

ABSTRACT

Tons of toxic wastes are dumped into world's oceans, rivers, and lakes by the industries. The government of every country allows the release of toxic industrial effluents into the water bodies only up to certain limits. Yet these standards fail in case of dynamically changing external factors, such as climatic conditions, which may cause unnatural changes in the chemical composition of water, resulting in aberrant levels of toxicity. In order to protect all life forms from environmental hazards and provide an ecological balance, it is necessary to have a robust Water Quality Monitoring System that can dynamically detect and certify the quality of the water accordingly.

Real-time measurements of the various contributing factors such as water, soil and ambient air are continually monitored. Such continuous monitoring will result in voluminous and complex data which is difficult to process using traditional data processing applications. Rapid study and analysis of this data needs to be performed to make faster and more intelligent decisions. Hence there arises a need to integrate this heterogeneous data and deal with it as and when it flows in using Data Fusion.

In our project, we have developed a Water Quality Monitoring System using features of water, air and soil. The relevant features that influence the water quality are derived using appropriate Feature Extraction methods. Consequently, the individual decision making process is executed by classification using Support Vector Machine. An ensemble learning is applied to determine the water quality. The accuracy of the overall system developed is 95.25%.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iv
	LIST OF TABLES	ix
	LIST OF FIGURES	x
	LIST OF ABBREVIATIONS	xii
1.	INTRODUCTION	1
	1.1 Industrial Water Pollution	1
	1.2 Water Quality and Human Health	2
	1.3 Current Water Quality Monitoring Systems	3
	1.4 Need for Real-Time Monitoring	4
	1.5 Data Fusion and Machine Learning	4
	1.6 Usability	5
2.	LITERATURE SURVEY	7
	2.1 Water quality and its Parameters	7
	2.2 Data Fusion	7
	2.3 Feature Extraction	8
	2.4 Machine Learning	9
3.	SYSTEM DESIGN	10
	3.1 Overall System Design	10
	3.2 System Architecture	11
	3.3 Data Acquisition	11

3.4	Local Fusion Modules - Water, Air and Soil	12
3.4.1	Data Cleaning and Pre-processing	13
3.4.2	Ground Truth Check	13
3.4.3	Feature Extraction	13
3.4.4	Classification and Decision Making	14
3.5	Central Fusion Module	15
4.	ALGORITHMS	16
4.1	Ground Truth Checking Methods	16
4.1.1	K-Means Clustering	16
4.1.2	Voting System	17
4.2	Feature Extraction Models	18
4.2.1	Chi-Square Test of Independence	18
4.2.2	Random Forest Model	19
4.2.3	Recursive Feature Elimination using Random Forest	20
4.3	Classification Models	21
4.3.1	Naive Bayes Model	21
4.3.2	Support Vector Machine	22
4.3.3	Decision Trees	23
5.	IMPLEMENTATION OF PROPOSED SYSTEM	24
6.	RESULTS AND DISCUSSIONS	32
6.1	Clustering Results	32
6.2	Feature Extraction Results	33
6.3	Local Fusion Module Decisions	35

6.4	Central Fusion Module Decisions	43
6.5	Advantages of Proposed System	44
7.	CONCLUSION AND FUTURE WORK	46
7.1	Conclusion	46
7.2	Future Work	46
	REFERENCES	47

LIST OF TABLES

NO.	TABLE NAME	PAGE NO.
6.1	Clustering of UCI Water Dataset with 37 parameters and their centroids for pH	33
6.2	Performance measures for Water using Naive Bayes Classification	36
6.3	Performance measures for Water using SVM Classification	37
6.4	Performance measures for Air using Naive Bayes Classification	38
6.5	Performance measures for Air using SVM Classification	39
6.6	Performance measures for Soil using Naive Bayes Classification	41
6.7	Performance measures for Soil using SVM Classification	42
6.8	Combination of Local Decisions for Final Decision	43

LIST OF FIGURES

NO.	FIGURE NAME	PAGE NO.
1.1	Volume of wastewater generated from different industries in India	2
3.1	Overall System Design	10
3.2	System Architecture	11
3.3	Structure of Local Fusion Module	12
3.4	Input to the Central Fusion Module	15
5.1	Water Local Fusion Module - Ground Truth Check using Clustering	24
5.2	Water Local Fusion Module - Feature Extraction using RandomForest	25
5.3	Water Local Fusion Module - Classification using Support Vector Machine	26
5.4	Water Local Fusion Module - Local Decision using Decision Tree	26
5.5	Air Local Fusion Module - Feature Extraction using Chi-Squared method	27
5.6	Air Local Fusion Module - Classification using Support Vector Machine	28
5.7	Air Local Fusion Module - Local Decision using Decision Tree	28
5.8	Soil Local Fusion Module - Feature Extraction using RFE	29
5.9	Soil Local Fusion Module - Classification using Support Vector Machine	30
5.10	Soil Local Fusion Module - Local Decision using Decision Tree	30
5.11	Final Decision Tree at Central Fusion Module	31

6.1	Elbow method to find optimal clusters	32
6.2	2D representation of cluster results for UCI Water Dataset	33
6.3	List of Attributes by Feature Selection methods for Water	34
6.4	List of Attributes by Feature Selection methods for Air	35
6.5	List of Attributes by Feature Selection methods for Soil	35
6.6	Classifiers Vs Accuracy for Water with Naive Bayes Classification	36
6.7	Classifiers Vs Accuracy for Water with SVM Classification	37
6.8	Decision Tree - Local Decision for Water Module	38
6.9	Classifiers Vs Accuracy for Air with Naive Bayes Classification	39
6.10	Classifiers Vs Accuracy for Air with SVM Classification	40
6.11	Decision Tree - Local Decision for Air Module	40
6.12	Classifiers Vs Accuracy for Soil with Naive Bayes Classification	41
6.13	Classifiers Vs Accuracy for Soil with SVM Classification	42
6.14	Decision Tree - Local Decision for Soil Module	43
6.15	Decision Tree - Central Decision determining water quality	44

LIST OF ABBREVIATIONS

1. SVM - Support Vector Machine
2. RFE - Recursive Feature Elimination
3. UCI - University of California, Irvine
4. DDT - Dichloro Diphenyl Trichloroethane
5. PCB - Polychlorinated Biphenyls
6. CPCB - Central Pollution Control Board
7. OOB - Out of Band

CHAPTER 1

INTRODUCTION

It is a known fact that water covers two-thirds of the Earth's surface, with over 97% present in the oceans and less than 1% in freshwater streams and lakes. Water is also present in the atmosphere in solid form in the polar icecaps and as groundwater in aquifers (water-bearing rocks) deep underground. Thus, we need to have a balance between usage and wastage of this water, which is an indispensable element of our lives.

1.1 Industrial Water Pollution

Water pollution may be defined as any chemical or physical change in water detrimental to living organisms. It is the leading worldwide cause of deaths and diseases, accounting for the death of more than 14,000 people daily - estimated 580 in India ill everyday.

There are several causes of water pollution in India namely unrelenting urbanization, growing population, hydraulic fracturing, agricultural runoff, withdrawal of water and religious and social practices. Water bodies are a major recipient of an extensive array of wastes produced by human activity. These may be discharged directly into watercourses by sewers or be washed down from agricultural or urban areas particularly after heavy rains.

One paramount cause of this catastrophe is unregulated Industrial Waste Discharge. In more developed countries, industrial pollutants, such as asbestos, lead, mercury, nitrates, phosphates, sulphur, oils, petrochemicals and heat add to the water pollution problem. From complex inorganic chemical industries, electrical power plants, food industries, iron and steel industries to nuclear

industries, mines and quarries, the total wastewater generated from all major industrial sources is 83,048 Mid which includes 66,700 Mid of cooling water generated from thermal power plants. Out of remaining 16,348 mid of wastewater, thermal power plants generate another 7,275 Mid as boiler blow down water and overflow from ash ponds. Engineering industries comprise the second largest generator of wastewater in terms of volume.^[11]

The other significant contributors of wastewater are paper mills, steel plants, textile and sugar industries. The major contributors of pollution in terms of organic load are distilleries followed by paper mills. Figure 1.1 shows the volume of wastewater from different industries in India.

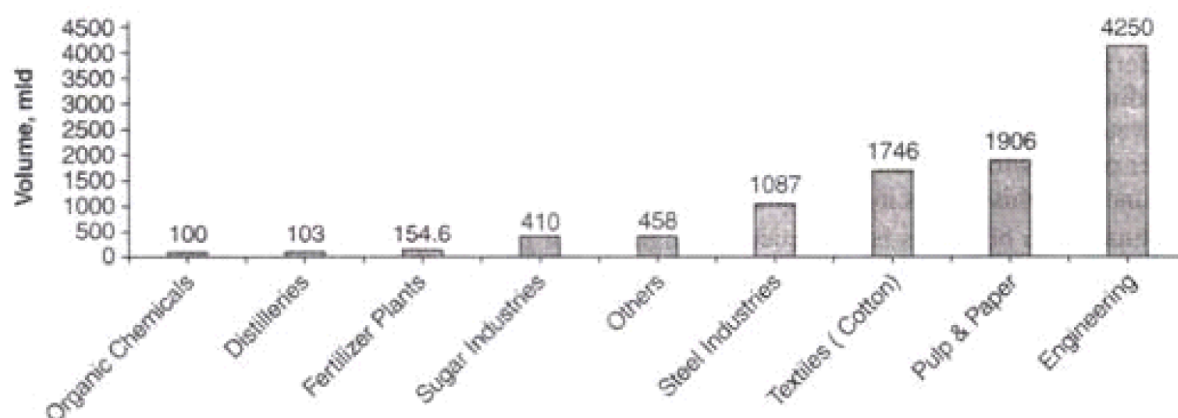


Fig. 1.1. Volume of wastewater generated from different industries in India

Courtesy-<http://www.yourarticlelibrary.com/water-pollution/5-major-causes-of-water-pollution-in-india/19764/>

Industries should not be allowed to discharge untreated chemicals into the water bodies. There should be a mechanism to ensure that only harmless substances are poured into rivers, lakes and oceans.^[11]

1.2 Water Quality and Human Health

Virtually all water pollutants are hazardous to humans as well as lesser species; sodium is implicated in cardiovascular disease, nitrates in blood disorders.

Mercury and lead can cause nervous disorders. Some contaminants are carcinogens. DDT is toxic to humans and can alter chromosomes. PCBs cause liver and nerve damage, skin eruptions, vomiting, fever, diarrhoea, and fatal abnormalities. Along many shores, shellfish can no longer be taken because of contamination by DDT, sewage, or industrial wastes.

Dysentery, salmonellosis, cryptosporidium, and hepatitis are among the maladies transmitted by sewage in drinking and bathing water. Beaches along both coasts, riverbanks, and lakeshores have been ruined for bathers by industrial wastes, municipal sewage, and medical waste. Water pollution is an even greater problem in India, where millions of people obtain water for drinking and sanitation from unprotected streams and ponds that are contaminated with human waste. This type of contamination has been estimated to cause more than 3 million deaths annually from diarrhoea in India, most of them children.

India has enacted extensive federal legislation to fight water pollution. Laws include Water (Prevention and Control of Pollution) Act (1974), The Water (Prevention and Control of Pollution) Cess Act (1977), The Maharashtra Prevention of Water Pollution Act (1969), The River Boards Act (1956) and The Orissa River Pollution Act (1953).

1.3 Current Water Quality Monitoring Systems

There are existing water quality monitoring stations developed and implemented in India by many organizations and boards such as Central Pollution Control Board (CPCB), Water Resources Information System of India (India - WRIS), etc. CPCB has established a nationwide network of water quality

monitoring comprising 2500 stations all over India. Water samples are being analysed for 28 parameters consisting of 9 core parameters.^{[1] [2]}

1.4 Need for Real-Time Monitoring

One setback of the existing systems is that the monitoring is done on monthly or quarterly basis in surface waters and on half yearly basis in case of ground water. This does not account for urban sewage discharges, extreme environmental conditions, high humidity coastal conditions, high temperature desert conditions, frequent climatic changes, natural disasters and its after effects such as after the Chennai Floods, etc. Water is being consumed on an everyday basis by millions of humans and lower species. Hence, regular checking of water quality in shorter intervals of time is crucial, having in mind all the extreme conditions also.

Moreover, in order to eliminate problems associated with manual water quality monitoring, installation of 'Real-Time Water Quality Monitoring Network' is essential. Hence, the need for Real-Time Water Quality Monitoring Systems.

1.5 Data Fusion and Machine Learning

Machine Learning is defined as "a field of study that gives computers the ability to learn without being explicitly programmed". It explores the study and construction of algorithms that can learn from and make predictions on data. We aim to develop an algorithm that allows us to implement a real-time system that predicts the water quality at a given place and time. Thus, a combination of various machine learning mechanisms is used.

Industrial effluents deteriorate not only the river water with their liquid wastes, but also deteriorate the air immediately above these waters and the river beds, ultimately affecting the water quality. So, it is absolutely necessary to include the air and soil quality parameters when predicting the water quality. This can be achieved by a phenomenon called Data Fusion.

Data fusion is the process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation. Fusion of the data from 2 sources (dimension 1 & 2) can yield a classifier superior to any classifiers based on dimension 1 or dimension 2 alone. Here, the dimensions used are water, ambient air and soil of the river to be tested. Along with Data Fusion, a methodology called the Feature Extraction can also be used for faster and effective results. Feature Extraction is a process which starts from an initial set of measured data and builds derived features intended to be informative and non-redundant, in our case, builds to dimensionality reduction.^{[4][10]}

An innovative combination of Machine Learning and Data Fusion techniques can be used to obtain accurate and smart results in lesser time. This also has a positive effect on the costs and can prove to be a cost-effective system compared to existing ones.

1.6 Usability

Once water has been labelled as potable or not, the industries nearby responsible for the deterioration of river waters must be notified about their unregulated discharge of effluents and must be asked to implement more filtering techniques with respect to the concentration of the elements discharged. The

industries shall be alerted at any time whenever there is a breach in the water quality in the neighbourhood. Strict measures must be brought in to regulate the industrial discharge and when this is done, the water becomes less polluted and usable for the people relying on it.^[9]

CHAPTER 2

LITERATURE SURVEY

2.1 Water Quality and its Parameters

In the report published by the CPCB, existing water quality monitoring systems, such as the National Water Quality Monitoring Programme, monitor water quality in a static manner based on 54 water parameters. With these 54 parameters the quality of water is statistically analysed on an half yearly or yearly basis. Depending upon the core 9 parameters' values of water they qualify water as drinking water source, Outdoor bathing, Irrigation, Industrial Cooling, Controlled Waste disposal etc.^{[8][18]}

However, in the proposed system, we consider not only the core parameters of water but also the conditions which may affect the quality of water like the soil attributes and air attributes which may influence the quality of water attributes one way or the other. We also dynamically analyse the quality of water by taking the data on a regular basis instead of half yearly or yearly basis with the help of big data tools like R. We fuse the data parameters of water, air and soil efficiently to foresee any possible quality issues and alert the system.

2.2 Data Fusion

In prevalent systems that use Multi-sensor Data Fusion for Water Quality Monitoring with the help of Wireless sensor networks, the parameters of water are not measured using readings from a single sensor but with the help of multiple sensors. Each attribute reading is valuated to a decision in the local fusion modules

and the decisions from each local fusion module is combined for an efficient decision in the central fusion module^{[5][6]}.

In our project also, we use the same idea, not only for measuring water attributes but for fusion too. We divide the local fusion modules into water module, air module and soil module. We arrive at a decision for each of the modules in the local nodes and send these individual decisions to a central fusion node. In the central fusion node, these decisions are fused to arrive at the final result.

2.3 Feature Extraction

Dimensionality reduction is a very important step in the data mining process. In the study ‘Feature Extraction for Classification in the Data Mining Process’, they consider feature extraction for classification tasks as a technique to overcome problems. Three different eigenvector-based feature extraction types of PCA approaches are discussed. In this, feature extraction is used as predecessor step before classifier. Also, a decision support system that integrates the feature extraction and classification processes. They conclude by stating that the result obtained with the parameters of original dataset can be obtained with the parameters reduced.^[7]

In our project, we implement the same sequential steps of feature extraction, classifiers and decision trees. Since there is no feature extraction method that would be the most suitable for all classification tasks, we have implemented Chi-square test, Random Forest, Random Forest with Recursive Feature Elimination, union of Chi-square and Random Forest, intersection of Chi-square and Random Forest. Since our data sources are water, air and soil, we used the different feature extraction methods pertaining to the data set. We used Random Forest for water, Chi-square for air and Random Forest with Recursive Feature Elimination for soil.

2.4 Machine Learning

Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled. If many variations of models are combined, the results are better often much better and at little extra effort for the user. Creating such a model is the simplest technique, called bagging, we simply generate random variations of the training set by resampling, learn a classifier on each, and combine the results by voting. In boosting, training examples have weights, and these are varied so that each new classifier focuses on the examples the previous ones tended to get wrong^[3]. In our system, there is a need for the implementation of machine learning algorithms in order to automate the system and to achieve dynamic performance. So, we have used Random Forest for Feature Extraction, Support Vector Machine for classifiers, Decision Trees for Decision support system. As a result, our system can be implemented to a wide scope of industrial pollution control system.

CHAPTER 3

SYSTEM DESIGN

3.1 Overall System Design

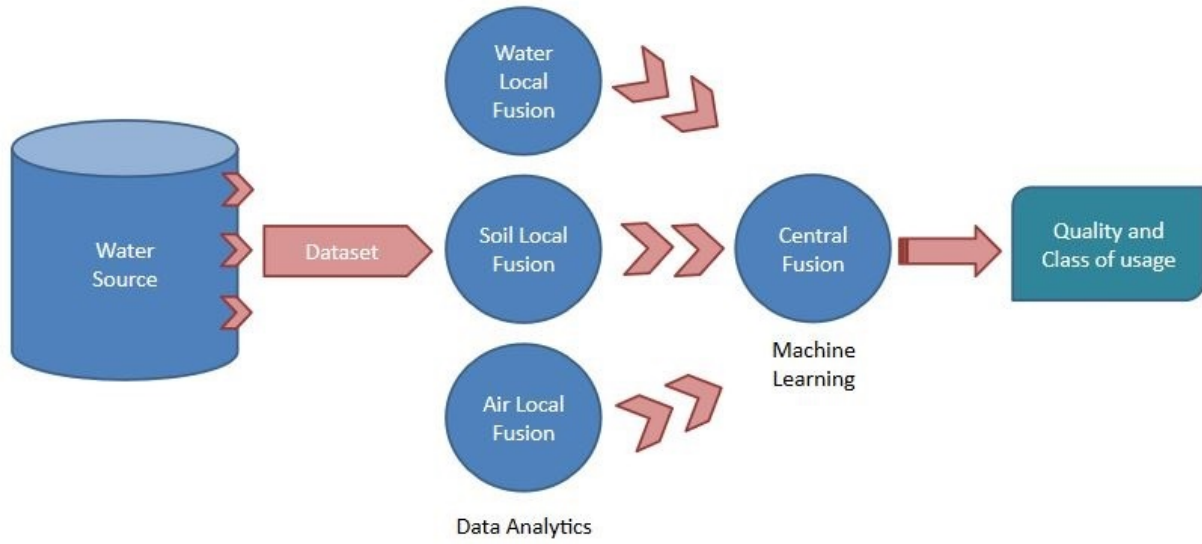


Fig. 3.1. Overall System Design

The overall design of the system can be broken down into three components as shown in figure 3.1. First, Data Acquisition module, Local Fusion modules and Central Fusion Modules. Water quality parameters are measured from the river in short intervals of time. They are fed into the system at regular intervals. The data that flows in is cleaned here and sent to the Local Fusion module for Water. This Local Fusion module computes the cleaned data to produce a local decision label for the water. Similarly Air and Soil Local Fusion modules process air and soil quality parameters respectively. On these three modules three local decisions are derived and sent to the Central Fusion Module. This component delivers the final class label denoting whether the water is safe or not. These labels are noted at regular intervals.

3.2 System Architecture

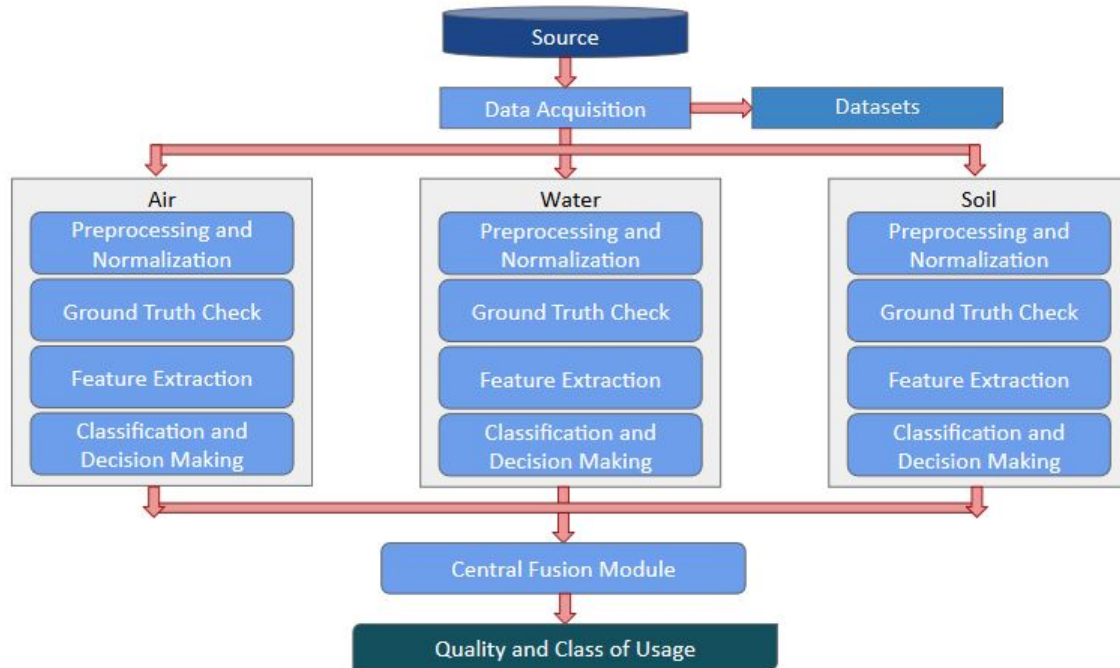


Fig. 3.2. System Architecture

Figure 3.2 describes the detailed architecture of the proposed system. Each of the components and modules are explained in the below sections.

3.3 Data Acquisition

Sensor data is streamed in from the river waters, riverbeds and the air above it in short intervals. This could be done with a wide sensor network distributed over various zones of the river. Although, it is essential that this sensor network be in close proximity to an industrial discharge area so that close monitoring of water quality can be done and damages can be identified and rectified immediately before it gets out of hand.

3.4 Local Fusion Modules - Water, Air and Soil

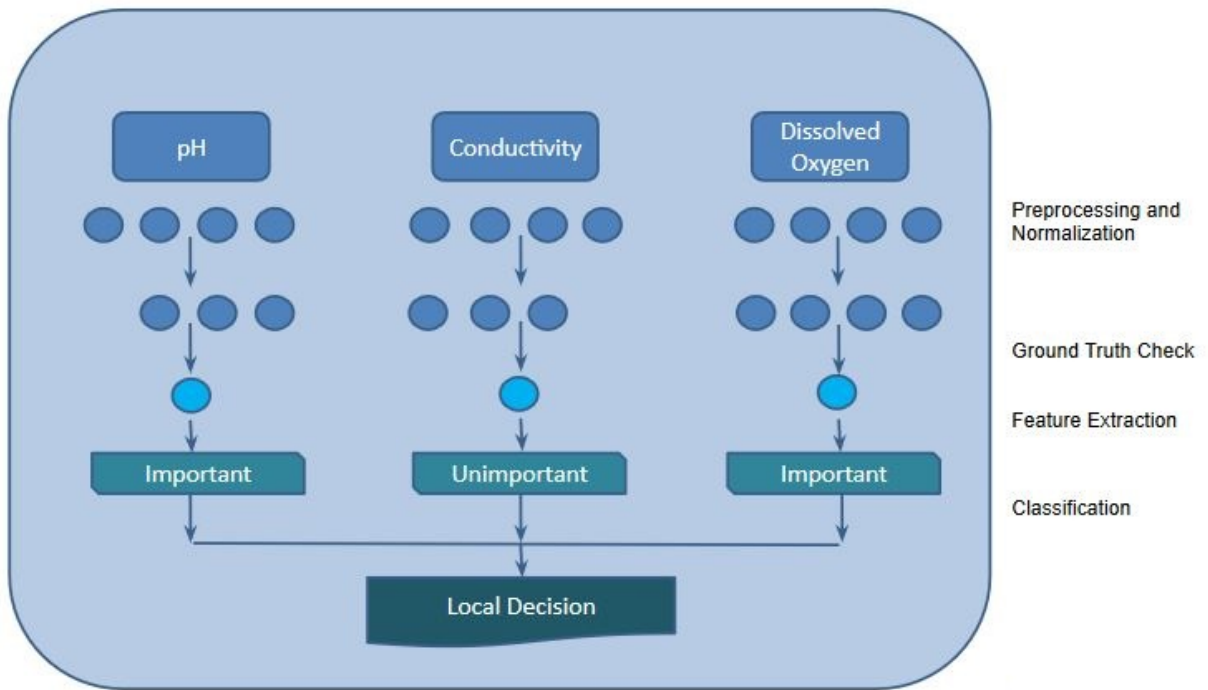


Fig. 3.3. Structure of Local Fusion Module

There are three Local Fusion modules assigned to water, air and soil quality parameters respectively. Each module executes the procedure to clean and pre-process the input data, to reduce its dimensionality and to yield a local decision as shown in figure 3.3.

There are four major steps enforced in these modules -

- Data Pre-processing and Cleaning
- Ground Truth Check
- Feature Extraction
- Classification and Decision Making

These steps are elaborated in the below sections.

3.4.1 Data Pre-processing and Cleaning

The input data arriving may not be of the expected format. Due to various factors such as faults in sensors, environmental conditions, location specific contamination due to other causes and more, some of the data may be erroneous or missing. The results will be skewed if the data is faulty, which may result in incorrect decisions. Thus, as a precautionary measure and a safety measure, it is desirable to always pre-process the data before computing it.

That being so, the data flowing into the local fusion modules is checked for missing and erroneous values and records having such values are removed.

3.4.2 Ground Truth Check

Every dataset contains a ground truth which describes the dataset and also specifies the different classes to which the records belong. It is a vital step to check whether our methods give results in alignment with the existing description. Thus, we use a few methods such as -

- K-Means Clustering
- Voting System

inside local fusion modules to cross verify the results and stay aligned.

3.4.3 Feature Extraction

Once we are sure the data is cleaned, we can use it for the decision making process. But, remember the quality parameter readings will be flowing into the system at regular intervals. As a result, a lot of information flows in and increases the overhead for the classification and the decision making activities. There will be a lesser control over the data. In regard to this, we will not be able to cut down on the number of features in the data acquisition operation initially since core attributes determining the quality vary with data and are not constant. Differences

exist due to different factors namely their origins, their intended usage, etc. To that account, it is wise to cut down the unimportant features at this stage.

When control over horizontal loading of data is hard to achieve, vertical control using Feature Extraction is advisable. This approach can greatly reduce time & space and improve the computing efficiency. It also reduces the amount of resources required to describe the large set of data.

The various Feature Extraction methods used in the system are -

- Chi - Squared method
- Neural Networks with Feature Extraction - By Importance
- Linear Vector Quantization (LVQ)
- Recursive Feature Elimination using Random Forest
- Random Forest
- Boruta method - Feature Selection Wrapper Algorithm
- Partial Least Squares Discriminant Analysis (PLSDA) - Resampling

These methods are performed on to the data and the best methods for every data type can be identified and used.

3.4.4 Classification and Decision Making

With the reduced Feature vector, the data in hand is refined and accurate. Partitioning it into training dataset and test dataset, we can commence the classification and prediction of data, setting rules in line with the Standards provided by the Indian Legislation. Many classification algorithms can be implemented for this purpose. Two of these algorithms used here are -

- Naive Bayes Classification
- Support Vector Machine (SVM)

The classifier labels the records into specific classes such as safe and unsafe. Once the class labels are assigned, Decision Trees can be used to take a wholesome decision of that module in that interval. A Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences which is used in decision analysis. With Decision Tree Learning, a decision tree can be seen as a predictive model to predict the quality.

3.5 Central Fusion Module

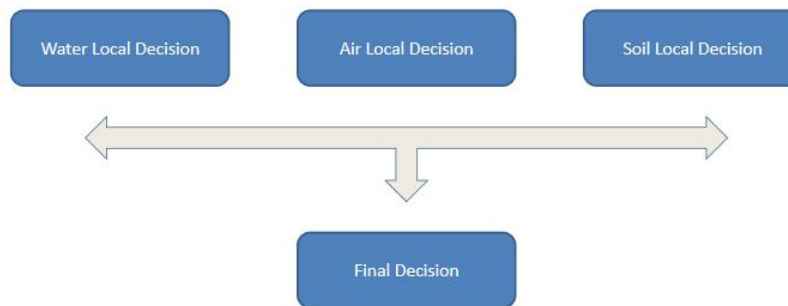


Fig. 3.4. Input to the Central Fusion Module

The above processes are performed on air and soil parameters also. Together, the three local decisions are sent to the Global or Central Fusion module where a central and final decision can be derived from the intermediate decisions which is illustrated in the figure 3.4. Decision Tree Learning, again, is used to conclude about the final decision.

CHAPTER 4

ALGORITHMS

In this chapter, the various algorithms implemented and their functionalities in this system are described in detail.

4.1 Ground Truth Checking Methods

As already mentioned, every dataset will contain a dataset description and Ground Truth. Accounting to that, we need to check whether our methods align with the original classification. If a dataset does not contain Ground Truth, it is desirable to use Voting System.

4.1.1 K-Means Clustering

It is a method of vector quantization that is popular for cluster analysis in Data Mining. K-Means Clustering aims to partition 'n' observations into 'k' clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given a training set $\{x(1), \dots, x(m)\}$, we want to group the data into a few cohesive 'clusters'. Here, $x(i) \in R^n$ as usual; but no labels $y(i)$ are given. So, this is an unsupervised learning problem.

The k-means clustering algorithm is as follows:

1. Initialize cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in R^n$ randomly.
2. Repeat until convergence: {

For every i , set

$$c^{(i)} := \arg \min_j ||x^{(i)} - \mu_j||^2.$$

For each j , set

$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\}x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

}

In the algorithm above, k (a parameter of the algorithm) is the number of clusters we want to find; and the cluster centroids μ_j represent our current guesses for the positions of the centres of the clusters.

To initialize the cluster centroids (in step 1 of the algorithm above), we could choose k training examples randomly, and set the cluster centroids to be equal to the values of these k examples. (Other initialization methods are also possible). The inner-loop of the algorithm repeatedly carries out two steps:

- (i) “Assigning” each training example $x^{(i)}$ to the closest cluster centroid μ_j , and
- (ii) Moving each cluster centroid μ_j to the mean of the points assigned to it.^[24]

In our project, we used this algorithm to check the cluster labels with the original class labels in the Ground Truth of the water dataset in order to check the accuracy. Also, we have attained the desired result which matched with a 95% accuracy. In order to implement this algorithm in R, we used the packages ‘stats’^[35] and ‘cluster’.^[29]

4.1.2 Voting System

In our project, we determined the class labels for each dataset using the voting system. This method is used when the ground truth is not available. According to this, 0's and 1's are assigned for parameter based on its ideal condition. 0's correspond to bad value and 1's correspond to good value. If the

count is greater than a desired value, then the class label is assigned to 'class1' which corresponds to good class if it is a binary class otherwise 'class2' which corresponds to bad class.

4.2 Feature Extraction Models

Feature Extraction models are used to have a vertical control over the data flowing in, i.e., cut down the irrelevant features for classification and hence bring in dimensionality reduction to bring accurate results.

4.2.1 Chi-Squared Test of Independence

The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. The test procedure is appropriate when the following conditions are met :

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps :

- State the hypotheses - Determine whether the two attributes are dependent or independent
- Formulate an analysis plan - The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the two elements which are Significance level and Test method.

- Analyse sample data - Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic.
- Interpret results - If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.^{[12][19]}

In our project, we used this algorithm in order to decide the primary attributes which influenced the class labels based on the attribute importance. Only those primary attributes were used for further processing. In order to implement this algorithm in R, we used the package ‘FSelector’.^[31]

4.2.2 Random Forest Model

Random forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This method combines bagging idea and the random selection of features.

The training algorithm applies the general technique of bootstrap aggregating or bagging to tree learners. Given a training set $X = x_1, \dots, x_n$ with responses $Y = y_1, \dots, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples.

Random forests can be used to rank the importance of variables in a regression or classification problem. The algorithm which we used, measures the variable importance in a data set is to fit a random forest to the data. During the fitting process the out-of-bag error for each data point is recorded and averaged

over the forest. To measure the importance of the j -th feature after training, the values of the j -th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j -th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. Features which produce large values for this score are ranked as more important than features which produce small values.^{[20][21]}

In our project, we used this algorithm in order to decide the primary attributes randomly based on the variable importance. Only those primary attributes were used for further processing. In order to implement this algorithm in R, we used the package ‘randomForest’.^[33]

4.2.3 Recursive Feature Elimination using Random Forest

Random Forest RFE algorithm is based on recursive elimination of variables. More precisely, they first compute RF variable importance. Then, at each step, they eliminate the 20% of the variables having the smallest importance and build a new forest with the remaining variables. They finally select the set of variables leading to the smallest OOB error rate. The proportion of variables to eliminate is an arbitrary parameter of their method and does not depend on the data.^{[15][16]}

In our project, we used this algorithm in order to decide the primary attributes randomly based on the variable importance. Only those primary attributes were used for further processing. In order to implement this algorithm in R, we used the packages ‘mlbench’^[32] and ‘caret’^[28].

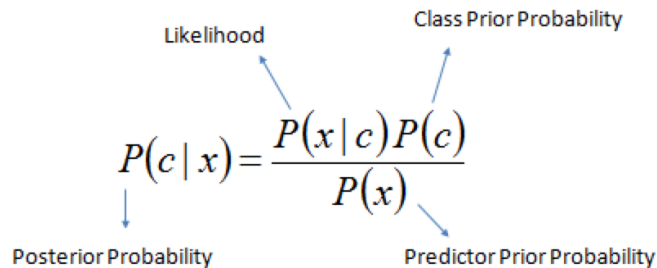
4.3 Classification Models

With the refined data, further classification in the local modules as well as the central modules can be performed using the below methods.

4.3.1 Naive Bayes Model

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from each term to a label: $P(c|x)$ points to 'Posterior Probability', $P(x|c)$ points to 'Likelihood', $P(c)$ points to 'Class Prior Probability', and $P(x)$ points to 'Predictor Prior Probability'.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

$P(c|x)$ is the posterior probability of class (target) given predictor (attribute).

$P(c)$ is the prior probability of class.

$P(x|c)$ is the likelihood which is the probability of predictor given class.

$P(x)$ is the prior probability of predictor.^[17]

In our project, we used this algorithm since Naive Bayes is efficient for binary classification. In order to implement this algorithm in R, we used the package ‘e1071’.^[30]

4.3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning method that can be used to carry out general regression and classification, as well as density-estimation. A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

In our project, we used C-Classification SVM.

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$$

subject to the constraints:

$$y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \dots, N$$

where C is the capacity constant, w is the vector of coefficients, b is a constant, and represents parameters for handling non separable data (inputs). The index i labels the N training cases. Note that represents the class labels and x_i represents the independent variables. The kernel is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C , the more the error is penalized. Thus, C should be chosen with care to avoid over

fitting.^{[22][23]} In order to implement this algorithm in R, we used the package ‘e1071’.^[30]

4.3.3 Decision Trees

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree. In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values.

This algorithm are greedy by nature and construct the decision tree in a top–down, recursive manner (also known as “divide and conquer”). In each iteration, the algorithm considers the partition of the training set using the outcome of a discrete function of the input attributes. The selection of the most appropriate function is made according to some splitting measures. After the selection of an appropriate split, each node further subdivides the training set into smaller subsets, until no split gains sufficient splitting measure or a stopping criteria is satisfied.^{[13][14]} In order to implement this algorithm in R, we used the package ‘rpart’.^[34]

CHAPTER 5

IMPLEMENTATION OF PROPOSED SYSTEM

This chapter describes the detailed explanation about the implementation process. The water dataset used here is ‘UCI Water Treatment Plant Dataset’ containing 37 attributes^[25]. The air dataset used is ‘Air Pollutant concentrations 2013’ which includes 6 parameters^[26]. ‘Soil data derived from SOTER’ is the soil dataset utilized consisting of 10 parameters^[27]. The software used is RStudio Version 0.99.491 with R version 3.2.3.

Water Local Fusion Module

Dataset should be retrieved and loaded onto workspace. Following that the dataset must be cleaned and pre-processed before further processing by throwing out the records with missing or erroneous data. Also, it must be checked with the Ground truth of the dataset. This is performed using K-Means Clustering in the Water Module as depicted in figure Fig.5.1

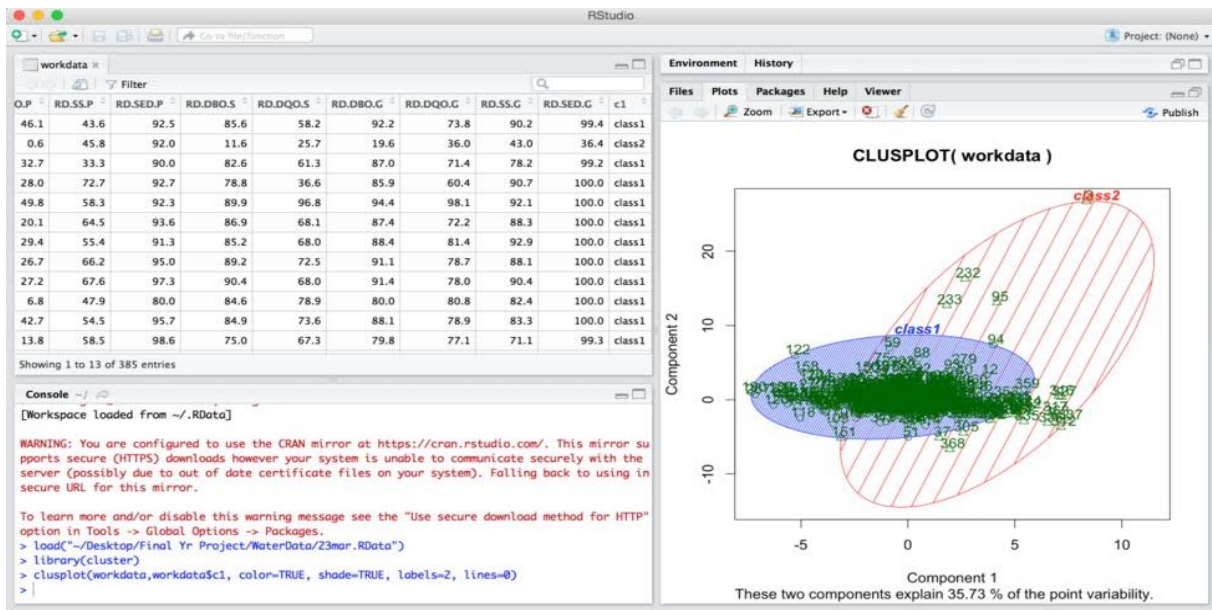


Fig. 5.1. Water Local Fusion Module - Ground Truth Check using Clustering

After that is done, Feature Extraction can be carried out to cut down the unimportant features and also to identify the attribute weights that determine their importance. Training models such as Random Forest can be used for this purpose such as in the below figure Fig.5.2.

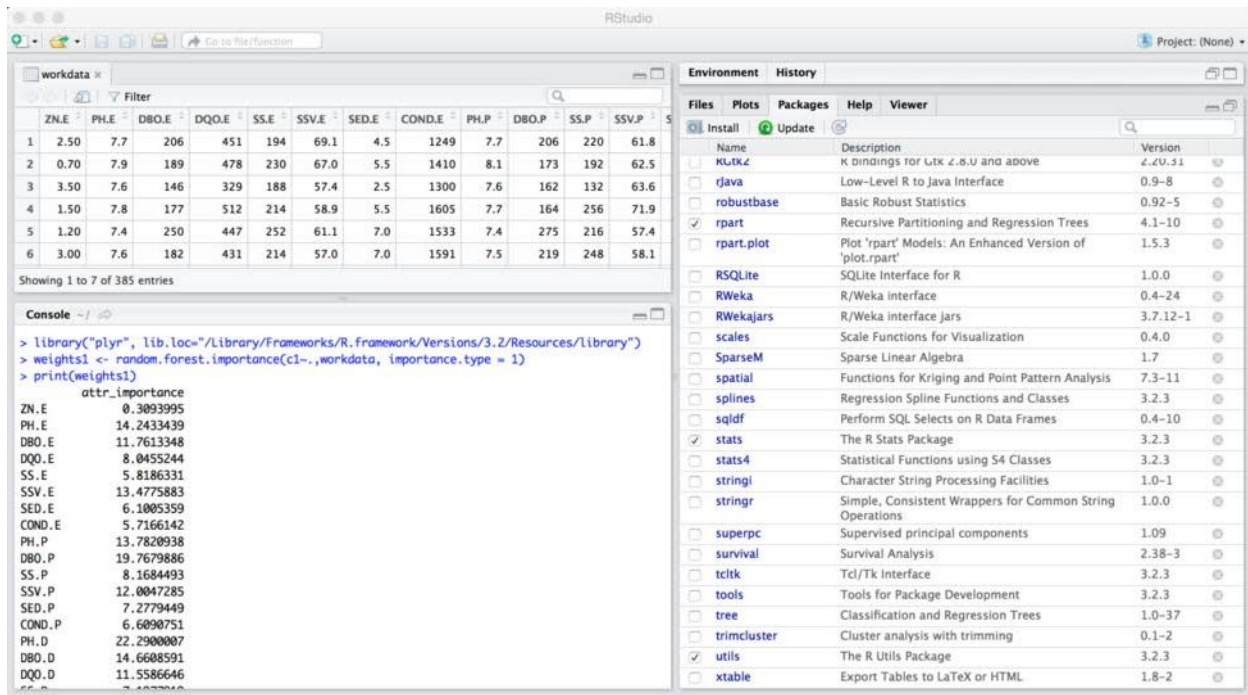
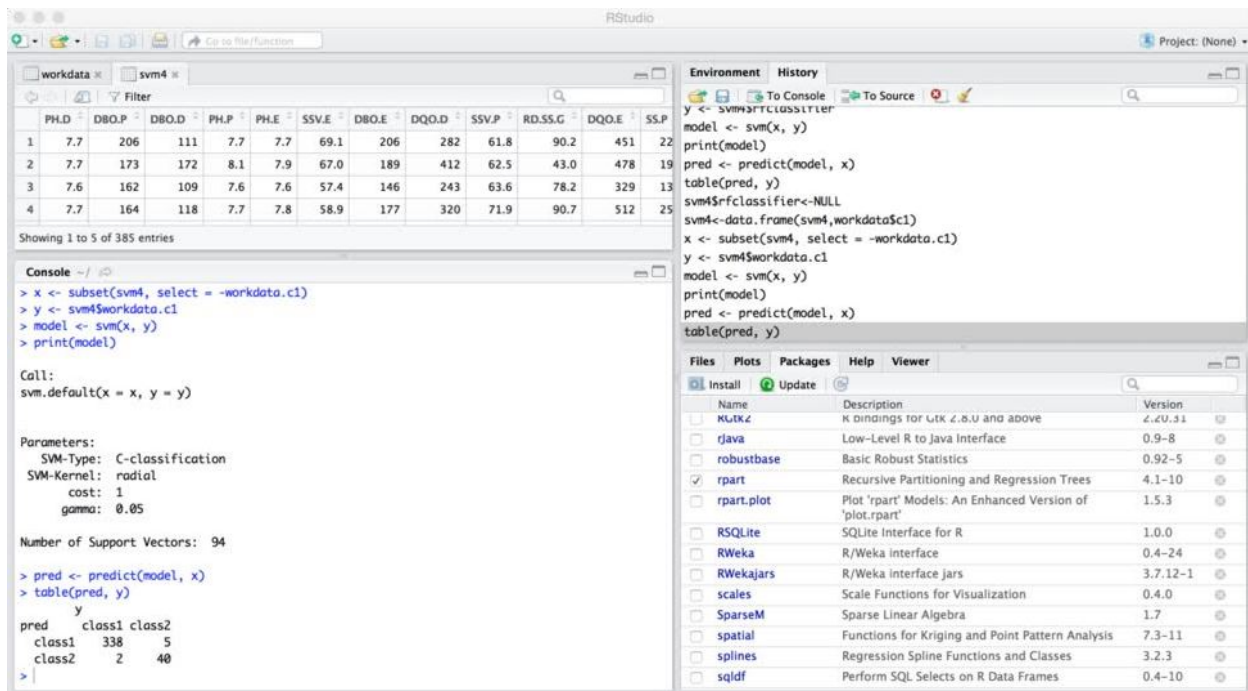
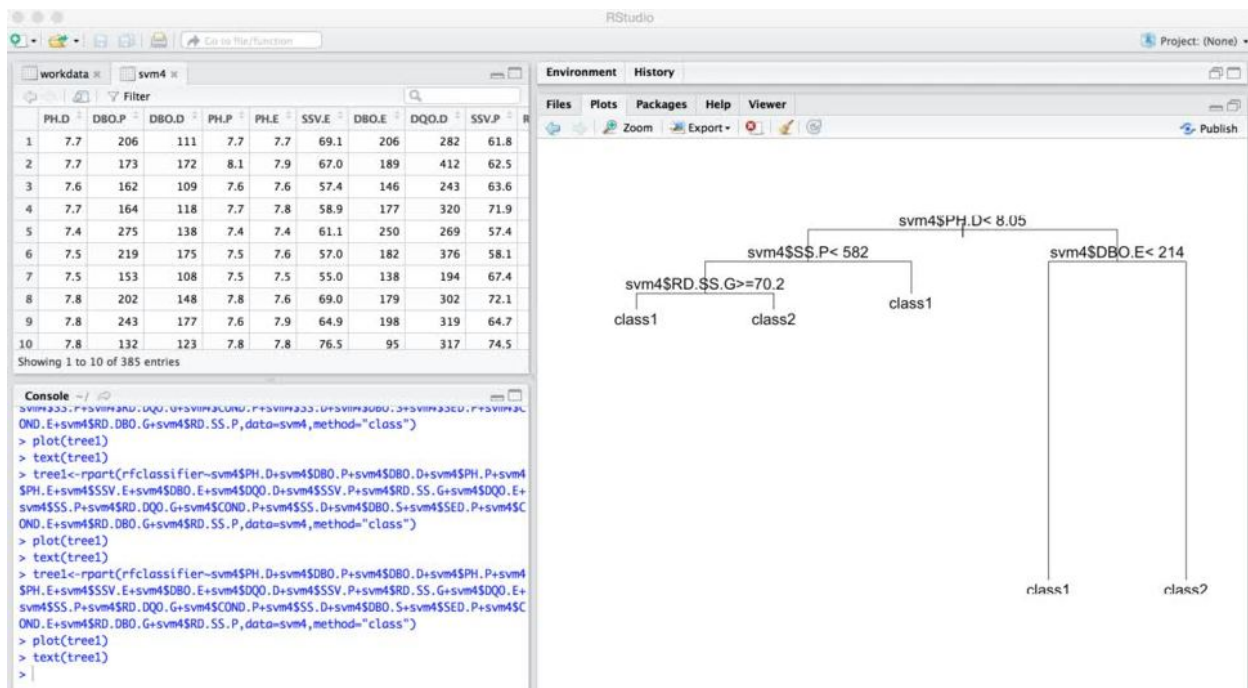


Fig. 5.2. Water Local Fusion Module - Feature Extraction using Random Forest

Using appropriate tools and libraries, one can obtain the attribute weights and derive the significant attributes. When the dataset is cut down, it can now be split into training dataset and test dataset and so, classification and prediction can be performed using appropriate classifiers as shown in the below figure Fig.5.3.



Consequently, the local decision tree can be given by Decision Tree Learning and this is shown in the diagram below Fig.5.4.



Local decision regarding the quality of water parameters alone can be sent to the Central Fusion Module.

Air Local Fusion Module

Similar processing is done for Air Module. Since there is no Ground Truth for the air dataset obtained, Voting System method is used after the dataset is pre-processed. K-Means Clustering can be used if the dataset contains Ground Truth. But here Voting System is performed. The remaining steps are similar to Water Module. Irrelevant features are removed here using Chi-Squared method, which proves to be efficient than other methods for this section as shown in Fig.5.5.

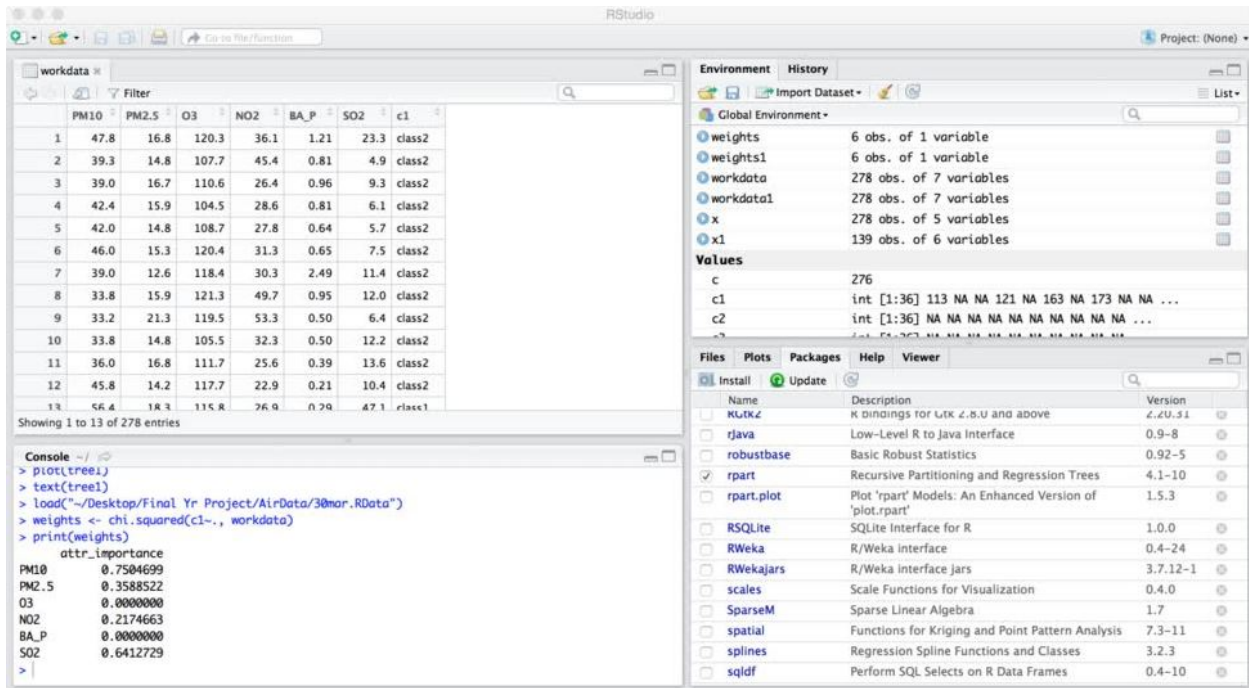


Fig. 5.5. Air Local Fusion Module - Feature Extraction using Chi-Squared method

Now, Classification and Prediction of records can be done using appropriate classifiers such as Naive Bayes or Support Vector Machine as in Fig. 5.6.

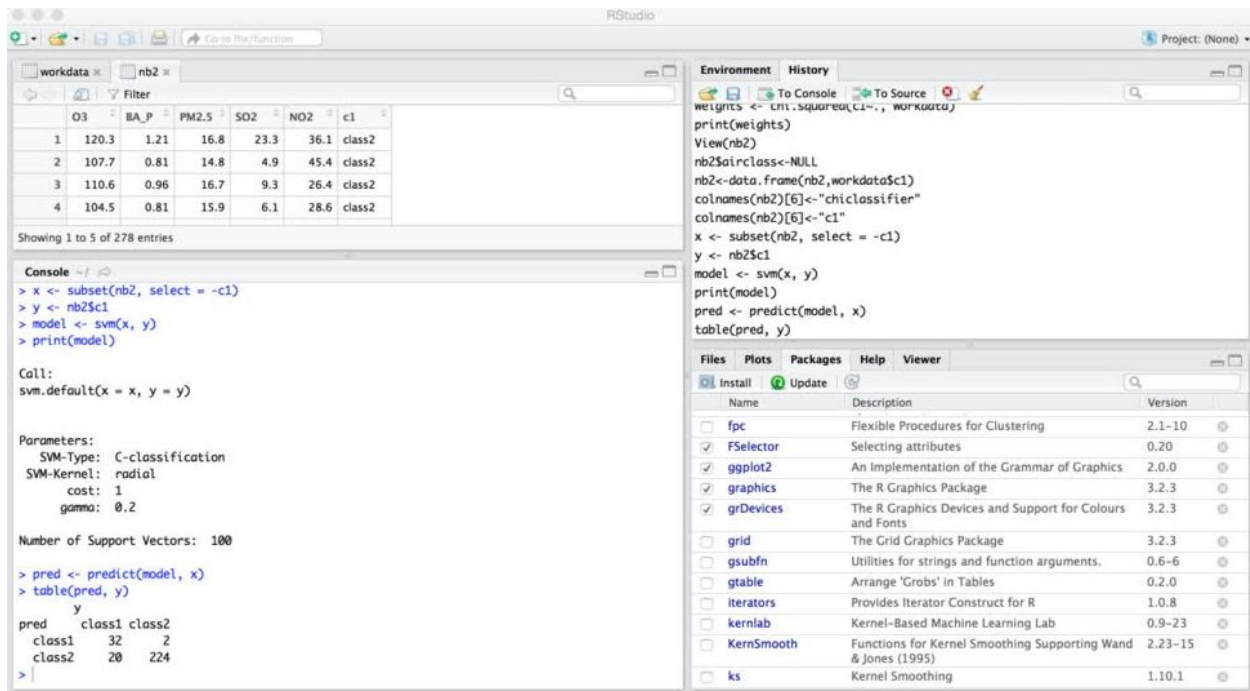


Fig. 5.6. Air Local Fusion Module - Classification using Support Vector Machine

The decision tree for Air quality parameters can be given in the below diagram Fig.5.7.

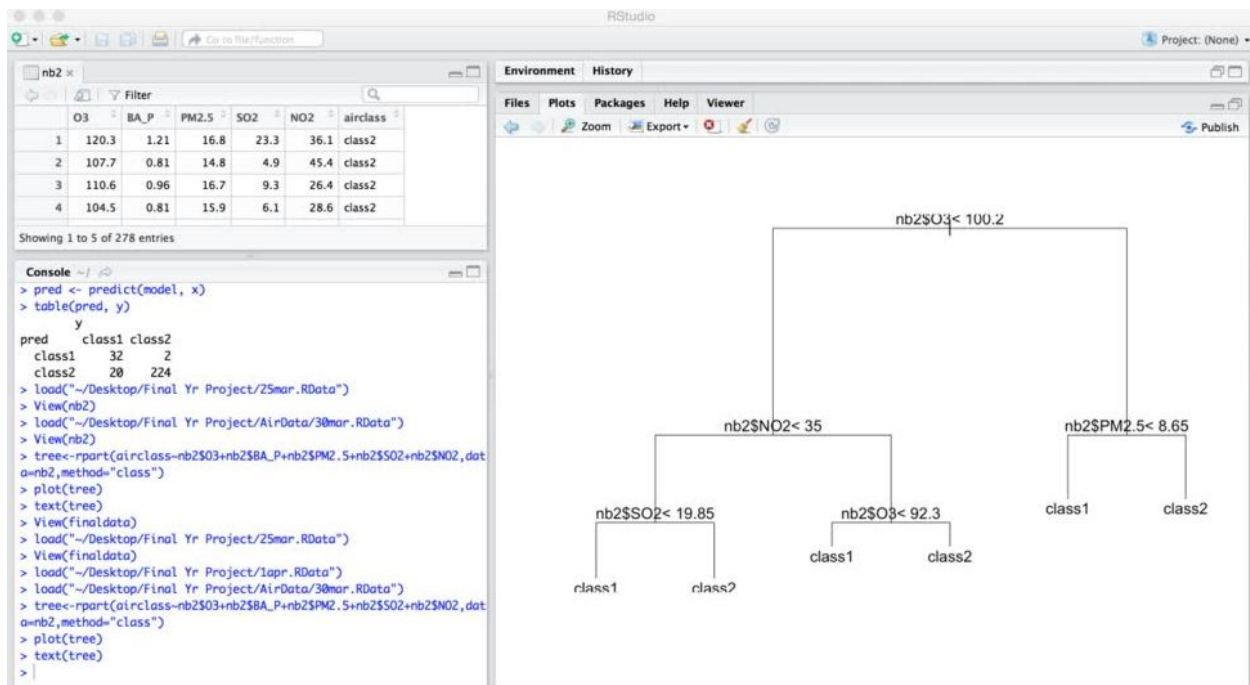


Fig. 5.7. Air Local Fusion Module - Local Decision using Decision Tree

Soil Local Fusion Module

Soil Module is very similar to the Air Module. The pre-processing step is followed by Voting System step here also, as there is no Ground Truth for this dataset obtained. Feature Extraction method used here is Recursive Feature Elimination using Random Forest, which proves to be efficient for this module as in Fig. 5.8.

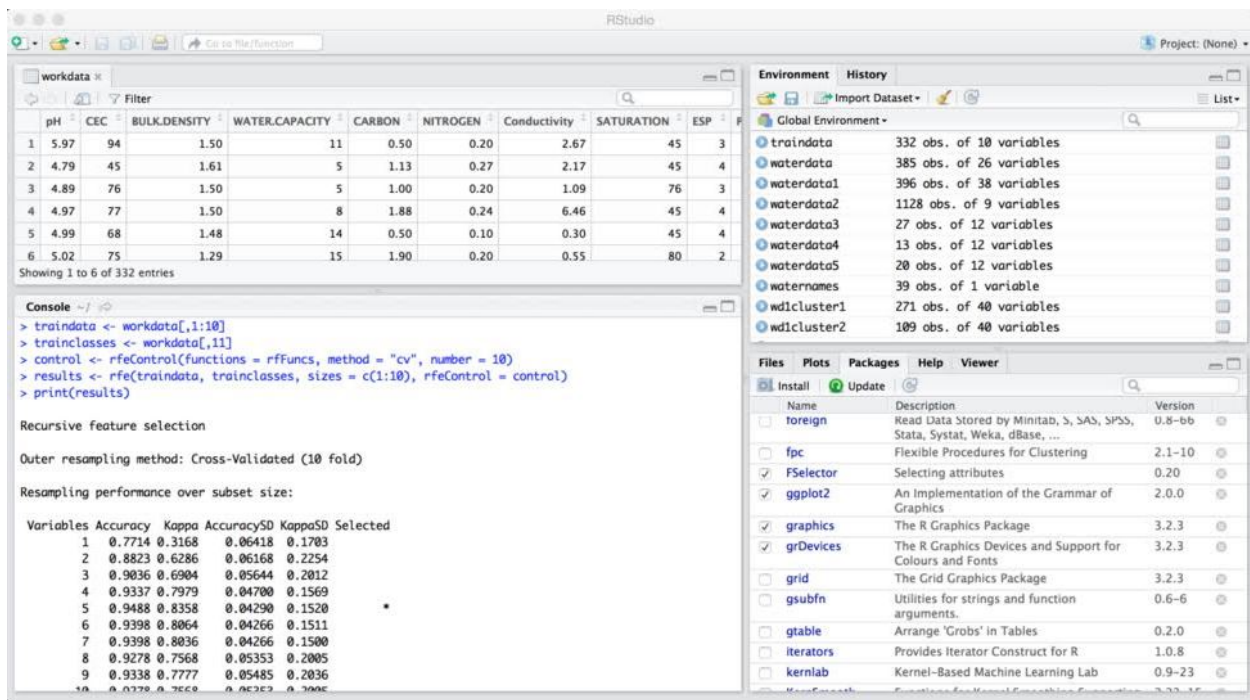


Fig. 5.8. Soil Local Fusion Module - Feature Extraction using RFE

Support Vector Machine Classification is used here for this module to classify data. Training datasets and Testing datasets are partitioned followed by classification as depicted in Fig.5.9.

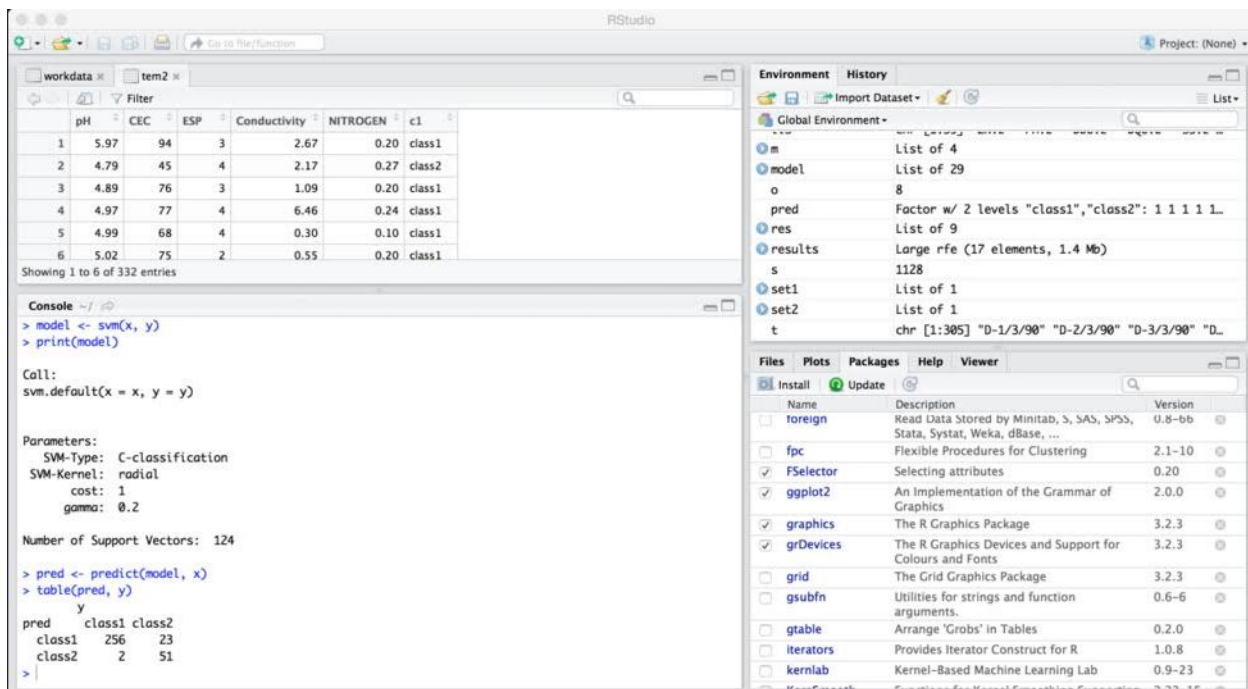


Fig. 5.9. Soil Local Fusion Module - Classification using Support Vector Machine

The decision tree for the Soil Module can be derived using the Decision Tree Learning which gives the local decision of that module as shown in Fig.5.10.

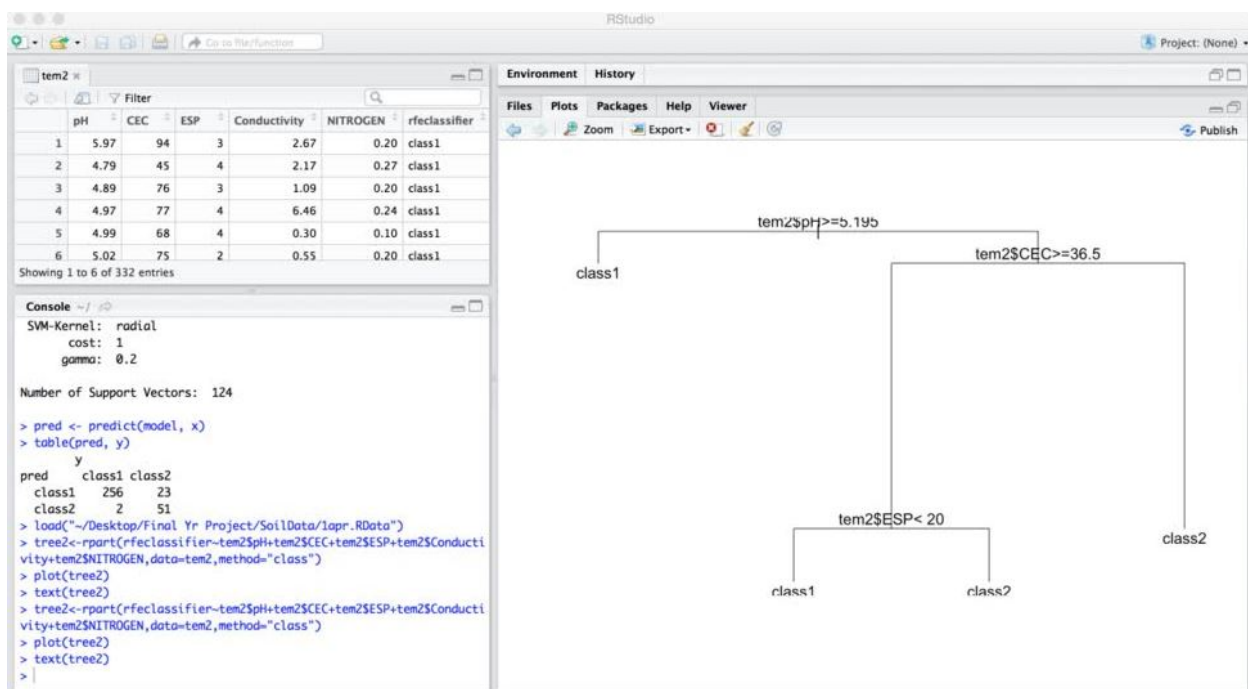


Fig. 5.10. Soil Local Fusion Module - Local Decision using Decision Tree

Central Fusion Module

The three local decisions can be given as input to the Decision Tree Learning at the Central Module to predict the final decision regarding the wholesome quality of water which is illustrated in the figure 5.11.

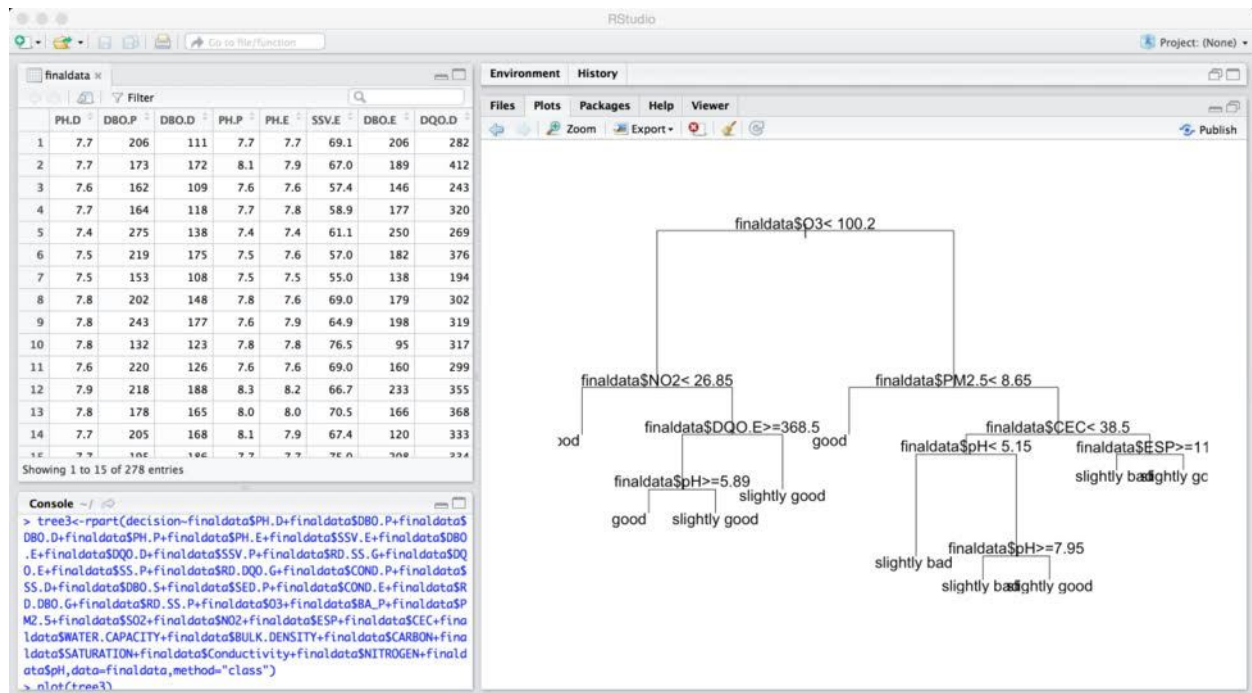


Fig. 5.11. Final Decision Tree at Central Fusion Module

CHAPTER 6

RESULTS AND DISCUSSIONS

The various experimental results that were recorded during the development of this system are presented in this chapter.

6.1 Clustering Results

K-Means clustering method was used for Water Fusion Module only to check for the Ground Truth of the dataset. The optimal number of clusters was decided by Elbow method for Clustering, which was two as given by the curve in Fig. 6.1.

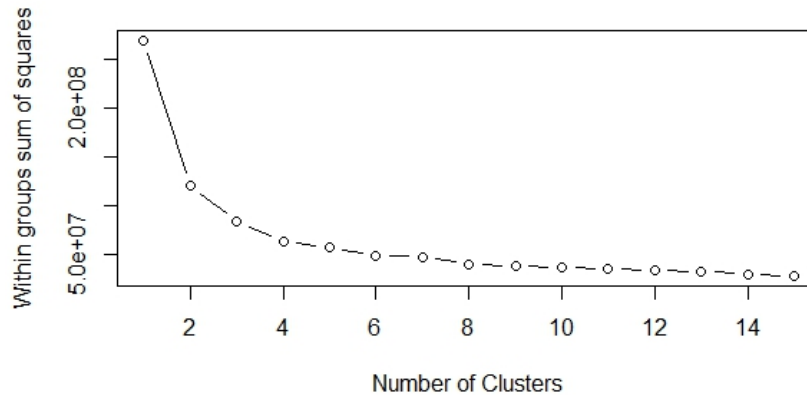


Fig. 6.1. Elbow method to find optimal clusters

After Elbow Method, K-Means method was applied and plotted for the water dataset obtained.

S.No	Name of the dataset	Number of Attributes	Number of clusters	Cluster means (For instance)
1	UCI Water Dataset	37	2 clusters Cluster1 - 271 entries Cluster2 – 109 entries	Cluster 1-pH(7.778544) Cluster 2-pH(7.927731)

Table 6.1. Clustering of UCI Water Dataset with 37 parameters and their centroids for pH

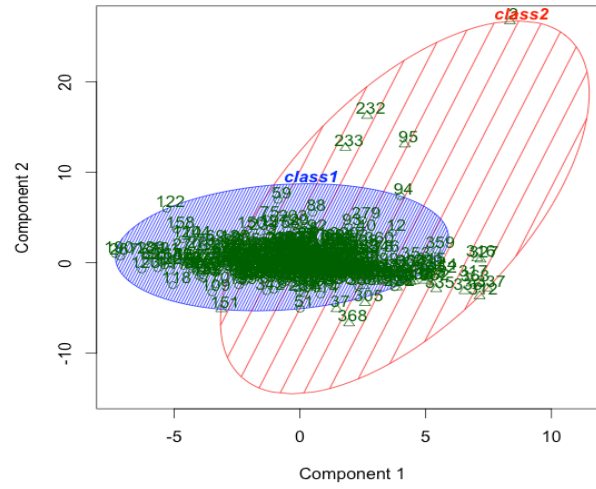


Fig. 6.2. 2D representation of cluster results for UCI Water Dataset - First two Principle components explain 35.73% of the point variability

6.2 Feature Extraction Results

A number of different Feature Extraction methods were applied, but only the best of them were chosen for further analysis. These are Chi-Squared, Random Forest and Recursive Feature Elimination using Random Forest.

All three methods were applied on all three local fusion modules and the following results were acquired.

Water

The important attributes selected by each method is given below in Fig.6.3.

```
Chi squared feature selection
[1] "PH.D"      "DBO.D"      "PH.E"      "PH.P"      "DBO.P"      "SSV.E"      "DQO.D"      "RD.DQO.G"
[9] "DBO.E"      "SSV.P"      "COND.S"     "COND.D"     "COND.E"     "DQO.E"      "COND.P"      "SED.D"
[17] "SED.P"      "SS.D"       "SED.E"      "ZN.E"

Random forest filter
[1] "PH.D"      "DBO.P"      "DBO.D"      "PH.P"      "PH.E"      "SSV.E"      "DBO.E"      "DQO.D"
[9] "SSV.P"      "RD.SS.G"     "DQO.E"      "SS.P"      "RD.DQO.G"   "COND.P"     "SS.D"       "DBO.S"
[17] "SED.P"      "COND.E"      "RD.DBO.G"   "RD.SS.P"

Automatic Feature Selection Methods using Recursive Feature Elimination (RFE) with Random Forest Algorithm
"PH.D"      "DBO.P"      "DBO.D"      "PH.P"      "PH.E"      "SSV.E"      "DQO.D"      "SSV.P"      "DBO.E"      "RD.DQO.G" "RD.SS.G"
"SS.P"      "DBO.S"      "SS.S"       "DQO.E"      "SS.D"      "COND.E"     "SED.P"      "COND.S"     "COND.P"
```

Fig. 6.3. List of Attributes given by Feature Selection methods for Water

Air

The important attributes in this module by each method are given here in Fig.6.4.

```
Chi-square
[1] "O3"      "BA_P"     "PM2.5"    "SO2"      "NO2"

Random Forest
[1] "O3"      "NO2"      "PM2.5"    "BA_P"     "SO2"      "PM10"

Union
[1] "O3"      "BA_P"     "PM2.5"    "SO2"      "NO2"      "PM10"

Intersection
[1] "O3"      "BA_P"     "PM2.5"    "SO2"      "NO2"

Random Forest RFE
[1] "O3"      "PM2.5"    "BA_P"     "NO2"      "SO2"
```

Fig. 6.4. List of Attributes given by Feature Selection methods for Air

Soil

Attributes considered significant are given here in Fig.6.5 for soil module.

Chi-square				
[1] "pH"	"CEC"	"Conductivity"	"NITROGEN"	
Random Forest				
[1] "pH"	"CEC"	"ESP"	"Conductivity"	
[5] "NITROGEN"	"WATER.CAPACITY"	"SATURATION"	"PHOSPHOROUS"	
[9] "CARBON"				
Random Forest RFE				
[1] "pH"	"CEC"	"ESP"	"Conductivity"	
[5] "NITROGEN"				

Fig. 6.5. List of Attributes given by Feature Selection methods for Soil

6.3 Local Fusion Module Decisions

Classification methods in combination with various feature extraction methods yielded in results with minute differences. The different performance metrics used are :

- Accuracy - a description of systematic errors given by the formula (True Positives + True Negatives) / (Positives + Negatives)
- Precision - Proportion of instances that are truly of a class divided by the total instances classified as that class
- Recall - proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
- F-Score - A combined measure for precision and recall calculated as $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

In order to attain higher accuracy and precision with lower number of features, the methods are analysed in detail and optimal combination is used for each module.

Water

For this module, a combination of the Random Forest Feature Extraction method and Support Vector Machine classifier gives efficient results when compared to other combinations. This is based on the values of the performance metrics listed in table 6.2.

RESULTS - WATER

Classifier used : Naive Bayes

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	37	0.9532	0.9824	0.9654	0.9484
2	Union (RandomForest + ChiSquared)	25	0.9636	0.9818	0.9759	0.9581
3	ChiSquared	20	0.9662	0.9967	0.9627	0.9595
4	RandomForest	20	0.9558	0.9849	0.9646	0.9500
5	RFE with RandomForest	20	0.9480	0.9702	0.9702	0.9412
6	Intersection (RF + Chi)	16	0.9558	0.9967	0.9504	0.9472

Table 6.2. Performance measures for Water using Naive Bayes Classification



Fig. 6.6. Classifiers Vs Accuracy for Water with Naive Bayes Classification

Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	37	0.9844	0.9883	0.9941	0.9824
2	Union (RandomForest + ChiSquared)	25	0.9844	0.9883	0.9941	0.9824
3	ChiSquared	20	0.9818	0.9826	0.9970	0.9796
4	RandomForest	20	0.9818	0.9854	0.9941	0.9795
5	RFE with RandomForest	20	0.9844	0.9883	0.9941	0.9824
6	Intersection (RF + Chi)	16	0.9740	0.9768	0.9941	0.9710

Table 6.3. Performance measures for Water using SVM Classification

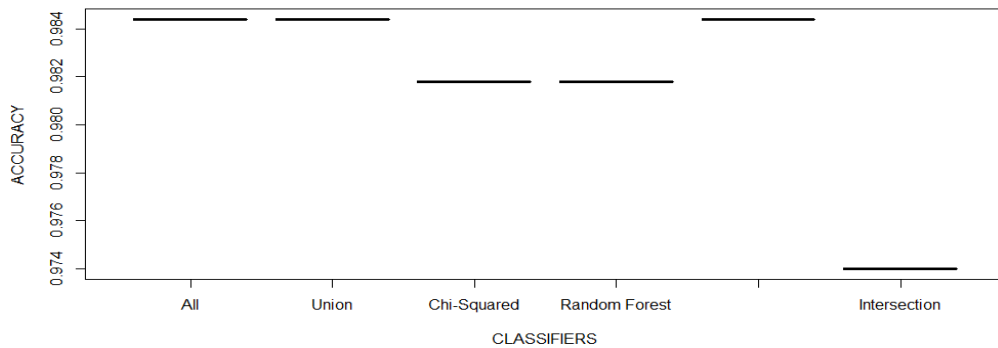


Fig. 6.7. Classifiers Vs Accuracy for Water with SVM Classification

The number of features must also be kept in mind when checking for accuracy and the methods must be chosen such that optimal results are obtained.

The decision tree for this module is given in the below diagram in Fig 6.8. This describes how records of the water dataset are classified.

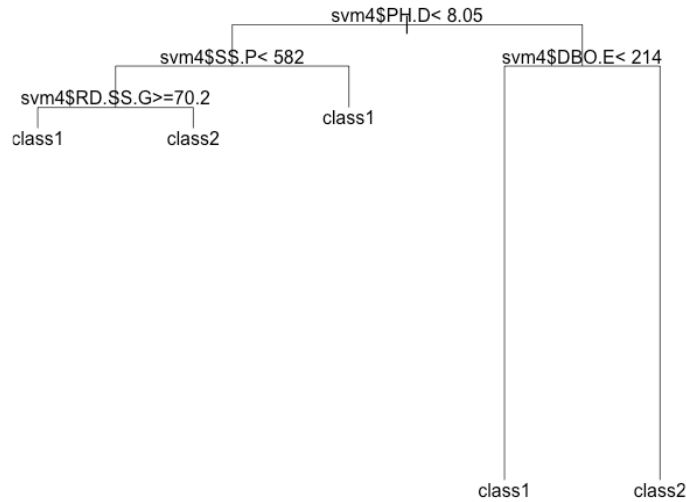


Fig. 6.8. Decision Tree - Local Decision for Water Module

Air

For this module, Chi-Squared Feature Extraction with Support Vector Machine gives the best decisions with optimally higher accuracies when compared to other combinations.

RESULTS - AIR

Classifier used : Naive Bayes

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	6	0.7589	0.8493	0.5254	0.4462
2	Union (RandomForest + ChiSquared)	6	0.7589	0.8493	0.5254	0.4462
3	ChiSquared	5	0.7841	0.8630	0.5575	0.4811
4	RandomForest	6	0.7589	0.8493	0.5254	0.4462
5	RFE with RandomForest	5	0.7841	0.8630	0.5575	0.4811
6	Intersection (RF + Chi)	5	0.7841	0.8630	0.5575	0.4811

Table 6.4. Performance measures for Air using Naive Bayes Classification

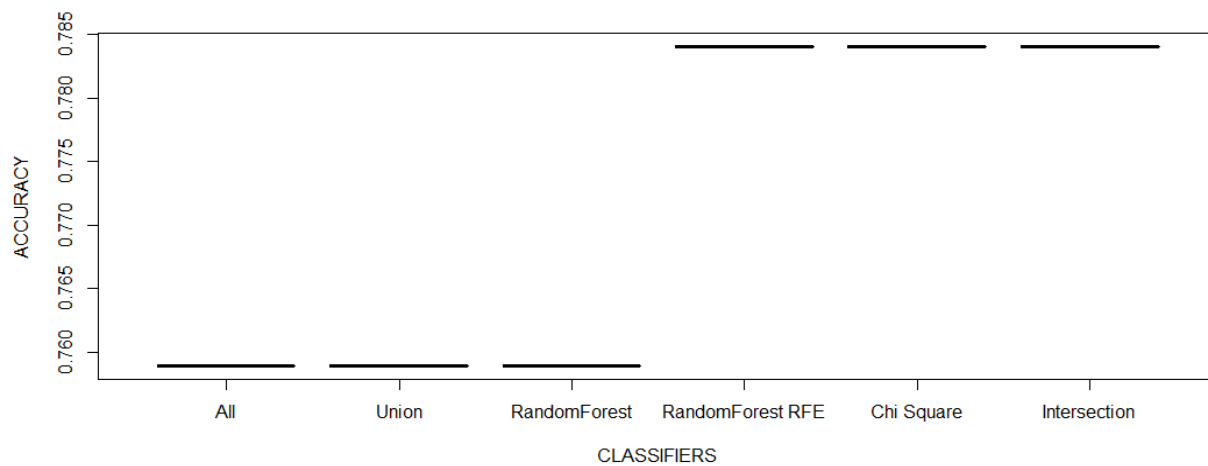


Fig. 6.9. Classifiers Vs Accuracy for Air with Naive Bayes Classification

Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	6	0.9352	0.8082	0.9365	0.7568
2	Union (RandomForest +ChiSquared)	6	0.9352	0.8082	0.9365	0.7568
3	ChiSquared	5	0.9388	0.8219	0.9375	0.7705
4	RandomForest	6	0.9352	0.8082	0.9365	0.7568
5	RFE with RandomForest	5	0.9388	0.8219	0.9375	0.7705
6	Intersection (RF + Chi)	5	0.9388	0.8219	0.9375	0.7705

Table 6.5. Performance measures for Air using SVM Classification

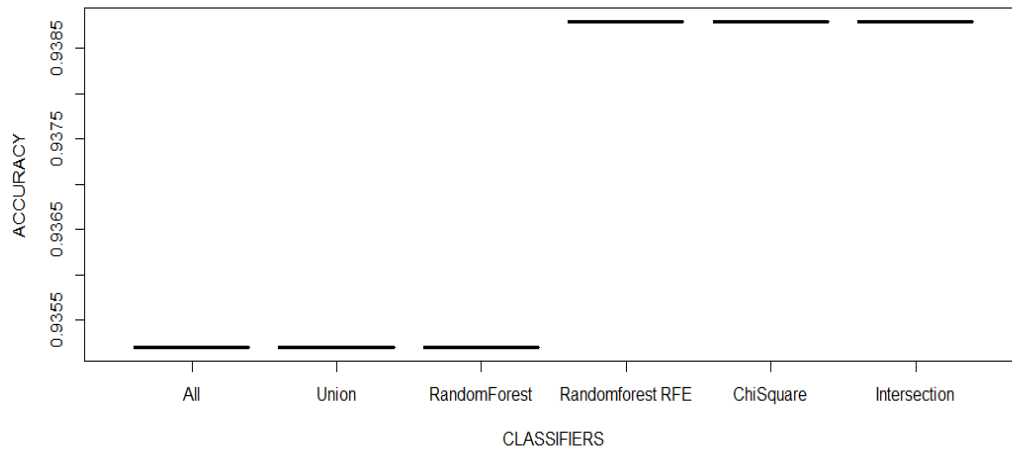


Fig. 6.10. Classifiers Vs Accuracy for Air with SVM Classification

The number of features must also be kept in mind when checking for accuracy and the methods must be chosen such that optimal results are obtained.

The decision tree for this module is given in the below diagram in Fig 6.11. This describes how records of the air dataset are classified.

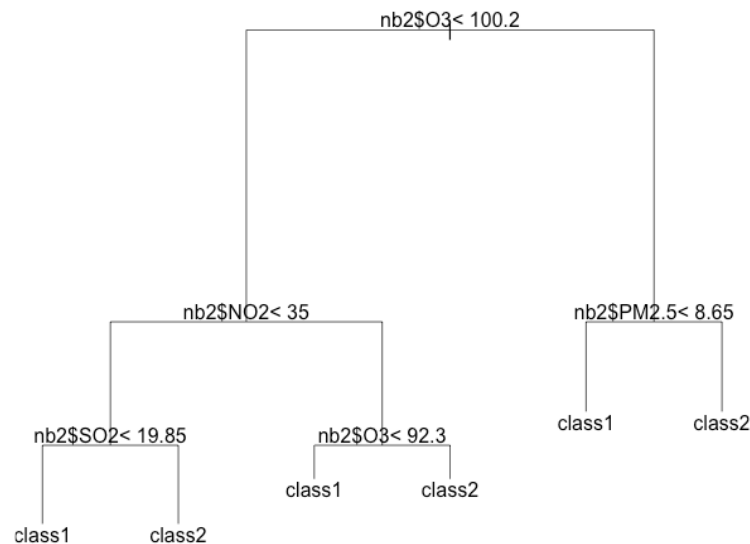


Fig. 6.11. Decision Tree - Local Decision for Air Module

Soil

For this module, Recursive Feature Elimination using Random Forest method with Support Vector Machine gives the best decisions with optimally higher accuracies.

RESULTS - SOIL

Classifier used : Naive Bayes

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	10	0.7831	0.8229	0.9186	0.7559
2	RandomForest	9	0.7740	0.8166	0.9147	0.7469
3	RFE with RandomForest	5	0.7861	0.8191	0.9302	0.7619
4	ChiSquared	4	0.8493	0.8489	0.9806	0.8324

Table 6.6. Performance measures for Soil using Naive Bayes Classification

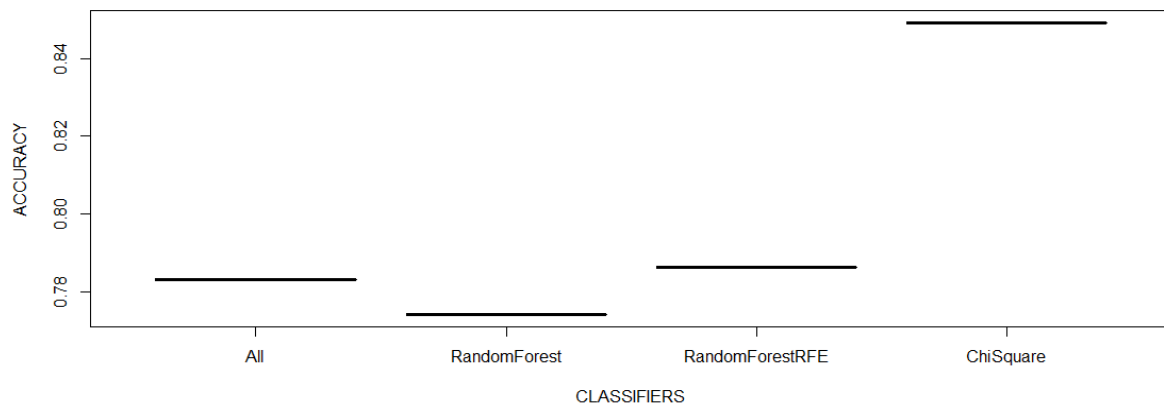


Fig. 6.12. Classifiers Vs Accuracy for Soil with Naive Bayes Classification

Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	10	0.9126	0.8989	1	0.8989
2	RandomForest	9	0.9126	0.8989	1	0.8989
3	RFE with RandomForest	5	0.9246	0.9175	0.9922	0.9104
4	ChiSquared	4	0.8825	0.8711	0.9961	0.8677

Table 6.7. Performance measures for Soil using SVM Classification

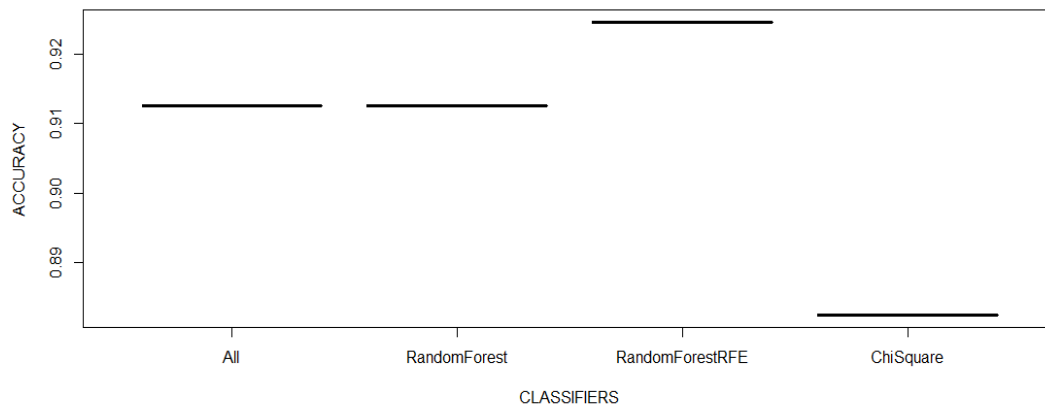


Fig. 6.13. Classifiers Vs Accuracy for Soil with SVM Classification

The number of features must also be kept in mind when checking for accuracy and the methods must be chosen such that optimal results are obtained.

The decision tree for this module is given in the below diagram in Fig 6.14. This describes how records of the air dataset are classified.

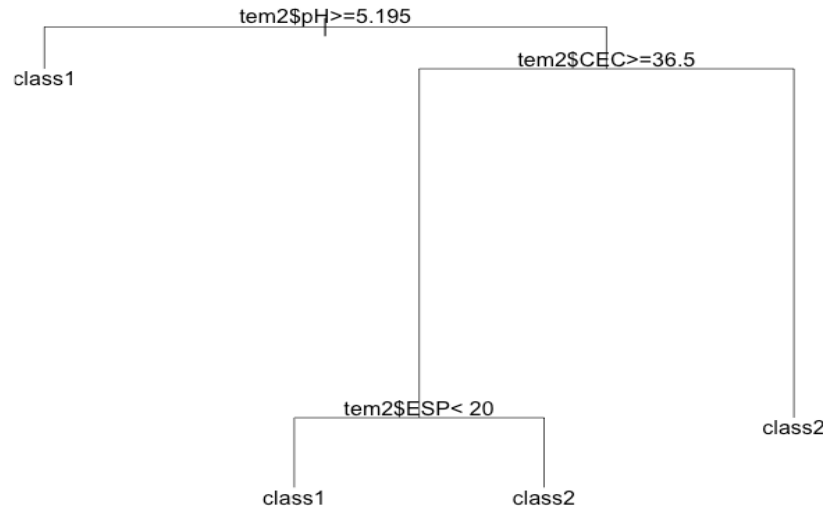


Fig. 6.14. Decision Tree - Local Decision for Soil Module

6.4 Central Fusion Module Decisions

The Decision Tree at the Central Module works based on the combination table of rules, which decides the ultimate decision about the water quality. The combination of decisions is presented below in table 6.8.

S.No	WATER	AIR	SOIL	FINAL
1	Bad	Bad	Bad	Bad
2	Bad	Bad	Good	Bad
3	Bad	Good	Bad	Bad
4	Bad	Good	Good	Bad
5	Good	Bad	Bad	Slightly Bad
6	Good	Bad	Good	Slightly Good
7	Good	Good	Bad	Slightly Good
8	Good	Good	Good	Good

Table 6.8. Combination of Local Decisions for Final Decision

The decision tree resulting in this final module yielding ultimate results regarding water quality is depicted in the Fig. 6.15.

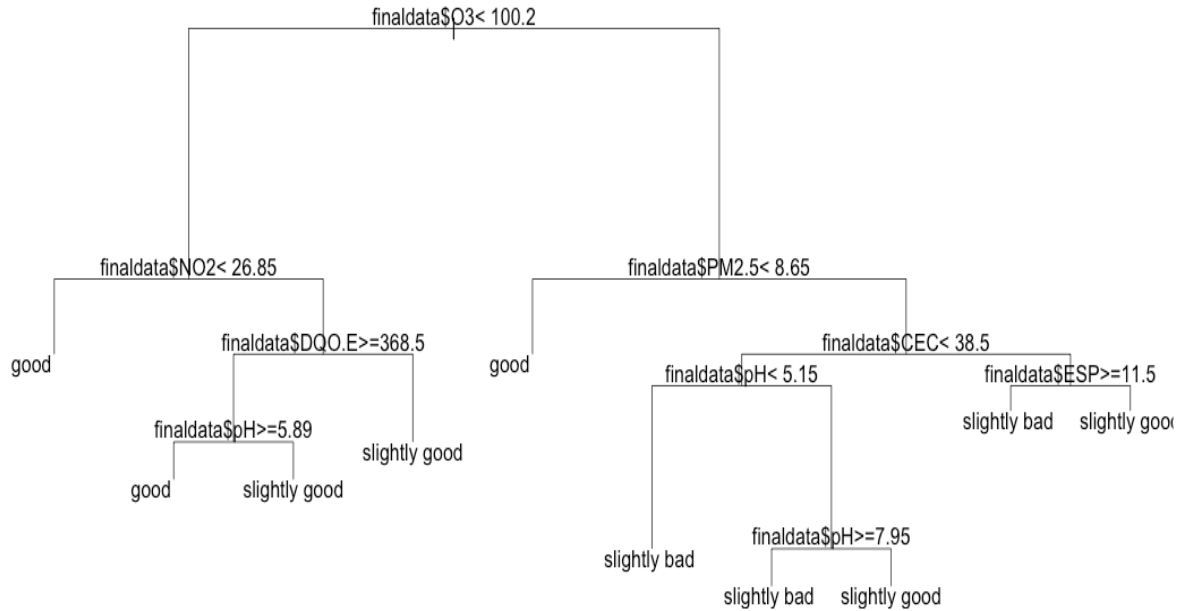


Fig. 6.15. Decision Tree - Central Decision determining water quality

The accuracy of the overall system is 95.25%. All the results in the chapter show how the data is taken in, processed at every stage and yield the desired outcomes.

6.5 Advantages of Proposed System

As opposed to existing water monitoring systems, there are several advantages to the proposed system. They are :

- This is a real time system. So, it includes working under extreme environmental conditions and supports dynamic decision making. This will have a lot of benefits on a large scale, improving the lives of people living nearby.

- Data Fusion of Water, Air and Soil parameters is a huge advantage. It covers all possible factors that influence the water quality. Contamination of river beds and ambient soil along with water deterioration can be identified immediately and actions can be taken. Fusing all the influential attributes gives precise decisions and accurate results.
- Compared to National Water Monitoring Programme (NWMP) and other existing systems which lacks real-time feature, this system provides accurate and faster decisions.
- The proposed system includes ensemble learning methods including many eclectic machine learning methods and feature extraction methods. Using the combination of Data Fusion, Analytics and Machine Learning has been a major advantage which is not present in other systems.
- The system is relatively cost-effective compared to other systems such as National Water Monitoring Programme (NWMP) and is more suited to Indian environments as it includes soil and air parameters.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

Water is a depleting resource and needs to be utilized to its fullest. There arises a need to classify water for various activities and make better use of it, rather waste it unnecessarily. Also, with the fast growing technical advances, the pollution and contamination of water is also growing at large. In fact, the paucity of clean water for domestic use has led to the increase in the number of deaths in both the urban and the rural parts of India.

Water borne diseases are the most common cause of deaths in developing countries such as India. Deaths due to water related diseases in India are nearly 76.7 percent. This needs to be mitigated immediately. This thought kindled us to take an initiative as engineers to save water resources. We felt that application and knowledge of Data Analytics can be put to better use as a contribution to the environment.

Incorporation of Data Fusion ideas in water monitoring is a fresh concept which is expected to grow greater heights in the future.

7.2 Future Work

This project could be further expanded by using a sensor network to acquire data from rivers. Usage of various sensors such as pH sensors, temperature sensors and Dissolved Oxygen sensors at various zones of the rivers can be used. Further, various other feature extraction methods and classification methods can be implemented giving more extensive classes of usage. It can also be upgraded to support any type of water source in the future.

REFERENCES

1. Bhardwaj, R. M. "Water quality monitoring in India—achievements and constraints." *IWG-Env, International Work Session on Water Statistics, Vienna* (2005): 1-12.
2. Byer, David, and Kenneth H. Carlson. "Expanded summary: Real-time detection of intentional chemical contamination in the distribution system." *Journal (American Water Works Association)* 97.7 (2005): 130-133.
3. Domingos, Pedro. "A few useful things to know about machine learning." *Communications of the ACM* 55.10 (2012): 78-87.
4. Han, Jiawei, Micheline Kamber, and Jian Pei. "*Data mining: concepts and techniques*". Elsevier, 2011.
5. Karami, Ebrahim, Francis M. Bui, and Ha H. Nguyen. "Multisensor data fusion for water quality monitoring using wireless sensor networks." *Communications and Electronics (ICCE), 2012 Fourth International Conference on*. IEEE, 2012.
6. Nakamura, Eduardo F., Antonio AF Loureiro, and Alejandro C. Frery. "Information fusion for wireless sensor networks: Methods, models, and classifications." *ACM Computing Surveys (CSUR)* 39.3 (2007): 9.
7. Pechenizkiy, M., Puuronen, S. and Tsymbal, A., 2003. "Feature extraction for classification in the data mining process."
8. Smith, Richard A., Gregory E. Schwarz, and Richard B. Alexander. "Regional interpretation of water- quality monitoring data." *Water resources research* 33.12 (1997): 2781-2798.
9. Standard, Indian. "Drinking water-specification." *1st Revision, IS 10500* (1991).

10. Stoianov, Ivan, et al. "Sensor networks for monitoring water supply and sewer systems: Lessons from Boston." *Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium*. 2006.
11. "5 Major Causes of Water Pollution in India",
<http://www.yourarticlelibrary.com/water-pollution/5-major-causes-of-water-pollution-in-india/19764>
12. "Chi-Square Test of Independence",
<http://www.stat trek.com/chi-square-test/independence.aspx?Tutorial=AP>
13. "Data Mining - Decision Tree Induction",
http://www.tutorialspoint.com/data_mining/dm_dti.htm
14. "Decision Trees",
<http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf>
15. "Feature Extraction Models",
http://topepo.github.io/caret/Feature_Extraction.html
16. "Feature Selection with the Caret R Package",
<http://machinelearningmastery.com/feature-selection-with-the-caret-r-package/>
17. "Naive Bayesian",
http://www.saedsayad.com/naive_bayesian.htm
18. "National Water Quality Monitoring Programme",
www.cpcb.nic.in/divisionsofheadoffice/pams/NWMP.pdf
19. "Pearson's Chi-Square Test for Independence",
<http://www.ling.upenn.edu/~clight/chisquared.htm>
20. "RandomForest",
https://en.wikipedia.org/wiki/Random_forest
21. "Random Forests:some methodological insights",
<http://arxiv.org/pdf/0811.3619.pdf>
22. "Support Vector Machines (SVM) Introductory Overview",
<http://www.statsoft.com/textbook/support-vector-machines>

23. “Support Vector Machines”,
<http://scikit-learn.org/stable/modules/svm.html>
24. “The k-means clustering algorithm”,
<http://cs229.stanford.edu/notes/cs229-notes7a.pdf>
25. “Water Treatment Plant Dataset”,
<http://archive.ics.uci.edu/ml/machine-learning-databases/water-treatment/>
26. “Air pollutant concentrations 2013”,
<http://www.eea.europa.eu/data-and-maps/data/air-pollutant-concentrations-at-station/pollutant-concentrations-by-city/air-pollutant-concentrations-2013-dataset-cities>
27. “ISRIC/WDC-Soil Dataset”
http://www.isric.org/content/download-form?dataset=SOTWIS_BR.zip
28. “Package ‘caret’”,
<https://cran.r-project.org/web/packages/caret/caret.pdf>
29. “Package ‘cluster’”,
<https://cran.r-project.org/web/packages/cluster/cluster.pdf>
30. “Package ‘e1071’”,
<https://cran.r-project.org/web/packages/e1071/e1071.pdf>
31. “Package ‘FSelector’”,
<https://cran.r-project.org/web/packages/FSelector/FSelector.pdf>
32. Package ‘mlbench’”,
<https://cran.r-project.org/web/packages/mlbench/mlbench.pdf>
33. Package ‘randomForest’”,
<https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
34. “Package ‘rpart’”,
<https://cran.r-project.org/web/packages/rpart/rpart.pdf>
35. “Package ‘stats’”,
<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html>