DATA WAREHOUSING & DATA MINING

SEMESTER: Spring 2021-2022

FINAL-TERM PROJECT DATA MINING WITH WEKA SUBMITTED BY:

STUDENT NAME	**ID**
Rimon Nath	18-38929-3
Md. Shahadat Ali	18-36092-1
Saila Sharmin Fariha	18-36071-1
Foysal Ahmed Pranto	18-36872-1

SECTION: C

DEPARTMENT: CSE

SUBMITTED TO: COURSE TEACHER: TOHEDUL ISLAM

INTRODUCTION

Data mining is the process of uncovering patterns and other valuable information from large data sets. It is also known as knowledge discovery in data (KDD). Data mining is used in many areas of research and business, including healthcare, education, weather forecasting, sales and marketing, product development, etc. It is a computer science and statistics multidisciplinary topic with the general purpose of extracting information from data collection and transforming the information into an accessible structure for subsequent use. KNN, Naive Bayes, and Decision Tree are some of the classification algorithms used in data mining. I have chosen "Seattle-weather" to classify the weather condition by using two different classifiers and find the best-suited classifier for the dataset. Another part of this dataset's classifications job is to predict which weather might be appropriate for predicting weather conditions with the same symptoms [1].

Supervised learning (SL) is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.[2]

Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that through mimicry, which is an important mode of learning in people, the machine is forced to build a compact internal representation of its world and then generate imaginative content from it.[3] I have chosen "Wholesale customers data" to classify the weather condition by using a k-means classifier for clustering the dataset.

Information about the supervised dataset: In this report, the used "Seattle-weather", a CSV dataset file, collected from Kaggle.com was used to predict the outcome of the condition that might be accurate for the predict the weather[2].

The targeted feature is:

• weather

The other features are:

• date

• precipitation

• temp_min

• temp_max

• wind

About the attribute: The dataset contains 5 attributes and 1 class attribute which is the targeted feature to predict. This class attribute refers to the condition of the weather.

Attribute	Representation in dataset	Data type
Date	dd/mm/yyyy(this formate) Numeric value	Numeric (ratio-scaled)continious
Precipitation(ratio of rainfall in wind)	Numeric value	Numeric (ratio-scaled)
temp_max(maximum temperature ratio)	Numeric value	Numeric (ratio-scaled)
temp_min(minimum temperature ratio)	Numeric value	Numeric (ratio-scaled)
Wind(pressure ration of wind)	Numeric value	Numeric (ratio-scaled)
weather(condition of weather: class attribute)	Drizzle, rain, snow, sun, fog	Class (nominal)

There is a total of 1461 instances of these 6 attributes and all these instances were used for classification. Here are the graphical details of the attributes:

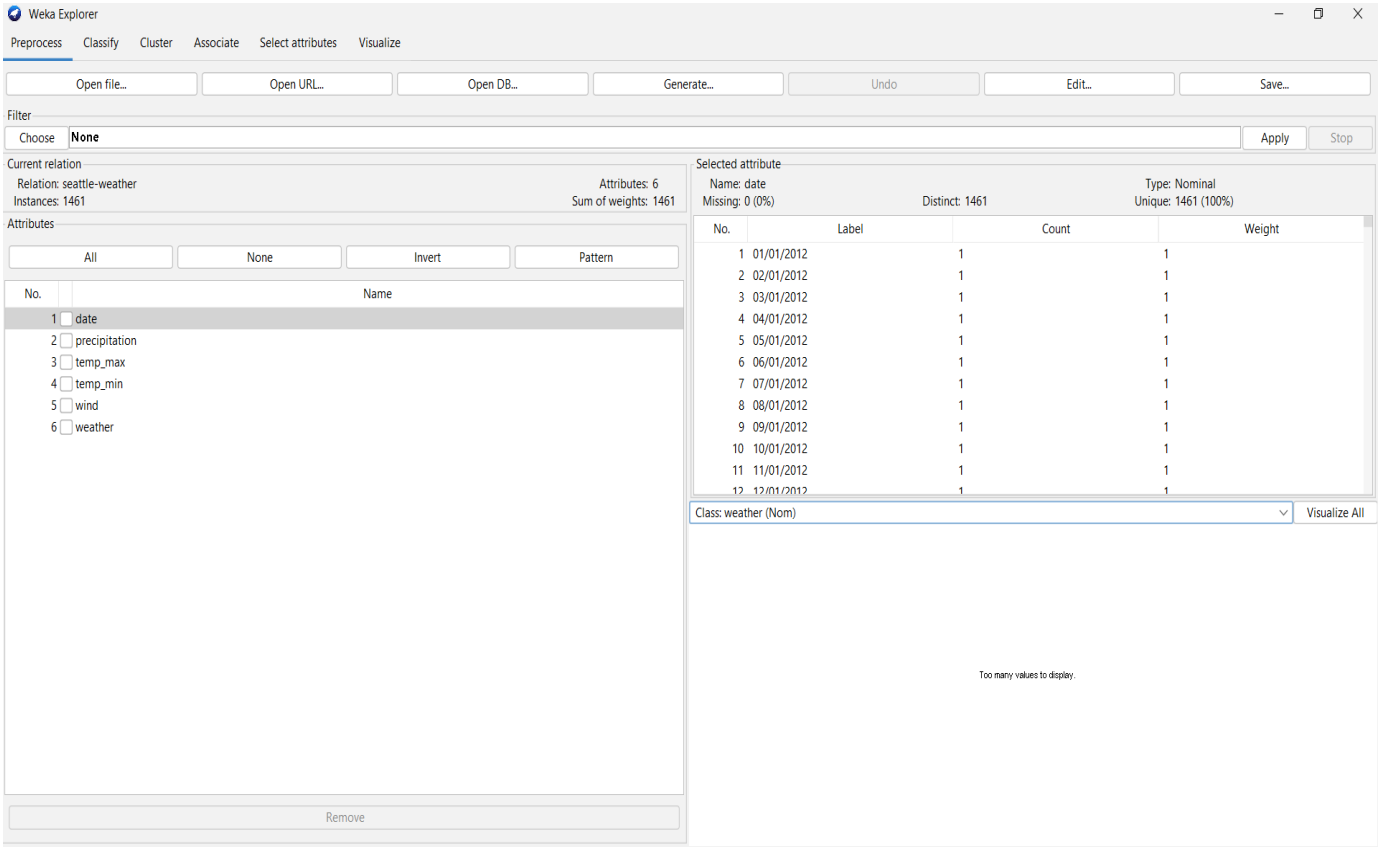

Figure 1: Import Data set

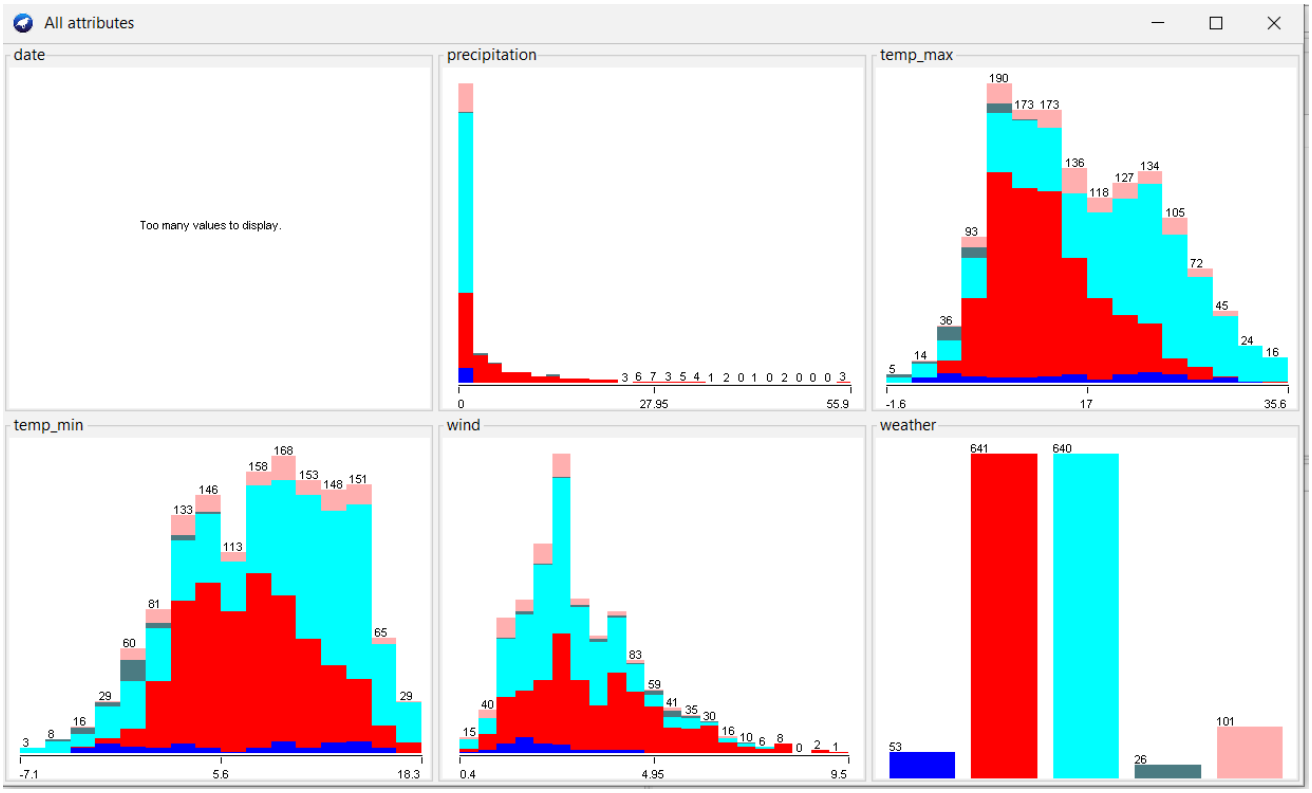

Figure 2: Graphically Represent all attributes

Classifier: A classifier is a machine learning model that is used to discriminate different objects based on certain features. Tow kinds of classification have been used with the same data to compare the result. In this process, Naïve Bayes and K-nearest Neighbour classifiers were used.

RESULTS OF THE CLASSIFIERS: Weka 3.8.5 version software was used to construct the classifier. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Applying naïve Bayes classifier: Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. While classifying the selected dataset, the NaiveBayes format was selected from the Bayes folder.

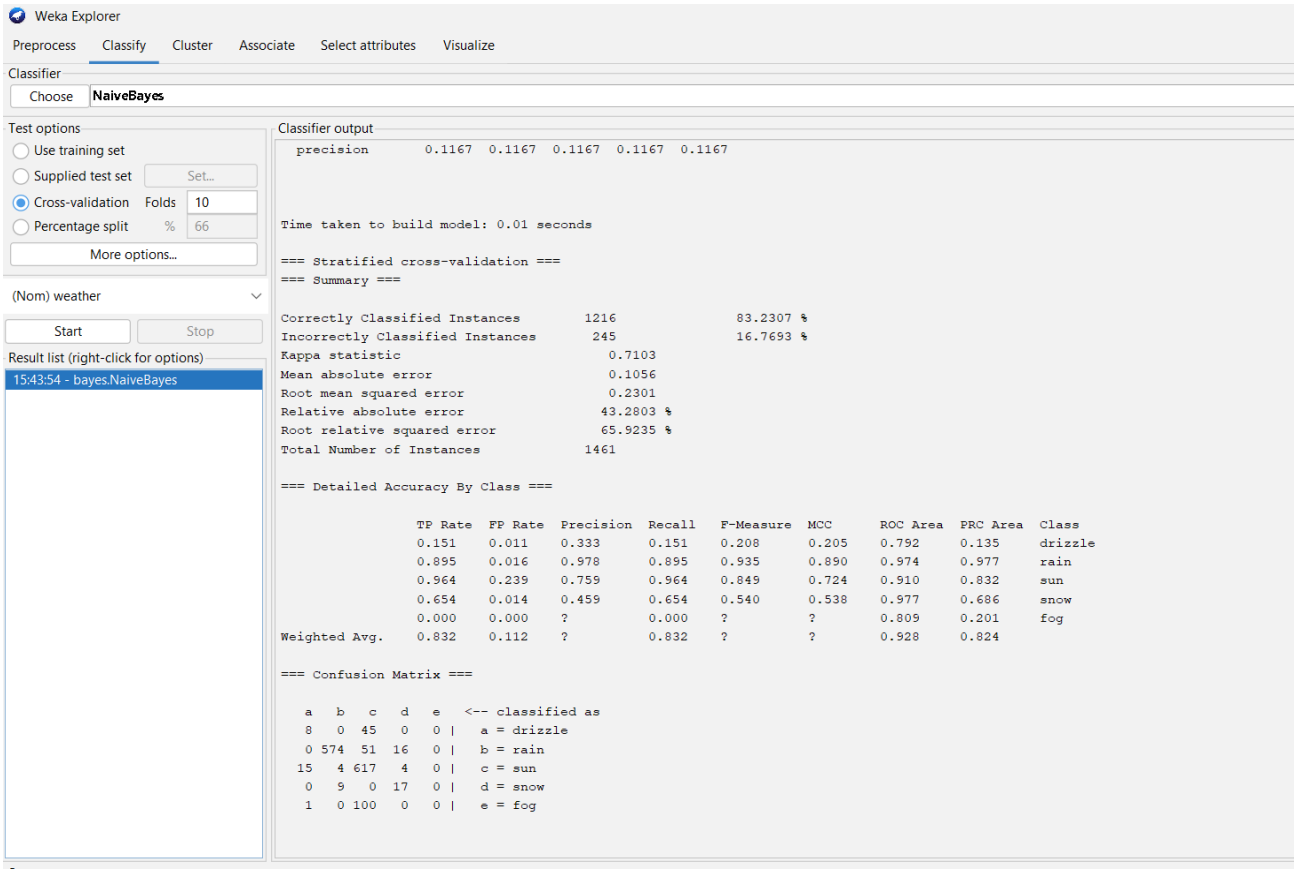

Figure 3: Naïve bayes classification

Applying K-nearest Neighbour classifier: K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. While classifying the selected dataset, the IBk format was selected for KNN classification.

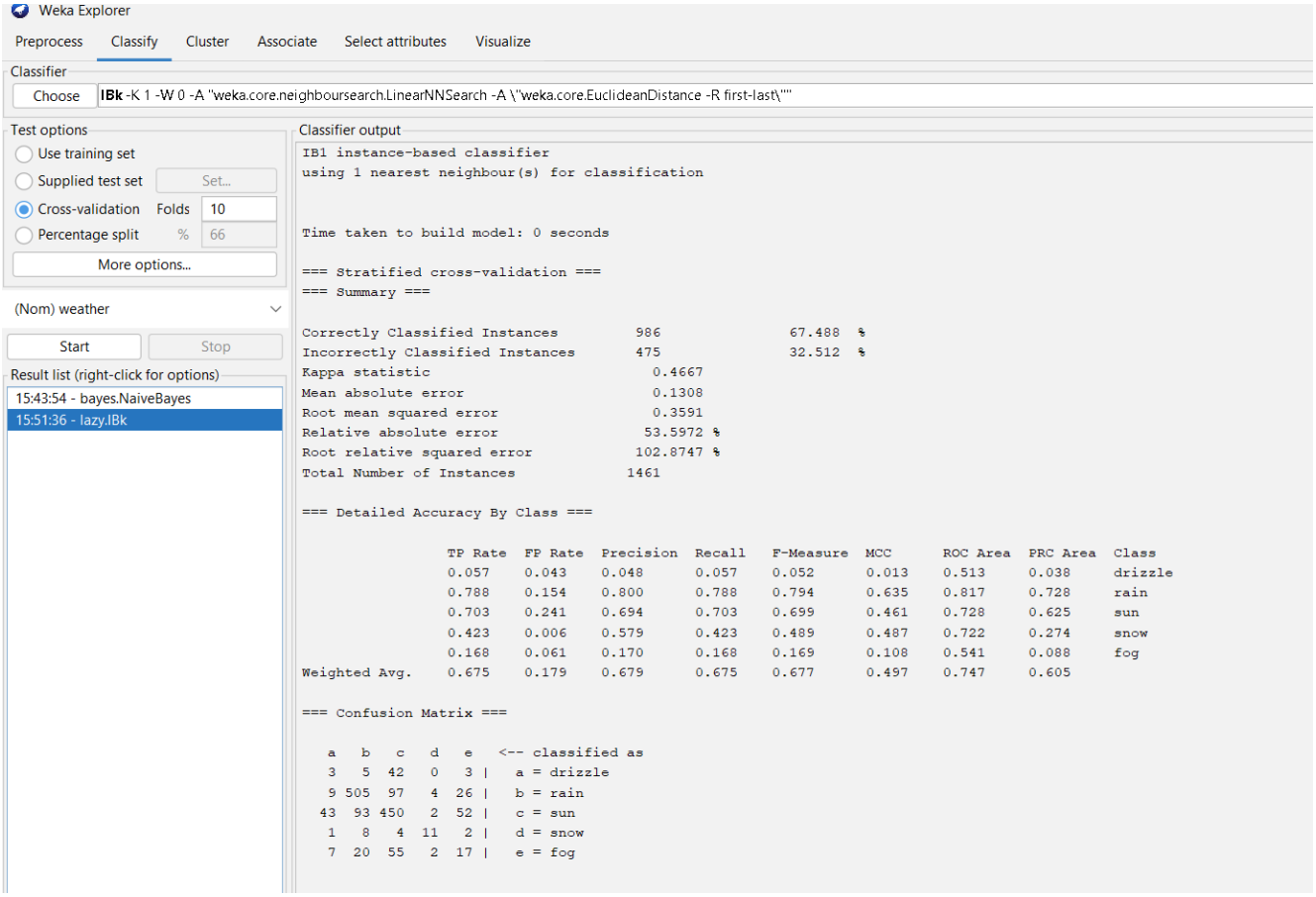

Figure 4: KNN classification

Reason to choose naïve Bayes classifier: From the obtained result of the two classifiers, it is clearly seen that **naïve Bayes** has the highest percentage of correctly classified instances which is 83.2307%. As it has the most accurate value so it would be a more suitable classifier for the dataset. One of the advantages of **naïve Bayes** is that Good results were obtained in most of the cases. Moreover, **naïve Bayes** is faster than KNN due to KNN's expensive real-time execution. It is easy to implement.

Here is the summary of naïve Bayes classifiers result:

=== Summary ===

- Correctly Classified Instances 1216 83.2307 %
- Incorrectly Classified Instances 245 16.7693 %
- Kappa statistic 0.7103
- Mean absolute error 0.1056
- Root mean squared error 0.2301
- Relative absolute error 43.2803 %
- Root relative squared error 65.9235 %
- Total Number of Instances 1461

PREPARING TEST-DATASET: One of the most fundamental methods in machine learning is to train our algorithm on a different and unique training set from the test set for which its correctness will be measured. A training dataset was created using a subset of the referred dataset in order to detect machine learning behavior. The model was then put to the test with a test dataset, which is a subset of the training dataset. Things that were ensured when constructing the test dataset were that it was large enough to give statistically relevant findings. It was also indicative of the entire data set. To put it another way, the test set that differed from the training set was not chosen. The suitable classifier is then used to predict the classification for the instances in the test set.

If the test set contains N instances of which C are correctly classified, C is correctly classified Predictive accuracy, P = C/N. There are 91 instances in this prepared test dataset.

date	precipitati	temp_max	temp_min	wind	weather
01/01/2012	0	12.8	5	4.7	drizzle
02/01/2012	10.9	10.6	2.8	4.5	rain
03/01/2012	0.8	11.7	7.2	2.3	rain
04/01/2012	20.3	12.2	5.6	4.7	rain
05/01/2012	1.3	8.9	2.8	6.1	rain
06/01/2012	2.5	4.4	2.2	2.2	rain
07/01/2012	0	7.2	2.8	2.3	rain
08/01/2012	0	10	2.8	2	sun
09/01/2012	4.3	9.4	5	3.4	rain
10/01/2012	1	6.1	0.6	3.4	rain
11/01/2012	0	6.1	-1.1	5.1	sun
12/01/2012	0	6.1	-1.7	1.9	sun
13/01/2012	0	5	-2.8	1.3	sun
14/01/2012	4.1	4.4	0.6	5.3	snow
15/01/2012	5.3	1.1	-3.3	3.2	snow
16/01/2012	2.5	1.7	-2.8	5	snow
17/01/2012	8.1	3.3	0	5.6	snow
18/01/2012	19.8	0	-2.8	5	snow
19/01/2012	15.2	-1.1	-2.8	1.6	snow
20/01/2012	13.5	7.2	-1.1	2.3	snow
21/01/2012	3	8.3	3.3	8.2	rain
22/01/2012	6.1	6.7	2.2	4.8	rain
23/01/2012	0	8.3	1.1	3.6	rain
24/01/2012	8.6	10	2.2	5.1	rain
25/01/2012	8.1	8.9	4.4	5.4	rain
26/01/2012	4.8	8.9	1.1	4.8	rain
27/01/2012	0	6.7	-2.2	1.4	drizzle
28/01/2012	0	6.7	0.6	2.2	rain

Figure 5: Prepare Test Data set

PROCEDURE OF TESTING THE TEST DATASET:

1. The open file option was selected and the extracted CSV file and training dataset was selected from the device.

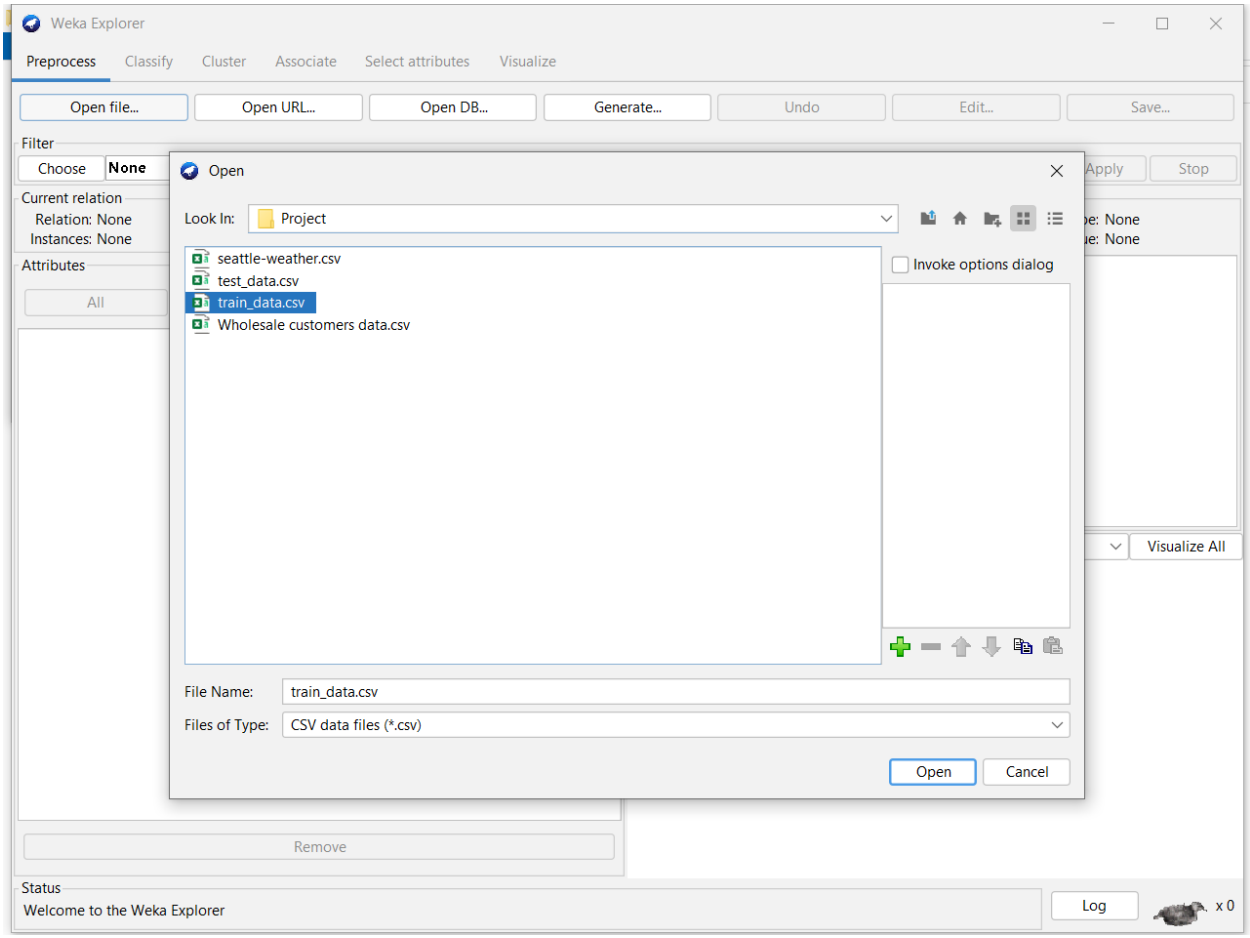

Figure: Training data select

2. **After the open option was clicked, the details of the dataset popped.**

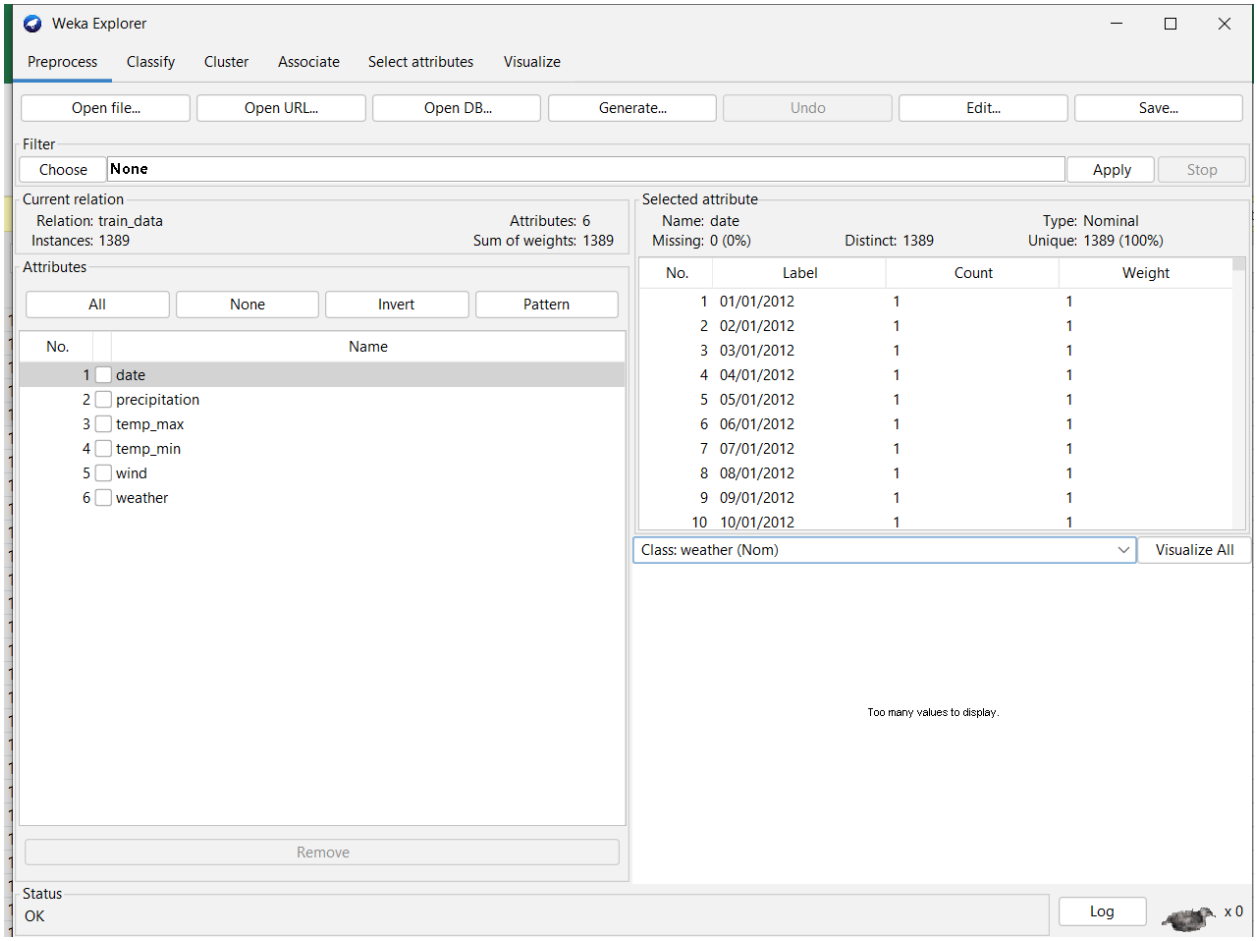

Figure: Train Data set Details

3. Then the preferred classifier (**Naïve Bayes**) for the training dataset was selected. Then from the test options, the use training set was chosen and the start option was selected.

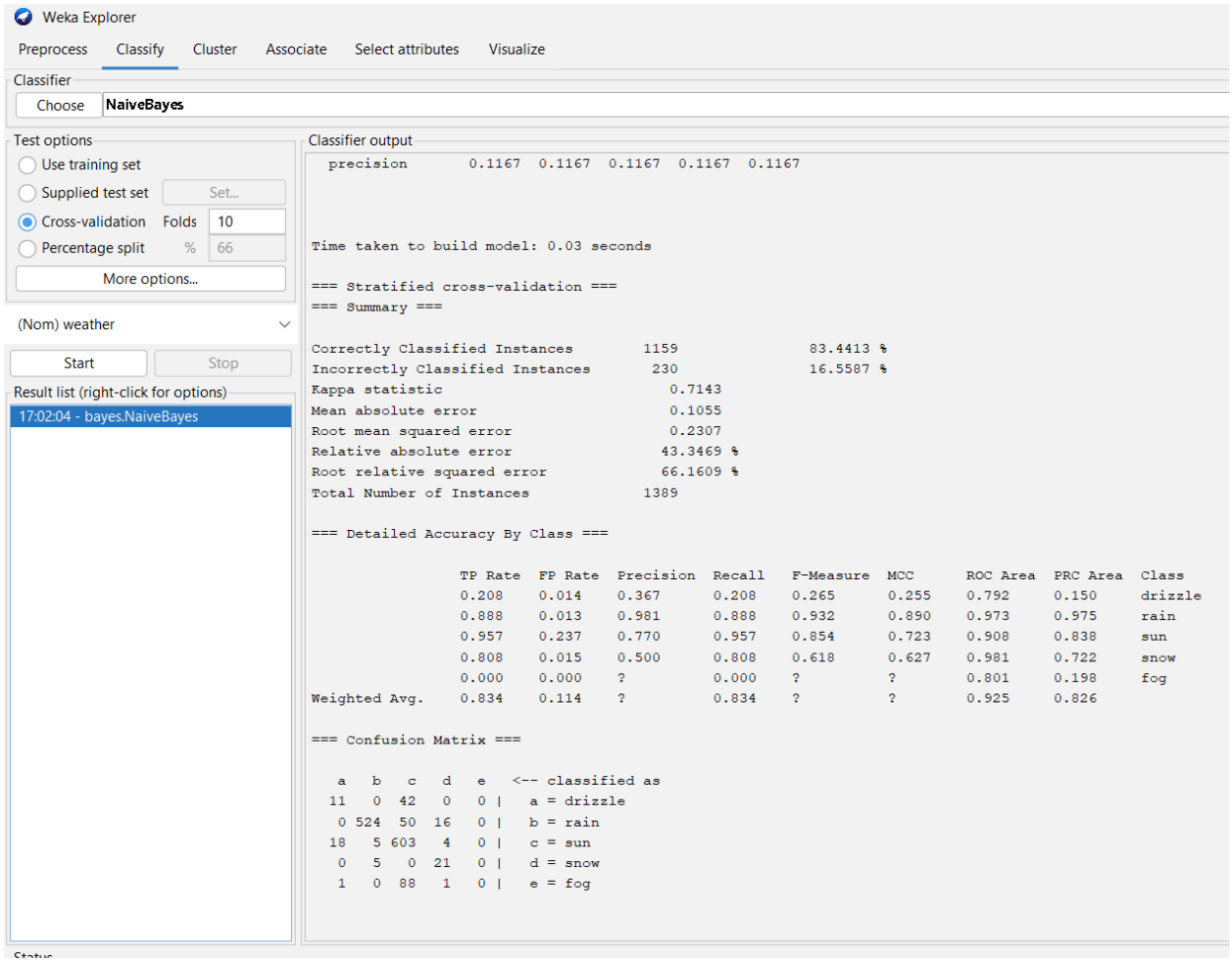

Figure: Result summary of the Train Data Set

4. Import Test Set

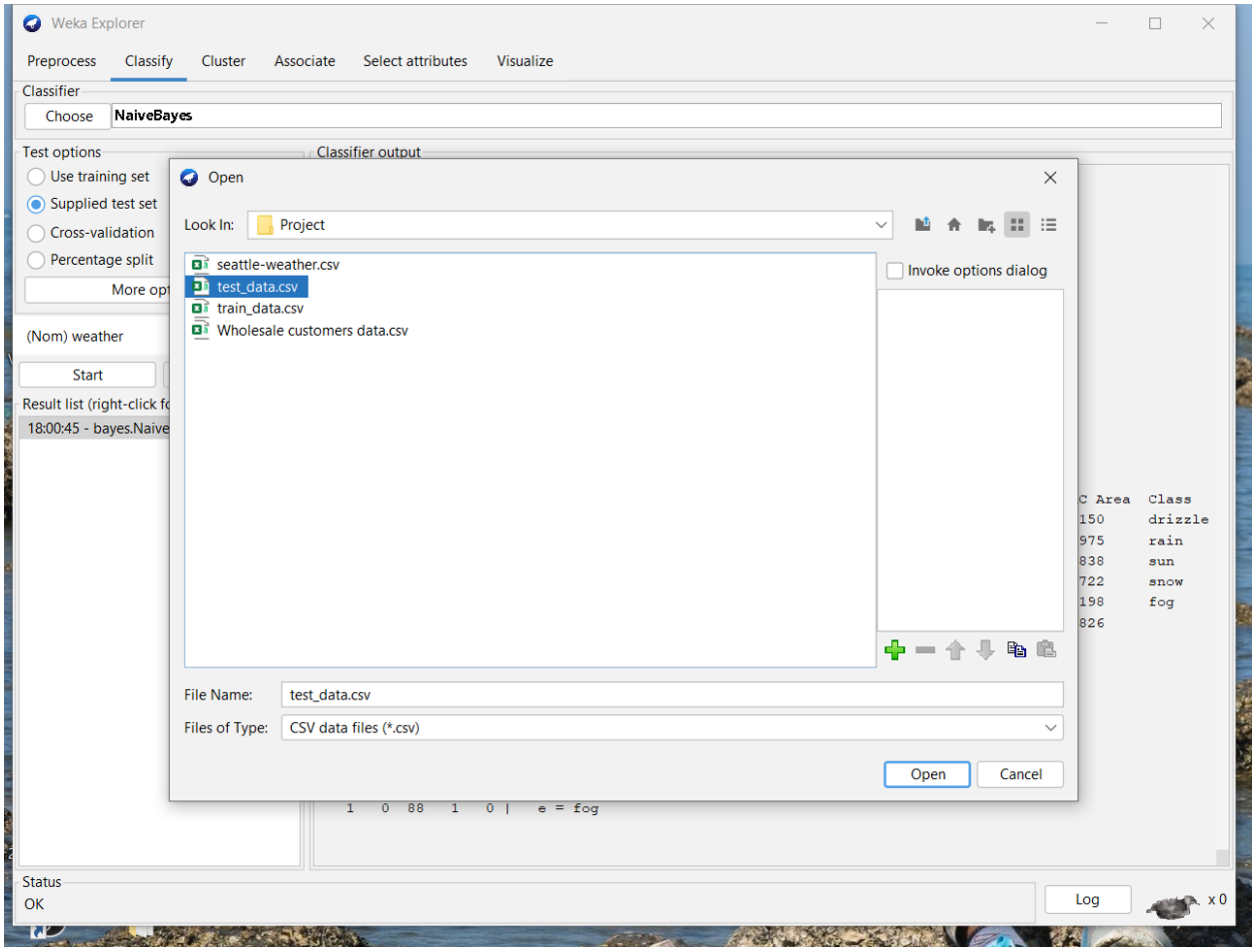

Figure: Import Test Data set

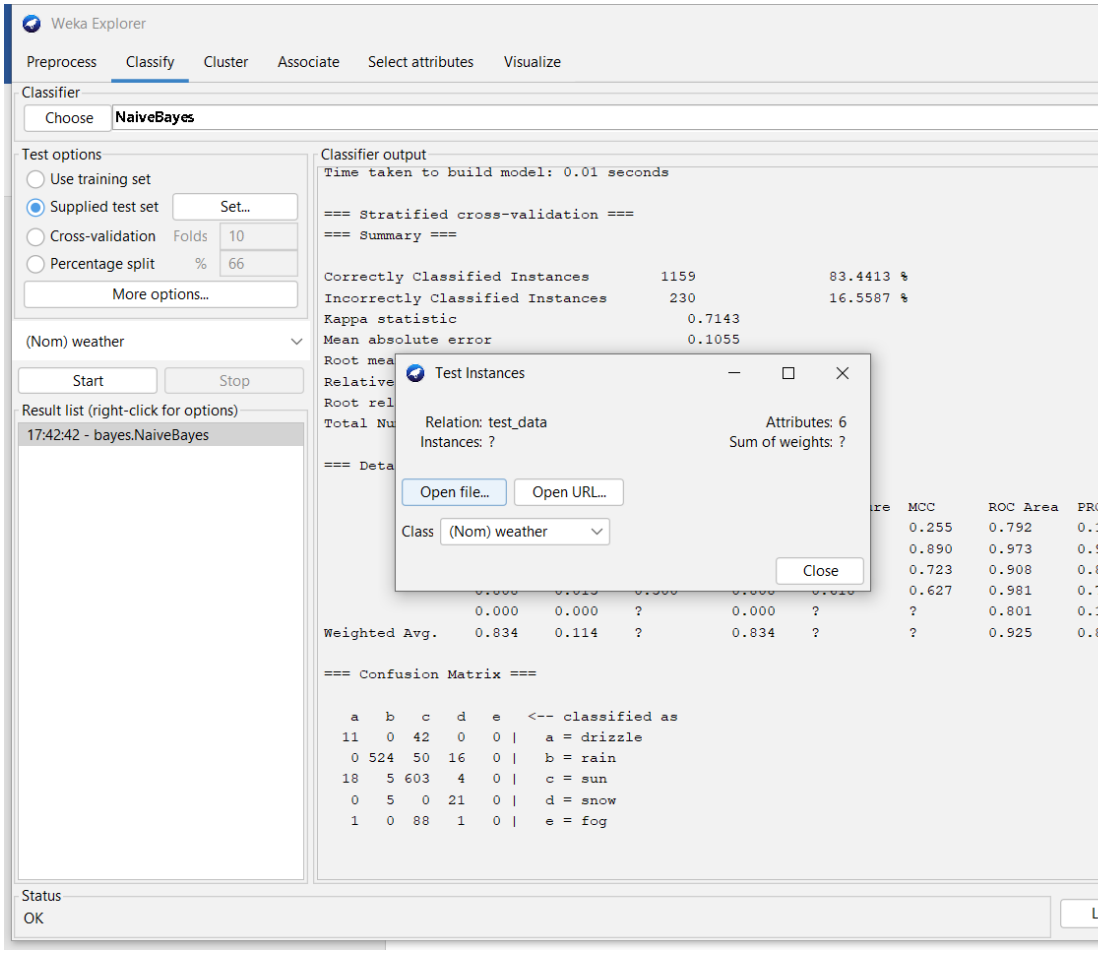

Figure: Open Test Set

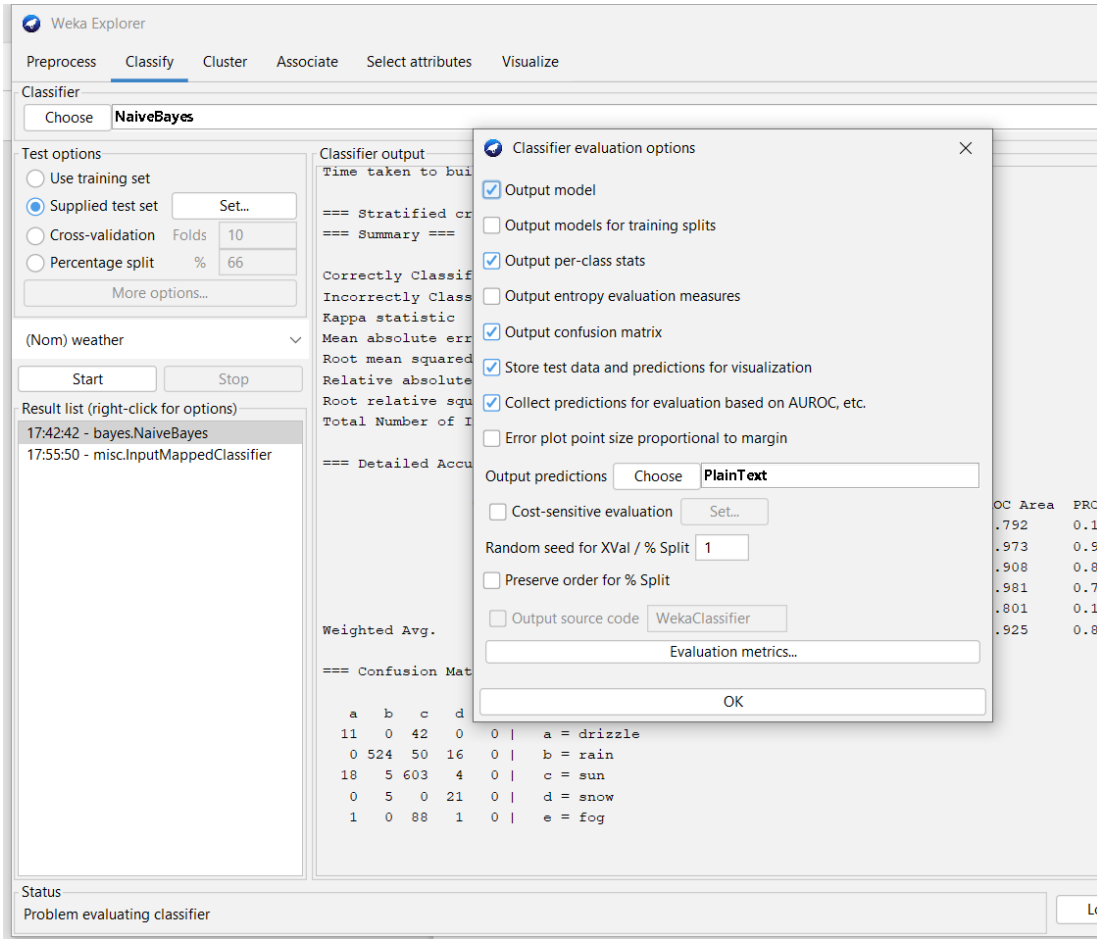

Figure: Plan Text Formate

5. Run the test data set

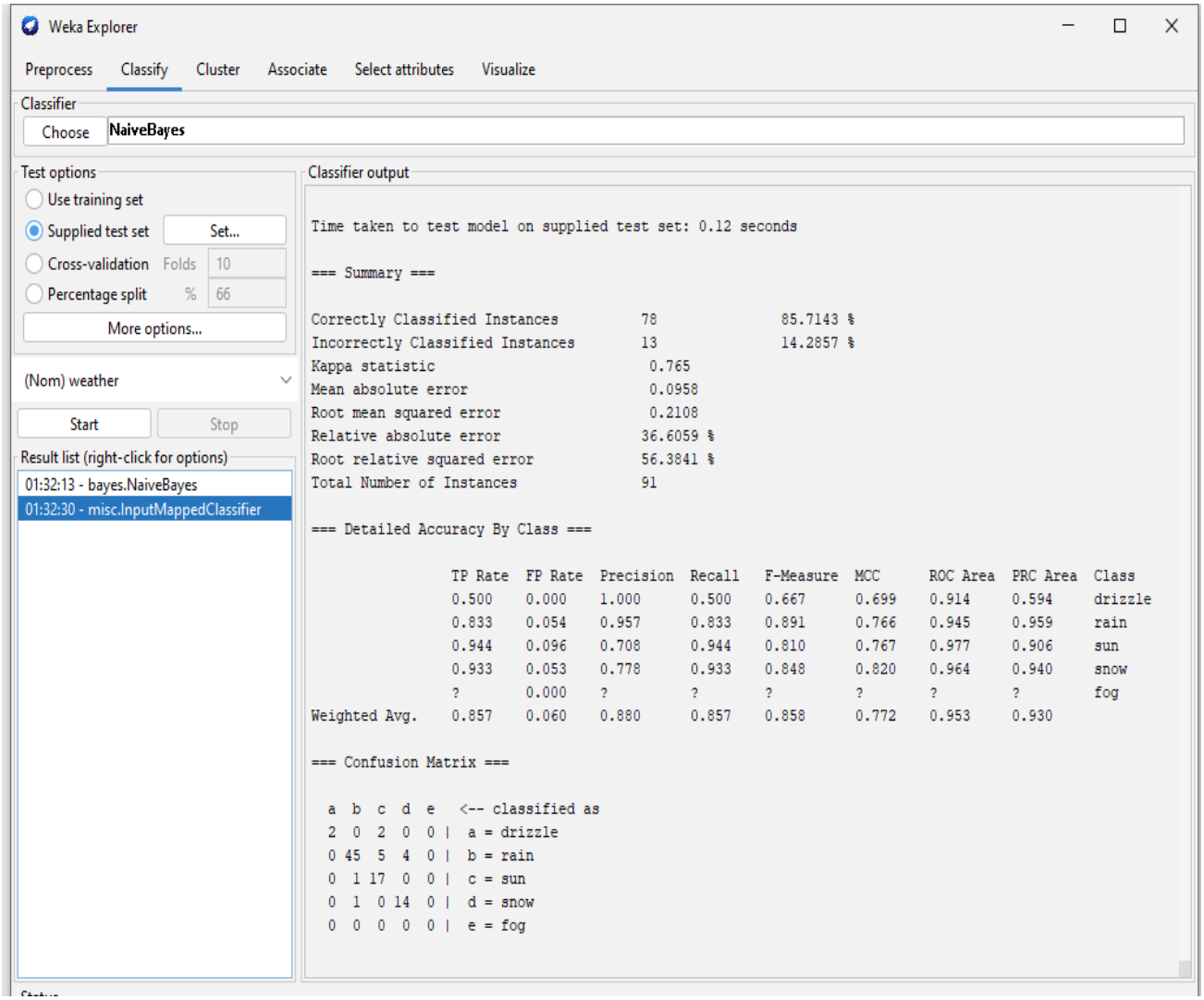

Figure: Test Set Run

RESULT OF Supervised TEST-DATASET MODEL: Once the start button was clicked, the output for the test dataset came. In the result, the test mode was 'user-supplied test set' which means the classifier is evaluated on how well it predicts the class of a set of instances loaded from a file that was inputted by the user. The total time taken to build the model was 0.12 seconds and among the 91 instances, there were 13 instances where the error was found. This means in the test model, those 13 instances were not properly classified and the Machine learning model predicted it after finding the error. In this model, the Correctly classified instances 85.7143 % Value describes the amount of accuracy of correctly classified instances provides by the algorithm. In this case, the percentage is 85.7143% which is quite good. Incorrectly classified instances 14.2857% Value describes how much incorrect instances are given by the algorithm. In this case, the percentage is 9%.

Mean Absolute Error (MAE): It can define as a statistical measure of how far an estimates e from actual values i.e. the average of the absolute magnitude of the individual errors. It is usually similar in magnitude but slightly smaller than the root means squared error. In this model, the MAE is 0.218

Root Mean-Squared Error (RMSE): The Root Mean Square Error (RMSE) calculates the differences between values predicted by a model / an estimator and the values observed from the thing being modeled/ estimated. RMSE is used to measure the accuracy. It is ideal if it is small. In this case, the RMSE is 0.2108 which is ideal.

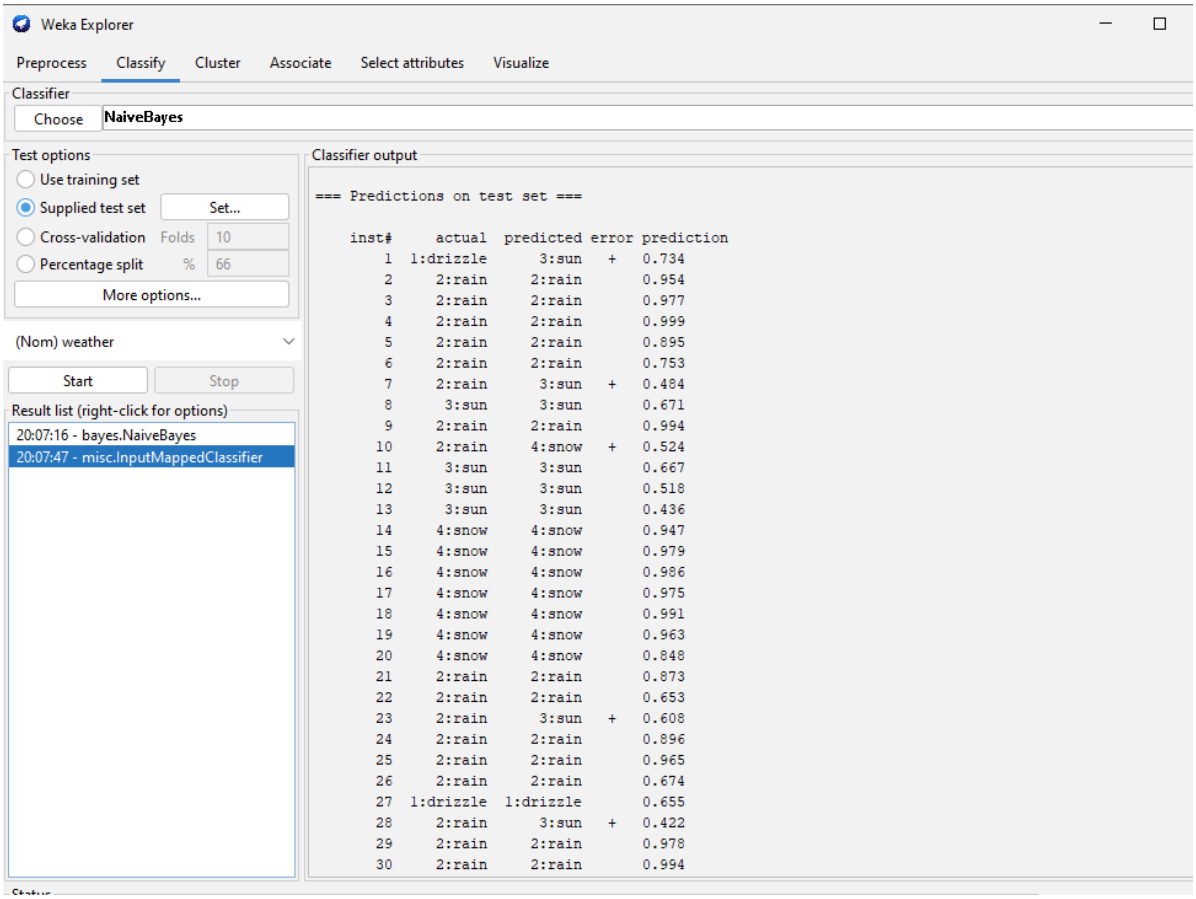

Figure: Predict The test data

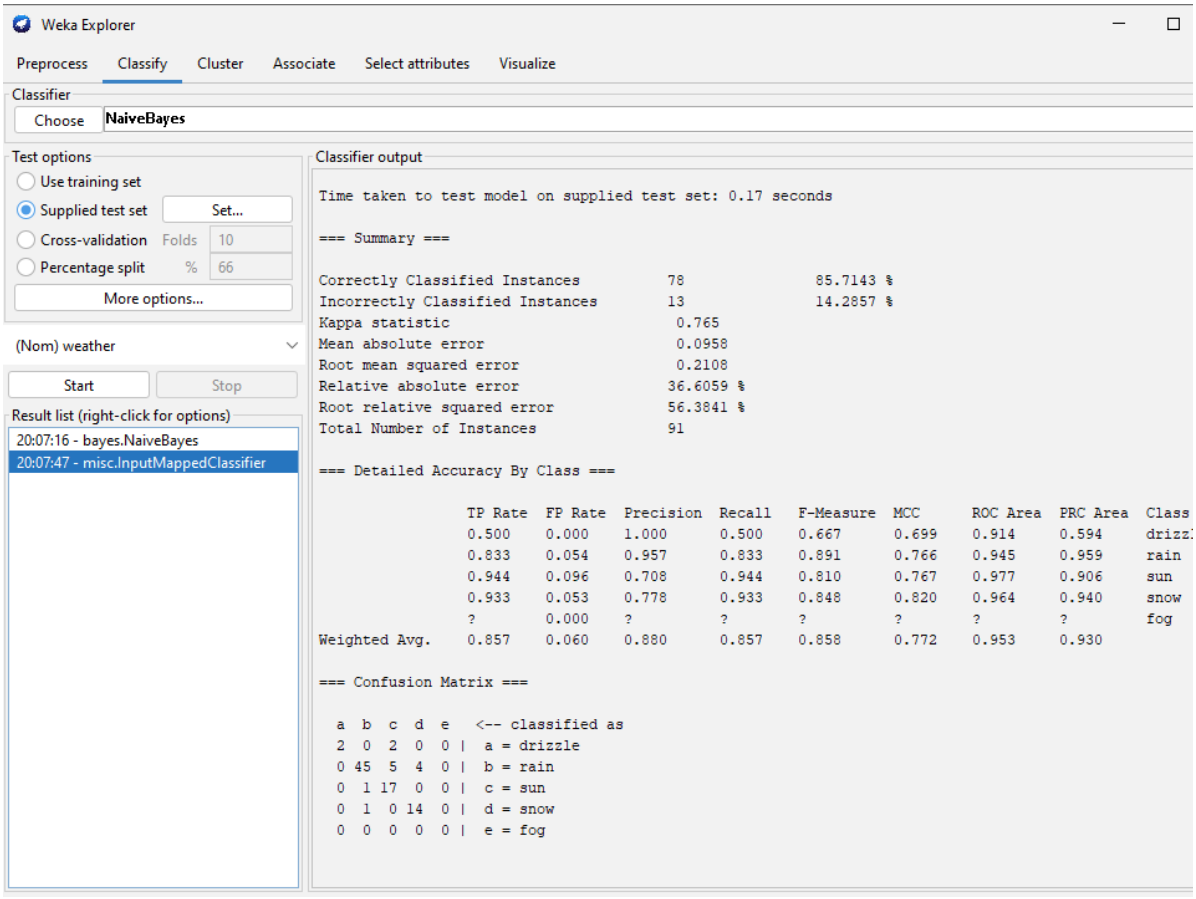

Figure: Result summary of Test Data Set

Information about the unsupervised dataset: In this report, the used "Wholesale customers data", a CSV dataset file, collected from Kaggle.com was used to predict the outcome of the condition that might be accurate for the predict the weather[4].

The features are:

- **Channel**
- **Region**
- **Fresh**
- **Milk**
- **Grocery**
- **Frozen**
- **Detergents_Paper**
- **Delicassen**

About the attribute: The dataset contains 8 attributes.

Attribute	Representation in dataset	Data type
Channel	Numeric value	Numeric Type
Region	Numeric value	Numeric Type
Fresh	Numeric value	Numeric Type

Milk	Numeric value	Numeric Type
Grocery	Numeric value	Numeric Type
Frozen	Numeric value	Numeric Type
Detergents_Paper	Numeric value	Numeric Type
Delicassen	Numeric value	Numeric Type

There is a total of 440 instances of these 8 attributes and all these instances were used for classification:

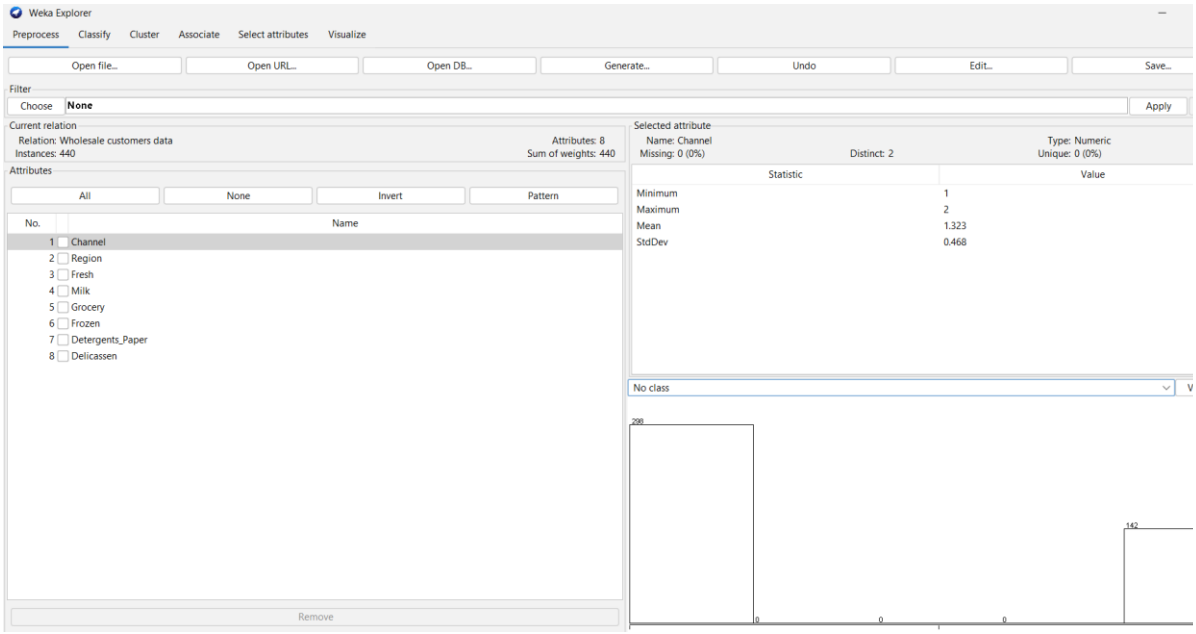

Figure: Unsupervised Data Imported

Classifier: A classifier is a machine learning model that is used to discriminate different objects based on certain features. In this process, k-means clustering was used.

Applying K-means Clustering: The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means. Here k value is 2.

RESULTS OF THE K-Means Clustering: Weka 3.8.5 version software was used to construct the classifier. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

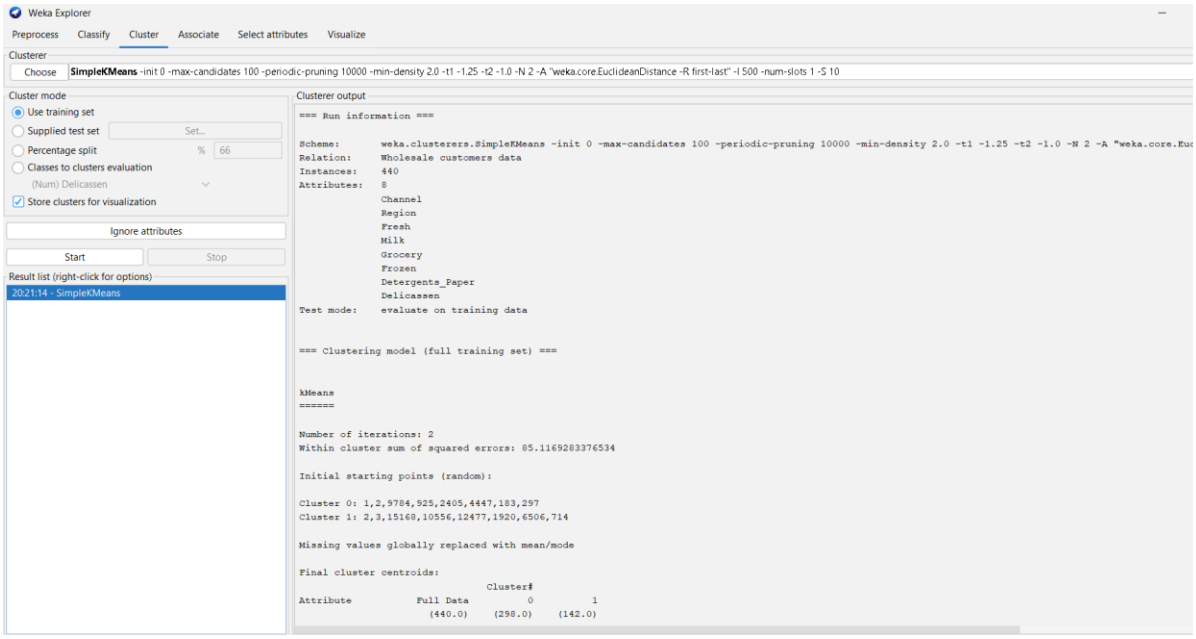

Figure: Result Of K-means Clustering

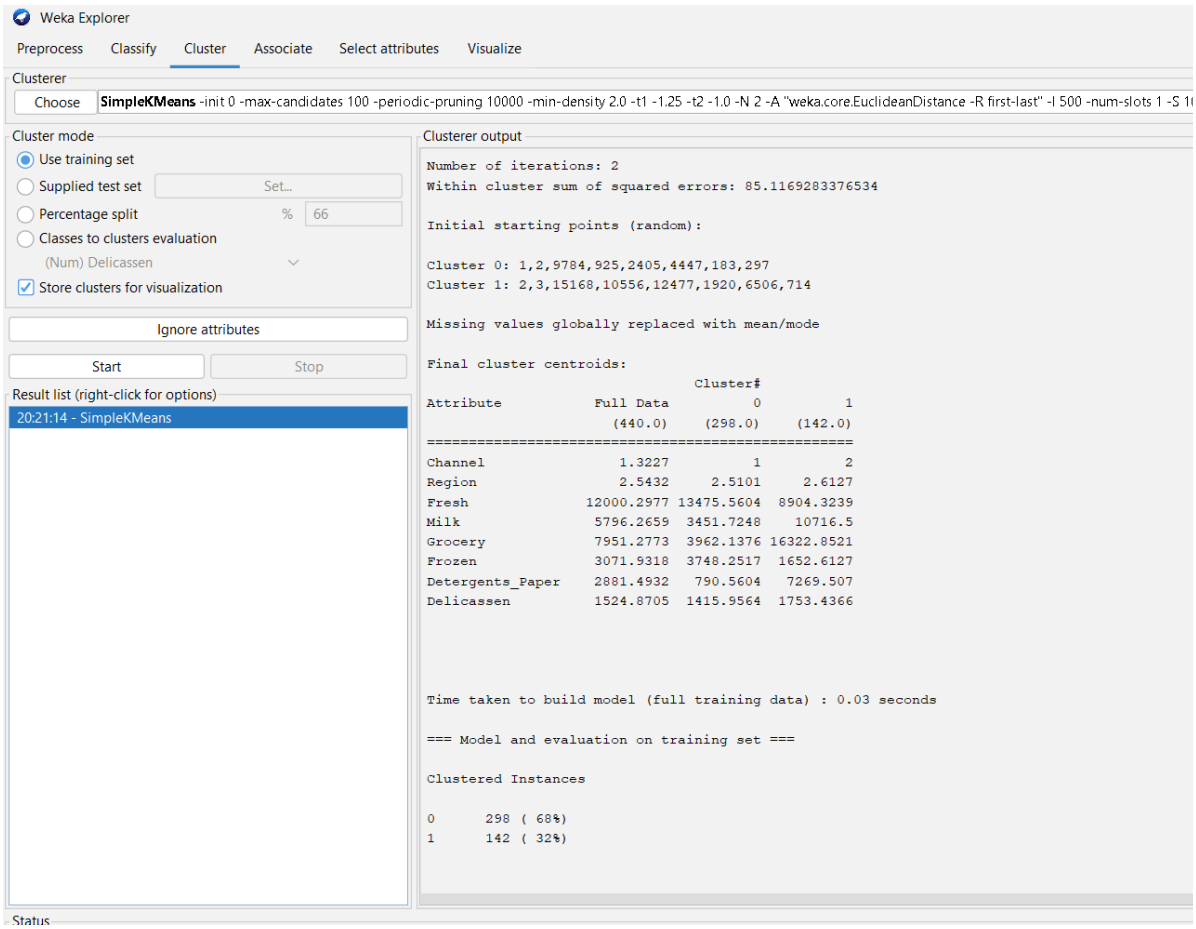

Figure: Summary Of k-means Clustering

Initial starting points (random):

Cluster 0: 1,2,9784,925,2405,4447,183,297

Cluster 1: 2,3,15168,10556,12477,1920,6506,714

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(440.0)	(298.0)	(142.0)
Channel	1.3227	1	2
Region	2.5432	2.5101	2.6127
Fresh	12000.2977	13475.5604	8904.3239
Milk	5796.2659	3451.7248	10716.5
Grocery	7951.2773	3962.1376	16322.8521
Frozen	3071.9318	3748.2517	1652.6127
Detergents_Paper	2881.4932	790.5604	7269.507
Delicassen	1524.8705	1415.9564	1753.4366

Within cluster sum of squared errors: 85.1169283376534

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(290.0)	(93.0)	(197.0)
Channel	1.3207	2	1
Region	2.531	2.5914	2.5025
Fresh	12647.6724	8708.9355	14507.0761
Milk	5864.7897	10543.7097	3655.9594
Grocery	7805.5103	15623.3226	4114.868
Frozen	3409.4724	1742.0108	4196.6497
Detergents_Paper	2804.2897	7024.0538	812.2183
Delicassen	1654.5483	1813.8602	1579.3401

Time is taken to build the model (percentage split) : 0 seconds

Clustered Instances

0 49 (33%)

1 101 (67%)

DISCUSSION: The goal of this study was to design an appropriate classifier for the weather classification dataset that could accurately categorize the condition and predict the class based on the test dataset. Following the application of two alternative classifiers, KNN and nave Bayes, the best-chosen classifier for the dataset is the Naive Bayes classifier, which has an accuracy of 83.03 percent. After that, a training set was chosen from the original dataset to create a Machine Learning Model. The model was tested using a prepared test dataset, and the accuracy of the model was 85.73 percent for the prepared test dataset. The concept of creating a training and testing dataset is crucial in data science since it is used to improve generalization and reduce overfitting. This also helps to give an unbiased evaluation of the accuracy of the model itself.

In the case of 8 attributes, we can think of the objects as being points in an eight-dimensional space, and visualizing clusters is generally straightforward too. And split 66% train, here Number of iterations 2. The K-means clustering algorithm calculates centroids and then repeats the process until the best centroid is discovered. The number of clusters is presumed to be known. The flat clustering algorithm is another name for it. The letter 'K' in K-means denotes the number of clusters found from data by the approach. Data points are assigned to clusters in this procedure in such a way that the sum of the squared distances between them and the centroid is as small as possible. It's important to remember that less cluster diversity leads to more identical data points within the same cluster.[5]

REFERENCES:

1. About data mining:
2. Supervised Data Set link (weather): WEATHER PREDICTION | Kaggle
3. Supervised Train and Test Data Set link (weather): WEATHER PREDICTION | Kaggle
4. Unsupervised Data set: Unsupervised Learning | Kaggle
5. Understanding K-Means Clustering Algorithm - Analytics Vidhya