Aravindan Srinivasan 2981707
Mukesh Sivathanu 2980821
Tarun Swarup 2988527

PREDICTIVE MODEL TO DETERMINE WINE QUALITY THROUGH DATA MINING TECHNIQUES

I. Objective

The main aim of this project is to categorise the various classes of wine originating from countries around the world using the wine dataset. We're looking forward to implement different available data mining techniques and machine learning algorithms such as classification , clustering and time series analysis on the dataset to achieve significant results in determining wine quality. After ample research and background work , we've selected the most suitable techniques for our dataset.

- **J48 Decision Tree**
- **Multi Layer Perceptron**
- **K-Means Clustering**
- **Time Series Analysis**

Still some research is going on to identify wines through all the information available. Ultimately , successful implementation of these algorithms would let us choose the most supporting attributes that would possibly help us create a predictive model to determine the quality of wines from around the world.

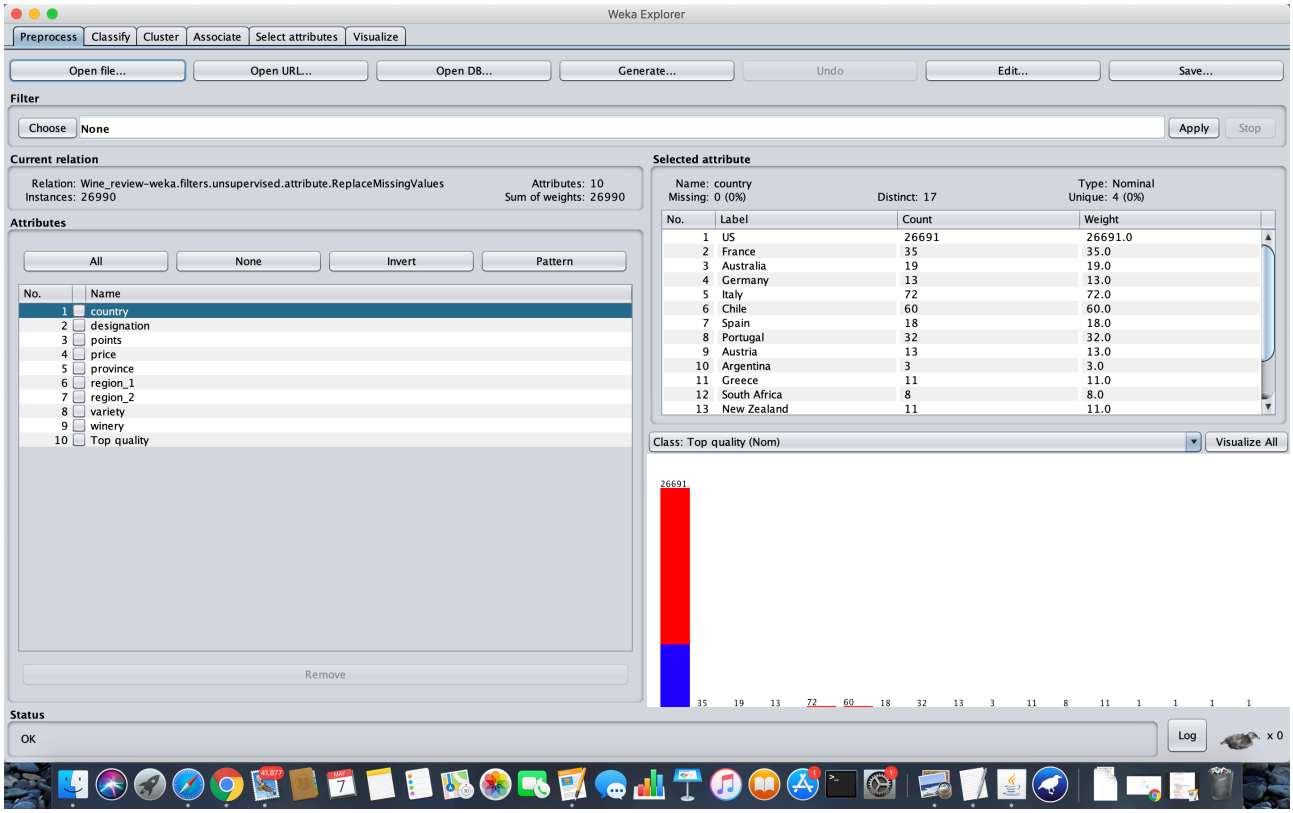
II. Description of Dataset

The data was scraped from WineEnthusiast during the week of June 15th, 2017. This dataset offers some great opportunities for sentiment analysis and other text related predictive models. The overall goal is to create a model that can identify the variety, winery, and location of a wine based on a sommelier's description. The dataset contains three separate files . https://www.kaggle.com/zynicide/wine-reviews. This company is the best in wine accessories, storage, information in the marketplace. And moreover, This website holds all the informations on wines like ratings, listing the best wines and still more information. There are totally 10 attributes and 27001 instances. Let me list out the attributes name and some information on the attributes.

- **Country:** This attribute defines in which country the wine is from. Totally there are 17 distinct countries.

- **Designation:** This attribute depicts the vineyard used for making the wines. Totally 9757 vineyards are there.
- **Points:** These are the scores given by the WineEnthusiast company. Rated from 1 to 100. (All the reviews are <=80 points)
- **Price:** Price of the wine in dollars (PRICE->Min : 4$, Max : 235$)
- **Province:** The state from which the wine is made. Totally there are 83 distinct states and 17 distinct countries.
- **Region1:** The wine growing area.
- **Region2:** More specific regions to make wines.
- **Variety:** The types of grapes used to make wine. 203 types of grapes were used.
- **Winery:** Company which made wine(3397 companies).
- **Top quality:** class variable deciding whether given wine is top quality or not.

Screenshot 1 : The wine dataset (.arff) is imported into the Weka environment.

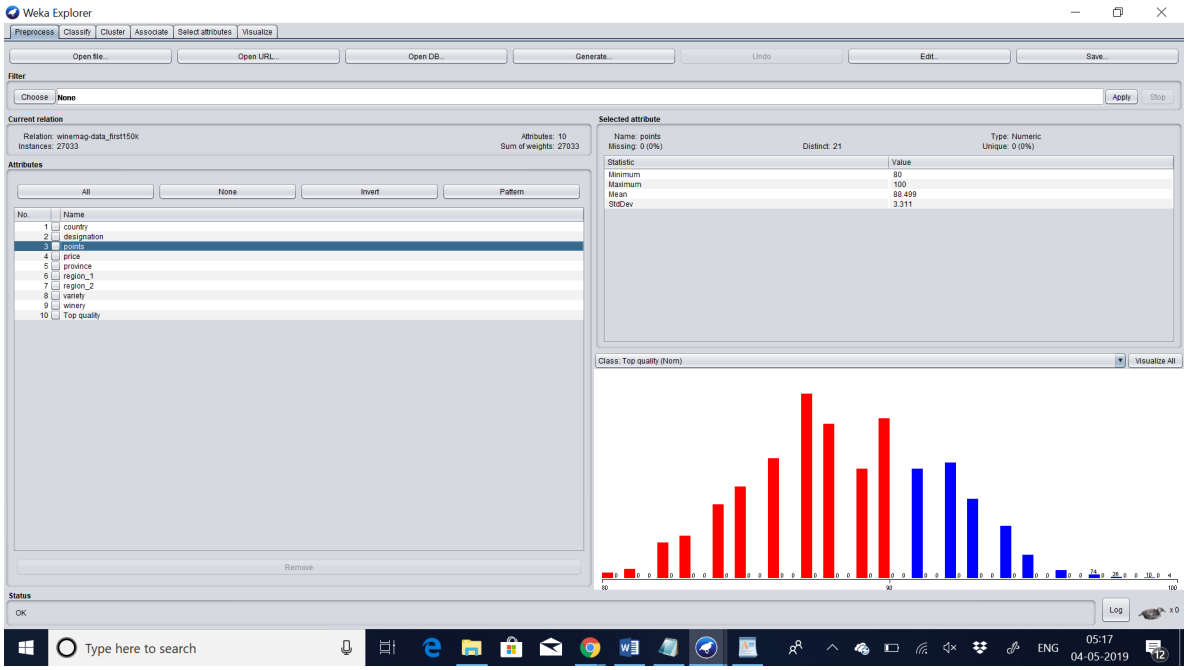


III. Preprocessing

1. Since all our attributes have only few unique values , but not a single attribute with 100% unique attributes , we are skipping the first step of preprocessing .
   (i.e) removing attribute which have 100% unique values.

2. Now we are going to perform **discretisation** on numeric values. In our dataset, points and score are the only two attribute which have numeric values. So we are going to convert them into nominal values.

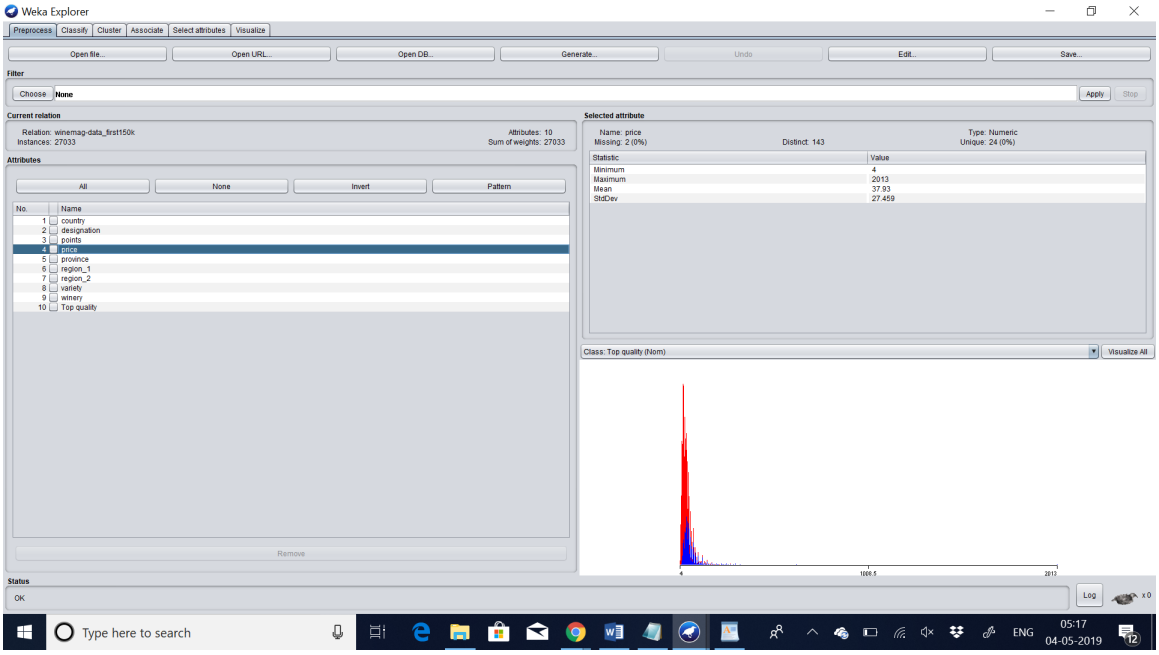
Screenshot 2 : The following screenshot shows the attribute **point** denoting as numeric .



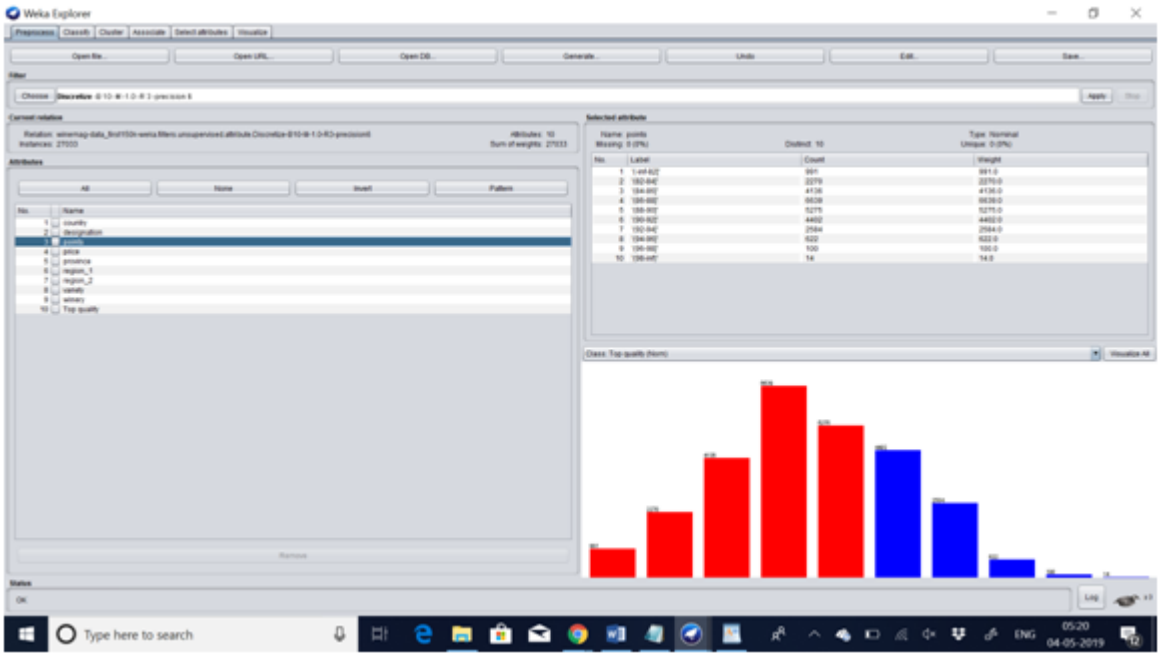
The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization. You can discretize your real valued attributes in Weka using the Discretize filter. Let us discretize the point and price attributes by choosing:

weka->filters->unsupervised->attribute->discretize.

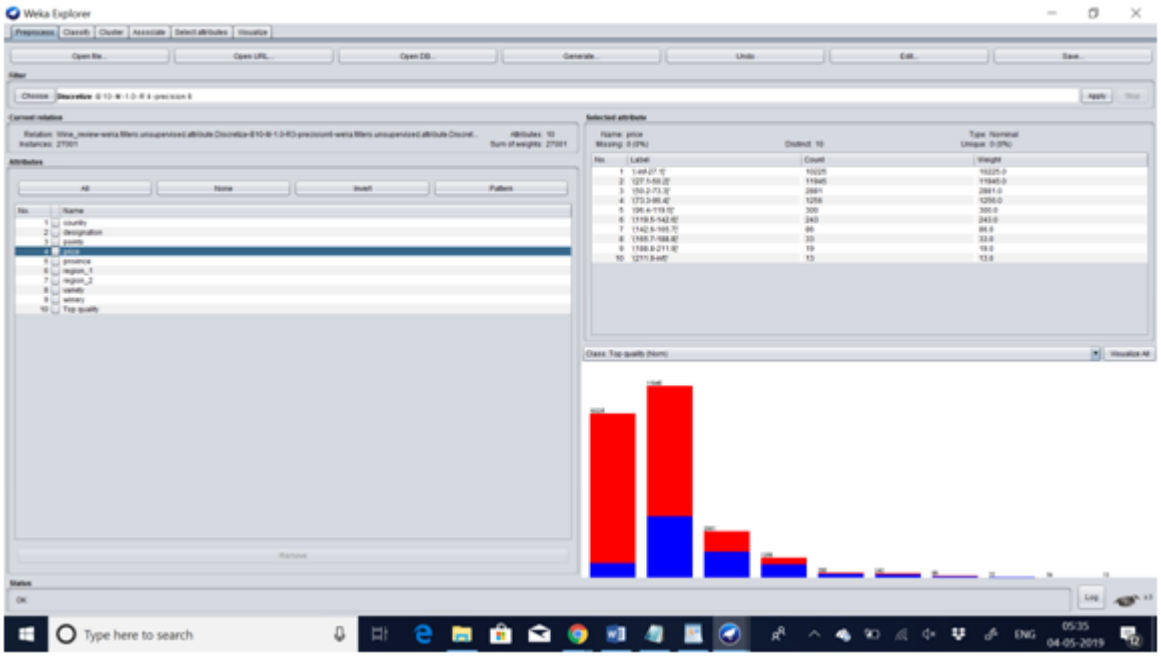
Screenshot 3 : Discretization



Screenshot 4 : Point and Price has been discretized (numeric to nominal)

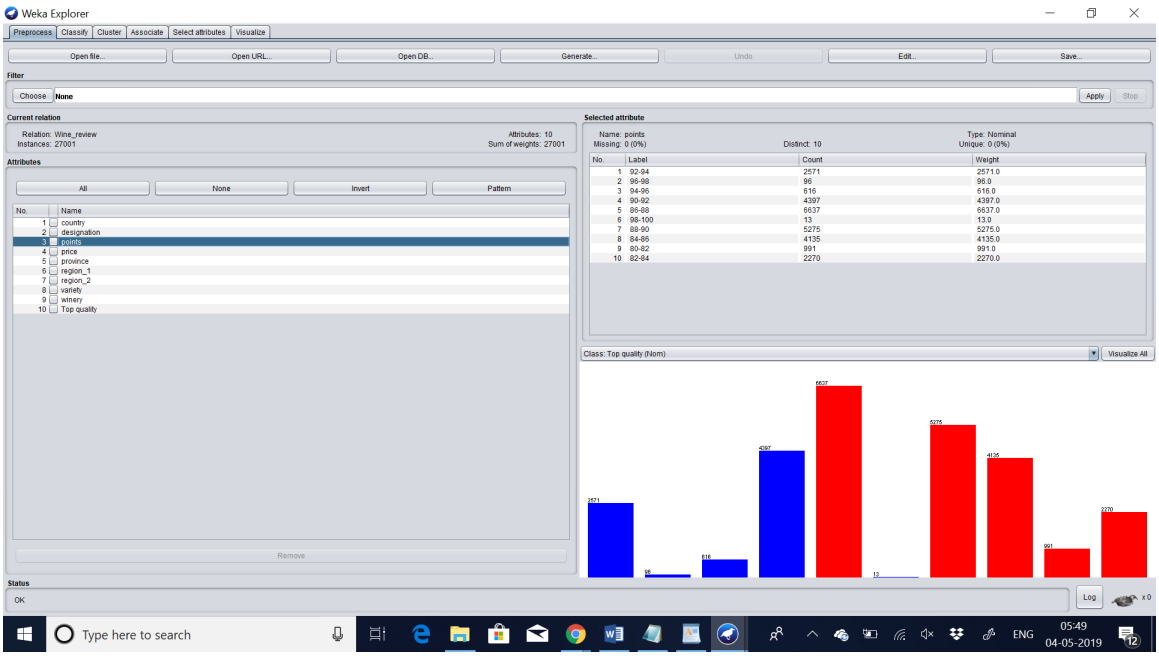


Screenshot 5 : Let us edit the labels accordingly to make it look nice.

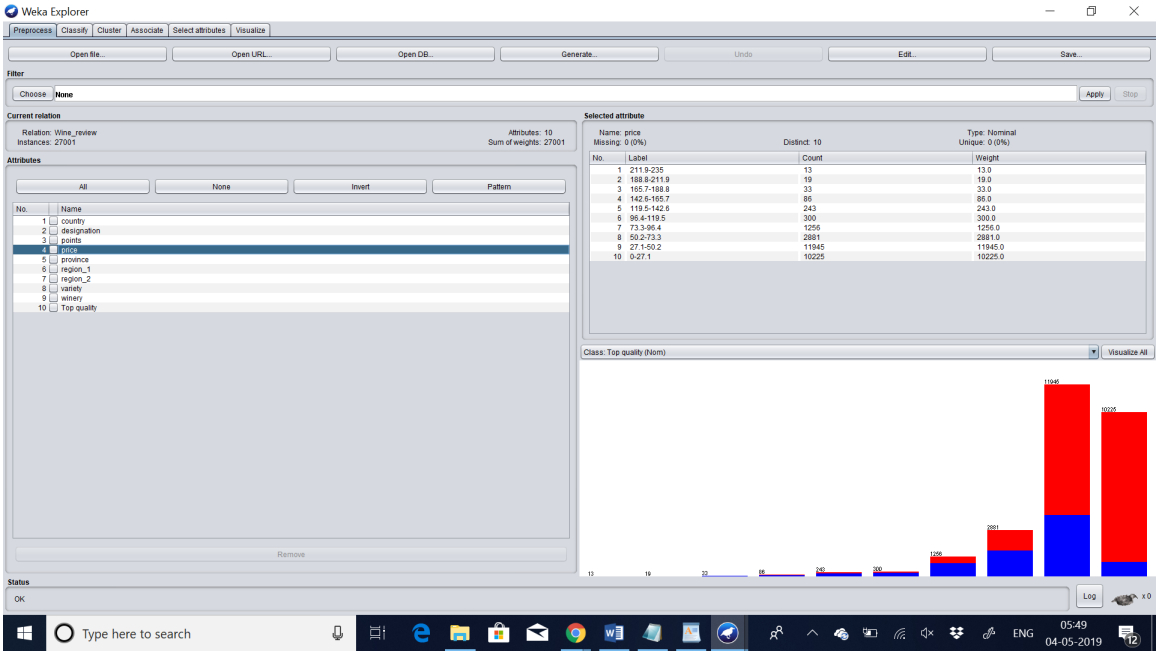


In our final preprocess, We are going to delete entire rows having missing values, because our missing values designation, region1, region2. So we cannot replace our missing values with mean values.

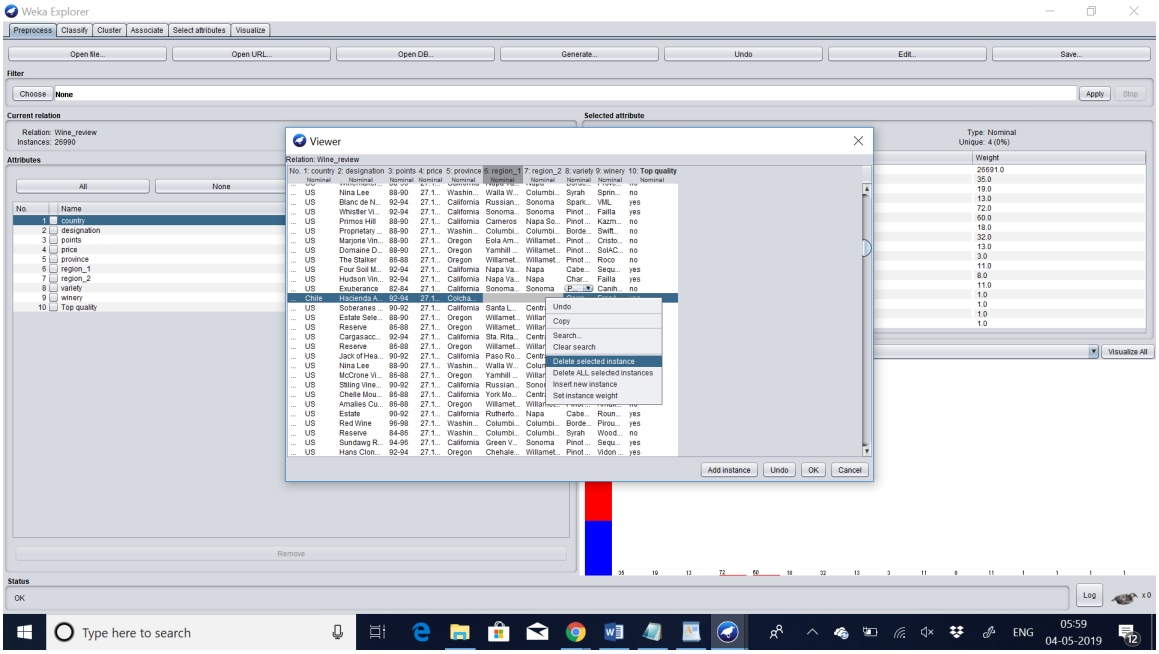
Screenshot 6 :



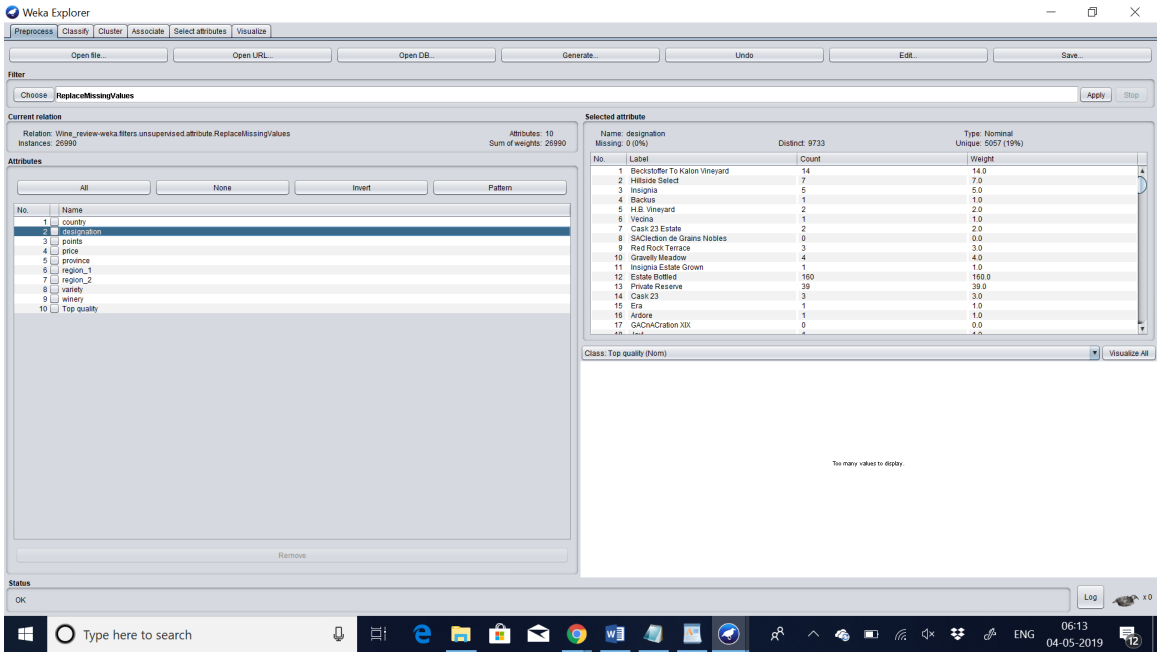
Screenshot 7 :



Screenshot 8 :



Screenshot 9 : Now we have deleted all the missing values.



Once after the necessary preprocessing steps are done , the **“dataset.arff”** file is generated.

IV. Classification

For any classification algorithm , the dataset is initially split into two , training setwhich is used to create the model and test set to assess its performance , in such a way to avoid common problems like overfitting and underfitting. Sometimes the model can turn out to be abnormally perfect that it learns the training dataset very well and negatively impacts the future performance of the model , when new data is introduced.

Training set is splitted as in **90%** of the dataset. The instance comes around **24291**.

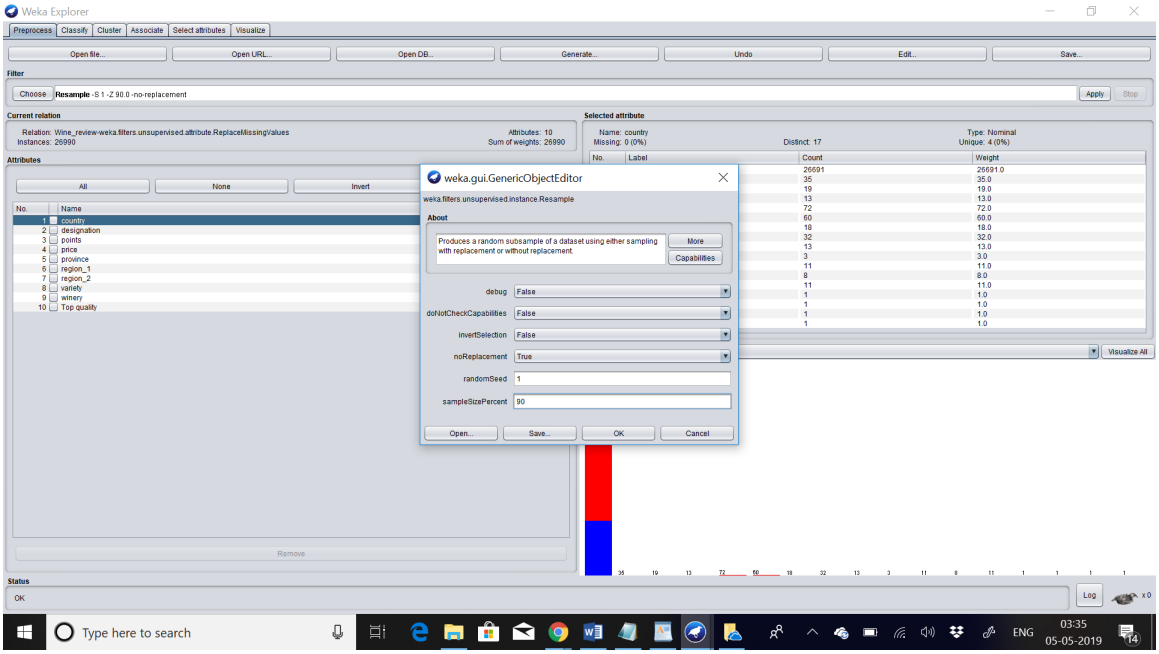
It is saved as **“trainingSet.arff”.**

Testing set is splitted as in **10%** of the dataset. The instance comes around **2699**

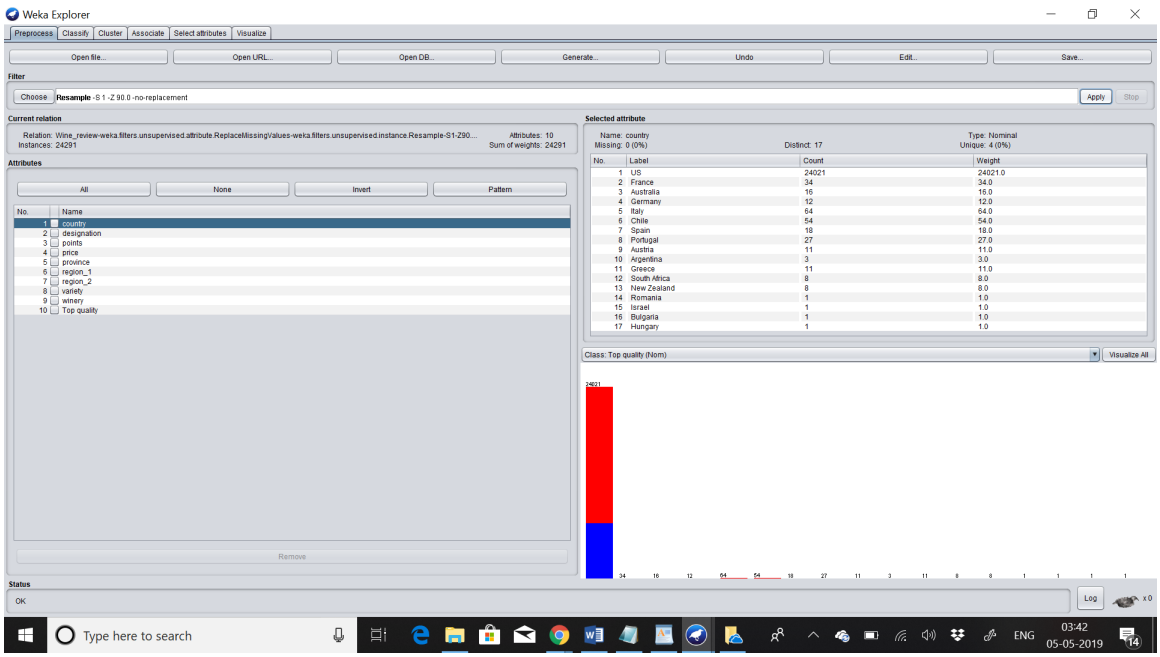
It is saved as **“testing.arff”.**

Screenshot 10 :

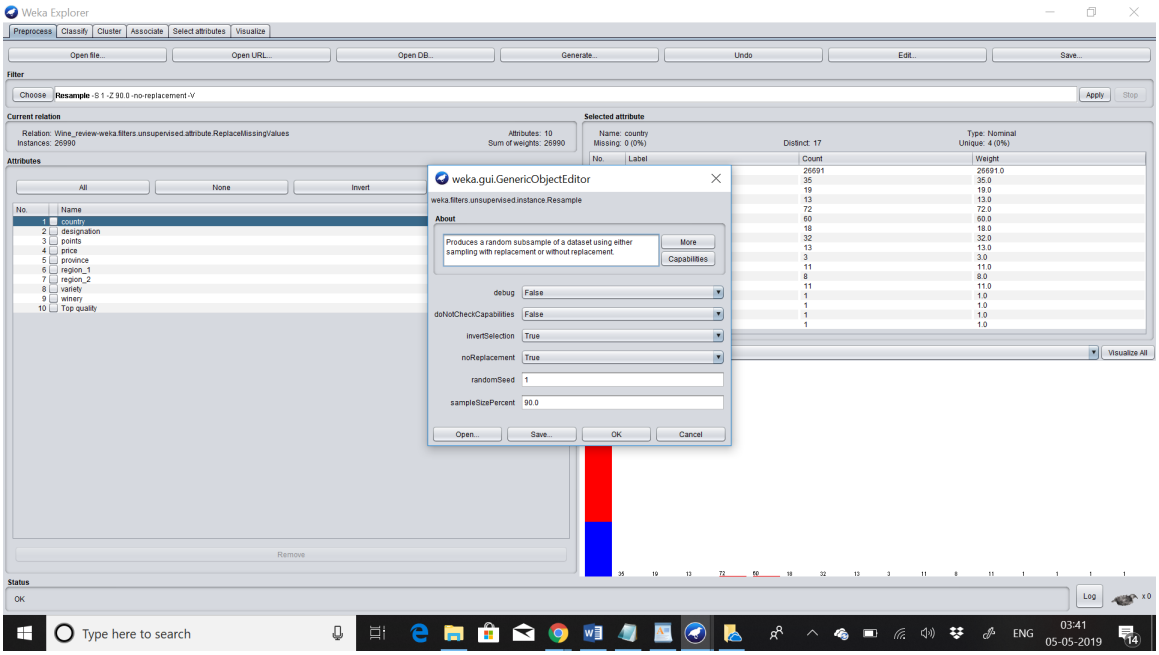
In this screenshot , the **Re-Sample** option is selected in order to split the data into training as well as testing dataset.



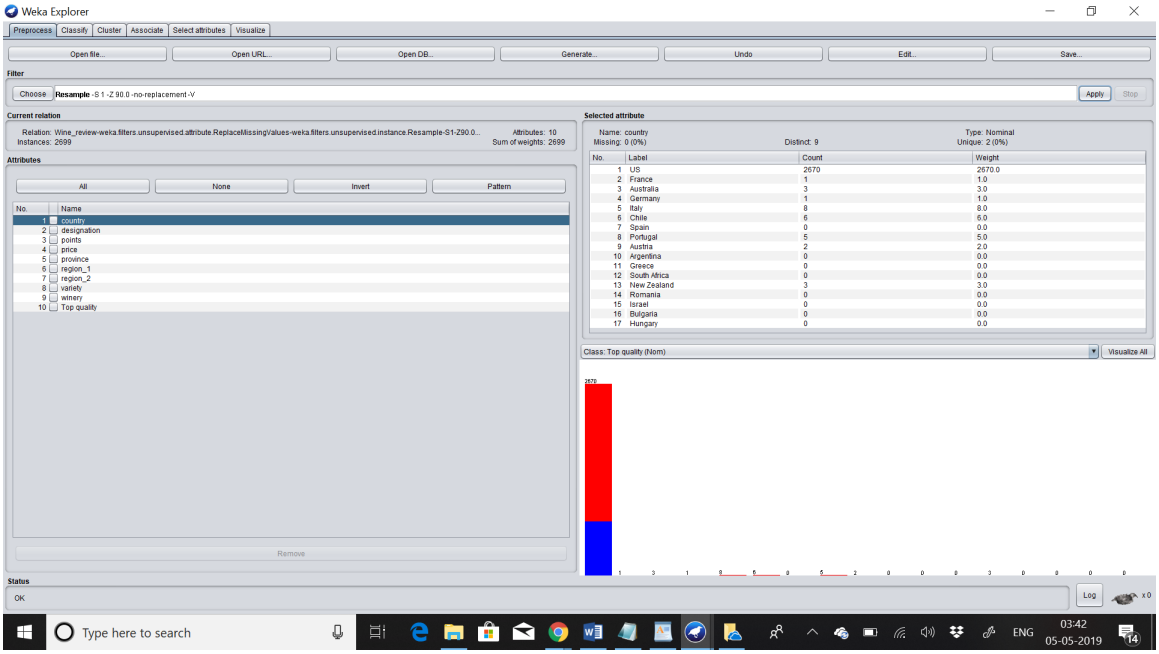
Screenshot 11 : This screenshot represents that 90% of the dataset has been taken for the training set.



Screenshot 12 : This screenshot displays how the testing data is actually divided .



Screenshot 13 : The testing data is splitted and stored as **“testingSet.arff” .**



J48

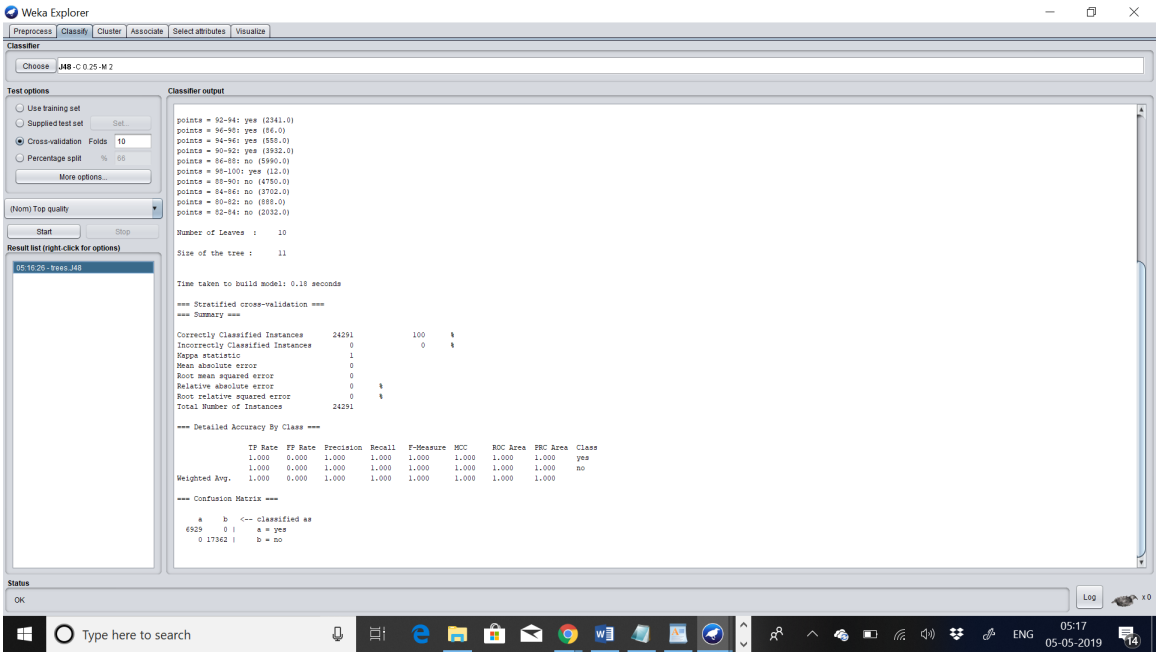
Training set:

The algorithm was run with **10-fold cross-validation.**

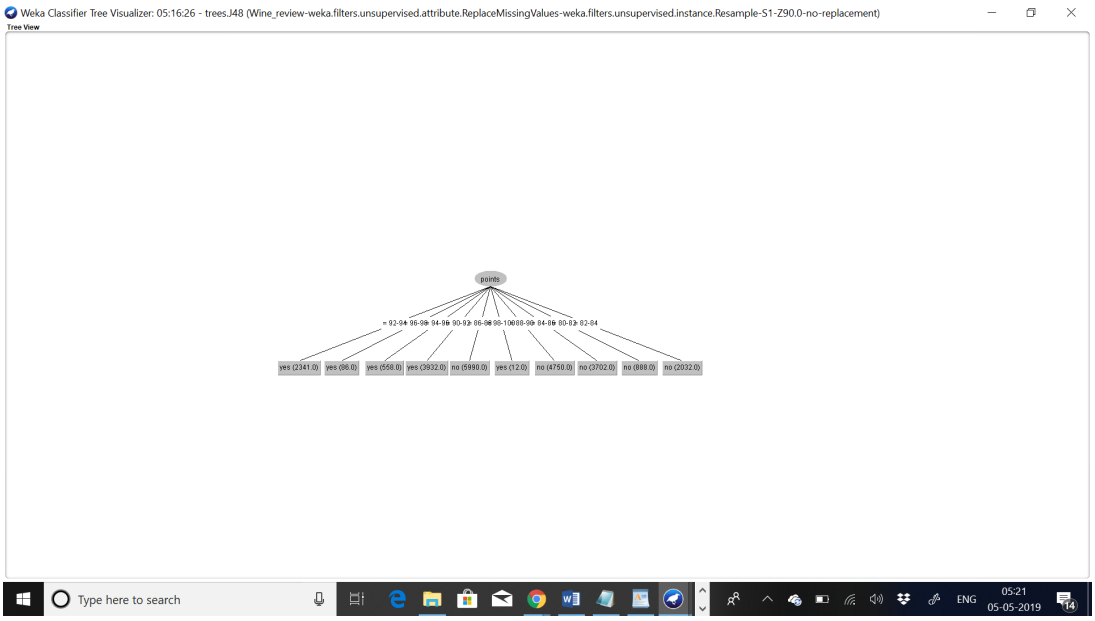
The model achieved a result of 24291/24291 correct or **100%**.

There are **no errors** in the confusion matrix.

Screenshot 14 : This displays how the J48 algorithm is applied to the **training** dataset.



Screenshot 15 : The result buffer displays the tree structure of the j48 algorithm applied to the training dataset.



Testing Set:

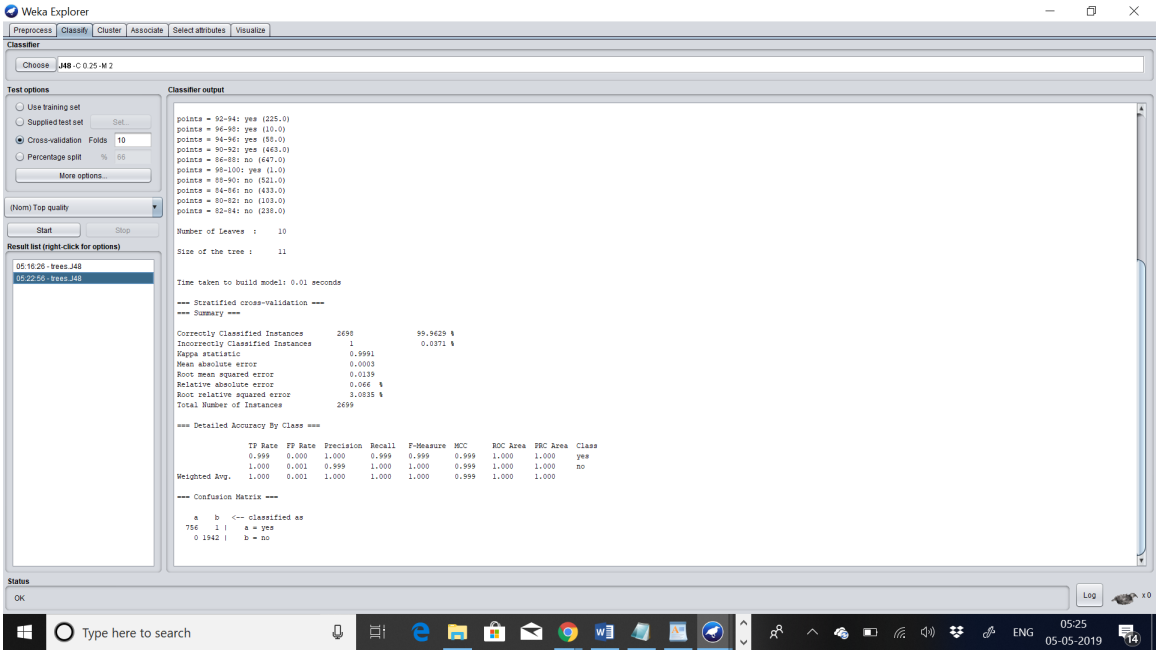
The algorithm was run with **10-fold cross-validation.**

the model achieved a result of 2698/2698 correct or **99.9629%.**

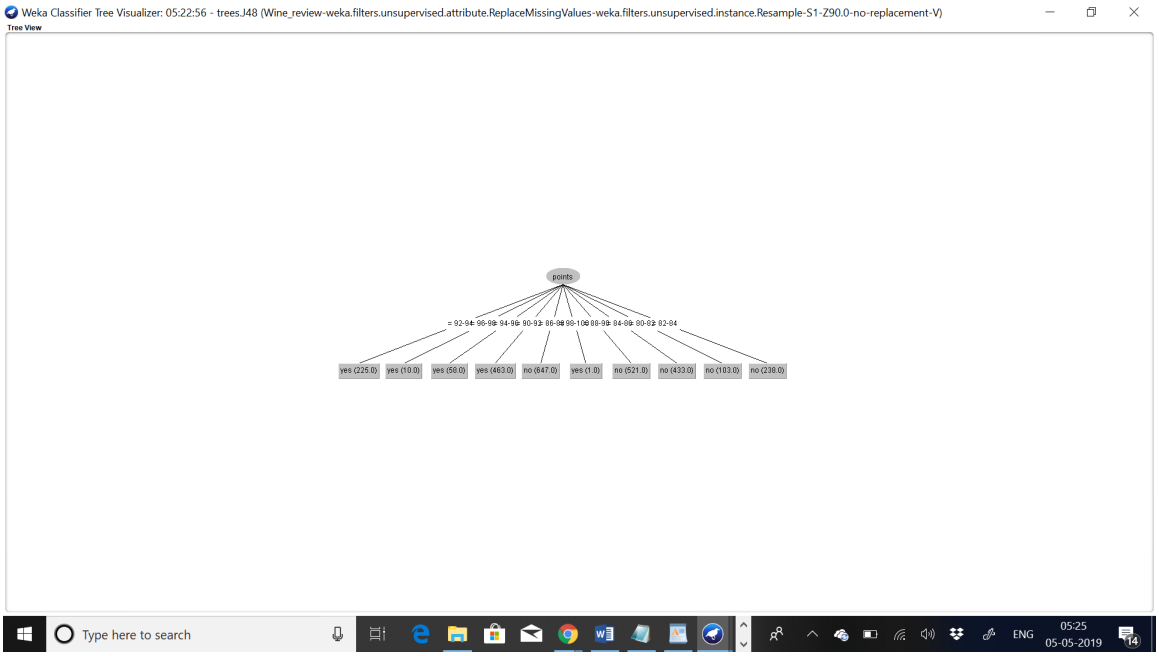
There is one error where one of the normal wine are classified as top quality wine.

Mean absolute error is **0.0003**

Screenshot 16 : This displays how the j48 algorithm is applied to the **testing** dataset.



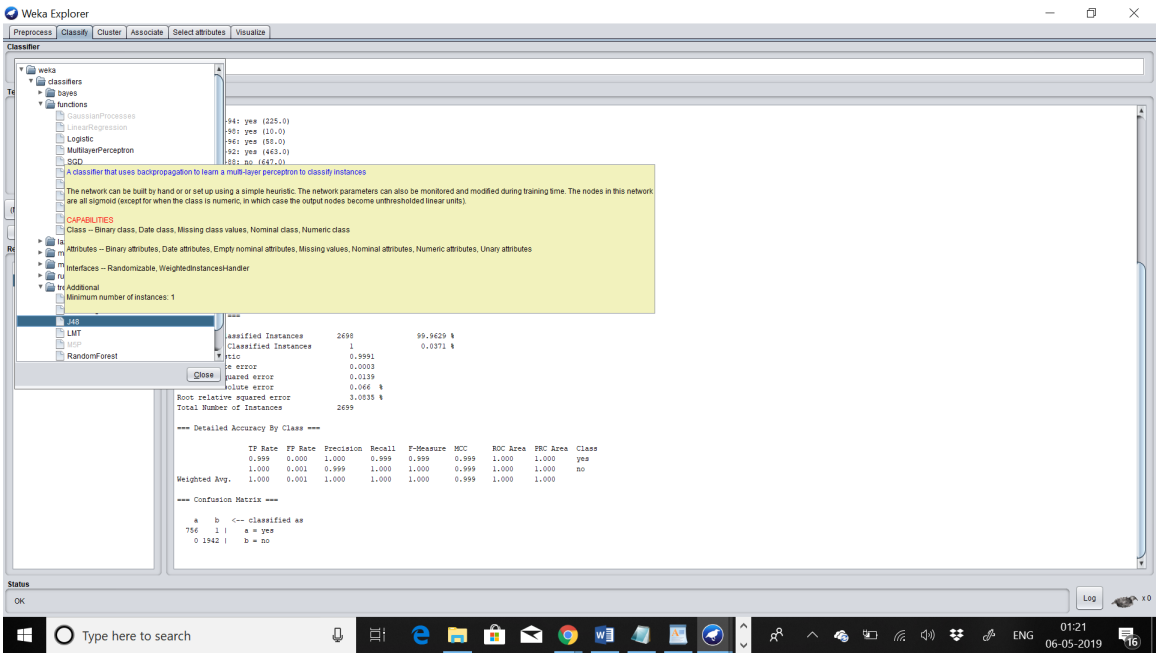
Screenshot 17 : The result buffer displays the tree structure of the j48 algorithm applied to the testing dataset.



V. Multi Layer Perceptron

Multilayer perceptrons are networks of perceptrons, networks of linear classifiers. Weka has a graphical interface that lets you create your own network structure with as many perceptrons and connections as you like.

Screenshot 18 : This screenshot shows how the MLP algorithm is applied to the dataset.

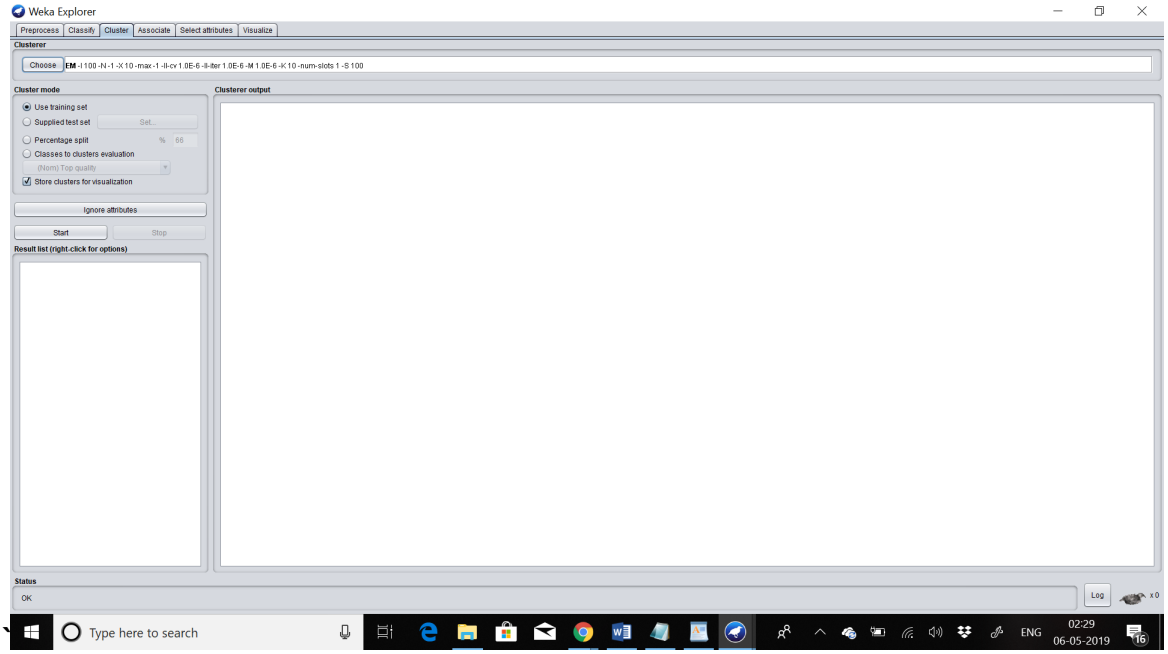


VI. Clustering

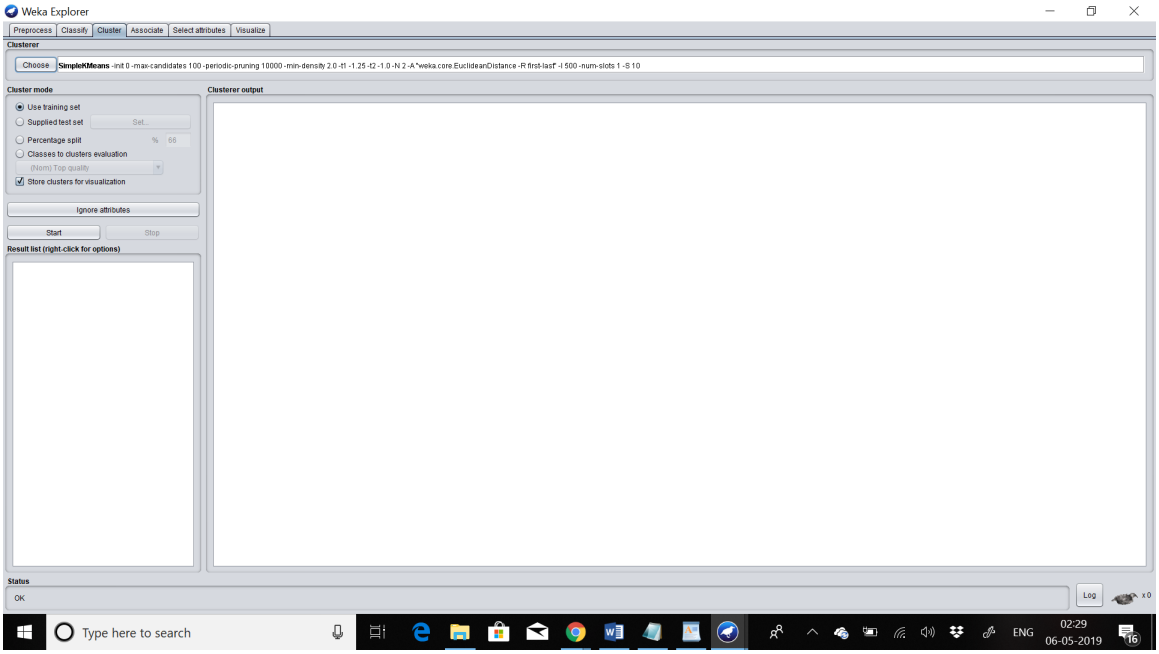
K means:

This technique is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belong to the cluster with nearest mean.The algorithm is called k-means, where k is the number of clusters we want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest. The action in the algorithm centres around finding the k-means

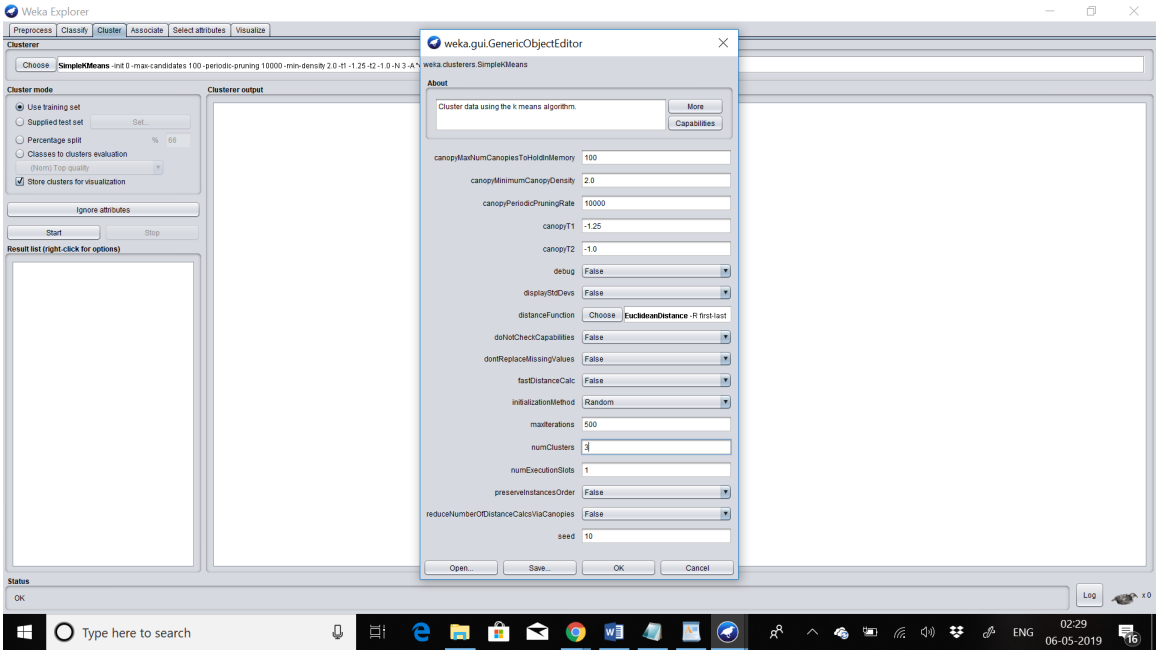
Screenshot 19 : Cluster section to select the Simple K means algorithm.



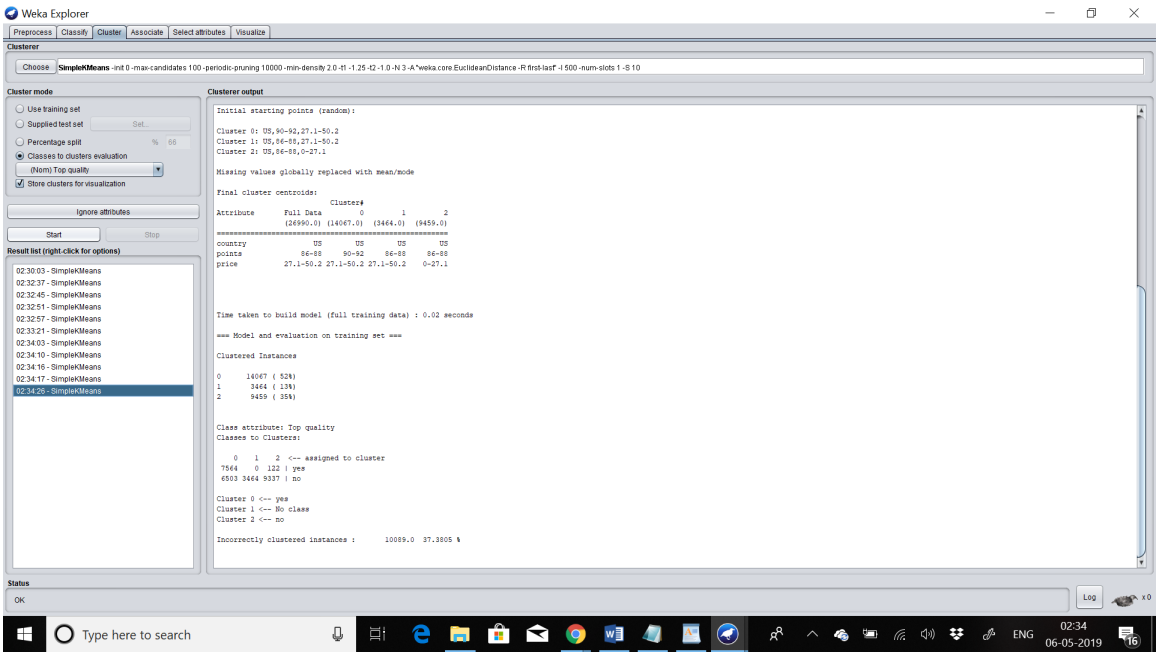
....



Screenshot 21 : Once the Simple K-means algorithm is selected, the number of clusters is set to 3.



Screenshot 22 : This screenshot displays the result buffer for the attribute “country” , “points” and “price”.



Explanation:

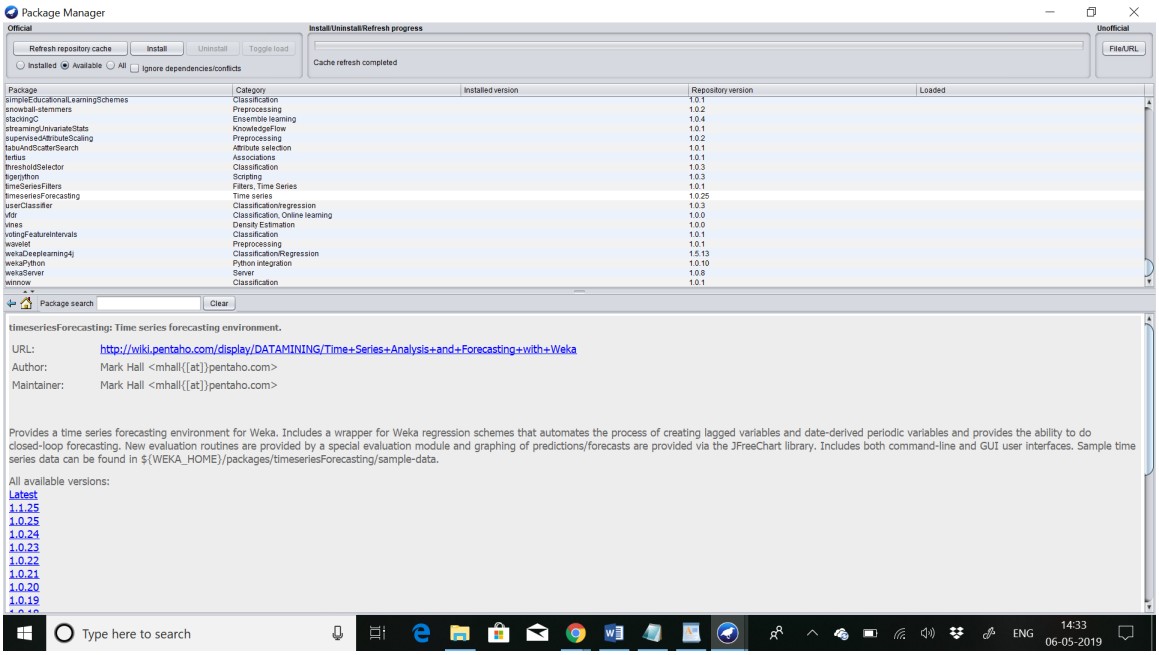
In this k-means algorithm, each cluster is classified in a separate manner. we compute the cluster means again, using the cases that are assigned to the cluster; then, we reclassify all cases based on the new set of means.In this the similar and the dissimilar values are separated and put into different clusters thereby the observations could be enhanced properly to maintain a stable output. We keep repeating this step until cluster means don't change much between successive steps.

VII. Time Series Forecasting – 15%

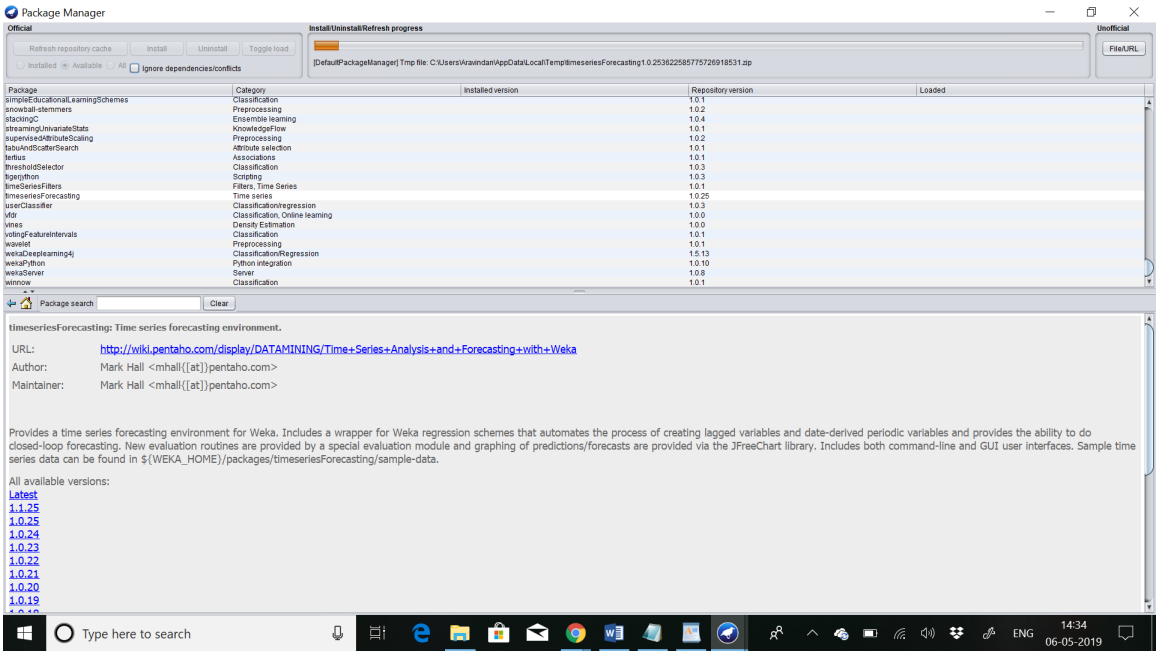
Time does play a role in normal machine learning datasets. Predictions are made for new data when the actual outcome may not be known until some future date. Though the future is being predicted, all prior observations are always treated equally. Time series adds an explicit order dependent dimension . This additional time dimension is both a constraint and a structure that provides a source of additional information.

Our goal is to analyse the Australian beer production dataset , predict what type of time series model may best fit the data and diagnose which model would be most appropriate to forecast beer production in Australia.

Screenshot 23 : Time Series forecasting environment of Weka



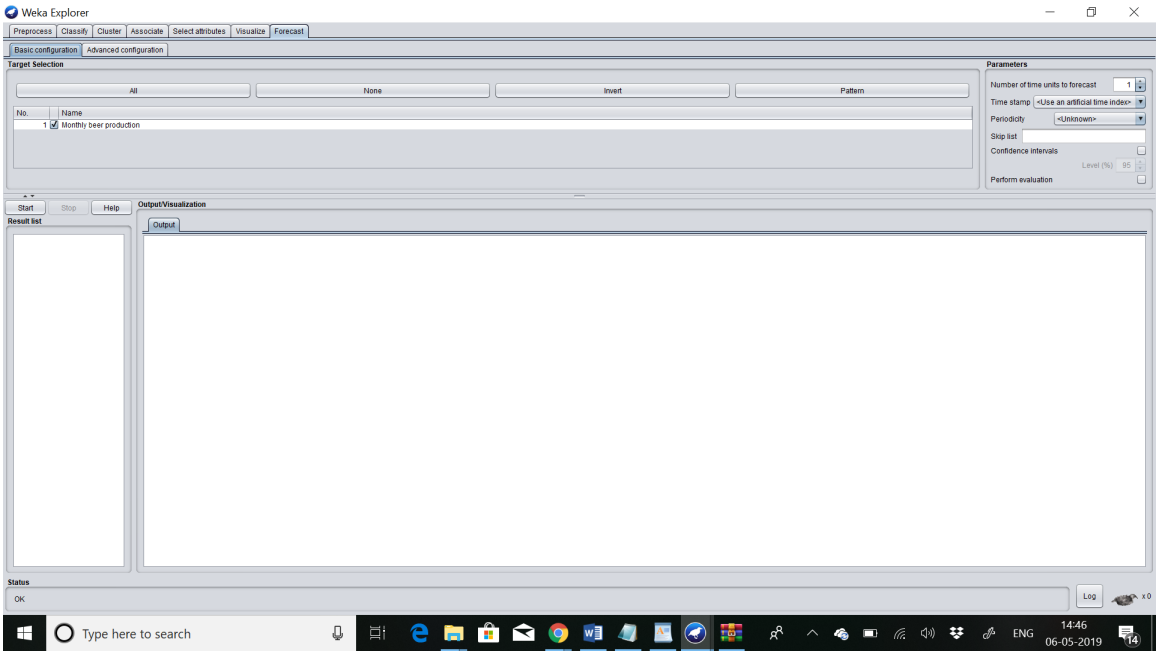
Screenshot 24 : Time Series forecasting environment of Weka



Target Selection :

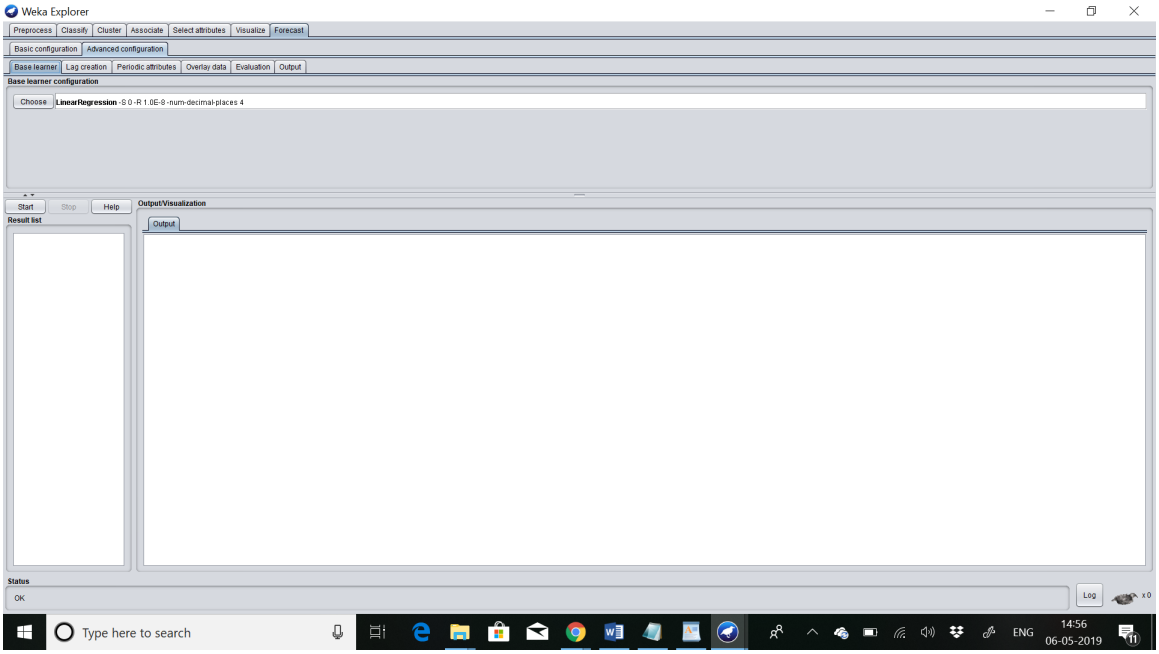
Modelling several series simultaneously can give different results for each series than modelling them individually. When there is only a single target in the data , the system selects it automatically.

Screenshot 25 : Target Selection

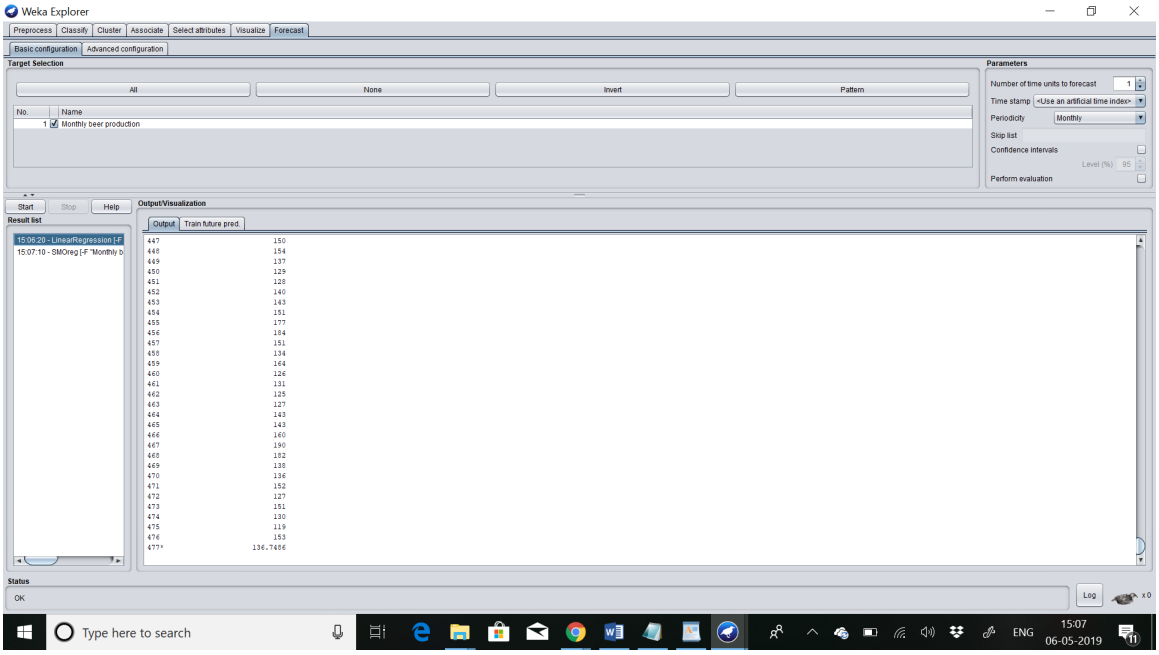


Regression is a **data mining** technique used to predict a range of numeric values (also called continuous values), given a particular dataset. Linear regression is used for finding linear relationship between target and one or more predictors.

Screenshot 26 : Linear Regression

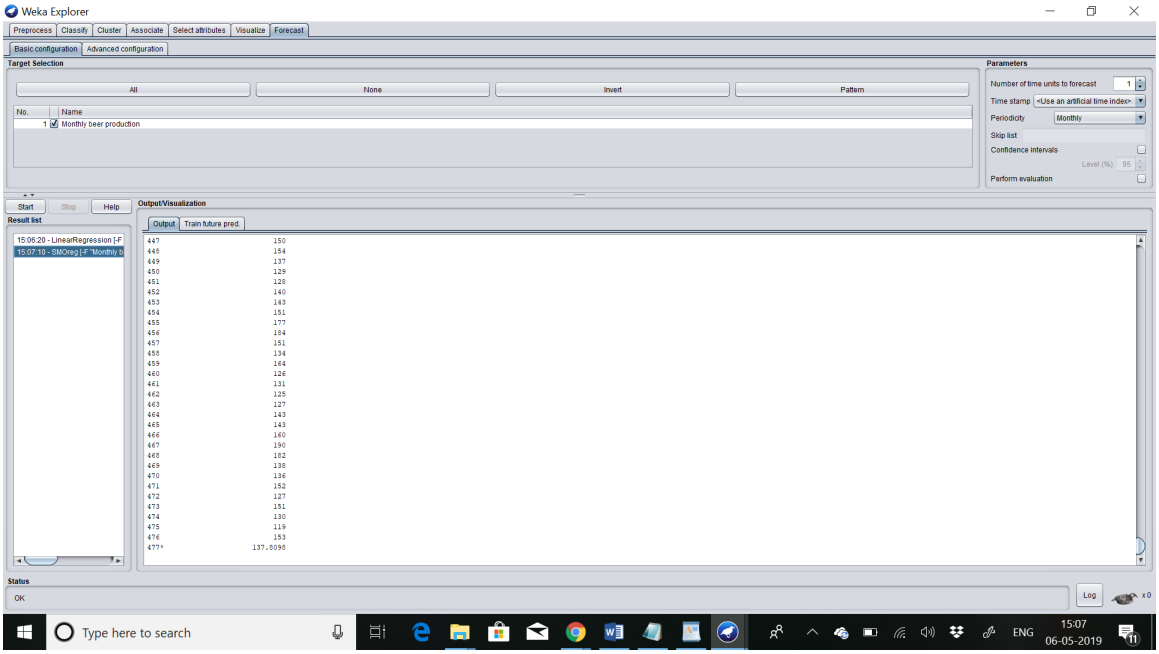


Screenshot 27 : Output / Visualisation

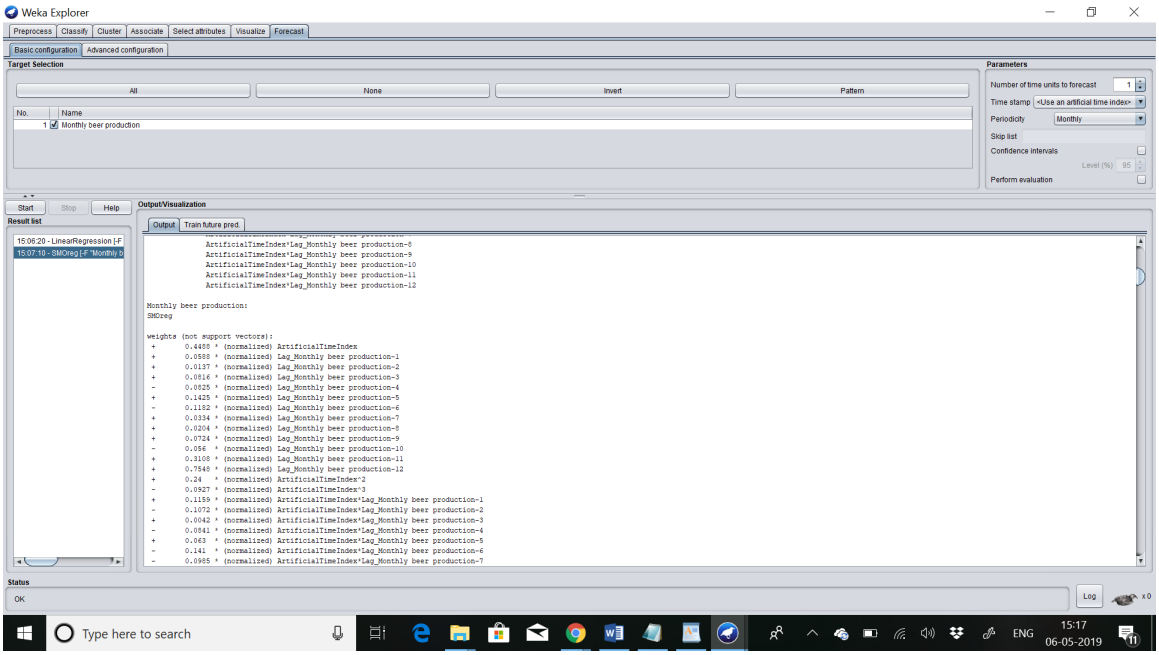


SMOreg implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer.

Screenshot 28 : SMOreg

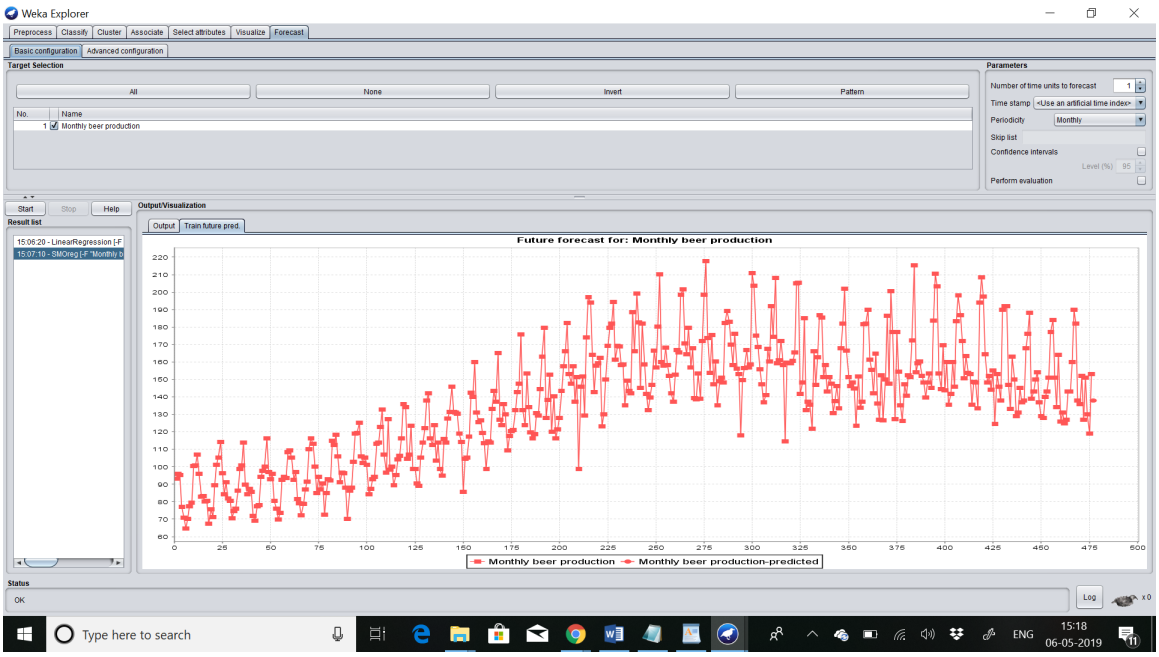


Screenshot 29 : Output / Visualisation



Selecting Output future predictions beyond the end of series will cause the system to output the training data and predicted values (up to the maximum number of time units) beyond the end of the data for all targets predicted by the forecaster.

Screenshot 30 : Future forecast for monthly beer production



VIII. Potential relevance

A couple of research documents on Data Mining Techniques were really helpful for our work namely

Topic 1 : Crime Prediction using Decision Tree (J48) Classification Algorithm (Tarun)

Topic 2 : A Comparative Study of Various Clustering Algorithms in Data Mining (Mukesh).

We have enclosed the necessary research documentations in the zip file.

IX. Division of Labor

We split the work based on our own strengths and weakness as we carried out everything together as a group. In the below table , we have given the percentage of work shared among us. In the end , it was all team work.

Filename / Task	Aravindan	Mukesh	Tarun
Selection of dataset	15	15	15
Cleaning of dataset	25	10	10
Classification	5.53	5.53	5.53
Clustering	5.53	5.53	5.53
Time Series Analysis	5.53	5.53	5.53
Paper selection	10	25	25
Paper Review	33.4%	33.3%	33.3%