COMP5310 Principles of Data Science

Assignment 1 - Project Stage 1

Student name: Ruijue Zou
Student ID: 500709979
Unikey: rzou3444

Problem

In the outer space, there is a special kind of neutron star called "Pulsar" that can rotate extremely fast. The radiation of the pulsar can be observed by large radio telescopes when its emission beam points to Earth ('Pulsar' 2020). Pulsars play an important role for astronomers to explore the universe, including interstellar medium and space-time investigation (UCI 2020). Researchers have detected large amount of radio signal through the years, however, most of them are noise or radio frequency interference. According to the pulsar's special signal emission pattern, researchers can identify whether the signal they observed is produced by the pulsar. Because of the large amount of noise, it is hard to find a real pulsar efficiently.

Therefore, the goal of this project is to classify candidates and identify which candidate has represented as a real pulsar on its given features automatically by using machine learning tools. This project will be focusing on the following aspects:

- Identify the relationship between each feature,
- Classify candidates by training machine learning models,
- Identify which feature is significant for classification.

Data

A dataset has been downloaded from the UCI website. The raw data was collected by the Parkes Observatory, and donated by Dr Robert Lyon from the University of Manchester in 2017. The original dataset, HTRU_2, is approximately 1.7 MB, contains 17,898 candidates, 8 attributes and corresponding class label. The dataset is presented in the CSV format. It also comes with a description file, which contains basic information of this dataset.

According to the description file, the 8 features in the dataset are statistics related to the integrated pulse profile and DM-SNR curve of each candidate, which can describe a candidate is a real pulsar or not. The astronomical information of the candidates is not included. Among of the 17,898 candidates, there are 1,639 candidates are positive and labelled as "1", which means actual pulsar, whereas 16,259 candidates are negative and labelled as "0", which are noise or radio frequency interference.

To explore and prepare the dataset that can be used for stage 2, the following preprocessing has been done:

- **Load dataset to Rstudio.** Exploration and preprocessing dataset will be implemented by using R code.
- **Add column names for each feature.** After loading the dataset into Rstudio, the dataset has no column names. Therefore, it is necessary to add corresponding column names to the dataset

based on the dataset description file. Since the attributes name is too long, it was decided to use a short name instead of the full name. The name mapping is provided in Appendix 1.

- **Check missing values.** There is no missing value in this dataset.
- **Data type transformation.** All values in the dataset were stored as numerical. Because of 0 and 1 in column "Class" should be treated as the category instead of numbers, this column has been changed to a factor.
- **Summarise statistical information.** The following table shows the basic statistical information of each attribute, including maximum value, minimum value, etc. It is clear that there is a huge difference of numerical between different attributes.

```
   Mean_IP            SD_IP           EK_IP              S_IP
Min.   :  5.812   Min.   :24.77   Min.   :-1.8760   Min.   :-1.7919
1st Qu.:100.930   1st Qu.:42.38   1st Qu.: 0.0271   1st Qu.:-0.1886
Median :115.078   Median :46.95   Median : 0.2232   Median : 0.1987
Mean   :111.080   Mean   :46.55   Mean   : 0.4779   Mean   : 1.7703
3rd Qu.:127.086   3rd Qu.:51.02   3rd Qu.: 0.4733   3rd Qu.: 0.9278
Max.   :192.617   Max.   :98.78   Max.   : 8.0695   Max.   :68.1016
    Mean_C             SD_C             EK_C             S_C
Min.   :  0.2132   Min.   :  7.37   Min.   :-3.139   Min.   :  -1.977
1st Qu.:  1.9231   1st Qu.: 14.44   1st Qu.: 5.782   1st Qu.:  34.961
Median :  2.8018   Median : 18.46   Median : 8.434   Median :  83.065
Mean   : 12.6144   Mean   : 26.33   Mean   : 8.304   Mean   : 104.858
3rd Qu.:  5.4643   3rd Qu.: 28.43   3rd Qu.:10.703   3rd Qu.: 139.309
Max.   :223.3921   Max.   :110.64   Max.   :34.540   Max.   :1191.001
```

- **Compute the correlation between attributes.** For better understand the distribution and the relationship of each feature, several related plots have been generated, provided in Appendix 2. From these plots, it is clear that the attribute "EK_IP" has the most significant correlation with "Class". Also, it can be seen that the dataset is imbalanced, and for some features, the data points are quite separable.
- **Check outlier.** Since the dataset is imbalanced, there are lots of outliers in class 0. An example has been provided in the Appendix 6. It is supposed that removing outliers will reduce a huge amount of data points, also might have an impact on the prediction model. So the outliers will not be removed at this stage.
- **Apply normalisation.** In stage 2, some machine learning tools, such as KNN and SVM, would like to be implemented. And they require to compute the distance. A normalised dataset can help to make the prediction faster and improve accuracy. Therefore, normalisation has been applied to the 8 attributes. After normalisation, the range of values of each attribute is from 0 to 1.

Approach

In stage 2, to solve the project problems, the following steps will be implemented:

- Split candidates into two subsets by 80% and 20% for training and testing.
- Training classification models to classify candidates. Several classification tools will be applied, such as KNN, Random Forest and Decision Tree, etc.
- Apply cross-validation to choose the best classification model with the best performance.
- Try to find which attribute is the most important on classification.

References

1. UCI 2020, *HTRU2 Data Set,* UCI, viewed 27 Sep 2020, https://archive.ics.uci.edu/ml/datasets/HTRU2
2. UCI 2020, (HTRU2 dataset), UCI, https://archive.ics.uci.edu/ml/machine-learning-databases/00372/HTRU2.zip
3. 'Pulsar', *Wikipedia,* viewed 29 Sep 2020, https://en.wikipedia.org/wiki/Pulsar

Appendices

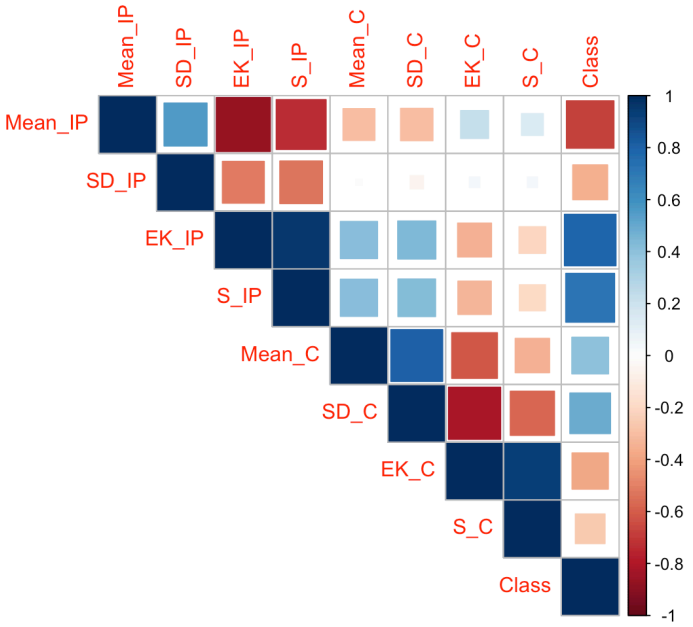
1. Column name mapping table
2. Correlation matrix and plots
3. Histograms of each attribute distribution
4. Bar chart of the "Class" distribution
5. Box plots of each attribute
6. An example of outliers
7. Statistical information of each attributes after normalisation

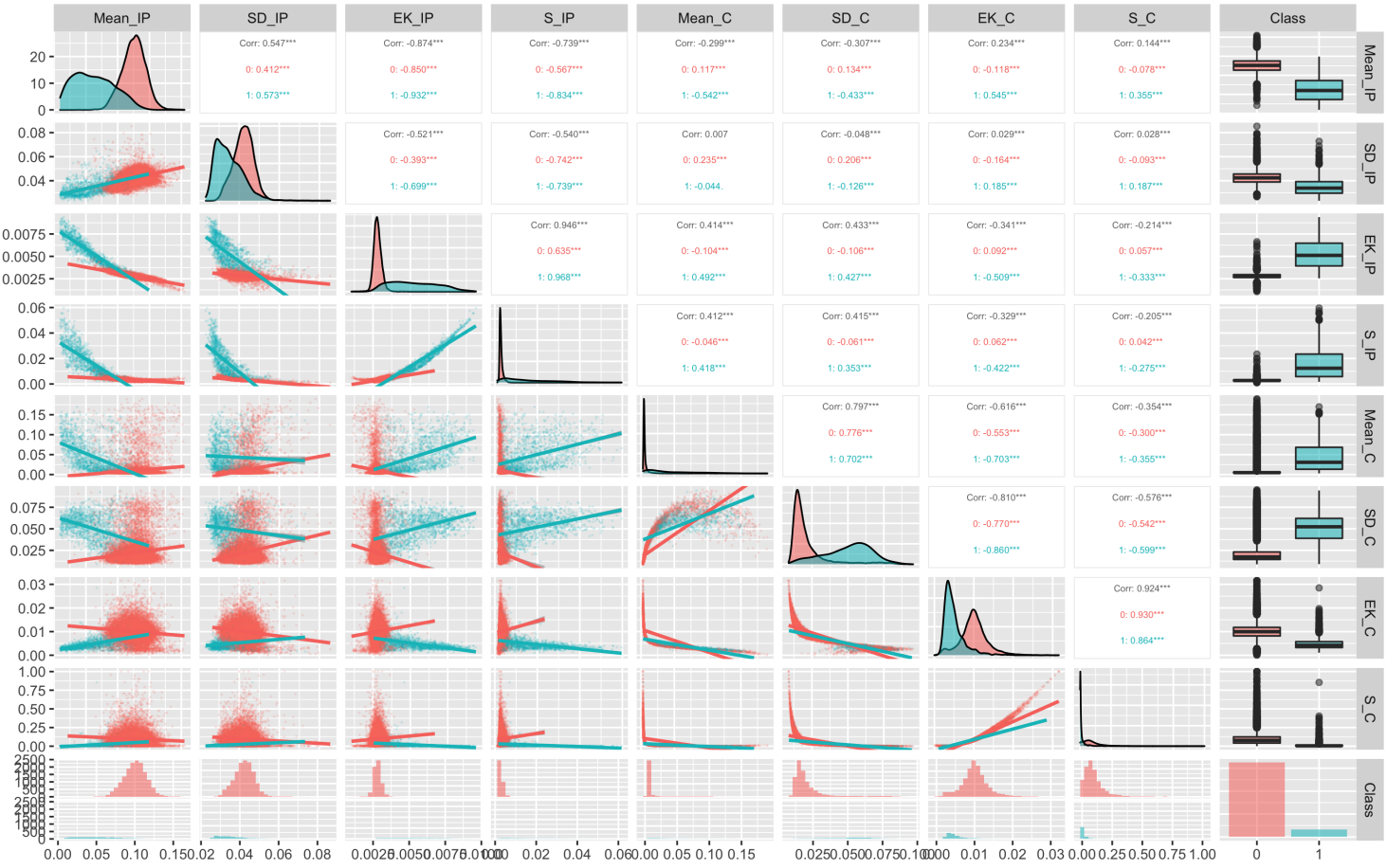
Appendix 1 - Column name mapping table

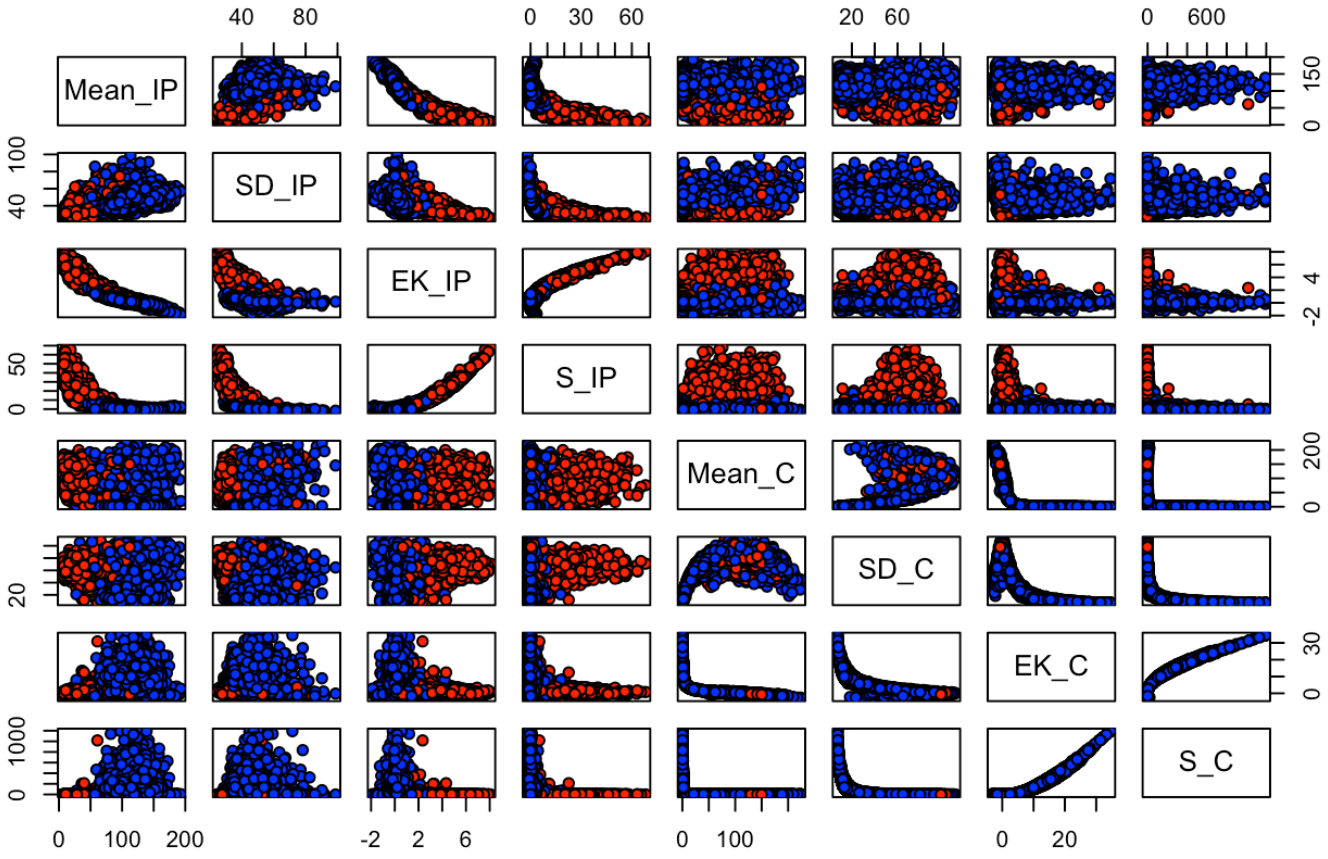
Column Name (Short)	Column Name (Full)
Mean_IP	Mean of the integrated profile
SD_IP	Standard deviation of the integrated profile
EK_IP	Excess kurtosis of the integrated profile
S_IP	Skewness of the integrated profile
Mean_C	Mean of the DM-SNR curve
SD_C	Standard deviation of the DM-SNR curve
EK_C	Excess kurtosis of the DM-SNR curve
S_C	Skewness of the DM-SNR curve

Appendix 2 - Correlation matrix table and plots

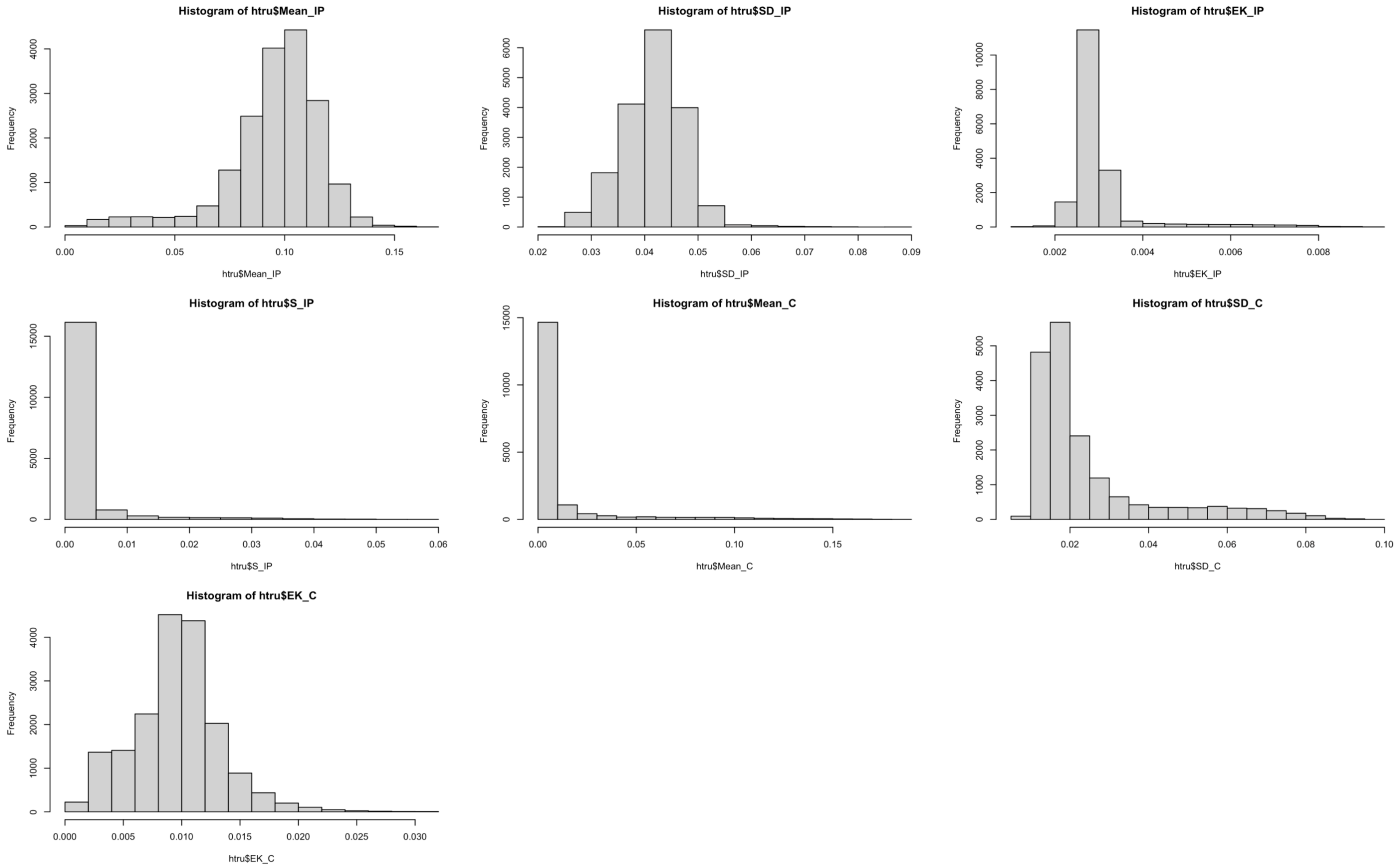
	Mean_IP	SD_IP	EK_IP	S_IP	Mean_C	SD_C	EK_C	S_C	Class
Mean_IP	1.00	0.55	-0.87	-0.74	-0.30	-0.31	0.23	0.14	-0.67
SD_IP	0.55	1.00	-0.52	-0.54	0.01	-0.05	0.03	0.03	-0.36
EK_IP	-0.87	-0.52	1.00	0.95	0.41	0.43	-0.34	-0.21	0.79
S_IP	-0.74	-0.54	0.95	1.00	0.41	0.42	-0.33	-0.20	0.71
Mean_C	-0.30	0.01	0.41	0.41	1.00	0.80	-0.62	-0.35	0.40
SD_C	-0.31	-0.05	0.43	0.42	0.80	1.00	-0.81	-0.58	0.49
EK_C	0.23	0.03	-0.34	-0.33	-0.62	-0.81	1.00	0.92	-0.39
S_C	0.14	0.03	-0.21	-0.20	-0.35	-0.58	0.92	1.00	-0.26
Class	-0.67	-0.36	0.79	0.71	0.40	0.49	-0.39	-0.26	1.00




Mean_IP
SD_IP
EK_IP
S_IP
Mean_C
SD_C
EK_C
S_C
Class

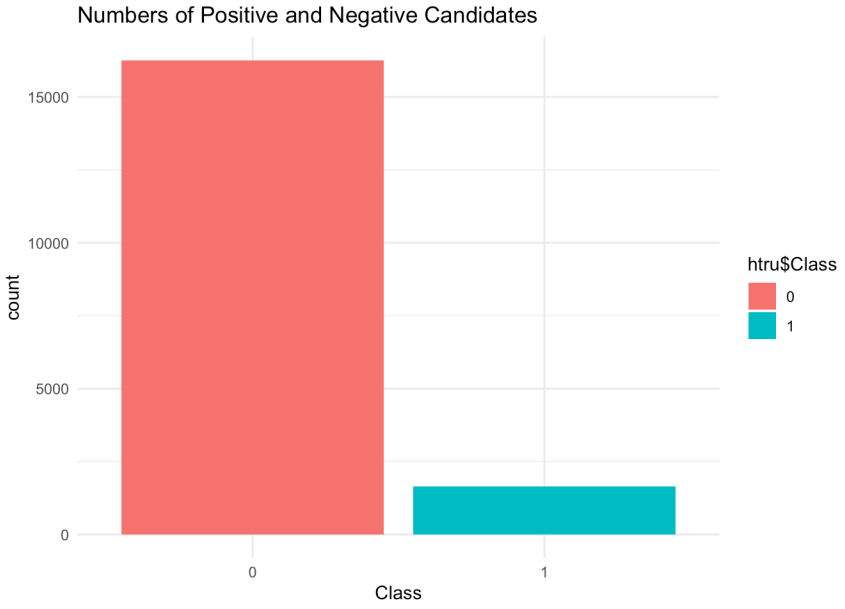

Mean_IP
SD_IP
EK_IP
S_IP
Mean_C
SD_C
EK_C
S_C

Appendix 3 - Histograms of each attribute distribution

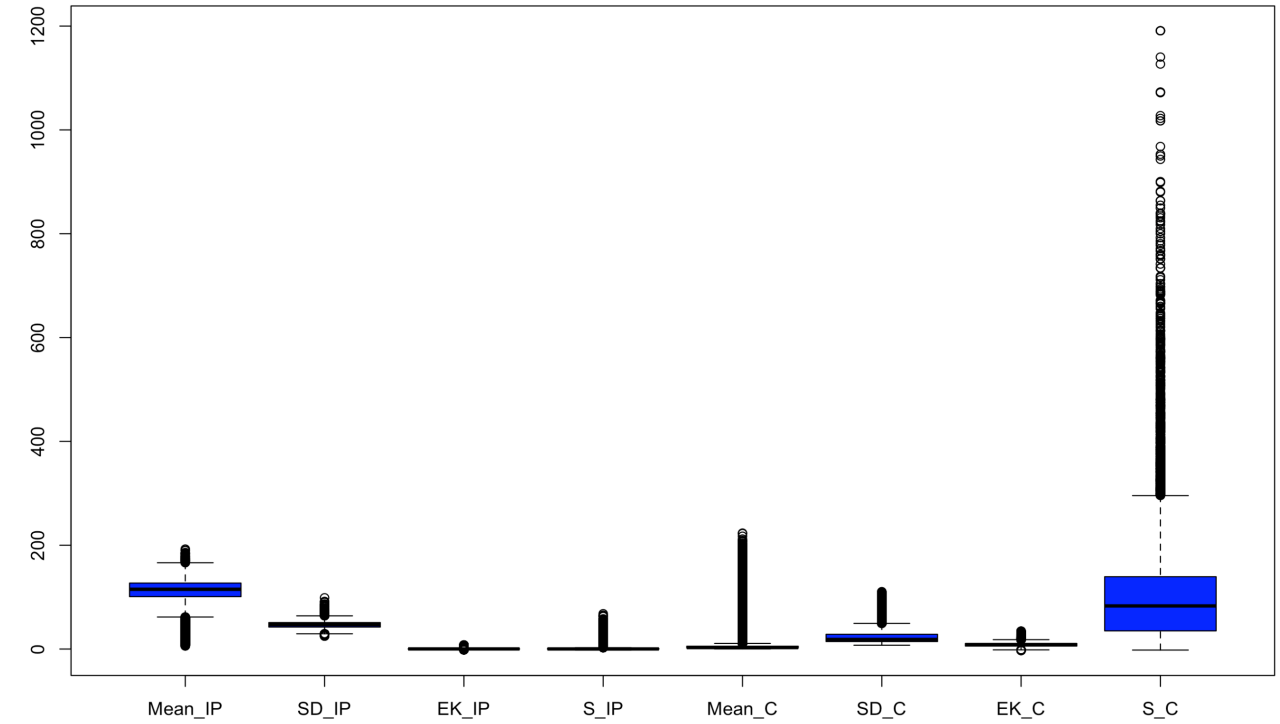


Appendix 4 - Bar chart of the "Class" distribution

It is clear that the dataset is imbalanced. There are only 9% of candidates are labelled as positive.

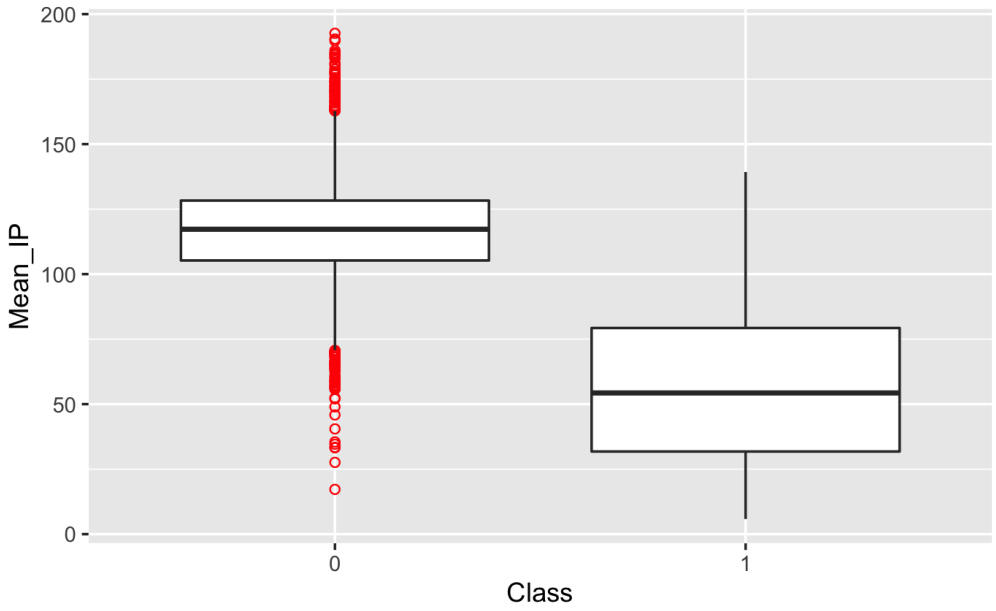


Appendix 5 - Box plots of each attribute



Appendix 6 - An example of outliers

In the following box plot, those red points are identified as outliers.



Appendix 7 - Statistical information of each attribute after normalisation

```
   Mean_IP            SD_IP             EK_IP              S_IP
Min.   :0.007496   Min.   :0.02337   Min.   :0.001058   Min.   :0.001128
1st Qu.:0.087150   1st Qu.:0.03812   1st Qu.:0.002652   1st Qu.:0.002471
Median :0.098998   Median :0.04194   Median :0.002816   Median :0.002795
Mean   :0.095650   Mean   :0.04161   Mean   :0.003029   Mean   :0.004111
3rd Qu.:0.109054   3rd Qu.:0.04536   3rd Qu.:0.003025   3rd Qu.:0.003406
Max.   :0.163931   Max.   :0.08535   Max.   :0.009386   Max.   :0.059659
    Mean_C             SD_C              EK_C               S_C
Min.   :0.002807   Min.   :0.008801   Min.   :0.000000   Min.   :0.0009733
1st Qu.:0.004239   1st Qu.:0.014719   1st Qu.:0.007470   1st Qu.:0.0319056
Median :0.004975   Median :0.018089   Median :0.009691   Median :0.0721890
Mean   :0.013192   Mean   :0.024675   Mean   :0.009582   Mean   :0.0904391
3rd Qu.:0.007205   3rd Qu.:0.026435   3rd Qu.:0.011592   3rd Qu.:0.1192897
Max.   :0.189703   Max.   :0.095283   Max.   :0.031553   Max.   :1.0000000
Class
0:16259
1: 1639
```