PLACEMENT PREDICTOR

Group Members

1.Deepak Kumar(19)
2.Kuldeep Saini(26)
3.Mayank Kharab(30)
4.Mohd Shohel(32)

Detailed Overview

01 PROBLEM STATEMENT

02 Data Preprocessing

03 DATA VISUALIZATION

04 MODELS

05 CONCLUSION & FUTURE WORK

PROBLEM STATEMENT

The primary challenge for learning institutions today is enhancing student placement performance. To address this, educational institutes seek more efficient technology, with a focus on improving the quality of knowledge related to educational processes and managerial systems. Machine learning techniques are proposed to extract insights from operational and historical data within educational databases, offering a promising approach to tackling this challenge.

DATASET INFORMATION

The dataset used for training as well as testing was obtained from the Kagle. It contains 10000 rows and 11 features.

Features

- Internship
- Projects
- Workshops/Certificates
- Soft Skills assessments
- Academic performance in Secondary School Certificate (SSC)
- ExtracurricularActivities
- PlacementTraining
- Aptitude Scores
- Higher Secondary Certificate Marks.
- CGPA

DATA PRE- PROCESSING

Data preprocessing is a crucial step, encompassing the cleaning and transformation of raw data
to render it suitable for analysis. The goal of data preprocessing is to improve the quality of
the data and to make it more suitable for the specific data mining task.

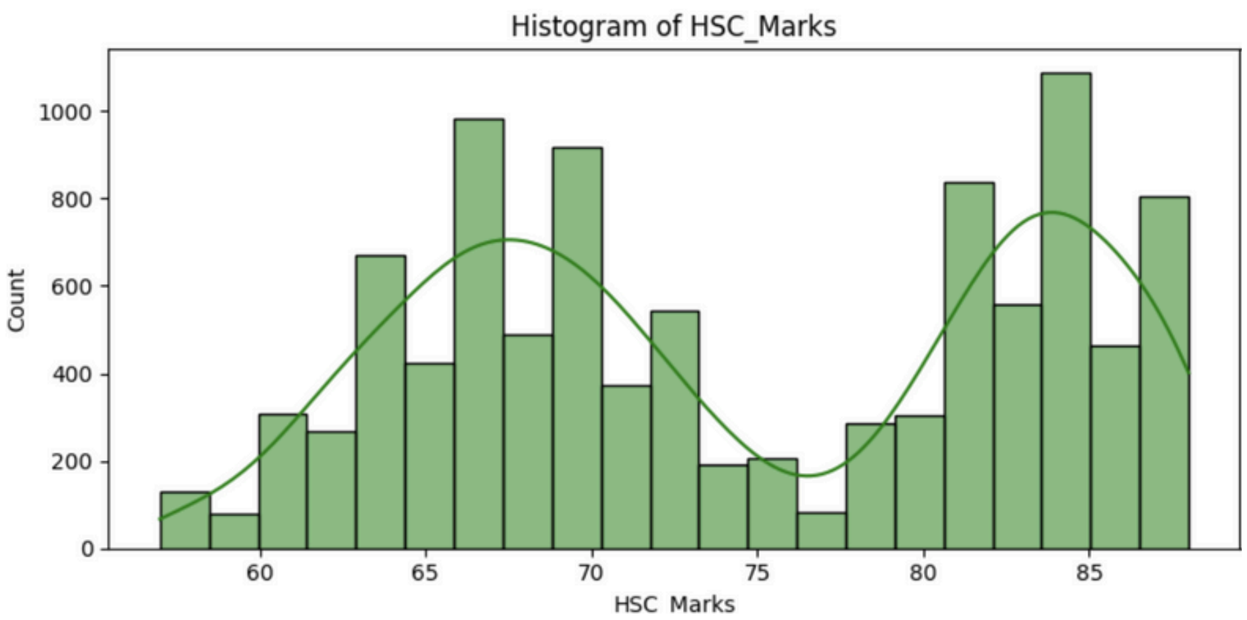
Dropped “StudentId” column because it not relevant for our data analysis.
StudentId does affect our model or is not significant for analysis

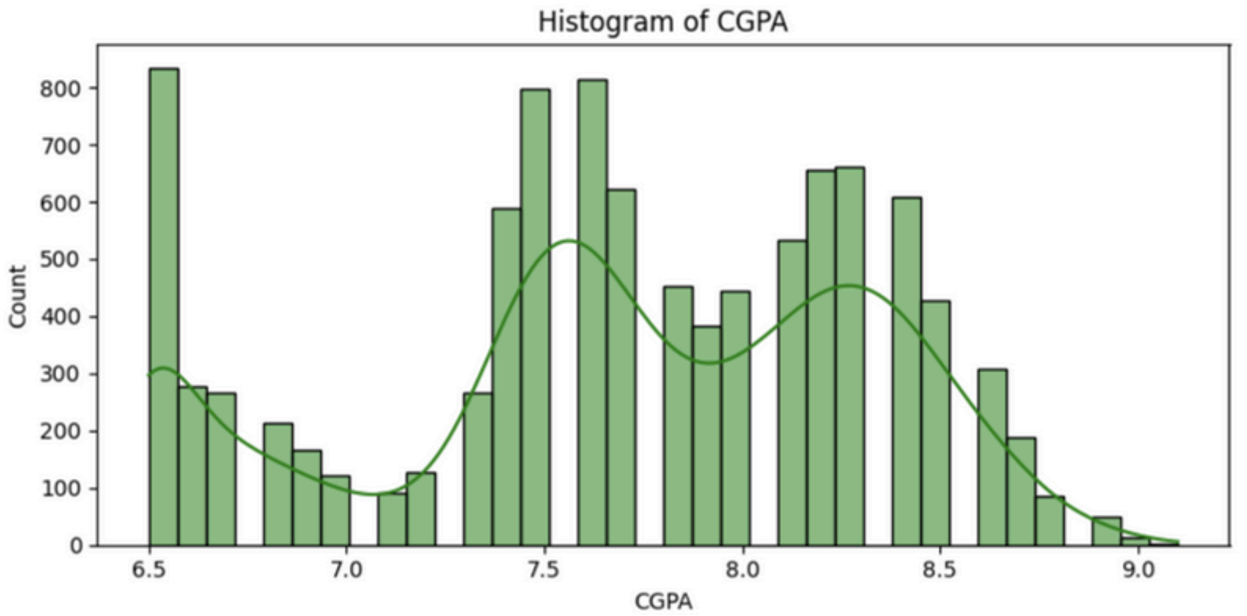
All the data in the dataset is in int or float except for 3 columns/variables that are of object data type:

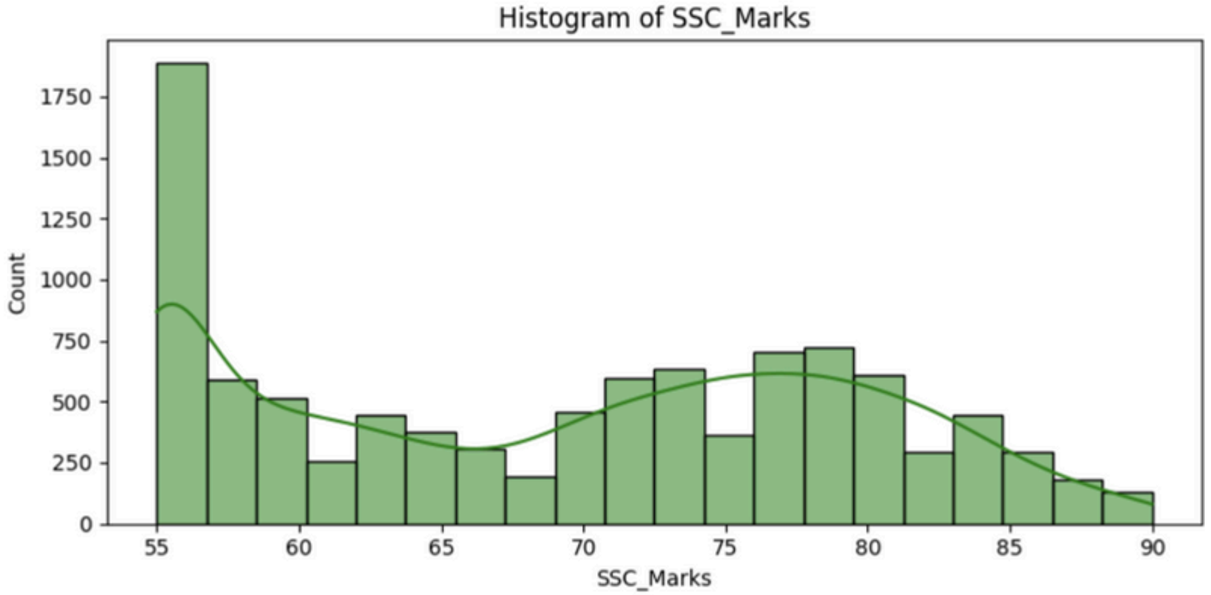
- **ExtracurricularActivities**
- **PlacementTraining**
- **Placement Status**

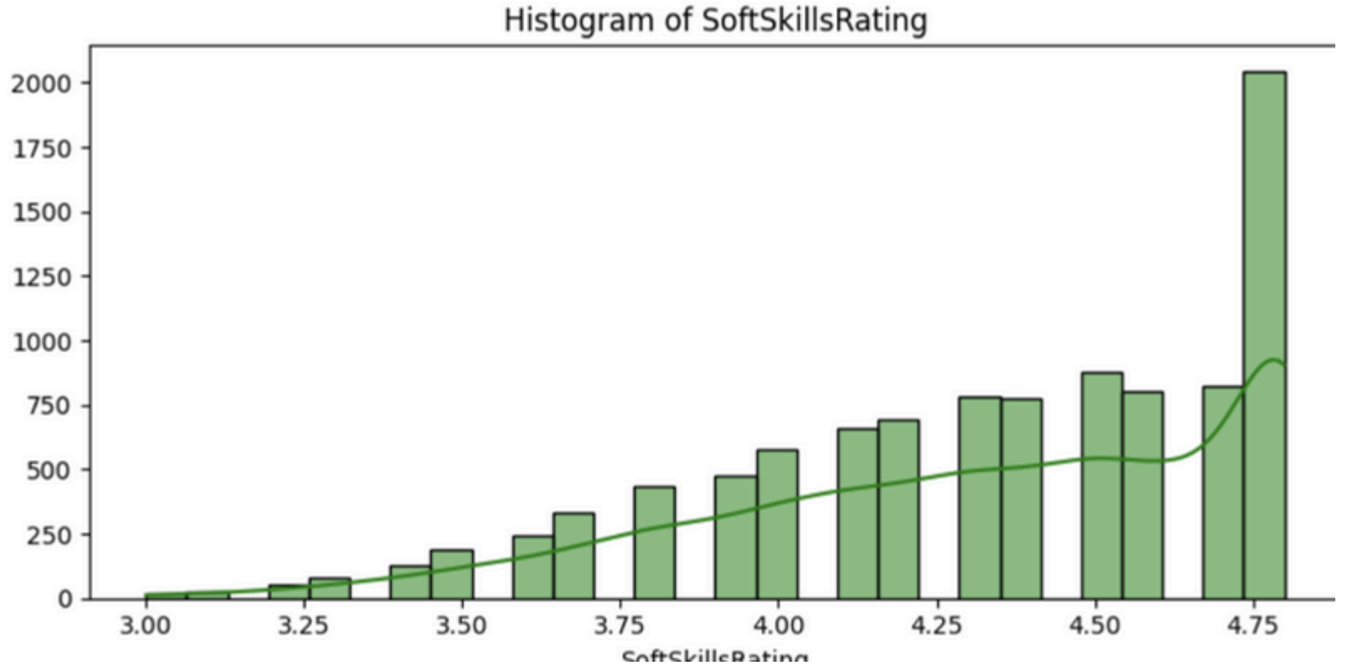
In sci-kit-learn, the LabelEncoder is a utility class used to encode categorical labels into numerical labels. It essentially converts categorical data (text labels) into numerical labels so that machine learning algorithms can handle them more effectively.
For instance, if you have a categorical feature like "Yes" and "No". The LabelEncoder would assign them numerical labels, such as 0 and 1, respectively.

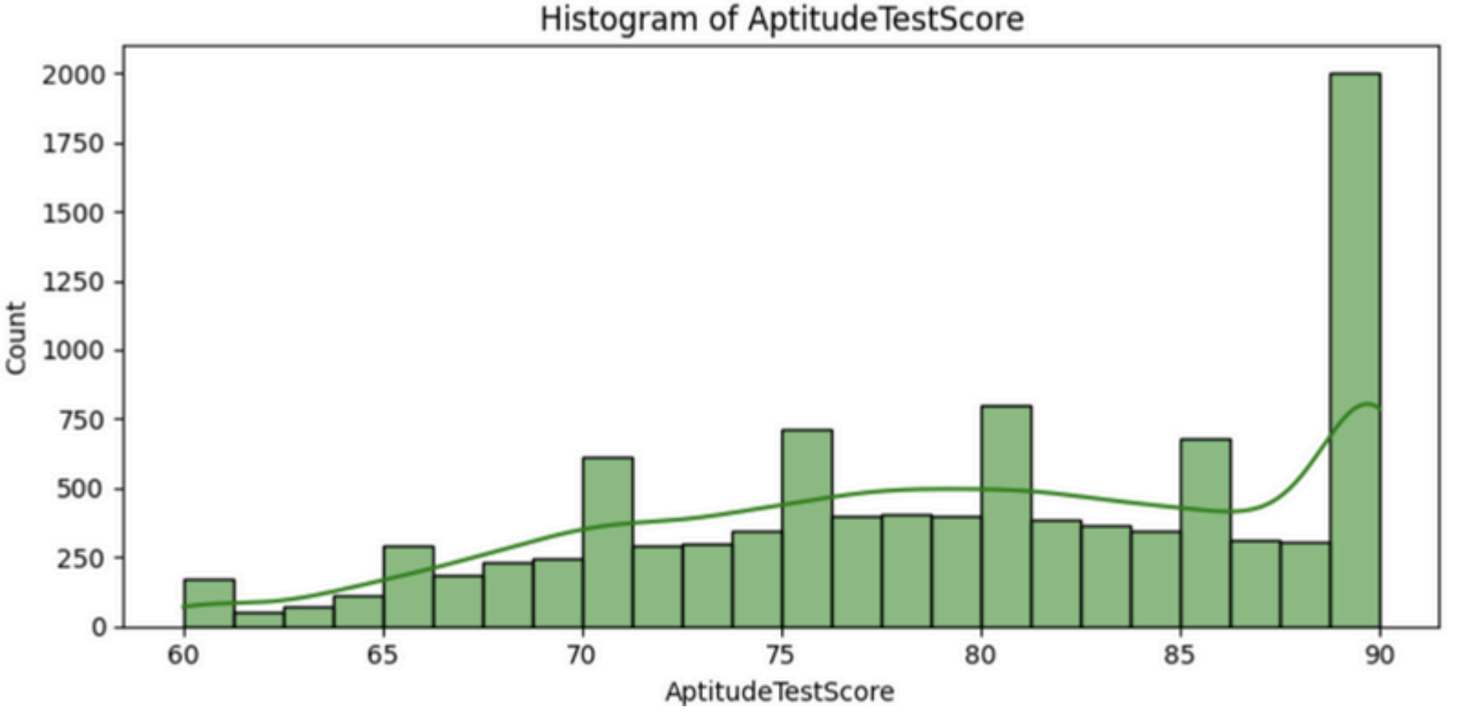
DATA VISUALIZATION

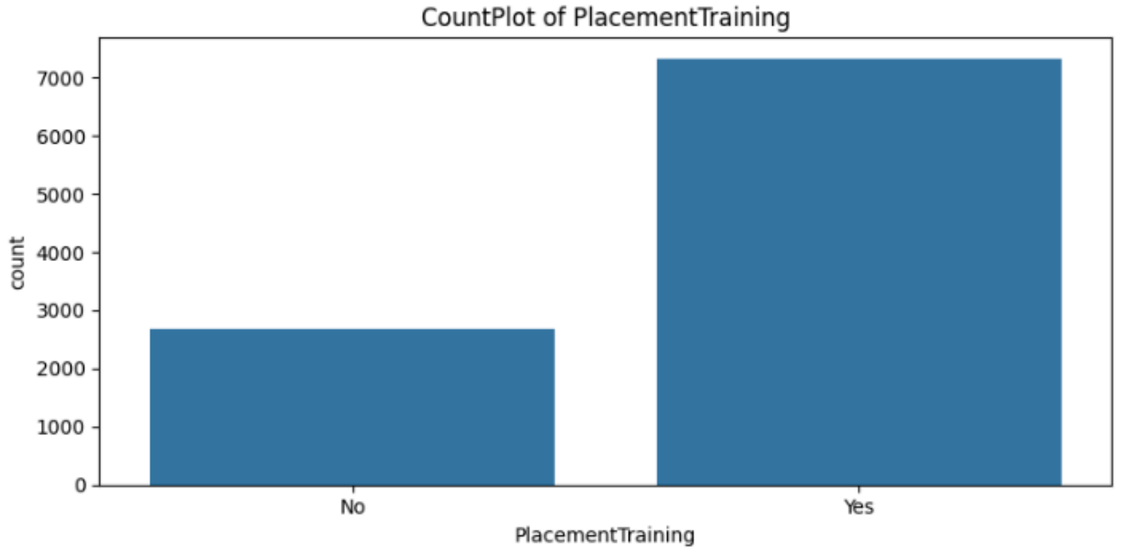


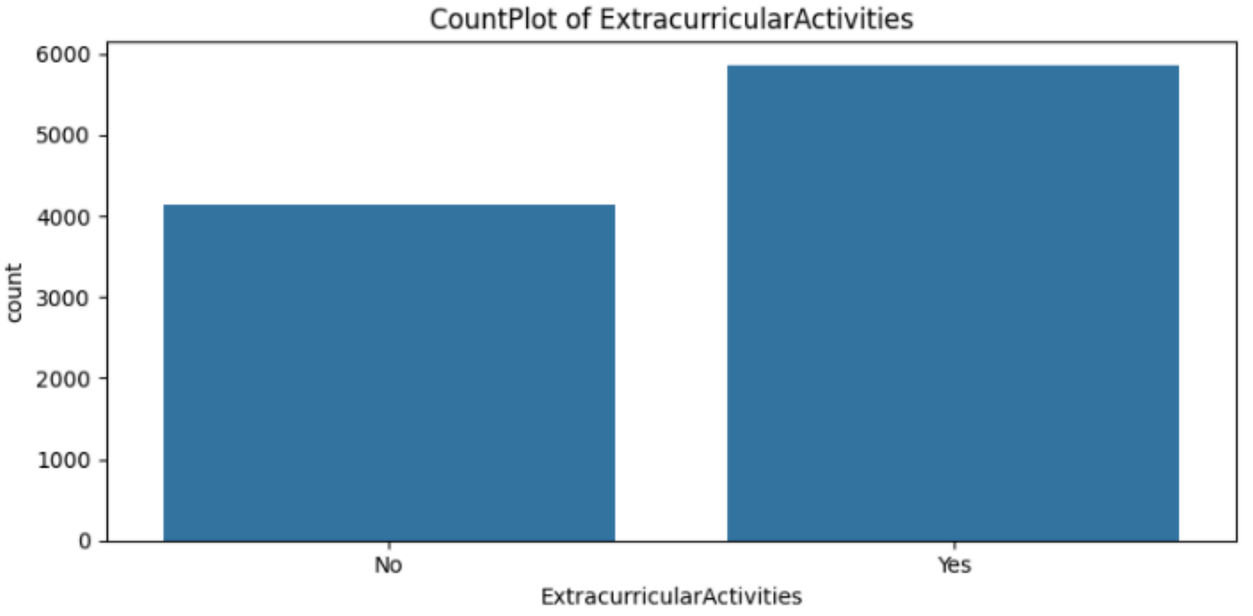




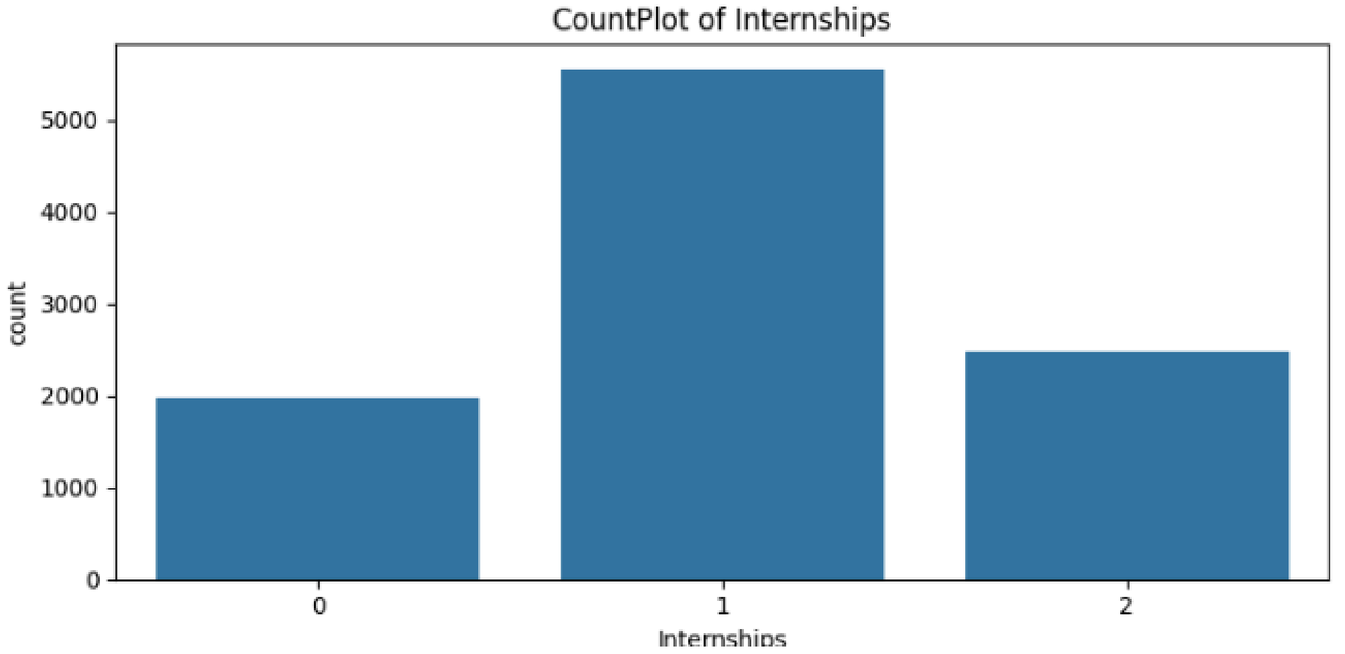

Histogram of SoftSkillsRating
2000
1750
1500
1250
1000
750
500
250
0
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75

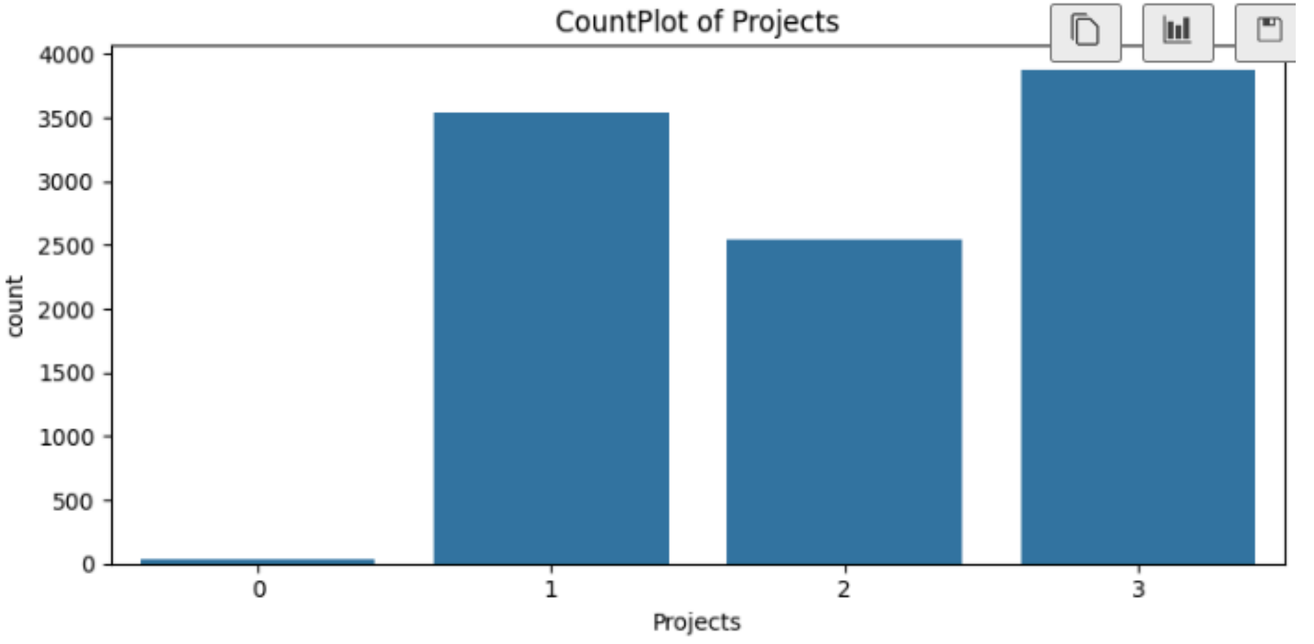

Histogram of AptitudeTestScore
Count
2000
1750
1500
1250
1000
750
500
250
0
60
65
70
75
80
85
90
AptitudeTestScore



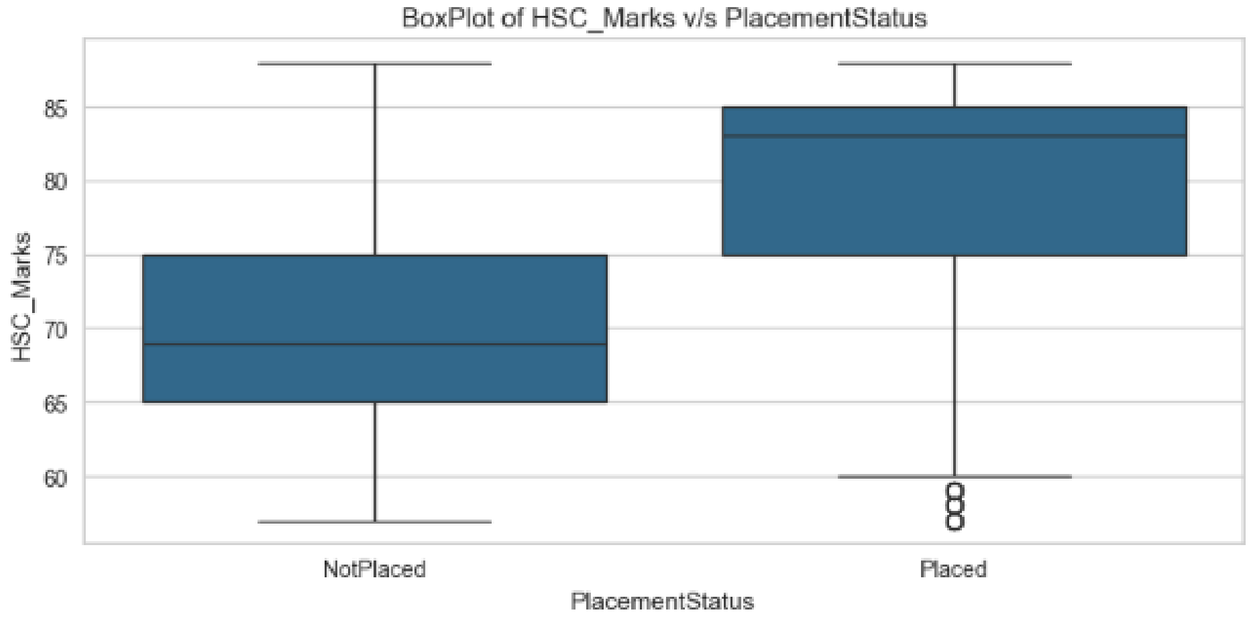


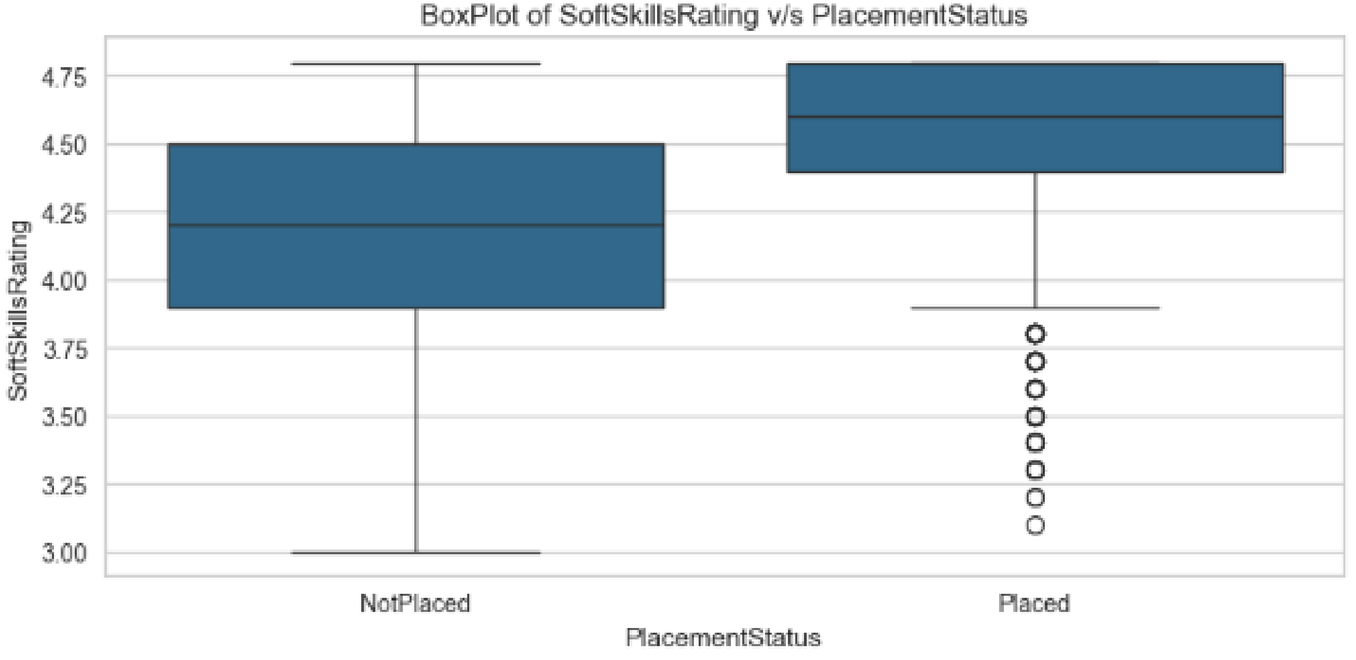
CountPlot of PlacementTraining

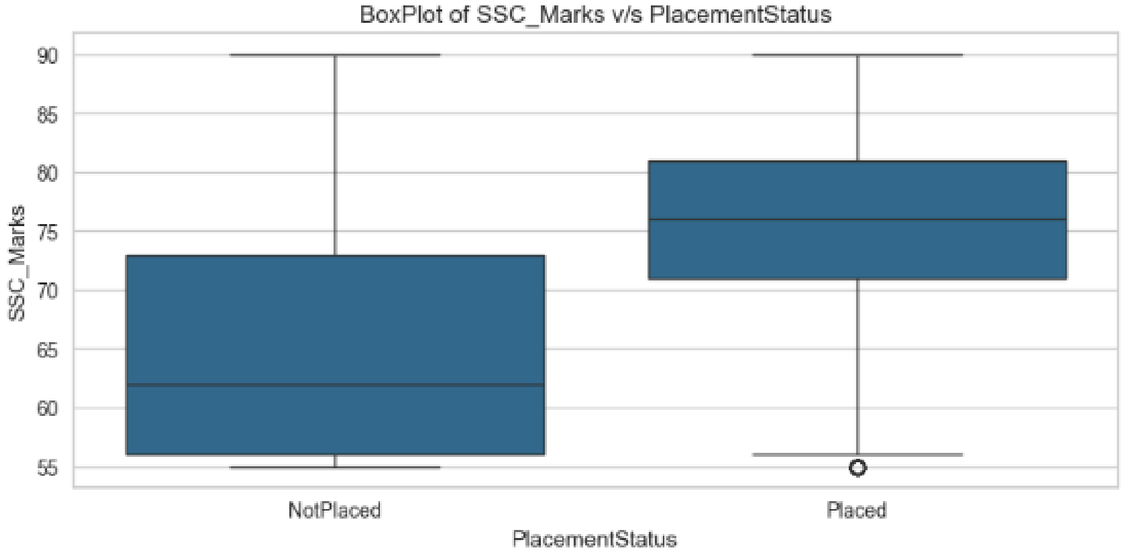

CountPlot of Internships
count
5000
4000
3000
2000
1000
0
0
1
2
Internships

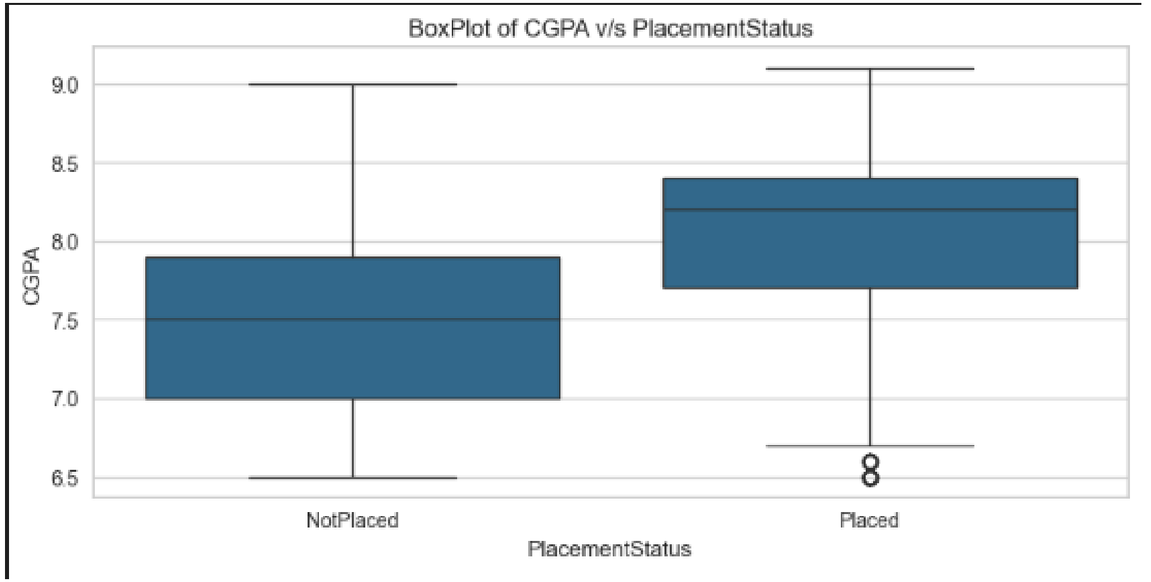

CountPlot of Projects
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
Projects

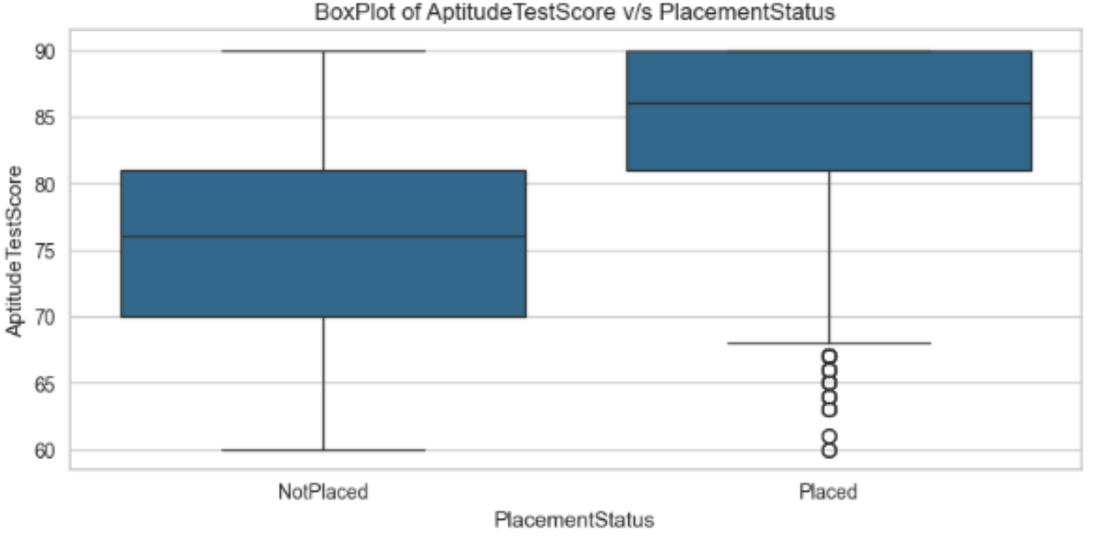

CountPlot of ExtracurricularActivities


BoxPlot of HSC_Marks v/s PlacementStatus
HSC_Marks
85
80
75
70
65
60
NotPlaced
Placed
PlacementStatus

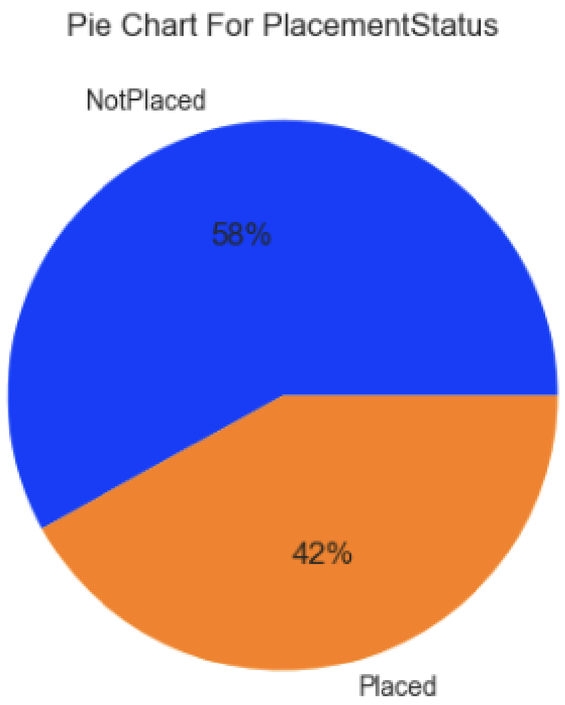

BoxPlot of SoftSkillsRating v/s PlacementStatus
SoftSkillsRating
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
NotPlaced
Placed
PlacementStatus

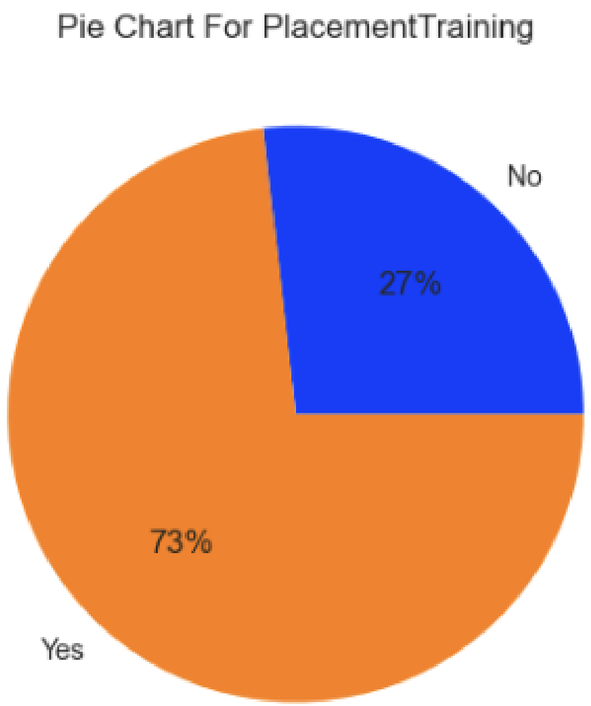

BoxPlot of SSC_Marks v/s PlacementStatus
SSC_Marks
90
85
80
75
70
65
60
55
NotPlaced
Placed
PlacementStatus

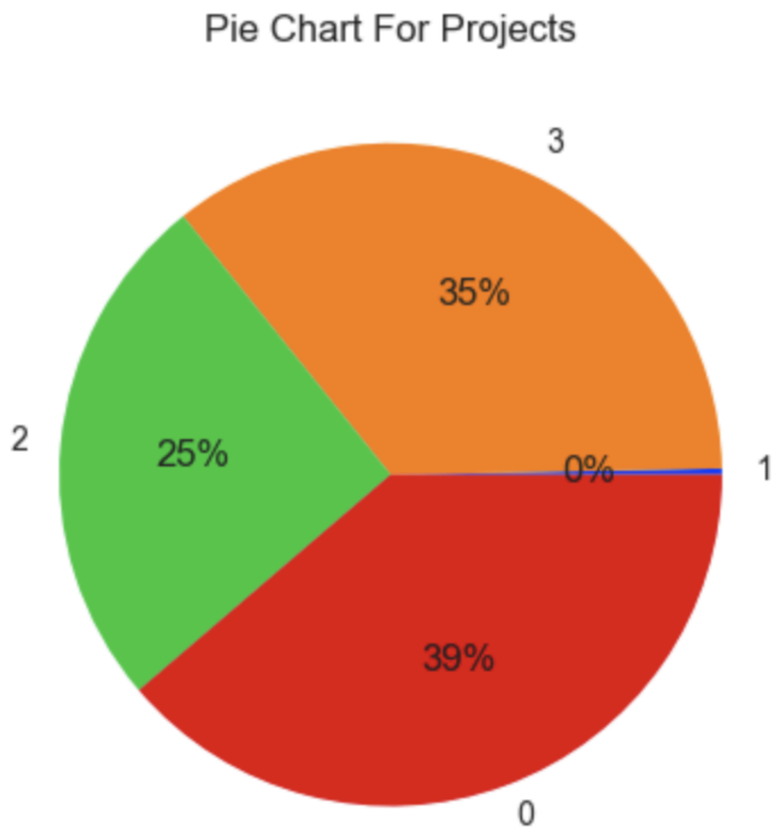

BoxPlot of CGPA v/s PlacementStatus
CGPA
9.0
8.5
8.0
7.5
7.0
6.5
NotPlaced
Placed
PlacementStatus

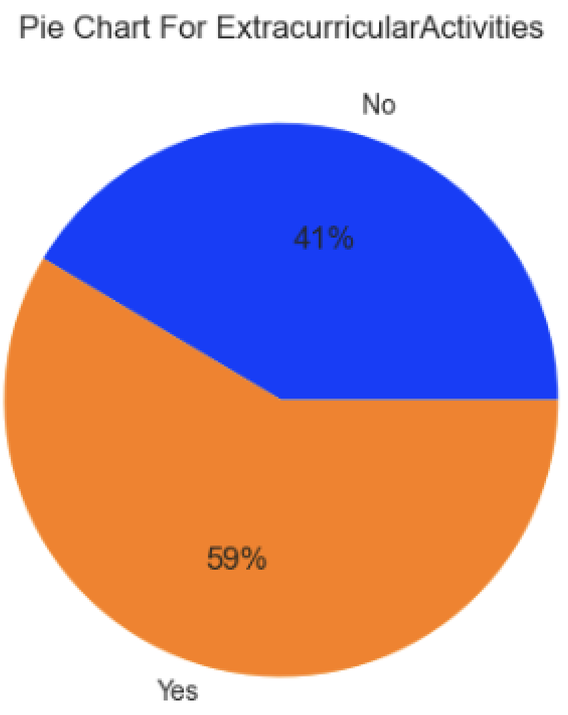

BoxPlot of AptitudeTestScore v/s PlacementStatus
AptitudeTestScore
90
85
80
75
70
65
60
NotPlaced
Placed
PlacementStatus

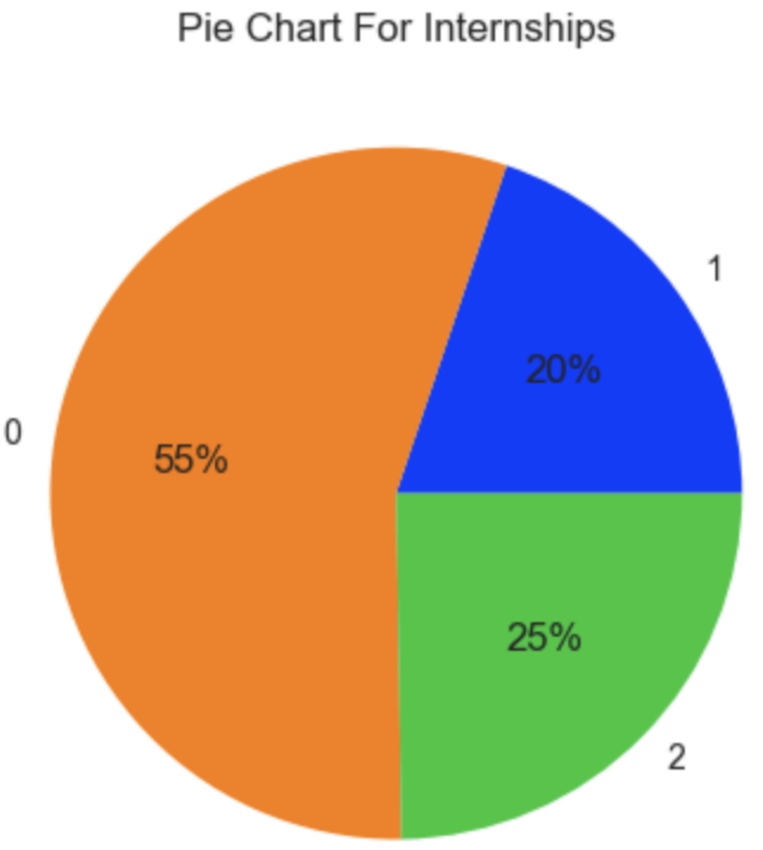
Percentage Distribution

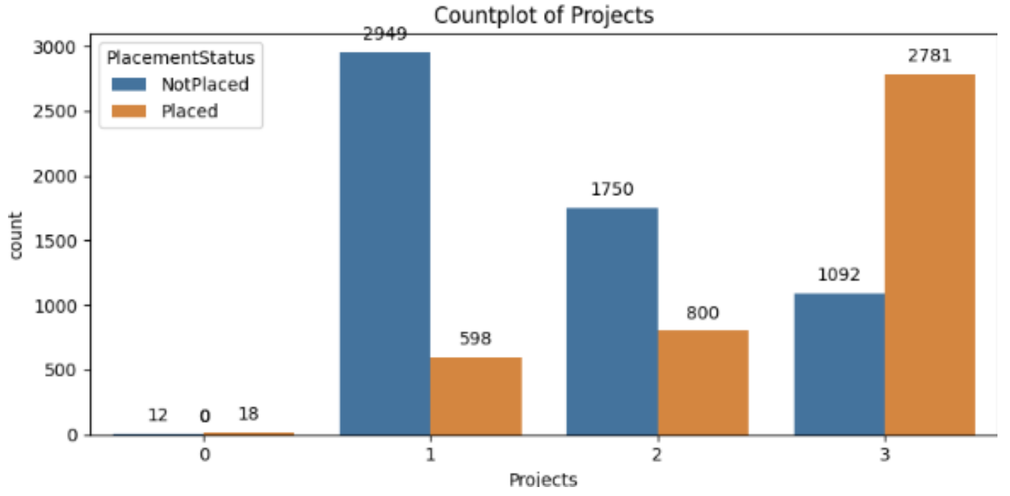


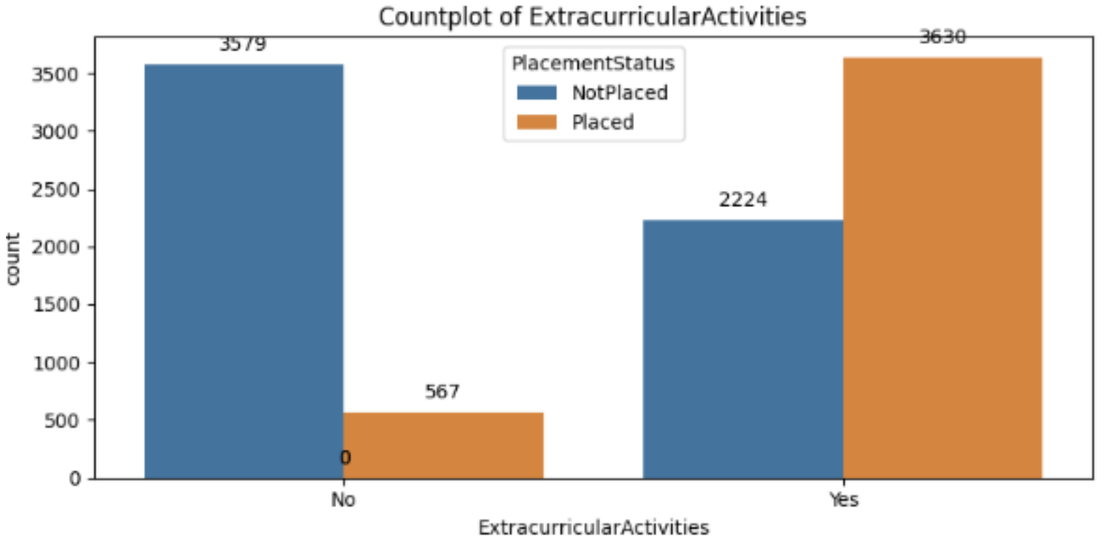


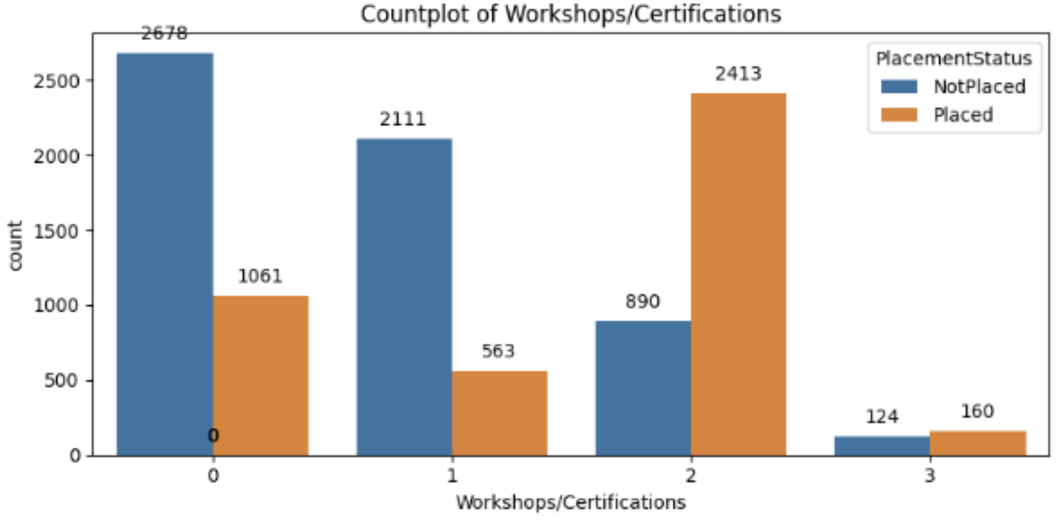


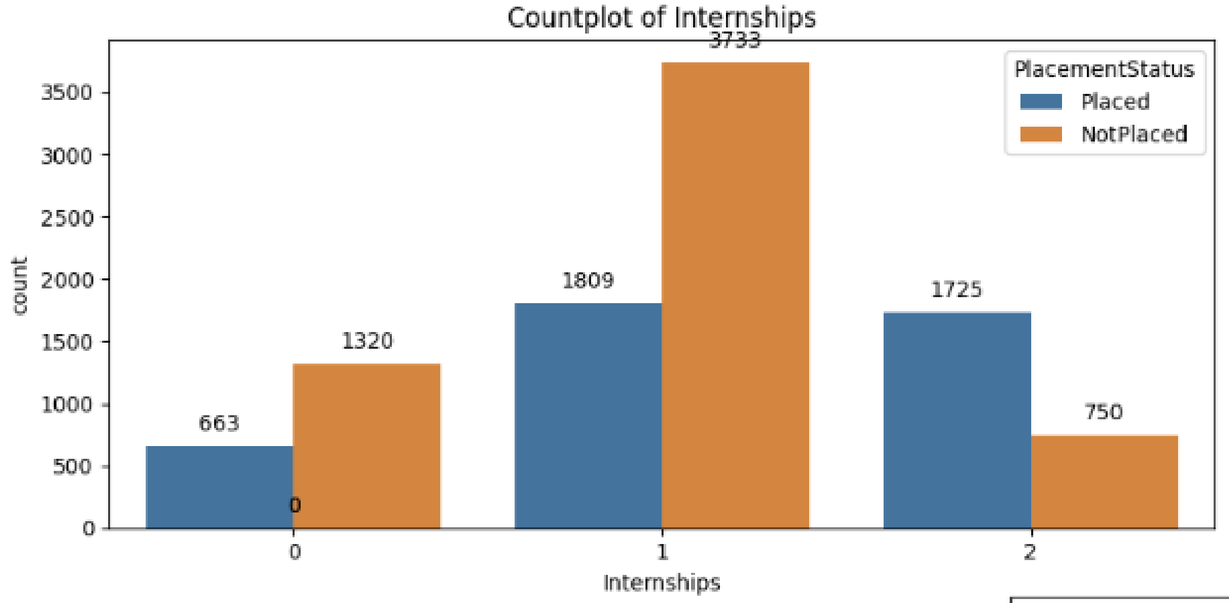

Pie Chart For ExtracurricularActivities
No
41%
59%
Yes

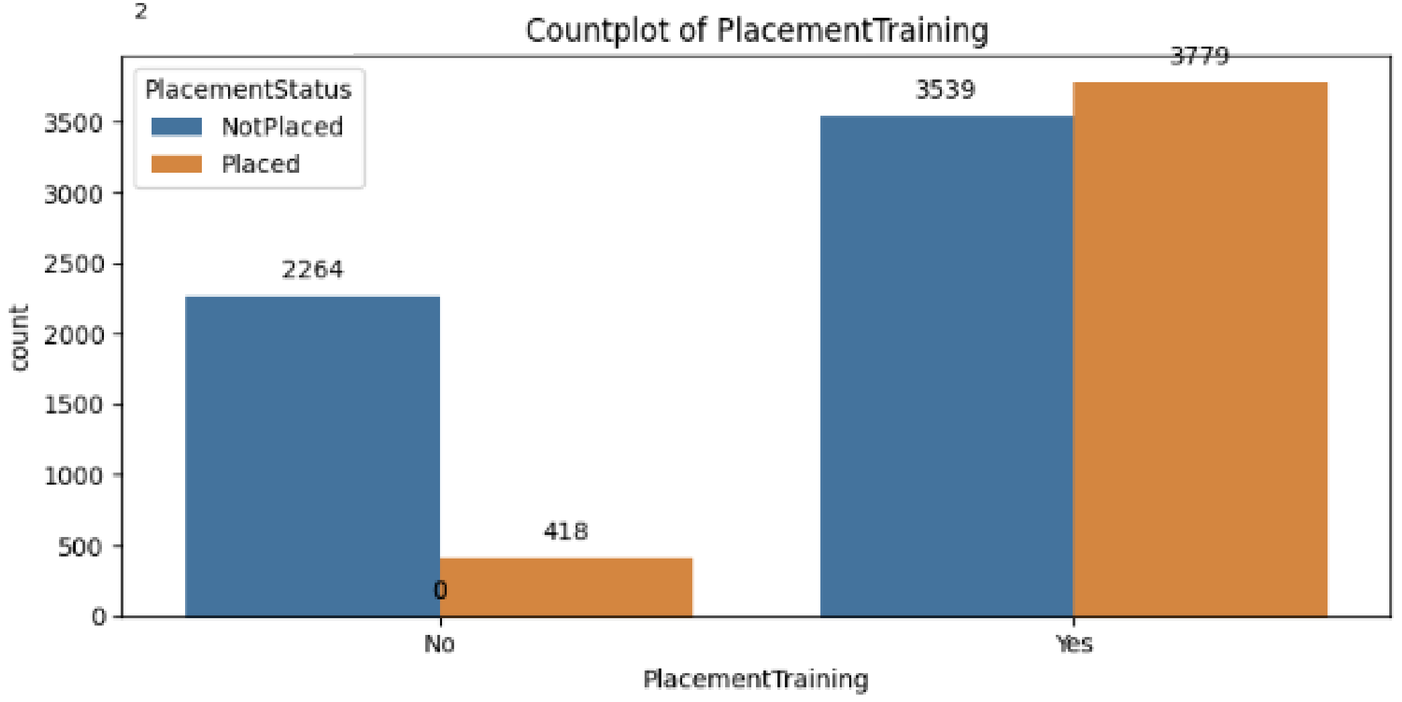

Pie Chart For Internships
1
20%
0
55%
25%
2


Countplot of Projects
PlacementStatus
NotPlaced
Placed
count
Projects
12
0
18
2949
598
1750
800
1092
2781

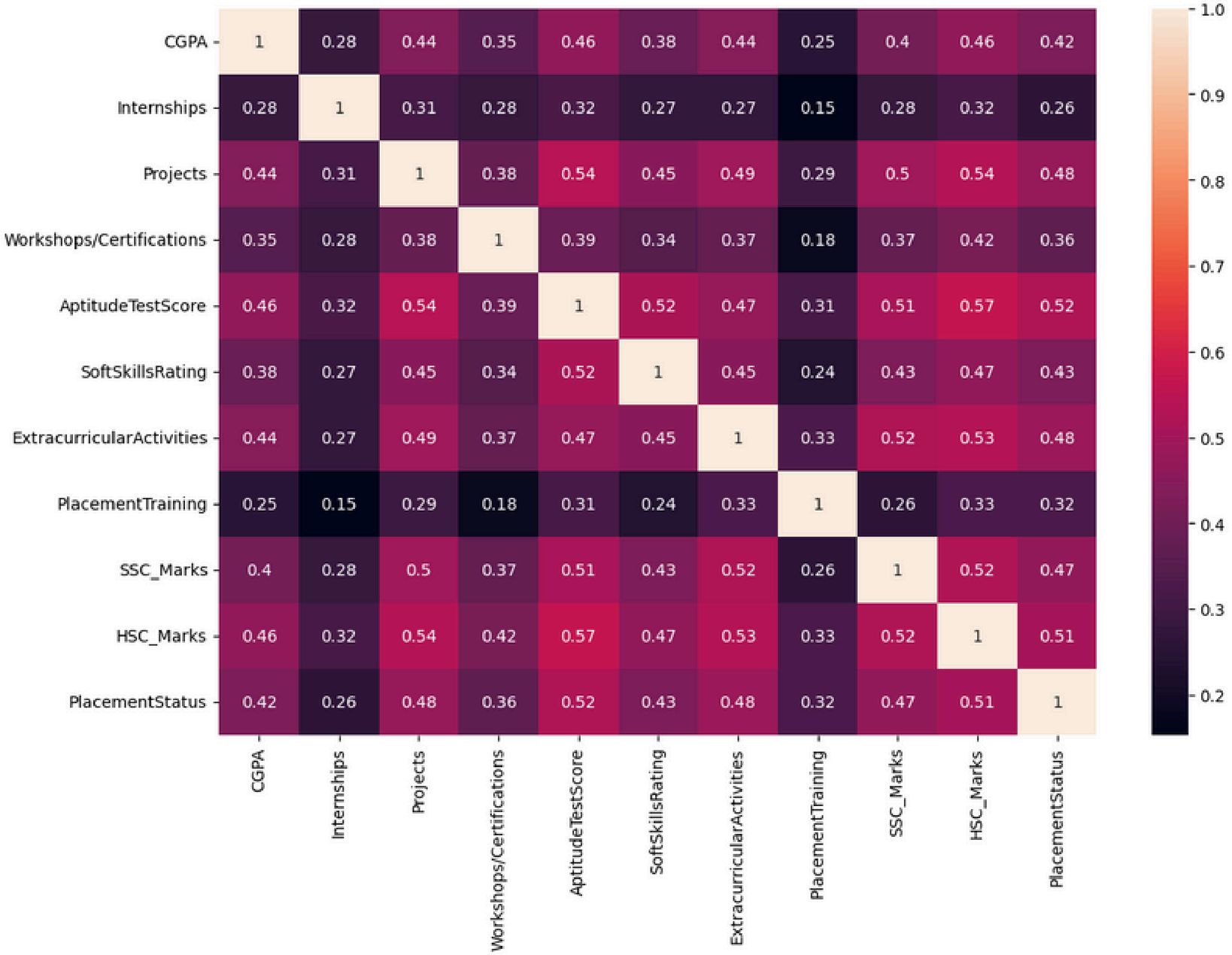

Countplot of ExtracurricularActivities
PlacementStatus
NotPlaced
Placed
count
ExtracurricularActivities
No
Yes
3579
0
567
2224
3630


Countplot of Workshops/Certifications
PlacementStatus
NotPlaced
Placed
count
Workshops/Certifications
2678
0
1061
2111
563
890
2413
124
160


Countplot of Internships
PlacementStatus
Placed
NotPlaced
count
Internships
0
1
2
663
1320
1809
3733
1725
750
0


Countplot of PlacementTraining
PlacementStatus
NotPlaced
Placed
count
PlacementTraining
No
Yes
2264
418
0
3539
3779

HeatMap


CGPA 1 0.28 0.44 0.35 0.46 0.38 0.44 0.25 0.4 0.46 0.42
Internships 0.28 1 0.31 0.28 0.32 0.27 0.27 0.15 0.28 0.32 0.26
Projects 0.44 0.31 1 0.38 0.54 0.45 0.49 0.29 0.5 0.54 0.48
Workshops/Certifications 0.35 0.28 0.38 1 0.39 0.34 0.37 0.18 0.37 0.42 0.36
AptitudeTestScore 0.46 0.32 0.54 0.39 1 0.52 0.47 0.31 0.51 0.57 0.52
SoftSkillsRating 0.38 0.27 0.45 0.34 0.52 1 0.45 0.24 0.43 0.47 0.43
ExtracurricularActivities 0.44 0.27 0.49 0.37 0.47 0.45 1 0.33 0.52 0.53 0.48
PlacementTraining 0.25 0.15 0.29 0.18 0.31 0.24 0.33 1 0.26 0.33 0.32
SSC_Marks 0.4 0.28 0.5 0.37 0.51 0.43 0.52 0.26 1 0.52 0.47
HSC_Marks 0.46 0.32 0.54 0.42 0.57 0.47 0.53 0.33 0.52 1 0.51
PlacementStatus 0.42 0.26 0.48 0.36 0.52 0.43 0.48 0.32 0.47 0.51 1
CGPA
Internships
Projects
Workshops/Certifications
AptitudeTestScore
SoftSkillsRating
ExtracurricularActivities
PlacementTraining
SSC_Marks
HSC_Marks
PlacementStatus
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Models used for making predictions

In our project, we have used 4 models to conduct a comparative study, aiming to analyze differences in accuracy. Models that we have used are:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. K- Nearest Neighbor

.

Logistic Regression

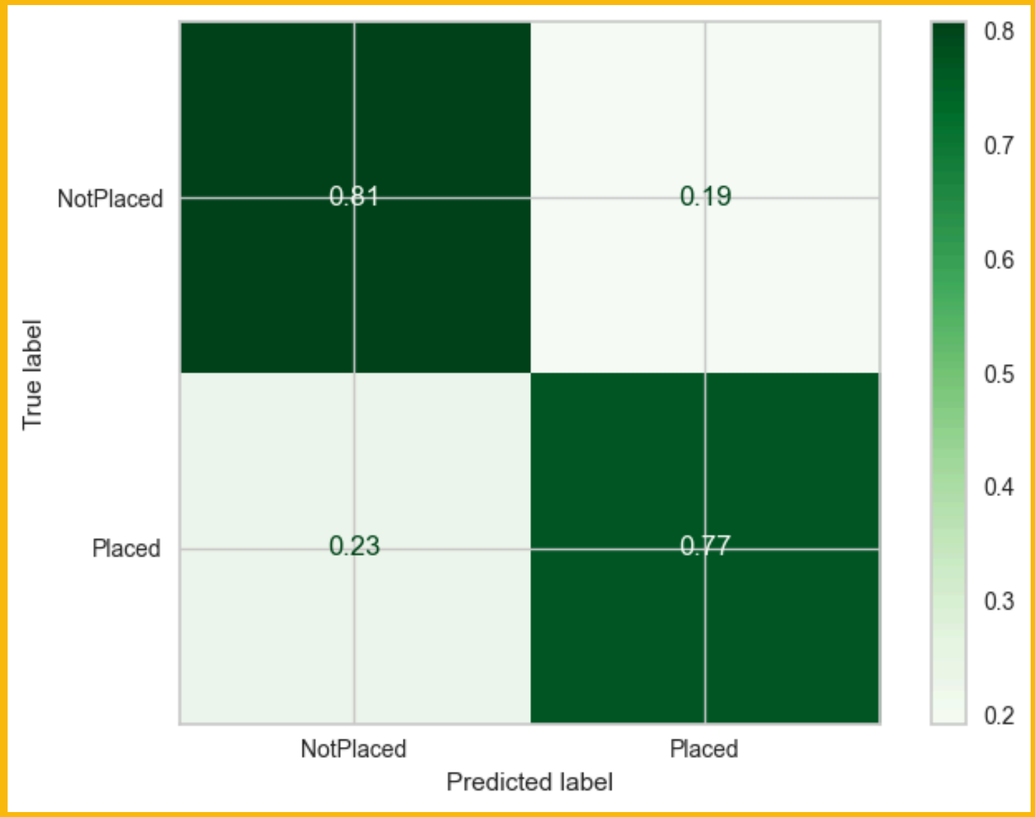
Logistic regression is a statistical method used for binary classification tasks. It predicts the probability of a categorical outcome by fitting data to a logistic function, transforming values into probabilities between 0 and 1.

.

This model scored an accuracy of approximately 79.28%.

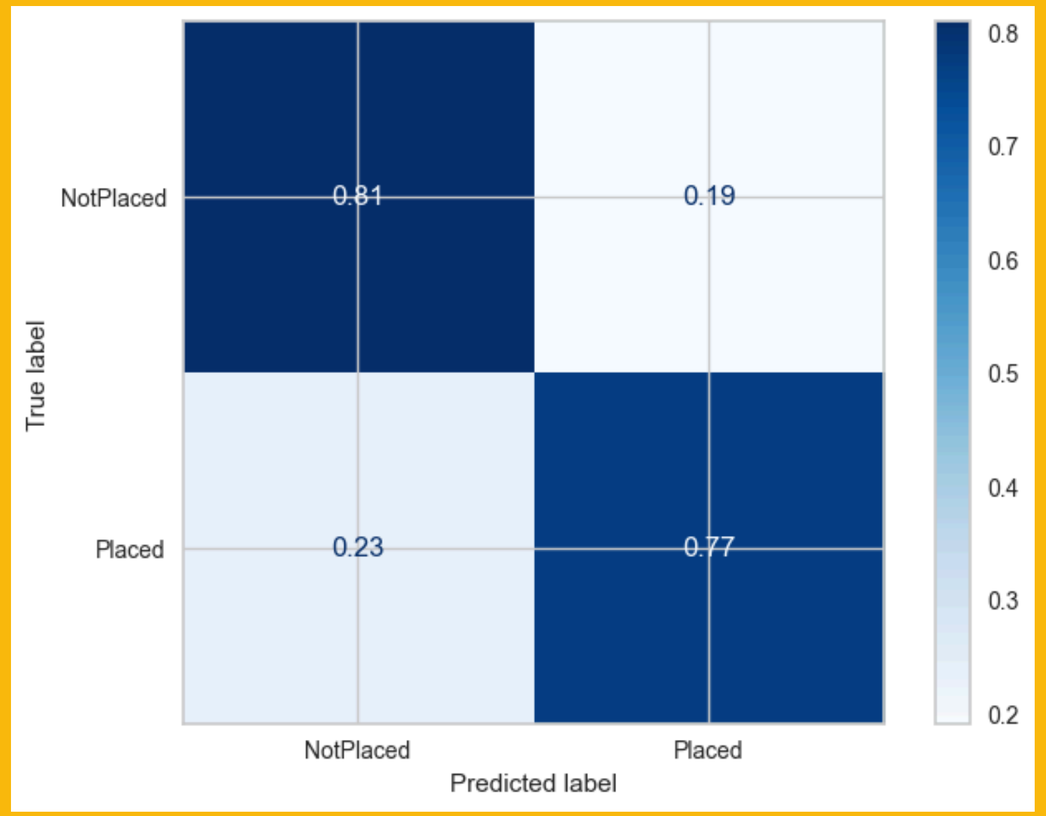
Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.83	0.81	0.82	1471
Placed	0.74	0.77	0.75	1029
Accuracy			0.7928	2500
Macro Avg	0.79	0.79	079	2500
Weighted Avg	0.80	0.79	079	2500



With Normalization Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	.0.83	0.81	0.82	1471
Placed	0.74	0.77	0.75	1029
Accuracy			0.7924	2500
Macro Avg	0.79	0.79	079	2500
Weighted Avg	0.79	0.79	079	2500



Accuracy of Model 79.24%

Decision Tree

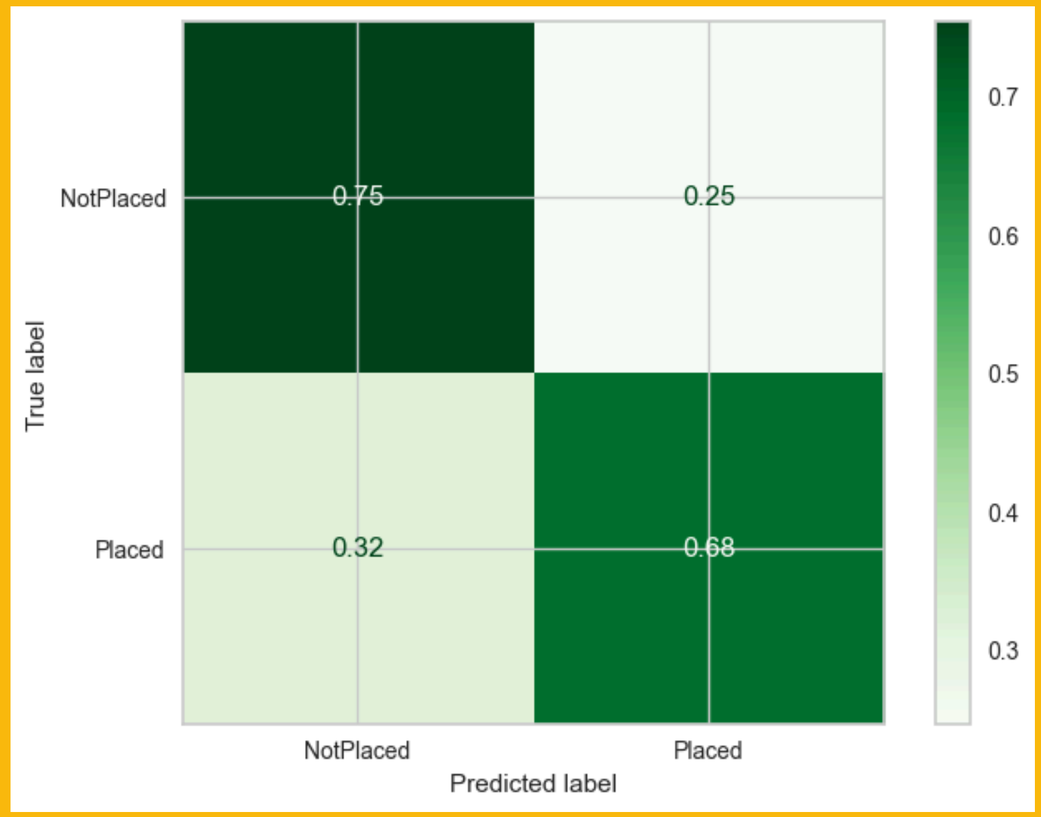
In machine learning, a decision tree is a model that resembles a tree-like structure used for classification and regression tasks. It starts with a root node that represents the entire dataset and splits the data into branches based on specific features and their values. This process continues down the tree until reaching leaf nodes, which provide final decisions or predictions.

.

This model scored an accuracy of approximately 72.52%.

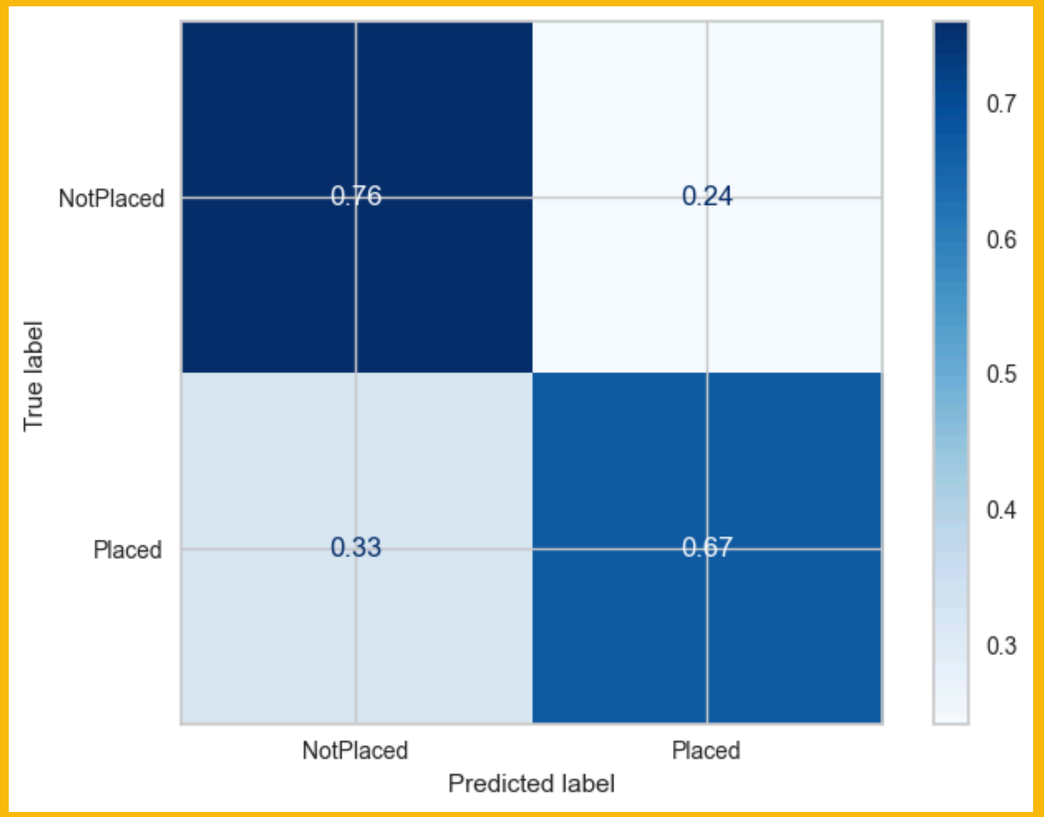
Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.77	0.75	0.76	1471
Placed	0.66	0.68	0.67	1029
Accuracy			0.7252	2500
Macro Avg	0.72	0.72	0.72	2500
Weighted Avg	0.73	0.72	0.73	2500



With Normalization Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.77	0.76	0.76	1471
Placed	0.66	0.67	0.67	1029
Accuracy			0.7248	2500
Macro Avg	0.72	0.72	0.72	2500
Weighted Avg	0.73	0.72	0.72	2500



Accuracy Of Model Is 72.48%

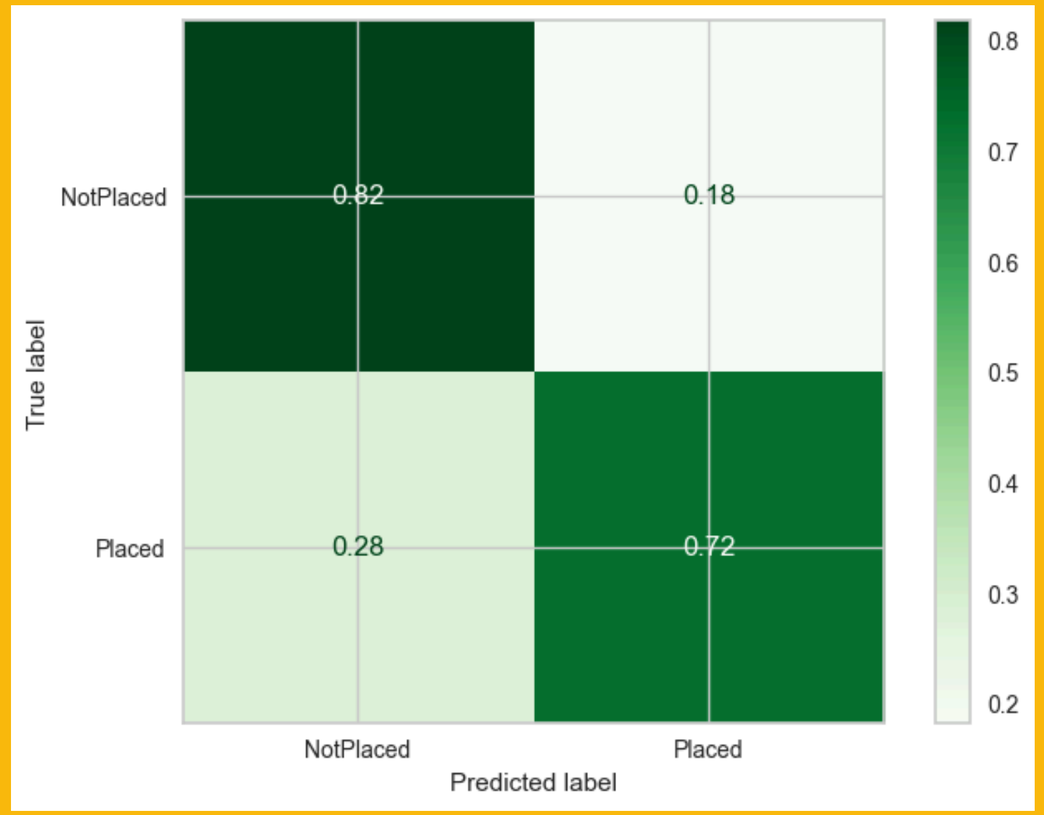
K Neighbour Classifier

The K-Nearest Neighbors (KNN) Classifier is a straightforward yet effective supervise machine learning algorithm for classification tasks. It operates on the principle of similarity classifying new data points based on their proximity to labeled data points in the trainin set. The algorithm determines the class of a new data point by identifying its K neares neighbors in the feature space using a distance metric (like Euclidean distance).

This model scored an accuracy of approximately 77.80%.

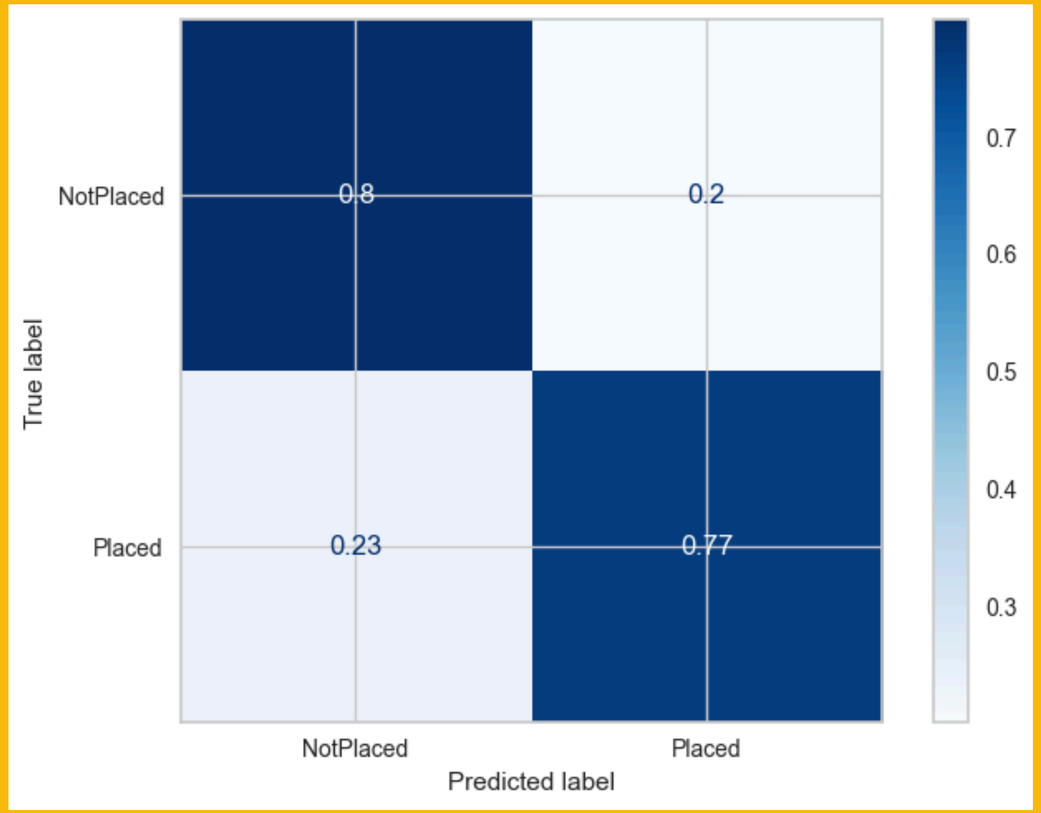
Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.81	0.82	0.81	1471
Placed	0.73	0.72	0.73	1029
Accuracy			0.7780	2500
Macro Avg	0.77	0.77	0.77	2500
Weighted Avg	0.78	0.78	0.78	2500



With Normalization Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.83	0.80	0.81	1471
Placed	0.73	0.77	0.75	1029
Accuracy			0.7864	2500
Macro Avg	0.78	0.78	0.78	2500
Weighted Avg	0.79	0.79	0.79	2500



Accuracy Of Model Is 78.64%

RANDOM FOREST

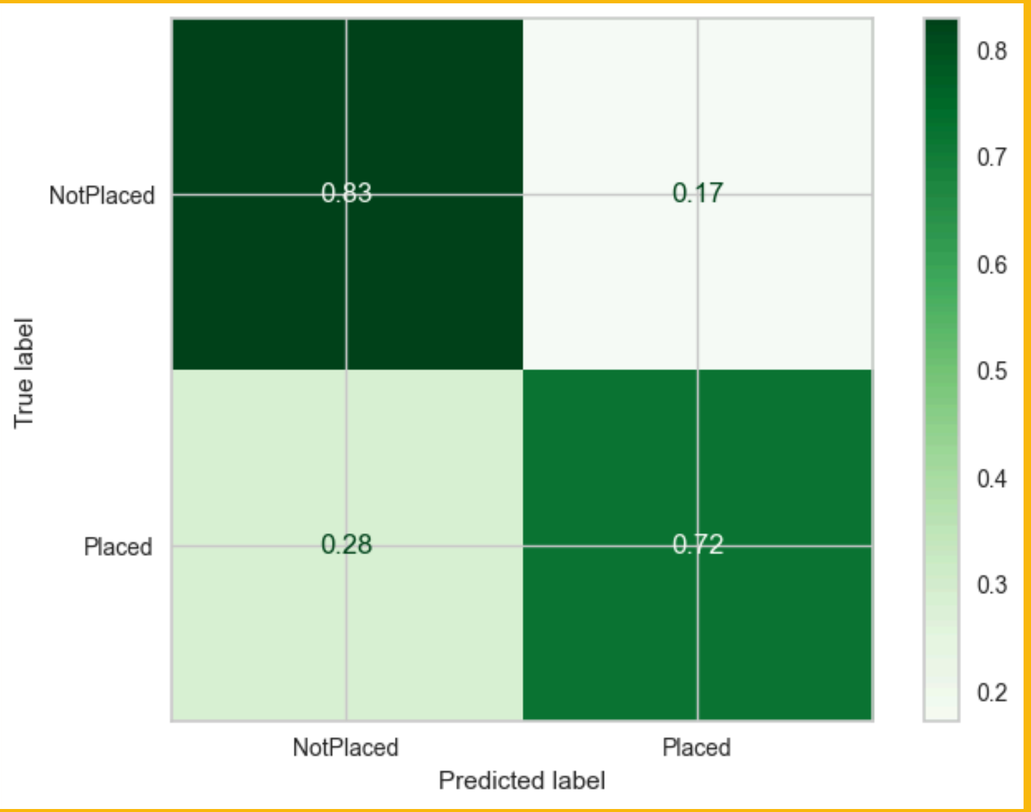
Random Forest is an ensemble learning method that constructs multiple decision trees during training.

- Each tree in the forest is trained on a random subset of the training data and a random subset of the features.
- During prediction, each tree in the forest independently makes a prediction, and the final prediction is determined by aggregating the predictions of all trees.
- Random Forest is robust to overfitting and tends to generalize well to unseen data due to the diversity of trees in the forest and the randomness introduced during training.

This model scored an accuracy of approximately 78.48%.

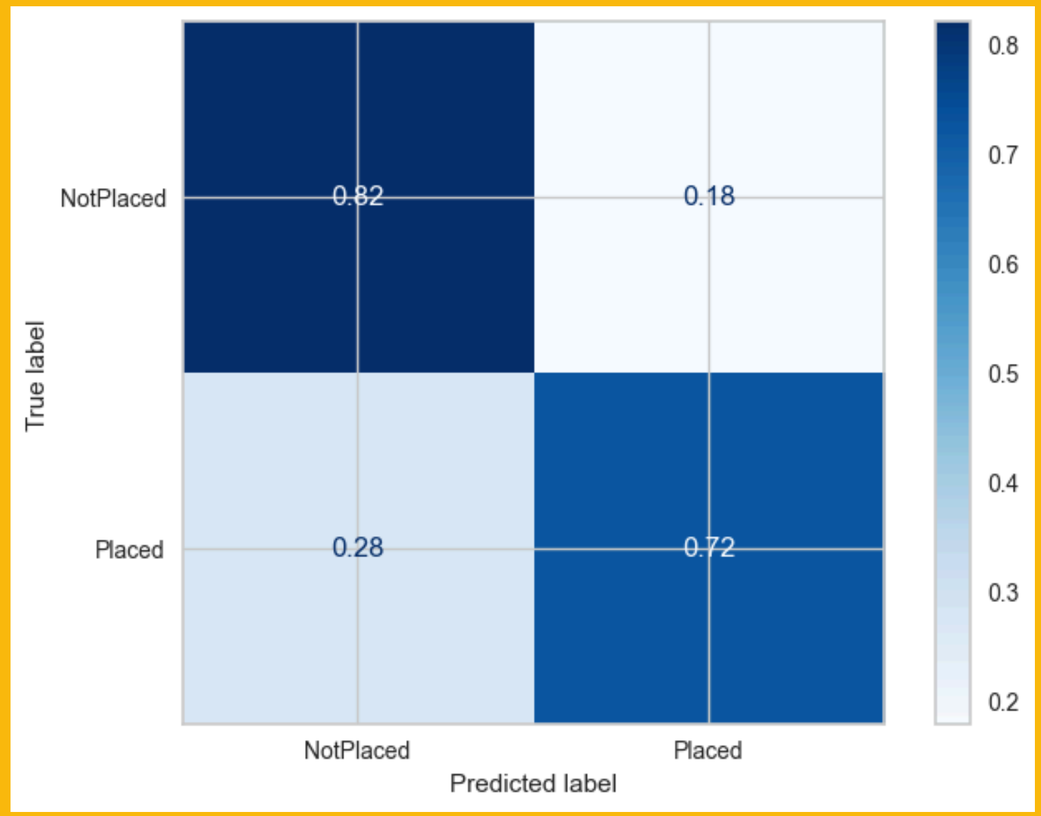
Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.81	0.83	0.82	1471
Placed	0.75	0.72	0.73	1029
Accuracy			0.7848	2500
Macro Avg	0.78	0.77	0.78	2500
Weighted Avg	0.78	0.78	0.78	2500



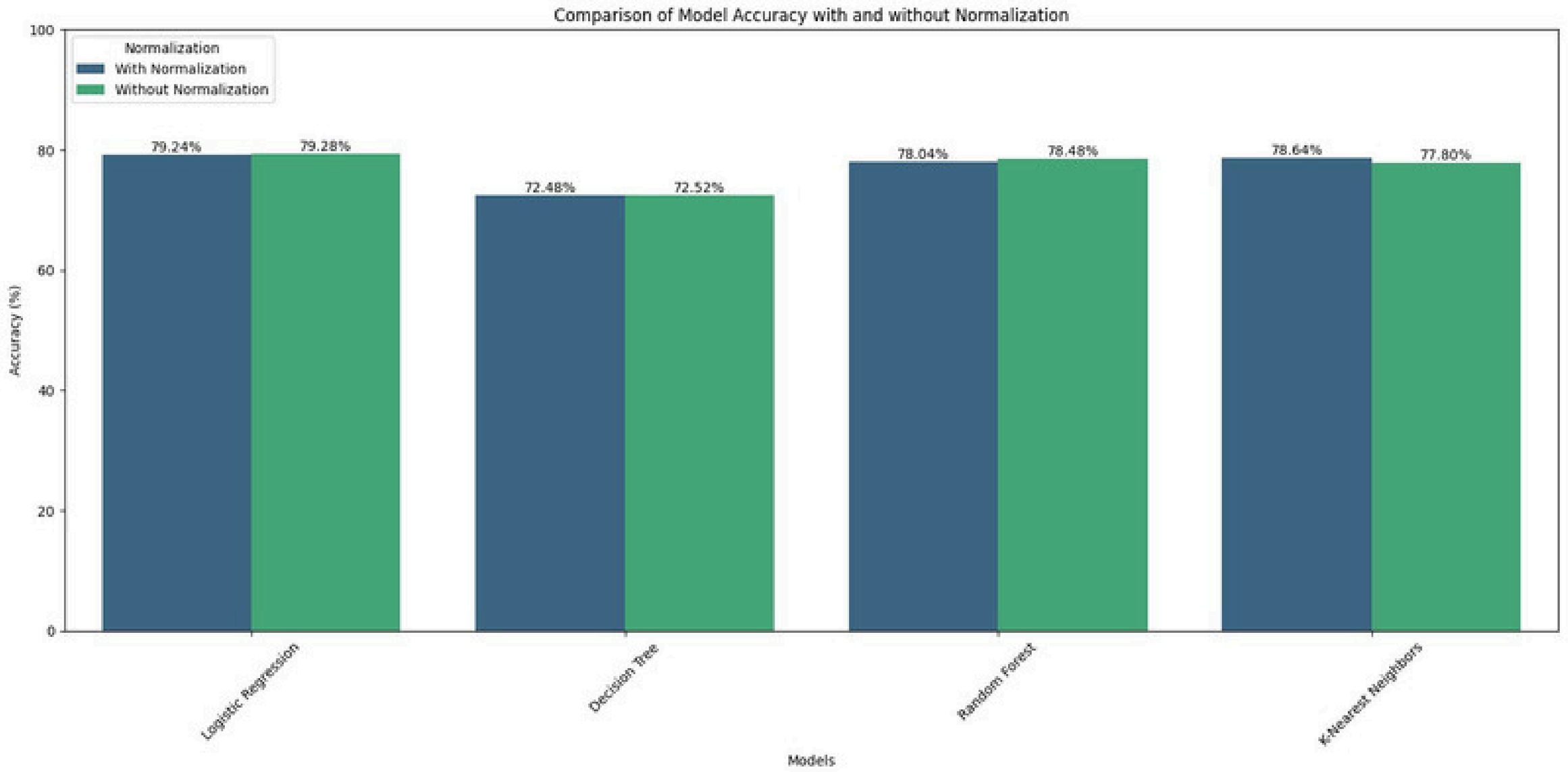
With Normalization Confusion Matrix

	Precision	Recall	F1-score	Support
Not Placed	0.81	0.82	0.82	1471
Placed	0.75	0.72	0.73	1029
Accuracy			0.7804	2500
Macro Avg	0.77	0.77	0.77	2500
Weighted Avg	0.78	0.78	0.78	2500



Accuracy Of Model Is 78.04%

Comparision Between Models



PyCaret Library

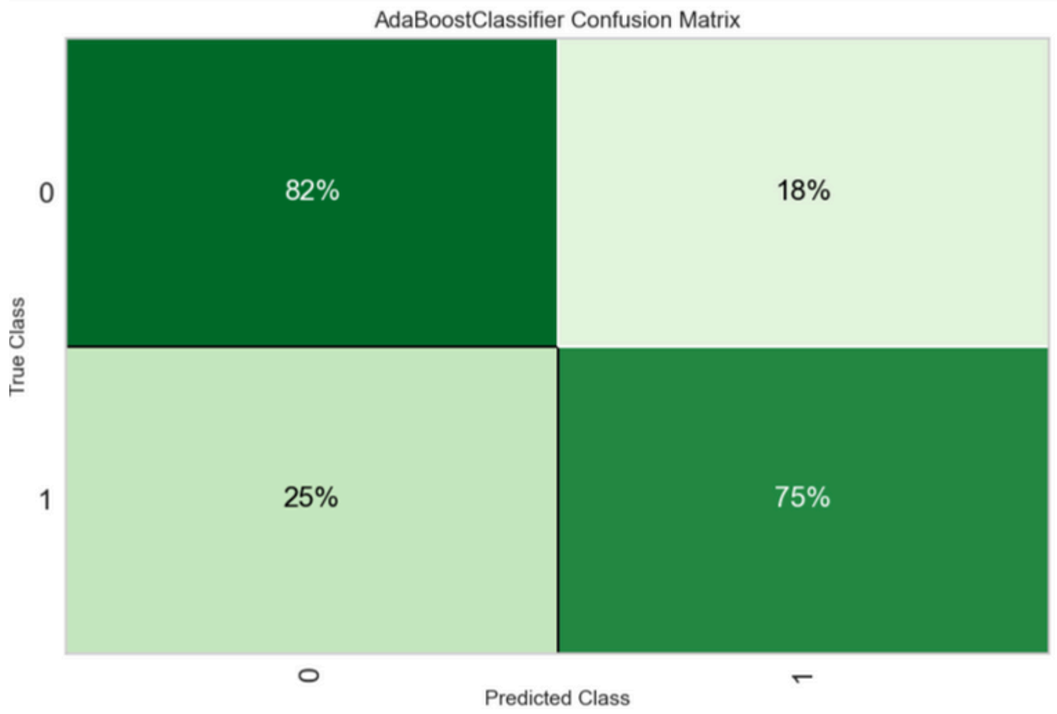
	Description	Value
0	Session id	123
1	Target	PlacementStatus
2	Target type	Binary
3	Original data shape	(10000, 11)
4	Transformed data shape	(10000, 11)
5	Transformed train set shape	(7000, 11)
6	Transformed test set shape	(3000, 11)
7	Numeric features	10
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	8625

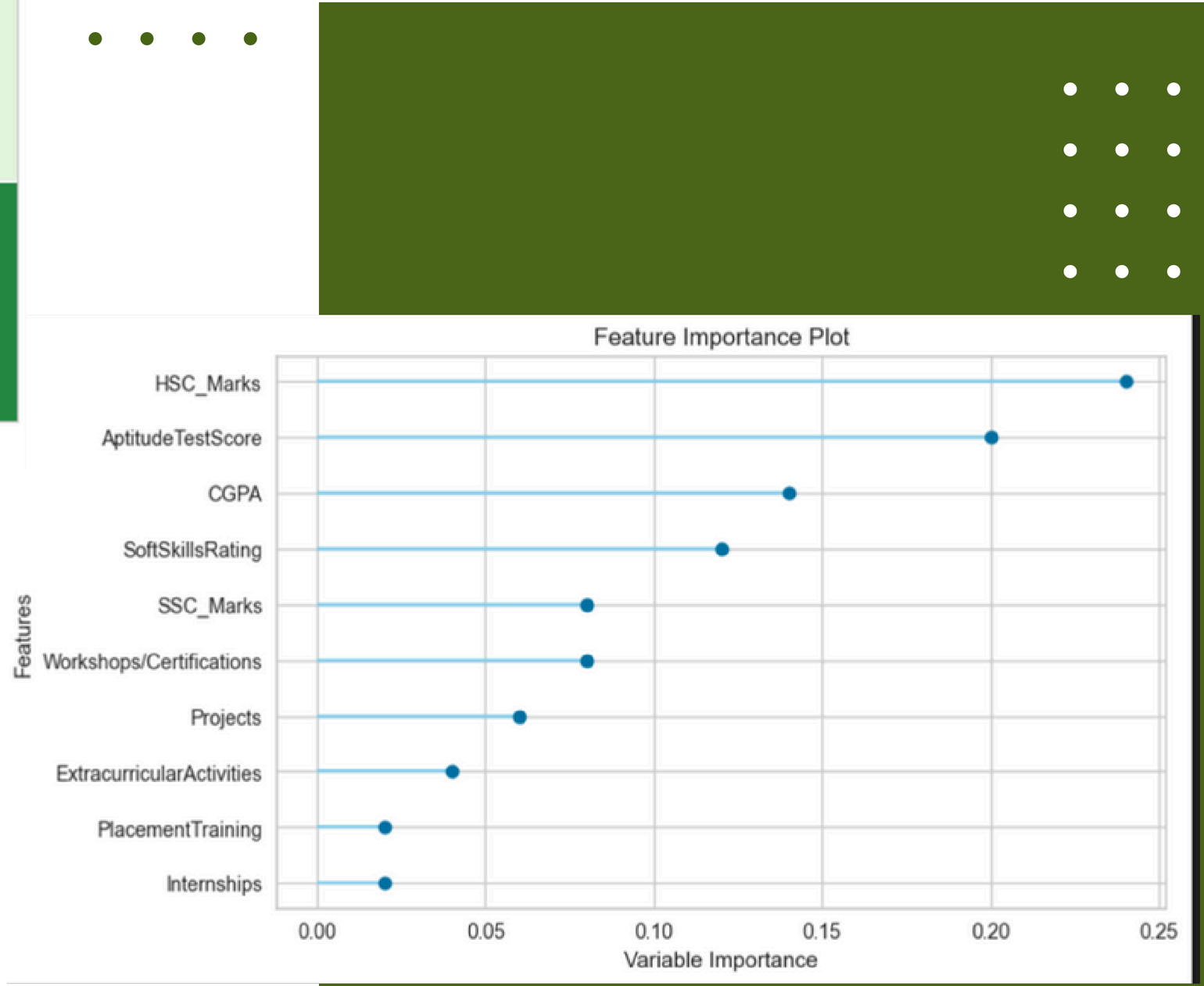
```
from pycaret.classification import *
pycrt = setup(df_min,target='PlacementStatus', session_id=123)
```

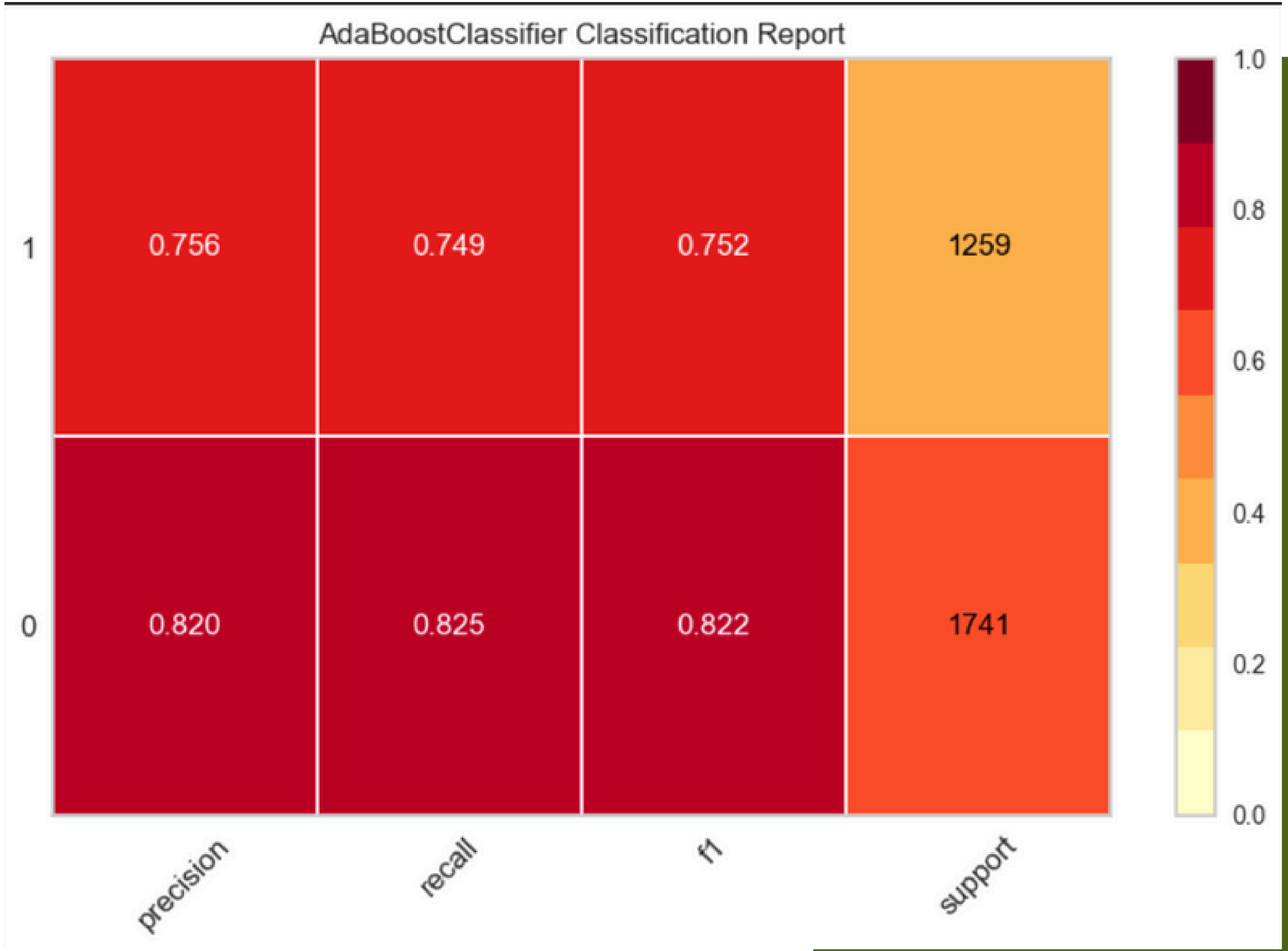
✓ 3.1s

PyCaret Library

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.8020	0.8813	0.7549	0.7692	0.7619	0.5925	0.5927	
lr	Logistic Regression	0.8019	0.8806	0.7685	0.7617	0.7650	0.5937	0.5939	
gbc	Gradient Boosting Classifier	0.8000	0.8801	0.7430	0.7719	0.7570	0.5872	0.5877	
lda	Linear Discriminant Analysis	0.7996	0.8795	0.7839	0.7502	0.7665	0.5911	0.5917	
ridge	Ridge Classifier	0.7993	0.0000	0.7828	0.7501	0.7660	0.5904	0.5910	
nb	Naive Bayes	0.7981	0.8781	0.7941	0.7432	0.7676	0.5896	0.5909	
lightgbm	Light Gradient Boosting Machine	0.7951	0.8747	0.7352	0.7673	0.7507	0.5770	0.5776	
rf	Random Forest Classifier	0.7934	0.8712	0.7223	0.7714	0.7457	0.5721	0.5733	
qda	Quadratic Discriminant Analysis	0.7929	0.8680	0.7454	0.7576	0.7511	0.5738	0.5742	
svm	SVM - Linear Kernel	0.7907	0.0000	0.7492	0.7622	0.7474	0.5698	0.5784	
et	Extra Trees Classifier	0.7880	0.8658	0.7134	0.7658	0.7383	0.5605	0.5619	
knn	K Neighbors Classifier	0.7749	0.8348	0.7260	0.7349	0.7301	0.5370	0.5374	
dt	Decision Tree Classifier	0.7097	0.7041	0.6658	0.6513	0.6582	0.4060	0.4063	
dummy	Dummy Classifier	0.5803	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	


AdaBoostClassifier Confusion Matrix
0
1
82%
18%
25%
75%
True Class
0
1
Predicted Class


Feature Importance Plot
HSC_Marks
AptitudeTestScore
CGPA
SoftSkillsRating
SSC_Marks
Workshops/Certifications
Projects
ExtracurricularActivities
PlacementTraining
Internships
Features
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance


AdaBoostClassifier Classification Report
1
0
0.756
0.749
0.752
1259
0.820
0.825
0.822
1741
precision
recall
f1
support
1.0
0.8
0.6
0.4
0.2
0.0

Conclusion & Future Work

- Primary objective achieved: Successfully predicted students' placement status post-final year using four classification algorithms (LR, DTC, KNN, RFC).
- In Future work we can apply this predictor Model on real time application instead of synthetic data.
- System's efficacy: Elevates institution's placement rates. Enhances institution's reputation.
- Signifies a substantial advancement in classification techniques for placement prediction.
- Stands as a pivotal tool to improve placement prediction methodologies significantly.

References and bibliography

1.Senthil Kumar Thangavel, Divya Bharathi P and Abijith Sankar: Student Placement Analyzer: A Recommendation System Using Machine Learning 2017 International Conference on Advanced Computing and Communication Systems.
2. Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System
Article in International Journal of Recent Technology And Engineering(IRJTE) September 2019:-By Abhishek Rao, NMAM Institute of Technology
3.2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) A Review on Student Placement Chance Prediction
4.2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) Campus Placement Predictive Analysis using Machine Learning
5.https://www.kaggle.com/datasets/chandhurubaskar/campus-placement-data-for-engineering-colleges/code?datasetId=3678158&sortBy=voteCount

THANK YOU