Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Survival Prediction and Comparison of the Titanic based on Machine Learning Classifiers

Tony Wayne Wang *

The High School Attached to Northeast Normal University, Changchun, China

* Corresponding Author Email: tony_wang@hsannu.com

Abstract. This study conducts a comparison of machine learning algorithms, including Logistic Regression (LR), Decision Tree Classifier (DT), and Random Forest Classifier (RF), to predict the survival outcomes of passengers on the Titanic. The dataset used in the study includes variables such as socio-economic status, age, gender, and family relationships; this paper meticulously prepares and analyzes the data to train and evaluate these models. The study's objective is to determine the impact of various passenger features on survival outcomes, employing machine learning algorithms to generate survival predictions. The findings demonstrate that the RF model, particularly with 45 or 75 trees, significantly outperforms LR and DT in terms of precision and recall, establishing it as a more robust classifier for this dataset. The research underscores the importance of the utility of different machine learning models for binary classification tasks and the role of parameter tuning in enhancing model performance. This comparative analysis not only contributes to the ongoing exploration of the Titanic disaster through data science but also highlights key considerations in the application of machine learning algorithms for predictive modeling.

Keywords: Machine learning; Random Forest Classifier; Logistic Regression; Decision Tree Classifier.

1. Introduction

The tragic sinking of the Titanic on April 15, 1912, remains one of the most profound maritime tragedies in history. During its maiden voyage, the ship collided with an iceberg, leading to a catastrophe that resulted in the loss of 1,502 lives from the 2,224 passengers and crew. It has been observed that survival rates differed across different groups, suggesting that certain demographic characteristics may have played a role in influencing the likelihood of surviving the tragic event [1]. This event has not only captivated public interest for over a century but also provided a rich dataset for statistical analysis and predictive modeling in the field of machine learning. Predicting survival on the Titanic involves analyzing various passenger features, such as age, sex, passenger class, and others, to determine their impact on survival outcomes. This problem is a classic example of a binary classification task, where the objective is to predict a binary outcome—survival or non-survival—based on a set of variables. Survival predictions are generated through machine learning algorithms, utilizing various combinations of features [2]. It serves as an excellent case study for applying and comparing different predictive modeling techniques, including Logistic Regression (LR), Decision Tree Classifier (DT), and Random Forest Classifier (RF).

Shawn, John, Clarke, and Muniswamaiah [3] conducted DT and Cluster Analysis. Their study indicates that of all the variables considered, gender was the predominant determinant in the survival outcomes of the passengers. Reference [4] utilized multiple LR to assess passenger survival outcomes. Through a comparative analysis of performance metrics across various scenarios, Chatterjee concluded that LR attained a maximum accuracy of 80.756%. While the DT and RF classifiers yielded 84% and 81% correctly classified instances, respectively. Feature engineering played a crucial role in these outcomes. Akriti Singh [5] compares the accuracy of survival predictions using LR, RF, DT, and Naive Bayes. They focused on several variables, including sex, Pclass, age, and whether passengers had children, to estimate survival rates. Their findings suggest that logistic regression is the most effective algorithm for predicting survival rates accurately, primarily due to its

lower false rate compared to the other evaluated algorithms. Using the Titanic dataset, Eric Lam, and Tang [6] conducted a comparative analysis of three distinct algorithms: Naive Bayes, Decision Tree Analysis, and Support Vector Machine (SVM). Their findings indicated that gender emerged as the most influential feature for accurately forecasting survival. Furthermore, they emphasized the significance of selecting crucial features to enhance prediction performance. Further, reference [7] introduced a "Gender-based model" heavily reliant on the 'sex' feature, which achieved the superior accuracy of 78.469%, marginally surpassing Random Forests by 2%. Research indicated in reference [8] that utilizing a variety of combinations, the application of Naïve Bayes, SVM, and Decision Trees was thoroughly explored. The outcomes demonstrated that the Decision Tree Algorithm outshined with an accuracy of 70.43%, positioning it as the most effective among the techniques evaluated. Conversely, Naïve Bayes was less effective, achieving a lower accuracy of 76.79%. The SVM method marked a middle ground, securing a 77.99% accuracy rate. the 'sex' attribute emerged as a critical determinant, enhancing the Decision Tree Classifier's accuracy to 81%.

This research primarily aims to forecast the survival outcomes of the Titanic's passengers by applying LR, DT, and RF and to evaluate the varying performances of these models. Specifically, first, the data set is pre-processed to make it available for machine learning. Second, for these three types of models, LR DT RF is built for the baseline models. Third, the predictive capabilities of the various models are assessed and contrasted. The experimental results demonstrate that LR and RF models excel with high precision and recall. However, DT model slightly lags behind in performance metrics. This paper could advance the field of data science by practicing and comparing different models to promote the application and innovation of machine learning techniques. What's more, this research can provide an understanding of the critical factors of passenger survival and sinking the rest.

2. Methodology

2.1. Dataset Description and Preprocessing

Table 1. Dataset description

Variable	Survival	P Class	Sex	Age	sibsp
Definition	0 = No, 1 = Yes	Ticket class	Sex	Age	number of siblings and spouses

The dataset is from Kaggle [9] which contain following variables: survival, pclass, sex, Age, sibsp, parch, Ticket number, Passenger fare, Cabin number and port of embarkation. The specific information is shown in Table 1. What's more, from the data set, the author can notice that the age peaks around the age of 20, and most of the people live in the three cabins. Figure 1 and Figure 2 showcase some instances from the dataset. As for the preprocessing, initially, the author undertook an exploratory data analysis to identify correlations and distributions among features. Pre-processing involves completing missing values in Age, Fare, and Cabin. As for the age fair, the author uses means to deal with it. Moreover, the study uses the value "U" to deal with the missing values of "Cabin", which is the most frequent occurrence. One-Hot Encoding is applied to convert categorical variables into a form that could be provided to models. Also remove "Passengerid" "Name" and "Ticket", because they are either unique identifiers that have no predictive relationship to the result or contain a high percentage of missing values [10].

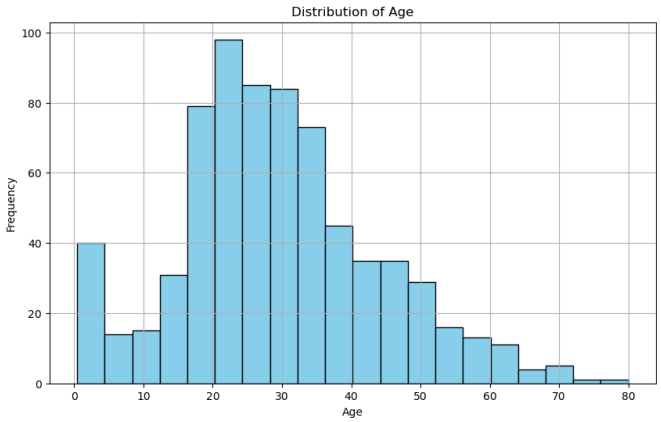

Figure 1. Frequency Distribution Across Age Groups

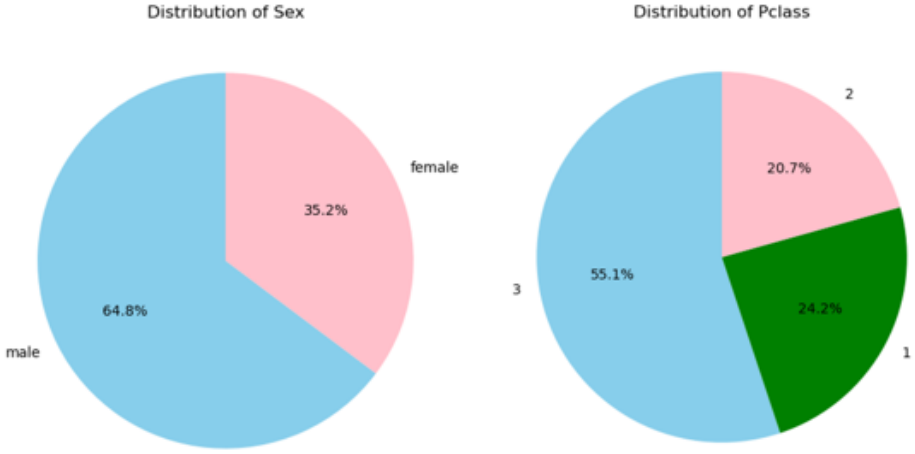

Figure 2. Data description

2.2. Proposed Approach

The author evaluates several machine learning models to ascertain their efficacy in survival prediction. The models include LR, DT, and RF. Each model is selected based on its suitability to handle the dataset's characteristics and its predictive performance in preliminary tests. It is worth noting that in order to ensure experimental consistency, the same training set and test set are used for all models. At last, all the models with different settings are compared. The pipeline is provided in Figure 3.

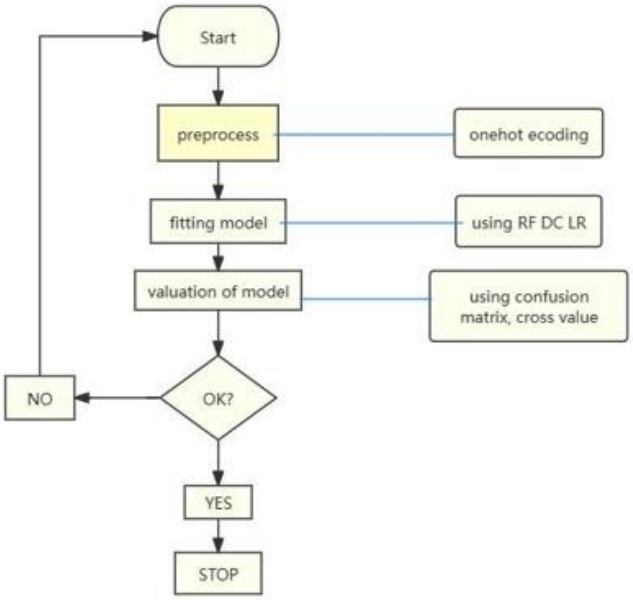

Figure 3. Pipeline

2.2.1. Logistic Regression (LR)

LR is an analytical method of choice when addressing situations where the dependent variable is binary, signifying outcomes that can be categorized distinctly, for instance, as 'success' or 'failure'[11]. Its utility spans various levels of measurement for independent variables, whether nominal, ordinal, interval, or ratio. In the finance sector, Logistic Regression (LR) can be employed to predict the probability of a loan applicant defaulting using a variety of financial indicators. The mathematical formulation of the LR model is expressed through the equation: $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$. In this formula, P represents the probability that a given input point belongs to the category labeled '1'. The term beta_0 is the intercept, and beta_1 is the coefficient of the model, while (x) is the independent variable. This logistic function is used to model the odds of the dependent variable being '1' for different values of the independent variables. As depicted in Figure 4, the logistic function curve (in red) illustrates the estimated probability as 'x' varies. The data points (in blue) reflect the observed outcomes, and the curve projects the changing likelihood of the event '1' as the independent variable increases, exemplifying the predictive power of the LR model."

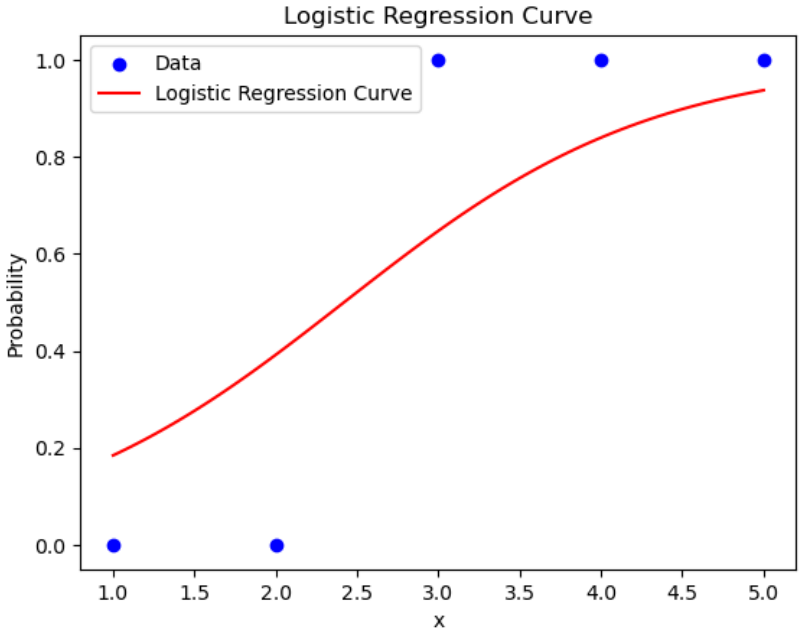

Figure 4. a LR curve

2.2.2. Decision tree (DT)

A DT is a supervised learning algorithm predominantly used for classification tasks. It is simple to understand and interpret [12]. It can simply be plotted, as shown in Figure 5. The formular can be expressed $\left[L_E(i) = -\sum_{j=1}^{m} f(i,j)\log_2 f(i,j)\right]$ For each outcome j, the product of the probability f(i, j) and the base-2 logarithm of that probability $\log_2 f(i,j)$ is calculated and then negated. This product represents the "information content" associated with the occurrence of outcome j within event i. Entropy is a measure of impurity or disorder. This equation is applied at each node to identify the optimal division by computing the information gain, which represents the change in entropy from before to after the split. In more detail, for each potential split, the DT algorithm will calculate the expected reduction in entropy - the information gain. The data is partitioned using the split that achieves the maximum information gain. This process repeats recursively, leading to the construction of a tree until the stopping criteria are met. Nevertheless, learners using decision trees may generate overly intricate trees that do not generalize well beyond the training data, a phenomenon referred to as overfitting [13].

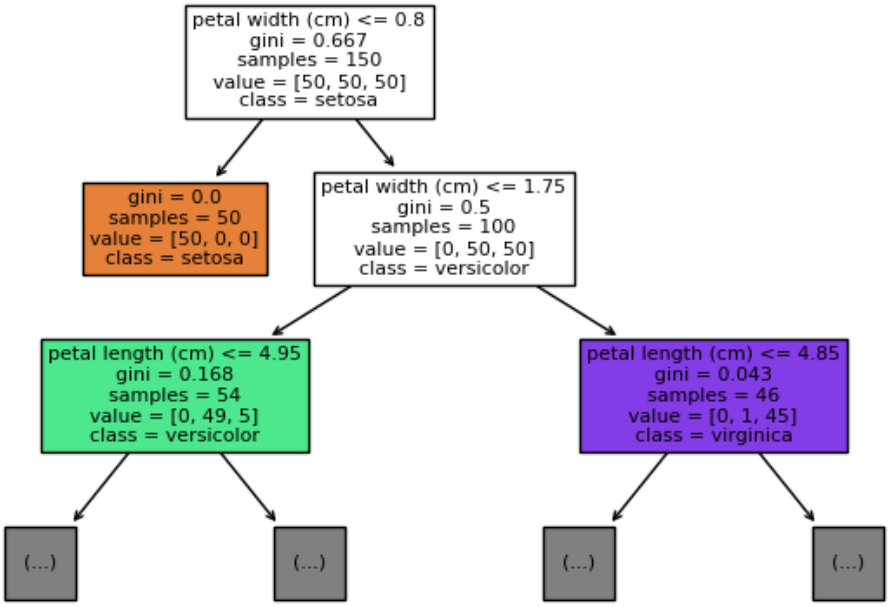

Figure 5. LR for Iris dataset

2.2.3. Random Forest (RF)

The RF algorithm is a supervised classification technique that builds a "forest" of DTs. In Random Forest (RF), every tree is constructed using a randomly chosen subset of the training data and a random selection of the features, such as in Figure 6. RF is effective for a wide range of applications. Random decision forests address the tendency of DTs to overfit their training data [14]. The formular of RF is$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} y_b(x)$. To be specific, 1/B represents the averaging operation. B is the total number of trees in the forest. By dividing by B, we compute the average prediction across all trees. ∑is the summation symbol, indicating that we sum over all B trees in the forest. $y\_b(x)$ represents the prediction made by the b^th tree for the input x. In summary, this formula computes the average prediction $\hat{y}(x)$ for an input x by summing up the predictions $y\_b(x)$ from all B trees in the forest and then dividing by the number of trees B. This method is used in RF regression to generate the final prediction, reducing sensitivity to overfitting of individual trees, and leveraging the power of ensemble learning.

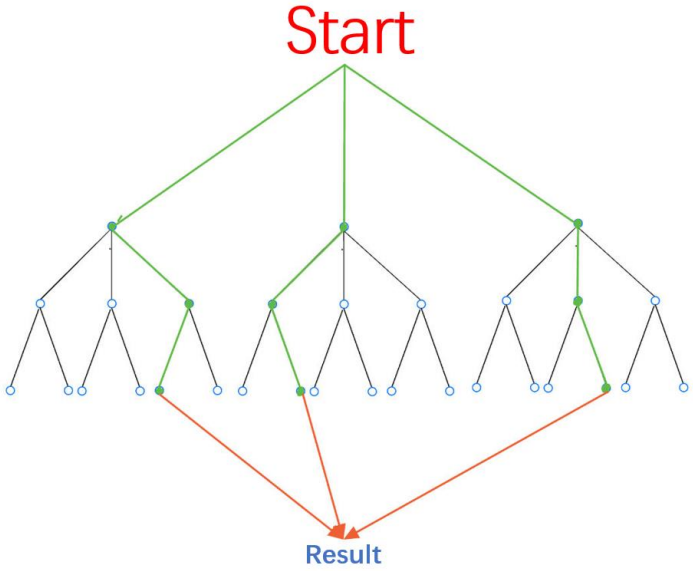

Figure 6. RF decryption

2.3. Implementation details

The study uses Python 3.11.5 and the sklearn library for LR DT RF models. Data visualization is done by using matplotlib.pyplot library. The study was run on a Windows 10 device with CPU Intel core i5-7200U. DT and RF models employ varying maximum depths: 5, 45, and 75 while keeping other parameters at their default values. For DT and RF, the default maximum depth is typically set to None. This means that the trees continue to grow until each leaf node is completely homogeneous (pure) or until the number of samples in a leaf node is less than the predefined threshold for splitting, typically set at 2.Additionally, other parameters such as criterion (which measures the quality of a split), splitter (which strategy to use for the split), and others are set to their defaults. As for the LR model, all settings are kept at their default values, including using an L2 penalty ('l2') with a regularization strength (C) of 1.0.

3. Results and Discussion

In this section, the author presents the results and discussion of the experiments comparing three machine learning models, LR, DT, and RF, on the classification dataset. The models were evaluated based on their cross-validation scores, confusion matrices, precision, recall, and ROC curves.

3.1. LR

The LR model achieved cross-validation scores of 0.789. The confusion matrix for the model is$\begin{bmatrix} 97 & 20 \\ 13 & 48 \end{bmatrix}$, seen at Figure 7. The precision of the model is 0.882, indicating that 88.2% of the positive predictions were correct. The recall of the model is 0.829, suggesting that the model correctly identified 82.9% of the actual positive cases. And the area of roc curve is 0.81.

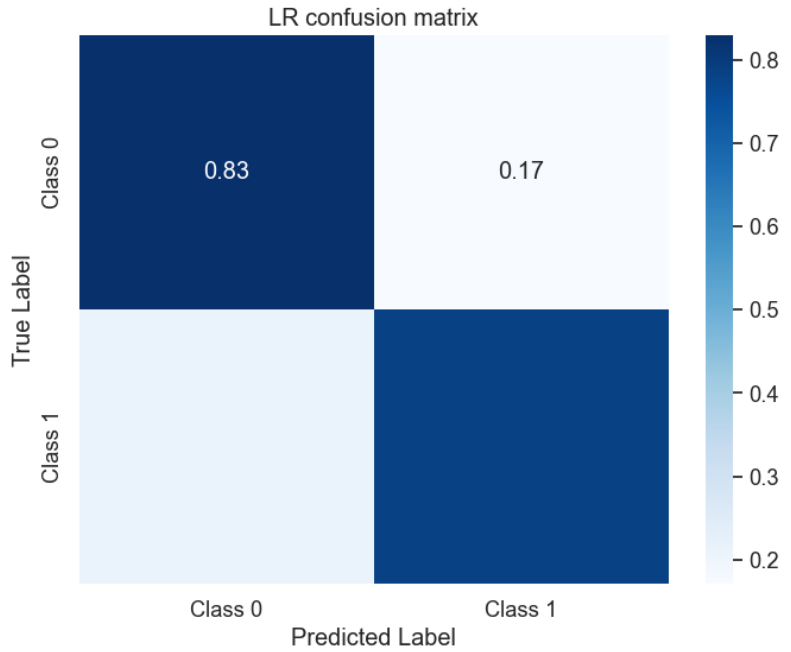

Figure 7. LR confusion matrix

3.2. DT

The author experimented with DT of different depths (5, 45, 75) and observed the following: The bar chart is showed as Figure 8. For depth 5, the cross-validation scores were 0.7924 with a precision of 0.905 and a recall of 0.812. For depth 45, the scores were 0.7724, with a precision of 0.868 and a recall of 0.846. For depth 75, the scores were identical to depth 45, with a precision of 0.868 and a recall of 0.846. The findings suggest that deepening the model beyond a certain threshold fails to markedly enhance its effectiveness.

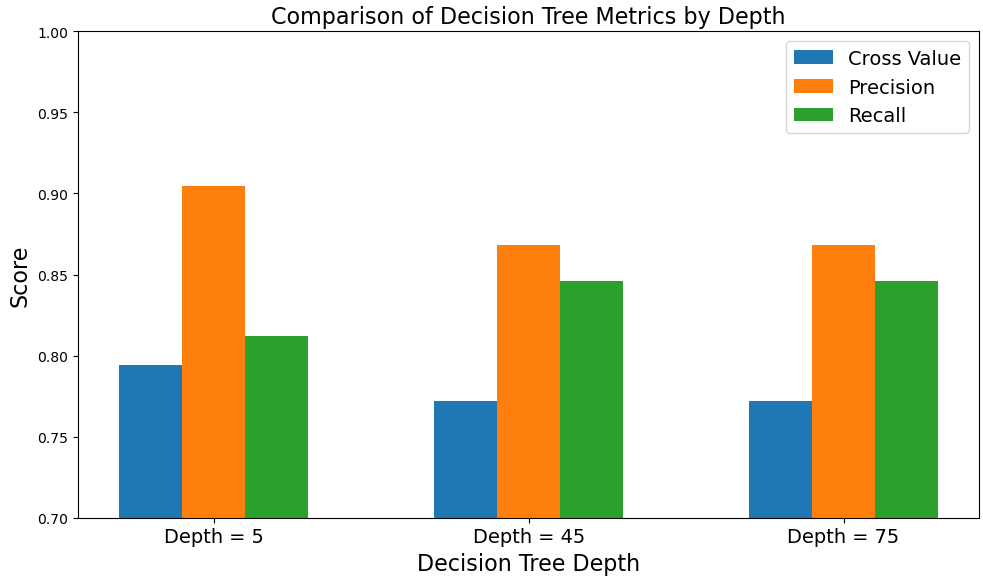

Figure 8. DT bar chart

3.3. RF

The author tested RF with different numbers of trees (5, 45, 75) and found the following: With 5 trees, the cross-validation score is [0.8133999999999999], with a precision of 0.885 and a recall of 0.923. With 45 trees, the scores were [0.798], with a precision of 0.919 and a recall of 0.872. With 75 trees, the scores were the same as with 45 trees, with a precision of 0.919 and a recall of 0.872. The RF model with 45 or 75 trees showed the best overall performance in terms of precision and recall. All of the comparison is placed in Figure 9.

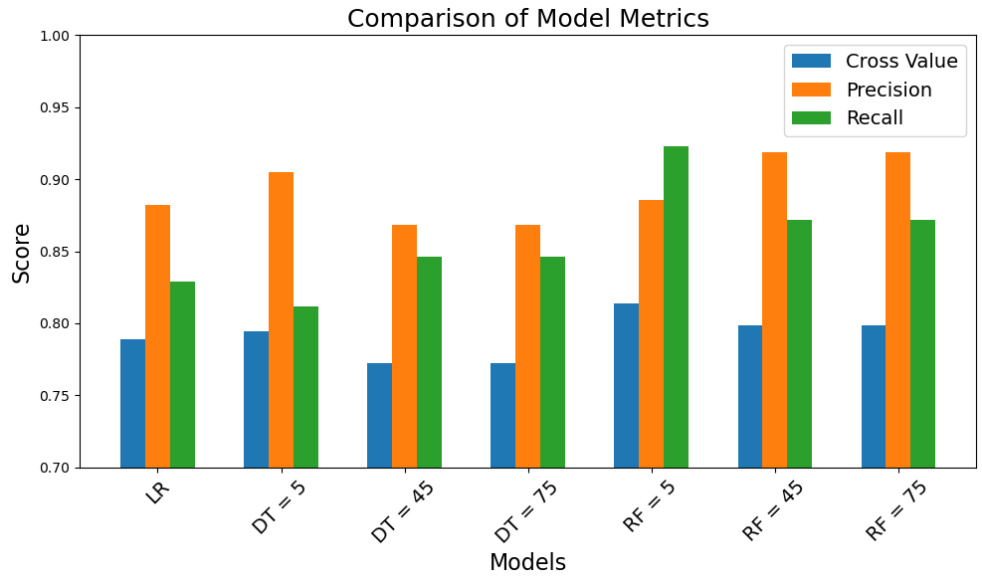

Figure 9. bar chart with all of the model

The results indicate that the RF model, particularly with 45 or 75 trees, outperforms the LR and DT in terms of precision and recall. This suggests that RF is a more robust classifier for this dataset, likely due to its ability to reduce overfitting by averaging multiple decision trees. The LR model also performed well, especially considering its simplicity compared to the other models. DT was sensitive to the depth parameter, with a moderate depth (e.g., 45) providing a good balance between precision and recall. These findings highlight the importance of model selection and parameter tuning in machine-learning tasks. Future work could explore other models and feature engineering techniques to further improve classification performance on this dataset.

4. Conclusion

This study introduces a comparative analysis to study survival conditions. Focus on the survival of passengers on the Titanic and predict survival outcomes using machine learning algorithms. The purpose is to understand the performance of different models in predicting passenger survival based on attributes such as socio-economic status, age, gender, and family relationships. This article introduces three machine learning algorithms for analysis. The Titanic dataset was carefully divided into training and testing sets to train these models and evaluate their predictive performance. This method includes metrics such as measurement accuracy, precision, and recall to evaluate the effectiveness of each algorithm in survival prediction. Experimental results show that the RF model significantly outperforms the LR and DT models in terms of precision and recall, demonstrating its superiority as a more robust classifier for this dataset. The simplicity and accuracy of the LR model are commendable, and the DT model's sensitivity to the depth parameter (with a moderate depth like 45 providing a good balance) is also a key finding. In future research, additional features will be considered, and the preprocessing phase will be refined to enhance the predictive accuracy of the models. The focus will be broadened to include advanced techniques such as ensemble methods, feature selection, and hyperparameter tuning to further improve the models' performance. This ongoing research aims to delve deeper into the complex dynamics of survival on the Titanic, striving to uncover nuanced insights for a more comprehensive understanding of this historical event.

References

[1] A. Singh, S. Saraswat, N. Faujdar. Analyzing Titanic disaster using machine learning algorithms, 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp: 406-411.

[2] A. Dasgupta,V. P. Mishra, S. Jha, et al. Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning, 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). IEEE, 2021 pp: 52-57.

[3] C. Shawn, et al. Classification of titanic passenger data and chances of surviving the disaster, Proceedings of student-faculty research day, csis, and 2014.

[4] T. Chatterjee. Prediction of survivors in titanic dataset: a comparative study using machine learning algorithms, Int J Emerg Res Manag Technol. Department of Management Studies, 2017.

[5] A. Singh, S. Saraswat and N. Faujdar, Analyzing Titanic disaster using machine learning algorithms, Computing Communication and Automation (ICCCA), 2017 International Conference, pp: 406-411.

[6] E. Lam and C. Tang, Titanic Machine Learning From Disaster, LamTang-Titanic Machine Learning From Disaster, 2012.

[7] K. Vyas, Z. Zheng, L. Li. Titanic-machine learning from disaster. Machine Learning Final Project, UMass Lowell, 2015, pp: 1-7.

[8] S. Cicoria, J. Sherlock, M. Muniswamaiah, et al. Classification of titanic passenger data and chances of surviving the disaster, Proceedings of student-faculty research day, csis, 2014, pp: 1-6.

[9] Information on: https://www.kaggle.com/competitions/titanic

[10] J. M. Bower, D. Beeman, The book of GENESIS: exploring realistic neural models with the GEneral NEural SImulation System, Springer Science & Business Media, 2012.

[11] A. Unwin, H. Hofmann, GUI and Command-line Conict or Synergy, In K Berk, M Pourahmadi, Computing Science and Statistics, 2019.

[12] G. James, D. Witten, T. Hastie, et al. An introduction to statistical learning. New York: springer, 2013.

[13] D. J. Hand, Principles of data mining, Drug safety, 30, 2007, pp: 621-622.

[14] H. Trevor, T. Robert, Friedman, the Elements of Statistical Learning, Springer, 2018, pp: 587-588.