Comparative study of algorithms for Term Deposit Prediction

Abstract— **This study examines the use of direct marketing in the finance domain and focuses on a typical bank direct marketing campaign dataset. This type of marketing helps financial institutions such as banks to target customers who could be interested in their products, offers, and packages. This study's primary goal is to forecast how bank direct marketing will be received by consumers utilizing a variety of algorithms, such as Gaussian Naive Bayes, Decision Tree, Logistic Regression, SVM, Random Forest, and a few more to develop a prediction model using the algorithm that achieved the highest accuracy (XGBoost). The second objective is to provide helpful analysis and visualization to the marketing team for strategizing and decision-making situations.**

Keywords— Bank telemarketing, Ensemble learning, Max voting, Bagging, Boosting, Customer targeting, Customer Segmentation.

I. INTRODUCTION

Marketing through telephones, a type of direct marketing, is typically achieved through several ways. Banks and financial institutions first identify the specific segment of customers they want to reach through their telemarketing campaign. This may involve analysing customer data and demographics, transaction history, previous marketing interactions, and other relevant information to determine the most promising customer segment for their products or services. services. Once the target customer segment is identified, banks create a database of potential customers to be contacted through telemarketing. Contact information for clients, including names, phone numbers, and additional relevant Facts, are maintained in this database. Then the Telemarketing agents, who are usually trained and employed by the bank or a third-party telemarketing agency, are provided with the necessary training on the bank's products, services, and telemarketing techniques. Telemarketing agents then initiate contact with potential customers from the telemarketing database through outbound calls. They introduce the bank's products or services, explain their benefits, and try to persuade customers to subscribe or show interest. The outcome of each call is documented and is stored, which forms the dataset, including whether the customer expressed interest, subscribed to the product or service, requested more information, or declined the offer.

- To aid in this process, banks and other financial institutions can utilise our prediction model to forecast the client responses.
- Financial organisations can optimise their marketing strategy, target the most promising clients, and boost the success of their direct marketing operations by harnessing the predictive capabilities of Ensemble methods.
- This can result in improved customer engagement, higher conversion rates, and better overall marketing performance.

II. RELATED WORK

The paper by S. Moro addresses a similar issue. The article is “A data-driven approach to predict the success of bank telemarketing”. The authors use the same dataset as we use in this research work of 5-year period, between 2008 and 2013, of the Portuguese retail bank. They implement four classification learning models for identifying the clients with increased probability of opening a long-term deposit: Logistic Regression, Decision Trees, Neural Network and Support Vector Machine. The best result is obtained by Neural Network. A telecommunication industry problem based on customer churn prediction is solved by T. Vafeiadis. Authors compare most popular classification methods. It includes Logistic Regression, Naive Bayes, Decision Trees learning, Support Vector Machine, and Artificial Neural Network. These models are compared by precision, recall, accuracy, and F-measure criteria. The article "Islamic versus conventional banks in the GCC countries" compares classification models in a similar way. The paper aims to predict suitability of financial ratios for distinguishing between conventional and Islamic banks. The paper again uses Neural Network, Classification tree, Logistic Regression, and Linear discriminant analysis. Authors consider the period between 2003 and 2010 years.

III. DATASET OVERVIEW

The dataset provides detailed information on the telemarketing initiatives of a Portuguese financial institution, for promoting the bank's term deposit product.

It includes demographic and transactional data for clients, including age, job type, marital status, education level, credit history, average yearly balance, housing and personal loan status, contact communication type, contact day and month, number of contacts performed during previous and current campaigns, number of days since the client was last contacted, and the outcome of the previous marketing campaign.

The target variable is whether or not the client subscribed to the bank term deposit product ('yes' or 'no'). It is important to note that the 'duration' attribute significantly affects the target variable, as a duration of 0 implies that the client did not subscribe. However, this attribute is only known after a call is made and therefore should only be used for benchmarking purposes.

IV. EXPLORATORY DATA ANALYSIS

Heatmaps can be used to visually represent the correlation between variables in a dataset. For example, a heatmap can show the pairwise correlation coefficients between different variables, with the colour intensity representing the strength and direction of the correlation. This can help identify patterns and relationships between variables, and can be particularly useful in exploring complex datasets with multiple variables

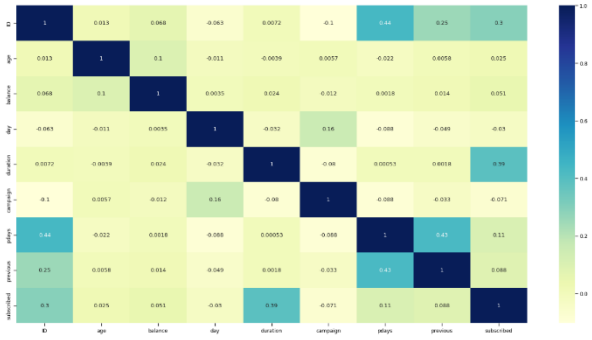

Fig 1. Heatmap of features

A. *Univariate Analysis*

Heatmaps can be used to visually represent the correlation between variables in a dataset. For example, a heatmap can show the pairwise correlation coefficients between different variables, with the colour intensity representing the strength and direction of the correlation. This can help identify patterns and relationships between variables, and can be particularly useful in exploring complex datasets with multiple variables.

In our case Univariate Analysis is used to analyse and understand the characteristics and distribution of a single variable that may impact the prediction of term deposit subscription.

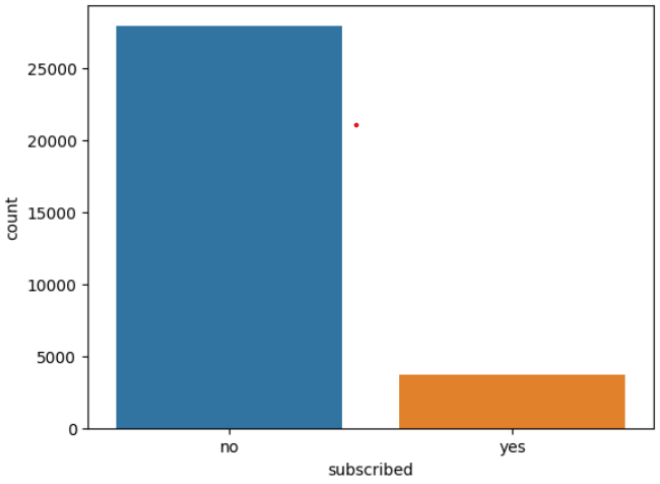

Fig 2. Count plot of Subscription status

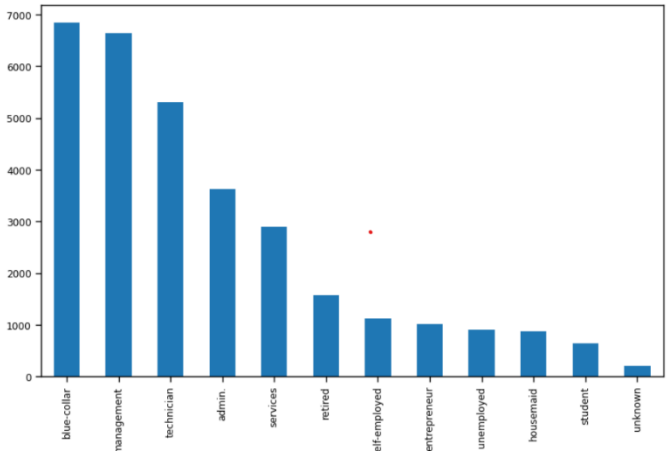

Fig 3. Count plot of Employment status

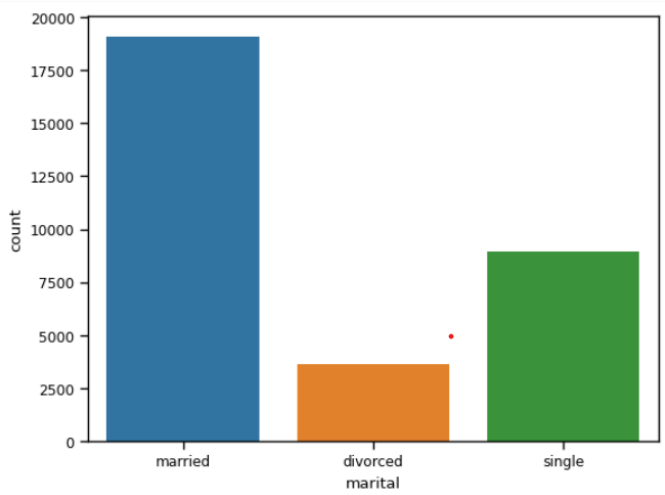

Fig 4. Count plot of Marital status

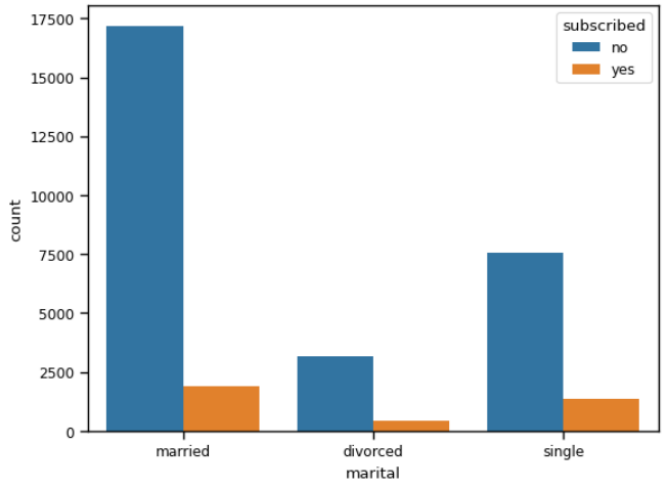

Fig 5. Count plot of Subscription and Marital status

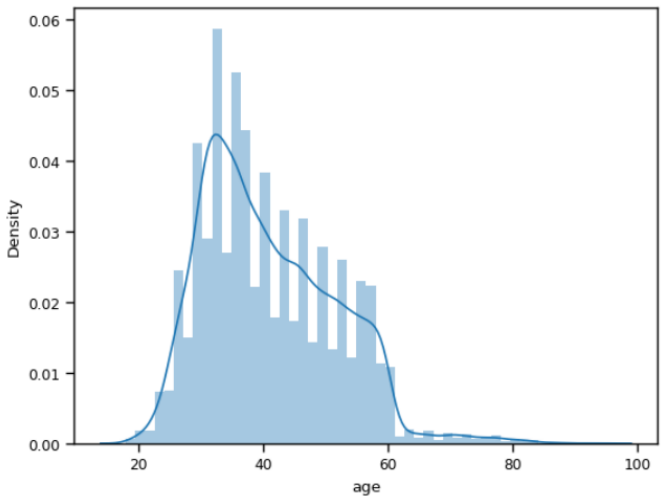

Fig 6. Distribution plot of age

B. *Bivariate Analysis*

Bivariate Analysis is a statistical method that involves the analysis of two variables simultaneously to understand the relationships or associations between them. It focuses on examining the joint behaviour, patterns, and interactions between two variables in a data set. Bivariate analysis can provide insights into how two variables are related or how they change together.

Here, Bivariate analysis is used to analyse and understand the relationships or associations between two variables that may impact the prediction of term deposit subscription.

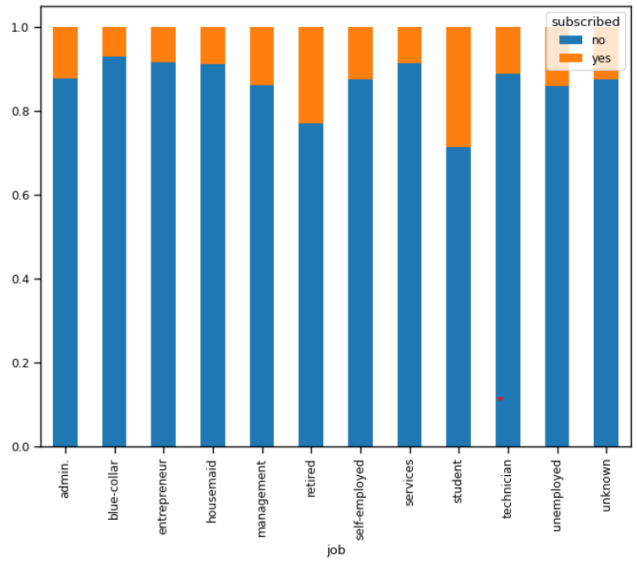

Fig 7. Bar graph of Job and Subscription status

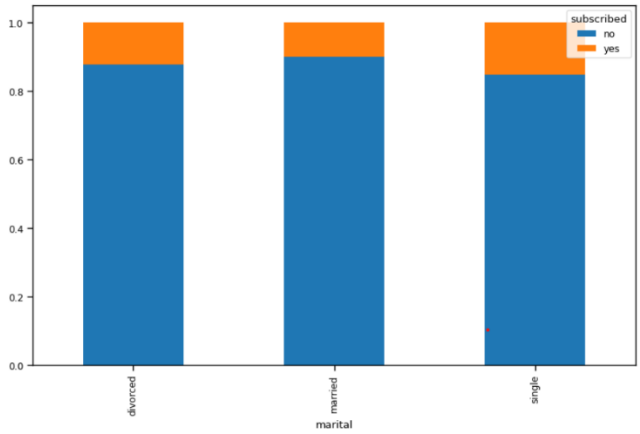

Fig 8. Bar graph of Job and Marital status

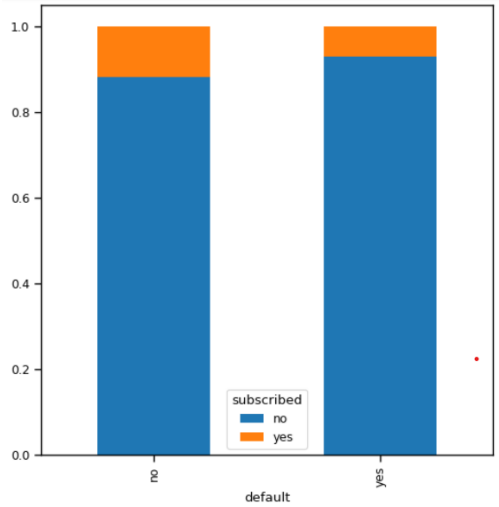

Fig 9. Bar graph of Previous Customers and Subscription status

V. PROPOSED SYSTEM

Our proposed system for predicting whether a customer will take a term deposit in-volves the analysis of a dataset of customer information, including demographic data, transaction history, and other relevant factors. By analysing this data, our system will be able to identify patterns and make predictions on which customers are most likely to take a term deposit. The insights gained from our system will help banks to better target their marketing efforts towards the customers who are most likely to take a term deposit. This will lead to more effective marketing campaigns and higher conversion rates.

A. *Machine Learning Algorithms*

In this project, we utilized several machine learning algorithms to classify and analyse data. They are –

- Gaussian Naive Bayes: is a fast probabilistic algorithm that calculates the probability of a sample belonging to a particular class based on the features' values. It works well with high-dimensional data and has been successful in text classification and spam filtering.

- Support Vector Classifier (SVC): is a binary classification algorithm that optimizes the boundary between two classes by maximizing the margin between them. It handles linearly separable data and non-linear data using kernel methods. SVC has been successful in a variety of applications, including bioinformatics, image recognition, and text classification.
- Logistic Regression: is a simple statistical algorithm that predicts the probability of a sample belonging to a specific class based on its feature values. It works well for binary and multi-class classification problems.
- K-Nearest Neighbour (KNN): is a non-parametric algorithm that classifies a sample based on its k-nearest neighbours in the feature space. It is a simple algorithm that works with linearly and non-linearly separable data and can be used for classification and regression problems.
- Decision Tree: is a hierarchical algorithm that recursively divides the data based on the most informative feature until a stopping criterion is met. It is interpretable and works well with both numerical and categorical data. Decision trees have been used in finance, marketing, and medicine.
- Random Forest: is an ensemble algorithm that combines multiple decision trees to increase the classification accuracy and decrease overfitting.
- XGBoost (Extreme Gradient Boosting): is a powerful machine learning algorithm used for supervised learning problems, including classification and regression. It is based on the gradient boosting framework and uses decision trees as base models.

Using one of the many ensemble techniques called Max voting, we develop a better model by in a way combining the outputs of the other trained models to provide a result that would in general be the most probable output of the situation. We first train different models with the training data, and evaluate them. Using these models we make a Voting classifier that will calculate the mode of all the outputs of these models and use that as its result. Since our case consists of a binary classification we are using mode for the ensemble technique. However for numerical predictions we can also use other ensemble methods such as mean or weighted average.

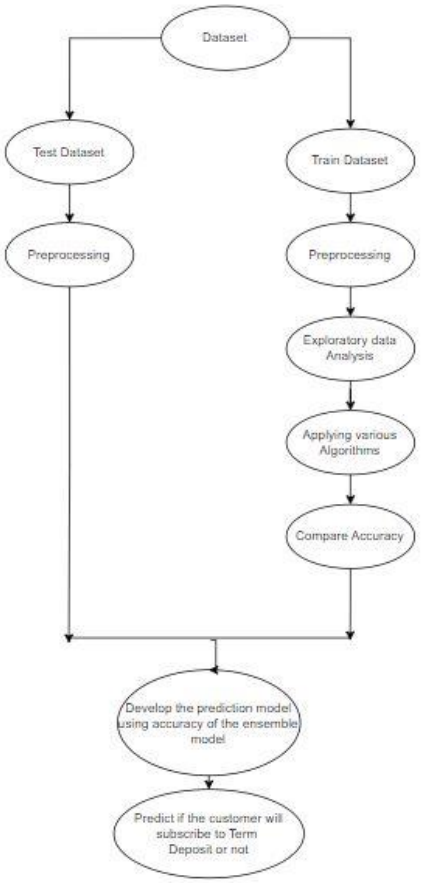

Fig 10. Flow diagram of proposed system

VI. EVALUATION

To determine the optimal machine learning model, various performance measures are evaluated during the execution of the models. The evaluation technique used to assess the performance of our model is accuracy, as indicated by the confusion metrics.

Accuracy: It represents how many instances are correctly predicted.

Accuracy = (True Positive Values + True Negative Values) / (True Positive + False Negative + True Negative + False Positive)

Based on the accuracy scores of the different machine learning models, the Gaussian Naive Bayes model had the lowest accuracy score of 84.78%, indicating underfitting. The Logistic Regression model performed better with an accuracy score of 89.51%. The Decision Tree model had an accuracy score of 89.92%, indicating that it was better at capturing more of the underlying patterns in the data. The Random Forest Classifier model had the highest accuracy score of 90.15%, suggesting that it may be slightly better at capturing the underlying patterns. The K-Nearest Neighbour and Support Vector Classifier models had an accuracy score slightly higher than Gaussian Naive Bayes.

However, these are the accuracies of individual models and we are using an ensemble model for the prediction. The Voting classifier model had the accuracy of 89.95%. This accuracy is lesser than some of the other individual models, but that is expected because it is a kind of mean accuracy of all the models.

VII. RESULTS

Random Forest Classifier is an ensemble method that combines multiple decision trees to make predictions, which can result in higher accuracy compared to individual classifiers like KNN, SVM, Naive Bayes, or Decision Tree. The accuracy of a Random Forest Classifier is higher than the other classifiers in the term deposit prediction system, indicating better prediction performance. The Voting Classifier also turned out to be good with 89.95% accuracy.

Overall, a term deposit prediction system model built and trained using the Random Forest model or the Voting Classifier model can lead to improved prediction performance, enhanced customer targeting, informed decision-making, reliable performance metrics, and robustness to data challenges, making it a valuable tool for financial institutions in optimizing their term deposit marketing campaigns. However, it's important to thoroughly validate and interpret the model's results in the context of the specific dataset and business requirements.

TABLE I. COMPARISON BETWEEN ALGORITHMS

Algorithm	Accuracy
Gaussian Naive Bayes	84.77
K-Nearest Neighbour	88.34
Support Vector Classifier	88.58
Logistic Regression	89.51
Decision Tree	89.92
Random Forest	90.16
XGBoost	89.89

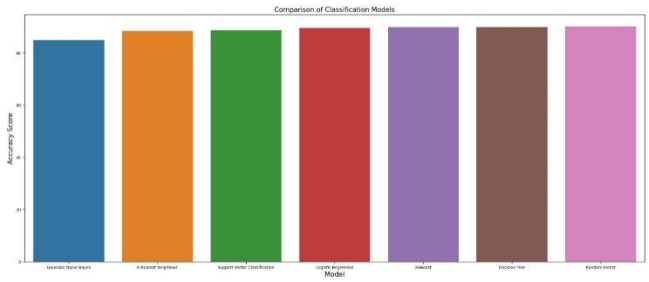

Fig 11. Model accuracies

VIII. CONCLUSION AND FUTURE WORK

This study aimed to compare the performance of six classifiers, including Gaussian Naive Bayes, K-Nearest Neighbour, Support Vector, Logistic Regression, Decision Tree, and Random Forest, to predict the likelihood of customers subscribing to term deposits. The results showed that Random Forest and Decision Tree performed the best in terms of prediction accuracy. The study also analysed key features that influenced a customer's decision to subscribe, including their occupation, age, and call duration. The findings can be used to improve the effectiveness of direct and telemarketing campaigns.

While our project has yielded valuable insights into term deposit prediction using machine learning algorithms, there are several avenues for future work that could be pursued. One possible direction would be to incorporate additional data sources, such as customer transaction histories or social media activity, to enhance the predictive power of our models. Additionally, we could explore the use of alternative machine learning algorithms, such as neural networks. Future research can focus on incorporating additional data sources, such as customer transaction histories and social media activity, to enhance the predictive power of models. Continued research in this domain using machine learning can help financial institutions make more informed decisions and better serve their customers.

REFERENCES

[1] Sérgio Moro and Raul M. S. Laureano, Paulo Cortez, “Using Data Mining for Bank Direct Marketing”, October 2011, European Simulation and Modelling Conference.

[2] E. Zeinulla, K. Bekbayeva and A. Yazici, "Comparative study of the classification models for prediction of bank telemarketing," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018, pp. 1-5, doi: 10.1109/ICAICT.2018.8747086..

[3] M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam and D. S. A. A. Reza, "Identifying Long-Term Deposit Customers: A Machine Learning Approach," 2021 2nd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/IISEC54230.2021.9672452.

[4] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), Kuala Lumpur, Malaysia, 2017, pp. 1-4, doi: 10.1109/ICE2T.2017.8215961.

[5] S. Moro, P. Cortez, and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing," Decision Support Systems, vol. 62, pp. 22-31, Jul. 2014. doi: 10.1016/j.dss.2014.03.001.

[6] T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1-9, 2015, doi: 10.1016/j.simpat.2015.03.003.

[7] K. B. Khediri, L. Charfeddine, and S. B. Youssef, "Islamic versus conventional banks in the GCC countries: A comparative study using classification techniques," Res. Int. Bus. Finance, vol. 33, pp. 75-98, Jan. 2015. DOI: 10.1016/j.ribaf.2014.07.002.

[8] K. Morani, E. K. Ayana and Ş. N. Engin, "Developement of Prediction in Clients’ Consent to a Bank Term Deposit Using Feature Selection," 2018 6th International Conference on Control Engineering & Information Technology (CEIT), Istanbul, Turkey, 2018, pp. 1-5, doi: 10.1109/CEIT.2018.8751816.

[9] Z. Yang, Y. Yang, D. Yin, M. Yang and L. Li, "Research on Data Analysis for Time Deposit of Bank Customers Based on Ensemble Learning," 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, 2019, pp. 1325-1330, doi: 10.1109/EITCE47263.2019.9094858.

[10] F. Huang, G. Xie and R. Xiao, "Research on Ensemble Learning," 2009 International Conference on Artificial Intelligence and Computational Intelligence, Shanghai, China, 2009, pp. 249-252, doi: 10.1109/AICI.2009.235.

[11] M. Chen, Q. Liu, S. Chen, Y. Liu, C. -H. Zhang and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," in IEEE Access, vol. 7, pp. 13149-13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

[12] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2017, pp. 65-68, doi: 10.1109/WCCCT.2016.25.

[13] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[14] P. Meel, P. Chawla, S. Jain and U. Rai, "Web Text Content Credibility Analysis using Max Voting and Stacking Ensemble Classifiers," 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Cochin, India, 2020, pp. 157-161, doi: 10.1109/ACCTHPA49271.2020.9213234.

[15] J. Suriya Prakash, K. Annamalai Vignesh, C. Ashok and R. Adithyan, "Multi class Support Vector Machines classifier for machine vision application," 2012 International Conference on Machine Vision and Image Processing (MVIP), Coimbatore, India, 2012, pp. 197-199, doi: 10.1109/MVIP.2012.6428794.

[16] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[17] T. Haifley, "Linear logistic regression: an introduction," IEEE International Integrated Reliability Workshop Final Report, 2002., Lake Tahoe, CA, USA, 2002, pp. 184-187, doi: 10.1109/IRWS.2002.1194264.