

# Student Placement Prediction and Analysis using Machine Learning

Pratik Pandey<sup>1</sup> Nikhil Khonde<sup>2</sup> Omkar Agre<sup>3</sup> Prof. Sonali Guhe<sup>4</sup>

<sup>1,2,3,4</sup>Department of Information technology  
<sup>1,2,3,4</sup>GHRCE, Nagpur, India

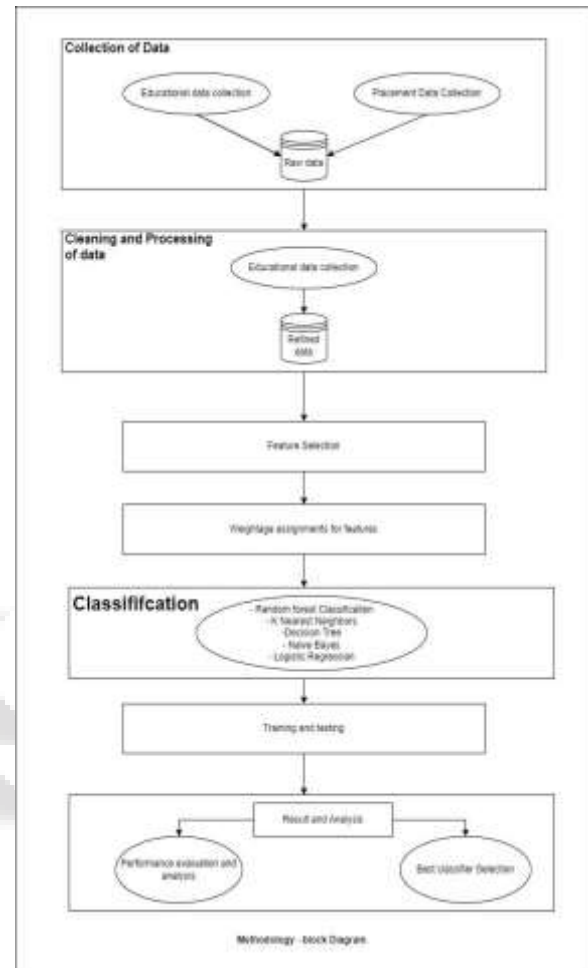
**Abstract** — Almost every student wishes to grab placement opportunity before completing their degree. There are various developing and developed placement opportunities in the market. However, a student must carefully choose his field and skills to satisfy the requirements set by employer. A placement prediction model helps students predict the probability of him/her being placed or not based on his academic and personal achievements. The K-Nearest Neighbors [KNN] algorithm, Decision tree, Logistic regression, Naïve Bayes, and Random Forest are the five distinct machine learning classification techniques that were considered for this project. These algorithms each independently forecast the outcomes, and we compare their efficacy based on the dataset. This prediction model can forecast the likelihood of the student being placed based on his qualification and work experience. Such prediction models could aid in a student's or an institution's future academic planning.

**Keywords:** Decision tree, Classification, Logistic regression, Placement Prediction, Random Forest, KNearest Neighbors [KNN], Naïve Bayes.

## I. INTRODUCTION

The placement of students into appropriate companies or startups is a critical goal for educators and educational institutions. Accurately predicting which course or programs a student is best suited for can help ensure that they receive the education and resources they need to succeed. In recent years, advances in data analytics and machine learning have made it possible to analyze student data and make more informed predictions about their academic performance and placement. Through a comprehensive review of analysis and processing of previous year dataset of students, this paper seeks to forecast if student will be placed or not. Different attributes like number of backlogs, CGPA, number of internship, and stream were taken into consideration to identify a pattern in placement criteria. Different classification algorithm will be used to process the data and after comparing the result of these algorithm, the algorithm with the best accuracy will be chosen to predict the result.

## II. PROPOSED METHODOLOGY:



### A. Collection of data

Collecting data is the foremost and one of the most important step in this module. For this student record from various sources were taken into account like educational institutions, employment data and surveys with student and employers. The dataset taken into account consist of parameters like gender, stream, CGPA, number of internship, number of backlogs and hostel accommodation.

### B. Cleaning and pre-processing

In this phase the obtained data will be carefully assessed for any missing values and it will cleaned or filled. This is done to transform and improve the dataset before using it, to make it compatible to further processing.

### C. Feature Selection

There can be number of elements in the dataset of students, however not all of these are required. For example, in our model the field hostel accommodation and gender are dropped as it has no relation to placement. This is done to

reduce the dimensionality of the dataset to make it easier to evaluate.

#### D. Weightage assignment for features

It is evident that not all features have to be taken into account to predict the placement. Also it is obvious that student can't excel in all the fields ,thus appropriate weightage was assigned for each field .CGPA , number of backlogs, and number of internship were assigned specific weightage to predict the results.

#### E. Classification Methods

Various types of algorithm were considered for this module, to compare and find out the best algorithm that would provide the maximum accuracy in forecasting the result. The algorithm used are:

##### 1) KNN algorithm:

In order to forecast the label of an input data point based on the majority label of its K neighbors, KNN first locates the K data points that are the closest to an input data point. The Euclidean distance formula is used to determine the distance between data points.

##### 2) Logistic Regression:

Logistic regression is a statistical method for examining the relationship between a dependent variable and one or more independent factors. It is commonly utilized when there is a binary dependent variable and the problem involves binary classification. Logistic regression determines the probability that the dependent variable will be in one of two possible states based on the values of the independent variables.

##### 3) Naïve Bayes:

The name "naive" comes from the fact that Naive Bayes makes the assumption that the features or attributes of the input data are independent of one another, which is frequently untrue in reality. Despite this oversimplifying presumption, Naive Bayes can frequently produce classification results with high precision, especially when the dataset is sizable and the features are largely independent.

##### 4) Decision tree:

Decision trees work by recursively splitting the dataset into subsets based on the most informative features, in order to create a tree-like structure that can be used to make predictions on new data. The decision is based on the value of a certain feature at each internal node of the tree, and the algorithm divides the data until a stopping requirement is satisfied.

##### 5) Random Forest algorithm:

Random forest algorithm is an ensemble learning technique that merges multiple decision trees and creates a forest of trees. The final forecast is based on the consensus of all the trees in the forest, and each tree is constructed using a random subset of features and data points.

#### F. Training and Testing:

Firstly, the data set and the object of the data set were preprocessed and label encoded by using the module preprocessing of sklearn library. After that, the dataset was divided in the proportion of 80:20 for training and testing phase with the help of train\_test\_split function. It was done using the sklearn.model\_selection.

#### G. Performance Evaluation and Analysis:

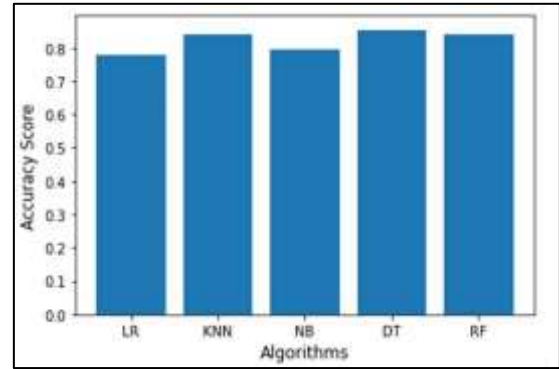


Fig. 1: Accuracy rate of algorithms

The K-Nearest Neighbors [KNN] algorithm, Decision tree, Logistic regression, Naïve Bayes, and Random Forest are the five distinct machine learning classification techniques that were used and later accuracy was calculated of every algorithm using confusion matrix. After comparing it was found that decision tree and random forest has nearly about same accuracy, however, decision tree was facing problem in overfitting the training data thus, making it too complex and losing its ability to generalize new unseen data. Finally, random forest algorithm was chosen over decision tree to predict the result.

### III. RESULT:

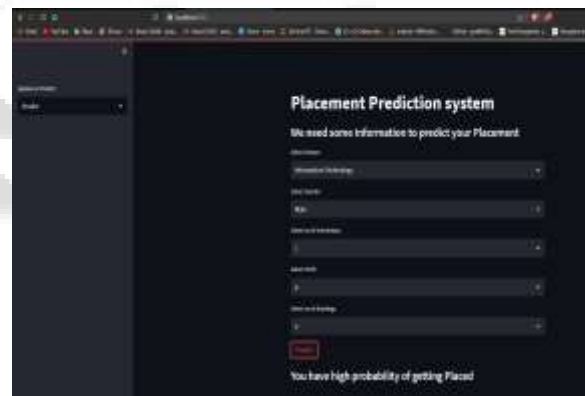


Fig. 2: Prediction output

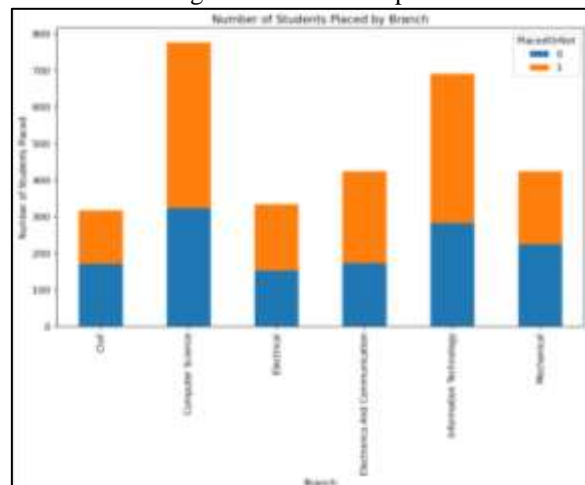


Fig. 3: Visualization output

#### IV. CONCLUSION:

In conclusion, the study and prediction of student placement can have a major effect on students' academic progress. After a thorough study of various algorithms, it was found that random forest has the most efficacy. Thus, it can be concluded that past dataset of student in an institution can be processed and used to predict the results. This model can help students recognize their probability of being placed and make the necessary improvements. Educators and administrators can learn more about a student's likelihood of succeeding in a particular academic setting by looking at numerous aspects such as past academic performance, internship status, and demographic data.

#### V. FUTURE SCOPE:

By utilising cutting-edge machine learning methods like neural networks and support vector machines, we can further enhance this model's performance and potentially increase prediction accuracy by more accurately spotting trends in the data set. We could also expand our dataset by collecting more detailed information on student backgrounds and experiences (such as student motivation, self-esteem, or socio-economic status). Additionally, we can foresee which business will choose which group of students. Make a list of the skills that a specific business is seeking, and then we can teach our student based on that. These characteristics will improve forecast accuracy.

#### REFERENCE

- [1] Mrs. J. Samatha D. Manjusha, B. Pooja, A. Usha: STUDENT PLACEMENT CHANCE PREDICTION. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2020
- [2] Naresh Patel K M Goutham N M Inzamam K A Suraksha V Kandi Vineet Sharan V R: Placement Prediction and Analysis using Machine Learning. *International Journal of Engineering Research & Technology (IJERT)*, 2022
- [3] Pratiksha Khamkar, Rutuja Lagad, Priyanka Shinde, Shubhangi Londhe, Prof. S.S. Bhosle: Students Placement Prediction System. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2022
- [4] Chinmay Deepak Chaugule, Kunal Prabhakar Temkar, Siddhant Sakharam Shinde, Neha Rupesh Thakur: Placement Prediction by Mining Student's Information.
- [5] Abhishek S. Rao, Aruna Kumar S V, Pranav Jogi, Chinthan Bhat K, Kuladeep Kumar B, Prashanth Gouda: Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System. *International Journal of Recent Technology and Engineering (IJRTE)*, 2019.
- [6] H. Sabnani, M. More, P. Kudale, S. Janrao, "Prediction of Student Enrolment Using Data Mining Techniques", *International Research Journal of Engineering and Technology (IRJET)*, 5(4), 1830-1833, 2018.
- [7] Thangavel, S. Bkaratki, P. Sankar, "Student Placement analyzer: A recommendation system Using machine learning", *Advances in Computing and Communication Systems (ICACCS-2017) International Conference on IEEE*, 2017.
- [8] Liu, Yang, et al. "The Application of Machine learning Techniques in College Students Information System." 2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018). Atlantis Press, 2018.
- [9] Sagardeep Roy, Anchal Garg, "Analyzing Performance of Students by Using Data Mining Techniques" (2017) 4<sup>th</sup> IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics
- [10] Manikandan, K., Sivakumar, S., Ashokvel, M. "A Classification Model for Predicting Campus Placement performance Class using Data Mining Technique" *International Journal of Advance Research in Science and Engineering* 7(6) 2018: 29-38.