

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6015 20904 Intermediate Analytics**

**Professor Name: ChuanLi Jiang,**

**Student Name: Dong Quoc Tuong (Lukas)**

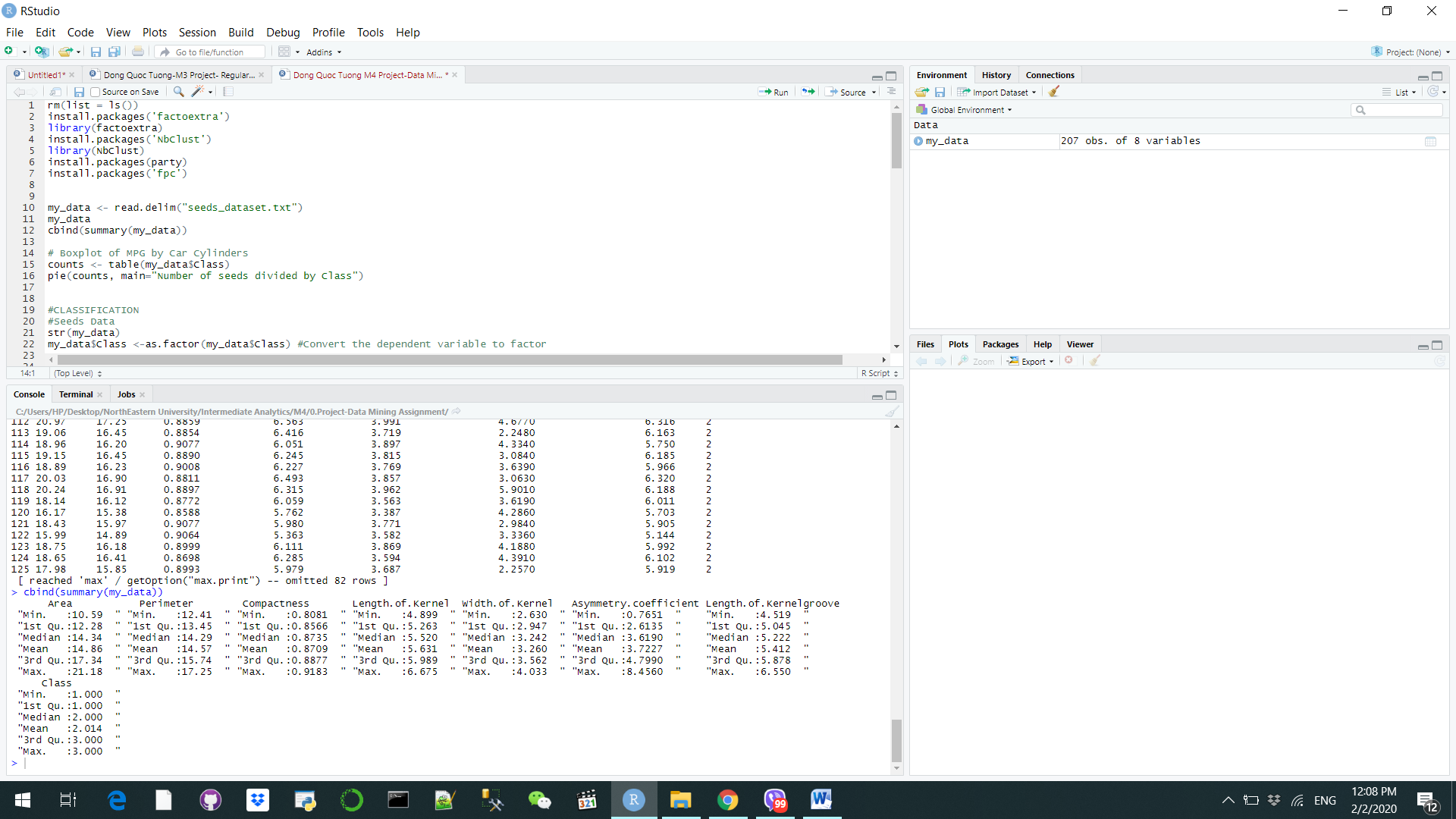
**Student Class: Fall 2019 CPS Term: A. 2020**

|  |
| --- |
| **Module 4: Data Mining Assignment**  **Completion Date: February 2nd Due Time:12:00am** |

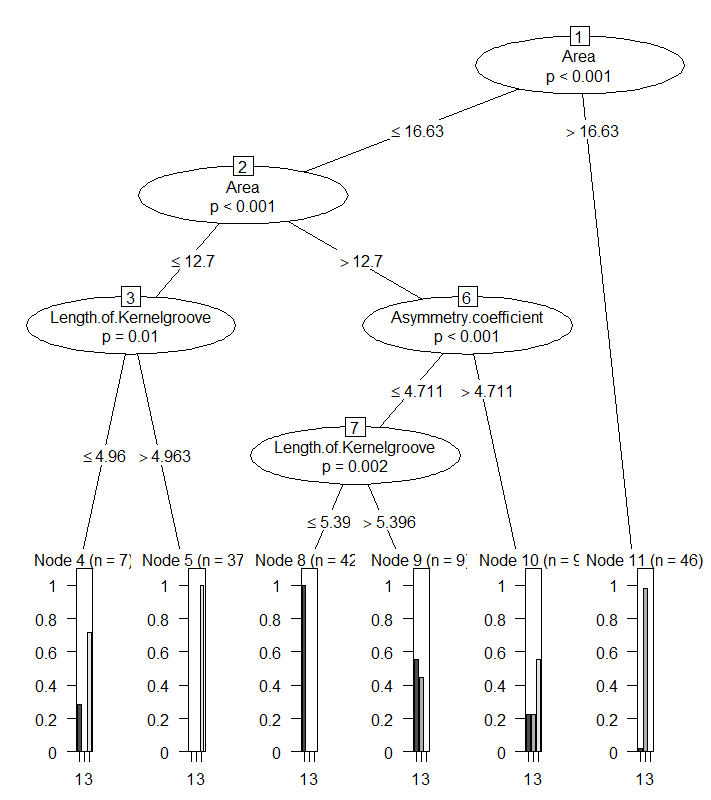
**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

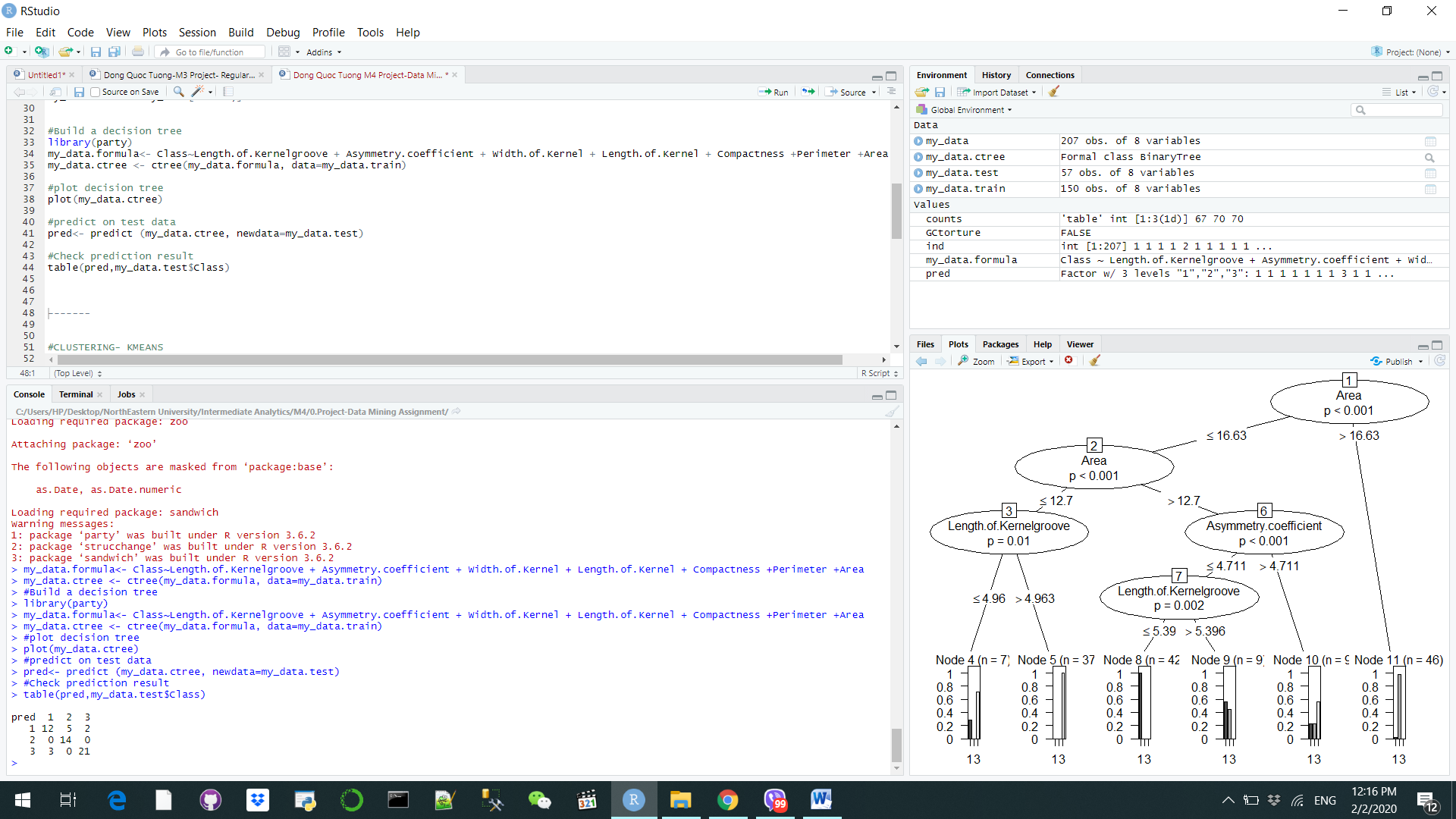
In this paper, we will master Data mining process though predicting the dataset with Classification and Clustering methods. First and foremost, we load the necessary packages in R and choose the dataset of wheat seeds with three different types: Kama, Rosa and Canadian. Each class contains 70 elements and is classified numerically. The experiment measured the 6 different factors: Area, Perimeter, Compactness, Length of Kernel, Width of Kernel, Asymmetry Coefficient, and Length of Kernel Groove. Summary is listed below.

**Classification**



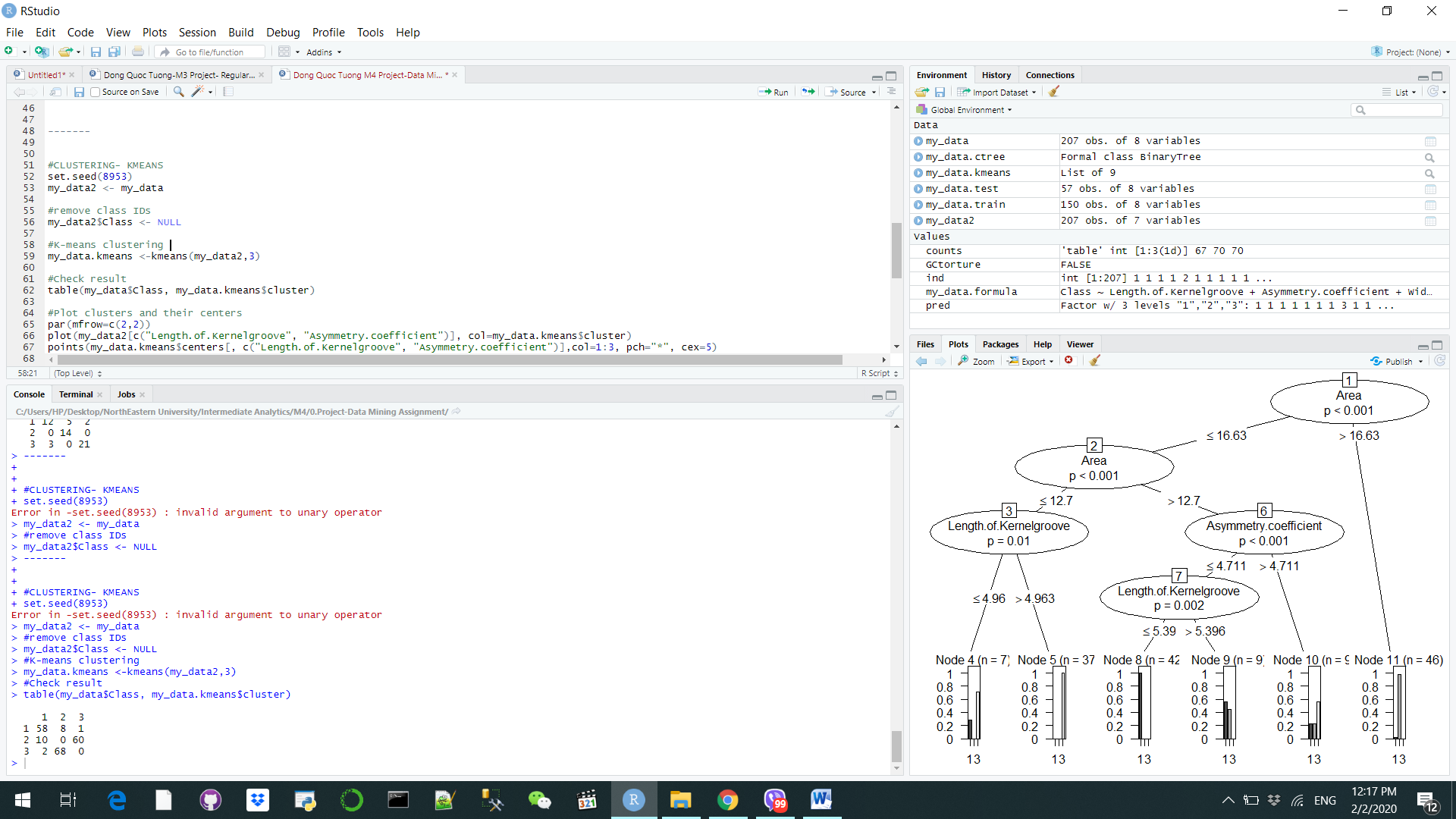
In machine learning, Classification is the problem where we look at the which set of the category that the variable belong to. (Garbade, M. J., 2018). First stop, we split the dataset into 2 parts, training set and test set at the ratio of the 7/3. Our aim is to see that if our prediction of the test’s predictors based on the training set will match with the test’s dependent variables. Next, we build our decision tree with the ‘party’ library and see how these 6 factors affect the Class of the seeds. The first and second split indicates that Area is the first factor that we need to have a look when categorizing these seeds. If the area is larger than 16.63 then it will be classified as node 11, which comprises of 46 inputs. If the area is between 12.7 and 16.63, they can be classified as node 8,9,10. Anything less than 12.7 would be node 4 and 5. The tree shows that length of Kermel groove and Asymmetry Coefficient were used for the splits but does not specify which feature is more important than the other. Nevertheless, we do know that length of Kermel groove does play an essential role at the end when we classifying node 8 and 9.

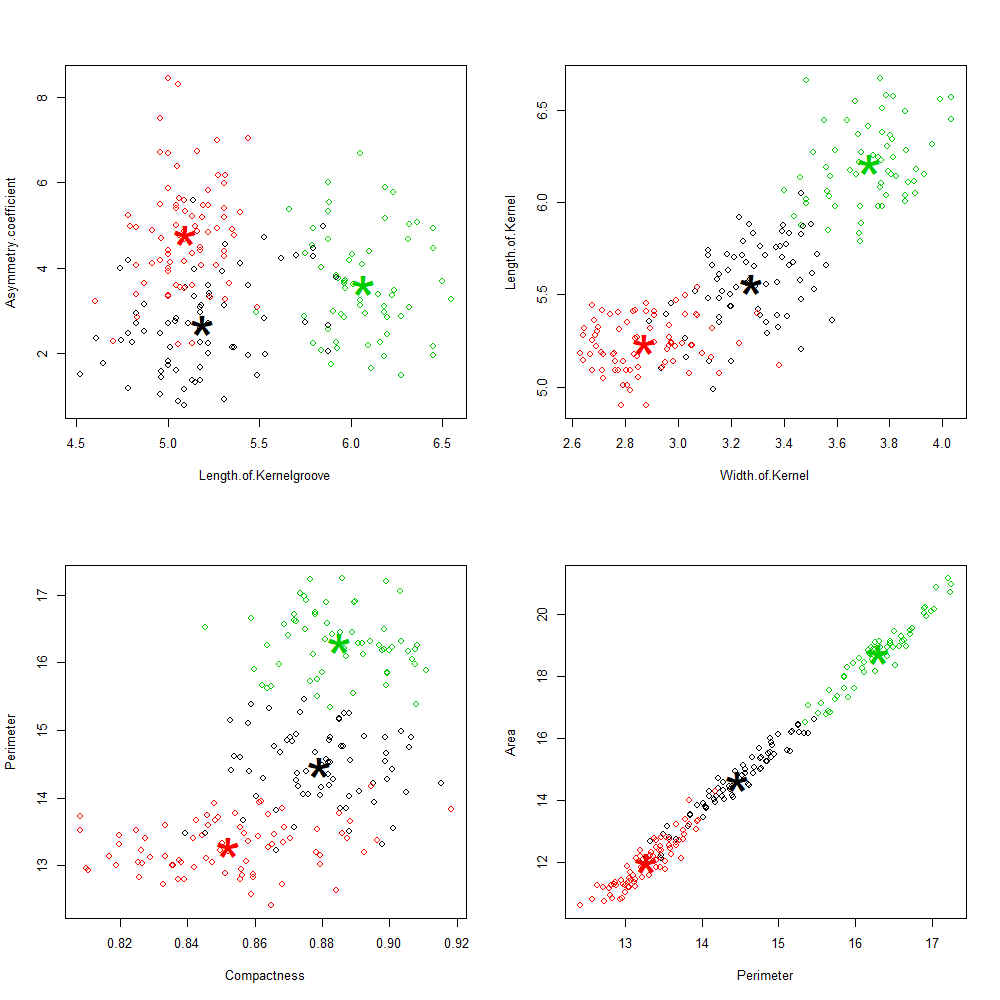
After that, we draw the comparison table between our prediction and the test set. As we see that there are 47 accurate predictions and 10 that are not correct while the rest is unknown. This is not a very good prediction and we will compare it with the Clustering method.



**Clustering- Kmeans**

Clustering is defined as a task of grouping a set of objects that shared similarity out from a tremendous amount inputs. There are many types of Clustering but we will focus on the Density based and K means. There are 58 correct predictions and 129 incorrect ones, with a few that are not detectable.



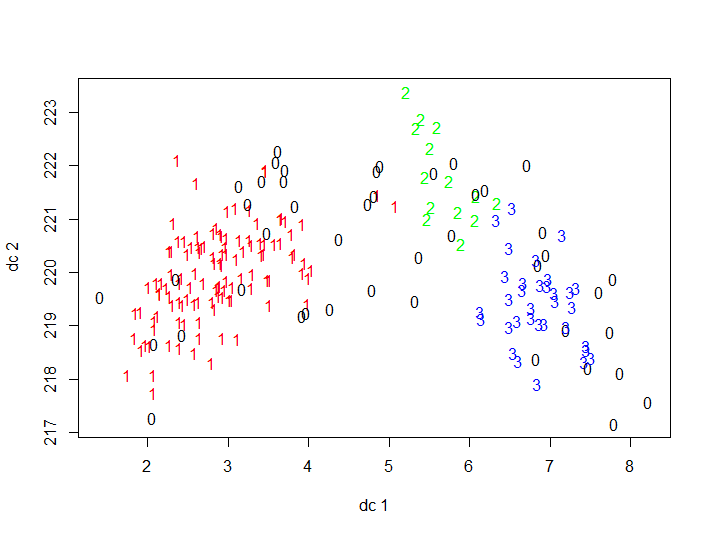


|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Class/ Factors | Area | Perimeter | Compactness | Length of Kernel | Width of Kernel | Asymmetry  Coefficient | Length of Kernel Groove |
| 1 | 12-14.5 | 12-14.5 | 0.82-0.88 | 5-5.5 | 2.6-3.2 | 4-7 | 4.5-5.5 |
| 2 | 14.5-16 | 14.5-15.5 | 0.86-0.9 | 5.5-6 | 3.2-3.4 | 0-4 | 4.5-5.5 |
| 3 | 16-22 | 15.5-18 | 0.86-0.9 | 6-6.5 | 3.6-4 | 2-6 | 5.5-6.5 |

We illustrated the dataset above in different groups and created centroids within them. Since there are only 3 classes, we will group them in 3 different groups. From the plots above we can see that the data is grouped different when it comes to different factors. For example, if the data is based on the Asymmetry coefficient and Length of Kemelgroove, they are untidy and chaotic. However, when it comes to the other, they resemble a positive curve from left to right with the Class 1 (red) at the bottom, Class 2 (black) in the middle and Class 3 (green) in at the top. The most prominent example of this are the final factors: Area and Perimeter.

**Clustering- Density based**

One of the major advantages of the Density based Clustering is the ability to detect anomaly and prevent them from pulling the centroids to them (Nandi M.,2015). We will choose to have try to group them into 3 neighborhood with the size of the epsilon neighborhood of 0.8 and keep the same defaults min points as 5. Then we plot our result and see that among the 3 groups that we have, there are a couple of outliners that could not be used to model an accurate prediction. We can eliminate them the next time that we do our analysis.



**References**

Garbade, M. J. (2018). Understanding K-means clustering in machine learning. Retrieved from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Nandi, M. (2015). Density-based clustering. Retrieved from https://blog.dominodatalab.com/topology-and-density-based-clustering/