CONTENT BASED DATA MINING AND ANALYSIS FOR WEATHER RELATED WEB DOCUMENTS

Ms. Thushika Nishatharan
169338 U

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka, for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology

February 2019

DECLARATION

We declare that, this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged. Due references have been provided on all supporting literatures and resources.

Thushika Nishatharan

Date: Signature of Student:

Supervised By:

Mr. S. C. Premaratne

Senior Lecturer

Faculty of Information Technology

University of Moratuwa

Date: Signature of Supervisor:

ACKNOWLEDGEMENT

I want to acknowledge and express my appreciation to those efforts and support in one way or other contributed to the successful completion of this research.

First, I am deeply grateful for the supervision throughout the one year of my research and the helps received from my supervisor Mr. S. C. Premaratne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa. I have learned so much from our project discussions. He is willing to motivate me to contributed tremendously to do my research.

Besides I would like to thank all academic staff from the Faculty of Information and technology, who shared their vast knowledge throughout these two years by providing me a good environment which influenced a lot to achieve this goal.

It is with great pleasure to thank the University of Moratuwa, Sri Lanka, for all its efforts and facilities that it has taken to contribute towards this postgraduate programme. Especially, I would like to thank my colleagues for their support to complete this research in many ways.

I am as ever, especially indebted to my parents, husband, and brother for their love and support throughout my life to improve my career

ABSTRACT

More than two decades, there is a number of weather-related websites are available which approximately predict the weather and climate. By extracting important data from the websites, a predictive data pattern can be produced to show the next day's weather is with rain or not.

By applying different types of web mining and analyzing techniques those extracted weather-related data can be visualized to a common pattern for weather forecasting with the main deciding factors of weather. With the use of these approaches, reasonably precise forecasts can be made up to about four to five days in advance. For the weather prediction analysis, we need to discover deciding factors of the next day's weather. Particularly, common weather dependent factors and the relationship of the prediction to the particular phenomenon

The solution proposed by this research can be used to analyze a large amount of weather data which are in different forms in each source. By using predictive mining task our solution allows us to make predictions for future instances according to the model what we have created. Evaluation measurements for the selected data mining technique such as accuracy percentage, TP & FP Rate, Precision, F-Measure, ROC area, SSE, and loglikelihood for classification and clustering leads to create a high quality model of prediction.

Knowledge flow interface provides the data flow to show the processing and analyzing data with precise association rules. In order to evaluate the model, SSE values and time to build the model, are considered in an effective manner.

TABLE OF CONTENTS

ABBREVIATIONS

KDD	Knowledge Discovery in Databases
WEKA	Waikato Environment for Knowledge Analysis
GUI	Graphical User Interface
TP	True Positive
FP	False Positive
SSE	Sum of Squared Error
ROC	Receiver Operating Characteristics

LIST OF FIGURES

LIST OF TABLES

CHAPTER 01

Introduction

1.1 Prolegomena

Extracting necessary and needed information from web pages are such an important task in this tech-covered world. However, the tools available for gathering, establishing, and distributing web content have not kept step with the rapid growth in information. So, the key analysis needed when web documents are in precisely content related. Nowadays, there is a number of weather-related websites which approximately predict the climate and weather. Both seasonal and reginal variability in weather directly influences in many fields like agriculture, tourism, disaster management, aircraft and shipping. By considering temperature, rainfall, evaporation, wind direction, wind speed, humidity, and air pressure most of the weather predictions are predicted and displayed. We have conducted a research to offer in depth analysis of weather prediction patterns using large data which are extracted from websites.

1.2 Background & Motivation

In this modern world, web development and its applications play an exciting role in everyone's day to day life. So, the world wide web is moving into a suitable atmosphere, where all kind of users can able to acquire important information quickly and easily as they need. Web mining is primarily about using the web, the web structure, and exploring web content. Web Structure Exploration attempts to discover the underlying model of web link structures. This template can be used to categorize web pages and is useful for generating information such as the similarity and relationship between different websites. Web usage mining using web effects usage derived data detect user behavior patterns to automatically access web services. Web content mining is a kind of mining technique, which mines extracted information from selected web page elements.

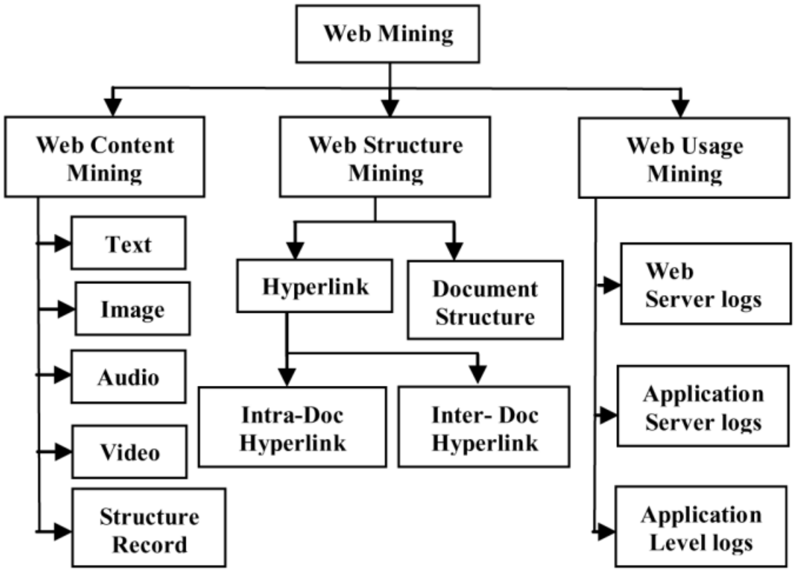


Figure 1.1 Web Mining Categories

Web content crawling is the analysis and exploration of text, images, and graphics on a web page to determine the importance of the content for the search query. This analysis is after the cluster in Web and the results with the huge amount of information available on the World Wide Web. Content mining provides search engine results lists in order of relevance to the keywords in the query [1]. Exploring web content with associated methodologies to extract information. These are the extraction of unstructured content, organized mining extraction, semi-organized content extraction and multimedia extraction.

The rapid development processes in web content mining have been initiated from the past few years. There are two approaches mainly considered within the web substance mining: Agent-based approach and Database approach [2]. The first approach is to improve the search and filtering of information the second approach to modeling data in the field of processing and analysis of data mining applications.

However, the processes available for gathering, forming and sharing web content have not kept step with the rapid growth in information. But the major complexity arises when web documents or information is not in a particular type of content. Due to the heterogeneity and the lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems

Extracting the content of the document and analyzing through text means is quite involved as both syntax and semantics are needed for this. The objective of this research is to achieve a concept-based term analysis regarding the weather on the sentence and document levels rather than a single-term analysis in the document set.

1.3 Research Problem Statement

The quality of web document clustering can be enhanced by dropping the noise in the data by pre-processing the structure of data representation and also by applying different clustering techniques.

1.4 Aim and Objective

1.4.1Aim

The aim is to achieve concept-based content analysis for web documents related to weather using web mining techniques and intend to afford an analysis study with our results of weather dataset.

1.4.2 Objectives

The main objectives of this research are:

- To collect planned weather-related web documents as data set.
- To identify the features using suitable procedures.
- To improve the clustering algorithm to identify weather-based content patterns from the dataset.
- To determine general weather content patterns at individual websites as well as across multiple sites.
- To improve classification algorithm to predict new weather content-based web sites on previous weather trends.
- To evaluate and analyze common weather-based patterns to access similar websites.

1.5 Overview of the Report

This chapter provides the introduction for web content mining for web pages using data mining techniques. The next chapter describes similar related work done by others. Chapter three explains technology adapted and Chapter four presents the approach of the research. Then chapter five elaborates the analysis and design of the research, meanwhile, chapter six is about the implementation part of the research component. Chapter seven presents the discussion part and the last section provides all list of references.

1.6 Summary

Weather prediction analysis is highly important in climatology and many another day to day life activities. So, this research has been conducted to check the most suitable model for the particular dataset.

CHAPTER 02

State of the art of exploring issues in weather-related web content mining

2.1 Introduction

This chapter describes different types of web mining and analyzing those extracted weather-related data to find out the common pattern for weather forecasting. Especially this chapter emphasizes on main deciding factors of weather on weather-content webpages. Furthermore, this chapter summarizes most of the web content mining approaches with their pros and cons, and list of different algorithms to apply data mining techniques also contained within this chapter.

2.2 Web mining techniques

In 1996, the idea of web mining was initiated by Etzioni. Web content mining has the associated procedures to explore necessary information. They are, Unstructured content mining, organized mining, semi-organized content mining and multimedia mining. The examination around applying information mining procedures to unstructured content is termed Knowledge Discovery in Texts (KDT), or content information mining.

Information extraction, topic tracking, summarization, categorization, clustering and information visualization are utilized methods in content mining. Grouping approaches have been classified. This is an old document two documents, many common words, it is possible that the two documents are very similar.

The requirement of retrieving various tools for data processing at web level which helps the user logically to transform these data into useful knowledge seems so important. Content mining techniques extract content models of Web object data, including plain text, semi-structured documents such as HTML, XML, structured documents, dynamic and multimedia documents. The extracted models are used to classify web objects, to recommend URLs or documents; to extract keywords for use in information retrieval and to infer the structure of objects semi-structured or unstructured [3].

Clustering is one of the techniques in data mining and web mining. The text-based clustering approaches characterize each document according to its word contained in it. The key idea is that if two documents contain many common words or phrase then it is obvious that, both documents are possible to be very similar. Partitioned, Hierarchical, Graph-Based, Probabilistic algorithms are additional categorized accounting to the clustering method used into the category mentioned above [4].

The most important of which are single-link clustering technique, complete link, k-means, Average link, medium Distance, Group Average link, ward's, clustering algorithm of LBC which can be as basic methods for clustering web pages according to data mining various algorithms and techniques have been represented to data mining. Various algorithms and techniques have been represented for clustering web pages based on data content of which are Hard and fuzzy algorithm of clustering web pages document tarries based on key-words inside web pages and Cosine likeness measure [5] and clustering web pages based on the behavioral models of the users [6], clustering of web pages based on link structure between them which is based on the textual information in links [7], clustering web documentaries using neural networks based on key-words inside the documentaries [8].

In the field of document clustering, k-means is a simple unsupervised learning platform. K centers are defined for every cluster that is newly formed. While dealing with all substrings or phrases within a page Yamamoto and Church [4] provided us with a method of computing term frequencies (tf) and document frequencies for a set of documents by using the concept of suffix array.

Later within the same platform K.M Hammoudoet.al. [9] came up with the idea of a phrase-based document indexing model which is later named as Document Index Graph (DIG). This method incorporates the construction of phrase-based indexes for the available document set but in an incremental manner. Indexing phrases repeatedly within the graph using the DIG model had a bottleneck of space complexity thus the STC algorithm introduced by Etzioni et-al. [10]. STC is basically a linear time clustering algorithm which exactly means that the size of the document set is linear. It works with identifying such phrases that are common to clusters of documents.

Singular value decomposition (SVD) creates good results but widely used over short texts. Term-based document model has the drawback as it ignores the relation among individual words that it works on and thus not efficient nowadays. DIG has also become slow with the results with large space complexity. STC being a better option for clustering web documents still need the help of other techniques too to enhance its feasibility [11].

WEB CONTENT MINING			
Year	**Author**	**Representation**	**Method Handled**
2000	Nahm & Mooney	Bag of words	Decision trees
1999	Freitag & McCallum	Bag of words	Hidden Markov Models
1999	Hoffmann	Bag of words	Unsupervised statistical method
1999	Junker	Relational	Inductive Logic Programming
1999	Yang	Bag of words and phrases	Clustering algorithms K-Nearest Neighbour Decision tree
1999	Billsus & Pazzani	Bag of words	TFIDE Naive Bayes
1987	Genersereth & Nilsson	Set of objects	Ontology

Table 2.1. Web Content Mining using different algorithms

Traditional methods are partitioned into four parts based on the documents in the web. The techniques which are used for mentioned four types of web documents are tabulated below in the *Table 2*.2.

Web Document	**Techniques**
Unstructured	Information Extraction
	Topic Tracking
	Summarization
	Categorization
	Clustering
	Information Visualization
Structured	Web Crawler
	Wrapper Generation
	Page Content Mining
Semi-Structured	Using OEM (Object Exchange Model)
	Top Down Extraction
	Web Data Extraction Language
Multimedia	SKICAT: Based on astronomical data analysis and cataloging system
	Color Histogram Matching
	Multimedia Miner
	Shot Boundary Detection

Table 2.2 Techniques for Web Content Mining

2.3 Approach for weather forecasting analysis

Nowadays, weather forecasting contains a hybrid computer model, observation, and information of patterns and trends. With the use of these approaches, reasonably exact forecasts can be made up to about three to four days in advance. Beyond this, detailed forecasts are less useful, since some atmospheric conditions such as wind gust direction, temperature and wind direction are very difficult. The capabilities of retrieving and storing have increased, due to the latest industrial updates which resulting in the accessibility of enormous climatology dataset in various arrangements. These data are generated both from the surface observation stations and aerial study stations. With the increase in the number of weather stations, a huge amount of data is available on daily, weekly, monthly and yearly basis and the data is stored exponentially [12].

Anyhow, to analyze the particular data from this huge data, mining techniques influence a dynamic role. To have an effective prediction for effective results, it is necessary to identify the relationship between the attributes of weather, which indirectly have a role in the weather changes which affects the climate.

2.4 Summary

This chapter elaborates the findings of our literature study and summary of the previous related work. It shows different web mining techniques and effective data mining algorithms which are used previously to analyze a large set of data to achieve dissimilar multipurpose goals.

CHAPTER 03

Technology adapted in weather data analysis

3.1 Introduction

This chapter presents data mining technology which we selected to analyze weather related data in an effective and efficient way in detail. Moreover, chapter three gives the impression of the selected technology which perfectly adapted for our research. And also it presents the usefulness of data mining techniques that distinguish from the technologies applied in existing literature.

3.2 Involvement of Data Mining

The rapid growth of Information Technology has made massive number of databases and large data in different applications and fields. Database and information technology research have led to an approach to store and manipulate these valuable data for further decision-making. Data mining is a process of exploring necessary information and patterns from a large dataset. It is also called a knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis [14].

From statistics and artificial intelligence with database management, data mining techniques extract patterns from large data sets by combining different methods. It is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

The whole procedure of discovering necessary knowledge in raw data includes the consecutive line up of steps such as, developing an understanding of the application domain, making a particular data set based on an intelligent way of selecting data by focusing on a subset of variables or data samples, data cleaning and preprocessing, data reduction and projection, choosing the data mining task, choosing the precise data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge. Hence, Knowledge Discovery in Databases (KDD) is the initial step of data mining with the process of applying data analysis and discovery algorithms. The steps attached in the KDD process is described in Figure 2.1.

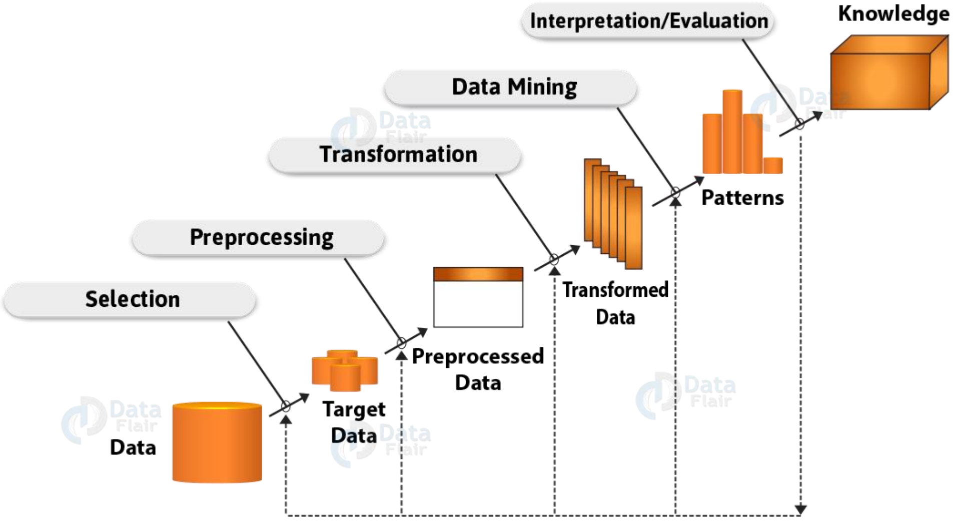


Figure 2.1 Overview of KDD process

i. Data Cleaning: It is defined as the removal of noise and irrelevant data from the collection.

- Cleaning in case of Missing values.
- Cleaning noisy data, where noise is a variance or noisy error.
- Cleaning with Data discrepancy detection and Data transformation tools.

ii. Data Integration: It is defined as heterogeneous data from multiple sources combined in a common source.

- With the aid of Data Migration tools.
- With the aid of Data Synchronization tools.
- With the aid of Extract-Load-Transformation process.

iii. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- By using Neural network.
- By using Decision Trees.
- By selection using Naive Bayes.
- By using Clustering, Regression.

iv. Data Transformation: Data transformation is defined as the process of transforming data into an appropriate form required by the scanning procedure. Data transformation is a two-step process:

- Data Mapping: Conveying elements from base source to destination to capture transformations.
- Code generation: Creation of the real transformation process.

v. Data Mining: Data mining is defined as best techniques that are applied to extract patterns potentially useful.

- Transforms task relevant data into patterns.
- Decides the purpose of the model using classification or characterization.

vi. Pattern Evaluation: Model evaluation is defined as the identification of strictly increasing patterns representing knowledge based on given measurements.

- Find the interestingness score of each pattern.
- Uses summarization and Visualization to make data understandable by the user.

vii. Knowledge representation: Knowledge representation is defined as a technique that uses visualization tools to represent the results of data mining.

- Reports generation.
- Tables generation
- Generation of classification, discriminant and characterization rules.

3.3 Major reasons for using data mining for weather prediction analysis

Data accompanied by weather patterns are large in size and very complexity, since deciding factors of weather are large in amount. Therefore, these data which kept in websites after doing different weather predictions time to time can be treated as Big Data. It is an important term given to data which is huge in volume, variety, and velocity. Previous implementations and analysis from literature review clearly indicate that data mining is the best option for data analysis among different other data analysis methods used.

While distinguishing data mining with traditional methods of querying a database, the latter require predefined variables and also poor visualization options available. Moreover, data mining is a complicated type of database querying which allows including new and large number of variables. Also, data mining allows to interpreting multiple areas simultaneously. That’s why the process of data mining consist of analyzing and summarizing dataset from many viewpoints and converting the whole data into valuable information. So, data mining techniques can be used as the most effective and suitable method for weather prediction data analysis according to our literature review.

3.4 Involvement of data mining for weather prediction analysis

For the weather prediction analysis, we need to discover deciding factors of the next day’s weather. Particularly, common weather dependent factors and the relationship of the prediction to the particular phenomenon. Weather factors represents the condition at a specific location on a specific day, while climate is the average conditions over many years. So, by analyzing weather prediction patterns, climate can be easily reported. It is very much necessary to find out major causes for rain and light rain shower. Therefore, this research utilizes different data mining techniques in training and testing category according to the research objective.

To analyze particular weather prediction primarily need to take data such as minimum temperature, maximum temperature, rainfall, evaporation, wind direction, wind speed, wind gust direction, wind gust speed, humidity, clouds, and air pressure which are available on the websites with the actual prediction. Then using data preprocessing techniques in data mining can reduce the uncertainty of data by removing erroneous data missing values. After that, by analyzing the deciding factors, major data fields can be chosen. To exactly find out the significant prediction and relationship between these data fields which contribute to rain day, light rain shower or no rain day, classification technique in data mining can be utilized. Revealing the hidden pattern behind rain cause can be effectively done through classification as it is a predictive mining task.

3.5 Summary

This chapter describes data mining as the technology proposed to analyze rain prediction to identify the patterns in weather. In this sense, it is pointed out how the data mining contributed an effective and precise solution for weather prediction analysis. The next chapter shows the approach of analyzing weather predictions through technology adopted here.

CHAPTER 04

A novel approach for analyzing weather predictions

4.1 Introduction

Chapter three discussed the technology adopted for our analyzing to identify the patterns in weather. This chapter presents our approach to analyze weather predictions on websites in detail using data mining techniques. Hypothesis, input, output, process, users, and features. This chapter emphasizes the key features that distinguish our novel approach from the existing approaches for weather predictions and analysis in various scenarios.

4.2 Hypothesis

Prediction or exploring issues attached to weather patterns can be done using data mining techniques. Predictive data mining can be used to predict the deciding factors which involve in different weather content issues. Descriptive data mining can be used to explore the current situation demonstrated in climate.

4.3 Input

As the initial input for this process, data obtained from websites which contain weather related details with deciding factors of weather forecasting is used.

4.4 Output

The output is obtained for this process as different data patterns related to weather prediction can be revealed according to the necessary factors identified. Prediction will be given as output attaches with predictive tasks. Summarization will be given according to the objective with descriptive tasks.

4.5 Process

In this process of weather analyzing with data mining techniques all standard steps in the KDD process which contains data selection to evaluation are carried out. Through-out the process, the data set is cleaned, formatted and prepared for mining and interpretation.

4.5.1 Data Selection

The weather analyzing can be made by gathering data about the current state of the atmosphere and using considerate of atmospheric processes to predict how the atmosphere will change. Before any forecast can be made, first it is mandatory to understand what the current weather conditions are and what is producing them. This is done by examining a large quantity of observation data.

The first set of data extracted from the weather content websites using python web scraping using BeautifulSoup. Following steps followed to extract actual data from websites – screen- shots and codes attached with Appendix A.

- Download the web page containing the weather details
- Create a BeautifulSoup class to parse the page
- Find the relevant style of HTML and assign to the index
- Inside the index, find each individual weather item.
- Extract and print the first weather item.
- Combine the data into Pandas DataFrame and analyze it

Then another set of data generated to add more instance for our dataset. Advanced data generator 3 for MySQL was used to generate the dataset. The process is shown in Appendix B. At last both data converted to CSV format in order to use and make analysis inside the Weka tool for preprocessing process.

4.5.2 Data Preprocessing

Our weather dataset contains 6588 instances and 20 attributes, in order to predict a pattern to find out the next day is a rainy or light shower or no rainy day. Noise inside the dataset should be removed using various preprocessing techniques. Always noise reduces the quality of the data. Filters should be applied to the dataset in a proper manner to get effect quality. Therefore, our dataset preprocessed before further analysis. Procedures are given in the implementation chapter.

4.5.3 Data Mining

This is the essential part of our research which used intelligent methods in order to determine exact data patterns. Not only that but also, discovering interesting knowledge associations, information gain, changes, anomalies and significant structures from our weather-related dataset. From association rules, the relationship between the selected attributes to the deciding factors can be determined.

In the classification part, different types of classification methods applied to find which outfit well for our model. Cluster analysis also was done with various techniques to visualize and get output with accuracy, true-false rate. Prediction and pattern analysis shows similar patterns which help to identify grouped attributes. According to the task we tried to achieve through our objectives.

4.5.5 Evaluation/ Interpretation

Evaluation of the research presented through graphs and time period which are the main two factors of analyzing. According to the goal evaluation should be precisely categorized to maximize efficiency. Many visualization packages and knowledge flow interfaces are available, including trees and distribute networks.

4.6 Features

The solution proposed by this research can be used to analyze a large amount of weather data which are in different forms in each source. By using predictive mining task, our solution allows us to make predictions for future instances. Moreover, descriptive tasks allow describing pattern by the dataset. So, our solution provides the particular output pattern according to the model we have taken.

4.8 Summary

This chapter presented an overview of our approach to analyze weather data to identify the rainy day. In this scenario, it is highlighted out our approach offers an efficient and precise solution for weather analysis using data mining techniques. The next chapter provides the design of our approach presented here.

CHAPTER 05

Implementation of the solution

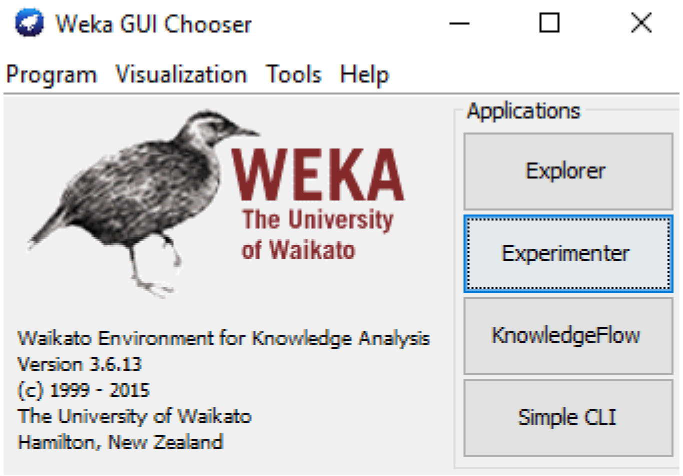
5.1 Introduction

This chapter provides implementation details of each and every process we have done. Moreover, this presents software and algorithms used in each process with sample outputs.

5.2 Weka

Weka tool is a collection of machine learning algorithms written in Java and enriched at the University of Waikato in New Zealand. It is free software licensed under the GNU General Public License. When considering a particular dataset, algorithms can be applied directly to the dataset or can be called through the own Java code. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, as well as graphical user interfaces for easy access to all functions. We have selected Weka since it includes many tools for preprocessing, classification, clustering, association ruled, regression and visualization as options.

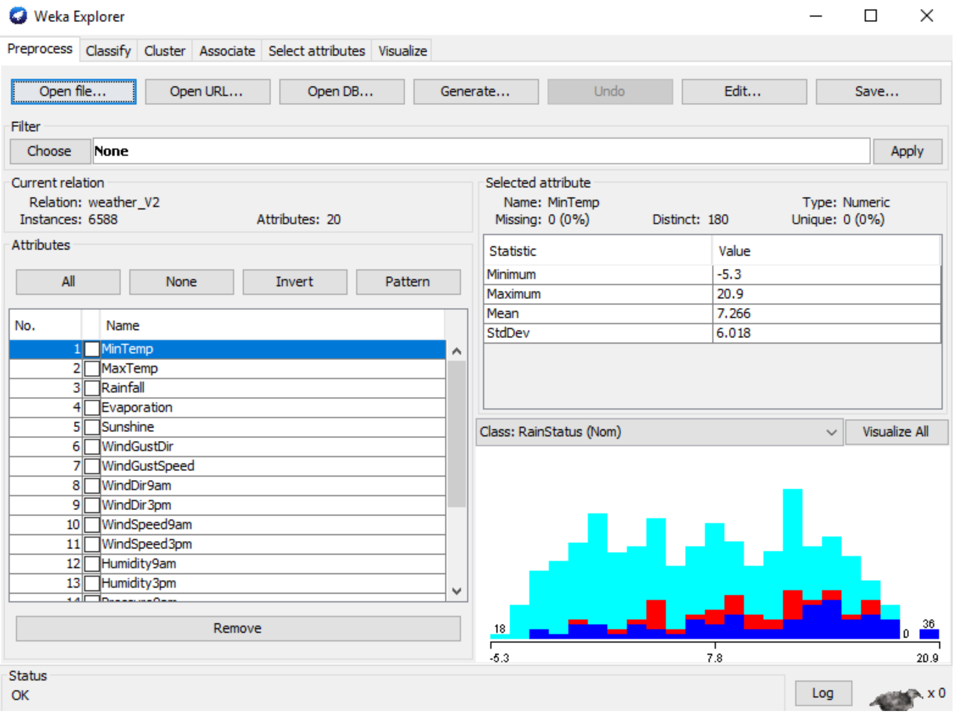
The 'weka.filters' package is attached with sub-classes that transform our dataset by ignoring or adding attributes, resampling the dataset, and removing samples. This package offers important support for data preprocessing, which is a necessary step in each machine learning research.



Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

Figure 5.1 Weka GUI Chooser

Figure 5.1 shows the Weka GUI Chooser window where all applications need to be chosen by the user and Figure 5.2 shows the Weka Explorer window where all the above-mentioned process actions reside.



Filter: None

Current relation

Relation: weather_V2
Instances: 6588
Attributes: 20

Selected attribute

Name: MinTemp
Missing: 0 (0%)
Distinct: 180
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	-5.3
Maximum	20.9
Mean	7.266
StdDev	6.018

Attributes

No.	Name
1	□ MinTemp
2	□ MaxTemp
3	□ Rainfall
4	□ Evaporation
5	□ Sunshine
6	□ WindGustDir
7	□ WindGustSpeed
8	□ WindDir9am
9	□ WindDir3pm
10	□ WindSpeed9am
11	□ WindSpeed3pm
12	□ Humidity9am
13	□ Humidity3pm

Class: RainStatus (Nom)

Status
OK

Figure 5.2 Weka Explorer

5.3 Data Collection and Preprocessing

The dataset was exported into CSV format in order to use inside the Weka tool for preprocessing process. The Weka suite contains a set of visualization tools and algorithms for data analysis and predictive modeling, as well as graphical user interfaces to facilitate access to this feature.

Since, Weka is a platform-independent Visualization of all attributes in the database can be shown after loading the dataset. Appendix A contains all the visualization of all attributes in Weka. Preprocessing is one the important process in the research which reduces the inaccuracy of the dataset. When the particular dataset is an effective one, then only we can get the best results from classification.

5.3.1 Replace Missing Values

There were some tuples that have no recorded values for several attributes; then the missing values can be filled in for the attribute by different methods. Some of the null values were replaced with the maximum number of attribute option. The process was done with the use of filters.

5.3.2 Data Transformation

String values were changed to nominal featured attribute by applying the ‘StingToBinary’ filter to the dataset. By maintaining all values as nominal make the dataset more effective one according to the analysis.

5.3.3 Data Normalization

Normalization is one of the scaling techniques in the preprocessing stage of any problem statement. Where, we can find a new range from an existing range. This can be very useful for prediction or forecasting. So, the standardization technique is necessary to bring.

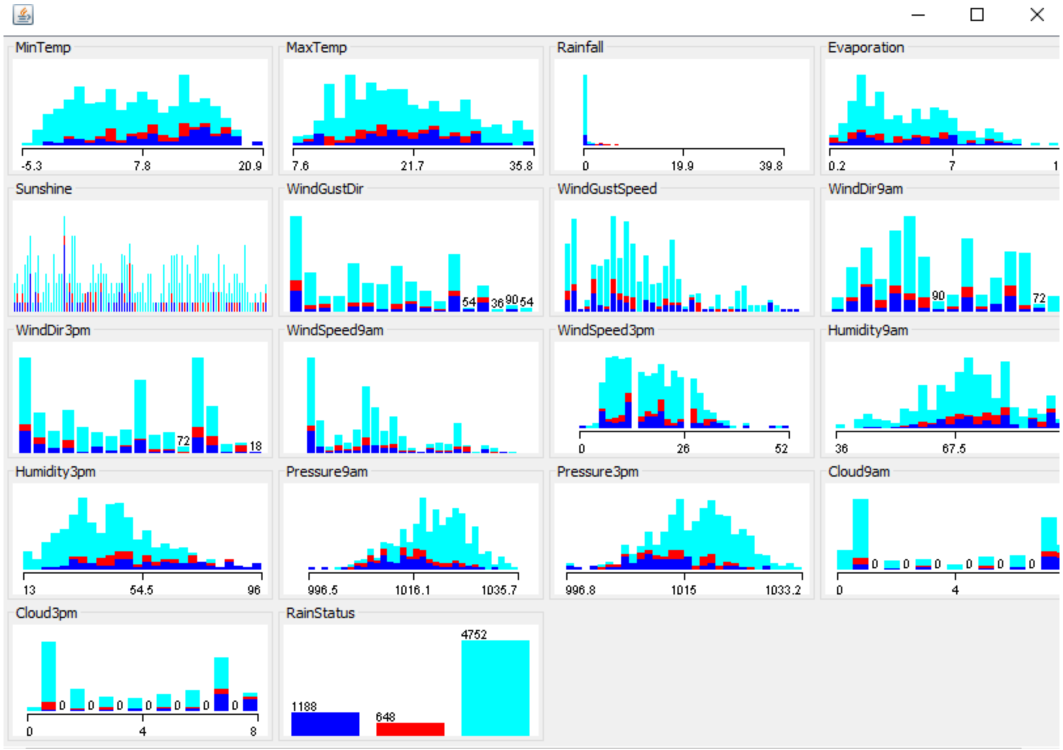


Figure 5.3 Visualization of attributes before preprocessing

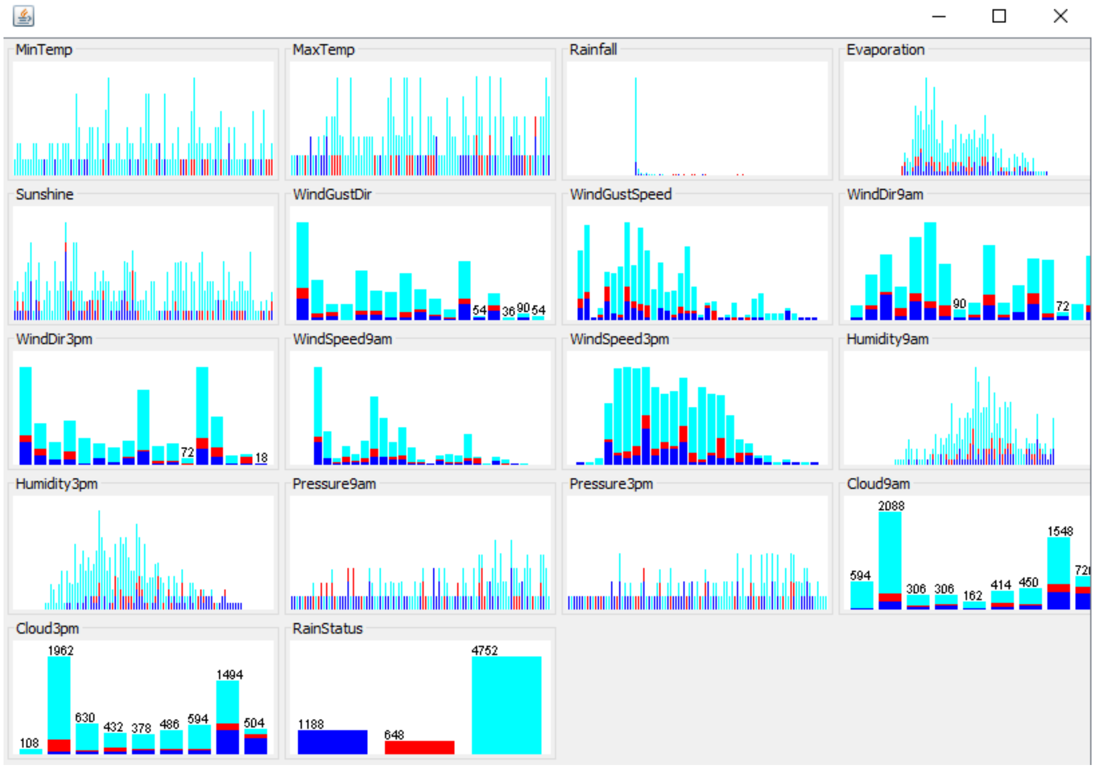


Figure 5.4 Visualization of attributes after preprocessing

Figure 5.3 and Figure 5.4 show the data distribution difference between the visualization of our selected attributes before and after the preprocessing respectively.

5.4 Weather Prediction Analysis

As the very first step, data exploration was done, where those 6588 instances dataset cleaned and transformed through several preprocessing techniques mentioned above. Significant factors and its dependent data were determined through the visualization patterns. Classification, Clustering, Association Rules, Decision Trees, Nearest Neighbour methods are some of the most important data mining techniques.

5.4.1 Using suitable classifier for classification

From the literature review we came up to the conclusion that we can’t define a particular classification algorithm is the best one to apply. The efficiency of the classification purely depends on the selected dataset. Through many types of research, it was proven that Naive Bayesian classifier frequently used because of its simplicity which performs fantastically for very large datasets. Since our dataset is also considerably large, we went through the classification model of Naïve Bayesian. J48 also shown the more accurate result, but we can’t able to visualize the tree- since we dealt with 19 attributes. There are three classes, ‘Yes’: there will be rain, ’LightRainShower’, ‘NoRain’.

5.4.2 Applying Appropriate Clustering Method

Clustering in a sense grouping such objects in particular or similar group related to each other. Clustering methods can be seen as divided into two types: hierarchical methods and partitioning methods. Partitioning methods relocate instances by moving them one cluster to another, starting from an initial partitioning. Those methods generally require that the count of clusters should be pre-set by the user. For this research finally, we have compared two algorithms belonged to portioning methods which are more relevant and popular for this data set.

5.4.2.1 Clustering using Simple KMean Algorithm

In KMean clustering, k number of centers, one for each cluster. Each center should be as far as possible from each. Then each point belonging to a given dataset and associated it to the nearest center. Initial step completed when there are no pending points. After that, k new centroids are calculated again as center taken from the previous step. Now there are k new centroids, it has to be related between same dataset values and the nearest new center. Iteratively same procedures are carried out. From the final result of these iteration, considerable change in k centers’ location are observed and it will be continued until no more changes are done.

When measuring cluster validity, several numerical measures are applied to keep track on various aspects of cluster validity. As an internal index, Sum of Squared Error (SSE) is displayed there. SSE is used to measure the effectiveness of a clustering structure without respect to external information. We choose three clusters where SSE is pretty high but by considering the sequence order of time taken and small changes in SSE value.

We have identified three clusters, the first cluster contain ‘yes’ in a sense heavy rain is expected for the next day. Expecting ‘light rain shower’ belong to the second cluster and ‘No rain’ is considered to be the third cluster. With the visualization of the cluster, we can understand cluster assignments.

Cluster Type	**No. of Instances**	**Percentage**
Cluster 0 : “No Rain”	3042	46%
Cluster 1: “Light Rain Shower”	1854	28%
Cluster 2: “Yes”	1692	26%

Table 5.1. Cluster instances in K-means clustering

5.4.2.2 Clustering using Simple Expectation-Maximization Algorithm

Expectation-Maximization (EM) algorithm is an iterative method for discovering maximum likelihood. According to the algorithm library, Gaussian distributions are usually modeled which are initialized randomly and whose parameters are iteratively optimized to fit better to the dataset.

From the EM algorithm, we have identified three clusters. Through the likelihood values, there are significant values changes are not seen after cluster three. The first cluster contains ‘yes’ in a sense heavy rain is expected for the next day. Expecting ‘light rain shower’ belong to the second cluster and ‘No rain’ is considered to be the third cluster.

Cluster Type	**No. of Instances**	**Percentage**
Cluster 0 : “No Rain”	2790	42%
Cluster 1: “Light Rain Shower”	1602	24%
Cluster 2: “Yes”	2196	33%

Table 5.2. Cluster instances in EM Clustering

5.4.4.2 Clustering using Density-Based Algorithm

The major importance of the density-based clustering algorithm is discovering nonlinear pattern structure based on the density. Identifying the clusters and distribution patterns are that main goal of this method. This method can be used for finding clusters of arbitrary shape which are not necessarily convex. From the Density Based algorithm, we have identified three clusters. Through the likelihood values, there are significant values changes are not seen after cluster three. The first cluster contains ‘yes’ in a sense heavy rain is expected for the next day. Expecting ‘light rain shower’ belong to the second cluster and ‘No rain’ is considered to be the third cluster.

Cluster Type	**No. of Instances**	**Percentage**
Cluster 0 : “No Rain”	1980	30%
Cluster 1: “Light Rain Shower”	2358	36%
Cluster 2: “Yes”	2250	34%

Table 5.3. Cluster instances in density-based clustering

5.5 Summary

This chapter presents the full path in implementing data models from the taken dataset. Furthermore, this chapter gives a detail description about the WEKA tool to build the data model and attribute selection. Next chapter provides the evaluation criteria.

CHAPTER 06

Evaluation

6.1 Introduction

This chapter focuses on how testing strategies carried out according to the objective in terms of evaluation measurements for the selected data mining technique such as accuracy percentage, TP & FP Rate, Precision, F-Measure, ROC area, SSE, and loglikelihood for classification and clustering.

6.2 Evaluation for classification

One of the important outputs from the classifier is a confusion matrix (Shown in Appendix D). By going through the confusion matrix, we can find a number of evaluation factors such as precision accuracy, and recall to evaluate data mining classifiers. These measurements and their definition are given in the following table 6.1.

Measure	Meaning	Relevant Formula
Precision	Percentage of positive predictions which are correct	TP/ (TP+FP)
Recall	Percentage of positive labeled instances that are predicted as positive	TP/ (TP+FN)
Accuracy	Percentage of predictions which are correct	(TP+TN)/ (TP+TN+FP+FN)

Table 6.1. Evaluation measurements for classifiers

Classifiers provide few sets of measures named, TP rate, FP rate, ROC area, and F-measure. TP rate is equal to sensitivity, while FP rate equal to one minus specificity. F measure calculated by precision and recall. The overall ability of the test to distinguish between usefulness and uselessness can be quantified by the ROC curve area.

A truly useless test has an area of 0.5. A perfect test has an area of 1.00. Usually better models are having higher TP rate, lower FP rate and ROC space close to 1.00. comparison of the confusion matrixes and weighted averages in the classification model used for a rainy day: "Yes" are given in the following Table 6.2.

Technique	NavieBayes	J48
TP Rate	0.723	0.987
FP Rate	0.034	0
Precision	0.823	1
F-Measure	0.777	0.994
ROC Area	0.956	0.999
Recall	0.723	0.987

Table 6.2. comparison of different classification methods for rainy day

Technique	NavieBayes	J48
TP Rate	0.935	1
FP Rate	0.021	0
Precision	0.832	1
F-Measure	0.881	1
ROC Area	0.994	0.999
Recall	0.935	1

Table 6.3. comparison of different classification methods for light rain shower day

Technique	NavieBayes	J48
TP Rate	0.956	1
FP Rate	0.15	0.008
Precision	0.943	0.997
F-Measure	0.949	0.998
ROC Area	0.972	0.999
Recall	0.956	1

Table 6.4. comparison of different classification methods for no rainy day

Table 6.3 shows the comparison of two classifiers for a light rain shower day. Here we can able to see that, J48 shows the TP rate is almost 1, where the model classifies almost more effective. But the drawback here is we can't be able to visualize the tree as it is a big data set with more branches and associations. Table 6.4 describes the comparison of different classification methods for no rainy day. Here also, the J48 FP rate shows 0.008, which means a very less false rate leads to an effective classification method. But comparatively Naïve Bayes classification also high in TP rate.

Knowledge flow interface provides the data flow in Weka. In order to show the processing and analyzing data knowledge flow helps a lot. There are a number of evaluations can be done through this component. Training Set Maker, Test Set Maker, Class Assigner, Class Value Picker, Train Test Split Maker, Classifier Performance Evaluator are most significant evaluation criteria in Knowledge flow. Figure 6.1 shows the knowledge flow carried out for the classification.

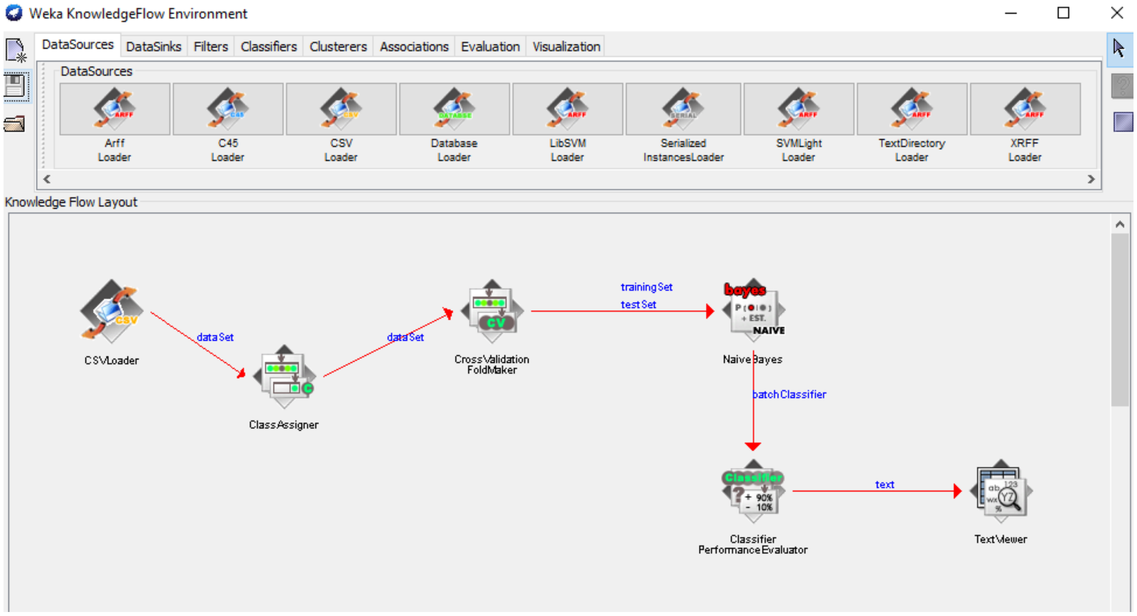


Figure 6.1. knowledge flow for the Naïve Bayes classification

6.3 Evaluation for Clustering

In order to evaluate the model, SSE values and time to build the model, are considered in the clustering. To evaluate the accuracy of the data model, datasets deployed in Weka tool and clustering algorithms are applied to the dataset with classes to cluster evaluation option. Table 6.5 shows the within cluster SSE values for a different number of clusters used in KMean algorithm with Euclidean Distance function. We have selected the seed value as -2 which made comparatively low SSE values among others.

No. of clusters	SSE
2	1958.2636
3	1311.10083
4	1269.08451
5	1167.59455
6	1092.65681
7	1021.33567
8	953.76428
9	923.93444

Table 6.5. Variation of SSE within clusters for different clusters

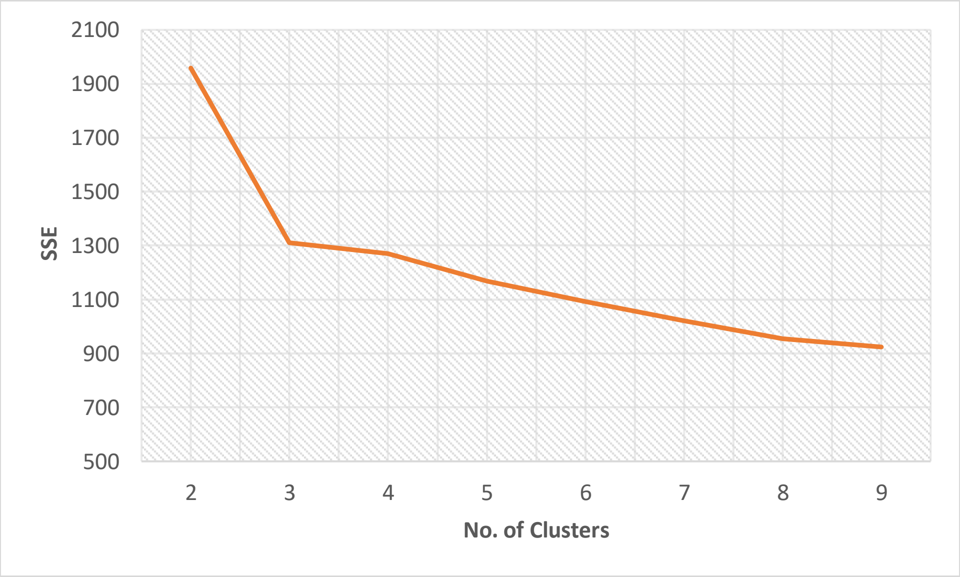


Figure 6.2. Graph of SSE within clusters versus different clusters

Intended for the evaluation of EM clustering, a number of clusters analyzed to find loglikelihood value effective time period. If loglikelihood of sample is greater under one model that other, we tend to infer that the former model is more likely than later. Table 6.6 shows the variations of loglikelihood with a number of clusters. And Figure 6.2 shows graph flow among these two factors.

No. of clusters	Log-likelihood
2	-55.906
3	-51.49529
4	-52.49095
5	-50.38657
6	-51.61624
7	-51.21999
8	-48.56916
9	-48.86482

Table 6.6. Variation of log-likelihood for different clusters

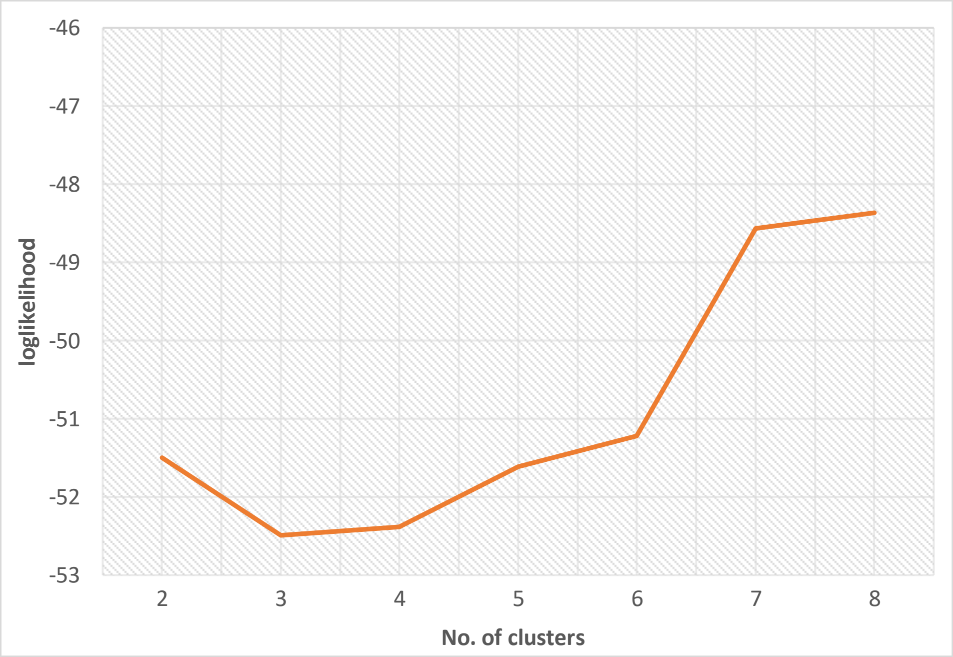


Figure 6.2. Graph of log-likelihood versus different clusters

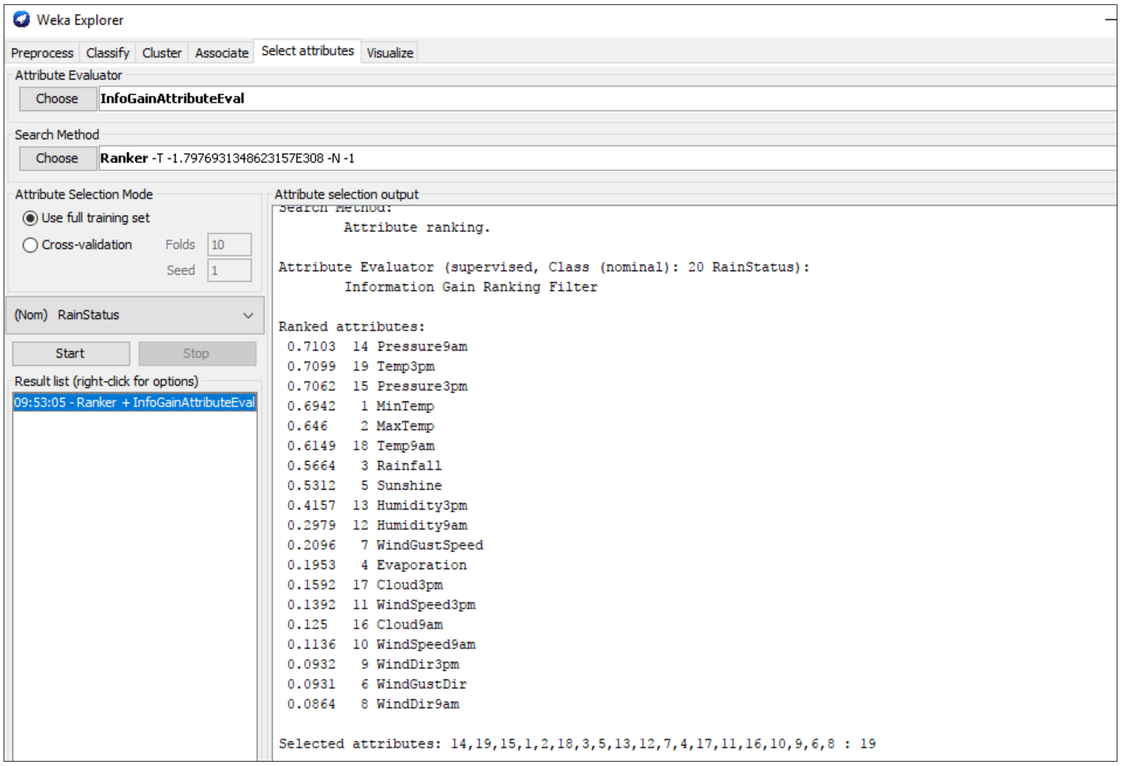
Table 6.7 indicates the time taken for K-Means, EM, Density-Based algorithms. From this analysis it is obvious that K-Means algorithm takes minimum time to make clusters in comparison with other clustering algorithms. Hence, time is more effective for K-means clustering algorithm.

Clustering Algorithm	**Time (sec.)**
K-means	0.19
EM	4.21
Density-Based	0.29

Table 6.7. Time taken for clustering by different clustering algorithms

6.3 Information gain from the dataset

Most necessary part of attribute evaluator in information gain. Through this evaluation how much attribute information gives about the class can be evaluated. Perfectly partition always produces a higher rate of information. Information gain for our dataset is shown in Figure 6.3.



Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) RainStatus

Result list (right-click for options)
09:53:05 - Ranker + InfoGainAttributeEval

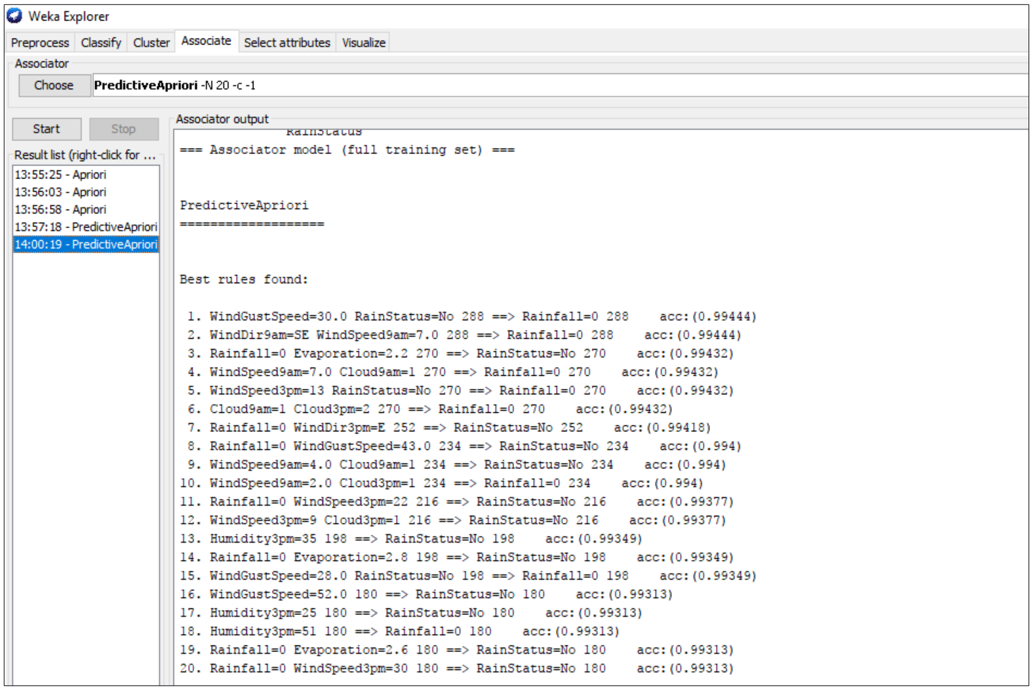
Attribute selection output

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 20 RainStatus):
        Information Gain Ranking Filter

Ranked attributes:
 0.7103  14 Pressure9am
 0.7099  19 Temp3pm
 0.7062  15 Pressure3pm
 0.6942   1 MinTemp
 0.646    2 MaxTemp
 0.6149  18 Temp9am
 0.5664   3 Rainfall
 0.5312   5 Sunshine
 0.4157  13 Humidity3pm
 0.2979  12 Humidity9am
 0.2096   7 WindGustSpeed
 0.1953   4 Evaporation
 0.1592  17 Cloud3pm
 0.1392  11 WindSpeed3pm
 0.125   16 Cloud9am
 0.1136  10 WindSpeed9am
 0.0932   9 WindDir3pm
 0.0931   6 WindGustDir
 0.0864   8 WindDir9am

Selected attributes: 14,19,15,1,2,18,3,5,13,12,7,4,17,11,16,10,9,6,8 : 19
```

Figure 6.3. Information Gain Evaluation

6.4 Generating Association Rules

Association rules look for the whole sets of items that have support larger than the minimum support, and then use huge sets of items to generate the desired rules that have confidence greater than minimum confidence. To find association rules we applied predictive apriori algorithm. Because, comparison of apriori & predictive apriori associators study gives that, predictive apriori is ended with higher accuracy[14]. Figure 6.4 shows 20 associate rules generated from our dataset.



PredictiveApriori -N 20 -c -1

13:55:25 - Apriori
13:56:03 - Apriori
13:56:58 - Apriori
13:57:18 - PredictiveApriori
14:00:19 - PredictiveApriori

```
              RainStatus
=== Associator model (full training set) ===


PredictiveApriori
===================


Best rules found:

 1. WindGustSpeed=30.0 RainStatus=No 288 ==> Rainfall=0 288    acc:(0.99444)
 2. WindDir9am=SE WindSpeed9am=7.0 288 ==> Rainfall=0 288    acc:(0.99444)
 3. Rainfall=0 Evaporation=2.2 270 ==> RainStatus=No 270    acc:(0.99432)
 4. WindSpeed9am=7.0 Cloud9am=1 270 ==> Rainfall=0 270    acc:(0.99432)
 5. WindSpeed3pm=13 RainStatus=No 270 ==> Rainfall=0 270    acc:(0.99432)
 6. Cloud9am=1 Cloud3pm=2 270 ==> Rainfall=0 270    acc:(0.99432)
 7. Rainfall=0 WindDir3pm=E 252 ==> RainStatus=No 252    acc:(0.99418)
 8. Rainfall=0 WindGustSpeed=43.0 234 ==> RainStatus=No 234    acc:(0.994)
 9. WindSpeed9am=4.0 Cloud9am=1 234 ==> RainStatus=No 234    acc:(0.994)
10. WindSpeed9am=2.0 Cloud3pm=1 234 ==> Rainfall=0 234    acc:(0.994)
11. Rainfall=0 WindSpeed3pm=22 216 ==> RainStatus=No 216    acc:(0.99377)
12. WindSpeed3pm=9 Cloud3pm=1 216 ==> RainStatus=No 216    acc:(0.99377)
13. Humidity3pm=35 198 ==> RainStatus=No 198    acc:(0.99349)
14. Rainfall=0 Evaporation=2.8 198 ==> RainStatus=No 198    acc:(0.99349)
15. WindGustSpeed=28.0 RainStatus=No 198 ==> Rainfall=0 198    acc:(0.99349)
16. WindGustSpeed=52.0 180 ==> RainStatus=No 180    acc:(0.99313)
17. Humidity3pm=25 180 ==> RainStatus=No 180    acc:(0.99313)
18. Humidity3pm=51 180 ==> Rainfall=0 180    acc:(0.99313)
19. Rainfall=0 Evaporation=2.6 180 ==> RainStatus=No 180    acc:(0.99313)
20. Rainfall=0 WindSpeed3pm=30 180 ==> RainStatus=No 180    acc:(0.99313)
```

Figure 6.4 Rules generated from Predictive Apriori algorithm

6.5 Summary

This chapter concludes with the evaluation results to evaluate the data model. The last chapter will summarize the overall research and highlights the significant findings of the research.

CHAPTER 07

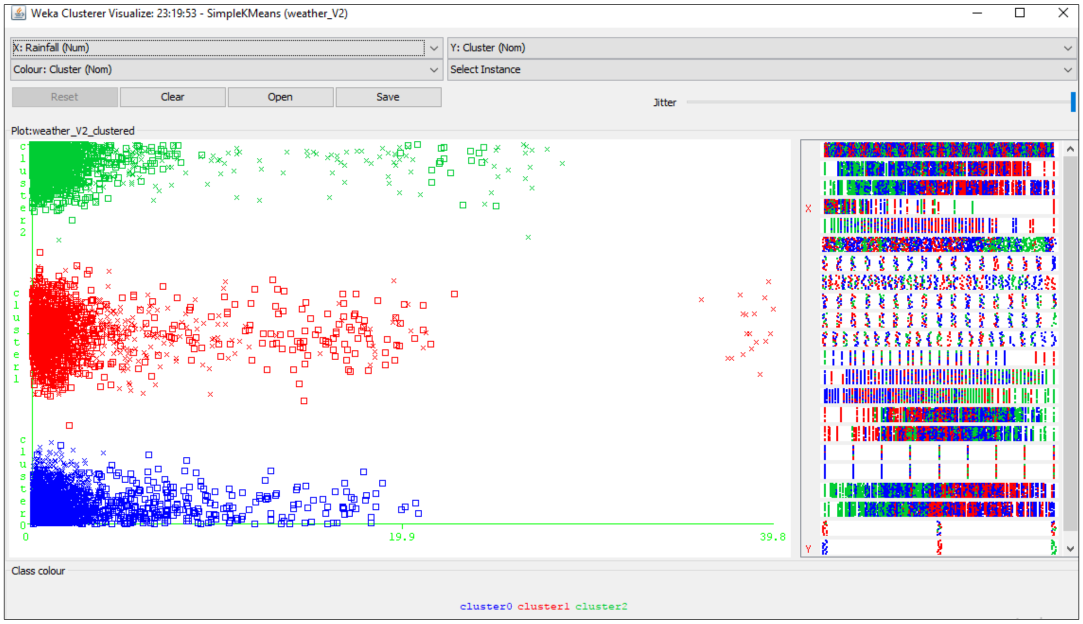
Conclusion and Further Work

7.1 Introduction

This chapter presents an overview of our research and how we provide the solution for analyzing weather prediction data which belong to Big data kind of category. And also, this chapter focuses on limitations and further work of this research.

7.2 Overview of the research

From the analysis of weather predictions, we found the major deciding factors of the rain prediction. Since there are many factors need to be considered for analyzing, according to the dataset extracted and generated we ignored some attributes to reduce the sum of squared errors. By ignoring several attributes incorrectly clustered attributes' count reduced tremendously. Around the overall world climate may differ with its geographical factors, so common deciding factors are considered and clustered. One sample attribute visualization with cluster variation is given in Figure 7.1. All Other attributes' visualization shown in Appendix E.


Weka Clusterer Visualize: 23:19:53 - SimpleKMeans (weather_V2)

X: Rainfall (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered

Figure 7.1 Cluster visualization with Rainfall factor

7.3 Major findings in our research

Major deciding factors of the weather prediction were listed down by ignoring attributes sequentially compared with SSE values (Shown in Appendix E). Figure 7.2 shows the reduction of SSE values with the selected attributes.

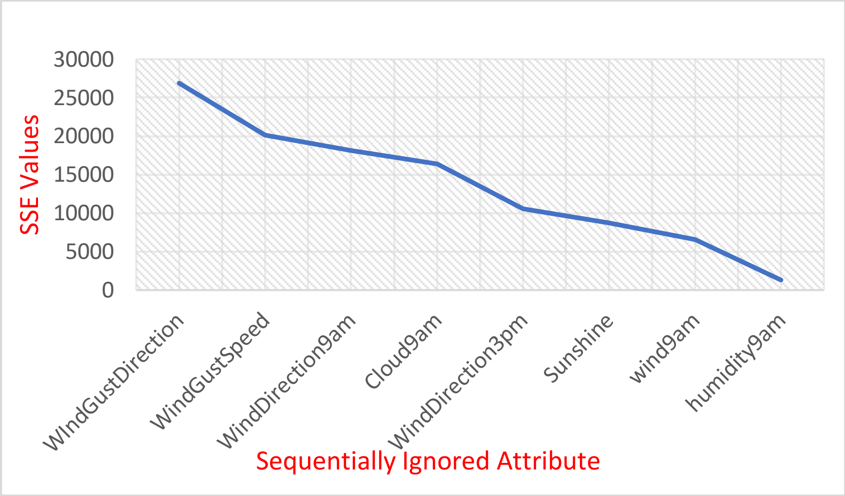


Figure 7.2 Graph of SSE values Vs Sequentially ignored attributes

From the analysis of weather predictions and ignorance of attributes, we found the major deciding factors of the rain prediction. Since there are many factors need to be considered for analyzing, according to the dataset extracted and generated we ignored some attributes to reduce the sum of squared errors. Our analysis reveals the significance of temperature and moisture among the influences such as Rain Fall, Evaporation, Pressure, Cloud status, and late-night Humidity. And also, Wind Gust Direction, Wind Direction, Minimum Temperature and Wind Gust Speed are low dependent factors in the rain prediction for the next day. These important patterns recognized from our research that can be used to model a climate pattern in a particular area.

7.4 Limitations

The overall least SSE error we got is considerably little high, it means incorrectly clustered rate is high. If it can be reduced the value of SSE, then the output will be more accurate to visualize the pattern. Accuracy of the tool highly depends on the dataset.

7.5 Further Developments

Integrated classification and clustering are expected to as a future work. That will be more precise to provide accurate predictive patterns. By using integrated algorithms, it will be a generic one for whole datasets, which helps to describe the efficiency of particular cluster algorithm in general.

7.6 Summary

This chapter concluded about the analysis, summary of our major findings, limitations and extended future work.

CHAPTER 08

REFERENCES

[1] B. Rajdeepa and Dr. P. Sumathi, "*An Analysis of Web Mining and its types besides Comparison of Link Mining Algorithms in addition to its specifications*," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3, issue 1, Jan. 2014.

[2] Kamlesh Patidar, Preetesh Purohit and Kapil Sharma, "*Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library*," Medicaps Institute of Technology & Management IJCST, Vol. 2, Issue 1, March 2011.

[3] M. Baglioni; U. Ferrara; A. Romei; S. Ruggieri; F. Turini, "*Preprocessing and Mining Web Log Data for Web Personalization*," LNCS, vol. 2829, pp. 237–249. Springer, Heidelberg (2003).

[4] Michael Azmy, "Web Content Mining Research: A Survey", DRAFT Version 1, - Nov. 2005.

[5] Menahem Friedman, Mark Last, Yaniv Makover and Abraham Kandel, "*Anomaly Detection in web documents using Crisp and fuzzy-based cosine clustering methodology*," Information sciences 177, pp. 467-475 (2007).

[6] Qinbao Song and Martin Shepperd, "*Mining web browsing patterns for E-commerce,"*, Computers in Industry 57, pp. 622-630 (2006).

[7] Xiaofeng He, Hongyuan Zha, Chris H.Q. Ding and Horst D. Simon, "*Web Document Clustering Using Hyperlink Structures*," Computational Statistics & Data Analysis 41, pp. 19-45 (2002).

[8] M. Shamim Khan and Sebastian W. Khor, "*Web clustering Using a hybrid neural network*," Applied Soft Computing 4, pp.423-432 (2004).

[9] K.M. Hammouda and M.S. Kamel, "*Effective Pharse-Based Document Indexing for Web Document Clustering*," IEEE Trans.Knowledge and Data Eng., vol. 16, no. 10, pp. 1279-1296, Oct. 2004.

[10] O.Zamir and O.Etziono, "*Web Document Clustering: A feasibility Demonstration*," Proc. Third Int'l Conf. Research and Development in Information Retrieval (SIGIR),1998.

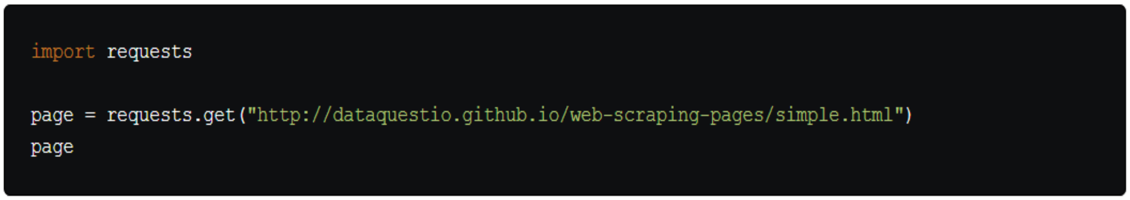
[11] S. K. Sahu and S. Srivastava, “*Review of Web Document Clustering Algorithms*,” Third Int’l Conf. IEEE Computing for Sustainable Global Development (INDIACom), 2016.

[12] Y. W. Dou, L. Lu, X. Liu and Daiping Zhang, “Meteorological Data Storage and Management System”, Computer Systems & Applications, vol. 20, no.7, (2011) July, pp. 116-12

[13] Kamber, M. and Pei, J. (2006). Data Mining:concepts and techniques. 2nd ed. heidellberg london: Morgan Kaufmann.

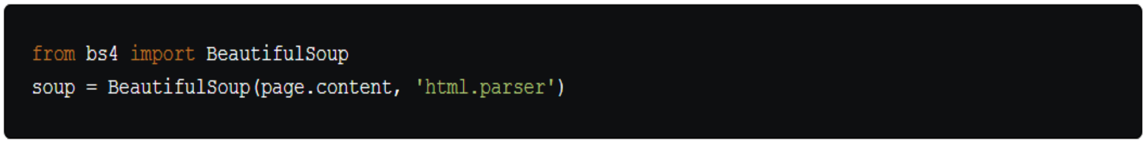
[14] Bharati, M. and Ramageri, A. (2013). Data Mining techniques and applications. Indian Journal of Computer Science and Engineering, 1(4), pp.301-305.

APPENDIX A

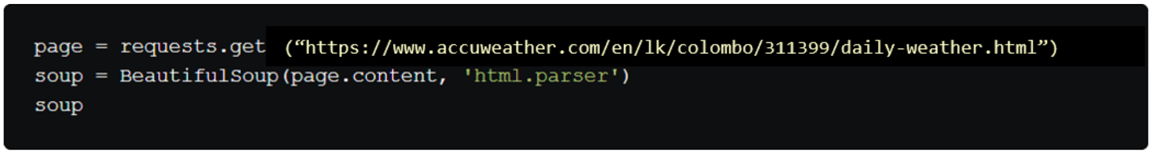


```python
import requests

page = requests.get("http://dataquestio.github.io/web-scraping-pages/simple.html")
page
```

```
<Response [200]>
```

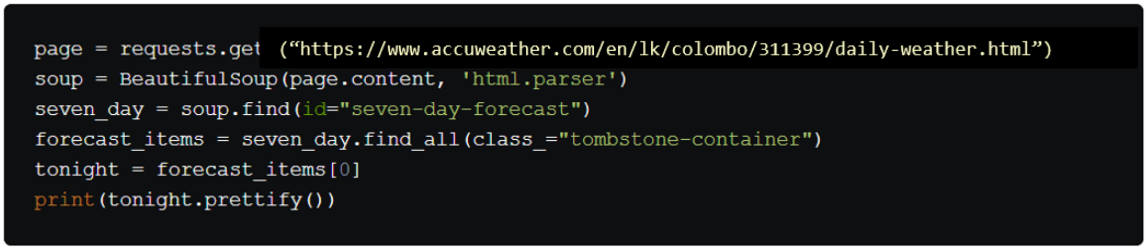


```python
from bs4 import BeautifulSoup
soup = BeautifulSoup(page.content, 'html.parser')
```

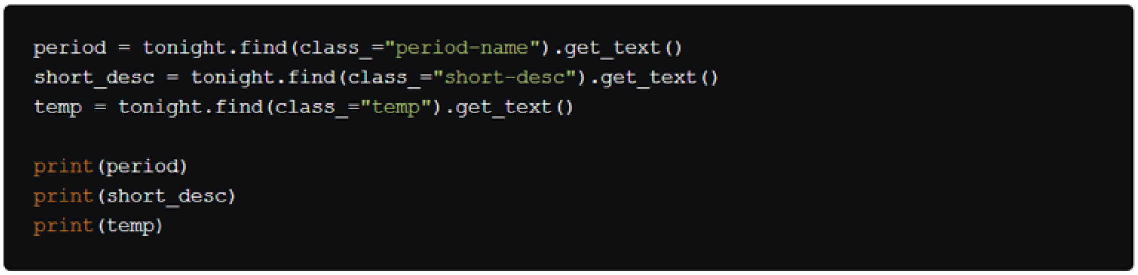


```python
page = requests.get (“https://www.accuweather.com/en/lk/colombo/311399/daily-weather.html”)
soup = BeautifulSoup(page.content, 'html.parser')
soup
```

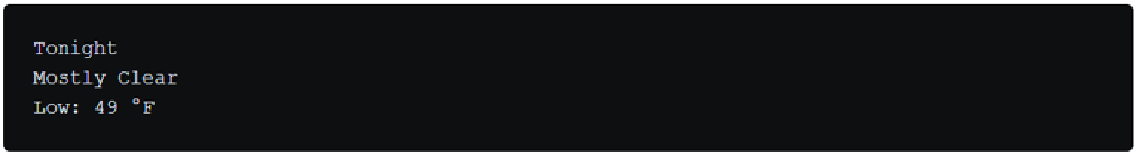
```python
soup.select("div p")
```



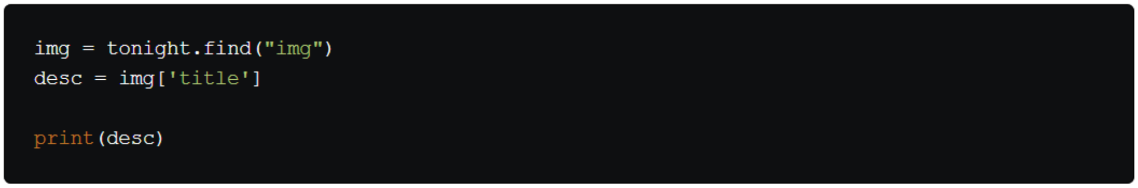
```python
page = requests.get (“https://www.accuweather.com/en/lk/colombo/311399/daily-weather.html”)
soup = BeautifulSoup(page.content, 'html.parser')
seven_day = soup.find(id="seven-day-forecast")
forecast_items = seven_day.find_all(class_="tombstone-container")
tonight = forecast_items[0]
print(tonight.prettify())
```



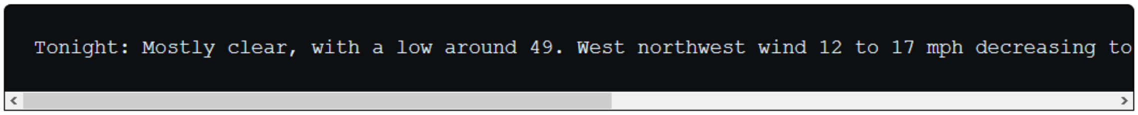
```
period = tonight.find(class_="period-name").get_text()
short_desc = tonight.find(class_="short-desc").get_text()
temp = tonight.find(class_="temp").get_text()

print(period)
print(short_desc)
print(temp)
```



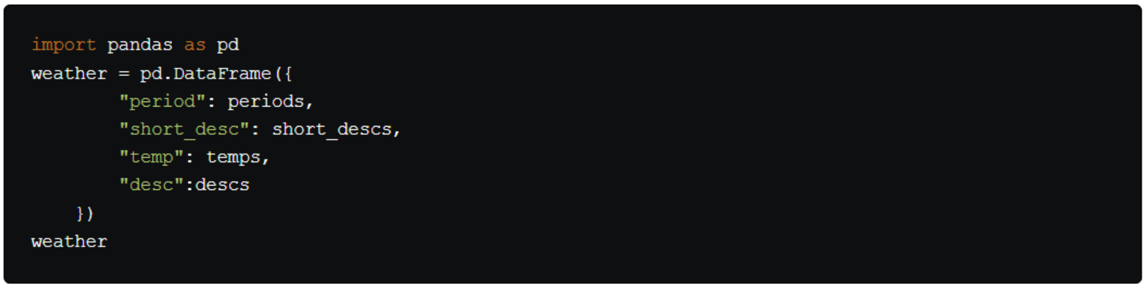
```
Tonight
Mostly Clear
Low: 49 °F
```



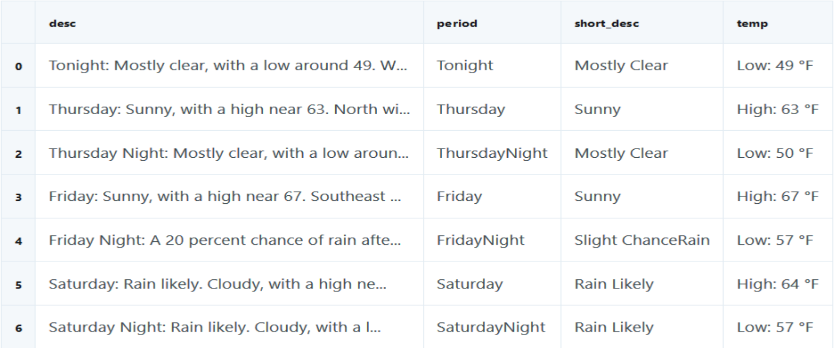
```
img = tonight.find("img")
desc = img['title']

print(desc)
```



```
Tonight: Mostly clear, with a low around 49. West northwest wind 12 to 17 mph decreasing to
```



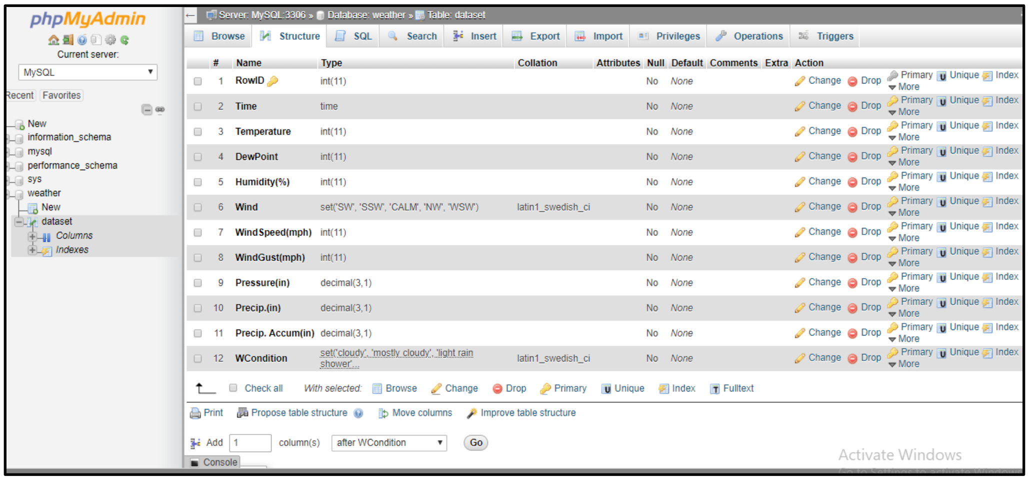
```
import pandas as pd
weather = pd.DataFrame({
        "period": periods,
        "short_desc": short_descs,
        "temp": temps,
        "desc":descs
    })
weather
```



	desc	period	short_desc	temp
0	Tonight: Mostly clear, with a low around 49. W...	Tonight	Mostly Clear	Low: 49 °F
1	Thursday: Sunny, with a high near 63. North wi...	Thursday	Sunny	High: 63 °F
2	Thursday Night: Mostly clear, with a low aroun...	ThursdayNight	Mostly Clear	Low: 50 °F
3	Friday: Sunny, with a high near 67. Southeast ...	Friday	Sunny	High: 67 °F
4	Friday Night: A 20 percent chance of rain afte...	FridayNight	Slight ChanceRain	Low: 57 °F
5	Saturday: Rain likely. Cloudy, with a high ne...	Saturday	Rain Likely	High: 64 °F
6	Saturday Night: Rain likely. Cloudy, with a l...	SaturdayNight	Rain Likely	Low: 57 °F

APPENDIX B

The sample attributes were taken to create the database in MySQL.



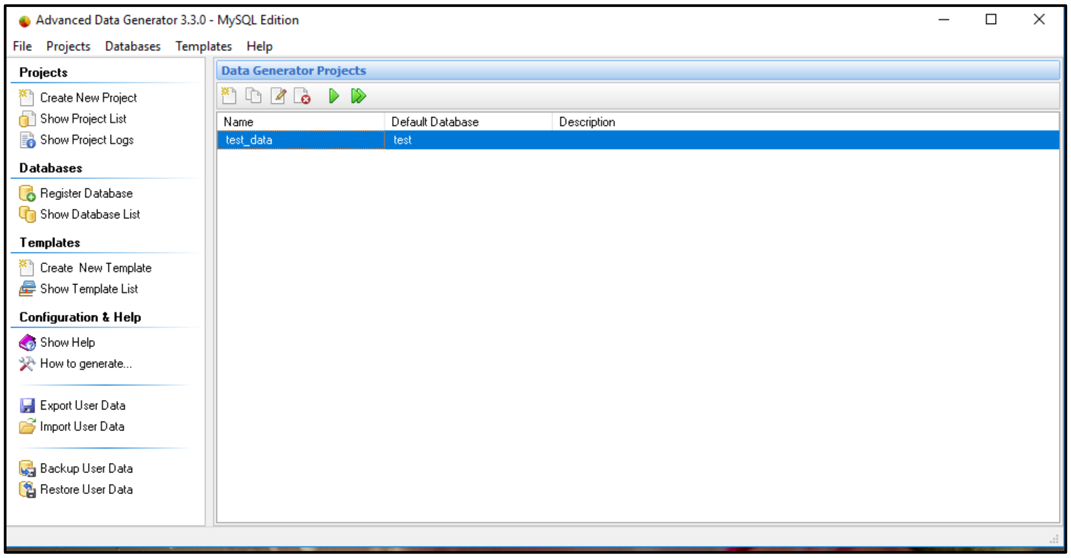
Server: MySQL:3306 » Database: weather » Table: dataset

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	RowID	int(11)			No	*None*			Change Drop Primary Unique Index More
2	Time	time			No	*None*			Change Drop Primary Unique Index More
3	Temperature	int(11)			No	*None*			Change Drop Primary Unique Index More
4	DewPoint	int(11)			No	*None*			Change Drop Primary Unique Index More
5	Humidity(%)	int(11)			No	*None*			Change Drop Primary Unique Index More
6	Wind	set('SW', 'SSW', 'CALM', 'NW', 'WSW')	latin1_swedish_ci		No	*None*			Change Drop Primary Unique Index More
7	WindSpeed(mph)	int(11)			No	*None*			Change Drop Primary Unique Index More
8	WindGust(mph)	int(11)			No	*None*			Change Drop Primary Unique Index More
9	Pressure(in)	decimal(3,1)			No	*None*			Change Drop Primary Unique Index More
10	Precip.(in)	decimal(3,1)			No	*None*			Change Drop Primary Unique Index More
11	Precip. Accum(in)	decimal(3,1)			No	*None*			Change Drop Primary Unique Index More
12	WCondition	set('cloudy', 'mostly cloudy', 'light rain shower'...	latin1_swedish_ci		No	*None*			Change Drop Primary Unique Index More

Add 1 column(s) after WCondition Go

Fig 5. MySQL database

Advanced data generator 3 for MySQL was used to create the sample dataset from the selected database.

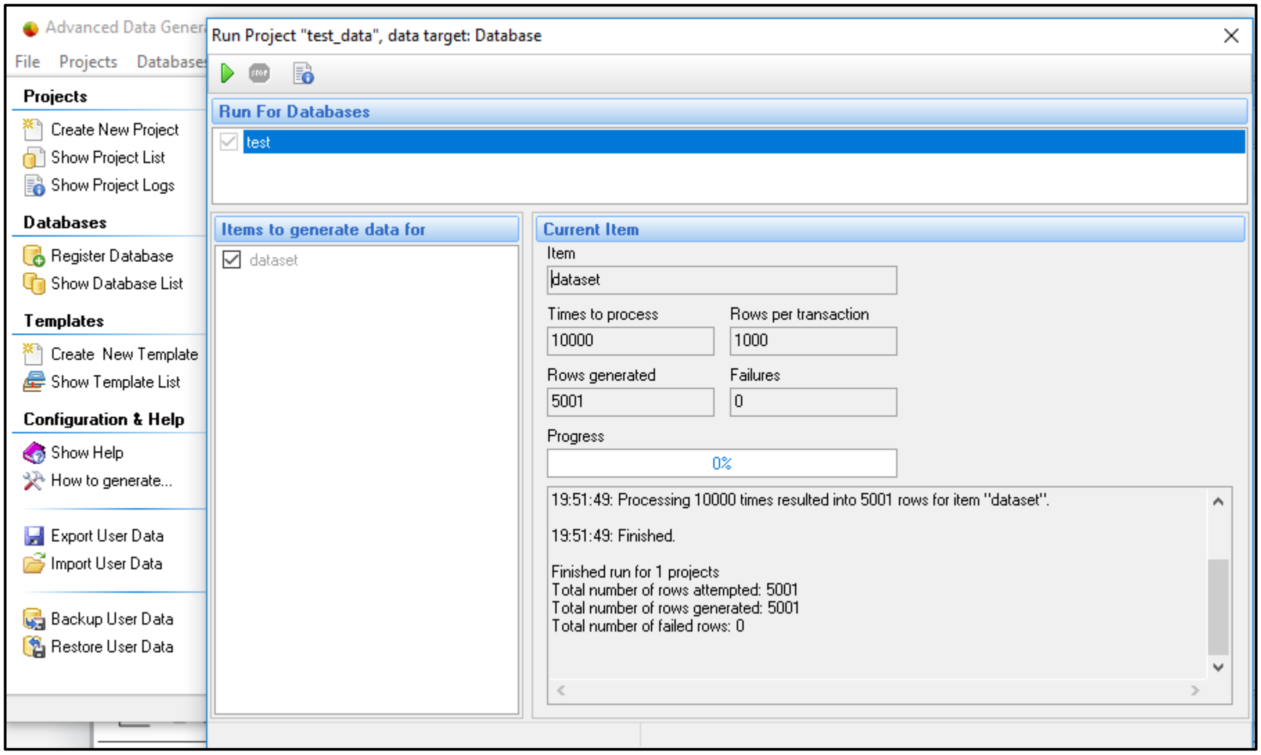


Advanced Data Generator 3.3.0 - MySQL Edition

Name	Default Database	Description
test_data	test	

Progress of sample data generation

Ten thousand sample rows were generated from the data generator tool.



Run Project "test_data", data target: Database

☑ test

☑ dataset

Item: dataset

Times to process: 10000

Rows per transaction: 1000

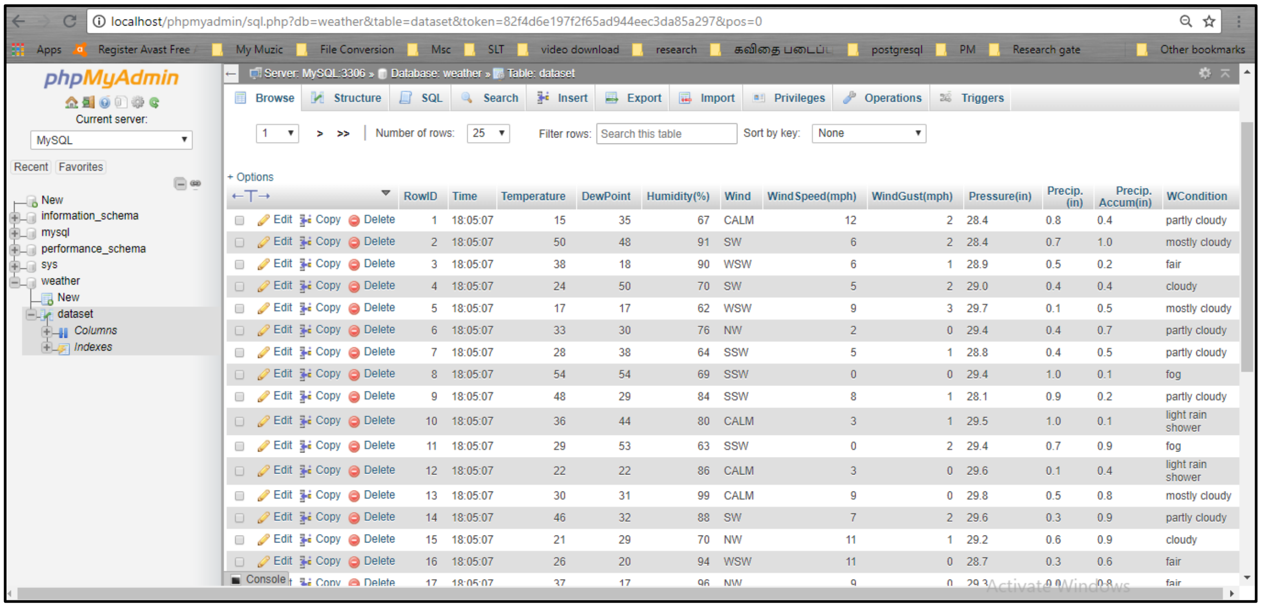
Rows generated: 5001

Failures: 0

Progress: 0%

```
19:51:49: Processing 10000 times resulted into 5001 rows for item "dataset".

19:51:49: Finished.

Finished run for 1 projects
Total number of rows attempted: 5001
Total number of rows generated: 5001
Total number of failed rows: 0
```

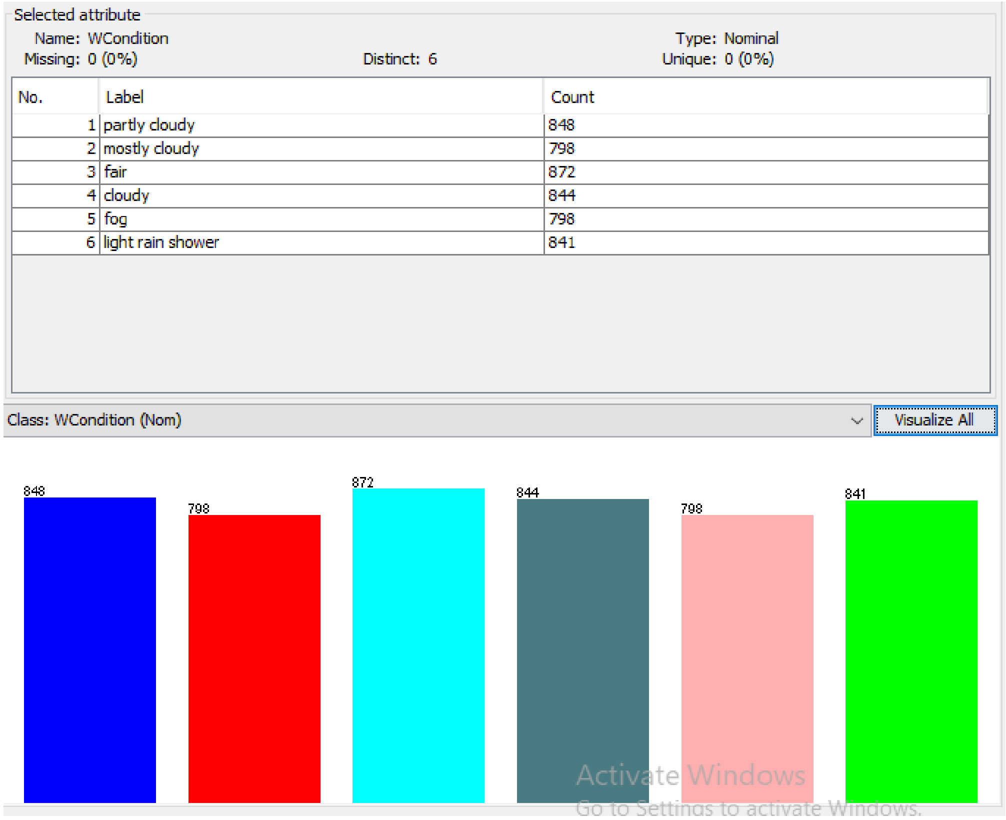
Completion of dataset



Number of rows: 25 · Filter rows: Search this table · Sort by key: None

RowID	Time	Temperature	DewPoint	Humidity(%)	Wind	WindSpeed(mph)	WindGust(mph)	Pressure(in)	Precip. (in)	Precip. Accum(in)	WCondition
1	18:05:07	15	35	67	CALM	12	2	28.4	0.8	0.4	partly cloudy
2	18:05:07	50	48	91	SW	6	2	28.4	0.7	1.0	mostly cloudy
3	18:05:07	38	18	90	WSW	6	1	28.9	0.5	0.2	fair
4	18:05:07	24	50	70	SW	5	2	29.0	0.4	0.4	cloudy
5	18:05:07	17	17	62	WSW	9	3	29.7	0.1	0.5	mostly cloudy
6	18:05:07	33	30	76	NW	2	0	29.4	0.4	0.7	partly cloudy
7	18:05:07	28	38	64	SSW	5	1	28.8	0.4	0.5	partly cloudy
8	18:05:07	54	54	69	SSW	0	0	29.4	1.0	0.1	fog
9	18:05:07	48	29	84	SSW	8	1	28.1	0.9	0.2	partly cloudy
10	18:05:07	36	44	80	CALM	3	1	29.5	1.0	0.1	light rain shower
11	18:05:07	29	53	63	SSW	0	2	29.4	0.7	0.9	fog
12	18:05:07	22	22	86	CALM	3	0	29.6	0.1	0.4	light rain shower
13	18:05:07	30	31	99	CALM	9	0	29.8	0.5	0.8	mostly cloudy
14	18:05:07	46	32	88	SW	7	2	29.6	0.3	0.9	partly cloudy
15	18:05:07	21	29	70	NW	11	1	29.2	0.6	0.9	cloudy
16	18:05:07	26	20	94	WSW	11	0	28.7	0.3	0.6	fair
17	18:05:07	37	17	96	NW	9	0	29.3	0.0	0.8	fair

APPENDIX C



Selected attribute

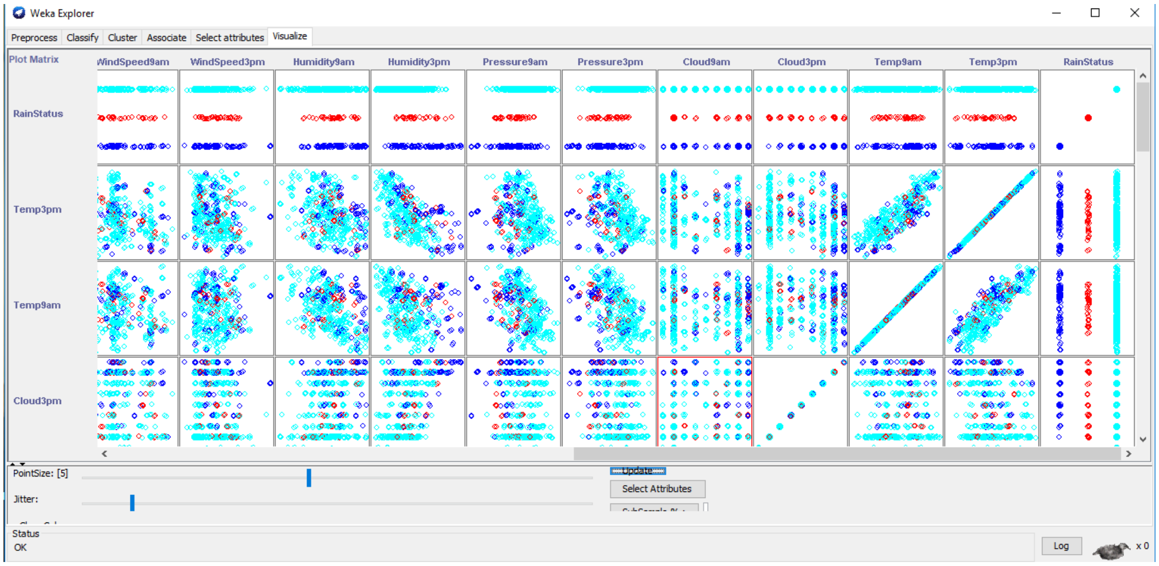
Name: WCondition Type: Nominal

Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

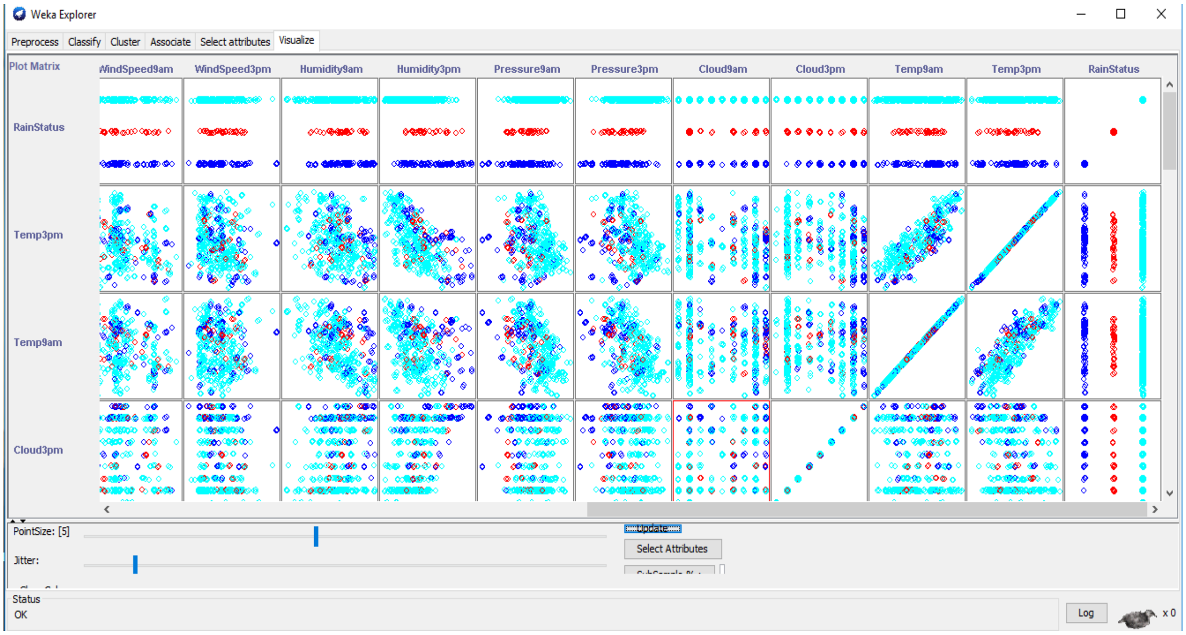
No.	Label	Count
1	partly cloudy	848
2	mostly cloudy	798
3	fair	872
4	cloudy	844
5	fog	798
6	light rain shower	841

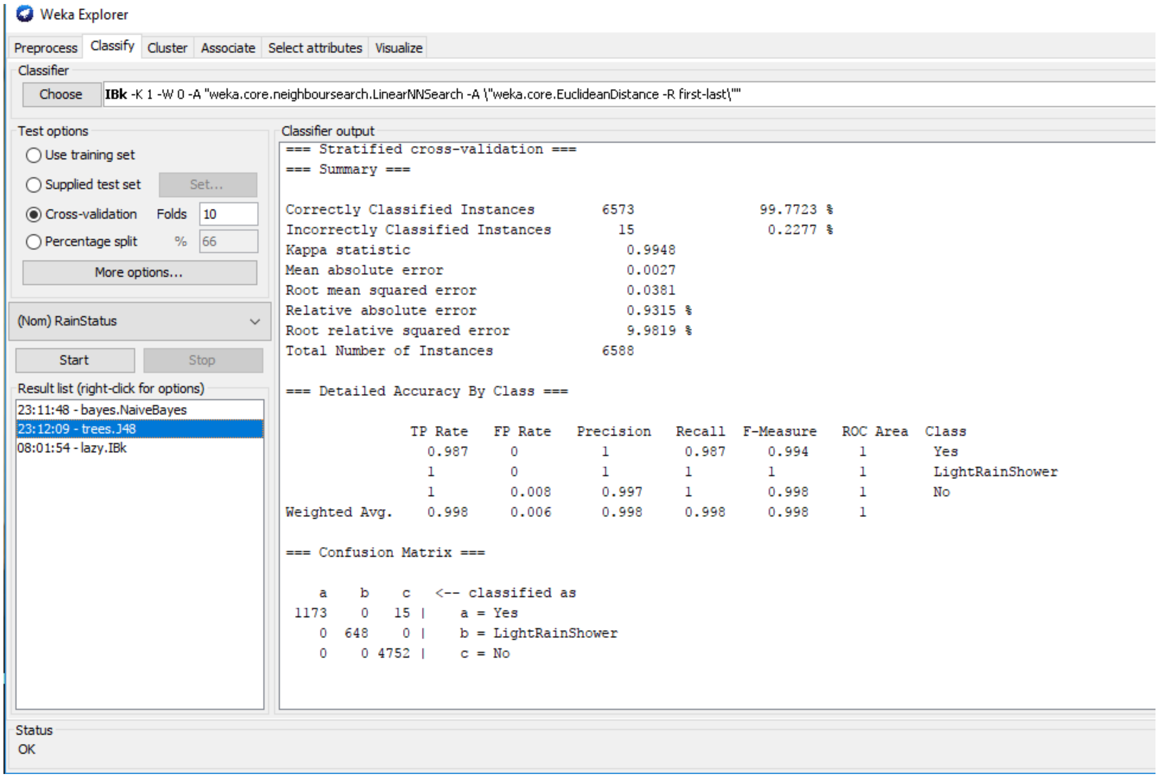
Class: WCondition (Nom)

Activate Windows
Go to Settings to activate Windows.



APPENDIX D





IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

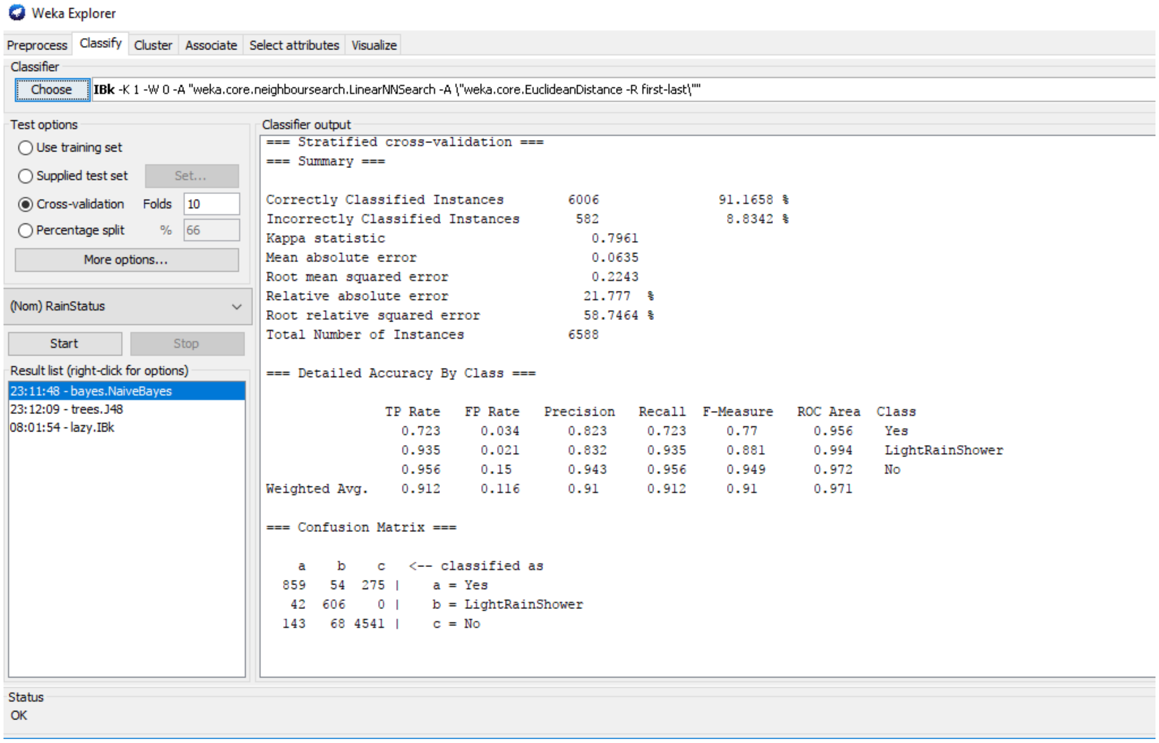
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) RainStatus

Result list (right-click for options)

- 23:11:48 - bayes.NaiveBayes
- 23:12:09 - trees.J48
- 08:01:54 - lazy.IBk

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6573               99.7723 %
Incorrectly Classified Instances        15                0.2277 %
Kappa statistic                          0.9948
Mean absolute error                      0.0027
Root mean squared error                  0.0381
Relative absolute error                  0.9315 %
Root relative squared error              9.9819 %
Total Number of Instances             6588

=== Detailed Accuracy By Class ===

              TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                0.987     0          1         0.987     0.994       1         Yes
                1         0          1         1         1           1         LightRainShower
                1         0.008      0.997     1         0.998       1         No
Weighted Avg.   0.998     0.006      0.998     0.998     0.998       1

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1173    0   15 |    a = Yes
    0  648    0 |    b = LightRainShower
    0    0 4752 |    c = No
```



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

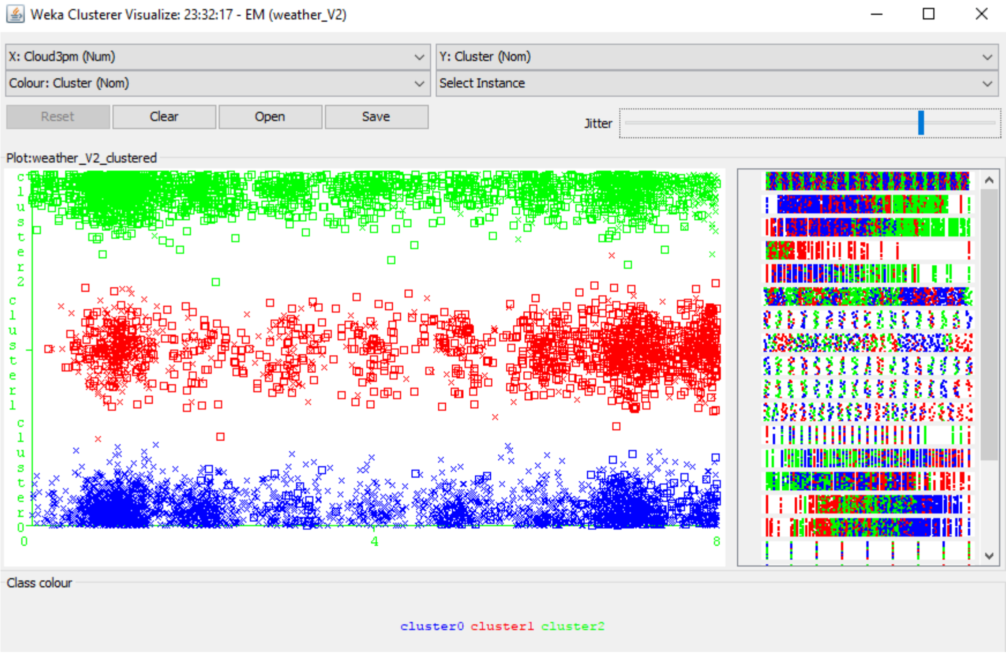
(Nom) RainStatus

Result list (right-click for options)
23:11:48 - bayes.NaiveBayes
23:12:09 - trees.J48
08:01:54 - lazy.IBk

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6006               91.1658 %
Incorrectly Classified Instances       582                8.8342 %
Kappa statistic                          0.7961
Mean absolute error                      0.0635
Root mean squared error                  0.2243
Relative absolute error                 21.777  %
Root relative squared error             58.7464 %
Total Number of Instances             6588

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                 0.723    0.034    0.823     0.723    0.77       0.956    Yes
                 0.935    0.021    0.832     0.935    0.881      0.994    LightRainShower
                 0.956    0.15     0.943     0.956    0.949      0.972    No
Weighted Avg.    0.912    0.116    0.91      0.912    0.91       0.971

=== Confusion Matrix ===

    a    b    c   <-- classified as
  859   54  275 |    a = Yes
   42  606    0 |    b = LightRainShower
  143   68 4541 |    c = No
```

Status
OK

APPENDIX E



Weka Clusterer Visualize: 23:32:17 - EM (weather_V2)

X: Cloud3pm (Num)

Y: Cluster (Nom)

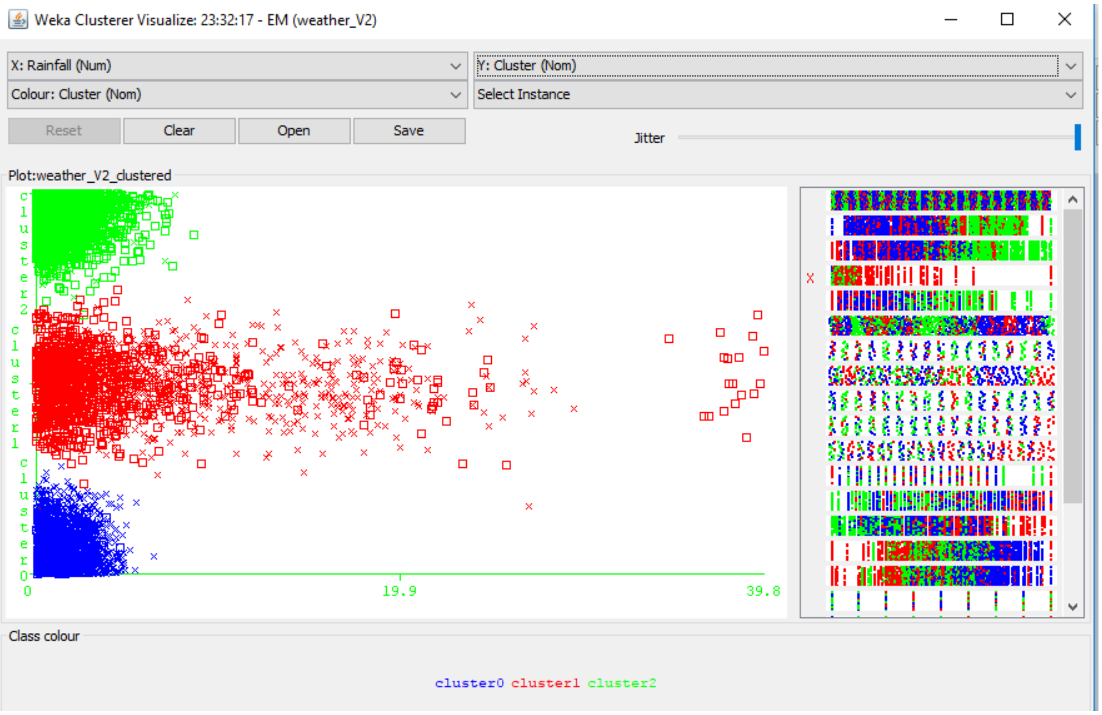
Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered

Class colour

cluster0 cluster1 cluster2



Weka Clusterer Visualize: 23:32:17 - EM (weather_V2)

X: Rainfall (Num)

Y: Cluster (Nom)

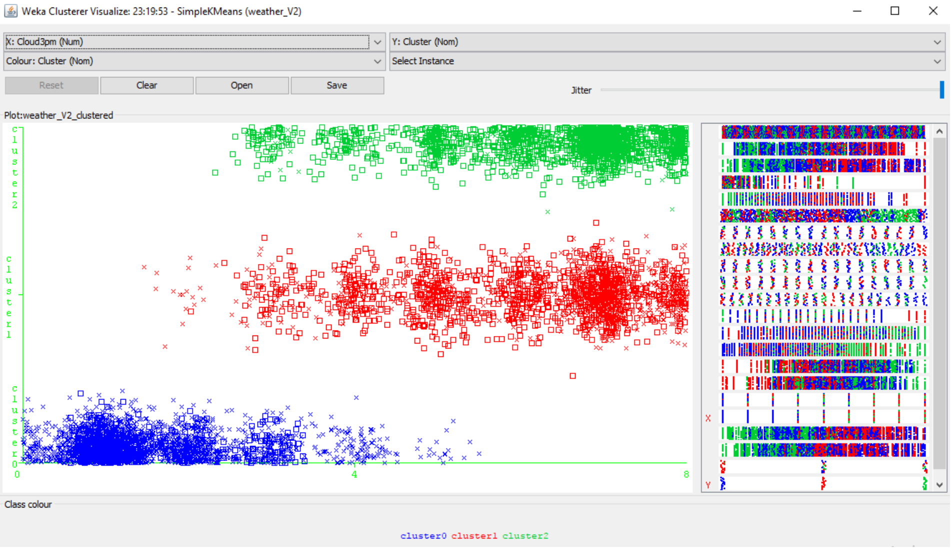
Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered

Class colour

cluster0 cluster1 cluster2


Weka Clusterer Visualize: 23:19:53 - SimpleKMeans (weather_V2)

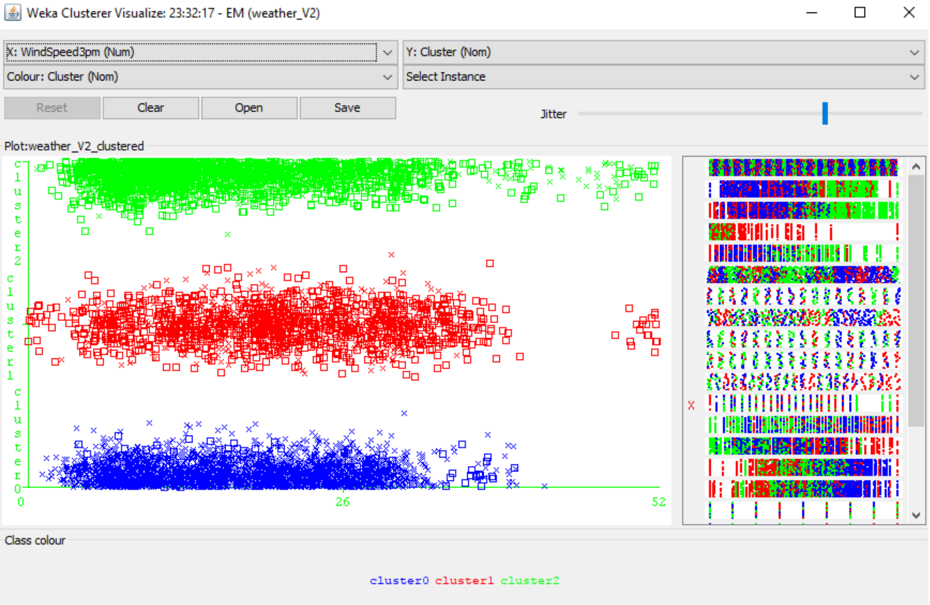
X: Cloud3pm (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered


Weka Clusterer Visualize: 23:32:17 - EM (weather_V2)

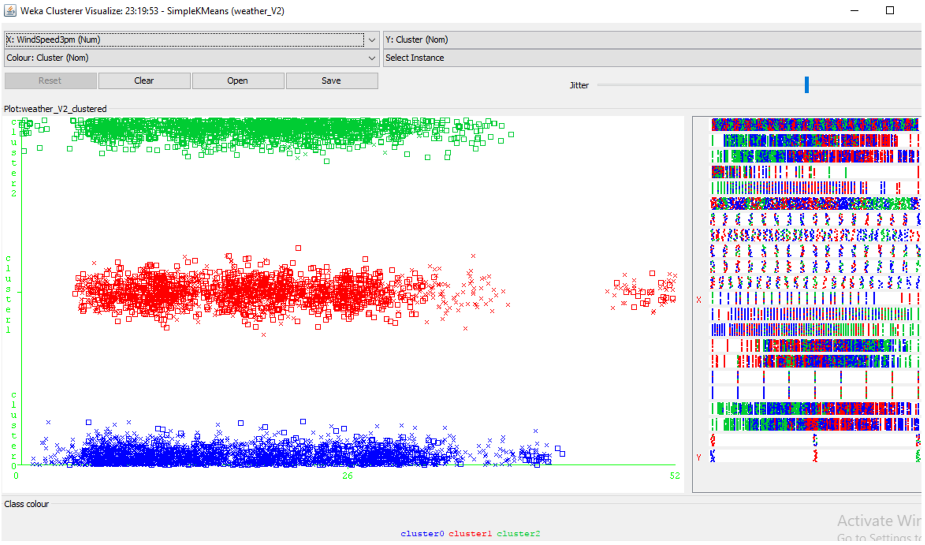
X: WindSpeed3pm (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered


Weka Clusterer Visualize: 23:19:53 - SimpleKMeans (weather_V2)

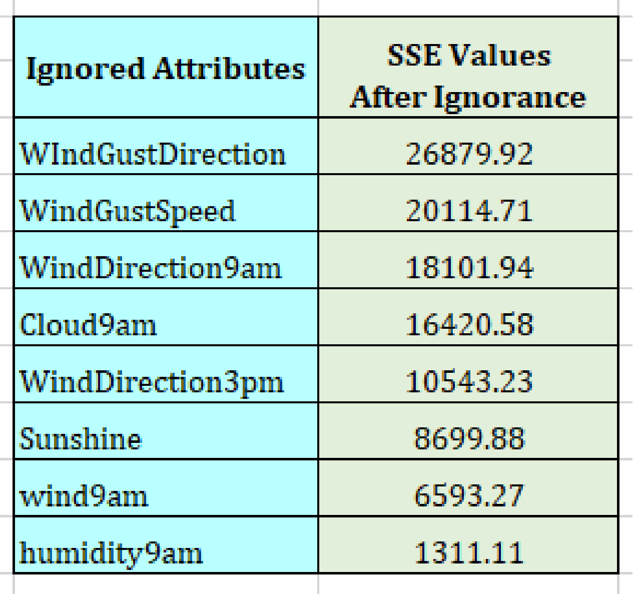
X: WindSpeed3pm (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered



Ignored Attributes	SSE Values After Ignorance
WIndGustDirection	26879.92
WindGustSpeed	20114.71
WindDirection9am	18101.94
Cloud9am	16420.58
WindDirection3pm	10543.23
Sunshine	8699.88
wind9am	6593.27
humidity9am	1311.11