University Recommendation System For Masters in USA

A Knowledge Discovery & Data Mining Project with Machine Learning Prediction

Date

Dataframe

Student Profile Dataset

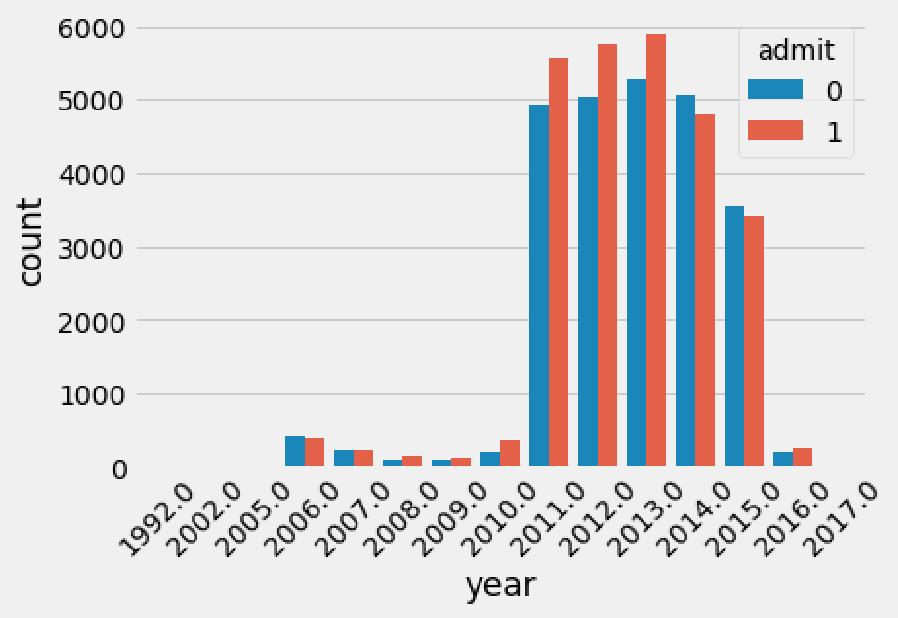
	userName	major	researchExp	industryExp	specialization	toeflScore	program	department	toeflEssay	internExp	greV	greQ	userProfileLink	journalPubs	greA	topperCgpa	termAndYear	confPubs	ugCollege	gmatA	cgpa	gmatQ	cgpaScale	gmatV	univName	admit
1750	hari_judee	Electrical and Computer Engineering	0	0	Digital VLSI Comp Arch	101.0	MS	EEE	Chemistry	NaN	3.0	152.0	168	http://www.edulix.com/unisearch/user.php?uid=2...	0.0	4.0	0	Fall - 2015	BITS Pilani	NaN	8.00	8.0	10	NaN	University of Wisconsin Madison	0
3646	Arun_Leo	Computer Science	0	0	Networks	100.0	MS	B.Tech	Information Technology	NaN	0.0	490.0	740	http://www.edulix.com/unisearch/user.php?uid=1...	0.0	3.5	8.89	Fall - 2012	Madras Institute of Technology	NaN	8.14	NaN	10	NaN	University of Texas Dallas	1
10495	hari_judee	Electrical and Computer Engineering	0	0	Digital VLSI Comp Arch	101.0	MS	EEE	Chemistry	NaN	3.0	152.0	168	http://www.edulix.com/unisearch/user.php?uid=2...	0.0	4.0	0	Fall - 2015	BITS Pilani	NaN	8.00	8.0	10	NaN	University of Southern California	0
13020	hari_judee	Electrical and Computer Engineering	0	0	Digital VLSI Comp Arch	101.0	MS	EEE	Chemistry	NaN	3.0	152.0	168	http://www.edulix.com/unisearch/user.php?uid=2...	0.0	4.0	0	Fall - 2015	BITS Pilani	NaN	8.00	8.0	10	NaN	University of Minnesota Twin Cities	1
17681	Arun_Leo	Computer Science	0	0	Networks	100.0	MS	B.Tech	Information Technology	NaN	0.0	490.0	740	http://www.edulix.com/unisearch/user.php?uid=1...	0.0	3.5	8.89	Fall - 2012	Madras Institute of Technology	NaN	8.14	NaN	10	NaN	University of Illinois Chicago	1
24034	hari_judee	Electrical and Computer Engineering	0	0	Digital VLSI Comp Arch	101.0	MS	EEE	Chemistry	NaN	3.0	152.0	168	http://www.edulix.com/unisearch/user.php?uid=2...	0.0	4.0	0	Fall - 2015	BITS Pilani	NaN	8.00	8.0	10	NaN	University of California Santa Barbara	0
26309	shrey911	environmental engineering	0	0	NaN	107.0	MS	Civil Engg	NaN	NaN	0.0	159.0	158	http://www.edulix.com/unisearch/user.php?uid=1...	0.0	NaN	81	Fall - 2014	Nagpur University	NaN	58.00	NaN	100	NaN	University of California Irvine	0
36456	hari_judee	Electrical and Computer Engineering	0	0	Digital VLSI Comp Arch	101.0	MS	EEE	Chemistry	NaN	3.0	152.0	168	http://www.edulix.com/unisearch/user.php?uid=2...	0.0	4.0	0	Fall - 2015	BITS Pilani	NaN	8.00	8.0	10	NaN	Purdue University	0
36794	shrey911	environmental engineering	0	0	NaN	107.0	MS	Civil Engg	NaN	NaN	0.0	159.0	158	http://www.edulix.com/unisearch/user.php?uid=1...	0.0	NaN	81	Fall - 2014	Nagpur University	NaN	58.00	NaN	100	NaN	Purdue University	0
37500	Arun_Leo	Computer Science	0	0	Networks	100.0	MS	B.Tech	Information Technology	NaN	0.0	490.0	740	http://www.edulix.com/unisearch/user.php?uid=1...	0.0	3.5	8.89	Fall - 2012	Madras Institute of Technology	NaN	8.14	NaN	10	NaN	Ohio State University Columbus	0

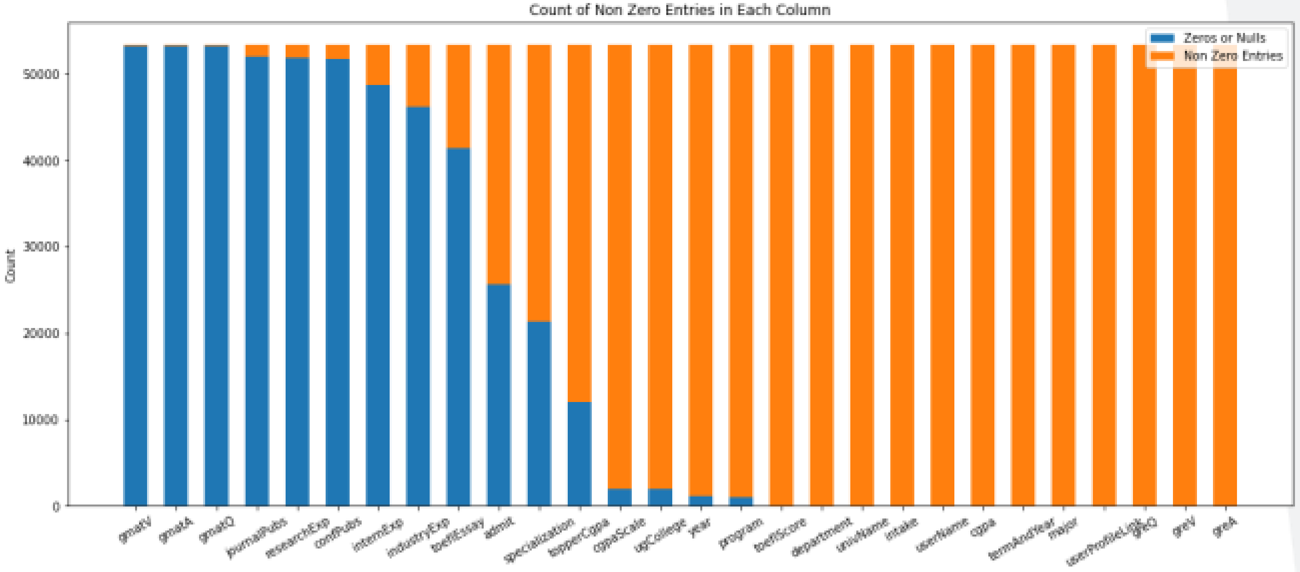
QS World Ranking Dataset

	Rank	institution	location code	location	ar score	ar rank	er score	er rank	fsr score	fsr rank	cpf score	cpf rank	ifr score	ifr rank	isr score	isr rank	irn score	irn rank	ger score	ger rank	score scaled
0	1	Massachusetts Institute of Technology (MIT)	US	United States	100.0	5	100.0	4	100.0	14	100.0	5	100.0	54	90.0	109	96.1	58	100.0	3	100
1	2	University of Cambridge	UK	United Kingdom	100.0	2	100.0	2	100.0	11	92.3	55	100.0	60	96.3	70	99.5	6	100.0	9	98.8
2	3	Stanford University	US	United States	100.0	4	100.0	5	100.0	6	99.9	9	99.8	74	60.3	235	96.3	55	100.0	2	98.5
3	4	University of Oxford	UK	United Kingdom	100.0	3	100.0	3	100.0	8	90.0	64	98.8	101	98.4	54	99.9	3	100.0	7	98.4
4	5	Harvard University	US	United States	100.0	1	100.0	1	99.4	35	100.0	2	76.9	228	66.9	212	100.0	1	100.0	1	97.6

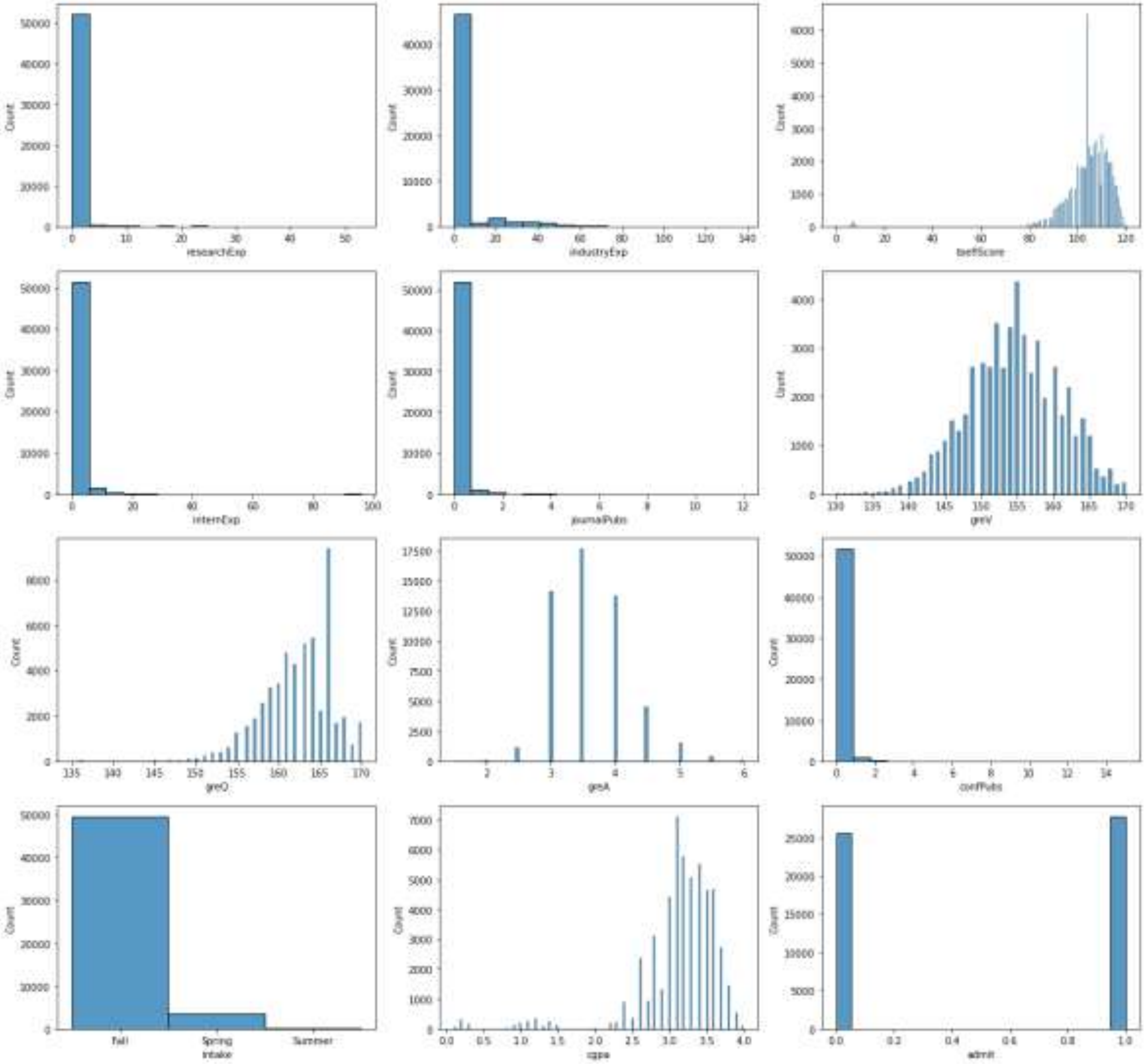
Exploratory Data Analysis (EDA)

- Null values in gmatA were 53203, gmatV were 53208 and gmatQ were significantly high.
- TOEFL essay had a total of 41448 null values.
- The data was shifted towards the left and was irregularly placed.
- CGPA was given in different scales like 0,4,5,10 and 100.
- Few GRE scores data followed previous marking system.

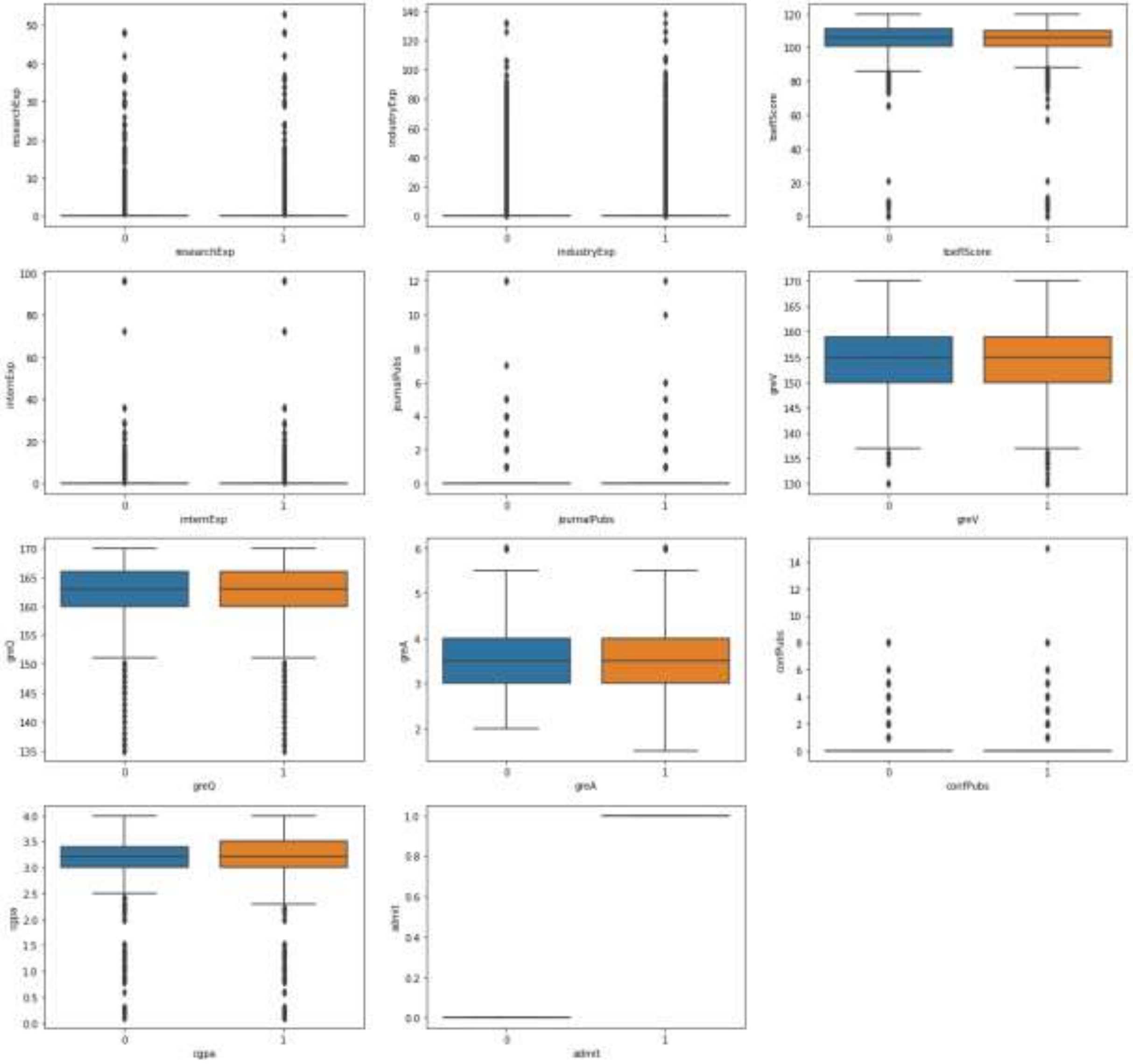




Exploratory Data Analysis (EDA)



Histogram of each feature



Box Plot of each feature

Data cleaning

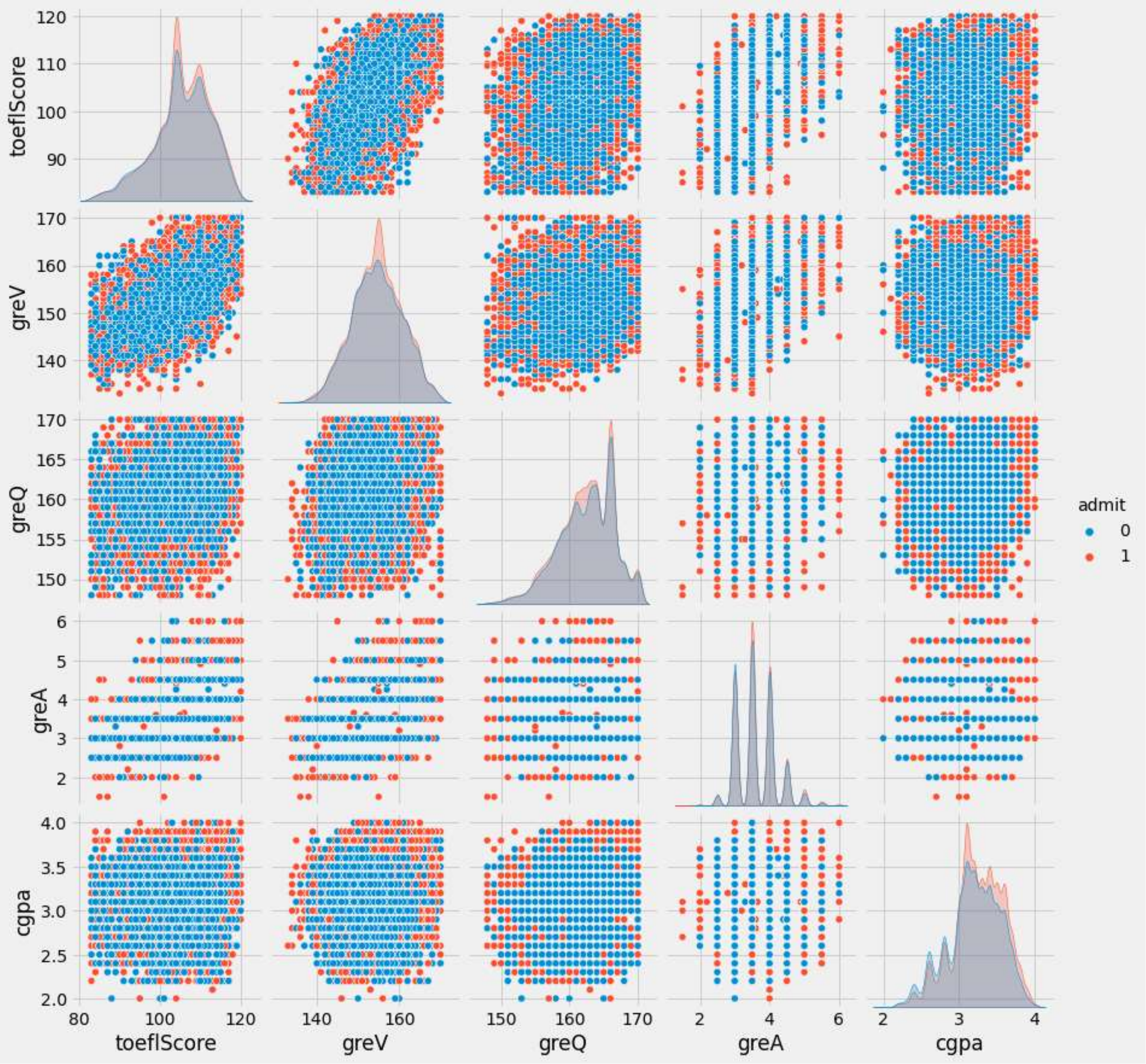
- Manual grouping of majors column and brought it down from 245 unique majors to 34.
- Converting GRE scores to the current scoring system.
- Shifting of misaligned data from columns.
- Identifying and removing irrelevant and less significant features.
- Filling of null values with either mean or mode.
- Restricting the dataset to MS values.
- Converting CGPA to 4 scale.
- Dropping rows with significant null values.
- Removing of outliers.

	old	newQ	newV
0	800	166	170
1	790	164	170
2	780	163	170
3	770	161	170
4	760	160	170

```
For Column:  toeflScore  ->  Range: 83.0  To  128.0
Total outliers removed:  597

For Column:  greV  ->  Range: 132.0  To  177.0
Total outliers removed:  8

For Column:  greQ  ->  Range: 148.0  To  178.0
Total outliers removed:  229

For Column:  greA  ->  Range: 1.0  To  6.0
Total outliers removed:  0

For Column:  cgpa  ->  Range: 2.0  To  4.5
Total outliers removed:  1877
```

Pair plots

- Pair plots visually plots each of the numeric feature against other features.
- Shows relationship between pairs of features.
- Represents data based on target feature.



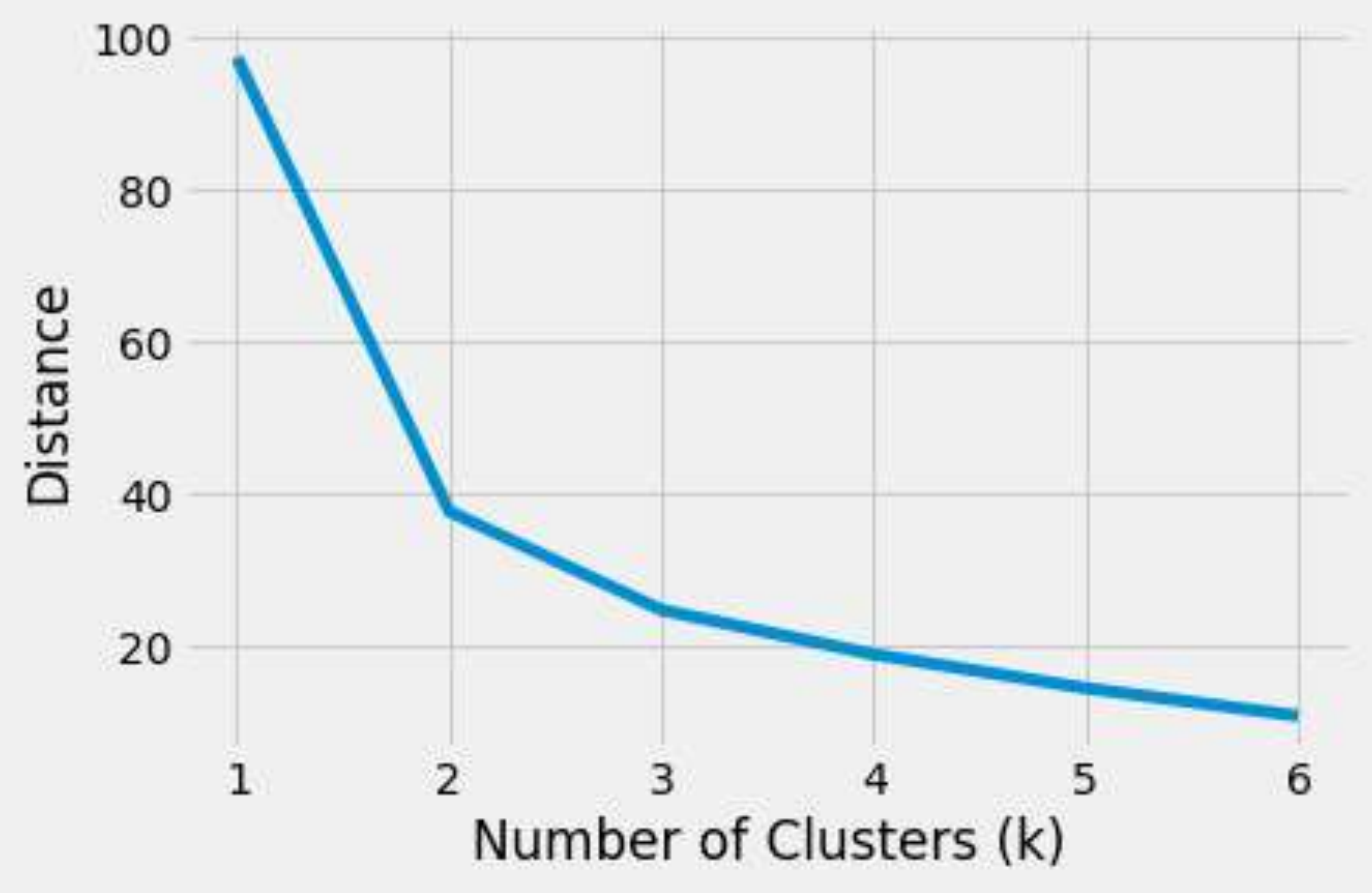
Data is linearly not separable

Clustering

- Used World ranking dataset from QS website.
- Filtered the data to universities in US.
- Dropped scores columns and kept ranks for better analysis.
- Performed Min-Max Scaling.
- Performed PCA on the dataset to help with visualisation.
- Schools were classified into tiers.

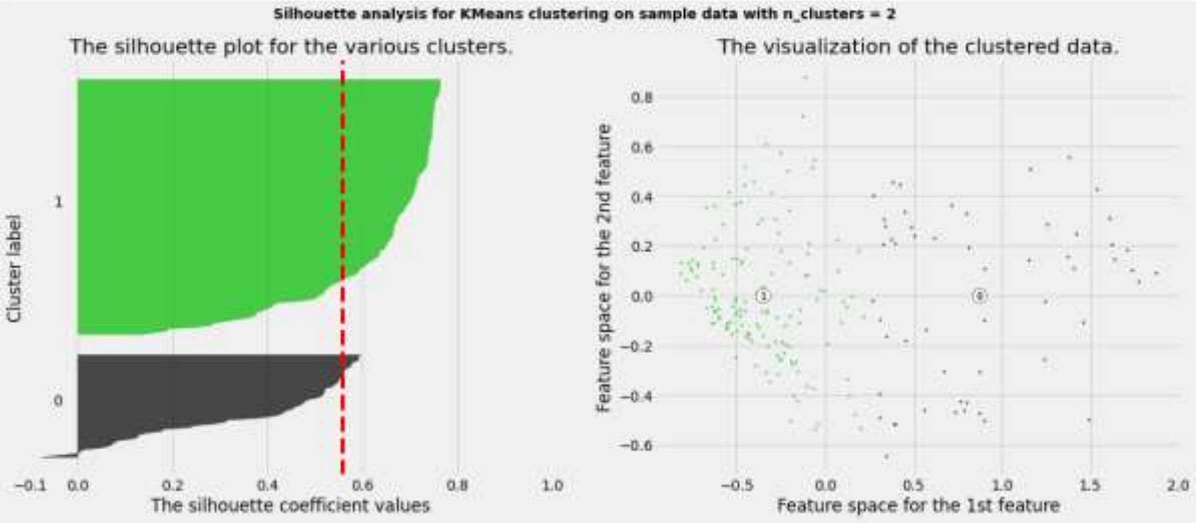
Cluster Analysis

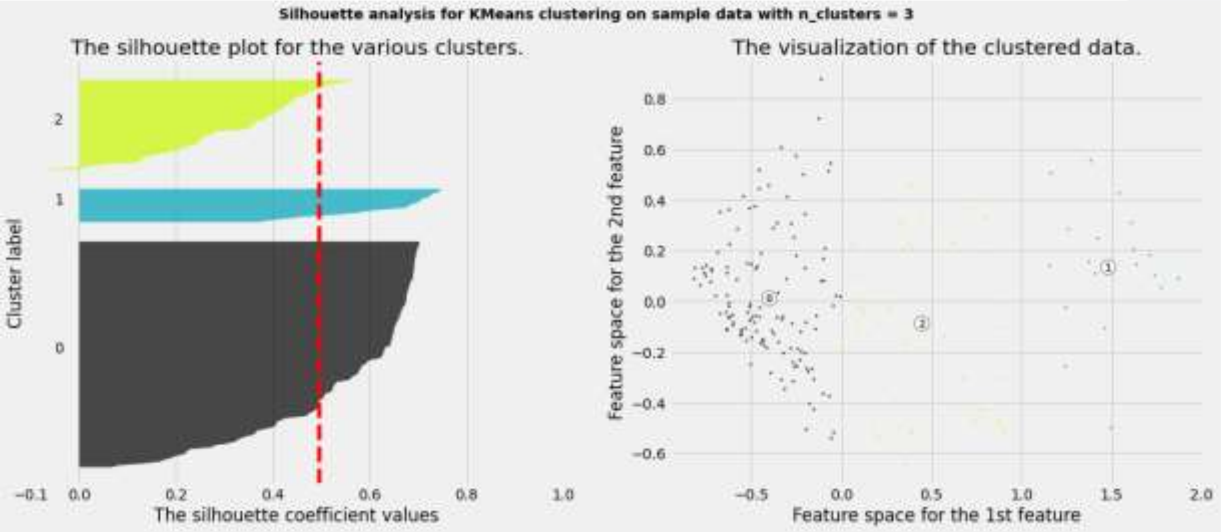
- Sum of squared distances of samples to their closest cluster center.
- Trade off between number of clusters and inertia.
- Elbow curve shows k=2 can produce better results.

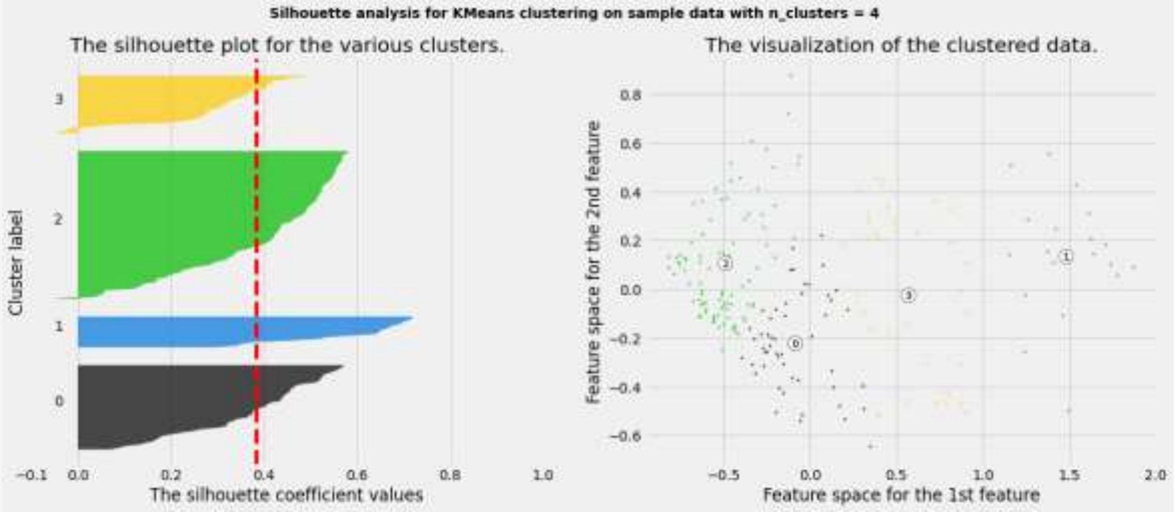


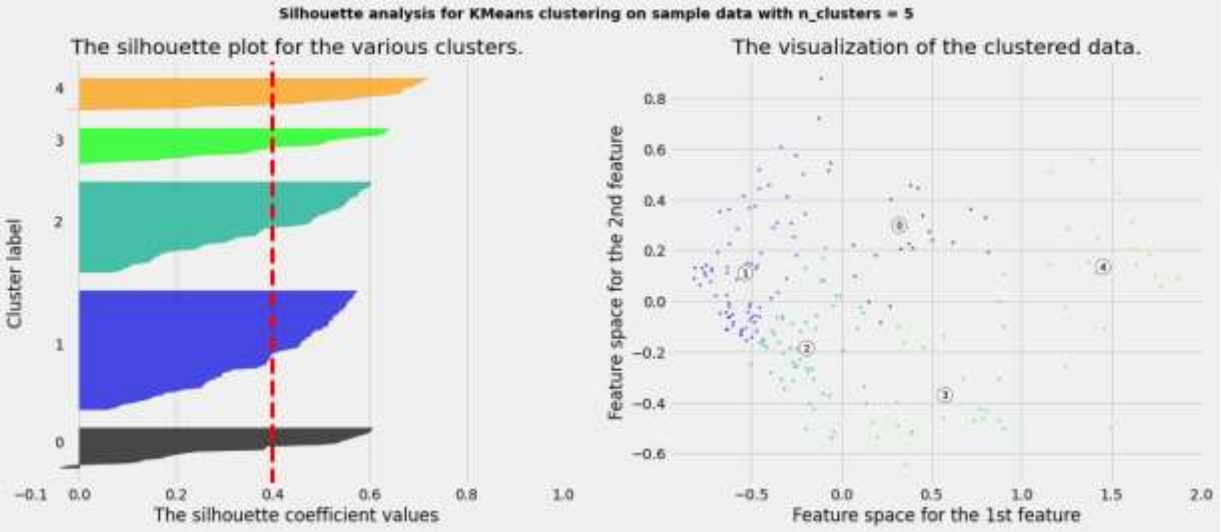
Elbow plot for K-Means inertia

Cluster Analysis







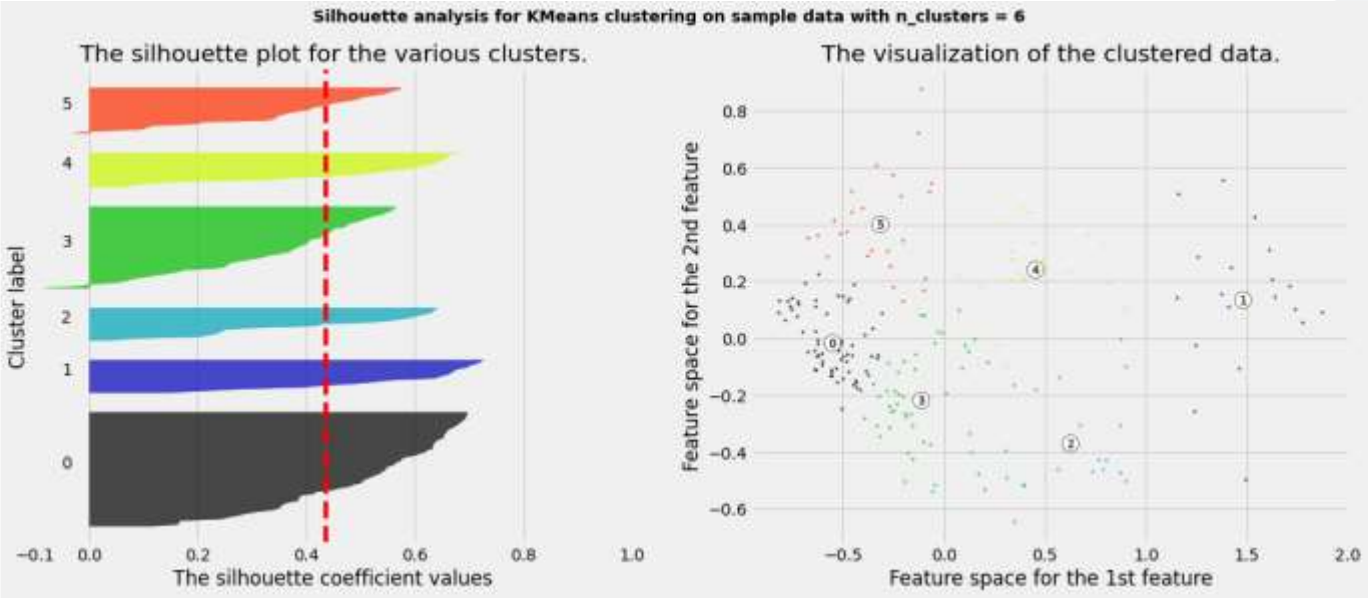


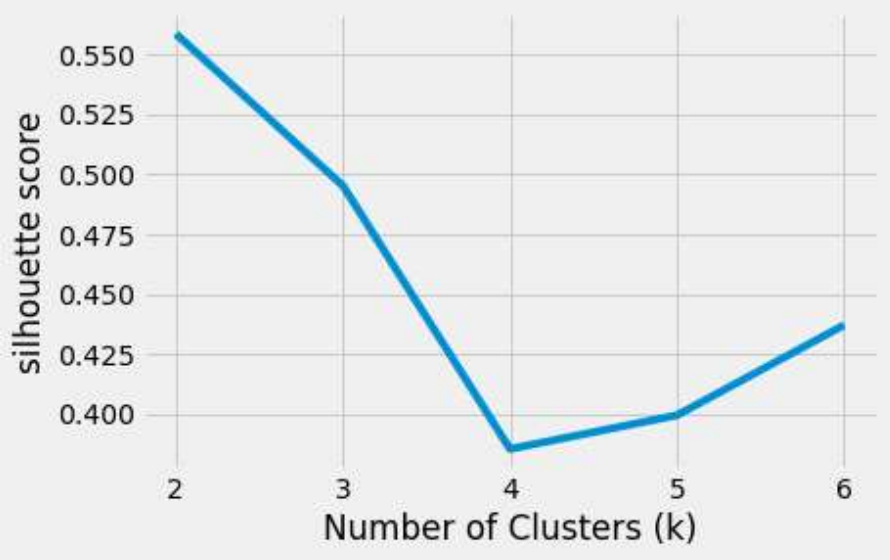
Cluster Analysis

- Compares how a point is similar to it's own cluster compared to other clusters.

- The value of the silhouette ranges between [1, -1]

- A high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.





Clustering Results

Tier 1:

Worcester Polytechnic Institute
Wayne State University
Virginia Polytechnic Institute and State University
University of Utah
University of Texas Dallas
University of Texas Arlington
University of North Carolina Charlotte
University of Illinois Chicago
University of Colorado Boulder
University of Cincinnati
University of California Santa Cruz
University of Arizona
Syracuse University
SUNY Stony Brook
SUNY Buffalo
Rutgers University New Brunswick/Piscataway
North Carolina State University
New Jersey Institute of Technology
George Mason University
Clemson University

Tier 2:

University of Wisconsin Madison
University of Washington
University of Texas Austin
University of Southern California
University of Pennsylvania
University of North Carolina Chapel Hill
University of Minnesota Twin Cities
University of Michigan Ann Arbor
University of Massachusetts Amherst
University of Maryland College Park
University of Illinois Urbana-Champaign
University of Florida
University of California Santa Barbara
University of California San Diego
University of California Los Angeles
University of California Irvine
University of California Davis
Texas A and M University College Station
Stanford University
Purdue University
Princeton University
Ohio State University Columbus
Northwestern University
Northeastern University
New York University
Massachusetts Institute of Technology
Johns Hopkins University
Harvard University
Georgia Institute of Technology
Cornell University
Columbia University
Carnegie Mellon University
California Institute of Technology
Arizona State University

Algorithms Applied

- KNN
- Random Forest
- Naive Baye’s
- Logistic Regression
- Support Vector Machines
- Linear Discriminant Analysis

Final transformations - I

Splitting

- Splitting the data based on it's cluster.
- Splitting the data into 30% test data and 70% train data.

Min-Max scaling

- Normalising data between 0 and 1.

Upscaling

- Uses bootstrapping with replacement.
- Creates a random resampling of data.
- Upscaling the admit rows since it's below the 65% of majority class.

Label encoding

- Converts text categorical data into numerical data.
- Each unique category gets a unique numerical value.

Initial testing results

Without university Name

PREDICTION FOR CLUSTER 0

	accuracy_score	precision_score	recall_score	f1_score
SVM	0.623638	0.629170	0.923096	0.748305
KNN	0.611077	0.656352	0.752060	0.700954
LR	0.628632	0.636220	0.904370	0.746958
LDA	0.629237	0.636427	0.905618	0.747527
Naive Bayes	0.622730	0.646568	0.832709	0.727928
Random Forest	0.598366	0.656257	0.708365	0.681316

PREDICTION FOR CLUSTER 1

	accuracy_score	precision_score	recall_score	f1_score
SVM	0.599161	0.607824	0.320443	0.419649
KNN	0.590766	0.572724	0.374581	0.452930
LR	0.565349	0.544701	0.237175	0.330460
LDA	0.565466	0.544917	0.237690	0.331000
Naive Bayes	0.554390	0.520940	0.182779	0.270611
Random Forest	0.572345	0.530911	0.467131	0.496983

With University Name

PREDICTION FOR CLUSTER 0

	accuracy_score	precision_score	recall_score	f1_score
SVM	0.751513	0.765919	0.851021	0.806231
KNN	0.739709	0.763194	0.828600	0.794553
LR	0.638620	0.644817	0.901844	0.751973
LDA	0.639074	0.644955	0.902840	0.752414
Naive Bayes	0.640738	0.657753	0.851769	0.742293
Random Forest	0.724425	0.749602	0.820379	0.783395

PREDICTION FOR CLUSTER 1

	accuracy_score	precision_score	recall_score	f1_score
SVM	0.731608	0.706752	0.669397	0.687568
KNN	0.721231	0.703715	0.635835	0.668055
LR	0.598461	0.571549	0.358879	0.440909
LDA	0.598694	0.571909	0.359408	0.441415
Naive Bayes	0.599977	0.583215	0.326903	0.418967
Random Forest	0.694415	0.666571	0.614958	0.639725

Final transformations - II

One hot encoding

- **Why**
  - Not to have any bias added to the data.
  - Improves accuracy of the model.
- **What**
  - It represents text data numerically.
  - Converts unique values into different features.
- **When**
  - We have nominal categorical data.
  - When curse of dimensionality doesn’t effect model accuracy.

Results with one hot encoding

- These algorithms were applied on both the clusters.
- SVM had the best accuracy for both cluster 0 and cluster 1.
- We also added KNN and random forest for further analysis.

PREDICTION FOR CLUSTER 0

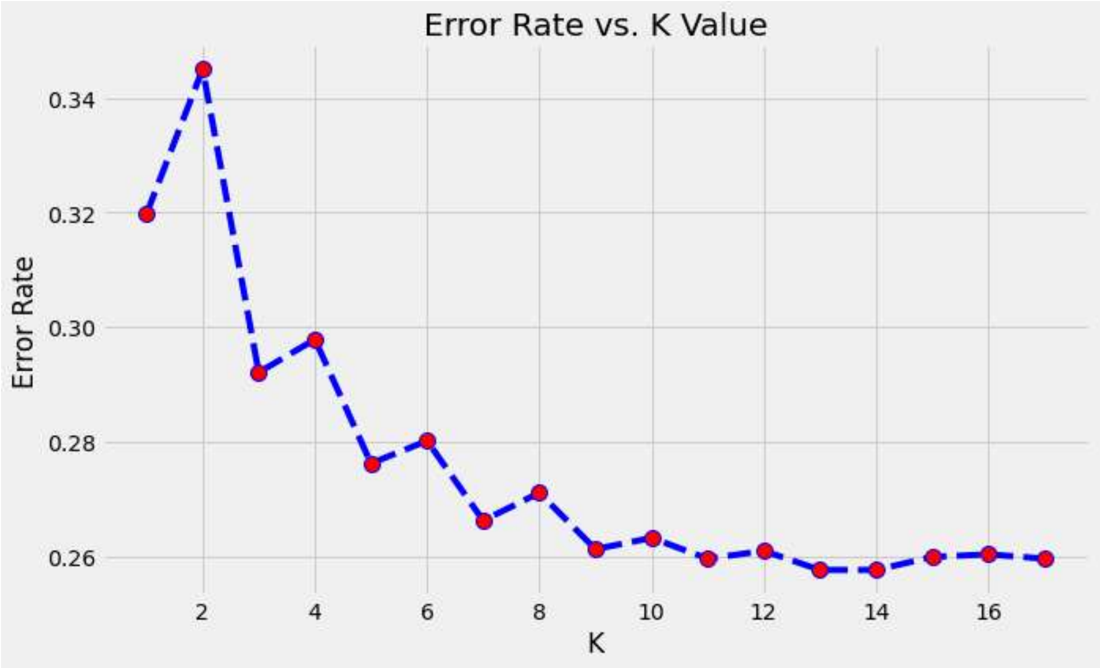
	accuracy_score	precision_score	recall_score	f1_score
SVM	0.756356	0.772953	0.849404	0.809377
KNN	0.743039	0.768715	0.826789	0.796695
LR	0.718069	0.740058	0.827783	0.781466
LDA	0.718069	0.739951	0.828032	0.781518
Naive Bayes	0.429782	0.849727	0.077286	0.141686
Random Forest	0.743493	0.768133	0.829026	0.797418

PREDICTION FOR CLUSTER 1

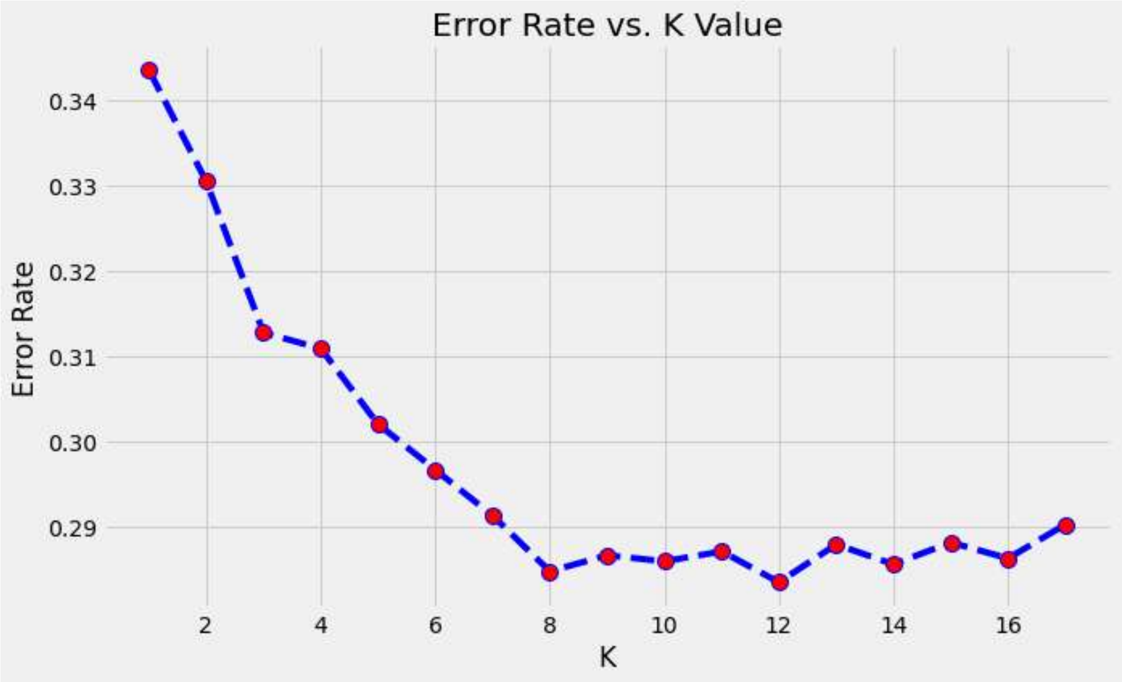
	accuracy_score	precision_score	recall_score	f1_score
SVM	0.726711	0.698038	0.656944	0.676868
KNN	0.719133	0.688209	0.649719	0.668410
LR	0.703743	0.676610	0.613059	0.643268
LDA	0.702227	0.674845	0.610918	0.641292
Naive Bayes	0.649878	0.615846	0.522077	0.565098
Random Forest	0.711321	0.679170	0.639550	0.658765

Hyper Parameter Tuning

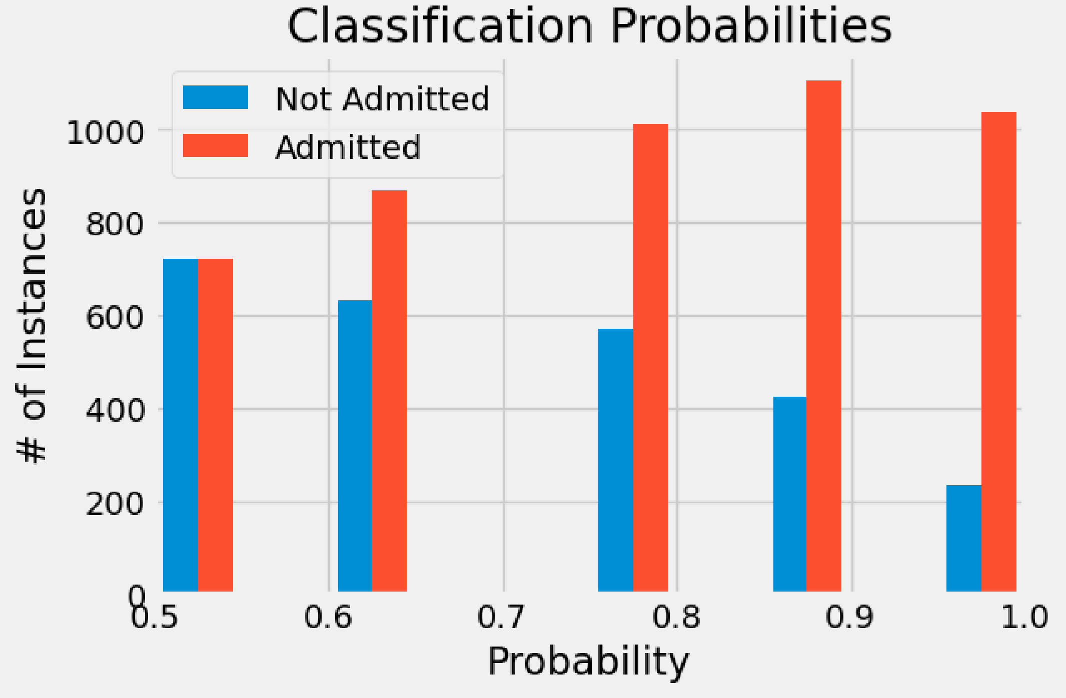
- **KNN**

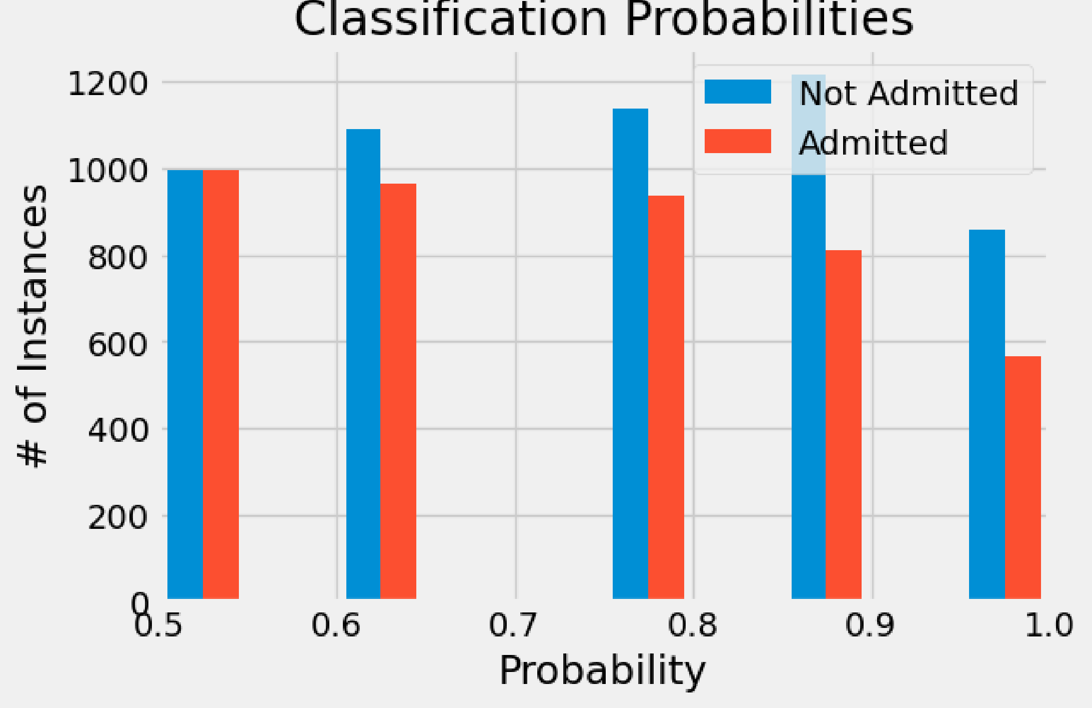


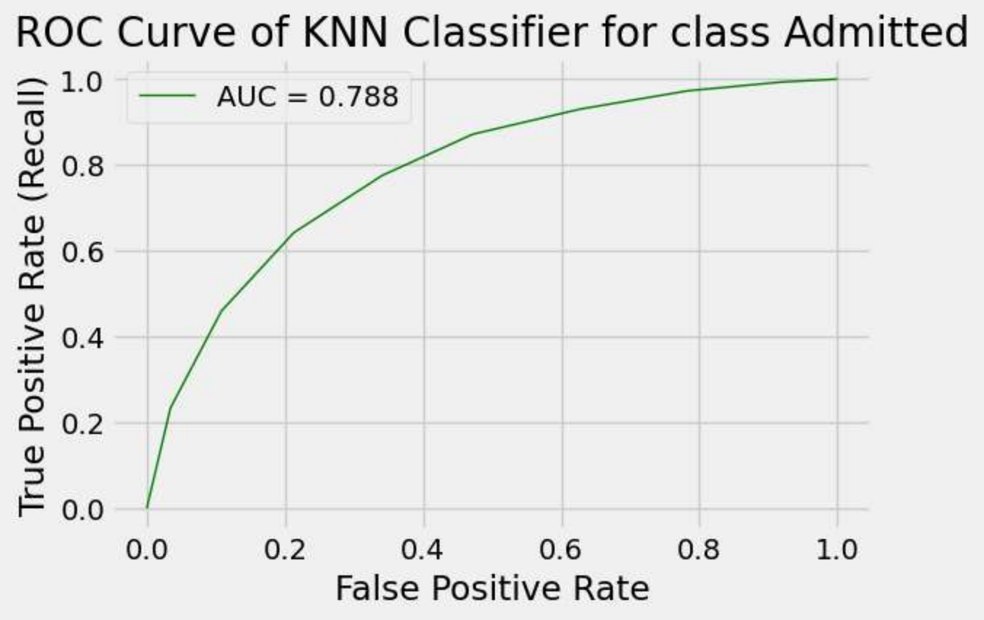
Tier 1

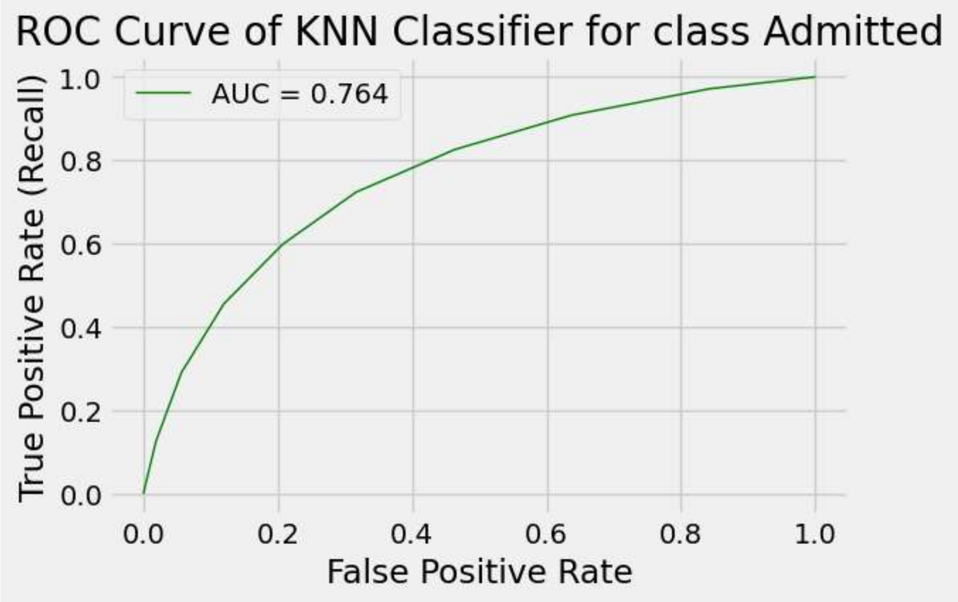


Tier 2


Classification Probabilities
Not Admitted
Admitted
# of Instances
0
200
400
600
800
1000
0.5
0.6
0.7
0.8
0.9
1.0
Probability


Classification Probabilities
Not Admitted
Admitted
# of Instances
0
200
400
600
800
1000
1200
0.5
0.6
0.7
0.8
0.9
1.0
Probability

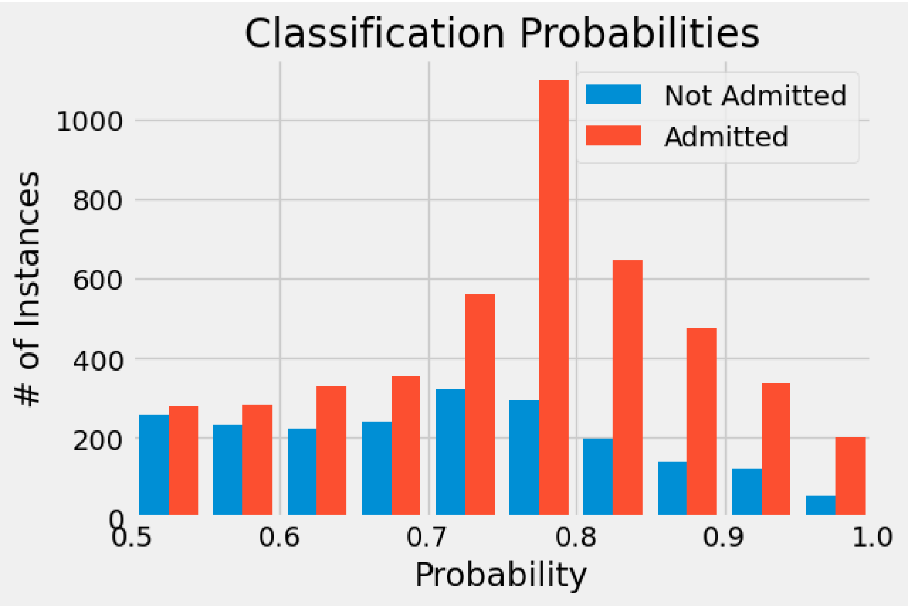

ROC Curve of KNN Classifier for class Admitted
AUC = 0.788
True Positive Rate (Recall)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

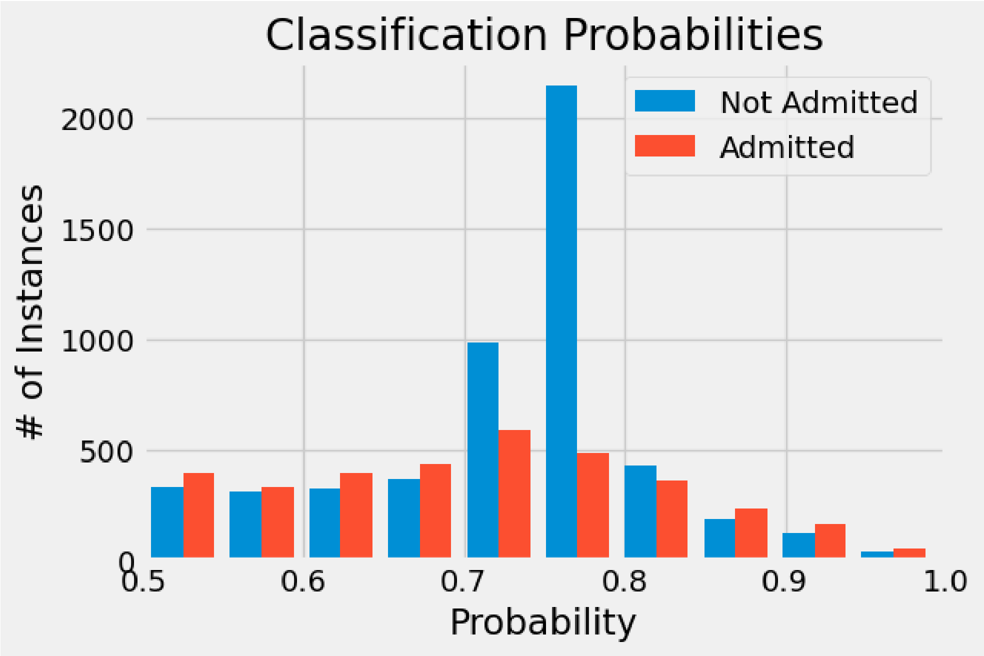

ROC Curve of KNN Classifier for class Admitted
AUC = 0.764
True Positive Rate (Recall)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

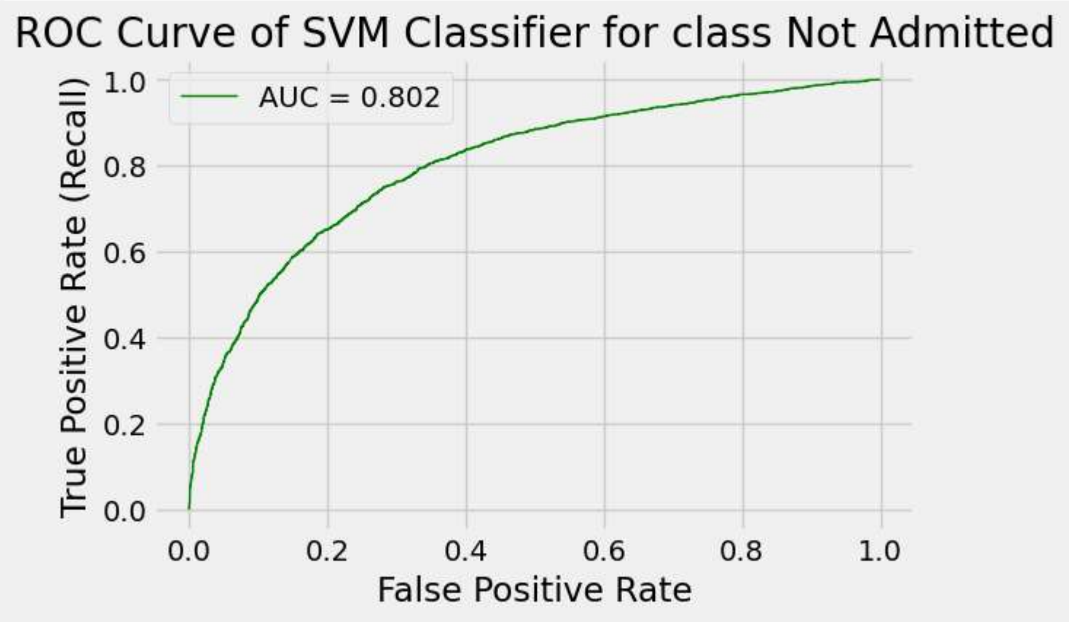
Tier 2

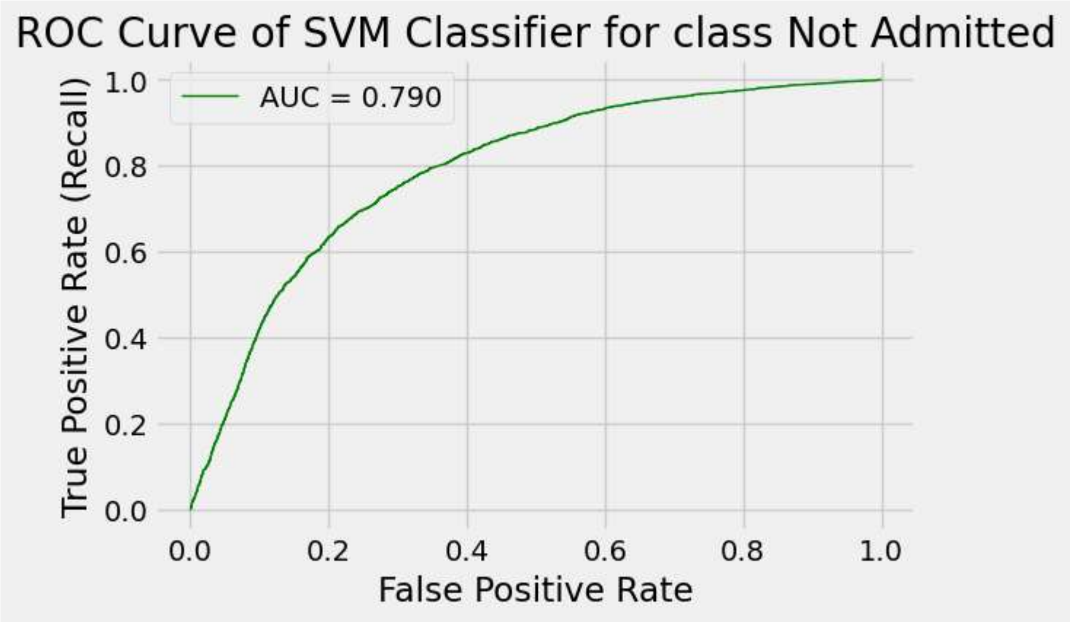
Tier 1

- **SVM**





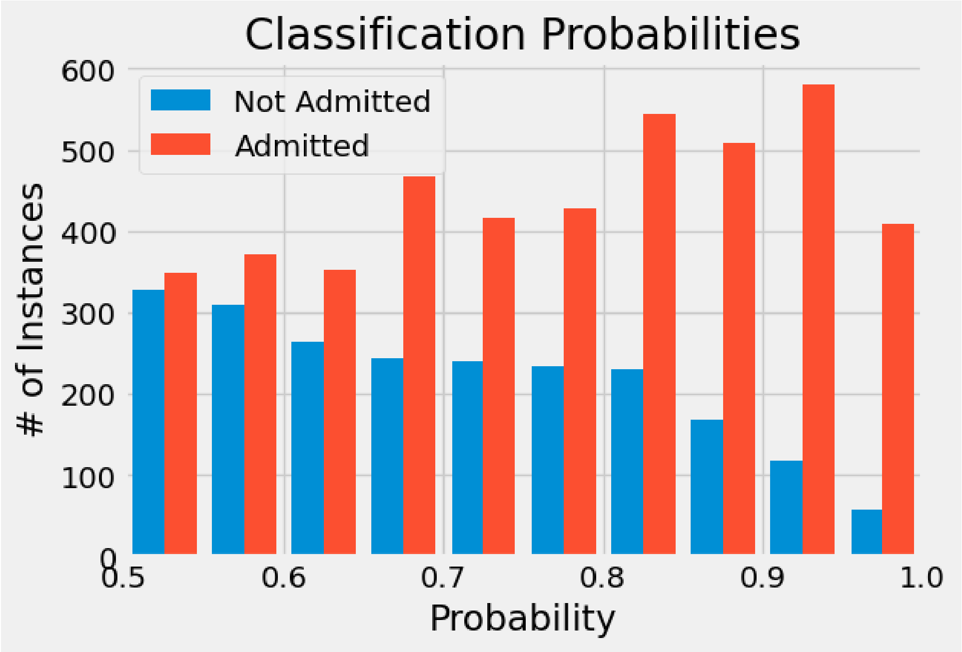


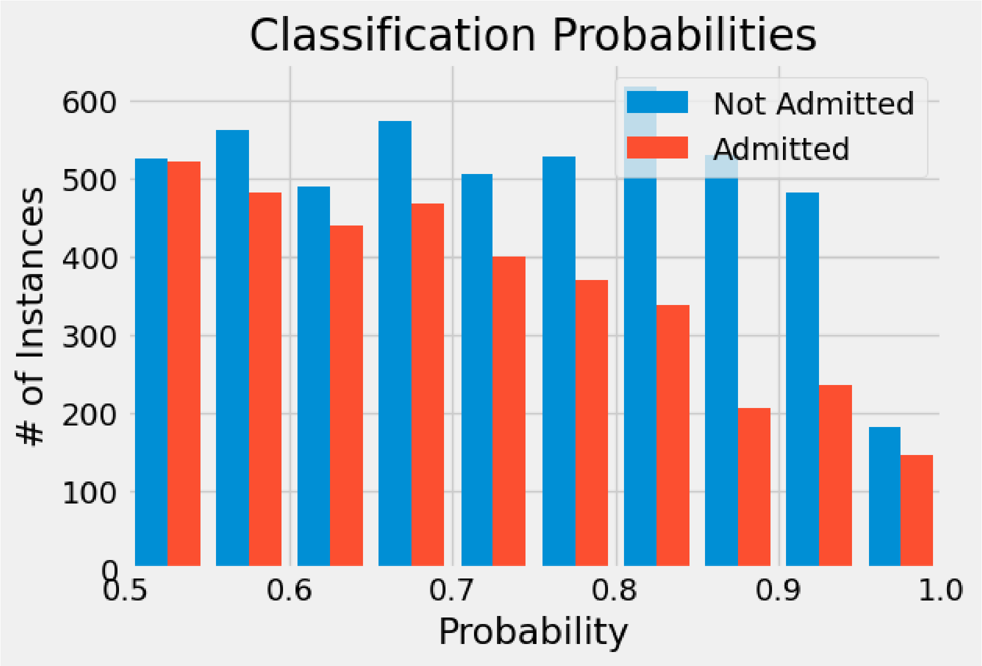


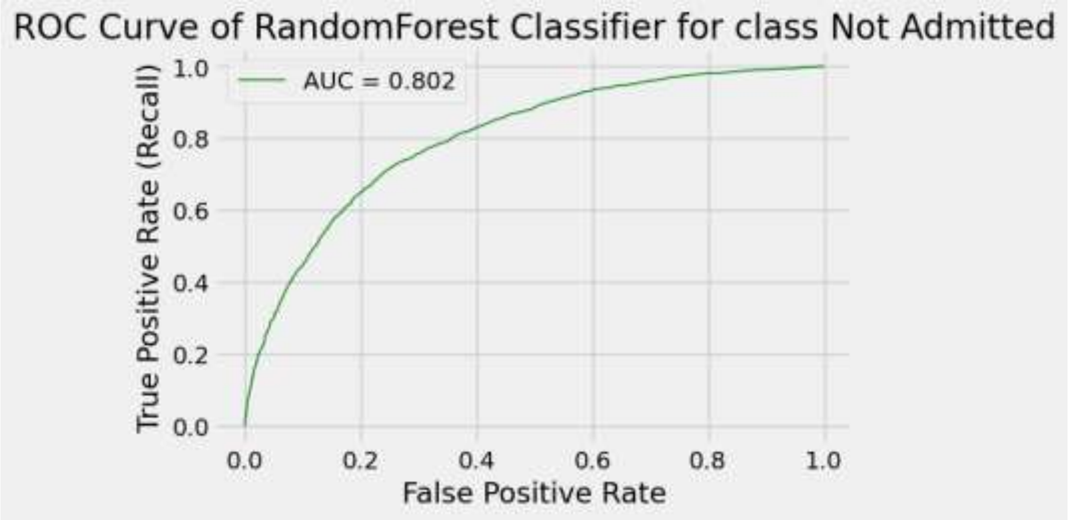
Tier 1

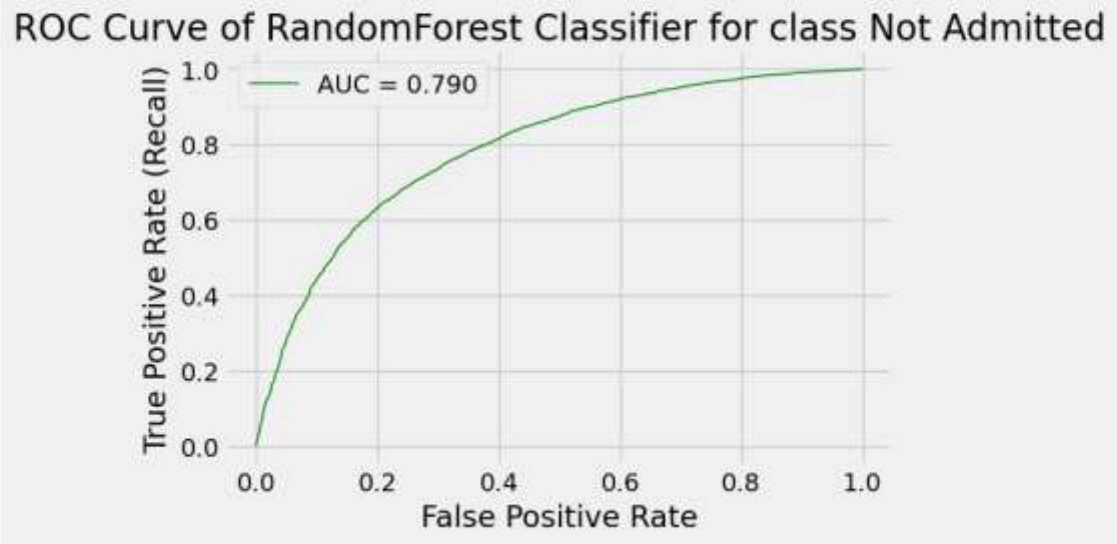
Tier 2

Random Forest





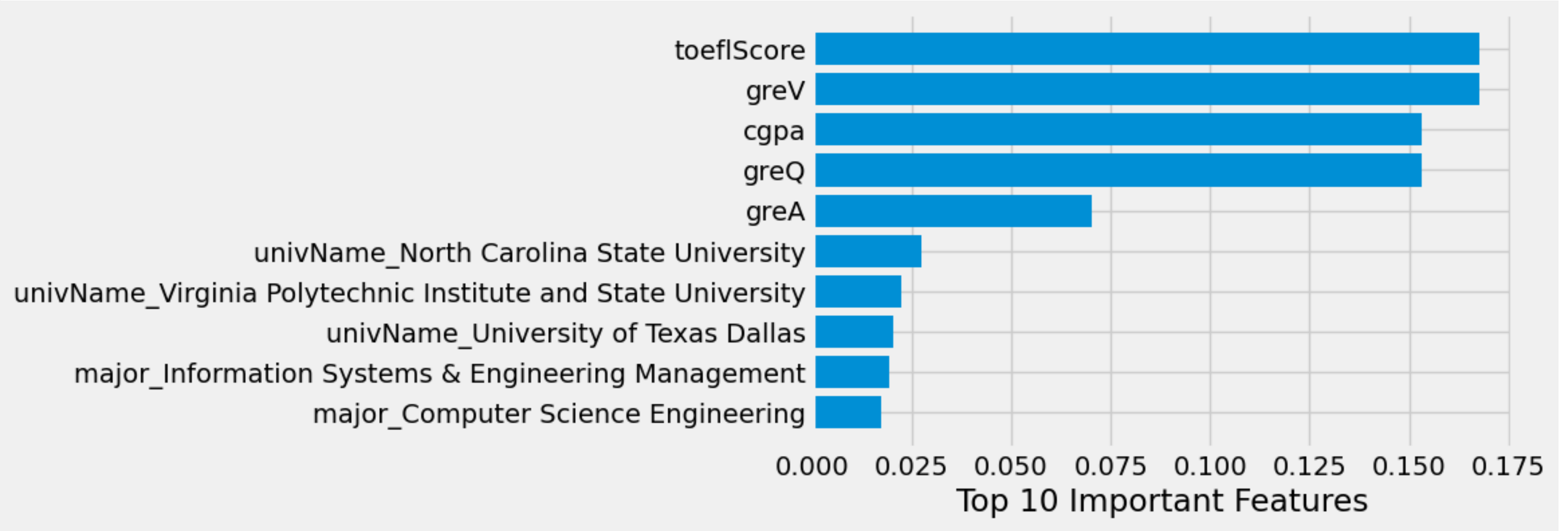




Tier 1

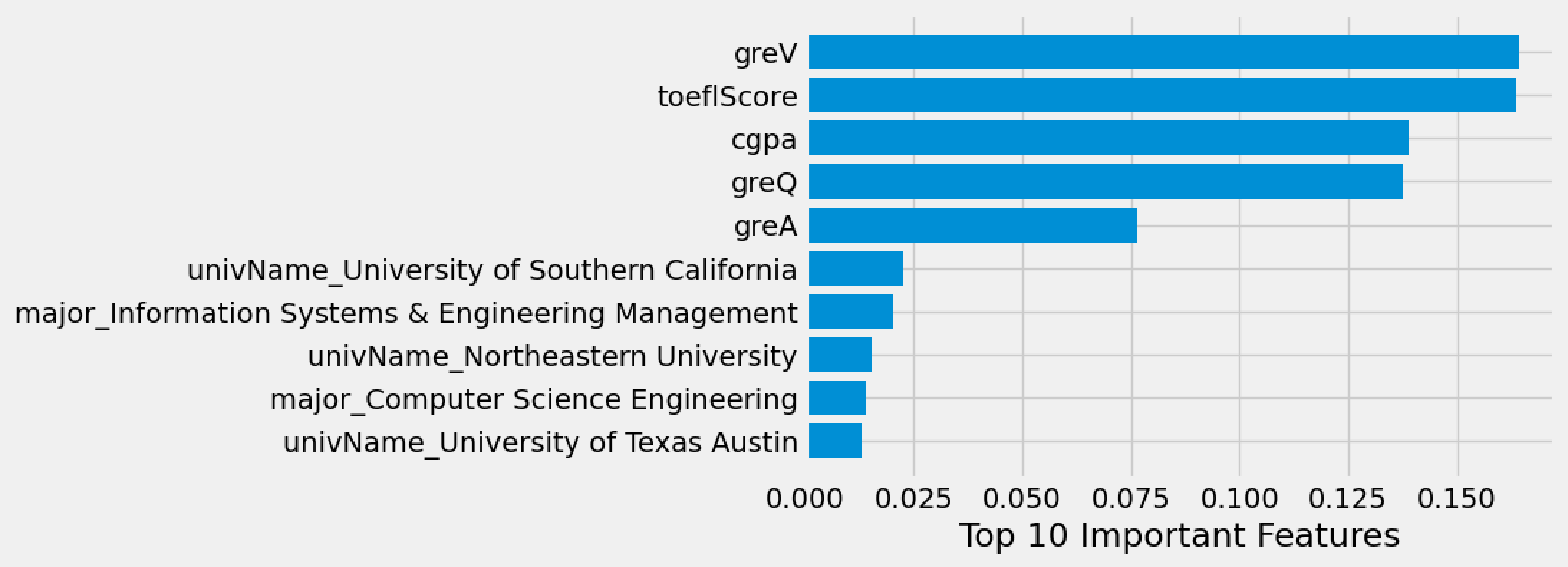
Tier 2

Top 10 important features



Tier 1

Tier 2



Final Results

KNN

Prediction for tier 1 Universities

	precision	recall	f1-score	support
0	0.65	0.66	0.65	2539
1	0.78	0.78	0.78	4069
accuracy			0.73	6608
macro avg	0.72	0.72	0.72	6608
weighted avg	0.73	0.73	0.73	6608

Prediction for tier 2 Universities

	precision	recall	f1-score	support
0	0.71	0.79	0.75	4741
1	0.70	0.60	0.65	3836
accuracy			0.71	8577
macro avg	0.70	0.70	0.70	8577
weighted avg	0.71	0.71	0.70	8577

Final Hyper parameters:
K=8,
leaf_size= 30,
weights=‘uniform',
p= 2

Final Results

SVM

Prediction for tier 1 Universities

	precision	recall	f1-score	support
0	0.70	0.60	0.64	2536
1	0.77	0.84	0.80	4072
accuracy			0.75	6608
macro avg	0.74	0.72	0.72	6608
weighted avg	0.74	0.75	0.74	6608

Prediction for tier 2 Universities

	precision	recall	f1-score	support
0	0.74	0.79	0.77	4785
1	0.71	0.65	0.68	3792
accuracy			0.73	8577
macro avg	0.73	0.72	0.72	8577
weighted avg	0.73	0.73	0.73	8577

Final Hyper parameters: C=1, gamma=1, kernel='poly'

Random Forest

Prediction for tier 1 Universities

	precision	recall	f1-score	support
0	0.69	0.60	0.64	2536
1	0.77	0.83	0.80	4072
accuracy			0.74	6608
macro avg	0.73	0.71	0.72	6608
weighted avg	0.74	0.74	0.74	6608

Prediction for tier 2 Universities

	precision	recall	f1-score	support
0	0.74	0.77	0.76	4785
1	0.70	0.66	0.68	3792
accuracy			0.72	8577
macro avg	0.72	0.72	0.72	8577
weighted avg	0.72	0.72	0.72	8577

Final Hyper parameters: 'bootstrap': True, 'max_depth': None, 'max_features': 'auto', 'n_estimators': 500

Final Insights

- Always analyse the data instead of directly dropping it.
- Necessary to visualise the data to gain better insights.
- Data pre-processing has the biggest impact in model accuracy.
- Test different encoding methods to check if it helps.
- Clustering of data can enhance performance.
- Always check if hyper parameter tuning helps your model.
- Our profiles are evaluated holistically, therefore your SOP and LOR matters.
- Each university is unique in their own way of evaluating candidates.

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK **YOU**