Social Networking Ads

This dataset is obtained by Kaggle. It contains the data about profiles of the users on the social network who on interacting with the advertisement either purchased the product or not. Goal is classify the customers based on their age, estimated salary and whether they made a purchase or not.

* Supervised or Unsupervised?
  + Supervised
* Classification or Regression?
  + Classification
* Prediction or Inference?
  + Prediction

About Data Features-

|  |  |
| --- | --- |
| **Data Columns** | **Type** |
| User ID | Quantitative |
| Gender | Qualitative |
| Age | Quantitative |
| Estimated Salary | Quantitative |
| Purchased | Quantitative  (0-Not Purchased 1- Purchased) |

Training Data- 300 Observation

Testing Data- 100 Observation

Predictors or Features considered- Age, Estimated Salary

Response or Prediction- Purchased (0,1)

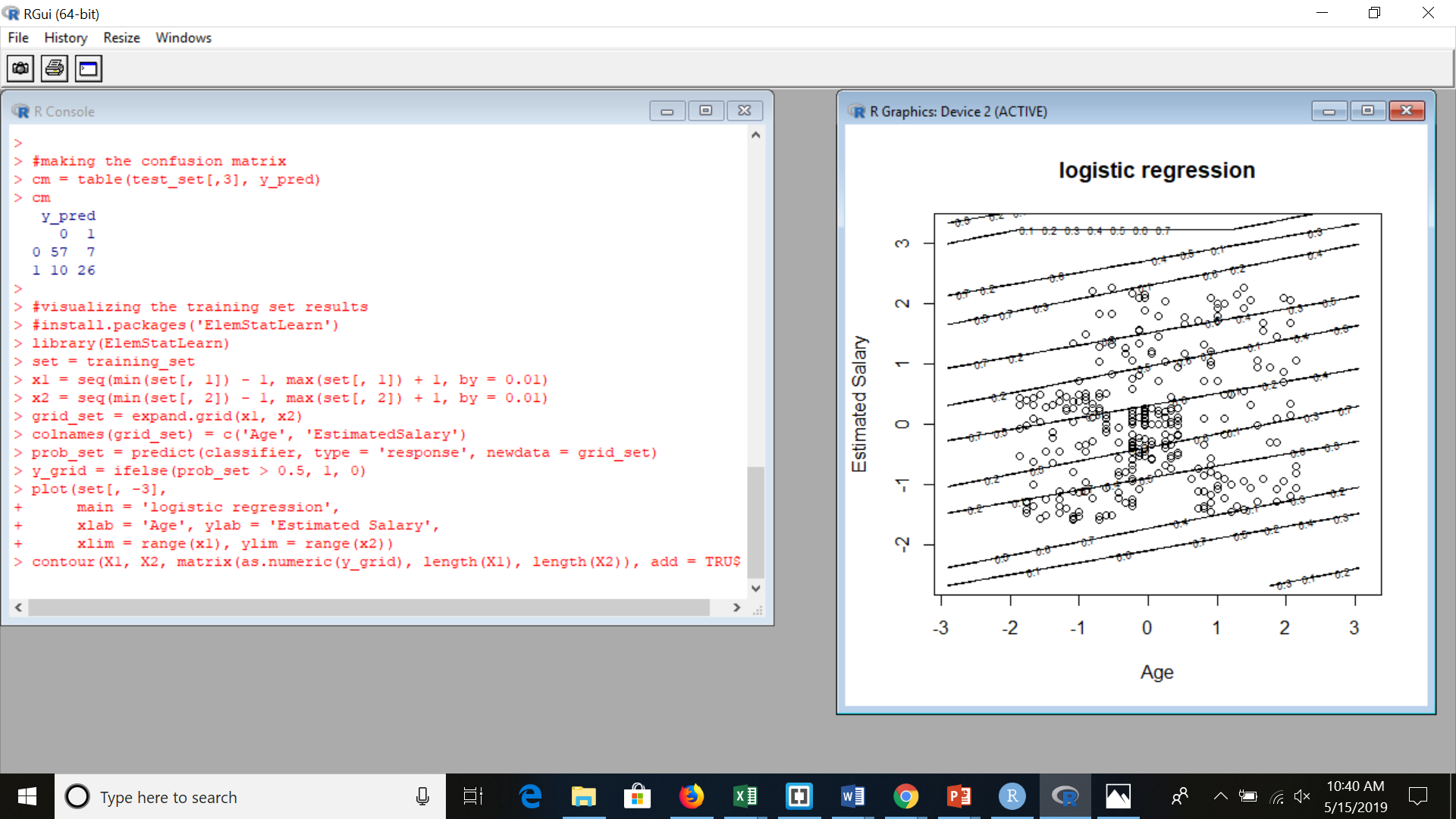
Approaches used for Prediction-

* Logistic Regression
* Classification Trees
* Support Vector Machines
* Random Forest
* Naïve Bayes

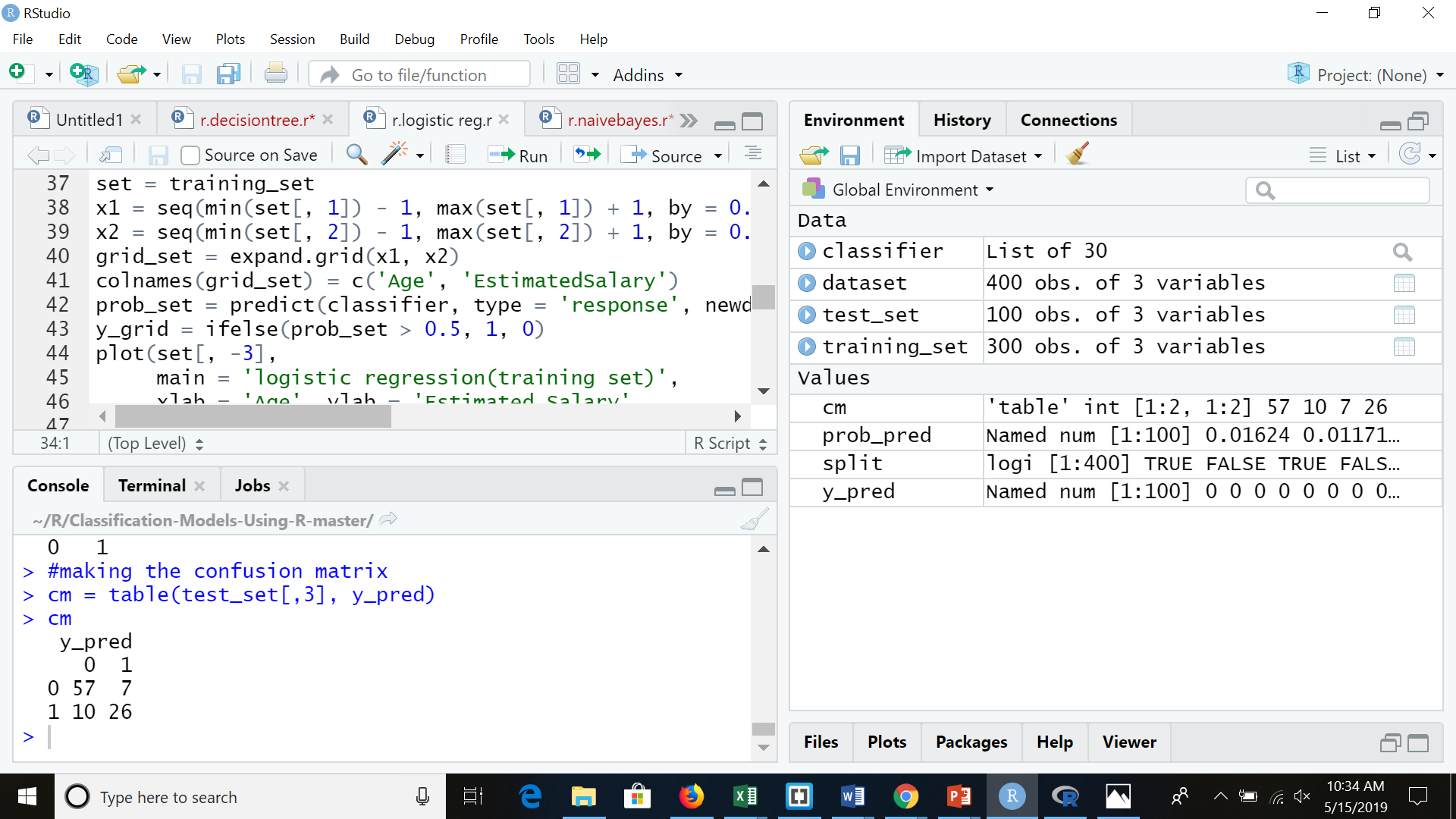
Going through all the approaches-

Logistic Regression-

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

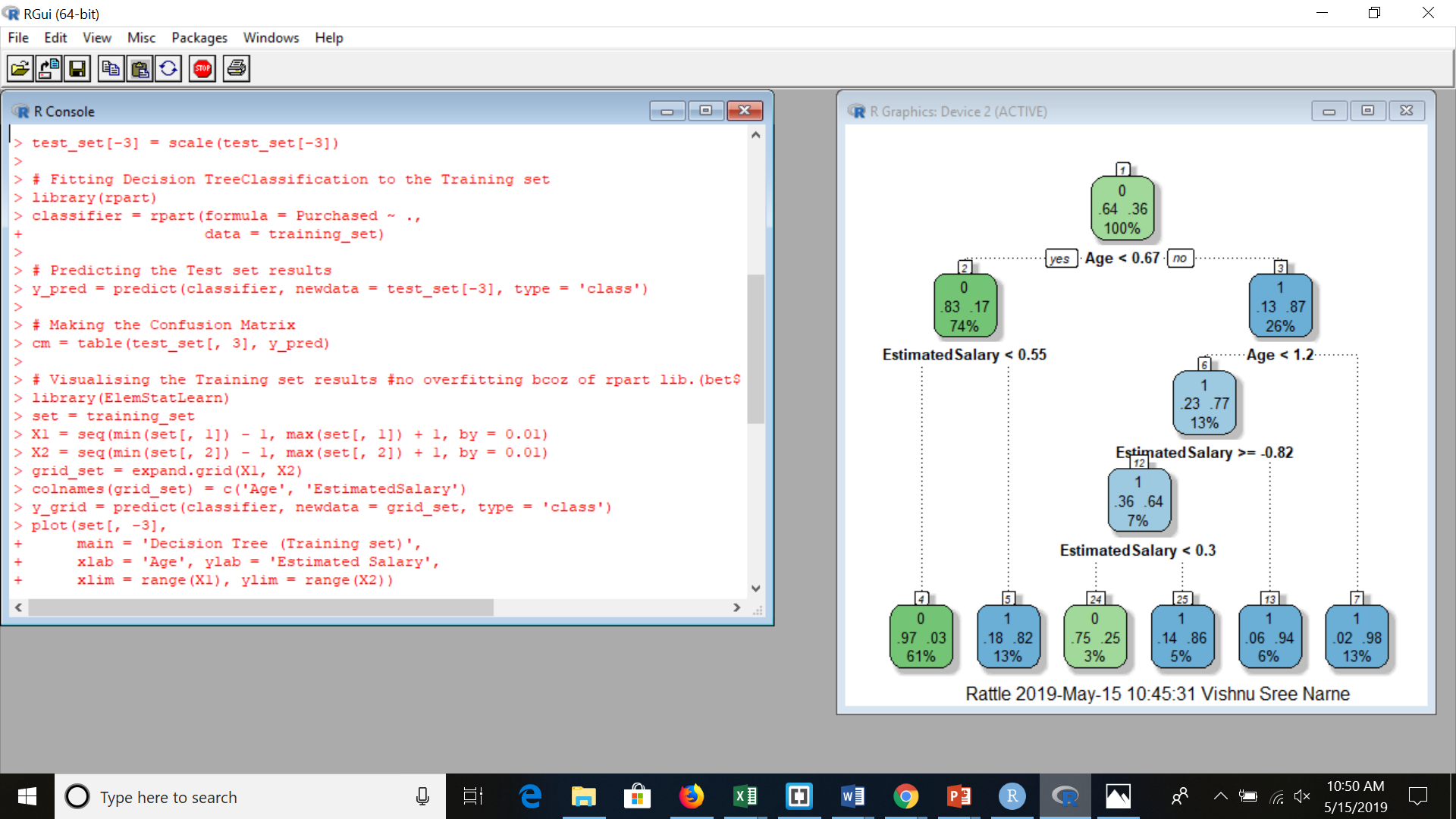


Confusion Matrix-



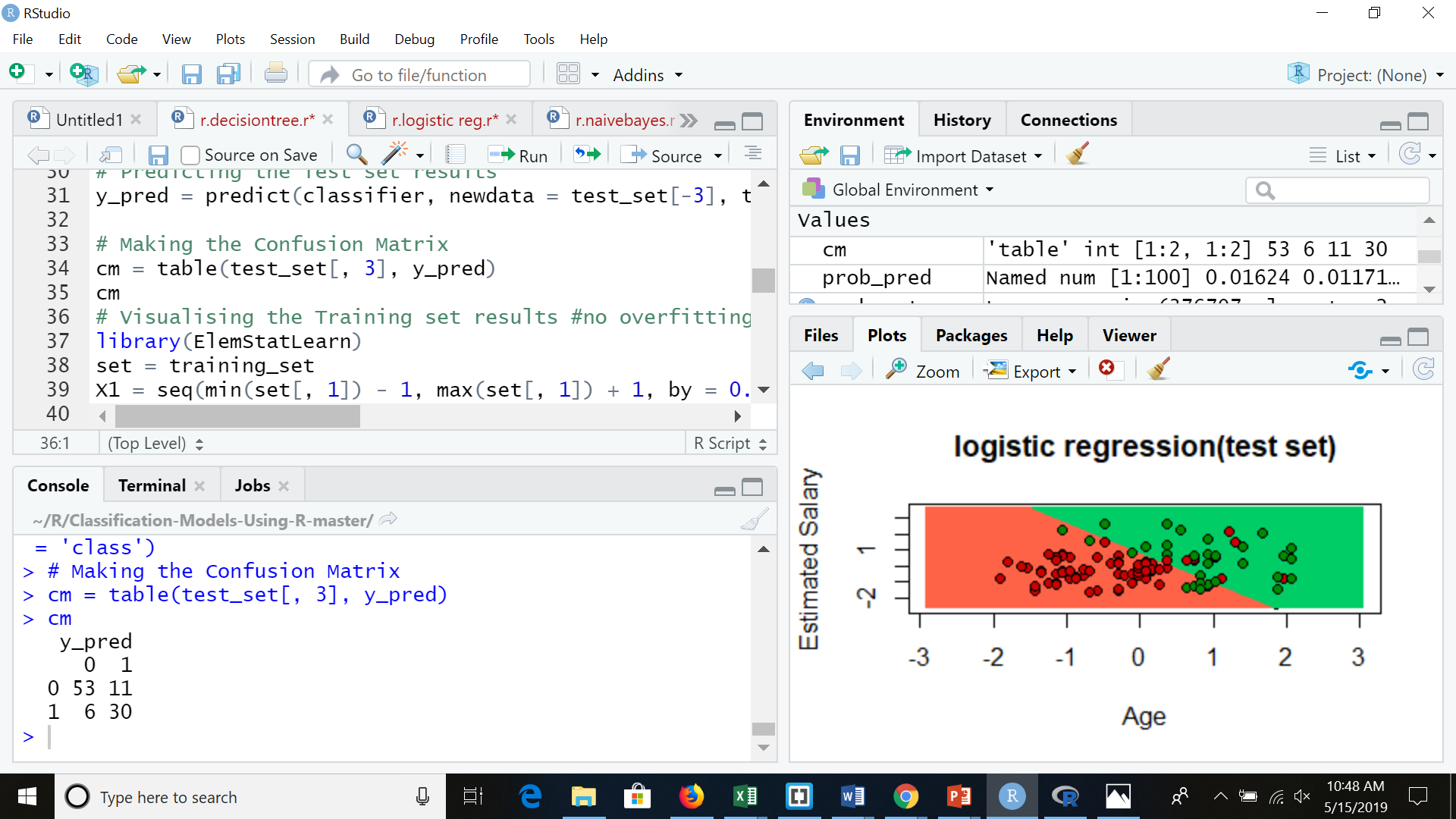
Accuracy- 0.83

Classification Decision Trees-



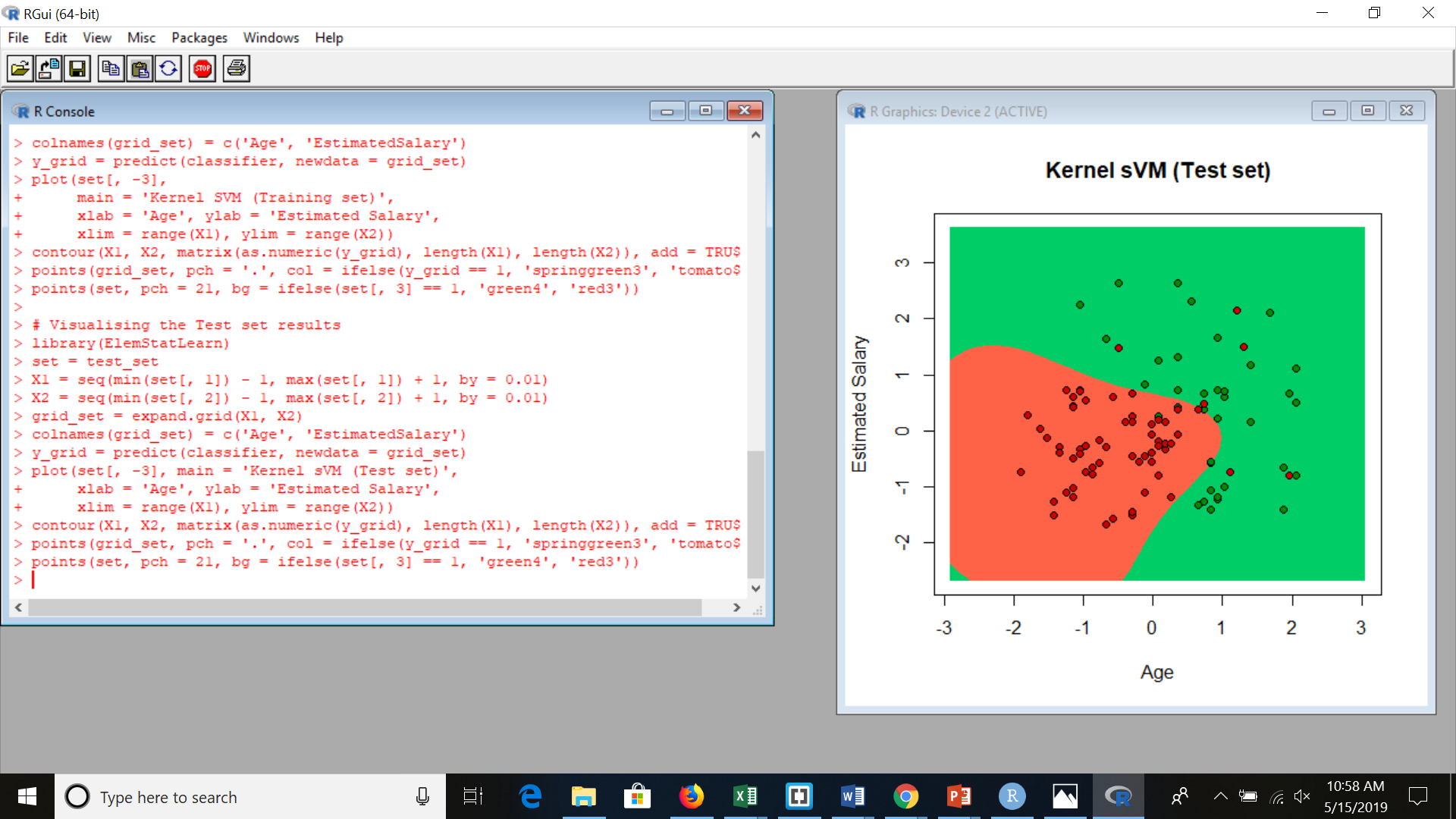
* Used rpart function to construct tree
* Visualized FanancyRpartPlot
* Tried changing weights accuracy was bad

Confusion Matrix-

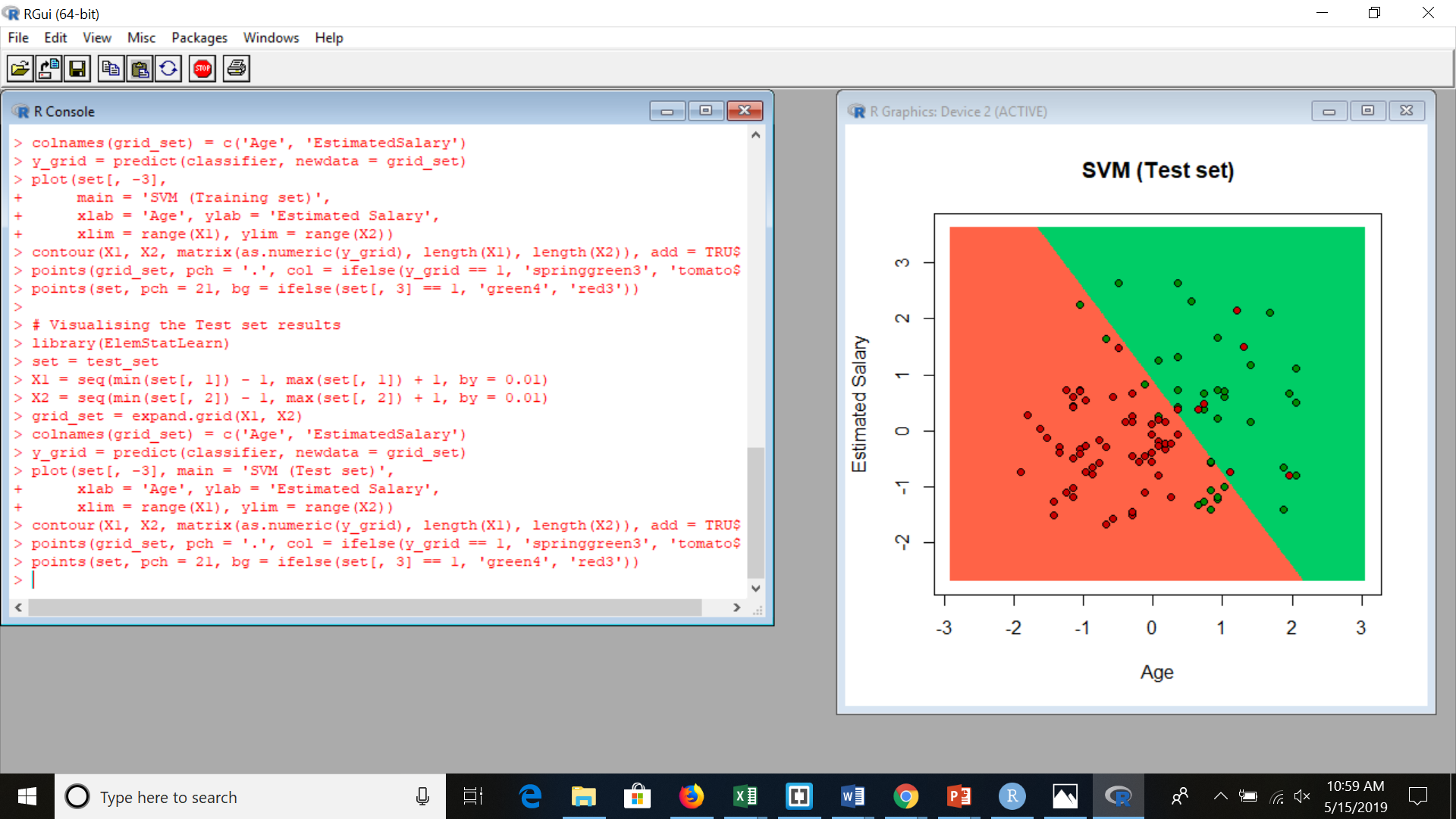


Accuracy -0.83

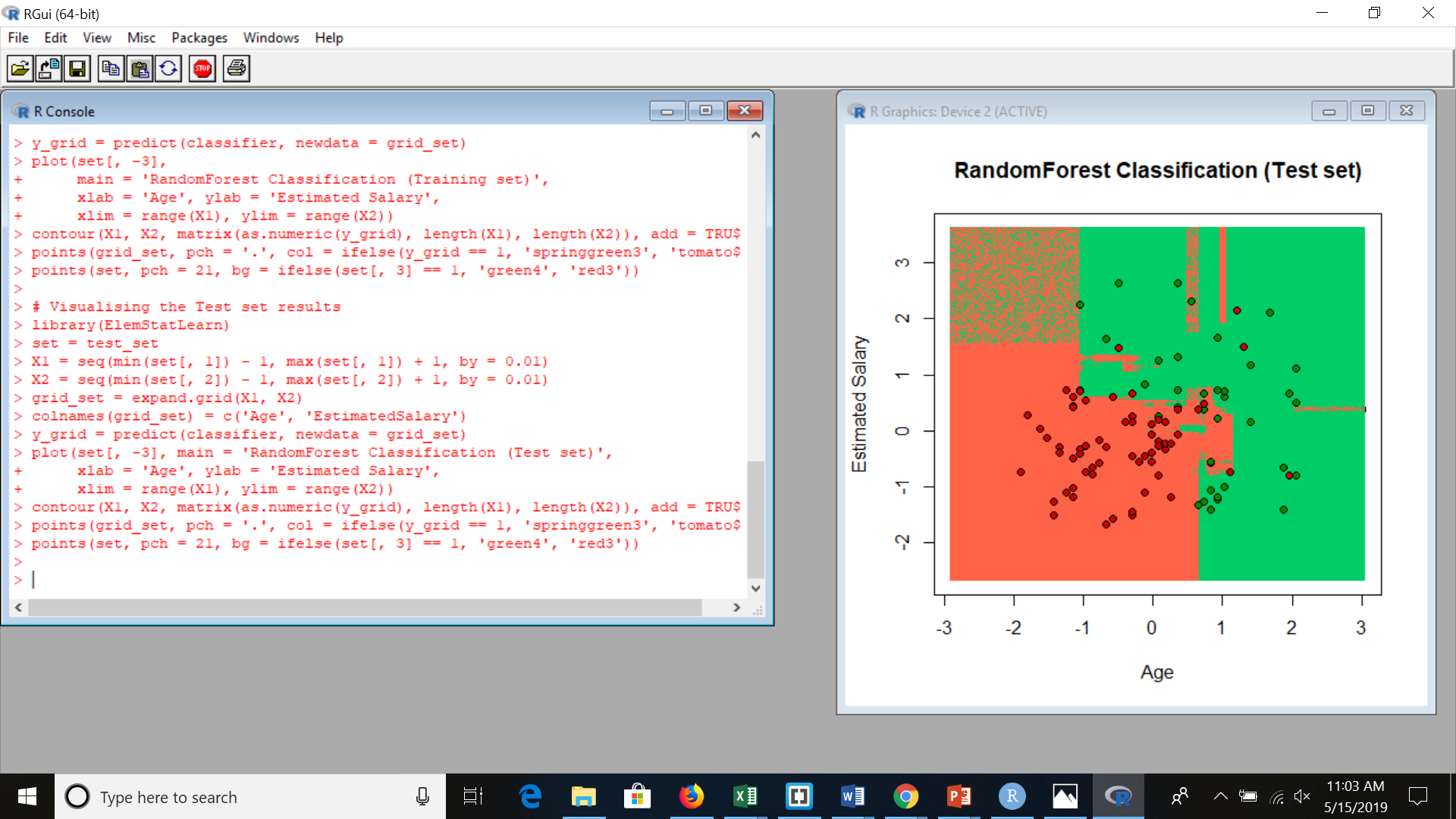
Kernel Support Vector Machine- Polynomial Cur



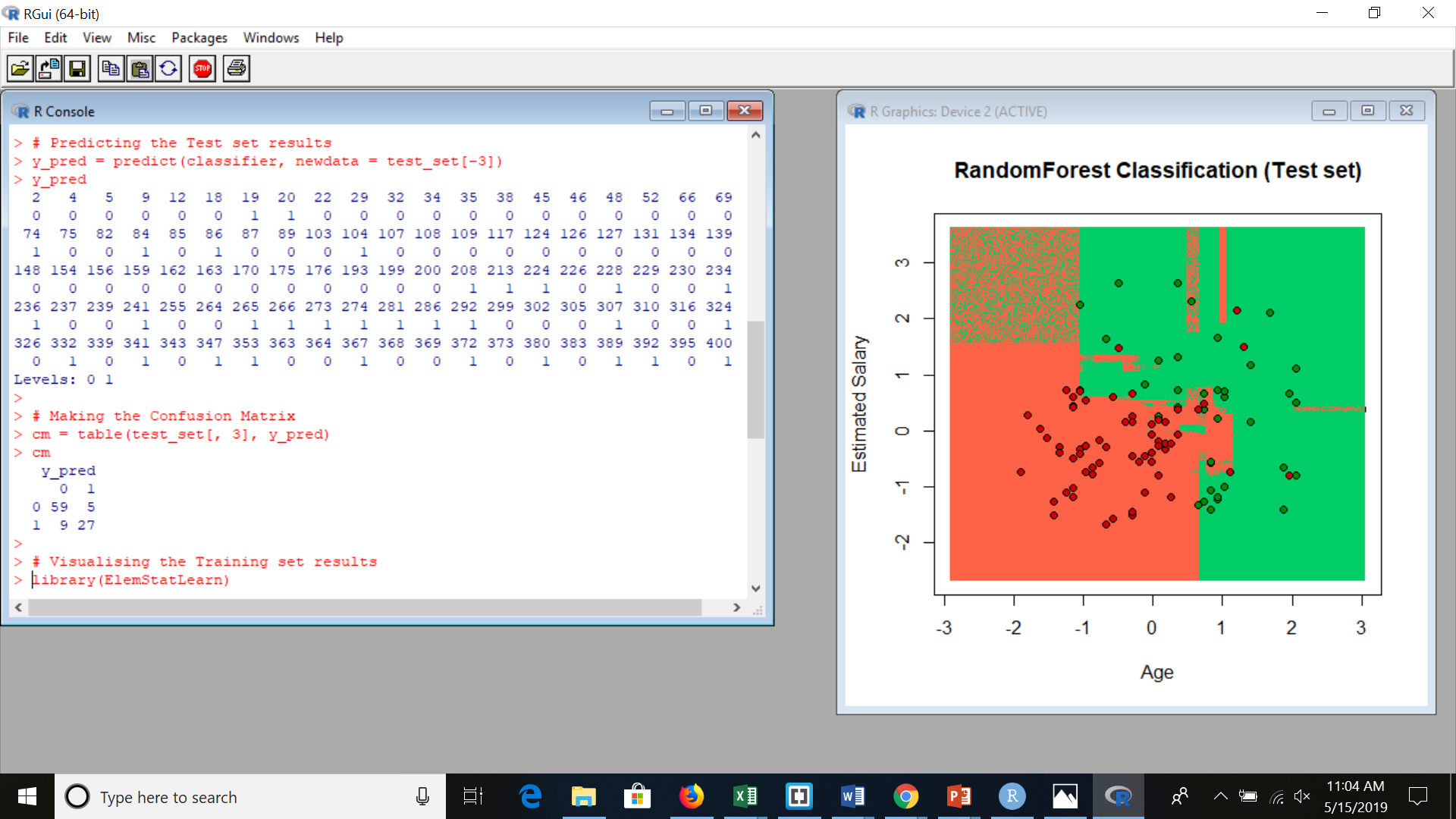
Support Vector Machine-



Random Forest-



Confusion Matrix-



Accuracy- 0.86

Naive Bayes Theorem-

* It implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:
* Accuracy -86
* Confusion Matrix

