

**VISVESVARAYA TECHNOLOGICAL UNIVERSITY**



**BELAGAVI-590018**

**“AN INTERNSHIP REPORT”**

(Subject Code: 17IS84)

ON

**“Water Purity Check using Machine Learning”**

Submitted in partial fulfillment for the requirements for the Award of Degree of

**BACHELOR OF ENGINEERING IN  
INFORMATION SCIENCE AND ENGINEERING**

**BY**

**ZEESHAN AHMED KHAN**

**1EP17IS053**

**UNDER THE GUIDANCE OF**

**VAISHALI SHESHRAO**

Ass. Professor

Dept. of ISE, EPCET



**Department of Information Science and Engineering  
Jnana Prabha Campus, Bidarahalli,**

**Bangalore – 560 049**

**2021-2022**



(Affiliated to Visvesvaraya Technological University, Belagavi)

**Jnana Prabha Campus, Bidarahalli,  
Bangalore – 560 049**

## **DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

### **CERTIFICATE**

This is to certify that the **Internship Work** entitled “**Water Purity Check Using Machine Learning**” is a bonafide work carried out by **ZEESHAN AHMED KHAN**, bearing USN **1EP17IS053** in partial fulfillment for the award of **Bachelor of Engineering in Information Science and Engineering** under **Visvesvaraya Technological University, Belagavi** during the year **2021-2022**. This report has been approved as it satisfies the academic requirements in respect of Seminar Work prescribed for the award of the said degree.

#### **GUIDE**

**Prof Vaishali Sheshrao**  
Professor  
Dept of ISE  
EPCET, Bangalore

#### **HOD**

**Dr. Lingaraju GM**  
Head of the  
Department Dept of  
ISE  
EPCET, Bangalore

#### **PRINCIPAL**

**Dr. Sateesh TK**  
Principal  
EPCET,  
Bangalore

#### **Examiners**

**Name of the Examiners**

**Signature with date**

**1.**

**2.**



REG NO:AAN8688

## **PraLoTech Solutions LLP**

PraLoTech Solutions Certifies

**ZEESHAN AHMED KHAN[IEPI7IS053]**

for successfully completing

Training and Internship Programme Conducted for a  
period of 8 weeks between **12<sup>th</sup> APRIL 2022** to **4<sup>th</sup> JUNE 2022**.

We take the pleasure in recognizing the achievement


with the award of

**"Internship Certificate"**

In

**"MACHINE LEARNING WITH PYTHON"**

given on the **4<sup>th</sup>** day of **JUNE 2022**.

  
**PraLoTech Solutions LLP**  
# 1, Silicon Plaza, Near T.C. Palya Signal,  
Subash Nagar, K.R. Puram,  
BANGALORE - 560 049.

Project Manager

PraLoTech Solutions LLP

#1, Silicon Plaza, Near T. C. Palya Signal, Subhashnagar, K. R. Pura, Bangalore -560 049



## DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

(Affiliated to Visvesvaraya Technological University, Belagavi)

Bangalore-560049

### DECLARATION

I GAJNEDRA G, student of 8<sup>th</sup> semester B.E, in Information Science and Engineering, East Point College of Engineering and Technology, Bengaluru, declare that the Internship Project entitled “**WATER PURITY CHECK**” has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree in Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University, Belagavi, during the academic year 2021-2022. The matter bodied in this report has not been submitted to any other university or institution for the award of any other degree.

**ZEESHAN AHMED KHAN**

**1EP17IS053**

## ACKNOWLEDGEMENT

Any achievement, be it scholastic or otherwise does not depend solely on the individual efforts but on the guidance, encouragement and cooperation of intellectuals, elders and friends. We would like to take this opportunity to thank them all.

First and foremost I would like to express our sincere regards and thanks to **Mr. Pramod Gowda** and **Mr. Rajiv Gowda**, CEO's, East Point Group of Institutions, Bangalore, for providing necessary infrastructure and creating good environment.

I express our gratitude to **Dr. Sateesh T K**, Principal, EPCET who has always been a great source of inspiration.

I express our sincere regards and thanks to **Dr. Lingaraju GM**, Professor and Head of Department of Information Science and Engineering, EPCET, Bangalore, for his encouragement and support.

I am obliged to **Mrs. Nandani Gowda**, Professor, who rendered valuable assistance as the internship coordinator.

I am grateful to acknowledge the guidance and encouragement given to us by **Vaishali Sheshrao**, Professor, Department of Information Science and Engineering, EPCET, Bangalore, who has rendered a valuable assistance.

I also extend our thanks to the entire faculty of the **Department of Information Science and Engineering**, EPCET, Bangalore, who have encouraged us throughout the course of the Seminar.

Last, but not the least, I would like to thank my family and friends for their inputs to improve the Seminar.

**ZEESHAN AHMED KHAN**  
**1EP17IS053**

## INDEX

Chapter No	Topics	Page No.
1	<b>About The Company</b> 1.1 Brief history of the Organization 1.2 Overall Organization Structure 1.2.1 Vision 1.2.2 Our Values 1.2.3 Our Goal 1.2.4 Mission 1.3 The Products and the Services Offered By Organization 1.3.1 Corporate Philosophy 1.4 Number of people working in the organization 1.5 Financial Details 1.5.1 Growth Record 1.5.2 Partnership 1.5.3 Experience Certainty 1.5.4 Overall Turnover Or Operational Cost Of Organization 1.6 Current Research And Development 1.6.1 New Technology Capability And Positions	1
2	<b>Department Profile</b> 2.1 Research And Development Centre 2.1.1 System Software And Programming Tools	11

	2.2 Department Centered Organizations	
3	<b>Task Performed</b> 3.1 Need For Artificial Intelligence 3.2 Machine Learning 3.3 Classification Of Machine Learning 3.4 Categorizing On The Basis Of Required Output 3.5 Data Preprocessing For Machine Learning In Python 3.6 Supervised Learning 3.6.1 Types Of Supervised Learning 3.7 Unsupervised Learning 3.8 Reinforcement Learning 3.9 Confusion Matrix in Machine Learning	17
4	<b>Introduction On Project Work</b> 4.1 Introduction 4.2 Problem Statement 4.3 Research Objectives 4.4 Dataset Details 4.5 Architecture	31
	4.6 SRS And Implementation 4.7 System Requirements 4.8 Methodology	34
	<b>Snapshots</b>	37
	<b>Conclusion</b>	45
	<b>Reference</b>	46

## LIST OF FIGURES

<b>Figure No.</b>	<b>Title</b>	<b>Page No.</b>
2.1	System Software and programming tools	12
2.2	IT Organization Layers	15
3.2	Machine Learning	19
3.5	Data Preprocessing	22
3.6	Supervised Learning	23
3.6.1	Types of Supervised Learning	24
4.4	Dataset Details	32
4.5	Architecture	33
4.6	Classification Process	35
4.7	Training Set Optimization Process	36
5.1	User Documentation in Jupyter Notebook	37
5.2	Displaying images in Jupyter Notebook	38
5.3	Handling missing data	39
5.4	Feature Selection	40
5.5	Renaming Attribute Names	41
5.6	Zen of Python	41



# ABSTRACT

The machine learning field, which can be briefly defined as enabling computers make successful predictions using past experiences, has exhibited an impressive development recently with the help of the rapid increase in the storage capacity and processing power of computers. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Python has a huge number of GUI frameworks (or toolkits) available for it, from Tk Inter (traditionally bundled with Python, using Tk) to a number of other cross-platform solutions, as well as bindings to platform-specific (also known as "native") technologies. Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, We are in need of implementing models which will record information about concentrations of air pollutants (so<sub>2</sub>, no<sub>2</sub>, etc). The deposition of these harmful gases in the air is affecting the quality of people's lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available. In this project, machine learning techniques are used to predict the concentration of so<sub>2</sub> in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. Models in time series are employed to predict the so<sub>2</sub> readings in near years or months.

# CHAPTER 1

## ABOUT THE COMPANY

### 1.1 Brief history of the Organization

With the active participation of its multi-disciplinary Assignment Execution Team, Pralotech Solutions has emerged as a leader in the ITES in India and has established itself in the field of software development, data processing, data conversion, digital printing, Digitization, System integration, smart card personalization, IT facility management and other IT enabled services. Pralotech Solutions LLP, incorporated in end of 2018, is a professionally managed, rapidly growing, multifaceted Information technology company. The company is actively involved in developing automation and e-Governance solutions for Transport, Social Security, Citizen Identity, Education, Public Distribution System, Retail Management and a host of other application areas.

Pralotech Solutions is a leading System integrator in India providing complete turnkey solutions on BOO & BOOT basis including facility management services, Smart Cards applications, Document Management System (DMS), Work Flow Management and Manpower Deployment.

Pralotech Solutions has successfully completed many e-governance projects for the various departments of Gov. of Karnataka and has won accolades for its superior service delivery, timely execution of projects and the quality of the deliverables. PRALOTECH SOLUTIONS is being trusted by many clients who are looking for reliable and quality services for their business. PRALOTECH SOLUTIONS is currently operating and managing in Bangalore and giving services to e-commerce business services.

Pralotech Solutions adopted project team and dedicated organization structure. In project based organization, the project managers directors have a high level of power to oversee and control the project assets. The project manager in this structure has downright power over the project and can secure assets expected to fulfill project targets from inside then again outside the parent organization, subject just to the extension, quality

, furthermore, budget constraints are identified in the project.

In the project based structure, staff is particularly relegated to the project and report specifically to the project manager. The project manager is in charge of the execution evaluation and vocation movement of all undertaking colleagues while on the project. This prompts expanded project faithfulness. Complete line power over undertaking endeavors bears the project manager solid undertaking controls and brought together lines of correspondence. This prompts quick response time and enhanced responsiveness. In addition, project work forces are held on a restrictive instead of shared or low maintenance premise. Project teams create an in number feeling of task recognizable proof and possession, with profound faithfulness efforts to the project and a decent comprehension of the way of project's exercises, mission, or objectives.

## **1.2 OVERALL ORGANISATION STRUCTURE**

Pralotech Solutions is completely dedicated to the success of our customers and does not permit external forces to diminish our focus and commitment. To achieve the highest level of customer satisfaction, we follow basic principles to deliver solutions with impact.

### **1.2.1 VISION**

Vision Statement: “To be the pioneer in e-commerce solutions by building and implementing robust and future proof systems which are efficient, transparent and accountable.”

### **1.2.2 OUR VALUES**

The core values which lay the foundation for Pralotech Solutions are:

- **Honesty & Integrity:** Pralotech Solutions individual and business relationships are governed by the highest standards of honesty and integrity. People at all levels adhere to the code of conduct and the highest standards of business ethics, as they believe in conducting their business with uncompromising integrity.

**Respect & Dignity:** Pralotech Solutions respects its customers, recognize that they have different needs and continuously strive towards satisfying those needs by improving the quality of its solutions and services. It trust and respect its people and recognize their contributions to Pralotech Solutions.

- **Terms Spirit & Camaraderie:** Pralotech Solutions believes that focus on Team Work is its competitive advantage. Teams act as catalysts for successful achievement of the organizational goals. Individuals are encouraged to interact with all levels of management, freely share their ideas and suggestions and work together as a cohesive unit.
- **Openness & Transparency:** Pralotech Solutions has an open and transparent culture. Openness facilitates informed decisions, shared understanding and builds an environment of trust in the organization.
- **Empowerment:** Pralotech Solutions employs high caliber people who take responsibility for their actions and exercise good judgment in an environment of mutual trust. Pralotech Solutions seeks to retain its entrepreneurial spirit and minimize bureaucracy.
- **Core Values:** When we take on your project, we take the stewardship of the project with you in the director's seat. As stewards of your project, we consider ourselves successful not when we deliver your final product but when the product meets your business objectives.
- **Integrity:** Honesty in how we deal with our clients, each other and with the world.
- **Candor:** Be open and upfront in all our conversations. Keep clients updated on the real situation. Deal with situations early; avoid last minute surprises.
- **Service:** Seek to empower and enable our clients. Consider ourselves successful not when we deliver our client's final product but when the product is launched and meets success.
- **Kindness:** Go the extra mile. Speak the truth with grace. Deliver more than is expected or promised.
- **Competence:** Benchmark with the best in the business. Try new and better things Never rest on laurels. Move out of comfort zones. Keep suggesting new things. Seek to know more.

### 1.2.3 OUR GOALS

Our company objectives as follows:

- To promote a profitable and sustainable business activity that meets the customer's needs.
- To increase the company's market share
- To gain the competitive edge
- To increase the company's role in relations to social responsibility
- To provide excellent customer service

### 1.2.4 MISSION

“To enable its customers to achieve total e-commerce through innovative solutions using the cutting edge technologies and to provide world class IT and ITES services at affordable costs to the customers with fast turnaround time and to continually improve the service delivery at the client service Centre's managed by us.”

Over the next few years our goal is to harness our talents and skills by permeating our company further with process-centered management. In this way, once a customer's project enters our quality oriented process, it will exit as a quality product.

We will also strive to add to our knowledge and enhance our skills by creating a learning environment that includes providing internal technology seminars, attending conferences and seminars, building a knowledge library and encouraging learning in every way. Our in-house Intranet portal makes sure that knowledge is shared within the organization.

With our beliefs, the future can only look promising as we continue to build our team with the best Indian talent and mould them into our quality-oriented culture. We will find our niche in a competitive world by excelling at what we do, following our guiding principles and most importantly, listening to the needs of our customers, to complete within deadline period is also our mission.

**Contact Information**

Company Name : Pralotech Solutions LLP  
Development &  
Data Processing Centre : #1 Silicon Plaza, Near Usha Dental clinic, subhashnagar,  
T C Palya , K R Puram, Bangalore-560049  
General Phone No : 9611431872, 9164884137  
Company Email : [info@pralotech.com](mailto:info@pralotech.com) , [pralotechsolutions@gmail.com](mailto:pralotechsolutions@gmail.com)  
Website : [www.pralotech.com](http://www.pralotech.com)  
Contact Person : Mr. Lohith.C (lohith@pralotech.com)

### **1.3 THE PRODUCTS AND THE SERVICES OFFERED BY ORGANISATION**

Company offer the key products and services you would expect from a leading Microsoft Gold and Oracle partner including Web, Software Development and Mobile application, Integration, Consultancy and Support Services.

What sets us aside is our focus, vision and capability to deliver. We are highly accredited, come highly recommended and invest heavily in both product development and our first class consultants.

Pralotech Solutions LLP is one of India most well-known and well-trusted solution provider. Today, Pralotech Solutions stands as a source of reliable and innovative products that enhance the quality of costumer's professional and personal lives.

Its employees in all the branches are active in the areas of production, software development, Implementation, system integration, and training.

Why Pralotech?

With a client list spanning nearly in all industries, and colleges, Pralotech Solutions product solutions have benefited customers of many different sizes, from non-profit organizations to companies.

By acquaintance with Pralotech Solutions, you'll have access to current IT research, tools, templates, and step-by-step action plans for completing Key projects. You'll also be provided full access to our research archives and knowledge base.

### **1.3.1 COPERATE PHILOSOPHY**

- **Quality Policy:** “Pralotech Solutions is committed to provide world class Information Technology Enabled Services (ITES) to its customers with high accuracy, unmatched quality and fast turnaround time”.
- **Security Policy:** “Pralotech Solutions understands that the trust of the client in it depends on how well it keeps their personal, business and accounts information secure. PRALOTECH SOLUTIONS follows international standards set under Information Security Management System (ISMS) policy guidelines.”

### **1.4 NUMBER OF PEOPLE WORKING IN THE ORANGISATION**

Pralotech Solutions employs more than 50 professionals with various skill set and professional competence. Have different project execution teams for different application areas in Information Technology Industry. The Human Resources available with Pralotech Solutions, their qualification and technical skills are depicted below.

Human capital is our most important asset. A qualified and highly specialized team with multi-disciplinary approach forms the technical core at Six Axis. This repository of talented and committed software developers has a proven track record to ensure success in IT solution implementation. With skills ranging from business process re-engineering to application development, Pralotech Solutions technical team seeks to constantly enhance and expand its technical knowledge. Capturing knowledge through procedures and processes is the premise on which the entire organization works. Pralotech Solutions resource base consists of IIT engineers (three including the Directors), management graduates, masters in computer applications and domain experts from various fields.

Pralotech Solutions have improved the quality of communication and satisfied customers. We have earned their respect by providing excellent products and services. In addition, we are flexible with services and financial structures for contracts aiming for mutually beneficial relationships with our customers.

Our customers are dynamic and diverse and include Large Corporate Offices, Universities, Educational Institutions, Factories, etc.

## **1.5 FINANCIAL DETAILS**

PRALOTECH SOLUTIONS provides flexible investment solutions, such as leasing, financing, utility programs and asset management services, for customers to enable the creation of unique technology deployment models and acquisition of complete IT solutions, including hardware, software and services from PRALOTECH SOLUTIONS and others.

Providing flexible services and capabilities that support the entire IT lifecycle, partners with customers globally to help build investment strategies that enhance their business agility and support their business transformation. PRALOTECH SOLUTIONS offers a wide selection of investment solution capabilities for large enterprise customers and channel partners, along with an array of financial options to SMBs, educational and governmental entities.

Corporate Investments Corporate Investments includes PRALOTECH SOLUTIONS Labs and certain cloud-related business incubation projects among others. Sales, Marketing and Distribution We manage our business and report our financial results based on the business segments described above.

### **1.5.1 GROWTH RECORD**

Since its inception and with initial small steps, Pralotech Solutions is now progressing by leaps and bounds. It has grown from a small venture to a medium scale enterprise with a strong 80+ workforce, our rate of more than 100%. The company is executing some of the prestigious projects and has earned a very respectable name in the Indian IT and e-commerce industry.

### **1.5.2 PARTNERSHIP**

Our innovative and highly integrated approach means customers benefit from working with specialists. Our continuous strive to be a technology leader in the industry means that our clients directly benefit from the huge expertise that our people possess.

We strive to be at the forefront of technology that enables us to provide you with highly effective and optimized solutions to all your problems.



Clients like to have a single point-of-contact for their solutions, and expect a complete solution from the vendor, which is not possible unless there are partnerships and alliances within and outside the company.

Pralotech Solutions fosters partnerships with companies with whom a value proposition can be offered to clients.

One of the key benefits that you receive by partnering with Pralotech Solutions is increased project completion certainty, project transparency, renewed customer confidence and credibility from our unparalleled track record, mature processes and quality recognition and customer endorsement.

### **1.5.3 EXPERIENCE CERTAINTY**

True certainty of success comes from working with a partner you trust to provide the insight, support and expertise that will propel your business forward. Experiencing certainty with Pralotech Solutions means you can count on results, partnership and leadership. When you work with us, your long-term success is our motivation. This is why we can offer you the ability to meet every challenge and the ability to capitalize on every opportunity. That's the power of certainty. And it is our promise to every client.

### **1.5.4 OVERALL TURNOVER OR OPERATIONAL COST OF ORGANISATION**

For most of our products, we have existing alternate sources of supply or such alternate sources of supply are readily available. However, we do rely on sole sources for laser printer engines, LaserJet supplies, certain customized parts and parts for products with short life cycles (although some of these sources have operations in multiple locations in the event of a disruption). We are dependent upon Intel and AMD as suppliers of x86 processors and Microsoft for various software products; however, we believe that disruptions with these suppliers would result in industry-wide dislocations and therefore would not disproportionately disadvantage us relative to our competitors.

See “Risk Factors—we depend on third-party suppliers, and our financial results could suffer if we fail to manage suppliers properly,” in Item 1A, which is incorporated herein by reference. Like other participants in the IT industry, we ordinarily acquire materials and

components through a combination of blanket and scheduled purchase orders to support our demand requirements for periods averaging 90 to 120 days. From time to time, we may experience significant price volatility or supply constraints for certain components that are not available from multiple sources.

Frequently, we are able to obtain scarce components for somewhat higher prices on the open market, which may have an impact on our gross margin but does not generally disrupt production. We also may acquire component inventory in anticipation of supply constraints or enter into longer-term pricing commitments with vendors to improve the priority, price and availability of supply. See “Risk Factors— we depend on third-party suppliers, and our financial results could suffer if we fail to manage suppliers properly,” in Item 1A, which is incorporated herein by reference. Research and Development Innovation is a key element of our culture. Our development efforts are focused on designing and developing products, services and solutions that anticipate customers’ changing needs and desires, and emerging technological trends.

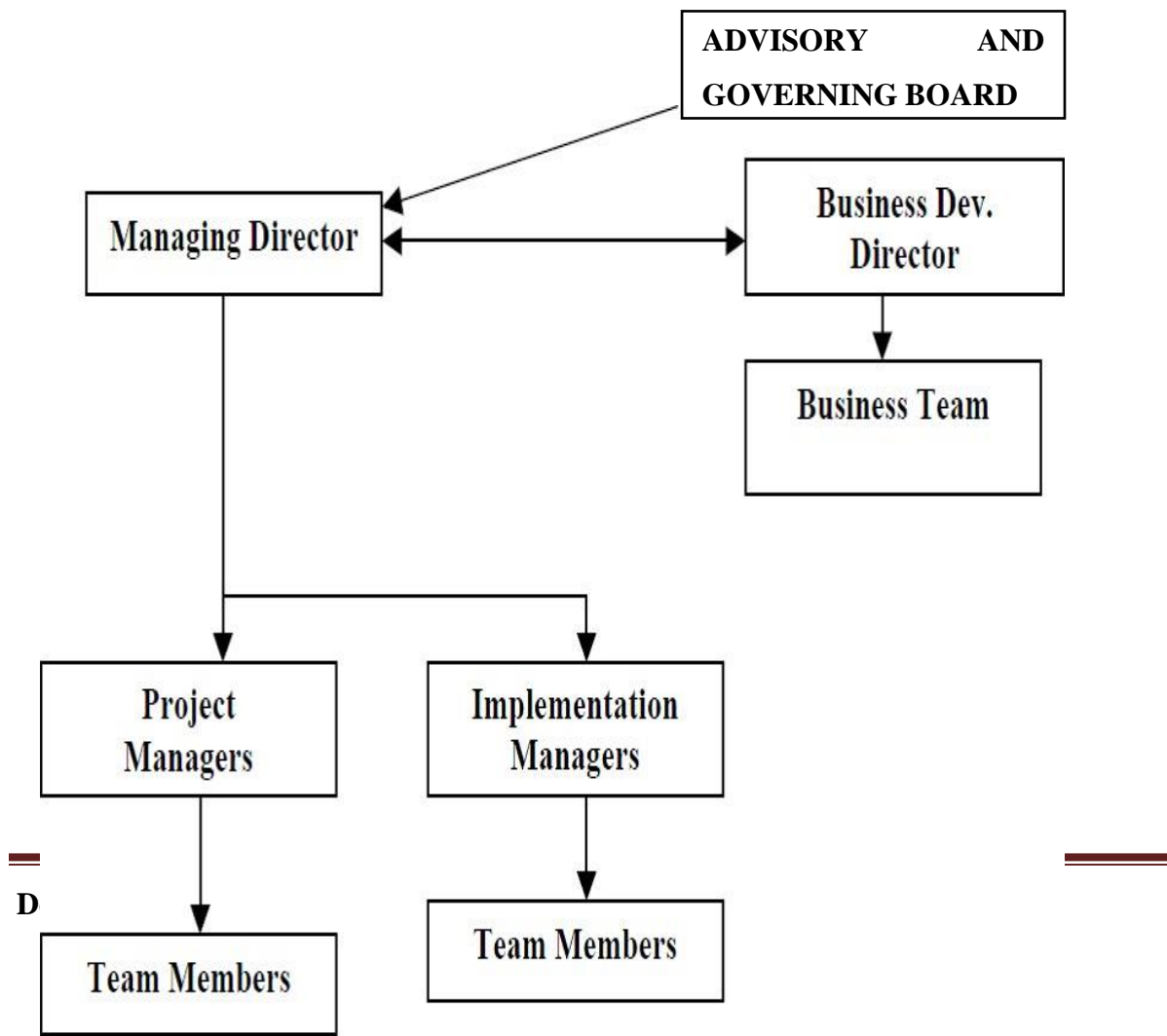
Our efforts also are focused on identifying the areas where we believe we can make a unique contribution and the areas where partnering with other leading technology companies will leverage our cost structure and maximize our customers’ experiences. PRALOTECH SOLUTIONS together with the various research and development groups within our business segments, are responsible for our research and development efforts. We anticipate that we will continue to have significant research and development expenditures in the future to support the design and development of innovative, high-quality products and services to maintain and enhance our competitive position. For a discussion of risks attendant to our research and development activities, see “Risk Factors— If we cannot successfully execute on our strategy and continue to develop, manufacture and market products, services and solutions that meet customer requirements for innovation and quality, our revenue and gross margin may suffer,” in Item 1A, which is incorporated herein by reference.

## 1.5.5 ORGANISATION STRUCTURE OF THE COMPANY

**Fig. 1.1 Organization Structure**

### 1.6.1 CURRENT RESEARCH AND DEVELOPMENT

- The current R & D efforts are primarily aimed at the following segments in the healthcare industry:
- Developing a system for integrating medical schools with major hospitals for knowledge gathering, sharing and learning
- Developing a Clinical Decision Support System to aid doctors in difficult to diagnose cases using Artificial Intelligence and Probabilistic Techniques.



## 1.6.2 NEW TECHNOLOGY CAPABILITY AND POSITIONS

The organization has a process in place, which addresses the issue of incorporating emerging technologies into the product design. The process is as follows:

- Core committee on new development evaluates and identifies new technology for the purpose of integration.
- The research and development department identifies the resource and people and formulates the process for working while setting key performance indicator.
- A thorough study of the new technology along the tools is made and documented.
- Estimates are made as to the impact of the new technology on the products developed by the company.
- Effort estimates are made for introducing the new technologies
- Client feedback is received about the efforts needed and the advantages of the new technology.
- The core committee takes a knowledgeable decision as to the advantages and efforts required and approve the introduction of the technology
- The affected personnel are trained in the new technologies
- The new technology is introduced and the product is enhanced
- The clients are informed about the enhancement and introduction of related documents are prepared for the changeover
- The clients are guided in implementing the new technologies

Being a technology driven company we are always exploring ways of enhancing our product capabilities and aim at providing the latest state-of-the-art products to our customers. We have incorporated the PDAs and the smart card already in the system. We are currently evaluating blue tooth capability and the Tablet PC relevance to the field.

## **CHAPTER 2**

### **DEPARTMENT PROFILE**

#### **2.1 RESEARCH AND DEVELOPMENT CENTRE**

Pralotech Solutions have the ability to architect, develop and maintain any complex software applications. Pralotech Solutions development team and research team is committed to continuing research and development in the rapidly evolving fields of software development and IT, so that informed decisions can be undertaken at the appropriate time regarding future technology choices and adoption and to help drive the continuing evolution of our software architecture.

Over the course of several years, PRALOTECH SOLUTIONS has used the benefit of its knowledge and experience of developing enterprise wide, web based applications coupled with its continuing research and development activity to develop its own in-house web based software architecture and supporting framework on which all of its current and future web based solutions are based. Conforming to the latest industry standards and best practice, PraLoTech software architecture has proven to be a reliable, robust, and scalable foundation on which to build its software products. A qualified and highly specialized team with multi-disciplinary approach forms the technical core at PraLoTech. This repository of talented and committed software developers has a proven track record to ensure success in IT solution implementation. Pralotech Solutions resource base consists of IIT engineers (three including the Directors), management graduates, masters in computer applications and domain experts from various fields.

##### **2.1.1 SYSTEM SOFTWARE AND PROGRAMMING TOOLS**

The software team at Pralotech Solutions has extensively worked on various flavors of UNIX and Windows based environment. Few of the UNIX operating systems, which have been used by the organization, are Sun Solaris, Tru64 UNIX & Linux.

The team at PRALOTECH SOLUTIONS has developed large scale and complex applications on Oracle, SQL Server and DB2. There is also substantial working expertise on MySQL and MS Access.

Pralotech Solutions stay relevant to their enterprise customers by helping them with transformational technology solutions utilizing SMAC structure and help in their growth journey and tailored offerings under various technology domains has been providing best quality, timely & cost effective solutions to its clients & has developed a long term strategic partnership with them all across the world.



**Fig.2.1 System Software and programming tools**

The web server experience extends to the following:

- Apache on Unix and NT Servers
- IIS on NT Platforms
- JDBS
- SQLYOG
- Net beans IDE with Servers

Operating Systems	:	Windows 7, Windows 8/8.1/10
Database Environment	:	DB2, Oracle, SQL Server, SQLYOG, MySQL.
Languages	:	Python, Java, EJB, XML, RMI, WAP, C/C++, CL/400, ASP and COBRA
Web Enabled Systems	:	MS-IIS, Visual Interdev, Web sphere, Web logic
Front End Tools	:	Tkinter, Visual Basic, Visual C++, Power Builder,
Web Designing Tool	:	FrontPage 2000, MySQL YOG
Data and Object Modeling	:	Rational Rose

Our Team members have extensive knowledge in Oracle Products ranging from Oracle

7.3 to 10i, Developer 6i, Oracle 10iAS and other oracle products.

Goal management is about more than just the annual assigning of goals and reviewing of employee performance. It's about getting every employee to use and develop their talents, skills and experience to help the organization. Too often, managers and employees set goals that are not SMART or in line with corporate strategies, resulting in lost productivity and disengaged workers. Goal management is more than just a once-a-year exercise – it is an opportunity to align the focus of the entire company, clarify performance expectations, and guide employees to success.

SET	ALIGN	MONITOR	ASSESS
Employee performance makes it easy for managers and employees to set SMART goals at the beginning of review cycle.	Easily cascade the company objectives to individuals or teams and link personal goals to organizational success	Log in year-round to update and comment on goals and track progress throughout the entire review cycle.	Measure goals using any rating scale that you choose. Comments & 360° feedback help ensure ratings are clear and accurate.

### Traditional IT Organizational Structure Issues:

Traditional IT organizations are typically structured to support vertical business units and applications. The roles, responsibilities, skills and budgets are focused on several discrete projects that address specifically business activities. In the traditional IT organization, projects are scoped and implemented without fully recognizing the core business processes that span business units.

Marketing department advances the business and drives offers of the items and administrations. It gives the vital exploration to recognize and target clients and different audiences. The marketing department consists of Marketing and Sales departments. The presales department further contains Healthcare, Education, Retail and Networking departments. Thus opportunities to radically improve business processes are overlooked.

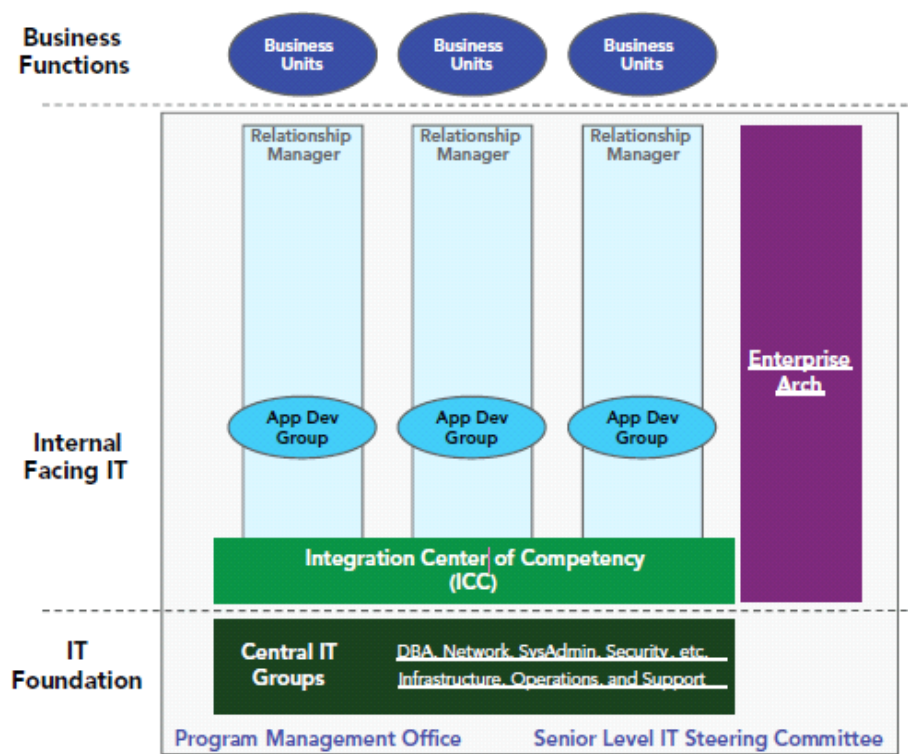


Fig. 2.2 IT Organization Layers

## 2.2 DEPARTMENTAL CENTERED ORGANISATION

Department centered development organizations start to become practical as a group grows above 25 developers or 5 projects. At these staffing levels, there are sufficient people to form multiple departments centered on particular software skills or life cycle areas. For instance, a 40-person group might have departments for:

- System and database administrators
- User interface programmers
- Application programmers

A common mistake in department-centered organizations is to break software architects into a separate department or group. We have found this can lead to elitism and be very counterproductive. First, it starts to separate the architects from the developers who are doing the actual implementation. If developers are too separated from architects, they may



have a built-in incentive to prove the architect's design was wrong by not working there hardest to implement it.

## CHAPTER 3

### TASK PERFORMED

Intelligence, as we know, is the ability to acquire and apply the knowledge. Knowledge is the information acquired through experience. Experience is the knowledge gained through exposure (training). Summing the terms up, we get **artificial intelligence** as the “copy of something natural (i.e., human beings) ‘WHO’ is capable of acquiring and applying the information it has gained through exposure.”

**Intelligence is composed of:**

- Reasoning
- Learning
- Problem Solving
- Perception
- Linguistic Intelligence

Many tools are used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics. The AI field draws upon computer science, mathematics, psychology, linguistics, philosophy, neuro-science, artificial psychology and many others.

### 3.1 NEED FOR ARTIFICIAL INTELLIGENCE

1. To create expert systems which exhibit intelligent behavior with the capability to learn, demonstrate, explain and advice its users.
2. Helping machines find solutions to complex problems like humans do and applying them as algorithms in a computer-friendly manner.

**Applications of AI include Natural Language Processing, Gaming, Speech Recognition, Vision Systems, Healthcare, Automotive etc**

Many times, students get confused between Machine Learning and Artificial Intelligence, but Machine learning, a fundamental concept of AI research since the field's

inception, is the study of computer algorithms that improve automatically through experience. The mathematical analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as a computational learning theory.

Stuart Shapiro divides AI research into three approaches, which he calls computational psychology, computational philosophy, and computer science. Computational psychology is used to make computer programs that mimic human behavior. Computational philosophy is used to develop an adaptive, free-flowing computer mind. Implementing computer science serves the goal of creating computers that can perform tasks that only people could previously accomplish.

**AI has developed a large number of tools to solve the most difficult problems in computer science, like:**

- Search and optimization
- Logic
- Probabilistic methods for uncertain reasoning
- Classifiers and statistical learning methods
- Neural networks
- Control theory
- Languages

## **3.2 MACHINE LEARNING**

Machine Learning (ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines (computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate.



### Basic Difference in ML and Traditional Programming?

- **Traditional Programming:** We feed in DATA (Input) + PROGRAM (logic), run it on machine and get output.
- **Machine Learning:** We feed in DATA (Input) + Output, run it on machine during training and the machine creates its own program (logic), which can be evaluated while testing.

## 3.3 CLASSIFICATION OF MACHINE LEARNING

Machine learning implementations are classified into three major categories, depending on the nature of the learning “signal” or “response” available to a learning system which are as follows:-

1. **Supervised learning:** When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples.
2. **Unsupervised learning:** Whereas when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new

useful inputs to supervised machine certainty logical

algorithms and commonly as a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

3. **Reinforcement learning:** When you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes comes under the category of Reinforcement learning, which is connected to applications for which the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. Errors help you learn because they have a penalty added (cost, loss of time, regret, pain, and so on), teaching you that a certain course of action is less likely to succeed than others. In this case, an application presents the algorithm with examples of specific situations, such as having the gamer stuck in a maze while avoiding an enemy. The application lets the algorithm know the outcome of actions it takes, and learning occurs while trying to avoid what it discovers to be dangerous and to pursue survival. You can have a look at how the company Google Deep Mind has created a reinforcement learning program that plays old Atari's videogames. When watching the video, notice how the program is initially clumsy and unskilled but steadily improves with training until it becomes a champion.
4. **Semi-supervised learning:** where an incomplete training signal is given: a training set with some (often many) of the target outputs missing. There is a special case of this principle known as Transduction where the entire set of problem instances is known at learning time, except that part of the targets are missing.

### 3.4 CATEGORIZING ON THE BASIS OF REQUIRED OUTPUT

Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system:

1. **Classification:** When inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example

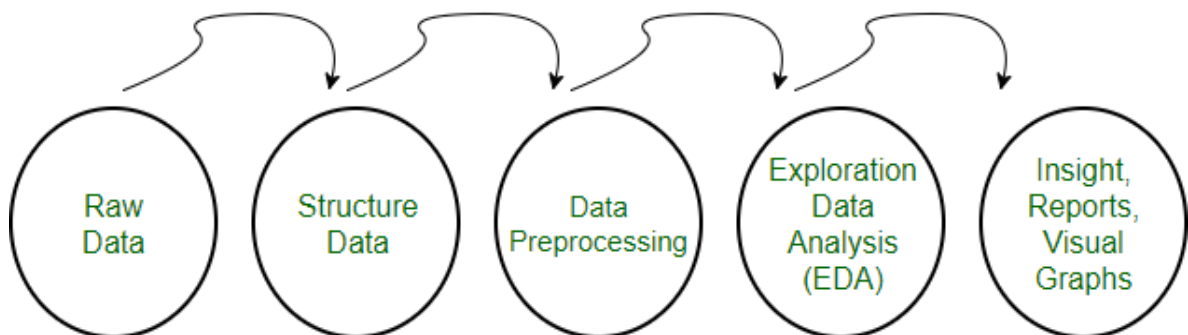
of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

2. **Regression:** Which is also a supervised problem, A case when the outputs are continuous rather than discrete.
3. **Clustering:** When a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task

### 3.5 DATA PREPROCESSING FOR MACHINE LEARNING IN PYTHON

- Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.



#### Need of Data Preprocessing

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original.

Another aspect is that data set should be formatted in such a way that more than one

Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

### 3 different data preprocessing techniques for machine learning

#### 1. Rescale Data

- When our data is comprised of attributes with varying scales, many machine learning algorithms can benefit from rescaling the attributes to all have the same scale.
- This is useful for optimization algorithms in used in the core of machine learning algorithms like gradient descent.
- It is also useful for algorithms that weight inputs like regression and neural networks and algorithms that use distance measures like K-Nearest Neighbors.
- We can rescale your data using skit-learn using the [MinMaxScaler](#) class.

#### 2. Binarize Data (Make Binary)

- We can transform our data using a binary threshold. All values above the threshold are marked 1 and all equal to or below are marked as 0.
- This is called binarizing your data or threshold your data. It can be useful when you have probabilities that you want to make crisp values. It is also useful when feature engineering and you want to add new features that indicate something meaningful.
- We can create new binary attributes in Python using scikit-learn with the [Binarize](#) class.

#### 3. Standardize Data

- Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.
- We can standardize data using scikit-learn with the [StandardScaler](#) class.

### 3.6 Supervised Learning:

Supervised learning is when the model is getting trained on a labelled dataset. **Labelled** dataset is one which has both input and output parameters. In this type of learning both training and validation datasets are labelled as shown in the figures below.

User ID	Gender	Age	Salary	Purchased	Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
15624510	Male	19	19000	0	10.69261758	986.882019	54.19337313	195.7150879	3.278597116
15810944	Male	35	20000	1	13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
15668575	Female	26	43000	0	17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
15603246	Female	27	57000	0	20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
15804002	Male	19	76000	1	22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
15728773	Male	27	58000	1	24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
15598044	Female	27	84000	0	24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
15694829	Female	32	150000	1	23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
15600575	Male	25	33000	1	22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
15727311	Female	35	65000	0	20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
15570769	Female	26	80000	1	17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
15606274	Female	26	52000	0	11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
15746139	Male	20	86000	1	14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
15704987	Male	32	18000	0	18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
15628972	Male	18	82000	0	22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
15697686	Male	29	80000	0	24.23155922	988.796875	19.74790765	318.3214111	0.329656571
15733883	Male	47	25000	1					

Figure A: CLASSIFICATION

Figure B: REGRESSION

Both the above figures have labelled data set –

- **Figure A:** It is a dataset of a shopping store which is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age and salary.

**Input :** Gender, Age, Salary

**Output :** Purchased i.e. 0 or 1 ; 1 means yes the customer will purchase and 0 means that customer won't purchase it.

- **Figure B:** It is a Meteorological dataset which serves the purpose of predicting wind speed based on different parameters.

**Input :** Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction

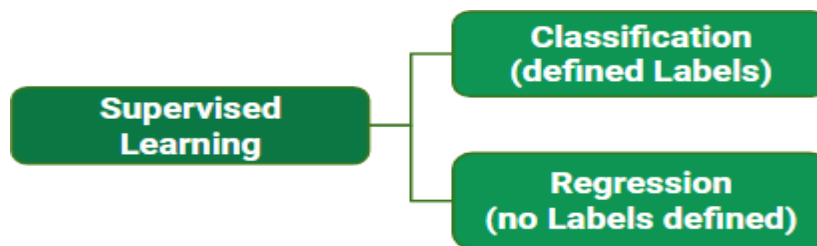
**Output :** Wind Speed

**Training the system:**

While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and rest as testing data. In training data, we feed input as well as output for 80% data. The model learns from training data only. We use different machine learning algorithms (which we will discuss in detail in next articles) to build our model. By learning, it means that the model will build our model.

Once the model is ready then it is good to be tested. At the time of testing, input is fed

from remaining 20% data which the model has never seen before, the model will predict some value and we will compare it with actual output and calculate the accuracy.



### Types of Supervised Learning:

1. **Classification:** It is a Supervised Learning task where output is having defined labels (discrete value). For example in above Figure A, Output – Purchased has defined labels i.e. 0 or 1; 1 means the customer will purchase and 0 means that customer won't purchase. The goal here is to predict discrete values belonging to a particular class and evaluate on the basis of accuracy.
2. It can be either binary or multi class classification. In **binary** classification, model predicts either 0 or 1; yes or no but in case of **multi class** classification, model predicts more than one class.

**Example:** Gmail classifies mails in more than one classes like social, promotions, updates, forum.

3. **Regression:** It is a Supervised Learning task where output is having continuous value. Example in above Figure B, Output – Wind Speed is not having any discrete value but is continuous in the particular range. The goal here is to predict a value as much closer to actual output value as our model can and then evaluation is done by calculating error value. The smaller the error the greater the accuracy of our regression model.

### Example of Supervised Learning Algorithms:

- Linear Regression



- Nearest Neighbor
- Gaussian Naive Bayes
- Decision Tree
- Support Vector Machine (SVM)

- Random Forest

### 3.7 Unsupervised learning

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore machine is restricted to find the hidden structure in unlabeled data our-self.

**For instance**, suppose it is given an image having both dogs and cats which have not seen ever.



Thus the machine has no idea about the features of dogs and cat so we can't categorize it in dogs and cats. But it can categorize them according to their similarities, patterns, and differences i.e., we can easily categorize the above picture into two parts. First may contain all pics having **dogs** in it and second part may contain all pics having **cats** in it. Here you didn't learn anything before, means no training data or examples.

Unsupervised learning classified into two categories of algorithms:

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

### 3.8 Reinforcement learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and

machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience.

**Example :** The problem is as follows: We have an agent and a reward, with many hurdles in between. The agent is supposed to find the best possible path to reach the reward. The following problem explains the problem more easily.



The above image shows robot, diamond and fire. The goal of the robot is to get the reward that is the diamond and avoid the hurdles that is fire. The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles. Each right step will give the robot a reward and each wrong step will subtract the reward of the robot. The total reward will be calculated when it reaches the final reward that is the diamond.

#### **Main points in Reinforcement learning –**

- Input: The input should be an initial state from which the model will start
- Output: There are many possible output as there are variety of solution to a particular problem
- Training: The training is based upon the input, The model will return a state and the user will decide to reward or punish the model based on its output.
- The model keeps continues to learn.
- The best solution is decided based on the maximum reward.

## Difference between Reinforcement learning and Supervised learning:

REINFORCEMENT LEARNING	SUPERVISED LEARNING
Reinforcement learning is all about making decisions sequentially. In simple words we can say that the output depends on the state of the current input and the next input depends on the output of the previous input	In Supervised learning the decision is made on the initial input or the input given at the start
In Reinforcement learning decision is dependent, So we give labels to sequences of dependent decisions	Supervised learning the decisions are independent of each other so labels are given to each decision.
Example: Chess game	Example: Object recognition

**Types of Reinforcement:** There are two types of Reinforcement:

### 1. Positive

Positive Reinforcement is defined as when an event, occurs due to a particular behavior, increases the strength and the frequency of the behavior. In other words it has a positive effect on the behavior.

Advantages of reinforcement learning are:

- Maximizes Performance
- Sustain Change for a long period of time

Disadvantages of reinforcement learning:

- Too much Reinforcement can lead to overload of states which can diminish the results

### 2. Negative

Negative Reinforcement is defined as strengthening of a behavior because a negative condition is stopped or avoided.

Advantages of reinforcement learning:

- Increases Behavior
- Provide defiance to minimum standard of performance

Disadvantages of reinforcement learning:

- It Only provides enough to meet up the minimum behavior

### 3.9 Confusion Matrix in Machine Learning

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix.

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

#### Confusion Matrix:

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

	<i>Class 1 Predicted</i>	<i>Class 2 Predicted</i>
<i>Class 1 Actual</i>	TP	FN
<i>Class 2 Actual</i>	FP	TN

Here,

- Class 1 : Positive

- Class 2: Negative

**Definition of the terms:**

- Positive (P): Observation is positive (for example: is an apple).
- Negative (N): Observation is not positive (for example: is not an apple).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

**Classification Rate/Accuracy:**

Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

**Recall:**

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

**Precision:**

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).

Precision is given by the relation:

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

**High recall, low precision:** This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

**Low recall, high precision:** This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

**F-measure:**

Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$\textbf{F - measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

## CHAPTER 4

### INTRODUCTION ON PROJECT WORK

In the developing countries like India, the rapid increase in population and economic upswing in cities have lead to environmental problems such as air pollution , water pollution ,noise pollution and many more. Water pollution has direct impact on human health. There has been increased public awareness about the same in our country. Contaminated water and poor sanitation are linked to transmission of diseases such as Cholera, diarrhea, hepatitis A, typhoid and polio patients are rapidly increase in our India. Precised water quality forecasting can reduce the effect of maximal pollution on the humans and other natural resources as well. Hence, enhancing water quality forecasting is one of the prime targets for the society.

Access to safe drinking water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water quality supply and sanitation can yield a net economic benefits, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions. The main pollutants include bacteria, fertilizer, pesticides, pharmaceutical, nitrates, phosphates, plastics, faecal waste and even radioactive substances. Every year, over 2 billion pounds of pollutants are dumped into our waterways by power plants- the largest source of toxic water pollution in our country. This wastewater contains heavy metals and chemicals which are very harmful for human health.

In this project we will evaluate or check the quality of water which is safe to drink the water. The contents of this project are: pH value, hardness, solids (total dissolved solids-TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, portability. The calculation of this content in the water can influence the habitat suitability for plant communities as well as animal life. Indicates if water is safe for human consumption where 1 means potable and 0 means not potable.The proposed system is capable of predicting the portability of water for forthcoming months/years.



# SYSTEM STUDY

## 4.1 SYSTEM ANALYSIS

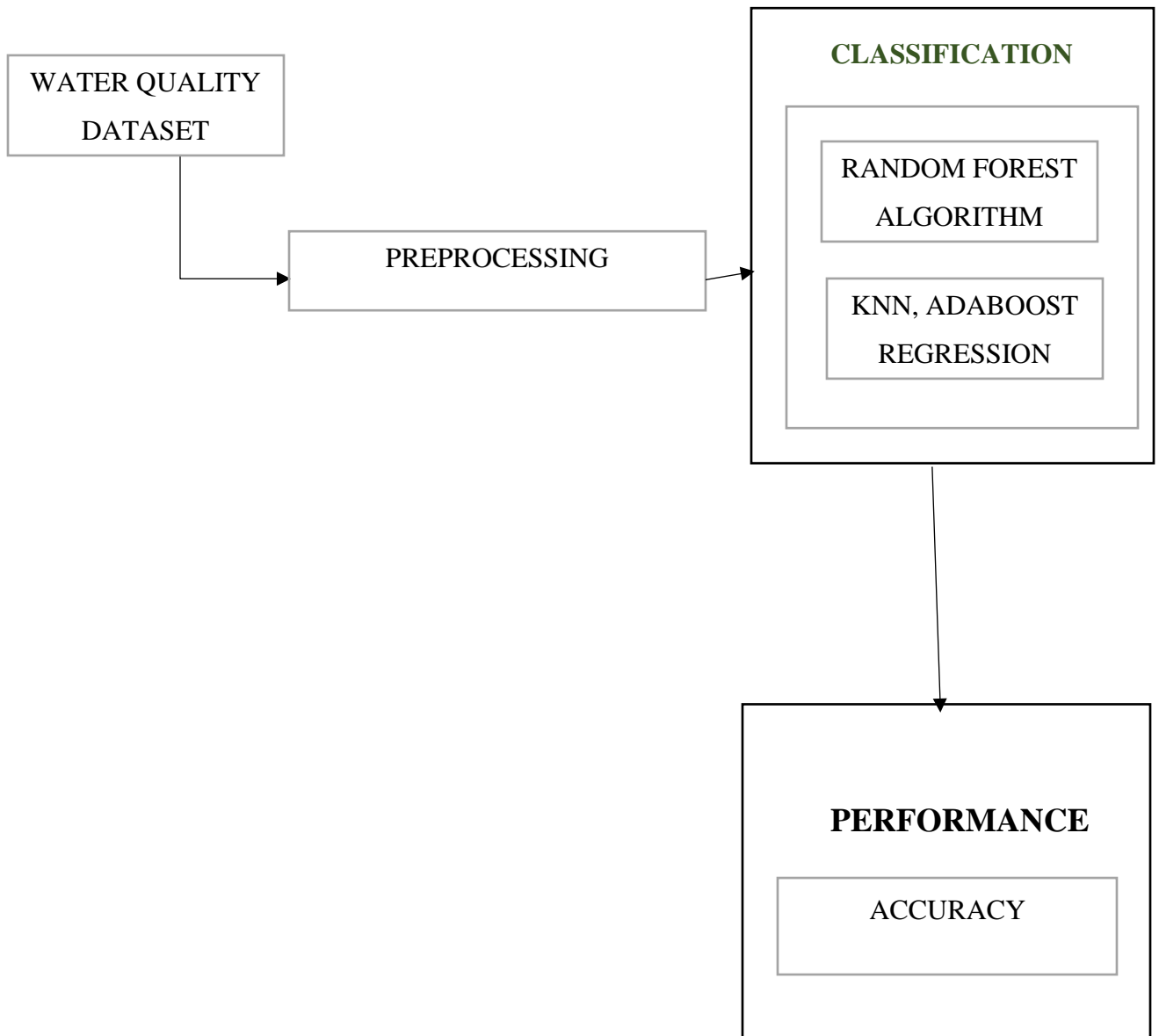
System analysis is a process of gathering and interpreting facts, diagnosing problems and the information to recommend improvements on the system. It is a problem solving activity that requires intensive communication between the system users and system developer's. System analysis or study is an important phase of any system development process. The system is studied to the minutest detail and analyzed. The system plays the role of the interrogator and dwells deep into the working of the present system.

### About Dataset

1. pH value: PH is an important parameter in evaluating the acid–base balance of water. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
2. Hardness: Hardness is mainly caused by calcium and magnesium salts. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
3. Solids (Total dissolved solids - TDS): Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.
4. Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
5. Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food.
6. Conductivity: Pure water is not a good conductor of electric current rather's a good insulator. According to WHO standards, EC value should not exceeded 400  $\mu\text{S}/\text{cm}$ .

7. Organic\_carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.
8. Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
9. Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
10. Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

## SYSTEM ARCHITECTURE



## SYSTEM REQUIREMENTS

### Hardware Requirements

Processors	:	Intel I3 2.2 Ghz
RAM	:	4 GB.
Storage	:	100 GB.
Monitor	:	15”
Keyboard	:	Standard 102 keys

### Software (Tools & Technologies) Requirements

Coding	:	Python
Platform	:	python 3.8
Tool	:	Spyder
OS	:	Windows 8

## IMPLEMENTATION

### Important source of the system

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.metrics import classification_report, accuracy_score

from sklearn.model_selection import GridSearchCV

from sklearn.preprocessing import MinMaxScaler

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import confusion_matrix

from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

import xgboost

from xgboost import XGBClassifier

import numpy as np

import warnings

warnings.filterwarnings('ignore')

data=pd.read_csv('water_potability.csv')

data.head()

data.info()

""" CHECKING THE PERCENTAGE OF MISSING VALUES IN DATASET """

for col in data.columns:
```

```
p=(data[col].isnull().sum()/len(data))*100

print('the column {0} have {1} percent of NAN values'.format(col,p.round(2)))

print()

# data.drop(['Sulfate'],axis=1,inplace=True)

""" REPLACING MISSING VALUE BY MEAN OF ALL VALUES IN RESPECTIVE
COLUMN """

def replace_nan_by_mean(info):

    for col in info.columns:

        info[col].fillna(np.mean(info[col]),inplace=True)

    return info

data=replace_nan_by_mean(data)

data.describe()

data.info()

inp =
['ph','Hardness','Solids','Chloramines','Sulfate','Conductivity','Organic_carbon','Trihalometha
nes','Turbidity']

X_train,X_test,y_train,y_test=train_test_split(data[inp],data['Potability'],test_size=0.2,rando
m_state=42)

#"DATA VISUALIZATION"

plt.figure(figsize=(15,12))

sns.heatmap(X_train.corr(),annot=True,vmin=-1)

plt.show()

plt.figure(figsize=(18,15))

sns.pairplot(X_train)

plt.show()

plt.figure(figsize=(20,20))

for i in range(8):

    plt.subplot(4,2,(i%8)+1)

    sns.distplot(X_train[X_train.columns[i]])
```

```
plt.title(X_train.columns[i],fontdict={'size':20,'weight':'bold'},pad=3)

plt.show()

plt.figure(figsize=(20,20))

for i in range(8):

    plt.subplot(4,2,(i%8)+1)

    sns.distplot(X_train[X_train.columns[i]])

    plt.title(X_train.columns[i],fontdict={'size':20,'weight':'bold'},pad=3)

plt.show()

#"SCALING DATA"

scaler=MinMaxScaler()

train_x_std=scaler.fit_transform(X_train)

test_x_std=scaler.transform(X_test)

models_scores=pd.DataFrame()

model_log = LinearRegression()

model_log.fit(train_x_std,train_data)

log_acc=accuracy_score(test_data,model_log.predict(test_x_std))

model_acc=pd.DataFrame({'Model name':['Linear Regression'],'Accuracy':[log_acc]})

models_scores=models_scores.append(model_acc,ignore_index=True)

print('Train          report          of          linear          Regression
\n',classification_report(train_data,model_log.predict(train_x_std)))

print('Test          report          of          linear          Regression
\n',classification_report(test_data,model_log.predict(test_x_std)))

plt.figure(figsize=(10,8))

sns.heatmap(confusion_matrix(test_data,model_log.predict(test_x_std)),annot=True,)

plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})

plt.show()

model_tree=DecisionTreeClassifier()

grid_tree=GridSearchCV(model_tree,param_grid={'max_depth':range(6,11)})
```

```
grid_tree.fit(X_train,y_train)

tree_acc=accuracy_score(y_test,grid_tree.predict(test_x_std))

model_acc=pd.DataFrame({'Model name':['Decision Tree classifier'],'Accuracy':[tree_acc]})

models_scores=models_scores.append(model_acc,ignore_index=True)

grid_tree.best_params_

print('Train          report          of          DecisionTreeClassifier
\n',classification_report(y_train,grid_tree.predict(train_x_std)))

print('Test          report          of          DecisionTreeClassifier
\n',classification_report(y_test,grid_tree.predict(test_x_std)))

plt.figure(figsize=(10,8))

sns.heatmap(confusion_matrix(y_test,grid_tree.predict(test_x_std)),annot=True)

plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})

plt.xlabel('Predicted value')

plt.ylabel('Actual value')

plt.show()

model_forest=RandomForestClassifier()

grid_forest=GridSearchCV(model_forest,param_grid={'max_depth':range(6,11)})

grid_forest.fit(X_train,y_train)

forest_acc=accuracy_score(y_test,grid_forest.predict(test_x_std))

model_acc=pd.DataFrame({'Model          name':['Random          Forest
Classifier'],'Accuracy':[forest_acc]})

models_scores=models_scores.append(model_acc,ignore_index=True)

print('best param',grid_forest.best_params_)

print('best score',grid_forest.best_score_)

print('Train          report          of          RandomForestClassifier
\n',classification_report(y_train,grid_forest.predict(train_x_std)))

print('Test          report          of          RandomForestClassifier
\n',classification_report(y_test,grid_forest.predict(test_x_std)))

plt.figure(figsize=(10,8))
```



```
sns.heatmap(confusion_matrix(y_test,grid_forest.predict(test_x_std)),a
nnot=True) plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})
plt.xlabel('Predicted value')
plt.ylabel('Actual value')
plt.show()
model_xgb=XGBClassifier(n_estimators=10)
# grid_xgb=GridSearchCV(model_forest,param_grid={'n_estimators':[25,50,75,100]})
model_xgb.fit(X_train,y_train)
xgb_acc=accuracy_score(y_test,model_xgb.predict(test_x_std))
model_acc=pd.DataFrame({'Model name':['XGBoost'],'Accuracy':[xgb_acc]})
models_scores=models_scores.append(model_acc,ignore_index=True)
print('Train          report          of          XGBClassifier
\n',classification_report(y_train,model_xgb.predict(train_x_std)))
print('Test          report          of          XGBClassifier
\n',classification_report(y_test,model_xgb.predict(test_x_std)))
plt.figure(figsize=(10,8))
sns.heatmap(confusion_matrix(y_test,model_xgb.predict(test_x_std)),annot=True)
plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})
plt.xlabel('Predicted value')
plt.ylabel('Actual value')
plt.show()
model_neighbor=KNeighborsClassifier()
grid_neighbor=GridSearchCV(model_neighbor,param_grid={'n_neighbors':range(4,12)})
grid_neighbor.fit(X_train,y_train)
neighbors_acc=accuracy_score(y_test,grid_neighbor.predict(test_x_std))model_acc=pd.Data
Frame({'Model name':['KNeighborsClassifier'],'Accuracy':[neighbors_acc]})
models_scores=models_scores.append(model_acc,ignore_index=True)
grid_neighbor.best_params_
```

```
print('Train          report          of          KneighborsClassifier\n',classification_report(y_train,grid_neighbor.predict(train_x_std)))

print('Test          report          of          KneighborsClassifier\n',classification_report(y_test,grid_neighbor.predict(test_x_std)))

plt.figure(figsize=(10,8))

plt.figure(figsize=(10,8))

sns.heatmap(confusion_matrix(y_test,model_svc.predict(test_x_std)),annot=True)

plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})

plt.xlabel('Predicted value')

plt.ylabel('Actual value')

plt.show()

model_adaboost=AdaBoostClassifier(n_estimators=70)

model_adaboost.fit(X_train,y_train)

adaboost_acc=accuracy_score(y_test,model_adaboost.predict(test_x_std))

model_acc=pd.DataFrame({'Model name':['Adaboost'],'Accuracy':[adaboost_acc]})

models_scores=models_scores.append(model_acc,ignore_index=True)

print('Train          report          of          AdaboostClassifier\n',classification_report(y_train,model_adaboost.predict(train_x_std)))

print('Test          report          of          ADAboostClassifier\n',classification_report(y_test,model_adaboost.predict(test_x_std)))

plt.figure(figsize=(10,8))

sns.heatmap(confusion_matrix(y_test,model_adaboost.predict(test_x_std)),annot=True)

plt.title('Confusion matrix of test data',fontdict={'size':22,'weight':'bold'})

plt.xlabel('Predicted value')

plt.ylabel('Actual value')plt.show()

models_scores.sort_values(by=['Accuracy'],ascending=False,ignore_index=True)
```

## RESULT/SNAPSHOTS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   ph                     2785 non-null   float64
1   Hardness               3276 non-null   float64
2   Solids                 3276 non-null   float64
3   Chloramines            3276 non-null   float64
4   Sulfate                2495 non-null   float64
5   Conductivity           3276 non-null   float64
6   Organic_carbon         3276 non-null   float64
7   Trihalomethanes        3114 non-null   float64
8   Turbidity              3276 non-null   float64
9   Potability             3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

```
the column ph have 14.99 percent of NAN values
the column Hardness have 0.0 percent of NAN values
the column Solids have 0.0 percent of NAN values
the column Chloramines have 0.0 percent of NAN values
the column Sulfate have 23.84 percent of NAN values
the column Conductivity have 0.0 percent of NAN values
the column Organic_carbon have 0.0 percent of NAN values
the column Trihalomethanes have 4.95 percent of NAN values
the column Turbidity have 0.0 percent of NAN values
the column Potability have 0.0 percent of NAN values
```

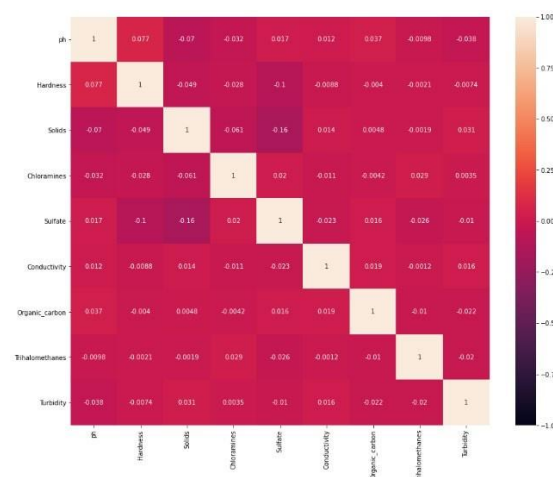
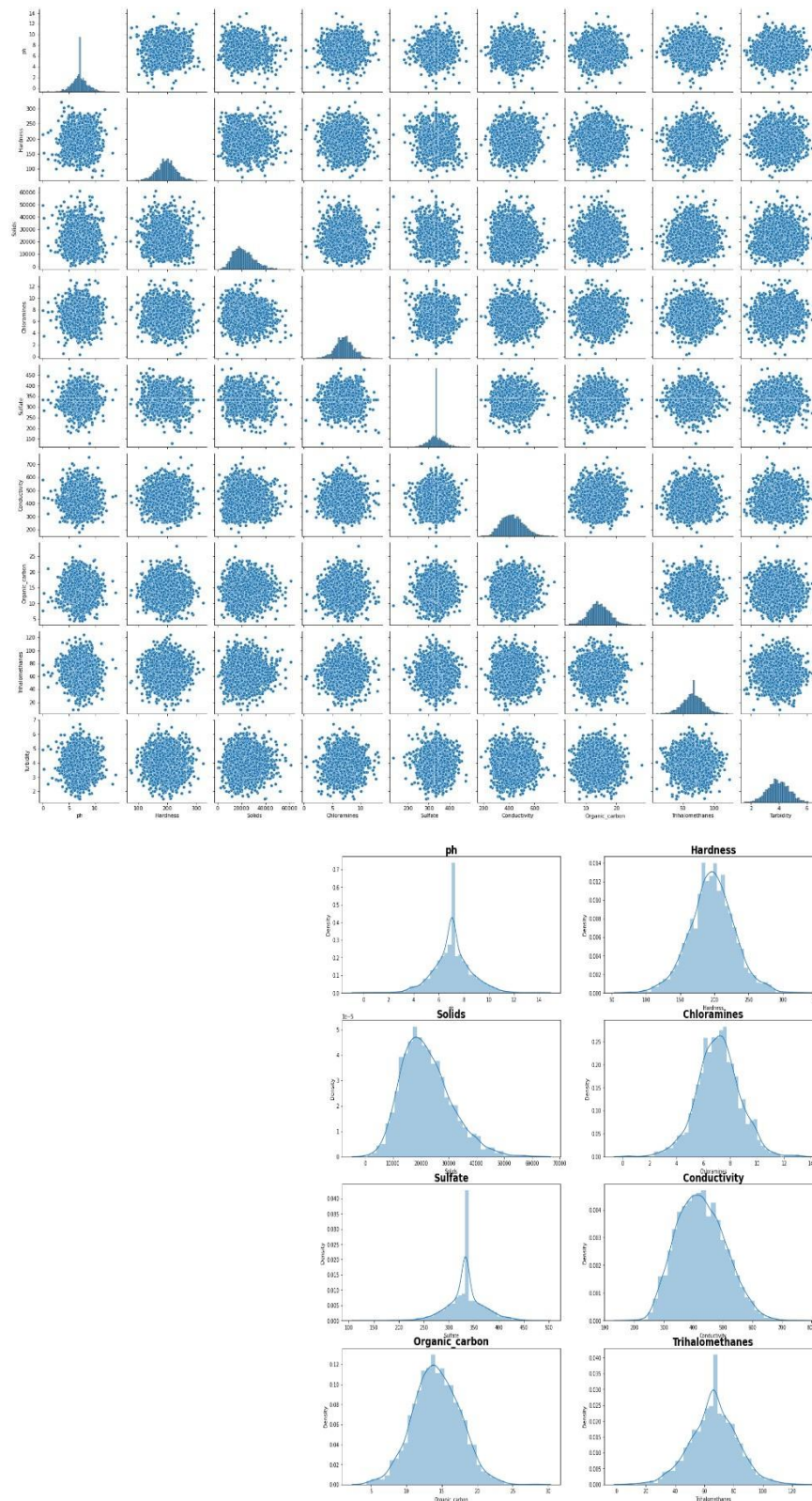


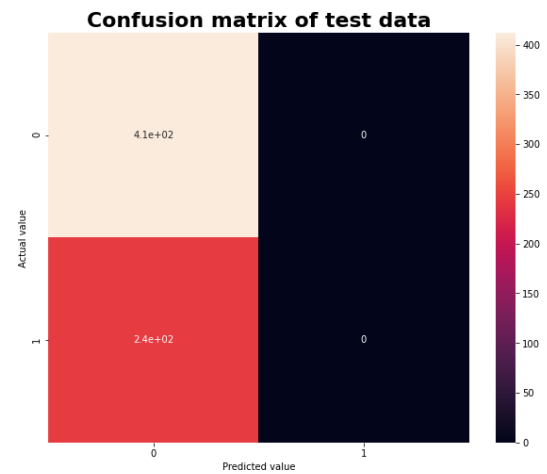
Figure 4.1

The above figure is the training dataset that is used for the training algorithm.

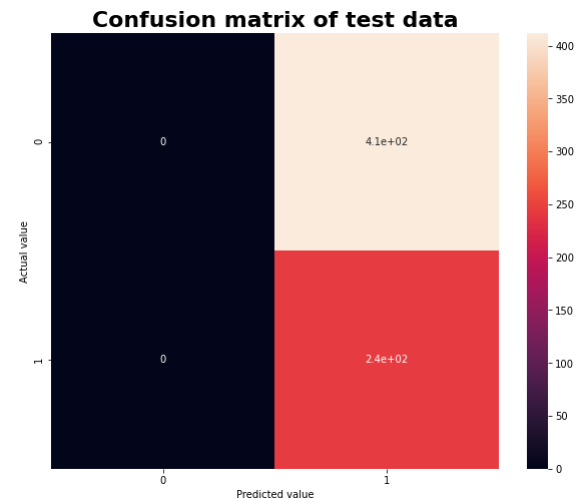


The above figure defines the sample of the water that is sent for the quality check.

Train report of DecisionTreeClassifier					
	precision	recall	f1-score	support	
0	0.61	1.00	0.75	1586	
1	0.00	0.00	0.00	1034	
accuracy			0.61	2620	
macro avg	0.30	0.50	0.38	2620	
weighted avg	0.37	0.61	0.46	2620	
Test report of DecisionTreeClassifier					
	precision	recall	f1-score	support	
0	0.63	1.00	0.77	412	
1	0.00	0.00	0.00	244	
accuracy			0.63	656	
macro avg	0.31	0.50	0.39	656	
weighted avg	0.39	0.63	0.48	656	

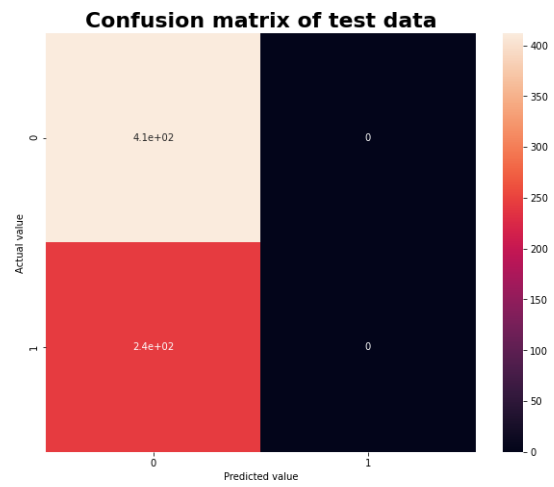


best param {'max_depth': 9}					
best score 0.662137404580153					
Train report of RandomForestClassifier					
	precision	recall	f1-score	support	
0	0.00	0.00	0.00	1586	
1	0.39	1.00	0.57	1034	
accuracy			0.39	2620	
macro avg	0.20	0.50	0.28	2620	
weighted avg	0.16	0.39	0.22	2620	
Test report of RandomForestClassifier					
	precision	recall	f1-score	support	
0	0.00	0.00	0.00	412	
1	0.37	1.00	0.54	244	
accuracy			0.37	656	
macro avg	0.19	0.50	0.27	656	
weighted avg	0.14	0.37	0.20	656	



The above figure shows the result obtained by the confusion matrix algorithm.

Train report of XGBClassifier					
	precision	recall	f1-score	support	
0	0.61	1.00	0.75	1586	
1	0.00	0.00	0.00	1034	
accuracy			0.61	2620	
macro avg	0.30	0.50	0.38	2620	
weighted avg	0.37	0.61	0.46	2620	
Test report of XGBClassifier					
	precision	recall	f1-score	support	
0	0.63	1.00	0.77	412	
1	0.00	0.00	0.00	244	
accuracy			0.63	656	
macro avg	0.31	0.50	0.39	656	
weighted avg	0.39	0.63	0.48	656	

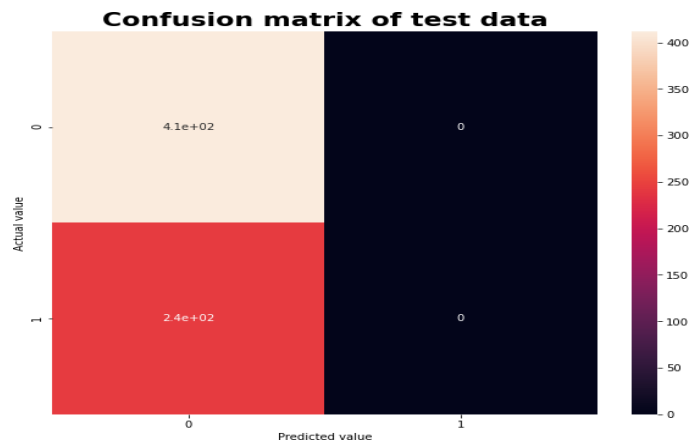




Train report of SVCClassifier					
	precision	recall	f1-score	support	
0	0.61	1.00	0.75	1586	
1	0.00	0.00	0.00	1034	
accuracy			0.61	2620	
macro avg	0.30	0.50	0.38	2620	
weighted avg	0.37	0.61	0.46	2620	

Test report of SVCClassifier					
	precision	recall	f1-score	support	
0	0.63	1.00	0.77	412	
1	0.00	0.00	0.00	244	
accuracy			0.63	656	
macro avg	0.31	0.50	0.39	656	
weighted avg	0.39	0.63	0.48	656	



**Figure 4.3**

The above figure shows the final output of the result.

## **CONCLUSION & FUTURE SCOPE**

### **Conclusion**

Globally, use of water has increased rapidly in recent years due to population growth and capita consumption. The present study finds that the right to clean water is not specifically guaranteed either by the constitution of India or by any other acts. Based on the bar plots plotted we come to the conclusion that some cities are highly polluted and need urgent attention. Also for cities like Bangalore, kashi, Varanasi, Ganga where concentration of pH, sulfate, chlorate etc. is increasing, we can take measures from now to not face problems later. We used linear Regression model for predicting values of pH, hardness, turbidity, organic carbon, trihalomethanes, conductivity, chloramines, solids (TDS). By predicting the values, the water is indicated as safe to drink or not

### **Future scope**

There is much scope to this application as it can also be included in other organizations in predictions of usable water and detection of water. Water quality can be measured by collecting water samples for lab analysis or using probes which can record data at a single point in time or logged at regular intervals over a extended period. Updated version of linear regression, k nearest neighbor, xgb classifier, svm classifier, random forest classifier and decision tree classifier models can make aware of the challenges in futures for further research.



## REFERENCES

- [1] <https://www.discoverCamps.co.in>
- [2] [atacamp.com/community/tutorials/feature-selection-python](https://atacamp.com/community/tutorials/feature-selection-python)
- [3] [www.kaggle.com](https://www.kaggle.com)
- [4] <https://towardsdatascience.com/machine-learning-562dd7df4d42>
- [5] <https://www.geeksforgeeks.org/parameters-feature-selection>
- [6] <https://www.kdnuggets.com/2019/04/text-preprocessing-learning.html> nlp-machine-
- [7] <https://youtube.com>
- [8] <https://google.com>