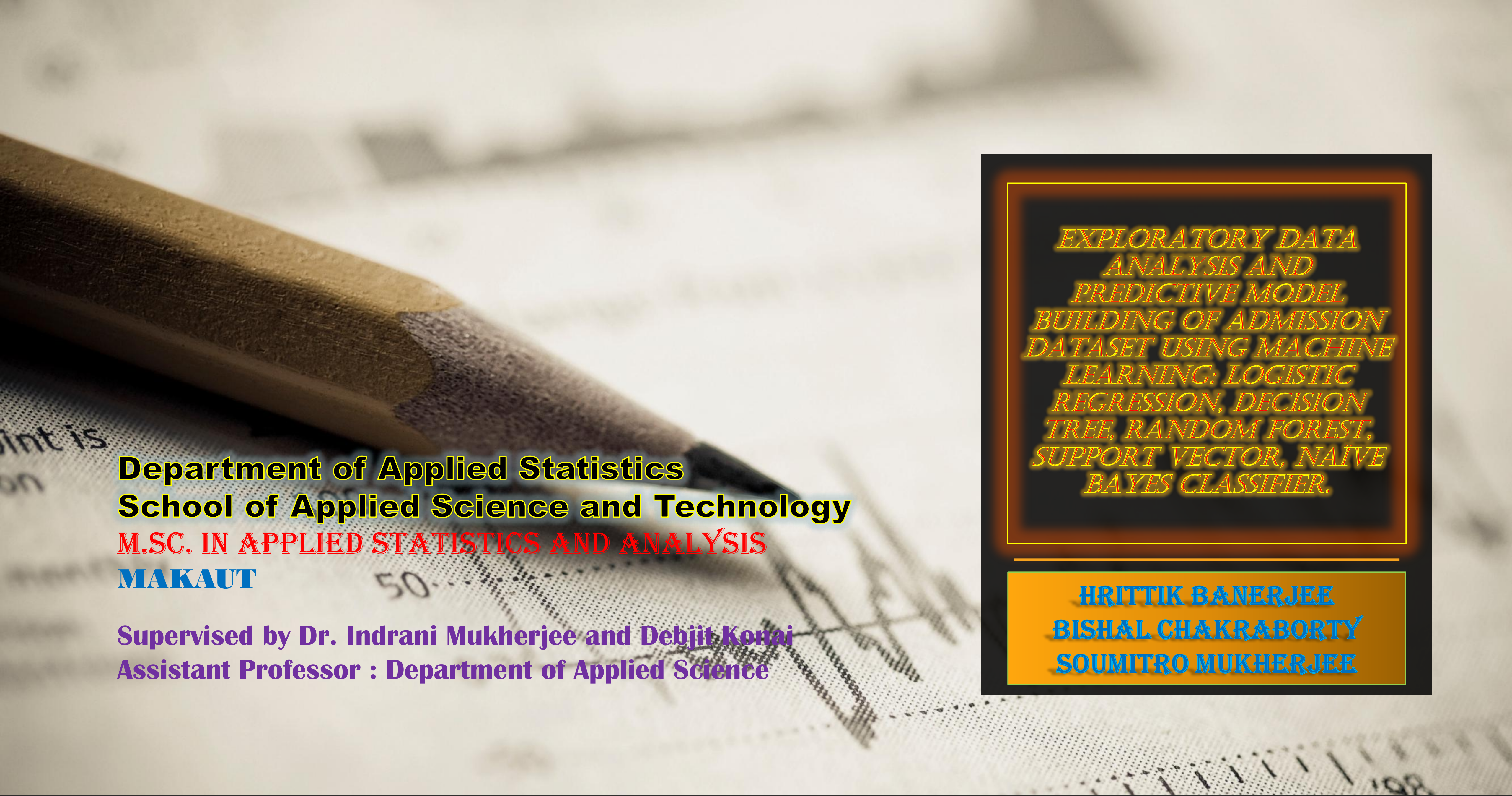


EXPLORATORY DATA ANALYSIS AND PREDICTIVE MODEL BUILDING OF ADMISSION DATASET USING MACHINE LEARNING: LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST, SUPPORT VECTOR, NAÏVE BAYES CLASSIFIER.

HRITTIK BANERJEE
BISHAL CHAKRABORTY
SOUMITRO MUKHERJEE

Department of Applied Statistics
School of Applied Science and Technology
M.SC. IN APPLIED STATISTICS AND ANALYSIS
MAKAUT

Supervised by Dr. Indrani Mukherjee and Debjit Konai
Assistant Professor : Department of Applied Science

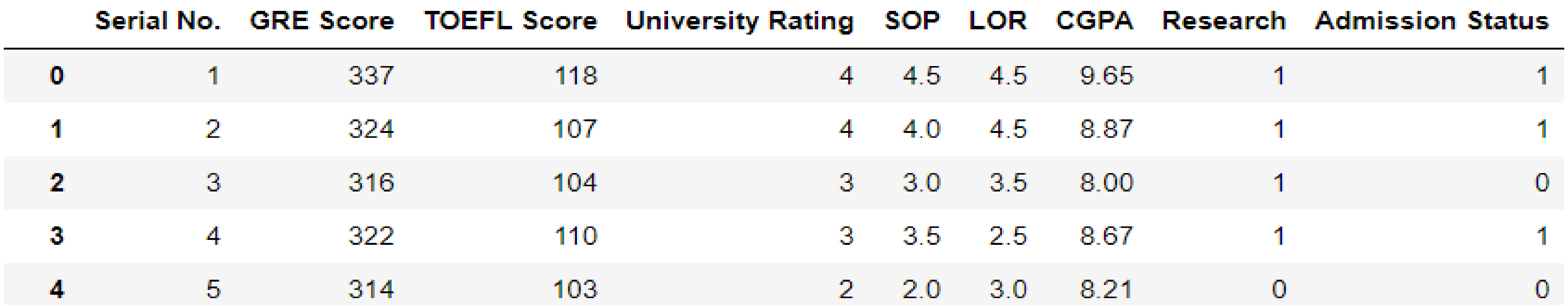
CONTENT

1. INTRODUCTION
2. DATA DESCRIPTION
3. METHODOLOGIES
4. DATA ANALYSIS
5. ML MODELLING
6. CONCLUSION
7. REFERENCE

INTRODUCTION

As we are studying MSc, we know the amount of pressure someone has to go through during admissions. There are different requirements for the admission process and after meeting them all one will get a chance into the MS course of a certain University and In this project, we have taken the dataset of graduate admissions and our goal is to find out some important features from this dataset through Explanatory Data Analysis (EDA) and then we applied Machine-learning classifiers on this data to observe which model best fits this data. We've applied five ML classifiers namely, Logistic Regression, Decision Tree, Random Forest, Support Vector (SVC), and Naive Bayes. Before moving to the next part of our project let's know more about the dataset.

DATA DESCRIPTION:



	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Admission Status
0	1	337	118	4	4.5	4.5	9.65	1	1
1	2	324	107	4	4.0	4.5	8.87	1	1
2	3	316	104	3	3.0	3.5	8.00	1	0
3	4	322	110	3	3.5	2.5	8.67	1	1
4	5	314	103	2	2.0	3.0	8.21	0	0

In the above picture, we are able to see a glimpse of the data set, in which there are 9 columns,

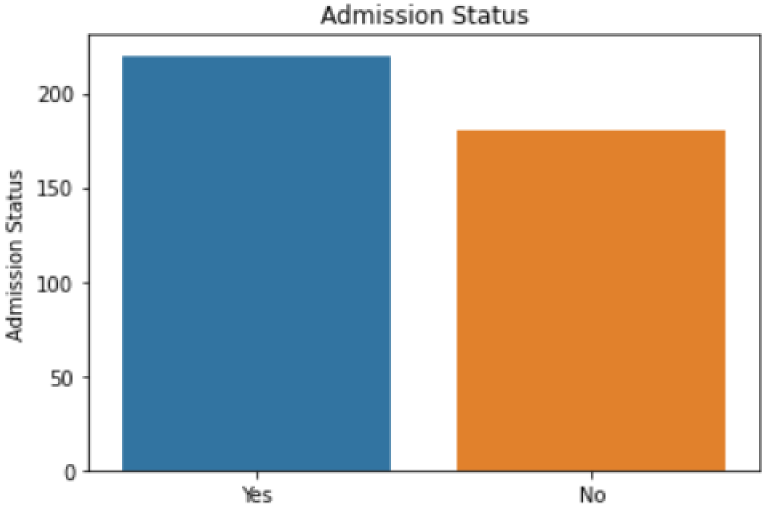
- The first one is for serial no.
- 2nd is for Graduate Record Examinations (GRE) scores ranging between 290 to 340.
- 3rd one is for the Test of English as a Foreign Language (TOEFL) scores ranging between 92 to 120.
- 4th column represents the university rating 1 to 5.
- 5th column represents the Statement of Purpose (SOP) ranging between 1 to 5.
- 6th column represents the Letter of Recommendation (LOR) strength ranging between 1 to 5.
- 7th column represents the CGPA score of the graduate students.
- 8th column represents the Research status of the students if they have researched on any topic then they have been assigned the value 1 otherwise 0 value is assigned.
- 9th column represents Admission status i.e. if the candidate gets admitted to the respective university, then he is assigned the value 1 otherwise 0 value is assigned to that row.

METHODOLOGY

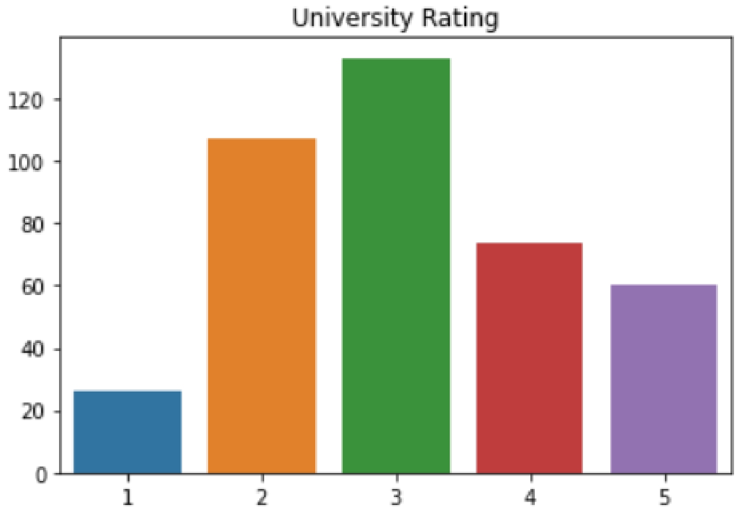
First, we have done EDA (tabular and graphical representation) for different columns of interest, we have standardized those columns by the MinMax Standardization method and then we have applied five machine learning classifiers to the dataset namely, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Naïve Bayes classifier. Then we obtained the Accuracy, Precision, Recall, F1 Score, and ROC-AUC Score of the respective model. Then based on the ROC-AUC score of a model we've decided on the best fitting model. The higher the value of ROC-AUC the better the model fit.

DATA ANALYSIS

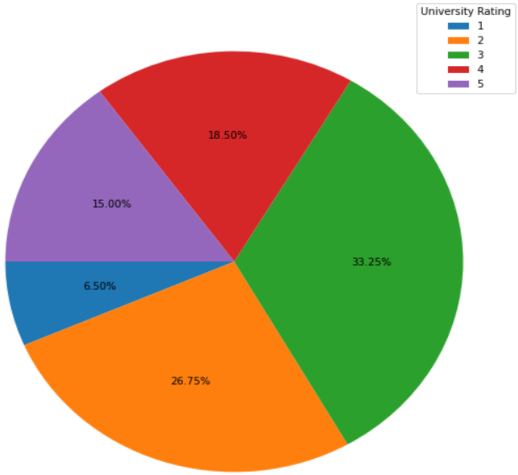
Bar plot of Admission Status



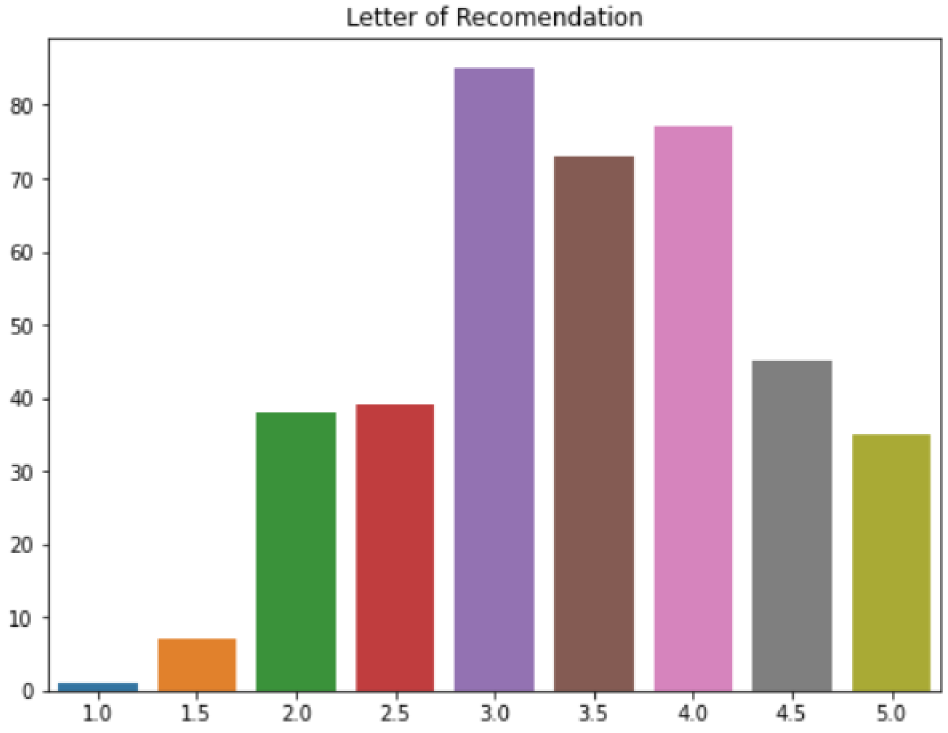
Bar Plot of University Rating



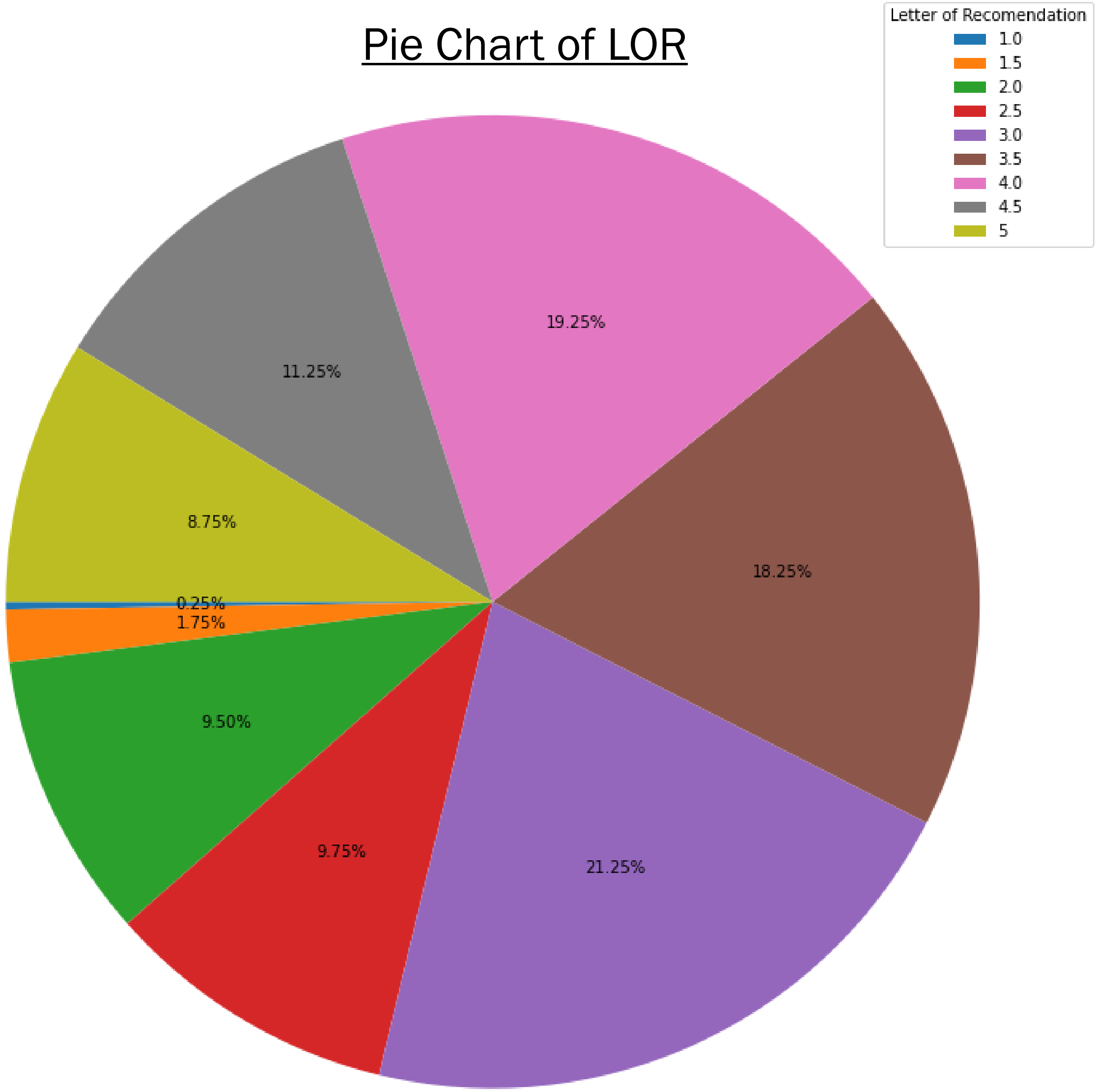
Pie Chart of University Rating



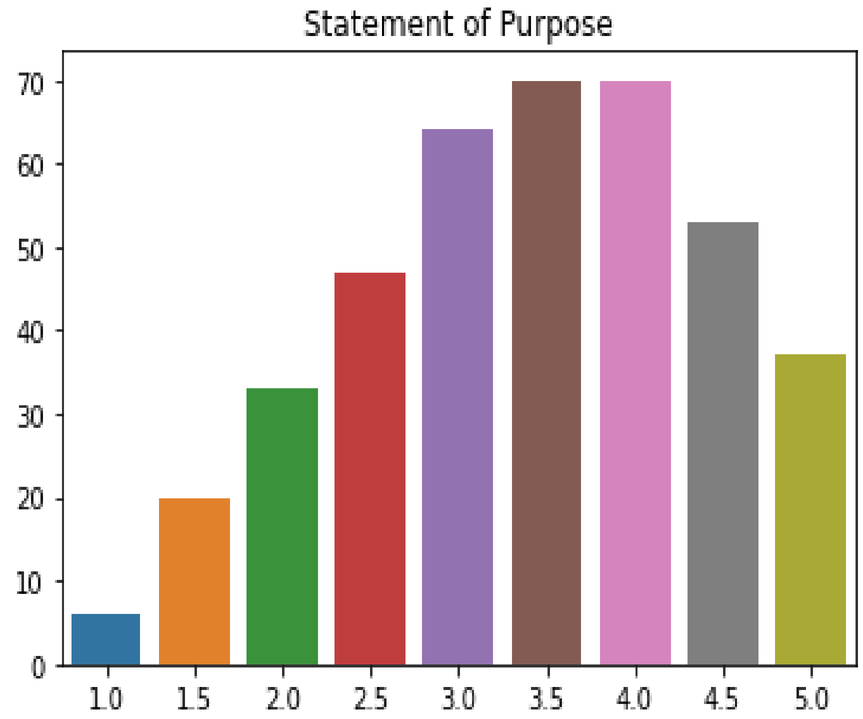
Bar plot of LOR



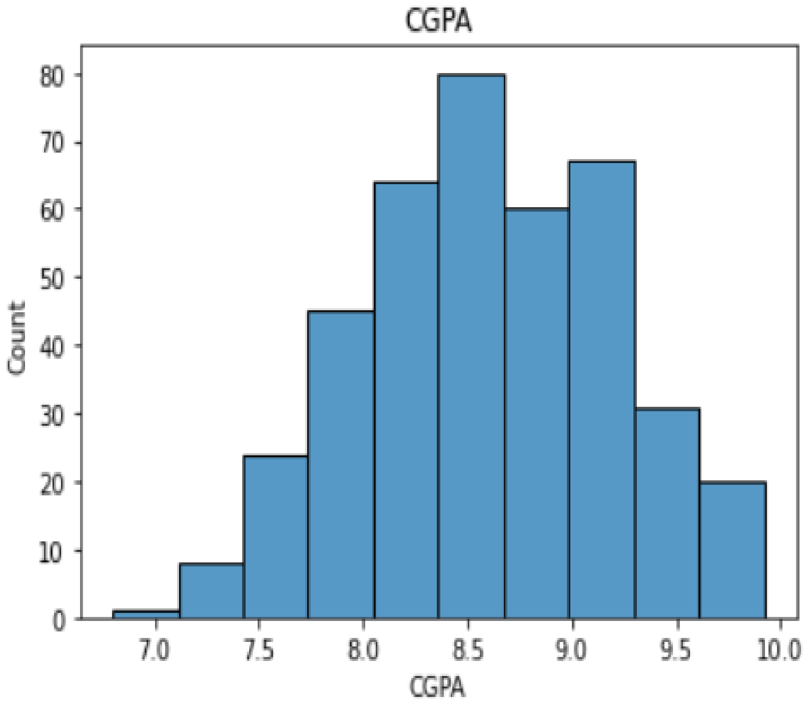
Pie Chart of LOR



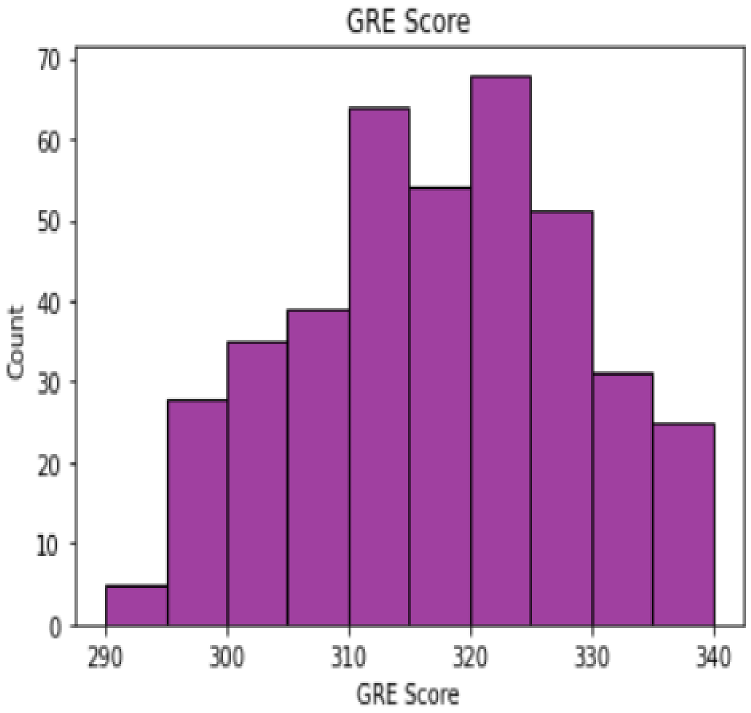
Bar Plot of SOP



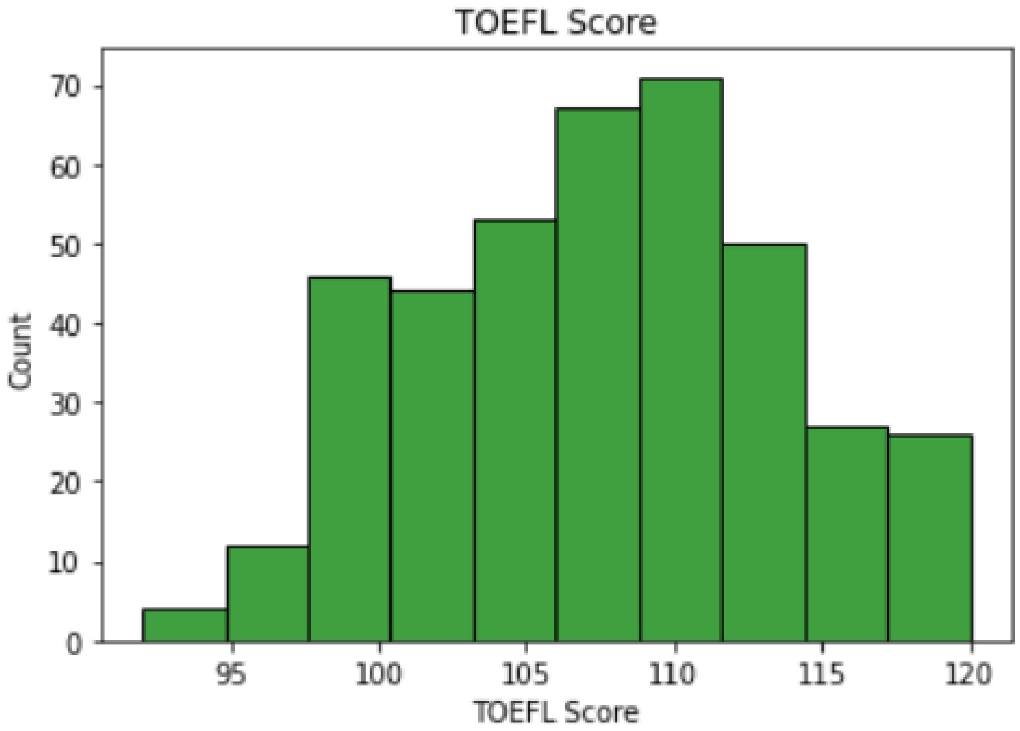
Histogram of CGPA



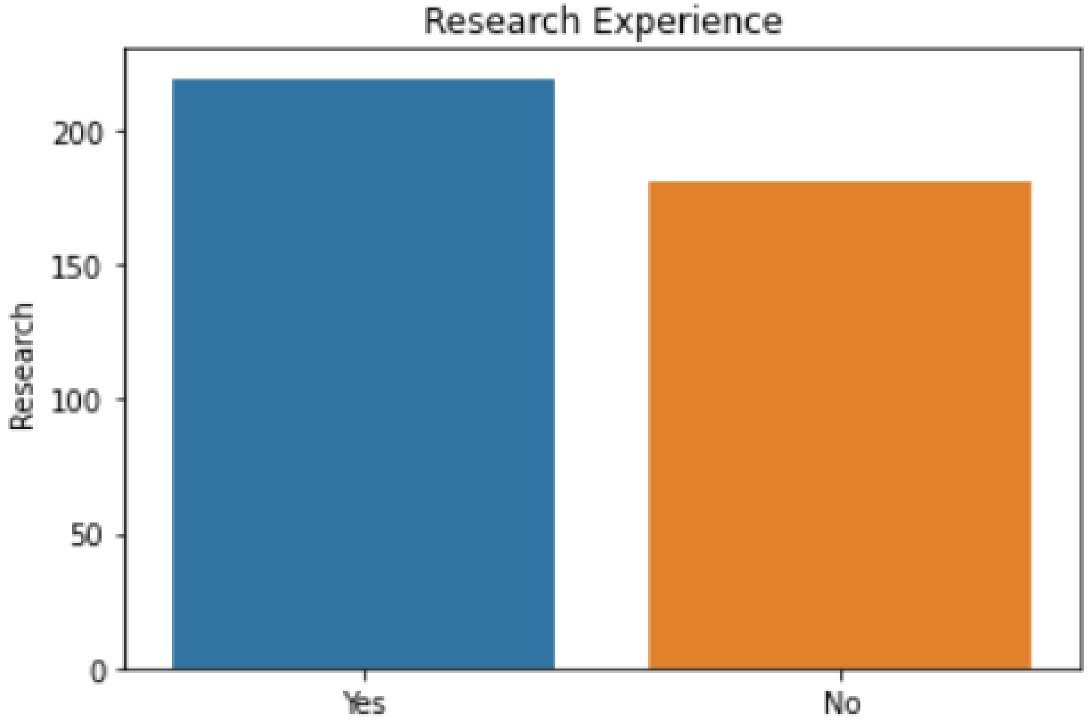
Histogram of GRE Score



Histogram of TOEFL Score



Bar plot of Research Experience



Correlation between the Covariates and Admission Status

Correlation of Admission Status with	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research
Correlation Coefficient Value	0.686138	0.672465	0.638983	0.612152	0.557481	0.737307	0.519441

From the above table, we can clearly observe that admission chance depends upon the Covariates among which, CGPA has maximum effect.

Maching Learning Modelling (SUPERVISED)

Min-Max Scalarization

After the graphical visualization of the data, we are going for some Supervised Machine Learning Models to fit our observations and validation.
But, before that, we have transformed the covariates within the range[0,10] using **Min-Max Scalarization.**

$$x_{scaled} = \frac{x - min\ (x)}{max(x) - min\ (x)}$$

This is done feature-wise in an independent way. The Min-MaxScaling compresses all inliers in a narrow range.

<u>LOGISTIC REGRESSION:</u> After fitting the Logistic regression classifier we got the confusion matrix as below.

Actual Pred	Negative	Positive
Negative	42	14
Positive	2	42

ML classifier	Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
LOGISTIC REGRESSION	0.84	0.75	0.954545	0.84	0.852272

DECISION TREE: After fitting the decision tree classifier we got the confusion matrix as below.

Actual Pred	Negative	Positive
Negative	40	16
Positive	4	40

ML Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
DECISION TREE	0.8	0.71428	0.9090	0.8	0.81168

RANDOM FOREST: After fitting the random forest classifier we got the confusion matrix as below.

Actual \ Pred	Negative	Positive
Negative	45	11
Positive	1	43

ML Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
RANDOM FOREST	0.88	0.796296	0.977272	0.8775510	0.89042

SUPPORT VECTOR MACHINE: After fitting the support vector classifier we got the confusion matrix as below.

Actual Pred	Negative	Positive
Negative	48	8
Positive	3	41

ML Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
SVM	0.89	0.836734	0.931818	0.88172	0.89448

Naïve Bayes: After fitting the naïve bayes classifier we got the confusion matrix as below.

Actual Pred	Negative	Positive
Negative	48	8
Positive	2	42

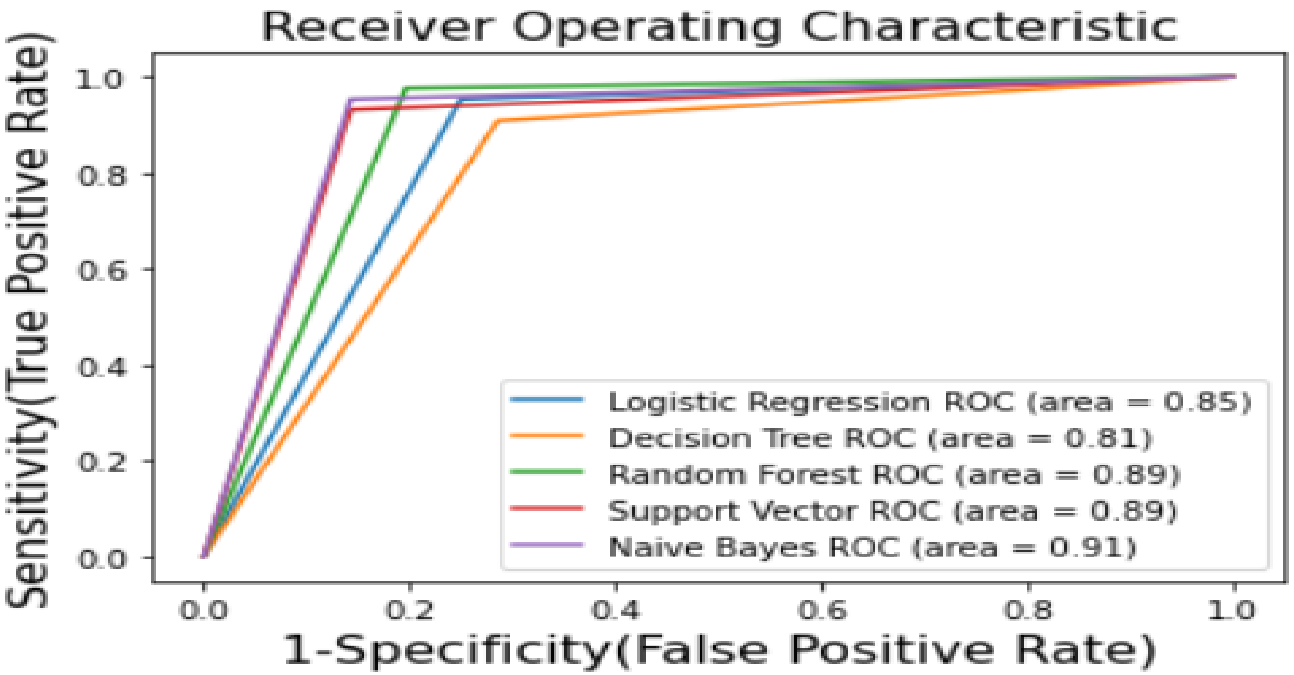
ML Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
NAÏVE BAYES	0.9	0.84	0.954545	0.863617	0.905844

SUMMARY

ML Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
Logistic Regression	0.84	0.75	0.954545	0.84	0.852272
DECISION TREE	0.8	0.71428	0.9090	0.8	0.81168
RANDOM FOREST	0.88	0.796296	0.977272	0.8775510	0.89042
SVM	0.89	0.836734	0.931818	0.88172	0.89448
NAÏVE BAYES	0.9	0.84	0.954545	0.863617	0.905844

CONCLUSION

Receiver Operating Characteristic graph



After fitting different models we found out The Naïve-Bayes classifier has the maximum value of the ROC-AUC Score. So, Naïve-Bayes is the best model for this data.

REFERENCE

- https://www.geeksforgeeks.org/decision-tree/
- https://www.sciencedirect.com/topics/computer-science/logistic-regression
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.
- https://en.wikipedia.org/wiki/Support-vector_machine
- https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79
- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Acknowledgment

We would like to express our profound and deep sense to our guide **prof. (Dr.) Indrani Mukherjee** and **prof. Debjit Konai** for their unending help, guidance, and suggestions without which this project would not have been a reality. They have acted as our philosopher, and guide. We owe great indebtedness for his untiring effort throughout the period of our research work.

We express our sincere thanks to the **prof. Prashanta Narayan Dutta** for his help in every corner of our study and we are gratefully indebted to **prof. (Dr.) Sukhendu Samajdar**, our respected director sir for his great inspiration and encouragement for our work.

We thank our teammate, **Mr. Soumitro Mukherjee**, for helping throughout the coding part and sparing his valuable time for project discussion and code writing.

We especially thank our teammate, **Mr. Bishal Chakraborty**, for sparing his valuable time for our academic project discussion.

Lastly, we thank all those concerned persons who have been directly or indirectly responsible for the completion of our project.

THANK YOU