**CONTENTS**

Contents..................................................................................................................... i

Student’s Declaration................................................................................................ ii

Certificate by the Guide............................................................................................ iii

Acknowledgement.................................................................................................... iv

Abstract..................................................................................................................... v

List of Tables............................................................................................................ vi

List of Figures......................................................................................................... viii

Chapter-1 INTRODUCTION ..............................................................................xv- xvii

* 1. Machine learning ............................................................................................. xv
  2. Significance of Machine learning....................................................................... xv
  3. Types of machine learning .............................................................................. xvi
  4. Project Relevance .......................................................................................xvii

Chapter-2 WEKA........................................................................................... xviii - xx

* 1. Introduction to WEKA ..............................................................................xviii
  2. User Interfaces...........................................................................................xvix

Chapter-3 CLASSIFIERS .............................................................................. xxi- xxiv

* 1. Decision Table Classifier......................................................................................xxi
  2. Decision Tree Classifier.................................................................................xxi
  3. Naïve Bayes Classifier Algorithm..................................................................xxii
  4. ZeroR Classifier in WEKA............................................................................xxiv

Chapter-4 IMPLEMENTATION .................................................................. xxv – xxix

* 1. Program Code.................................................................................................xxv
  2. Output ......................................................................................................xxviii

Conclusion................................................................................................................xxx

References...............................................................................................................xxxi

**CANDIDATE’S DECLARATION**

I hereby, declare that the work presented in this report, entitled “**COMPARING CLASSIFIERS ON WEATHER DATASET USING WEKA** ”, in fulfilment of the requirement for the award of the degree Bachelor of Engineering in Computer Science and Engineering, submitted in CSE Department, Chandigarh College of Engineering and Technology (Degree Wing), affiliated to Panjab University, Chandigarh, is an authentic record of our own work carried out during my degree under the guidance of Dr. Ankit Gupta The work reported in this has not been submitted by me for award of any other degree or diploma.

Aayushi Aggarwal

(CO17502)

Abhijeet Baruah

(CO17302)

Date: 27/04/2020

Place: CCET

**CERTIFICATE**

This is to certify that the Project work entitled “**COMPARING CLASSIFIERS ON WEATHER DATASET USING WEKA**”, submitted by **Aayushi Aggarwal (CO17502) and Abhijeet Baruah (CO17302)** infulfillment for the requirements of the award of Bachelor of Engineering Degree in Computer Science & Engineering at Chandigarh College of Engineering and Technology (Degree Wing), Chandigarh is an authentic work carried out by him/her under my supervision and guidance. To the best of my knowledge, the matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree.

Dr. Ankit Gupta

Dept of CSE

CCET(DegreeWing), Chandigarh

Date: 27/04/2020

Place: CCET

DATE: 17/12/2019

PLACE: CCET

Date : 17/12/19

**ACKNOWLEDGEMENT**

“Any serious and lasting achievement or success, one can never achieve without the help, guidance and co-operation of so many people involved in the work.

I would like to express deep gratitude to Dr. Sunil K. Singh, Head of Department (Computer Science & Engineering), submitted in CSE Department, Chandigarh College of Engineering & Technology(Degree wing),and affiliated to Punjab University, Chandigarh, without whose permission the subject models on real-world problems would not be possible. I would also like to thank Dr. Ankit Gupta, CSE Department, who encourages me for taking up this Project.

I have tried my best to keep report simple yet technically correct. I hope I succeed in my attempt.

**ABSTRACT**

WEKA is open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that one can develop machine learning techniques and apply them to real-world data mining problems. Weka is tried and tested open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research, and industrial applications, contains a plethora of built-in tools for standard machine learning tasks, and additionally gives transparent access to well-known toolboxes such as scikit-learn, R, and Deeplearning4j.

Depending on the kind of Machine Learning model that is desired to be developed, one can opt for any algorithm out of the available options on the given dataset. Each algorithm develops a unique model with a different accuracy. Especially, when it comes to weather prediction, we know that no algorithm is there which can make accurate predictions. Thus, the following discussion describes a project considering the dataset of weather details of a country and comparing the accuracy of various models that can be developed using different algorithms on that dataset.

**LIST OF TABLES**

**DATASET**

The dataset taken here, contains the information of weather conditions of a western country.

* The attributes of the dataset are as follows:-

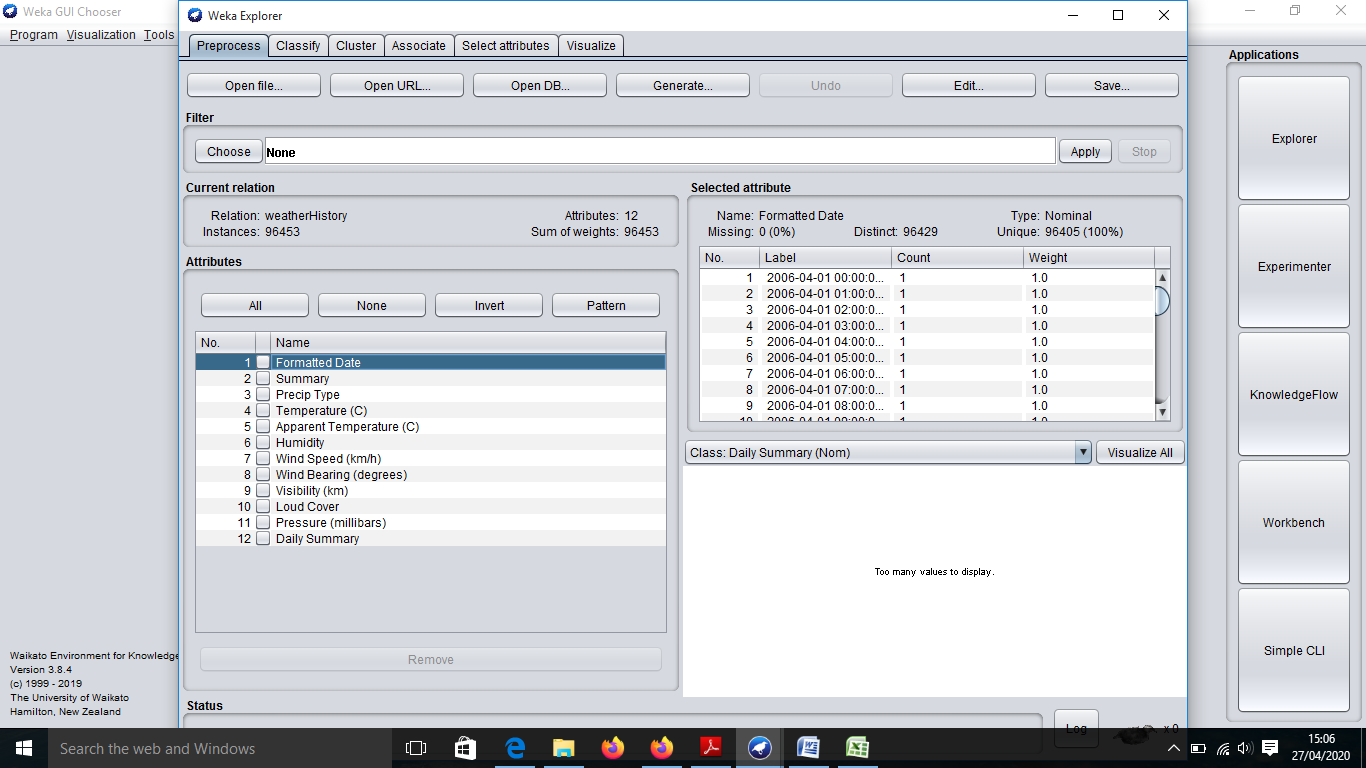
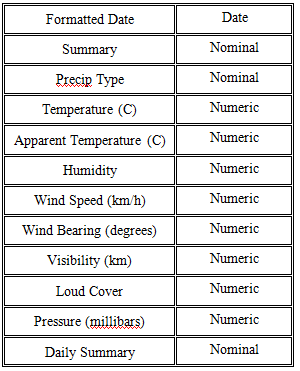


FIG-1 Attributes in the Dataset used.

* FIG-2 The dataset contains 12 attributes such that they have the following types:-



* The Dataset contains 96454 tuples filled in the following manner.

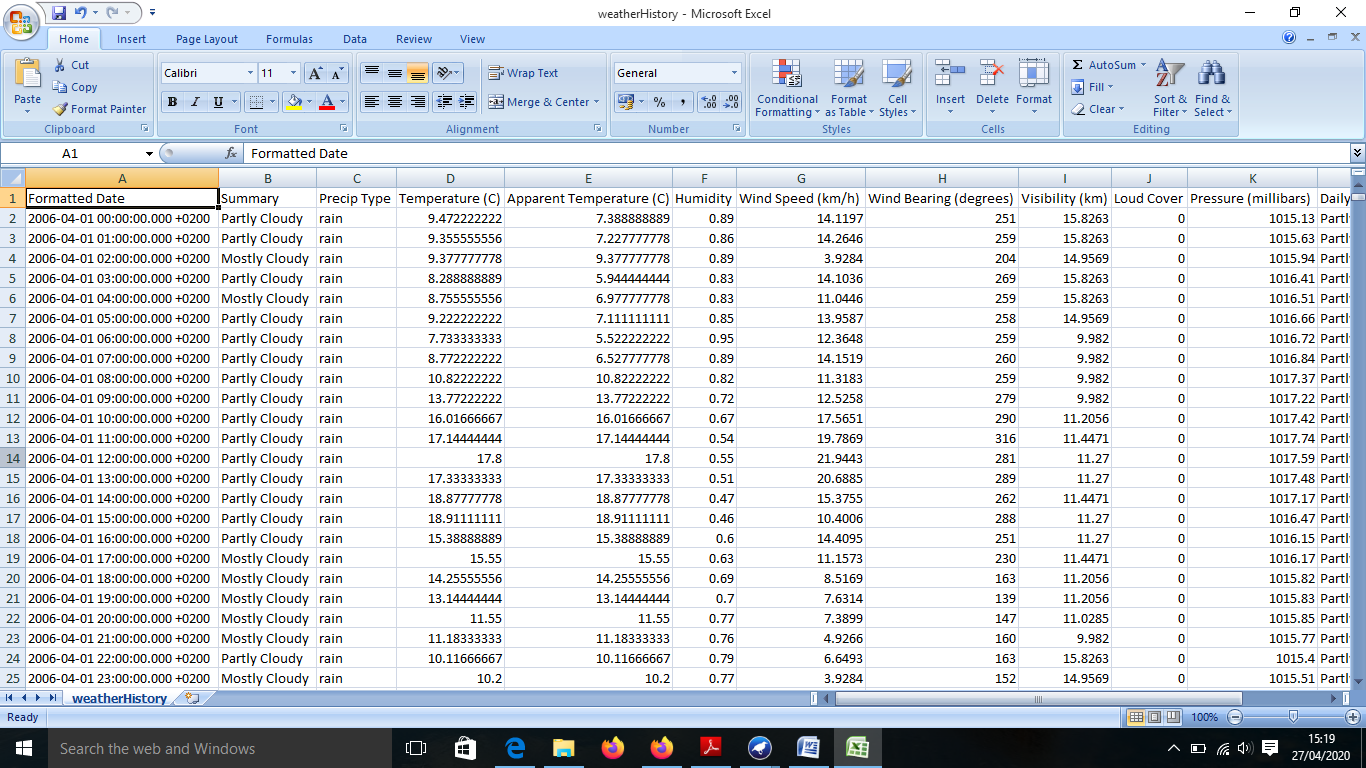


FIG-3 First seven attributes in the Dataset used.

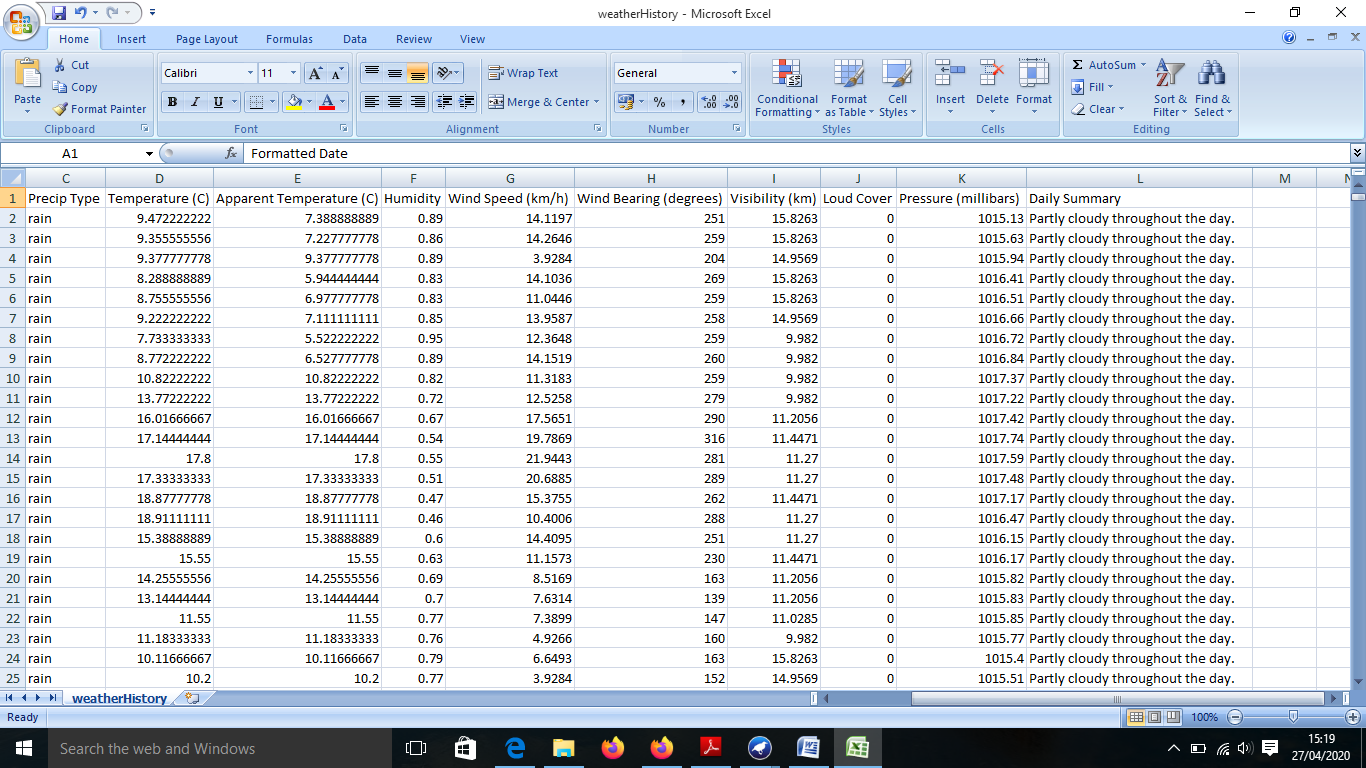
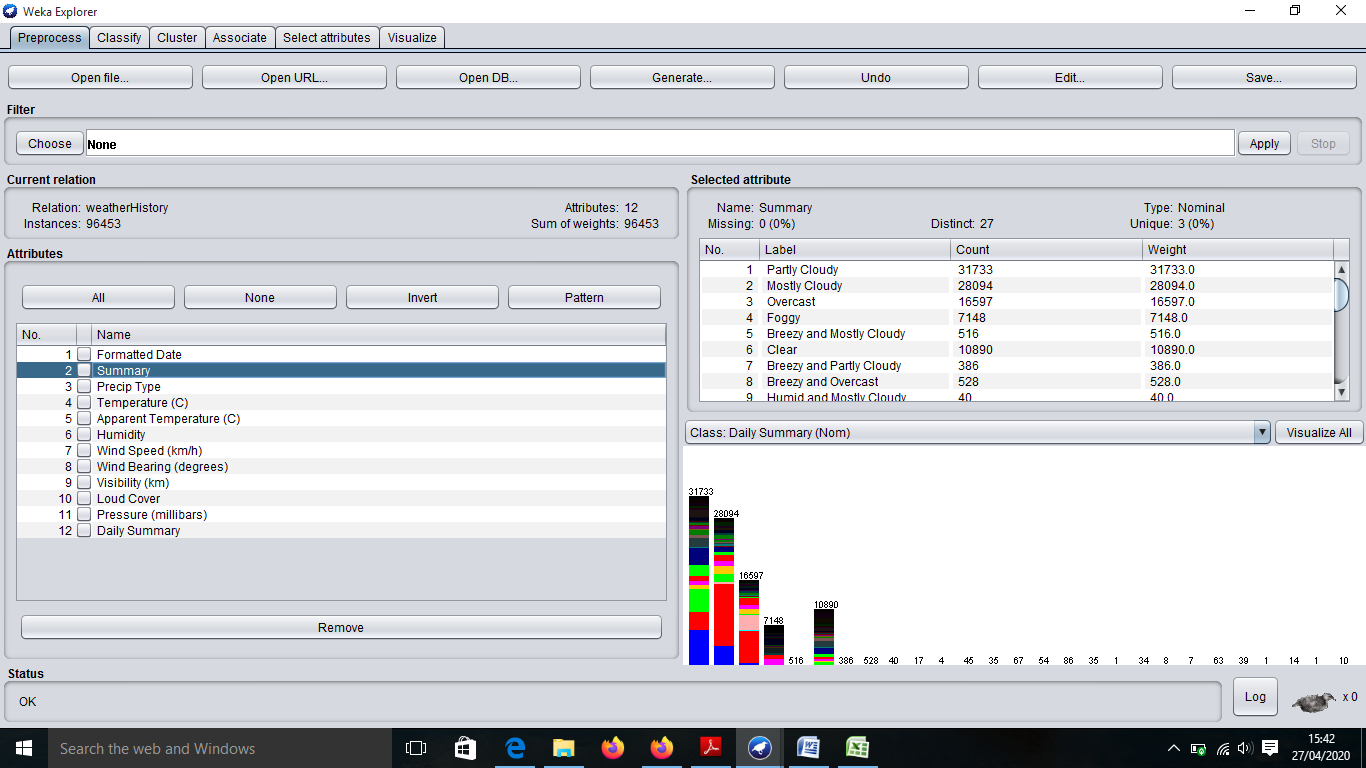
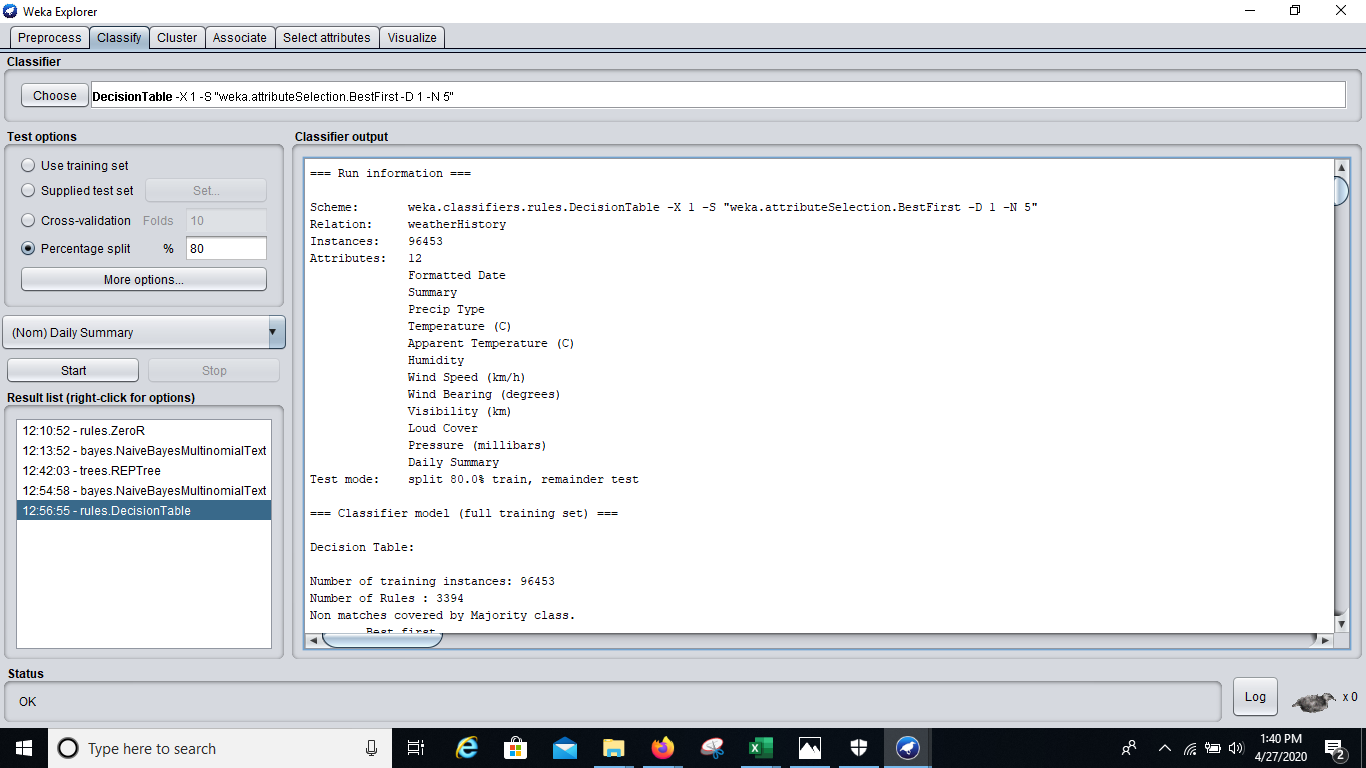


FIG-4Last seven attributes in the Dataset used.

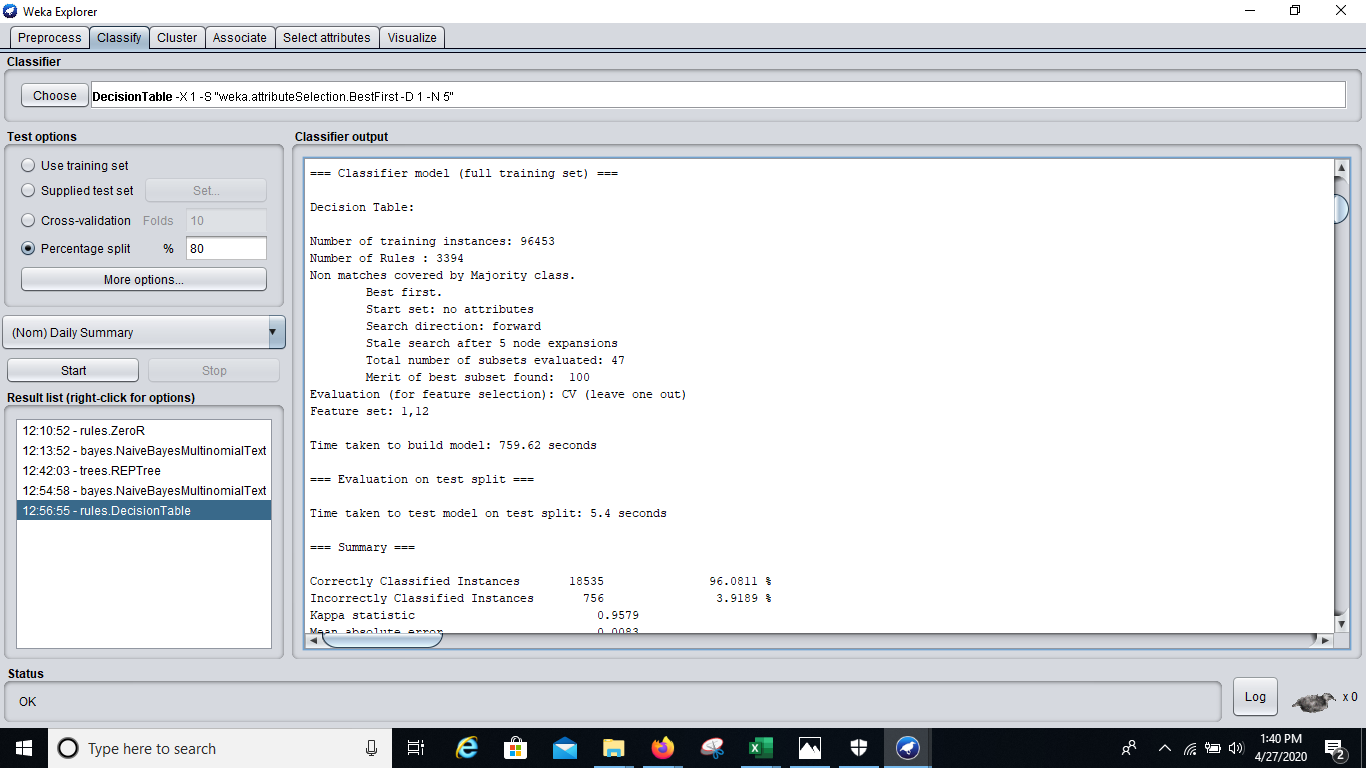
**LIST OF FIGURES**

**Figure:- 1** Preprocessing data in WEKA Explorer.

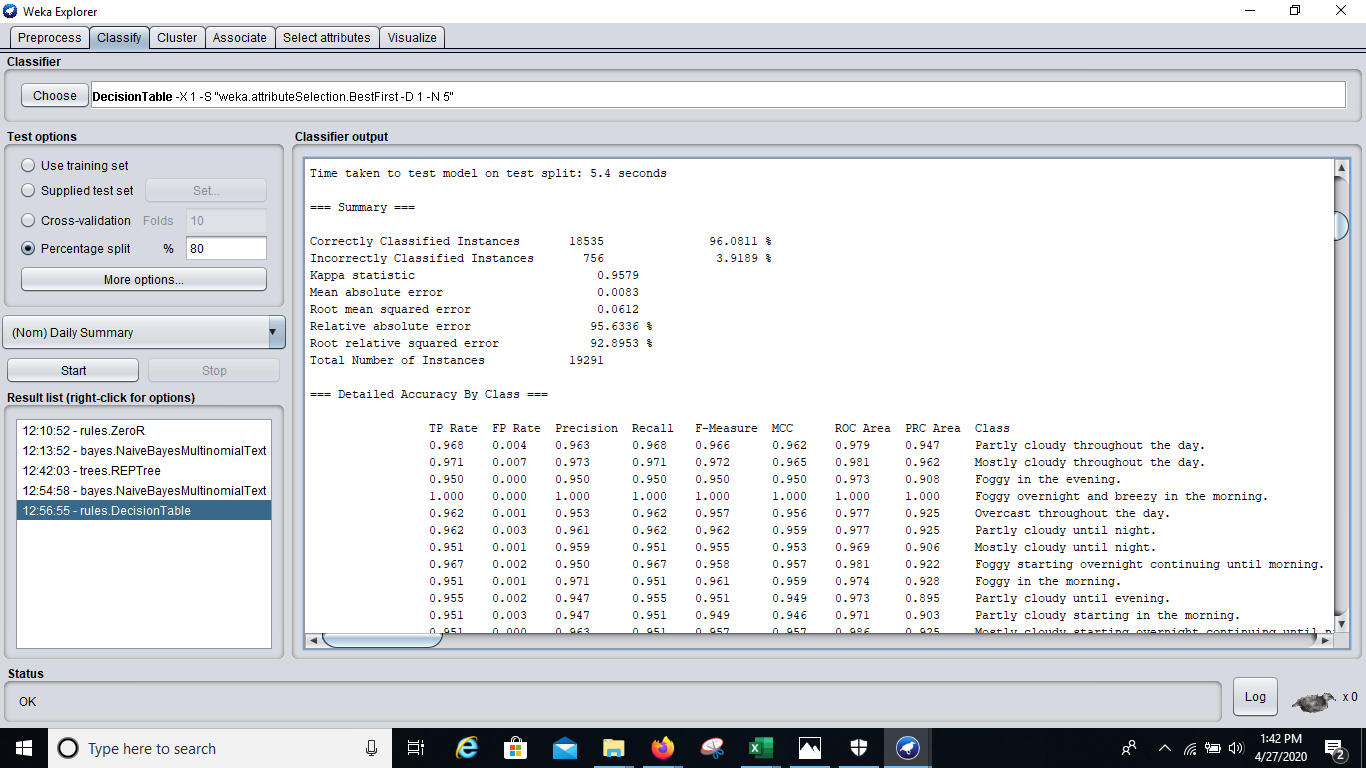
****

**Figure:-2** Output of decision table

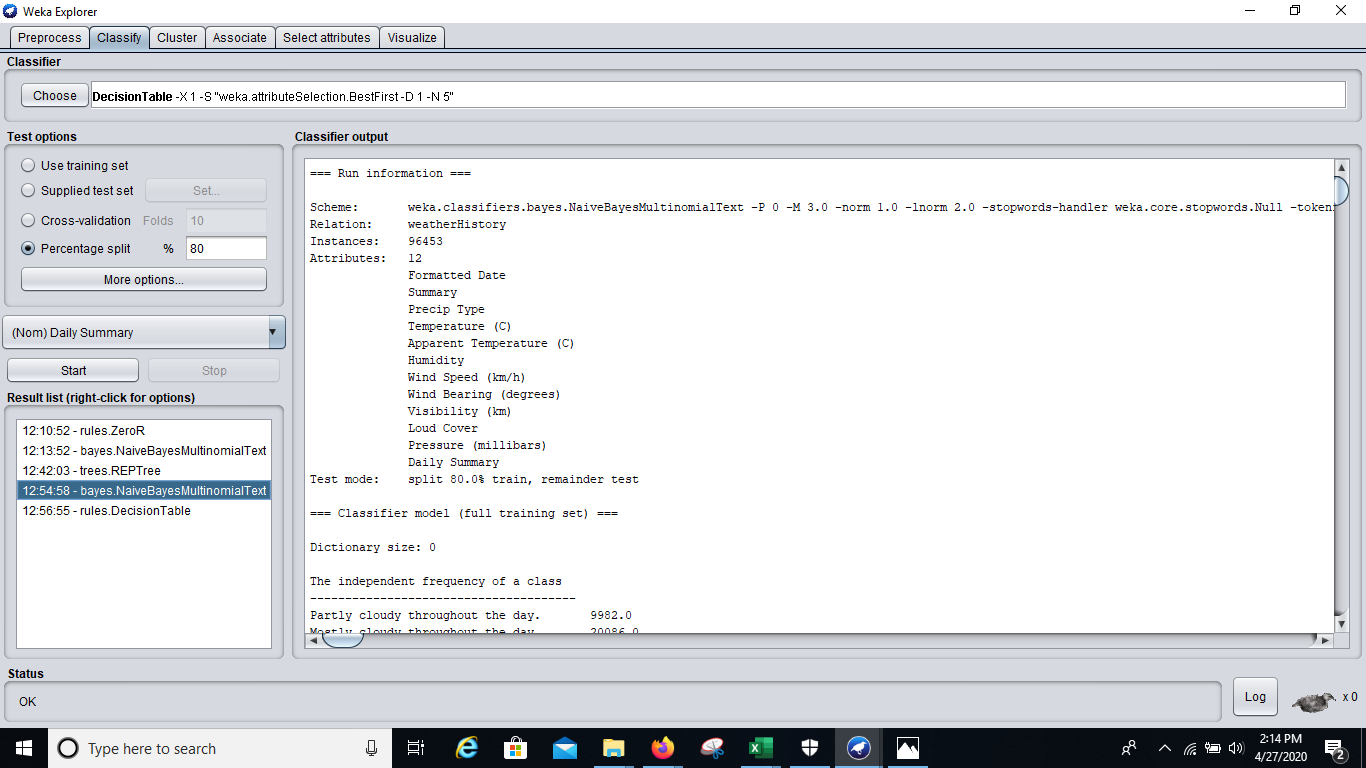
*Fig 2.1 : This shows the attributes, number of instances, scheme and the split in the dataset*

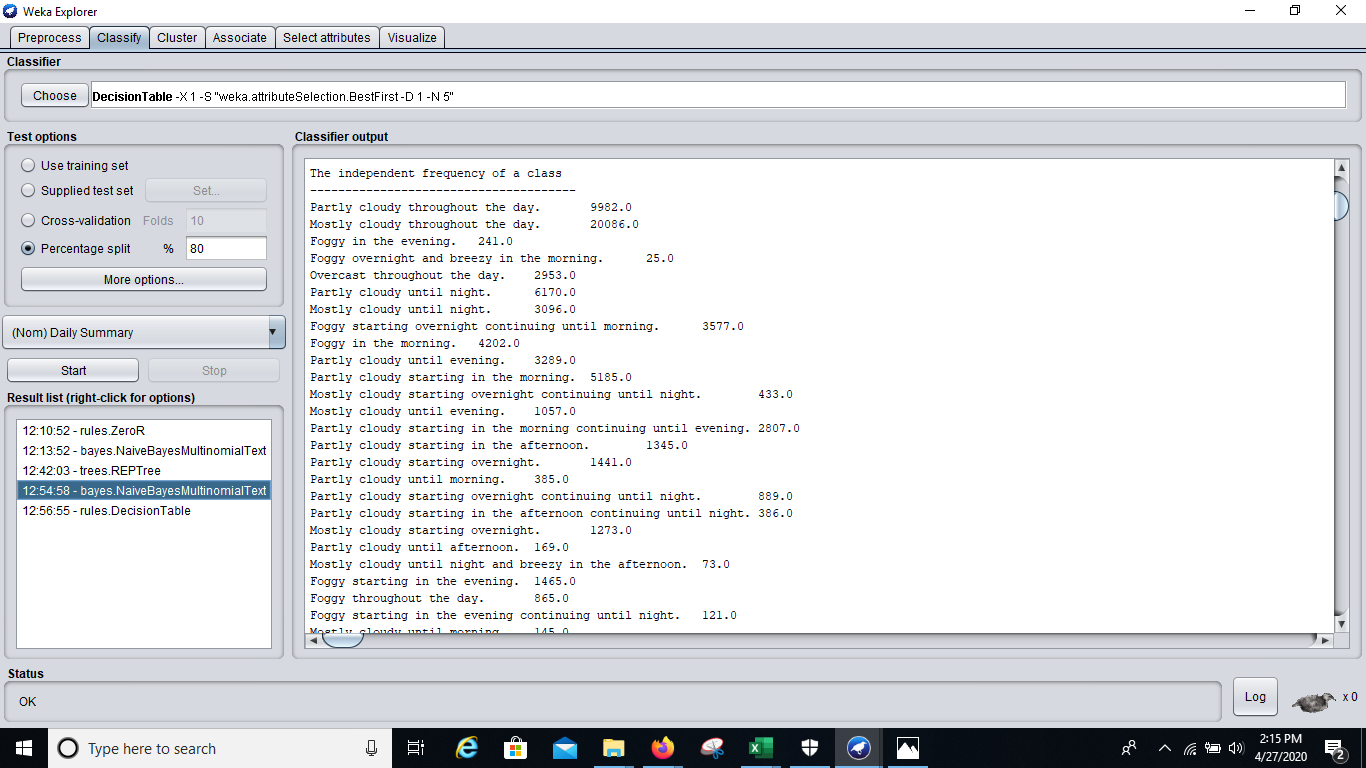


*Fig 2.2 : This shows the Output summary*

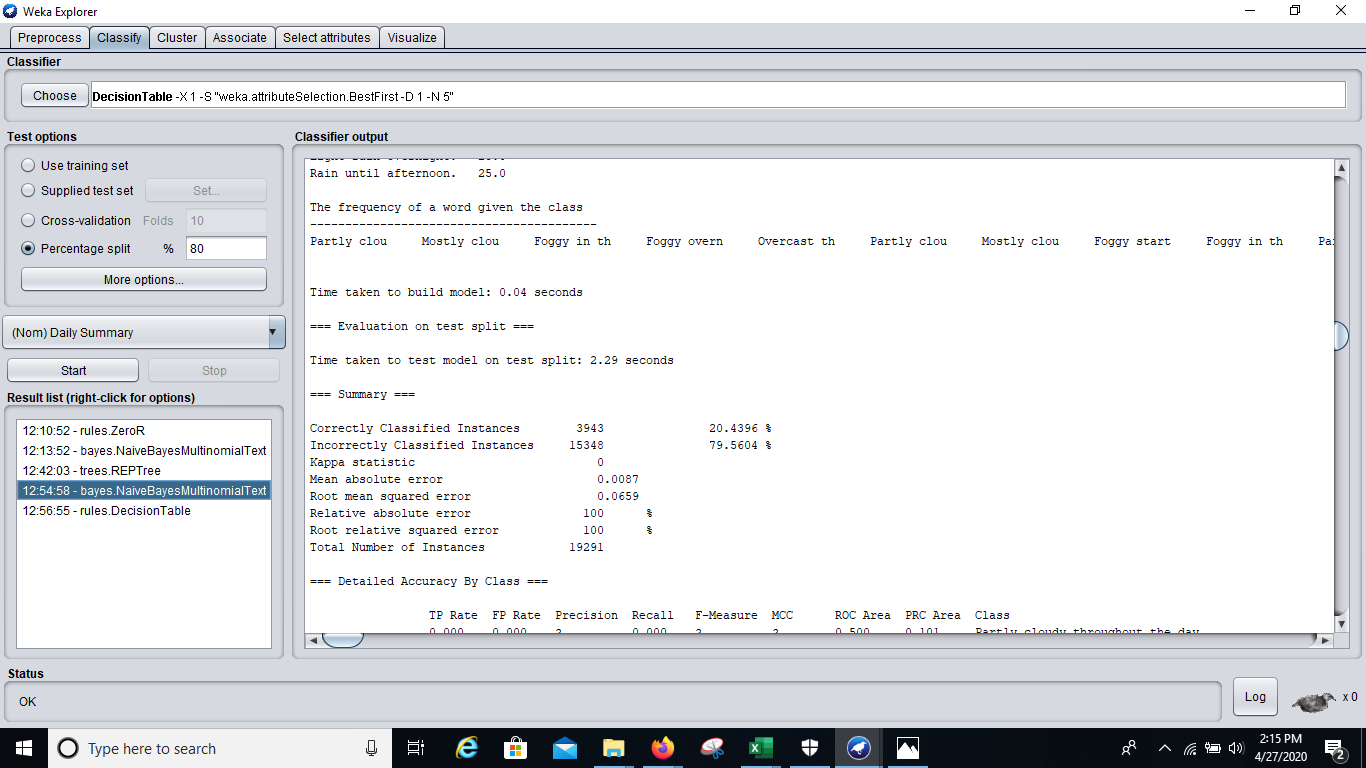


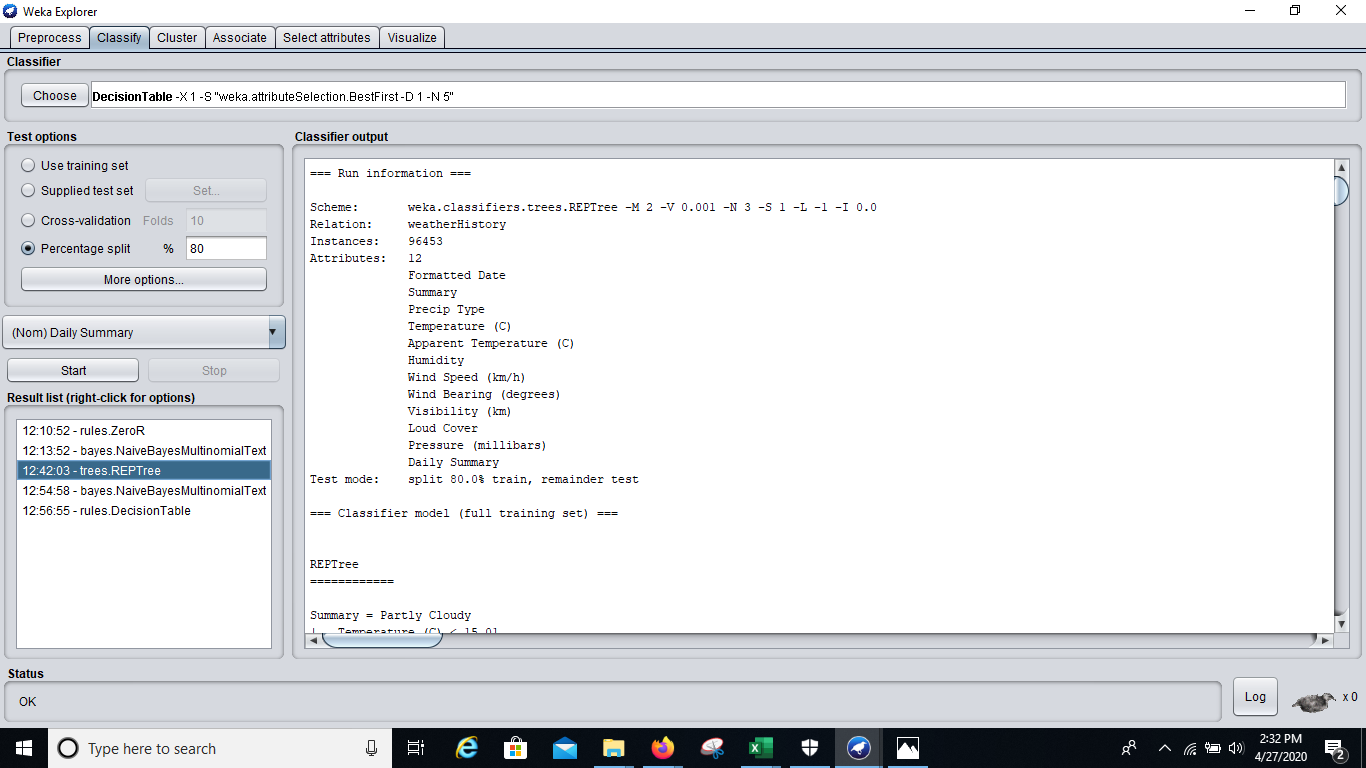
**Figure:-3** Output of Naive Bayes Multinomial Classifier Model

*****Fig:3.1 This shows the attributes, number of instances, scheme and the split in the dataset*

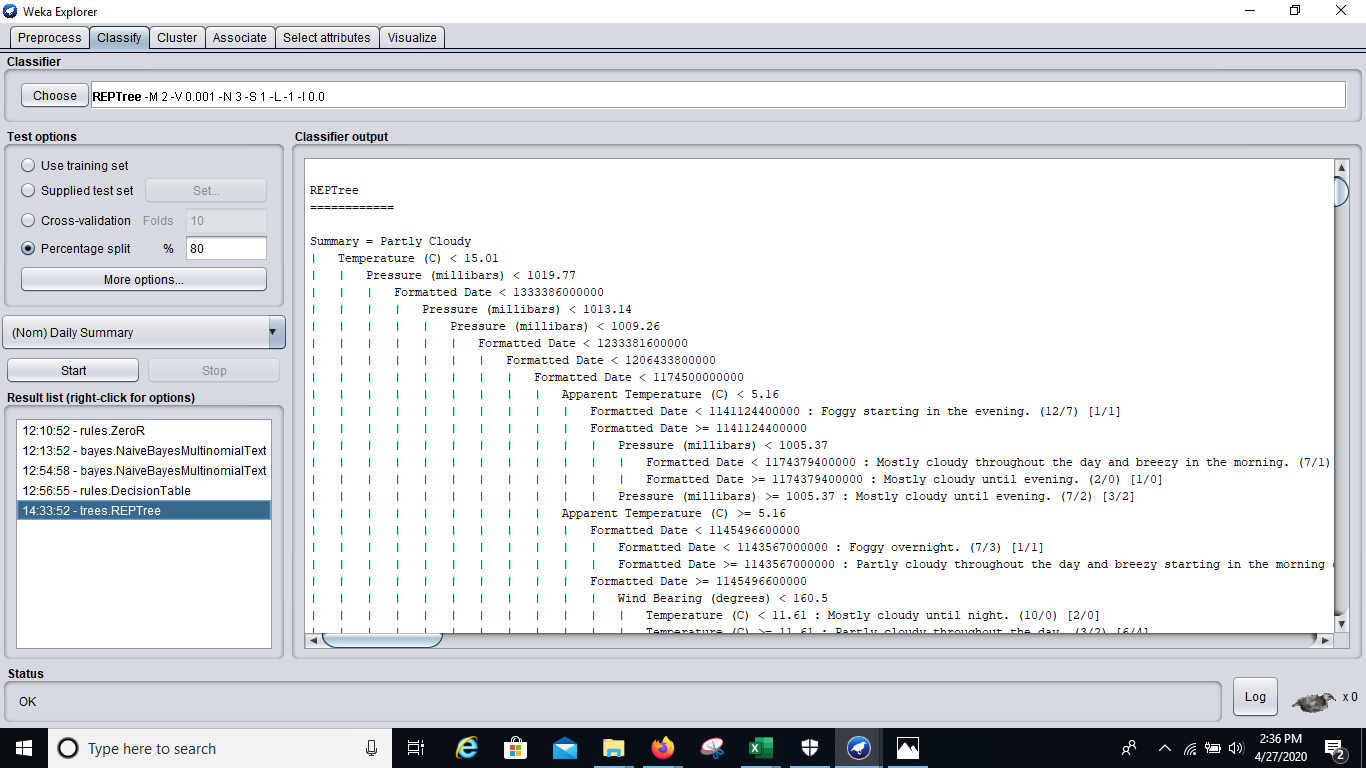
****

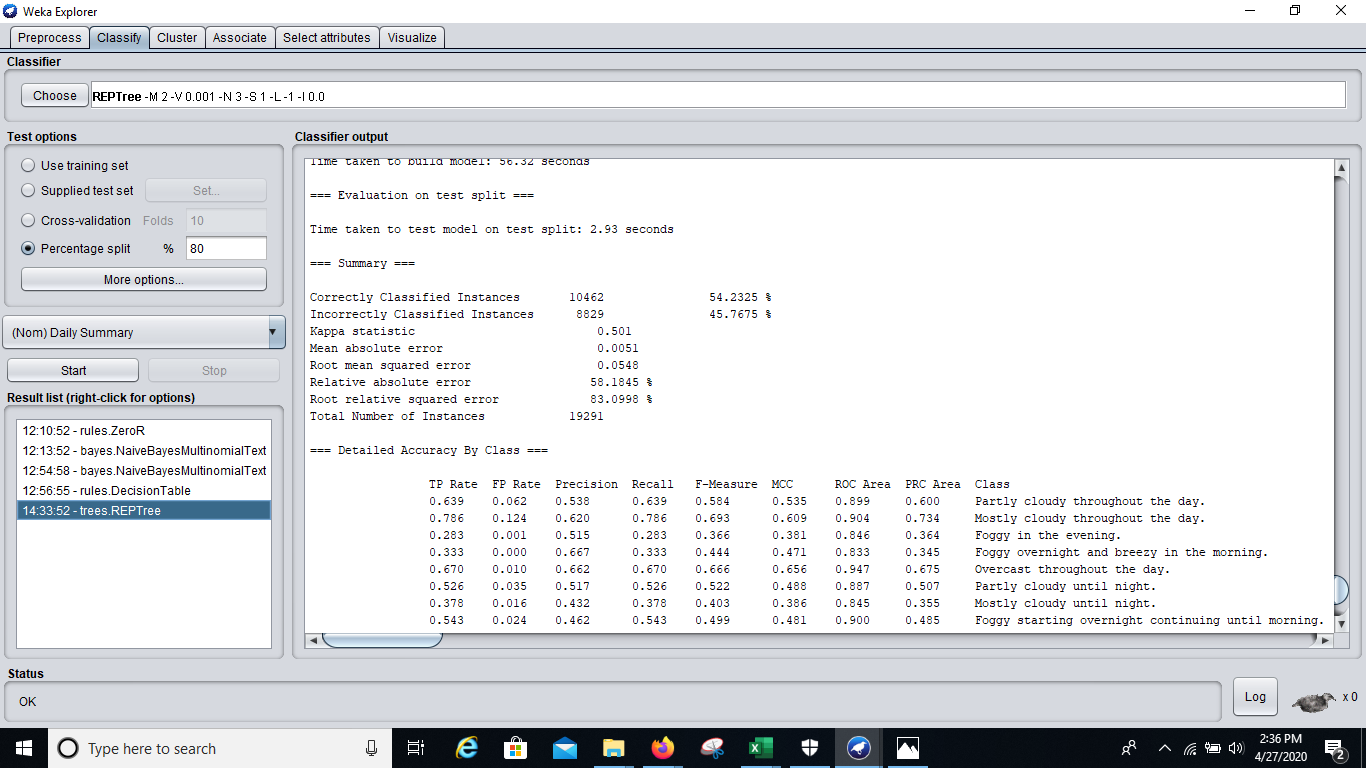
*Fig:3.2 This shows the frequency of each class in the ‘Daily Summary’*

*****Fig 3.3: This shows the Output summary*

**Figure:-4** Output of Decision Tree. ****

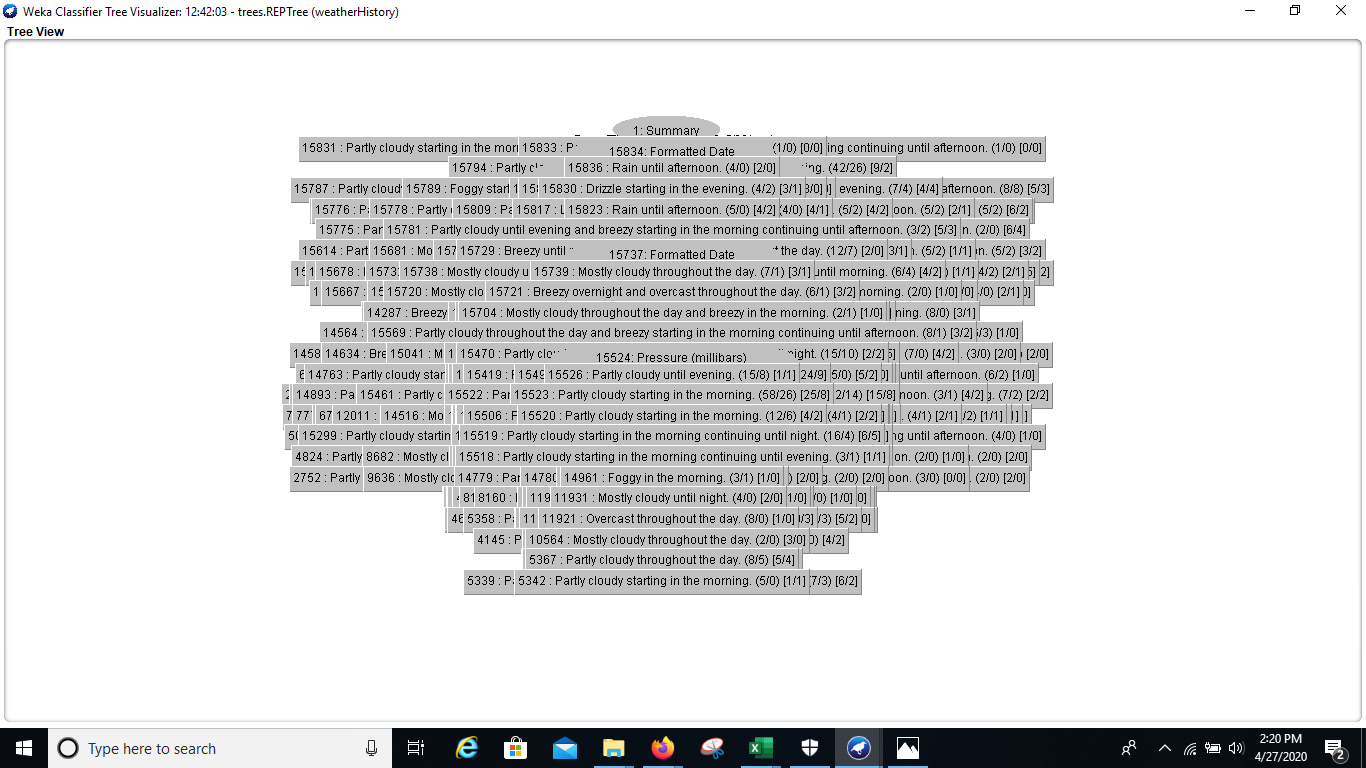
*Fig:4.1 This shows the attributes, number of instances, scheme and the split in the dataset*

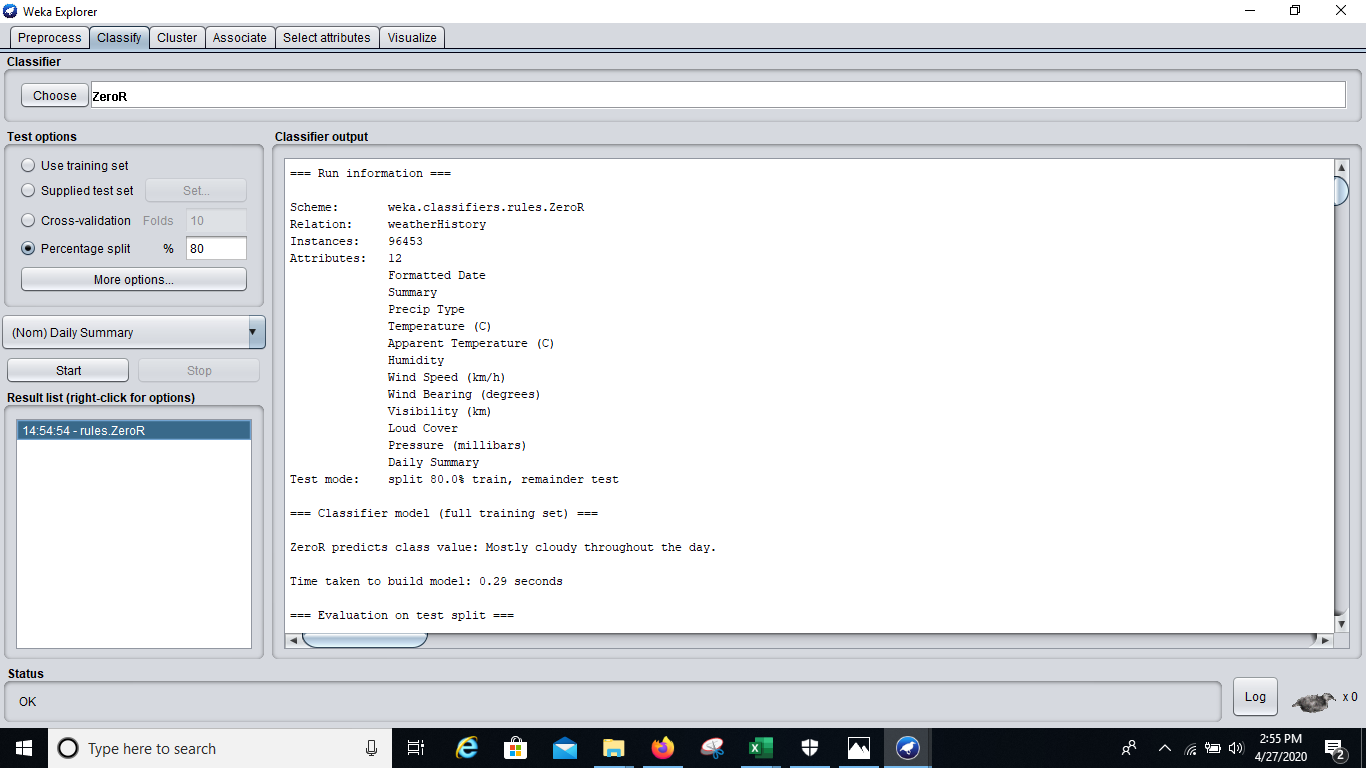
****

****

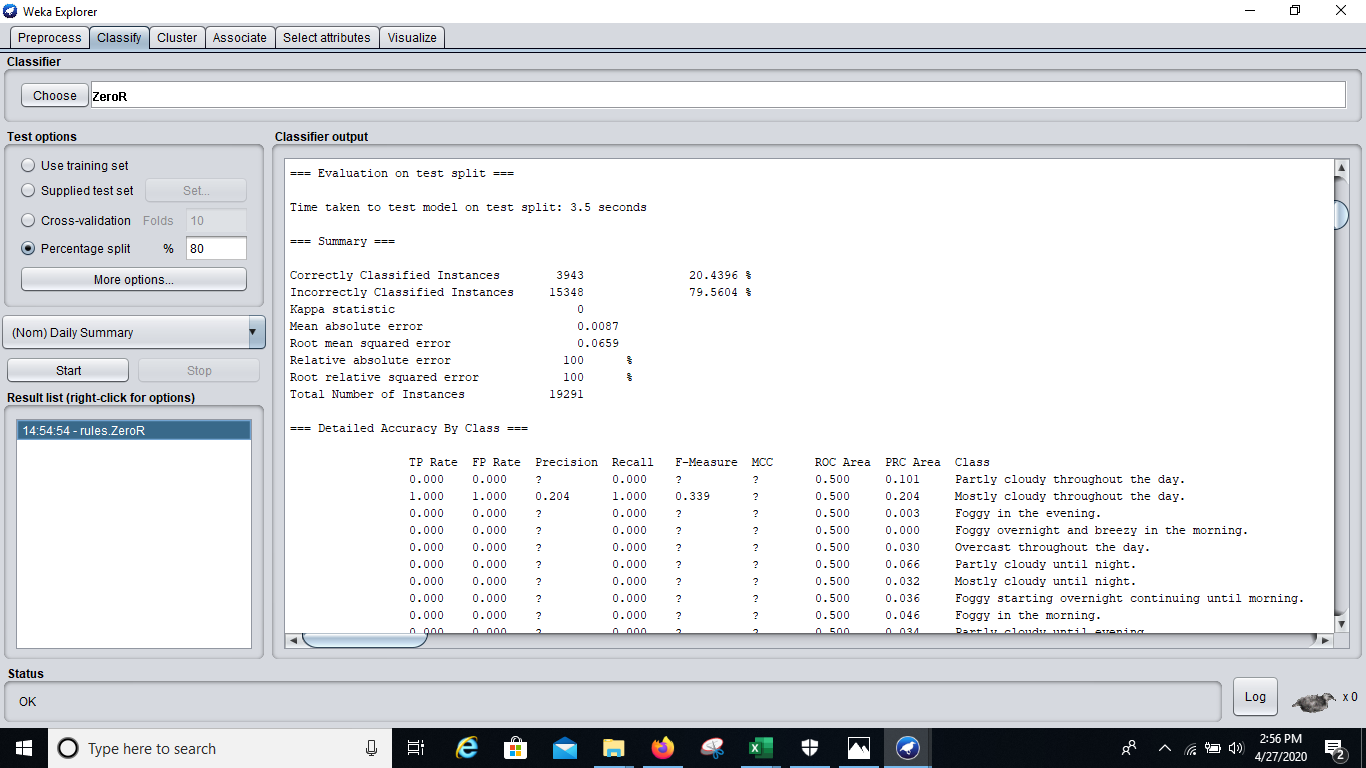
*Fig 4.2 : This shows the Output summary*

**Figure:-5** Tree View of Decision Tree.

****

**Figure:- 6** Output of ZeroR.****

*Fig:6.1 This shows the attributes, number of instances, scheme and the split in the dataset*

**** *Fig 6.2: This shows the Output summary*

**Chapter – 1 INTRODUCTION**

**1.1 MACHINE LEARNING**

Machine Learning is now one of the most important topics around the world. Well, it can even be said as the new electricity in today’s world. But to be precise what is Machine Learning, well it’s just one way of teaching the machine by feeding the large amount of data. Machine learning is defined in 90’s by Arthur Samuel described as the,” it is a field of study that gives the ability to the computer for self-learn without being explicitly programmed”, that means imbuing knowledge to machines without hard-coding it. And also “A computer algorithm/program is said to learn from performance measure P and experience E with some class of tasks T if its performance at tasks in T, as measured by P, improves with experience E.” -Tom M. Mitchell.

Machine learning is mainly focused on the development of computer programs which can teach themselves to grow and change when exposed to new data. Machine learning studies algorithms for self-learning to do stuff. It can process massive data faster with the learning algorithm. For instance, it will be interested in learning to complete a task, make accurate predictions, or behave intelligently.

**1.2 SIGNIFICANCE OF MACHINE LEARNING**

Data is growing day by day, and it is impossible to understand all of the data with higher speed and higher accuracy. More than 80% of the data is unstructured that is audios, videos, photos, documents, graphs, etc. Finding patterns in data on planet earth is impossible for human brains. The data has been very massive, the time taken to compute would increase, and this is where Machine Learning comes into action, to help people with significant data in minimum time.

It also helps in predicting the outcomes that were not possible in earlier days. For example weather forecasting has become more accurate and also we can make humanoids that can work in case of disaster where humans cannot enter. The ability to process very large amount of data within seconds cannot be done by a human. Decision making has also been improved.

**1.3 TYPES OF MACHINE LEARNING**

1. **Supervised learning: -** Supervised Learning is the first type of machine learning, in which labelleddata used to train the algorithms. In supervised learning, algorithms are trained using marked data, where the input and the output are known. We input the data in the learning algorithm as a set of inputs, which is called as Features, denoted by X along with the corresponding outputs, which is indicated by Y, and the algorithm learns by comparing its actual production with correct outputs to find errors. It then modifies the model accordingly. The raw data divided into two parts. The first part is for training the algorithm, and the other region used for test the trained algorithm. Supervised learning uses the data patterns to predict the values of additional data for the labels. This method will commonly use in applications where historical data predict likely upcoming events.

Ex:- It can anticipate when transactions are likely to be fraudulent or which insurance customer is expected to file a claim. We will focus more in this in this project.

1. **Unsupervised learning :-**  Unsupervised Learning is the second type of machine learning, in which unlabeled data are used to train the algorithm, which means it used against data that has no historical labels. What is being showing must figure out by the algorithm? The purpose is to explore the data and find some structure within. In unsupervised learning the data is unlabeled, and the input of raw information directly to the algorithm without pre-processing of the data and without knowing the output of the data and the data cannot divide into a train or test data. The algorithm figures out the data and according to the data segments, it makes clusters of data with new labels. This learning technique works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the primary qualities that separate customer segments from each other. These algorithms are also used to segment text topics, recommend items and identify data outliers.
2. **Reinforcement learning: -** Reinforcement Learning is the third type of machine learning in which no raw data is given as input instead reinforcement learning algorithm have to figures out the situation on their own. The reinforcement learning frequently used for robotics, gaming, and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the most significant rewards. This type of training has three main components which are the agent which can describe as the learner or decision maker, the environment which described as everything the agent interacts with and actions which represented as what the agent can do. The objective is for the agent to take actions that maximise the expected reward over a given measure of time. The agent will reach the goal much quicker by following a good policy. So the purpose of reinforcement learning is to learn the best plan.

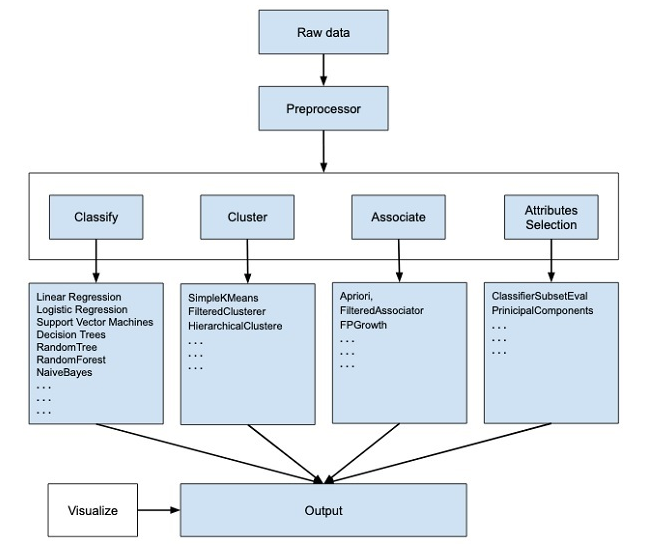
**1.4 Project Relevance**

In this project we have used the supervised learning, and in that we have used classification. Classification is used when we have different classes to which a tuple in a dataset may belong to. We have used a weather dataset from kaggle and in that there is an attribute ‘Daily Summary’ which will be predicted by our model once it gets trained on the dataset. The used dataset has many classes around 214. So since this dataset also contains around 96K instances from 1 Jan 2006 to 1 Jan 2017 of hourly weather report, it gives us a very good chance to work with such large dataset. In this project we will apply 4 different classifiers on this dataset to understand how classifiers work and what type of classifiers work the best for this type of large dataset thus will help us to achieve maximum accuracy on weather reporting.

**Chapter – 2 WEKA Tool**

**2.1 Introduction to WEKA**

WEKA is open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram –



While observing the beginning of the flow of the above figure, one can easily understand that there are many stages in dealing with Big Data to make it suitable for machine learning −

* Firstly, raw data is collected from the field. This data may contain several null values and irrelevant fields. There are various data preprocessing tools provided in WEKA to cleanse the data.
* Then, one could save the preprocessed data in some local storage for applying ML algorithms.
* Next, depending on the kind of ML model that is desired to be developed, one can select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.
* Note that under each category, WEKA provides the implementation of several algorithms. Selection of an algorithm of required choice is done, the desired parameters are set and it is run on the dataset.

Then, WEKA would give the statistical output of the model processing. It provides a visualization tool to inspect the data. The various models can be applied on the same dataset. One can also compare the outputs of different models and select the best that meets the desired purpose. Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

## 2.2 User Interfaces

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

* The ***Preprocess*** panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for preprocessing this data using a so-called ***filtering*** algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
* The ***Classify*** panel enables applying classification and regression algorithms (indiscriminately called ***classifiers*** in WEKA) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
* The ***Associate*** panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
* The ***Cluster*** panel gives access to the clustering techniques in WEKA, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
* The ***Select attributes*** panel provides algorithms for identifying the most predictive attributes in a dataset.
* The ***Visualize*** panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

**Chapter – 3 Classifiers**

* 1. **Decision Table Classifiers**

Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees. Selection to explore decision tables because it is a simpler, less compute intensive algorithm than the decision-tree-based approach.

* The algorithm, decision table, is found in the WEKA classifiers under Rules. The simplest way of representing the output from machine learning is to put it in the same form as the input.
* It summarizes the dataset with a “decision table‟ which contains the same number of attributes as the original dataset.
* The use of the classifier rules decision table is described as building and using a simple decision table majority classifier.
* The output will show a decision on a number of attributes for each instance. The number and specific types of attributes can vary to suit the needs of the task.
* Decision Table classifier algorithm is used to summarize the dataset by using a decision table containing the same number of attributes as that of the original dataset.
* A new data item is allocated a category by searching the line in the decision table that is equivalent to the values contained in the non-class of the data item.
* Decision Table are one of the simplest hypothesis spaces possible and usually they are easy to understand. Decision Table builds a decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation.
  1. **Decision Tree Classifier**

Decision Tree algorithm is another machine learning algorithm, which belongs to the supervised learning algorithm. This is one of the most popular machine learning algorithms. It can be used for both classification and regression problems. In the decision tree algorithm, we can solve the problem, by using tree representation in which, each node represents a feature, each branch represents a decision, and each leaf represents the outcome.

In the decision tree, we start from the root of the tree and compare the values of the root attribute with record attribute. On the basis of this comparison, we follow the branch as per the value and then move to the next node. We continue comparing these values until we reach the leaf node with predicated class value.

## Advantages of the Decision Tree

* It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
* It can be very useful for solving decision-related problems.
* It helps to think about all the possible outcomes for a problem.
* There is less requirement of data cleaning compared to other algorithms.

## Disadvantages of the Decision Tree

* The decision tree contains lots of layers, which makes it complex.
* It may have an overfitting issue, which can be resolved using the **Random Forest algorithm.**
* For more class labels, the computational complexity of the decision tree may increase.
  1. **Naïve Bayes Classifier Algorithm**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

* It is mainly used in *text classification* that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

* **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of colour, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
* **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

## Bayes' Theorem:

* Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm **Where,**

* **P (A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.
* **P (B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.
* **P (A) is Prior Probability**: Probability of hypothesis before observing the evidence.
* **P (B) is Marginal Probability**: Probability of Evidence.

### Advantages of Naïve Bayes Classifier:

* Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
* It can be used for Binary as well as Multi-class Classifications.
* It performs well in Multi-class predictions as compared to the other Algorithms.
* It is the most popular choice for **text classification problems**.

### Disadvantages of Naïve Bayes Classifier:

* Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

### Applications of Naïve Bayes Classifier:

* It is used for **Credit Scoring**.
* It is used in **medical data classification**.
* It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner.
* It is used in Text classification such as **Spam filtering** and **Sentiment analysis**.
  1. **ZeroR Classifier in WEKA**

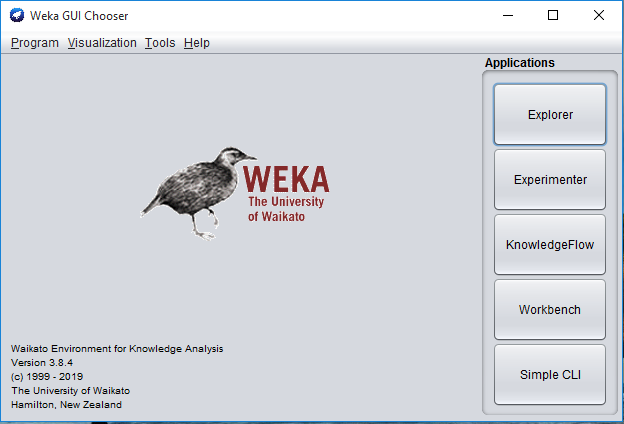
The simplest of the rule based classifiers is the majority class classifier, called 0-R or ZeroR in WEKA.

* The 0-R (zero rule) classifier takes a look at the target attribute and its possible values.
* It will always output the value that is most commonly found for the target attribute in the given dataset.
* 0-R as its names suggests; it does not include any rule that works on the non target attributes. So more specifically it predicts the mean (for a numeric type target attribute) or the mode (for a nominal type attribute).
* Thus, we can say, ZeroR is the simplest classification method which relies on the target and ignores all predictors.
* ZeroR classifier simply predicts the majority category (class).
* Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

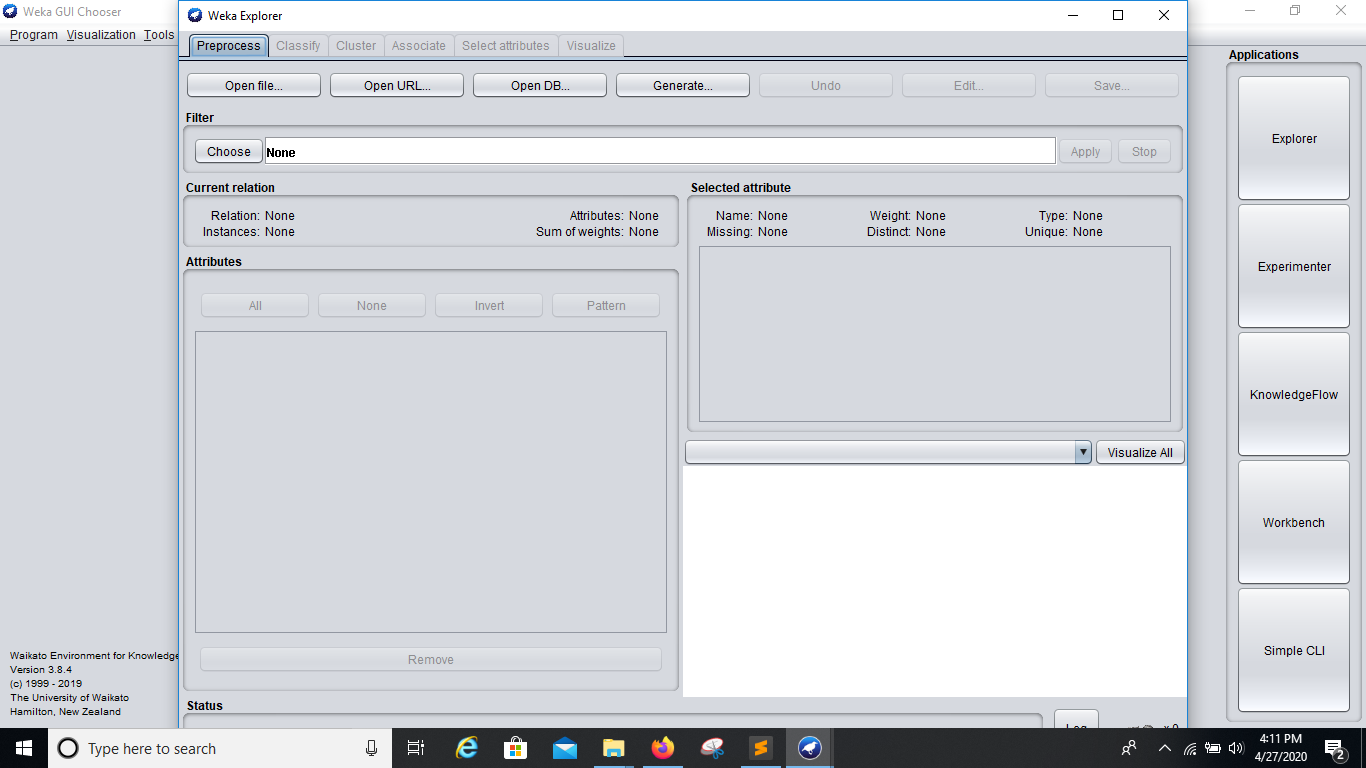
**Chapter – 4 IMPLEMENTATION**

**4.1 Steps to be performed in WEKA software**

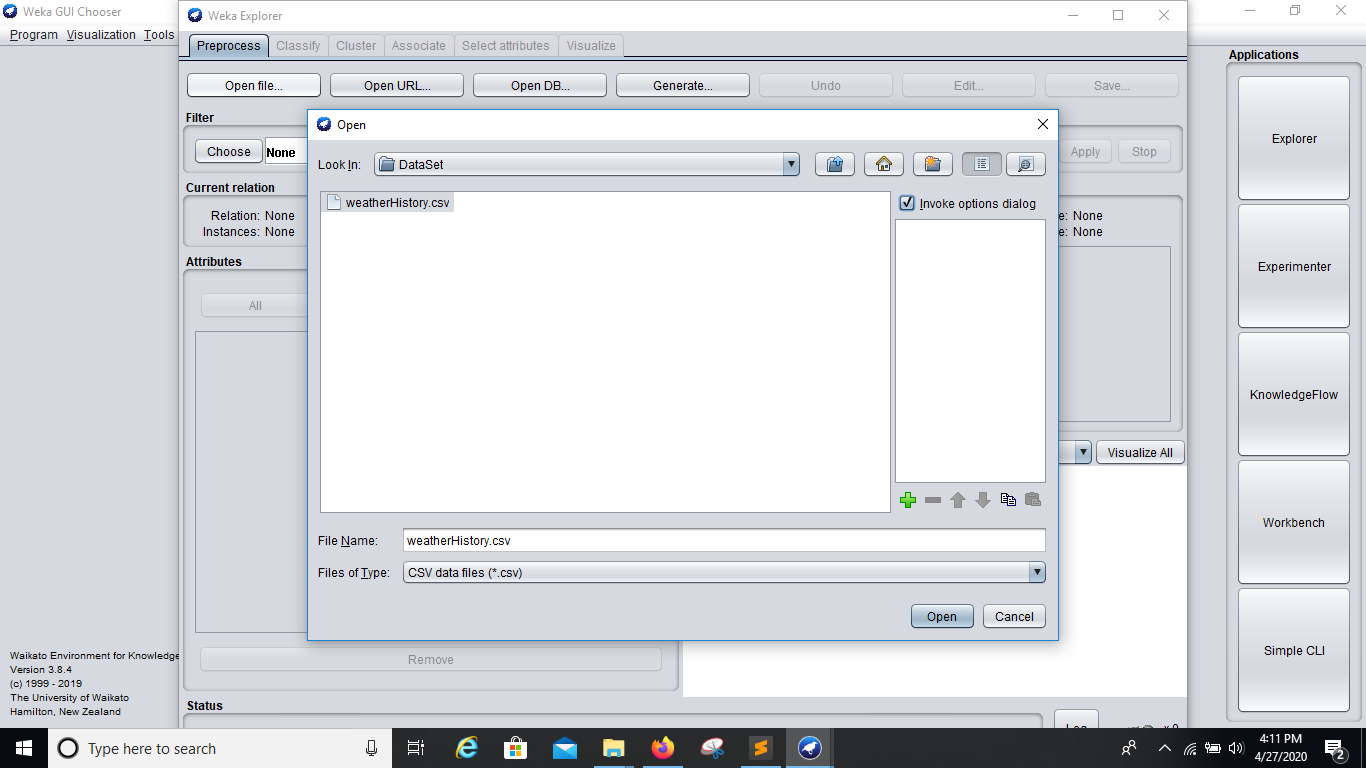
**Steps 1: -** The very first step is to open the WEKA application and go to the Explorer and click on that.



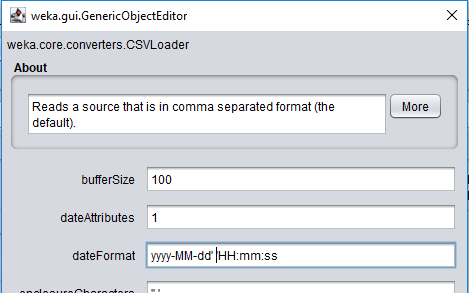
**Step 2: -** The next step is of preprocessing and that includes importing the dataset and also do the necessary formatting in the attributes or in the data.



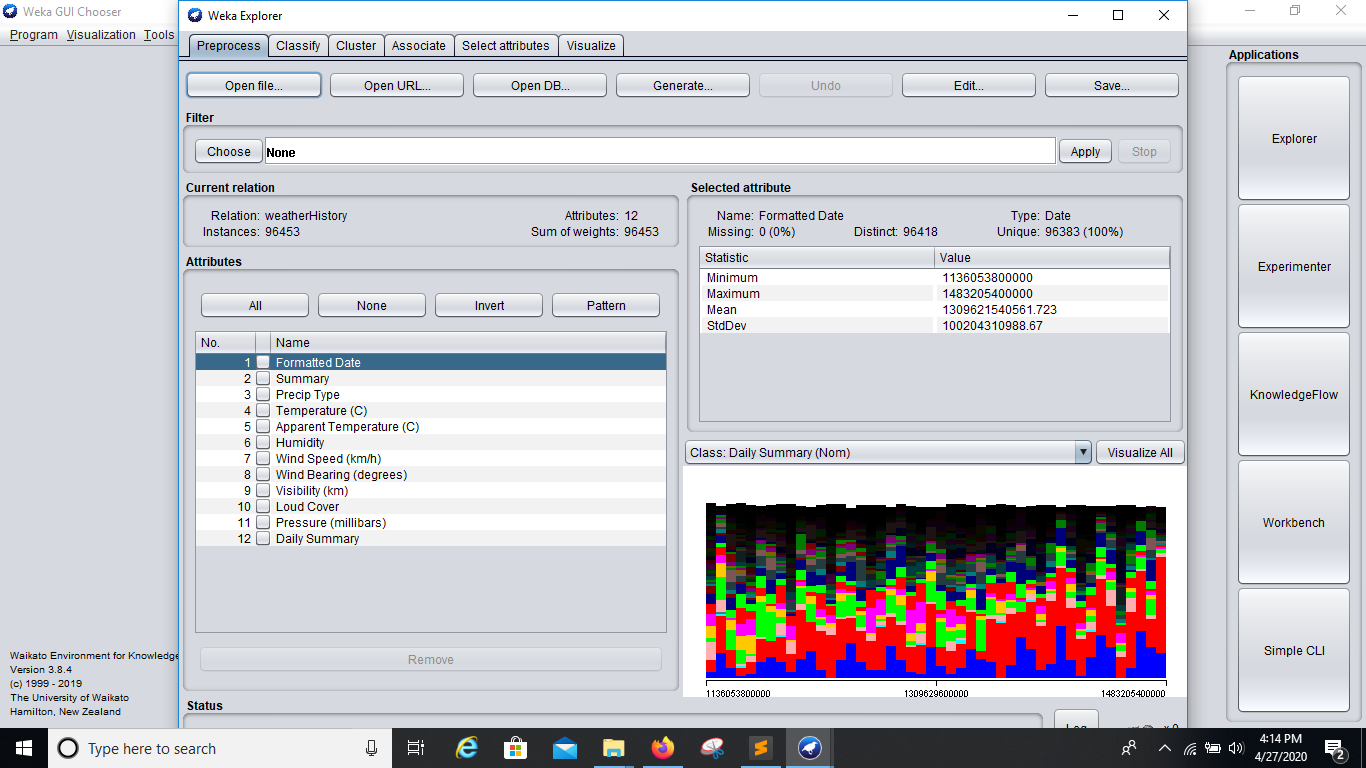
**Step 2.1: -** As shown in the pic go the file location and choose the file, we have chosen the weatherHistory.csv dataset and before clicking the OK button, click on the ‘invoke options dialog’. This is an important step as if you don’t click this option then the dataset will directly get imported and we will not be able to do some required changes. Thus now click on the OK button.



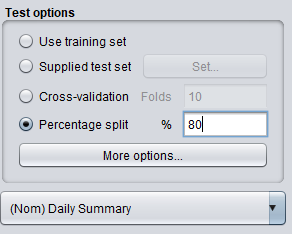
**Step 2.2: -** A new window will get opened up and in this window we will be doing two changes. These changes are done so that WEKA parser can detect the date Time stamp in the dataset as Type: date and not Nominal. So as shown in the Image do the required changes.



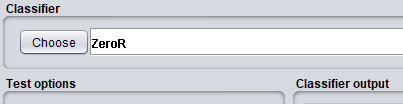
**Step 2.3: -** On clicking the Ok button, a window would appear as shown in the figure below:



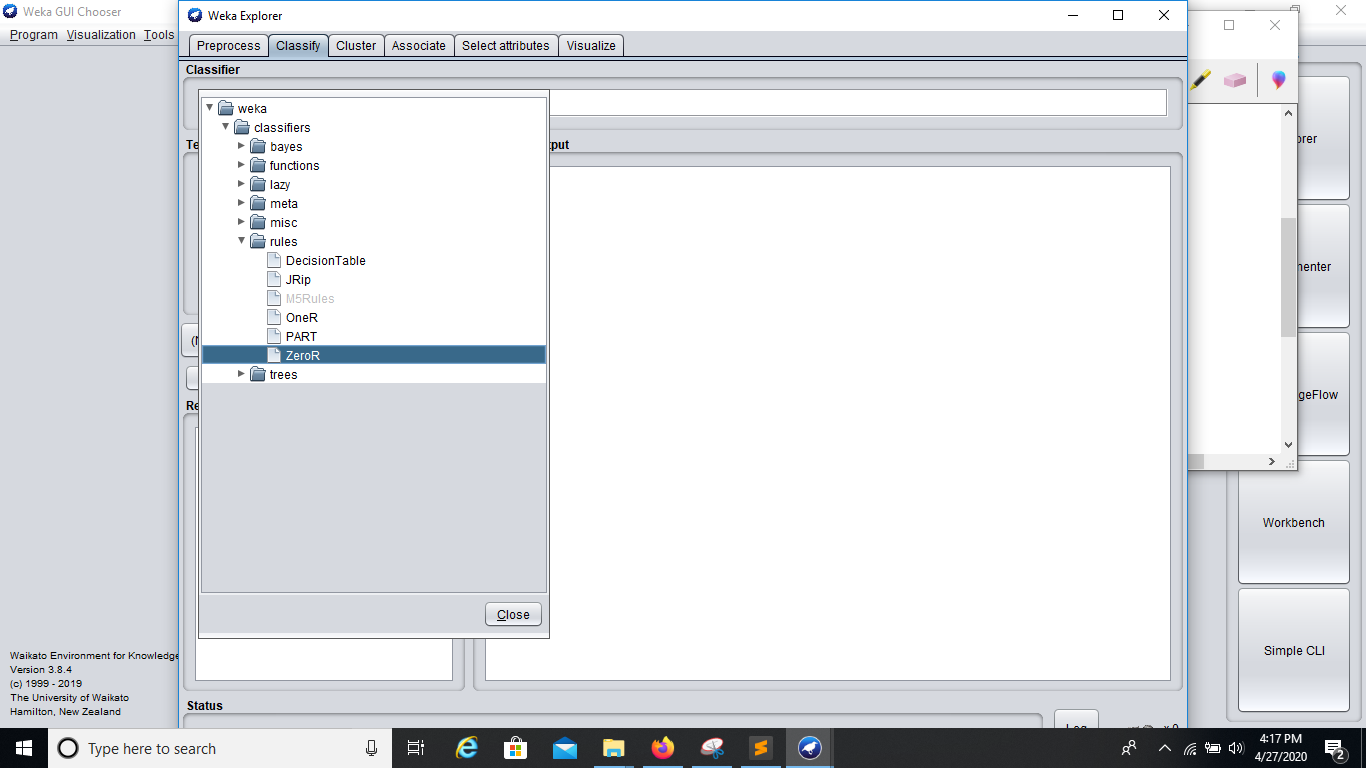
**Step 3: -** Now the preprocessing is done, the next step is to classify the data. So click on the classify tab in WEKA. We have used the last column ‘Daily Summary’ as our class. So set the Class as ‘Daily Summary’. Also change the Test Options. These changes are shown in the below Image.



**Step 3.1: -** In this step we choose the classifier for the data. In this project we have used 4 classifiers; these 4 classifiers are already explained in the previous chapters. So go to choose option under the Classifier, as shown in the image.



**Step 3.2: -** Choose the classifiers one at a time as shown in the image.



**Step 4: -** After all this step click on the start button and let the things happen normally. Outputs will be shown in the Classifier Output window on the right hand side. These outputs are shown in the List of Figures shown earlier.

**4.2 Description of Output summary**

On the bases of correctly classified instances and incorrectly classified instances, we have split the data in 80-20 ratio between train-test. The total number of instances is 96453. Out of these 19291 instances are reserved for the testing of model as according to the split.

* For Naive Bayesian classifier, the
  + Correctly classified instances : 20.44%
  + Incorrectly classified instances : 79.56%
* For Decision table classifier,
  + Correctly classified instances : 96.08%
  + Incorrectly classified instances : 3.92%
* For Decision Tree (REPTREE) classifier , the
  + Correctly classified instances :54.23%
  + Incorrectly classified instances :45.57%
* For ZeroR classifier , the
  + Correctly classified instances :20.44%
  + Incorrectly classified instances :79.56%

So according to the results the Decision Table classifier works best out of these classifier by giving a very high level of accuracy in the test instances.

**Conclusion**

Weather is something that cannot be predicted accurately even if we choose the highly capable computers in the world in parallel. The main reason is that it depends on so many factors that it is very difficult to get those factors counted. For example, the happening of rain does not only depend on the movement of clouds or density of clouds but it also depends on the altitude of the region from where clouds are passing, the speed of wind, atmospheric pressure, temperature and so on. So, research and computational models in this field is very common and it always gives new results. In this project we took this challenge of obtaining the accuracy as high as possible, as well as; at the same time comparing the different classifiers that are often used in classification. We used 4 classifiers that are Naive Bayesian, Decision Tee (REPTREE), Decision Table and ZeroR.  These classifiers were given the task to predict the ‘Output Summary’, for example ‘partly cloudy throughout the day’. Out of these classifiers, ZeroR and Naive Bayesian performed very unsatisfactorily with only 20% accuracy whereas the Decision Tree has given about 54% accuracy and Decision Table worked extremely well with an accuracy of about 96%. Thus, we conclude that Decision table classifier worked the best out of these 4 classifiers. This conclusion does not depict that this classifier will always work the best but for this type of data, it has shown the best result.

Our future work will be based on working with the random forest , random tree and J48. These algorithms of classification are expected to work very well with large set of data and are quit advance.

**REFRENCES**

* *https://www.tutorialspoint.com/weka/what\_is\_weka.htm*
* *https://www.kaggle.com/muthuj7/weather-dataset/data*
* *https://forums.pentaho.com/threads/160812-how-to-deal-with-quot-a-date-attribute-quot-in-weka/*
* *https://www.tutorialspoint.com/weka/weka\_classifiers.htm*
* *https://www.javatpoint.com/classification-algorithm-in-machine-learning*
* *https://www.javatpoint.com/data-science*
* *https://www.geeksforgeeks.org/basic-concept-classification-data-mining/*
* *https://www.javatpoint.com/machine-learning-naive-bayes-classifier*
* *http://www.ijesi.org/papers/Vol(7)i2/Version-3/A0702030109.pdf*
* *https://www.saedsayad.com/zeror.htm*