Student's Placement Prediction using Machine Learning

Shamjad Mazood Nazer
Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirappally, Kerala
shamjadmazoodnazer2023b@mca.ajce.in

Mr. Rony Tom
Department of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, Kerala
ronytom@amaljyothi.ac.in

***Abstract*— *At academic institutions, student placement is critical. It is the decisive element in admission and the reputation of the school. To strengthen their placement department, all organizations attempt to improve their methods. In this essay, we hope to assess past student data, forecast job prospects for current students, and offer assistance in improving the institutions' placement rates. In this research, we present a method for predicting a student's placement status. We also anticipate the business to which the student will be sent based on data from previously placed students. Our method entails implementing two independent machine learning classification algorithms: the Random Forest Regressor and the K Nearest Neighbors [KNN] algorithm. We assess the performance of both algorithms by comparing their forecasting outcomes on the dataset. Using this model, a company's position cell could identify promising students and focus on developing both their technical and soft skills.*

Keywords— Data science, Classification Techniques, Machine learning, Random Forest Regressor model, K Nearest Neighbor.

I. INTRODUCTION

The Training and Placement activity at college is an important component of a student's academic life. Consequently, it is crucial to have a streamlined procedure so that students can access the needed data whenever they do. A good system would allow the Training and Placement cell personnel to inform students promptly and efficiently, decreasing their burden. The "Student's Placement Prediction using Machine Learning" solution is made to deal with the issues with manual systems. This program tries to minimize or lessen the difficulties encountered by the current system and make company operations more effective and efficient.

Organisations must manage information on student training, placement, and performance. Our system is designed to satisfy individual training on the institute, and our remote access feature allows you to manage students from anywhere and at any time. Predicting final-year students' placement status will encourage them to study more and make adequate progress. It will also aid teachers and placement cells in providing enough attention to students' progress throughout the course. Building an educational institution's reputation necessitates a high placement rate. As a result, this system is critical to the instructional programme at every higher education institution.

The Random Forest Regression technique is an ensemble learning technique used in the regression area. The process entails combining several decision trees to build a forest, with each individual tree suited to a random partition of both the data and the training set's attributes. To create a forecast, the unique data point is submitted to each individual tree inside the forest, and the aggregate prediction of all trees is generated as the final output. This method addresses the issue of overfitting while also improving model accuracy. The optimisation of performance necessitates the thorough tweaking of hyperparameters, which are critical in defining the number of trees in forests and the size of subsets used for training.

The Random Forest Regression approach is commonly used to solve regression issues like estimating student performance or placement. Its utility stems from its capacity to manage non-linear interdependencies between input features and output variables, adjust for missing data, and lessen the influence of irrelevant or noisy information.

The Random Forest algorithm has the advantage of providing accurate forecasts for student performance and placement based on their training test and past academic success on final examinations. This method can handle complicated data relationships as well as missing data. It can also give feature significance rankings, which can aid in identifying the most influential aspects influencing student performance and placement.

II. LITERATURE REVIEW

A. Research Paper 01

Rathi Viram, Swati Sinha, Bhagyashree Tayde, and Aakshada Shinde conducted a study on placement prediction system using machine learning [1]. The authors utilised Google forms to obtain information from students. They employed a variety of machine learning algorithms to train and assess their prediction model, including Regression and Support Vector Machine. This approach will show pupils how to overcome their weaknesses. This technique will also benefit college growth since students will generate new project skills and the total percentage of placement will rise.

B. Research Paper 02

Rosemary Vargheese, Adlene Peraira, Aswathy Ashok and Bassant Johnson proposed a system model for predicting students' performance using knowledge mining technique which is classification [2]. The authors collected data from 114 students' records and analysed it using machine learning techniques to calculate a student's performance in future semesters. Their new model, Support Vector Machine, outperformed previous machine learning models such as Decision Tree, Naive Bayes, and Random Forest, with the maximum accuracy of 81.82%.

C. Research Paper 03

Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, and Tojo Mathew developed students' placement prediction using machine learning tools [3]. They predicted using two machine learning models, Naive Bayes and K Nearest Neighbour, on data from the prior year. They increased the number of parameters in their system to improve its accuracy.

D. Research Paper 04

Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju Joel James, and Mereen Thomas Vadakkel created placement prediction using various machine learning models and their efficiency comparison for make prediction that a student will get placed or not [4]. The authors trained machine learning models such as random forest, logical regression, KNN, and SVM on datasets with multiple parameters. The SVM model was the best of the bunch, with 100% accuracy.

E. Research Paper 05

Pothuganti Manvitha and Neelam Swaroopa conducted a study on campus placement prediction using supervised machine learning technique [5]. The scientists collected data from the US Department of Transportation and applied numerous machine learning models, including linear regression, decision trees, random forests, and artificial neural networks. The random forest model fared better than the other models, obtaining an accuracy of 92.68%.

III. MOTIVATION

According to estimates, 1.5 million engineers graduate each year in India. The need for competent graduates in any business is increasing by the day. One of the most difficult problems that students confront is campus placement. Providing maximal placement drives is one of an institution's responsibilities, as is using the drives by the student. Today, the majority of companies choose students based on an aptitude round. Each student performed differently in the aptitude round.

So, one thing we can do is identify students who have low aptitude performance on the quizzes administered within our system and assist them in passing the aptitude test.

IV. METHODOLOGY

The goal of this project is to create a machine learning model using the available dataset that can reliably predict a student's performance. The data for the dataset was obtained from the placement cell and is utilised in the system for training and testing the model. It is required to train the model on a bigger amount of data in order to enhance its accuracy. In this work, the Random Forest method is employed to create predictions without the requirement to establish structure or rms. The model's results can be used to forecast future student performance.

A. Data Collection: The training and testing datasets for this investigation were collected from the placement cell. These databases provide critical information on numerous elements that impact student placement prediction, such as PG Percentage, PG CGPA, UG Percentage, UG CGPA, and Percentage of 12th grade, Percentage of 10th, and Percentage secured in quiz related with these variables. All of these characteristics are used to provide an accurate forecast of placement.

B. Data Pre-processing: This is the initial stage of every machine learning algorithm. This process includes data cleaning, transformation, and reduction. All of this is done to increase the efficacy of the data. The data may be analysed to increase the model's accuracy. To ensure that the categorisation is valid.

a. Cleaning Data – The training dataset was cleaned by deleting any null values that were no longer needed for the feature selection procedure. In addition, a few columns in the dataset were removed. Following data processing, new columns with numerical values were created and saved for prediction. Furthermore, the columns containing categorical data were removed from the dataset. As a consequence, an appropriate training dataset was created, which comprised attribute columns required for the investigation.

b. Splitting of Data – Following formatting, the data is divided into two distinct datasets, the training dataset and the testing dataset. The testing dataset is used to evaluate the performance of the machine learning model after it has been trained with the training dataset.

C. Machine Learning: This is designed to assist students predict their chances of being placed with the greatest degree of accuracy feasible. Machine learning methods are used to forecast placement probability using the available dataset. Several learning algorithms are available for predicting placement chance, and their performance is dependent on how they are trained. Several factors influence the best algorithm to use, including the type of problem to be solved, available computer resources, and data type.

1. **K-Nearest Neighbors -** K-Nearest Neighbours (KNN) is a basic and adaptable machine learning technique that may be used for classification and regression. It is a non-parametric method, which means it makes no assumptions about the underlying distribution of the data. KNN predicts a new instance based on the k-nearest examples in the training set. The value of k is normally selected by the user and denotes the number of neighbours to examine. KNN predicts the output of a regression issue as the average of the target values of the k-nearest neighbours. The equation for KNN regression is as follows:

$$y = (1/k) * \Sigma (yi)$$

Where y is the projected output, yi is the goal value of the i^{th} neighbor, and k is the number of neighbors evaluated. KNN predicts the outcome of a classification issue based on the class that appears most frequently among the k-nearest neighbours. The equation for KNN classification is as follows:

$$y = \text{argmax}(\Sigma I(yi = c)),$$

Where y is the predicted class, c is a class label, yi is the i^{th} neighbor's class label, and I is an indicator function that returns 1 if yi equals c and 0 otherwise. KNN provides various advantages in high-dimensional spaces, including simplicity, interpretability, and efficacy. It does, however, have several shortcomings, including being sensitive to noisy or irrelevant characteristics, being computationally expensive during testing, and requiring vast amounts of memory to retain the training data. The R2 score of the model's prediction accuracy is 81.13%.

2. **Random Forest Regressor -** Random Forest is a versatile machine learning approach that may be used for classification as well as regression. It is an ensemble learning approach that builds numerous decision trees during the training phase, with each tree trained on a random subset of the data and features. During prediction, the random forest model combines predictions from all trees to get the final forecast. This method reduces overfitting while increasing the model's accuracy and generalisation capabilities. Random Forest is a popular choice in many industries, including banking, healthcare, and e-commerce, since it is particularly successful at managing high-dimensional datasets and noisy data, and it may give insights into the value of characteristics. And the R2 score of this model's prediction accuracy is 89.04%.

V. BUILDMODEL

The main phase in Placement Prediction is model construction. Use the algorithms for creating the model.

1. Import the relevant packages.

```
import os
import pickle
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
```

2. Add the data to a Data Frame [Fig1] to determine the form of the data.

	planAfterGraduate	sslcPer	sslcYoP	sslcBoard	hsePer	hseYoP	hseBoard	nameOfUG	ugPer	ugCgpa	ugYoP	col
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	Job	88.0	2016	State Board of Kerala	73.5	2018	VHSE	BCA	72.8	7.19	2021	MES College, Er
2	Job	95.0	2016	State Board of Kerala	58	2018	HSE	BCA	59	5.8	2021	Sree Sabareesh
3	Job	80.0	2016	State Board of Kerala	65	2018	VHSE	B.Sc Electronics	66	6.6	2021	MES College, Er
4	Job	45.0	2016	State Board of Kerala	48	2018	HSE	B.Sc CS	45	4.6	2021	Sree Sabareesh

Fig 1-Table used for prediction

3. The dataset was then divided into training and testing datasets.

```
data = pd.read_csv("static/csv/2020-Student-DB.csv")

X = data.drop("output", axis=1)
y = data["output"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)

model.fit(X_train, y_train)
```

4. Save the trained model as a pickle file, which may be used to generate predictions without retraining the model with the dataset each time.

```
with open(pickle_file_path, 'wb') as f:
    pickle.dump(model, f)
```

VI. RESULT

The outcome demonstrates that the table indicates a study of Placement Prediction for a student using the quiz results and also the prediction of outcomes. The Random Forest method was one of the regression algorithms utilised in the study to predict placement. It outperformed the other algorithms in terms of accuracy, with an R-square value of 0.89 indicating that it explained a bigger share of the variability in placement prediction. Its MSE and MAE values were likewise lower than those of the other algorithms, indicating that its projected values were closer to the actual ticket prices.

Analysis of Algorithms:

Model	R2 Score	Prediction
K Nearest Neighbor	81.13	91.60
Random Forest	89.04	93.35

Random Forest Regressor

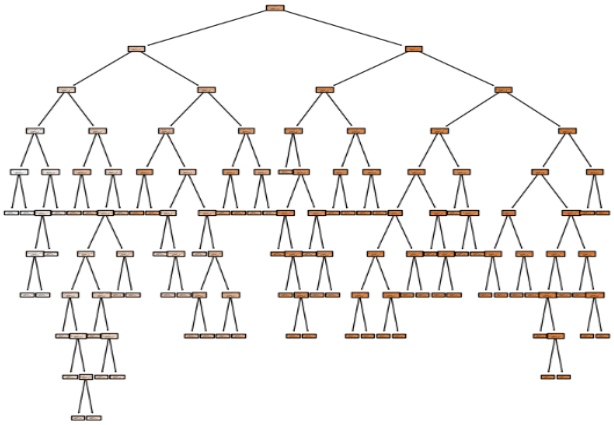

Fig2-Visualization of Random Forest.

```
mca_percentage     float64
mca_cgpa           float64
ug_percentage      float64
ug_cgpa            float64
hse_percentage     float64
sslc_percentage    float64
quiz_percentage      int64
output             float64
dtype: object
R2 Score:  0.8904021922937554
```

R-squared (R2), Mean Squared Error (MSE), and Mean Absolute Error (MAE) are metrics used to assess regression model accuracy.

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
print("R2 Score: ", r2)
print("Mean Squared Error: ", mse)

mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error: ", mae)
```

```
R2 Score:  0.8904021922937554
Mean Squared Error:  0.005671001562499995
Mean Absolute Error:  0.056246875000000064
```

The estimated prediction value [Fig 3] was more than a threshold. If the quiz result is high and good academic background, the probability of getting placed will be displayed.

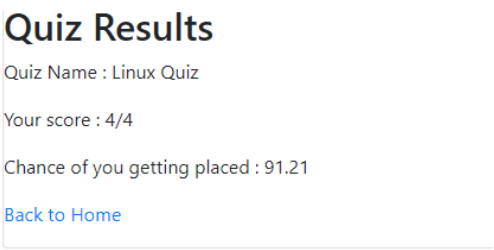

Fig 3 - Interface will show the placement chances of a student.

VII. CONCLUSION

modeled, and investigated to test the algorithmic rule. Machine learning methods are used to predict getting placed on a company accurately and provide the accurate value of the chance of getting placement from the performance of their Academic. Student's data is obtained from the database. As indicated in the above analysis, the Random Forest Regressor achieves the highest accuracy in forecasting placement prediction. The R-squared value is used to predict the model's accuracy, and high values are frequently obtained within the system

REFERENCES

[1] Rathi Viram, Swati Sinha, Bhagyashree Tayde, and Aakshada Shinde (2020). Placement prediction system using machine learning. International Journal of Creative Research Thoughts, vol. 08, issue 04, ISSN: 2320-2882.

[2] Rosemary Varghese, Adlene Peraira, Aswathy Ashok and Bassant Jhonson. (2022). Students' performance analysis using machine learning algorithms. International Journal of Research in Engineering and Science, vol. 10, issue 61, PP 1804-1809.

[3] Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew. (2019). Students' placement prediction using machine learning. International Research Journal of Engineering Technology, vol 06, issue 04, ISSN: 2395-0056.

[4] Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju, Joel James, and Mereen Thomas Vadakkel. (2020). Placement prediction using various machine learning models and their efficiency comparison. International Journal of Innovative Science and Research Technology, vol. 05, issue 05, ISSN: 2456-2165.

[5] Pothuganti Manvitha and Neelam Swaroopa. (2019). Campus placement prediction using supervised machine learning techniques. International Journal of Applied Engineering Research, vol. 14, issue 09, PP 2188-2191.