IHE-7510-01 - Data Mining

Project 1

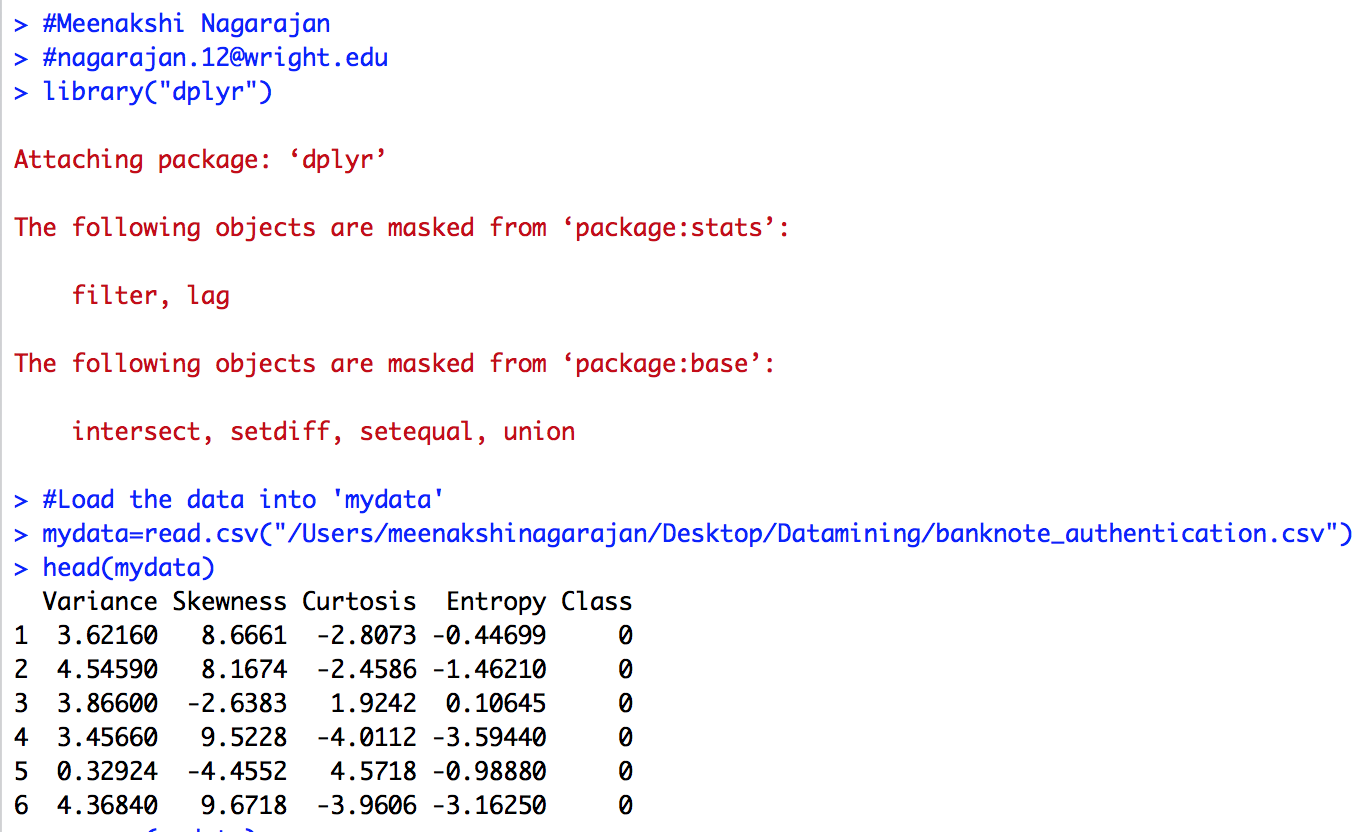
Meenakshi Nagarajan

[Nagarajan.12@wright.edu](mailto:Nagarajan.12@wright.edu)

#Dataset – Banknote Authentication

# <http://archive.ics.uci.edu/ml/datasets/banknote+authentication>

#Initially the dataset has been loaded into a variable called ‘mydata’



***Fig. 1*** *Read the data*

#Once the data has been loaded, the summary of data has been analyzed.

**Background of data**

The image of a bank note is pre-processed and the classification features were extracted. Later the note is classified as genuine or forged note based on the features extracted. In this project, the correct combination of features used to determine the authenticity of the banknote is being identified statistically.

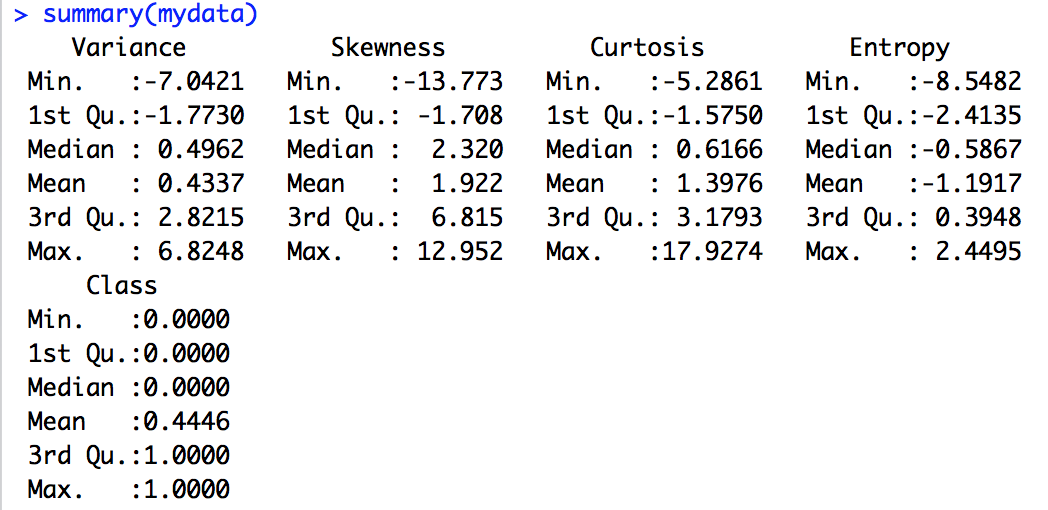
The dataset used in this study is obtained from the UCI Machine learning repository. ([**http://archive.ics.uci.edu/ml/datasets.html**](http://archive.ics.uci.edu/ml/datasets.html) **)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | **Dataset Characteristics** | Multivariate | 4 | **Number of instances** | 1372 |
| 2 | **Attribute characteristics** | Real | 5 | **Number of Attributes** | 5 |
| 3 | **Date Donated** | 2013/ 04/ 16 | 6 | **Missing values** | None |

**Attributes**

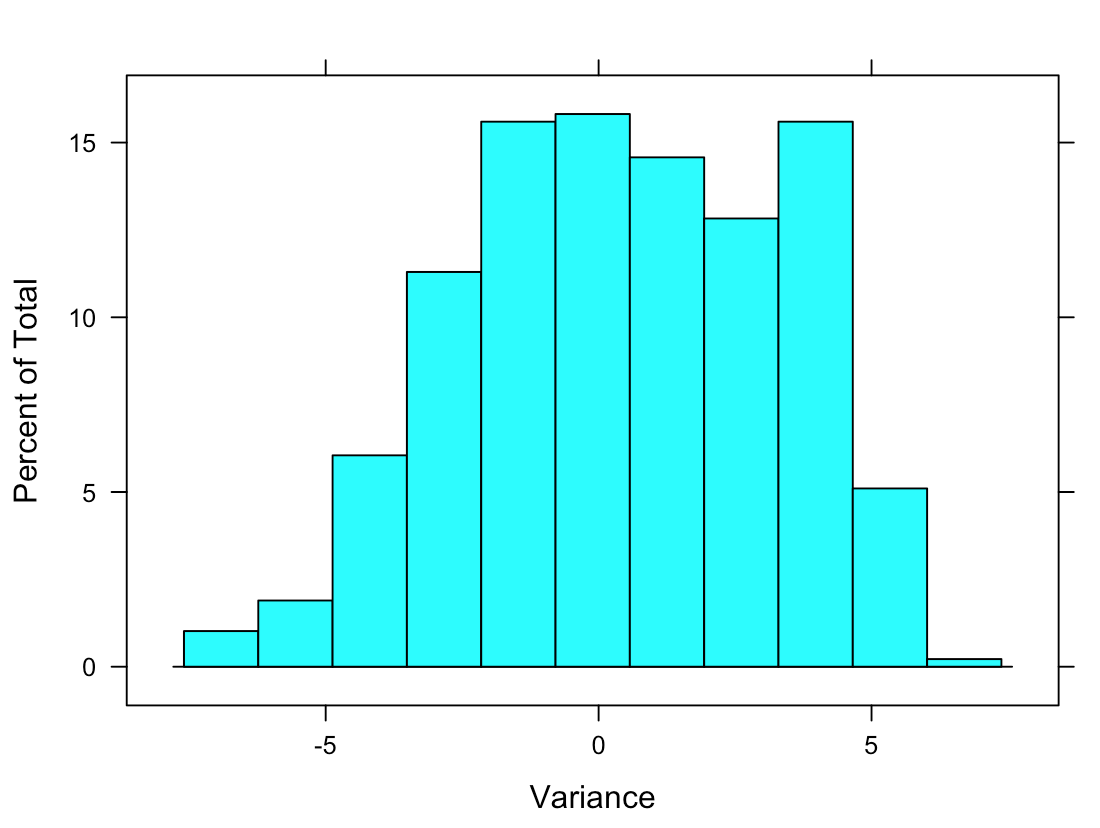
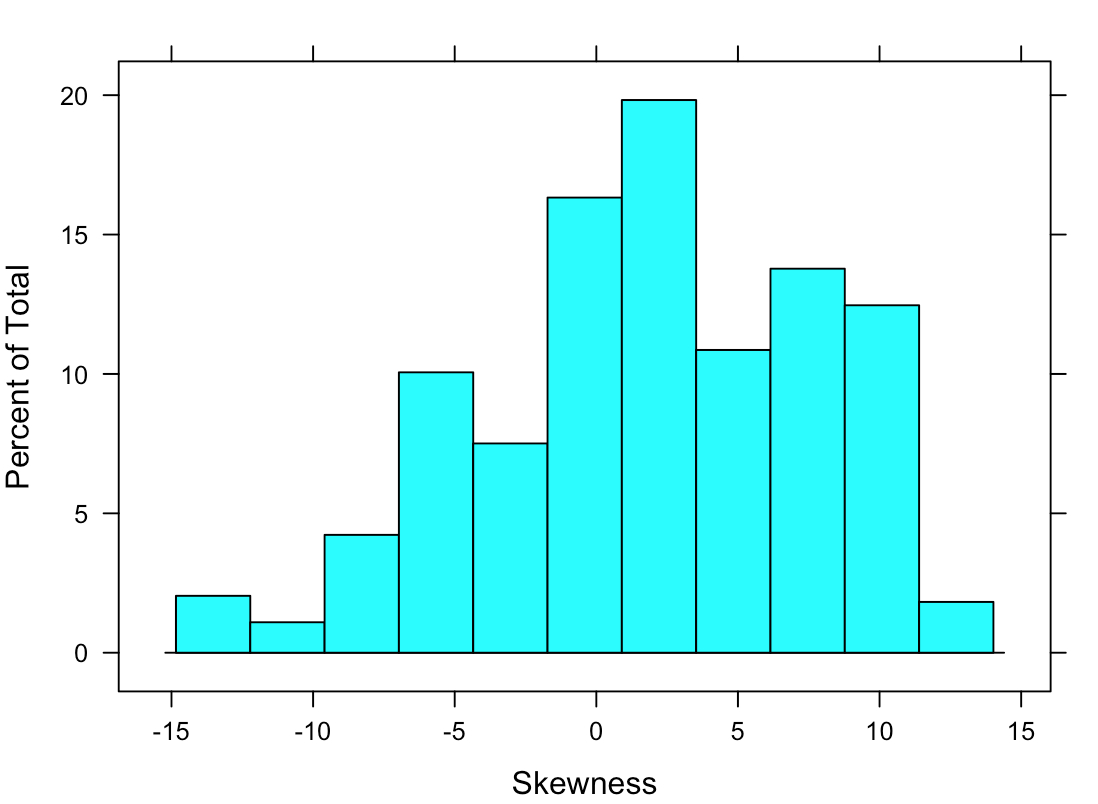
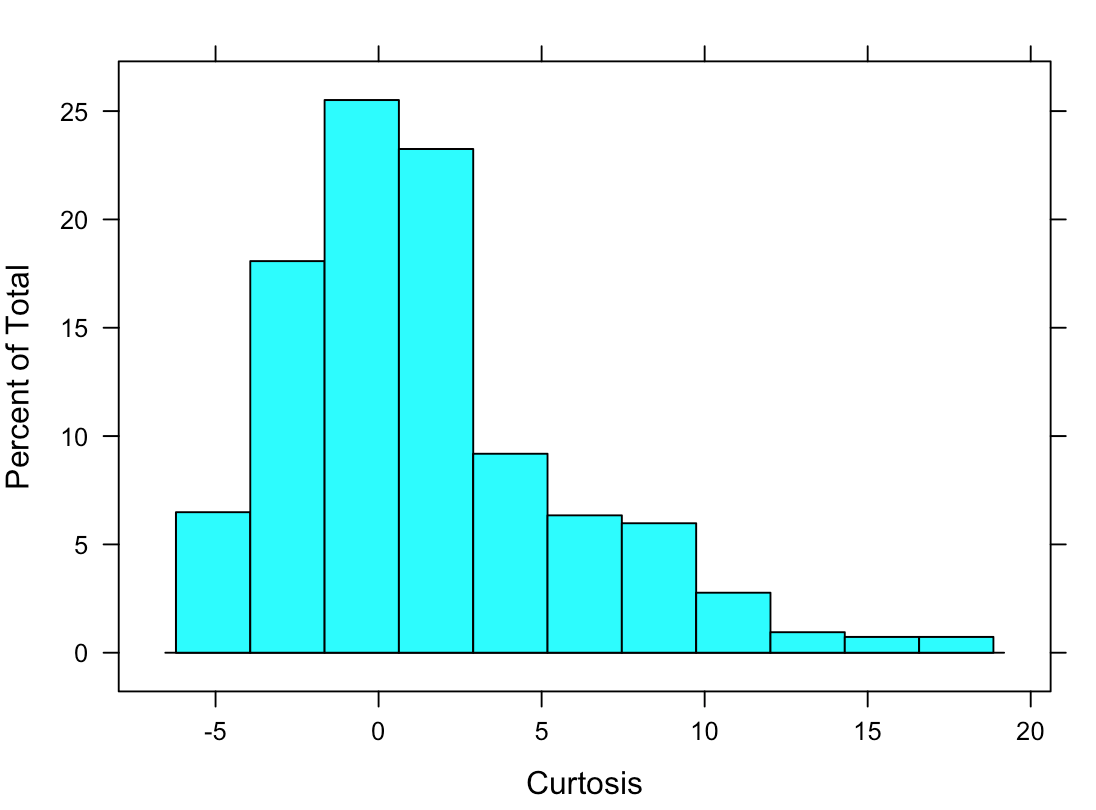
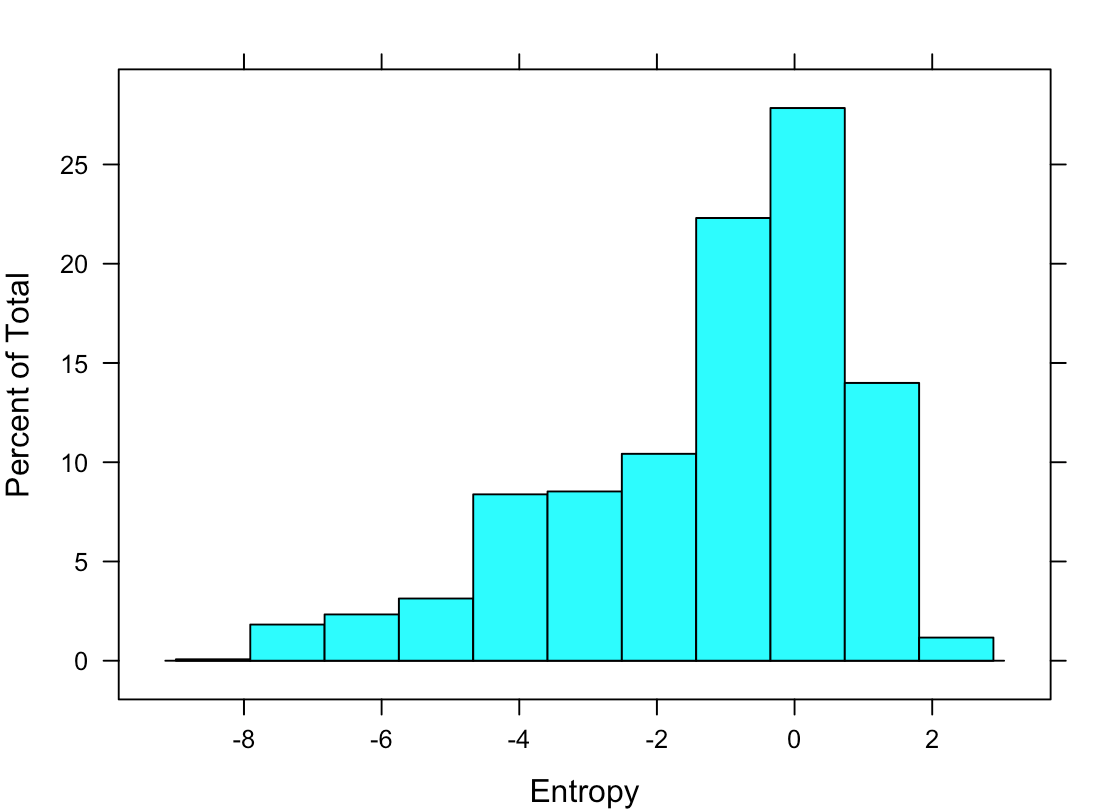
|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Datatype** | **Meaning** |
| Variance | Numerical | Variance gives the amplitude distribution of the Wavelet coefficients around the center of histogram. |
| Skewness | Numerical | Skewness is the symmetry of the distribution of data around the center. |
| Kurtosis | Numerical | Kurtosis describes the deviation relative to the Gaussian distribution. |
| Entropy | Numerical | Entropy/ average information of an image. |
| Class | Binary | Authenticity (Output 0 means genuine and Output 1 means forged). |

**Insights**



***Fig. 2*** *Summary of data*

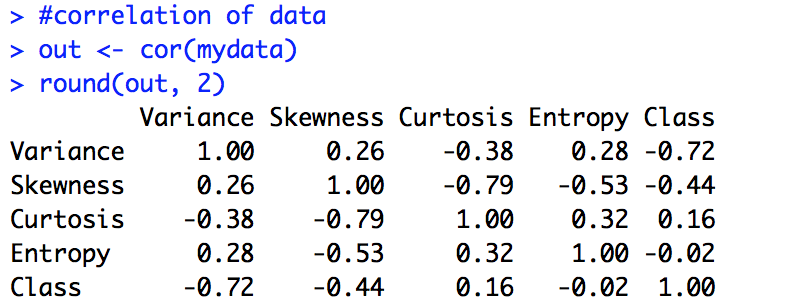
#Histogram for all the attributes has been constructed to check the data distribution

   .. 

***Fig. 3*** *Histogram*

#From the distribution of data, some outlying observations could be seen for Kurtosis and Entropy.

#Here, it is not evident that the outliers are due to recording errors. Therefore, they are not removed for the data analysis

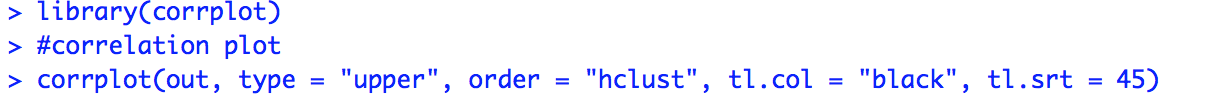


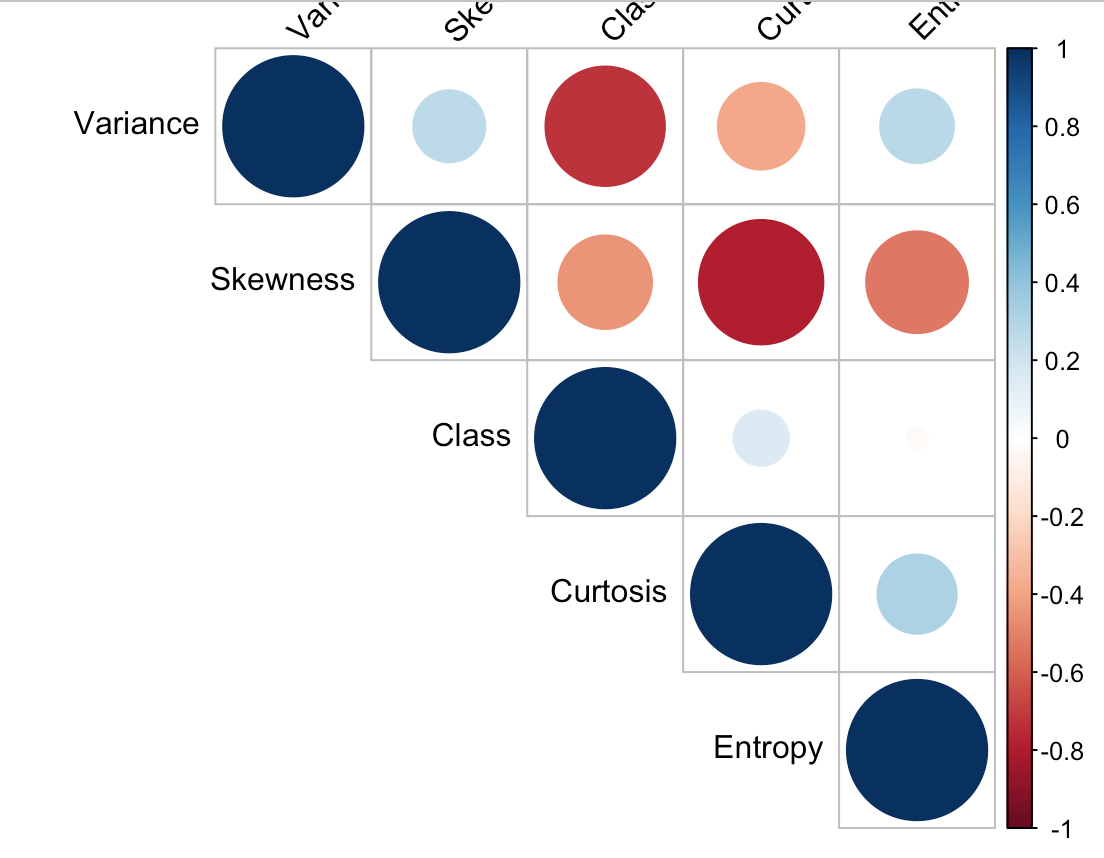
***Fig. 4*** *Correlation matrix*

#Correlation matrix has been constructed to see how well the variables are correlated with each other.

#Here, high negative correlation is observed between Skewness and a decent positive correlation is observed between kurtosis and entropy.

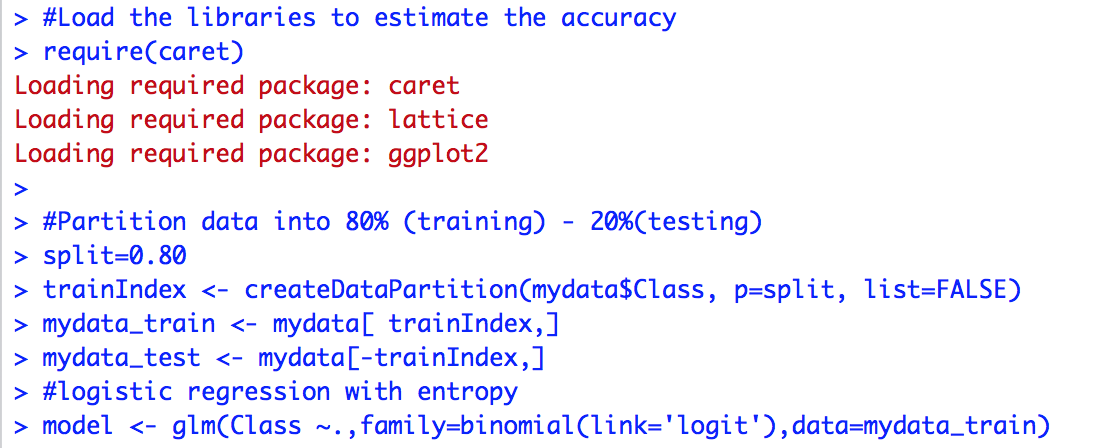
#using the library ‘corrplot’, a graphical display of correlation matrix is obtained, which can be used for quick interpretation





***Fig. 5*** *Correlation plot*

#From the plot, it is clear that there is zero correlation between the class attribute and Entropy. Therefore, the use of entropy in the classification, can be determined by classifying the dataset into training and testing data and computing the classification accuracy of the model with and without entropy.

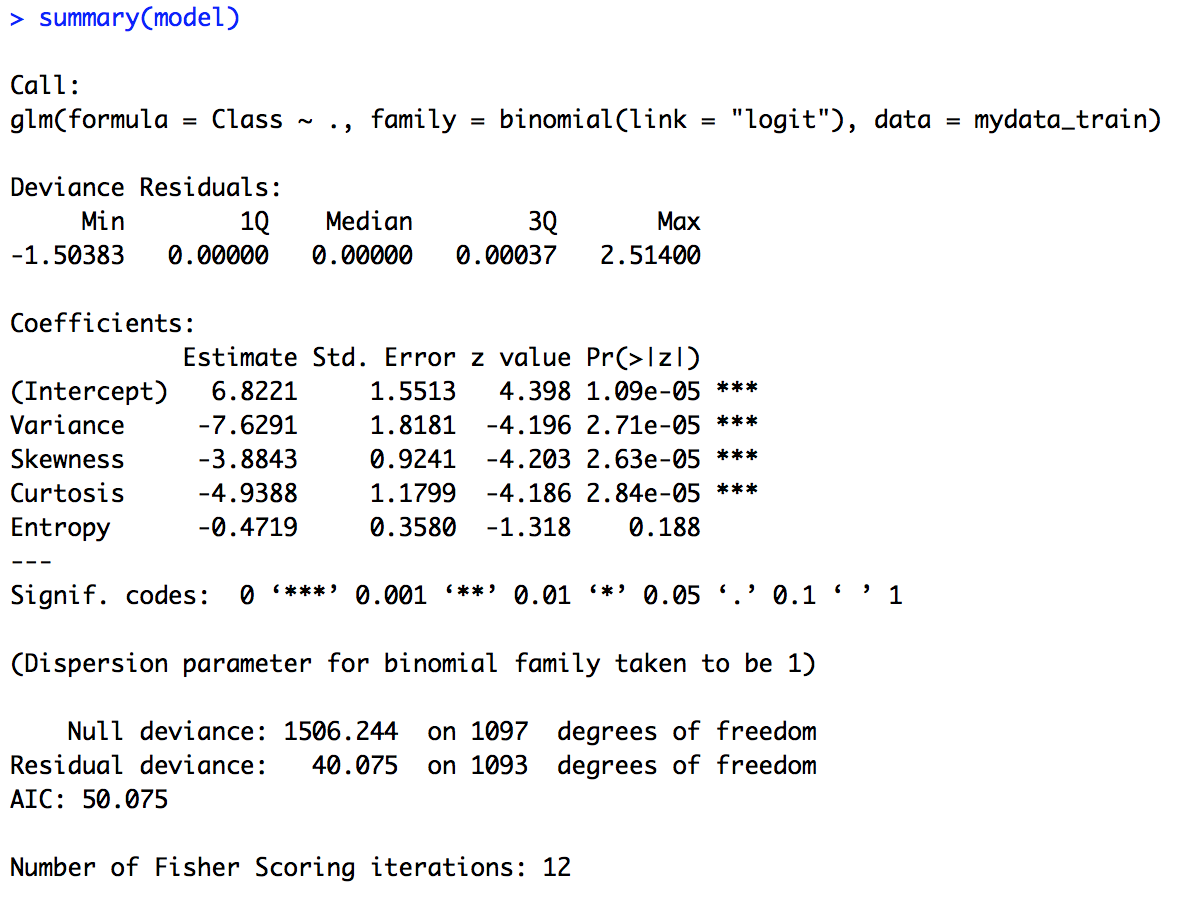


***Fig.6*** *Logistic regression model*

#The dataset is classified into 80 and 20%. 80% data has been used for training and 20% of data was used for testing.

#Since the datase has two possible outcomes, 0 or 1, logistic regression is appropriate.

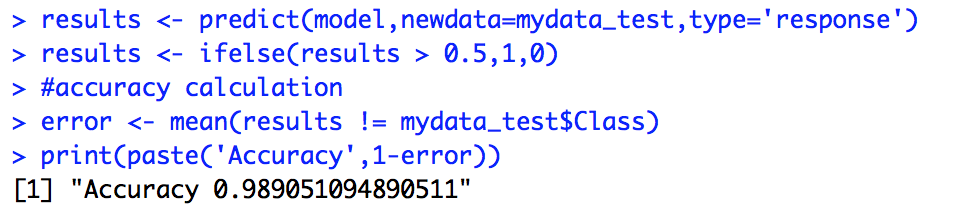
#To perform logistic regression, glm() function is used.



***Fig.7*** *Model summary*

#From the model, we could see that Entropy has higher p value >0.05.

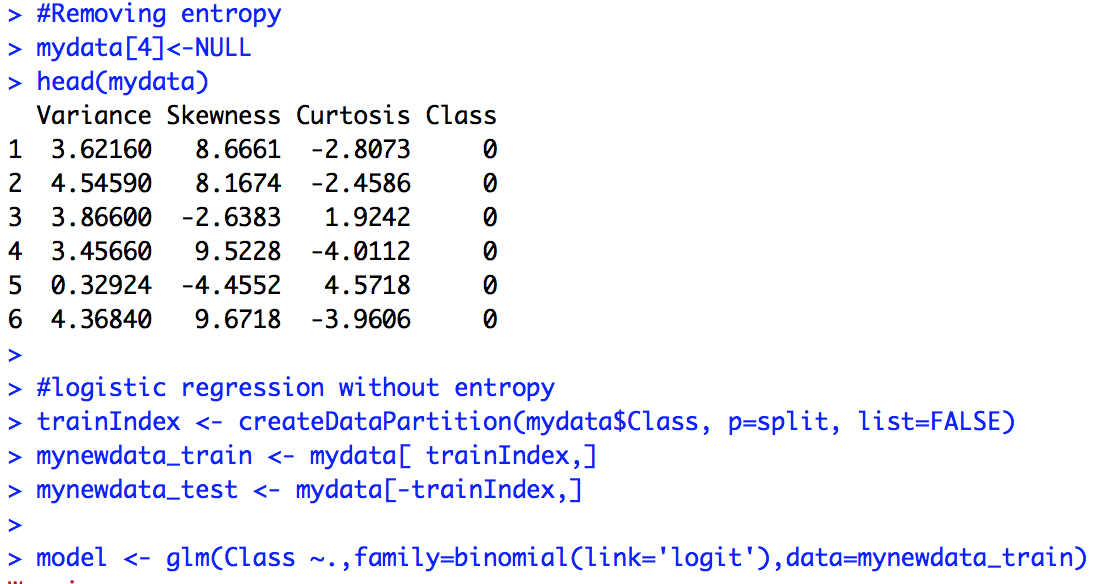
#It suggests that there is no strong association between entropy and class attribute.



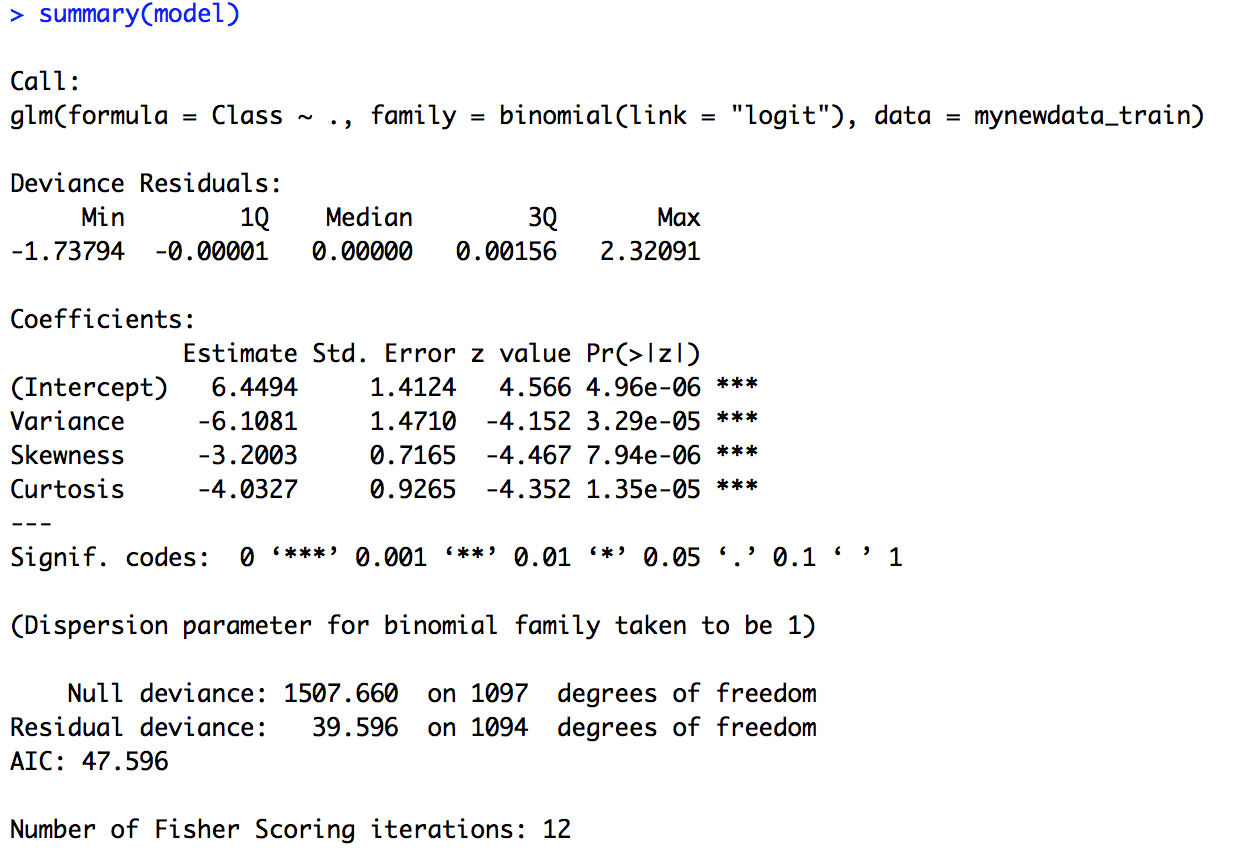
***Fig.8*** *Predictive ability of model*

#From the accuracy obtained, it could be seen that the model is good for predicting the authenticity of bank notes.

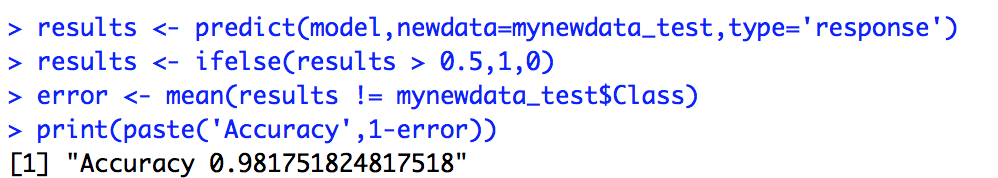
#Again, a model is constructed with Entropy removed and the predictive ability of that model will be assessed.



***Fig.9*** *New model with Entropy removed*

**

***Fig.10*** *Summary of new model*



***Fig.11*** *Predictive ability of new model*

#We see that there is not much difference in the accuracy obtained for the new model compared to the old model.

#Therefore, it could be concluded that, Variance, Skewness, Kurtosis together can effectively classify between genuine and forged notes.

Reference

<http://archive.ics.uci.edu/ml/datasets.html>

IHE-7510-01 - Data Mining- Lecture