CST8390 BI & Data Analytics

Assignment I: Analysis of Seeds Dataset

Elaiza Rivera

(040839516)

&

Abdullah Zeki Ilgun

(040991363)

1&2. Selecting Dataset

a. **Exasens Data Set**
Exasens is a multivariate data set made for providing classification for different respiratory diseases, which are COPD, asthma, infected, and HC with **399 instances** in it. The data are collected for a joint research project, and the set can be analyzed by classification or clustering. There are **6 attributes** that are mostly set to be integers, and each one provides information about a different patient. The attributes are patient's ID, age, gender (1 for male, 0 for female), smoking status (1 for non-smoker, 2 for ex-smoker, 3 for smoker), and saliva permittivity as a complex number with imaginary and real parts and min/max values for each part. The only non-integer attribute is the diagnosis of the patient which is a nominal attribute with 4 categories. Two for outpatients without acute respiratory infections: COPD and asthma. Patients without CODP and asthma but having infections: Infected. Lastly, the healthy controls: HC.

b. **Seeds Dataset**
Clustering was a method used with various data mining. In this this study, there were three types of wheat varieties: Kama, Rosa and Canadian each with **70 elements** to be clustered. With the use of soft X-ray technique, kernel structures were visualized compare to using sophisticated imaging techniques. Complete Gradient Clustering Algorithm (CGCA) was used to get an accurate clustering result. The clustering method was determined by its seven attributes area, perimeter, compactness, length of the kernel, width of the kernel, asymmetry coefficient and length kernel groove. With these attributes, the imaging will characterize a kernel by its density estimator. It will divide high densified kernels from a sparser object and organized in ascending manner. The result of the said study shows that there were 6 incorrectly classified kernels.

c. **For this assignment we have chosen Seeds Dataset**
Even though we think that Exasens data set is interesting regarding its content that allows us to see how smoking marijuana may affect the possibility of having certain diseases, compare to the Exasens data set, the seeds data was more straight-forward, organized regarding the columns and was easy to comprehend. In Exasens, there are columns that have missing with data, and some ID attributes are inconsistent with the format – some IDs are in date format. It also has attributes that contain sub-attributes. Finally, the seeds dataset became our choice since the data given were complete, and consistent with its values and easy to comprehend.

3. Analyzing data

There was a total of **210 samples** collected randomly. **70 kernels variety** which were the class nominal instance as 1 – Kama, 2 – Rosa, and 3 – Canadian. There are 7 attributes area, perimeter, compactness, length, width of the kernel, asymmetry coefficient and the length of kernel groove all are in numeric data type.

4. Loading file to Weka

Top of the raw data file, the relation's name and the attributes were defined in order to make the data loadable to Weka by using an arff file. All attributes are numeric except for the variety, which is nominal.
After loading the file to Weka from the arff file, the class attribute for each type of kernel was in type numerical. To understand and the properties of each kernel it was changed to nominal data type for better understanding of classification.

5. Tabulate Statistics

	Max	Min	Mean	StrdDev	Label	Count	Weight
Varieties	N/A	N/A	N/A	N/A	**1**	**70**	**70.0**
					2	**70**	**70.0**
					3	**70**	**70.0**
Area	**21.18**	**10.59**	**14.848**	**2.91**	N/A	N/A	N/A
Perimeter	**17.25**	**12.41**	**14.559**	**1.306**	N/A	N/A	N/A
Compactness	**0.918**	**0.808**	**0.871**	**0.024**	N/A	N/A	N/A
Length of Kernel	**6.675**	**4.899**	**5.629**	**0.443**	N/A	N/A	N/A
Width of Kernel	**4.033**	**2.63**	**3.259**	**0.378**	N/A	N/A	N/A
Asymmetry coefficient	**8.456**	**0.765**	**3.7**	**1.504**	N/A	N/A	N/A
Length of kernel groove	**6.55**	**4.519**	**5.408**	**0.491**	N/A	N/A	N/A

- The table gives us organized information on the attributes’ statistical values.
- Varieties are put as nominal attributes, so that there is an inner table used to show the count and weight values for different seed types.
- For the varieties, meaning of each label: 1-> Kama, 2 -> Rosa, 3 -> Canadian.
- The seed types are used in the data set with the same amount (70).
- Max, min, mean, and standard deviation values are not applicable for the varieties since they only represent the seed type rather than a specific statistic.
- Label, count and weight values are applicable for none of the statistics as they are only used for the seed varieties.

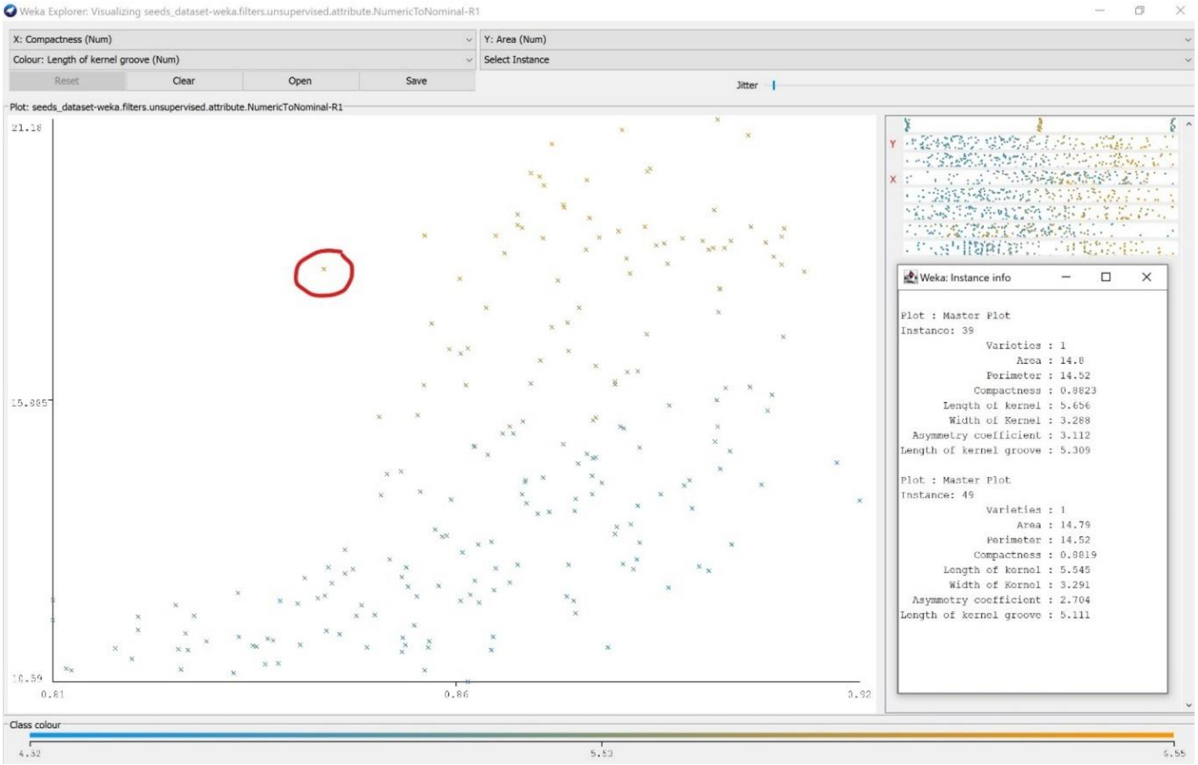
6. Data Preparation

The downloaded datasets text file was edited first by removing spaces to come data that are out of its column. This text file is then loaded as csv file, removing the space delimiter to place them according on each column. Since the class column was located at the end part of the table it was moved to the beginning. Then a header to each attribute was added to make the table more readable. After these preparations, the file is ready to load to Weka and it was successful. After it was loaded to Weka, the Varieties attribute was in numeric data type, so it was then changed to nominal by using the filter feature of the system. It was then saved it as an arff file as the final step.

7. Data Visualization

a) X – Compactness, Y – Area

Image below shows that each Classes/Varieties of the kernels are closely alike to each other with regards to their compactness and area. With the jitter set to normal, there aren't much of space between each class. Instance 95 is the only data that can be determine clearly as a certain variety (2) only by looking, unlike the others which are almost like the other class basing on its color. With the same jitter setting, it has given us instances that overlaps each other for example variety 39 and 49 which falls on the same category, where their areas only differ by .01 and their perimeter the same.


Weka Explorer: Visualizing seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1

X: Compactness (Num)

Y: Area (Num)

Colour: Length of kernel groove (Num)

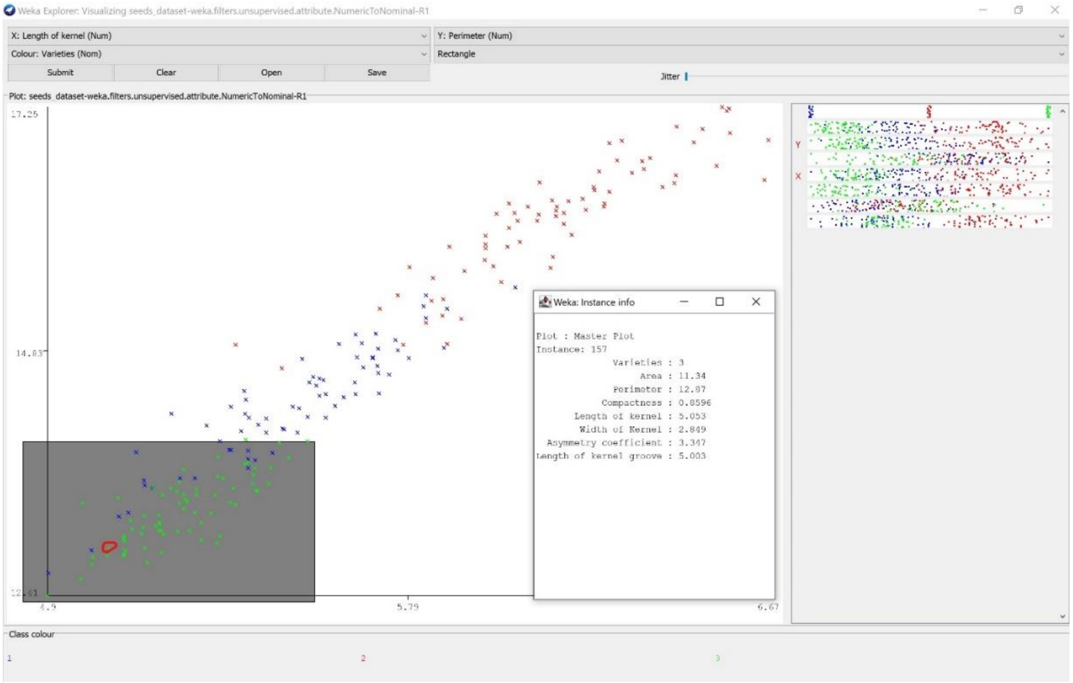
Select Instance

Plot: seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1

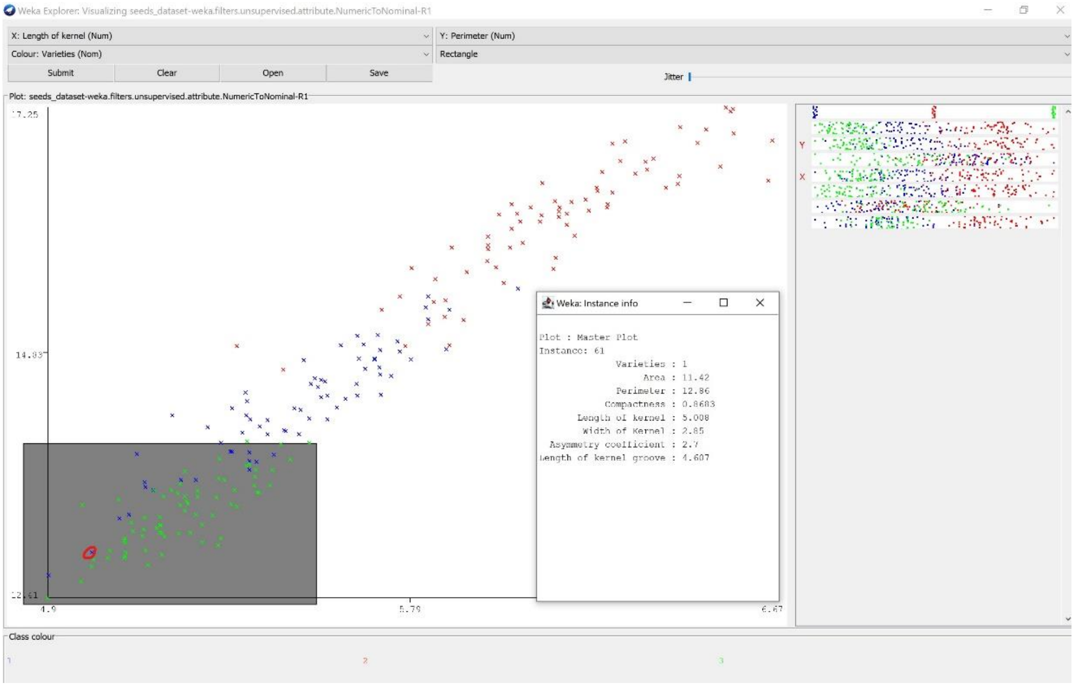
```
Plot : Master Plot
Instanco: 39
                Varictics : 1
                     Area : 14.8
                Porimeter : 14.52
              Compactness : 0.8823
         Length of kernel : 5.656
          Width of Kernel : 3.288
  Asymmetry coefficient : 3.112
Length of kernel groove : 5.309

Plot : Master Plot
Instance: 49
                Varieties : 1
                     Area : 14.79
                Perimeter : 14.52
              Compactness : 0.8819
         Length of kornol : 5.545
          Width of Kornol : 3.291
  Asymmotry coefficiont : 2.704
Length of kernel groove : 5.111
```

b) X – Length of Kernel, Y – Perimeter

On the images attached below, we select 2 samples. They are clearly with separate classes but contained similar properties or closely alike with slight difference. Instance 157 – Variety 3 (Canadian) and instance 61 – Variety 1 (Kama) are almost alike. Looking at their area, perimeters, compactness, length, and width of their kernel have a very slight difference. Without the other properties those samples can be identify as one of each variety.


Weka Explorer: Visualizing seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1
X: Length of kernel (Num)
Y: Perimeter (Num)
Colour: Varieties (Nom)
Rectangle
Submit
Clear
Open
Save
Jitter
Plot: seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1
17.25
14.83
12.41
4.9
5.79
6.67
Y
X
Class colour
1
2
3
Weka: Instance info

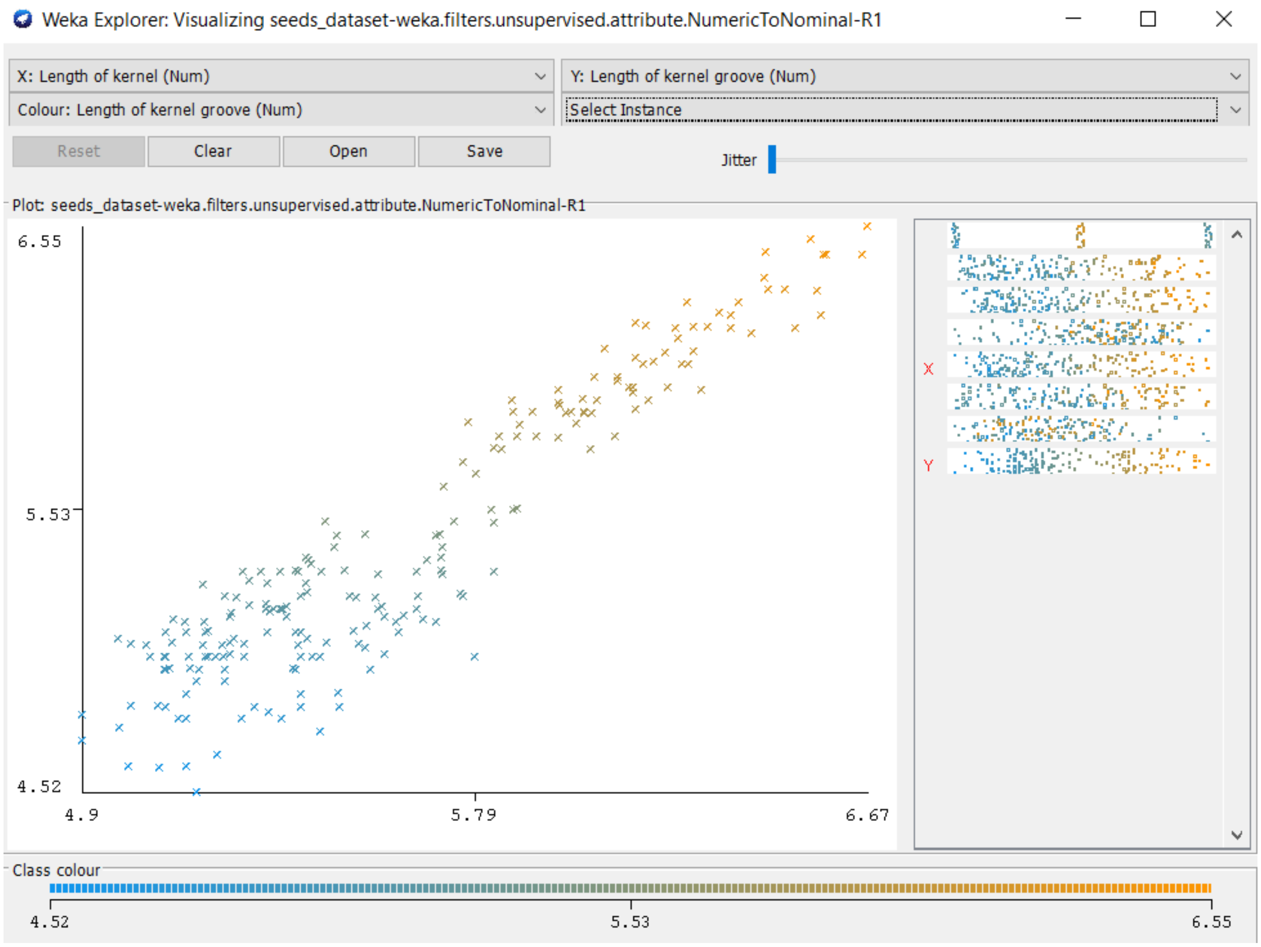
```
Plot : Master Plot
Instance: 157
                Varieties : 3
                     Area : 11.34
                Perimeter : 12.87
              Compactness : 0.8596
         Length of kernel : 5.053
          Width of Kernel : 2.849
   Asymmetry coefficient : 3.347
Length of kernel groove : 5.003
```


Weka Explorer: Visualizing seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1
X: Length of kernel (Num)
Y: Perimeter (Num)
Colour: Varieties (Nom)
Rectangle
Submit
Clear
Open
Save
Jitter
Plot: seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1
17.25
14.83
12.41
4.9
5.79
6.67
Y
X
Class colour
1
2
3
Weka: Instance info

```
Plot : Master Plot
Instance: 61
                Varieties : 1
                     Area : 11.42
                Perimeter : 12.86
              Compactness : 0.8683
         Length of kernel : 5.008
          Width of Kernel : 2.85
   Asymmetry coefficient : 2.7
Length of kernel groove : 4.607
```

c) X – Length of Kernel, Y – Length of Kernel Groove

For the last chart, we chose the one that compares length of kernel and the kernel groove. Although there is a right proportion between those values, there is no exact ratio between those two parameters. There are some instances which the length of the kernel groove length is higher than the ones with higher kernel length.



Weka Explorer: Visualizing seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1

X: Length of kernel (Num)

Y: Length of kernel groove (Num)

Colour: Length of kernel groove (Num)

Select Instance

Plot: seeds_dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1

8. kNN classification with 10-fold cross validation

K	**Percentage of correctly classified instances**	**True Positive Rate (TPR)**	**False Positive Rate (FPR)**	**Number of instances misclassified in each class**
3	92.8571%	1 – 0.857 2 – 0.971 3 – 0.957 **Ave. – 0.929**	1 – 0.036 2 – 0.021 3 – 0.050 **Ave. – 0.036**	1 – 10 (2&3) 2 – 2 (1) 3 – 3 (1)
5	92.381%	1 – 0.857 2 – 0.971 3 – 0.943 **Ave. – 0.924**	1 – 0.043 2 – 0.029 3 – 0.043 **Ave. – 0.038**	1 – 10 (2&3) 2 – 2 (1) 3 – 4 (1)
7	92.8571%	1 – 0.871 2 – 0.957 3 – 0.957 **Ave. – 0.929**	1 – 0.043 2 – 0.021 3 – 0.043 **Ave. – 0.036**	1 – 9 (2&3) 2 – 3 (1) 3 – 3 (3)
9	92.381%	1 – 0.857 2 – 0.957 3 – 0.957 **Ave. – 0.924**	1 – 0.043 2 – 0.021 3 – 0.050 **Ave. – 0.038**	1 – 10 (2&3) 2 – 3 (1) 3 – 3 (1)
11	91.4286%	1 – 0.843 2 – 0.957 3 – 0.943 **Ave. – 0.914**	1 – 0.050 2 – 0.029 3 – 0.050 **Ave. – 0.043**	1 – 11 (2&3) 2 – 3 (1) 3 – 4 (1)
13	92.8571%	1 – 0.857 2 – 0.971 3 – 0.957 **Ave. – 0.929**	1 – 0.036 2 – 0.021 3 – 0.050 **Ave. – 0.036**	1 – 10 (2&3) 2 – 2 (1) 3 – 3 (1)

- In order to do the kNN classification, firstly numeric to nominal filter was applied.
- Then, kNN method was chosen in classify section, and the cross-validation was set to 10.
- Finally, the classification was applied k value in 3.
- Classification was done also for 3,5,7,9,11,13, and the accuracy statistics were recorded to the table for each one.
- The table shows the classification ratings, and number of misclassified instances from the information gained from the confusion matrix.
- Guide to understand the last column:
  1 – 10 (2&3) -> means there are 10 misclassified instances, and those instances were classified as class 2 and 3

9. KNN classification with 10-fold cross validation for Percentage split of 70%

K	**Percentage of correctly classified instances**	**True Positive Rate (TPR)**	**False Positive Rate (FPR)**	**Number of instances misclassified in each class**
3	95.2381 %	1 – 1.000 2 – 0.917 3 – 0.958 **Ave. – 0.952**	1 – 0.063 2 – 0.000 3 – 0.000 **Ave. – 0.015**	1 – 0 2 – 2 (Class 1) 3 – 1 (Class 1)
5	96.8254%	1 – 1.000 2 – 0.958 3 – 0.958 **Ave. – 0.968**	1 – 0.882 2 – 1.000 3 – 1.000 **Ave. – 0.010**	1 – 0 2 – 1 (Class 1) 3 – 1 (Class 1)
7	95.2381%	1 – 1.000 2 – 0.917 3 – 0.958 **Ave. – 0.952**	1 – 0.063 2 – 0.000 3 – 0.000 **Ave. – 0.015**	1 – 0 2 – 2 (Class 1) 3 – 1 (Class 1)
9	92.0635%	1 – 0.933 2 – 0.917 3 – 0.917 **Ave. – 0.921**	1 – 0.083 2 – 0.000 3 – 0.026 **Ave. – 0.030**	1 – 1 (Class 3) 2 – 2 (Class 1) 3 – 2 (Class 1)
11	93.6508%	1 – 1.000 2 – 0.917 3 – 0.917 **Ave. – 0.937**	1 – 0.083 2 – 0.000 3 – 0.000 **Ave. – 0.020**	1 – 0 2 – 2 (Class 1) 3 – 2 (Class 1)
13	92.0635%	1 – 0.933 2 – 0.917 3 – 0.917 **Ave. – 0.921**	1 – 0.083 2 – 1.000 3 – 0.026 **Ave. – 0.030**	1 – 1 (Class 3) 2 – 2 (Class1) 3 – 2 (Class 1)

- KNN validation method was done again, but with percentage split of 70% here. In the last table, the percentage split was 66% by default.
- With higher percentage split, the classification accuracy is higher for most of them, except for the one that k is 7.

10. Repeat step 8 for 2 other seeds

a. Seed of 7

K	**Percentage of correctly classified instances**	**True Positive Rate (TPR)**	**False Positive Rate (FPR)**	**Number of instances misclassified in each class**
3	92.8571%	1 – 0.843 2 – 0.971 3 – 1.971 **Ave. – 0.929**	1 – 0.029 2 – 0.029 3 – 0.050 **Ave. – 0.036**	1 – 11 (Class 2 – 4, Class 3 - 7) 2 – 2 (Class 1) 3 – 2 (Class 1)
5	93.8095%	1 – 0.871 2 – 0.971 3 – 0.971 **Ave. – 0.938**	1 – 0.029 2 – 0.025 3 – 0.043 **Ave. – 0.031**	1 – 9 (Class 2 – 3, Class 3 - 6) 2 – 2 (Class 1 - 2) 3 – 2 (Class 1)
7	92.8571%	1 – 0.857 2 – 0.957 3 – 0.971 **Ave. – 0.929**	1 – 0.036 2 – 0.021 3 – 0.050 **Ave. – 0.036**	1 – 10 (Class 2 – 3, Class 3 - 7) 2 – 3 (Class 1) 3 – 2 (Class 1)
9	92.381%	1 – 0.857 2 – 0.957 3 – 0.957 **Ave. – 0.924**	1 – 0.043 2 – 0.021 3 – 0.050 **Ave. – 0.038**	1 – 10 (Class 2 – 3, Class 3 - 7) 2 – 3 (Class 1) 3 – 3 (Class 1)
11	92.8571%	1 – 0.857 2 – 0.957 3 – 0.971 **Ave. – 0.929**	1 – 0.036 2 – 0.029 3 – 0.043 **Ave. – 0.036**	1 – 10 (Class 2 – 4, Class 3 - 6) 2 – 3 (Class 1) 3 – 2 (Class 1)
13	92.381%	1 – 0.857 2 – 0.957 3 – 0.957 **Ave. – 0.924**	1 – 0.043 2 – 0.021 3 – 0.050 **Ave. – 0.038**	1 – 10 (Class 2 – 3, Class 3 - 7) 2 – 3 (Class 1) 3 – 3 (Class 1)

- KNN method was repeated with the split percentage back to the default 66%. This time a 7% random seed was used for another test run.
- By looking and comparing Seed of 7 to Seed of 1 (Number 8) with similar split percentage of 66%. We receive same value of the Percentage of correctly classified instances for Ks = 3, 7 and 9. On the other note, TPR and FPR are not giving same result, yet the number of misclassified in each class vary by an average of 1 instance per classes.

b. Seed of 9

K	**Percentage of correctly classified instances**	**True Positive Rate (TPR)**	**False Positive Rate (FPR)**	**Number of instances misclassified in each class**
3	93.8095%	1 – 0.871 2 – 0.971 3 – 0.971 **Ave. – 0.938**	1 – 0.029 2 – 0.021 3 – 0.043 **Ave. – 0.031**	1 – 9 (2&3) 2 – 2 (1) 3 – 2 (1)
5	92.381%	1 – 0.843 2 – 0.971 3 – 0.957 **Ave. – 0.924**	1 – 0.036 2 – 0.029 3 – 0.050 **Ave. – 0.038**	1 – 11 (2&3) 2 – 2 (1) 3 – 3 (1)
7	92.8571%	1 – 0.871 2 – 0.971 3 – 0.943 **Ave. – 0.929**	1 – 0.043 2 – 0.021 3 – 0.043 **Ave. – 0.036**	1 – 9 (2&3) 2 – 2 (1) 3 – 4 (1)
9	91.4286%	1 – 0.843 2 – 0.957 3 – 0.943 **Ave. – 0.914**	1 – 0.050 2 – 0.021 3 – 0.057 **Ave. – 0.043**	1 – 11 (2&3) 2 – 3 (1) 3 – 4 (1)
11	90.9524%	1 – 0.829 2 – 0.957 3 – 0.943 **Ave. – 0.910**	1 – 0.050 2 – 0.029 3 – 0.057 **Ave. – 0.045**	1 – 12 (2&3) 2 – 3 (1) 3 – 4 (1)
13	91.9048%	1 – 0.843 2 – 0.957 3 – 0.957 **Ave. – 0.919**	1 – 0.043 2 – 0.029 3 – 0.050 **Ave. – 0.040**	1 – 11 (2&3) 2 – 3 (1) 3 – 3 (1)

- KNN method was repeated with the split percentage back to the default 66%. This time random seed was used for another test run was set to 9%.
- By looking and comparing Seed of 9 with 66% split percentage to Seed of 1 with split percentage 70% (Number 9). We notice similar result with the Percentage of correctly classified instances for Ks = 5 and 7. TPR and FPR did not generate same result, yet the number of misclassified instances in each class vary by an average of 1 instance per classes.