**Exploratory Data Analysis of Apple iOS App Store**

1. **Problem definition**

With the great number of apps available in the Apple app store, developers of new apps need to know some statistical information about apps in the app store such as which app category is preferred by users, which apps have highest rating., and what are the most paid apps. This valuable information is of great importance to developers to get insights about the factors contributing to a successful app.

1. **Data description**

Apple iOS app store dataset [1] will be used in this project. The dataset consists of statistics of about 7200 apps. the dataset consists of 16 attributes such as app ID, name, user rating, price, genre, and size. The dataset was collected from Apple iOS app store in July 2017.

The dataset consists of the following attributes:

Table 1 Attribute description

|  |  |
| --- | --- |
| Attribute | Description |
| ID | App ID |
| track\_name | App Name |
| size\_bytes | App size in bytes |
| currency | Currency type |
| price | Price amount |
| rating\_count\_tot | User Rating counts (for all version) |
| rating\_count\_ver | User Rating counts (for current version) |
| user\_rating | Average User Rating value (for all version) |
| user\_rating\_ver | Average User Rating value (for current version) |
| ver | Latest version code |
| cont\_rating | Content rating |
| prime\_genre | Primary Genre |
| sup\_devices.num | Number of supported devices |
| lang.num | Number of supported languages |
| ipadSc\_urls.num | Number of screenshots showed for display |
| Vpp\_lic | Vpp Device Based Licensing Enabled |

1. **Data preprocessing**

* Loading data into Weka

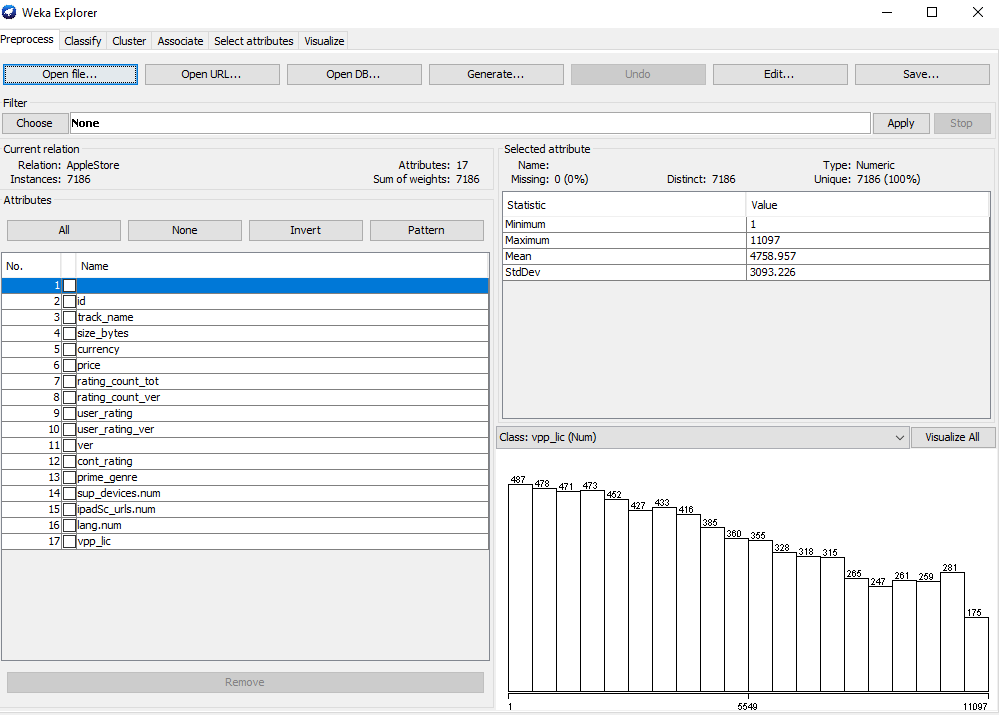
****

Figure 1 Loading data

* Removing irrelevant attributes such as id, app name, currency, … etc.
* Transforming size attribute from bytes into Megabyte

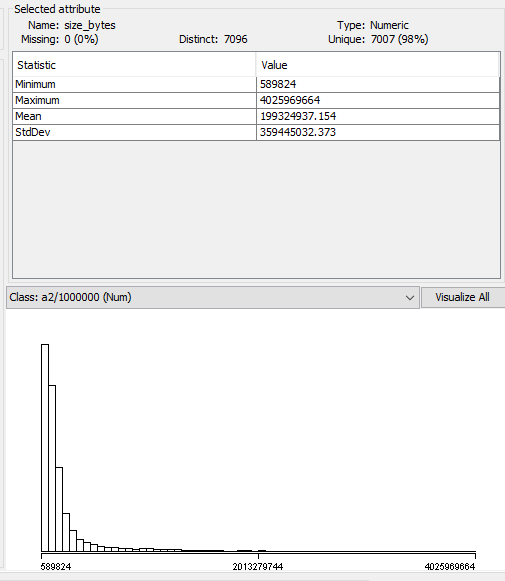
****

Figure 2 Size (Bytes)

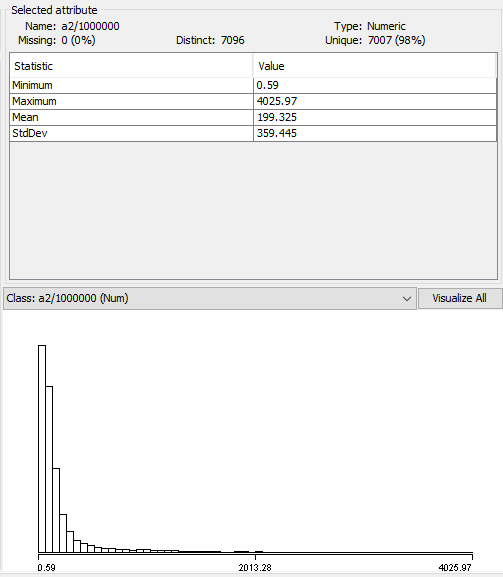


Figure 3 Size (Mega bytes)

**Exploring attributes**

* Size attribute

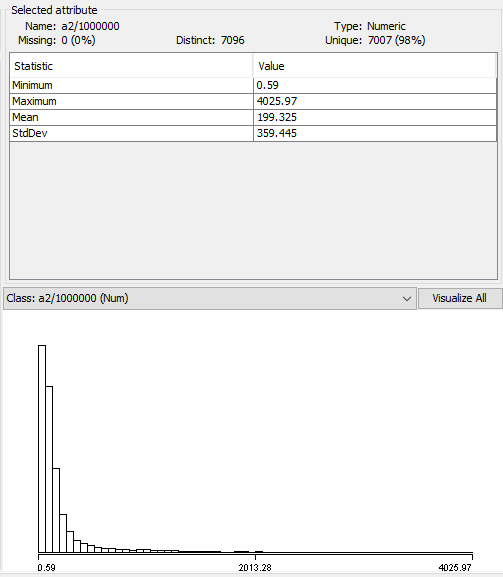


Figure 4 Size attribute

* Price

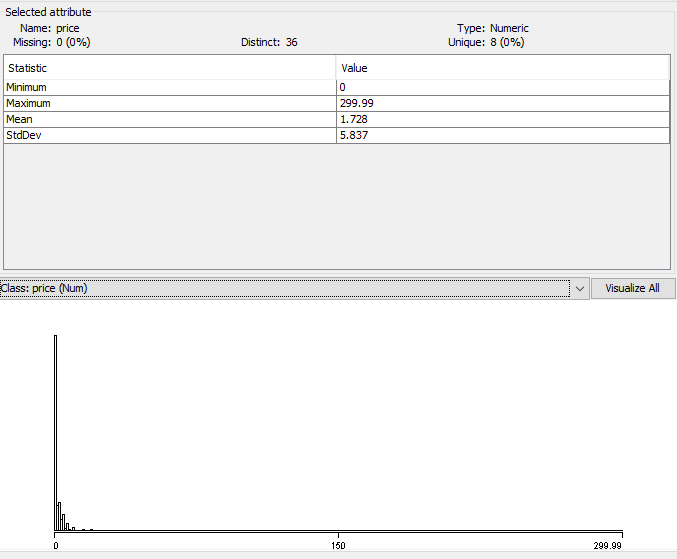


Figure 5 Price attribute

* User rating

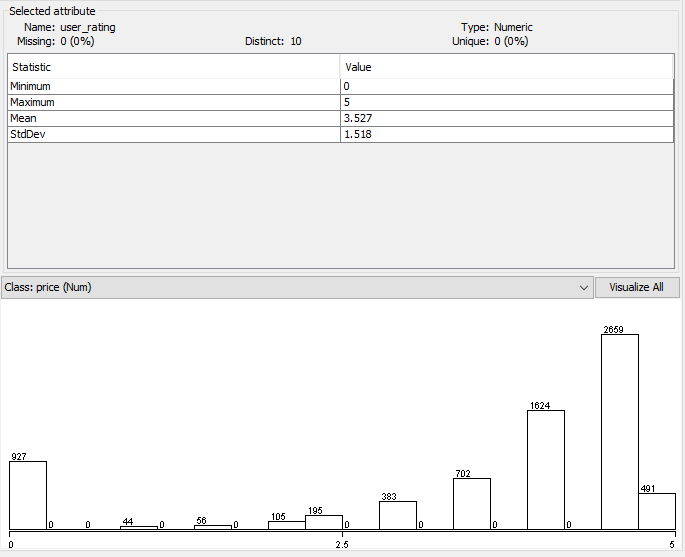


Figure 6 User rating

User rating for the current version

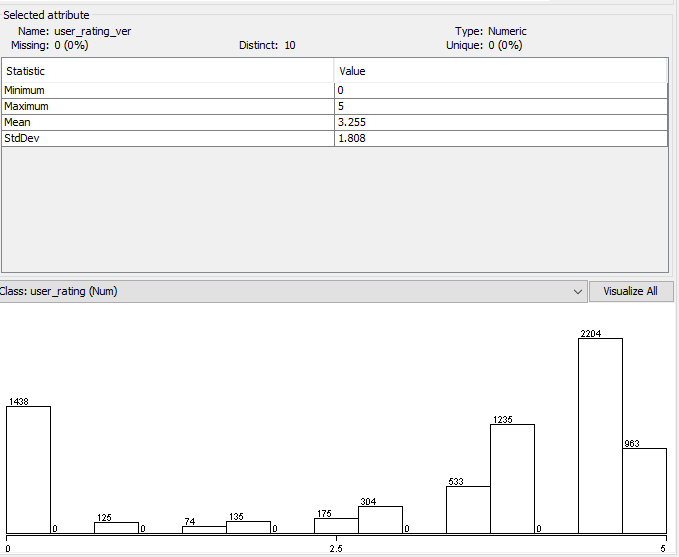


Figure 7 user rating (current version)

* Content rating

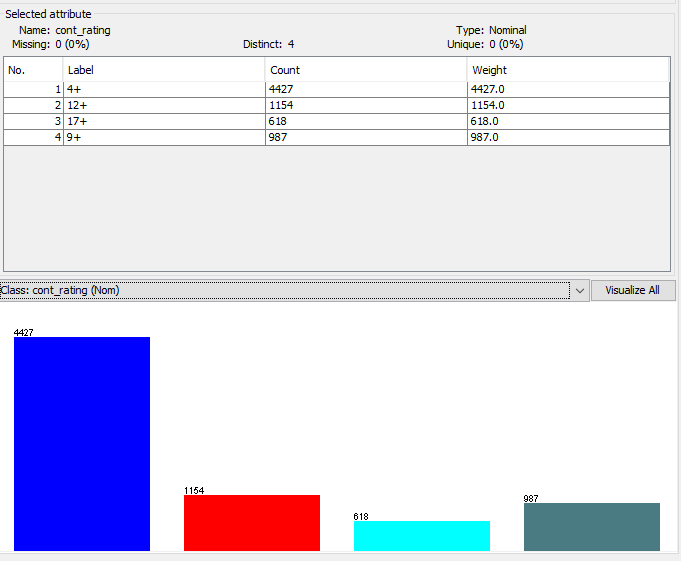


Figure 8 Content rating

* App genre

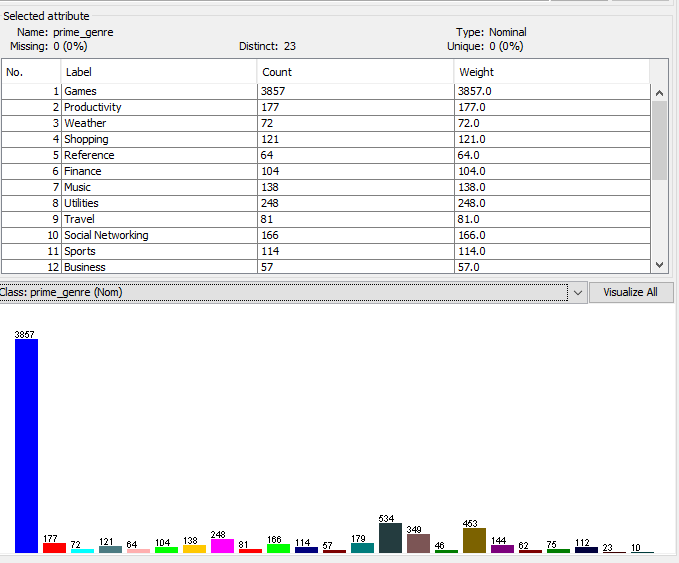


Figure 9 App genre

* Number of supported devices

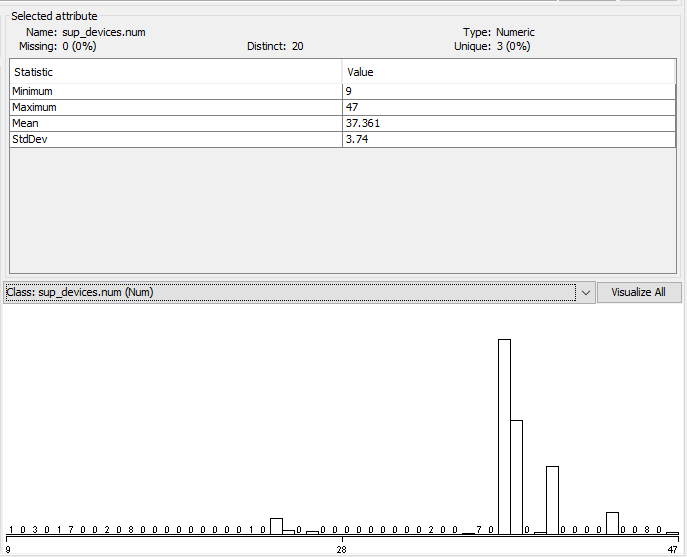


Figure 10 Number of supported devices

* Number of supported languages

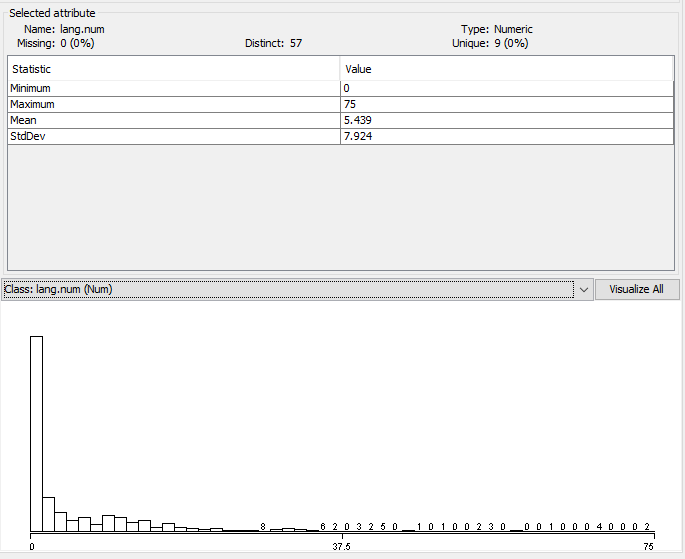


Figure 11 Number of supported languages

1. **Data mining algorithms**

We will use classification algorithms to predict user ratings from other attributes.

* First, user rating attribute is discretized into five levels using Discretize filter in Weka.

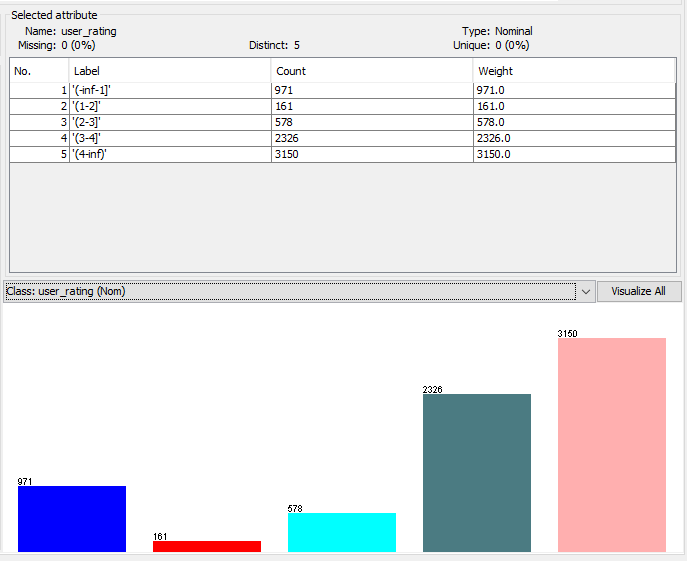
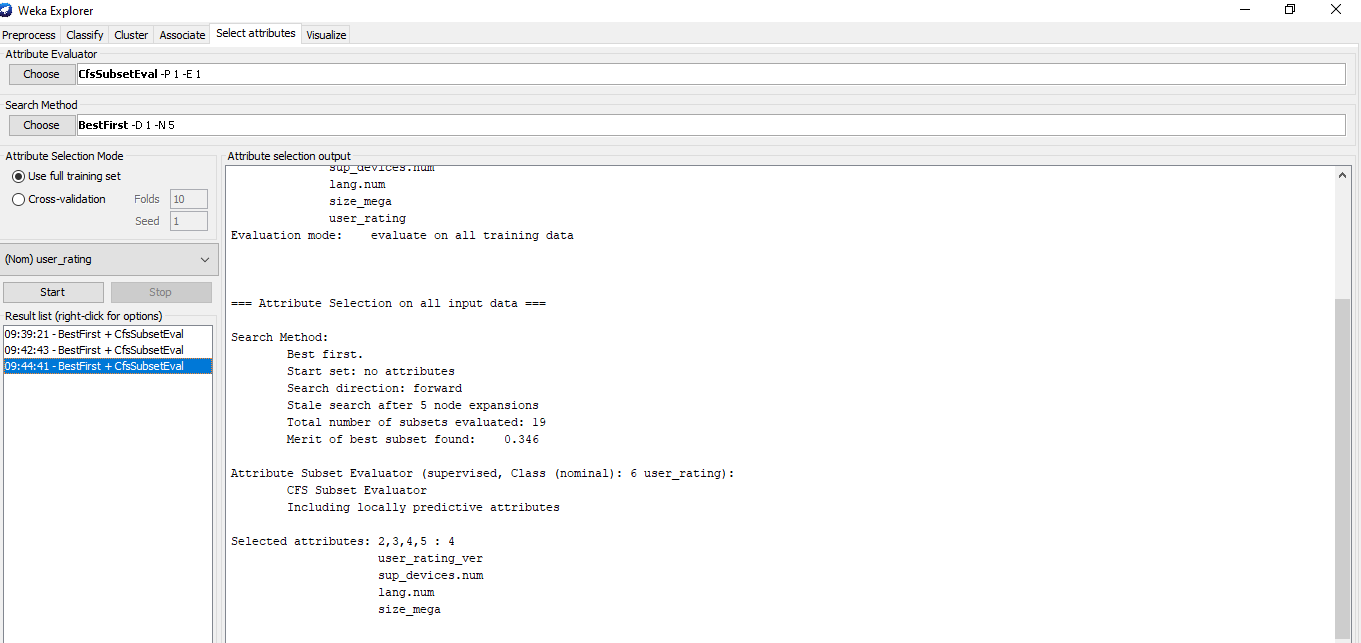


Figure 12 Discretizing user rating attribute

* Second, important attributes are selected.

****

* Then, classification algorithms are applied.

1. Naïve Bayes Classifier

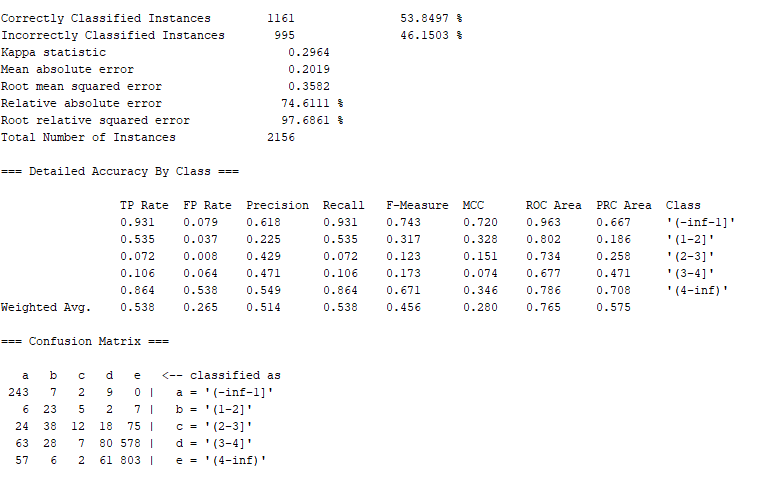


Figure 13 Naive Bayes classifier

1. Logistic Regression

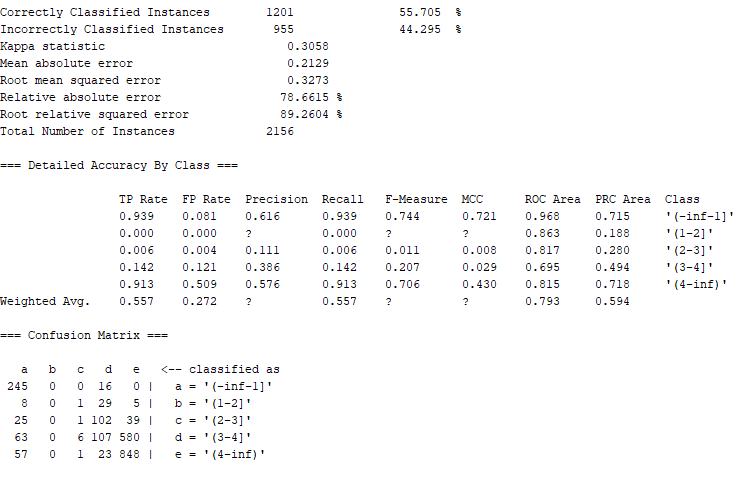


Figure 14 Logistic regression

1. Multilayer Perceptron (MLP)

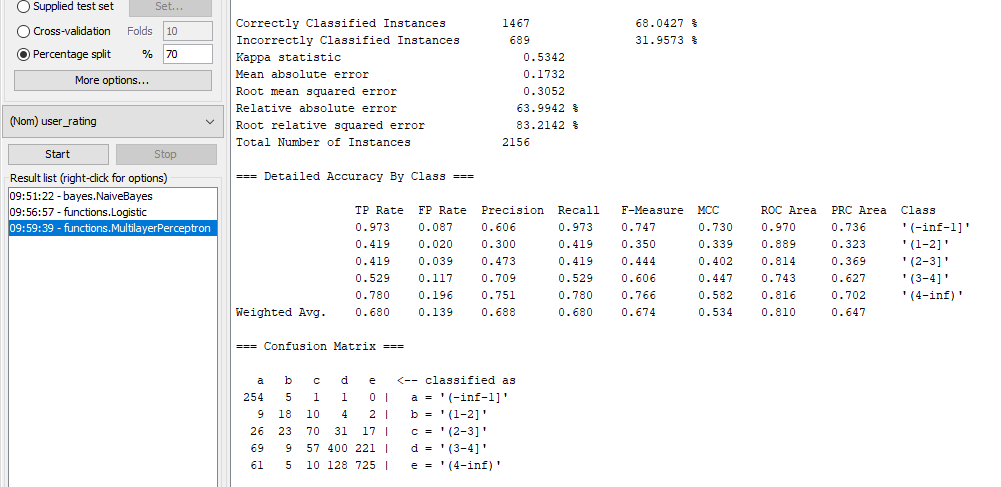


Figure 15 Multilayer Perceptron

1. Random Forest

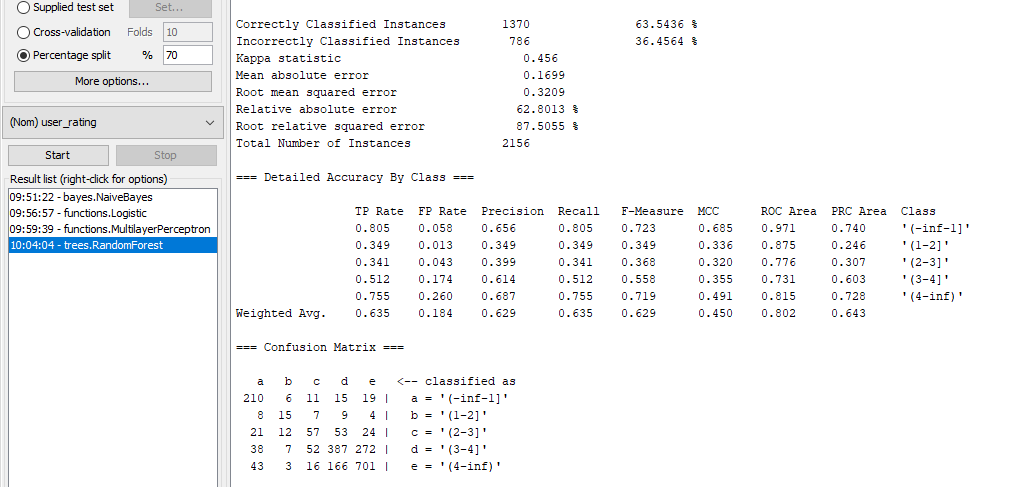


Figure 16 Random Forests

1. **Conclusions**

* The number of free apps is greater than the number of paid apps.
* Most Apps are priced at less than $10.
* The most popular App types are Games and Entertainment
* Most Apps support between 25 and 35 devices.
* Most Apps support only one language which is English language.
* Most Apps' size is under 200MB.
* App rating can be predicted from other attributes such as number of supported devices and number of supported languages and app size with a maximum accuracy of 68%.
* Multilayer Perceptron (MLP) achieves highest accuracy of 68% then Random Forest of 63.54%, Logistic Regression (55.7%), Naïve Bayes (53.9%).

References

<https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps>