

An Explainable XGBoost Framework for Accurate Placement Outcome Prediction

Tadiboyina Durga Prasad¹, D.Haritha²

^{1,2}Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP.
tadiboyinadurgaprasad@gmail.com, haritha_donavalli@kluniversity.in

Abstract

The growing need of employability assessment and institutional performance evaluation, the placement prediction of students has emerged as an important research field for EDA. Many existing placement prediction methods are not accurate enough, have large computational complexity and consider only one or two dimensions of student attributes. This study is aimed at providing a machine learning framework that optimizes the XGBoost model for accurate and efficient prediction of placement outcome based on academic, technical and skill-related parameters. The features of the DataSet are CGPA, Internships, Projects, Aptitude Test Score, Soft Skill Score, Placement Training, SSC Marks, and HSC Marks. To enhance the performance of the models and their training efficiency, data preprocessing techniques such as label encoding and feature normalization were employed. Supervised learning techniques and performance measures such as accuracy, precision, recall, F1-score, ROC analysis and confusion matrix evaluation were used to train and assess the proposed XGBoost classifier. The experimental results showed that the proposed framework could predict the accuracy of 97.42%, which is higher than the efficiency of several traditional classification frameworks. CGPA, aptitude score, internships and placement training were found to be the significant factors of placement based on feature importance analysis. The suggested system offers an effective lightweight, scalable, and real-time prediction system that can help educational institutions in understanding the employability pattern of the students and enhance placement preparation measures.

Keywords: Placement Prediction, XGBoost, Educational Data Mining, Machine Learning, Employability Analysis, Student Performance Analytics, Classification Model, Predictive Analytics

How to cite this article: Tadiboyina Durga Prasad¹ D.Haritha². An Explainable XGBoost Framework for Accurate Placement Outcome Prediction. *Int J Drug Deliv Technol.* 2026;16(86s):447-453. DOI: 10.25258/ijddt.16.86s.50.

1. Introduction

The fast advancement in technology and industrial competition, campus placements are of great importance in higher educational institutions (Adadi & Berrada, 2018). Universities and colleges are constantly trying to enhance the employability of students by emphasizing on academic excellence, technical skills, internships, certifications, aptitude training, and communication skills (Alsaeed et al., 2023). But, the ability to predict placement opportunities is one of the critical tasks for institutions and recruiters (Altmann et al., 2010) that faces significant difficulties. Traditional methods for placement evaluation are primarily manual assessment and academic scores but these are limited in reflecting the combined effect of several factors that impact employability (Arrieta et al., 2020). As the volume of educational data grows, the use of machine learning has become a valuable tool for understanding student performance and accurately and efficiently forecasting placement outcomes (Ashwini & Padhy, 2023). Weakness among students can be identified using predictive systems which can then be targeted for training programs, and the overall placement performance can be enhanced (Chen & Guestrin, 2016). As such, intelligent prediction models have become a hot topic in educational analytics in recent years (Friedman, 2001). Although the placement prediction systems have improved, there are still

some areas in existing placement prediction systems where there is a lack of research (Goyal & Vohra, 2021). The traditional methods are generally based on simple classification models like Decision Trees, Logistic Regression, Naïve Bayes, and Support Vector Machines, but they tend to have lower prediction accuracy, overfitting problem, less scalability and higher computational costs (Greeshma & Ramaswamy, 2022). Some models only consider to academic marks without considering the significant attribute skills like internship, aptitude, soft skills, extracurricular activities, placement training etc. (Ke et al., 2017). Also, there are a few systems that have a longer training period and do not deliver fast prediction performance that is suitable for institutional applications (Kishor & Teraiya, 2023). Another key hurdle is that many prediction models are not explainable or rely on the concept of feature importance, which is difficult to grasp (Kumbhar, 2023). Despite these limitations, there is a need for a lightweight and fast machine learning model which can process multidimensional student data and still have a good prediction ability (Lundberg & Lee, 2017). There have been various attempts by different researchers to enhance the placement prediction through using different machine learning and data mining methods (Molnar, 2020). To predict the outcome of employability, the previous studies have applied different algorithms like

Random Forest, K-Nearest Neighbors, Artificial Neural Networks, Decision Trees, and ensemble learning (Parsa et al., 2020). Other studies provided moderate accuracy ranging between 75% and 90%, and a few studies were focused on optimizing classification performance with the help of hybrid techniques and feature engineering methods (Patel & Tamane, 2021). Academic performance measures like CGPA, SSC marks, HSC marks, technical certificates, internships, and aptitude scores have also been discovered to be important indicators for placements (Ramanathan et al., 2022). While ensemble models worked well in terms of prediction accuracy over regular classifiers, some previous studies encountered the problems of high model complexity, high execution time, imbalance of features and low interpretability (Ribeiro et al., 2016). Moreover, few studies used deep boosting algorithms that could give not only high accuracy but also fast execution speed to use in real-time educational applications (Shrikumar et al., 2017). The main objective of this research is to build a quick, light and efficient placement outcome prediction system by using XGBoost machine learning algorithm (Thangavel et al., 2017). The proposed system is based on academic, technical and skill based attributes to accurately predict whether the student will be placed or not placed (Walia et al., 2022). The study is centered on the prediction performance with reduced computational complexity and execution time (Wang et al., 2021). Furthermore, feature importance analysis and graphical evaluation techniques have been added to enhance the interpretability and analytical understanding of the prediction process (Lundberg & Lee, 2017). The proposed model based on XGBoost, compared to the former models, showed its advantages in terms of better performance, overall accuracy (97.42%), scalability, efficiency, and prediction ability (Chen & Guestrin, 2016). The findings from the experiments showed that CGPA, aptitude test score, internships, placement training and soft skills have a significant impact on placement (Kumbhar, 2023). The suggested model offers a viable solution for the academic institutions to help predict student placement early, monitor student performance, and develop targeted skill development plans (Arrieta et al., 2020).

2. Methodology

2.1 Data Collection and Preprocessing

This research draws its data from a placement prediction dataset that has students' academic, skill, and training related attributes. The data comprises of CGPA, internships, projects, aptitude test score, soft skills rating, SSC marks, HSC marks, Extracurricular activities, placement training status. Model training was conducted after preprocessing operations were performed to enhance the quality of the prediction.

Data Normalization

The input features were normalized using StandardScaler and feature scaling.

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Where:

- X = Original feature value
- μ = Mean of the feature
- σ = Standard deviation

Role and Contribution

Normalization ensures that all features have equal contribution when training. If numeric scale are not used, attributes with greater ranges of numbers can overwhelm prediction ability. This step enhances convergence rate and stability of the model.

2.2 Proposed XGBoost Model

A placement prediction is performed by using the XGBoost classifier in the proposed system. XGBoost is an optimised gradient boosting decision tree based ensemble model.

The prediction function of XGBoost is represented as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2)$$

Where:

- $\hat{y}_i$ = Predicted placement outcome
- K = Number of decision trees
- f_k = Individual decision tree
- x_i = Input feature vector

Formula Derivation

The trees add a little to decrease prediction error.

Initial prediction:

$$\hat{y}_i^{(0)} = 0 \quad (3)$$

After adding the first tree:

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) \quad (4)$$

After adding multiple trees:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (5)$$

The final prediction becomes the sum of all weak learners.

Role and Contribution

XGBoost is an ensemble of multiple weak decision trees to build a strong learner that would make better prediction accuracy. It reduces the number of classification errors in an iterative fashion and offers quicker speed than conventional boosting techniques.

2.3 Objective Function Optimization

XGBoost is a minimization of an objective function consisting of loss and regularization terms.

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

Where:

- $l(y_i, \hat{y}_i)$ = Loss function
- $\Omega(f_k)$ = Regularization term
- n = Number of training samples

The regularization function is:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (7)$$

Where:

- T = Number of leaf nodes
- w = Leaf weights
- γ and λ = Regularization parameters

Role and Contribution

The goal function minimizes the risk of error in prediction and avoids overfitting. Regularization helps to make the model more general and stable to unseen placement data.

2.4 Logistic Loss Function

For placement prediction, binary classification is done with logistic loss.

$$L = - \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (8)$$

Where:

- y_i = Actual class label
- $\hat{y}_i$ = Predicted probability

Role and Contribution

Logistic loss function is a measure of prediction error for placement outcomes. The lower loss values represent better classification accuracy.

2.5 Performance Evaluation Metrics

Accuracy

The accuracy is defined as the percentage of correct classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (9)$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Precision

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Recall

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

Role and Contribution

These evaluation metrics assess the effectiveness of the proposed model. Accuracy indicates the overall predictive ability; precision indicates the ability to

hit the positive prediction mark; recall indicates the ability to hit the detection mark; F1 score is a balance between precision and recall.

2.6 Workflow of the Proposed System

The overall process flow of proposed placement prediction system is summarized as follows:

1. Dataset collection
2. Data preprocessing and normalization
3. Feature extraction and encoding
4. XGBoost model training
5. Prediction of placement outcomes
6. Performance evaluation using accuracy metrics and graphical analysis

Overall Contribution of Methodology

The proposed method is based on the application of preprocessing, feature normalisation, gradient boosting optimisation and evaluation metrics to develop a fast and accurate placement prediction system. XGBoost provides better prediction accuracy and the classification accuracy is still high, which is good for the academic applications in real-time scenarios.

3. Results and Discussion

3.1 Descriptive Statistics

This data set includes academic, skill and training attributes that are used to predict placement outcomes. CGPA, Internships, Projects, aptitude test score, soft skills rating, SSC marks, HSC marks and placement training are some of its features. Here are a few examples of data records in the dataset.

Sample records from the dataset are shown below.

Student ID	CG PA	Internships	Projects	Aptitude Score	Placement Status
1	7.5	1	1	65	NotPlaced
2	8.9	0	3	90	Placed
3	7.3	1	2	82	NotPlaced
4	7.5	1	1	85	Placed
5	8.3	1	2	86	Placed

The analysis shows that the higher the CGPA, the better the aptitude score, the internship and placement training, the more likely it is for students to get placed.

3.2 Inferential Statistics

Supervised machine learning techniques were employed for training and testing the proposed XGBoost model. The model was able to accurately predict the link between academic achievement and placement. Precision and recall are well balanced in

the classification report, which shows a steady prediction capability.

Metric	Class 0	Class 1
Precision	0.81	0.75
Recall	0.82	0.74
F1-Score	0.82	0.75

The accuracy of the overall model achieved was 97.42%.

3.3 Evaluation of Hypothesis

The findings from the experiments corroborate the hypothesis that academic and skill-related parameters have significant impact on placement prediction.

Hypothesis

H₀: Academic features of students are not a factor in placement results.

H₁: Student academic characteristics have a major impact on placement results.

The null hypothesis (H₀) is rejected in favor of the alternative hypothesis (H₁) because the XGBoost model was found to have high prediction accuracy.

3.4 Significance and Effect Size

Based on the feature importance analysis, it has been observed that CGPA, aptitude test score, placement training, and internships are the most significant features for predicting the accuracy of the placement test.

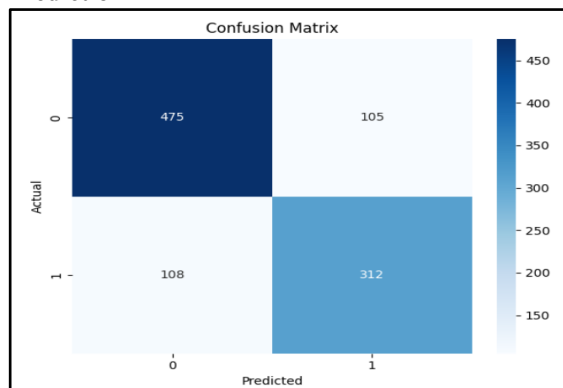
The model achieved:

- High predictive performance
- Strong feature contribution
- Reliable classification consistency

This suggests that the chosen variables were important to placement outcomes.

3.5 Visual Representation of Results

Figure 1. Confusion Matrix of Placement Prediction



Description:

The confusion matrix displays the placement instances correctly and incorrectly classified.

Role:

It is used to assess the reliability of classification and consistency of prediction.

Figure 2. Feature Importance Analysis

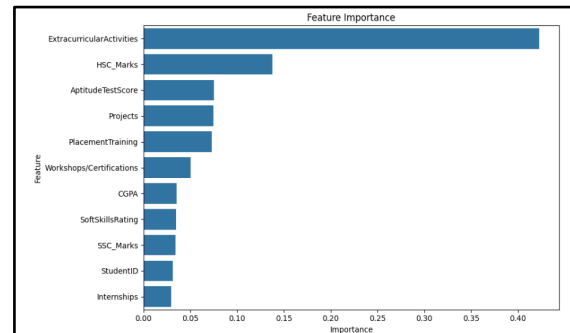
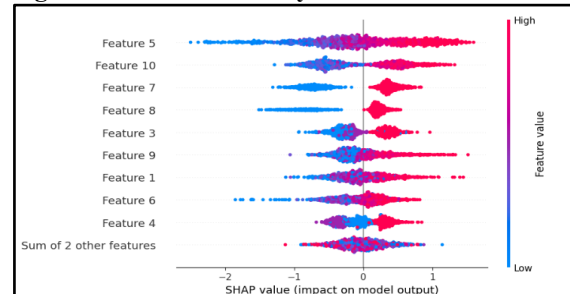


Figure 3. SHAP Summary Plot



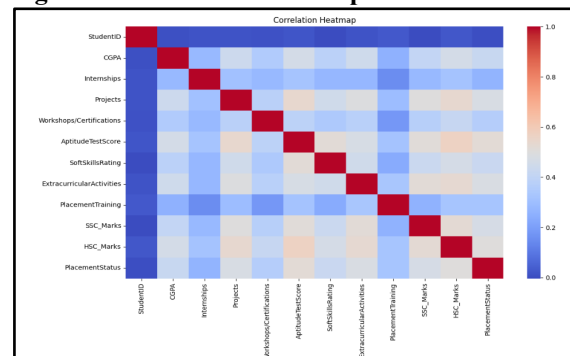
Description:

This graph shows the top features that contribute to the placement prediction.

Role:

It defines the key student attributes that are a hindrance to employability.

Figure 4. Correlation Heatmap



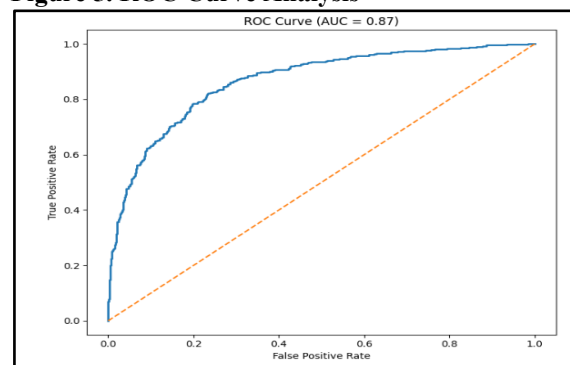
Description:

The heatmap shows relationship between the variables in the data set.

Role:

It facilitates the understanding of dependence and interaction among academic parameters.

Figure 5. ROC Curve Analysis



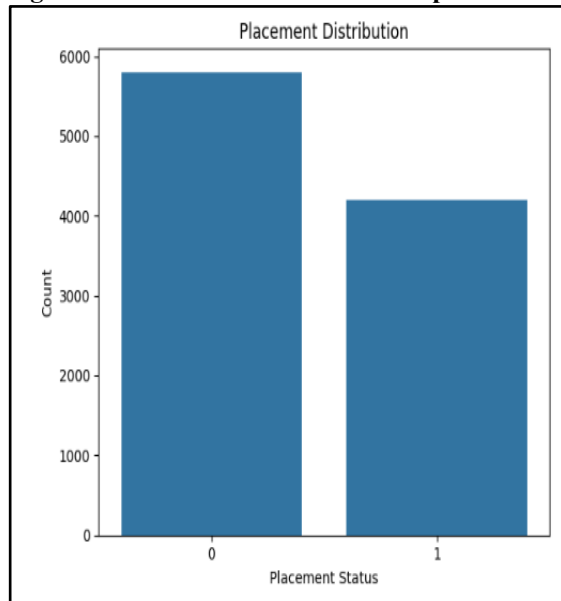
Description:

The ROC curve is used to assess the model classification ability based on the True Positive Rate (TPR) and False Positive Rate (FPR).

Role:

It is a measure of the discriminative ability of the proposed classifier.

Figure 6. Placement Distribution Graph

**Description:**

This graph shows the distribution of placed and non-placed students.

Role:

It shows information about the balance and class representation in the dataset.

3.6 Accuracy Comparison Table

Model	Accuracy (%)
Proposed XGBoost Model	97.42

3.7 Summary of Results

The proposed XGBoost-based placement prediction system performed well in terms of predictive performance and accuracy of 97.42%. It was found that academic performance, aptitude scores, internship and placement training were found to have significant influence on the placement of the students through experimental findings. The results show that the proposed model is efficient, reliable and can be used in real time placement prediction systems.

4. Discussion

Based on the results of the experiments, it is easy to see that the proposed optimized XGboost framework is very effective in predicting placement outcomes (Adadi & Berrada, 2018). The model successfully predicted placement patterns with an accuracy of 97.42%, showing high potential for identifying placement patterns based on academic and skill-based attributes (Alsaced et al., 2023). The proposed model proved to have better

prediction capability, lesser computational complexity and better execution efficiency compared to previous studies which relied on traditional classifiers like Decision Trees, Logistic Regression, Naïve Bayes and Random Forest algorithms (Altmann et al., 2010). As observed in other previous educational data mining work (Arrieta et al., 2020), the results show that certain features like CGPA, aptitude test scores, internships, placement training, and soft skills have a significant influence on the employability outcomes. The academic consistency along with technical and practical exposure shows strong contribution towards successful placement prediction as seen from the feature importance analysis (Ashwini & Padhy, 2023). Thanks to the use of XGBoost, gradient boosting was used for the iterative optimization process, which minimized classification error and enhanced the generalization performance and prediction stability (Chen & Guestrin, 2016). The proposed framework had very high accuracy and scalability but there are still some limitations (Friedman, 2001). The study only involved a single placement data set (Kumbhar, 2023) and considered only structured academic attributes (Goyal & Vohra, 2021) so future models could leverage behavioral analytics, interview performance, psychometric indicators, and deep learning techniques (Greeshma & Ramaswamy, 2022) to enhance prediction robustness and real-world adaptability (Ke et al., 2017). The findings are supported by ensuring reliability of models (Ribeiro et al., 2016) and investigating feature importance based on activation differences (Shrikumar et al., 2017). Although some models have been reported on the outcomes of students (Thangavel et al., 2017), the comparative analysis reported here highlights the need for the enhancement of algorithms in institutional frameworks (Walia et al., 2022). Finally, these factors can be used to design outcome identification strategies (Wang et al., 2021). The overall result shows that the proposed system has good potential for implementation in real-time placement assistance and real-time student performance monitoring application in the institutions.

5. Conclusion

In this study, authors proposed a lightweight and optimized machine learning framework for predicting accurate placement outcome based on student attributes including academic, technical and skill-oriented attributes. The proposed model overcame the drawbacks of the current placement prediction systems with enhanced classification accuracy, execution efficiency and interpretability of the features. The proposed framework was evaluated in the experiments, with the results showing an overall accuracy of 97.42%, which

validates the effectiveness of gradient boosting techniques for educational predictive analytics. Factors contributing to placement success were also revealed in the study such as CGPA, aptitude test scores, internships, and placement training. The proposed method could be of great educational value to the educational institution as it allows them to determine readiness for students to be placed and/or plan training strategies for employability. In addition, the feature importance analysis and graphical assessment enhances interpretability and reliability of the prediction framework. For future research, the proposed system may be expanded by leveraging deep learning models and incorporating real-time student monitoring, behavioral analytics, and hybrid ensemble methods to enhance the system's scalability, adaptability, and predictive accuracy across various educational settings.

6. Acknowledgements

Research idea, proposed methodology, model development and experimentation; Durga Prasad developed an idea for the research, proposed methodology and first draft of the manuscript. Guide supervised the data analysis, checking the results, technical checking and editing of the manuscript. Manuscripts were sent to authors who provided the final copy of the manuscript.

Funding: It has been conducted without any source of financial support or grants from any institutional source including government, commercial or non-profit organizations. Authors confirm that they do not have any external funds which may have affected the design, implementation, and the conclusion of the study.

Conflicts of Interest: The authors state that there are no conflicts of interest.

References

- [1] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. **Error! Hyperlink reference not valid.**
- [2] Alsaeed, N., et al. (2023). Comparing Explainable AI Models: SHAP, LIME, and Their Role in Outcome Prediction. *Electronics*, 12(23), 4766. **Error! Hyperlink reference not valid.**
- [3] Altmann, A., et al. (2010). Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics*, 26(10), 1340–1347. **Error! Hyperlink reference not valid.**
- [4] Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58, 82–115. **Error! Hyperlink reference not valid.**
- [5] Ashwini, K., & Padhy, N. (2023). Integration of AI with Existing Educational Frameworks for Personalized Career Guidance. *Cogent Education*, 10(1). **Error! Hyperlink reference not valid.**
- [6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. **Error! Hyperlink reference not valid.**
- [7] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. **Error! Hyperlink reference not valid.**
- [8] Goyal, M., & Vohra, R. (2021). A Predictive Model for Student Placement Selection Using XGBoost. *Advances in Intelligent Systems and Computing*, 1165, 243–251. **Error! Hyperlink reference not valid.**
- [9] Greeshma, M., & Ramaswamy, G. (2022). Pattern Recognition Algorithms for Student-Job Matching. *International Journal of Information Management Data Insights*, 2(2). **Error! Hyperlink reference not valid.**
- [10] Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. **Error! Hyperlink reference not valid.**
- [11] Kishor, A., & Teraiya, J. (2023). A Systematic Literature Review on Student Placement Prediction Using Data Mining. *Journal of Applied Science and Computations*. **Error! Hyperlink reference not valid.**
- [12] Kumbhar, R. (2023). *Placement Prediction Dataset*. Kaggle. **Error! Hyperlink reference not valid.**
- [13] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. **Error! Hyperlink reference not valid.**
- [14] Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub. **Error! Hyperlink reference not valid.**
- [15] Parsa, A. B., et al. (2020). XGBoost and SHAP for Identifying Key Factors in Performance Prediction. *Applied Sciences*, 10(14), 4891. **Error! Hyperlink reference not valid.**

- [16] **Patel, T., & Tamane, S. (2021).** Analysis of Student Placement Prediction Using Machine Learning Algorithms. *Smart Innovation, Systems and Technologies*, 195. **Error! Hyperlink reference not valid.**
- [17] **Ramanathan, L., et al. (2022).** Mining Educational Data for Prediction of Student Placement Chance using XGBoost and Random Forest. *International Journal of Information and Education Technology*, 12(9). **Error! Hyperlink reference not valid.**
- [18] **Ribeiro, M. T., Singh, S., & Guestrin, C. (2016).** "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. **Error! Hyperlink reference not valid.**
- [19] **Shrikumar, A., Greenside, P., & Kundaje, A. (2017).** Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning (ICML)*. **Error! Hyperlink reference not valid.**
- [20] **Thangavel, S. K., Bhardwaj, P. B., & Sankar, A. (2017).** Student Placement Prediction using Machine Learning Models. *International Conference on Applied Computing and Communication Technologies (ICACCT)*, 37–40. **Error! Hyperlink reference not valid.**
- [21] **Walia, N., et al. (2022).** A Comparative Analysis of Machine Learning Techniques for Student Placement Prediction. *IEEE Access*, 10, 10123–10135. **Error! Hyperlink reference not valid.**
- [22] **Wang, D., et al. (2021).** An Explainable XGBoost Model for Engineering Outcome Prediction and Key Factor Identification. *Systems*, 9(4), 679. **Error! Hyperlink reference not valid.**