iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.70090

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Enhancing Salary Predictions with Ensemble Learning Techniques

V Ashok Kumar [1], Majji Swathi[2], Neyyala Mahesh[3], Sasumahanthi Mahesh[4], Adavikolanu Ravishankar[5]
Department of ECE, Aditya Institute of Technology and Management

Abstract: ***Accurate salary prediction is essential for informed decision-making in various industries. This study applies ensemble learning techniques using five models—Decision Tree, Logistic Regression, XGBoost, LightGBM, and Random Forest—to predict salary levels through binary classification. By combining these models, we enhance prediction accuracy, stability, and reliability. The models are evaluated using accuracy, precision, recall, and F1-score. Results show that ensemble learning significantly improves prediction performance, offering a more reliable approach to salary forecasting. This method provides valuable insights for practical applications in salary prediction and human resource management. Additionally, the findings suggest that the ensemble approach can be applied to other classification tasks beyond salary prediction. The study also highlights the importance of model selection and evaluation metrics in optimizing performance. Overall, the research contributes to the growing field of machine learning applications in human resources and compensation management.***
Keywords: Ensemble Learning, Linear Regression, Logistic Regression, Random Forest, Salary Prediction, XGBoost

I. INTRODUCTION

Accurate salary prediction is a vital component in fostering transparency, equity, and efficiency within the labour market. With the increasing availability of workforce data and advancements in machine learning, researchers have explored intelligent systems to predict salaries based on various influencing factors. Babasaheb S. Satpute et al. pioneered a model that utilizes socio-demographic attributes such as age, gender, education, race, and country, along with experience, to predict employee salaries. Their findings uncovered significant income disparities across demographic groups, highlighting the need for data-driven strategies to promote income equality.[1] Building upon the pursuit of fairness and accuracy, Rukiye Kaya et al. evaluated multiple classification algorithms and identified the Majority Voting Classifier as the most effective, underscoring the value of ensemble methods in enhancing prediction consistency.[2] In a similar vein, Habibu Aminu et al. integrated Principal Component Analysis (PCA) for feature selection with a Deep Neural Network (DNN), achieving superior performance compared to traditional models like Decision Trees and Random Forests, and demonstrating the power of deep learning in salary classification tasks.[3] These foundational studies collectively highlight the multidimensional nature of salary determination. They also underline the growing relevance of predictive modelling in human resource analytics. As the demand for fair and evidence-based compensation structures increases, machine learning offers promising tools to support data-driven decision-making. Moreover, the integration of ensemble techniques and deep learning has significantly advanced prediction accuracy and model robustness. With continuous innovations in feature selection, model tuning, and algorithmic design, the landscape of salary prediction is becoming increasingly intelligent and equitable. These developments pave the way for automated systems that not only forecast earnings but also help reduce systemic bias in compensation frameworks.[4]

II. LITERATURE REVIEW

I Jawad Hussain developed a regression model for salary prediction using machine learning to enhance fairness and accuracy. Data processing techniques such as exploratory data analysis (EDA), feature correlation, and engineering were implemented to refine model inputs. The study deployed Random Forest, Gradient Boosting, and Light Gradient Boosting Regressor models, achieving an impressive 99% accuracy. This underscored the potential of machine learning to mitigate bias and enhance transparency in salary estimation. The findings emphasized the value of AI-driven approaches in ensuring equitable salary outcomes and promoting accountability in compensation frameworks. [5]Yasser Matbouli and Suliman M. Alghamdi introduced a salary prediction framework that leveraged statistical machine learning models on Saudi labour market data. Five supervised algorithms were tested, with Bayesian Gaussian Process Regression outperforming traditional linear regression methods for predicting salaries by economic activity. Artificial Neural Networks yielded the highest accuracy for occupational group predictions, reinforcing the capability of deep learning in handling complex classification problems.

Their research highlighted the superior accuracy, adaptability, and reliability of advanced ML models over conventional techniques in wage forecasting. [6]Manisha Joshi et al. proposed a cutting-edge enhancement to neural networks by developing a fractional-order derivative-based backpropagation algorithm. They evaluated variants like Riemann-Liouville, Caputo, and Caputo-Fabrizio derivatives against classical integer-order methods. Incorporating momentum into a three-layer feed-forward neural network enabled greater learning stability and faster convergence. Their approach significantly improved generalization in salary predictions, particularly in experience-based models, demonstrating the promise of fractional calculus in neural network optimization. [7]Ziyuan Feng et al. applied Convolutional Neural Networks (CNNs) to a structured dataset from Kaggle by reshaping tabular data into two-dimensional matrices, thereby allowing CNNs to capture spatial and hierarchical patterns within the input. The CNN architecture achieved a lower error rate and variance compared to Random Forest models, signifying high robustness and prediction confidence. Their study validated CNN's suitability for salary prediction tasks, especially in scenarios where data exhibits complex interdependencies. [8]Sayan Das et al. focused on long-term salary forecasting by designing a data-driven system that analysed historical organizational salary data to project future trends. The system featured a visual component for trend analysis and salary growth prediction, making it a practical tool for HR analytics and strategic workforce planning. The model enabled organizations and employees to make informed decisions based on anticipated salary progressions, illustrating the relevance of predictive analytics in financial planning. [9]Phuwadol Viroonluecha and Thongchai Kaewkiriya leveraged Deep Learning models on a large-scale job portal dataset comprising over 1.7 million user records in Thailand. Their framework integrated sophisticated feature selection techniques to improve model accuracy and runtime performance. Among various models, Deep Learning delivered the best results, achieving an R^2 score of 0.462 while maintaining computational efficiency. The study highlighted the real-world applicability of Deep Learning in employment analytics, particularly when combined with robust preprocessing pipelines. [10]Lastly, Pornthep Khongchai and Pokpong Songmuang developed a salary prediction system targeting student career guidance. Using a dataset of 13,541 student records, they applied machine learning algorithms—Decision Trees, Naive Bayes, K-Nearest Neighbour, SVMs, and Neural Networks—evaluated using 10-fold cross-validation. A post-study survey with 50 users confirmed the system's usability and value. The model empowered students to explore realistic salary expectations based on their academic profiles, demonstrating its potential to assist in career planning, motivation, and goal-setting. [11]

III.METHODOLOGY

The figure-1 shows the flowchart illustrates the end-to-end process of building a machine learning model to predict salaries using ensemble techniques. It begins with data collection, where relevant salary-related features are gathered. This is followed by data preprocessing, which includes cleaning, transforming, and preparing the data for modelling. Once the data is ready, ensemble techniques such as Random Forest or Gradient Boosting are applied to build robust predictive models. The next step involves evaluating the model's performance using appropriate metrics to ensure accuracy and reliability. Finally, the trained model is used to predict salaries on new, unseen data. This structured approach ensures a systematic and efficient pipeline from data acquisition to deployment.

Fig.1. Flow Chart for Enhancing Salary Prediction

1) Dataset Description

The dataset comprises diverse employee information, including demographics (age, gender, location, education) and professional details (years of experience, job role, industry, previous salary). The salary, originally continuous, is re-categorized into three classes—low, medium, and high—based on predefined thresholds. This transformation into a multi-class classification problem enhances its applicability in hiring platforms and workforce analytics, while the variety of occupations ensures model robustness and generalizability.

2) Data Preprocessing

To ensure the dataset was clean and suitable for modelling, several preprocessing techniques were applied. Missing values were handled using imputation, with the mean used for numerical attributes and the mode for categorical ones. Categorical variables were encoded using Label Encoding and One-Hot Encoding, depending on the compatibility with specific algorithms. To standardize the feature space, Min-Max normalization was performed, ensuring that all numerical features were scaled within the same range. Additionally, outliers were detected and addressed using Z-score analysis and the Interquartile Range (IQR) method, improving the quality and consistency of the input data. Data distribution visualizations and box plots were used to verify normalization effectiveness and detect anomalies.

3) Model Implementation

The models implemented in this study include both traditional regression and advanced ensemble methods. The primary models are:

- Decision Tree: The baseline model, which predicts salary as a decision tree function of the input features:

$$y^{\wedge} - \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (1)$$

where $y^{\wedge}$ is the predicted salary, β_0 is the intercept, and β_i are the coefficients of the feature x_i.

- Logistic Regression: This model is used for predicting salary categories by estimating the probability of each class using the sigmoid function:

$$p(y = k|x) = \frac{1}{1+e^{-z}}, z = \beta_0^k + \sum \beta_i^k x_i \quad (2)$$

where $p(y = k|x)$ represents the probability of the salary belonging to category k, and z is the linear combination of features for class k.

- Random Forest: An ensemble method consisting of multiple decision trees. Each tree is trained on a bootstrapped sample of the data, and the final prediction is obtained by averaging or voting across the trees:

$$y^{\wedge} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (3)$$

where T is the number of trees and $h_t(x)$ is the prediction from the t-th tree.

- XGBoost: A gradient boosting algorithm that sequentially builds decision trees, with each tree correcting the errors of previous trees. The objective function is minimized by adding a regularization term to prevent overfitting:

$$L = \sum_{i=1}^{n} l(y_i, y_i^{\wedge}) + \sum_{k=1}^{k} \Omega(f_k) \quad (4)$$

where l is the loss function, y_i is the true value, $y_i^{\wedge}$ is the predicted value, and $\Omega(f_k)$ is the regularization term for the k-th tree.

- LightGBM: Another gradient boosting framework, optimized for efficiency. It builds trees leaf-wise and uses histogram-based splitting to reduce computation time and memory usage. The loss function is similar to XGBoost but adapted for speed:

$$L = \sum_{i=1}^{n} l(y_i, y_i^{\wedge}) + \lambda ||\omega||^2 \quad (5)$$

where $\lambda ||\omega||^2$ is the L2 regularization term.

4) Model Evaluation

Given that the salary is categorized into three levels, the models are evaluated using classification metrics rather than regression metrics. The evaluation metrics are from equation (6) to (8) as shown:

- Accuracy: Measures the overall proportion of correct predictions:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

- Precision: Reflects the exactness of the positive predictions:

$$\text{Precision}=\frac{TP}{TP+FP} \quad (7)$$

- Recall: Measures the model's ability to detect actual positives:

$$\text{Recall}=\frac{TP}{TP+FN} \quad (8)$$

- F1-Score: Provides a balance between Precision and Recall:

$$\mathbf{F1}=2\times\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \quad (9)$$

Cross-validation is used to validate the models' performance, ensuring that the results are not biased by any particular data split.

IV. RESULTS AND DISCUSSIONS

Ensemble learning techniques significantly enhance salary prediction accuracy by combining the strengths of multiple models to reduce bias and variance. By leveraging models such as Decision Trees, Logistic Regression, Random Forest, XGBoost, and LightGBM, the ensemble approach aggregates individual predictions to produce more robust and reliable results. Random Forest, for example, minimizes overfitting by averaging predictions from multiple decision trees, while XGBoost and LightGBM further refine accuracy through gradient boosting, optimizing model performance iteratively. Logistic Regression contributes by modelling linear relationships and aiding in feature selection. Evaluation metrics such as accuracy, precision, recall, and F1-score demonstrate the effectiveness of these techniques in capturing complex patterns in salary data, providing more accurate and generalizable predictions. The use of ensemble methods in salary prediction offers a significant improvement in handling various data complexities, ensuring better performance than individual models alone.

Table-1: Evaluation metrics comparison for ensemble learning techniques

Model	Accuracy	Precision	Recall	F1 Score
Logistic regression	71.62%	71.44%	7056%	71.00%
Decision tree	80.36%	78.77%	82.28%	80.49%
RF	80.90%	79.84%	81.87%	80.84%
XGBoost	83.12%	81.53%	84.95%	83.21%

The table-1 shows comparing the performance of the models, XGBoost stands out with the highest accuracy (83.12%), precision (81.53%), recall (84.95%), and F1 score (83.21%). This model shows the best balance in identifying positive cases and minimizing false positives, making it the most effective for the task. Random Forest follows closely with an accuracy of 80.90%, precision of 79.84%, recall of 81.87%, and an F1 score of 80.84%. It performs very well across all metrics but slightly lags behind XGBoost. Decision Tree provides strong performance with accuracy of 80.36%, precision of 78.77%, and recall of 82.28%, offering a solid middle ground, although it is less powerful than Random Forest and XGBoost. Finally, Logistic Regression lags significantly with accuracy of 71.62%, precision of 71.44%, recall of 70.56%, and F1 score of 71.00%, performing the least well among all models. In conclusion, XGBoost is the top performer, while Logistic Regression shows the weakest results.

The figure-10 shows comparing the performance of Logistic Regression, Decision Tree, Random Forest, and XGBoost models based on their confusion matrices, XGBoost clearly emerges as the most effective. Logistic Regression, being a linear model, performs the weakest among the four, with an accuracy of approximately 71.3%, and relatively lower precision and recall values. This indicates that it struggles to capture the non-linear relationships present in the data. The Decision Tree model shows a notable improvement, achieving around 80.4% accuracy with better precision and recall, as it can handle more complex patterns by splitting the data hierarchically. Random Forest, which is an ensemble of multiple decision trees, further enhances performance by reducing overfitting and improving generalization. It achieves an accuracy of about 80.9%, along with balanced precision and recall scores, making it a strong candidate for binary classification. However, XGBoost surpasses all the others by leveraging gradient boosting, which focuses on correcting previous errors in a sequential manner. As a result, it delivers the highest accuracy of around 82.3%, along with the best precision and recall scores, indicating its robustness and efficiency in handling classification tasks. Overall, while tree-based models significantly outperform Logistic Regression, XGBoost stands out as the best-performing model in this comparison.

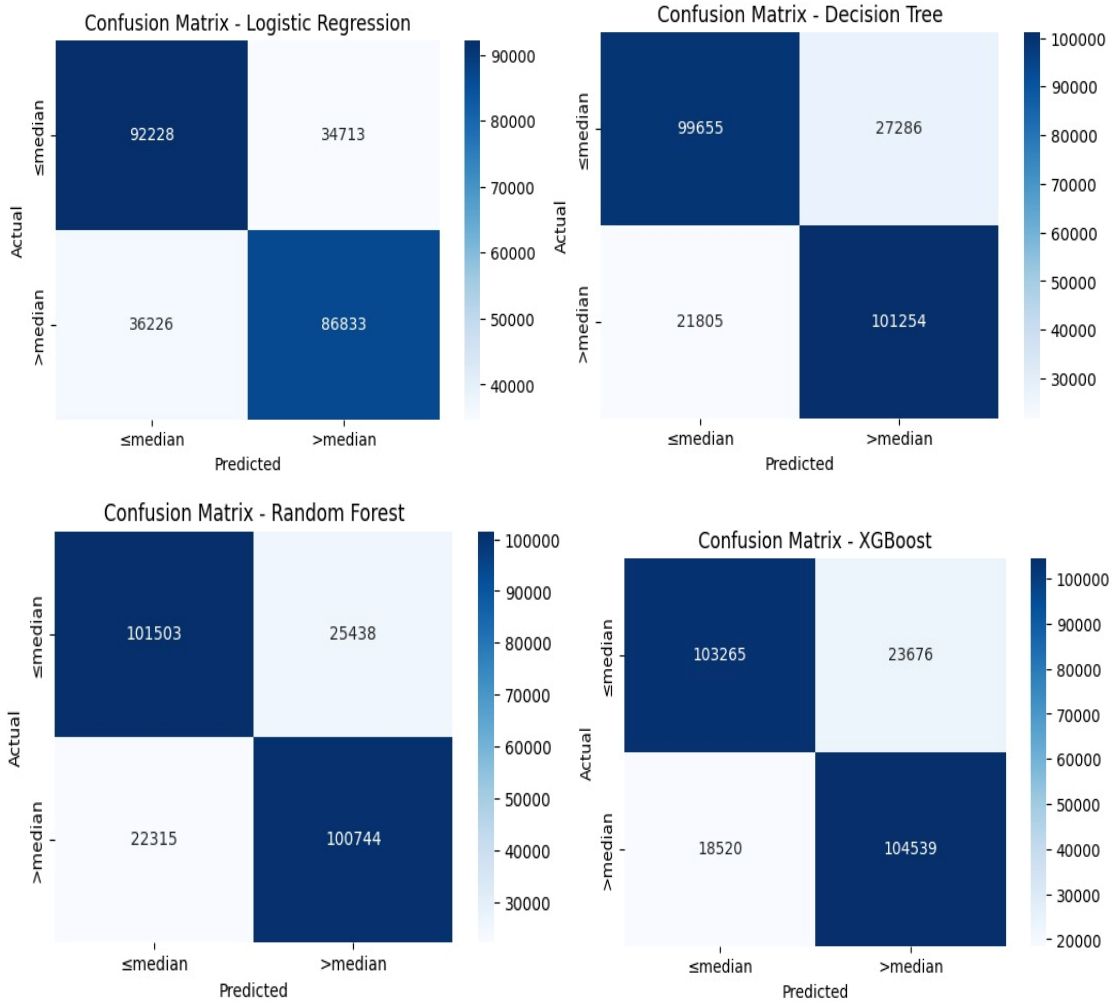


Fig.2.Comparsion Of Confusion Matrices For Ensemble Methods

The figure-11 shows comparing the ROC curves of Logistic Regression, Decision Tree, Random Forest, and XGBoost, we observe clear differences in their ability to distinguish between the two salary classes. The Logistic Regression model shows the weakest performance, with a ROC curve that stays closer to the diagonal line, indicating limited discriminative power and a relatively low AUC (Area Under the Curve). The Decision Tree performs better, with its ROC curve bending further away from the diagonal, showing improved sensitivity and specificity. Random Forest outperforms both, with a more pronounced curve and a higher AUC, thanks to its ensemble approach that captures more complex patterns and reduces variance.

Finally, XGBoost achieves the best ROC curve, staying closest to the top-left corner of the plot, which signifies high true positive rates and low false positive rates across all thresholds. Ic

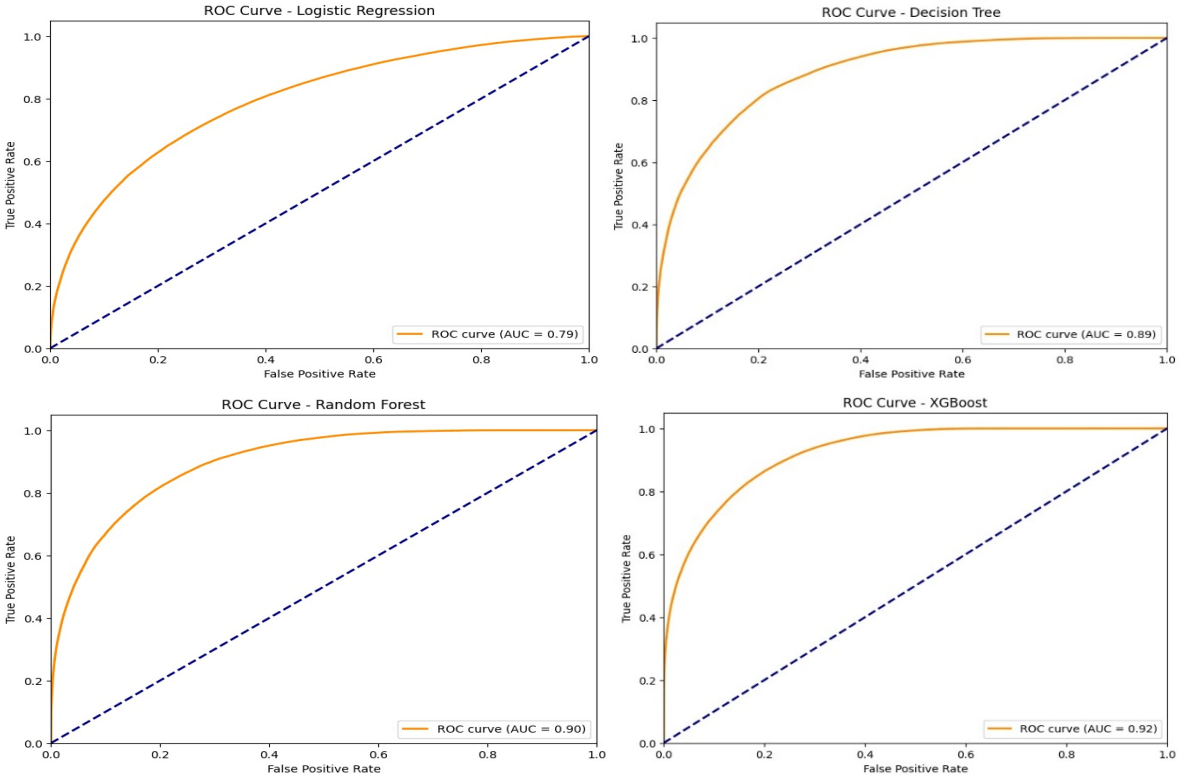


Fig.3. Comparison Of Roc Curves For Ensemble Methods

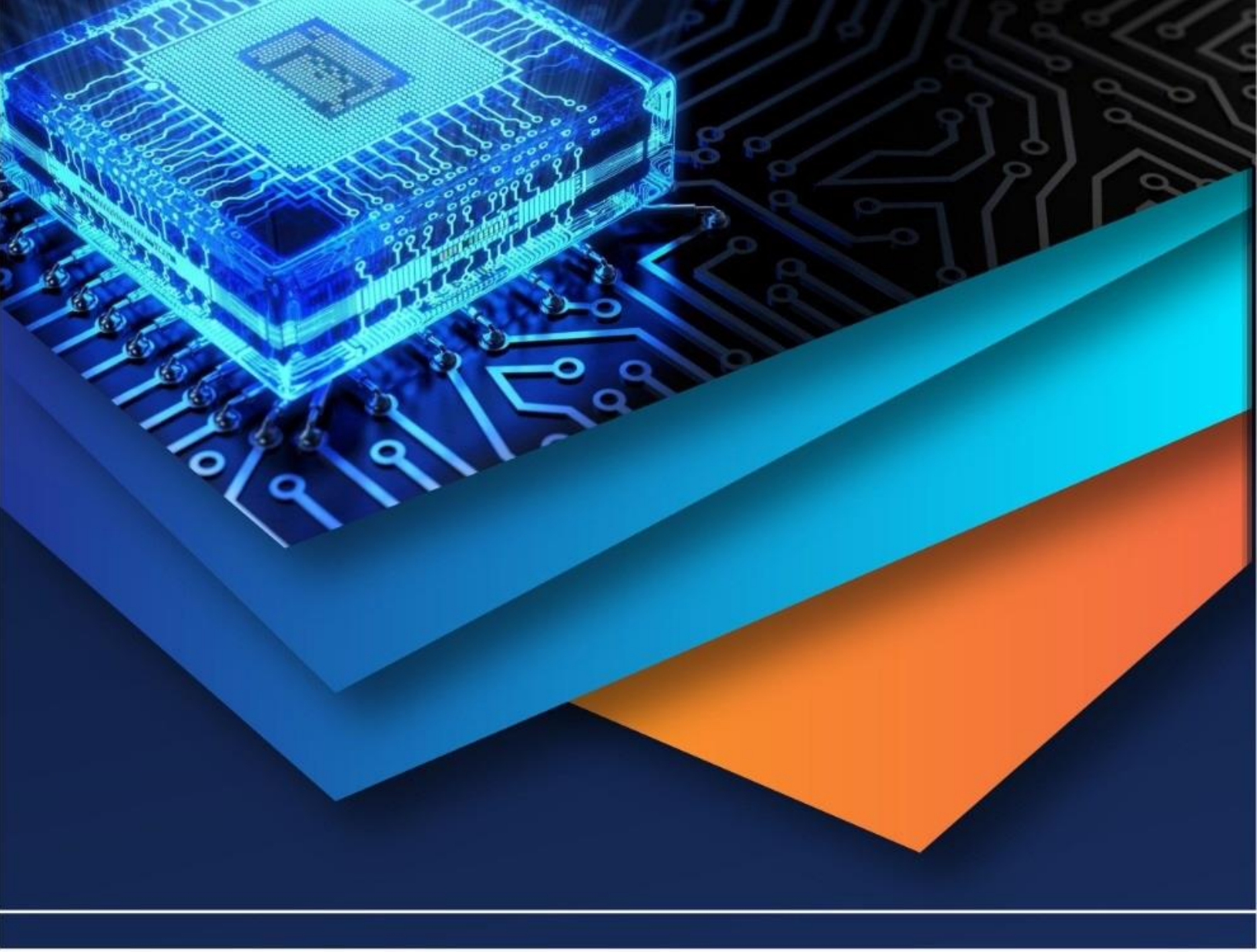
V. CONCLUSIONS

In conclusion, when comparing the four models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—for binary classification based on salary prediction, XGBoost clearly stands out as the most effective. Logistic Regression, being a linear model, falls short in handling complex patterns, resulting in lower accuracy and a less favourable ROC curve. The Decision Tree improves upon this by capturing non-linear relationships, offering better performance with fewer misclassifications. Random Forest enhances the results further by using multiple decision trees to reduce overfitting and increase stability, leading to a well-balanced and accurate model. However, XGBoost outperforms all others with the highest accuracy, precision, recall, and the best ROC curve. Its gradient boosting mechanism allows it to learn from mistakes and refine predictions efficiently. Therefore, XGBoost is the most powerful, accurate, and reliable model for this binary classification task

REFERENCES

[1] Satpute, Babasaheb S., Raghav Yadav, and Pramod K. Yadav. "Machine Learnig Approach for Prediction of Employee Salary using Demographic Information with Experience." 2023 4th IEEE Global Conference for Advancement in Technology (GCAT). IEEE, 2023.
[2] Kaya, Rukiye, Mehtap Saatçi, and Mehmet Gökhan Bakal. "Improving Salary Offer Processes With Classification Based Machine Learning Models." 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, 2024.
[3] Aminu, Habibu, et al. "Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm." International Journal of Innovative Science and Research Technology 8.12 (2023): 1-11
[4] Wang, Guanqi. "Employee Salaries Analysis and Prediction with Machine Learning." 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE). IEEE, 2022.
[5] Hussain, Jawad. "Employee Salary Prediction in HRMS Using Regression Models." Journal of Innovative Computing and Emerging Technologies 4.2 (2024).
[6] Matbouli, Yasser T., and Suliman M. Alghamdi. "Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations." Information 13.10 (2022): 495.
[7] Joshi, Manisha, Savita Bhosale, and Vishwesh A. Vyawahare. "Using fractional derivative in learning algorithm for artificial neural network: Application for salary prediction." 2022 IEEE Bombay Section Signature Conference (IBSSC). IEEE, 2022.
[8] Feng, Ziyuan, Zixian Liu, and Yibo Yin. "Comparison of deep-learning and conventional machine learning algorithms for salary prediction." Applied and Computational Engineering 6 (2023): 643-651.

[9] Das, Sayan, Rupashri Barik, and Ayush Mukherjee. "Salary prediction using regression techniques." Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19) (2020).
[10] Viroonluecha, Phuwadol, and Thongchai Kaewkiriya. "Salary predictor system for thailand labour workforce using deep learning." 2018 18th International Symposium on Communications and Information Technologies (ISCIT). IEEE, 2018.
[11] Khongchai, Pornthep, and Pokpong Songmuang. "Implement of salary prediction system to improve student motivation using data mining technique." 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS). IEEE, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY