A DATASET DESCRIPTION

The Titanic dataset consists of information about passengers on the ill-fated maiden voyage of the RMS Titanic in 1912. It contains various attributes for each passenger, such as age, sex, ticket class, fare, cabin, and whether they survived or not. The original dataset is based on information gathered by the British Board of Trade from the official inquiry into the Titanic disaster [6]. The dataset was made publicly available by Encyclopedia Titanica, which provides extensive resources related to the Titanic [7]. Kaggle, a popular data science platform, hosts a version of the Titanic dataset for public use. The Kaggle dataset includes the same information as the original dataset but is preprocessed and standardized for easy analysis. The dataset consists of 891 data points (passengers) and 12 features describing various aspects of each passenger's profile and circumstances. It can be accessed at Titanic: Machine Learning from Disaster [8]

In the given dataset, the features represent various attributes of the passengers onboard the Titanic.

1. PassengerId: It is a unique identifier assigned to each passenger, helping to distinguish one passenger from another.

2. Survived: This feature serves as an indicator variable, where 0 represents "No" indicating that the passenger did not survive, and 1 represents "Yes" indicating that the passenger survived.

3. Pclass: It denotes the ticket class, with values 1, 2, and 3 representing the first, second, and third class respectively. The ticket class provides information about the socio-economic status of the passenger.

4. Name: This feature represents the name of the passenger.

5. Sex: It indicates the gender of the passenger, with values "female" and "male" specifying the respective genders.

6. Age: It represents the age of the passenger in years, providing information about their age group.

7. SibSp: This feature signifies the number of siblings/spouses the passenger had aboard the Titanic, indicating the presence of family members.

8. Parch: It represents the number of parents/children the passenger had aboard the Titanic, indicating the presence of family members.

9. Ticket: This feature denotes the ticket number assigned to the passenger.

10. Fare: It represents the fare paid by the passenger for their ticket, reflecting the cost of their journey.

11. Cabin: This feature represents the cabin number assigned to the passenger, indicating their allocated cabin on the Titanic.

12. Embarked: It denotes the port of embarkation, with "C", "Q", and "S" representing Cherbourg, Queenstown, and Southampton respectively. The port of embarkation signifies the location where the passenger boarded the Titanic.

These features provide insights into various aspects of the passengers, such as their demographics, family size, ticket information, and cabin details. Researchers and data enthusiasts can use these features to analyze the factors that influenced the survival of passengers and to develop predictive models to determine the likelihood of survival based on these attributes.

B PROPOSED METHODOLOGY

In this report, we utilize the Titanic dataset to investigate the prediction of passenger survival using a decision tree classifier. The methodology involves several steps. Firstly, we perform data preprocessing by handling missing values and encoding categorical variables. Additionally, we drop three columns, namely "Embarked," "Name," and "Cabin," based on their limited relevance to the prediction task. We apply feature selection techniques to identify the most informative features for the decision tree classifier. Subsequently, we split the dataset into training and testing sets. The decision tree classifier is then trained with pruning techniques, such as cost complexity pruning or reduced error pruning, to enhance its generalization capability and avoid overfitting. Model evaluation is conducted using various metrics, such as accuracy, precision, recall, and F1-score, to assess the classifier's performance. We explore the decision tree's interpretability by visualizing its structure, gaining insights into the most influential features for predicting passenger survival. This approach allows us to build a more interpretable and optimized decision tree model for predicting passenger survival on the Titanic.

Algorithm 1 Titanic Dataset with Decision Tree Classifier

Require: Titanic dataset
Ensure: Survival prediction using decision tree
1: Import necessary libraries
2: Load train and test data from CSV files
3: Preprocess the data by dropping irrelevant columns (Name, Cabin, Embarked) and handling missing values
4: Convert categorical variables to numerical labels
5: Split the data into training and test sets
6: Train decision tree classifier with pruning
7: Train decision tree classifier without pruning
8: Predict target variable for the test set
9: Evaluate classifier performance
10: Visualize the decision tree

C MATHEMATICAL FORMULAE

These formulas are used in decision trees to evaluate the impurity or disorder of a set of samples (using entropy or Gini impurity), the classification error rate, and the usefulness of a feature for splitting the data (using information gain and gain ratio).

Entropy (H): The formula for entropy measures the impurity or disorder in a set of samples. For binary classification:

$$H(S) = -p_0 \log_2(p_0) - p_1 \log_2(p_1) \tag{1}$$

Where p_0 is the probability of class 0 and p_1 is the probability of class 1. Also, the Generalized formula:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{2}$$

Where c represents the number of classes, and p_i is the probability of class i.

Gini Impurity (G): The Gini impurity measures the probability of misclassifying a randomly chosen element in a set. For binary classification:

$$G(S) = 1 - p_0^2 - p_1^2 \tag{3}$$

Where p_0 is the probability of class 0 and p_1 is the probability of class 1. Also, the Generalized formula:

$$G(S) = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{4}$$

Where c represents the number of classes, and p_i is the probability of class i.

Classification Error (CE): The classification error is the fraction of misclassified instances in a set. For binary classification:

$$CE(S) = 1 - \max(p_0, p_1) \tag{5}$$

Where p_0 is the probability of class 0 and p_1 is the probability of class 1. Also, the Generalized formula:

$$CE(S) = 1 - \max(p_1, p_2, \ldots, p_c) \tag{6}$$

Where c represents the number of classes, and p_i is the probability of class i.

Information Gain (IG): Information gain measures the reduction in entropy achieved by splitting the data on a particular feature. The formula for information gain is:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{7}$$

Where $IG(S, A)$ represents the information gain of feature A in dataset S, $H(S)$ is the entropy of the original dataset S, $Values(A)$ represents the set of all possible values for feature A, $|S_v|$ denotes the number of samples in dataset S where feature A has value v, and $H(S_v)$ is the entropy of the subset S_v where feature A has value v.

Gain Ratio: The gain ratio is an adjustment to information gain that takes into account the intrinsic information of a feature. The formula for the gain ratio is:

$$\text{GainRatio}(S, A) = \frac{\text{IG}(S, A)}{\text{SplitInfo}(S, A)} \tag{8}$$

Where $\text{GainRatio}(S, A)$ represents the gain ratio of feature A in dataset S, $\text{IG}(S, A)$ is the information gain of feature A in dataset S, and $\text{SplitInfo}(S, A)$ is the split information of feature A in dataset S. The split information measures the potential information generated by the feature A in dataset S.

Now the formulae used for the performance evaluation are as follows:

Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

where: - TP: True Positives (correctly predicted positive instances) - TN: True Negatives (correctly predicted negative instances) - FP: False Positives (incorrectly predicted positive instances) - FN: False Negatives (incorrectly predicted negative instances)
Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

where: - TP: True Positives (correctly predicted positive instances) - FP: False Positives (incorrectly predicted positive instances)
F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where: - Precision: Ratio of correctly predicted positive instances to the total predicted positive instances - Recall: Ratio of correctly predicted positive instances to the total actual positive instances
Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

where: - TP: True Positives (correctly predicted positive instances) - FN: False Negatives (incorrectly predicted negative instances)

D INSTRUMENTATION DETAILS

The problem was tackled using Jupyter Notebook as the The problem at hand was addressed using the Jupyter Notebook as the computational environment, which offered an interactive and iterative approach to the analysis. To tackle different aspects of the problem, various Python libraries were utilized. The NumPy library was employed for efficient numerical operations and array manipulation, allowing for seamless handling of numerical data. Additionally, Pandas played a crucial role in data handling and preprocessing tasks, providing convenient functions for data manipulation and cleaning.
In order to gain insights and explore the dataset, visualization and exploratory data analysis were performed using Seaborn and Matplotlib. These libraries enabled the generation of informative plots and graphs, aiding in the identification of patterns and trends within the data. For the classification task, the scikit-learn library was leveraged to implement the DecisionTreeClassifier model. This model is widely used for classification problems and offers flexibility in terms of parameter tuning. The model's performance was evaluated using various metrics, including the confusion matrix, F1 score, accuracy, precision, and recall. These metrics provide a comprehensive assessment of the model's effectiveness in predicting the target variable.
By combining these powerful libraries within the Jupyter Notebook environment, a robust and efficient toolset was created for addressing the problem at hand. The interactive nature of Jupyter Notebook allowed for seamless exploration, analysis, and experimentation, ultimately leading to valuable insights and conclusions.

E SYSTEM BLOCK DIAGRAM

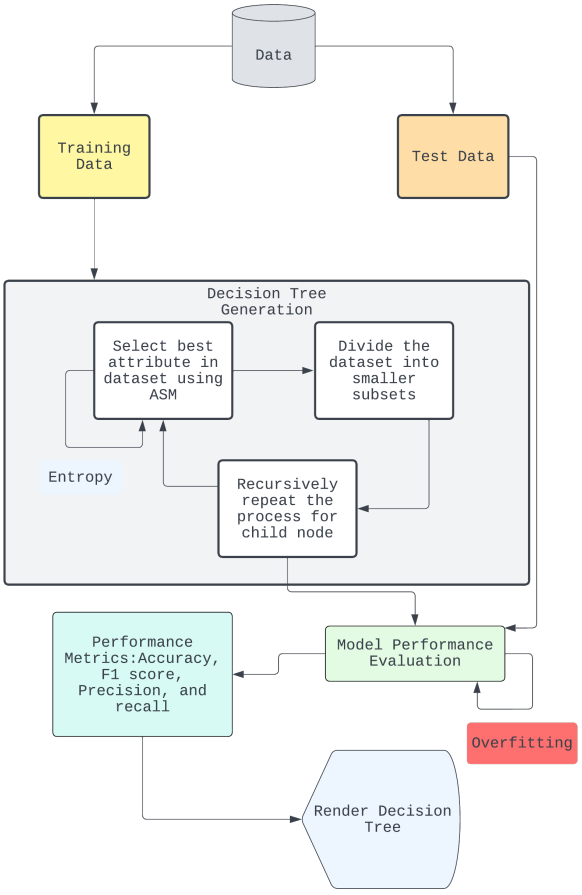

Figure 1: Block Diagram for Decision Tree

IV EXPERIMENTAL RESULTS

In this study, we performed Decision Tree analysis on the Titanic dataset to predict the survival outcomes of passengers. We first prepared the dataset by applying one-hot encoding to the categorical features, resulting in a one-hot encoded dataset as shown in Table 2. Next, we conducted Decision Tree analysis with different pruning settings to assess the model's performance. The results are summarized in Table 3.
When using pruning with a maximum depth of 8, we achieved an F1 score of 0.7317, indicating a reasonable balance between precision and recall. The accuracy of the model was 0.7953, indicating a high level of correct predictions. The precision value of 0.8 suggests that the model correctly identified the positive instances, while the recall value of 0.6742 indicates the model's ability to capture the positive instances from the dataset.
Similarly, when applying pruning with a maximum depth of 4, we obtained an F1 score of 0.7303, demonstrating consistent performance in balancing precision and recall. The accuracy achieved was 0.7767, indicating a strong predictive capability. The precision and recall values of 0.7303

highlight the model's ability to correctly classify positive instances.

For the pruning setting with a maximum depth of 3, the F1 score was 0.6871, indicating a relatively lower balance between precision and recall. The accuracy achieved was 0.7628, indicating a satisfactory level of correct predictions. The precision value of 0.7568 suggests the model's ability to accurately classify positive instances, while the recall value of 0.6292 indicates the model's ability to capture the positive instances effectively.

Finally, when applying pruning with a maximum depth of 2, the F1 score achieved was 0.6761, indicating a slightly lower balance between precision and recall. The accuracy achieved was 0.7860, indicating a high level of correct predictions. The precision value of 0.9057 indicates the model's ability to correctly identify positive instances, while the recall value of 0.5393 suggests that the model captured a moderate number of positive instances.

Additionally, we visualized the results using figures. Figure 2 contains 5 subfigures (a-e) representing the confusion matrices for the different pruning settings. These matrices provide insights into the performance of the models in terms of true positive, true negative, false positive, and false negative predictions.

Furthermore, figures 3 to 7 depict the decision trees for different pruning depths: 17, 8, 4, 3, and 2, respectively. These decision trees illustrate the hierarchical structure of the models, showcasing the splitting criteria and resulting classifications at each node.

These figures and results collectively demonstrate the effectiveness of Decision Tree analysis in predicting survival outcomes for the Titanic dataset.

V DISCUSSION AND ANALYSIS

The experimental results obtained from the Decision Tree analysis on the Titanic dataset provide valuable insights into the performance and behavior of the models. Let us delve into the discussion and analysis of these results.

Firstly, the F1 scores obtained for different pruning depths indicate the overall effectiveness of the models in achieving a balance between precision and recall. Higher F1 scores signify better performance in correctly classifying both positive and negative instances. In our experiments, the F1 scores ranged from 0.6761 to 0.7317, suggesting that the Decision Tree models exhibit reasonably good predictive capabilities.

The accuracy values obtained for the models range from 0.7628 to 0.7953, indicating a relatively high level of correct predictions. Higher accuracy values imply a better overall performance of the models in correctly classifying the survival outcomes. It is worth noting that the model with a maximum depth of 8 achieved the highest accuracy of 0.7953, indicating its superior performance compared to the other models.

Precision and recall values further shed light on the models' predictive abilities. Precision represents the proportion of correctly predicted positive instances out of the total instances predicted as positive. Recall, on the other hand, indicates the proportion of correctly predicted positive instances out of the actual positive instances in the dataset. The precision values range from 0.8 to 0.9057, highlighting the models' ability to accurately classify positive instances. The recall values range from 0.5393 to 0.6742, indicating the models' effectiveness in capturing positive instances from the dataset.

Considering the confusion matrices depicted in Figure 2, we can observe the distribution of true positive, true negative, false positive, and false negative predictions for each pruning depth. These matrices provide insights into the models' performance and highlight potential areas for improvement. By examining the confusion matrices, we can identify the specific types of errors made by the models and further refine the decision-making process.

Moreover, the visualizations of the decision trees (Figures 3-7) offer a comprehensive understanding of the hierarchical structure of the models. These decision trees illustrate the splitting criteria and the resulting classifications at each node, allowing us to interpret the decision-making process. By analyzing the decision trees, we can gain insights into the most influential features for predicting survival outcomes and identify potential patterns and rules.

Overall, the experimental results and analysis demonstrate the viability and effectiveness of Decision Tree analysis for predicting survival outcomes in the Titanic dataset. The models achieved reasonably high accuracy, precision, and recall values, indicating their potential for practical applications. However, further optimization and analysis can be conducted to enhance the models' performance and generalization capabilities.

VI CONCLUSION

The application of Decision Tree analysis on the Titanic dataset has provided valuable insights into the behavior and performance of the models. The obtained results shed light on the predictive capabilities and effectiveness of the Decision Tree models in predicting survival outcomes. The F1 scores achieved by the models indicate a reasonably good balance between precision and recall, suggesting satisfactory predictive capabilities. The accuracy values obtained for the models indicate a relatively high level of correct predictions. The precision and recall values further emphasize the models' predictive abilities, demonstrating their ability to accurately classify positive instances and effectively capture positive instances from the dataset.

The confusion matrices provided insights into the models' performance and identified potential areas for improvement. Analyzing the decision trees allowed for the interpretation of influential features and the identification of patterns and rules that contributed to the survival outcomes. Overall, the experimental results and analysis confirm the viability and effectiveness of Decision Tree analysis for predicting survival outcomes in the Titanic dataset. Further optimization and analysis can be conducted to enhance the models' performance and generalization capabilities in practical applications.

REFERENCES

[1] Kaggle.com. Titanic - machine learning from disaster. [Online]. Available: https://www.kaggle.com/c/titanic/data

[2] B. Durmuş and İşçi Güneri, "Analysis and detection of titanic survivors using generalized linear models and decision tree algorithm," *International Journal of Applied Mathematics Electronics and Computers*, vol. 8, no. 4, pp. 109 – 114, 2020.

[3] A. Singh, S. Saraswat, and N. Faujdar, "Analyzing titanic disaster using machine learning algorithms," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 406–411, 2017.

[4] Y. Kakde and S. Agrawal, "Predicting survival on titanic by applying exploratory data analytics and machine learning techniques," *International Journal of Computer Applications*, vol. 179, no. 44, pp. 32–38, May 2018. [Online]. Available: http://www.ijcaonline.org/archives/volume179/number44/29430-2018917094

[5] A. M. Barhoom, S. Abu-Naser, B. Abu-Nasser, A. Khalil, and M. Musleh, "Predicting titanic survivors using artificial neural network," 09 2019.

[6] "Report on the loss of the 'titanic.'." [Online]. Available: https://www.titanicinquiry.org/BOTInq/BOTReport/botRep01.php

[7] E. Titanica. [Online]. Available: https://www.encyclopedia-titanica.org/

[8] Kaggle.com. Titanic - machine learning from disaster. [Online]. Available: https://www.kaggle.com/c/titanic/

Rijan Ghimire is currently pursuing his undergraduate degree in Electronics, Communication, and Information at IOE, Thapathali Campus. His research interests encompass various areas, including data mining, network communications, and optimization theory and technology.(THA076BEI022)

Sujan Bhattrai is currently pursuing his undergraduate degree in Electronics, Communication, and Information at IOE, Thapathali Campus. His research interests cover various areas, including data mining, deep learning, operating system, robotics, and power electronics.(THA076BEI037)