International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Early Placement Prediction in Higher Education Institutions Using Machine Learning

Namita Satpute[1], Dr. Ayesha Siddiqui[2]

[1]Computer Application Department of Computer Science, JSPM University, Pune, Maharashtra, India
Email: *namitasatpute21[at]gmail.com*

[2]Associate Professor Department of Computer Science, JSPM University, Pune, Maharashtra, India
Email: *ais.scos[at]jspmuni.ac.in*

Abstract: ***In recent times, campus placement prediction has emerged as an important area of research owing to its direct influence on the future careers of students as well as on the success of educational institutes. The conventional approaches employed for such predictions are restricted in nature and are unable to derive any insights. For this reason, in this paper, an approach is proposed to predict the outcome of campus placement using machine learning techniques based on students' academics and other information. For this research, the dataset used has a set of important features including gender, scorecard grades from secondary and higher secondary classes, degree percentage, work experience, specialization, and MBA marks. Preprocessing steps such as removing unwanted features, handling missing values, and encoding categorical features are taken to prepare the dataset. Furthermore, feature engineering has been done using the creation of some new features such as total marks and average marks. Data is split into training and testing data with a split ratio of 70:30. Three machine learning algorithms, Gradient Boosting, LightGBM, and CatBoost, are built and compared based on their performance. Based on the experimental findings, it is observed that LightGBM performs better with a maximum accuracy of 86%, whereas Gradient Boosting and CatBoost produce the same accuracy rate of 80%. In order to ensure better understanding, various visualization techniques, including confusion matrix, heat maps, and accuracy charts, have been applied. As seen from the above discussion, it can be concluded that the suggested method proves successful in utilizing machine learning for prediction purposes.***

Keywords: Machine Learning, Student Placement Prediction, LightGBM, Gradient Boosting, CatBoost, Feature Engineering, Classification, Educational Data Mining

1. Introduction

Use of machine learning technology in higher education has been growing tremendously over the last few decades, helping institutions in adopting data-driven approaches in making their decisions [1] [2]. Among the important applications of machine learning is the process of predicting student placements based on factors related to students' background and academic achievements. Prediction of the probability of getting placement would help identify those who may need additional assistance, hence improving placement results. Predicting student placements, however, is far from being an easy undertaking. There are several factors that affect the performance of students including academic results, educational background, specialization, as well as previous working experience [3]. In most cases, those factors tend to influence one another, hence making it hard to predict student placement based on conventional analysis approaches alone.

In order to solve these issues, in this study, a framework for prediction of campus placement is developed with the help of the machine learning algorithms which uses a structured dataset containing features like gender, secondary and higher secondary education marks, degree percentage, work experience, specialization, and marks obtained in MBA [4] [17]. After data cleaning, pre-processing, and creation of new attributes like total marks, average marks, and performance score, three machine learning algorithms including gradient boosting, light GBM, and catboost have been implemented [5] [6] [7].

These models are compared to each other by splitting the data into a training set (70%) and test set (30%). Finally, the performances of all models are measured on the basis of their accuracy. It is observed that among these models, LightGBM gives best results for the problem [6].

Moreover, various visualization techniques like confusion matrix, heat map, and graph depicting accuracies of various models are used in order to understand the working of models in a more effective way [8] [20].

2. Problem Statement

Predicting successful placement of students in higher educational organizations is considered a difficult process since it involves several factors [1]. Existing approaches are typically focused on evaluating a student according to simple criteria and manual analysis of performance indicators that do not reflect the complexity of interrelated parameters . This approach leads to a high number of students who cannot be effectively placed after undergoing certain programs, resulting in a low placement rate. Moreover, there is a necessity to implement an efficient prediction system in order to increase the probability of positive placement outcomes [4]. Thus, there is a need to create an efficient system that will provide accurate prediction of employment through analyzing the required data and employing advanced machine learning approaches [9] [19]. It should be noted that the future system needs to consider several features of each student. This project seeks to solve this problem through the implementation of machine learning algorithms that predict placements, namely Gradient Boosting, LightGBM, and CatBoost [5] [6]. The purpose of this is to establish the best model for predicting placements, which will help institutions

optimize their placements.

3. Literature Review

Student placement prediction has become one of the most researched topics in recent years, mainly because the amount of data related to students has increased significantly [2]. Machine learning and data mining techniques are employed to process the data and discover patterns that can be useful in predicting future placements and career success of a graduate [2]. The main disadvantage of traditional statistical approaches is the difficulty to model relationships between multiple academic and personal factors [1].

Machine learning classification approaches were suggested by many scholars as means of improving accuracy of prediction [9] [19]. The choice of algorithm depends on specific properties of the dataset. Boosting algorithms like Gradient Boosting, LightGBM, and CatBoost became popular among predictive models because of their flexibility and efficiency in predicting structured data. Features used for prediction include scores and grades, educational background, specialization, work experience, etc. It was proven that the combination of several important features improves the accuracy of prediction. Visualization tools to examine the performance of the models used in previous studies are yet another vital component that should be considered when discussing past research. Confusion matrices, heatmaps, and even accuracy comparison graphs are often employed to test and assess the efficiency of particular algorithms [10].

Nevertheless, certain aspects related to the research on machine learning models are still unexplored. For instance, a considerable number of studies tend to analyze only one algorithm and thus fail to provide a comparison between several models based on the same data set [4]. What is more, some papers are characterized by the usage of limited data sets. Additionally, there seems to be an absence of applicable models for actual educational settings due to the complexity of frameworks developed in previous studies. By addressing these problems, the current study attempts to apply and compare several boosting-based machine learning algorithms: Gradient Boosting [5], LightGBM [6], and CatBoost [7] to a data set preprocessed in the same way. Moreover, the primary purpose of the study is to identify the optimal model among these three in terms of application in the field of placement prediction.

4. Objectives of Research

The major objective of this research is to develop an efficient and effective machine learning model for predicting student placements using academic records and background information [1].

The objectives of the research are as follows:

- To analyze the database structure which includes student attributes such as gender, academic scores (SSC, HSC, degree, MBA), specialization, and experience.
- To perform data cleaning through elimination of irrelevant variables, treatment of missing observations, and conversion of categorical variables into numerical form [8].
- To use feature engineering techniques where further features like total marks, average marks, and performance grade will be generated [11].
- To use various machine learning algorithms such as Gradient Boosting [5], LightGBM [6], and CatBoost [7] for student placements prediction.
- To calculate and compare the accuracy of various machine learning algorithms and then choose the appropriate one [10].
- To present the results in graphical format through confusion matrix, heatmap, and accuracy plots.
- To identify the best performing algorithm (LightGBM) and determine whether it can be used for student placements predictions [6].
- To develop an effective model of prediction for educational institutes to adopt their future placements policies [1] [2].

5. Related Work

Recently, the topic of predicting students' placements using different machine learning classification models was investigated in the field of educational data mining. Researchers applied different machine learning models on structured data with features including grades, education history, specialization, and previous work experience to determine which type of patterns could indicate a likelihood of student placement [1] [4].Nowadays, boosting algorithms, such as Gradient Boosting [5], LightGBM [6], and CatBoost [7], receive significant popularity because of their high efficiency in working with complex and nonlinear data dependencies. Moreover, such models demonstrate higher levels of performance when being used with structured datasets [11]. On the other hand, data preprocessing became one of the crucial factors determining the success of applying these algorithms. Feature selection and data preparation steps also gained popularity recently [11]. These actions include deleting irrelevant attributes from a dataset, dealing with missing values in the dataset, and converting categorical features into numerical ones [8]. Furthermore, creating new features based on available data is considered helpful in improving prediction [18]. Visualization methods are popular when it comes to checking how the models perform [8]. For example, one can use confusion matrices, heat maps, and charts that allow comparing the accuracy of various models and thus determine their strengths and weaknesses [10]. Nevertheless, there are several drawbacks that should be mentioned when discussing recent research on the topic. First of all, some research papers consider only one model without giving any comparison with other approaches [9]. Second, many researchers have been working with small samples [2]. Lastly, there is a lack of research works which offer an easy-to-use framework [1]. The purpose of the current study is to fill this gap and conduct a comparison between three models (Gradient Boosting [5], LightGBM [6], and CatBoost [7]) and find out the best of them for predicting placements on a structured dataset [1].

6. Research Gap

Though there is an increasing application of machine learning algorithms to make accurate predictions about

students' placement, a number of gaps can be noted within this field of research [3]. For instance, many authors tend to concentrate on enhancing prediction accuracy by neglecting the necessity to make comparisons among various models [9]. Thus, it is rather difficult to determine which algorithm will show the most effective results [10].

Furthermore, there is no agreement on using the same data sets, measures of effectiveness, or set of features in order to be able to conduct a fair comparison between results. What is more, in certain cases, the amount of data provided is insufficient for achieving reliable predictions and applying models in practice [2].

Moreover, not enough attention is paid to the preparation stage, such as data cleaning and creating new valuable features based on already available information [8]. If not processed in the appropriate way, missing values or categorical data may have a negative influence on further stages [11]. Finally, there are no practical approaches to implementing algorithms in educational organizations [1]. The main emphasis is always placed on theory and its validation, whereas the process itself requires more attention to be paid to real applications [9]. Thus, there is an evident research gap in the need to develop a single machine learning algorithm which uses uniform preprocessing, takes advantage of feature engineering, and utilizes several different models, including Gradient Boosting [5], LightGBM [6], and CatBoost [7], on a single set of data for comparison. Closing this gap will allow us to determine the most appropriate model for our task.

7. Proposed System Architecture

A multi-layer architecture has been proposed in this system, which involves the use of machine learning techniques for predicting the result of placement for the students [4]. With the help of multi-layer architecture, it will become easy to manage the workflow in a systematic manner. The proposed architecture comprises four different layers which include Data Layer, Processing Layer, Model Layer, and Visualization Layer. Data Layer: The data related to students consisting of gender, SSC/HSC Percentage, Degree percentage, work experience, specialization, and MBA Performance will be gathered [4]. These pieces of information can act as inputs to construct the predictive model.Processing Layer: The process of preprocessing involves the deletion of redundant columns, handling missing values, and conversion of categorical values to numeric [8]. New features will be engineered with names such as Total Marks, Average marks, and performance score [11]. Model Layer: In this section, application of machine learning algorithms takes place. Gradient boosting [5], Light GBM [6], and Catboost [7] models have been implemented for this particular case study, and their performances have been evaluated in terms of accuracy.

Visualization Layer serves the purpose of analysis as well as visualization of the results [8] This process involves the usage of various graphical methods like confusion matrix, heatmap, and accuracy comparison plots [10]. Finally, predictions about the optimal location are made by the system, which will be beneficial for both parties.

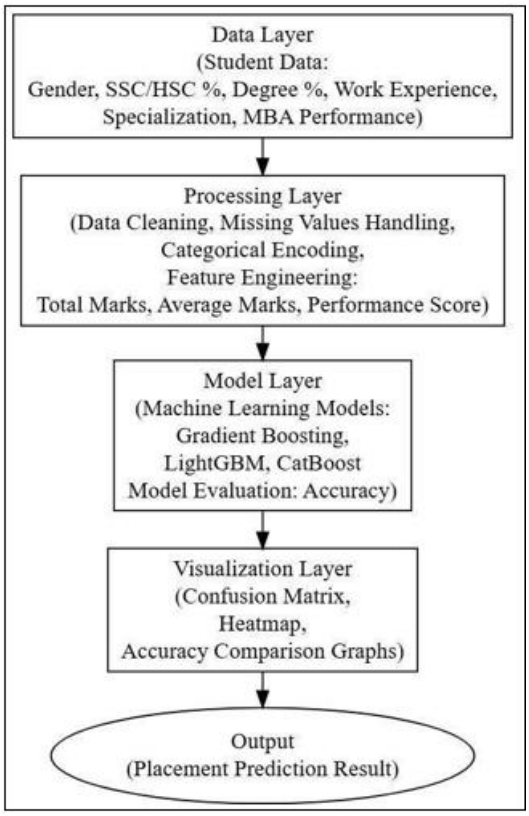

Figure 1: Multi-Layer Architecture for Student Placement Prediction System

8. Research Methodology

1) Dataset Used

In the study, the dataset used comprises student characteristics such as gender, SSC percent, HSC percent, degree percent, work experience, specialization, and MBA percent [4]. The aforementioned attributes function as input parameters to predict the placement status of the students.

2) Tools and Software

The development process is performed in the programming language Python [12] [20]. For data manipulation purposes, the project uses Pandas and NumPy libraries. In addition, the Scikit-learn library will be used to develop the predictive model [8] [15]. The environment of the project development is Jupyter notebook. Visualization in the current project is achieved through the application of Matplotlib and Seaborn libraries.

3) Algorithms/Models

Different algorithms are employed to create the predictive model [10] [19]. For implementing the algorithmic model in the current study, classifiers such as Gradient Boosting [5], LightGBM [6],and CatBoost [7] are selected. Such models were selected because of their ability to deal effectively with classification problems.

4) Experimental Setup

To perform experiments, the used dataset is split into training and testing datasets by the ratio of 70-30 [4].

5) Evaluation Measures

The performance of the models is measured through the use of measures like accuracy, precision, recall, and F1 measure

[10]. The confusion matrix is also used for analysis of classification outcomes [8].

9. Results and Analysis

1) Classification Performance

The classification performance is the measure used to assess the efficiency and effectiveness of the models in predicting the output [10]. Several models were used in this study to forecast the output depending on different features contained in the datasets. Some of the models include Gradient Boosting [5], LightGBM [6], and CatBoost [7]. When using the models in terms of accuracy, the LightGBM appears more efficient and effective in comparison to the other two models [6]. Based on the obtained result, its accuracy rate stands at 86.15%. In contrast, the accuracy for Gradient Boosting is 80% while that of CatBoost is 80%%.

Table 1: Accuracy Evaluation of Gradient Boosting, LightGBM, and CatBoost Models

Model	Accuracy
Gradient Boosting	80%
LightGBM	86%
CatBoost	80%

2) Confusion Matrix Analysis

Confusion matrices provide information concerning the performances of each model in forecasting the output [10]. It provides information regarding the number of wrong and correct predictions made by the models. The confusion matrix contains the number of the right and wrong predictions made depending on whether the data was placed or not [8]. From the analysis of the confusion matrix, the model which makes the least incorrect predictions is LightGBM [6]. This is because of its effectiveness in making true positive and true negative predictions.

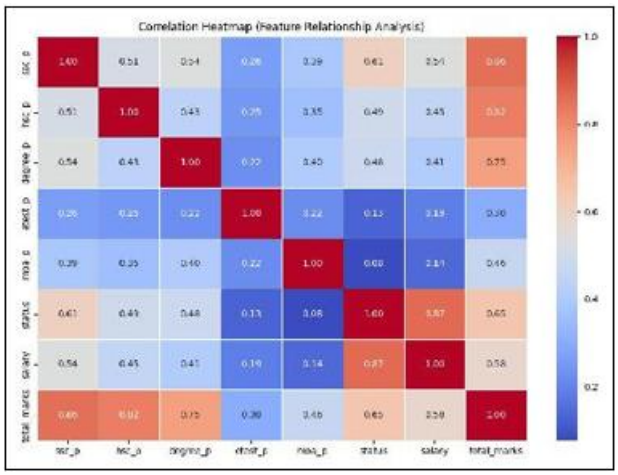

Figure 2: Correlation Heatmap

3) ROC Curve Analysis

The receiver operating characteristic (ROC) curve technique is used to determine the efficiency of the classification performance between models using the performance of True Positive Rate vs. False Positive Rate [10]. The model that achieves a larger Area Under the Curve (AUC) is deemed to perform well.

LightGBM is found to be a better performing model than other models in terms of classification performance in this study [6].

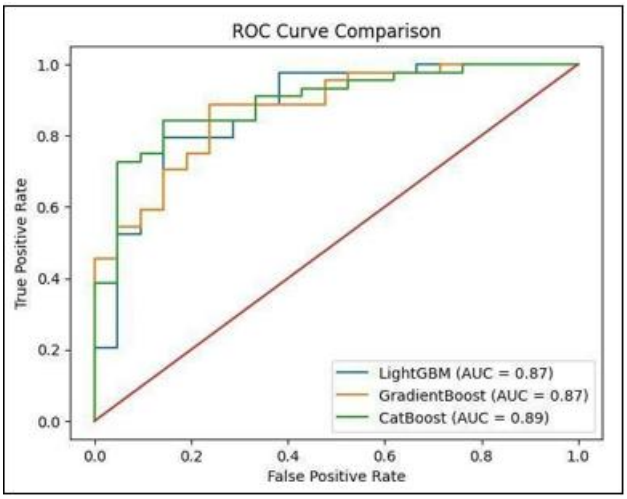

Figure 3: ROC Curve Comparison

4) Precision, Recall, and F1-Score Analysis

The sole use of accuracy as a metric does not suffice in evaluating the performance of the models [10]. Hence, additional metrics like Precision, Recall, and F1-score are taken into consideration. Precision refers to how many of the placements predicted by the model are accurate while Recall is the measure of how many students who get placed are accurately identified by the model. F1-score is the balance point of Precision and Recall. It can be concluded that the lightGBM model is more balanced than gradient boosting and catboost models based on the above-mentioned three parameters [6].

Table 2: Precision, Recall, and F1-Score Comparison of Classification Models

Model	Class	Precision	Recall	F1- Score
Gradient Boosting	0	0.79	0.52	0.63
	1	0.8	0.93	0.86
	Weighted Avg	0.8	0.8	0.79
LightGBM	0	0.93	0.62	0.74
	1	0.84	0.98	0.91
	Weighted Avg	0.87	0.86	0.85
CatBoost	0	0.83	0.48	0.61
	1	0.79	0.95	0.87
	Weighted Avg	0.81	0.8	0.78

5) Feature Importance Analysis

Feature importance analysis helps to identify which are the important features from the set of inputs that affect the outputs [11]. From the analysis, it can be inferred that academic performance metrics like SSC, HSC, Degree percentage, and MBA score make up for a significant portion of the factors affecting the placements [4]. Besides that, work experience and specialization are also significant factors that affect prediction accuracy [9].

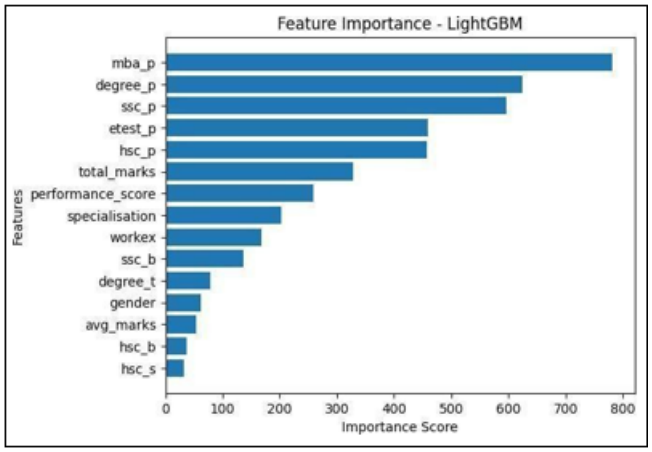

Figure 4: Feature Importance - LightGBM

10. Discussion

As seen from the experimental results, the accuracy of prediction obtained by using LightGBM was the highest at 86% in comparison with other models used for this research task [6]. It can be explained by the ability of LightGBM to process efficiently structured datasets while being capable of modeling more complex relationships between variables [6].The features chosen in this research such as scores in various examinations (SSC, HSC, degree, MBA), work experience, and specialization are essential for determining the probability of placement since they contribute significantly to obtaining useful information [4]. Relevant academic and background characteristics were included into the dataset used for this research allowing for building the relationships between input features and placements successfully [9]. Also, proper preprocessing and engineering of features, including calculating the average and total number of marks, have helped improve the results [11]. The application of different evaluation techniques (accuracy, confusion matrix, and others) contributes to obtaining accurate results while enabling researchers to compare several models [10].

Benefits:
Offers excellent predictive accuracy with LightGBM giving the best results when compared with all other models [6]. Employs many machine learning models for analysis and comparison, making it easier to select the optimal methodology [10]. Involves feature engineering which improves data quality and enhances model accuracy [11]. Helps identify key variables affecting placement decisions [4]. Predicts student placement outcomes in advance, allowing for proper action to be taken at an early stage . The proposed framework is easy to implement, scalable and practical [1].

Limitations:
Even though the proposed model provides good results for the student placement prediction problem, there are several limitations that should be taken into account. The first limitation is the size of the used dataset [9]. Being rather limited in size, the dataset may negatively impact the generalization power of the model [2].The next limitation concerns the choice of features to use in the model [4]. The current study is based only on the analysis of some basic attributes and their influence on student success, such as academic performance, education scores, specialization, and work experience. Nevertheless, other aspects that also have a significant effect, including economic situations, special requirements of a company, and personal characteristics, have been omitted. Furthermore, even though the model makes use of more sophisticated algorithms, such as Gradient Boosting [5], LightGBM [6], and CatBoost [7], its efficiency is dependent on the input data, and therefore, may be influenced by a lack of representation of some particular features.Finally, the current implementation of the model does not include additional elements, such as visualization techniques, and cannot be used in real time [8].

11. Conclusion

In this paper, our main focus is on prediction of student placement based on their performance in higher educational institutes [1]. Traditional approaches have limitations regarding capturing the inter-relation between academic success and employment-related variables in a structured dataset [3] [17]. Therefore, this research paper proposes an approach based on machine learning by employing gradient boosting [5], LightGBM [6], and CatBoost [7] algorithms to predict students' placement status. It has been observed from the analysis that LightGBM provides the best accuracy score of 86% [6]. Gradient boosting algorithm also has reasonable accuracy [5], whereas CatBoost lags behind the rest of the two algorithms regarding predictive accuracy [7]. The better accuracy score obtained by LightGBM implies its efficiency in handling structured data for capturing inter-relation between features [6]. Data preparation plays a crucial role in increasing predictive efficiency of a model [8]. In addition to SSC, HSC, degree, MBA, work experience, and specialization, total marks and average marks also serve as vital features in increasing predictive efficiency of LightGBM [11]. In order to further analyze the performance of the model, visualization methods such as confusion matrix, heatmap, and accuracy comparison were employed [8]. This helped in analyzing the data and ensured that the analysis done would be accurate and clear [10] [11]. On the whole, the framework suggested herein highlights that machine learning is an efficient method that could be applied for predicting student placements. The proposed approach offers an effective way of solving this problem, which may help educational institutions improve their methods of student placement and development through predictive analytics [9]. In future research, efforts could be made towards working with bigger datasets, implementing more sophisticated models and creating solutions in the form of real-time systems [2].

12. Future Work

In summary, the presented project offers a solid solution to the problem of predicting student placement outcomes [4]. Nevertheless, there are a few improvements that can be made in order to improve the performance and accuracy of predictions in the future. First, data collection should not be limited only to the current dataset as the predictions obtained using such a limited amount of information will not be accurate and reliable enough [9]. Therefore, it may be worth extending the dataset by gathering information about

students from other colleges or universities [2]. Another interesting direction in terms of research is experimenting with even more sophisticated techniques of machine learning. As mentioned above, models of Gradient Boosting [5], LightGBM [6], and CatBoost [7] were implemented successfully; however, the performance of the system may be further enhanced through proper hyperparameter tuning and employing more complex techniques of ensemble learning [14] [15]. Furthermore, it may be beneficial to add even more features to the dataset in order to obtain a wider range of information about students' characteristics. The application can be built to have a real-time feature or even a dashboard [8]. With such an application, there is a guarantee that institutions will get timely predictions about their students' performance and know those who need extra care and training [1]. The last way of improving the prediction process is through combining the model with the database of the institution, where it can automatically collect all the required data [4].

References

[1] S. Kumar and R. Singh, "Student placement prediction using machine learning techniques," International Journal of Computer Applications, vol. 182, no. 10, pp. 1–6, 2018.
[2] M. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601–618, 2010.
[3] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63–69, 2011.
[4] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.
[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.
[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[9] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica, vol. 31, pp. 249–268, 2007.
[10] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York: Springer, 2013.
[11] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[12] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann, 2016.
[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1145.
[17] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Hoboken, NJ: Pearson, 2018.
[18] H. Han, X. Guo, and H. Yu, "Variable selection using mean decrease accuracy and mean decrease Gini based on random forest," in Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC), 2016.
[19] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Machine learning: A review of classification and combining techniques," Artificial Intelligence Review, vol. 26, no. 3, pp. 159–190, 2006.
[20] J. Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-End. Machine Learning Mastery, 2016.