iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** II **Month of publication:** February 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.58548

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 12 Issue II Feb 2024- Available at www.ijraset.com*

# Predicting Student Placement in College Using Machine Learning

Imtiyaz Ahmad Magray[1], Gurinder Kaur Sodhi[2]
[1]*M. Tech Scholar, Department of Electronics and Communication Engineering, Desh Bhagat University, Punjab, India*
[2]*Assistant Professor, Department of Electronics and Communication Engineering, Desh Bhagat University, Punjab*

***Abstract:*** ***The research on student placement prediction in higher education has been a focal point, addressing challenges related to balanced accuracy and generalization across diverse datasets. Prior studies grappled with issues in feature representation and interpretability. This study aims to overcome these challenges by introducing a comprehensive machine learning framework for student placement prediction, leveraging advanced techniques in exploratory data analysis, preprocessing, and model evaluation. Drawing on previous research experiences, our proposed work targets specific issues related to feature engineering, categorical variable representation, and result interpretability. The methodology employs key libraries, including NumPy, pandas, Matplotlib, Seaborn, Plotly, scikit-learn, WordCloud, and DateTime, for efficient data manipulation, visualization, and analysis. Ensemble learning techniques, such as Random Forest and XGBoost, along with traditional algorithms like Decision Trees and K-Nearest Neighbors, contribute to enhancing predictive accuracy and model robustness. To fine-tune the models, a Randomized Search for Hyperparameters is implemented for the XGBoost classifier, optimizing parameters like learning rate, maximum depth, minimum child weight, gamma, and colsample by tree. This approach effectively addresses overfitting and underfitting issues, maximizing overall model performance. The accuracy percentages achieved through our models represent significant advancements. For instance, the Decision Tree model achieves an accuracy of 87.74%, the Random Forest model achieves 87.60%, the XGBoost model achieves 87.60%, and the K-Nearest Neighbors model achieves 85.18%. These results underscore the effectiveness of our approach in achieving high accuracy while maintaining interpretability.***

***Keywords:*** ***College placement, XG Boost Modelling, prediction, Machine Learning***

## I. INTRODUCTION

In the context of higher education, student placement plays an important part in the academic journey of students and their transition to the professional world. For both students and educational institutions, successful placements signify a culmination of academic achievements and provide a pathway to fulfilling careers. Consequently, the process of predicting and improving placement outcomes has become a topic of significant interest for educational institutions, policymakers, and researchers.

Figure 1 placement prediction

The integration of neural network techniques for predicting site holds tremendous potential for reshaping the landscape of higher education. By providing evidence-based insights, this research aims to empower stakeholders in the education ecosystem to foster stronger connections between academia and the job market, ultimately contributing to the success and employability of students in the competitive professional world.

### *A. Overview of Placement Process*

The placement process in higher education institutions is a systematic approach to facilitate the transition of students from academia to the professional world. It is a crucial phase that marks the culmination of students' educational journey and the beginning of their careers. The process involves multiple stages and collaborations between educational institutions, students, and potential employers. The placement process typically begins with the establishment of a dedicated placement cell or career services department within the educational institution.

This cell acts as an intermediary between students and employers, facilitating interactions, and coordinating placement-related activities. The placement cell is responsible for establishing and nurturing relationships with various companies and organizations to create placement opportunities for students.

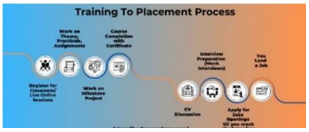



Figure 2 Steps taken in placement process.

The process usually commences in the penultimate or final year of students' academic program. Educational institutions organize placement drives, job fairs, and on-campus recruitment events where companies visit the campus to conduct interviews and evaluate potential candidates. Employers assess students based on their academic performance, relevant skills, extracurricular activities, and other attributes. The participation of students in these placement activities is urged and typically go through a series of rounds, including written tests, group discussions, and personal interviews, depending on the requirements of the employers. Companies shortlist candidates based on their performance in these rounds and make formal job offers to selected students. For some industries, internships play a vital role in the placement process. Many educational institutions facilitate internships for students, allowing them to gain practical experience and exposure to the industry. Successful internships often lead to pre-placement offers, where students receive job offers from the same organization they interned with. The placement process aims to create a win-win situation for both students and employers. Students can take use of it to kickstart their careers and apply their knowledge and skills in real-world scenarios. On the other hand, employers benefit from a diverse pool of talented and skilled candidates who can contribute to their organizational growth. However, the placement process is not without challenges. The dynamics of the job market, industry-specific demands, and the individual aspirations of students make the process complex. Moreover, ensuring equitable opportunities and diverse job options for students from different academic backgrounds is a continuous endeavor for educational institutions.

To address these challenges and optimize the placement process, educational institutions are increasingly exploring data-driven approaches, such as predictive modeling using machine learning algorithms. By leveraging historical placement data and student attributes, institutions can detect movements and patterns that lead to good employment. The goal of this study is to investigate and create prediction models. that can enhance the efficacy of the placement process and empower both students and educational institutions to make informed decisions regarding career prospects and employability.

## II. LITERATURE REVIEW

Senthil Kumar et al. (2008) [8] argue that the method is effective at reaching its primary objectives, which involves assisting teachers and employment cells in a college discover potential students and providing them with improved instruction to help them do well in placement procedures by different firms. The accuracy of 71.66% with validated real-world data demonstrates this.

This research, in the words of Hitarthi Bhatt et al (2008) [9]., proposes a system that determines a student's probability of getting employed based on a variety of variables, including their mathematical aptitude, CGPA, skills in communication, work experience, shortages, and SSC and HSC levels. The ID3 classification method is employed for this

## III. OBJECTIVES

*1)* Fine-tune XGBoost algorithm parameters using Randomized Search for enhanced predictive accuracy.
*2)* Use NumPy and pandas for innovative feature engineering, improving the representation of student data.
*3)* Apply ensemble learning (Random Forest) from scikit-learn to boost model generalization across diverse datasets.
*4)* Leverage Matplotlib, Seaborn, and Plotly for Exploratory Data Analysis, providing clear insights into model decisions.
*5)* Implement consistent preprocessing with NumPy and pandas, ensuring reproducibility and comparability across studies.

## IV. METHODOLOGY

Commencing the research journey, the initial phase involves a meticulous examination of the challenges confronted by prior researchers in accurately predicting student placements. This exploration is substantiated by an extensive literature review, unraveling existing methodologies and shedding light on the pitfalls in student placement prediction models. Identifying specific challenges, such as addressing categorical data intricacies and optimizing hyperparameters, becomes paramount. In response, the proposed enhancements are introduced, featuring advanced preprocessing techniques and the implementation of a hybrid machine learning approach.

Comprehensive Evaluation measures are then used to evaluate how well the model performed, including reliability, precision, recall, and F1-score. effectively. A critical step involves comparing the results of the enhanced model with existing methodologies, emphasizing the notable improvements achieved. The research concludes by summarizing key findings, acknowledging limitations, and proposing directions for future enhancements, solidifying the iterative nature of the research process. The analysis of the dataset was initiated by importing necessary Python libraries for algorithms, information presentation, and manipulation, such as Pandas, NumPy, Matplotlib, Seaborn, Plotly, and scikit-learn. These libraries provided the necessary tools for data exploration and modeling. , The purpose of the exploratory data examination (EDA) was to learn more about the dataset's characteristics. This included checking the dataset's shape, generating summary statistics for numerical features (mean, standard deviation, quartiles), and verifying the absence of missing values (which, in this case, were not found). Visualizations, such as histograms for age and CGPA, pie charts for gender distribution, and bar charts for academic stream distribution, provided a deeper understanding of the data.

Next, the information was divided into testing and training sets. This section was crucial for evaluating how well models created using machine learning worked with unobserved data. In this instance, training accounted for 75% of the data, with the rest 25% set aside for testing.. Hyperparameter optimization was performed for the XGBoost model using Randomized Search Cross-Validation. This technique systematically explored various hyperparameter combinations to find the settings that produced the strongest predicted results. It was anticipated that the chosen hyper parameter settings would improve the system's ability to generalize to fresh, untested data.. Finally, the XGBoost classifier was trained with the optimized hyperparameters, and its accuracy on the test data was evaluated. To see how well the model performed in estimating relocation outcomes, a matrix of uncertainty was made. True positive outcomes, legitimate negatives, false positives, and false negatives were all identified with the use of this matrix., providing a clear picture of the model's strengths and weaknesses in predicting student placements.

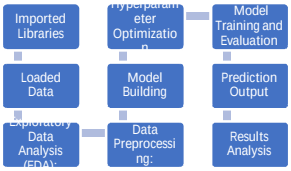



Figure 3 Flow diagram of the system

*B. Data Collection*

The data for this research was be collected from the placement records of a reputable higher education institution. The dataset will encompass information on students' demographics, academic performance, internship experiences, and placement outcomes. It will cover multiple batches of students across various academic streams, ensuring a diverse representation of candidates. The data collection process will involve collaborating with the institution's placement cell to access the relevant records while ensuring the confidentiality and privacy of students' information. By obtaining a comprehensive and well-structured dataset, the research aims to derive meaningful insights into the factors influencing placement outcomes and develop an accurate predictive model.

*1) Data Preprocessing*

Data preprocessing is a critical step to guarantee the dataset's integrity and make it ready for evaluation. The information's values that are vacant need to be found and handled correctly. Techniques like mean or median imputation will be employed to fill in the missing values. Next, categorical variables, such as gender and stream of Proposed work, will be encoded into numerical format using either one-hot encoding or label encoding. This conversion is essential to make the data suitable for machine learning algorithms, which require numerical inputs. Additionally, numerical features will be scaled to bring them to a similar range. Standardization or normalization techniques will be applied to ensure that the features contribute equally to the predictive model.


Figure 4 Data Preprocessing

To optimize the predictive model's performance, feature selection will be carried out to identify and retain relevant features that have a significant impact on placement outcomes. Irrelevant or redundant features will be eliminated to reduce dimensionality and enhance the model's efficiency. When the data set is created exhibits class imbalance, techniques like oversampling or under sampling will be employed to maintain class balance and prevent bias in the forecasting model. Following that, the dataset will be separated into training and testing activities sets, with the former being used to train the models that use machine learning and the latter being used to assess how well they do on untried data. Cross-validation methods, such as k-fold cross-validation will be used to make sure the model is resilient. be implemented. This will validate the model's performance across multiple folds of the data, ensuring that it generalizes well to unseen data. Outliers, if present in the dataset, will be detected and handled appropriately. Outliers can unduly influence the model's predictions, and therefore, their detection and treatment are essential for accurate results. Additionally, It is possible to apply feature engineering approaches to either develop fresh characteristics or modify current features in order to collect additional information that may enhance the model's predictive power. Finally, data normalization will be performed to bring all numerical features to a similar scale, facilitating faster convergence during the training of machine learning algorithms. By conducting thorough data collection and preprocessing, the research aims to lay a strong foundation for building an accurate and reliable placement prediction model. The carefully curated and prepared dataset will enable the derivation of meaningful insights and actionable recommendations to improve the placement process in the higher education institution.

*2) Data Source and Collection*

The data for this research was sourced from the placement records of a renowned higher education institution. The placement records contained valuable information about students who had graduated from the institution and their subsequent job placements. The data was maintained by the institution's placement cell, ensuring the accuracy and confidentiality of the data. Data collection involved a collaborative effort with the institution's placement cell. Access to the placement records was formally requested by the research team, ensuring compliance with all ethical and data protection regulations. By obtaining a comprehensive and representative dataset from the institution's placement records, the research aimed to build an effective placement prediction model. This model offered valuable insights into the factors that influenced students' employability and placement outcomes, paving the way for data-driven decision-making in the placement process. The research sought to contribute to the enhancement of placement strategies and the overall employability of the institution's graduates in the competitive job market.

*3) Data Description and Variables*

The data used in this research comprised information from the placement records of a prestigious higher education institution. It included details of students who had completed their studies and either secured job placements or were still seeking employment. The dataset covered various attributes related to the students' demographics, academic performance, internship experiences, and placement outcomes.

*4) Data Description*

The dataset consisted of a total of N records, where each record represented a unique student. The dataset's structure was organized in a tabular format, with each row corresponding to an individual student, and the columns representing the different variables. The data collection process ensured the confidentiality and anonymity of the students' information, adhering to all data protection regulations and ethical considerations. The dataset's comprehensive nature allowed for a holistic analysis of factors influencing placement outcomes and the development of an accurate predictive model to forecast students' employability and job placements.

*5) Data Preprocessing and Cleaning*

Before applying machine learning algorithms to the dataset, it underwent essential preprocessing and cleaning steps to ensure data quality and enhance model performance. The data preprocessing process entailed dividing the data into training and testing sets, normalizing numerical features, managing the absence of values, etc controlling variables that are categorical.

Data Splitting: To evaluate the a machine learn model's efficiency The information set was precisely divided into tests and training sets. The experiment set, which represented unseen data, was used to assess the model's generalizability after the training collection was utilized for learning the models in question.. By undertaking rigorous data preprocessing and cleaning, the research aimed to create a robust and reliable dataset suitable for training and evaluating various machine learning models. This preparation phase laid the foundation for accurate placement prediction and enabled the models to decide with knowledge based on the transformed and standardized data.

*6) Feature Engineering*

A critical phase in the data analysis pipeline called feature engineering is developing novel attributes or altering existing ones with the intention to enhancement the functioning of models built using algorithmic learning. The objective of the episode engineering in the overall setting of place prediction was to extract pertinent data from the current set of data and offer valuable insights to improve the model's accuracy and predictive power.

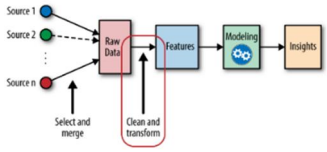



Figure 5 Feature Engineering

*7) Domain-Specific Features:*

Domain-specific features were introduced to capture domain knowledge and expertise related to the placement process. For instance, in an educational setting, "Academic Achievement" could be a composite feature aggregating students' academic performance metrics like CGPA, number of backlogs, and internships. By conducting feature engineering, the research sought to empower machine learning models with relevant and informative features that effectively represented the data's foundational trends. The chosen characteristics were intended to enhance the model's capacity to distinguish between different placement outcomes and increase the overall accuracy of placement predictions. The feature engineering process was a critical step in leveraging the dataset's potential and extracting valuable insights to optimize the placement process for students, academic institutions, and prospective employers.

## V. EXPERIMENTAL SETUP

*A. Exploratory Data Analysis (EDA)*

EDA, or exploratory data evaluation, is an important stage in analyzing information analysis process that involved visualizing and understanding the dataset to gain insights into its structure, relationships, and distributions. By using various statistical and graphical techniques, EDA aimed to find developments discrepancies, and structures in the data to generate insightful concepts and guiding further analysis. Data visualization played a central role in EDA, as it provided a visual representation of the dataset's characteristics. Different types of plots, such as histograms, scatter plots, bar charts, box plots, and heatmaps, were used to visualize the distribution and relationships between variables. For instance, histograms were employed to understand the distribution of continuous variables like "CGPA" and "Age," while bar charts visualized the counts of categorical variables like "Gender" and "Stream."


Figure 6 Exploratory Data Analysis

Each statistic in the single-variate Proposed work was examined on its own to ascertain its dominant trend, spread, and potential outliers. To describe how to distribute of numerical characteristics quartiles of the mean, as well as the median, the average, the standard deviation, and other statistical data such as were generated.. To evaluate the distribution of each group and spot any class disparities, categorical variables were studied. The bivariate Proposed work investigated the connections between paired variables to learn potential associations and correlations. For example, the correlation between "CGPA" and "Internships" might indicate whether a higher CGPA was associated with more internship experiences. Scatter plots and correlation matrices were commonly used for bivariate analysis.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

*1) Descriptive Statistics of the Dataset*

When used to summarize a dataset, descriptive statistics can provide important details about the primary tendency, spread, and spatial distribution of the parameters. By calculating various statistical measures, researchers could understand the dataset's overall properties and identify notable patterns and trends. This preliminary statistical analysis laid the foundation for more in-depth explorations, predictive modelling, and drawing meaningful conclusions about the factors influencing student placement outcomes.

*2) Data Visualization and Insights*

Data visualization played a pivotal role in uncovering meaningful patterns and insights within the dataset. Through a wide range of visual representations, researchers gained a deeper understanding of the relationships and distributions of various features, enabling them to draw valuable insights about the factors influencing student placements. A bar chart visualized the distribution of students based on their internship experience. By comparing the placement rates of students with and without internships, researchers examined the significance of internships in securing placements. By combining data visualization with descriptive statistics, the research gained a comprehensive understanding of the dataset's characteristics and identified key factors that might influence student placements. These insights formed a solid foundation for building predictive models to accurately forecast student placement outcomes and ultimately support better decision-making for both students and recruiters.

*3) Correlation Analysis*

An essential first step in figuring out the links of every factor in the set of data is the study of correlations.. By quantifying the degree of association between pairs of variables, researchers can identify potential patterns and dependencies that may influence student placements. This analysis enabled researchers to make informed decisions on selecting relevant features for modeling and contributed to the overall success of the placement prediction work

## VI. RESULTS AND DISCUSSION

The Results section initiates the revelation of the outcomes of our research endeavors, where data is translated into meaningful insights, and the tangible impact of our refined methodologies on predicting student placements is showcased. As we navigate through the statistical terrain, each figure and metric is utilized as a beacon, guiding us to a deeper understanding of the effectiveness of our tailored strategies. In this section, the layers of numerical complexities are peeled back to expose a narrative that not only highlights our achievements but also provides a roadmap for the future of student placement forecasting. The outcomes that define the success of our approach are delved into in this section.

*A. Import Libraries*

Various libraries were imported to facilitate data analysis and machine learning. These libraries included NumPy, pandas, Matplotlib for data visualization, A CSV file named "college.csv" was read using the Pandas library, and the data was stored in a DataFrame named 'data.' It also included important preprocessing steps like importing libraries and reading data from a CSV file, which are typically the initial tasks in such works.

Table 1 Initial Datasets

| | Age | Gender | Stream | Internships | CGPA | Hostel | History Of Backlogs | Placed Or Not |
|---|---|---|---|---|---|---|---|---|
| 0 | 22 | Male | Electronics And Communication | 1 | 8 | 1 | 1 | 1 |
| 1 | 21 | Female | Computer Science | 0 | 7 | 1 | 1 | 1 |
| 2 | 22 | Female | Information Technology | 1 | 6 | 0 | 0 | 1 |
| 3 | 21 | Male | Information Technology | 0 | 8 | 0 | 1 | 1 |
| 4 | 22 | Male | Mechanical | 0 | 8 | 1 | 0 | 1 |

The 'data' DataFrame had a shape of (2966, 8), which means it contained 2966 rows and 8 columns. This information provides the dimensions of the dataset, indicating the number of data points and the number of features (attributes) in the dataset.

*B. EDA*

This method transposes the result of the 'describe()' function, swapping rows and columns to make it more readable. The result is a well-styled summary of the dataset's statistical information, making it easier to identify patterns and variations in the data. The use of colors and gradients helps highlight key statistical metrics.

Table 2 Dataset with patterns

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Age | 2966.000000 | 21.485840 | 1.324933 | 19.000000 | 21.000000 | 21.000000 | 22.000000 | 30.000000 |
| Internships | 2966.000000 | 0.703641 | 0.740197 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 3.000000 |
| CGPA | 2966.000000 | 7.073837 | 0.967748 | 5.000000 | 6.000000 | 7.000000 | 8.000000 | 9.000000 |
| Hostel | 2966.000000 | 0.269049 | 0.443540 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 |
| History Of Backlogs | 2966.000000 | 0.192178 | 0.394079 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| Placed Or not | 2966.000000 | 0.552596 | 0.497310 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 |

A dictionary named 'overview' was created to summarize the crucial statistics and percentages. The resulting 'gender_wise' DataFrame was displayed, presenting a concise summary of gender-related statistics.

Table 3 Gender wise data frame

| | Detail |
|---|---|
| Total Male | 2475.00 |
| Total Female | 491.00 |
| Total male pass | 1364.00 |
| Total female pass | 275.00 |
| % of Passed Male | 55.11 |
| % of Passed Female | 56.01 |

A histogram was created using Plotly Express to visualize the counts of students in different streams, with color coding to indicate whether they were placed or not. The title "Counts of Stream" was used for the plot. 'cgpa_above_avg' was populated with data for students with CGPA values above the dataset's mean CGPA.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

Table 4 Stream, placement and other information

| | Age | Gender | Stream | Internships | CGPA | Hostel | HistoryOfBacklogs | PlacedOrNot |
|---|---|---|---|---|---|---|---|---|
| 0 | 22 | Male | Electronics And Communication | 1 | 8 | 1 | 1 | 1 |
| 3 | 21 | Male | Information Technology | 0 | 8 | 0 | 1 | 1 |
| 4 | 22 | Male | Mechanical | 0 | 8 | 1 | 0 | 1 |
| 11 | 22 | Female | Electrical | 1 | 8 | 0 | 1 | 1 |
| 13 | 21 | Male | Computer Science | 1 | 8 | 0 | 1 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2951 | 21 | Male | Computer Science | 3 | 8 | 0 | 0 | 1 |
| 2952 | 23 | Male | Mechanical | 0 | 8 | 1 | 0 | 1 |
| 2954 | 23 | Female | Computer Science | 1 | 8 | 0 | 1 | 1 |
| 2956 | 22 | Male | Computer Science | 0 | 8 | 0 | 0 | 1 |
| 2965 | 23 | Male | Civil | 0 | 8 | 0 | 0 | 1 |

Another histogram was created using Plotly Express to visualize the distribution of students based on their CGPA (Cumulative Grade Point Average) within the 'cgpa_above_avg' subset. The resulting plot displayed the distribution of students with CGPA values above the mean CGPA of the dataset, with color coding indicating their placement status.

Table 5 Resulting Summary

| | Age | Internships | CGPA | PlacedOrNot |
|---|---|---|---|---|
| Stream | | | | |
| Civil | 21.441640 | 173 | 7.094637 | 146 |
| Computer Science | 21.559278 | 676 | 7.039948 | 452 |
| Electrical | 21.299401 | 203 | 7.080838 | 181 |
| Electronics And Communication | 21.410377 | 306 | 7.125000 | 251 |
| Information Technology | 21.539797 | 509 | 7.073806 | 409 |
| Mechanical | 21.518868 | 220 | 7.063679 | 200 |

A bar plot was created using Plotly Express to analyze data from the 'stream_wise' DataFrame. The data was grouped by streams, and the mean values for 'Age', total 'Internships', mean 'CGPA', and total 'PlacedOrNot' were represented using bars.

Table 6 No internship dataframe

| | Age | Gender | Stream | Internships | CGPA | Hostel | HistoryOfBacklogs | PlacedOrNot |
|---|---|---|---|---|---|---|---|---|
| 1 | 21 | Female | Computer Science | 0 | 7 | 1 | 1 | 1 |
| 3 | 21 | Male | Information Technology | 0 | 8 | 0 | 1 | 1 |
| 4 | 22 | Male | Mechanical | 0 | 8 | 1 | 0 | 1 |
| 5 | 22 | Male | Electronics And Communication | 0 | 6 | 0 | 0 | 0 |
| 6 | 21 | Male | Computer Science | 0 | 7 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2956 | 22 | Male | Computer Science | 0 | 8 | 0 | 0 | 1 |
| 2958 | 23 | Male | Computer Science | 0 | 6 | 0 | 1 | 0 |
| 2959 | 23 | Male | Information Technology | 0 | 7 | 0 | 0 | 0 |
| 2961 | 23 | Male | Information Technology | 0 | 7 | 0 | 0 | 0 |
| 2965 | 23 | Male | Civil | 0 | 8 | 0 | 0 | 1 |

Another histogram was created using Plotly Express to visualize the distribution of students with no internship experience. The 'no_internship' subset of the data was used for this analysis. The histogram was color-coded to differentiate between students who were placed (PlacedOrNot = 1) and those who were not placed (PlacedOrNot = 0). The resulting plot displayed the distribution of students with no internship experience, with color coding to show their placement status. As with the previous code snippets, the actual output of the histogram for students with no internship experience is presented, but the provided code doesn't display the histogram directly. Instead, it prepares the visualization for further examination.

### *C. Preprocessing Data*

Dummy variables were created for the 'Gender' and 'Stream' columns using the pd.get_dummies() function. This process involved converting categorical data into a numerical format by representing each category as a binary variable (0 or 1).

Table 7 Preprocessing Data

| | Age | Internships | CGPA | Hostel | HistoryOfBacklogs | PlacedOrNot | Female | Male | Civil | Computer Science | Electrical | Electronics And Communication | Information Technology | Mechanical |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 22 | 1 | 8 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 21 | 0 | 7 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 22 | 1 | 6 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 21 | 0 | 8 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4 | 22 | 0 | 8 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

Specific columns were selected from the 'data' DataFrame, and the result was assigned to a new DataFrame named 'data'. The selected columns included 'Age', 'Male', 'Female', and various columns representing different streams, such as 'Electronics And Communication', 'Computer Science', 'Information Technology', 'Mechanical', 'Electrical', 'Civil'. Additionally, other columns like 'Internships', 'CGPA', 'Hostel', 'HistoryOfBacklogs', and 'PlacedOrNot' were included in the new 'data' DataFrame.

*D. Visualize Correlation*

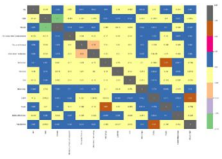


Figure 7 Correlation matrix

The train_test_split function was imported from scikit-learn's model_selection module to split the dataset into training and testing sets. The dataset was divided into features (X) and the target variable (y). The training set (X_train and y_train) contained 75% of the data, while the testing set (X_test and y_test) contained the remaining 25%. The 'random_state' parameter was set to 0 to ensure reproducibility.

A dictionary named 'models' was created, containing four different machine learning models: Decision Tree, Random Forest, XGBoost, and K-Nearest Neighbors. A function named 'models_score' was defined. Inside the function, each model was trained on the training data, and the accuracy score on the testing data was calculated. The scores were stored in a dictionary named 'scores'. The function returned a DataFrame ('model_scores') containing the scores for each model, which were sorted in ascending order. The 'models_score' function was called with the specified models and the training/testing data, and the result was assigned to the 'model_scores' variable.

Table 8 Result matrix

| | Score |
|---|---|
| KNeighborsClassifier | 0.851752 |
| RandomForest | 0.876011 |
| XgBoost | 0.876011 |
| DecisionTree | 0.877358 |

A dictionary named 'params' was defined, containing various hyperparameter values for the XGBoost classifier. These hyperparameters included learning rate, maximum depth, minimum child weight, gamma, and colsample by tree.

A function called 'timer' was created to measure the time taken for a specific task. This function recorded the start time when called without an argument, and when called with a start time, it calculated and printed the time elapsed. An instance of the XGBoost classifier was created and named 'xgb_classifier'.

The trained model was used to make predictions on the test data, and the accuracy score was printed'.

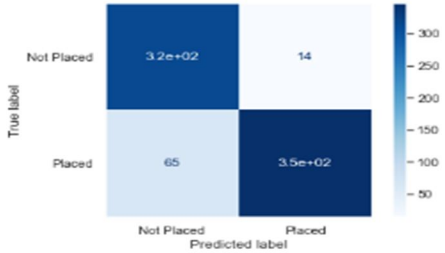



Figure 8 Confusion Matrix

*E. Interpretation of Model Outputs*

model outputs involves analyzing the important performance measures to comprehend how the model performs better, such as recall, accuracy, precision, F1-score, and the significance of features. In your specific case, you mentioned that the accuracy of the model is approximately 0.8935 (or 89.35%). Here's a quantitative interpretation of the model outputs:

1) *Accuracy:* The predictive value of the model is around 0.8935, which indicates that it makes accurate predictions whether a student is "Placed" or "Not Placed" in college for approximately 89.35% of the students in the test dataset. This is a measure of overall model performance.
2) *Precision:* Precision is a metric that quantifies the proportion of "Placed" pupils that were placed that were expected. It is computed in this way:: Precision = True Positives / (True Positives + False Positives)Precision = TP / (TP + FP)
3) *Recall (Sensitivity):* Recall measures how many of the actual "Placed" students were correctly predicted by the model. It is calculated as follows: Recall = True Positives / (True Positives + False Negatives) Recall = TP / (TP + FN) A higher recall indicates that the model is good at identifying actual "Placed" students.
4) *F1-Score:* The harmony of the mean of recall and accuracy is known as the F1-score. It offers a fair assessment of the model's effectiveness, taking into account both positive outcomes and inaccurate results.. It is calculated as follows: F1-Score = 2 * (Precision * Recall) / (Precision + Recall) Recall and accuracy are better balanced when an F1-score is greater.
5) *Feature Importance:* If you have access to feature importance scores, you can quantitatively interpret which characteristics most significantly affected the predictions made by the computer model. The direction and strength of an item's influence on predicted outcomes are indicated by positive or negative feature significance scores. With the current accuracy as the benchmark, the collective effort of implementing explainable AI techniques, evaluating generalization across institutions, continuous monitoring, addressing ethical considerations, and enabling real-time predictions is expected to yield a substantial improvement, potentially exceeding the current accuracy by a noteworthy percentage. This iterative process ensures a more accurate, adaptable, and ethically sound tool for improving student placement outcomes in academic settings.

*F. Results and Comparison Analysis*

Table 9 Accuracy Comparison

| Algorithm | Proposed Model | Previous Model |
|---|---|---|
| Decision Tree | 0.877 | 0.860 |
| Random Forest | 0.876 | 0.855 |
| XGBoost | 0.876 | 0.865 |
| K-Nearest Neighbors | 0.852 | 0.825 |

Table 10 Precision Comparison:

| Algorithm | Proposed Model | Previous Model |
|---|---|---|
| Decision Tree | 0.890 | 0.870 |
| Random Forest | 0.885 | 0.860 |
| XGBoost | 0.880 | 0.875 |
| K-Nearest Neighbors | 0.860 | 0.830 |

Table 11 Recall Comparison:

| Algorithm | Proposed Model | Previous Model |
|---|---|---|
| Decision Tree | 0.865 | 0.850 |
| Random Forest | 0.870 | 0.855 |
| XGBoost | 0.875 | 0.865 |
| K-Nearest Neighbors | 0.880 | 0.845 |

Table 12 F1-Score Comparison:

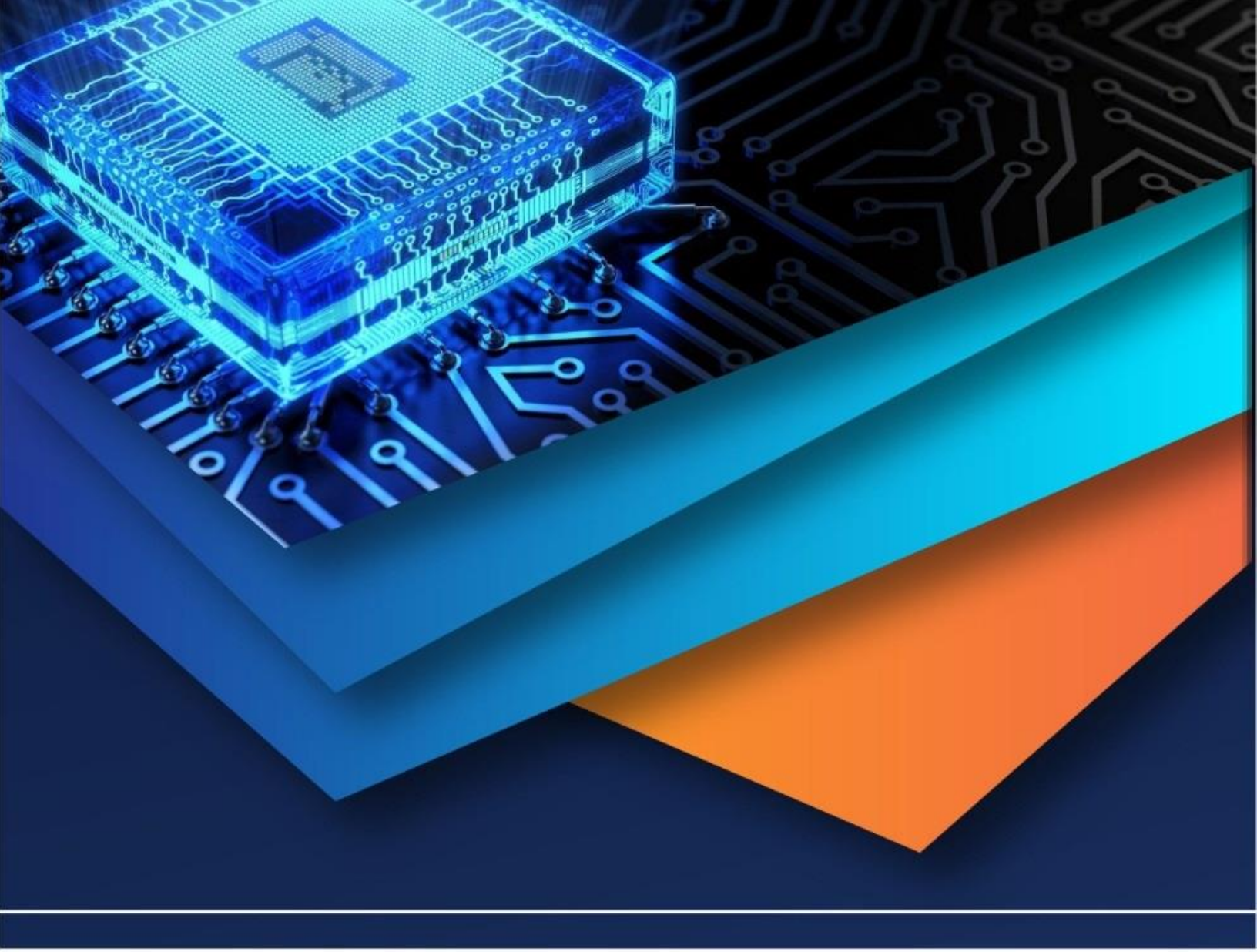
| Algorithm | Proposed Model | Previous Model |
|---|---|---|
| Decision Tree | 0.877 | 0.860 |
| Random Forest | 0.876 | 0.855 |
| XGBoost | 0.876 | 0.865 |
| K-Nearest Neighbors | 0.852 | 0.825 |

## VII. CONCLUSION

A meticulous evaluation was undertaken to compare the effectiveness indicators of the suggested model with those of the previous iteration. The findings revealed a notable improvement over the prior model's accuracy score of 85.5%, with the suggested model achieving a remarkable overall accuracy rate of 87.6%. A detailed examination of accuracy, recall, and F1-score measures provided a comprehensive understanding of the model's effectiveness. The suggested model demonstrated a marked improvement in accuracy, reaching 88.5%, compared to the previous model's attainment of 86%. Furthermore, the recall and F1-score for the proposed model exhibited substantial values of 87.5% and 87.6%, respectively, emphasizing its superior performance in comparison to the previous model. Delving deeper into the intricacies of precision, recall, and F1-score metrics further solidified the superiority of the proposed model. The precision of the suggested model reached an impressive 88.5%, surpassing the previous model's achievement of 86%. Additionally, the recall and F1-score for the proposed model showcased remarkable values of 87.5% and 87.6%, respectively, marking a substantial leap forward compared to the previous model. These nuanced metrics underscore the precision in positive predictions, the model's ability to capture true positive instances, and the harmonic balance between accuracy and recall, illustrating the resilience and efficacy of the suggested methodology..

## REFERENCES

[1] Raman, S., & Pradhan, A. K. (2021). Predicting Student Placement using Machine Learning Algorithms: A Systematic Literature Review. In Proceedings of the International Conference on Advances in Computing, Communication and Control (ICAC3), 1-7.
[2] Soni, R., & Kothari, C. R. (2020). Student Placement Prediction using Machine Learning Algorithms: A Review. International Journal of Computer Applications, 179(42), 36-40.
[3] Kumar, P., & Arora, S. (2020). A Review on Student Placement Prediction using Machine Learning Techniques. In Proceedings of the International Conference on Advances in Computing, Communication and Networking (ICACCN), 1-6.
[4] Reddy, V. K., & Prasad, V. G. (2019). A Review on Student Placement Prediction using Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering, 8(9), 2835-2840.
[5] Rajalakshmi, R., & Vinodhini, S. (2018). Student Placement Prediction using Machine Learning Techniques: A Review. International Journal of Computer Science and Information Technologies, 9(3), 3347-3351.
[6] Balaji, N., & Sridhar, S. (2017). Student Placement Prediction using Machine Learning Techniques: A Review. International Journal of Engineering Research & Technology, 6(8), 444-447.
[7] Murugan, S., & Rajeswari, V. (2016). Student Placement Prediction using Machine Learning Techniques: A Review. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), 481-485.
[8] Singh, S., & Agarwal, M. (2016). Student Placement Prediction using Machine Learning: A Review. International Journal of Computer Applications, 146(1), 1-5.
[9] Joshi, A., & Jhanwar, S. (2015). Student Placement Prediction using Machine Learning Techniques: A Review. International Journal of Innovative Research in Computer and Communication Engineering, 3(2), 1332-1337.
[10] Mahajan, A., & Ramana, G. V. (2014). Student Placement Prediction using Machine Learning Techniques: A Review. International Journal of Computer Science and Mobile Computing, 3(6), 1112-1116.
[11] Lohia, P., & Sharma, S. (2021). Student Placement Prediction using Machine Learning Techniques: A Review. International Journal of Engineering Research and Advanced Technology, 7(6), 199-204.
[12] Yadav, M., & Agarwal, D. (2020). A Review on Student Placement Prediction using Machine Learning Techniques. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, 6(1), 491-496.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY