WATER QUALITY MONITORING USING MACHINE LEARNING AND INTERNET OF THINGS (IoT)

By
Naveed Rahman
16321016
Riaz Uddin Ahmed Nir
16221009
Saila Hasan Tithi
16121143
Baishakhi Rani Das Shupti
16321098

A thesis submitted to the Department of Electrical and Electronic Engineering in partial fulfillment of the requirements for the degree of
Bachelor of Science in Electrical and Electronic Engineering

Department of Electrical and Electronic Engineering
BRAC University
April 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our original work while completing the degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material that has been accepted or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Naveed Rahman **16321016**	**Riaz Uddin Ahmed Nir** **16221009**
Saila Hasan Tithi **16121143**	**Baishakhi Rani Das Shupti** **16321098**

Approval

The thesis/project titled “WATER QUALITY MONITORING USING MACHINE LEARNING AND INTERNET OF THINGS (IoT)”
submitted by

1. Naveed Rahman (16321016)
2. Riaz Uddin Ahmed Nir (16221009)
3. Saila Hasan Tithi (16121143)
4. Baishakhi Rani Das Shupti (16321098)
   of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Electrical and Electronic Engineering on 13th January 2021.

Examining Committee:

Supervisor:
(Member)

_____
Abu S.M. Mohsin, PhD
Assistant Professor, Department of Electrical and Electronic Engineering
BRAC University

Program Coordinator:
(Member)

_____
Abu S.M. Mohsin, PhD
Assistant Professor, Department of Electrical and Electronic Engineering
BRAC University

Departmental Head:
(Chair)

_____
Md. Mosaddequr Rahman, PhD
Professor and Chairperson, Department of Electrical and Electronic Engineering
BRAC University

Abstract/ Executive Summary

In this project, we built and developed an IoT based system that can monitor the water quality of various places in real-time and provides future predictions regarding the water quality in each place. For this project, we developed a physical device that collects various data of water. This data was collected by various sensors built-in with this device. This data includes the water's pH level, turbidity level, TDS (Total dissolved Solid) level, Rain level, Sunlight level, etc. The physical device consists of a microcontroller that gathers these data and sent it to a secured website using a Wi-Fi module. This hardware device is wireless, and it is water-resistant as it was placed close to water sources. It consists of a big battery or solar panel to charge the device. Afterwards the hardware device sends data to the website, stores the information & collects data of water quality every day. Each day it collects data 2 times (once every 12 hours. We collected the data for 2 weeks and analyzed the data and performed future prediction. As the sample size was small therefore, we observed larger error rate, however the error was reduced increasing the number of data set. The proposed system will not only be helpful to observe the real-time monitoring of water quality but also to develop a better water management system for the local community.

Keywords: Water quality monitoring(WQM), water pollution, pH, TDS, Turbidity, machine learning and future prediction, and Internet of Things (IoT), .

Dedication (Optional)

This paper is dedicated to all to the fishermen who provide fish for our country. Also dedicated to our beloved parents and respective supervisor, Dr. Abu S.M. Mohsin, who guided us through this project.

Acknowledgment

Firstly, we are extremely grateful to the Almighty Allah for allowing us to complete this research.

We would like to express our gratitude towards our supervisor Dr. Abu S.M. Mohsin, Assistant Professor, Department of Electrical and Electronic Engineering, BRAC University, for his support and motivation. We are most grateful for his utmost dedication and for believing in us.

We would also like to mention our parents for their support and for providing us with love and care. It was their encouragement that helped us reach our goal even through these hard times.

Table of Contents

List of Tables

List of Figures

List of Acronyms

TDS	Total Dissolved Solids
ROC	Receiver Operating Characteristic
TP	True Positive
TF	True Negative
FP	False positive
FN	False Negative
WQM	Water Quality Monitoring

Chapter 1

Introduction

1.1 Background

Water is one of the most vital natural elements. We cannot think of a day in our life without water. From the very beginning of our life to the survival on earth, in every aspect directly or indirectly we are dependable on it. However, as industrialization occurs, we, humans are polluting our water sources. So, we must identify the pollutants and remove them from the water, not only that but also we need a fast system that can notify the authority.

In Bangladesh, many of the people are dependent on water sources as it is a riverine country. So we must find a way to detect clean water or to see if the water is toxic or not. However, that is not enough. We need something that works very fast and gives very accurate data and to get this fast and accurate data we need IoT technology.

So, the main objective of our research is to build up a system that can detect any type of pollutant in the water, give real-time data and also notify the authorities. The challenge for our research is to develop a secured platform where companies and authorities can get their water source data in real-time. However, that will not be enough. To tackle water pollution we need future predictions of water sources contamination level. Our research must include this future to get better WQM of water sources.

1.2 The motivation & Objective of this project

We have identified 5 key motivations for this project. They are:

- Developing an IoT based device for Water Quality Monitoring
- Water Quality Prediction
- Detecting the level of contamination
- Detecting the most suitable Algorithm for WQM prediction
- Detecting the error rate of the prediction

1.3 Literature review

Bangladesh being a riverine country, there are much research is related to its water, rivers, and lakes regarding WQM (water quality monitor). However, only quite recently IoT has been introduced in this field of research. Most of the research done before has been done manually or without any IoT system. Only very recently some research has been done in Bangladesh regarding IoT-based water quality monitoring. For instance, a research paper by

Abhijit Pathak (2020) [9] shows that Bangladesh has implemented an IoT-based WQM system on the Karnaphuli River. The IoT based system that they used is called ANN (artificial neural network) analysis. They used the BP algorithm (Backpropagation algorithm) for their project as well as Esp8266 (Node MCU) for their hardware. They have been successful in implementing a low-cost IoT based WQM system on a river with the ability to conduct future predictions. On the other hand, a similar project was conducted by BRAC University regarding sensor-based water quality monitoring system, Paul (2018) [10]. For their project, they used GSM shield (SIM 808) and ARTIK Cloud for their IoT services and Arduino Mega-2560 for their hardware. Even though they made a reliable WQM system, they lacked IoT services as there were online services only to give data and could not predict future patterns. Another paper regarding WQM by Daigavane & Gaikwad (2017) [11] conducted a similar project based on IoT services. Using Arduino Atmega238p, ESP8266 Wi-Fi module & BLYNK app this project is conducted. However, this project also lacks the future predictability of IoT services. Since this technology is quite new, our government also has not implemented any vast IoT based water quality monitoring system in Bangladesh. Till now the water quality monitoring has been done manually in Bangladesh. So we are greatly in need of a low-budget, IoT based water quality monitoring system. Water-quality monitoring is used to alert us about current, ongoing, and emerging problems; to determine compliance with drinking water standards, and to protect other beneficial uses of water. Bacteria viral and parasitic diseases like typhoid, cholera, encephalitis, poliomyelitis, hepatitis, skin infection, and gastrointestinal are spreading through polluted water. It is recommended that we should examine the water quality on regular basis to avoid its destructive effects on human health.

1.3.1 Detecting toxic wastewater near industrial factories

Water quality is vital for a solid environment and livelihood. Bangladesh, as a downstream portion of three major streams, is dominated by floodplain and deltaic environments that are increasingly being undermined by surface water contamination and excessive groundwater deliberation. Populace increment, discharge of untreated mechanical effluents and sewage, spontaneous strong squander dumping along the river banks, infringement, and non-point source contamination from agricultural exercises are major reasons for destitute water quality and disintegration of the stream ecosystem. Office of Environment (DoE) has been monitoring surface water quality since its foundation in 1973. DoE's surface water quality checking arrangement incorporates major rivers and lakes. Checking data give water quality of major waterways of Bangladesh. The water quality of waterways encompassing Dhaka, Chattagram, Khulna city, and close other major urban zones was below the Natural Quality Standard. The water quality of streams has been seriously influenced by the dumping of civil strong squander, the coordinate connection of sewerage lines, and toxic mechanical discharge. Buriganga has gotten to be greatly polluted and is near to natural death. The huge increment in pollution has made colossal environmental issues. Industrial Clusters – Commitment to Pollution. The Buriganga and adjoining waterways around Dhaka City are being contaminated by emanating streaming in from two directions: – the surge of sewage from the Dhaka City & – from the clusters of industry. Nine hotspots of mechanical contamination have been distinguished by IWM as a major source of mechanical contamination. These are (1) Tongi, (2) Hazaribag, (3) Tejgaon, (4) Tarabo, (5) Narayanganj, (6) Savar, (7) DEPZ, (8) Gazipur, and (9) Ghorashal. As an impact, Individuals are moved from the waterway side to other places. They are socially confined to their relatives. Rural arrive losing their fertility. Exceedingly contaminated stream water may not be utilized for irrigation. As the Effect on Economy, Stream-based tourism industry has not developed. Anglers ended up unemployed. Numerous greenery and Fauna are nearly lost. Nourishment chains are irritated and biological adjust are beneath debilitating due to noxious chemicals. Destitute water quality disheartens recreation. Buriganga misplaced her fascination & excellence.

1.3.2 Drinking Water as a Vehicle of Diseases

As water is an essential element for life, a satisfactory, secure, as well as open supply must be accessible to all. Moving forward to secure drinking-water can result in noteworthy benefits to wellbeing. Each exertion ought to be made to attain a drinking water quality as secure as conceivable. Many individuals battle to get to secure water. The supply of clean and treated water to each house may be the standard in Europe and North America, however, in creating nations, get to both clean water and sanitation are not the run the show, and waterborne diseases are common. Two and a half billion individuals have not got to make strides in sanitation, and more than one and a half million million children pass on each year from diarrheal maladies. Agreeing with the World Health Organization (WHO), the mortality of water-related infections surpasses 5 million individuals per year. From these, more than 50% are microbial intestinal infections, with cholera standing out within the, to begin with, put. In common terms, the most noteworthy microbial dangers are related to the ingestion of water that's sullied with human or creature feces. Wastewater releases in new waters and coastal seawaters are the major sources of fecal microorganisms, counting pathogens. Acute microbial diarrheal infections are a major open wellbeing issue in creating nations. Individuals influenced by diarrheal infections are those with the most reduced monetary assets and poorest clean offices. Children beneath five, fundamentally in Asian and African countries, are the foremost influenced by microbial maladies transmitted through water. Microbial waterborne illnesses to influence created nations. Within the USA, it has been evaluated that each year 560,000 individuals endure serious waterborne maladies, and 7.1 million endure from mellow to moderate diseases, coming about in evaluated 12,000 passing a year. Cholera flare-ups can happen sporadically in any portion of the world where water supplies, sanitation, nourishment security, and cleanliness hones are lacking. Stuffed communities with destitute sanitation and risky drinking-water supplies are most regularly affected. Cholera is caused by the bacterium Vibrio cholera. Crude or undercooked fish may be a source of contamination in zones where cholera is predominant and sanitation is destitute. Vegetables and natural products that have been washed with water sullied by sewage may moreover transmit the contamination if V. cholera is present. Control of cholera may be a major issue in a few Asian countries as well as in Africa. Within the year 2000, a few 140 000 cases coming about in roughly 5000 passing were authoritatively informed to WHO. Africa accounted for 87% of these cases. Diarrhea illness may be a leading cause of child mortality and horribleness within the world, and for the most part, comes about from

sullied nourishment and water sources. Worldwide, 780 million people need to get to progressed drinking water. The runs could be an indication of diseases caused by a wave of bacterial, viral, and parasitic living beings, most of which are spread by feces-contaminated water. The disease is more common when there's a deficiency of satisfactory sanitation and cleanliness and secure water for drinking, cooking, and cleaning. Rotavirus and Escherichia coli are the two most common etiological operators of moderate-to-severe diarrhea in low-income nations. Other pathogens such as cryptosporidium and Shigella species may be critical. Location-specific etiologic designs too got to be considered. Water sullied with human defecation, for illustration, from sewage, septic tanks, and toilets, is of specific concern. Creature defecation too contains microorganisms that can cause the runs.

1.3.3 Microbiological Water Analysis

There are many types of micro bacteria in water bodies across the world. Cholera, shigellosis and salmonellosis are some of the common micro bacteria that are effacing people all across the globe. These infections are transmitting through water bodies. They also spread through feces of infected people. Drinkable water sometimes gets infested with these microscopic organisms which are pathogenic, and typically can cause great hard to a wind number of people. In any case, the nearness of deadly microbes inside water bodies is scattered and whimsical. That is why, a scheduled investigation regarding microbiological of water does not incorporate the discovery of these dangerous microbes. Be that as it may, secure water is mandatory to be free of theses deadly pathogens. This water quality monitoring system can be able to detect these pathogens if equipped with the right sensors.

1.3.4 Protecting Fishes

Whether in new or marine waters, angle requires untainted nourishment, appropriate environment, and satisfactory oxygen to outlive. Any component, whether chemical or characteristic, that upsets this adjust is considered water contamination, or essentially a toxin. Water poisons are wide-ranging and depend on the locale of the world the angle lives in, but there are a few that are common to numerous parts of the world. Fertilizer Supplements Drain Oxygen Nitrogen and phosphorus are supplements that ended up water poisons when

they enter waterways, lakes, and seas through runoff, such as rain washing overabundance fertilizer from grass into a lake, or a coordinate release when a sewage treatment plant pumps handled sewage into a waterway. As these abundance supplements construct up in a body of water, plants and green growth develop at quickened rates causing plant abundance and hurtful algal blossoms. When plants kick the bucket, the rot preparation brings down the broken down oxygen level within the water to a level too low for the angle to outlive, coming about in angle slaughters. When an angle bolsters on hurtful green growth, it ingests poisons that gather in its body and are passed on to another angle that eat them. Manufactured pesticides, such as weed and bug executioners, are poisonous to angle in moo concentrations coming about in angle mortality and a decrease in angle populaces. A few angles are more delicate than others and pass on at lower concentrations. Pesticides enter new and marine waters when connected to a grass or farming field, and overabundance is washed into the water when it downpours, or on the off chance that the splash floats when connected. Angle bolsters on spineless creatures that live within the water. Take absent this nourishment source and they either pass on from starvation or move to an unused environment. These spineless creatures incorporate waterborne creepy crawlies; pesticides are harmful to them in moo concentrations. Be that as it may, on the off chance that the pesticide does not slaughter the creepy-crawly, it is exchanged when an angle eats it. Over time, pesticide builds up within the angle until it comes to a deadly level. Dregs are another toxin that slaughters spineless creatures. A thick layer of sediment can cover bottom-dwelling spineless creatures. Overwhelming silt can moreover cover angle eggs, decreasing their populaces. Residential wastewater contains toxins extending from biological dangers and microplastic particles to cleansers and fats. Rural wastewater contains biological risks, salts, pesticides, and fertilizers. Urban wastewater incorporates household and mechanical wastewater but moreover contains runoff from storm channels. Storm channels carry toxins from yards and parks (earth, pet squander, pesticides, herbicides, and fertilizers) as well as from boulevards and stopping parcels (oil, gasoline, soil, and junk). Mechanical wastewater contains a wide extend of chemicals that incorporate petrochemicals and other chemicals, acids, radioactive materials, and salts. Later discoveries appear that an assortment of drugs to sully wastewaters. Mechanical squanders frequently pass through the same sewer treatment offices as household squanders. Mechanical squander frequently contains an assortment of chemicals and may too contain overwhelming metals like lead, mercury, cadmium, and arsenic. Not all of these chemicals are evacuated in sewage treatment plants, so the chemicals are discharged into streams, lakes, and marine waters. In expansion, a few squander may be discharged or spilled

into sea-going biological systems without any treatment. The impacts of sewage contamination on marine life affect life forms all through the nourishment chain. Heavy metals construct up in angle tissues as the angle devour tiny fish, green growth, and littler prey containing the metals. This preparation is called biomagnification. Like other creatures, counting people, eat these angles, the overwhelming metals can reach adequate concentrations to harm the buyer. These overwhelming metals may amass in poisonous sums for angle as well.

1.3.5 Detecting clean drinkable water

Water quality is vital for a solid environment and livelihood. Bangladesh, as a downstream portion of three major streams, is dominated by floodplain and deltaic environments that are increasingly being undermined by surface water contamination and excessive groundwater deliberation. Populace increment, discharge of untreated mechanical effluents and sewage, spontaneous strong squander dumping along the river banks, infringement, and non-point source contamination from agricultural exercises are major reasons for destitute water quality and disintegration of the stream ecosystem. Office of Environment (DoE) has been monitoring surface water quality since its foundation in 1973. DoE's surface water quality checking arranges incorporates major rivers and lakes. Checking data give water quality of major waterways of Bangladesh. The water quality of waterways encompassing Dhaka, Chattagram, Khulna city, and close other major urban zones was below the Natural Quality Standard. The water quality of streams has been seriously influenced by the dumping of civil strong squander, the coordinate connection of sewerage lines, and toxic mechanical discharge. Buriganga has gotten to be greatly polluted and is near to natural death. The huge increment in pollution has made colossal environmental issues. Industrial Clusters – Commitment to Pollution. The Buriganga and adjoining waterways around Dhaka City are being contaminated by emanating streaming in from two directions: – the surge of sewage from the Dhaka City & – from the clusters of industry. Nine hotspots of mechanical contamination have been distinguished by IWM as a major source of mechanical contamination. These are (1) Tongi, (2) Hazaribag, (3) Tejgaon, (4) Tarabo, (5) Narayanganj, (6) Savar, (7) DEPZ, (8) Gazipur, and (9) Ghorashal. As an impact, Individuals are moved from the waterway side to other places. They are socially confined to their relatives. Rural arrive losing their fertility. Exceedingly contaminated stream water may not be utilized for

irrigation. As the Effect on Economy, Stream-based tourism industry has not developed. Anglers ended up unemployed. Numerous greenery and Fauna are nearly lost. Nourishment chains are irritated and biological adjust are beneath debilitating due to noxious chemicals. Destitute water quality disheartens recreation. Buriganga misplaced her fascination & excellence.

1.3.6 Detecting arsenic in water

Based upon the various forms such as chemical, physical, and biogeochemical and environmental conditions, different mode of arsenic showed in water. There are two classes' arsenic water types. First one is inorganic arsenic (IAS) and other one is natural arsenic (OAS) species. All kinds of acidic species have their own ionic and atomic features in water and chemically well-balanced. The smearing of IAS and OAS modes of arsenic could be an estimated by working with pH of water which is found in lakes, groundwater, sea and streams. The WHO estimated temporary level of arsenic 10 μg L−1 for drinking water. The delicate and chief particular strategies for consolation of accumulation of arsenic and its species in the water are coupled methods counting optical strategies, chromatography, and mass spectrometry. Identifying the arsenic species is of central significance for the fixation of arsenic evacuation innovation. Best acquirable benefits are depended on precipitation retention, half breed layer and layer. Metals soil, water, minerals, silt, rocks and living beings that contained large amounts of arsenic disintegrates and through this Arsenic enters into the water. Usually the water from surface water layers, contain a little concentration of arsenic and that marked as drinking water. Highly concentrated arsenic water has been found within the deep water. The elements of arsenic in the nature are altered and impacted by exercises of Human. There is peculiar coordination between the uses of arsenic compounds for distinctive purposes with the impact. There is the too backhanded impact that influences the portability of arsenic from distinctive common sources. The natural source of arsenic compounds such as AsC, AsB, TMAO, TETRA, arsenic-containing lipids and arsenosugar are mainly found in the marine life forms, even though a little of these elements have been found in earthbound species. The level of arsenic concentration in drinking can be identified by advanced explanatory procedures like HG-AAS, GF-AAS and ICP-MS.

1.3.7 Detecting weather patterns and temperature

Temperature is a critical factor to consider when evaluating water quality. In expansion to its claim impacts, temperature impacts a few other parameters and can change the physical and chemical properties of water. In this respect, water temperature ought to be accounted for when deciding metabolic rates and photosynthesis production – Compound toxicity – Dissolved oxygen and other broken up gas concentrations – Conductivity and salinity – Oxidation decrease potential (ORP) – pH – Water Thickness. Considered alone, the water temperature can influence the metabolic rates and natural movement of oceanic organisms. As such, it impacts the chosen territories of an assortment of sea-going life. A few life forms, especially sea-going plants prosper in hotter temperatures, whereas a few angles such as trout or salmon favor colder streams. Studies have appeared a coordinated relationship between metabolic rates and water temperature. This happens as numerous cellular chemicals are more dynamic at higher temperatures. For most angle, a 10°C increment in water temperature will around twofold the rate of physiological work. This increment in metabolic rate can be taken care of by a few species ways better than others. Expanded metabolic work can be taken note of in-breath rates and stomach related reactions in most species. Expanded breath rates at higher temperatures lead to expanded oxygen utilization, which can be inconvenient in case rates stay rose for an expanded period. Plants are moreover influenced by water temperature. Whereas a few oceanic plants endure cooler waters, most lean toward hotter temperatures. Tropical plants in specific will appear confined development and torpidity in water temperatures underneath 21°C. Whereas torpidity is fitting for surviving a cold winter, hotter temperatures are required for most plants to flourish. Temperature can moreover hinder plant breath and photosynthesis. In common, algal photosynthesis will increment with temperature, even though distinctive species will have diverse crest temperatures for ideal photosynthetic movement. Over and underneath this temperature, photosynthesis will be decreased. pH is calculated by the number of hydrogen particles in the arrangement. At a pH of 7, the hydrogen and hydroxyl particles have a break-even with concentrations, 1 x 10-7 M, keeping the arrangement unbiased. In any case, these concentrations were held genuine at 25°C. As the temperature increments or diminishes, the particle concentrations will moreover move, in this way moving the pH esteem. Be that as it may, that does not cruel that temperature changes will make an arrangement more acidic or fundamental. Since the proportion of hydrogen and hydroxyl particles remains the same, the corrosiveness of water does not alter with temperature. Instep, the whole pH extends shifts, so that unbiased water

will have esteem other than 7. Unadulterated water will stay unbiased at 0°C (pH is 7.47), 25°C. (pH is 7.00) or 100°C. (pH is 6.14). Lastly, expanded turbidity will too increment water temperature. Turbidity is the sum of suspended solids in water. These suspended particles retain warmth from sun oriented radiation more proficiently than water 47. The warmth is at that point exchanged from the particles to water atoms, expanding the temperature of the encompassing water.

1.4 Types of water

We can categorize water based on its location, pollution level, etc. Some of these water sources are described below:

- Drinking water: For drinking water, the parameters include color, pH, TDS, Iron, Arsenic, etc. The parameters have some average values and we get to know about the quality by measuring these per liter whether the water is drinkable or not. Drinking water is so far the most important to notice for us in Bangladesh as so many cases of water polluted diseases are found.
- Hospital wastewater: Because of the use of several types of chemicals used and for the human wastes the hospital water gets highly toxic and contaminated so this is also very important to make that water renewable again to use. The parameters are usually TCOD, BOD, NH_4, NO_3, TSS, and pH and turbidity is essential as well.
- Industrial wastewater: Industrial wastewater is the discard that results from substances having being dissolved or suspended in water, typically during the use of water in an industrial manufacturing process or the cleaning activities. This water is hugely toxic and full of knitted goods. When they get directly thrown into the river the fishes and the environment also get polluted. So, it will be better if this water can be purified to a minimum level and be used again for the same purpose. The parameters are pH, BOD, COD, TSS, TDS, Oil, and grease, and also the temperature. If we can minimize the parameters according to the index we can use the same water in the industrial work.

If we can categorize the water sources we can also categorize sensors of measurement and the monitoring further. If we do so, we won't need to check the waste materials for every water source because domestic water won't have any industrial or hospital waste. Our time and work will be very efficient then. Real-time data of water quality will be stored in a cloud

server which can be accessed through the internet of things from any place of the world. The data obtained from the sensors are recorded in the database and further sent for analysis. The neural network algorithm is used for predicting the result. It is used to obtain non-linear relationships for predicted output. The system sends the alert message to the user when any of the parameters are lower than the standard values. The data updated at intervals within the server may be retrieved or accessed from anyplace within the world. If the sensors do not work or get into abnormal conditions, then a buzzer will be ON. This helps the user to know beforehand about the contamination of water in their residential tanks. We will be using Dissolved Oxygen Sensor, Flow Sensor, CO_2 Sensor, PH sensor, TDS sensor, Turbidity Sensor, Temperature Sensor, Moisture Sensor, LDR Sensor, MQ137 Ammonia Detection Sensor, Rain Sensor to detect water quality and analyze the data overtime to fix the increase rate of pollution ingredients.

1.5 Water Quality of Bangladesh

In Bangladesh, the main source of usable water is the river since it is a riverine country. We are supplied water from rivers and after the usage, this water goes back to different water sources like lakes, ponds, rivers and make them polluted. For example, in houses, the water gets mixed with various types of soaps, detergents, and so on. Hence, when this water gets mixed with the water from a nearby lake or river, the water of the source becomes polluted. In the same way, industries, garments factories, hospitals, etc. are also liable for water pollution.

Some chemical-physical composition of wastewater from different sources is given below:

1.5.1 Hospital wastewater

Several types of chemicals are used in hospitals Amir Mohammad Mansouri (2015) **[12]**

Therefore, the wastewater from hospitals is composed of these things:

Table-1: Type of chemicals in Hospital water [12]

Parameter	Range
TCOD (mg/L)	450-654
BOD (mg/L)	220-345
TSS (mg/L)	259-520
TKN (mg/L)	81-120
NH_4-N (mg/L)	18-41
NO_2-N (mg/L)	0.03-0.3
NO_3-N (mg/L)	0.08-0.36
N-organic (mg/L)	59-72
TN (mg/L)	81.1-120.7
TP (mg/L)	14-19
Turbidity (NTU)	50-71
pH	7.5-7.9
Alkanity (mg/L)	376-509

1.5.2 Textile Wastewater

Haque Md Mahbubul (2008) **[13]** says in his article "Treatment of textile wastewater in Bangladesh" that nearly 150-200 liters of water are required to dye one kg of knitted goods. According to the Department of Environment, Government of Bangladesh the standard characteristics of textile wastewater is given below:

Table-2: Type of chemicals in the water near Textile Industries [13]

Parameters	Unit	The standard value for Discharge into Inland River	The standard value for Discharge into On Land for irrigation
pH		6-9	6-9
BOD	mg/L	<50	<100
COD	mg/L	<200	<400
TSS	mg/L	<150	<200
TDS	mg/L	<2100	<2100
Oil & Grease	mg/L	<10	<10
Color	Co-pt Unit	<150	<150
Temperature	C	<30	<30

1.5.3 Buriganga River

The minimum water quality should be maintained to save the ecosystem of that particular source. A few days back the river Buriganga was not suitable for any living organisms. In recent days it is becoming more suitable for the survival of organisms. However, according to DOE, the standard ratios of the elements are given for maintaining the ecosystem of the river.

3) Hoque, Muhammad. (2016). **[14]**

Table-3: Type of chemicals in Buriganga River [14]

Parameters	DOE Standards to maintain the aquatic ecosystem	Winter	Rainy Season
Temperature	$20\text{-}30^0$ C	Yes	Yes
pH	6.5-8.5	Yes	Yes
DO	5 mg/L	No	Yes

BOD_5	2 mg/L	No	No
COD	4 mg/L	No	No
TDS	450-2000 mg/L	Yes	Yes
CL ion	5 mg/L	Yes	Yes

1.5.4 Dhanmondi Lake

The sample analysis was done from the water sample collected from Rabindra Sharobor. Jan - E – Alam (2017)**[15].** From the analysis, we have come to know that the water quality of this portion is:

Table-4: Type of chemicals in Dhanmondi Lake [15]

Water Quality Parameter	**Concentration Present**
pH	6.8
Turbidity	6.2 (NTU)
BOD_5	16.5 mg/L
COD	13.2 mg/L
DO	9.1 mg/L
Color	51 pt-Co

1.5.5 Hatirjheel and Gulshan Lake

The Hatirjheel Lake and the Gulshan Banana lake are polluted by both domestic and industrial wastes according to Tariquzzaman (2016) **[16].** Through Hatirjheel Lake, the polluted water flows from Gulshan to Rampura. The values of water quality in these areas are given below:

Table-5: Type of chemicals in Hatirjheel and Gulshan Lake [16]

Water Quality Parameter	Concentration Present
pH	7.1
TDS (Total Dissolved Solids)	288 mg/L
TSS (Total Suspended Solid)	60 mg/L
Phosphate (PO_4)	3.4 mg/L
Ammonia (NH_3)	`12.5 mg/L
Nitrate (NO_3)	.33 mg/L

1.6 Conclusion

In the end, to overview this chapter at a glance we can say that we have mentioned our motivation behind doing the WQM based research. There is a brief discussion about the water types, water sources, and pollutants of our country. The necessity of monitoring the water quality not only for the usage of the human being but also for the living of the fishes. Moreover, a little bit about the weather pattern, the water quality of some well-known water sources is also included. Details of the project planning and schedule are added in this chapter. In the next chapter, the hardware implementation process and the details of the sensors we used will be described so that it becomes clearer how the physical device has been constructed.

Chapter 2

Hardware Device

2.1 Introduction

To gain data from a water source a proper device must be made which can gather data and send it to the internet. It also must be capable of being portable and light. This is because it has to float on water to gather data. It also needs to be wireless to be portable. Also, it has to have enough energy to collect data for a long period. The solar panel can be used. However, we instead used a rechargeable battery to supply power. The major components are given below:

2.2 List of components

The major components for our device are:

1. Arduino Mega 2560
2. pH Sensor
3. LDR Sensor
4. Turbidity Sensor
5. TDS Sensor
6. Rain Sensor
7. Esp8266 Wi-Fi module
8. Rechargeable Battery

Other components include tape, wires, floating devices, etc.

2.3 Arduino mega 2560

Arduino Mega 2560 is a microcontroller based on Atmega 2560 microcontroller. It is an open-source board that is very easy to use. It is a small and compact and very user-friendly board. These boards are fully functional computers that can extend separate interactive objects. [17]

Specifications:

- Run by ATmega2560
- 5 volts operating voltage
- Recommended Input range from 7V to 12V
- Actual input voltage range from 6V to 20V
- The digital I/O pins are 54
- 15 pins with PWM
- 16 Analog Input Pins
- 40 mA DC Current for I/O pins
- 50 mA DC Current for 3.3V Pins
- Flash Memory 256 KB
- Boot loader 8KB
- SRAM 8 KB
- EEPROM 4 KB
- CLK speed 16 MHz
- The USB host chip used in this is MAX3421E
- Length 101.52 mm
- Width 53.3 mm
- Weight 36 g (very light weighted)

2.3.1 Arduino mega 2560 pinout

Pin configuration [17] for Arduino mega 2560 is given below:

- Pin 3.3V & 5V – provides voltage
- Reset (RST) Pin- resets the microcontroller
- Vin Pin- external voltage provider (7V to 20V)
- Serial Communication- TX & RX pins for serial communication

- External Interrupts-6 pins for interrupt
- Led- pin no 13 for built-in LED
- AREF- for reference voltage
- Analog Pins-A0-A15 for I/O pins
- I2C- pin 20 & 21 for I2C
- SPI Communication-4 pins for SPI

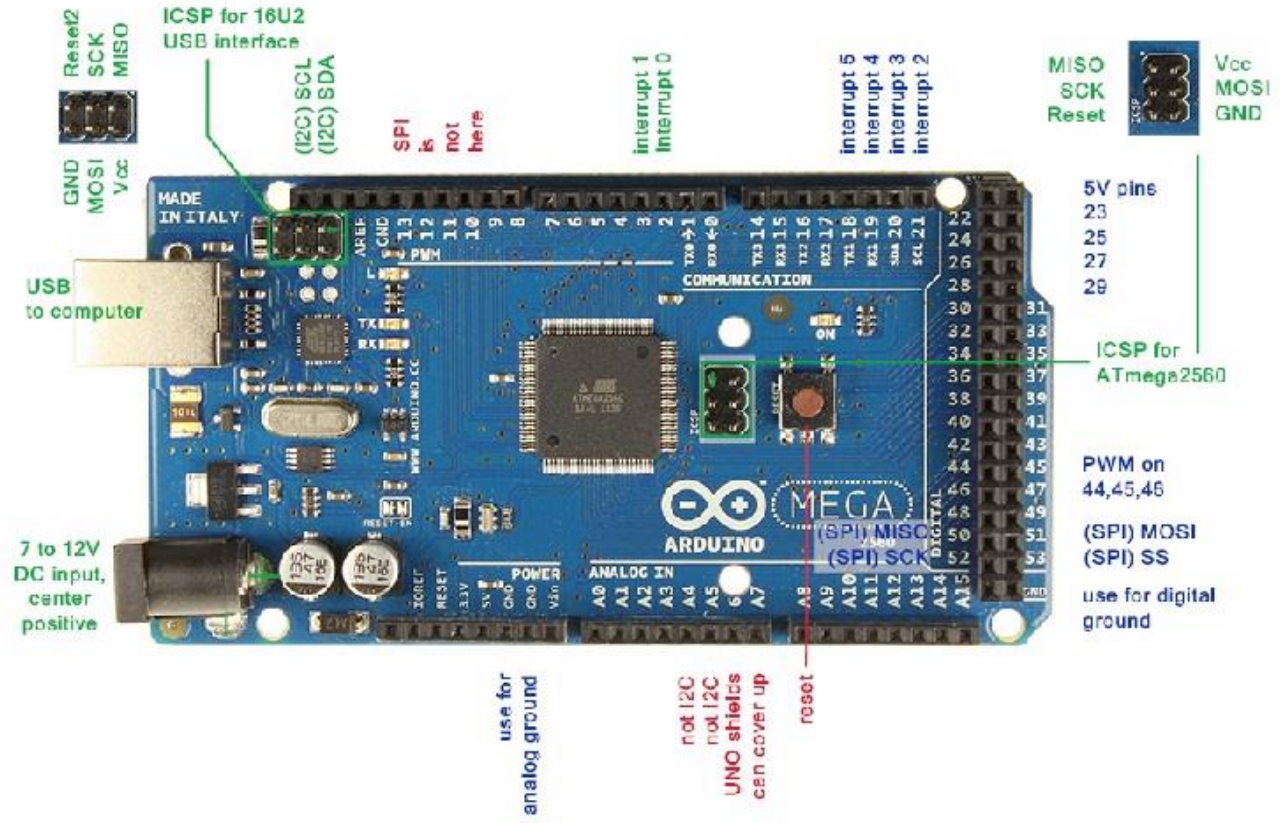


Figure-1: Arduino mega 2560 with pinout [17]

2.4 pH Sensor

The pH meter measured the activity of the hydrogen ion of the liquid. Different scale of hydrogen ion signifies the pH of the water-based solution. It's usually told that the solution is either acidic or alkaline. The scale of pH is 0-14, where pH 0-6 defined acidic and 8-14 defined alkaline. pH 7 is a neutral solution. In different scientific researches, pharmaceutical

industry, and chemical industry, the dye industry, it's required to test the pH of the solution. pH sensor is widely used for that purpose. pH sensor with Arduino or Raspberry Pi is a very cost-friendly pH sensor and that is very easy to handle. The driver board of the kit support 3.3V and 5V both. Arduino and Raspberry Pi are very sufficient to use with the Grove connector and stander BNC probe interface.[18]

Figure-2: pH Sensor [18]

Features:

- All ever used in scientific researches, pharmaceutical industry and chemical industry, dye industry.
- Budge friendly and easy to handle.
- Support with Arduino and Raspberry pi.
- Resolution most ±0.15PH (STP)
- Prove replaceable.

Specification: the specification of a pH meter is given below:

Table-6: Specification of pH meter [18]

Items	Values
Operating voltage	3.3V/5V
Range	0-14 pH
Resolution	±0.15 pH (STP)
Response time	<1min
Probe interface	BNC
Measure temperature	0-60℃
Internal resistance	≤250MΩ（25℃）
Alkali error	0.2PH（1mol/L）Na+，PH14)（25℃）

Cautions:

- The electrode should be marked with a stander buffer solution with a known pH value. The known pH value need to be feasible and intimated to the measured value.
- After the ending of measurement, the electrode protective sleeve needs to be closed and its needs to be wet with 3.3mol / L potassium chloride solution.
- To prevent short circuits, the leading end of the electrode should be dry and clean.
- Gradient inaccuracy may be arising due to long term use. In this case, the lower end of the electrode needs to be soaked under 4% of hydrofluoric acid for 3-5 seconds. Then washed with distill water and soak under potassium chloride solution.

Part list: the part list of pH kit is given below:

Table-7: Parts of pH kit [18]

Items	Quantity
pH prober	1
Grove cable	1
Driver board	1
Saturated sodium chloride solution	1

2.5 LDR Sensor

Light Dependent resistor (LDR) or photoresistor is a kind of electronic component which is sensitive to light. The resistance of LDR changes when light falls upon it. There is a different magnitude of changing resistance depending on the level of light that falls on it. The values of resistance of LDR usually are several ohms in darkness and bright light it's a few hundred ohms. However, LDRs are easy to use in a wide range of resistance and there are a lot of LDRs are available. The wavelength of incident light also affects the sensitivity of the resistors.

LDRs are composed of semiconductor material to have its light-sensitive properties. Cadmium sulfide, CdS is one of those materials. However, it’s forbidden for Europe because of the negative effects of cadmium on the environment. Lead sulfide, PbS, and indium antimonide, InSb is used for other options. Semiconductor materials are passive devices as they possess a PN junction. This property separates the LDRs from the other photodetectors like phototransistors and photodiodes [19].

Figure-3: LDR sensor [19]

The working mechanism of LDRs is not complicated. The movement of the electron produces movement of current inside the materials. A large number of free electrons of a good conductor moves in a given direction under the action of potential difference. On the other hand, insulators have high resistance and very few free electrons. Thus no current flows through them. LDR is made with high resistance semiconductors so that they have electrons locked in the crystal lattice. This state is called high LDR resistance. When light falls on top of it, photons are absorbed by the semiconductor materials. This causes the electron of the semiconductors to gain energy and goes to higher states. Thus free electrons start flowing, which creates electricity. As a result, the resistance of the LDR drops. Depending on the amount of lowered resistance, we can determine whether is it morning, noon, afternoon, or night.

2.6 Turbidity Sensor

The gravity Arduino turbidity sensor is a water quality monitor that measures the level of turbidity of the water. By transmitting and scattering the light it measures the suspended particle in water. Suspended solid that can change the light scattering and transmitting rate by which the amounts of solid can be measures. The increase of Total suspended solids (TSS) is defined to increase the turbidity level. In the wastewater, Stream, and river water, settling ponds, effluent and laboratory measurement, sediment transport, turbidity sensors are used. Analog and digital signal output modes can be found by this liquid sensor. In a digital signal, the mode threshold is adjustable [20].

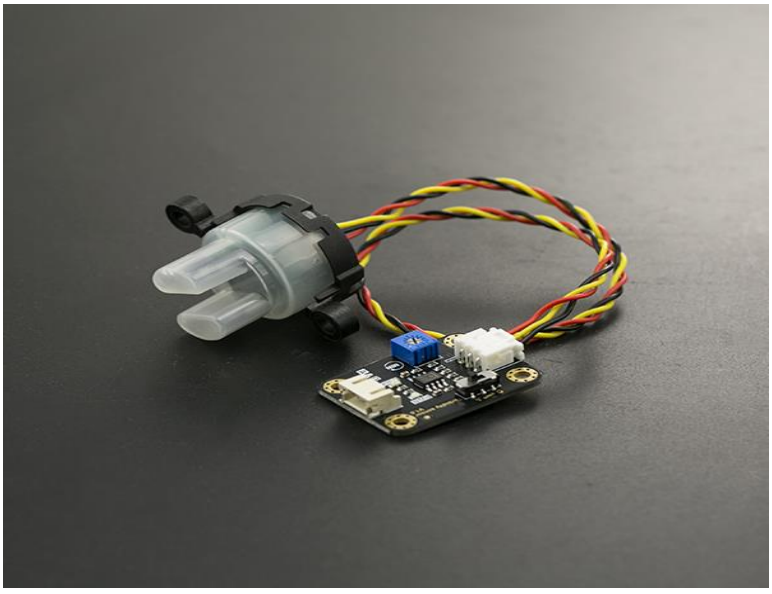


Figure-4: Turbidity sensor [20]

Specification:

- Operating Voltage: 5V DC
- Operating Current: 40mA (MAX)
- Response Time : <500ms
- Insulation Resistance: 100M (Min)
- Output Method:
    - Analog output: 0-4.5V
    - Digital Output: High/Low-level signal (you can adjust the threshold value by adjusting the potentiometer)
- Operating Temperature: 5℃~90℃
- Storage Temperature: -10℃~90℃
- Weight: 30g
- Adapter Dimensions: 38mm*28mm*10mm/1.5inches *1.1inches*0.4inches

2.7 TDS Sensor

It is an Arduino compatible kit for TDS meter that measured the cleanliness of the water. It is used to measure hydroponic, domestic water, and another purpose of testing water quality. Total Dissolved Solids (TDS) means the amount of dissolved solid in one liter of water. The higher TDS value indicates a higher amount of solids dissolve in one liter of water and lower cleanliness. TDS value can be used as a reference for measuring water purity.

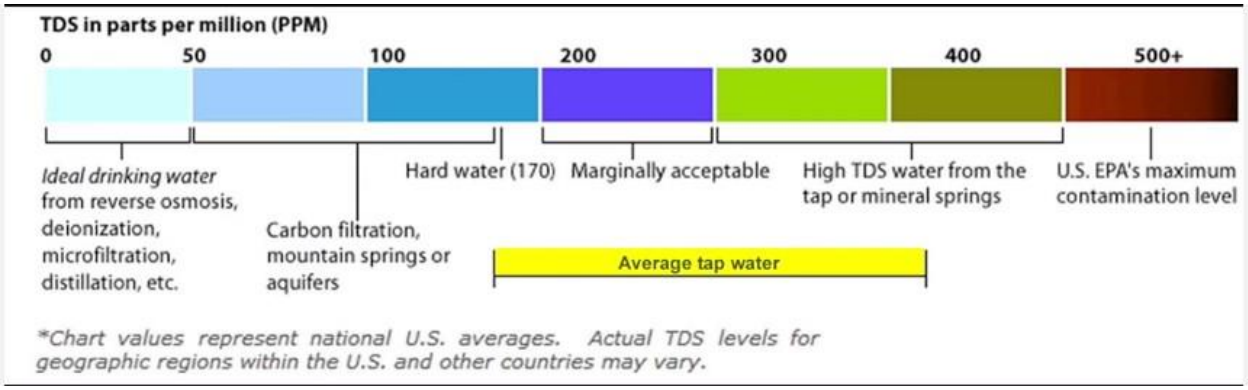


Figure-5: TDS range of contamination level [21]

To specify TDS value, a TDS pen is used that is very cost-effective and easy to handle. However, in some water analysis, it cannot transmit data to the control system for online monitoring. To send data to the control system, high accuracy instrument is required in which the price is out of the budget of the ordinary people. To solve this problem, an analog TDS sensor kit that is compatible with Arduino and that is easy to use. TDS measurement is easy with matching with Arduino controller.

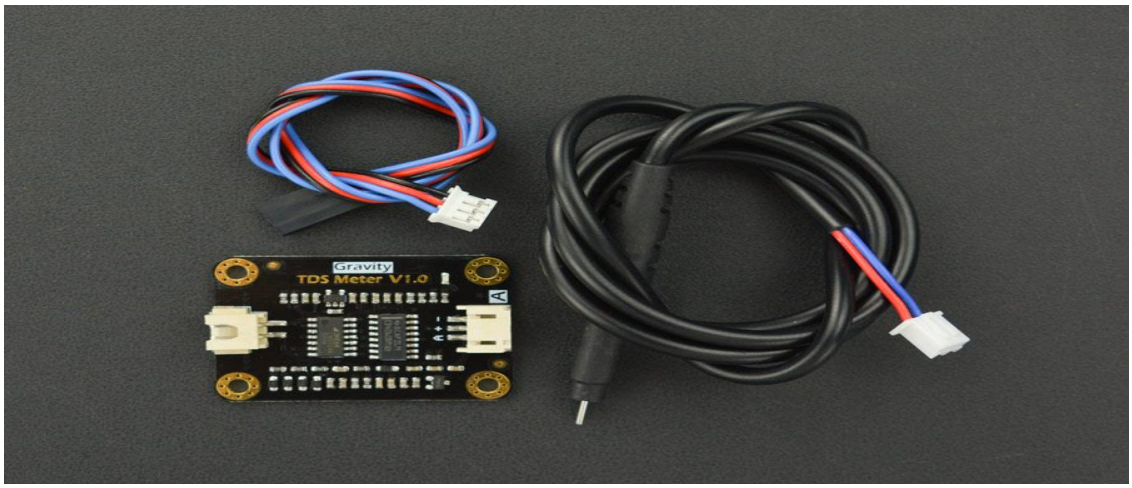


Figure-6: TDS sensor [21]

This sensor support 3.3 ~ 5.5V wide voltage input and 0 ~ 2.3V analog voltage output which is compatible with a 5V or 3.3V control system or board. To prevent the probe from polarization excitation of the source of the AC signal is required. Meanwhile, the output signal becomes more stable and longer the life of the probe. The measurement can be continuing for a long time as the TDS probe is waterproof [21].

Attention:

- The probe is not applicable above 55 degrees centigrade water temperature.
- For an accurate reading, the probe needs to be kept far from the edge of the container.
- Care must be taken in handling the connector and the signal transmitter boards as those are not waterproof.

Features:

- Wide Voltage Input: 3.3~5.5V
- Good Compatibility Output: 0~2.3V analog signal output, compatible with 5V or 3.3V controller
- AC Excitation Source: effectively prevent the probe from polarization
- Waterproof Probe
- Easy to Use: Arduino compatible, simple connection, plug, and play without soldering

Specification:

- Signal transmitted board
- Input Voltage: 3.3 ~ 5.5V
- Output Voltage: 0 ~ 2.3V
- Working Current: 3 ~ 6mA
- TDS Measurement Range: 0 ~ 1000ppm
- TDS Measurement Accuracy: ± 10% F.S. (25 ℃)
- Module Size: 42 * 32mm
- Module Interface: PH2.0-3P
- Electrode interface: XH2.54-2P probe

- Number of Needle: 2
- Total Length: 83cm
- Connection Interface: XH2.54-2P
- Color: Black
- Other: Waterproof Probe

2.8 Rain Sensor

A rain sensor is made of a rain detector plate having a comparator that manages the intelligence.

There are tapes in the printed circuit and when rain falls on top of it, it can detect it because of short-circuiting. Its action is changing the variable resistance status. When the sensor gets wet, the resistance increase, and when it is dry the resistance is lowed. The digital output (0/1) and analog output (0 to 1023) are the two comparator outputs that are connected to the rain sensor [22].

Figure-7: Rain sensor [22]

Sensor characteristics:

- Voltage : 3, 3v-5v
- Sensor Dimension : 3.9 x 5.4 cm
- Sensitivity potentiometer
- 2 control led

2.9 Esp8266 Wi-Fi module

The ESP8266 is a low-cost Wi-Fi microchip, with a full TCP/IP stack and microcontroller capability, created by Espressif Frameworks in Shanghai, China. The chip first came to the consideration of Western creators in Admirable 2014 with the ESP-01 module, made by a third-party producer Ai-Thinker. The ESP8266 could be a truly valuable cheap WiFi module for controlling gadgets over the Internet. It can work with a micro-controller just like the Arduino or it can be modified to work on its claim. The ESP8266 comes with factory introduced firmware permitting you to control it with standard "AT commands".

In an ordinary working mode, the ESP module expends almost 70mA which leads to depleting three AAA batteries within a couple of hours. It is exceptionally unreasonable for inaccessible IoT gadgets. All things considered, it is conceivable to put the ESP8266 into a deep sleep and wake up intermittently to enact Wi-Fi and send information as it were when required. [23]

Pin configurations:

1. -(Ground)- Connected to the ground of the circuit
2. -(TX)- Connected to Rx pin of programmer/uC to upload program or Can act as a general-purpose input/output pin when not used as TX
3. -(GPIO-2)- General purpose Input/output pin
4. -(CH-EN)- Chip Enable – Active high
5. -(GPIO-0)- General purpose Input/output pin or Takes module into serial programming when held low during start-up
6. -(Reset)-Resets the module
7. -(GPIO-3)- General purpose Input/output pin
8. -(VCC)- Connect to +3.3V only

ESP8266 has three operational modes. Those are the followings:

- Access Point (AP) — In AP, the Wi-Fi module works as a Wi-Fi network or access point. It permits other gadgets to associate with it. And builds up two-way communication between the ESP8266 and the gadget that's associated with it through Wi-Fi.
- Station (STA) — In STA mode, the ESP-01 can be interfaced to an AP, such as the Wi-Fi network from your house. This permits any gadget associated with that network to communicate with the module.
- Both — in this model ESP-01 act as both an AP as well as in STA mode.

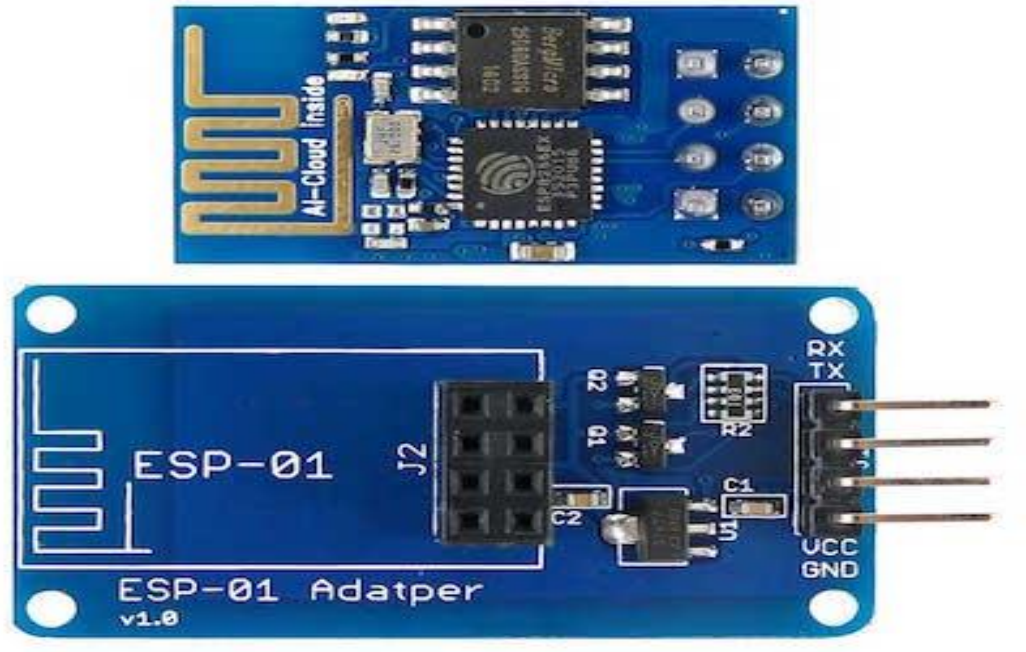


Figure-8: ESP8266 Wi-Fi module [23]

2.10 Rechargeable Battery

For our device to be portable, it needs its power source. For this reason, we used a Lithium-Ion Battery. This battery is called an **11.1V 1100mAh 3S 30C LiPo battery.** [24]

Features:

1) High Capacity Battery
2) Higher voltage capacity for single battery cells
3) Pollution-free
4) Longer cycle life
5) Has no memory-effect
6) Time consistency is good.
7) Safe and reliable
8) Lightweight and small.
9) Low self-discharge

Specifications:

1. Model: Li-po 3S
2. Max Capacity: 1100 mAh
3. Voltage level: 11.1V
4. Continuous Discharging rate: 30C
5. Max charging rate: 3C

Figure-9: Rechargeable Battery [24]

2.11 Schematics

To build the device, we 1st created a schematic diagram using Proteus Software. That way we can create a general idea regarding the connections of our device. In the hardware section, the major component is the Arduino mega 2560. Everything else is connected to it. All the sensors (pH Sensor, LDR Sensor, Turbidity Sensor, TDS Sensor & Rain Sensor) have a Ground pin & a 5V power pin. These pins are connected to the Ground pin & a 5V power output pin of the Arduino Board respectively. The connection is made through jumping wires and breadboard. The ESP8266 Wi-Fi module has 5V & 3.3V power pins which are connected to the Arduino Mega's 5V & 3.3V power output pins.

The major pin connections are given below:

1. ESP8266 Wi-Fi Module Tx pin - Arduino mega 2560 Tx pin
2. ESP8266 Wi-Fi Module Rx pin - Arduino mega 2560 Rx pin
3. pH Sensor Data pin - Arduino mega 2560 A0 pin
4. LDR Sensor Data pin - Arduino mega 2560 A3 pin
5. Turbidity Sensor Data pin - Arduino mega 2560 A1 pin
6. TDS Sensor Data pin - Arduino mega 2560 A4 pin
7. Rain Sensor Data pin - Arduino mega 2560 A2 pin
8. LiPo Battery's positive pin - Arduino mega 2560 Power Pin
9. LiPo Battery's Ground pin - Arduino mega 2560 Ground pin

The Proteus Schematic is given below:

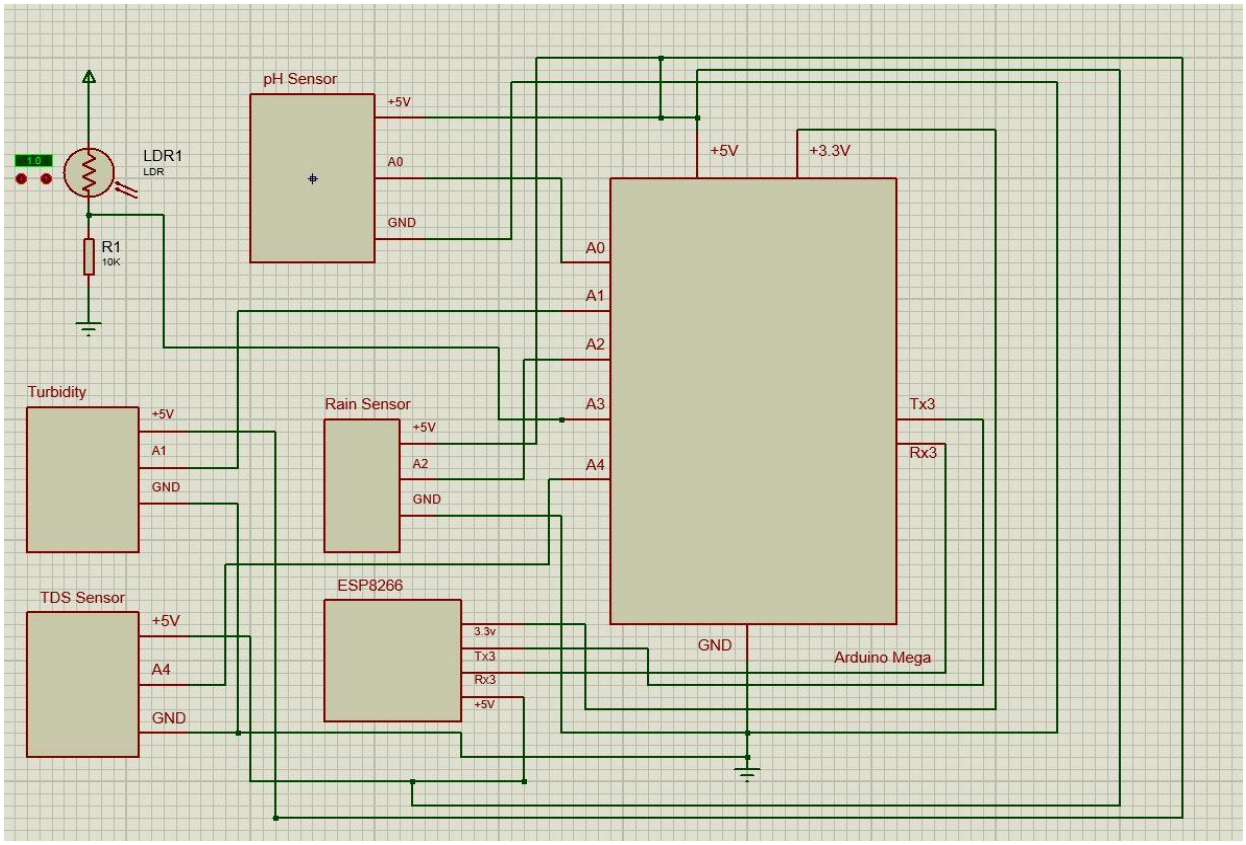


Figure-10: Schematic of the Device & connection points

2.12 Floating mechanism

Our device needs a floating mechanism to work. This is because the sensors need to be submerged in water for a minimum of an hour the gather a good amount of Data from the water. Also, the device needs to be Water-resistant so that the Arduino mega 2560 & ESP8266 Wi-Fi Module do not get wet. If they get wet, they will become non-functional. So we put the Arduino mega 2560, ESP8266 Wi-Fi Module, and the sensors modules inside a plastic box and kept the sensors & probes outside of the box. Then we sealed the box using tape. We kept the battery outside also for ease of Powering on & off the device but we made sure it was warped in tape properly so that water cannot get inside. That is how we made the device water-resistant. For the floating part, we used the ESP sheet (Polystyrene sheet). These sheets are very light weighted and can float above water. We used a big ESP box and cut the center of it. We placed the device in the center of the ESP box and connected the box with some rope so that we can send it to the center of our water source. An image of our floating mechanism and device is given below:

Figure-11: Floating mechanism & device

2.13 Hardware Limitations

Despite having such high preparations, we faced some issues regarding our hardware device. These limitations are given below:

- One of the major issues we faced was regarding data collection speed. Initially, we wanted to gather data from each sensor every second. However, some of these sensors (pH, TSD & Rain) cannot read data every second. So we had to gather data every minute instead of every second. So data collection became limited.
- Some of the sensors we equipped were not compatible with the Arduino Mega 2560. For instance, we connected a water flow sensor which did not give accurate readings rather caused our sensors to malfunction.
- The pH sensor being chemical-based kept on malfunctioning. So we had to buy 3 of them just to provide data correctly.
- In WQM two of the major factors are CO2 and O2 levels. However, we could not buy sensors of them as they were extremely expensive and non-reliable.

- The sensors did not provide exactly accurate data which caused the issue in predicting data.
- ESP8266 Wi-Fi Module sometimes dropped connection for unknown reasons.
- Data collecting was limited to 2 hours per day due to charge capability.

2.14 Conclusion

In this chapter, the overall description of the sensors we have used: pH sensor, rain sensor, LDR, TDS sensor and, turbidity sensor has been given. By looking into this chapter, one can easily find the mechanism, properties as well as safety measures for the sensors. The schematic diagram which indicates the connection of every sensor with the Arduino Mega board can be found here also. Since the device is implemented for monitoring the water quality; it should be kept in such housing so that it can float on water sources easily. Therefore, the floating mechanism of our device has been described here. Moreover, there are still some limitations of the sensors as we are using the available sensors during this tough time, the challenges we have faced regarding the sensors are also provided. In the coming chapter, the process of sending data to the server will be provided in detail.

Chapter 3
IoT & Live Data

3.1 Introduction of IoT

The IoT is a system of interconnected devices with unique identifiers that can transfer data over a network without the help of human-to-human interaction. In other words, the Internet of Things refers to the uncountable number of electrical devices around the globe that connects to the internet and constantly collects and shares data.[25] An IoT system has various parts. It normally includes a web-enabled smart device or multiple devices. These devices use embedded systems for functioning and are equipped with sensors and communicative chips or hardware. These devices collect and send data that they get from their environments. IoT devices can share their sensory data to the cloud via an IoT gateway. Then that data can be analyzed. Sometimes, these IoT devices communicate with other IoT devices and can act on the information they acquire from each other. The devices can function without any human intervention, even though anyone can interact with the devices. For instance, people can set these devices up and give instructions to them or access their data, etc. The term “IoT” is used mainly for devices that normally would not have any internet connection but has internet connection so that it can communicate with a network without any human action. That is why a PC or smartphone is not an IoT device because it is expected to have internet connectivity even if it is filled with sensors.

3.2 Thingspeak

The analysis, visualization, and aggregation of live data can be instantly performed by an IoT analytic platform which is called Thingspeak. Thingspeak gives a visualization of posted data from the smart devices connected to it. Any smart devices connected to thingspeak can provide data analysis through a graph or plot. Online processing and analysis of data can be performed by executing MATLAB code in it. The system which requires analytics often uses thingspeak for proving of concept and prototyping.[26] With thingspeak sending data of sensors to the cloud can be done seamlessly. In thingspeak analyzing and visualizing data is possible via MATLAB. Moreover, a certain action can be provoked using thingspeak. It also

provides basic security and is accessed via a password. In short, thingspeak is a Gateway for IoT devices.

Features:

- Real-time data visualization.
- Sending data to the thinkspeak is easy.
- Can use MATLAB for data analysis.
- Schedules can be run automatically.
- It does not require developing web software or setting up servers for building and prototyping the loT system.
- Action on data and connection with third-party services are automatic.

3.3 Implementation of Thingspeak & Esp8266 Wi-Fi module

An IoT system mainly has 2 major components. They are IoT devices and IoT gateway. In our case, the Esp8266 Wi-Fi module acts as a part of the IoT device, and thingspeak acts as the IoT gateway. So basically, the Arduino mega 2560 gathers the data from the sensors and sends them to the ESP8266 Wi-Fi Module. The Wi-Fi Module sends data to the cloud using thingspeak. The ESP8266 Wi-Fi Module is connected to a smartphone Mobile Hotspot. That is how the module connects to thingspeak. In thingspeak, the Graphs of each sensor can be seen. They are the live stream data that is plotted in thingspeak. As we gather more data, the more accurate these graphs get. This helps to get a more accurate prediction of future data and to get lower error rates.

3.4 Code

Code and its description is given below:

Wifi access points and passwords:

```
String AP = "net dimu na";

String PASS = "lollollolString

//String AP = "iPhone";
```

```
//String PASS = "NirAhmed";

API = "ADX4VA06O7ZPO4YP";
```

Initializing ports and variables:

```
String HOST = "api.thingspeak.com";

String PORT = "80";

String field = "field1";

String field2 = "field2";

String field3 = "field3";

String field4 = "field4";

String field5 = "field5";

String field6 = "field6";

int countTrueCommand;

int countTimeCommand;

boolean found = false;

float valSensor = 1;

int valSensor2 = 2;

int valSensor3 = 3;

int valSensor4 = 4;

int valSensor5 = 5;

int valSensor6 = 6;

int x = 0;

int y = 15;
```

Flow sensor initialization:

```
volatile int NbTopsFan;
```

```
int Calc;

int hallsensor = 2;

void rpm (){

 NbTopsFan++;

}
```

Esp8266 pin connection & variable declaration:

```
const int analogInPin = A0;

int sensorValue = 0;

unsigned long int avgValue;

float b;

int buf[10];

int temp=0;
```

Water flow sensor connection:

```
void setup()

{

   pinMode(hallsensor, INPUT);

  attachInterrupt(0, rpm, RISING);

  Serial.begin(9600);

  Serial3.begin(115200);

 sendCommand("AT",5,"OK");

  sendCommand("AT+CWMODE=1",5,"OK");

  sendCommand("AT+CWJAP=\""+ AP +"\",\""+ PASS +"\"",20,"OK");

}
```

Code for printing sensor data:

```
void loop()
{
ph();
 Serial.print("valSensor = ");
 Serial.println(valSensor);
turb();
 Serial.print("valSensor2 = ");
 Serial.println(valSensor2);
israin();
 Serial.print("valSensor3 = ");
 Serial.println(valSensor3);
  daynight();
 Serial.print("valSensor4 = ");
 Serial.println(valSensor4);
 tds();
 Serial.print("valSensor5 = ");
 Serial.println(valSensor5);
SendToServer();
}
```

Sending sensor Data to cloud:

```
void SendToServer(){
String getData = "GET /update?api_key="+ API +"&"+ field +"="+String(valSensor);
```

```
sendCommand("AT+CIPMUX=1",5,"OK");
 sendCommand("AT+CIPSTART=0,\"TCP\",\""+ HOST +"\","+ PORT,15,"OK");
 sendCommand("AT+CIPSEND=0," +String(getData.length()+4),4,">");
Serial3.println(getData);
 for(x = 0; x<y; x++){
 delay(1000);
 }
  String getData2 = "GET /update?api_key="+ API +"&"+ field2 +"="+String(valSensor2);
sendCommand("AT+CIPMUX=1",5,"OK");
 sendCommand("AT+CIPSTART=0,\"TCP\",\""+ HOST +"\","+ PORT,15,"OK");
 sendCommand("AT+CIPSEND=0," +String(getData2.length()+4),4,">");
  Serial3.println(getData2);


for(x = 0; x<y; x++){
 delay(1000);
 }
 String getData3 = "GET /update?api_key="+ API +"&"+ field3 +"="+String(valSensor3);
sendCommand("AT+CIPMUX=1",5,"OK");
 sendCommand("AT+CIPSTART=0,\"TCP\",\""+ HOST +"\","+ PORT,15,"OK");
 sendCommand("AT+CIPSEND=0," +String(getData2.length()+4),4,">");
  Serial3.println(getData3);


for(x = 0; x<y; x++){
 delay(1000);
```

```
 }


 String getData4 = "GET /update?api_key="+ API +"&"+ field4 +"="+String(valSensor4);
sendCommand("AT+CIPMUX=1",5,"OK");
 sendCommand("AT+CIPSTART=0,\"TCP\",\""+ HOST +"\","+ PORT,15,"OK");
 sendCommand("AT+CIPSEND=0," +String(getData2.length()+4),4,">");
  Serial3.println(getData4);


for(x = 0; x<y; x++){
 delay(1000);
 }


  String getData5 = "GET /update?api_key="+ API +"&"+ field5 +"="+String(valSensor5);
sendCommand("AT+CIPMUX=1",5,"OK");
 sendCommand("AT+CIPSTART=0,\"TCP\",\""+ HOST +"\","+ PORT,15,"OK");
 sendCommand("AT+CIPSEND=0," +String(getData2.length()+4),4,">");


 Serial.println(getData2.length());
 Serial.println(".............");
  Serial3.println(getData5);


 for(x = 0; x<y; x++){
 delay(1000);
 }
```

```
 countTrueCommand++;
  sendCommand("AT+CIPCLOSE=0",5,"OK");
}
```

Code for LDR sensor:

```
void daynight(){
  int sun = analogRead(A3);
  if(sun<100){
    Serial.println("Morning");
    valSensor4 = 1;
  }
  else if(sun<500){
    Serial.println("Afternoon");
    valSensor4 = 2;
  }
  else{
    Serial.println("Night");
    valSensor4 = 3;
  }
  delay(10);
}
```

Code for rain sensor:

```
void israin(){
  int rain = analogRead(A2);
  if(rain<500){
    Serial.println("Raining");
    valSensor3 = 1;
  }else{
 Serial.println("No Rain");
    valSensor3 = 0;
  }
  delay(10);
}
```

Code for water flow sensor:

```
void flow(){
 NbTopsFan = 0;
 sei();
 delay (1000);
 cli();
 Calc = NbTopsFan*8;
 Serial.println(Calc);


}
```

Code for turbidity sensor:

```
void turb(){
```

```
   Serial.print("Turbidity = ");
   Serial.println(analogRead(A1));
   delay(10);
   valSensor2 = analogRead(A1);
}
```

Code for pH sensor:

```
void ph(){
 for(int i=0;i<10;i++)
 {
 buf[i]=analogRead(analogInPin);
 delay(10);
 }
 for(int i=0;i<9;i++)
 {
 for(int j=i+1;j<10;j++)
 {
  if(buf[i]>buf[j])
  {
   temp=buf[i];
   buf[i]=buf[j];
   buf[j]=temp;
  }
 }
 }
```

```
avgValue=0;
for(int i=2;i<8;i++)
avgValue+=buf[i];

float pHVol=(float)avgValue*5.0/1024/4.3;
float phValue = -5.70 * pHVol + 22.8;
phValue=14.2-phValue;
  Serial.print("sensor = ");
 Serial.println(phValue);

delay(50);
valSensor = phValue;
}
```

Code for TDS sensor:

```
void tds(){
  valSensor5 = analogRead(A4);
}
int getSensorData(){
  }
```

Code for control & error detection:

```
void sendCommand(String command, int maxTime, char readReplay[]) {
  Serial.print(countTrueCommand);
  Serial.print(". at command => ");
  Serial.print(command);
```

```
Serial.print(" ");
while(countTimeCommand < (maxTime*1))
{
  Serial3.println(command);//at+cipsend
  {
    found = true;
    break;
  }
   countTimeCommand++;
}
if(found == true)
{
  Serial.println("OYI");
  countTrueCommand++;
  countTimeCommand = 0;
}
if(found == false)
{
  Serial.println("Fail");
  countTrueCommand = 0;
  countTimeCommand = 0;
}
found = false;
}
```

3.5 Selected Water Source

For our thesis project, we have used the water from a small lake in Mymensingh. To analyze the quality of water, we have not chosen any domestic water rather we chose a source where we would be able to see variations of toxicity in the water from time to time. The water of that small lake is not always pure. Various causes are possible to pollute the lake water. Mainly, the local people who live beside that lake use it to wash clothes, to take shower, and for all the household chores daily. So, these incidents pollute the water and as a result, we expect to get higher pH, turbidity, and TDS values. Also the lake is at the border of a fishery research institution named **BFRI (Bangladesh Fisheries Research Institute)**. While researching fishes, they sometimes use chemicals that directly dissolve in the water of that lake. Thus, the water gets highly polluted. Some pictures of the lake are given below:

Figure-12: Selected Lake for Data collection located at Mymensingh picture-1

Figure-13: Selected Lake for Data collection located at Mymensingh picture-2

3.6 Limitations of Thingspeak

There are some limitations of thingspeak such as,

- The minimum time to send two consecutive data in thingspeak is 15 seconds. Since we wanted to collect data every second, it became an issue that we could not overcome.
- Sometimes it crashes and causes Data uploading to stop.
- There was a bug in thingspeak which causes the time zone to change automatically.

3.7 Conclusion

The intelligent connectivity of physical devices with the internet is known as the Internet of Things or IoT. Sending the data collected by the device to the server was not an easy task. At first, NodeMCU was used for sending the data to the server. However, it was a difficult task to implement, and then we have used the ESP8266 Wi-Fi module for sending the data. As the

cloud, ThingSpeak has been used. ThingSpeak is an open platform for IoT based projects and a very user-friendly website. Therefore, we have gone for this website however, there are some limitations of this website, such as; it cannot be used for free for a lifetime. And for additional data, the storage must be purchased or there should be a subscription to this website. Apart from these, it is a very helpful website. We had to implement an algorithm that can predict the future of the water of the source means the probability of the water's being pure or impure. In the upcoming chapter, there will be a vast discussion about the algorithms.

Chapter 4

Algorithm & Future Prediction

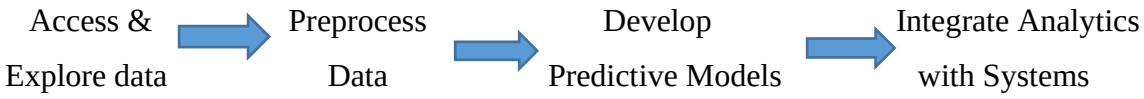
4.1 Introduction

Using an algorithm to predict future data captures an important trend in new technology. To predict future events, predictive analytics uses previous and present data. Different algorithms have different models to predict a new data point. To get an optical outcome, these statistics of historical data are used to predict what is going to happen next or they can suggest what actions should take to get necessary outcomes.

In the last few years, using an algorithm to predict future data has received a lot of attention because of having a lot of advantages in supporting new technology especially in the areas of big data and machine learning. There are millions of ways for us to share our intentions, opinions, and beliefs with words. The problem is that one person could express themselves in a completely different way to another. With this in mind, a rule-based machine can't understand what we say. This is where technology has always failed to understand the human language -- until now.

Predicting future in different industries like automotive, aerospace, energy production, financial services, industrial automation, medical devices, and so on.

Predictive Analytics Workflow



In our paper, we have used an algorithm, for predicting future trends and detecting period for the future. Giving a repeating pattern, our process can automatically find out its period and make predictions on its future values. To the best of our knowledge, this is such an algorithm which can automatically find the period of that inputs and further use it to predict the future trend. The theoretical analysis and simulation results have also been discussed in this paper. [27]

4.2 Machine Learning Algorithms

Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. Machine learning helps the IT system to verify new sequences based on existing algorithms and sets the data to develop the solution concepts. In machine learning, we can generate artificial knowledge based on experiences.

Machine learning and Artificial intelligence (AI) are currently trending topics in the IT sector. But it's not just hype — it's a technology, many organizations in every industry are already stated adopting and getting value from it. The reason stems from the unlimited use cases where machine learning can play a vital role from self-driving cars to fraud detection, and even face recognition.

Machine learning has evolved from the study of computational learning theory and pattern recognition. No other method is so effective as Machine Learning used in the field of data analysis to predict something by devising some models and algorithms. These analytical models allow researchers, engineers, data scientists, and analysts to produce reliable and valid results and decisions. It also helps to discover some hidden patterns or features through historical learnings and trends in data. Feature selection is the most important task of machine learning. Model is created based on the results gathered from the training data that is why machine-learning algorithms are non-interactive. It studies the past observations to make precise predictions. It is a very difficult task to make an accurate prediction rule based on which algorithm can be developed. [29]

A system can perform the following tasks by Machine Learning:

- Finding, extracting, and summarizing relevant data
- making predictions based on the analysis data
- Calculating probabilities for specific results
- Adapting to certain developments autonomously
- Optimizing processes based on recognized patterns

4.3 KNN

The k-nearest neighbors (KNN) algorithmic program could be an easy, easy-to-implement supervised machine learning algorithmic program which will be accustomed to solving each classification and regression issues.

KNN (k-nearest neighbor) is an associate degree extensively used classification algorithmic program as a result of its simplicity, easy implementation, and effectiveness. It is the best in every of the highest 10 data processing algorithms, that has been widely applied in variety of fields. KNN has few shortcomings poignant and its accuracy of classification is awesome. It's massive memory necessities moreover as time quality and many techniques are projected to enhance these shortcomings in literature also. During this paper, we've 1st reviewed some enhancements created within the KNN algorithmic program. Then, we've projected our thesis using improved algorithmic program. It's a mix of dynamic hand-picked, attribute weighted and distance weighted techniques. We have, by experimentation, tested our projected algorithmic program in web Beans IDE, employing a common place UCI dataset-Iris. The accuracy of our algorithmic program is improved with a mix of classification and clump techniques. Experimental results have proved that our projected algorithmic program performs higher than the traditional KNN algorithmic program.

Nowadays the exponential growth of the generation of matter documents and also the nascent have to be compelled to structure them to increase the eye to the automatic classification of documents into predefined classes. There is a large variety of supervised learning algorithms that modify text classification. This paper deals with associating degree approach for building a machine learning system in R that uses the K-Nearest Neighbors (KNN) technique for the classification of matter documents. The experiment was dedicated to difficult things of the KNN algorithmic program that to seek out the right price of k that represents the quantity of neighbors. KNN algorithmic program will be used for each classification and regression issues. The KNN algorithmic program uses 'feature similarity' to predict the values of any new knowledge points. This suggests that the new purpose is allotted a price supported algorithm. However, closely it resembles the points within the coaching set.

4.4 Linear regression

Straight relapse endeavors to show the relationship between two factors by fitting a direct condition to watched information. One variable is considered to be an informative variable, and the other is considered to be a subordinate variable. For illustration, a modeler might need to relate the weights of people to their statures employing a straight relapse model. Before endeavoring to fit a direct show to watched information, a modeler ought, to begin with determining whether or not there's a relationship between the factors of intrigued. This does not essentially suggest that one variable causes the other (for illustration, higher SAT scores don't cause higher college grades), but that there are a few noteworthy affiliations between the two factors. A scatterplot can be an accommodating apparatus in deciding the quality of the relationship between two factors. On the off chance that there shows up to be no affiliation between the proposed informative and subordinate factors (i.e., the scatterplot does not show any expanding or decreasing trends). at that point fitting, a direct relapse demonstrates to the information likely will not give a valuable show. An important numerical degree of affiliation between two factors is the relationship coefficient, which may be esteem between -1 and 1 showing the quality of the affiliation of the watched information for the two variables. [30]

A straight relapse line has a condition of the frame Y = a + bX, where X is the illustrative variable and Y is the subordinate variable. The slope of the line is b, and a is the captured (the esteem of y when x = 0).

1. Simple Linear Regression

With straightforward direct relapse when we have a single input, ready to utilize measurements to gauge the coefficients. This requires simply calculate factual properties from the information such as implies, standard deviations, relationships, and covariance. All of the information must be accessible to navigate and calculate statistics.

This is fun as work out exceed expectations, but not truly valuable in the hone.

2. Ordinary Least Squares

When we have more than one input able to utilize Conventional Slightest Squares to gauge the values of the coefficients.

The Ordinary Least Squares method looks to play down the entirety of the squared residuals. This implies that given a relapse line through the information we calculate the remove from each information point to the relapse line, square it, and whole all of the squared blunders together.

This approach treats the information as a lattice and employments direct variable based math operations to appraise the ideal values for the coefficients. It implies that all of the information must be accessible and you must have sufficient memory to fit the information and perform framework operations.

It is unusual to actualize the Standard Slightest Squares strategy yourself unless as a work out in straight variable based math. It is more likely that you simply will call a strategy in a direct polynomial math library. This strategy is exceptionally quick to calculate.

3. Gradient Descent

When there are one or more inputs you'll utilize a handle of optimizing the values of the coefficients by iteratively minimizing the blunder of the show on your preparing data. This operation is called Angle Plunge and works by beginning with irregular values for each coefficient. The entirety of the squared mistakes is calculated for each combination of input and yield values. A learning rate is utilized as a scale factor and the coefficients are upgraded within the course towards minimizing the blunder. The method is rehashed until a least whole squared mistake is accomplished or no advance change is possible.

When utilizing this strategy, you must select a learning rate (alpha) parameter that decides the measure of the change step to require each emphasis of the strategy.[31]

There are a lot of benefits of using linear regression. Some are:

Linear Assumption

Direct relapse expects that the relationship between your input and yield is direct. It does not back anything else. This may be self-evident, but it is nice to keep in mind once you have a lot of traits. You will get to change information to form the relationship direct (e.g. log change for an exponential relationship).

Remove Noise

Straight regression assumes that your input and yield factors are not boisterous. Consider utilizing information cleaning operations that let you way better uncover and clarify the flag

in your information. Typically most vital for the yield variable and you want to expel exceptions within the yield variable (y) if possible.

Remove Collinearity

Straight relapse will over-fit your information after you have exceedingly related input factors. Consider calculating pairwise relationships for your input information and evacuating the foremost correlated. [32]

Gaussian Distributions

Straight relapse will make more solid forecasts on the off chance that your input and yield factors have Gaussian dissemination. You'll get a few advantages utilizing your factors to create their dispersion more Gaussian looking.

4.5 ROC curve

ROC stands for Receiver Operating Characteristic. Showing graphically the connection between clinical sensitivity and specificity for every possible cut-off for a test or having a combination of tests. The optimization area under the ROC curve gives an idea about the benefits of using the tests. For collecting the most appropriate cut-off for a test, mostly ROC curves are used in clinical biochemistry. With the lowest false positive rate, we get the best cut-off for the highest true positive rate. For the usefulness of a test in general, the area under a ROC curve is measured. A greater area indicates a more useful test. For comparing the usefulness of tests, the area under ROC curves is used. In 1940, for detecting the Radio signals in a noisy area, ROC curves were used following Pearl Harbor. For showing the connections between any clinical sensitivity and any specificity for every possible cut-off or any combination of tests, ROC curves are used. [33]

There is another term AOC (area under the ROC curve) that presents the concept of all positive classification of thresholds.

4.5.1 How to make a ROC curve

We need to have a clear concept of true negative, true positive, false positive, and false negative. For comparing the result in a test with clinical truth, we need to establish the use of diagnostic procedures by not involving the test. [33]

ROC curve (receiver operating characteristic curve) shows the performance of classification thresholds. The curve below plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR=TP/(TP+FN)

False Positive Rate (FPR) is defined as follows:

FPR=FP/(FP+TN)

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

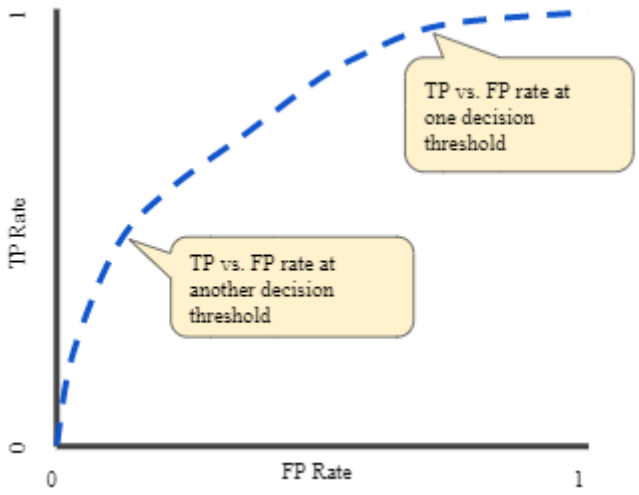


Figure-14: ROC Curve [33]

4.6 Confusion matrix

A confusion matrix represents the summary of predicting results in a classification problem.

We need to summarize the number of correct and incorrect predictions with count values by breaking them down with each other. This is the main factor to the confusion matrix.

It gives both in sighting the errors made by our classifiers and the types of errors that we made.

It is the breakdown that overcomes the limitation by using classification accuracy alone.

A confusion matrix (N*N) is used to evaluate the performance for the classification model when N will be the number of all the targeted classes. By using the machine learning model, this matrix can compare the actual targeted values with the help of those predicted values. This matrix gives us a model with that predicted values and a holistic view of how good our classification model.

In a binary classification problem, we might have a 2*2 matrix having 4 values as shown below:

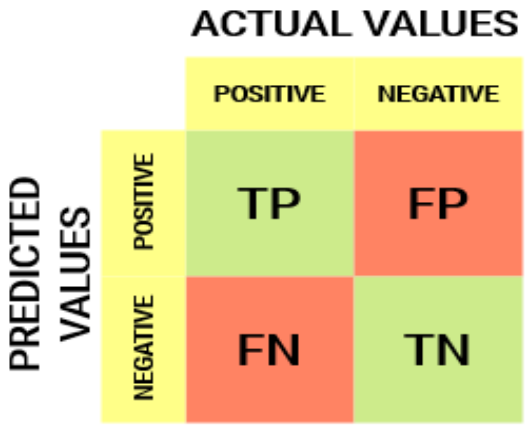


Figure-15: confusion matrix [36]

Let’s understand the matrix:

- The target variable has two values: Positive or Negative
- The columns represent the actual values of the target variable
- The rows represent the predicted values of the target variable [36]

Understanding True Positive, True Negative, False Positive, and False Negative in a Confusion Matrix

True Positive (TP)

The predicted value matches the actual value

The actual value was positive and the model predicted a positive value

True Negative (TN)

The predicted value matches the actual value

The actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error

The predicted value was falsely predicted

The actual value was negative but the model predicted a positive value

Also known as the **Type 1 error**

False Negative (FN) – Type 2 error

The predicted value was falsely predicted

The actual value was positive but the model predicted a negative value

Also known as the **Type 2 error** [36]

4.6.1 How to Calculate a Confusion Matrix

Below is the process for calculating a confusion Matrix.

- You need a test dataset or a validation dataset with expected outcome values.
- Predict each row in your test dataset.
- From the expected outcomes and predictions count:

1. The number of correct predictions for each class.
2. The number of incorrect predictions for each class, organized by the class that was predicted. [35]

Accuracy:

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally, the definition of accuracy is-

Accuracy = (number of correct predictions) / (Total number of predictions)

For binary representation, we can explain accuracy by positive and negative as below:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where TP= True Positive, TN= True Negative, FP = False Positive , FN= False Negative. [38]

4.7 Machine Learning

Without being explicitly programmed, Machine learning is a type of application of AI (artificial intelligence) which provides the ability for a system to learn automatically and improving from experience. Machine learning mainly focuses on the development of computer programming.

Machine learning focuses on the observation of the data, such as direct experience, instruction, or looking for patterns in data, or making better decisions for the future based on the data we provide.

Our focus is mainly allowing the computers to learn automatically by human intervention or any assistance and adjusting accordingly. [37]

Some Machine Learning Methods:

Machine learning algorithms are mainly categorized as supervised and unsupervised.

- **A supervised Machine learning algorithm** can be applied to any past value or data or using labeled examples for predicting future events. From the analysis, this algorithm produces the inferred function for predicting the output values. This system can provide our targets for any of our input after sufficient training. This algorithm

can compare any output with the original, intended output and finding errors by modifying the models accordingly.

- On other hand, an **unsupervised machine learning algorithm** refers to getting information that is used to train neither classified nor labeled. Unsupervised learning shows how the system infers a function describing any hidden structure from any unlabeled data. Though this system can't figure out the original output it can explore the data as well as can draw inferences from the dataset for describing the hidden structure from unlabeled data.
- **Reinforcement learning algorithms** refer between supervised learning and unsupervised learning. This learning uses both labeled and unlabeled data to train the typical and small amount of labeled data or any large amount of unlabeled data. When the acquired data can require any skilled or relevant resources to train it or learn from it, we can use Reinforcement Learning. Acquiring unlabeled data doesn't require any additional resources. [37]

4.8 Classification of Machine Learning

The classification of Machine learning is given below:

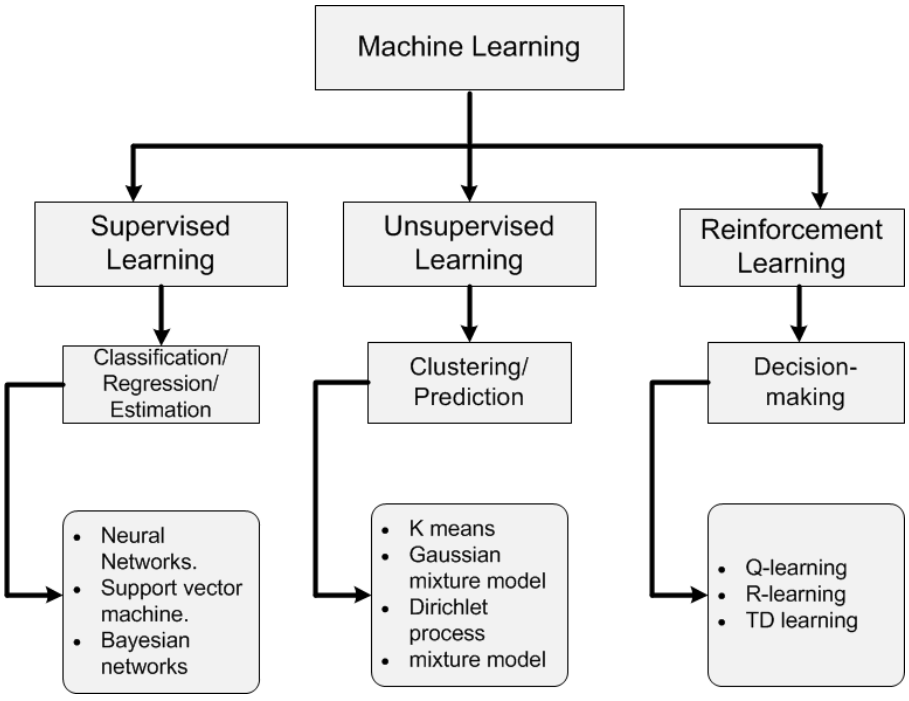


Figure-16: classification of Machine learning [37]

4.9 Evaluation Matrix

The evaluation matrix refers to evaluate different ideas or rating them which is based on a set of some defined criteria and identifying the most promising ones. The common sets of criteria include all the levels of complex and related to idea implementations and setting the values that would bring to the users and the organizations.

For any planning or organizing an evaluation, it's very essential to use the evaluation Matrix. It is a table with a row for each evaluation matrix and a column that addresses evaluation design issues like data collection ideas or data sources or data analysis methods or criteria for any comparisons. This evaluation matrix connects all the evaluations to the means.

This application and its different levels raise the risk of the question that how to weigh the criteria as the main user or the main purpose of the declared data to use the application and connected object.

Evaluation matrix was made and crossing and on the one hand or the type of users. And on the other hand, it shows the "purpose of use". And here, the most critical application should have the maximum attention. [39]

4.10 Cost Function

The cost function is a function measuring the Machine Learning performance model for a given data. Quantifying the error, cost function determines error between two predicted and expected values as well as presents in a single real number form.

Cost function might be formed in different ways depending on the problems.

Cost Function purpose:

- **Minimized** - Then the returned value is usually called **cost**, **loss**, or **error**. The goal is to find the values of model parameters for which Cost Function return as as a small number as possible.
- **Maximized** - Then the value it yields is named a **reward**. The goal is to find the values of model parameters for which returned number is as large as possible. [40]

4.11 Selected Algorithms

For our project, we have used two algorithms. 1st one is polynomial regression and the 2nd one is Probabilistic Analysis. The reason for choosing these two algorithms is because they are much suitable for our data set. Our dataset does not have any specific classification. So polynomial regression is more suitable for this type of Data. Also in other future predictions of weather, water, air, etc. has been conducted with this algorithm. So that is why we choose linear regression as our algorithm. The 2nd algorithm is Probabilistic Analysis. When any binary type data is given Probabilistic Analysis is very suitable for calculation. Our Rain sensor only gives 2 types of data 0 & 1. On the other hand, our Day-Night sensor gives 3 types of dates 1, 2 & 3. So these two sensors are very compatible with Probabilistic Analysis.

4.12 Conclusion

Machine learning has become a very much popular medium for predicting the future of anything with the help of previous data. There are so many different ways or algorithms which can be used for this purpose such as polynomial regression, cost function, evaluation matrix, etc. in our research work; we have used two of them: Polynomial Regression analysis and Probabilistic through which we have predicted the quality of the water of our selected source in upcoming days. In the future, these processes can be used largely to predict for even years. The final findings of all the processes will be found in the next chapter.

Chapter 5

Data analysis and prediction

5.1 Introduction

To simplify the process of future prediction and calculation of error rate, we decided to work on only one sensor. This is because since all the major sensors are done with the same polynomial regression algorithm they will all give similar types of prediction. Most of the error rate is dependable on the given Dataset. That is why we will only conduct the pH data prediction and check its error rate. However, to check if our algorithm is working, we need another data set to compare. We acquired another data set [41] and ran it through our code. Thus now we can compare data and also check our algorithm. Here, our data is denoted as "Experimental Data" and the other data [41] set that we will compare to be denoted as "Collected Data".

5.2 types of Error

For our project we detected two types of error. These errors will indicate how accurate our future predictions are. There are given below:

5.2.1 Average Percentage Difference

The first type of error we detected is normal error also known as Average Percentage Difference. This type of error is calculated by taking the difference between an actual value and a predicted value of a given time, then dividing it with the actual value. If there are lots of iterations, then calculate all of them, then sum them up and divide it with the no of iterations. The formula is:

Error (Average Percentage Difference) = $1/N \sum (A-P/A)*100\%$ …….. (1)

Here,

N= Number of iterations

A= Actual Data

P= Predicted Data

In our project we used the N as number of data samples we predicted, A as the actual pH value which we got from our hardware device, P as the predicted pH value which we got from the machine learning Algorithm.

The significance of this error is:

- Give us a rough idea of the difference level (error) across the data set.
- Does us an idea of the accuracy level of our algorithm.
- Does not account for sudden change (spikes) in data set.

5.2.2 RMSE (Root Mean Square Error)

RMSE stands for root mean square error. This error takes the difference between an actual value and a predicted value at any given moment, squares it and sums up all the iterations. Then it is divided by the number of iterations and then square rooted. The formula for Root mean Square Error is:

RMSE error = $\sqrt{}$ (1/N*$\sum$ (A-P)^2)…… (2)

Here,

N= Number of iterations

A= Actual Data

P= Predicted Data

In our project we used the N as number of data samples we predicted, A as the actual pH value which we got from our hardware device, P as the predicted pH value which we got from the machine learning Algorithm.

The significance of this error is:

- Calculates of how well a regression line will fit in a given data set.
- RMSE is best when big errors are undesirable.
- RMSE is good for forcasting or future prediction.

5.3 Experimental pH Data prediction & error rate of 480 samples

We plotted & used 450 pH Data samples of our Experimental Data to Predict & plot 30 Data samples in the figure-17.

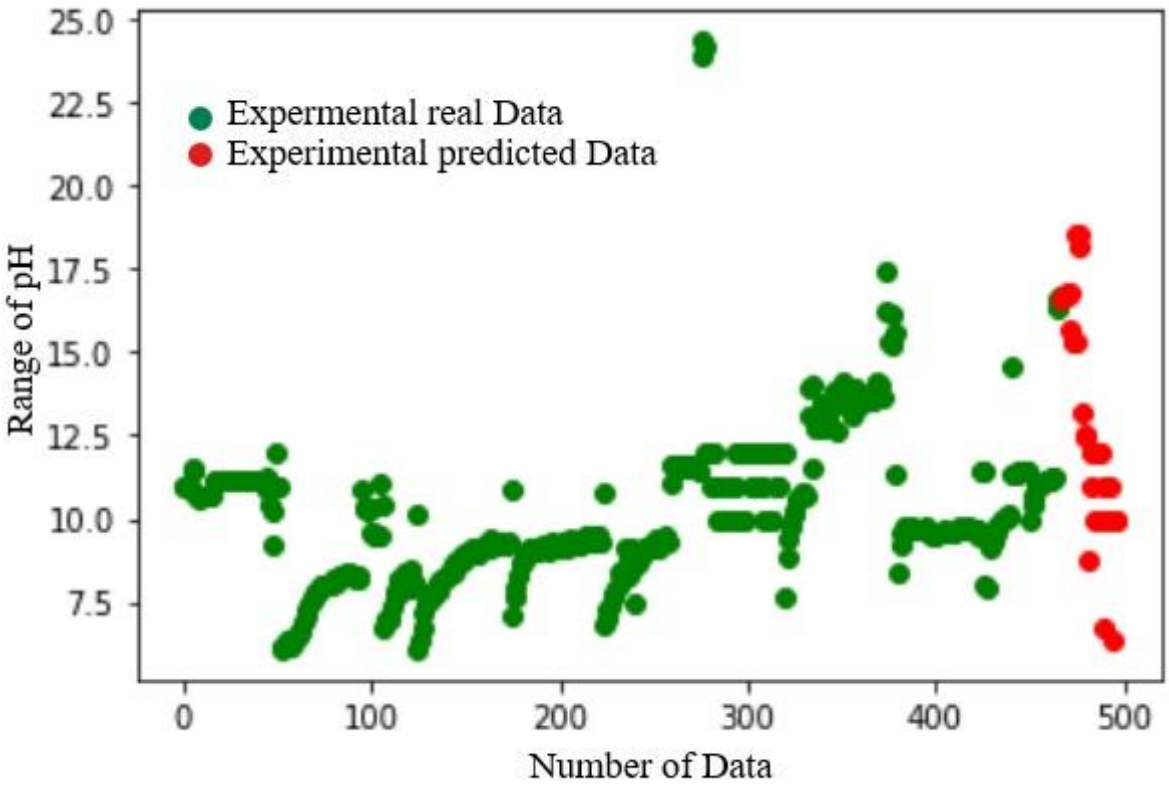


Figure-17: Experimental pH Data prediction of 480 samples

Error rate **23.827415708321816 %**

RMSE error **67.24272361064077 %**

We can see that our dataset is a lot scattered than other datasets. There may be some reasons for that. For example,

- Sudden change in pH due to waste dumping during data collection
- A sudden rise in water flow will change pH drastically
- During data collection sudden change in pH due to dumping detergent type elements
- Error due to sensor malfunction

5.4 Collected pH Data prediction & error rate of 480 samples

We plotted & used 450 pH Data samples of Collected Data [41] to Predict & plot 30 Data samples in the figure-18.

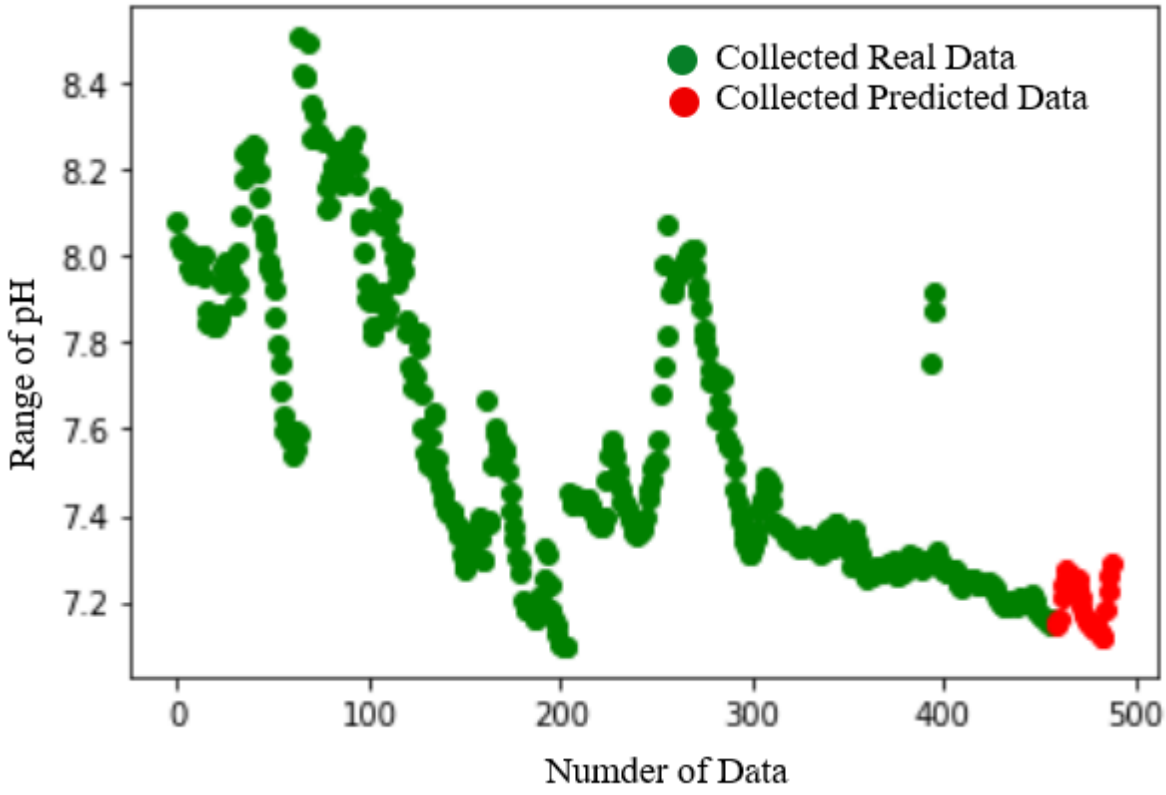


Figure-18: Collected pH Data prediction of 480 samples

Error rate **0.8832788248716351 %**

RMSE error **0.033008372040608226 %**

With our algorithm and Collected data [41] we can see that their data set provides a very low error rate. This means that their data set is much more accurate. From this information, we can say that they have a better water source and a better sensor to detect pH. That is why their data is much more accurate than ours.

5.5 Combined view of 480 samples

We plotted & used 450 pH Data samples of both Experimental and Collected [41] Data to Predict & plot 30 Data samples each in the figure-19.

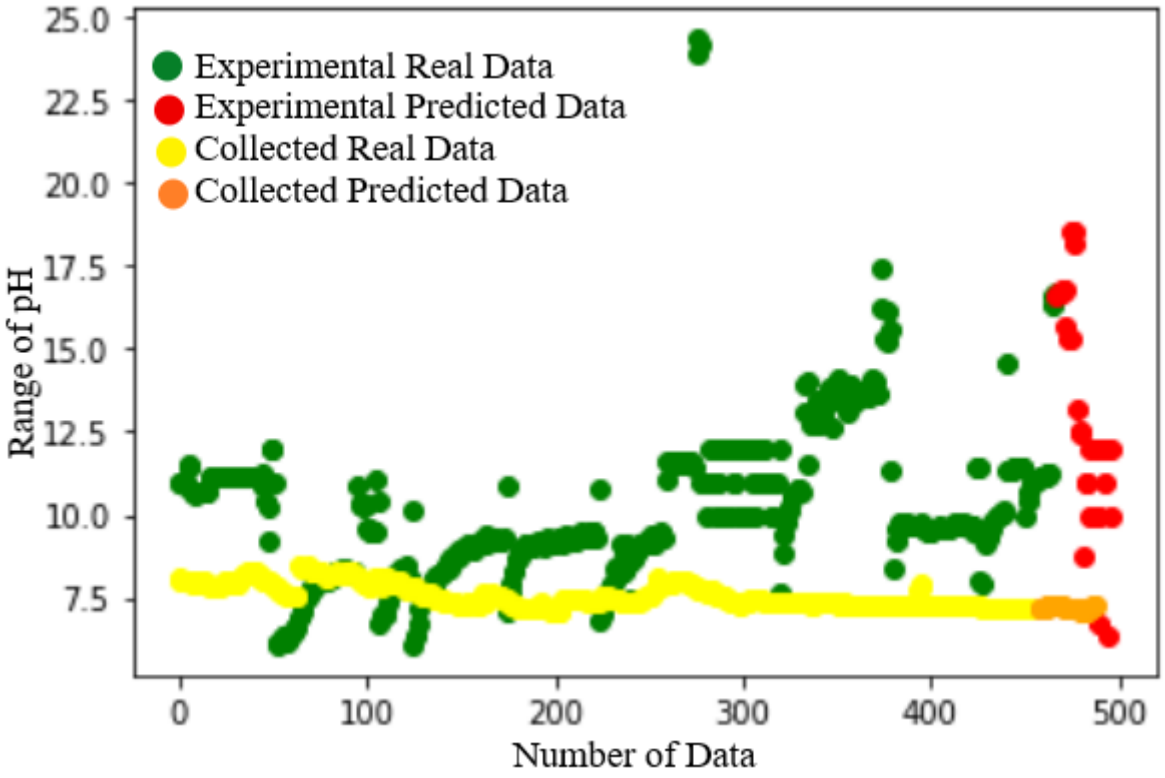


Figure-19: Experimental & Collected pH Data prediction of 480 samples in the same graph

Error rate of Experimental Data **23.827415708321816 %**

RMSE error of Experimental Data **67.24272361064077%**

Error rate of Collected [41] data **0.8832788248716351 %**

RMSE error of Collected [41] data **0.033008372040608226 %**

Here, we can see both Experimental and Collected data [41] on the same graph. From this graph, we can conclude that our data set is more scattered. This may be because there was a constant supply of waste from villages and a research facility nearby. This addition of waste may cause the pH level to give spikes of data. We can also see that the Collected dataset [41] to be much accurate at prediction. That is because their dataset is much more accurate than

ours. However, since both datasets were run by the same algorithm, it is safe to say that the algorithm works perfectly.

5.6 Explanation of higher error rate

From the previous figure-19, we can see that there is a huge difference in error rate between Experimental & collected [41] data set. The Experimental data error rate is significantly higher than the collected data. This is cased for two main reasons. One of them is that their data is taken on clean water which has a constant like value. In other words, they have less fluctuation in data. Since the Experimental data is taken on polluted water, there is a lot more fluctuation. The second reason is that our pH sensor is not high quality. Being a chemical sensor, it has an acceptable error range. However, in long term usage (2 3 days) the sensitivity drops and the sensor does not give accurate value. For example we given the first day of data (60 samples) in our machine learning algorithm and got a significantly lower error rate shown in figure-20.

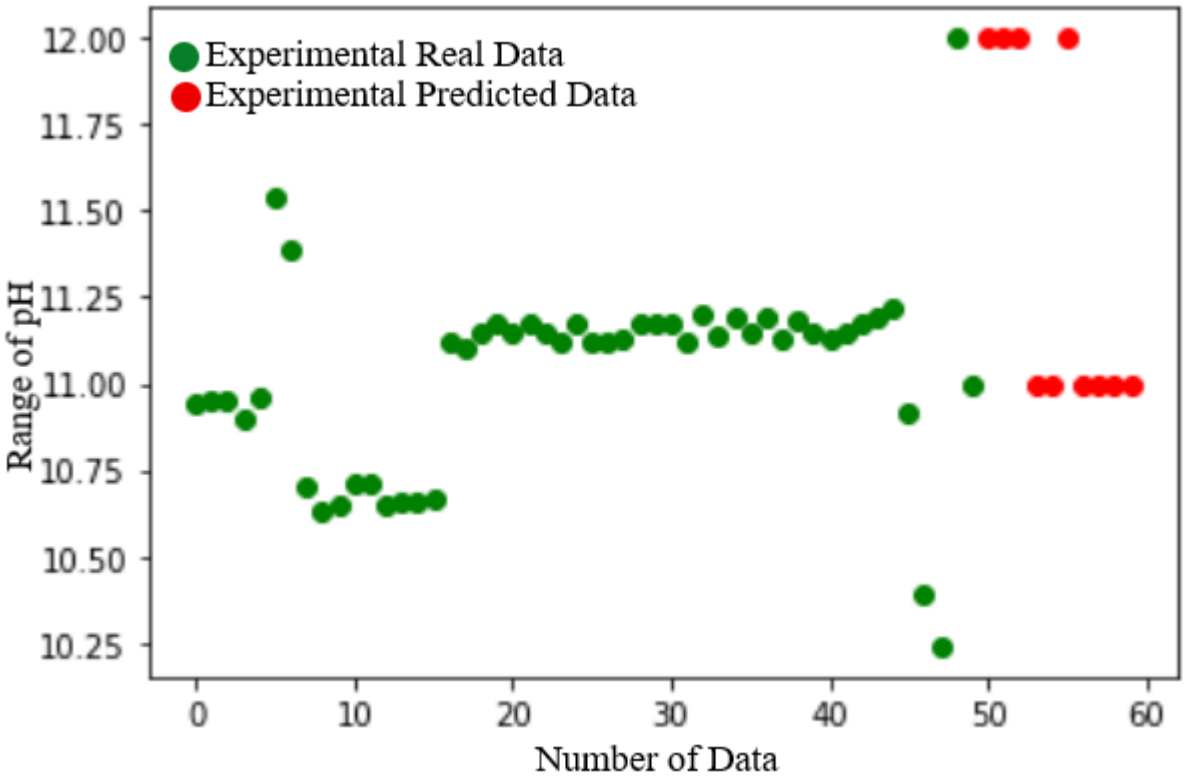


Figure-20: Experimental pH Data prediction of 60 samples

Error rate **4.1996501145912895 %**

Since our data is more accurate in the 1st day, we got a much lower error rate despite having very few data points. So from this figure-20 we can conclude that the high error rate is due to the sensitivity issue of the pH sensor.

5.7 Experimental pH data prediction of average 10 days with an error rate

Experimental pH data prediction of an average of 10 days is given below:

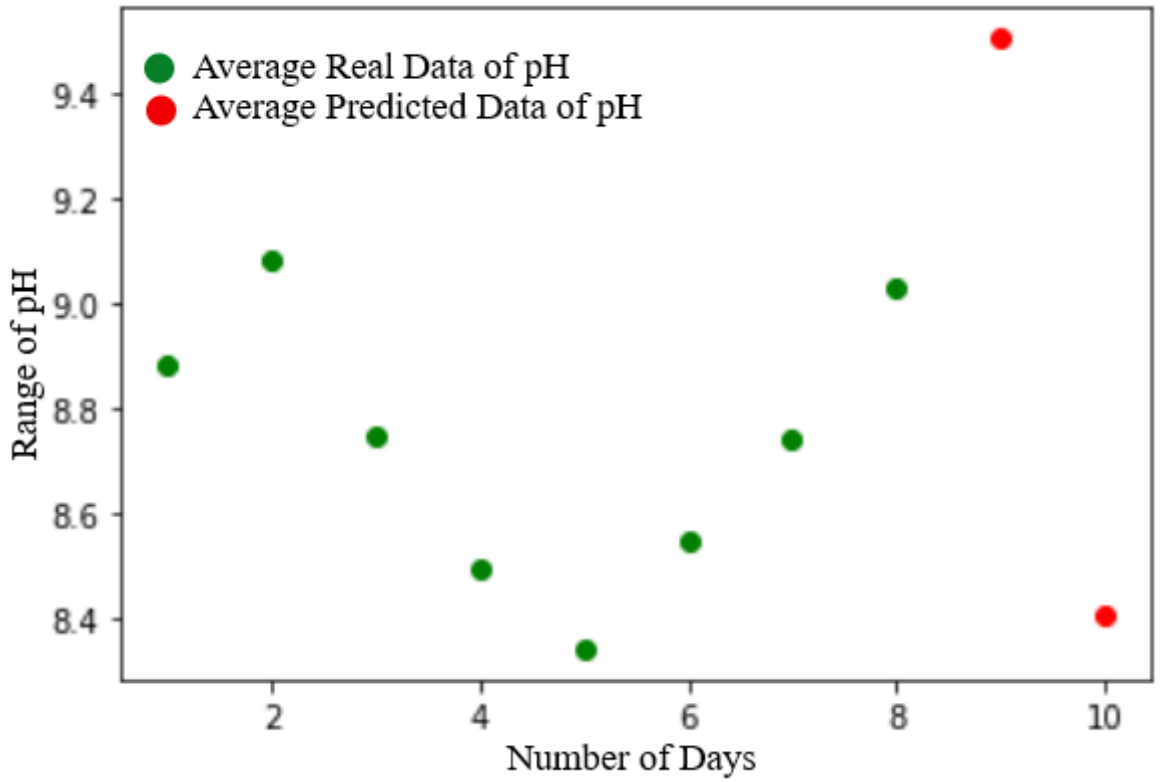


Figure-21: Experimental pH data prediction of an average of 10 days

Error rate **9.245474962305341 %**

With our data sample, without average, the error rate was **23.827415708321816 %**

With our data sample, with average, the error rate was **9.245474962305341 %**

So, we can say that by getting the average value we can get a more accurate result in terms of error rate. This is because the average pH value of one day is a more optimum value which is near the predicted value.

5.8 Experimental turbidity data prediction of average 10 days with an error rate

Experimental turbidity data prediction of an average of 10 days is given below:

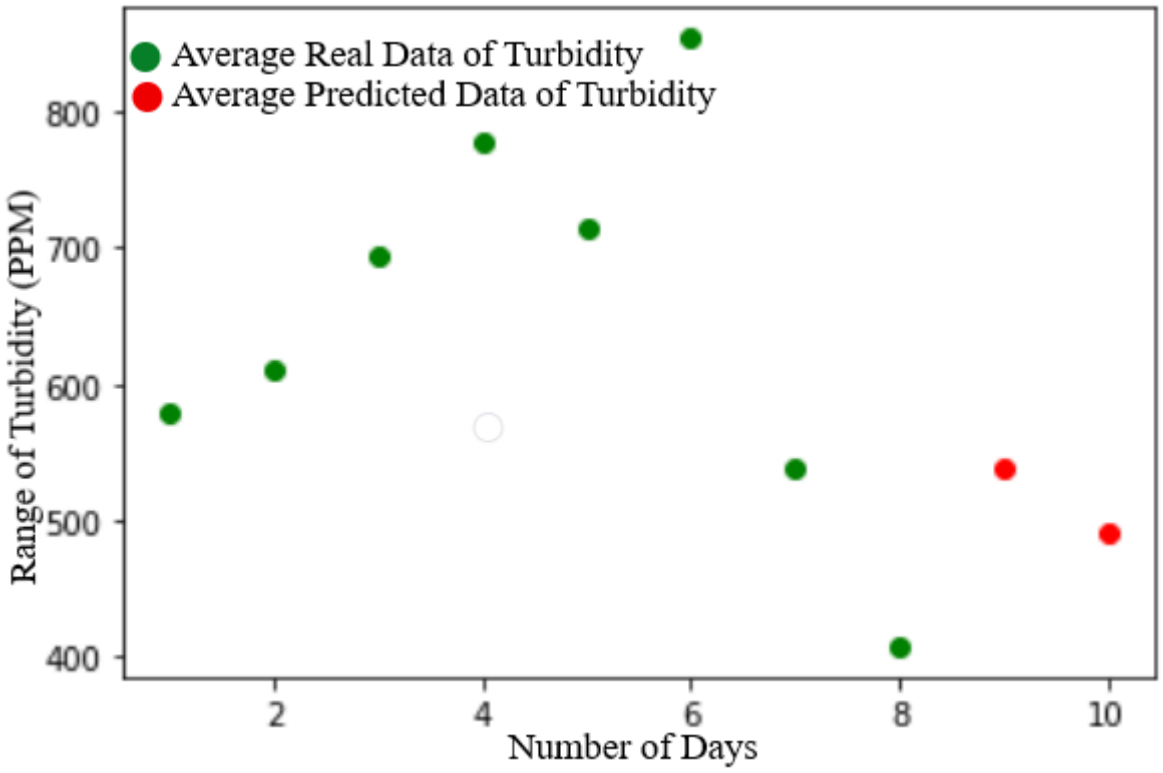


Figure-22: Experimental turbidity data prediction of average 10 days

Error rate **12.686137724743102 %**

Here, the average predicted value is given to show the error rate of turbidity of our water source. Since we are using the average method of polynomial regression, we can get a more accurate result in terms of error rate. This is because the average turbidity value of one day is a more optimum value which is near the predicted value.

5.9 Experimental TDS data prediction of average 10 days with an error rate

Experimental TDS data prediction of an average of 10 days is given below:

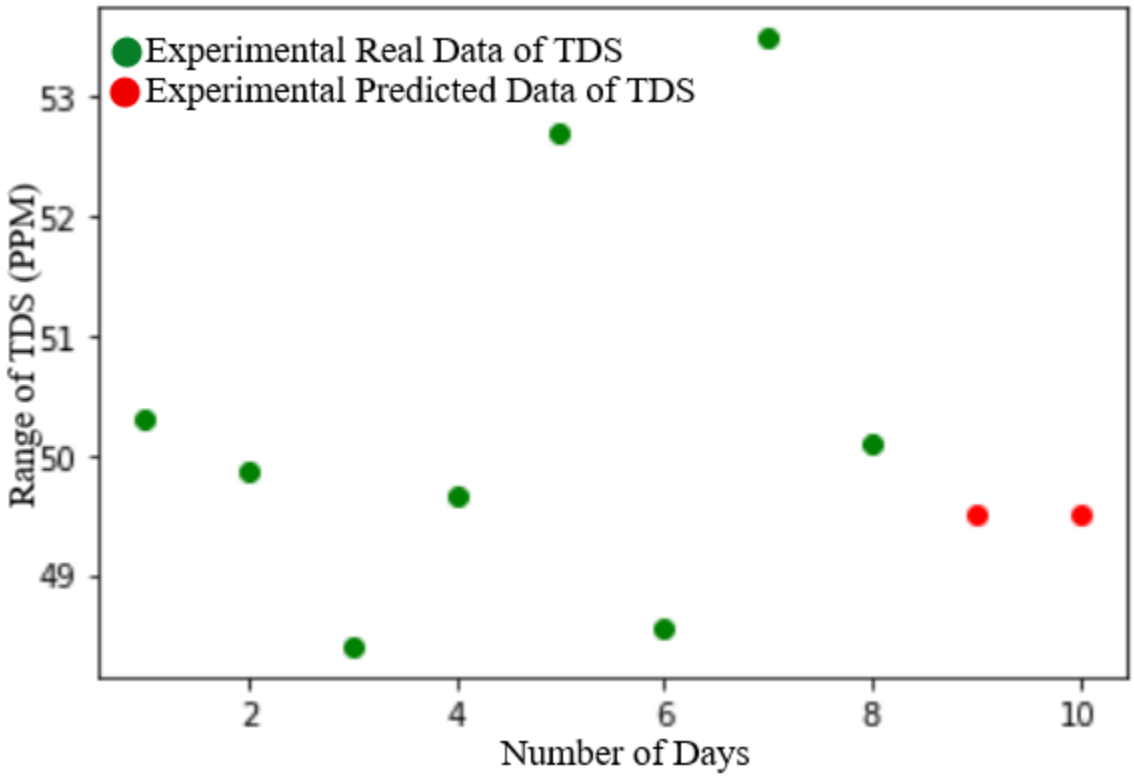


Figure-23: Experimental TDS data prediction of an average of 10 days

Error rate **4.460299421520006 %**

Here, the average predicted value of TDS is given to show the error rate of TDS of our water source. Since we are using the average method of polynomial regression, we can get a more accurate result in terms of error rate. This is because the average TDS value of one day is a more optimum value which is near the predicted value.

5.10 Collected pH data prediction of 2000 data with an error rate

We plotted & used 1970 pH Data samples of Collected [41] Data to Predict & plot 30 Data samples each in the figure-23.

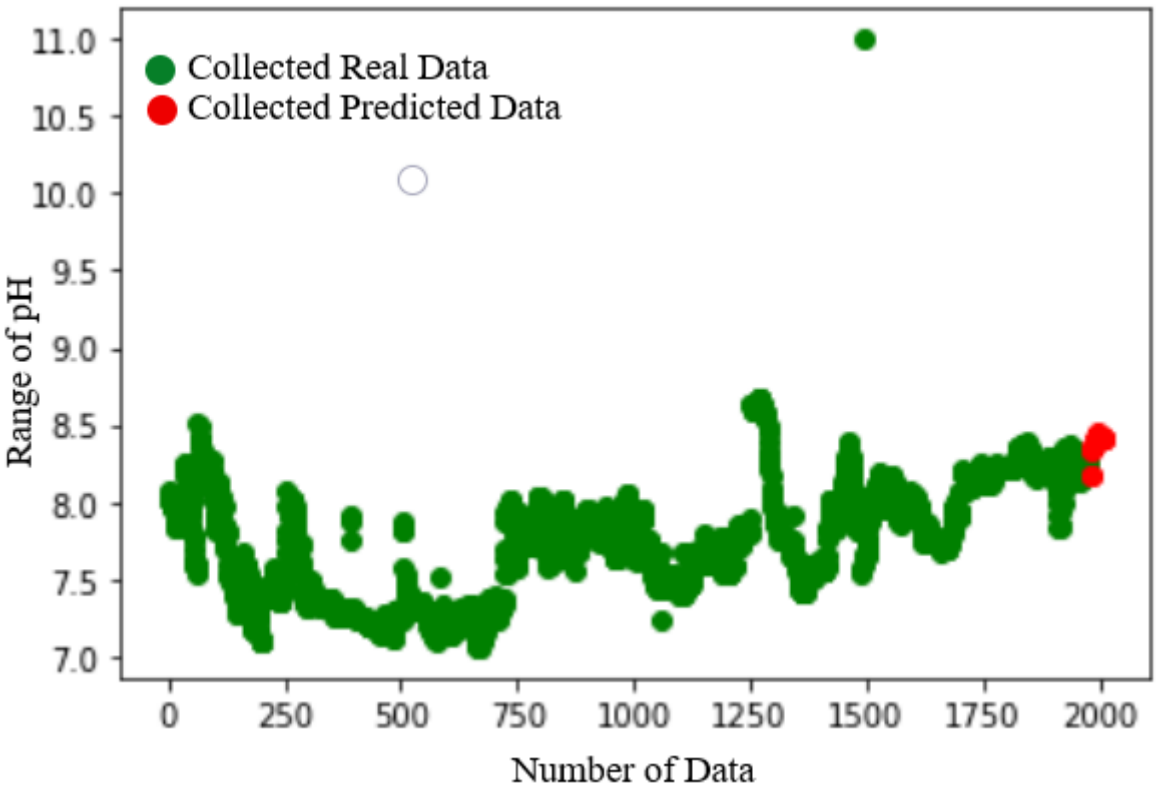


Figure-24: Collected pH data prediction of 2000 data collection

Error rate **0.3712671393716075 %**

Here, we can see that the error rate is lowest. This is due to the fact that the more data we have the more accurate prediction can be measured.

5.11 Rain & Day-Night Sensors error rate

For the rain & day-night sensor, a different algorithm was used. This algorithm is called probabilistic Analysis.

Rain sensor:

For the rain sensor, there are only 2 values to consider which are 0 & 1.

Probability of being 1 on rain sensor **0.35106382978723405 %**

Day-Night sensor:

For the rain sensor, there are 3 values to consider which are 1, 2 & 3.

Probability of being 2 on Day-Night sensor **0.6413043478260869 %**

This is a very simple algorithm that has a high accuracy level. So as a result the error rate is very low.

5.12 Error Rates

In this paper, there are lots of error rates of lots of parameters. For ease of understanding, a list of these error rates is given below:

Table-8: List of Error Rates

Type of Data set	Error Rate (%)	Comments
Experimental pH Data of 480 samples	**23.827415708321816**	High fluctuation in Data
Collected pH Data of 480 samples	**0.8832788248716351**	Data Stable and possibility of clean water
Experimental pH Data of 50 samples	**4.1996501145912895**	Detecting sensor's sensitivity issue
Collected pH Data of 2000 samples	**0.3712671393716075**	Larger Data size decreases error
Experimental pH Data of 480 samples (RMSE)	**67.24272361064077**	High rate of change in Data
Collected pH Data of 480 samples (RMSE)	**0.033008372040608226**	Low rate of change in Data
Experimental pH data of average 10 days	**9.245474962305341**	Average Data gives more accurate results
Experimental turbidity data of average 10 days	**12.686137724743102**	Average Data gives more accurate results
Experimental TDS data of average 10 days	**4.460299421520006**	Average Data gives more accurate results

5.13 Conclusion

In this chapter, mainly after doing all the regression and analysis for our Experimental data, we compared the pH dataset with another paper's [41] pH dataset. We have found out the error rate of the pH of both datasets to compare them. We also predicted the average value of pH, turbidity, TDS of our dataset.

Here, we could see that our error rate was a bit high because we could only collect a minimum number of data. We know that for machine learning algorithms, the more amounts of data we get, the prediction becomes more accurate. So, if we want to implement this in the future, we need a huge number of data. Then the algorithm can give us a better prediction.

Chapter 6

Conclusion & Future Work

6.1 Thesis Conclusion

First of all, our project is on IoT based water quality monitoring system and future prediction using machine learning. We declared our objectives for this project. We have gathered information about the importance of water quality monitoring and how the water usually gets polluted. We have also gathered info about some contaminants that are harmful to using the water in daily life, drinking water, and fishes as well in the first chapter. In the 2nd chapter, we discussed the sensors we have used and the Arduino ESP8266 and battery was discussed vastly. Then we gave a detailed idea of the construction of our hardware device step by step with a schematic diagram. Thus we built the IoT based hardware device. In the third chapter, the data was sent and collected to a server named thinkspeak using a Wi-Fi-module called ESP8266. ThinkSpeak is an open platform for IoT based projects and a very user-friendly website. We used ESP8266 as the Wi-Fi module. Next, we studied different types of machine learning algorithms and decided which to use. We have used polynomial regression and probabilistic analysis as we have found them to be preferable for our Dataset. Furthermore, in the 5th chapter, we have also compared our dataset to another paper's dataset that was published beforehand. Then, we finally showed the difference in the error rate of future predicted data. Last but not least, we gave some future ideas and concluded our thesis paper.

It is very much possible to implement an IoT based water quality monitoring system in any type of water source. With very little human interaction data can be gathered and get a live view of the data stream. Also for future prediction, good quality algorithms can be made & implemented to get very accurate results. However, there are some issues to be resolved. For example, there must be a way to get a highly accurate data set. This can be done by getting better & high-quality sensors. The better the Dataset, the better prediction a machine learning Algorithm can do.

6.2 Future Work

This project can be implemented on a larger scale in the upcoming days. Some scopes regarding the future implementation are here:

- **Notify via SMS or Email:** Since, the project is an IoT-based project, the maximum and minimum criteria can be set and a system can be developed in such a way if the water does not fulfill the criteria the device will immediately notify its user via email or mobile SMS.
- **Online forecasting:** The future we are calculating now by using an algorithm can be integrated into the website. Therefore, as we can see the weather updates and the probable updates for the coming few days, the quality of water can also be predicted for the coming few days in the same way on the website.
- **Mobile Application:** Mobile application can also be built for this device so that everyone with the app and selective device can monitor the real-time data in the app.
- **Auto shut-off technology:** We generally use water coming from WASA or some other sources. Sometimes, dirty and polluted water is being supplied. Using some interrupt operation, the device can be integrated with the pump we are using to collect water. So that whenever the dirty water is detected in the water tank, the device will notify the user and instantly turn off the water-pump so that dirty water cannot fill the tank.
- **Using other algorithms:** We have already mentioned that we have used Polynomial regression and probabilistic analysis. If somebody wants, they can also use other algorithms as well like evaluation matrix, cost function, etc.

Finally, it can be said that our project has a large scale of future scopes. It can be extended further and can be modified to make it more modern.

References

1. Syed Nazmul Ahsan (2019). State of Water Pollution from Industrial Effluents in Bangladesh. Department of Environment; Dhaka.
2. Cabral JPS. Water Microbiology. Bacterial Pathogens and Water. "*International Journal of Environmental Research and Public Health"*. 2010; 7(10):3657-3703
3. WHO (World Health Organization) ,Water sanitation hygiene, Water-related Diseases, Cholera.
4. WHO (World Health Organization) ,Newsroom, Fact sheets, Diarrhea disease**.**
5. K., Diana. (2018) "How Does Water Pollution Affect Fish?" sciencing.com.
6. Bidaisee, Satesh. (2018). The Importance of Clean Water. Biomedical Journal of Scientific & Technical Research. 8. 10.26717/BJSTR.2018.08.001719.
7. Ljubinka Rajakovic and Vladana Rajakovic-Ognjanovic (July 25th, 2018). Arsenic in Water: Determination and Removal, Arsenic - Analytical and Toxicological Studies, Margarita Stoytcheva and Roumen Zlatev, IntechOpen, DOI: 10.5772/intechopen.75531.
8. Fondriest Environmental, Inc. "Water Temperature." Fundamentals of Environmental Measurements. 7 Feb. 2014. Web. < https://www.fondriest.com/environmental-measurements/parameters/water-quality/water-temperature/ >.
9. Abhijit Pathak, Abrar Hossain Tasin, Umma Salma, Likhon Barua, Md. Sabbir Hossain, Sanjay Datta "*American Journal of Engineering Research*" (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-x, Issue-x, pp-xxx-xxx.
10. Abhijit Pathak, Abrar Hossain Tasin, Umma Salma, Likhon Barua, Md. Sabbir Hossain, Sanjay Datta American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-x, Issue-x, pp-xxx-xxx.
11. Abhijit Pathak, Abrar Hossain Tasin, Umma Salma, Likhon Barua, Md. Sabbir Hossain, Sanjay Datta American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-x, Issue-x, pp-xxx-xxx.

12. Process modeling and optimization of biological removal of carbon, nitrogen, and phosphorus from hospital wastewater in a continuous feeding & intermittent discharge (CFID) bioreactor – Scientific Figure on Research Gate.
13. Haque, Md.Mahbubul. (2008). TREATMENT OF TEXTILE WASTE.WATER IN BANGLADESH. 8. 18.
14. Hoque, Muhammad. (2016). Seasonal Variation of Physicochemical Properties of Water in the Buriganga River, Bangladesh. World Applied Sciences Journal. 34. 24=36. 10.5829/idosi.wasj.2016.34.1.22871.
15. Jan - E - Alam, Md & Reza, Prince & Hossain, Saddam & Hossain, Md. (2017). WATER QUALITY ASSESSMENT OF DHANMONDI LAKE IN DHAKA CITY. Multidisciplinary Journal of the European University of Bangladesh. 2. 43-47.
16. Tariquzzaman, s.M. & Nishu, Susmita & Saeed, Tanveer & Reday, Rifat. (2016). WATER QUALITY AND EIA OF SIMPLE HATIRJHEEL LAKE.
17. elprocus.com "Arduino Mega 2560 Board" arduino.cc.
18. Seeedstudio.com. Home, sensors, Environment. Grove - PH Sensor Kit (E-201C-Blue). © 2008-2021 Seeed Technology Co.,Ltd.
19. Electronics-notes.com articles, electronic components, LDR. "Light Dependent Resistor LDR: Photo resistor"
20. wiki.dfrobot.com, "Turbidity sensor SKU SEN0189" DFRobot, Sensors & Modules, Sensors , Liquid Sensors.
21. dfrobot.com, "Gravity: Analog TDS Sensor/Meter for Arduino" Sales, Clearance Gravity, Analog TDS Sensor/Meter for Arduino
22. MisterBotBreak (2019). "How to Use a Rain Sensor", project hub, create.arduino.cc.
23. Robin Kanattu Thomas (2018). "How to Program the ESP8266's Onboard GPIO Pins", maker.pro, esp8266/tutorial.
24. Bdspeedytech.com. "11.1V 1100mAh 3S 30C lipo battery", BdSpeedytech © 2009-2016 All Rights Reserved, SWOT System Ltd.
25. E. P. Yadav, E. A. Mittal, and H. Yadav, "IoT: Challenges and Issues in Indian Perspective," *2018 3rd International Conference On Internet of Things: Smart*

Innovation and Usages (IoT-SIU), Bhimtal, 2018, pp. 1-5, doi: 10.1109/IoT-SIU.2018.8519869.

26. Thingspeak.com. “About ThingSpeak”, © 2021 The MathWorks, Inc.
27. S. Ray, *"A Quick Review of Machine Learning Algorithms"*, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.
28. D. Bajpai and L. He, "*Evaluating KNN Performance on WESAD Dataset*", 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 2020, pp. 60-62, doi: 10.1109/CICN49253.2020.9242568.
29. S. Angra and S. Ahuja, "*Machine learning and its applications: A review*", 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, 2017, pp. 57-60, doi: 10.1109/ICBDACI.2017.8070809.
30. B. Ilias, S. A. A. Shukor, A. H. Adom, N. A. Rahim, M. F. Ibrahim, and S. Yaacob, *"Indoor mobile robot localization using KNN"*, 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 211-216, doi: 10.1109/ICCSCE.2016.7893573.
31. Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.
32. Hoffmann, John. (2008). “*Linear Regression Analysis: Assumptions and Applications”*.
33. Suzanne Ekelund (2011). “ROC curves – what are they and how are they used?” home, articles, acutecaretesting.org
34. developers.google.com. “Classification: ROC Curve and AUC” Home, Products, Machine Learning, Courses.
35. Jason Brownlee (2016). “What is a Confusion Matrix in Machine Learning”, machinelearningmastery.com, Code Algorithms from Scratch.
36. ANIRUDDHA BHANDARI (2020). “Everything you should know about Confusion Matrix for Machine Learning”, analyticsvidhya.com.

37. Expert.ai Team (2017). “What is Machine Learning? A Definition.” Home, Insights & Resources, What is Machine Learning? A Definition, www.expert.ai.

38. developers.google.com. “Classification: Accuracy”, Home, Products, Machine Learning, Courses.

39. Karima Bourquard, Daniel Israël, Robert Picard, Hugues Brouard, Virginie Delay, Matthieu Faure, Bastien Fraudet, Thierry Gatineau (2018). “Evaluations and Effectiveness”, sciencedirect.com.

40. Kamil Krzyk (2018). “Coding Deep Learning for Beginners — Linear Regression (Part 2): Cost Function”, MyRoadToAI, towardsdatascience.com.

41. Hongqian Qin (2019). “Time-series-data-prediction”, github.com. available from: https://github.com/qin67/Time-series-data-prediction/blob/master/Predict%20Dissolved%20Oxygen%20based%20on%2060000%2B%20%20data%20.pdf

Appendix A

Code for 480 & 2000 Data plot and Prediction with error rate

```
import pandas as pd
import matplotlib.pyplot as plt
import math
import random
import numpy
def getNewAnalysis(dataset,field,deg):
    dataset[field] = dataset[field].fillna(-1)

    np_ = dataset.to_numpy()

    a = numpy.array([0])

    for i in range(len(np_)):
        if np_[i] == -1.0:
            #print(i)
            a = numpy.append(a, [i]).astype(int)

    np_ = numpy.delete(np_,a)

    for i in range(len(np_)):
        if np_[i] >= 25 or np_[i] <= 6 :
            np_[i] = random.randint(10,12)



    X = [ x for x in range(len(np_)-30)]
    y = [ float(x) for x in np_[0:len(np_)-30] ]

    rx = [ x for x in range(len(np_)-30, len(np_))]
    ry = [ float(x) for x in np_[len(np_)-30: len(np_)] ]
```

```
    mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))
    pred = 0

    for i in range(len(np_)-30, len(np_)):
        print("prediction at ["+str(i)+"] :"+str(mymodel(i)))
        print("realvalue at ["+str(i)+"] :"+str(np_[i]))

        pred = pred + (abs(mymodel(i) - np_[i])) / np_[i]

    print("error ",float(pred/30*100),"%")

    myline = numpy.linspace(1, len(np_) , 100)

    plt.scatter(X, y, color='green')
    plt.scatter(rx, ry, color='red')
    plt.show()



def getNewAnalysisComparison(dataset1,field1, dataset2,field2,deg):
    dataset1[field1] = dataset1[field1].fillna(-1)

    dataset2[field2] = dataset2[field2].fillna(-1)

    np_ = dataset1.to_numpy()
    np_2 = dataset2.to_numpy()

    a = numpy.array([0])

    for i in range(len(np_)):
        if np_[i] == -1.0:
            #print(i)
```

```
        a = numpy.append(a, [i]).astype(int)

np_ = numpy.delete(np_,a)

a = numpy.array([0])
for i in range(len(np_2)):
    if np_2[i] == -1.0:
        #print(i)
        a = numpy.append(a, [i]).astype(int)

np_2 = numpy.delete(np_2,a)

for i in range(len(np_)):
    if np_[i] >= 25 or np_[i] <= 6 :
        np_[i] = random.randint(10,12)



X = [ x for x in range(len(np_)-30)]
y = [ float(x) for x in np_[0:len(np_)-30] ]

rx = [ x for x in range(len(np_)-30, len(np_))]
ry = [ float(x) for x in np_[len(np_)-30: len(np_)] ]



X2 = [ x for x in range(len(np_2)-30)]
y2 = [ float(x) for x in np_2[0:len(np_2)-30] ]

rx2 = [ x for x in range(len(np_2)-30, len(np_2))]
ry2 = [ float(x) for x in np_2[len(np_2)-30: len(np_2)] ]



mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))
```

```
mymodel2 = numpy.poly1d(numpy.polyfit(X2, y2, deg))

pred = 0
pred2 = 0

sqrt = 0
sqrt2 = 0

for i in range(len(np_)-30, len(np_)):
    print("prediction at ["+str(i)+"] :"+str(mymodel(i)))
    print("realvalue at ["+str(i)+"] :"+str(np_[i]))

    pred = pred + (abs(mymodel(i) - np_[i])) / np_[i]
    sqrt = sqrt + (mymodel(i) - np_[i])**2

print("error ",float(pred/30*100),"%")
print("RMSE error", sqrt/30**0.5)

for i in range(len(np_2)-30, len(np_2)):
    print("prediction at ["+str(i)+"] :"+str(mymodel2(i)))
    print("realvalue at ["+str(i)+"] :"+str(np_2[i]))

    pred2 = pred2 + (abs(mymodel2(i) - np_2[i])) / np_2[i]
    sqrt2 = sqrt2 + (mymodel2(i) - np_2[i])**2

print("error ",float(pred2/30*100),"%")
print("RMSE error", sqrt2/30**0.5)

myline = numpy.linspace(1, len(np_) , 100)
myline2 = numpy.linspace(1, len(np_2) , 100)

plt.scatter(X, y, color='green')
```

```
    plt.scatter(rx, ry, color='red')

    plt.scatter(X2, y2, color='yellow')
    plt.scatter(rx2, ry2, color='orange')
    plt.show()



print("For field 1\n")

dataset = pd.read_csv('Nir_Project.csv', usecols=[2], engine='python')
getNewAnalysis(dataset, 'ph', 2)

print("For field 8\n")
dataset = pd.read_csv('Nir_Project.csv', usecols=[8], engine='python')
getNewAnalysis(dataset, 'PH4', 2)

print("For field 7\n")
dataset = pd.read_csv('Nir_Project.csv', usecols=[7], engine='python')
getNewAnalysis(dataset, 'PH', 2)



dataset1 = pd.read_csv('Nir_Project.csv', usecols=[2], engine='python')
dataset2 = pd.read_csv('Nir_Project.csv', usecols=[8], engine='python')
getNewAnalysisComparison(dataset1,'ph', dataset2,'PH4',2)
```

Code for 60 Data plot and Prediction with error rate

```
import pandas as pd
import matplotlib.pyplot as plt
import math
import random
import numpy



def getAnalysisOf10(dataset,field,deg):
    dataset[field] = dataset[field].fillna(-1)



    np_ = dataset.to_numpy()


    a = numpy.array([0])


    for i in range(len(np_)):
        if np_[i] == -1.0:
            a = numpy.append(a, [i]).astype(int)


    np_ = numpy.delete(np_,a)



    for i in range(60):
        if np_[i] >= 25 or np_[i] <= 10 :
            np_[i] = random.randint(11,12)
```

```
X = [ x for x in range(50)]

y = [ float(x) for x in np_[0:50] ]

rx = [ x for x in range(50, 60)]
ry = [ float(x) for x in np_[50: 60] ]



mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))
pred = 0

for i in range(50, 60):
    print("prediction at ["+str(i)+"] :"+str(mymodel(i)))
    print("realvalue at ["+str(i)+"] :"+str(np_[i]))

    pred = pred + (abs(mymodel(i) - np_[i])) / np_[i]

print("error ",float(pred/10*100),"%")

myline = numpy.linspace(1, 60 , 100)

plt.scatter(X, y, color='green')
plt.scatter(rx, ry, color='red')
plt.show()
```

```
print("For field 1\n")


dataset = pd.read_csv('Nir_Project.csv', usecols=[2], engine='python')
getAnalysisOf10(dataset, 'ph', 1)
```

Code for Average Data plot and Prediction with error rate

```
import pandas as pd
import matplotlib.pyplot as plt
import math
import random
import numpy

date_num=[85,146,10,278,88,489,519,503,182,73]
date_cons=[85,231,241,519,607,1096,1615,2118,2300,2373]

def accuracy2(test_y,a_list,distribution):

    err = 0
    for i in test_y:
        if i != random.choices(a_list, distribution):
            err += 1

    print(err/len(test_y))
```

```
def analysis4(dataset,field):
    dataset.dropna(subset = [field], inplace=True)


    np_array = dataset.to_numpy()


    df = pd.DataFrame(np_array, columns = [field])


    X = [[x] for x in range(1,int(((len(np_array)+1))*.8))]
    y = df.iloc[1:int((len(np_array)+1)*.8), 0].values


    test_X = [ x for x in range(int(((len(np_array)+1))*.8),(len(np_array)+1)) ]
    test_y = df.iloc[int((len(np_array)+1)*.8):(len(np_array)+1), 0].values


    no_1 = 0
    no_2 = 0
    no_3 = 0
    for i in y:
        if i == 1:
            no_1 += 1
        elif i == 2:
            no_2 += 1
        elif i == 3:
            no_3 += 1


    a_list = [1,2,3]
    distribution = [no_1/len(y), no_2/len(y), no_3/len(y)]
```

```
    accuracy2(test_y,a_list,distribution)


def analysis3(dataset,field):
    dataset.dropna(subset = [field], inplace=True)


    np_array = dataset.to_numpy()
    df = pd.DataFrame(np_array, columns = [field])


    X = [[x] for x in range(1,int(((len(np_array)+1))*.8))]
    y = df.iloc[1:int((len(np_array)+1)*.8), 0].values


    test_X = [ x for x in range(int(((len(np_array)+1))*.8),(len(np_array)+1)) ]
    test_y = df.iloc[int((len(np_array)+1)*.8):(len(np_array)+1), 0].values


    no_1 = 0
    no_0 = 0
    for i in y:
        if i == 1:
            no_1 += 1
        elif i == 0:
            no_0 += 1


    a_list = [0, 1]
    distribution = [no_0/len(y), no_1/len(y)]
```

```
    accuracy2(test_y,a_list,distribution)



def getAverage(dataset,field,deg):
    dataset[field] = dataset[field].fillna(-1)



    np_ = dataset.to_numpy()

    #print((np_))



    avg_value = []
    sum_ = 0
    k = 0

    times = 0

    list_ = np_.tolist()

    for i in range(len(np_)):

        if np_[i] != -1.0:
            sum_ += np_[i]
            times += 1
            #print(np_[i],sum_,k,i)
```

```
    if i == date_cons[k]-1 or i == len(np_)-1:
        x = sum_/times
        avg_value.append(x)
        k += 1
        sum_= 0
        times = 0
X = [ x for x in range(1,9)]

y = [ float(x) for x in avg_value[0:8] ]

rx = [ x for x in range(9,11)]

ry = [ float(x) for x in avg_value[8:10] ]

mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))

print("prediction at 9 ", mymodel(9))
print("prediction at 10 ", mymodel(10))

pred1 = mymodel(9)-avg_value[8]
pred2 = mymodel(10) - avg_value[9]

if pred1 < 0 : pred1 = -pred1
if pred2 < 0 : pred2 = -pred2

error = ((pred1/avg_value[8]) + (pred2/avg_value[9]))/2
```

```
    print("error ",float(error*100),"%")


    myline = numpy.linspace(1, 11 , 100)


    plt.scatter(X, y, color = 'green')
    plt.scatter(rx, ry, color = 'red')


    plt.show()



def getAverage1(dataset,field,deg):
    dataset[field] = dataset[field].fillna(-1)


    np_ = dataset.to_numpy()




    for i in range(len(np_)):
        if np_[i] >= 25 or np_[i] <= 6 :
            np_[i] = random.randint(7,10)






    avg_value = []
    sum_ = 0
```

```
k = 0



times = 0


list_ = np_.tolist()



for i in range(len(np_)):


    if np_[i] != -1.0:
        sum_ += np_[i]
        times += 1
        #print(np_[i],sum_,k,i)
    if i == date_cons[k]-1 or i == len(np_)-1:
        x = sum_/times
        avg_value.append(x)
        k += 1
        sum_= 0
        times = 0




X = [ x for x in range(1,9)]
```

```
y = [ float(x) for x in avg_value[0:8] ]


rx = [ x for x in range(9,11)]


ry = [ float(x) for x in avg_value[8:10] ]




mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))


print("prediction at 9 ", mymodel(9))

print("prediction at 10 ", mymodel(10))


pred1 = mymodel(9)-avg_value[8]

pred2 = mymodel(10) - avg_value[9]


if pred1 < 0 : pred1 = -pred1

if pred2 < 0 : pred2 = -pred2


error = ((pred1/avg_value[8]) + (pred2/avg_value[9]))/2

print("error ",float(error*100),"%")


myline = numpy.linspace(1, 11 , 100)


plt.scatter(X, y, color='green')

plt.scatter(rx, ry, color="red")
```

```
    plt.show()









def getAverage5(dataset,field,deg):
    dataset[field] = dataset[field].fillna(-1)


    np_ = dataset.to_numpy()


    for i in range(len(np_)):
        if np_[i] > 0 and np_[i] < 5:
            np_[i] = random.randint(40,60)
        if np_[i] < 20 :
            np_[i] = random.randint(40,60)




    avg_value = []
    sum_ = 0
    k = 0



    times = 0
```

```
list_ = np_.tolist()



for i in range(len(np_)):


    if np_[i] != -1.0:
        sum_ += np_[i]
        times += 1
        #print(np_[i],sum_,k,i)
    if i == date_cons[k]-1 or i == len(np_)-1:
        x = sum_/times
        avg_value.append(x)
        k += 1
        sum_= 0
        times = 0




X = [ x for x in range(1,9)]



y = [ float(x) for x in avg_value[0:8] ]


rx = [ x for x in range(9,11)]


ry = [ float(x) for x in avg_value[8:10] ]
```

```
    mymodel = numpy.poly1d(numpy.polyfit(X, y, deg))

    print("prediction at 9 ", mymodel(9))
    print("prediction at 10 ", mymodel(10))

    pred1 = mymodel(9)-avg_value[8]
    pred2 = mymodel(10) - avg_value[9]

    if pred1 < 0 : pred1 = -pred1
    if pred2 < 0 : pred2 = -pred2

    error = ((pred1/avg_value[8]) + (pred2/avg_value[9]))/2
    print("error ",float(error*100),"%")

    myline = numpy.linspace(1, 11 , 100)

    plt.scatter(X, y, color='green')
    plt.scatter(rx, ry, color="red")
    # plt.plot(myline, mymodel(myline))
    plt.show()


print("For field 1\n")
```

```
dataset = pd.read_csv('Nir_Project.csv', usecols=[2], engine='python')
getAverage1(dataset, 'ph', 2)



print("\nFor field 2\n")



dataset = pd.read_csv('Nir_Project.csv', usecols=[3], engine='python')
getAverage(dataset, 'turbidty', 1)


print("\nFor field 3\n")



dataset = pd.read_csv('Nir_Project.csv', usecols=[4], engine='python')
lin3 = analysis3(dataset,'field3')


print("\nFor field 4\n")


dataset = pd.read_csv('Nir_Project.csv', usecols=[5], engine='python')
```

```
lin3 = analysis4(dataset,'field4')



print("\nFor field 5\n")



dataset = pd.read_csv('Nir_Project.csv', usecols=[6], engine='python')

getAverage5(dataset, 'tds', 2)
```