Predicting Admission Status

Using Machine Learning & Data Analysis

(ICSI 536 Machine Learning)

Guided by:

DR. SIWEI LYU

Authored by:

Dhruv Patel & Rajat Gupta

ABSTRACT

Machine Learning is a system which predicts instances based on Data analysis. When analyzed data is provided, the system executes various algorithms such as K-Mean and LDA. These algorithms help us predict desired results. The results obtained are implicitly dependent on the accuracy of analyzed data and selected algorithm. In this report, we have tried to solve the problem of college acceptance for different colleges in United Sates. Students applying in different colleges, don't know about the university life as well as the minimum SAT score required to have a chance to get accepted in the applied university. The basic solution of this problem is to give the applicants a possible predication of college acceptance so that they don't miss out any good College in which they could have got admitted into. Here we are using past data of different College in United States of America, with the help of which we have trained data to fit Logistic Regression Model to give a prediction of Admission Status.

INTRODUCTION

The above problem mentioned in the abstract signifies one of the major roadblocks faced by the students after high school. The question is to select a proper college for their undergraduate studies. In the dilemma of choice of available options, we always want to go a reputed college so that we have a bright and successful future. In the above confusion, many of us miss out the second-best university we could have got into but, we missed to apply into it thinking the university having high SAT score than what we have scored. The primary reason of selecting the above problem is that many of us have faced the same problem during under graduate admission, and then regret not being able to change the college and send the rest four years with the regret of not being able to get into the deserving environment of study.

To solve the problem in hand we have used the raw data provided by UCI dataset named university. Data. This data consists of different columns specifying SAT Scores according to name of University, Admission Percentage, Social Life, State and Student Faculty Ratio. On Analyzing the raw data, we have Formulated to covert the data in Lisp format to excel format. Later, by analyzing data we formulated admission chances based on the admission percentage and university Life based on 3 different factors on a scale of 1-5. Than by training the machine learning model we form a function prediction that will be able to predict the admission status i.e Accept/Reject when SAT scores are being given as an input.

RELATED WORK

The website www.usnews.com provides us with the college rankings that tells the users which are the top universities in the System right now and what their aim should be. This website is open of the top recognized resource for the any information regarding university details and college ranking for students. This website doesn’t give the rank but it helps its user to view all colleges based on their ranking.

There is another website named veritasprep.com that takes in details such as grade, GPA and extracurricular activities done and gives a percentage score that will say whether the user will be able to get admit in a university. This helps the user, to have some basic information, also giving them valuable advice to start working on their academic score and try to increase their GPA and start participating in different extra-curricular activities. But no information regarding the SAT score is being provided which acts as one of the major factors effecting the admission process in College across different states in the US.

The above-mentioned products each give the details about the university based on their reputation or GPA and athletic activities. There are also some of the website affiliated with ETS that give the prediction about GRE score, that later comparing with usnews.com gives them the desired output, there is no combination work that take score such as SAT and predicts whether the user will get an admit or a reject.

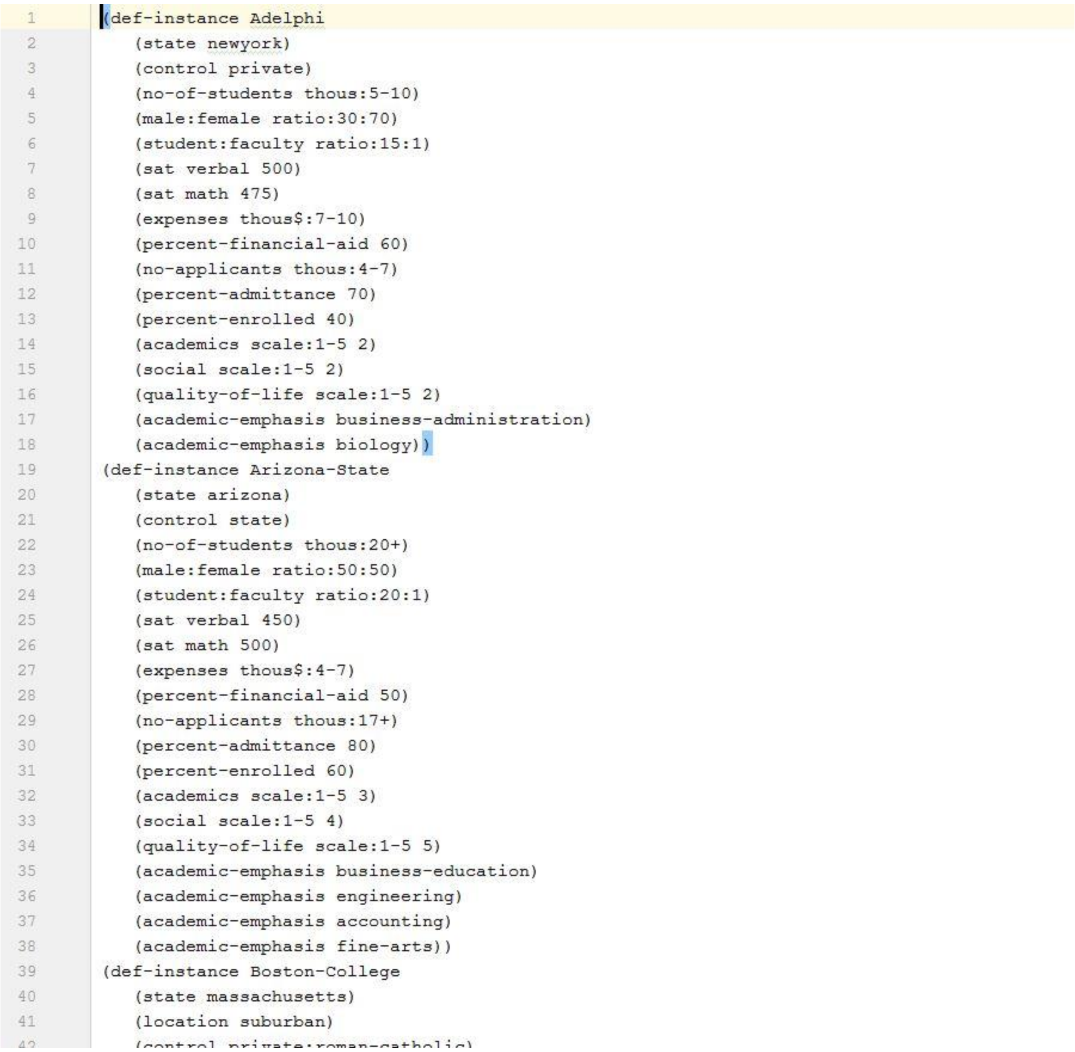
Methods

For Pre-Data Processing:

The Data was in lisp format, therefore for its evaluation we had to write some code in python to convert the data from Lisp format to CSV format. We had to do the above process because analysis based on Lisp format was not possible and python (as well as MATLAB) can only form confusion matrix (will be discussed later in this report) based on CSV files or json format data. Hence, we used the following code to covert data from lisp to csv. The data provided had some missing values which we tried to manually find and inset in the csv file.

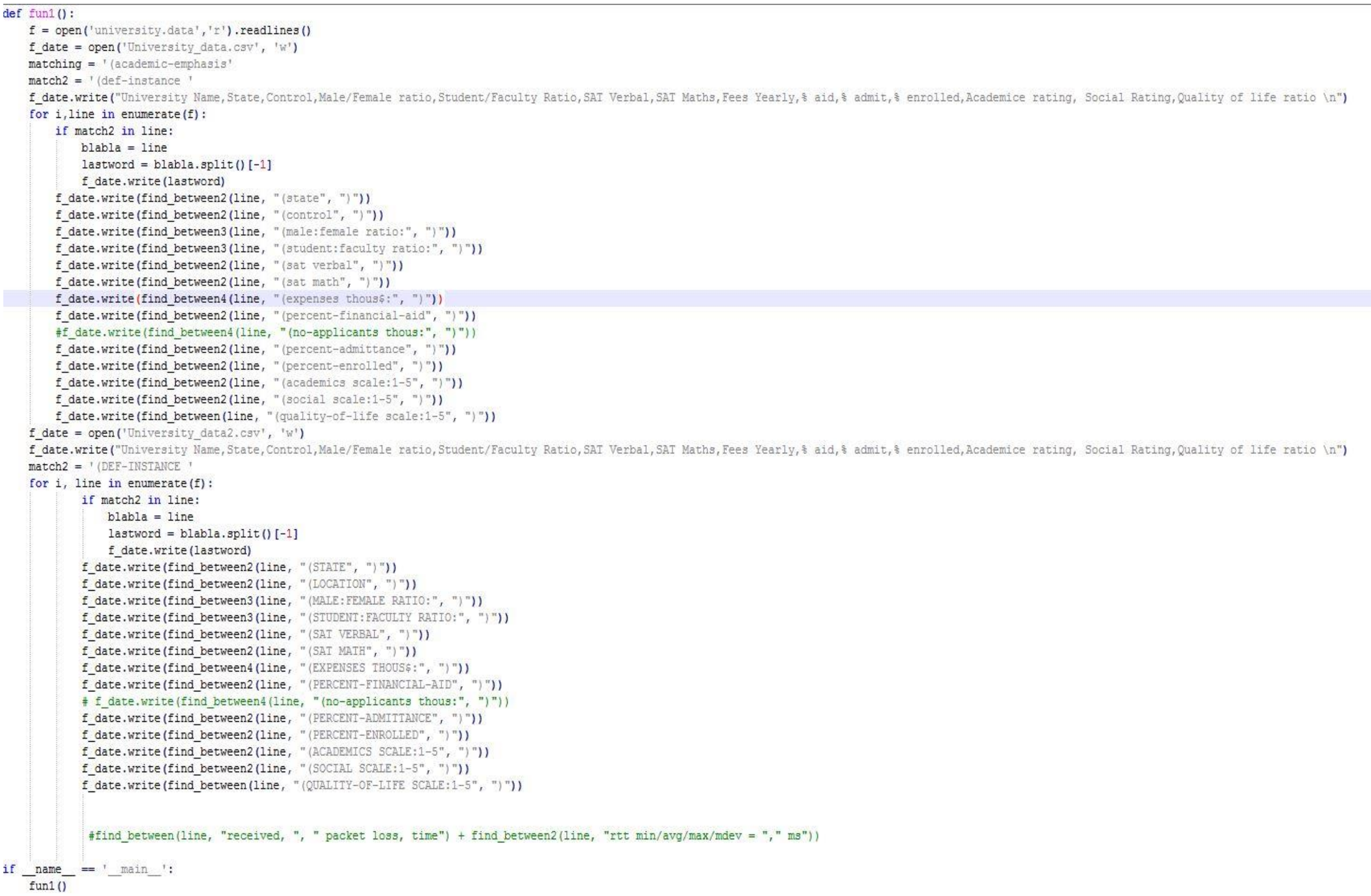
Step by Step Data Processing:

RAW data in LISP format:



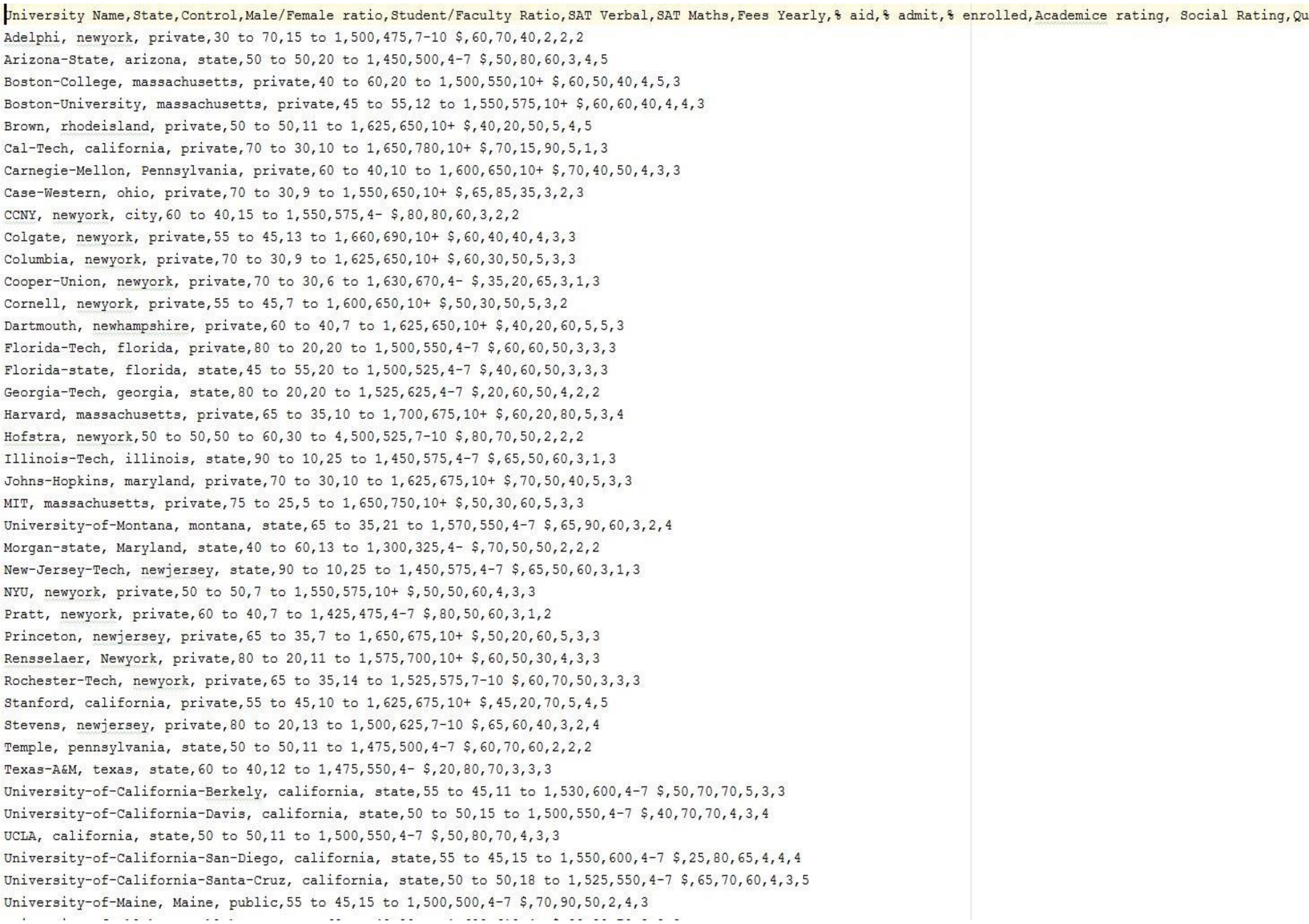
```
(def-instance Adelphi
   (state newyork)
   (control private)
   (no-of-students thous:5-10)
   (male:female ratio:30:70)
   (student:faculty ratio:15:1)
   (sat verbal 500)
   (sat math 475)
   (expenses thous$:7-10)
   (percent-financial-aid 60)
   (no-applicants thous:4-7)
   (percent-admittance 70)
   (percent-enrolled 40)
   (academics scale:1-5 2)
   (social scale:1-5 2)
   (quality-of-life scale:1-5 2)
   (academic-emphasis business-administration)
   (academic-emphasis biology))
(def-instance Arizona-State
   (state arizona)
   (control state)
   (no-of-students thous:20+)
   (male:female ratio:50:50)
   (student:faculty ratio:20:1)
   (sat verbal 450)
   (sat math 500)
   (expenses thous$:4-7)
   (percent-financial-aid 50)
   (no-applicants thous:17+)
   (percent-admittance 80)
   (percent-enrolled 60)
   (academics scale:1-5 3)
   (social scale:1-5 4)
   (quality-of-life scale:1-5 5)
   (academic-emphasis business-education)
   (academic-emphasis engineering)
   (academic-emphasis accounting)
   (academic-emphasis fine-arts))
(def-instance Boston-College
   (state massachusetts)
   (location suburban)
   (control private:roman-catholic)
```

Python Coding that converts LISP to CSV format:



```python
def fun1():
    f = open('university.data','r').readlines()
    f_date = open('University_data.csv', 'w')
    matching = '(academic-emphasis'
    match2 = '(def-instance '
    f_date.write("University Name,State,Control,Male/Female ratio,Student/Faculty Ratio,SAT Verbal,SAT Maths,Fees Yearly,% aid,% admit,% enrolled,Academice rating, Social Rating,Quality of life ratio \n")
    for i,line in enumerate(f):
        if match2 in line:
            blabla = line
            lastword = blabla.split()[-1]
            f_date.write(lastword)
        f_date.write(find_between2(line, "(state", ")"))
        f_date.write(find_between2(line, "(control", ")"))
        f_date.write(find_between3(line, "(male:female ratio:", ")"))
        f_date.write(find_between3(line, "(student:faculty ratio:", ")"))
        f_date.write(find_between2(line, "(sat verbal", ")"))
        f_date.write(find_between2(line, "(sat math", ")"))
        f_date.write(find_between4(line, "(expenses thous$:", ")"))
        f_date.write(find_between2(line, "(percent-financial-aid", ")"))
        #f_date.write(find_between4(line, "(no-applicants thous:", ")"))
        f_date.write(find_between2(line, "(percent-admittance", ")"))
        f_date.write(find_between2(line, "(percent-enrolled", ")"))
        f_date.write(find_between2(line, "(academics scale:1-5", ")"))
        f_date.write(find_between2(line, "(social scale:1-5", ")"))
        f_date.write(find_between(line, "(quality-of-life scale:1-5", ")"))
    f_date = open('University_data2.csv', 'w')
    f_date.write("University Name,State,Control,Male/Female ratio,Student/Faculty Ratio,SAT Verbal,SAT Maths,Fees Yearly,% aid,% admit,% enrolled,Academice rating, Social Rating,Quality of life ratio \n")
    match2 = '(DEF-INSTANCE '
    for i, line in enumerate(f):
            if match2 in line:
                blabla = line
                lastword = blabla.split()[-1]
                f_date.write(lastword)
            f_date.write(find_between2(line, "(STATE", ")"))
            f_date.write(find_between2(line, "(LOCATION", ")"))
            f_date.write(find_between3(line, "(MALE:FEMALE RATIO:", ")"))
            f_date.write(find_between3(line, "(STUDENT:FACULTY RATIO:", ")"))
            f_date.write(find_between2(line, "(SAT VERBAL", ")"))
            f_date.write(find_between2(line, "(SAT MATH", ")"))
            f_date.write(find_between4(line, "(EXPENSES THOUS$:", ")"))
            f_date.write(find_between2(line, "(PERCENT-FINANCIAL-AID", ")"))
            # f_date.write(find_between4(line, "(no-applicants thous:", ")"))
            f_date.write(find_between2(line, "(PERCENT-ADMITTANCE", ")"))
            f_date.write(find_between2(line, "(PERCENT-ENROLLED", ")"))
            f_date.write(find_between2(line, "(ACADEMICS SCALE:1-5", ")"))
            f_date.write(find_between2(line, "(SOCIAL SCALE:1-5", ")"))
            f_date.write(find_between(line, "(QUALITY-OF-LIFE SCALE:1-5", ")"))


             #find_between(line, "received, ", " packet loss, time") + find_between2(line, "rtt min/avg/max/mdev = "," ms"))

if __name__ == '__main__':
    fun1()
```

Converted CSV data:

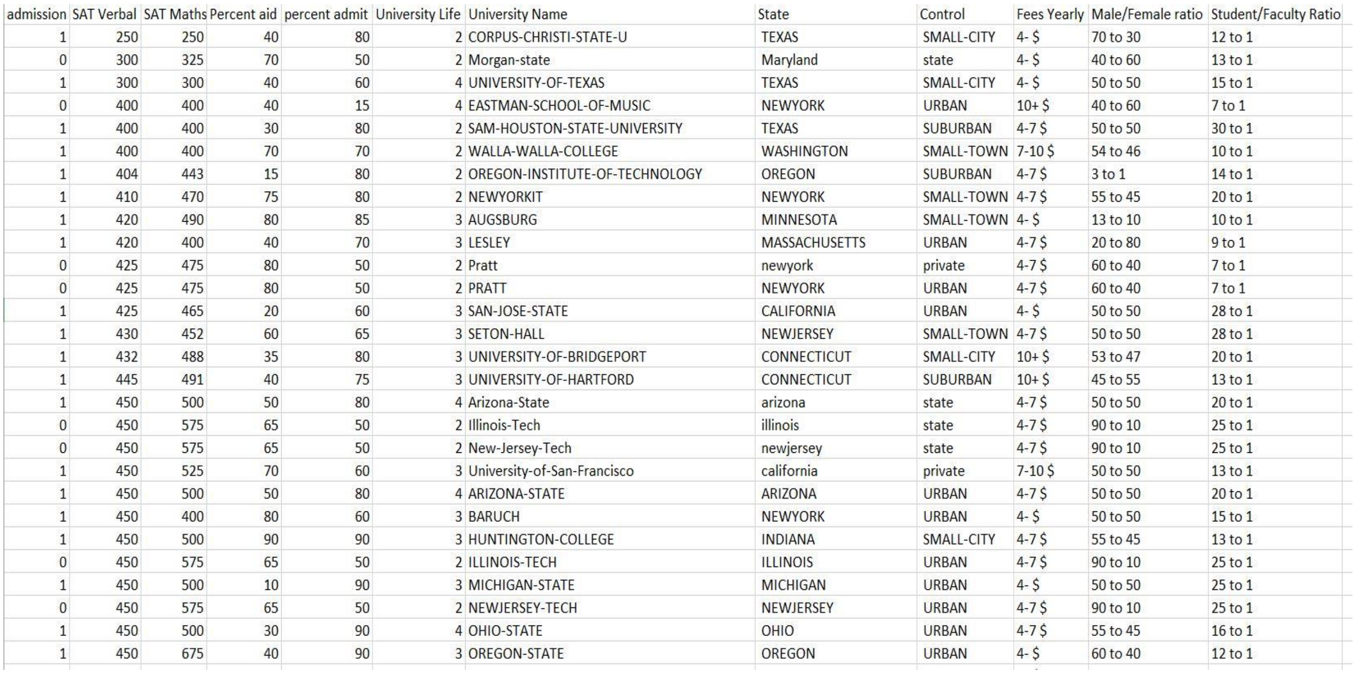


```
University Name,State,Control,Male/Female ratio,Student/Faculty Ratio,SAT Verbal,SAT Maths,Fees Yearly,% aid,% admit,% enrolled,Academice rating, Social Rating,Qu
Adelphi, newyork, private,30 to 70,15 to 1,500,475,7-10 $,60,70,40,2,2,2
Arizona-State, arizona, state,50 to 50,20 to 1,450,500,4-7 $,50,80,60,3,4,5
Boston-College, massachusetts, private,40 to 60,20 to 1,500,550,10+ $,60,50,40,4,5,3
Boston-University, massachusetts, private,45 to 55,12 to 1,550,575,10+ $,60,60,40,4,4,3
Brown, rhodeisland, private,50 to 50,11 to 1,625,650,10+ $,40,20,50,5,4,5
Cal-Tech, california, private,70 to 30,10 to 1,650,780,10+ $,70,15,90,5,1,3
Carnegie-Mellon, Pennsylvania, private,60 to 40,10 to 1,600,650,10+ $,70,40,50,4,3,3
Case-Western, ohio, private,70 to 30,9 to 1,550,650,10+ $,65,85,35,3,2,3
CCNY, newyork, city,60 to 40,15 to 1,550,575,4- $,80,80,60,3,2,2
Colgate, newyork, private,55 to 45,13 to 1,660,690,10+ $,60,40,40,4,3,3
Columbia, newyork, private,70 to 30,9 to 1,625,650,10+ $,60,30,50,5,3,3
Cooper-Union, newyork, private,70 to 30,6 to 1,630,670,4- $,35,20,65,3,1,3
Cornell, newyork, private,55 to 45,7 to 1,600,650,10+ $,50,30,50,5,3,2
Dartmouth, newhampshire, private,60 to 40,7 to 1,625,650,10+ $,40,20,60,5,5,3
Florida-Tech, florida, private,80 to 20,20 to 1,500,550,4-7 $,60,60,50,3,3,3
Florida-state, florida, state,45 to 55,20 to 1,500,525,4-7 $,40,60,50,3,3,3
Georgia-Tech, georgia, state,80 to 20,20 to 1,525,625,4-7 $,20,60,50,4,2,2
Harvard, massachusetts, private,65 to 35,10 to 1,700,675,10+ $,60,20,80,5,3,4
Hofstra, newyork,50 to 50,50 to 60,30 to 4,500,525,7-10 $,80,70,50,2,2,2
Illinois-Tech, illinois, state,90 to 10,25 to 1,450,575,4-7 $,65,50,60,3,1,3
Johns-Hopkins, maryland, private,70 to 30,10 to 1,625,675,10+ $,70,50,40,5,3,3
MIT, massachusetts, private,75 to 25,5 to 1,650,750,10+ $,50,30,60,5,3,3
University-of-Montana, montana, state,65 to 35,21 to 1,570,550,4-7 $,65,90,60,3,2,4
Morgan-state, Maryland, state,40 to 60,13 to 1,300,325,4- $,70,50,50,2,2,2
New-Jersey-Tech, newjersey, state,90 to 10,25 to 1,450,575,4-7 $,65,50,60,3,1,3
NYU, newyork, private,50 to 50,7 to 1,550,575,10+ $,50,50,60,4,3,3
Pratt, newyork, private,60 to 40,7 to 1,425,475,4-7 $,80,50,60,3,1,2
Princeton, newjersey, private,65 to 35,7 to 1,650,675,10+ $,50,20,60,5,3,3
Rensselaer, Newyork, private,80 to 20,11 to 1,575,700,10+ $,60,50,30,4,3,3
Rochester-Tech, newyork, private,65 to 35,14 to 1,525,575,7-10 $,60,70,50,3,3,3
Stanford, california, private,55 to 45,10 to 1,625,675,10+ $,45,20,70,5,4,5
Stevens, newjersey, private,80 to 20,13 to 1,500,625,7-10 $,65,60,40,3,2,4
Temple, pennsylvania, state,50 to 50,11 to 1,475,500,4-7 $,60,70,60,2,2,2
Texas-A&M, texas, state,60 to 40,12 to 1,475,550,4- $,20,80,70,3,3,3
University-of-California-Berkely, california, state,55 to 45,11 to 1,530,600,4-7 $,50,70,70,5,3,3
University-of-California-Davis, california, state,50 to 50,15 to 1,500,550,4-7 $,40,70,70,4,3,4
UCLA, california, state,50 to 50,11 to 1,500,550,4-7 $,50,80,70,4,3,3
University-of-California-San-Diego, california, state,55 to 45,15 to 1,550,600,4-7 $,25,80,65,4,4,4
University-of-California-Santa-Cruz, california, state,50 to 50,18 to 1,525,550,4-7 $,65,70,60,4,3,5
University-of-Maine, Maine, public,55 to 45,15 to 1,500,500,4-7 $,70,90,50,2,4,3
```

For Data-Analysis:

In this phase, we performed some analysis based on the available data in the **csv** file where we convert the admission percentage into a new variable by doing some analysis and research over the internet and calculated the admission chances also, we programmed it in python to calculate University Life considering 3 different values namely, Social Scale, Academic Life, and Quality of life. This data is used for giving using information about the university Life which is one of the important factors which guides the students to select a college.

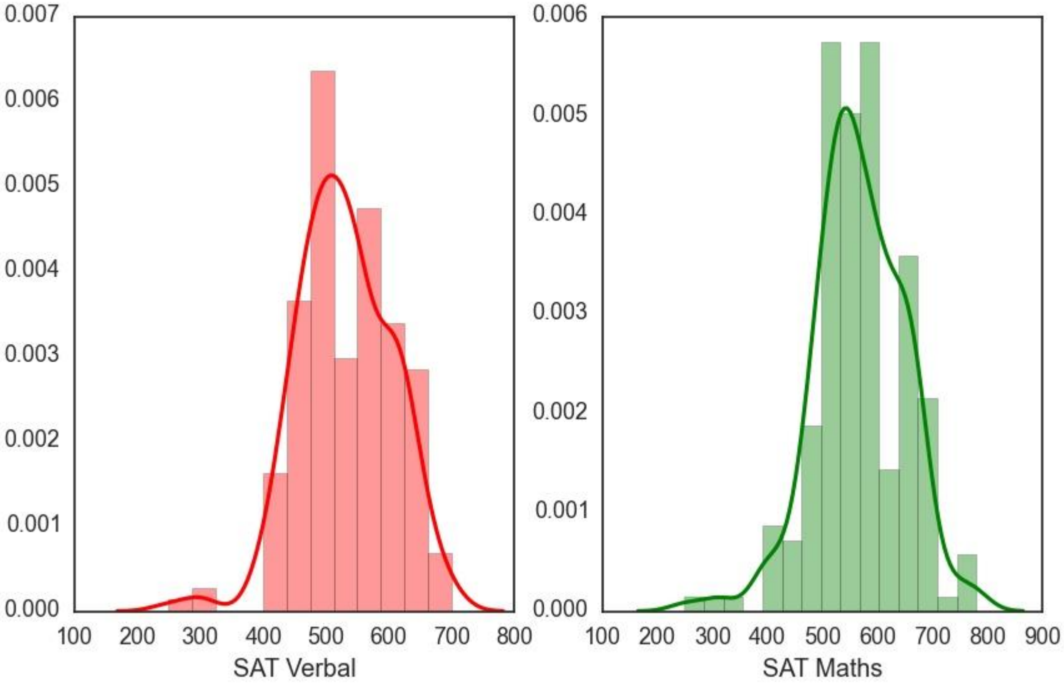
.



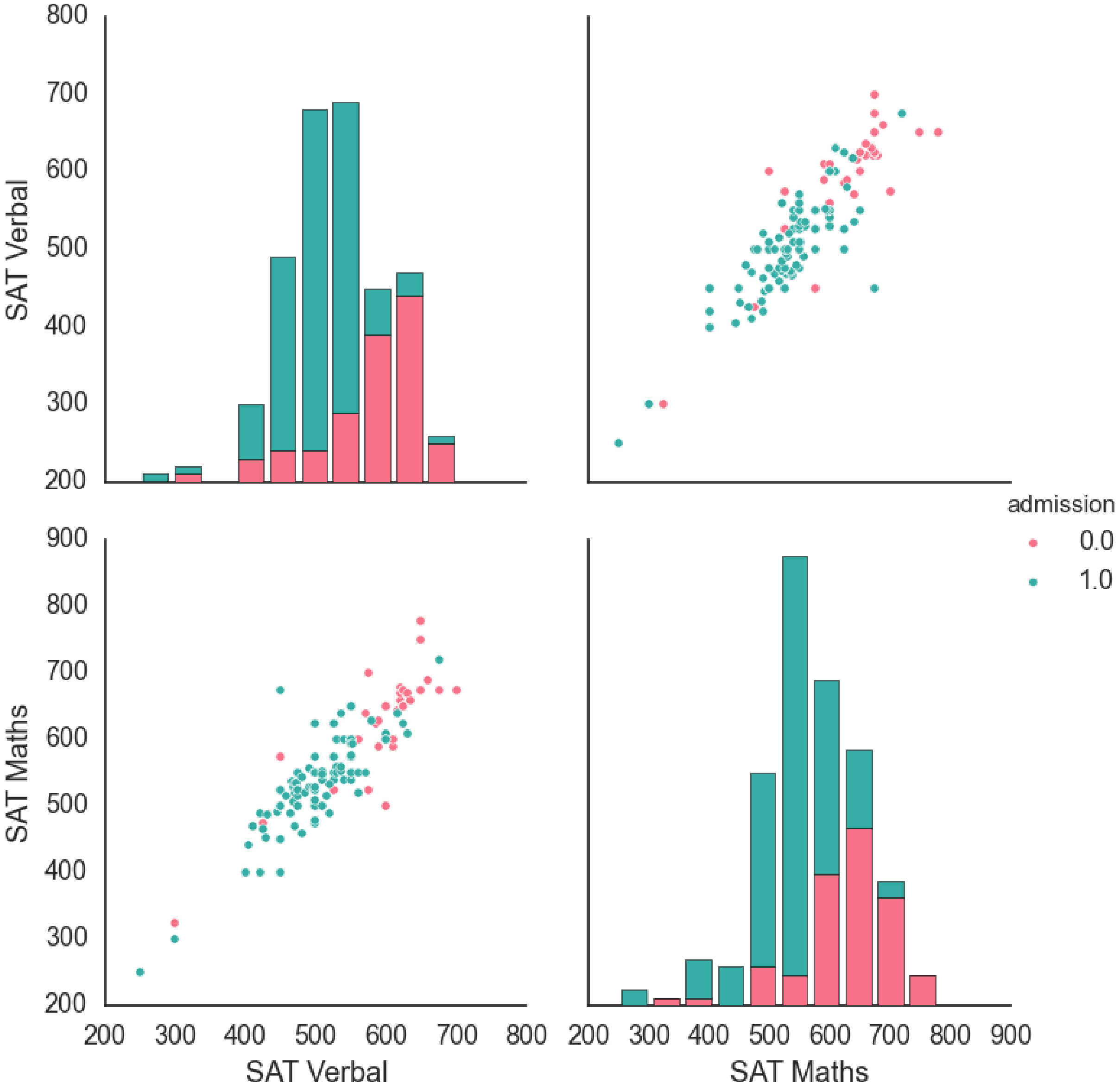
admission	SAT Verbal	SAT Maths	Percent aid	percent admit	University Life	University Name	State	Control	Fees Yearly	Male/Female ratio	Student/Faculty Ratio
1	250	250	40	80	2	CORPUS-CHRISTI-STATE-U	TEXAS	SMALL-CITY	4- $	70 to 30	12 to 1
0	300	325	70	50	2	Morgan-state	Maryland	state	4- $	40 to 60	13 to 1
1	300	300	40	60	4	UNIVERSITY-OF-TEXAS	TEXAS	SMALL-CITY	4- $	50 to 50	15 to 1
0	400	400	40	15	4	EASTMAN-SCHOOL-OF-MUSIC	NEWYORK	URBAN	10+ $	40 to 60	7 to 1
1	400	400	30	80	2	SAM-HOUSTON-STATE-UNIVERSITY	TEXAS	SUBURBAN	4-7 $	50 to 50	30 to 1
1	400	400	70	70	2	WALLA-WALLA-COLLEGE	WASHINGTON	SMALL-TOWN	7-10 $	54 to 46	10 to 1
1	404	443	15	80	2	OREGON-INSTITUTE-OF-TECHNOLOGY	OREGON	SUBURBAN	4-7 $	3 to 1	14 to 1
1	410	470	75	80	2	NEWYORKIT	NEWYORK	SMALL-TOWN	4-7 $	55 to 45	20 to 1
1	420	490	80	85	3	AUGSBURG	MINNESOTA	SMALL-TOWN	4- $	13 to 10	10 to 1
1	420	400	40	70	3	LESLEY	MASSACHUSETTS	URBAN	4-7 $	20 to 80	9 to 1
0	425	475	80	50	2	Pratt	newyork	private	4-7 $	60 to 40	7 to 1
0	425	475	80	50	2	PRATT	NEWYORK	URBAN	4-7 $	60 to 40	7 to 1
1	425	465	20	60	3	SAN-JOSE-STATE	CALIFORNIA	URBAN	4- $	50 to 50	28 to 1
1	430	452	60	65	3	SETON-HALL	NEWJERSEY	SMALL-TOWN	4-7 $	50 to 50	28 to 1
1	432	488	35	80	3	UNIVERSITY-OF-BRIDGEPORT	CONNECTICUT	SMALL-CITY	10+ $	53 to 47	20 to 1
1	445	491	40	75	3	UNIVERSITY-OF-HARTFORD	CONNECTICUT	SUBURBAN	10+ $	45 to 55	13 to 1
1	450	500	50	80	4	Arizona-State	arizona	state	4-7 $	50 to 50	20 to 1
0	450	575	65	50	2	Illinois-Tech	illinois	state	4-7 $	90 to 10	25 to 1
0	450	575	65	50	2	New-Jersey-Tech	newjersey	state	4-7 $	90 to 10	25 to 1
1	450	525	70	60	3	University-of-San-Francisco	california	private	7-10 $	50 to 50	13 to 1
1	450	500	50	80	4	ARIZONA-STATE	ARIZONA	URBAN	4-7 $	50 to 50	20 to 1
1	450	400	80	60	3	BARUCH	NEWYORK	URBAN	4- $	50 to 50	15 to 1
1	450	500	90	90	3	HUNTINGTON-COLLEGE	INDIANA	SMALL-CITY	4-7 $	55 to 45	13 to 1
0	450	575	65	50	2	ILLINOIS-TECH	ILLINOIS	URBAN	4-7 $	90 to 10	25 to 1
1	450	500	10	90	3	MICHIGAN-STATE	MICHIGAN	URBAN	4- $	50 to 50	25 to 1
0	450	575	65	50	2	NEWJERSEY-TECH	NEWJERSEY	URBAN	4-7 $	90 to 10	25 to 1
1	450	500	30	90	4	OHIO-STATE	OHIO	URBAN	4-7 $	55 to 45	16 to 1
1	450	675	40	90	3	OREGON-STATE	OREGON	URBAN	4- $	60 to 40	12 to 1

Data Visualization:

It’s time to get into colorful charts and graphs. **Data Visualization** is the most important skill every Data Scientist needs. This is because only from graphs and charts, our brain could visualize those 400 rows of raw data. If you were asked to go through each row *line-by-line* and compute what is the range of SAT scores that students get, you are bound to get confused. It is not that much harder (i.e. you could work it out manually, which is time-consuming), but instead, Data Visualization tools give us the power to explore **hidden patterns** in the dataset. For prediction of Analysis



In the above two plots, we could see that SAT Verbal score mostly fall within the range **(250-700)** and SAT Maths grade mostly fall within the range **(250-700)**. This could not be guessed just by looking at the *raw csv file*. That is why we go for Data Visualization.

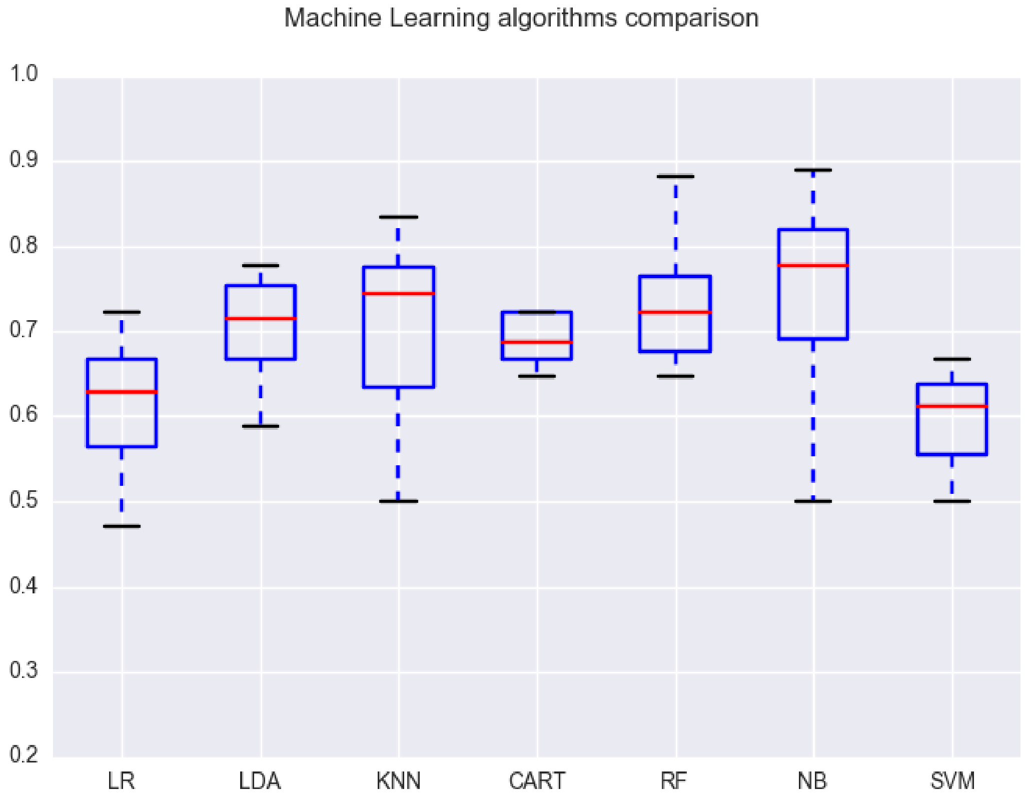


From the above scatterplot matrix, we could see that (SAT Verbal vs SAT Maths) is **highly correlated,** therefore we use the power of machine learning models such as Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Trees, Random Forests, Gaussian Naive Bayes and Support Vector Machine.

Predictive Analysis:

After getting enough information and visualizing the dataset, we tried implementing **Machine Learning models** to perform predictive analysis. We explored different kinds of ML models that gives different accuracies on the same dataset. After that we selected the Linear Discriminant Analysis Model because the following boxplot shows the accuracy of different Machine Learning Algorithm.

.



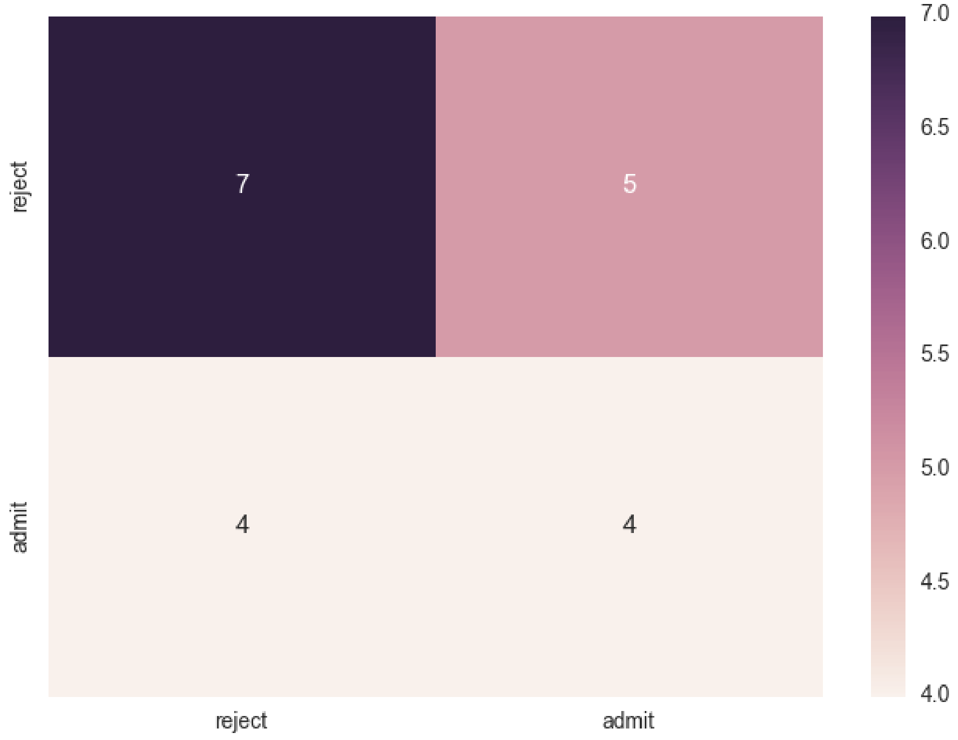
LDA has been shortly explained in the below few lines.

Listed below are the 5 general steps for performing a linear discriminant analysis; we will explore them in more detail in the following sections.

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors (e_1, e_2 ,...,e_d) and corresponding eigenvalues (λ_1,λ_2,...,λ_d) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W (where every column represents an eigenvector).
5. Use this d × k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $Y=X \times W$ (where X is a n × d-dimensional matrix representing the n samples, and y are the transformed n × k-dimensional samples in the new subspace).

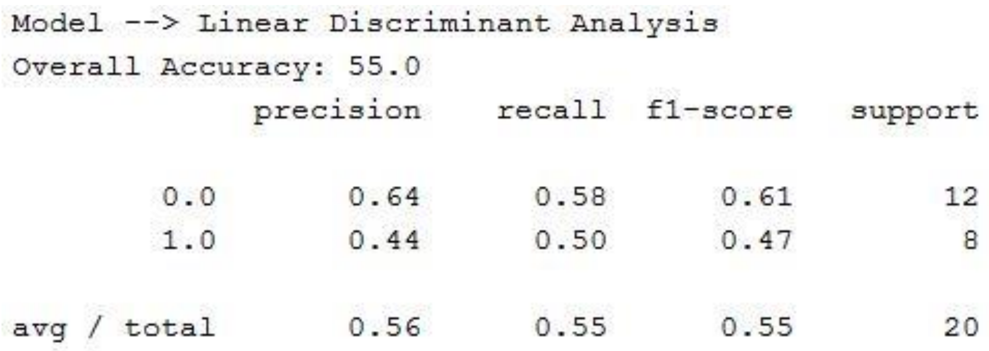
Coming back to our project, we also explore how to split training and testing data from the original dataset, so that we can make the model learn from the training data and predict on the test data. On performing we get the following Confusion Matrix Shown as a heatmap.

.



Experiment Results:

After training the Model we had set some testing data to test the data predicted using a testing array of data. This was then compared with the actual results which the dataset already consisted of. On comparing both the values we get the following information about the algorithm executed, its precision and support.



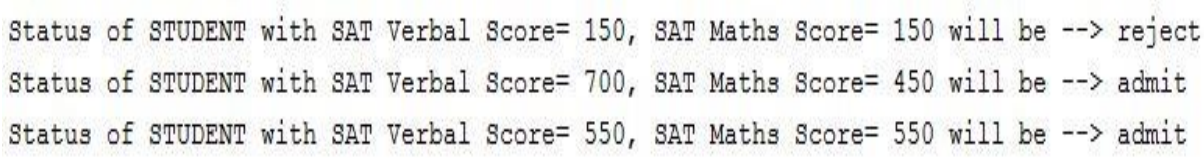
```
Model --> Linear Discriminant Analysis
Overall Accuracy: 55.0
             precision    recall  f1-score   support

        0.0       0.64      0.58      0.61        12
        1.0       0.44      0.50      0.47         8

avg / total       0.56      0.55      0.55        20
```

Than later we provided information of 3 students and ran the prediction algorithm we got the final input as follows:

Data provided

- Student 1: - SAT Verbal Score= 150, SAT Math's Score= 150
- Student 2: - SAT Verbal Score= 700, SAT Math's Score= 450
- Student 3: - SAT Verbal Score= 550, SAT Math's Score= 550

Output is as Follows:



```
Status of STUDENT with SAT Verbal Score= 150, SAT Maths Score= 150 will be --> reject
Status of STUDENT with SAT Verbal Score= 700, SAT Maths Score= 450 will be --> admit
Status of STUDENT with SAT Verbal Score= 550, SAT Maths Score= 550 will be --> admit
```

As you can see, a student with a high SAT score, will eventually get an **REJECT**. While, the other two test cases got **ADMIN** due to low SAT scores and low GPA grades.

Problems:

In the above report as we only had a limited amount of data i.e. the data of 400 universities we could not give students information of all the universities. Also, while analyzing Box-plot we found that LDA was one of the best as it had an accuracy of 70 % but while executing the algorithm the accuracy decreased to 55%. According to our analysis, the main reason for this failure can be less training data available to the model therefore there was a 15% decrease in accuracy.

Discussion

As we had described the problem that is faced by the students for selection of colleges, after getting the results we were able to predict the admission status based on SAT scores. As we don’t know what the final output would have been, if we had selected a different Algorithm for prediction the prediction accuracy is a question to be answered. The answer remains in the grey zone for selection of the most appropriate algorithm.

This report can be used by High schools to guide their students as to, in what range their scores need to be in order to get an admission in their dream college. This can also be used by students on their own as once they have their SAT Scores they can have an idea about different universities they need to apply to have good chance to get admitted in that university.

This report can be used as a base product to analyze further prediction on large data as it is able to predict the admit status. This report can be further used for analysis of Male/ Female ratio and Student/Faculty Ratio and further provide an accurate prediction for university selection based on our work.

References:

1. Python Software Foundation
   https://www.python.org/

2. The pandas community
   http://pandas.pydata.org/community.html

3. UCI machine learning repository
   https://archive.ics.uci.edu/ml/datasets/University

4. Linear Discriminant Analysis for Machine Learning by Jason Brownlee
   http://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

5. U.S.News and world report
   https://www.usnews.com/

6. Veritas, LLC.
   https://www.veritasprep.com/college/free-college-admissions-calculator/

7. The College Board and PSAT/NMSQT
   https://collegereadiness.collegeboard.org/