

Jomo Kenyatta University of Science and Technology

SCT313-2780/2023

Philip Okisegere Adung'o

Dr. Mulang

MSc. Software Engineering

10<sup>th</sup> January, 2024

**ICS 3102 : Foundations of Logic and Symbolic Reasoning.**

Question 2

## Introduction

Linear regression is a widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables. In the context of placement prediction, linear regression allows us to understand the linear associations between various features and the likelihood of placement.

Given the binary nature of placement outcomes (placement or no placement), classification models are imperative. Neural networks, a subset of machine learning models, have shown great promise in capturing complex patterns in data.

In the rapidly evolving landscape of higher education, the efficient transition of students from academia to the workforce is a critical aspect. Predicting Student placements has profound implications for students, educational institutions, and employers. To address this challenge, we embark on developing a Student placement predictor leveraging machine learning techniques and incorporating domain-specific knowledge.

In the process of developing Student Placement prediction machine language, the first step is data collection. Such data include historical placement data. The data to be used should have been collected over a sufficiently long period so as to obtain a comprehensive representation of Student placements, especially when considering variables like GPA and IQ.

Once the data related to Student placements has been collected, it can be utilized to train a machine learning algorithm for predicting placement outcomes based on various factors such as academic performance, extracurricular activities, and industry demands.

Once the model is trained, it can be applied to make predictions about future placement outcomes based on the current or anticipated academic and industry conditions. These predictions can then be used to guide decisions for students, educational institutions, and employers, assisting in career planning, curriculum development, and recruitment strategies. For example, if the model predicts that students with certain academic profiles are likely to have

higher placement rates in specific industries, educational institutions can tailor their programs to better align with industry needs.

In conclusion, machine learning models for Student placement prediction can be developed using diverse factors such as academic performance, extracurricular activities, and industry demands. By employing machine learning algorithms like Random Forest, SVMs, Neural Networks, and Gradient Boosting, it is possible to build models that provide valuable insights into placement probabilities, assisting students in making informed decisions about their career paths and aiding institutions in aligning their programs with industry requirements for improved placement outcomes

## Problem statement

The challenge at hand in developing a Student placement predictor is to identify the optimal academic and personal attributes that will maximize a student's likelihood of securing a placement. In other words, the problem is to determine the factors contributing to successful Student placements, taking into account variables such as GPA, IQ, and extracurricular activities. Students encounter numerous challenges when navigating the Student placement process, including academic performance and cognitive abilities.

Additionally, the placement predictor needs to consider the changing dynamics of the job market and industry-specific demands. This decision-making process is intricate, requiring a thorough understanding of the multifaceted factors influencing Student placement outcomes. Moreover, the placement decision is not a one-time event; it has repercussions on a student's career trajectory and the long-term success of their professional journey. An incorrect prediction can impact a student's future opportunities and career sustainability.

In summary, the problem statement for the Student placement predictor is to optimize the prediction of successful placements by considering the potential impact of academic, personal, and industry-specific factors on a student's placement outcomes, and ensuring the long-term success of their professional endeavors.

## Proposed Solution

Creation of an effective Student placement predictor model that leverages available data to accurately forecast students' likelihood of securing placements upon graduation. This model aims to assist students in making informed career choices, guide educational institutions in enhancing career services, and aid employers in understanding and meeting future workforce needs.

Furthermore, the deployment of such a model could contribute to optimizing the allocation of resources, improving the alignment between education and industry demands, and ultimately fostering a more efficient and responsive higher education system.

## Data Modelling

To develop a Student placement predictor, student data such as academic performance, extracurricular activities, IQ scores, and demographics are collected for a specific institution over an extended period. This data is then combined with placement outcomes for the same period and institution. The merged dataset is utilized to train a machine learning model.

Several machine learning algorithms can be employed for building a Student placement predictor based on student-related factors, such as:

- i. **Random Forest:** A decision tree algorithm leveraging multiple trees to enhance accuracy and stability in predicting placement outcomes.
- ii. **Support Vector Machines (SVMs):** An algorithm suitable for classification and regression tasks, potentially adept at discerning patterns in student data.
- iii. **Neural Networks:** An algorithm inspired by the structure and function of the human brain, capable of handling diverse features relevant to Student placements.
- iv. **Gradient Boosting:** An algorithm that boosts prediction accuracy by combining multiple weak models, which can be beneficial in understanding complex relationships in student profiles.

In this project, the Random Forest algorithm is selected as the preferred choice due to the following advantages:

- a. **High Accuracy:** Among various classification methods, Random Forests consistently demonstrate superior accuracy in predicting Student placement outcomes.
- b. **Efficient Variable Handling:** The algorithm efficiently handles a myriad of student-related variables, making it well-suited for the complexity of predicting placement success.
- c. **Scalability to Big Data:** Random Forests are adept at handling large datasets with numerous student variables, which is particularly advantageous in dealing with extensive historical placement and student profile data.

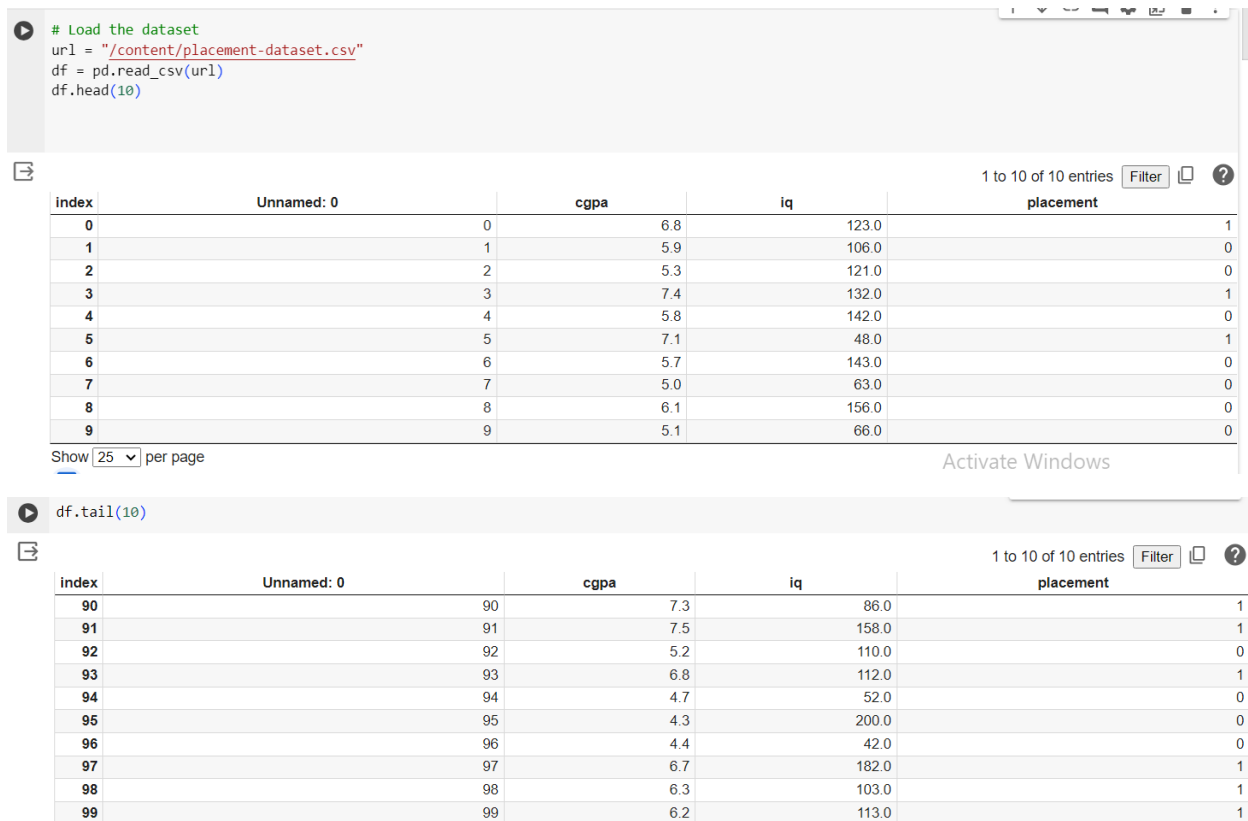
Once the model is trained, it can be utilized to make predictions about future Student placements based on current or forecasted academic and extracurricular achievements of students. These predictions can then be used to inform decisions about educational strategies and career counseling to optimize placement outcomes. In addition, the model can be fine-tuned and improved by incorporating more student data and updating it with new profiles as they become available. This continuous improvement process can help increase the accuracy and reliability of placement predictions over time.

## Process Workflow

A student placement prediction machine learning model typically follows a structured process flow:

### a. Data Collection:

The initial stage involves gathering comprehensive data related to students' academic performance, extracurricular activities, internship experiences, and industry demands over an extended period. This includes historical records of placement outcomes, demographic information, and IQ scores for individual students. The data collection process spans multiple academic years to ensure a diverse representation of student profiles and evolving industry requirements.



The screenshot displays a Jupyter Notebook interface. The top cell contains Python code to load a CSV dataset. The bottom cell shows the tail of the dataset, displaying rows 90 through 99. The data table has five columns: index, Unnamed: 0, cgpa, iq, and placement.

```
# Load the dataset
url = "/content/placement-dataset.csv"
df = pd.read_csv(url)
df.head(10)
```

index	Unnamed: 0	cgpa	iq	placement
0	0	6.8	123.0	1
1	1	5.9	106.0	0
2	2	5.3	121.0	0
3	3	7.4	132.0	1
4	4	5.8	142.0	0
5	5	7.1	48.0	1
6	6	5.7	143.0	0
7	7	5.0	63.0	0
8	8	6.1	156.0	0
9	9	5.1	66.0	0

df.tail(10)

index	Unnamed: 0	cgpa	iq	placement
90	90	7.3	86.0	1
91	91	7.5	158.0	1
92	92	5.2	110.0	0
93	93	6.8	112.0	1
94	94	4.7	52.0	0
95	95	4.3	200.0	0
96	96	4.4	42.0	0
97	97	6.7	182.0	1
98	98	6.3	103.0	1
99	99	6.2	113.0	1

Figure 1- Dataset Preview

### b. Data Pre-processing:

The gathered data for the student placement predictor underwent meticulous pre-processing to prepare it for effective use in machine learning models. This phase involved thorough cleaning and formatting to address missing values, outliers, and other potential issues that could impact the precision of the placement prediction model. This ensured that the input data was of high quality, contributing to the robustness and reliability of the placement predictor. No missing

values were found, and the 'Unnamed: 0' column was dropped as it serves as an index. The data types were appropriate for analysis.

```
[44] df.shape
```

```
(100, 4)
```

```
# Check for missing values
print("Missing Values:")
print(df.isnull().sum())
```

```
Missing Values:
Unnamed: 0      0
cgpa            0
iq             0
placement      0
dtype: int64
```

Figure 2- Data Processing

### c. Data Splitting and Model Training:

Prior to model development, the dataset was strategically partitioned into training (80%) and testing (20%) sets, leveraging 'IQ' and 'CGPA' as features and designating 'Placement' as the target variable. This meticulous data splitting process lays the foundation for robust model training and evaluation.

The pre-processed Studentplacement data is then utilized to train a machine learning model. This involves employing algorithms such as Random Forest, Support Vector Machines (SVMs), Neural Networks, or Gradient Boosting to develop a predictive model for forecasting students' likelihood of securing placement after graduation.



```
# Splitting the data
X = df[['iq', 'cgpa']]
y = df['placement']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print("X_train: ", X_train.shape)
print("X_test: ", X_test.shape)
print("y_train: ", y_train.shape)
print("y_test: ", X_test.shape)
```

X\_train: (80, 2)  
X\_test: (20, 2)  
y\_train: (80,)  
y\_test: (20, 2)

```
✓ [84] # Model Training
0s model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```

RandomForestClassifier  
RandomForestClassifier(random\_state=42)

Figure 3- Model Training

## b. Model Evaluation

Upon training the student placement predictor model, a thorough evaluation is conducted using a set of metrics to assess its accuracy and performance. Key measures considered in the evaluation process include mean squared error, mean absolute error, and the coefficient of determination (R-squared). These metrics serve as essential indicators of the model's effectiveness in predicting placement outcomes, allowing for a comprehensive understanding of its predictive capabilities.

The developed Studentplacement predictor underwent rigorous evaluation using an independent testing set. Performance metrics, including accuracy, precision, recall, F1-score, and a detailed confusion matrix, were computed to assess the model's effectiveness in forecasting students' placement outcomes

```
✓ [85] # Model Evaluation
      y_pred = model.predict(X_test)

✓ [86] # Test the Model Accuracy
      accuracy = accuracy_score(y_test, y_pred)
      print("\nModel Accuracy:", accuracy)

      Model Accuracy: 0.85

✓ [87] # Classification Report and Confusion Matrix
      print("\nClassification Report:\n", classification_report(y_test, y_pred))
      print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))

      Classification Report:
              precision    recall  f1-score   support

         0       0.82      0.90      0.86         10
         1       0.89      0.80      0.84         10

   accuracy          0.85
  macro avg          0.85
 weighted avg          0.85

      Confusion Matrix:
      [[9 1]
       [2 8]]
```

Figure 4- Model Evaluation

## Challenges

Developing a college placement prediction model presents a complex and multifaceted challenge. The placement outcomes of students are influenced by various factors, including academic performance, extracurricular activities, industry demands, and individual student profiles. Furthermore, these factors can vary significantly across academic years, institutions, and geographical locations. This complexity makes it difficult to create a model that accurately predicts college placement outcomes in all conditions.

One of the primary challenges in building a college placement prediction model is the availability of diverse and comprehensive data. To construct an accurate model, it is necessary to gather substantial data on student profiles, academic achievements, extracurricular involvements, and industry demands. Obtaining this data can be challenging, as it may not be uniformly available for all institutions, and the format may not be standardized for use in a prediction model.

Another challenge arises from the intricate interplay of factors influencing college placements. Academic performance, extracurricular achievements, industry demands, and individual preferences can all significantly impact placement outcomes. Moreover, these factors may interact in complex ways, similar to the complexities found in crop prediction. For example, a student's choice of major may be influenced by industry demands, which, in turn, can affect placement opportunities.

In addition to these challenges, economic and social factors must be considered. A student's decision on the choice of major or specialization may be influenced by job market trends and career prospects. Similarly, economic conditions, such as job market stability, may impact employers' decisions to hire graduates. Incorporating these economic and social factors is essential for developing a holistic college placement prediction model.

Despite the challenges, creating a college placement prediction model is crucial for enhancing educational outcomes and career opportunities for students. Accurate predictions can aid students in making informed decisions about their academic and extracurricular pursuits, leading to improved placement outcomes. Furthermore, such a model can assist educational institutions in adapting their programs to align with industry demands, ultimately contributing to improved overall educational quality and students' career success.

In conclusion, developing a college placement prediction model is a complex undertaking that requires extensive and varied data, coupled with a deep understanding of the factors influencing placement outcomes. Despite the challenges, this endeavor is indispensable for optimizing educational pathways and facilitating successful transitions from academia to the workforce. A well-crafted college placement prediction model can empower students and institutions, fostering better-informed decisions and improved overall outcomes in the realm of higher education and career placement.

## Reference

<https://www.kaggle.com/datasets/sameerprogrammer/college-placement/data>

Vinayak Hegde, Abhinav M R, Roshin C, "Predicting Student Placement using PCA and Machine Learning Technique", *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp.1-5, 2023.