# Identifying patterns and trends in campus placements using Machine learning

The project submitted to Smart internz

By

Chilamakuri Sushma

Routhu Neeraja

Kambhampati Ramatulasi

Ganganapalli Indupriya

Department of Computer Science & Engineering


SVCE KADAPA

EDUCATION FOR A BETTER SOCIETY

# 1INTRODUCTION

NOWADAYS the number of educational institutes is growing day by day. The aim of each higher educational institute is to help their students to get a well-paid job through their placement cell. One of the biggest challenges that higher learning institutes face these days is to uplift the placement performance of scholars. The goal of this system is to predict whether the student will get a campus placement or not based on various parameters such as gender, SSC percentage, HSC percentage, HSC stream, degree percentage, degree type, work experience & e-test percentage. This research focuses on various algorithms of machine learning such as Logistic Regression, Decision Tree, K-Nearest Neighbours and Random Forest in order to produce economical and correct results for campus placement prediction. This system follows a supervised machine learning approach as it uses class labelled data for training the classification algorithm.

## 2.Overview

Campus recruitment is a strategy for sourcing, engaging and hiring young talent for internship and entry-level positions. College recruiting is typically a tactic for medium- to large-sized companies with high-volume recruiting needs, but can range from small efforts (like working with university career centers to source potential candidates) to large-scale operations (like visiting a wide array of colleges and attending recruiting events throughout the spring and fall semester).Campus recruitment often involves working with university career services centers and attending career fairs to meet in-person with college students and recent graduates.Our solution revolves around the placement season of a Business School in India. Where it has various factors on candidates getting hired such as work experience,exam percentage etc., Finally it contains the status of recruitment and remuneration details.
We will be using algorithms such as KNN, SVM and ANN. We will train and test the data with these algorithms. From this the best model is selected and saved in .pkl format. We will be doing flask integration and IBM deployment.

## 1.2 Purpose

The use of this project. What can be achieved using this.

## LITERATURE SURVEY

Jain, S., & Kumar, R. (2021). A Review on Student Placement Prediction Models using Machine Learning. International Journal of Computer Applications, 179(8), 19-22. This review paper provides an overview of different machine learning techniques used for student placement prediction. It discusses various features and algorithms employed in

the prediction models, along with their advantages and limitations. Bhatia, N., & Singh, V. (2020). Student Placement Prediction using Machine Learning: A Review. In Proceedings of the International Conference on Recent Innovations in Computing (ICRIC), 1-5. The authors present a review of different machine learning algorithms used for student placement prediction. They analyze the performance of various models and highlight the factors influencing placement outcomes, such as academic performance, skills, and internships.

## THEORITICAL ANALYSIS

The steps involved in this system are as follows,

A. Data Acquisition:

The campus placement dataset is collected from Kaggle website. Here is the link for the dataset: https://www.kaggle.com/benroshan/factors-affecting-campus placement?select=Placement_Data_Full_Class.csv The dataset consists of various attributes such as Serial Number, Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E -test Percentage, Degree Specialization, Degree Percentage, Placement Status & Salary. The size of dataset is 19.71 KB & it has total 215 records.

1) Handling missing values:

In our dataset missing values are present only in the salary column as these values correspond to the students who didn't get placed in any placement drive. So it is assumed that the missing values in Salary Column are Zero & replaced them by zero using fillna(0,inplace=True) function in Python.

2) Handling categorical data:
since we cannot deal with categorical values directly, mapping is done for attributes having categorical values. Gender attribute has values M (Male) & Female (M). Here, M is replaced by 0 & F is replaced by 1. SSC & HSC Board attributes has values 'Central' & 'Other.' Here, Central is replaced by 1 & Other is replaced by 0. Work Experience attribute has values 'Yes' & 'No'. Here, 'Yes' is replaced by 1 and 'No' is replaced by 0. Degree specialization attribute has values 'Marketing & Finance' & 'Marketing & HR'. Here, 'Marketing & Finance' is replaced by 1 and 'Marketing & HR' is replaced by 0. Status attribute has values 'Placed' and 'Not Placed'. Here, 'Placed' is replaced by 1 and 'Not Placed' is replaced by 0. This is achieved through map function in Python. For e.g., x df['gender']=df['gender'].map({'M':0,'F':1}) x

df['ssc_b']=df['ssc_b'].map({'Central':1,'Others':0}) x
df['workex']=df['workex'].map({'Yes':1,'No':0})

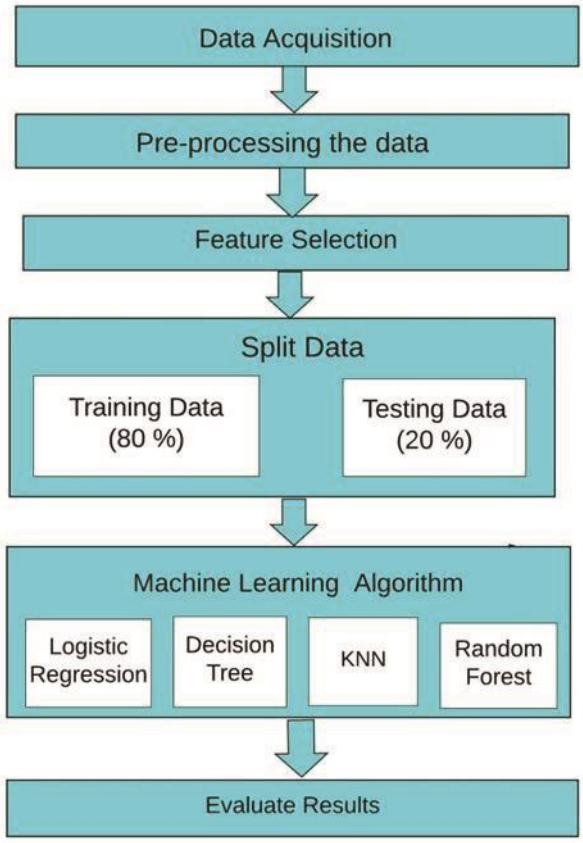



Fig. 1. Architecture Diagram

*3) Feature Selection:*

Here, various features are visualized to understand their correlation with the target feature.

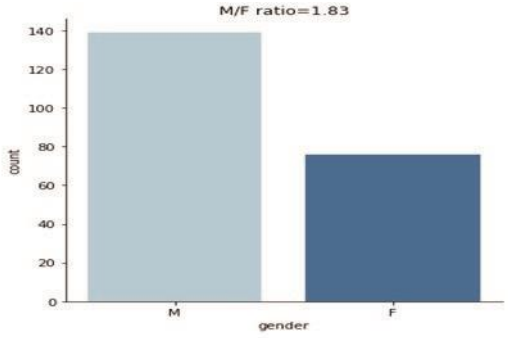



Fig. 2. M/F ratio

Here, male : female ratio for one batch of students is approximately equal to 2. It means that there are 2 male candidates appearing for placement drives for every 1 female candidate.

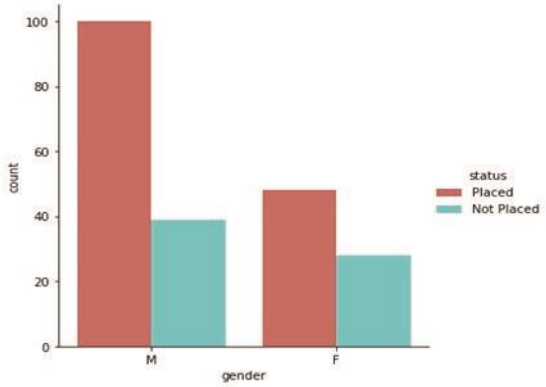



Fig. 3. Placement count vs. gender

From the above graph it can be concluded that the count of placed male candidates in a batch is higher as compared to female candidates & the placement count is dependent on gender.

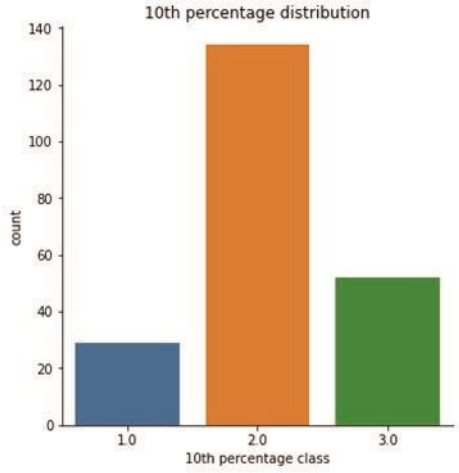



Fig. 4. 10th standard percentage distribution

In the above graph, class 1 represents students having scores between 80-100%, class 2 represents students having scores
between 60-80% and class 3 represents students having less than 60 % score in 10th standard.

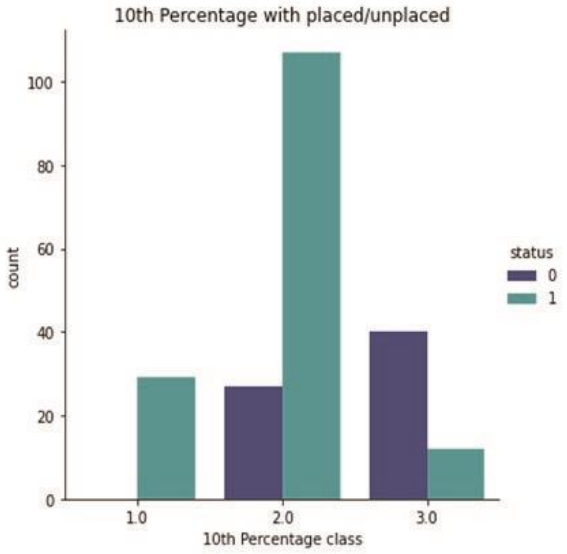



Fig. 5. Placement count vs. 10th percentage

From the above graph, it's observed that all the students having scores between 80-100% in 10th standard got placed. Very few students having scores between 60-80% in 10th standard couldn't get placed. Whereas, most of the students having below 60% score in 10th standard couldn't get placed.

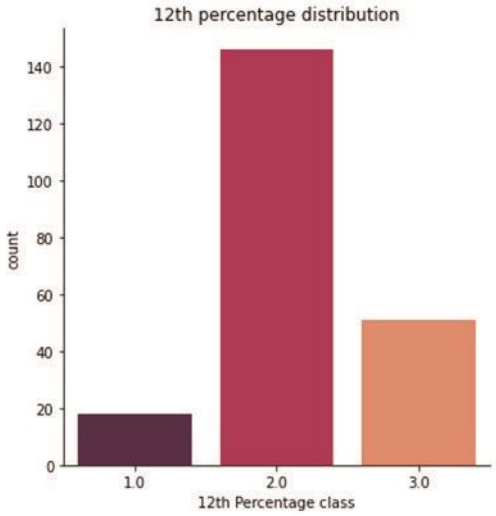



Fig. 6. 12th standard percentage distribution

In the above graph, class 1 represents students having scores between 80-100% , class 2 represents students having scores between 60-80% and class 3 represents students having less than 60 % score in 12th standard.

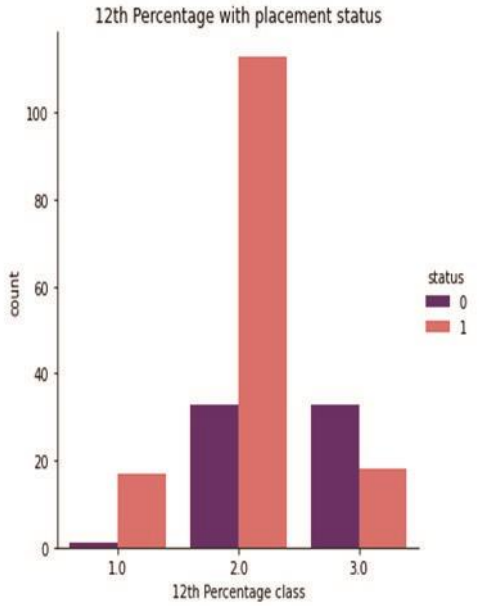



Fig. 7. Placement count vs. 12th percentage

From the above graph, it's observed that all the students having scores between 80-100% in 12th standard got placed. Very few students having scores between 60-80% in 12th standard couldn't get placed. Whereas, most of the students having below 60% score in 12th standard couldn't get placed.

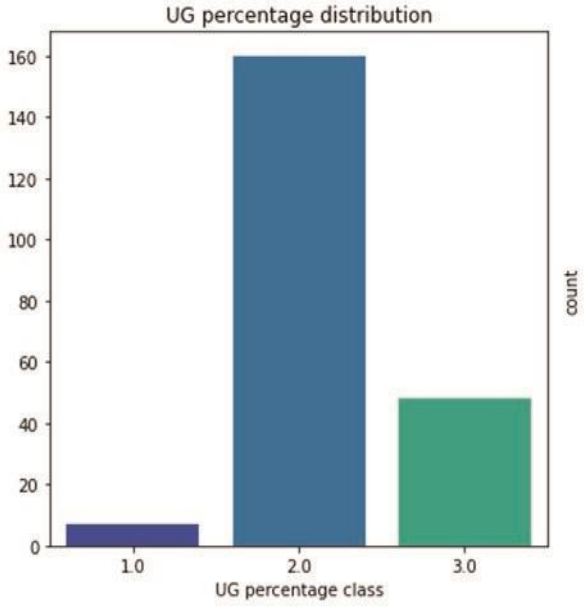



Fig. 8. UG percentage distribution

In the above graph, class 1 represents students having scores between 80-100%, class 2 represents students having scores between 60-80% and class 3 represents students having less than 60 % score in UG degree.

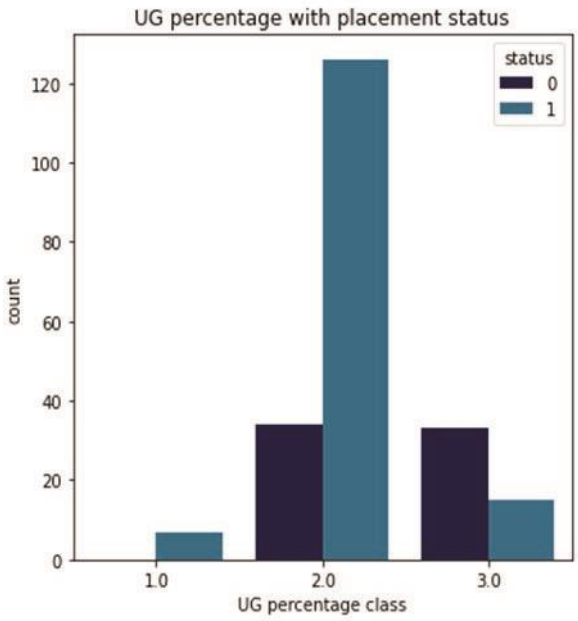



Fig. 9. Placement count vs. UG percentage

From the above graph, it's observed that most of the students having scores between 80-100% in UG got placed. Very few students having scores between 60-80% in UG couldn't get placed. Whereas, most of the students having below 60% score in UG couldn't get placed.

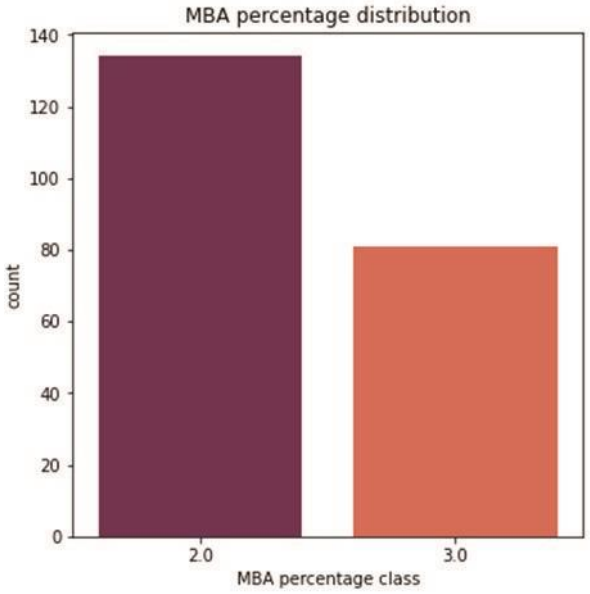



Fig. 10. MBA percentage distribution

After studying MBA percentage data it is observed that no student has secured more than 80% marks. So the class 1 data isn't available for percentage of MBA.

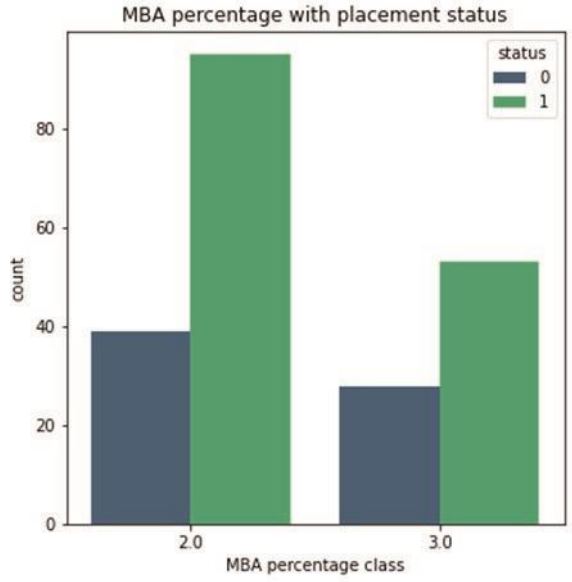



Fig. 11. Placement count vs. MBA percentage

In the above graph we can see that more students from class 2 got placed as compared to class 3.

Hence, it is clear that placement count of the students is dependent on various features such as Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E - test Percentage, Degree Specialization, Degree Percentage.

*4) Split data:*

Here, data is divided into two parts i.e. training data & testing data. Where 80 % data is taken for training our machine learning algorithm and remaining 20 % data is used for testing whether our trained machine learning model is working correctly or not.

*5) Machine Learning Algorithm:*

*a) Logistic Regression:*

Logistic regression is a statistical method used to determine the outcome of a dependent variable (y) based on the values of independent variable (x).

In our problem dependent variable is placement status and independent variables are the features selected by us in the previous step.

This algorithm is mostly used for the problems of binary classification.

*b) Decision Tree:*

A decision tree is a graph like a tree where nodes represent the position where we select the feature and ask a question, edges represent the answers of the question; and the leaves represent the final output or label of the class.

*c) KNN:*

K-NN stores all the training data into different classes based on the class labels and classifies new data by checking its similarity with data in the available classes.

*d) Random Forest:*

Random Forest classifier consists of a number of decision trees which apply on different subsets of our dataset and the average of outputs of all the decision trees is taken to improve the accuracy of output prediction.

*6) Evaluate results:*

Accuracy is calculated by following formula,

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Where,

TP: True Positive (the number of cases correctly identified as placed)

TN: True Negative (the number of cases correctly identified as unplaced).

FP: False Positive (the number of cases incorrectly identified as placed)

FN: False Negative (the number of cases incorrectly identified as unplaced)

TABLE I. TP, FP, FN & TN VALUES OF DIFFERENT MODELS

| Model | TP | FP | FN | TN |
|---|---|---|---|---|
| Logistic Regr ession | 16 | 1 | 1 | 25 |
| Decision Tree | 13 | 3 | 4 | 23 |
| KNN | 14 | 1 | 3 | 25 |
| Random Forest | 13 | 2 | 4 | 24 |

TABLE II. CAMPUS PLACEMENT PREDICTION ACCURACY OF DIFFERENT MODELS.

| Model | Accuracy |
|---|---|
| Logistic Regression | 95.34 % |
| Decision Tree | 83.72 % |
| KNN | 90.69 % |
| Random Forest | 88.67 % |

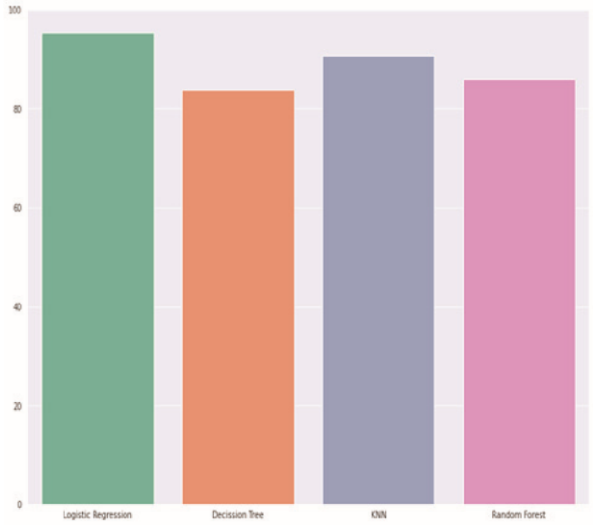


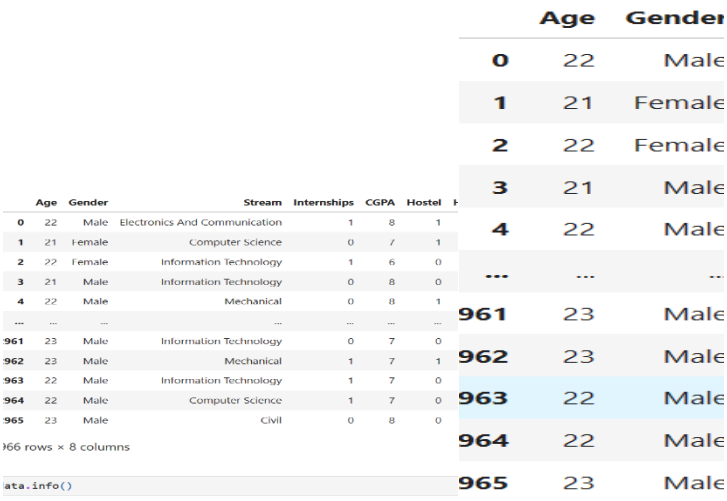

Fig. 12. Comparison of Campus placement prediction accuracy of different

# RESULT

Final findings (Output) of the project along with screenshots.

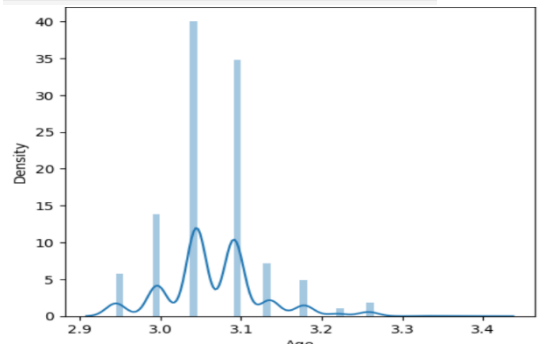
| | Age | Gender | Stream | Internships | CGPA | Hostel |
|---|---|---|---|---|---|---|
| 0 | 22 | Male | Electronics And Communication | 1 | 8 | 1 |
| 1 | 21 | Female | Computer Science | 0 | 7 | 1 |
| 2 | 22 | Female | Information Technology | 1 | 6 | 0 |
| 3 | 21 | Male | Information Technology | 0 | 8 | 0 |
| 4 | 22 | Male | Mechanical | 0 | 8 | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| 961 | 23 | Male | Information Technology | 0 | 7 | 0 |
| 962 | 23 | Male | Mechanical | 1 | 7 | 1 |
| 963 | 22 | Male | Information Technology | 1 | 7 | 0 |
| 964 | 22 | Male | Computer Science | 1 | 7 | 0 |
| 965 | 23 | Male | Civil | 0 | 8 | 0 |

966 rows × 8 columns



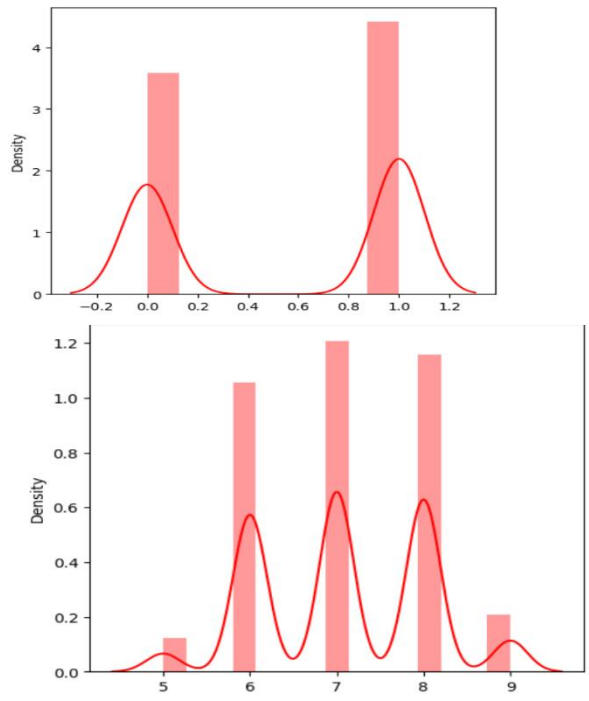

```
ata.info()
```

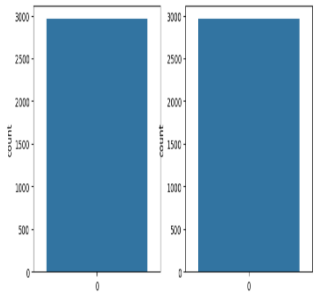
```
RangeIndex: 2966 entries, 0 to 2965
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Age                2966 non-null   int64
 1   Gender             2966 non-null   object
 2   Stream             2966 non-null   object
 3   Internships        2966 non-null   int64
 4   CGPA               2966 non-null   int64
 5   Hostel             2966 non-null   int64
 6   HistoryOfBacklogs  2966 non-null   int64
 7   PlacedOrNot        2966 non-null   int64
dtypes: int64(6), object(2)
memory usage: 185.5+ KB
```



| | Age | Gender | Stream | Internships | CGPA | Hostel | HistoryOfBacklogs | PlacedOrNot |
|---|---|---|---|---|---|---|---|---|
| **0** | 22 | 1.0 | 3.0 | 1 | 8 | 1 | 1 | 1 |
| **1** | 21 | 0.0 | 1.0 | 0 | 7 | 1 | 1 | 1 |
| **2** | 22 | 0.0 | 4.0 | 1 | 6 | 0 | 0 | 1 |
| **3** | 21 | 1.0 | 4.0 | 0 | 8 | 0 | 1 | 1 |
| **4** | 22 | 1.0 | 5.0 | 0 | 8 | 1 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **2961** | 23 | 1.0 | 4.0 | 0 | 7 | 0 | 0 | 0 |
| **2962** | 23 | 1.0 | 5.0 | 1 | 7 | 1 | 0 | 0 |
| **2963** | 22 | 1.0 | 4.0 | 1 | 7 | 0 | 0 | 0 |
| **2964** | 22 | 1.0 | 1.0 | 1 | 7 | 0 | 0 | 0 |
| **2965** | 23 | 1.0 | 0.0 | 0 | 8 | 0 | 0 | 1 |

2966 rows × 8 columns






| | Age | Gender | Stream | Internships | CGPA | Hostel | HistoryOfBacklogs | PlacedOrNot |
|---|---|---|---|---|---|---|---|---|
| **0** | 0.388131 | 1.0 | 3.0 | 1 | 8 | 1 | 1 | 1 |
| **1** | -0.366752 | 0.0 | 1.0 | 0 | 7 | 1 | 1 | 1 |
| **2** | 0.388131 | 0.0 | 4.0 | 1 | 6 | 0 | 0 | 1 |
| **3** | -0.366752 | 1.0 | 4.0 | 0 | 8 | 0 | 1 | 1 |
| **4** | 0.388131 | 1.0 | 5.0 | 0 | 8 | 1 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **2961** | 1.143013 | 1.0 | 4.0 | 0 | 7 | 0 | 0 | 0 |
| **2962** | 1.143013 | 1.0 | 5.0 | 1 | 7 | 1 | 0 | 0 |
| **2963** | 0.388131 | 1.0 | 4.0 | 1 | 7 | 0 | 0 | 0 |
| **2964** | 0.388131 | 1.0 | 1.0 | 1 | 7 | 0 | 0 | 0 |
| **2965** | 1.143013 | 1.0 | 0.0 | 0 | 8 | 0 | 0 | 1 |

2966 rows × 8 columns

| | PlacedOrNot |
|---|---|
| **0** | 1 |
| **1** | 1 |
| **2** | 1 |
| **3** | 1 |
| **4** | 1 |
| **...** | ... |
| **2961** | 0 |
| **2962** | 0 |
| **2963** | 0 |
| **2964** | 0 |
| **2965** | 1 |

2966 rows × 1 columns

SVC

```
SVC(kernel='linear')
```

```
k= 5
accuracy= 88.04713804713805
```

## CONCLUSION

The problem of campus placement prediction can be solved with the help of different machine learning algorithms such as Logistic regression, Decision Tree, KNN & Random Forest.

Here, the Logistic Regression algorithm gave the highest accuracy of  95. 34 % for campus placements prediction.

The selected features i.e. Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E -test Percentage, Degree Specialization & Degree Percentage lead to higher classification accuracy.

## FUTURE SCOPE

Accuracy may further increase by application of more advanced techniques such as deep learning & experimenting with different activation functions of neural networks such as linear, sigmoid, tan h & ReLU.

We can also experiment with different cross validation techniques such as 3 Fold, 5 Fold, 10 Fold, 15 Fold cross validation in order to analyze the change in accuracy.