Comparison of Different Classification Algorithm Based on Significant Water Parameters

ومن يؤت الحكمة
فقد أوتي خيرا كثيرا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

By
Tayyaba Aurangzeb

Department of Statistics
Faculty Of Natural Sciences
Quaid-i-Azam University, Islamabad

2018

بسم الله الرحمن الرحيم

In the name of Allah the most merciful and the most beneficent

Comparison of Different Classification Algorithm Based on Significant Water Parameters

ومن يؤت الحكمة
فقد أوتي خيرا كثيرا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

By

Tayyaba Aurangzeb

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MPHIL IN STATISTICS

Supervised By

Dr. Ijaz Hussain

Department of Statistics
Faculty Of Natural Sciences
Quaid-i-Azam University, Islamabad
2018

CERTIFICATE

Comparison Of Different Classification Algorithms Based On Significant Water Parameters

by

TAYYABA AURANGZEB

(Reg.No.02221611009)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF THE MASTER OF PHILOSOPHY IN STATISTICS

We accept this thesis as conforming to the required standard.

1. 3/8/18
Dr. Zahid Asghar
(Chairman)

2. ____________________
Dr. Ijaz Hussain
(Supervisor)

3. ____________________
Dr. Muhammad Hanif
(External)

Date. ____________________

Department of Statistics
Faculty Of Natural Sciences
Quaid-i-Azam University, Islamabad
2018

Declaration

I "Tayyaba Aurangzeb" hereby solemnly declare that this thesis entitled "Comparison of different classification algorithm based on significant water parameters ", submitted by me for the partial fulfillment of Master of Philosophy in Statistics, is the original work and has not been submitted concomitantly or latterly to this or any other university for any other Degree.

Dated: 2-08-2018

Signature:

Dedication

I am feeling great honour and pleasure to dedicate this research work to

My Beloved Parents

My Mamoo Malik Hanif, Malik Nasir, Kashif, Malik Sherbaz and My Siblings

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period.

Acknowledgment

First and foremost I praise and acknowledge Allah Almighty, the Lord and creator of the heavens and earth. All respect and gratitude goes to the Holy Prophet Muhammad (Peace be upon him) who enlightens our hearts with the light of Islam and whose way of life has been a continuous guidance for us.

I deem it my utmost pleasure to avail this opportunity and express gratitude and deep sense of obligation to my supervisor **Dr. Ijaz Hussain** for his valuable and dexterous guidance, scholarly criticism, untiring help, compassionate attitude, kind behavior and moral support.

I offer my deepest sense of gratitude, profound respect and tribute to all those teachers who have enlightened my statistical knowledge through out my academic career and guidance throughout research work.

My sincere thanks go to my parents for their love and support throughout my life. I owe my loving thanks to my brothers, cousins and my whole family for their great love and support. Without their support and encouragement it would have been impossible for me to complete this work. I would like to thank all my friends and classmates for their cooperation and help especially *Hafiza Memona Nazir, Neelum Bibi, Lubna, Saba Abbasi, Fatima Junaid, Sultana, Amna and Kiran Shahzadi.* Finally, I would say thanks to all who have been supportive and cooperative to me during my research work.

Abstract

In recent years, water issues have come to claim a firm position among the top challenges facing globally. Through the Sustainable Development Goals, it becomes a global concern where a goal dedicated specifically to water and its sanitation system. It is said in World Economic Forum, that water issue will become a major issue in coming years. Like many other developing countries, Pakistan is also facing great public health challenge due to unhygienic and polluted water. People in our country are forced to buy bottled water because of poor quality. The intake of bottled water has been increasing constantly over the last era, even in countries wherever tap water quality is reported excellent. But it is found that many mineral water companies are selling contaminated water. The present study aimed to monitor Bottled water quality of 538 commercially available brands of mineral water from local market of 15 cities of Pakistan, from January-2011 to September-2015. All mineral water brands data were analyzed by using physio-chemical parameters named; pH, Electrical Conductivity (EC), Total Dissolve Solids (TDS), Calcium (Ca), Magnesium (Mg), Hardness, Bicarbonate (HCO3), Chloride (Cl), Sodium (Na) and Sulfate (SO4). In order to communicate the quality of water, it is needed that all parameters should be compressed in a standard format to interpret the quality of water precisely. Three Water Quality Indices have been used to evaluate the quality of bottled water, named Weighted Arithmetic Water Quality Index, National Sanitation Foundation Water Quality Index and Water Quality Index (WQI) . Classes of Bottled WQI falls in three groups i.e. Excellent class of water, Good class of water and Poor class of water. Out of 538 brands, 56 were found to supply very poor quality of water. Moreover, the objective of this research is two fold, first Bottled water data is analyzed by using supervised machine learning algorithms such as artificial neural network, support vector machine by using different kernel functions and random forest Model and its improved version i.e C4.5, C5.0. Each algorithm is trained using 80% of data and remaining 20% of data is used for testing purpose. Comparison is made within and between algorithms to demonstrated the most useful classification method to classify the quality of Bottled water with smallest error rate. Results reveal that although random forest methods showed highest accuracy rate, however C.50 is the most useful to classify test data with high accuracy level within minimum time and required less memory as compared to remaining algorithms. Support vector machine with complex polynomial kernel revealed significant result. The second objective is to reduce the dimension of data in such a way that classification of bottled water quality is preserved. To accomplish this goal, Principal Component Analysis (PCA) and t-SNE is used to determine the lower dimension of data.

PCA determines the optimal linear combinations that appropriately explain the data. It is concluded that four component explain the 80.4% variation of the data. The t-SNE is another method which is more efficient dimension reduction in such a way that similarity of classes is remain in lower dimension well. Graphs of both methods are displayed and it is observed that t-SNE is more stable algorithm by showing clustering of different classes accurately. Pakistan Council of Research in Water Resources (PCRWR) should review the quality of Bottled water consistently, and it is also suggested that PCRWR should take forward steps to ban those brands which found by saling contaminated water.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Water is a chemical substance which is the most important constituent of Earth. Its chemical formula is H_2O, which means that each water molecule consists of two Hydrogen atoms and one oxygen atom that are connected by covalent bonds. Water is found in three different forms on Earth like liquid, solid and gas. Water flows as liquid in streams, rivers and oceans; its solid form is ice and in gas form is vapor and steam in the atmosphere. Water covers about 71% of the Earth's surface. It is very important for all known forms of life. On Earth, about 96.5% water is found in seas and oceans, 1.7% in groundwater, 1.7% in glaciers and the ice caps of Antarctica and Greenland, 0.001% in the air as vapor and clouds. From them only 2.5% of water is found fresh and useable without any treatment process. Less than 0.3% of all fresh water is in lakes, rivers and smaller amount of the Earth's freshwater (0.003%) is contained within biological bodies and manufactured products. A larger quantity of water is found in the underground of surface.

The amount of water in the human body is 50 to 75%. The percentage of water in newborns is much higher, usually around 75 to 78% water, dropping to 65% by one year of age. To work properly, the body requires between 1 to 7 glass of water per day to avoid dehydration, apprehension, weakness and exhaustion; the precise amount depends upon the level of activity, humidity, temperature and other factors. Water, in human body, controls the working of liquids, tissues, lymph, blood, cells and glandular emissions.

Water quantity is certainly a major problem in all over the world. Scarcity resulted when the physical quantity of water is low or water scarcity refers to the situation wherein fresh water supply is not sufficient to fulfill the needs of the people. Water scarcity typically occurs because of low rainfall. Rain water is the most significant source of fresh water. In regions with low rainfall, the ground water is recharged; rivers get dried up, thereby resulting in water scarcity. The most common cause of water scarcity is overexploitation and mismanagement of water resources. Large scale urbanization and industrialization which have increased the consumption of water, thereby resulting in water scarcity. Two other factors that add to water scarcity is the expansion of agriculture and deforestation, the number of trees decreases, that's why the water keeps flowing towards the sea and the ground water is not recharged. Consequently, the use of water in a way that its consumption exceeds its replenishment is the cause of water scarcity.

World Health Organization (WHO) stated that, approximately three in ten people suffering lack of water. Within this, 844 million people have lack of access to basic water and 159 million people are forced to use untreated surface water direct from rivers and lakes for their daily needs. Due to poor and drainage water, diseases such as typhoid, cholera, polio and diarrhea spread rapidly across communities. The WHO estimates that every year 361,000 children under the age of five die due to using the contaminated water. Ali et al. (2012) analyzed that, situation in developed countries is not serious, as 95% of the population has access to clean drinking water. But, the situation is worse in under developed countries where most of the people consume the contaminated water with unacceptable levels of toxic chemicals, pathogenic microorganisms or suspended solids. The (UNICEF and WHO 2012) United Nations has estimated that, at least 780 million people in developing countries have no access to clean drinking water and over 2.5 billion people have no proper sanitary system. Like other developing countries, Pakistan is also suffering same issue related to purify clean drinking water. According to World Bank Report, Pakistan is currently among the seventeen countries that encountering water shortage. The major reason is the absence of proper water policy in the country. Kamal (2009) investigate that, Pakistan has a 165 million of population, out of which approximately 41 million (25 percent) are under the poverty area, 50 million have not access to safe and clean drinking water; and 74 million people have sanitation problem. Jabeen et al. (2015) highlights an important contribution in water contamination is industrial wastage (including rubbish, chemical, insecticides, pesticides, leather, tanneries and pharmaceuticals). It is roughly estimated that from 6634 registered industries from which 1228 industries are considering extremely polluted by directly releasing their wastes into rivers. Round about 2000 million gallon of sewage is discharged into surface water without any refining treatment in Pakistan. Only one percent water is processed by industries before releasing it to channel or canal. There are various troubles arising due to beyond mentioned ecological pollution. According to Asian Development Bank (ADB), approximately 3.1% death happens due to the unsafe water. Low quality of drinking waterand sanitation lead to main outbreaks of waterborne diseases and main outbreaks swept the cities of Lahore, Faisalabad, Karachi and Peshawar in 2006. However, results indicate that every year, more than three million Pakistanis affected with waterborne diseases. Furthermore, Farid et al. (2012) investigate that agriculture production is also affecting water pollution by contaminated with fertilizers and pesticides which are mixed in water, irrigation is effecting by seeping into underground water. As rapidly increasing the population growth is also posing great threat to water as well.

Due to above mentioned factors of pollution, unpured quality of underground drinking water has forced a several citizens to purchase bottled water. Bottled water quality is usually good compared to tap drinking water but there is possibility that it can also suffer from the same contamination risk as tap water. El Aal et al. (2015) evaluated the quality of various mineral water brands from Egypt and Saudi Arabia. They collect sample from twenty four bottled water brands from different areas of Egypt and Saudi Arabia. Every brand is analyzed for thirty four chemical parameters such as Li, Ta, Ti, Mo, Nb, Si, P, Re, V, Ca, Ge, Cr, K, B, Al, Na, Zn, Ni, Se, Sr, Ag, Ga, Fe, Pb, Co, Cu, Cs, Be, Cd, As, Bi, Mg, Mn, Ba,. They used chemical or compound method to categorized and identify the concentration level of considered sample. Their results revealed unsatisfactory quality of water. Therefore, to improve bottled water quality, companies should monitored to assess the quality of water on regular basis. It is necessary that consumers have access to important information directly on the bottles quality labels, that is the type of water (purified water or not, etc.).

The management of Pakistan has assigned the Pakistan Council of Research in Water Resources (PCRWR) as focal point agency to watch over the quality of mineral bottled water. To enhance the nature of filtered water quality, the government through the Ministry of Science and Technology has assigned the undertaking to PCRWR for quarterly observing of mineral/packaged water marks and broadcast the outcomes. As indicated in the report of PCRWR from October to December (2015) in which a further 100 specimens of filtered/mineral water brands were collected from Bahawalpur, D.I.Khan, Faisalabad, Gujranwala, Islamabad, Karachi, Lahore, Multan, Peshawar, Quetta, Rawalpindi, Sargodha, Sialkot, Sahiwal and Tandojam. Comparison of investigative discoveries with acceptable points of confinement of Pakistan Standards and Quality Control Authority (PSQCA) showed that thirteen brands, for instance, Al-Haider, Al-Sahar, Aqua National, Deep Pure Water, Eau Water, Eden Pure Water, Mazan, New Nation, NGFresh Water, Oriel, PureLife, Premier and Water Icon were observed to be dangerous because of microbiological sullying or compound. Another three filtered water brands, Eden Pure Water, Al-Sahar and Water Icon were observed to be unsafe or risky because of microbiological tainting. Microbiological pollution, as identified by the PCRWR, may bring out diarrhea, looseness of the bowels, cholera, typhoid, Hepatitis and so forth. The remaining ten brands stated above, were discovered hazardous because of the nearness of more elevated amounts of potassium and sodium. The PRCWR suggested that the national quality control powers should to make legitimate move against the companies which were selling defiled filtered water to the citizens.

1.1 Literature Review

Water is one of the most significant compound of the ecosystem. All living organisms on the surface require water for their survival and growth. Rapidly increasing the population, industrialization, use of fertilizers in the agriculture and man-made activity, it is highly polluted with different harmful contaminants. Therefore, it is necessary that the drinking water quality should be checked at regular basis, because due to use of contaminated drinking water, human suffers from water borne diseases. According to Disease Control Center (CDC), waterborne diseases are mostly due to pathogenic microbes that directly spread from contaminated water. Most of waterborne diseases spread due to diarrhea and 88 percent of diarrhea cases are related to unclean water, insufficient hygienic or inadequate sanitation. These cases result in 1.5 million deaths every year, mostly in young children and usual cause of death is dehydration.

International Bottled Water Association (IBWA, 2000) discussed about the purified water, taken from lakes, rivers or underground springs that have passed through some cleaning process. It is treated by reverse osmosis, deionization, distillation or other suitable processes. It is also chemically treated in order to disappear various components. Considering the way it is treated, IBWA suggested that there is a slight difference between municipal tap water and purified water (Klein and Huang (2008)). The study mentioned the significance of water quality review in India. For that purpose, they studied various physic-chemical parameters that are used for testing the quality of water. For this testing process, selection of parameters completely depends upon the purpose for which we are going to use that water. They suggested that drinking water have to proceed through chemical and physical tests for attaining purity and high-quality of water (Patil et al. (2012)).

Samuel et al. (2016) analyzed the quality of bottled water of several brands in Awka, Nigeria. They collected samples of 21 different brands of water from seven main markets in Awka. Samples have been referred to microbiological laboratory for further examination analysis. Samples are collected frequent basis in order to test the quality and microbiological tests were performed. They reported that the readings of temperature have been taken using a thermometer and for pH meter is used for readings. There result indicates that water quality of several brands was not suitable according to the standards recommended by World Health Organization.

Water Quality Index is one of the most efficient tools to communicate information on the quality of water to the concerned policy makers and citizens according to Atulegwu and Njoku (2004). WQI is defined as, a rating reflecting the composite affect of different water quality parameters. WQI is computed from the point of view of the suitability of surface water for human utilization. Kalavathy et al. (2011) used the WQI to evaluate the quality of river water in tamilnadu, India. For this purpose, they collect sample from 4 stations of river over different time period. 10 physico chemical parameters, such as pH, sulphate, nitrate, total dissolved solids, total hardness, chloride, biological oxygen demand, total alkalinity, calcium and dissolved oxygen are used. They used weighted technique to find WQI and compared results of WQI with Central Pollution Control Board and Bureau of Indian Standards and concluded that, station 4 is unsafe for drinking, third station is extremely polluted while station 1 and 2 are moderately polluted. Sufficient treatment is required before its consumption. Furthermore, Karavoltsos et al. (2008) evaluate the quality of water from different areas of Greece. Physico chemical parameters were used such as conductivity, TDS, bromide, pH, chloride, magnesium, ammonium, potassium, sodium, nickel, copper, calcium, cadmium, lead, chromium, fluoride, nitrates, sulfates, phosphates and nitrites. They classified these parameters according to their nature like physico-chemical parameters, 10 heavy metals and anions.

The researcher discussed different WQI such as Canadian Council of Ministers of the Environment Water Quality Index (CCME-WQI), British Columbia Water Quality Index, US National Sanitation Foundations Water Quality Index (NSF-WQI), Florida Stream Water Quality Index and Oregon Water Quality Index for evaluating the water quality. Hence, he suggest continuous periodical checking of water quality is necessary so that suitable steps are taken for water resource management practices (Kannel et al. (2007)). The supervised learning machine methods provide a better understanding of water quality for the interpretation and prediction of complex datasets. Supervised learning methods, which permits the user to examined the data set from different dimensions and to classify them, (Gupta and Agarwal (2010)) and (JIAWEI et al. (2007)). Many methods are used in supervised learning to extract information from large amount of database. Classification is a supervised method used to predict the group membership for data attributes. The classification methods include decision trees (DTs), k-Nearest Neighbor (kNN), neural networks and logistic regression. Decision trees are widely used for classification because of their ability to manage noisy data and which concludes the value of a dependent attribute given the values of the independent attributes. Yadav et al. (2012) performed a comparative analysis of decision tree techniques such as CART, ID3 and C4.5. This analysis was performed on forty eight students of MCA course from a Jaunpur university. The result of this analysis showed that CART technique had better accuracy than the other two decision tree techniques. In another analysis the recommendation for promotion (yes/no), is considered as the target class in classification procedure. For human talent data set, Jantan et al. (2009) used employees data set from one of Malaysian higher learning institutions as a training data set. The C4.5 and C5.0 classifier are used to generate the talent performance knowledge from yearly performance evaluation database. They conclude that, C5.0 decision tree is better than C4.5 decision tree in term of memory and efficiency. Furthter, random forest method has been effectively used in the past, providing accurate land cover maps. Analysis using the hyperspectral data for forest sciences revealed that random forest is effectively used to detect insect infestations, extract physiological plant characteristics and plant biomass (Ghimire et al. (2010)).

The support vector machine, fundamentally a kernel-based process, is a relatively new machine learning technique, Singh et al. (2011) that recently emerged as one of the leading techniques for function approximation purposes and pattern classification.

Support vector machine can simultaneously reduce the model dimensions and estimation error it has better generalization ability and is less at risk of over-fitting. SVMs have been efficiently applied for regression and classification purposes. The study used a support vector regression and classification models to surface water quality data to improve the monitoring program (Asefa et al. (2005)). More than that, the probabilistic neural network (PNN), artificial neural networks (ANN), k-nearest neighbor (KNN) and support vector machine (SVM) techniques were also used to classify water quality datasets (Modaresi and Araghinejad (2014a)). Furthermore, artificial neural network (ANN) is an information processing device that is inspired by the way like biological nervous system such as brain. The main purpose of neural network is to calculate output values from input values by some inner calculations (Delgrange et al. (1998)). These techniques have different characteristics and contribute to analyzing the water quality classification and recognition; yet, there are still debates on the best techniques for the evaluation of water quality.

Furthermore, the multivariate statistical techniques also evaluate water quality datasets. Traditional multivariate techniques, including cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) have been extensively studied. These multivariate techniques are generally used to evaluate spatio-temporal variation in river pollution and classify the related sources (Tsakovski et al. (2009); Zhou et al. (2007b); Zhang et al. (2009) and Zhou et al. (2007a)). Cloutier et al. (2008) used the principal component analysis and cluster analysis are applied to hydrochemical data to demonstrate the regional impact of human activities on ground water composition. An alternative of PCA is t-stochastic neighborhood embedding (t-SNE) that maps high dimensional data to a lower dimension although preserving the local character of the dataset. t-SNE has revealed its built-in advantages in capturing local data characteristics and showing subtle data structures in visualization (Maaten and Hinton (2008) and van der Maaten (2009)). In recent years, various studies have used a number of relatively sophisticated techniques; including the Hasse diagram technique (HDT), N-way principal component analysis (N-Way PCA) and support vector machines (SVMs) for the classification, modeling, and interpretation of many environmental compartments. But, these sophisticated techniques have not been completely explored in river studies in China (Tsakovski and Simeonov (2011) and Astel et al. (2008)).

1.2 Objective of study

The main objectives of this study are detailed as follows:

1. To evaluate the quality of bottled water of various brands in Pakistan

2. To find the significant parameters for classification of bottled water quality

3. To reduce the dimensionality of bottled water data set and to identify the meaningful parameterse quality of Bottled water data.

4. To visualize the groups of different bottled water classes effectively by using t-Stochastic Neighbor Embedding

1.3 Composition of the Thesis

This thesis consists of 6 chapters. Chapter 1 contains the brief introduction, review of literature and objective of the study. In chapter 2, brief description of data and by using the water quality indices. Chapter 3, based on the C4.5, C5.0 and random forest methods used to find the significant parameters of bottled water data set. Chapter 4, based on the artificial neural network using the backpropagation and multi-class support vector machine by using different kernel function, both are used to detect the accuracy rate of bottled water quality. Chapter 5 based on PCA and t-SNE methods. Chapter 6 describe the brief conclusion about the study.

Chapter 2

Water Quality Index

2.1 Introduction

Water a precious national asset and prime natural resource, forms the main constituent of ecosystem. Water sources may be generally in the form of lakes, rivers, glaciers, ground water and rain water etc. Furthermore the need of water for drinking, water resources play a crucial role in many sectors of economy such as livestock production, agriculture, forestry, hydropower generation, fisheries, industrial activities and other creative activities. The quality and availability of water either ground or surface, have been deteriorated due to some important factors such as an increasing population, urbanization and industrialization etc.

Quality of water of any specific area or specific source can be assessed using biological, chemical and physical parameters. The values of such parameters are harmful for human health if they occurred more than defined limits (Sharma and Tyagi (2013)). Thus, the suitability of water sources for human consumption has been described in terms of Water Quality Index (WQI), which is one of the most effective method to describe the water quality. WQI is used to observe the water quality and helps in the modification of the rules and policies, which are formulated by several environmental monitoring agencies. For this reason, WQI has the capability to minimize the bulk of the information into a single value to represent the data in a simplified and logical form. In present study, we review some of the crucial indices used in water quality evaluation and provide their mathematical structure, set of parameters and calculations, which are being used worldwide.

2.2 Study Area and Data

The main purpose of study is to monitor the bottled water quality of 538 commercially available brands of mineral water from local market of fifteen cities of Pakistan named as: Islamabad, Lahore, Rawalpindi, Sialkot, Sargodha, Bahawalpur, Karachi, D.I Khan, Peshawar, Faisalabad, Sahiwal, Gujranwala, Multan, Quetta and Tandojam. Data is collected quarterly from Jan-Mar (2011) to July-Sep (2015) regularly except Oct- December (2011). Sample contained 1495 observations of water brands. It is an unbalanced data that not all brands were analyzed frequently due to some reasons. But, some brands which are frequently observed over this time period are Springley, Nestle, Hydra, Isberg, Sufi and Kinley; and frequencies of these brands in sample are 18,18,17,16,16 and 16 respectively. Samples were collected from senior staff member of Pakistan Council of Research in water resources. Sample of four bottles are collected from every brand and its quality is analyzed by measuring following parameters. Brief description of every parameter is discussed as follows:

1) Physio-chemical: (pH, Electrical Conductivity (EC), Total Dissolve Solids (TDS), Calcium (Ca), Magnesium (Mg), Hardness (Hard), Bicarbonate (HCO3),Chloride (Cl), Sodium(Na), Sulfate (SO4)).

2) Microbiological: (Total Coliforms. E-Coliforms).

pH: The pH is determined the hydrogen ion concentration in water. According to the different standards recommended through WHO, CPCB, BIS, and ICMR, the range of pH lies between (6.5-7) in the water. When the value is less than 6.5, it discontinues the making of minerals and vitamins in the human body and pH greater than 7 cause the flavor of water more salty. Further, if pH is more than 11 then it causes of eye inflammation and skin ailment. The rainwater which has no minerals useful for human body has a pH of (5.56) and not hazardous on used for drinking purpose.

Total Dissolved Solids: Total Dissolved solids (TDS) is checked for measuring the amount of solid materials dissolved within the water (ground, surface). Sources of TDS in drinking water are sewage, industrial waste water, natural sources and urban runoff. Its concentration level in water varies from area to area, depending at the solubility of minerals. Due to high value of TDS, that is greater than 1000mg/l, drinking water causes detrimental effect to the peoples health such as the irritability, dizziness, central nervous system, provoking paralysis of the tongue, lips, and face.

Electrical conductivity: Water capacity to transmit electric current is known as Electrical Conductivity (EC) and served as a device to evaluate the purity of water (Murugesan et al. (2006)). This capability depends on the presence of ions, their total concentration, valence, mobility, relative concentrations and temperature of measurement. Standard value of EC in drinking water is 1400/cm. Heart and Kidney patients, kids under age of one year and individuals having chronic diarrhoea are highly affected where EC in drinking water exceed from the standard values.

Calcium: Calcium also denoted by Ca, is the most abundant ions in fresh water. It plays a vital role in blood clotting, cell signaling, muscle contraction and bone structure. Standard value of calcium in drinking water is 75ml/l. In human body ninety nine percent of calcium is present in teeth and bone, and the remaining is found in soft tissues. Its utilization in drinking water decreases the threat of kidney stones.

Magnesium: Magnesium is often related with calcium in all types of waters, but its concentration remains usually lower than the calcium. Its permissible level in water is 150mg/l. There is a significant protecting effect of magnesium taking from drinking water on the risk of cerebrovascular illness.

Hardness: The hardness of water is not a particular constituent but is a variable and complex combination of anions and cations. Mostly the water hardness is changed by ions which includes calcium and magnesium. According to WHO its standard value in drinking water is 500mg/l. It is very significant for human physical condition, due to its presence, threat of heart disease reduced.

Hydrogen Carbonate: Hydrogen carbonate is denoted by HCO3 and a crucial parameter of water. It is a natural factor, in human body it formed by itself. An element of the salts of carbonic acid, but it is not a mineral. According to WHO standard value for it is 500mg/l. HCO3 also plays significant role in the digestive system. It expanded the interior pH level of the stomach, after surprisingly acidic digestive juices have done in their processing of nourishment.

Chlorine: It is usually denoted by Cl and naturally occurring element. It combines with chemical dissolved in water. It is used to kill certain bacteria and other microbes in water, as chlorine is extremely toxic and used to prevent the spread of waterborne diseases which are cholera, typhoid and dysentery. It also eliminates molds, algae and slime bacteria that usually develop in water storage bottles or tanks. In drinking water its standard level is 250mg/l.

Sodium: Sodium is denoted by Na and it can occur naturally or be the result of water treatment chemicals, road salt application and ion-exchange softening units. Sodium is not taken consideration to be toxic. The human body needs sodium in order to preserve blood pressure, muscle function, control fluid levels and for normal nerve. Permissible level of sodium above 200mg/l may change the taste of drinking water.

Sulfate: Sulfate (SO4) can be found in approximately all natural water. The basis of most sulfate compounds is the oxidation of sulfite ores, the industrial wastes, or the presence of shales. Sulfate is one of the main dissolved components of rain. High concentrations of sulfate in the water we drink can have a laxative impact when mixed with magnesium and calcium, the two most frequent constituents of hardness. Its permissible level by WHO is 250mg/l. When its level exceeds in blood, it quickly removed through urinary excretion. Its concentration level in drinking water increasing from 500 to 700mg/l caused of dehydration and diarrhea.

2.3 Methodology

Different types of Water Quality Index

Firstly, WQI was proposed by Horton in United States by selecting 10 most commonly used water quality variables like alkalinity, pH, dissolved oxygen (DO), coliforms, specific conductance and chloride etc. It has been widely use and accepted in African, Asian and European countries. The assigned weight reflects the importance of a parameter for a particular use and has great impact on the index. The main advantage of water quality indices is that, they efficiently give the overall water quality of a particular area. Different water quality indices developed worldwide are Oregon Water Quality Index (OWQI), US National Sanitation Foundation Water Quality Index (NSFWQI), British Columbia Water Quality Index (BCWQI) and Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI). These indices provide the water quality in a single value by comparing parameters as per the standard values. Standard values, weights and relative weights of each parameters in drinking water, recommended by WHO are listed in Table 2.1.

Table 2.1: Standards values, Weights and Relative weights recommended by WHO

Parameters	Standard Values	Weights	Relative Weight
pH	7	4	0.1333
TDS	1000	4	0.1333
Cl	250	3	0.1000
$SO4$	250	4	0.1333
Ca	75	2	0.0667
Mg	150	2	0.0667
$Hardness$	500	2	0.0667
EC	400	2	0.0667
$HCO3$	500	3	0.1000
Na	50	4	0.1333

2.3.1 Weighted Arithmetic Water Quality Index Method

Weighted arithmetic water quality index method (WAWQI) categorized the water quality according to the degree of clarity and purity by using the most frequently measured water quality variables. The technique has been extensively used by the many scientists (Chowdhury et al. (2012); Rao et al. (2010)) . The calculation WAWQI is measured by using the following equation:

$$WQI = \sum Q_i W_i / \sum W_i \tag{2.1}$$

where Q_i is the quality rating scale for each parameter, and calculated by using following expression:

$$Q_i = 100[(V_i - V_0)/(S_i - V_0)] \tag{2.2}$$

where V_i is estimated concentration of i^{th} parameter in the analyzed water, S_i is recommended standard value of i^{th} parameter and Vo is the ideal value of this parameter in pure water. All the ideal values V_0 are taken as zero for drinking water except pH and dissolved oxygen that is ($V_0 = 0$, except pH =7.0 and DO = 14.6 mg/l), because values of these parameter lies in point (Tripathy and Sahu (2005)). Thus, the unit weight (W_i) for every water quality parameter is calculated by using the following equation:

$$W_i = K/S_i \tag{2.3}$$

here, K = proportionality constant and can be measured by using the this expression:

$$K = \frac{1}{\sum(\frac{1}{S_i})} \tag{2.4}$$

But the drawbacks of weighted arithmetic water quality index are, may not carry enough information about the real quality of water and a single parameter value change the whole story of WQI.

2.3.2 National Sanitation Foundation Water Quality Index Method

In the current study, the National Sanitation Foundation Water Quality Index (NSFWQI) is used to compute the water quality index. NSFWQI assign weights to the selected water parameters. The work done by the Brown was supported by National Sanitation Foundation and that is why it become referred as NSFWQI. It is seems to be most powerful method, and was followed by many researcher in water quality. WQI is express in following equation:

$$WQI = \sum SI_i \tag{2.5}$$

where

$$SI_i = w_i \times q_i \tag{2.6}$$

SI_i is the sub-index of the i^{th} water parameter, w_i is the relative weight and is calculated by using this equation:

$$w_i = 1/S_i \tag{2.7}$$

and S_i indicates the standard values of the each parameters.

$$q_i = \frac{C_i}{S_i} \times 100 \tag{2.8}$$

where q_i is indicates the quality rating of i^{th} parameter and C_i indicates the concentration value. Moreover, the disadvantage of this method is that represents the general water quality, it does not represent specific use of the water.

2.3.3 Water quality index

According to Lefebre and Couillard, a WQI is an algorithm that represents a measure of the qualitative state of the water. The final result is obtained in a simple combination of numeric and alphanumeric variables. Water Quality Index is assigns the weights to the parameters, given by,

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{2.9}$$

where, W_i is relative weights and n is number of parameters. Further, quality rating q_i is measured by following expression,

$$q_i = \frac{C_i}{S_i} \times 100 \tag{2.10}$$

Where, C_i is the concentration value of each water parameter, S_i is the standard value of each parameter, which are given in Table 2.2. Finally, WQI is calculated by the following expression,

$$WQI = \sum (W_i q_i) \tag{2.11}$$

2.4 Results and Discussions

In this chapter, three water quality indices have been computed to estimate the water quality of 1495 samples. Range of these WQI are varies from 13.16 to 133.80 which are classified into three classes mentioned in Table 2.2

Table 2.2: Water quality classification based on three WQI methods

WQI levels	Class	WAWQI	NSFWQI	WQI
<50	Excellent	193	1496	1324
$50-100$	Good	1217	0	170
$100-150$	Poor	86	0	2
$150-200$	Very Poor	0	0	0
200>	Unsuitable	0	0	0

The results from Table 2.2 showed that majority of the water categories fall under the excellent and good class. As indicated by the highest values of WQI are 121.81 and 101.15, which represents that bottled water quality is poor, according to Table 2.2. Asal Pia Na brand is classified for highest poor quality of water for both values. According to, Weighted Arithmetic WQI, 56 brands have poor quality and has highest value is 133.80, which represents that Aquwa Plus brand has poor quality among all the brands.

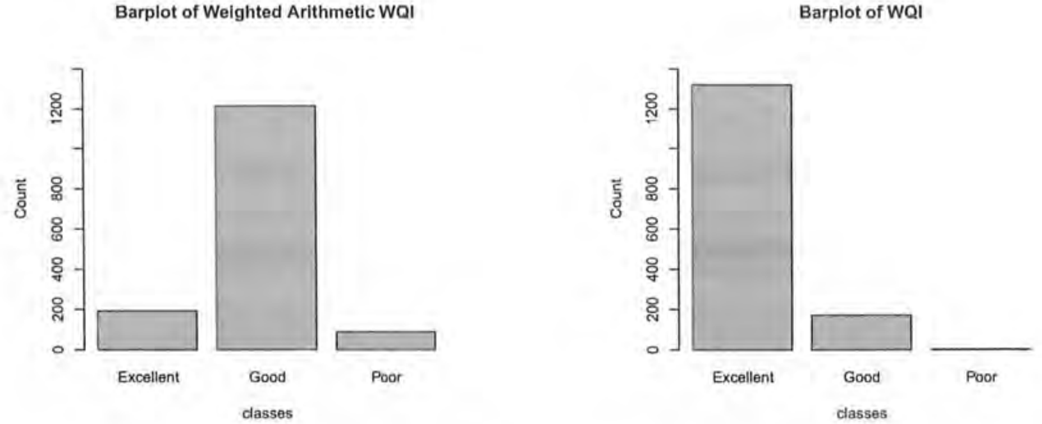


Figure 2.1: Visualisation of classes of WQI through bar graph.

In Figure 2.1 contribution of each class to the samples, in percentage is given. It can be observed that poor class have the lowest percentage that is 0.1% and good class have the second lowest percentage 11.36%, while excellent class of water have the highest percentage i.e 88.50% from WQI and in Weighted Arithmetic QWI, good class have highest percentage other than excellent and poor class respectively. Furthermore National Sanitation Foundation WQI classify that all brands water are excellent. But these indices have the disadvantage that they may not carry enough information about the quality of water, over-emphasizing of a single bad parameter value and many water data cannot met with such indices.

2.5 Conclusion

Acceptable water quality criterion relies upon the prevailing conditions. Therefore, water quality index plays a main role in water quality assessment\evaluation of a given source, as a function of time and other influencing factors. But, it is very difficult to develop a universally acceptable common water quality index. In this chapter we evaluate the quality of drinking water of various brands of Pakistan. Ten physcio chemical parameters such as pH, TDS, Cl, SO4, Ca, Mg, Hardness, EC, HCO3 and Na, are evaluated for every brand and concentration of these values are compared with recommended values of World Health Organization. As it is a difficult task for common man and specialists to understand these parameters one by one so we computed WQI to understand the quality of water. Minimum and maximum values of three WQI for our sampled data are 13.16 to 133.80 respectively. These values are classified into five classes; Excellent, Good, Poor, Very Poor and Unsuitable for drinking. Asal Pia Na and Aquwa Plus brands contain the highest value of indices and classified as Poor and other brands are good for human utilization.

Chapter 3

Comparative Analysis of Decision Tree Data Mining Algorithms for Bottled Water Classification

3.1 Introduction

The fast economic expansion and population growth; along with improper urban planning have caused huge environmental pressures on water. The accretion of high levels of pollutants in water may cause negative effects for humans, such as reproductive disorders, disruption of the immune system, damage to the nervous system and cancer (Huang and Chang (2003)). For this purpose supervised classification methods are applied. Supervised classification methods will be trying to classify data with high accuracy rate for classifying future. Classification method includes Decision Trees, k-nearest neighbor, neural network, logistic regression, support vector machine and so on, but here we used Decision Tree methods. Decision Trees (DTs) are extensively used for classification because of their ability to handle the noisy data and return well organized results that are computationally efficient, easily interpreted and robust. Many techniques, such as ID3 and C4.5, C5.0, CART and Random Forest have been devised for the construction of DTs and are applied in many areas, including risk management, text categorization, medical diagnosis, customer relationship management and credit rating of loan applicants (Abbas (2004)). In traditional DT techniques, a target variable (label) of a tuple is either Boolean or categorical that is, the techniques operate under the assumption that the labels are flat and nominal. However, various practical situations contain more complex classification scenarios, wherein the label to be predicted can occur in a variety of types, for instance, hierarchically related variable, continuous variable, or both.

Kaur et al. (2015) review the decision tree data mining algorithms based on ID3 and its improved version C4.5. They estimated that C4.5 is more benefit over ID3 algorithm in terms of accuracy by selecting attributes with highest gain ratio. C4.5 has also enjoyed continuous attributes which are not used by ID3. It also facilitates us by requiring less memory; by the way one can get results that are robust in terms of over-fitting. Later, they evaluated the performance of C4.5 with its improved version C5.0 by applying customer database classification. The optimality of C5.0 lies in the fact that this algorithm does not require too many investigations to split variables as it directly used maximum information gain comparative to C4.5. C5.0 requires only 200Mb data whereas C4.5 take 3Gb.

Barbieri et al. (2001) used cluster analysis to detect the classes of fresh water of the southern plain of Italy. Henley and Hand (1996) used application of K-NN for consumer scoring problem and modified it by contributing adjusted version of distance method into class separation information of K-NN. They suggested that K-NN performed better than discriminant analysis and logistic regression for classification of credit worthiness of consumer loan application. Random forest method is considering more efficient learning algorithm to perform classification proposed by (Ho (1995)) and later by Random (2001) independently. It works with a large collection of decorrelated decision trees by creating random subsets of samples. Hartfield et al. (2013) mapped crops data by using classification and regression based algorithms. They made comparison between classification and regression tree models with other classification methods and concluded that these methods outperformed the other classification methods.

3.2 Methods

Data mining algorithms used for classification of water quality

Data mining is the extraction of hidden predictive information from large databases. It is a most versatile emerging technology with capacity to help analysts focus on the most important information in their data. Mining tools predict future trends and behaviors to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. By the data mining tools, one can answer those questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Random Forest (RF) is a powerful classification algorithms used to handle large data sets effectively. In sciences data, RF models have efficient prediction accuracy because not all data is useful in prediction. As in case of water data, we have multiple parameters but these all parameters neither all have equal weightage in water significance nor all parameters useful to predict and classify data well (Touw et al. (2012)). Whereas, the remaining algorithms (Nave Bayes, Probabilistic Classifier and K-Nearest Neighbor algorithms) techniques take into account all variables at equal level. Here algorithms are described which are ue for classification of water quality of Bottled water.

3.2.1 Random Forest Model

Random Forest is a non-parametric method, in which a forest of Decision Tree is generated by boosting the training data sets, so instead of using a single classifier, which train by using training data set one time, RF used boosting of training data set and then aggregate (bagging) it. The whole procedure of boosting and bagging is called machine learning ensemble meta- algorithm. The purpose of using meta-algorithm, in comparison to single classifier, is to improve the accuracy and stability of RF by reducing variances so one can use this trained algorithm for classification purpose with high accuracy rate. By meta-algorithm, the problem of over fitting is also avoided by improving prediction. In this technique, we split the data into two or more subsets of full data based on most significant splitter (parameter) in input variables.The advantage of RF is that it used the information of differences in local importance between samples to uncover the most significant variable which split the classes appropriately, structure of RF is defined in Figure 3.1.

Algorithm is listed below:

- For a given dataset, select ntrees bootstrap samples.
- For each n_{trees} samples that generated from above step, grow classification tree, with the following consideration: for every node instead of choosing the most significant splitter, random sample of predictors is drawn i.e. m_{try} called RF parameter and choose the best splitter from among of m_{try}.
- Predict new observation from aggregation of ntrees, class label is assigned having maximum votes.

The main concept of variable importance is an implicit variable selection performed by RF, and it is assessed by the Gini impurity criterion index. The Gini index is a measure of prediction power of variables in classification or regression, based on the principle of impurity reduction (Ceriani and Verme (2012)). Thus, the Gini index can be used to rank the importance of variables for a classification problem.

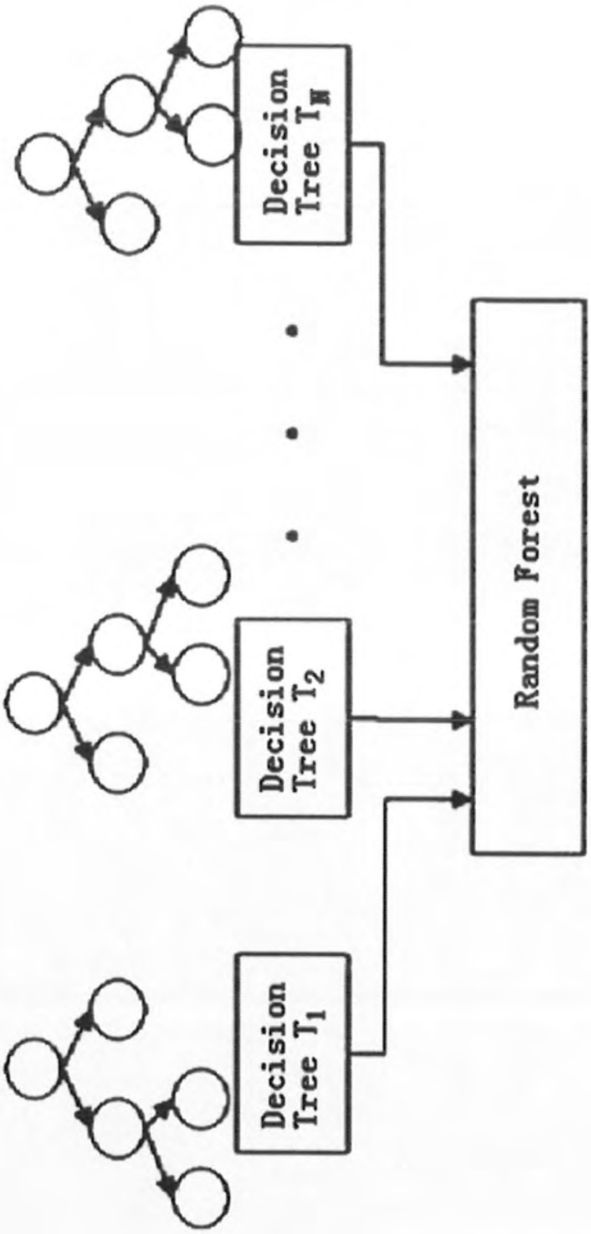


Figure 3.1: General structure of Random Forest model with ntrees trees

3.2.2 C4.5; modified algorithm of ID3

C4.5 is basically the extension of ID3 (Iterative DiChaudomiser), proposed by Ross Quinlan in 1975. To overcome the drawback of ID3, that is only for categorical independent variables not for continuous variables, C4.5 is proposed. The whole algorithm of C4.5 is same as RF, except the splitting criteria of nodes. As the algorithm of C4.5 is based on various types of entropies like Shannon entropy, Havrda entropy, Charvt entropy and quadratic entropy. Entropy measures the degree of randomness in data. Most widely used entropy is Shannon entropy (Kaur et al. (2015)):

$$Entropy(P) = -\sum_{i=1}^{k} p_i \times logp_i \tag{3.1}$$

where p_i is the probability of selecting ith class and k is the total classes, then information gain is calculated by following formula :

$$Gain(p,T) = Entropy(P) - \sum_{j=1}^{k}(p_j \times Entropy(p_j)) \tag{3.2}$$

where Entropy(p_j) is calculated for all parameters. Splitting criteria is based on information gain as parameter with highest gain is selected for splitting criterion attribute. ID3 used information just based on information gain. But C4.5 parameter with the highest gain ratio is selected for splitting criterion attribute because information gain is biased in parameters with more values and the formula for gain ratio is calculated as:

$$\text{Gain Ratioe}(p,T) = \frac{Gain(p,T)}{-\sum_{j=1}^{k}(p_j \times Entropy(p_j))} \tag{3.3}$$

The Gain ratio measure is a parameter selection criterion which is less biased towards selecting attributes with more number of values. The disadvantage of C4.5 are, small variation in dataset can lead to different decision trees especially when the variables are close to each other in values. In small training data set it does not work very well.

3.2.3 C5.0; modified algorithm of C4.5

As C4.5 persuade the principles of ID3, similarly C5.0 persuade the modified version of C4.5. C5.0 is much faster than C4.5. Memory usage is more efficient in C5.0 than C4.5. C5.0 made smaller decision trees in comparison with C4.5 by selecting the most helpful predictors for splitting purposes. The C5.0 have lower error rates on unseen cases and it is more efficient in classifying testing data as C5.0 automatically allows removing unhelpful attributes. C5.0 also split data based on gain information ratio but this time it depends on maximum information gain ratio which can be achieved by setting more than one rule for each node. Splitted nodes are then again split and the process is continued until the subsets cant be split further (Ahmadi et al. (2017)).
A model of C5.0 again splitting the data in training set depend on parameter that get the maximum information ratio. Every subset describe by the first split is then again splited, and the process is recursively repeated until the subsets cannot be splited any more further and measuring the importance of the attribute in the nodes that it has splited and also executed as a surrogate (Kaur et al. (2015)) and define as:

$$\text{Gain Ratio}(p|T) = \frac{Gain(p|T)}{Splitratio(p|T)} \tag{3.4}$$

C5.0 automatically allows to removing the unhelpful attributes.

3.3 Results and Discussions

3.3.1 Results of Random Forest, C4.5 and C5.0

In this research, Random Forest, C4.5 and C5.0 models are trained by using R and Weka softwares to classify the Brands according to their status of water quality. Models are trained by splitting data 80% in training and 20% in testing. As water contains different parameters each have different unit of measurement. The whole algorithms of all three methods depend on two things, one is how to calculate the forest decision tree by using splitting criteria and second its single parameter m_{try}. The value of m_{try} tells the number of parameters used in searching of significant parameter used for classification. The most significant parameters identified by RF, C4.5 and C5.0 are Na, EC and TDS with highest Gini index among other predictors that contributes very effective in classifying brands with high accuracy.

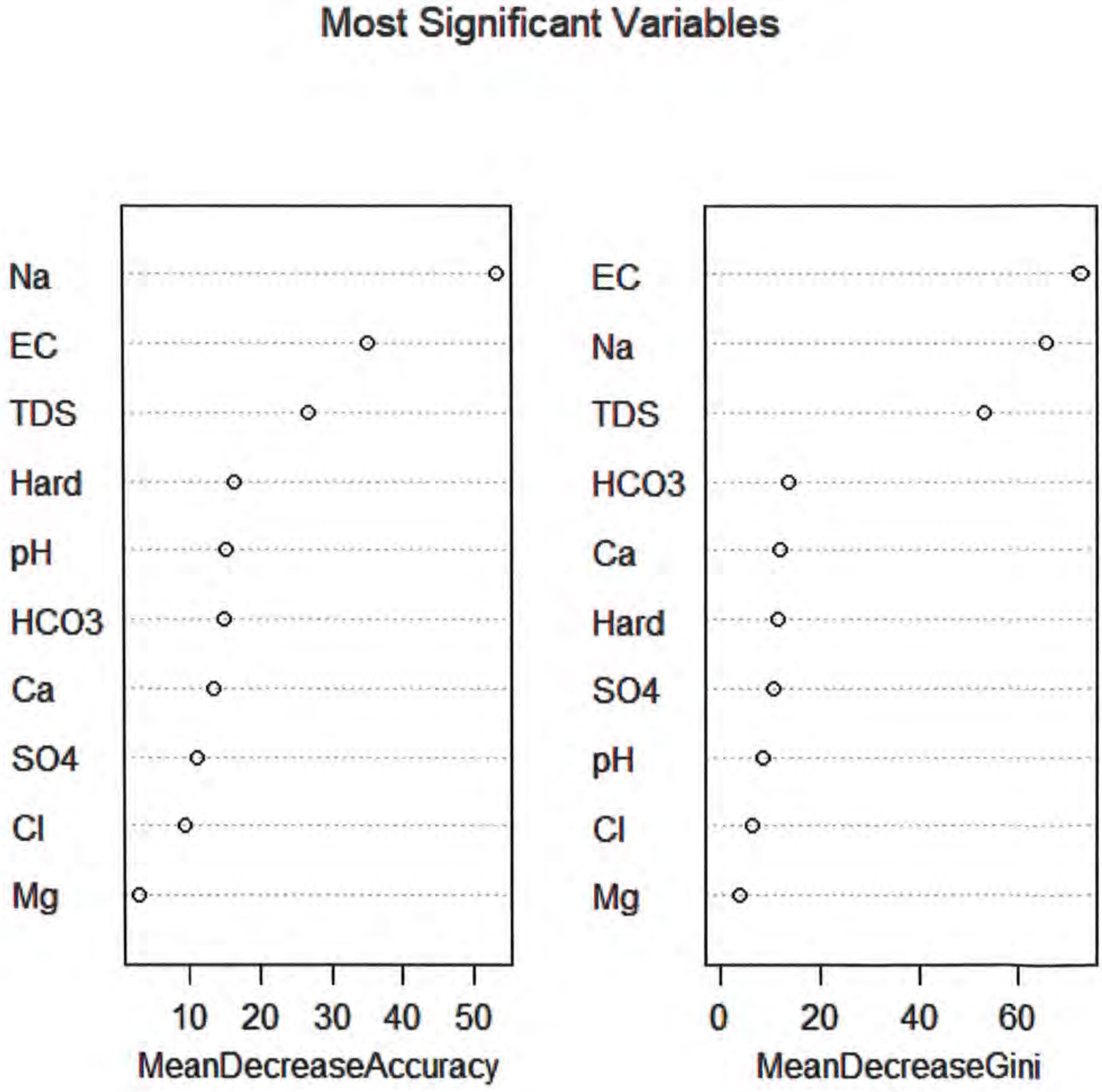


Figure 3.2: Shows the most significant predictors produced by RF, C4.5 and C5.0.

Table 3.1: Experimental Results of Random Forest, C4.5 and C5.0 by splitting 80 and 20 percent for training and testing respectively.

	80% training			20% test		
	Random Forest	C4.5	C5.0	Random Forest	C4.5	C5.0
E.Time	0.3	0.06	0.02	0.034	0.02	0.00
T.Instances	1196	1196	1196	299	299	299
C.Classified	1196	1183	1185	288	281	291
I.Classified	0	13	11	11	18	8
M.A Error	0.010	0.013	0.01	0.03	0.04	0.02
Accuracy	100.0	98.91	99.08	96.32	93.97	97.00

Mean decrease accuracy and mean decrease gini both measures are used for describing variable importance with high mean decrease accuracy and mean decrease Gini, see in Figure 3.2. The experimental results are shown in Table 3.1. All algorithms predicts well in terms of accuracy in 80% training data set. Random forest predicts 100% accurately in 80% training data sets. But C4.5 did not predict more better as compared to RF and C5.0 on training data and effected by small data. C5.0 has highest accuracy of 97% followed by C4.5 with 93.97% in 20% test data. While RF has highest error rate on unseen data set. Therefore accuracy rate among these algorithms are moderate. C5.0 is much faster and require less memory. Overall C5.0 is better than with respect to time and accuracy among all algorithms.

3.4 Conclusion

The present study was conducted to analyze the water quality of Bottled water. Data was collected by Pakistan Council of Research in Water Resources (PCRWR) from 538 easily available commercial Brands of all districts of Pakistan. The study revealed the usefulness of Random Forest, C4.5 and C5.0 which are improved version of decision tree by generating multiple trees using subsets of data to get low error rate. Purpose of all these data mining algorithms is to compare the performance of each algorithm to predict WQI of bottled water. Water data is splitted into 80, 20 percentages for all algorithms. By comparing the performance of all, it is concluded that RF is useful in large data sets whenever it is more efficient in 80% with 100% accuracy. Overall C5.0 is optimal in terms of time and prediction of unseen classes with 97% accuracy in 20% data sets. The findings of water quality might be helpful for decision makers in establishing the guidelines for PCRWR to improve the quality of water in selected Brands.

Chapter 4

Application of Artificial Neural Network and Multiclass-Support Vector Machine

4.1 Introduction

Analysis of water data is a very wide field that includes modeling of water demand, water quality and water reticulation networks. As water affect human health and aquatic ecosystems, so there is a need to check rivers, as main inland water resources used for irrigation, municipal and industrial purposes, which are particularly vulnerable to damage (Su et al. (2011)). The fast population growth and economic expansion, along with improper urban planning have caused huge environmental pressures on rivers. The accretion of high levels of pollutants in water may cause negative effects for wildlife and humans, such as reproductive disorder, damage to the nervous system, cancer and disruption of the immune system (Liu and Xia (2004)). Moreover contaminated water can also lead to some waterborne diseases and also influences child mortality. Due to water pollution, about 25 million people die each year and it become a main problem in several countries around the world (Pimpunchat et al. (2009)). Buck et al. (2004) stated that, the quality of water is always influenced by many factors, including climate change, atmospheric chemistry, the underlying geology and anthropogenic activities. Human activities such as the release of domestics and industrial effluents, the use of agricultural chemicals, cover changes and land use are the main factors that influence water quality. However, source apportionment analyses can be used as a noticeably fast, accurate and cost effective technique for identifying and targeting pollution sources and their relative contributions to the entire pollution load. Water pollution has been a crucial concern in several developing countries and Pakistan is also among them. In Pakistan, only 30% of urban and 23.5% of rural population have access to drinking water, while each year 200,000 children die due to diarrheal disease (Rosemann (2005)). Therefore in order to protect water quality, the government has performed a series of governance programs, one of which is the vast monitoring and evaluation of basic water quality.

4.2 Literature Review

Water quality assessment can synthetically quantify the water quality state by using an appropriate evaluation method and is gradually an important tool for management of water resources and scientific utilization (Nazeer et al. (2014)). However, in recent decades more nonlinear classification solvers were developed for pattern recognition, such as artificial neural network, fuzzy mathematics, grey relation analysis, and support vector machines (SVM). Moreover, k-nearest neighbor (KNN), probabilistic neural network (PNN) and artificial neural networks (ANN) methods were also used to classify water quality (Modaresi and Araghinejad (2014b); Taormina et al. (2012)). In a recent analysis of forecasting approaches for the building sector, Support Vector Machines (SVM) and artificial neural networks models (ANN) are the most common tools, to develop energy prediction approaches, which in turn support physical improvement strategies (Çelik et al. (2016)) . Particularly, SVMs have been used for time series predictions, mostly in financial time series and electrical load forecasting (Sapankevych and Sankar (2009)). As for appropriate choice of optimization model, Artificial Neural Networks (ANNs) which act on proposal received from biological neural networks are located amongst prediction tools and yield suitable results in many fields such as water quality, flood prediction and land use (Alizadeh and Kavianpour (2015)). These networks perform like the human brain system. The process received data and depend on various elements called neurons that have severely high continuity of neuronal output. Like human beings, instruction or information takes place in these networks through identifying the many examples and by making adjustments in the synaptic connections between nerves. However, in AAN these adjustments are formed by changing the weights of every neuron present in the network (Ma et al. (2014); Vakili et al. (2017)). ANN is use in groundwater studies to determine the aquifer parameters (Aziz et al. (1992)). The most important application of ANNs has been in studies conducted on quantitative and qualitative changes in water resources to help in decision making and management of water resources (Aish et al. (2015); Banerjee et al. (2011)). These techniques have different characteristics and contribute to studying the water quality recognition and classification; yet, there are still some debates on the best methods for the assessment of water quality. Among them, SVM, a statistical learning theory, has been used recently as a popular method for pattern recognition and problem classification and obtained a series of satisfactory results in classification problems. Many researches have showed that SVM performance was comparable with or even superior to artificial neural network. Support Vector Machine (SVM) is based on statistical learning theory and has the main purpose of determining the location of decision boundaries that make the optimal separation of classes. In the case of a two-class pattern recognition problem in which the classes are linearly separable, the SVM chooses the one among the infinite number of linear decision boundaries that reduce\minimizes the generalization error. This trouble of maximizing the margin can be solved by using the standard quadratic programming (QP) optimization methods. The training data points that are nearest to the hyperplane are used to measure the margin; hence these training data points are termed support vectors. Therefore, the number of support vectors is small (Vapnik (2013)). Baldi and Pollastri (2002); Ganapathiraju et al. (2004) discussed if the two classes are not linearly separable, then SVM tries to seek the hyperplane that maximizes the margin while, at the same time, minimizing a quantity proportional to the number of misclassification errors. The trade-off between misclassification error and margin is controlled by a user-defined constant (Cort. Support vector machine can also be extended to address nonlinear decision surfaces. As, real-world problems often require the distinction for more than two classes. Therefore, the multi-class pattern recognition has a extensive range of applications including, intrusion detection optical character recognition, bioinformatics and speech recognition. In exercise, the multi-class classification problems (k>2) are usually decomposed into a series of binary problems such that the standard SVM can be directly carried out.

4.3 Methodology

4.3.1 Artificial Neural Network (ANN)

ANNs are most powerful tool for pattern recognition and classification due to their nonlinear nonparametric adaptive-learning properties. Artificial neural networks are mostly used multi-layer networks which are trained using optimization algorithms or back propagation (Khodadadi et al. (2016)). Neural network observed the values in the past that are the inputs and accumulated these inputs in a neuron considering the weight of every variable and a constant value or bias. These inputs affect the objective function and generate the output, this is prediction of future measurements (Çelik et al. (2016)). Training an ANN contains the weight matrix that minimizes the prediction error for a set of training observation for which there is knowledge of what the output vector ought to be (JIAWEI et al. (2007)). Every connection between the input and output nodes carries a connection weight (w_{j_0}), which identifies the strength of that connection. These weights may be zero, negative, positive; zero weights represent the connections that do not exist in the network, negative weights represent inhibitory signals and positive weights represent excitory signals in the network. Artificial neural networks do not give an accurate physical model but learn to show the relationship in terms of the squashing or activation functions of the neurons (Rosemann (2005)). Activation function is also known as sigmoid function which is defined as a strictly increasing function that exhibits smoothness and asymptotic properties. The general graph of network is shown in Figure 4.1

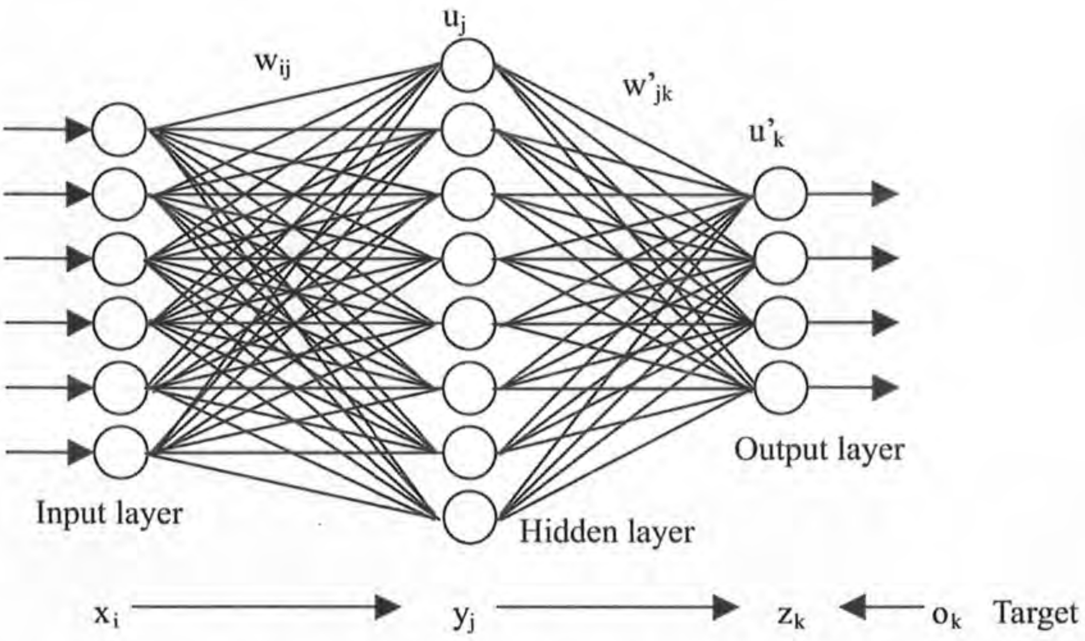


Figure 4.1: Neural network structure

Mathematically, the algorithm of backpropagation is to derive an updated form for the connection from the m^{th} input node to the j^{th} hidden node. At the i^{th} iteration, let

$$u_{i,j} = w_{i,j_0} + x_i^{\tau} w_{i,j} \tag{4.1}$$

$$u_{i,j} = \sum_{m \in M} w_{i,jm} x_{i,m} \tag{4.2}$$

be the sum of weighted inputs to the j^{th} hidden node, where

$$x_i = (x_{i,1}, ..., x_{i,r})^{\tau}, \quad w_{i,j} = (w_{i,j1}, ..., w_{i,jr})^{\tau} \qquad j \in J \tag{4.3}$$

The resultant output is:

$$z_{i,j} = f_j(u_{i,j}) \tag{4.4}$$

Where $f_j(.)$ is the activation function and logistic activation function is written as;

$$f_j(u_{i,j}) = \frac{1}{1 + exp^{(-u_{i,j})}} \tag{4.5}$$

4.3.2 Support Vector Classification (SVC) Model

Support vector machine are more powerful and usefuthe optimal classification hyperplanel method for data classification. A SVM classification model is outlined below, according to earlier studies (Chih-Chung (2011);Canu (2005) and Guardiola et al. (2014)). The training data set is denoted as,

$$(x_i, y_i)_{i=1,2,...,n} \tag{4.6}$$

where x_i is the input vector, y_i the class label, and n is the total number of training data. Then, an SVM model in high dimension feature space can be expressed as follows:

$$f(x) = sgn(\langle w, \varphi(x)\rangle + b) \tag{4.7}$$

Where w is a weight vector, $\varphi(x)$ is a nonlinear that mapped the input variable into high dimension feature space, b is a bias and $\langle\rangle$ denotes the dot product (or inner product). Then get from the following initial formulation:

$$min_{w,b,\xi}\frac{1}{2}\|w\|^2 + C\sum_{i\in 1}^{n}\xi_i \tag{4.8}$$

subject to

$$y_i(\langle w, \varphi(x)\rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, ..., n \tag{4.9}$$

where $\frac{1}{2}\|w\|^2$ controls the complexity of the model and ε_i is a slack variable measuring the error on inputs. $C > 0$ is a the penalty parameter of the error term (or regularization parameter), which determines the tradeoff between the complexity of the model and the empirical error. A Lagrangian function corresponding to equation (1.3) can be define as follow by introducing Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$.

$$l(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\xi_i - \sum_{i \in 1}^{n} \xi_i \alpha_i [y_i(\langle w, \varphi(x)\rangle + b) - 1 + \xi_i] - \sum_{i \in 1}^{n} \beta_i \xi_i \tag{4.10}$$

Hence, the following equation can state the equation (1.5)

$$f(x) = sgn(\sum_{i \in 1} \alpha_i y_i k(x_i, x) + b) \tag{4.11}$$

where $k(x_i, x) = \langle \varphi(x_i), \varphi(x)\rangle$ There are four fundamental kernels suggested by research Hsu et al. (2003).

Table 4.1: Kernel Functions

Kernel	K(x,y)
Linear	$x_i^T x_j$
Quadratic	$(\gamma x_i^T x_j + r)^2$
Polynomial	$(\gamma x_i^T x_j + r)^d$
Radial	$exp(-\gamma \|x_i - x\|^2)$
Sigmoid	$tanh(\gamma x_i^T x_j + r)$

Where $\gamma \geq 0$, r and d are kernel parameters.

Multi-class SVM

One Against One (OAO) and One Against All (OAA) are two popular algorithms to solve the multi-classification tasks (Knerr et al. (1990)). The One Against All SVM required consensus among all SVMs: a information factor or data point could be classified under a certain class if and only if that classs SVM confirmed or accepted it and SVMs rejected it all other classes, whereas an One Against One SVM technique trains a SVM for any two classes of data and attains a decision function. There are $(k(k-1)/2)$ decision functions for a k-class problem. For training data set from the i^{th} and the j^6th classes, the two class classification problem will be solved Chih-Chung (2011):

$$min_{w,b,s}\frac{1}{2}\|w\|^2+C(\sum_t(\xi)_t) \tag{4.12}$$

Subject to;

$$(\langle w, \varphi(x)\rangle + b) \geq 1 - \xi_t, \text{if} x_t \text{in the} i^{th} \text{class} \tag{4.13}$$

$$(\langle w, \varphi(x)\rangle + b) < 1 - \xi_t, \text{if} x_t \text{in the} j^{th} \text{class} \tag{4.14}$$

$$\xi \geq 0. \tag{4.15}$$

Now we use a voting strategy in classification technique, every binary classification is considered a voting where the testing point is certain to be in a class with the highest number of votes.

Model and parameters selection

There are five suggested steps for multiclass- SVM procedure that are;

(1) Split the data into training and testing set.

(2) Consider the different kernel functions $k(x_i, x)$, which are linear, polynomial, sigmoid etc.

(3) Use cross validation method to avoid the over fitting problem and Grid-search to find the best parameter C and γ.

(4) Choose the best parameter C and γ to train the whole training data set.

(5) Check the performance on test data set.

At step (3), among the linear, quadraic, polynomial, sigmoid and the radial kernel functions, which kernel function gave the best result. All kernel functions are nonlinearly separable except linear kernel (Hsu et al. (2003)). The main objective is to identify the best (C,γ), so that the classifier can accurately predict the test data (i.e unknown data). The prediction accuracy gained from the test data set more precisely shows the performance on classifying the training data set. After solving the above steps, one can apply the decision function to predict class labels (target values, here the class of water quality is Excellent, Good, Poor) of testing data, and we calculate the predictions by the accuracy.

4.4 Results and discussion

4.4.1 Results and discussion of ANN

Selection of the appropriate number of hidden layers in the neural network is very important issue. According to previous study, neural network with more than three or four hidden layers rarely show optimal performance in prediction problems (Aish et al. (2015)). However, selection of too many hidden layers may cause the over-fitting problem. Since neural network split the entered data into two parts, 80% is considered as the training set remaining 20% for the test set. According to rule thumb method, the number of hidden neurons should be in the range between the size of the input layer and the size of the output layer. There are two hidden layers where 10 neurons in the first hidden layer and five in the second hidden layer, see in Figure 4.2. These layers gave the large weights, bias and effect the results and represent the "under-fitting" problem, did not gave a precise results.

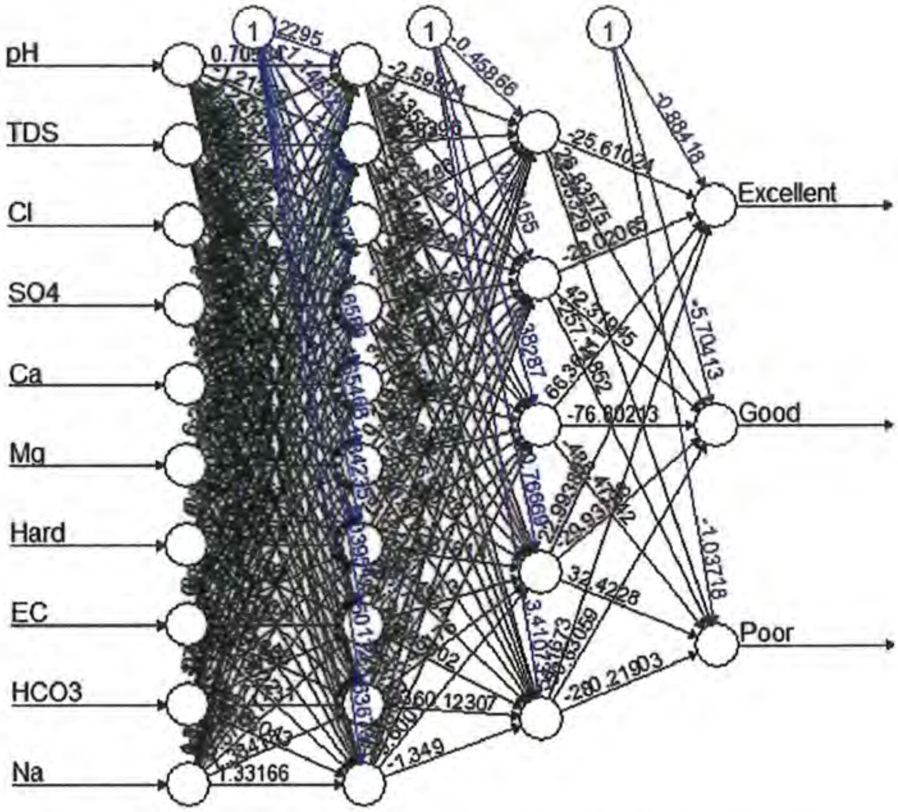


Figure 4.2: Feed Forward Neural network two hidden layers

In study, run different size of hidden layers to get the optimal numbers of hidden layers, neurons and possible scenarios are studied to get the desirable result. According to Aish et al. (2015), three or four hidden layers are optimal for neural network, and more than these hidden layers cause the over-fitting problem. Select the three hidden layers where 10 neurons in the first hidden layer, five neurons in the second layer and three neurons in the third hidden layer. As shown in Figure 4.3:

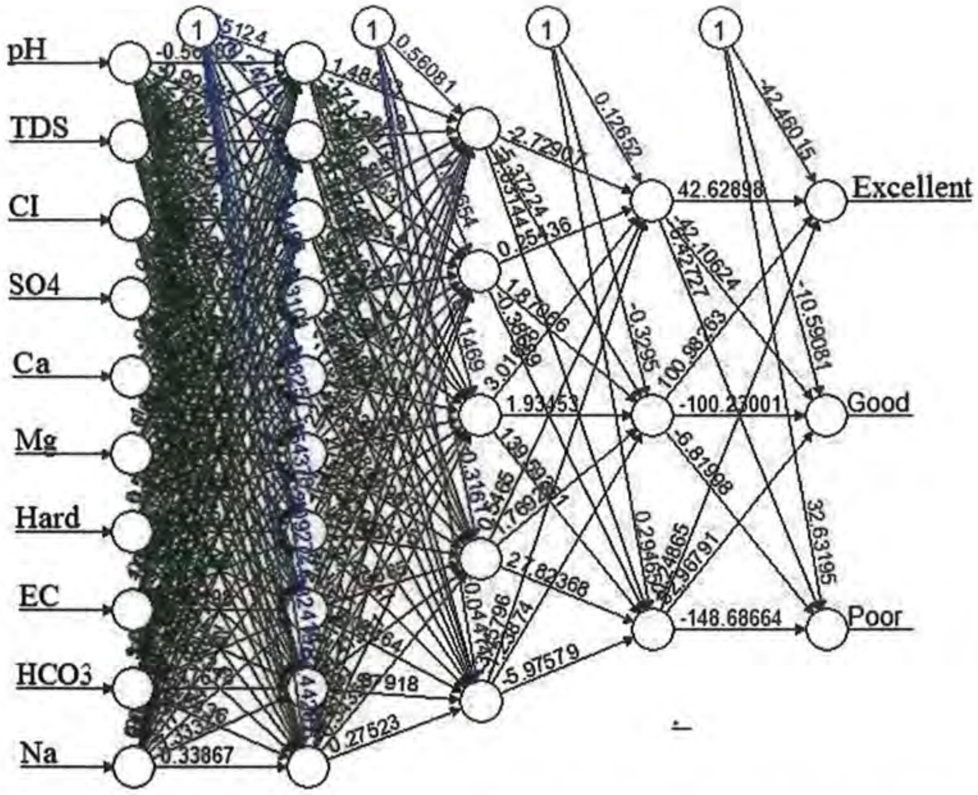


Figure 4.3: Feed Forward Neural network with three hidden layers

The goal is to explain the parameters that influence the bottled water data set and check the accuracy rate of ANN. Furthermore, all the variables of water quality used in this study and multiplied by an optimized coefficient based on the trained model set that contribute to each neuron through a specific constant value also known as bias listed in Table 4.2 along with all coefficients.

Table 4.2: The coefficient matrix for each variables and bias values for each neuron

layer1	pH	TDS	Cl	SO4	Ca	Mg	Hard	EC	HCO3	Na	Bias
1	0.56	0.94	0.70	0.52	-0.54	-1.01	0.58	-0.07	1.08	0.04	-2.65
2	-0.99	2.14	-0.98	1.34	-0.04	0.00	0.25	1.68	1.05	-0.25	-10.2
3	2.21	3.81	2.67	3.59	4.66	1.29	3.10	1.26	3.66	2.07	1.22
4	1.50	0.98	95.2	-9.89	-1.35	4.48	-14.22	-1.64	-20.78	38.45	-0.04
5	1.43	0.03	0.05	-1.09	-1.26	-4.47	1.99	-0.82	-0.85	-1.09	4.43
6	-2.04	-0.93	-1.41	-14.5	2.83	-1.76	-0.48	-1.20	-8.12	-4.97	-4.08
7	0.89	-1.09	0.95	0.65	-2.02	5.67	0.47	-1.05	-0.35	0.70	1.54
8	-2.70	0.14	-0.50	0.45	0.78	0.70	0.92	0.48	0.78	3.17	-2.09
9	2.61	3.57	0.36	1.37	1.52	-1.15	0.85	0.69	1.17	3.33	-.82
10	0.29	2.81	-0.79	-1.83	8.70	-13.3	-0.74	-0.35	-2.84	0.33	-1.44
	$N1_1$	$N1_2$	$N1_3$	$N1_4$	$N1_5$	$N1_6$	$N1_7$	$N1_8$	$N1_9$	$N1_10$	Bias
1	1.48	2.06	0.13	-1.12	-2.51	22.41	-1.52	2.01	0.37	1.16	0.56
2	-171.2	-4.99	1.78	1.57	-5.16	-163.6	103.1	145.1	2.71	-0.87	1.64
3	2.31	-1.77	0.04	1.01	1.02	-1.43	2.09	-2.37	0.21	-0.33	0.11
4	-2.76	0.34	1.93	0.32	32.33	-53.59	-2.10	-1.19	0.41	-0.54	-0.31
5	41.91	2.16	0.40	0.51	-2.63	27.53	-2.47	2.04	-1.87	0.27	-0.04
	$N2_1$	$N2_2$	$N2_3$	$N2_4$	$N2_5$	Bias					
1	-2.72	0.25	3.01	0.54	-3.45	0.12					
2	-5.37	1.87	1.93	1.76	-1.23	-0.32					
3	-1.83	-0.38	139.5	27.82	-5.97	0.29					
	$N3_1$	$N3_2$	$N3_3$	Bias							
	42.62	100.9	-0.24	-42.46							
	-42.10	-100.2	52.96	-10.59							
	6.42	-6.81	-148.6	32.63							

The final neural network model of this study was attained by replacing Table 4.2 coefficients in function of the model (equation 1.2). The negative values in Table 4.2, represents inhibitory which means less significant, positive values represent excitatory which means significant values and zero represent that do not exist in network. However, as more hidden layers are added, then neural network becomes more complex which may cause the network to learn noises in addition to the underlying patterns or rules. Confusion matrix Table 4.3 shows the test result for each class (Excellent, Good, Poor) of Bottled water data.

Table 4.3: The confusion matrix of ANN on test bottled water dataset

	Predicted		
Observed	Excellent	Good	Poor
Excellent	264	1	0
Good	8	23	0
Poor	0	1	2

Table 4.3 represents the performance of ANN on test data set. The ANN model achieving the 96.65% of accuracy rate on test data set with minimum error rate. Hence, based on the trained neural network, it is analyzed that chlorine, calcium, sodium, total dissolved solids and magnesium ions has influence the Bottled water quality and companies put more concentration on these ions to achieve the standard values or limits that are fixed by PCRWAR.

4.4.2 Results and discussion of multi-class SVM

The bottled water dataset is used in this study and split the entire data into two data sets. These are 80% training set and 20% unseen set. The goal is to identify, which kernel function is the best for SVM that can accurately predict the unknown data and best kernel to choose the optimal value of C and γ.

Note that it could not be useful to attain the highest training accuracy (i.e. a classifier which accurately predicts training data set whose class labels are indeed known). Usual strategy is to split the entire data set into two parts, of which one is considered unknown. Therefore, an improved version of this technique is known as cross-validation. It suggested a grid-search on C and γ using 10-fold cross-validation. Several pairs of (C,γ) values are tried and then the picked the one which has the best 10-fold cross-validation accuracy. We found that trying tune function of C and γ is a practical method to identify good parameters (for example, C = 0.01, 0.1, 1, 10, 100, 500, 1000, $\gamma = 0.1,...,2$). Moreover, its save the computational cost, time (which avoid doing an exhaustive parameter search by approximations or heuristics) and tune function of C is shown in Figure 4.5 (right side). In addition, Figure 4.4 presents the different kernel function using multi-class SVM model on unknown data set. Different kernel parameters (C,γ) would cause different accuracy rate. It shows that which kernel has well separate the water quality class (Excellent, Good, Poor).

Table 4.4: Experimental results of multi-class SVM by using Kernel functions

	Different kernel functions				
	Linear	Quadratic	Polynomial	Radial	Sigmiod
$Cost$	1	4	4	1	4
$Gamma$	0	0.5	0.5	2	0.5
$SupportVector$	116	102	99	416	277
$CorrectClassified$	289	290	291	283	248
$IncorrectClassified$	10	9	8	16	51
$AccuracyRate$	0.9665	0.9698	0.9732	0.9464	0.8294
$ErrorRate$	0.0334	0.0301	0.0267	0.0535	0.1705

Table 4.4 presents the experimental result using the multi-class SVMs on different kernel functions for the bottled water quality data in Pakistan. In this study, basic aim to find the best kernel function which shows the better performance for classification of bottled water quality. Study has achieved the best classification performance by applying the 97% on Polynomial and 96% on (Linear, Quadratic) kernel functions on test data set. The multiclass-SVM model with sigmoid kernel function shows the poorest performance on test data set and has highest error rate.

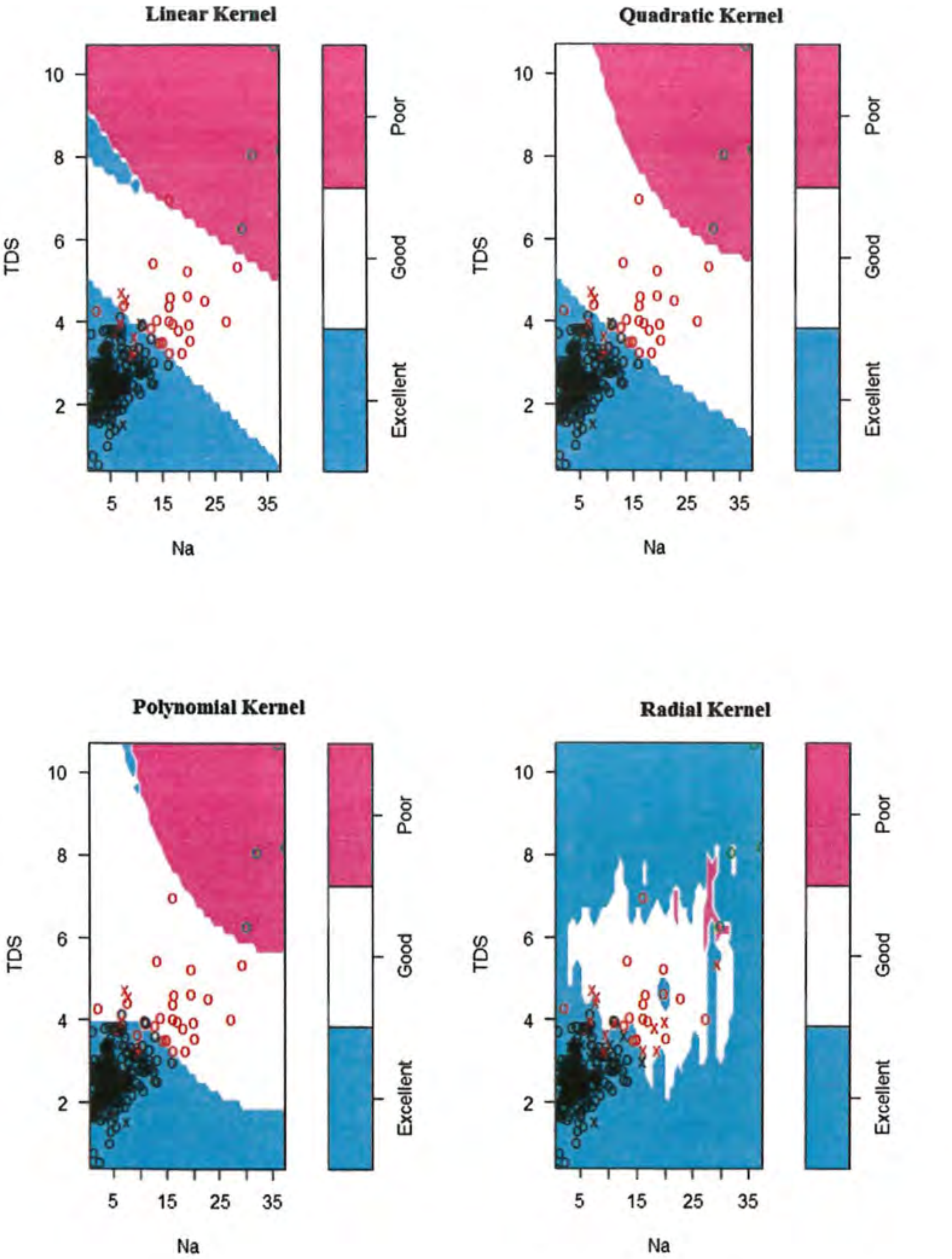
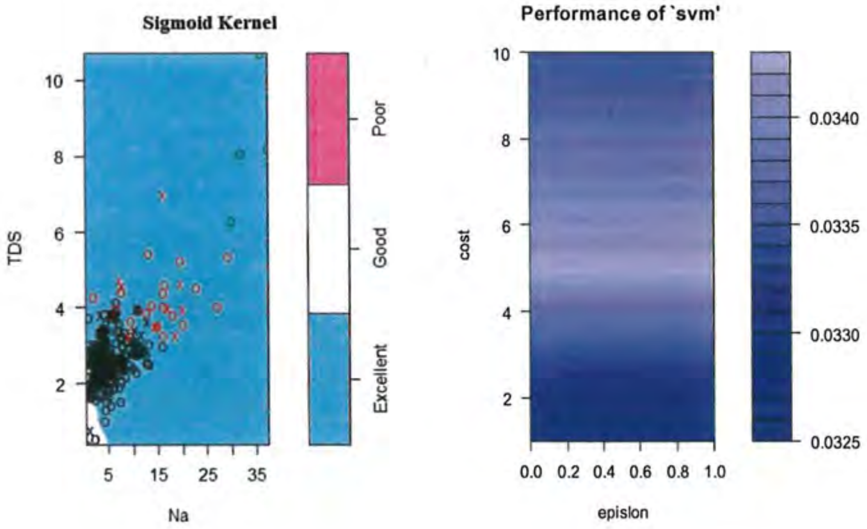


Figure 4.4: The SVM classification plot of test SVM fit based on four kernel functions on bottled water data

Figure 4.5: SVM plot visualizing the test data. Support vectors are shown as X, true classes are highlighted through symbol color, predicted class regions are visualized using colored background and Optimal value of C



4.5 Conclusion

In this chapter, we studied the classification of Bottled water quality in Pakistan using the two machine learning techniques. These are artificial neural network (ANNs) and multi-class support vector machine (SVM). An approach is chosen to perform two parallel experiments one for ANNs and other for multi-class SVM. The experimental results of ANN using back propagation is compared to different kernel functions of multi-class SVM. The multi-class SVM is comprised with kernel functions and some of these kernel functions has different additional argument such as coefficient and degree. These kernels are compared to each other in order to determine the appropriate kernel function. The overall performance criteria used to determine the Bottled water quality in Pakistan from each experiment, is the accuracy rate of the target class of the test set. The experimental results reveal that polynomial kernel has the better classification performance in term of four other kernel function of multi-class SVM and also compared with ANN classifier in detection of Bottled water quality. The accuracy rate of polynomial kernel 97.32% which is higher than 96.65%, 96.98%, 96.64%, 82.94% and 96.65% obtained through multi-class SVM kernel functions and ANN using back propagation respectively.

The results of this chapter demonstrate that multi-class SVM technique has better performance than the ANN back propagation technique. Multi-class SVM using polynomial kernel is more effective and applicable in Bottled water quality in Pakistan.

Chapter 5

Principal Component Analysis and t-SNE

5.1 Introduction

The regular monitoring and quality of drinking water supplied to public are important in terms of health. Public water distribution systems are generally analyzed and monitored by way of taking samples regularly from various sampling sites. Assessment of quality of the drinking water is primarily based on the collected and monitored chemical, physical and biological data from drinking water brands. But, settlement of the drinking water treatment plants and water infrastructures, and the maintenance or protection of their operations are costly and have need of significant investment. Water resource management is another significant factor, especially in the areas where the quality of the drinking water is very poor or the improvement of the quality of the drinking water is crucial (Aydin et al. (2015)). In evaluating the quality of drinking water, characterization of several variables representing the water composition is an elementary force leading to the evaluation of many data, mostly not normally distributed and collinear, including errors and outliers. The data finds a hidden pattern and multidimensional space in which the composition and the resources of the drinking water should be brought into light. Multivariate techniques such as principal component analysis (PCA), discriminant analysis and hierarchical cluster analysis (HCA) can be used for this purpose (Brereton (2003); Lavrač and Zupan (2009)). Brereton (2009) stated that, PCA is the main component of the explanatory data analysis technique and considered as the most extensive and robust method for the classification and visualization of data. Results of PCA are discussed in terms of the loadings and scores matrices. While loading matrices can be used for the correlation of the main components (variables) of the constituents of the data and score matrices can be used for the classification of data. Another study discuss that both PCA and HCA can be used together for the assessment of surface and groundwater quality (Omo-Irabor et al. (2008); Parizi and Samani (2013)). Evaluation of the quality of surface water and groundwater is well discussed in the literature by applying multivariate techniques. Although the use of PCA is mostly satisfactory, but any approach cannot be superior. Most extensively, PCA is extremely affected by the presence of outliers and cannot capture local data characteristics well. Then, t-SNE is a famous choice in the analysis of single-cell RNA-seq data, but it has not been applied extensively to water quality data (Poirion et al. (2016);Wagner et al. (2016)). In the single preceding publication wherein t-SNE was applied to genetic data, the main conclusion is, if t-SNE is considered as a clustering approach, it performs better than PCA (Platzer (2013)).

5.2 Methodolgy

5.2.1 Principal Component Analysis

PCA is a multivariate statistical technique which permits the representation of the original dataset in a new reference system characterized with the help of new orthogonal variables called Principal Components (PCs). Let $x^T = (x_1, x_2, \ldots, x_p)$ be a random vector with mean and covariance matrix (μ, Σ) respectively. It is a technique for dimensionality reduction from p dimension to k, where k $<$ p dimensions. It tries to find the most informative k linear combination of a set of variables $y_1, y_2, \ldots, y_k$. Information is interpreted as a percentage of the total variation in Σ. PC's may be express in term of sample (using S) or population (using Σ). Let

$$\begin{aligned} y_1 &= a_1^T x \\ y_2 &= a_2^T x \\ &\vdots \\ y_p &= a_p^T x \end{aligned}$$

where, $y_j = a_{1j}x_1 + a_{2j}x_{+\ldots} + a_{pj}x_p$ are the linear combination of the x^T that is $(x_1, x_{2\ldots}, x_p)$ such that $a_j^T a_j = 1$, $a_j^T a_k = 0$ for j $\neq$ k and $a_1, a_2, \ldots, a_p$ are the weights that show how much each original variable (j) contributes to the linear combination forming this component (k). The weights of the original variables on each PC, are also called loadings. The graphical depiction of the scores allows the identification of groups of samples displaying similar behaviors (samples near one to the opposite in the graph) or different characteristics (samples far from each other). By viewing at the corresponding loading plot, it is feasible to identify the variables which can be responsible for the analogies or the differences detected for the data in the score plot. From this point of view, PCA allows the representation of multivariate data sets by means of only a few PCs, identified as the most significant.

5.2.2 t-Distributed Stochastic Neighbor Embedding

PCA is a linear technique, it will now not be able to interpret complex polynomial relationship among features. On the other hand, t-Distriuted stochastic neighbor embedding (t-SNE) is primarily based on probability distributions with random walk on neighborhood graphs to get the structure within the data. t-SNE minimizes the kullback-Leiler divergence among two distributions, a distribution that measures pairwise similarities of the input objects in high dimensional space and another heavy-tailed

Students t-distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding space. t-SNE has confirmed its built-in advantages in capturing local data characteristics and revealing precise data structures in visualization. We are mostly interested in t SNEs claimed capability to reveal structure at various different scales, as main population stratification co-exists with other small-scaled shared evolutionary history among objects (Maaten and Hinton (2008)).

Model of t-SNE:

t-SNE describe the joint probabilities p_{ij} that calculate the pairwise similarity between objects x_i and x_j by symmetrizing two conditional probabilities as given below:

$$p_{i|j} = \frac{exp(-d(x_i, x_j)^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-d(x_i, x_k)^2/2\sigma_i^2)}, p_{i|i} = 0 \tag{5.1}$$

$$p_{ij} = \frac{p_{j|i}}{p_{i|j}} \tag{5.2}$$

Above equation, the bandwidth of the Gaussian kernels, σ_i is set in such a manner that the perplexity of the conditional distribution p_i equals a predefined perplexity u. Then, measured the similarities between two points y_i and y_j (i.e., the low-dimensional models of x_i and x_j) by using a normalized heavy-tailed kernel. Particularly, the embedding similarity q_{ij} between the two points y_i and y_j is calculated as a normalized Student-t kernel with a single degree of freedom:

$$q_{ij} = \frac{(1+ \parallel y_i - y_j \parallel^2)^{-1}}{\sum_{k \neq l}(1+ \parallel y_i - y_j \parallel^2)^{-1}} \tag{5.3}$$

Then, the locations of the embedding points y_i are considered by minimizing the Kullback-Leibler divergence between the joint distributions P and Q:

$$C(\varepsilon) = KL(P \| Q) = \sum_{i \neq j} log \frac{p_{ij}}{q_{ij}} \tag{5.4}$$

where ε is output and equal to $(y_1, y_2, ..y_N)$. Because of the asymmetry of the Kullback-Leibler divergence, the objective function focuses on modeling high values of p_{ij} (similar objects) by high values of q_{ij} (nearby points in the embedding space). The objective function is non-convex in the embedding E. It is typically minimized by descending along the gradient.

5.3 Results and discussions

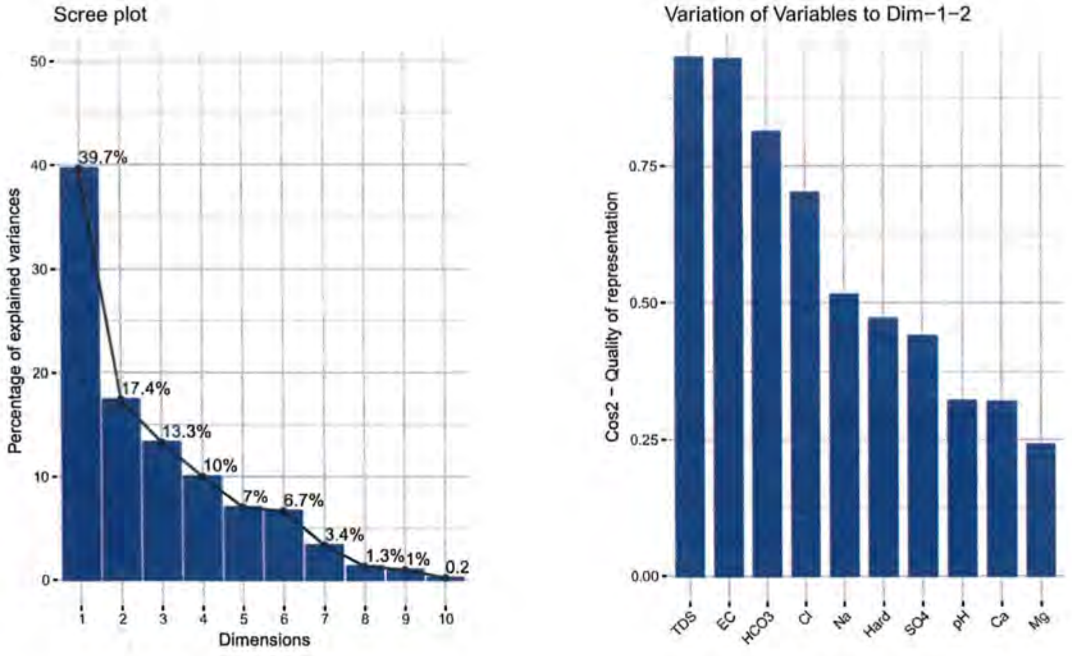
5.3.1 Results of PCA

PCA for the parameters is calculated by specifying the 2 dimensions. Analysis is completed by using R software. Table.5.1 represents the loading of each parameter for two dimension. The well worth point of the table noted here is that, it shows a small value loading, which demonstrates, these parameters are not contributing significantly to the PC. pH, Magnesium (Mg), Calcium (Ca) and sulfate (SO4), all these parameters are not extremely affect the quality of bottled water.

Table 5.1: Discrimination Measures of the two Dimension in PCA

Parameters	Dim.1	Dim.2
pH	0.316	-0.470
TDS	0.954	0.197
Cl	0.100	0.831
$SO4$	0.584	0.313
Ca	0.501	-0.262
Mg	0.488	-0.050
$Hard$	0.646	-0.231
EC	0.953	0.194
$HCO3$	0.637	-0.638
Na	0.624	0.355

Figure 5.1 (a) displayed the percentage of explained variances for each dimension. This shows that first PC explains 39.7% variance, second PC explain 17.4% variance, third PC explain 13.3% variance, fourth PC explain 10% and so on. So, four PC explain around 80.4% variance in the data set.



(a)

(b)

Figure 5.1: plots of Scree and Variation of variables

From the right side of Figure 5.1 shows that, by using PCA we have reduced 10 parameters to five parameters without compromising on explained variance. Total Dissolved Solids (TDS), Electrical conductivity (EC), Hydrogen carbonate (HCO3),Chlorine (Cl) and Sodium (Na) are most significant parameters as compared to other parameters in the bottled water data. If these parameter values are increased by the standard limits, then its extremely affect the human health.

5.3.2 Results of t-SNE

We first examine the ability of t-SNE to separate Bottled water quality (Excellent, Good, Poor) and Figure 5.2, Figure 5.3 shows the results from PCA and t-SNE respectively, with the first and second major dimensions or generate 2D plots that show obvious clusters of water quality class. t-SNE generated low-dimensional clusters represent the classes of water quality. The number of observation in each class 1314 Excellent, 162 Good and 19 Poor; and these classes are represented by 1, 2 and 3 respectively. Both methods are able to separate water quality, PCA (1-2 dimensions) shows an overlap among Excellent, Good and Poor class (though poor is split into excellent and good class), whereas t-SNE shows Poor has minor overlap with Good class. Hence, PCA is not separating the classes well, but t-SNE is separating the classes well.

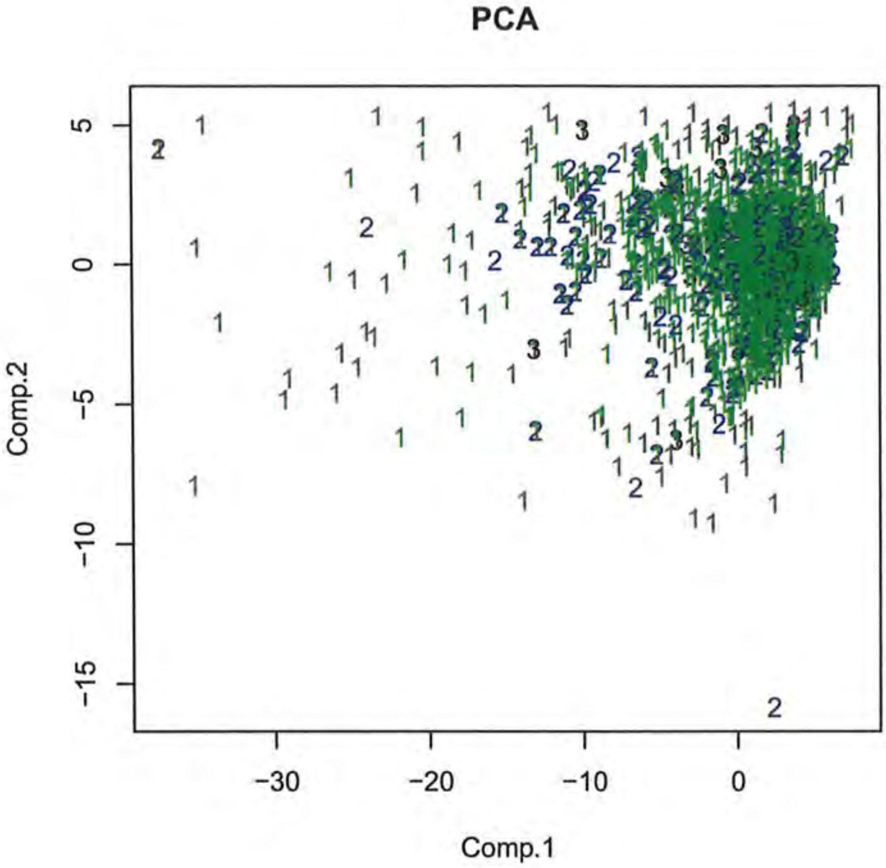


Figure 5.2: visualization of two dimension by PCA

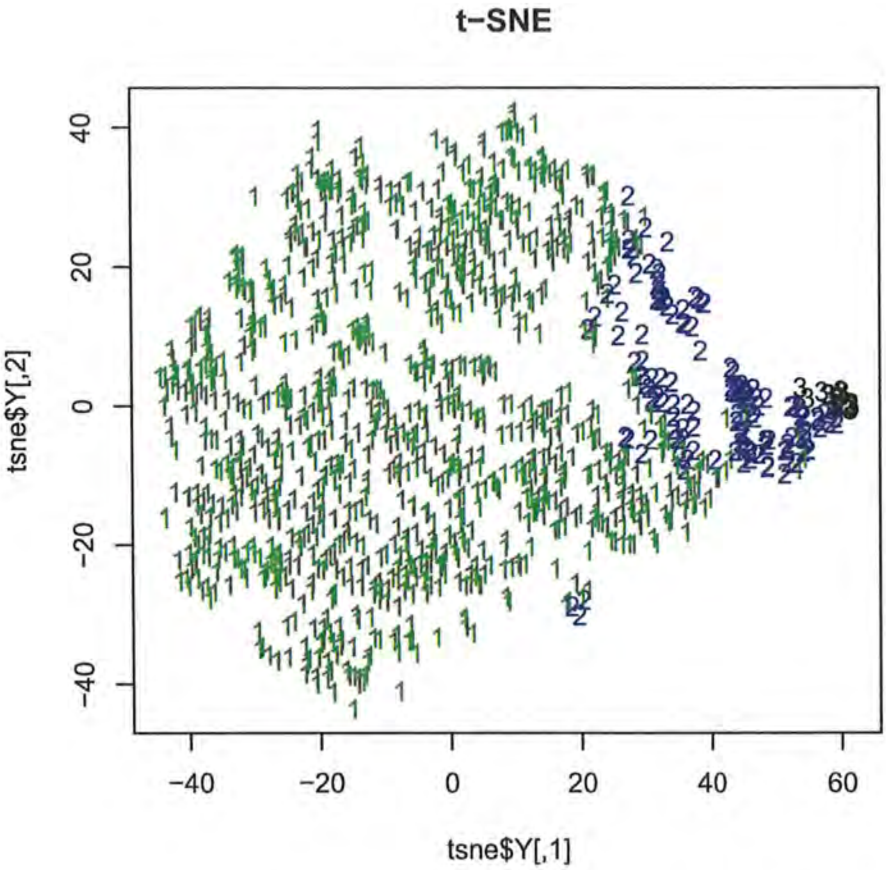


Figure 5.3: visualization of two dimension by t-SNE

5.4 Conclusion

Chapter five demonstrates, the application of PCA to Bottled water data to display water quality stratification is a common practice, and its plot is easily interpreted but using t-SNE is new. We show in Figure 5.3 that t-SNE can separate classes well as compared to PCA. Even though some problems in the application of PCA to revealing class stratification remain in t-SNE, but t-SNE in order to best show the class structure. In other words, the display of the pattern in t-SNE is more robust than PCA.

Chapter 6

Summary and Conclusion

The main purpose of this research is to analyze the quality of bottled water in Pakistan. To attain this purpose data of 538 brands of bottled water from April 2011 to September 2015 is used. Every brand is analyzed by using ten physcio chemical parameters such as pH,TDS, Cl, SO4, Ca, Mg, Hardness, EC, HCO3 and Na. As it is a difficult task for common man as well as for experts to analyze quality of water from all these parameters one by one. So for this purpose there are three water quality indices have been calculated. Their outcomes can be compared with standard values for instance with World Health Organization(WHO). In this regard we find three water quality indices and selected one for our data set and used statistical technique such that Decision trees, Artificial neural network and Support vector machine for analysis has done to attain the aim of study.

6.1 Detailed Summery of the Chapters

In Chapter 1, detailed information related to importance of water quality, significance of Bottled water and its quality is given. Moreover, literature review related to drinking water quality is discussed in detail.

In Chapter 2, detail data description and study area is given. Three water quality indices are discussed in detailed. As there are many water quality indices used by many researchers to assess the quality of water, we find the water quality index of every brand

and compare results with recommended values. According to their quality, each brand is categorized into one of five classes. There are two brands whose quality is highly poor and unsuitable for drinking. These brands are named as Asal Pia Na and Aquwa Plus. Authorities and s government need to pay attention to improve the quality of these brands. These brands have to warned to improve their quality otherwise must be highly fined or banned to safe health of people. Incentives ought given to those companies whose quality is in best state for example, according to three indices, Nestle, Aqua Life, Aqua Pure, Aquana and many others deserve appreciation.

In Chapter 3, used three decision tree technique. These are C4.5, C5.0 and Random forest, and are prefer for non-linear relationship of predictors and response. From these approaches we can extract the most important variables, these are EC, Na, TDS and HCO3 as given in figure 3.2. Prediction of test data set has been done by using these techniques and compare on the basis of accuracy rate and time. From the results it is conclude that C5.0 perform higher than other two techniques with respect to accuracy and time.

In Chapter 4, compare the performance of two techniques. One is artificial neural network using the back propagation method and other is multi-class support vector machine having linear, quadratic, polynomial, sigmoid and radial kernel functions. Predictions are done on the basis of test data set and compare the results with accuracy rate. Results conclude that multi-class svm using the polynomial kernel performs better as compared to other kernel functions and artificial neural network.

In Chapter 5, Principal Component Analysis and t-SNE are applied to get the information about the data setss. In this regard, Principal Components Analysis can be performed by specifying two dimensions. Based on the ten parameters, one may conclude that four components are responsible for 80.4% variation in data and in which three components are mainly Total Dissolved Solids (TDS), Electrical conductivity (EC), Hydrogen carbonate (HCO3). T-SNE is another method used here to compare with the two dimensional graph of PCA. In other words, the display of the pattern in t-SNE is more robust than

PCA. We conclude that the ability for t-SNE to reveal classes clustering are useful for water quality association studies.

References

Abbas, A. E. (2004). Entropy methods for adaptive utility elicitation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 34(2):169–178.

Ahmadi, E., Weckman, G. R., and Masel, D. T. (2017). Decision making model to predict presence of coronary artery disease using neural network and c5. 0 decision tree. *Journal of Ambient Intelligence and Humanized Computing*.

Aish, A. M., Zaqoot, H. A., and Abdeljawad, S. M. (2015). Artificial neural network approach for predicting reverse osmosis desalination plants performance in the gaza strip. *Desalination*, 367:240–247.

Ali, S. S., Anwar, Z., Khattak, J. Z. K., et al. (2012). Microbial analysis of drinking water and water distribution system in new urban peshawar. *Current Research Journal of Biological Sciences*, 4(6):731–737.

Alizadeh, M. J. and Kavianpour, M. R. (2015). Development of wavelet-ann models to predict water quality parameters in hilo bay, pacific ocean. *Marine pollution bulletin*, 98(1):171–178.

Asefa, T., Kemblowski, M., Urroz, G., and McKee, M. (2005). Support vector machines (svms) for monitoring network design. *Groundwater*, 43(3):413–422.

Astel, A., Tsakovski, S., Simeonov, V., Reisenhofer, E., Piselli, S., and Barbieri, P. (2008). Multivariate classification and modeling in surface water pollution estimation. *Analytical and Bioanalytical Chemistry*, 390(5):1283–1292.

Aydin, N. Y., Zeckzer, D., Hagen, H., and Schmitt, T. (2015). A decision support system for the technical sustainability assessment of water distribution systems. *Environmental Modelling & Software*, 67:31–42.

Aziz, A., Abd, R., and Wong, K.-F. V. (1992). A neural-network approach to the determination of aquifer parameters. *Groundwater*, 30(2):164–166.

Baldi, P. and Pollastri, G. (2002). A machine learning strategy for protein analysis. *IEEE Intelligent Systems*, 17(2):28–35.

Banerjee, P., Singh, V., Chatttopadhyay, K., Chandra, P., and Singh, B. (2011). Artificial neural network model as a potential alternative for groundwater salinity forecasting. *Journal of Hydrology*, 398(3):212–220.

Barbieri, P., Adami, G., Favretto, A., Lutman, A., Avoscan, W., and Reisenhofer, E. (2001). Robust cluster analysis for detecting physico-chemical typologies of freshwater from wells of the plain of friuli (northeastern italy). *Analytica Chimica Acta*, 440(2):161–170.

Brereton, R. G. (2003). *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons.

Brereton, R. G. (2009). *Chemometrics for pattern recognition*. John Wiley & Sons.

Buck, O., Niyogi, D. K., and Townsend, C. R. (2004). Scale-dependence of land use effects on water quality of streams in agricultural catchments. *Environmental Pollution*, 130(2):287–299.

Canu, S. (2005). Svm and kernel methods matlab toolbox. *http://asi. insa-rouen. fr/enseignants/˜ arakoto/toolbox/index. html*.

Çelik, Ö., Teke, A., and Yıldırım, H. B. (2016). The optimized artificial neural network model with levenberg–marquardt algorithm for global solar radiation estimation in eastern mediterranean region of turkey. *Journal of Cleaner Production*, 116:1–12.

Ceriani, L. and Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, 10(3):421–443.

Chih-Chung, C. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27–1.

Chowdhury, R. M., Muntasir, S. Y., and Hossain, M. M. (2012). Water quality index of water bodies along faridpur-barisal road in bangladesh. *Glob Eng Tech Rev*, 2(3):1–8.

Cloutier, V., Lefebvre, R., Therrien, R., and Savard, M. M. (2008). Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. *Journal of Hydrology*, 353(3):294–313.

Delgrange, N., Cabassud, C., Cabassud, M., Durand-Bourlier, L., and Laine, J. (1998). Neural networks for prediction of ultrafiltration transmembrane pressure–application to drinking water production. *Journal of membrane science*, 150(1):111–123.

El Aal, S. A., El-Saftawy, A., Alkadi, M., Salama, S., and Kandil, S. (2015). Quality evaluation of several brands of bottled mineral water from egypt and saudi arabia. *Asian Journal of Chemistry*, 27(9):3494.

Farid, S., Baloch, M. K., and Ahmad, S. A. (2012). Water pollution: Major issue in urban areas. *International journal of water resources and environmental engineering*, 4(3):55–65.

Ganapathiraju, A., Hamaker, J. E., and Picone, J. (2004). Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52(8):2348–2355.

Ghimire, B., Rogan, J., and Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the getis statistic. *Remote Sensing Letters*, 1(1):45–54.

Guardiola, J., García-Rubio, M. A., and Guidi-Gutiérrez, E. (2014). Water access and subjective well-being: The case of sucre, bolivia. *Applied Research in Quality of Life*, 9(2):367–385.

Gupta, M. and Agarwal, N. (2010). Classification techniques analysis. In *Proceedings of National Conference on Computational Instrumentation*.

Hartfield, K. A., Marsh, S. E., Kirk, C. D., and Carrière, Y. (2013). Contemporary and historical classification of crop types in arizona. *International journal of remote sensing*, 34(17):6024–6036.

Henley, W. and Hand, D. J. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *The statistician*, pages 77–95.

Ho, T. K. (1995). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huang, G. and Chang, N. (2003). The perspectives of environmental informatics and systems analysis. *Journal of Environmental Informatics*, 1(1):1–7.

Jabeen, A., Huang, X., and Aamir, M. (2015). The challenges of water pollution, threat to public health, flaws of water laws and policies in pakistan. *Journal of Water Resource and Protection*, 7(17).

Jantan, H., Hamdan, A. R., and Othman, Z. A. (2009). Classification techniques for talent forecasting in human resource management. In *International Conference on Advanced Data Mining and Applications*, pages 496–503. Springer.

JIAWEI, H., MICHELINE, K., and DATA, M. (2007). Concepts and techniques.

Kalavathy, S., Sharma, T. R., and Sureshkumar, P. (2011). Water quality index of river cauvery in tiruchirappalli district, tamilnadu. *Archives of Environmental Science*, 5:55–61.

Kamal, S. (2009). Pakistans water challenges: Entitlement, access, efficiency, and equity. *Running on Empty.*

Kannel, P. R., Lee, S., Lee, Y.-S., Kanel, S. R., and Khan, S. P. (2007). Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. *Environmental monitoring and assessment*, 132(1-3):93–110.

Karavoltsos, S., Sakellari, A., Mihopoulos, N., Dassenakis, M., and Scoullos, M. J. (2008). Evaluation of the quality of drinking water in regions of greece. *Desalination*, 224(1-3):317–329.

Kaur, D., Bedi, R., and Gupta, S. K. (2015). Review of decision tree data mining algorithms: Id3 and c4. 5. In *Proceedings of international conference on Information Technology and Computer Science*, pages 11–12.

Khodadadi, M., Mesdaghinia, A., Nasseri, S., Ghaneian, M. T., Ehrampoush, M. H., and Hadi, M. (2016). Prediction of the waste stabilization pond performance using linear multiple regression and multi-layer perceptron neural network: a case study of birjand, iran. *Environmental Health Engineering and Management Journal*, 3(2):81–89.

Klein, C. A. and Huang, L.-Y. (2008). Cultural norms as a source of law: The example of bottled water. *Cardozo L. Rev.*

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Neurocomputing: algorithms, architectures and applications.*

Lavrač, N. and Zupan, B. (2009). Data mining in medicine. In *Data Mining and Knowledge Discovery Handbook*, pages 1111–1136. Springer.

Liu, C. and Xia, J. (2004). Water problems and hydrological research in the yellow river and the huai and hai river basins of china. *Hydrological Processes*, 18(12):2197–2210.

Ma, Z., Song, X., Wan, R., Gao, L., and Jiang, D. (2014). Artificial neural network modeling of the water quality in intensive litopenaeus vannamei shrimp tanks. *Aquaculture*, 433:307–312.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Modaresi, F. and Araghinejad, S. (2014a). A comparative assessment of support vector machines, probabilistic neural networks, and k-nearest neighbor algorithms for water quality classification. *Water resources management*, 28(12):4095–4111.

Modaresi, F. and Araghinejad, S. (2014b). A comparative assessment of support vector machines, probabilistic neural networks, and k-nearest neighbor algorithms for water quality classification. *Water resources management*, 28(12):4095–4111.

Murugesan, A., Ramu, A., and Kannan, N. (2006). Water quality assessment from uthamapalayam municipality in theni district, tamil nadu, india. *Pollution Research*.

Nazeer, S., Hashmi, M. Z., and Malik, R. N. (2014). Heavy metals distribution, risk assessment and water quality characterization by water quality index of the river soan, pakistan. *Ecological indicators*, 43:262–270.

Omo-Irabor, O. O., Olobaniyi, S. B., Oduyemi, K., and Akunna, J. (2008). Surface and groundwater water quality assessment using multivariate analytical methods: a case study of the western niger delta, nigeria. *Physics and Chemistry of the Earth, Parts A/B/C*, 33(8):666–673.

Parizi, H. S. and Samani, N. (2013). Geochemical evolution and quality assessment of water resources in the sarcheshmeh copper mine area (iran) using multivariate statistical techniques. *Environmental earth sciences*, 69(5):1699–1718.

Patil, P. N., Sawant, D. V., and Deshmukh, R. (2012). Physico-chemical parameters for testing of water-a review. *International Journal of Environmental Sciences*, 3(3).

Pimpunchat, B., Sweatman, W. L., Wake, G. C., Triampo, W., and Parshotam, A. (2009). A mathematical model for pollution in a river and its remediation by aeration. *Applied Mathematics Letters*, 22(3):304–308.

Platzer, A. (2013). Visualization of snps with t-sne. *PloS one*, 8(2):e56883.

Poirion, O. B., Zhu, X., Ching, T., and Garmire, L. (2016). Single-cell transcriptomics bioinformatics and computational challenges. *Frontiers in genetics*, 7.

Random, B. L. (2001). Random forests. *Mach. Learn*, 45:5–32.

Rao, C. S., Rao, B. S., Hariharan, A., and Bharathi, N. M. (2010). Determination of water quality index of some areas in guntur district andhra pradesh.

Rosemann, N. (2005). Drinking water crisis in pakistan and the issue of bottled water: The case of nestlés pure life.. *Actionaid Pakistan*.

Samuel, O., Florence, N., and Ifeanyi, O. (2016). Microbial quality assessment of commercial bottled water brands in major markets in awka, nigeria.

Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*.

Sharma, B. and Tyagi, S. (2013). Simplification of metal ion analysis in fresh water samples by atomic absorption spectroscopy for laboratory students. *Journal of Laboratory Chemical Education*, 1(3):54–58.

Singh, K. P., Basant, N., and Gupta, S. (2011). Support vector machines in water quality management. *Analytica chimica acta*, 703(2):152–162.

Su, S., Zhi, J., Lou, L., Huang, F., Chen, X., and Wu, J. (2011). Spatio-temporal patterns and source apportionment of pollution in qiantang river (china) using neural-based modeling and multivariate statistical techniques. *Physics and Chemistry of the Earth, Parts A/B/C*, 36(9):379–386.

Taormina, R., Chau, K.-W., and Sethi, R. (2012). Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the venice lagoon. *Engineering Applications of Artificial Intelligence*, 25(8):1670–1676.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., and van Hijum, S. A. (2012). Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, 14(3):315–326.

Tripathy, J. and Sahu, K. (2005). Seasonal hydrochemistry of groundwater in the barrier spit system of the chilika lagoon, india. *Journal of Environmental Hydrology*, 13.

Tsakovski, S., Kudłak, B., Simeonov, V., Wolska, L., Garcia, G., Dassenakis, M., and Namieśnik, J. (2009). N-way modelling of sediment monitoring data from mar menor lagoon, spain. *Talanta*, 80(2):935–941.

Tsakovski, S. and Simeonov, V. (2011). Hasse diagram technique as exploratory tool in sediment pollution assessment. *Journal of Chemometrics*, 25(5):254–261.

Vakili, M., Sabbagh-Yazdi, S. R., Khosrojerdi, S., and Kalhor, K. (2017). Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. *Journal of Cleaner Production*, 141:1275–1285.

van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. *RBM*, 500(500):26.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160.

Yadav, S. K., Bharadwaj, B., and Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815.*

Zhang, Y., Guo, F., Meng, W., and Wang, X.-Q. (2009). Water quality assessment and source identification of daliao river basin using multivariate statistical methods. *Environmental Monitoring and Assessment*, 152(1-4):105.

Zhou, F., Guo, H., Liu, Y., and Jiang, Y. (2007a). Chemometrics data analysis of marine water quality and source identification in southern hong kong. *Marine Pollution Bulletin*, 54(6):745–756.

Zhou, F., Liu, Y., and Guo, H. (2007b). Application of multivariate statistical methods to water quality assessment of the watercourses in northwestern new territories, hong kong. *Environmental Monitoring and Assessment*, 132(1):1–13.