DATA WAREHOUSING AND DATA MINING

SEMESTER: SUMMER 2021-2022

FINAL-TERM PROJECT DATA MINING WITH WEKA SUBMITTED BY

STUDENT NAME	ID
Arpita Saha	19-41363-3
Jakia Sultana Nupur	19-41405-3
Abdullah Mutasaddik	19-41373-3
Saifullah	19-41367-3

SECTION: A

DEPARTMENT: CSE

SUBMITTED TO: TOHEDUL ISLAM

INTRODUCTION

Data mining is the process of discovering knowledge and pattern from large amounts of data. It is also known as discovering knowledge or knowledge discovery in database (KDD). Data is preprocessed into a use form through which information can be claimed. The things that we can gain from this information is called knowledge. Hence knowledge is derived from the information which is gathered from the data sets. Data mining is used in many areas of research and business including education, healthcare, marketing, product development etc. KNN, Naïve Bayes, and Decision Tree are some of the classification algorithms used in data mining. We chose room occupancy of a hotel to predict weather a room will me occupied or not. To do this we have used two different classifiers to find the best suited classifier for the data set.

Supervised learning (SL) is the machine learning problem of learning a function that translates an input to an output based on sample input-output pairs. It infers a function from labeled training data consisting of a collection of training instances.

Unsupervised learning is a type of algorithm that learns patters from untagged data. Through mimicry which is also an important learning method of humans, the machine is forced to build a compact internal representation of its world and then generate imaginative content from it. We chose "Facebook Live Sellers In Thailand Data Set" to classify by using K means clustering dataset.

Information about the supervised set: To make a classification-based model we need a proper dataset
Dataset Name: Room Occupancy Detection Data Set
Link: http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
This dataset is a classification-based dataset which is based on real data. Which means our model will have a real-life impact. It has total 8143 instances or row of data. It has total 7 columns or attributes. From them 5 are important and can be called the feature matrix and the last one is the class attribute which is used for prediction. Our dataset is about predicting in a office will a certain room be occupied or not based on some parameters or attributes. Now let us know about those parameters and why they are important for a room to be occupied or not.

The class attribute is:

- **Occupancy**

The other attributes are:

- **Room number**
- **Temperature:** Our first attribute is temperature which is given in Celsius. When you choose a room always you want a room not to be too hot or cool. If a room always has high temperature or very low temperature it will not be suitable for work.

- **Humidity:** Humidity is also another factor to choose your room in the office. When the humidity is between 30 to 50% it is better. If the humidity is high the outside feels wetter and if it is higher than 60% it will cause health issues.

- **Light:** Light can be said as the most important factor for a room to be occupied or not. If a room does not have proper lighting, then there is a very high chance an employee will not use that room for work.

- **CO2:** CO2 Level is also an important factor for a room. High CO2 levels indicate a potential problem with air circulation and fresh air in a room or building. It should be lower than 1000 ppm in all the case for a room to be in fresh condition.

- **Humidity Ratio:** It is also known as the Specific or Absolute Humidity, the finite quantity of moisture in each volume of air. It should be well balanced in a room otherwise a room will not be suitable for working.

There is a total of 8143 instances of 7 attributes and all these instances were used for classification. Here are the graphical details of the attributes:

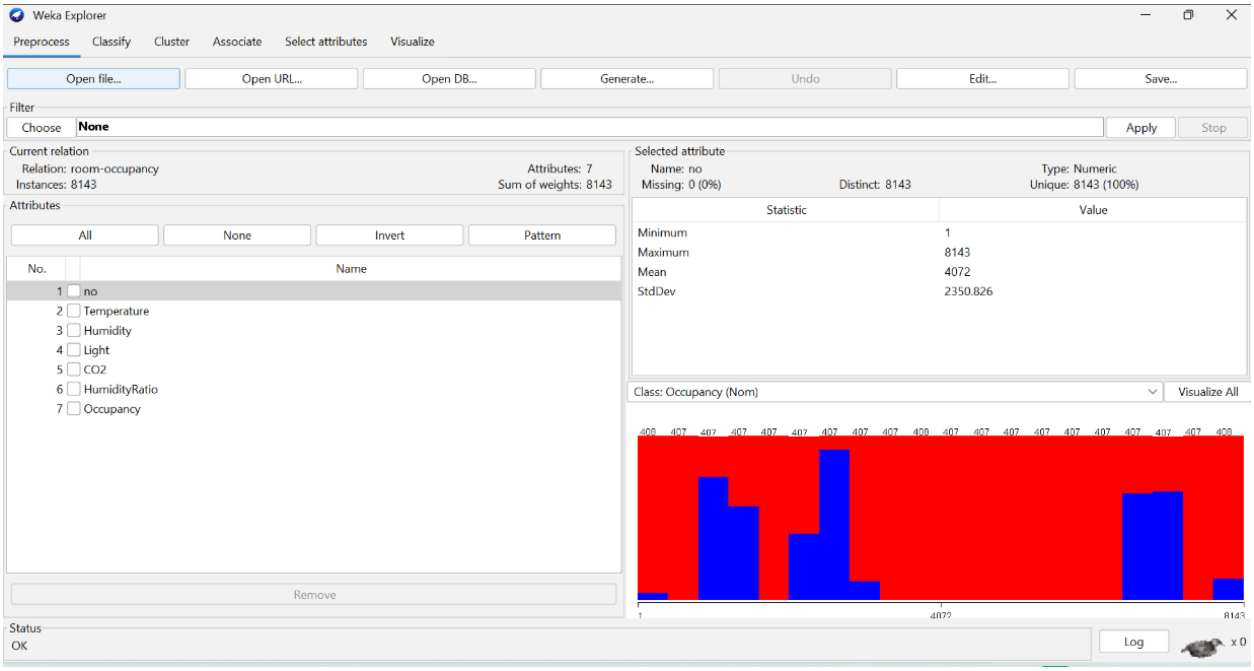

Figure 1: Imported Data Set

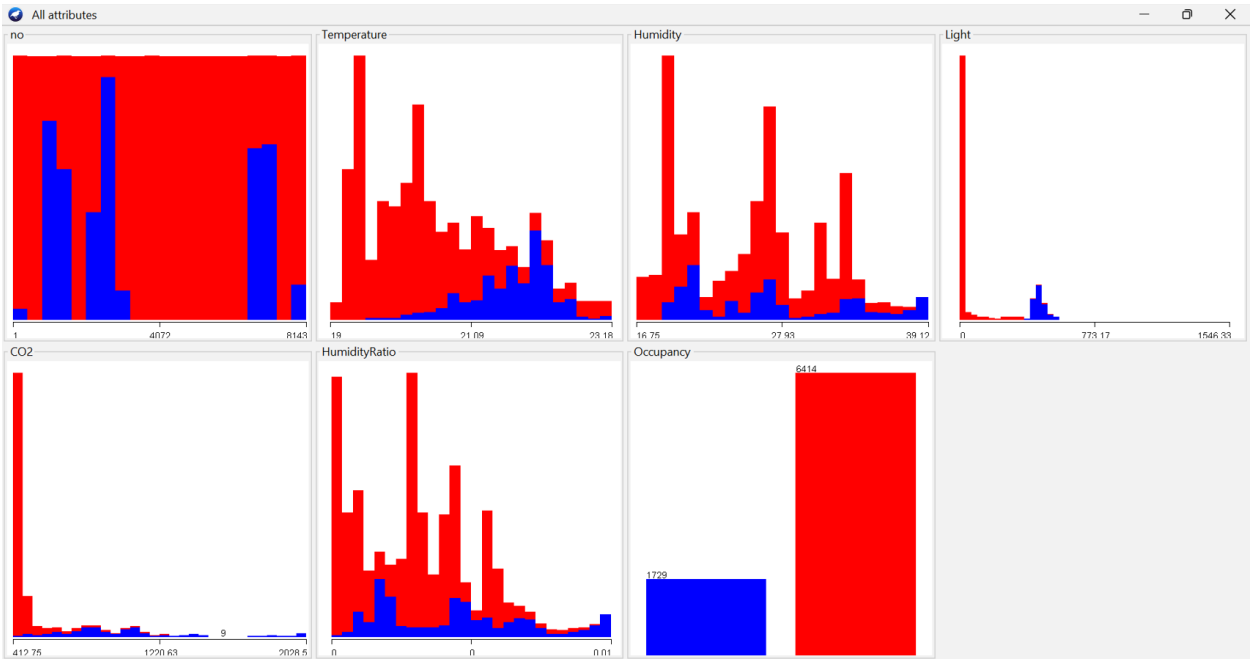

Figure 2: Graphically Represent all attributes

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

Results of the classifiers: Weka 3.8.6 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.

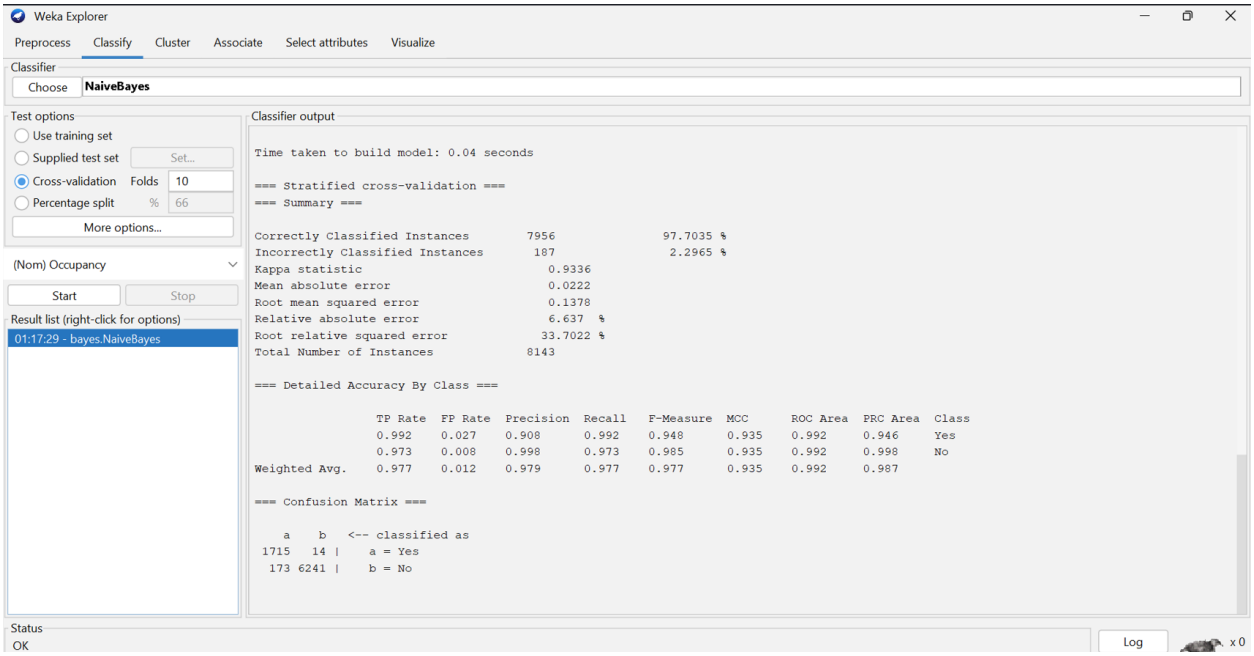

Figure 3: Naïve bayes classification

Applying K-nearest Neighbor classifier: K-Nearest Neighbor 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.

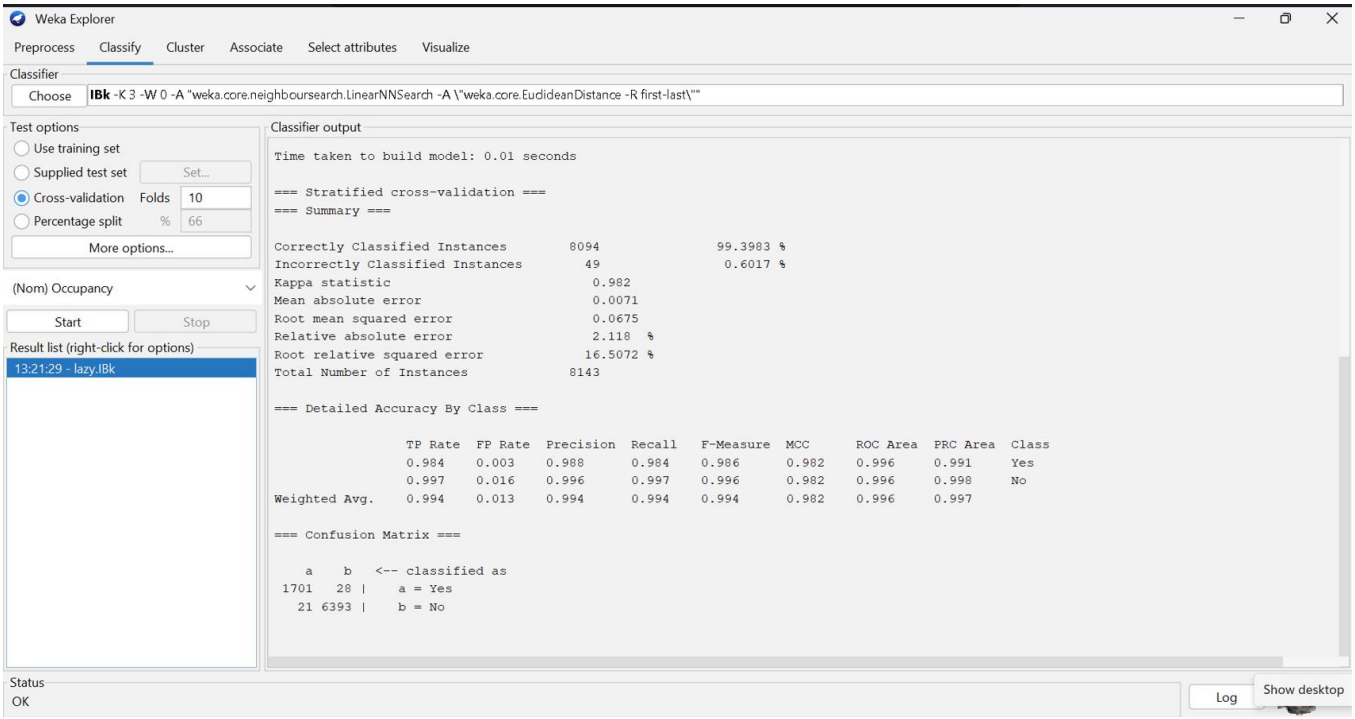

Figure 4: KNN Classification

Reason to choose KNN classifier: From the obtained result of the two classifiers, it is clearly seen that KNN the highest percentage of correctly classified instances which is 99.39% whereas naïve Bayes had 97.7%. As KNN has the most accurate value so it would be a more suitable classifier for the dataset. One of the advantages of KNN that good results were obtained in most of the cases. Moreover, KNN faster than naïve Bayes due to naïve Bayes's expensive real-time execution. It is easy to implement.

Here is the summary of the KNN classifiers result:

==Summery==

- Correctly Classified Instances 8094 99.3983 %
- Incorrectly Classified Instances 49 0.6017 %
- Kappa statistic 0.982
- Mean absolute error 0.0071
- Root mean squared error 0.0675
- Relative absolute error 2.118 %
- Root relative squared error 16.5072 %
- Total Number of Instances 8143

Preparing Test Dataset: One of the fundamentals of machine learning is to separate the training set from the test set. A portion of the original dataset was used to construct a training dataset for ML behavior detection. Subsets of the training dataset were used to evaluate the model. When creating the test dataset, we made sure it was big enough to provide reliable results. Also, it was representative of the whole database. In other words, data that would have been most dissimilar to the training set was not selected as the test data. For the test set, we utilize the best classifier to make predictions about their labeling.

no	Temperat	Humidity	Light	CO2	HumidityR	Occupancy
1	23.18	27.272	426	721.25	0.004793	Yes
2	23.15	27.2675	429.5	714	0.004783	Yes
3	23.15	27.245	426	713.5	0.004779	Yes
4	23.15	27.2	426	708.25	0.004772	Yes
5	23.1	27.2	426	704.5	0.004757	Yes
6	23.1	27.2	419	701	0.004757	Yes
7	23.1	27.2	419	701.6667	0.004757	Yes
8	23.1	27.2	419	699	0.004757	Yes
9	23.1	27.2	419	689.3333	0.004757	Yes
10	23.075	27.175	419	688	0.004745	Yes
11	23.075	27.15	419	690.25	0.004741	Yes
12	23.1	27.1	419	691	0.004739	Yes
13	23.1	27.16667	419	683.5	0.004751	Yes
14	23.05	27.15	419	687.5	0.004734	Yes
15	23	27.125	419	686	0.004715	Yes
16	23	27.125	418.5	680.5	0.004715	Yes
17	23	27.2	0	681.5	0.004728	No
18	22.945	27.29	0	685	0.004728	No
19	22.945	27.39	0	685	0.004745	No
20	22.89	27.39	0	689	0.00473	No
21	22.89	27.39	0	689.5	0.00473	No
22	22.89	27.39	0	689	0.00473	No
23	22.89	27.445	0	691	0.004739	No
24	22.89	27.5	0	688	0.004749	No
25	22.89	27.5	0	689.5	0.004749	No
26	22.79	27.445	0	689	0.00471	No
27	22.79	27.5	0	685.6667	0.00472	No
28	22.79	27.5	0	687	0.00472	No
29	22.79	27.5	0	688	0.00472	No
30	22.745	27.5	0	670	0.004707	No
31	22.7	27.46333	0	668.6667	0.004688	No
32	22.7	27.5	0	670	0.004694	No
33	22.7	27.5	0	667	0.004694	No
34	22.66667	27.42667	0	664.5	0.004672	No
35	22.7	27.6	0	670	0.004711	No
36	22.6	27.42667	0	670.3333	0.004653	No

Figure 5: Prepare Test Data set

Procedure of testing the test dataset:

1. Open the file and extract CSV file and training data set and select it from the device

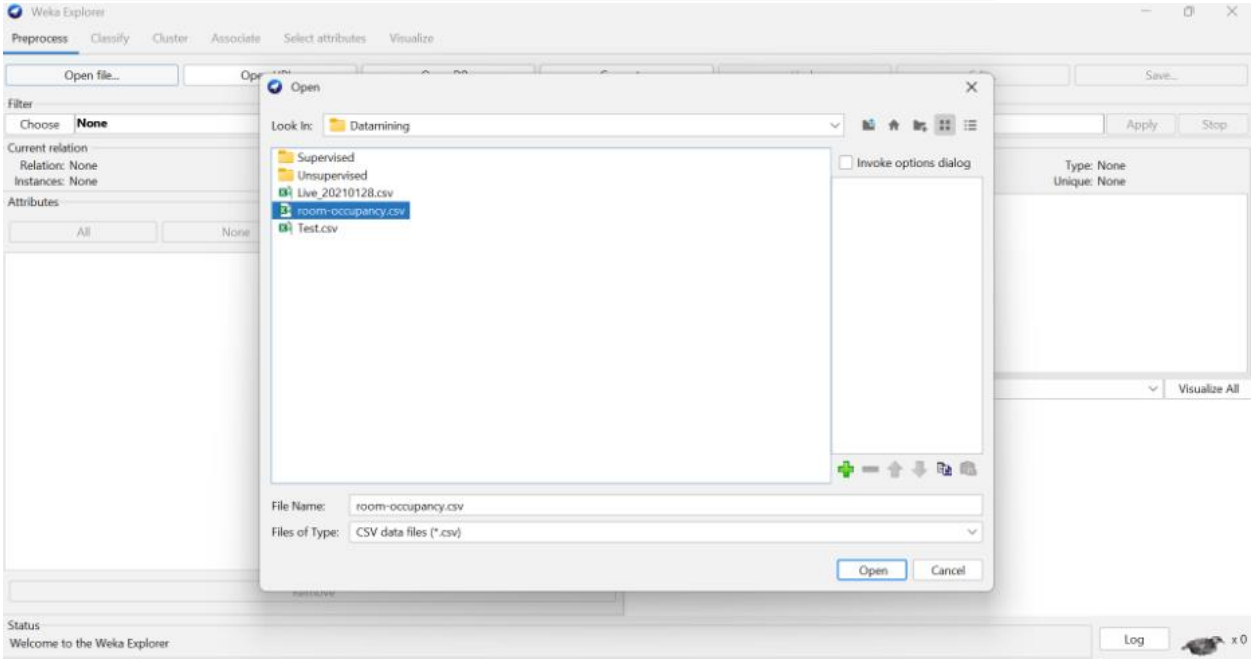

Figure 6: Training data set

2. After that click the open option and the details of the dataset will be popped

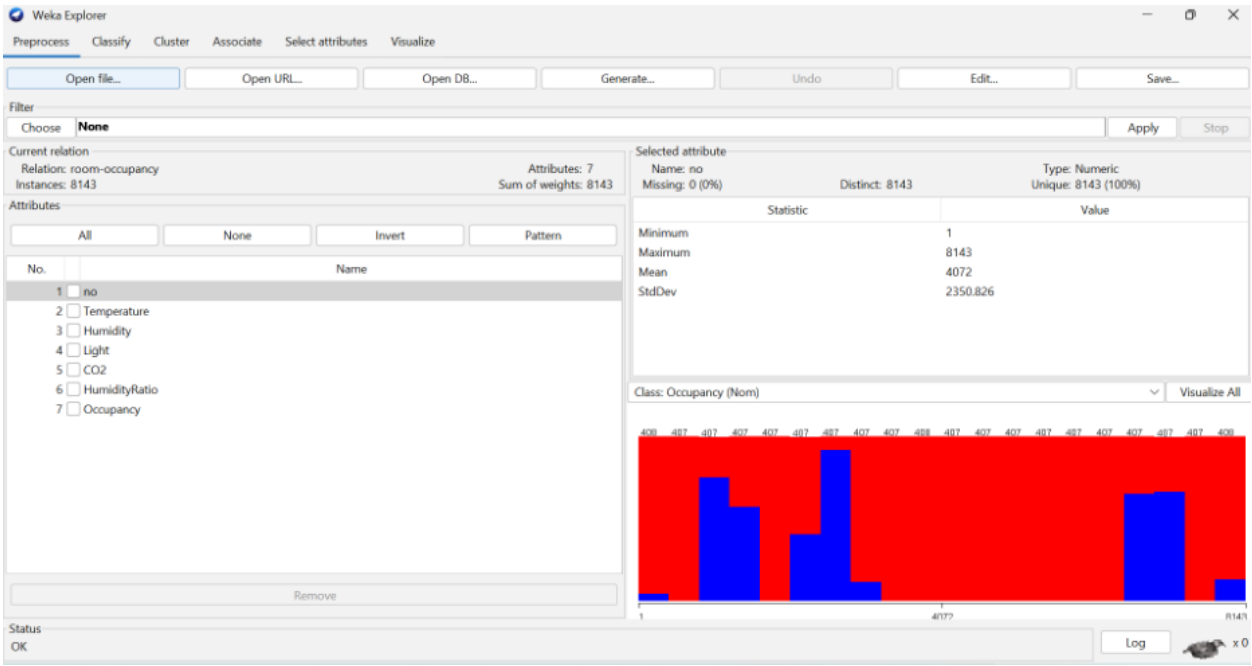

Figure 7: Train Data set details

3. Then the preferred classifier for the training dataset was selected and then from the test options we chose the training set and the start option was selected.

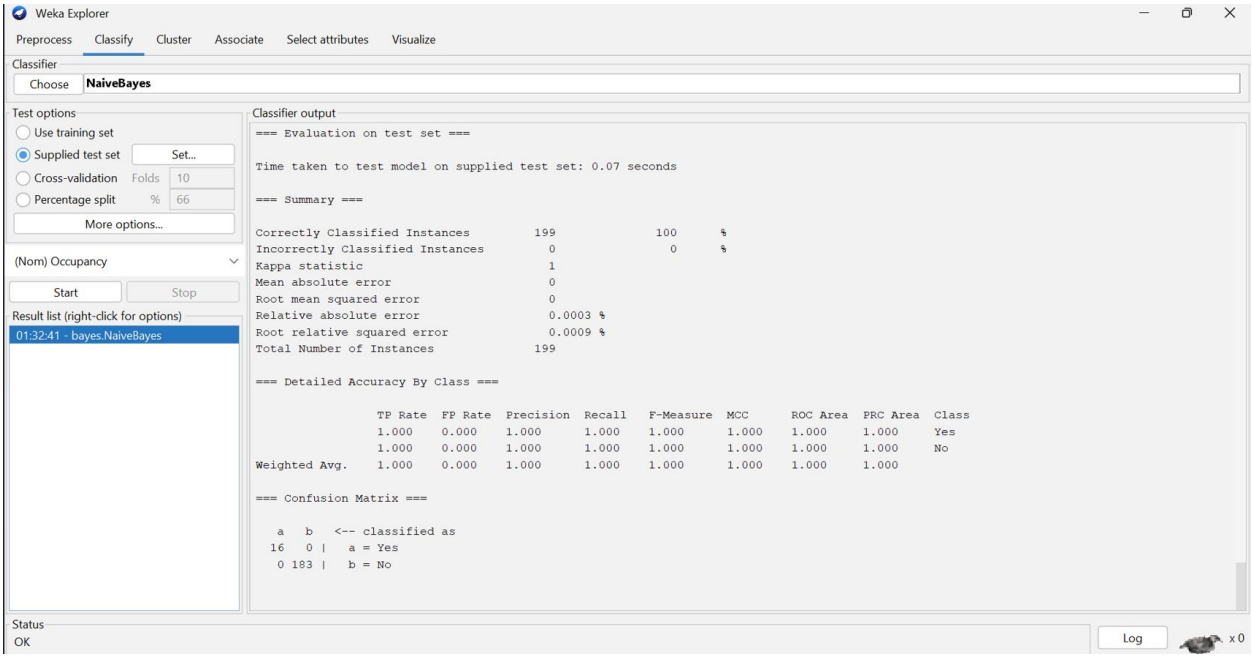

Figure 8: Result summary of the Train Data Set

4. Import Test Set

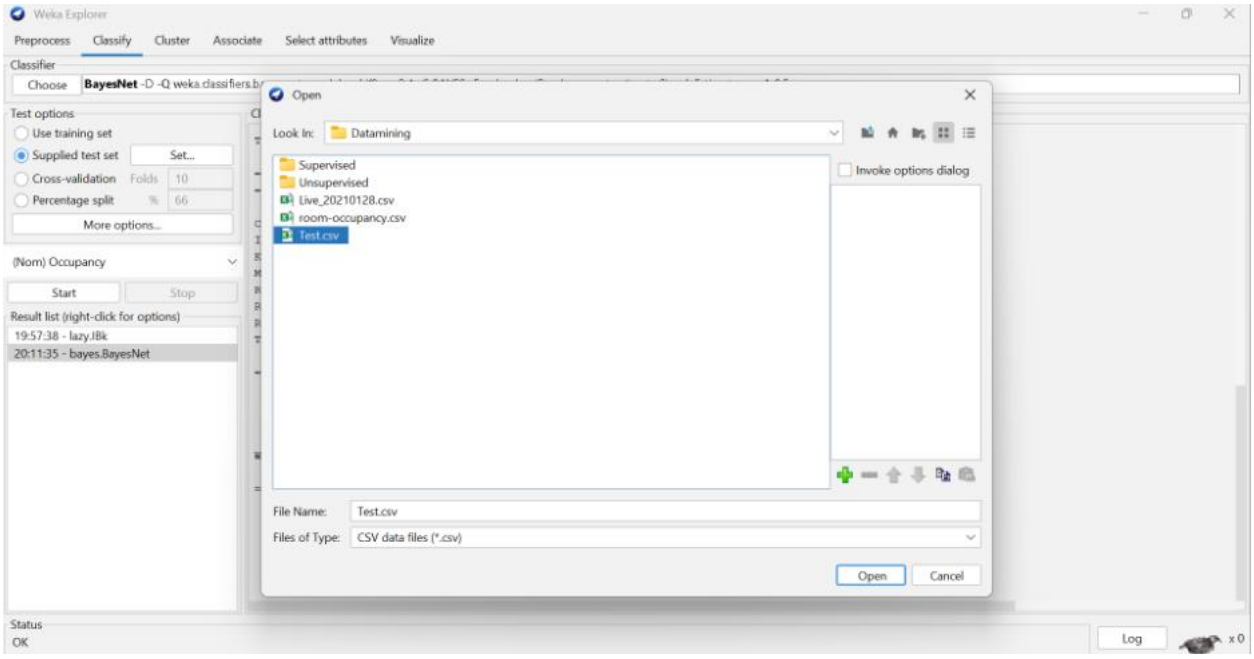

Figure 9: Import Test Data Set

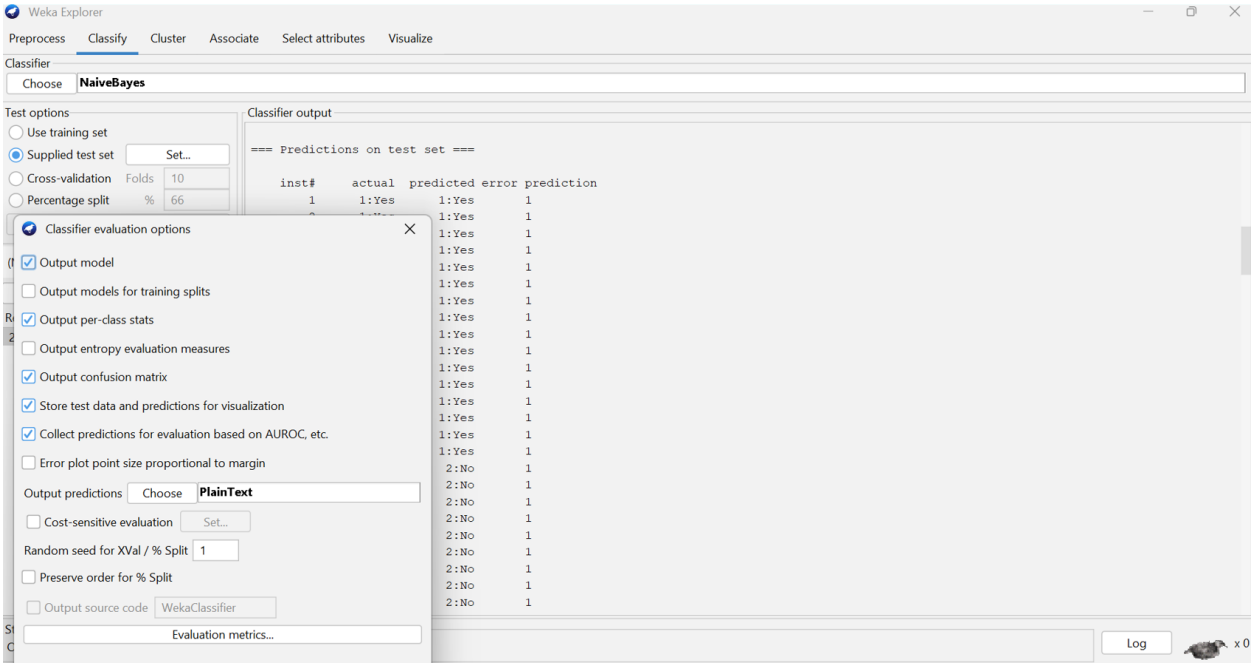

Figure 10: Plain Text Format

5. Run the test data set

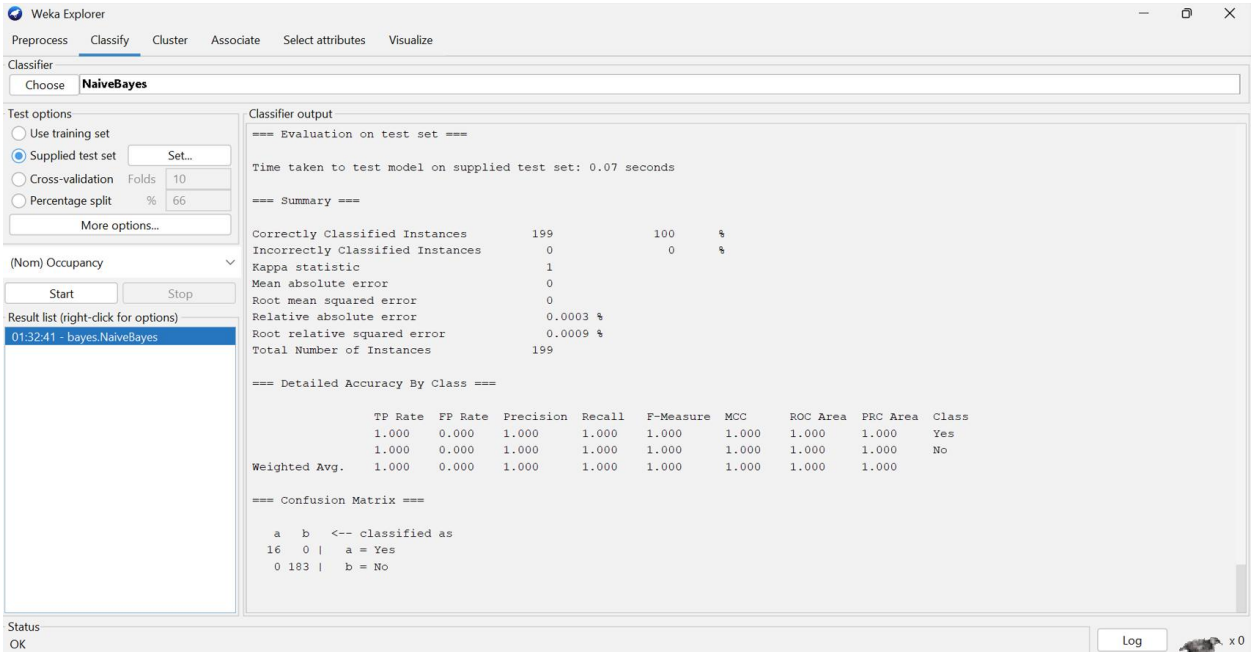

Figure 11: Test Set Run

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 '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 0.07 seconds and among the 199 instances, there were 0 instances where the error was found. In this model, the Correctly classified instances 100 % Value describes the amount of accuracy of correctly classified instances provides by the algorithm. In this case, the percentage is 100% which is perfect. In this case, the percentage is 0% 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 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 0.0009% which is ideal. RAE 0.0003%.

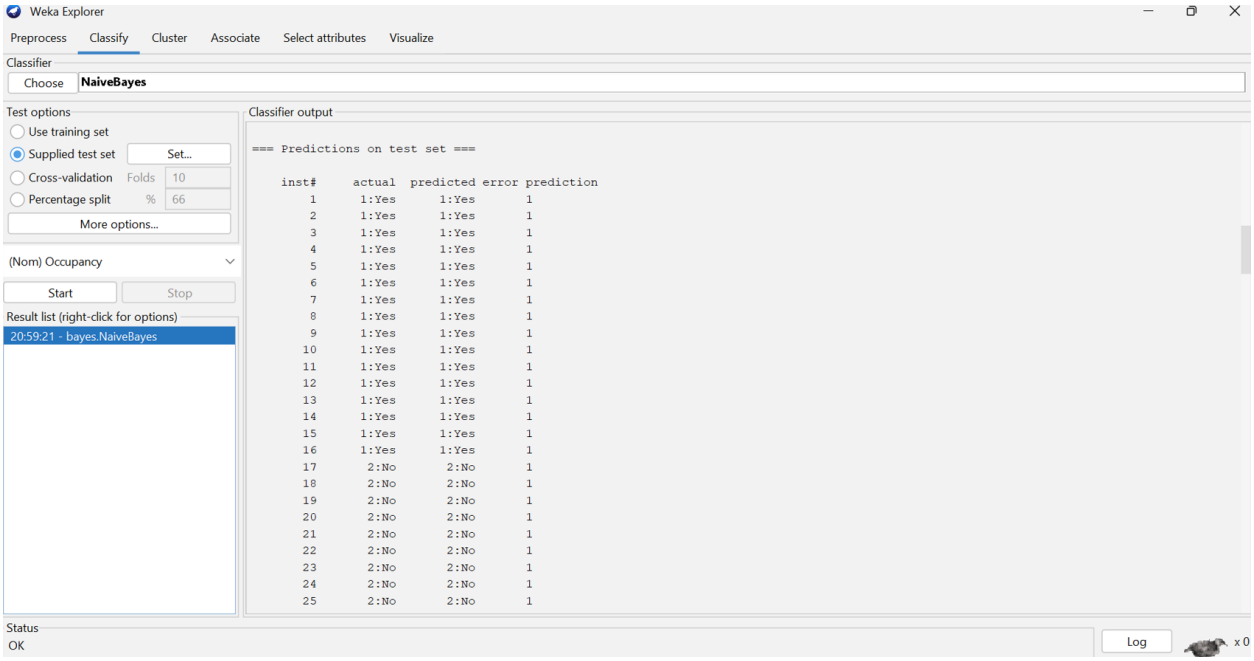

Figure 12: Predict The test data

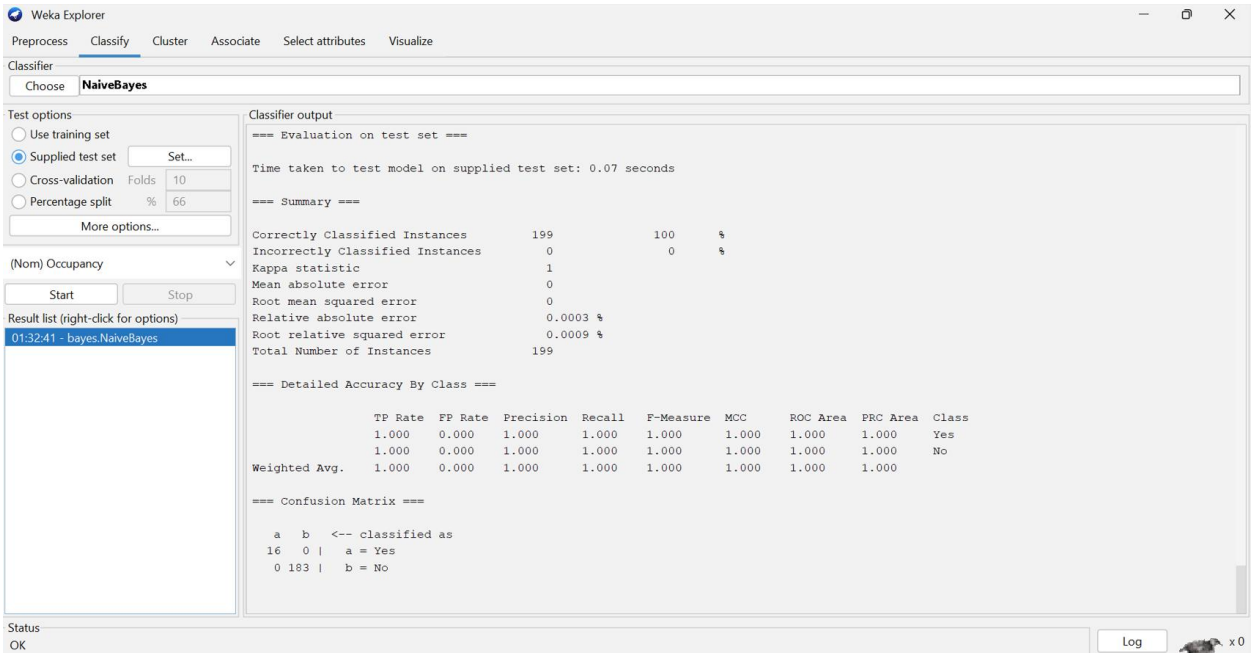

Figure 13: Result Summery of Test Data Set

Information about the unsupervised dataset: In this report, the used 'Facebook live sellers in Thailand' a CSV dataset file, collect from uci.com was use used to predict the outcome of the condition that might be accurate. This dataset would serve as a basis for research on customer engagement with the noble sales channel that is Facebook live, through comparative studies through other forms of content.

The features are:

- Status_id
- Status_type
- Status_published
- Num_reactions
- Num_comments
- Num _shares
- Num _likes
- Num _haha
- Num _sad
- Num _angry

About the attribute: The dataset contains 10 attributes.

Attribute	Representation in Dataset	Data type
Status_id	Numeric value	Numeric Type
Status_type	Numeric value	Numeric Type
Status_published	Numeric value	Numeric Type
Num_reactions	Numeric value	Numeric Type
Num_comments	Numeric value	Numeric Type
Num _shares	Numeric value	Numeric Type
Num _likes	Numeric value	Numeric Type
Num _haha	Numeric value	Numeric Type
Num _sad	Numeric value	Numeric Type
Num _angry	Numeric value	Numeric Type

There is a total number of 7050 instances of these 10 attributes and all these instances were used for the classification

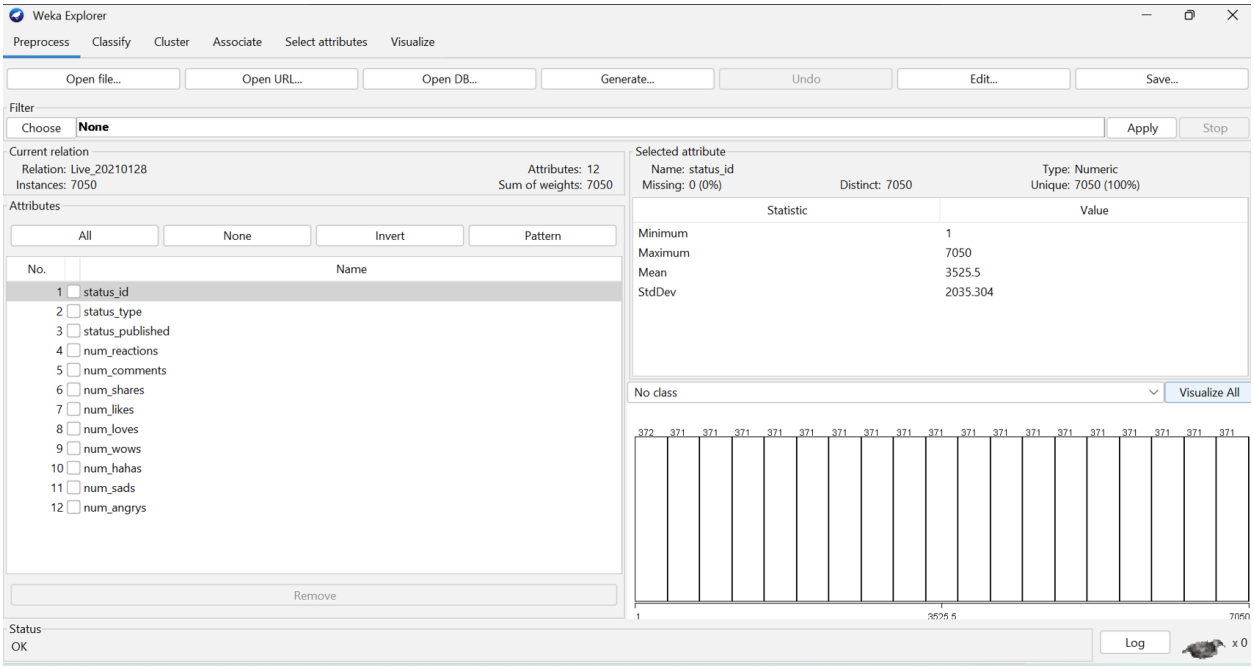

Figure 14: Unsupervised Data Imported

Applying K-means Clustering: The K-means approach for clustering performs repeated centroids computations until the best possible centroid is found. Number of clusters is assumed to be known. It's also referred to as the flat clustering method. The 'K' in 'K-means' stands for the total number of clusters that can be extracted from a dataset using this technique. In this case, k equals 2.

Results of the K-Means clustering: Weka 3.8.6 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.

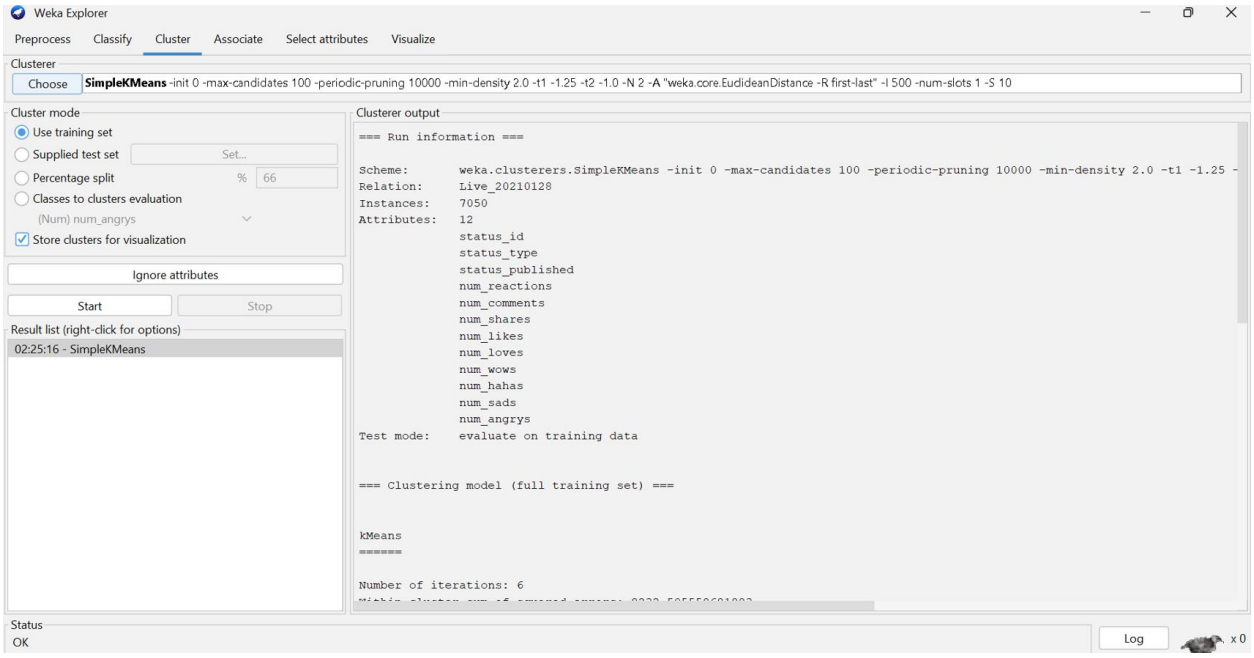

Figure 15: Results of K means clustering

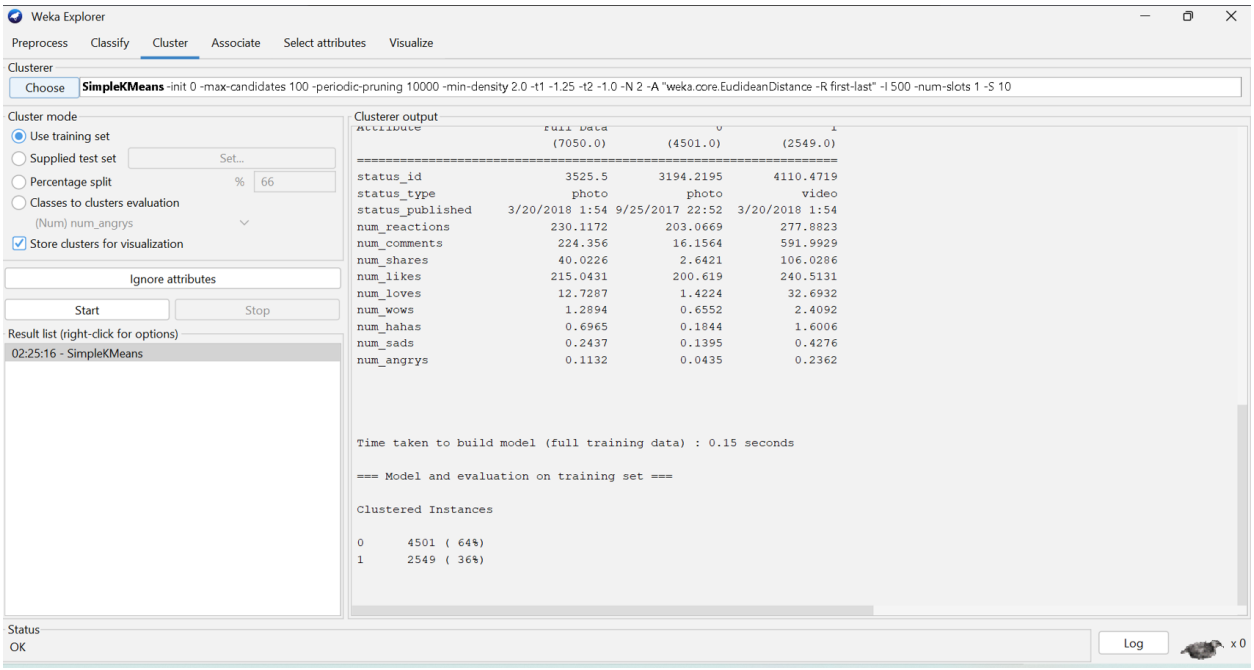

Figure 16: Summery of K Means Clustering

Number of iterations: 6

Within cluster sum of squared errors: 8232.505558681883

Initial starting points (random):

Cluster 0: 4414,video,'11/14/2017 3:17',5,0,0,5,0,0,0,0,0

Cluster 1: 6043,video,'11/23/2017 2:04',42,1,0,41,1,0,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(7050.0)	(4501.0)	(2549.0)
status_id	3525.5	3194.2195	4110.4719
status_type	photo	photo	video
status_published	3/20/2018 1:54	9/25/2017 22:52	3/20/2018 1:54
num_reactions	230.1172	203.0669	277.8823
num_comments	224.356	16.1564	591.9929
num_shares	40.0226	2.6421	106.0286
num_likes	215.0431	200.619	240.5131
num_loves	12.7287	1.4224	32.6932
num_wows	1.2894	0.6552	2.4092
num_hahas	0.6965	0.1844	1.6006
num_sads	0.2437	0.1395	0.4276
num_angrys	0.1132	0.0435	0.2362

Time taken to build model (full training data) : 0.11 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 4501 (64%)

1 2549 (36%)

Discussion: The purpose of this project was to create a classifier for the Room occupancy classification dataset that could properly identify the condition and forecast the class based on the test dataset. We know that naïve bayes best works with the categorical based dataset whereas our whole dataset was numeric, and its only class variable was categorical. Still this algorithm gave us over 97.70% accuracy. For this reason, we can also say naïve bayes is not a bad algorithm to apply here. Overall, to draw the conclusion we will choose KNN as the most suitable algorithm because KNN gave us 99.39%.

In unsupervised case there are total instances are 7050 and total attributes are 10. After applying k-means clustering, here Number of iterations were 6. after final iteration 64% of instances belong to cluster 0 and 36% instances belongs to cluster 1. The K-means clustering technique calculates centroids and then iterates until the optimal centroid is identified. 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 approached. 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.

References:

1. Supervised Data Set link: http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
2. Unsupervised Data set link: https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand