

# **WATER QUALITY PREDICTION MODEL USING DATAMINING TECHNIQUES**

**A PROJECT REPORT**

*Submitted By*

**AMIRTHA VARSHINI A(810015104701)**

**THILSHATH S (810015104719)**

*In partial fulfillment of the award of the degree  
of*

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND ENGINEERING**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**UNIVERSITY COLLEGE OF ENGINEERING – BIT CAMPUS**

**ANNA UNIVERSITY TIRUCHIRAPPALLI - 620 024**

**ANNA UNIVERSITY : CHENNAI- 600 025**

**APRIL 2019**

# **ANNA UNIVERSITY : CHENNAI-600 025**

## **BONAFIDE CERTIFICATE**

Certified that this project report "**WATER QUALITY PREDICTION MODEL USING DATAMINING TECHNIQUES**" is the bonafide work of Ms. **A.AMIRTHAVARSHINI (810015104701)** and Ms. **S.THILSHATH (810015104719)** who carried out the work under my supervision

**SIGNATURE**

**DR. D. VENKATESAN**

**HEAD OF THE DEPARTMENT**

Professor & Head

Department of Computer Science

University College of Engineering

BIT Campus

Tiruchirappalli – 620024

*Senthilkumar*  
*1/4/2019*

**SIGNATURE**

**DR.D. SENTHILKUMAR**

**PROJECT GUIDE**

Assistant Professor

Department of Computer Science

University College of Engineering

BIT Campus

Tiruchirappalli - 620024

Certified that **A.AMIRTHA VARSHINI** and **S.THILSHATH** was examined in a Project Viva Voice examination held on

---

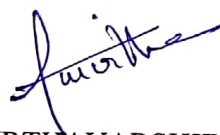
**Internal Examiner**

**External Examiner**

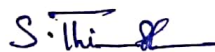
## DECLARATION

We hereby declare that the work “**WATER QUALITY PREDICTION MODEL USING DATAMINING TECHNIQUES**” is submitted in partial fulfillment of the requirement for the award of the degree in B.E., University College of Engineering (BIT Campus), Tiruchirappalli is a record of own work carried out by us during the academic year 2018-2019 under the supervision and guidance of **Dr.D.SENTHILKUMAR**, Assistant Professor, Department of Computer Science and Engineering, University College of Engineering (BIT Campus), Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any other degree or diploma, either in this or any other universities.

SIGNATURE OF THE CANDIDATES

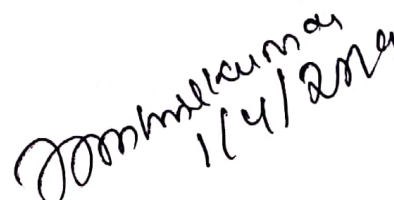


A.AMIRTHAVARSHIINI(810015104701)



S.THILSHATH(810015104719)

I certify that the declaration made above by the candidates is true.



SIGNATURE OF THE GUIDE

**Dr.D.SENTHILKUMAR**

Assistant Professor

Department of Computer Science and Engineering

University College of Engineering (BIT Campus),

Tiruchirappalli-620 024.

## **ACKNOWLEDGMENT**

We would like to thank the Almighty for all the blessings he bestowed on us, which drove us to the successful completion of this project.

We would like to extend our heartfelt gratitude to our respected Dean of Anna University-BIT Campus, Tiruchirappalli **Prof. Dr. T. SENTHILKUMAR**, who is the guiding light for all the activities in our college.

We would like to express our special thanks to our beloved Head of the Department **Dr. D. VENKATASAN**, Head of Department/CSE for his kind guidance towards the success of this project.

We would like to thank and express our deep sense of gratitude to our project Guide **Dr. D. SENTHILKUMAR**, Assistant Professor, Department of Computer Science and Engineering, for his valuable guidance, encouragement and constant support throughout our work.

We also thank all the teaching and non-teaching staffs of the Department of CSE, our beloved parents and friends, for their help and support to complete our project successfully.

**AMIRTHA VARSHINI A  
THILSHATH S**

## **ABSTRACT**

Water quality is the measure of chemical, physical, biological and radiological characteristics of water. The factors that influence water quality are water temperature, pH, specific conductance, turbidity, dissolved oxygen, salinity, hardness and suspended sediments. The quality of water decides whether the water is suitable for aquatic organisms, human consumption and other water based activities. Predicting the quality of water is a very important issue in an ecosystem and it can be used to control the increase of water contamination. The quality of water is not accurately predicted using existing techniques. Therefore, in this project, two techniques namely CART (Classification And Regression Trees) and MARS (Multivariate Adaptive Regression Splines) are proposed to predict the water quality. Experimental results show that the proposed methods CART and MARS is able to predict the future values of the variables more precisely when compared with the existing techniques. This project will be helpful for the authorities, to take necessary precautionary measures to control the pollution level.

## திட்டச்சுருக்கம்

நீர் தரம் இரசாயனத்தின் வேதியியல், உடல், உயிரியல் மற்றும் கதிரியக்க குணங்களின் அளவாகும். நீரின் தரத்தை பாதிக்கும் காரணிகள் நீர் வெப்பநிலை, பிஎச், குறிப்பிட்ட கடத்துதல், குழப்பம், கரைந்த ஆக்ஸிஜன், உப்புத்தன்மை, கடினத்தன்மை மற்றும் இடைநீக்கம் செய்யப்பட்ட வண்டல்கள். நீரின் தரமானது நீர்வாழ் உயிரினங்கள், மனித நுகர்வு மற்றும் பிற நீர் சார்ந்த நடவடிக்கைகள் ஆகியவற்றிற்கு நீர் ஏற்றதா என்பதை முடிவு செய்யும். நீர் தரத்தை முன்னறிவித்தல் என்பது ஒரு சுற்றுச்சூழலில் மிகவும் முக்கியமான சிக்கலாகும், மேலும் அது நீர் மாசுபாடு அதிகரிப்பதை கட்டுப்படுத்த பயன்படுகிறது. நீரின் தரம் தற்போதுள்ள நுட்பங்களைப் பயன்படுத்தி துல்லியமாக கணிக்கப்படவில்லை. எனவே, இந்த திட்டத்தில், CART (கிளாசிஃபிகேஷன் அண்ட் ரிக்ரஷன் ட்ரீஸ்) மற்றும் MARS (மல்டிவேரியேட் அடாப்டிவ் ரிக்ரஷன் ஸ்ப்லைன்ஸ்) ஆகிய இரண்டு உத்திகள் நீர் தரத்தை முன்னறிவிக்க முன்மொழிகின்றன. பரிசோதனை முடிவுகள், முன்மொழியப்பட்ட முறைகள் CART மற்றும் MARS ஆகியவை ஏற்கனவே இருக்கும் நுட்பங்களுடன் ஒப்பிடும் போது எதிர்கால மதிப்பீடுகளை இன்னும் துல்லியமாக கணிக்கின்றன. இந்த திட்டம் மாசு அளவை கட்டுப்படுத்த தேவையான முன்னெச்சரிக்கை நடவடிக்கைகள் எடுக்க, அதிகாரிகள் உதவியாக இருக்கும்.

## TABLE OF CONTENTS

S.NO	TITLE	PAGE NO
	<b>ABSTRACT (ENGLISH)</b>	v
	<b>ABSTRACT (TAMIL)</b>	vi
	<b>LIST OF TABLES</b>	x
	<b>LIST OF FIGURES</b>	xi
	<b>LIST OF ABBREVIATION</b>	xii
<b>1.</b>	<b>INTRODUCTION</b>	1
	1.1 Overview	1
	1.2 Problem description	1
	1.3 Purpose of this project	2
	1.4 Existing techniques	3
	1.5 Contribution of this project	3
	1.6 Organization of the project	4

<b>2.</b>	<b>LITERATURE SURVEY</b>	5
	2.1 Review of water quality prediction(with andro dataset)	5
	2.2 Review of water quality prediction(using of other datasets)	7
	Summary	10
<b>3.</b>	<b>PROPOSED SYSTEM</b>	11
	3.1 Proposed techniques	11
	3.1.1 Cart	12
	3.1.2 Mars	12
	3.2 System module	14
	3.2.1 Module 1 dataset collection	14
	3.2.2 Module 2 dataset collection	15
	3.2.3 Module 3 Implementation and execution	16
	3.2.3.1 Cart	16
	3.2.3.2 Mars	17
	3.2.4 Module 4 Algorithm comparison	17



3.2.4.1 Serial predictions	18
3.2.4.2 Parallel predictions	18
3.2.5 Module 5 Performance evaluation	19
3.2.5.1 Comparison of performance evaluation metrics	20
<b>4. REQUIREMENT SPECIFICATIONS</b>	<b>23</b>
4.1 Hardware and Software Specifications	23
4.1.1 Hardware specifications	23
4.1.2 Software specifications	23
<b>5. EXPERIMENTAL RESULTS AND DISCUSSIONS</b>	<b>24</b>
5.1 Data collection module	24
5.2 Comparison with existing techniques	28
<b>6. CONCLUSION</b>	<b>38</b>
<b>REFERENCES</b>	<b>39</b>

## LIST OF TABLES

TABLE NO	NAME OF THE TABLE	PAGE NO
5.1	Water Quality Dataset description	24
5.2	Comparison of R-Squared values	25
5.3	Comparison of Error values	26
5.4	Comparison of average values of CART and MARS	27
5.5	R-Squared values of existing techniques	28
5.6	Comparison of average R-Squared values	29
5.7	Time difference of serial vs. parallel computation	36

## LIST OF FIGURES

FIGURE NO	NAME OF THE FIGURE	PAGE NO
3.1	Proposed Techniques	11
3.2	System Architecture	13
3.3	Data collection	14
3.4	Data preprocessing	15
5.1	Comparison of MSE	30
5.2	Comparison of RMSE	31
5.3	Comparison of R-Squared values	32
5.4	Comparison of Standard Error	33
5.5	Comparison of MARS and CART	34
5.6	Comparison of avg. R-Squared values with existing techniques	35
5.7	Comparison of time difference in MARS and CART	36

## LIST OF ABBREVIATIONS

CART	Classification And Regression Trees
MARS	Multivariate Adaptive Regression Splines
SMO	Sequential Minimal Optimization
SLR	Simple Linear Regression
IB $k$	Instance Based
M5P	M5 Pruning
MLP	Multi Layer Pruning
RW	Random Walk
$k$ NN	Nearest Neighbor
PSO	Particle Swarm Optimization
R-sq	R-square value
Std.error	Standard error
MSE	Mean squared error
RMSE	Root mean square error
AR	Auto Regressive
MI	Multivariate Model

## **CHAPTER 1**

### **INTRODUCTION**

#### **1.1 Overview**

Water quality is defined by the physical, chemical, biological and radiological characteristics of water. It is the measure of the condition of water with respect to the requirements of the biotic species and or any human need or purpose. The term “Water Quality” is used to express the suitability of water to sustain various uses.

Many factors affect the quality of water. The various factors that influence the quality of water are sedimentation, runoff , erosion , Dissolved Oxygen, pH , Temperature , Decayed organic materials , Pesticides , Toxic and Hazardous substances , oils , grease , Detergents , litter, rubbish and other chemicals. The key factors affecting the water quality are water temperature, pH, specific conductance, turbidity, dissolved oxygen, salinity, hardness and suspended sediments. [1][3]

This chapter discusses the motivation for the design of efficient prediction algorithm for Water Quality datasets using data mining techniques. It also presents various issues and the major contributions of this project. Finally the structure of this project is outlined.

#### **1.2 Problem description**

One of the important issues is the ability to predict one or more days ahead the quality of water in an ecosystem. Deploying such a predictive model will be able to predict an increase of the pollution levels in the sea water and provide alerts

for the precaution measures. Water quality prediction can be used in some major fields, such as irrigation and piscicultures. [4]

Water Quality plays a key role in environmental monitoring. Poor quality of water affects not only aquatic organisms, but also the surrounding ecosystem as well. The need for water quality predictions urges from the fact that, distinct compositions of water serves different purposes.

### **1.3 Purpose of this project**

The quality of water is analyzed in order to predict whether the water is suitable for various industrial and domestic purposes. The quality of water decides whether the water is suitable for aquatic organisms, human consumption, and other water based activities. The water quality prediction is mainly used for monitoring purposes. The water quality analysis is used to verify whether the water is suitable for the expected use or not. Using the perfect composition and quality of water ensures the yield of best results. [5]

The water quality analysis provides best advices on the precise use of water. For example, in industries, hard water is not suited for boilers. Hard water decreases the boiler efficiency, produces corrosion, foam, limescale deposits and can even lead to bursting of boilers. It also increases the amount of energy used and thereby increases the cost of production. As considered in household, hard water does not produce foam on washing the clothes with detergents. Clothes washed in hard water are harsh and scratchy. Therefore, estimation and prior prediction of water quality is very important.

## 1.4 Existing techniques

The quality of water is predicted using several data mining techniques. Some of the existing forecasting techniques include the Simple Linear Regression (SLR), Sequential Minimal Optimization (SMO) , IB $k$  (Instance Based) , M5P (M5 Pruning) , MLP (Multi Layer Pruning) , RW (Random Walk),  $k$ NN (Nearest Neighbor) etc. [2] [4] [5]

The IB1, IB3 and MLP are not strongly affected by the increase of the lead values. But in RW and SLR, their errors linearly increase with increasing values of lead. The error of SMO and M5P has an increasing value of lead too, but with a smaller rate compared to RW and SLR. Among the different learning algorithms, the nearest neighbor classifier achieved the best overall performance. Predicting the quality of water at the early stages reduces the risk of water contamination. Prediction of the quality of water several days before leads to the utmost usage of the water.

## 1.5 Contribution of this project

This project deals with the effective prediction of future values of the factors influencing the water quality. The CART (Classification And Regression Trees) and the MARS (Multivariate Adaptive Regression Splines) algorithm are used in this project. The CART and MARS are effective Regression analysis techniques. The Regression analysis is used to predict the values of continuous attributes. Since the data used is a daily measurement of the factors affecting the water quality, the data used is a continuous data. Therefore the two effective Regression techniques are considered. Furthermore, the targets of the dataset are processed in both serial and parallel manner to analyze the performance.

## **1.6 Organization of the report**

- Chapter 1 starts with a brief introduction about the Water quality prediction, its purpose and existing techniques.
- Chapter 2 briefly discusses the literature review related to this project.
- Chapter 3 presents the proposed work of this project and is detailed explanation.
- Chapter 4 shows software and hardware require in this project.
- Chapter 5 presents the experimental results and its related discussions. It is then compared with the results of existing techniques.
- Chapter 6 discuss end of this project with a comprehensive summary, conclusion and result of this project.



## **CHAPTER 2**

### **LITERATURE SURVEY**

This section deals with the discussion of literature review related with the current work. Initially, the techniques are used for the Water Quality data is discussed. Then the review about water quality prediction is approached.

#### **2.1 REVIEW OF WATER QUALITY PREDICTION (WITH ANDRO DATASET)**

E.hatzikos et al. [1] addressed the problem of predicting the future values for a number of water quality variables in “An Empirical Study on Sea Water Quality Prediction”. It investigates the ability to predict future values for a varying number of days ahead. The data is collected based on the measurements from underwater sensors. It performed exploratory analysis using various linear and non-linear modeling methods. It used Simple Linear Regression (SLR), Sequential Minimal Optimization (SMO), Instance Based (IBk), and M5 Pruning (M5P). The result of this analysis showed that the machine learning algorithms help in accurate predictions several days ahead. The nearest neighbor classifier achieved the overall best performance among different learning algorithms.

E.Hatzikos et al. [2] discussed “An Intelligent System For Monitoring And Predicting Water Quality” for decision making process to fight against pollution of the aquatic environment. Two sensor-telematic networks for collecting water quality measurements are deployed. The intelligent system monitors sensor data, reasons, using fuzzy logic, about the current level of water suitability for various

aquatic uses, such as swimming and piscicultures, and flags out appropriate alerts. It employs Machine Learning and Adaptive Filtering Techniques for more precise prediction one day ahead and is better than naïve prediction that the value is similar to today.

E. Hatzikos et al. [3] addressed a solution for “Applying Adaptive Prediction to Sea-Water Quality Measurements” using Projection Algorithm and Least Squares Algorithm for predicting water quality a day ahead. It explores the possibility of using adaptive filtering to predict water quality indicators such as temperature, pH, oxygen etc. The results indicate that the measurements remain reasonably stationary. The variable temperature is more precise for prediction. Substantially better predictions can be obtained for temperature factor. The value of certain quality variable the next day is equal to the value of today.

Ioannis Vlahavas et al. [4] presented a solution for the application of neural networks in “Applying Neural Networks With Active Neurons To Sea-Water Quality Measurements”. This is used to predict a number of water quality variables produced by a under-water measurement set-up is possible. It applies Active Neural Networks for one step ahead prediction models for water temperature, pH, amount of dissolved oxygen and turbidity. A Variety of linear and non-linear modeling techniques are applied in the studies. The predicted values are compared against a Random Walk model which serves as a bench mark model in prediction tasks. Active neurons are performing better than the random walk model. Neural networks with active neurons are chosen because they do not require a large number of training data and they perform better for under determined and noisy tasks.

Grigorios Tsoumakas et al. [5] suggested “Ensemble Selection For Water Quality Prediction” which applies Greedy Ensemble Selection. This actively performs evaluation of a dataset using forward , backward and active neural networks. Experimental comparisons of various parameters are performed on a application domain. It uses two parameters of the algorithm. The direction of the algorithm (forward, backward) and the performance evaluation dataset(training set, validation set). Using a separate unseen set for evaluation leads the algorithm to improve its performance.

Nick Bassiliades et al. [6] presented “Monitoring Water Quality Through a Telematic Sensor Network and a Fuzzy Expert System” in which the data is processed in a Expert System. This is used to determine the suitability of water for various purposes like the swimming, shell-culture, pisiculture and also alerts the user. The expert system aims at “decision making” to battle against pollution of aquatic environment. The expert system is equally flexible and extensible. A new sensor for a variety of environmental readings can be easily added to the system.

## **2.2 REVIEW OF WATER QUALITY PREDICTION (USING OTHER DATASETS)**

Dzeroski et al. [7] suggested a solution for the prediction of chemical parameters of river water in “Predicting chemical parameters of River water quality from Bio indicator Data”. This system focuses on inferring the chemical parameters of river water quality from biological parameters. Machine Learning techniques, in particular regression tree inductors are made use for his inference. The Regression trees predict values of the chemical parameters from data on the

presence of Bio indicator taxa at the species and family levels. The experiments indicate that the ammonia concentration, biological oxygen demand and chemical oxygen demand can be predicted relatively from bio indicator data. A properly aggregated species level data with family level data does not improve the predictive performance of regression trees. In some chemical parameters, cumulative effects are more pronounced and their average values are easier to predict. For other parameters the maximum and minimum values maybe more relevant.

Chau [8] presented “A split-step PSO algorithm in prediction of water quality pollution” to predict the water quality pollution. It allows the stake holders to have more float time to take precautionary and predictive measures. The accuracy of the prediction of water quality pollution is important. Various existing techniques apply exogenous input and different algorithms, but the Artificial Neural Networks has the ability to be a cost-efficient solution. This system applies the split-step Particle Swarm Optimization (PSO) for training the perceptrons in forecasting real time algal blooms. The PSO algorithm and Levenberg-Marquardt algorithm are combined together. On comparison with the benchmark backward propagation and the usual PSO algorithm, it attains a higher accuracy in a shorter time. The chlorophyll-a output from the 1 week time-lagged chlorophyll-a input is a effective forewarning and decision-support tool. The split-step PSO-based perceptron outperforms the other commonly used optimization techniques in algal bloom prediction.

Longqin et al.[9] discussed “Short-term water quality prediction model based on wavelet neural network” to improve the prediction accuracy. The data is proposed in a intensive freshwater pearl breeding ponds. On comparison with the BP neural network and Elman neural network,a high learning speed, improve predict accuracy, and strong robustness. The experiments indicate that the solar radiation, water temperature, dissolved oxygen, pH, humidity and wind speed. Wavelet neural network is based on topology and structure of the BP neural network. The water quality prediction based on the wavelet neural network algorithm, it can fit the complex nonlinear relationship between the ecological environment factors and dissolved oxygen. To improved prediction of the wavelet neural network algorithm to correspond the real intensive freshwater pearl aquaculture water quality prediction. The water quality is heavily affected by hydrological and meteorological factors. To establishing different predictive models according to the weather conditions and combining the prediction model to improve the prediction accuracy.

Guohua Tan et al. [10] presented “Prediction of water quality time series data based on least squares support vector machine” to predict the least squares support vector machine parameters. To applied the river water quality measurement data, after training the LS-SVM model for predicting the water quality monitoring systems, under the BP neural network and RBF network prediction. The least squares support machine method is better than multi-layer BP and RBF neural network, to the requirements of water quality prediction. The prediction model that the LS-SVM based water quality prediction model, root mean square error and mean relative error than the BP network method and RBF

network method is smaller, indicate that the LS-SVM method has a high prediction accuracy model and more applicable to the real-time water quality.

## **Summary**

The prediction of the future values of the factors affecting water quality is targeted in this project. The WQD is imported for the prediction. There are different existing algorithms used in the forecasting of the water quality. The existing techniques include the Simple Linear Regression (SLR), Sequential Minimal Optimization (SMO) , IBk (Instance Based) , M5P (M5 Pruning) , MLP (Multi Layer Pruning) , RW (Random Walk) kNN (Nearest Neighbor). This project focuses on improving the prediction accuracy.

## CHAPTER 3

### PROPOSED SYSTEM

This chapter deals with the detailed discussion of the proposed data mining techniques. The proposed data mining techniques are CART and MARS. Both the algorithm uses regression method for the prediction.

#### 3.1 PROPOSED TECHNIQUES

Regression is one of the techniques used for prediction in data mining. The Regression techniques are specially used for prediction and forecasting. Regression is used for predicting numeric or continuous range of values. In general, Regression is a statistical measurement used to analyze the strength of the relationship between one dependant variable (target variable) and one or more independent variables (input attributes). The independent variable is the cause and the dependant variable is the effect. This kind of relationship is called the cause and effect relationship. A decision tree is generated when each node in the decision tree contains a set of test on the input variable's value. The two types of regression techniques used are Classification And Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS).

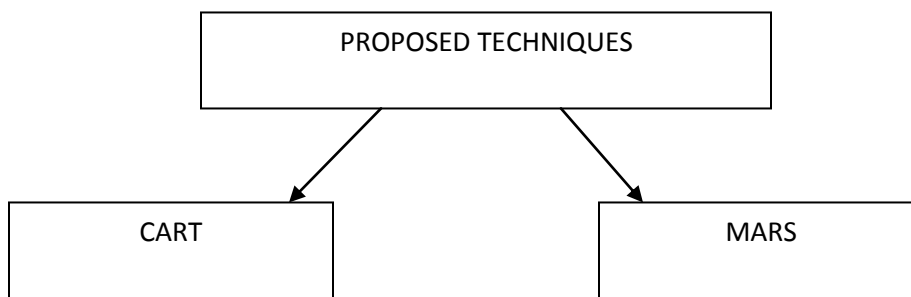


Fig 3.1 proposed techniques

### 3.1.1 CART

The Classification And Regression Trees or generally CART is a decision tree used for constructing prediction model from the data. It is a non-parametric technique which produces classification or regression based on the given data, i.e., categorical or numerical. A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The model is generated by recursively partitioning the data and within each partition fitting a simple prediction model. In this system, the prediction is performed in R language. The CART algorithm is implemented on the importation of a specific package called the *rpart*, which is a specific package for implementing the CART algorithm. The algorithm performs well for the applied data.

### 3.1.2 MARS

The Multivariate Adaptive Regression Splines or commonly called MARS is a effective Regression technique. The salford systems trademarked and licensed the term “MARS”. The open source implementations of MARS are called *Earth*. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables. MARS constructs this kind of relation from coefficients and basis functions. The MARS algorithm uses divide and conquer technique for splitting the data into regions using its own regression techniques. MARS uses two-sided truncated functions as basis functions for linear or nonlinear expansion.



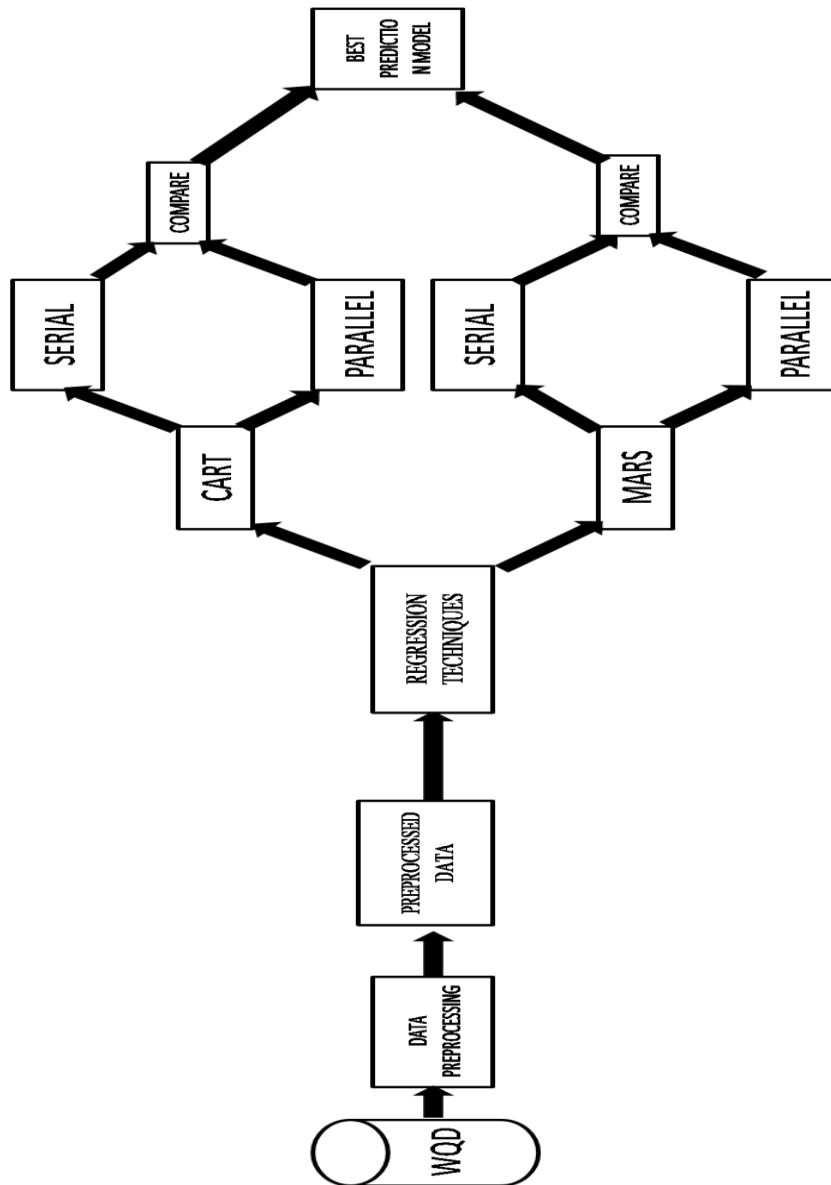


Fig 3.2 System Architecture

The fig 3.1 describes the flow of the system architecture. The data that is used in this project is a Water Quality Dataset. The Water Quality Dataset is obtained from the UCI machine learning repository. The data obtained from the UCI repository was already preprocessed for noise and missing values. But the data collected was unstructured. The collected data is further preprocessed in this project as a structured dataset. The preprocessed data is imported for implementation. The regression techniques are applied to the preprocessed dataset. The CART and MARS algorithms are applied in serial and parallel for computation.

## 3.2 SYSTEM MODULES

### 3.2.1 Module 1 dataset collection

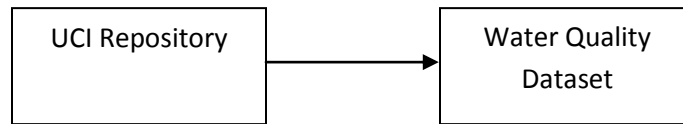


Fig 3.3 dataset collection

The WQD refers to Water Quality Dataset. The data is collected from UCI Repository. The data used consists of 6 targets and 30 attributes. The data is collected using Andromeda analyzer. The data used was collected from April 14,2003 to June 11,2003 on an hourly basis. There are totally 49 observations, which are the recorded values of each day for each attributes. Each attribute has 5 windows and 5 leads. The windows are the average measure of each day. The leads are the days between the last window and the target value. The system is installed in the Thermaikos, Gulf of Thessaloniki, Greece. It consists of 3 local measurement stations and one central station.

The central station acts as the master and collects the data from all the local measurement stations. The communication between the central and local

stations is through hand-shake protocol. The electricity needed by the sensors are provided by the solar collectors and batteries. The central data collection station uses a Pentium computer operating system in SCADA environment.

The Andromeda dataset is related to the prediction of future values of six water quality factors. The water quality factors predicted in this dataset are temperature pH, conductivity, salinity, oxygen and turbidity. The data is collected using under-water sensors with a sampling interval of 9 seconds and is averaged to get a single measurement for each factor over each day. The andro dataset corresponds to the use of a window of 5 days and a lead of 5days. The attribute of the factors corresponds to the value of six water quality factors up to 5 days in the past called the window. The lead is used to predict the value of each factors 6 days ahead.

### 3.2.2 Module 2 data preprocessing

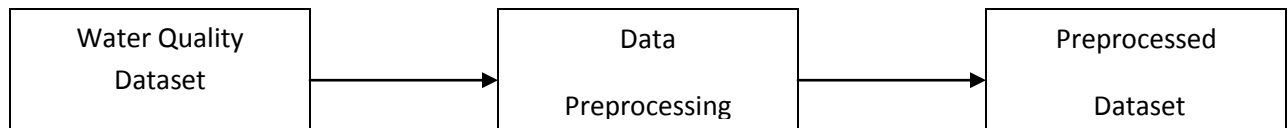


Fig 3.4 data preprocessing

The data preprocessing is the initial phase in data mining. The data collected may have noise and missing values. These noisy data may lead to inaccurate prediction of the targets. The data used for prediction is already preprocessed for missing values and outliers before uploaded in the UCI Repository.

The two problems that aroused during the collection of data are (a) There was some missing values due to the inefficiency of the sensors and problems in the transmission of data (b) The occurrence of some special events near the local

measurement stations like the crossing of boats, led to the recording of some outliers. These temporary problems are solved automatically by the daily averaging process. The missing values on a day are in a range of 0 to 3, and so the rest of the measurements can provide the mean estimate of the day.

The data obtained from the UCI Repository is further preprocessed. The data collected was in Attribute-Relation File Format (.arff) , which was then converted into Comma Separated values (.csv) format. The ARFF format is a ASCII text file that describes a list of instances sharing a set of attributes. The CSV file stores tabular data as plain text. Each line of the data is record which may have two or more fields separated by commas. The conversion of the format of the data is done through *Weka* tool.

### **3.2.3 Module 3 Implementation and execution**

Regression is one of the techniques used for prediction in data mining. The Regression techniques are specially used for prediction and forecasting. Regression is used for predicting numeric or continuous range of values. Many techniques are developed for Regression predictions. The regression technique is used to calculate the relationship between the target variable and the input attributes.

#### **3.2.3.1 CART**

A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The model is generated by recursively partitioning the data space and fitting a simple prediction model within each partition.

### 3.2.3.2 MARS

It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables. MARS constructs this relation from a set of coefficients and basis functions. The MARS algorithm uses divide and conquer technique for splitting the data into regions using its own regression techniques.

### 3.2.4 Module 4 Algorithm comparison

The proposed algorithms in this project are the CART and MARS. The CART is a Classification and Regression decision tree. The CART algorithm splits the giving dataset using regression method. The MARS is a regression analysis technique that models the non linearities and interactions between the variables. The executions of these algorithms are performed in RStudio. The RStudio is a tool specifically used for data mining. The RStudio uses the R Script for its execution. The RStudio can be installed from the official website [www.rstudio.com](http://www.rstudio.com) . The RStudio uses the R language. The R language is an efficient statistical and data mining language. R is a clear and easily accessible tool. The R consists of various libraries specifically designed for statistical data mining.

The Preprocessed dataset is loaded primarily as an object. The dataset is then divided into training and testing datasets using inbuilt partitioning techniques. After the partitioning, the model is generated and the future values are predicted. The prediction is done in two phases. The first phase is the serial prediction, where each targets are predicted individually. The second phase is the parallel prediction, where the targets are predicted parallel. All the targets are predicted

simultaneously. The predicted values are then calculated their efficiency using the performance evaluation metrics.

#### **3.2.4.1 Serial predictions**

The preprocessed dataset is used to generate the model. The model can be generated using the specific package for the specific algorithm. Since this project proposes two algorithms, two packages are used for model generation. For CART algorithm, the package specified is *rpart*. The *rpart* function can be used to generate the model. The *rpart* function is available in the library *rpart*. For MARS, the required package is *earth*. The *earth* package is available in the library *earth*. The additional packages that are required for the implementation are the *doSNow*, *plotrix*, *foreach*, *Parallel*.

After the generation of model, the prediction is implemented. The prediction is done serially, where each input attributes with their related target variables are predicted individually. This kind of prediction is generally slower than the parallelized predictions. However, the prediction values and the values of the performance metrics still remains the same. The major advantage of moving from serial to parallel predictions is that, the processing time is reduced, and hence the processors are used effectively.

#### **3.2.4.2 Parallel predictions**

The parallel predictions are similar to the serial predictions, but the only difference is that the inputs and the targets are processed simultaneously. This is implemented using a function called *cbind*. The *cbind* function is used to predict

the targets parallel, whereas the *foreach* function is used to process the inputs parallel. The *foreach* function is implemented on importing the libraries parallel and *doSNow*.

The reason for considering parallel predictions is that, when the dataset is a single target dataset, then it can be processed in serial method. Since the dataset used in this project is a multi-target dataset, processing of every target individually consumes more time. If such multi targeted datasets are processed in parallel, the processing time will be minimized. Hence it improves the efficiency and enhances the performance of the algorithm.

The parallel predictions are much faster than the serial predictions, because the processor cores are efficiently used for better results. In this project, the number of available cores is 2. Hence, both the cores are efficiently used for the processing. This reduces the processing time of the objects.

### **3.2.5 Module 5 Performance evaluation**

The performance evaluation is the measure of the strength of the algorithms that are applied to a dataset. The performance evaluation or the performance measure is calculated by several distinct metrics. Some of the metrics are Logarithmic loss, confusion matrix, Mean Absolute error, Mean Square Error, Root Mean Square Error, Standard Error, F1 Score, Area Under Curve, Classification Accuracy etc. The Performance evaluation metrics that are considered in this project are Mean Square Error, Root Mean Square Error, Standard error and R-Squared Value.

### 3.2.5.1 Comparison of performance evaluation metrics

This module discusses the comparison of performance evaluation metrics for the prediction of water quality factors using data mining techniques. The evaluation metrics used in this project are briefly explained below.

#### R-square value

R squared is an analytical measure of how adjacent the data are to the fitted regression line. It is also referred as the coefficient of multiple determinations or for the coefficient of determination. **0%** specifies that the model explains nothing variability of the target data around its average.

$$R^2 = 1 - \frac{SSE}{SSrr}$$

$$\text{Where } SSE = \sum (x - \hat{x})^2,$$

$$SSrr = \sum (x - \bar{x})^2,$$

$x$  is the actual value

$\hat{x}$  is the predicted value of  $x$ ,

and  $\bar{x}$  is the mean of the  $x$  values.



**Standard error**

It is the measure of the analytical accuracy of an evaluation, equivalent to the standard deviation of the theoretical distribution of the huge populations of such evaluations.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where,

$\sigma$  is the standard deviation of the population.

$n$  is the size (number of observations) of the sample.

**Mean squared error**

The mean squared error is the estimation of the average of squares of the errors. It is the difference of the square of the estimated value and what is estimated.

$$MSE = \frac{1}{i} \sum_{m=1}^i (Y_m - \hat{Y}_m)^2$$

Where,

$i$  is the number of data points,

$Y_m$  represents observed values,

$\hat{Y}_m$  represents predicted values.

**Root mean squared error**

The root mean squared error is the measure of the difference between the values predicted by a model and the values observed. It is simply the square root value of the mean squared error value.

$$RMSE = \sqrt{\frac{\sum_{n=1}^I (predicted_n - Actual_n)^2}{I}}$$

## **CHAPTER 4**

### **REQUIREMENT SPECIFICATIONS**

#### **4.1 HARDWARE AND SOFTWARE SPECIFICATIONS**

##### **4.1.1 HARDWARE REQUIREMENTS**

RAM	:	2 GB
Hard Disk	:	500 GB
Processor	:	AMD Processor
Monitor	:	13' LCD Monitor

##### **4.1.2 SOFTWARE SPECIFICATIONS**

Tools	:	R, R studio
Back End	:	Java, MS Excel
Operating System	:	Windows 7
Language	:	R

## CHAPTER 5

### EXPERIMENTAL RESULTS AND DISCUSSIONS

This section discusses the inferences of the proposed CART and MARS techniques that are adopted in this work. In this experiment, four different evaluation metrics namely R-squared value, MSE, RMSE, and standard error are calculated.

#### 5.1 Data collection module

A real world dataset for water quality prediction is used in this project. The water quality dataset consists of 6 water quality factors. Since there are 6 targets, the dataset is a Multi-Target dataset. The data is collected from the UCI Repository and is also available in ([mulan.sourceforge.net](http://mulan.sourceforge.net)).

Table 5.1 Water Quality Dataset description.

S.NO	CONTENT	DESCRIPTION
1	DATASET NAME	ANDRO
2	SAMPLES	49
3	ATTRIBUTES	30
4	TARGETS	06

The table 5.1 describes the details of the obtained dataset. The ANDRO dataset is obtained from the UCI Repository. It is the observations of sea water in Greece. The dataset consists of 49 samples with a sampling interval of 9 seconds. There are 30 attributes in the obtained dataset. The attributes consists of 5 windows

of each factors affecting the quality of water. The targets are Temperature, pH, conductivity, salinity, dissolved oxygen and turbidity. These are he factors affecting water quality used in this dataset.

Table 5.2 comparison of R-Squared values

<b>TARGETS</b>	<b>R-SQUARED VALUE</b>	
	<b>CART</b>	<b>MARS</b>
<b>Y0</b>	0.47228	0.96961
<b>Y1</b>	0.99362	<b>0.99373</b>
<b>Y2</b>	0.79608	0.96896
<b>Y3</b>	0.74912	0.95122
<b>Y4</b>	0.25998	0.81301
<b>Y5</b>	0.55631	0.8548

From the table 5.2, for the ANDRO dataset, the target Y1 has achieved the highest R-Squared value (0.99373) for MARS algorithm. The second largest R-Squared value is achieved for the target Y1 (0.99362) for the CART algorithm. The R-squared value for target Y1 has achieved 99.37% on the application of MARS. This promotes that MARS has higher accuracy in prediction than CART algorithm.

The R-Squared values of other targets on MARS algorithm are Y0 (0.96961) 96.9%, Y2 (0.96896) 96.8%, Y3 (0.95122) 95.1%, Y4 (0.81301) 81.3%, Y5 (0.8548) 85.4%. The R-Squared values of the targets on CART algorithm are Y0 (0.47228) 47.2%, Y1 (0.99362) 99.3%, Y2 (0.79608) 79.6%, Y3 (0.74912) 74.9%, Y4 (0.25998) 25.9%, Y5 (0.55631) 55.6%.

Table 5.3 comparison of Error values

TARGETS	MARS			CART		
	MSE	RMSE	Std error	MSE	RMSE	Std error
<b>Y0</b>	0.34038	0.58342	1.01298	1.99208	1.41141	0.66643
<b>Y1</b>	0.00241	0.0491	0.17497	<b>0.00182</b>	<b>0.04266</b>	<b>0.15603</b>
<b>Y2</b>	0.39216	0.62623	1.00175	2.12031	1.45613	0.69222
<b>Y3</b>	0.32456	0.5697	0.6814	1.4391	1.19962	0.44442
<b>Y4</b>	74.3903	8.62498	6.07556	213.528	14.6126	3.99707
<b>Y5</b>	0.26862	0.51828	0.38311	0.52794	0.72659	0.27331

From the table 5.3, the MSE and the RMSE values of target Y1 for CART algorithm has the lowest error values of (0.00182 , 0.04266) 0.1% and 4.26% respectively. The standard error value of target Y1 for CART algorithm has achieved the lowest of (0.15603) 15.6%.

The MSE values of the remaining targets are Y0 (0.34038) , Y1 (0.00241), Y2 (0.39216), Y4 (74.3903), Y5 (0.26862). The RMSE values are Y0 (0.58342), Y1 (0.0491), Y2 (0.62623), Y3 (0.5697), Y4 (8.62498), Y5 (0.51828). The standard error values are Y0 (1.01298), Y1 (0.17497), Y2 (1.00175), Y3 (0.6814), Y4 (6.07556), Y5 (0.38311). In CART algorithm, the MSE values of the remaining targets are Y0 (1.99208), Y2 (2.12031), Y3 (1.4391), Y4 (213.528), Y5 (0.52794). The RMSE values are Y0 (1.41141), Y2 (1.45613), Y3 (1.19962), Y4 (14.6126), Y5 (0.72659). The standard error values are Y0 (0.66643), Y2 (0.69222), Y3 (0.44442), Y4 (3.99797), Y5(0.27331).

Table 5.4 Comparison of average values of CART and MARS

<b>Evaluation metrics</b>	<b>MARS</b>	<b>CART</b>
R-Squared value	<b>0.925222</b>	0.6379
MSE	<b>12.61974</b>	36.60146
RMSE	<b>1.828618</b>	3.2415
Standard Error	1.554962	<b>1.038248</b>

From the table 5.4, the average values of each performance metrics of CART and MARS are compared with each other. The average R-squared value of MARS is found to be (0.925222) 92.5% whereas the average R-Squared value of CART is (0.6379) 63.7%. The MSE and RMSE values are comparatively low in MARS algorithm. These values estimate that MARS tends to perform well for this dataset.

## 5.2 Comparison with existing techniques

Table 5.5 R-Squared values of existing techniques

MODEL	TARGETS					
	Y0	Y1	Y2	Y3	Y4	Y5
RW	0.8164	0.7872	-0.050	0.0737	-	-
AR	0.6887	0.7872	-0.050	0.0737	-	-
MI	0.6887	<b>0.8618</b>	0.2222	-	-	-

From the table 5.5, the existing R-Squared values of the targets are observed. It is seen that, the Random Walk (RW) and the Auto Regressive models (AR) have R-Squared values for 4 of the 6 targets. Whereas the Multivariate Model (MI) has R-Squared values for 3 out of the 6 targets. On the analysis of these values, the R-Squared value of target Y1 in the MI model has found to have the best performance value. But the NRMSE values of the RW and AR are found to be low. The RW model performs efficiently, and is suggested as a best model by the researchers.



Table 5.6 Comparison of average R-Squared values

<b>ALGORITHM</b>	<b>AVERAGE R-SQUARED VALUES</b>
RW	0.406825
AR	0.3749
MI	0.5909
CART	0.6379
MARS	<b>0.925222</b>

On comparing the average values of R-Squared error with the existing techniques, the MARS algorithm achieves the highest value of (0.925222) 92.5%. But the exceptional case is that, the R-Squared values of the existing techniques RW and AR were computed only using 4 targets and for MI only 3 targets were considered. The proposed CART and MARS algorithms are computed with all the 6 targets. Therefore the CART and MARS average values are calculated for all the six target variables. Both the algorithms tend to perform efficiently on comparison with the existing values. On an average, the proposed MARS algorithm functions efficiently.

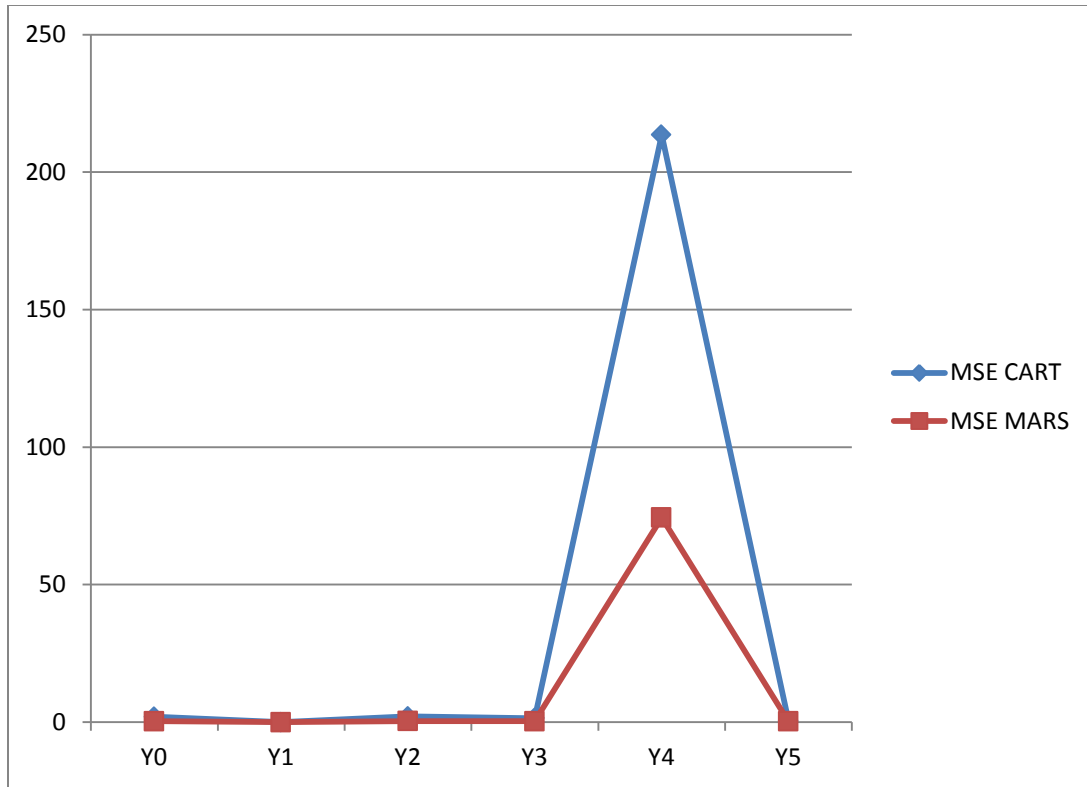


Fig 5.1 comparison of MSE

The fig 5.1, is a line chart that illustrates the comparison of the values of MSE achieved on the application of CART and MARS algorithm on the considered dataset. Both the algorithm performs efficiently, but MARS algorithm tends to have a low MSE error rate when compared with CART algorithm. For each target, on an average, MARS tends to have the low error rate when compared with CART algorithm.

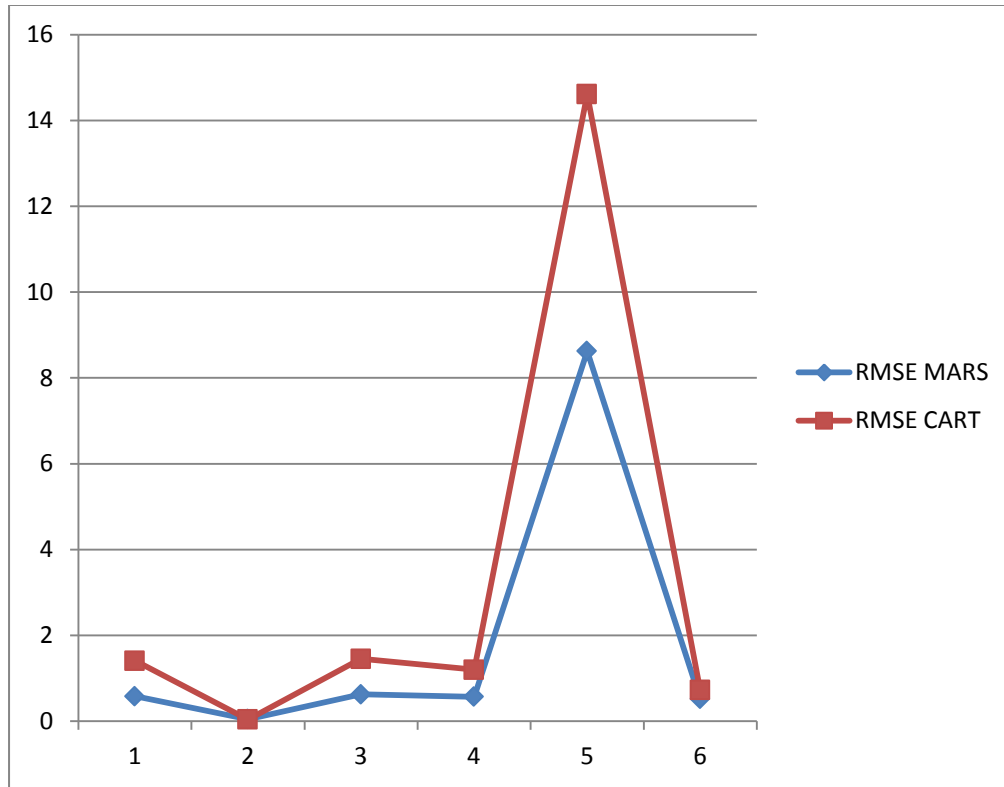


Fig 5.2 comparison of RMSE

The fig 5.2 is the illustration of comparison of RMSE values achieved for both CART and MARS algorithm. The MARS algorithm is observed to have reduced RMSE when compared with CART algorithm. The MARS algorithm has the reduced RMSE error rate. The reduced error in an algorithm defines that the predicted values have less error. This ensures the low error and enhanced prediction of future values.

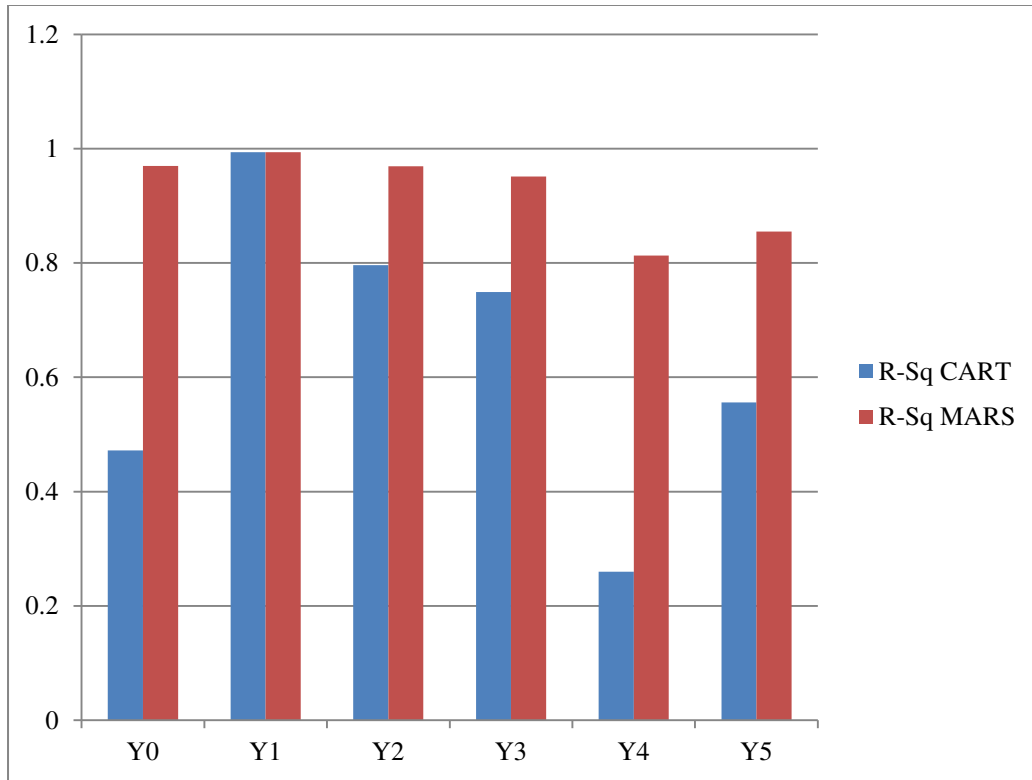


Fig 5.3 comparison of R-Squared values

The fig 5.3, is the comparison of R-Squared values of both the algorithms. The R-Square value determines the accuracy of the predicted future values. The MARS algorithm is found to have the highest R-Squared value. The highest value is achieved for target Y1 of (0.99373) 99.3% and an average R-Squared value of (0.925222) 92.5%. The higher the R-Squared value the greater will be the accuracy. Hence MARS algorithm is found to have the highest prediction accuracy when compared with CART algorithm.

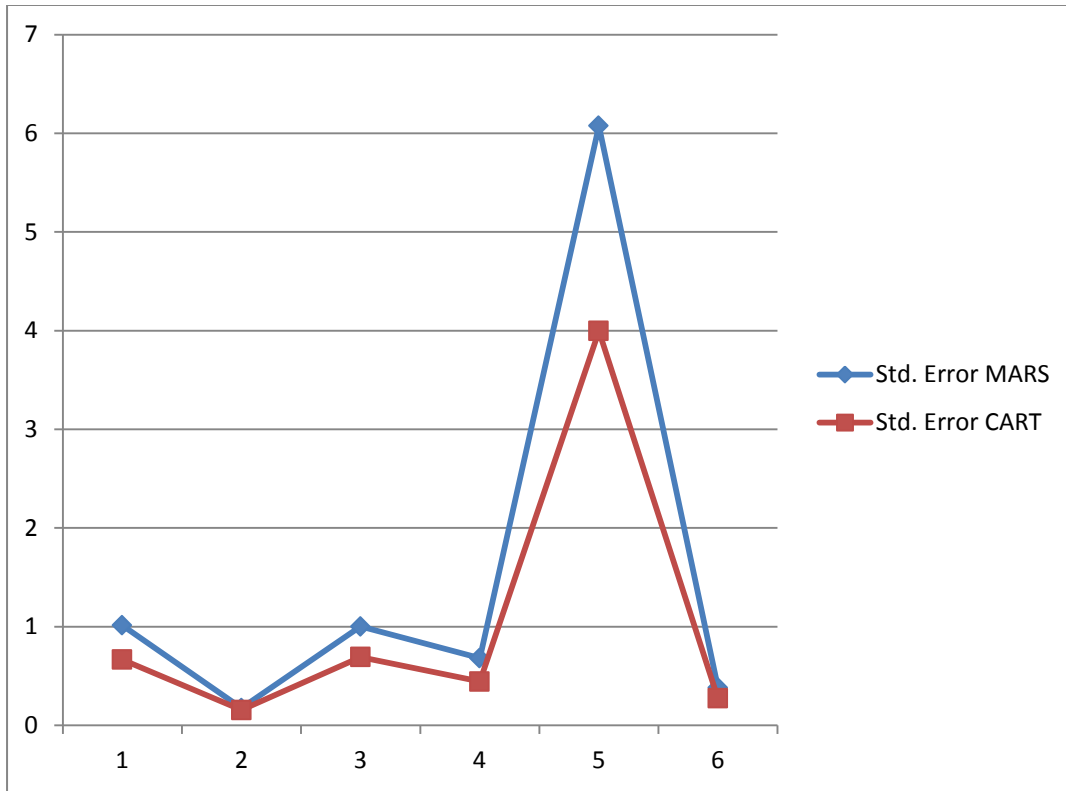


Fig 5.4 comparison of Standard Error

The fig 5.4 is the representation of the comparison of standard error values of each target in CART and MARS algorithm. The standard error is one of the errors that are calculated to measure the error rate of an algorithm on application on a dataset. The standard error value is found to be low in CART algorithm. The variation in the values of both the algorithms are more or less similar. Even MARS has low error rate but when compared with CART, the CART algorithm is found to have less standard error rate.

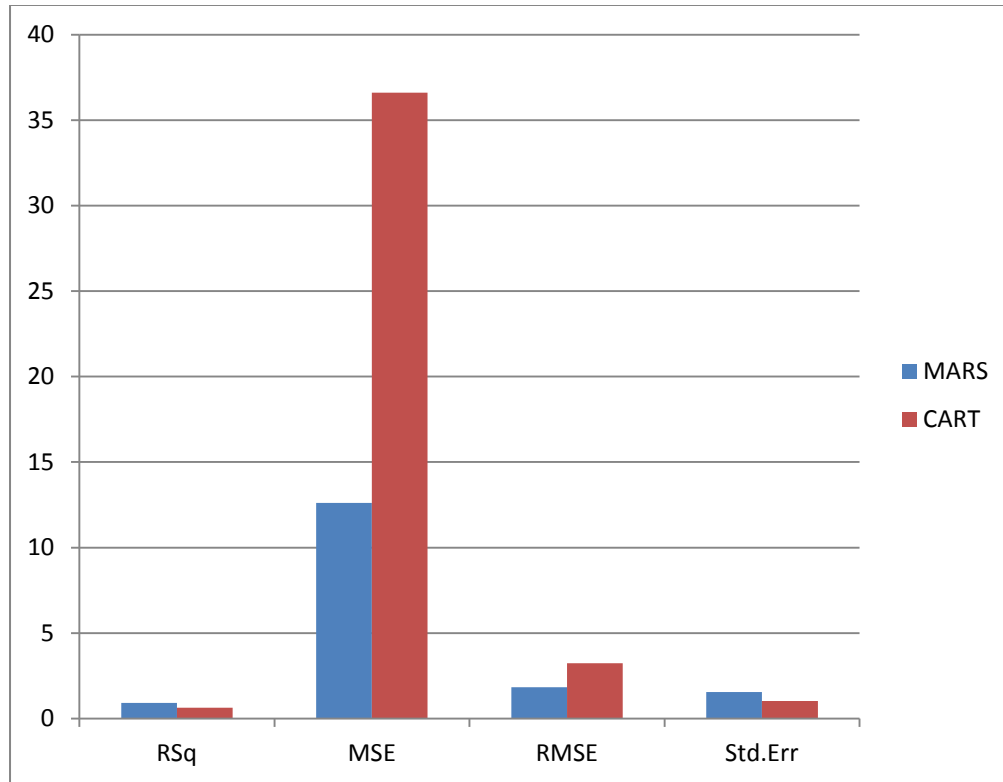


Fig 5.5 comparison of MARS and CART

The fig 5.5 illustrates the comparison of the average values of the performance metrics for both CART and MARS. The R-Squared value of MARS has achieved the highest value. The MSE and RMSE error rates have found to be low on the application of MARS algorithm. The standard error value is found to be low in CART algorithm. On overall estimation, three out of four considered evaluation metrics, the MARS algorithm has achieved the best performance.

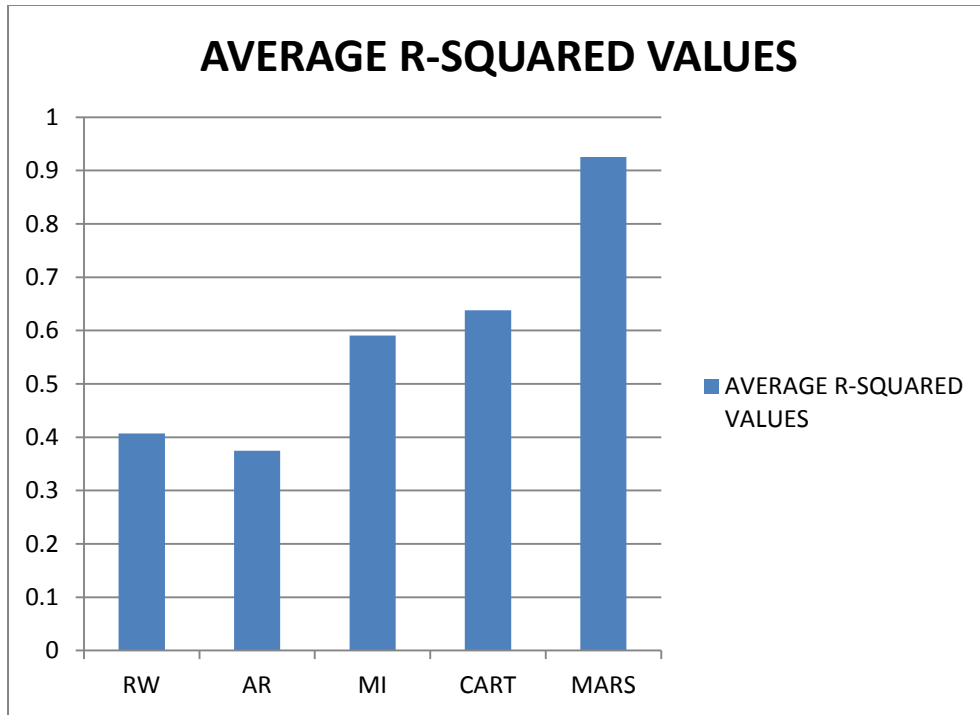


Fig 5.6 comparison of avg. R-Squared values with existing techniques

The fig 5.6 is the comparison of the average R-Squared values of the existing algorithms with the considered CART and MARS algorithms. On comparing, MARS algorithm achieved the highest R-Squared value among all the considered techniques.

Table 5.7 Time difference of serial vs. parallel computation

TIME DIFFERENCE	MARS	CART
SERIAL	1.827104	1.727099
PARALLEL	<b>1.355078</b>	<b>0.736043</b>

The table 5.6 illustrates the comparison of serial vs. parallel prediction time difference of MARS and CART algorithms. It is observed that the amount of time required for processing is relatively low when processed parallel. When the targets and the inputs are processed parallel, it minimizes the time required for computation.

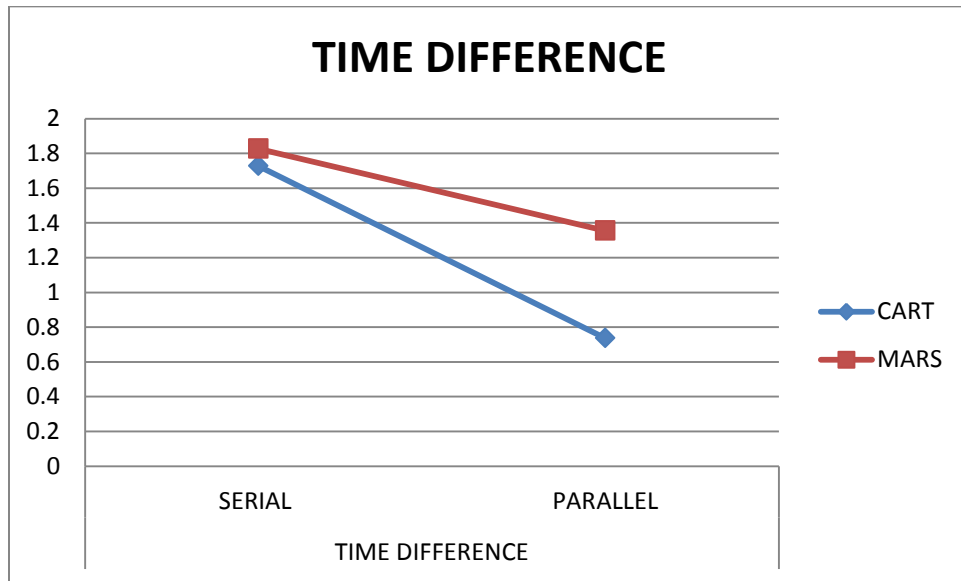


Fig 5.7 comparison of time difference in MARS and CART



The fig 5.7 represents the comparison of time difference of MARS algorithm when processed in serial and in parallel. The resulting time difference shows that the algorithm performs well when it is computed in parallel.

## **CHAPTER 6**

### **CONCLUSION**

The prediction of the quality of water has a significant value in various commercial applications like irrigation and piscicultures. It can help in avoiding undesirable environmental situations. Several techniques are available to predict the water quality. ANDRO is a multi targeted dataset and processing of each target individually can consume more time for processing and analysis. The CART and MARS algorithms are used for prediction of water quality. Both the algorithms are executed in serial and parallel. The parallel execution of the algorithms reduces the time complexity. On comparing the proposed techniques with the existing techniques, experimental results shows that MARS has been performing effectively. The error rates are comparatively low, and the R-Squared value is relatively high. Therefore, the outcomes suggest that MARS could be a best algorithm for water quality prediction in a multi targeted dataset. The CART algorithm has achieved the second best performing algorithm from the experimental results. This project suggests the use of MARS and CART algorithms for water quality prediction and moreover the processing of multiple targets in parallel is found to increase the performance of the proposed algorithms. It provides insights on the protection of ecosystem and helps the authorities to instruct the precautionary measures to control the increase of pollution level.

## REFERENCES

- [1] Hatzikos, E. V., Tsoumakas, G., Tzanis, G., Bassiliades, N., & Vlahavas, I. (2008). An empirical study on sea water quality prediction. *Knowledge-Based Systems*, 21(6), 471-478.
- [2] Bassiliades, N., Antoniadis, I., Hatzikos, E., Vlahavas, I., & Koutitas, G. (2009, March). An intelligent system for monitoring and predicting water quality. In *Proceedings of the European Conference Towards eENVIRONMENT, Prague, Czech Republic* (pp. 534-542).
- [3] Hatzikos, E., Hättönen, J., Bassiliades, N., Vlahavas, I., & Fournou, E. (2009). Applying adaptive prediction to sea-water quality measurements. *Expert Systems with Applications*, 36(3), 6773-6779.
- [4] Hatzikos, E. V., Anastasakis, L., Bassiliades, N., & Vlahavas, I. (2005). Applying neural networks with active neurons to sea-water quality measurements. In *Proceedings of the 2nd International Scientific Conference on Computer Science, IEEE Computer Society, Bulgarian Section* (pp. 114-119).
- [5] Partalas, I., Hatzikos, E., Tsoumakas, G., & Vlahavas, I. (2007). Ensemble selection for water quality prediction. In *Proceedings of 10th International Conference on Engineering Applications of Neural Networks* (pp. 428-435).

- [6] Hatzikos, E. V., Bassiliades, N., Asmanis, L., & Vlahavas, I. (2007). Monitoring water quality through a telematic sensor network and a fuzzy expert system. *Expert Systems*, 24(3), 143-161
- [7] Džeroski, S., Demšar, D., & Grbović, J. (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, 13(1), 7-17.
- [8] Chau, K. (2005, May). A split-step PSO algorithm in prediction of water quality pollution. In *International Symposium on Neural Networks* (pp. 1034-1039). Springer, Berlin, Heidelberg.
- [9] Xu, L., & Liu, S. (2013). Study of short-term water quality prediction model based on wavelet neural network. *Mathematical and Computer Modelling*, 58(3-4), 807-813.
- [10] Tan, G., Yan, J., Gao, C., & Yang, S. (2012). Prediction of water quality time series data based on least squares support vector machine. *Procedia Engineering*, 31, 1194-1199.
- [11] H. Blockeel, L. De Raedt, Top-down induction of first order logical decision trees, *Artificial Intelligence* 101 (1–2) (1998) 285–297.
- [12] H. Blockeel, S. Dzeroski, J. Grbovic, Simultaneous prediction of multiple chemical parameters of river water quality with tilde, in: *Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery*, vol. 1704 of LNAI, Springer-Verlag, 1999.
- [13] F. Fdez-Riverola, J. Corchado, Cbr based system for forecasting red tides, *Knowledge-Based Systems* 16 (321–328) (2003).
- [14] F. Fdez-Riverola, J. Corchado, Fsftr: Forecasting system for red tides, *Applied Intelligence* 21 (251–264) (2004).

- [15] quality analysis of the River Elbe, Germany, *Water Research* 35 (9) (2001) 2153–2160.
- [16] K. Reckhow, Water quality prediction and probability network models, *Canadian Journal of Fisheries and Aquatic Sciences* 56 (1999) 1150–1158.
- [17] C. Romero, J. Shan, Development of an artificial neural network-based software for prediction of power plant canal water discharge temperature, *Expert Systems with Applications* 29 (2005) 831–838.
- [18] A. Smola, B. Scholkopf, A tutorial on support vector regression, Technical report, NeuroCOLT2 Technical Report NC2-TR-1998-030, 1998.
- [19] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, second ed., Morgan Kaufmann, 2005.