Vol.2 No.1 February 2025: xx-xx e-ISSN: 3064-4372
International Journal of Computer Science and Humanitarian AI (IJCSHAI) DOI: 10.21512/ijcshaijournal.v2i1.12163

Comparison of Machine Learning Classification Models in Predicting The Titanic Survival Rate

Andika Elok Amalia[1], Cindy Rahayu[2*]

[1,2] Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
andika.elok@binus.ac.id; cindy.rahayu@binus.ac.id

*Correspondence: cindy.rahayu@binus.ac.id

***Abstract*— The tragic sinking of the Titanic in 1912 has been a subject of great interest, particularly in analyzing the factors that influenced passenger survival rates. This study applies machine learning techniques to predict the survival of Titanic passengers based on various attributes. The dataset used includes demographic details and passenger-specific features such as age, gender, ticket class, number of siblings/spouses, number of parents/children traveling, ticket fare, and departure location. An exploratory data analysis is conducted to understand patterns within the dataset, followed by data preprocessing steps, including handling missing values and encoding categorical variables. To develop the predictive model, multiple machine learning algorithms are implemented, including Logistic Regression, Random Forest, Extra Trees, Decision Tree, LGBM Classifier, and XGBoost Classifier. The results indicate that the Random Forest model achieves the highest accuracy at 0.815, while the LGBM Classifier attains the highest cross-validation score of 0.821. Feature importance analysis highlights gender and ticket class as the most significant factors affecting survival probability. This study demonstrates the effectiveness of machine learning classification techniques in analyzing historical data and predicting binary outcomes. The insights gained from this research can be applied to other domains involving historical data analysis and classification tasks, such as risk assessment, medical prognosis, and social science research. By leveraging machine learning, this approach provides a data-driven perspective on historical events, enabling better decision-making in similar predictive modeling scenarios.**

Keywords—machine learning, classification model, survival prediction

I. INTRODUCTION

One of the most well-known maritime catastrophes in history is the 1912 sinking of the RMS Titanic. It has long been a concern to comprehend the elements that led to the Titanic passengers' survival. Machine learning algorithms have become a potent tool for evaluating and forecasting results based on previous data in recent years. This study explores the application of the Random Forest algorithm to predict the survival probability of Titanic passengers. The primary objective is to develop a reliable predictive model that classifies passengers as survivors or non-survivors based on various factors, such as age, gender, ticket class, and other relevant attributes. To achieve our goal, we use publicly available datasets that contain information about the passengers of the Titanic, including their attributes and survival outcomes. In order to handle missing values, encode category variables, and normalize numerical features, we preprocess the data. The machine learning classifier is then trained and assessed using the processed dataset.

Classification is to attempt predicting output label based on input attribute with the highest accuracy, classification algorithm goal is to dig the relation of output attribute and then construct model based on that from its training process [1]. As other metric of measurement for our research, we will survey the result of other studies with same topic. An analytical approach has been conducted to forecast the survival rate of people on the Titanic ship and it is stated that random forest have the highest accuracy among another classifier algorithms, with accuracy of 86.29%[2]. Another study with testing 13 different machine learning algorithms found that Voting (GB, ANN,kNN) algorithm is the best, with accuracy score of 86,9% then followed by gradient boosting (GB) with the same accuracy score but has a bit lower f-measure, while random forest in this study are placed at 4th with accuracy of 84,8% [3]. While another study that use 5 algorithms found that decision tree works the best with accuracy of 93,6%[4], which are super high and surprising since other's work rarely exceed 90% accuracy threshold.

Additionally, Huang's research compared Random Forest, Gradient Boosting (GB), and XGBoost, finding that GB performed best on the original model with 83% accuracy, Random Forest achieved the highest accuracy on the tuned model with 84%, and XGBoost consistently performed the worst on both models [5]. A similar comparative study also examined performance between random forest, decision tree and Ada boost model. The random forest exhibited high accuracy and performed better prediction than decision tree, while Ada boost beat decision tree in accuracy [6]. Another research investigated the performance of logistic regression, decision tree and random forest to predict the likelihood of the Titanic passengers. The difference with the two studies

mentioned previously is in the feature engineering carried out on the dataset and result 80,4% random forest's accuracy [7] .

In all studies were conducted in predicting the Titanic survivor above, we have not found the use of the light gradient boosting machine (LGBM) classifier model even though this model provides good performance in other cases. A research that aims to classify invoice deduction observed that LGBM provided the optimum result compared to the random forest [8]. Another research compared various classifier models such as logistic regression, XGBoost, GB, decision tree, extra trees, random forest and LGBM to predict diabetes mellitus disease. It shows that LGBM had the highest accuracy with 95,20% followed by random forest and extra trees while logistic regression performed the lowest accuracy [9].

In this paper, we propose examining LGBM classifier to predict the Titanic survival rate and its performance will be compared with other machine learning algorithms commonly used for classification tasks. We choose the good performance models from previous studies such as Extra Trees, Decision Tree, Random Forest and XGB classification and also the worst one that is logistic regression, to be compared with LGBM. We evaluated the model using accuracy and cross validation score to assess its predictive ability. Our experimental results demonstrate the position of LGBM classifier in predicting survival on the Titanic. We uncover critical features that significantly impact survival, providing valuable insights into disaster dynamics. In addition, our findings contribute to a broader understanding of the factors that play a role in determining survival during marine accidents.

II. Literature Review

A. Logistic Regression

Evaluating the parameters of a calculated show (the coefficients within the straight combination) is known as calculated relapse. While the autonomous factors can each be a twofold variable (two classes, coded by a marker variable) or a ceaseless variable (any true esteem), the binary calculated relapse formally consists of a single twofold subordinate variable, coded by a pointer variable, with the two values labeled "0" and "1". The probability equation used to describe calculated relapses is condition (1) and condition (2) [10].

$$P(x) = \frac{\exp(wx+b)}{1+\exp(wx+b)} \quad (1)$$

$$P(x) = \frac{1}{1+\exp(wx+b)} \quad (2)$$

The input is $x \in R^n$, then the label vector is $y \in \{0,1\}$, w is weight and b is the offset value.

B. Decision Tree

A decision tree are popular approach used in data mining and knowledge discovery which are used to explore large and complex dataset to find pattern, it's powerful, efficient in its task [11]. Additionally, it is a supervised learning approach used for regression modeling and classification. These trees are used to either categorize data or forecast future events because regression is a technique used in predictive modeling [12].

Decision trees resemble flowcharts, with a root node containing a particular data question and branches that may contain responses. The decision (internal) nodes that follow the branches pose more queries and produce more results. This continues until the data terminates at what is known as a terminal (or "leaf") node.

C. Random Forest Algorithm

An ensemble of decision trees is constructed using a Random Forest, and each tree is trained using a randomly selected subset of characteristics and data [12]. Through a process of bagging and majority voting, the ensemble combines predictions from several decision trees to make a final prediction. This approach helps reduce overfitting and increases the generalizability of the model. Classification data that perfomed by Random Forest Gini index (3) or entropy (4) to select or decide nodes on decision tree branch.

$$Gini = 1 - \sum_{i=1}^{C}(Pi)^2 \quad (3)$$

$$Entropy = \sum_{i=1}^{C} -Pi * log_2(pi) \quad (4)$$

D. Extra Trees

Extra trees (brief for amazingly randomized trees) are a gathering administered machine learning strategy that employments choice trees and is utilized by the Prepare Utilizing AutoML instrument.

The Extra Trees algorithm, like the Random Forest algorithm, builds multiple decision trees. However, unlike Random Forest, Extra Trees selects features for each tree entirely at random and without replacement. Consequently, each tree is trained on a unique subset of the dataset with distinct testing conditions. Furthermore, a predetermined number of features are randomly chosen from the overall feature set for each tree [13]. The defining feature of the Extra Trees algorithm is its random selection of split values for features. Unlike traditional methods that determine an optimal split using Gini impurity or entropy, this algorithm assigns split values randomly. This randomness results in a more diverse set of trees with lower correlation, improving the model's robustness and reducing variance.

E. XGB Classifier

Extreme Gradient Boosting, or XGBoost, is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable and tailored for contemporary data and science tools and issues. Its advantages include being extremely scalable, fast to implement, and using regularized model formalization, which makes it superior than other algorithms [14].

F. LGBM Classifier

The free and open-source distributed gradient-boosting framework for machine learning, known as LGBM (light gradient-boosting machine), was first created by Microsoft. Based on decision tree methods, its development prioritizes scalability and performance through parallelized learning, accuracy improvement, memory reduction, and compatibility with discrete data classes [15], [16].

The Light GBM algorithm makes use of two innovative methods, Exclusive Feature Bundling (EFB) and Gradient-Based One-Side Sampling (GOSS), which enable the algorithm to operate more quickly without sacrificing accuracy. Different data instances in GOSS have distinct functions when determining information gain; an instance with higher gradients can increase the information gain [16]

III. Proposed Method

A. Data Gathering

Data is gathered using kaggle [17], where there's already existing open dataset about the titanic accident. The dataset include Passenger ID, name, survival status, class of passenger (Pclass) which are divided into 3 group, Age, SibSp which define sibling and spouse they are traveling with, parch with define the number of parent and child they have, ticket, fare, cabin number, and port of embarkment (embarked). The total count of the data is 1309 passenger. Table I describes the data attributes clearly.

TABLE I. Data Attributes

Attribute	Definition	Values
survival	Survival	0/1
pclass	Ticket class	1/2/3
sex	Sex	male/female
age	Age in years	
sibsp	Of sibling/spouse aboard	0/1
parch	Of parent/children aboard	0/1
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

B. Data Analysis

This part is done using matplotib and seaborn, both are python library which are used to make statistic graph, bar, diagram, and heat corelation map, so that the data can be analysed more efficiently and in shorter amount of time. Figure 2 shows that most of the people that die is people of the third class and Figure 3 shows that most casualty is male and most female survived.

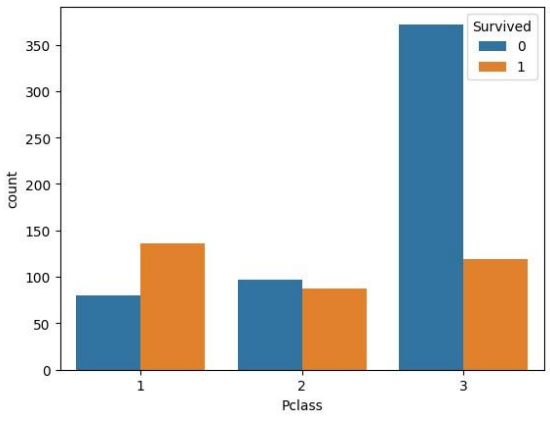

Fig. 1. Bar graph survival distribution by class

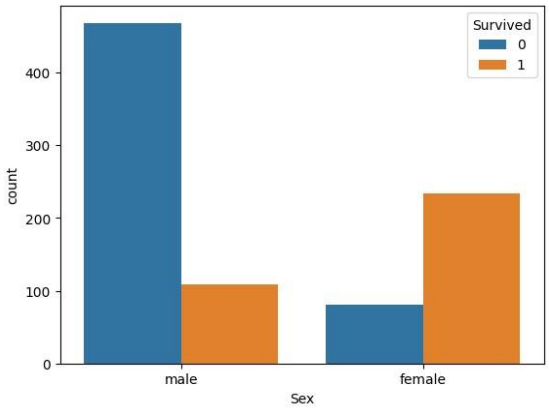

Fig. 2. Bar graph survival distribution by gender

Data also show most people that survived is people aged below 35 and people with 0 or 1 sibling, we can see it on Figure 4 and Figure 5.

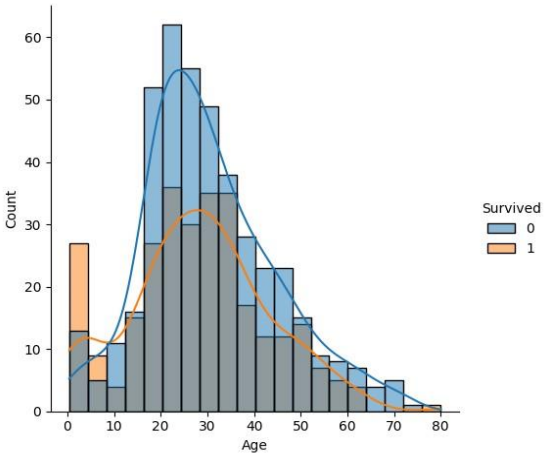

Fig. 3. Bar graph survival distribution by gender

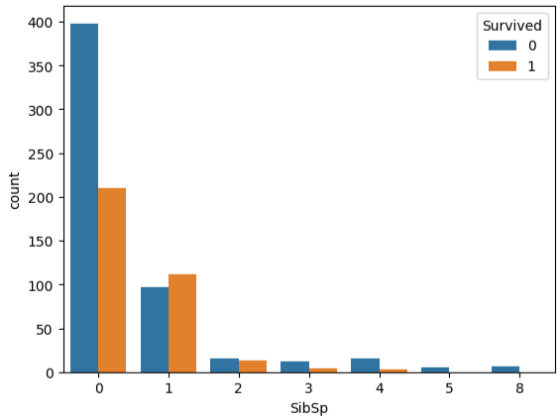

Fig. 4. Bar graph survival distribution by amount of sibling/spouse

C. Data Cleansing

After doing the analysis and determining which data table are not important for the model, we drop it, those dropped table is name, passenger id, cabin, embarked, ticket. Those attribute deemed unnecessary by the writer and thus not used in model training later on. After dropping unnecessary tables, data is cleansed more by finding the null value and filling it with the mean of the data, most of the null value comes from the Age column, and thus resolved by filling it in with the

mean of the other's passenger age data. there's also 1 other null value within the fares and resolved in the same manner, which are filled with the mean.

Since the fare data is unbalanced, the data is thus balanced by using log on it with +1 value added to every row. making the data more balanced and can be used better in the machine learning.

D. *Model Training*

After cleaning the data, we first use label encoding to turn sex data into value that can be used, data then split into output and input, then data is split into 80:20 ratio size where the 80% is the training set, and the 20% is the testing size with a random state of 15. Then the train data is fed into the model, which later can be inserted into a different kind of classification algorithm that is already explained above.

IV. Experimental Result

The result of the model training gives us the accuracy and cross validation score using *k-fold*. Table 1 describes the output of the training with logistic regression (LR), decision tree (LR), random forest (RF), XGB classifier (XGB) and LGBM classifier (LGBM). It's shown that the accuracy of the random forest is the highest, followed by logistic regression, and then decision trees and extra trees. While the highest cross validation (CV) score is LGBM classifier, followed by XGB classifier and then random forest.

TABLE II. Result of Training

Models	Accuracy	Cross Validation
Random Forest	0.815	0.814
Logistic Regression	0.787	0.783
Decision Tree	0.776	0.771
XGB	0.776	0.817
Extra Trees	0.765	0.790
LGBM	0.765	0.821

One thing that is interesting about the results is that the CV value of LGBM is the highest even though it has the lowest accuracy. This can be an indication that underfitting or overfitting may occur.

V. Conclusion

In this study, a survival prediction model is developed to predict the survival of Titanic passengers. The process begins with data collection, preprocessing, and analysis, followed by integrating the dataset into the relevant program. Several machine learning models are applied, including Logistic Regression, Decision Tree, Random Forest, Extra Trees, XGBoost, and LGBM. While that's the case, the model can be improved with some hyper parameter tuning that can be made using the data to prevent class unbalance, underfitting or overfitting . This work can still be improved upon.

Be that as it may, this paper as it were pointed at six models to anticipate, information include choice is constrained, and the exactness of the demonstrate still has to be moved forward. Within the future, more prescient models will be outlined, and information highlights will be included, information sets and distinctive information will be included, and different variables will be considered to create more precise and logical expectations.

References

[1] Jijo, B, T., and Abdulazeez, A, M. (2021) "Classification Based on Decision Tree Algorithm for Machine Learning", *JASTT*, vol. 2, no. 01, pp. 20 - 28,. https://doi.org/10.38094/jastt20165

[2] Shekhar, S., Arora, D., Sharma, P. (2021). Classifying Titanic Passenger Data and Prediction of Survival from Disaster. In: Goar, V., Kuri, M., Kumar, R., Senjyu, T. (eds) Advances in Information Communication Technology and Computing. Lecture Notes in Networks and Systems, vol 135. Springer, Singapore. https://doi.org/10.1007/978-981-15-5421-618

[3] Ekinci, Ekin & Omurca, Sevinc & Acun, Neytullah. (2018). A Comparative Study on Machine Learning Techniques Using Titanic Dataset.

https://www.researchgate.net/publication/324909545_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset

[4] Singh, K., Nagpal, R., & Sehgal, R. (2020). Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). https://doi:10.1109/confluence47617.2020.

[5] Huang, S. (2024). Processing and Comparison of GBoost, XGBoost, and Random Forest in Titanic Survival Prediction. Applied and Computational Engineering, 102(1), 175–182. https://doi.org/10.54254/2755-2721/102/20241195

[6] Huang, Y. (2024). Comparative Analysis of Models Based on Titanic Survival Predictions (pp. 146–153). https://doi.org/10.2991/978-94-6463-540-9_17

[7] Dasgupta, A., Mishra, V. P., Jha, S., Singh, B., & Shukla, V. K. (2021). Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning. Proceedings of 2nd IEEE International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2021, 52–57. https://doi.org/10.1109/ICCIKE51210.2021.9410757

[8] Tutica, L., Vineel, K., Mishra, S., Mishra, M.K., Suman, S. (2021). Invoice Deduction Classification Using LGBM Prediction Model. In: Mallick, P.K., Bhoi, A.K., Chae, GS., Kalita, K. (eds) Advances in Electronics, Communication and Computing. ETAEERE 2020. Lecture Notes in Electrical Engineering, vol 709. Springer, Singapore. https://doi.org/10.1007/978-981-15-8752-8_13

[9] Ahamed, B. S., Arya, M. S., Sangeetha, S. K. B., & Auxilia Osvin, N. v. (2022). Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers. Applied Computational Intelligence and Soft Computing, 2022. https://doi.org/10.1155/2022/7899364

[10] Al-Hadhrami, S., Al-Fassam, N., Benhidour, H. (2019). "Sentiment Analysis Of English Tweets: A Comparative Study Of Supervised And Unsupervised Approaches", . In 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyad, Suudi Arabaistan, 1-3 Mayıs.

https://doi.org/10.1016/j.procs.2021.12.187

[11] Priyanka, N. A., & Kumar, D. (2020). Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences, 12(3), 246. https://doi:10.1504/ijids.2020.108141

[12] Y. Huang, "Comparative Analysis of Models Based on Titanic Survival Predictions," in *International Conference on Image, Algorithms and Artificial Intelligence*, China, 2024.

[13] Abhishek, L. (2020). Optical Character Recognition using Ensemble of SVM, MLP and Extra Trees Classifier. 2020 International Conference for Emerging Technology (INCET) doi: https://10.1109/incet49848.2020.915405

[14] Sinha, N. K., Khulal, M., Gurung, M., & Lal, A. (2020). Developing a web based system for breast cancer prediction using xgboost classifier. *International Journal of Engineering Research Technology (IJERT)*, 9(6), 852-856.

[15] Wang, Y., Liu, Y., Zhao, J., & Zhang, Q. (2023). Low-Complexity Fast CU Classification Decision Method Based on LGBM Classifier. *Electronics*, *12*(11), 2488. MDPI AG. Retrieved from http://dx.doi.org/10.3390/electronics12112488

[16] Osman, M., He, J., Mokbal, F. M. M., Zhu, N., & Qureshi, S. (2021). ML-LGBM: A Machine Learning Model Based on Light Gradient Boosting Machine for the Detection of Version Number Attacks in

RPL-Based Networks. IEEE Access, 9, 83654–83665. https://doi.org/10.1109/ACCESS.2021.3087175

[17] Will Cukierski. Titanic - Machine Learning from Disaster. https://kaggle.com/competitions/titanic, 2012. Kaggle.