Mikias Berhanu | 2021280115
Assignment IV
Supervised and Unsupervised Learning

Lab Report on Supervised Learning

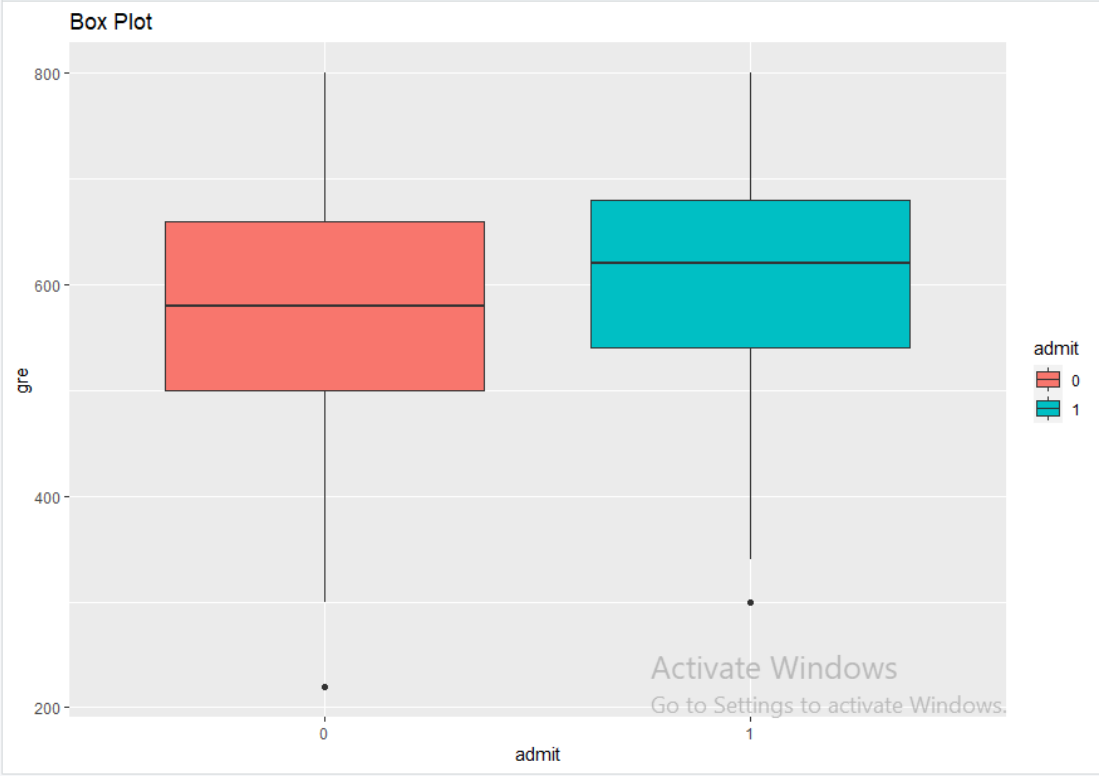
The first lab report is performed on a dataset composed of numeric values. The dataset contains 4 columns each containing individual students' records like GRE, GPA, RANK and also a binary valued column ADMIT. In total the dataset has 400 records and based on this dataset we will perform classification task. Given a student's GRE, GPA and RANK values, find out if a student is admitted to a university or not.

admit	gre	gpa	rank
0	380	3.61	3
1	660	3.67	3
1	800	4.00	1
1	640	3.19	4
0	520	2.93	4
1	760	3.00	2
1	560	2.98	1
0	400	3.08	2
1	540	3.39	3
0	700	3.92	2
0	800	4.00	4
0	440	3.22	1
1	760	4.00	1
0	700	3.08	2
1	700	4.00	1
0	480	3.44	3
0	780	3.87	4
0	360	2.56	3
0	800	3.75	2
1	540	3.81	1

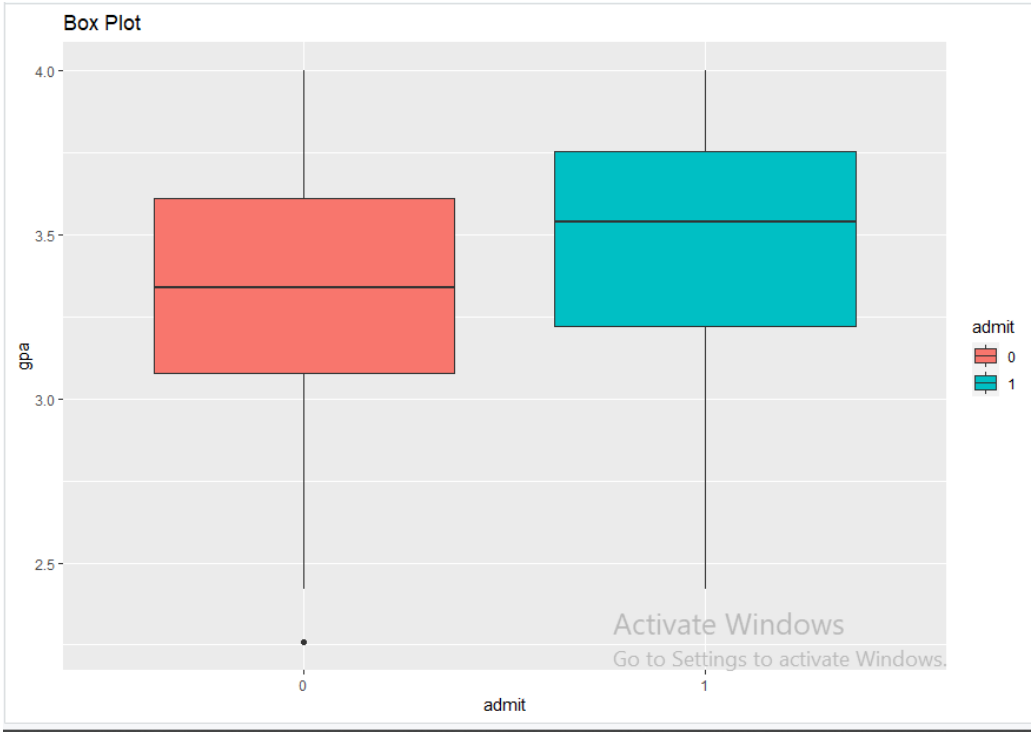
Dataset table view

Once the data is loaded it's always a good idea to visualize the data and see what kind of data we are working with. The first kind of visualization is the box plot which is very effective and easy to read. It will help us to generalize data from multiple sources and display the results in a single graph, this will help us later to make effective decisions. The first box plot is done between the GRE score of students and the ADMIT column of the dataset. From the box plot it's clear to see that there is an

imbalance between the number of admitted and not admitted students. The number of admitted students is slightly higher than those who are not admitted. This will create some baise later in the model and the performance of the model. To solve this issue we can use Stratified Sampling which is used to tackle imbalance of data points in training and testing data. The imbalance in the data is also show on the plot of the GPA vs ADMIT values on the second plot.

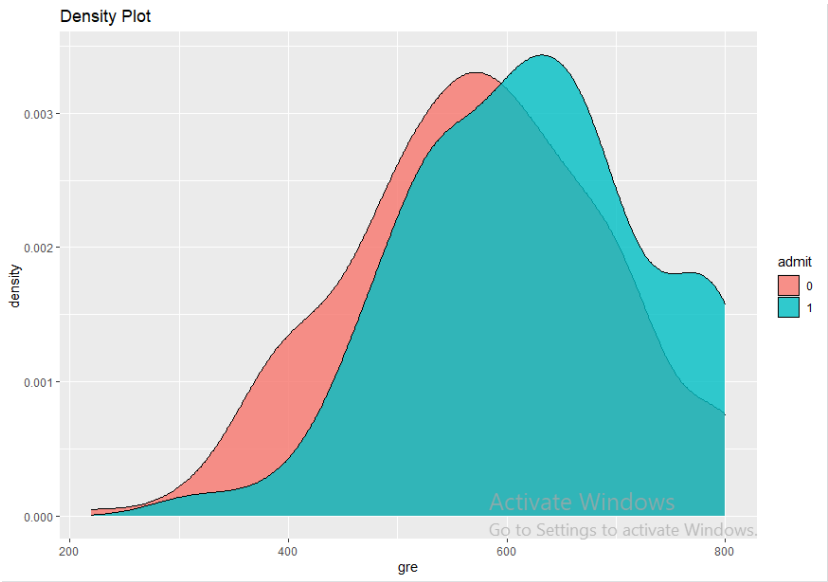


Box plot for ADMIT Vs GRE score

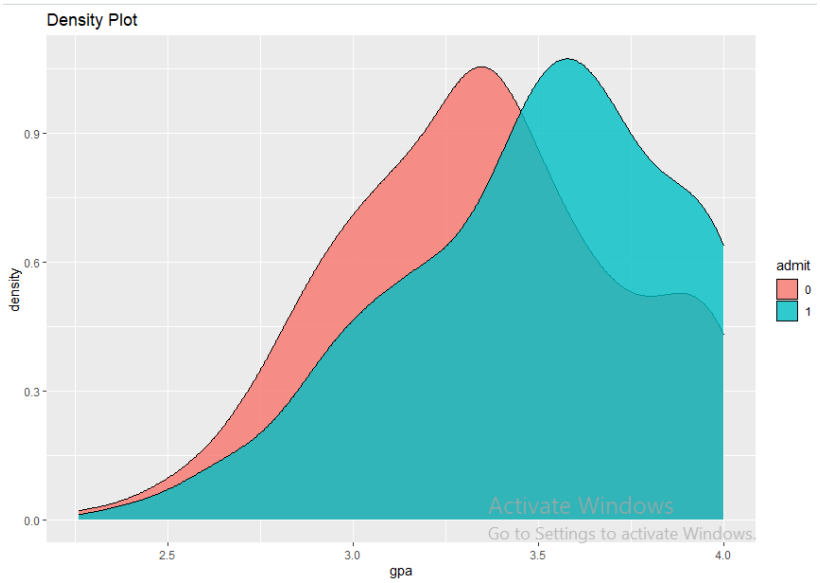


Box plot for ADMIT Vs GPA score

The next visualization to use is density plot which will help us to see the distribution of values in our dataset. The first density plot is between GRE score and ADMIT values and the second density plot is between GPA score and ADMIT values. The distribution of variables on both plots is not uniform and there is an overlap between these values.



Density plot GRE Vs Admit



Density plot GRE Vs Admit

Once we visualize our dataset the next thing to do is split the data into training and testing sets. The training set is used to train our model and the testing set is used to test our model. The dataset is splitted in 80:20 ratio meaning 80% of the data is used for training the dataset and 20% of the training set is used for testing the dataset. The dataset can be summarized as well the mean of the GRE score when the value of ADMIT is equal to 0 is 578.6547 and the standard deviation is 116.325 and when the value of ADMIT is equal to 1 is 622.9412 for the mean value 110.924 for the standard deviation.

Once the data is splitted into a training and testing set, we use the Naive Bayes model and train it with our training set. The Naive Bayes model is a supervised learning model algorithm based on applying the bayes theorem with “naive” assumption of conditional independence between every pair of features given the value of the class variable. Once the model is trained we can do our prediction and also calculate the confusion matrix. Confusion matrix is like an error matrix which tells us about the performance of our model and mainly used in supervised learning. From our confusion matrix on the training set we can see that 229 values are

correctly classified whereas 96 values are misclassified. When we calculate the confusion matrix on our testing set we get 51 values correctly classified and 24 values misclassified.

The second experiment is done using Logistic Regression on the same dataset. Logistic Regression is a regression analysis algorithm which is used when the class we want to predict or when the dependent variable is binary. Despite the fact that the name says Logistic Regression, this algorithm is used for classification tasks. Logistic Regression is used to describe the relationship between one dependent variable in our case the ADMIT column and other independent variables like the GRE, GPA etc… The dataset is still having the same content and load the same way as the first experiment. The summary of the model is shown below. The model is trained on the training set and the prediction is done both on the training set and testing set.

```
Call:
glm(formula = admit ~ +gpa + rank, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5156  -0.8880  -0.6318   1.1091   2.1688

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.7270     1.2918  -3.659 0.000253 ***
gpa           1.3735     0.3590   3.826 0.000130 ***
rank2        -0.5712     0.3564  -1.603 0.108976
rank3        -1.1645     0.3804  -3.061 0.002203 **
rank4        -1.5642     0.4756  -3.289 0.001005 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 404.39  on 324  degrees of freedom
Residual deviance: 371.81  on 320  degrees of freedom
AIC: 381.81

Number of Fisher Scoring iterations: 4
```

Model summary for Logistic Regression

Summary

The experiment was based on Supervised Machine Learning specifically classification tasks. We used a single dataset with 400 entries and 4 columns, 1 dependent variable and 3 independent variable. We used two supervised learning algorithms Naive Bayes and Logistic Regression for the purpose of classification. We also explored a popular method of looking at the performance of our models called confusion matrix.