North South University

Department of Electrical and Computer Engineering

CSE445 – Machine Learning

Section: 6

Project Topic

Risk Factor Prediction of Chronic Kidney Disease

Group Members

Md. Rifat Ahmed	Faija Islam Oishe	Sumiya Sultana
1931725042	1821720042	1931277642

1. Pre-processing

To Pre-process the Dataset, we've used Pandas on the Local system using the following Codes:

```
import pandas as pd

#Importing the Dataset
df = pd.read_csv('D:\Varsity Documents\Spring 23\CSE445\Works\Project\CKD
Prediction Dataset.csv')
```

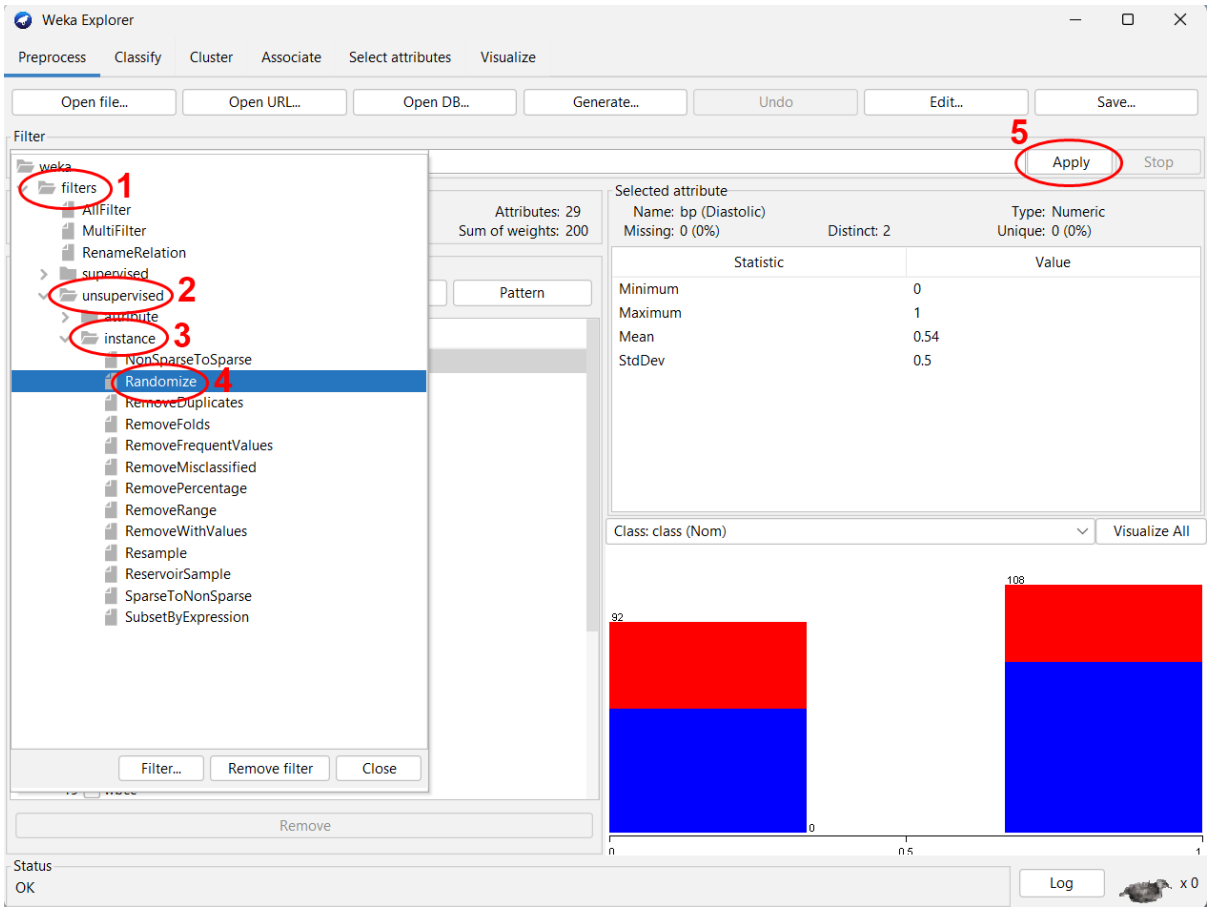
```
# Viewing the first 5 rows
df.head()
```

```
# Since the first two rows doesn't have any values needed for our work
# We'll Drop the first two rows
df = df.drop([0, 1])

# Resetting the index of the DataFrame
df = df.reset_index(drop=True)

df.head()
```

```
# Exporting the newly updated Dataframe as a new Dataset
df.to_csv('D:\Varsity Documents\Spring 23\CSE445\Works\Project\CKD Prediction
Dataset (Pre-Processed).csv', index=False)
```

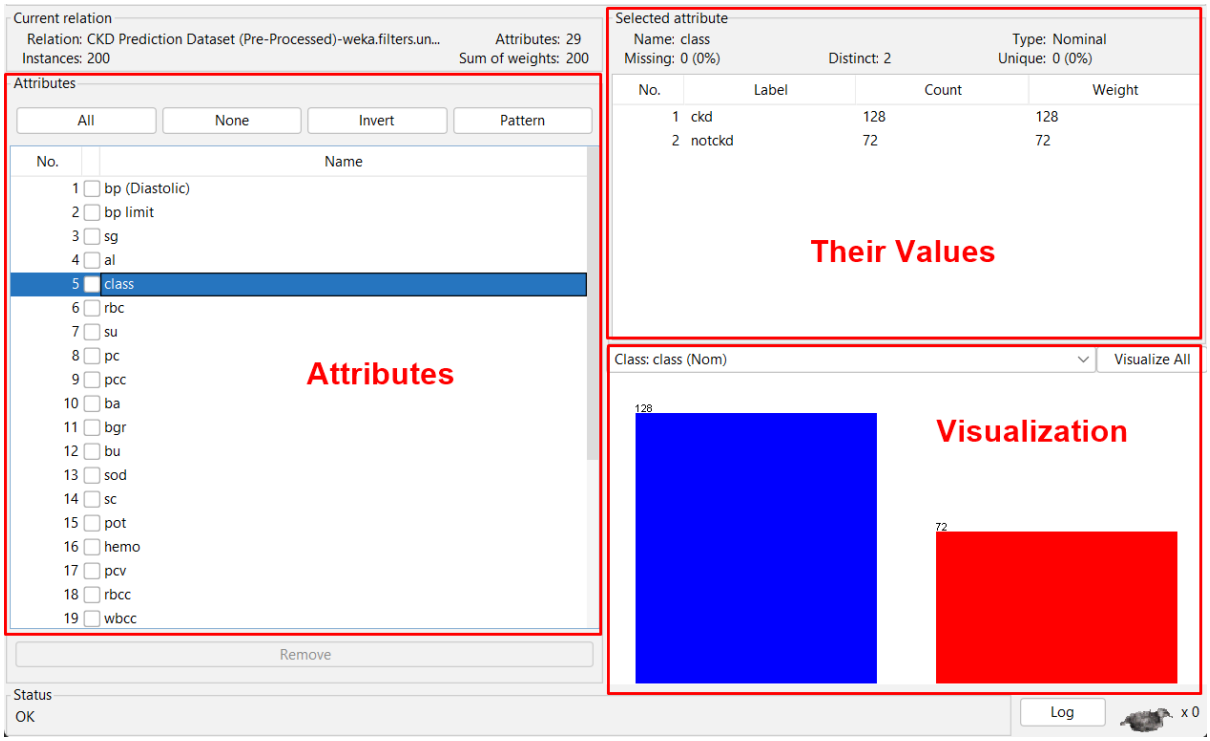
➢ Now, we'll move on to Weka for the Remaining Parts:

After opening the new CSV file on Weka we'll use the Randomize filter from Filter section to Randomly shuffle the Order of the instances in our Dataset:



Now, since we don't have any missing values in the dataset, we don't have to deal with any.

2. EDA & Visualization



On the left side we can see all the attributes (total 29 in our dataset) and clicking on them we can see their values on the right side and their visualization at the bottom right corner.

For this instance, we can see that the attribute “class” holds the information “ckd” and “notckd” meaning either someone has ckd or they do not. And the visualization is shown based on the number of instances with these values.

3. Classification

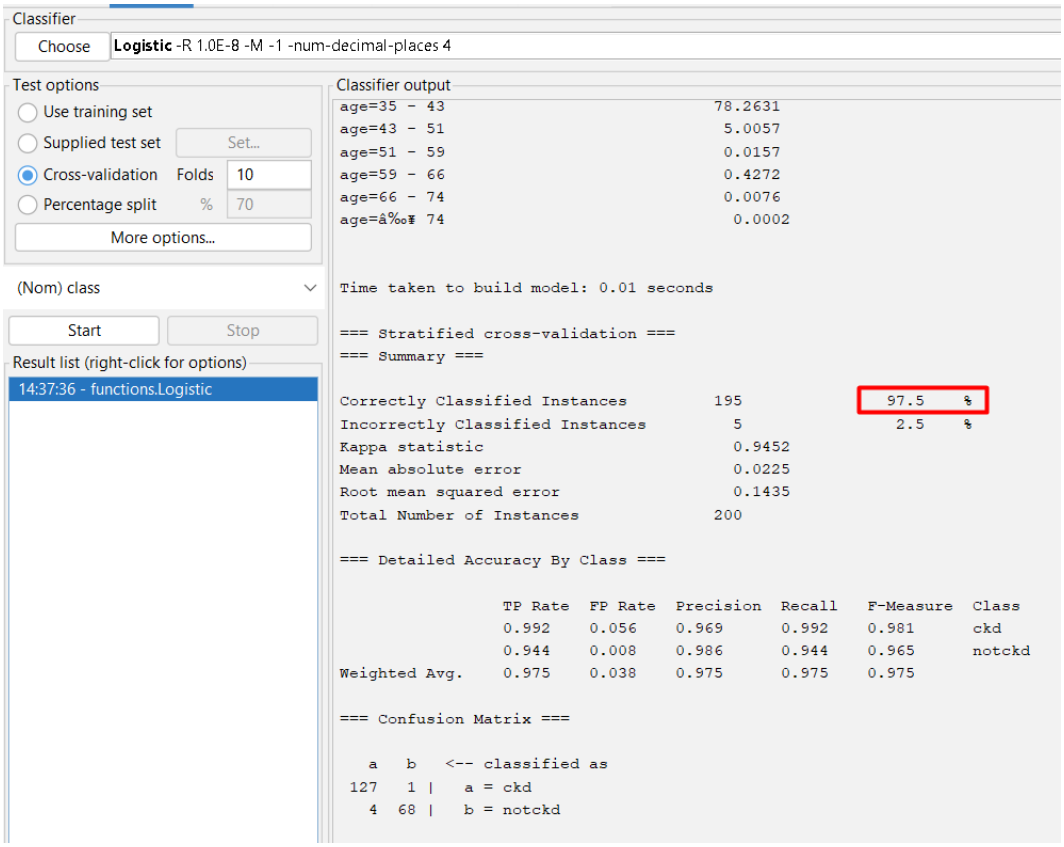
We'll be classifying the dataset using two algorithms

i) Logistic Regression
ii) Random Forest

& we'll be using both cross-validation and dataset splitting technique to get the results to see which gives the best result.

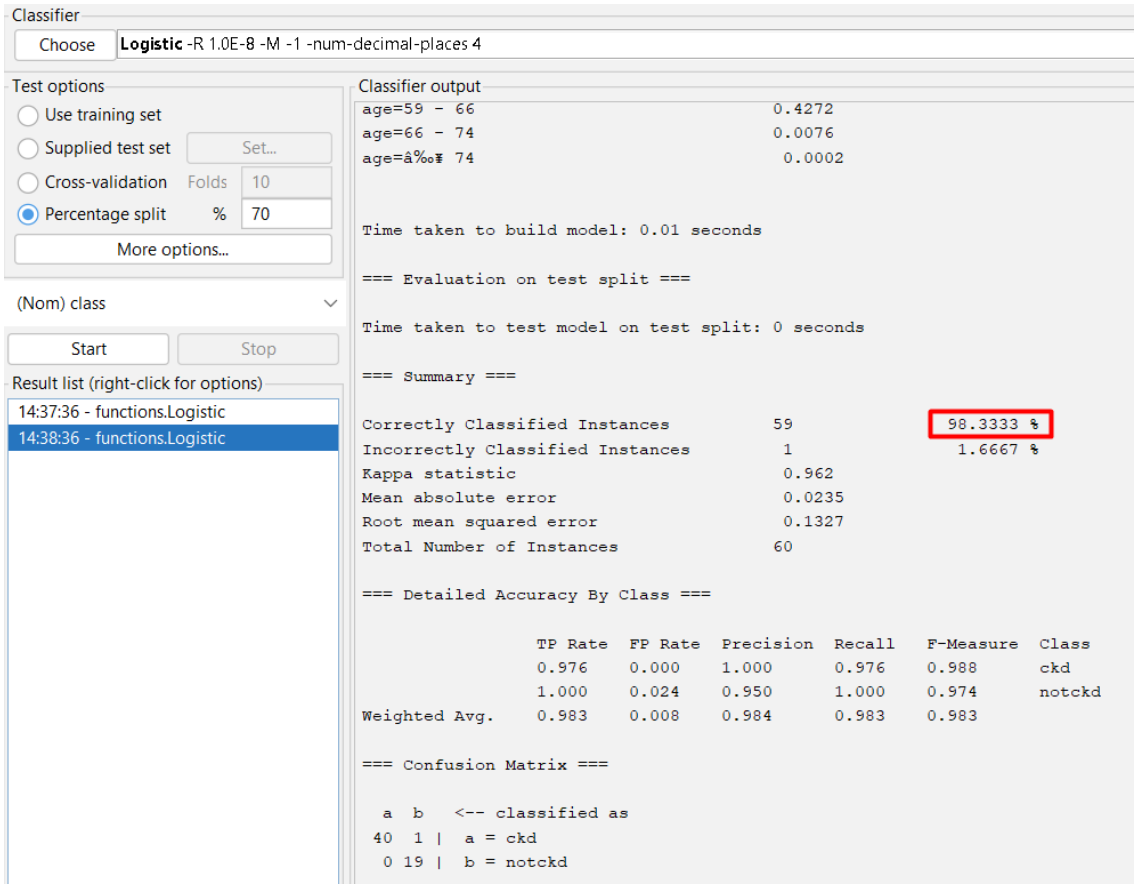
And as for the Evaluation metrices we'll be keeping MAE, RMSE, TP Rate, FP Rate. Precision, Recall & F-score.

Method 1: We'll be using Logistic Regression using Cross-validation technique with 10 folds:



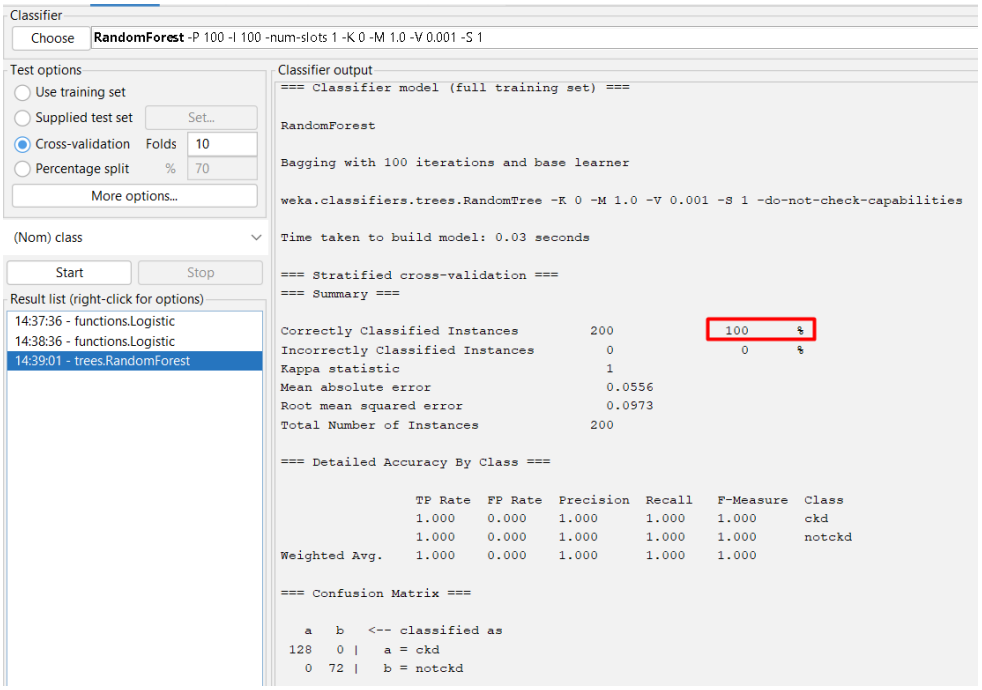
We got an accuracy of 97.5%.

Method 2: Again, using Logistic Regression but this time we split the training dataset into 70%:



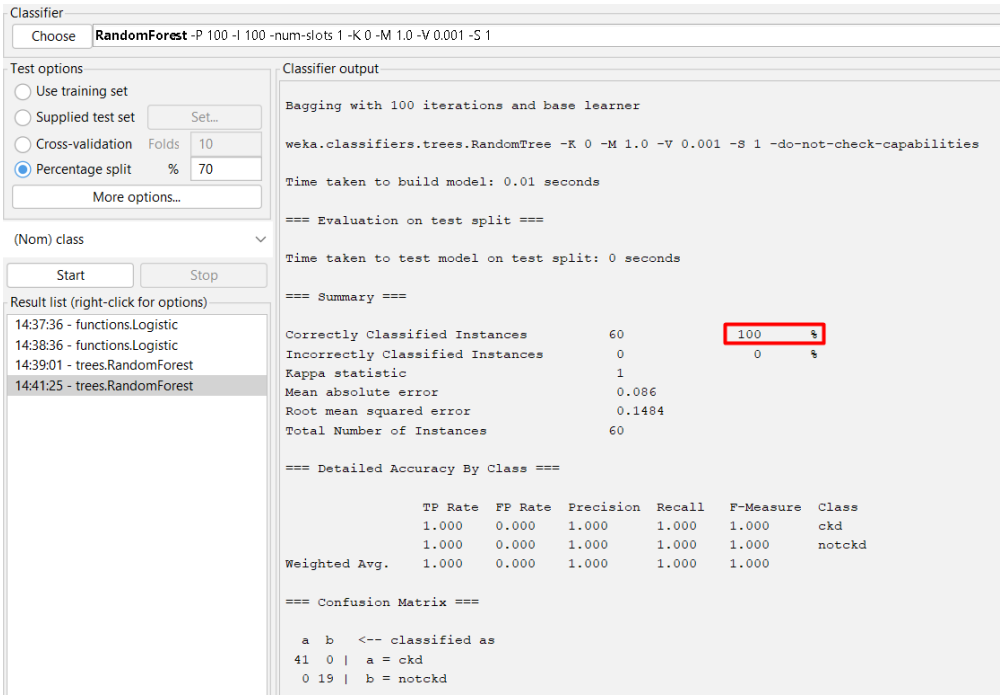
This time we got an accuracy of 98.3333%.

Method 3: Now, using Random Forest algorithm using Cross-validation technique with 10 folds:



We got an accuracy of 100%.

Method 4: Again, using Random Forest but this time we split the training dataset into 70%:



And again, we've got an accuracy of 100%.

4. Result & Conclusion

Even though the Mean Absolute Error and the Root Mean Squared Error are bit varying. Using the 1st method, we get the lowest MAE while using the 3rd method we get the lowest RMSE.

But, to classify the dataset of Chronic Kidney Disease the best algorithm would be the Random Forest since we're getting a very small error in Logistic Regression and also it's giving an accuracy of 100% using both cross-validation and splitting technique.