## CHAPTER 1 INTRODUCTION TO THE PROJECT

* 1. **Introduction**

The project is based on testing the quality of water. The data set consisted of 900 samples that was monitored for 15 years from 2001 to 2015 based on 5 parameters once a month at 5 different sampling points i.e Devprayag, Rishikesh, Haridwar, Jwalapur and Roorkee. The parameters which were analyzed were: Temperature, pH value, Dissolved Oxygen(DO), Biochemical Oxygen Demand(BOD) and Total Coliform .

The objectives of the study were to estimate the Total Coliform value in order to check the acceptability of water. The machine is trained on the dataset and it produces the hypothesis line which predicts value based on which data can be classified into various divisions like household usage, industrial usage, unfit for any use.

### Water quality

Water quality refers to the chemical, physical and biological characteristics of water. It is a measure of the condition of water relative to the requirements of one or more biotic species and or to any human need or purpose. It is most frequently used by reference to a set of standards against which compliance can be assessed. The most common standards used to treat water quality related to the health of ecosystems, safety of human contact and drinking water. Different properties were analyzed & compared during the course of the project.

Some of the properties analyzed are as follows –

* + - Temperature
    - pH Value
    - Dissolved Oxygen(DO)
    - Biochemical Oxygen Demand(BOD)
    - Total Coliform

# LITERATURE REVIEW

## Temperature:

Temperature impacts both the chemical and biological characteristics of surface water. It affects the dissolved oxygen level in the water, photosynthesis of aquatic plants, metabolic rates of aquatic organisms, and the sensitivity of these organisms to pollution, parasites and disease.

Thermal pollution is the introduction of water that is warmer than the body of water into which it flows. It generally occurs near power plants.

Warm water is less capable of holding dissolved oxygen. For this reason, temperature should be measured at the same place within the stream at which dissolved oxygen is measured. This allows the correlation between the two parameters to be observed.

## pH value:

pH is a logarithmic scale used to specify the acidity or basicity of an aqueous solution. Solutions with a pH less than 7 are acidic and solutions with a pH greater than 7 are basic. Pure water is neutral, at pH 7 (25 °C), being neither an acid nor a base. Contrary to popular belief, the pH value can be less than 0 or greater than 14 for very strong acids and bases respectively.

Measurements of pH are important in agronomy, medicine, chemistry, water treatment, and many other applications.

### Importance of pH

* + - The solubility and biological availability of chemical constituents such as nutrients (phosphorus,nitrogen, and carbon) and heavy metals (lead, copper, cadmium, etc.) can be determined by pH of water.
    - pH also determines whether aquatic life can use it. Metals are generally more toxic at lower pH as they are more soluble.
    - Extremely low and high pHs can be significant for the use of water. High pH causes a bitter taste, water pipes and water-using appliances become encrusted with deposits, and it also depresses the effectiveness of the disinfection of chlorine, thereby generating the need for additional chlorine when pH is a bit high. Low-pH water might corrode or dissolve metals and other substances.

## Dissolved Oxygen(DO):

It is the measure of the concentration of oxygen that is dissolved or carried in a given medium as a proportion of the maximal concentration that can be dissolved in that medium. It can be measured with a dissolved oxygen probe such as an oxygen sensor or an optode in liquid media, usually water. The standard unit of oxygen saturation is percent (%).

Oxygen saturation can be measured regionally and noninvasively. Arterial oxygen saturation (SaO2) is commonly measured using pulse oximetry. Tissue saturation at peripheral scale can be measured using NIRS. This technique can be applied on both muscle and brain.

## Biochemical Oxygen Demand(BOD):

Biochemical Oxygen Demand (BOD, also called Biological Oxygen Demand) is the amount of dissolved oxygen needed (i.e. demanded) by aerobic biological organisms to break down organic material present in a given water sample at certain temperature over a specific time period. The BOD value is most commonly expressed in milligrams of oxygen consumed per litre of sample during 5 days of incubation at 20 °C and is often used as a surrogate of the degree of organic pollution of water. BOD can be used as a gauge of the effectiveness of wastewater treatment plants.

### Total Coliform

Coliform is a rod-shaped bacteria which are always present in the digestive tract of warm-blooded animals, including Coliforms are found in human and animal waste, and are also found in water, plants and soil.

Most coliform bacteria don’t cause disease, but some such as a few strains of E. Coliform can cause illness. Positive coliform bacteria tests of food and water indicate the possibility of fecal pollution, and are often associated with the outbreak of disease.

Groundwater pollution caused by fecal contamination is a serious problem, and can be associated with disease.

# 1.3 PROBLEM DEFINITION

Machine learning algorithms are the applications used to build a smarter AI. In the traditional method used a software based machine wouldn't be able to predict the water quality with more than 70% accuracy. With the help of previous year data we trained our machine on 15 years of Ganga river data and with the help of the learning algorithm i.e. Support Vector Classifier we classified the data into usable and non-usable. Now it could predict the water quality for the future data of the Ganga river. This project can be deployed as an external application for the analysis of the water by an AI machine.

# 1.4 Methodology

Iterative process starts with a simple implementation of a subset of the software requirements and iteratively enhances the evolving versions until the full system is implemented. At each iteration, design modifications are made and new functional capabilities are added. The basic idea behind this method is to develop a system through repeated cycles iterative and in smaller portions at a time.

Iterative and Incremental development is a combination of both iterative design or iterative method and incremental build model for development. "During software development, more than one iteration of the software development cycle may be in progress at the same time." This process may be described as an "evolutionary acquisition" or "incremental build" approach."

In this incremental model, the whole requirement is divided into various builds. During each iteration, the development module goes through the requirements, design, implementation and testing phases. Each subsequent release of the module adds function to the previous release. The process continues till the complete system is ready as per the requirement.

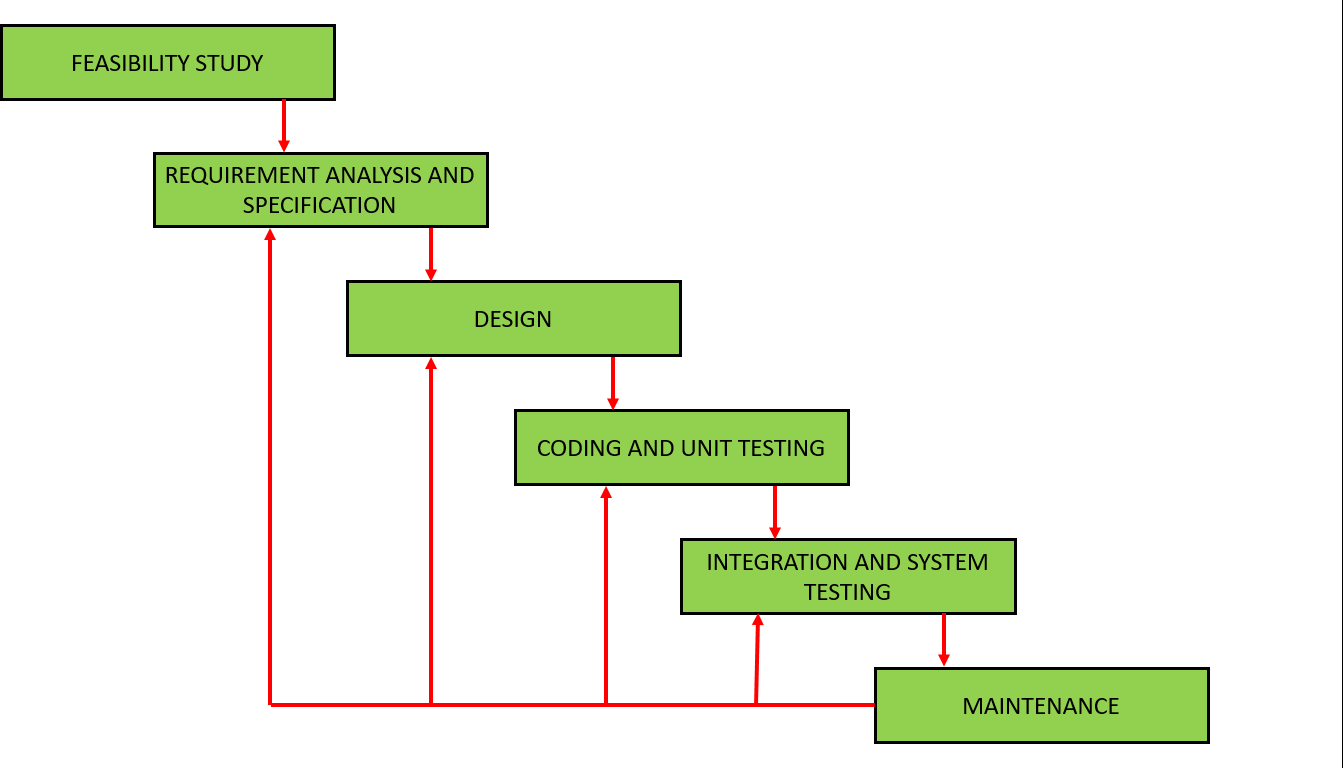


Fig1.1: SDLC Model

# FEASIBILITY STUDY

Feasibility study is made to see if the project on completion will serve the purpose of the organization for the amount of work, effort and the time that is spent on it. Thus when a new application is proposed it normally goes through a feasibility study before it is approved for development. The feasibility study of our project is as follows:

## Technical Study:

The technical study of the website determines whether the software is technically feasible or not.

Table 1.1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Available | Used | Remarks |
| Technology | R, Python, SVM | Support Vector Classifier , Python | Technically |
| Feasible |
| Tools | Anaconda, Jupyter Notebook , Spider | Spider | Technically |
| Feasible |

## Financial Study :

The financial study of the software determines whether it is financially feasible or not.

Table 1.2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Used | Funds | Remarks |
| Required |
| Software Tools | Python, Anaconda, | Cost=0 | Financially |
| (Free and open | Feasible |
| source) |  |

## Time Constraint :

time constraint determines the required time needed for completing the project:

Table 1.3

|  |  |  |
| --- | --- | --- |
| Required Time | Available Time | Remarks |
| 4 month | 4 months | Possible |

# CHAPTER 2

**PROJECT REQUIREMENTS ANALYSIS AND SPECIFICATION**

## INTRODUCTION:

Project Requirement Analysis and Specification is the starting point of the development. Before designing any model, it is very important to define precise requirements of the water quality analyzer which specify the software function and performance, indicate software interface with other system elements, and set up constraints that software meet and Requirement Specification is produced at the end of the analysis work.

* + - The aim of the requirements analysis and specification phase is to understand the exact requirements of the customer and to document them properly. This phase consists of two distinct activities, namely
* Requirements gathering and analysis
* Requirements specification
  + - The goal of the **requirements gathering** activity is to collect all relevant information from the customer regarding the product to be developed. This is done to clearly understand the customer requirements so that incompleteness and inconsistencies are removed.
    - The **requirements analysis** activity is begun by collecting all relevant data regarding the product to be developed from the users of the product and from the customer through interviews and discussions.
    - The data collected from such a group of users usually contain several contradictions and ambiguities, since each user typically has only a partial and incomplete view of the system. Therefore all ambiguities, inconsistencies, and incompleteness are resolved and all the requirements properly understood, the requirements specification activity can start. During this activity, the user requirements are systematically organized into a Software Requirements Specification (SRS) document.
    - The customer requirements identified during the requirements gathering and analysis activity are organized into a SRS document.

# FUNDAMENTAL REQUIREMENTS:

## Functional System Requirements:

* + - A Water Quality Analyzer will be a helpful tool in determining the quality of water.
    - It will predict the quality of water based on the test data provided.

The sample data contains 900 samples and each sample’s 5 perimeters have been noted that are Temperature, pH, Dissolved Oxygen, Biochemical Oxygen Demand

* + - Will help in categorizing weather the water source can be used for household work, bathing, industry or waste .

Training data has been provided to the machine and it has been trained by a dataset of 10 years. Any new data item that is added to the model will predict the accuracy of the model and the water quality is predicted, hence classification of that source is done into usable or unusable.

**Outcome of Project:** Water source whose dataset is provided, the model will predict whether the water source can be used or not and for which purpose. This will really be helpful in pre-determining the future of any water source so that actions can be taken in advance.

**Response:** Response is according to the training data as well as the new dataset.

# REQUIREMENTS ELICITATION

Requirements elicitation is the practice of collecting the requirements of a system from users, customers and other stakeholders. The practice is also sometimes referred to as "requirement gathering".

### Data Requirements.

The Water Quality Analyzer requires a data set on which we will train the machine. In our project we used the dataset of Ganga river, five of its perimeters - Temperature ,pH, Dissolved Oxygen(DO), Biochemical Oxygen Demand (BOD) and Total Coliform were collected upto 14 years that are observed and noted from 5 different locations at different times of the year.

These observation data is used to establish the model of water quality indicator i.e analysing the quality of water i.e usable for bathing or for washing or industry or waste .

This Dataset is Taken from Ganga Tech Action Plan for the last 15 years.

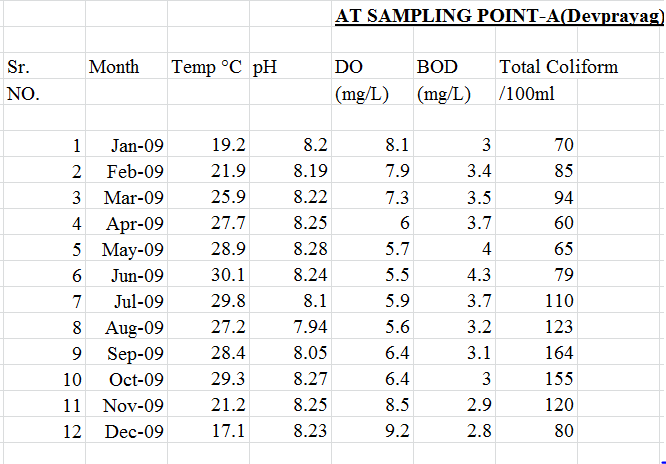


Fig2.1:Sample Dataset

# CHAPTER 3

# PROJECT DESIGN

## INTRODUCTION:

The goal of this design is to transform the requirement specification into a structure that is suitable for implementation in some programming language. In technical terms, during the design phase the software architecture is derived from the SRS document. Two distinctly different approaches are available: the traditional design approach and the object-oriented design approach.

### Traditional design approach

Traditional design consists of two different activities; first a structured analysis of the requirements specification is carried out where the detailed structure of the problem is examined. This is followed by a structured design activity. During structured design, the results of structured analysis are transformed into the software design.

### Object-oriented design approach

In this technique, various objects that occur in the problem domain and the solution domain are first identified, and the different relationships that exist among these objects are identified. The object structure is further refined to obtain the detailed design.

## Preliminary design:

It is the first phase of the design process .It is also known as conceptual design or architectural design .During this phase the high level design concept is created ,which will implement the complex electronic requirement .The objective of the design phase is to create a design that will correctively and completely implement the requirement. The main design activities for the preliminary design phase are :

* + - Create the high level design description
    - Identify major component.
    - Any emission or errors are resolved
    - Include reliability, maintenance, and test features that are necessary to meet performance and quality requirements.
    - Identify constraints on system elements that are result of high level design

## Project Design:

### Use Case Diagram:

### This Diagram shows the flow of data from client to server and vice versa

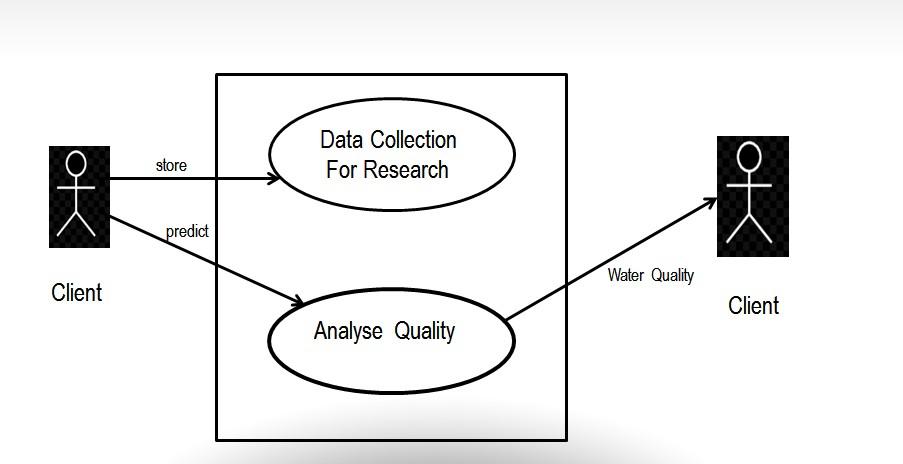


Fig3.1:Used Case

**Data Flow Diagram(DFD) :**

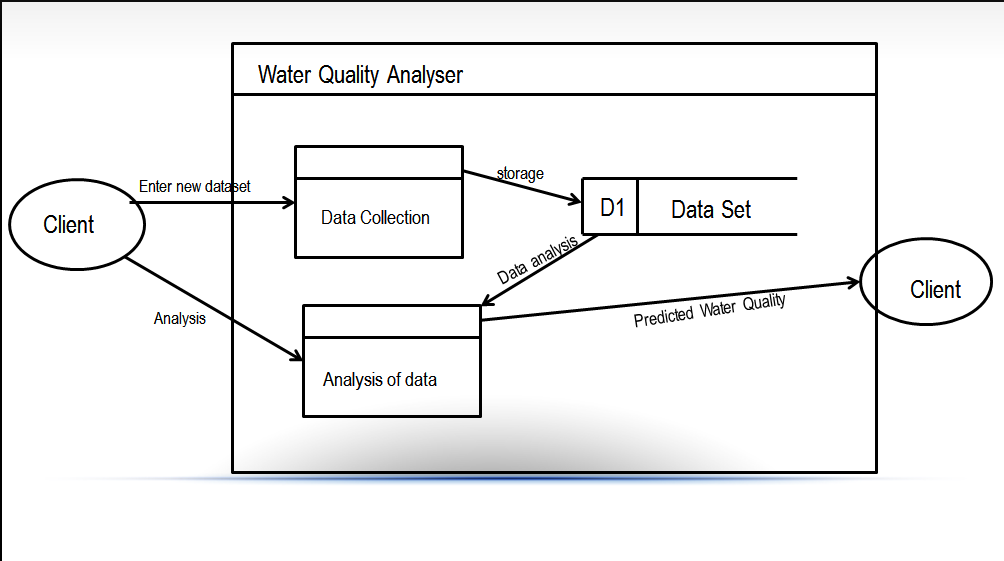
This Data Flow Diagram Shows the 2nd Phase of Analysis in which Database D1 and flow of data is elaborated. 

Fig3.2: Data Flow Diagram

**CHAPTER 4 PROJECT IMPLEMENTATION**

* 1. **MANAGING IMPLEMENTATIONS**

**Support Vector Machine**

It is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes.

### Python:

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that can be used for both research and development and developing production systems.

There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming**.**

### Anaconda (Python distribution)

Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system conda.

### 

### Python Data Analysis Library : Pandas

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

It is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

### NumPy

NumPy is a library for the Python programming language, adding support for large, multi- dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is open-source software .It contains among other things:

* a powerful N-dimensional array object
* sophisticated (broadcasting) functions.

# 4.2 PRACTICAL CONSIDERATIONS

## 

## Screenshots:

### DataSet:

### This data set is taken directly from excel sheet for last 15 years and calculated WQ

### Table4.1

### 

### 

### 

**X Train:**

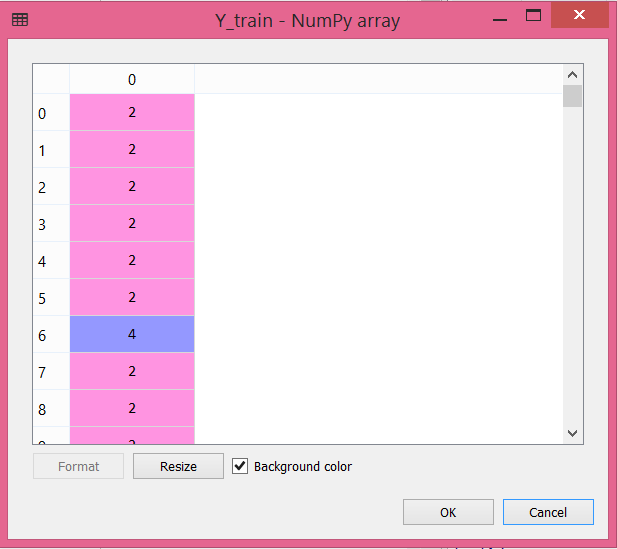
This Data Set is the Data through which we will predict the accuracy.

**Table4.2**

**YTrain:**

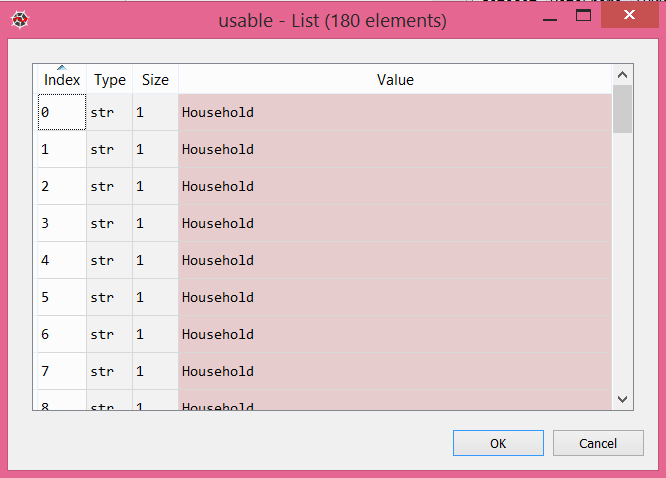
This Data Set is the Data set in which we have predicted the missing values and WQ .

**Table 4.3**



**Accuracy:**



**Classified Result: Table 4.4**

**CHAPTER 5**

**PROJECT TESTING**

* 1. **INTRODUCTION**
* Testing is an activity performed for evaluating product quality, and for improving it, by identifying defects and problems.
* It is a process used to help identify the correctness, completeness and quality of developed computer software.
* It is not a distinct phase in system development but should be applicable throughout all phases i.e. design, development and maintenance phase.

**Some important principles of software testing are as given:-**

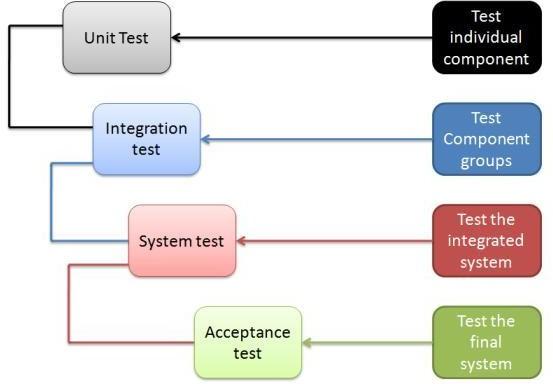
* Testing time and resources should be limited i.e. avoid redundant testing.
* It is impossible to test everything.
* Use effective resources to test.
* Tests should be planned long before testing begins i.e. after the requirement phase.
* Test for invalid and unexpected input conditions as well as valid conditions.
* For the most effective testing should be conducted by an independent party.
* Document test cases and test results.
* Examining what the software not doing which it expected to do and also checking

# TESTING LEVELS

There are different levels during the process of testing. Levels of testing include different methodologies that can be used while conducting software testing. The main levels of software testing are:

* Unit testing.
* Integration testing.
* System testing.
* Acceptance testing

## The testing sequence

Fig5.1: Testing

There are five steps that are involved while testing an application for functionality:

### Steps Description

1. The determination of the functionality that the intended application is meant to perform.
2. The creation of test data based on the specifications of the application.
3. The output based on the test data and the specifications of the application. IV The writing of test scenarios and the execution of test cases.

V The comparison of actual and expected results based on the executed test cases.

## Unit Testing:

* Unit testing focuses verification effort on the smallest unit software design- the module.
* It is basically done by the developers to make sure that their code is working fine and meet the user specifications. They test their piece of code which they have written like classes, functions, interfaces and procedures.
* Every module is tested fully to check its syntax and logical correctness. Error handling was implemented into relevant modules so that the code doesn’t crash on errors.

## Integration Testing:

* Integration testing is done when two modules are integrated, in order to test the behavior and functionality of both the modules after integration.
* In integration testing the two programs are integrated and then testing is done by running the programs. This works properly and therefore integration testing is done.
* Below are few types of integration testing: Top down

Bottom up

Functional incremental

## System Testing:

* In system testing the testers basically test the compatibility of the application with the system.
* System testing for this project is done by running the program and then checking whether it works properly or not. To run the program command prompt is used and then the program is run by the user.

## Acceptance Testing:

* Acceptance testing is basically done to ensure that the requirements of the specification are met.
* Here all the requirements of specification are met because all functions are working properly according to the need.

# TESTING TECHNIQUES

There are different methods that can be used for software testing.

## Black-Box Testing:

The technique of testing without having any knowledge of the interior workings of the application is called black-box testing. The tester is oblivious to the system architecture and does not have access to the source code. Typically, while performing a black-box test, a tester will interact with the system's user interface by providing inputs and examining outputs without knowing how and where the inputs are worked upon.

This testing is done by simply running the source code program and observing that it shows the output. Means that no error in running the source code.

## White-Box Testing:

White-box testing is the detailed investigation of internal logic and structure of the code. White-box testing is also called glass testing or open-box testing. In order to perform white-box testing on an application, a tester needs to know the internal workings of the code.

The tester needs to have a look inside the source code and find out which unit/chunk of the code is behaving inappropriately.

This testing is done by running the source code program and then checking every function of the project. If some function does not work properly then the source code is checked for correct this.

## Technique Used: White Box Testing

To perform testing we used White box testing techniques. We investigated internal logic and structure of the code. When the accuracy of the model was not very good we look for internal problems i.e. data not properly cleaned, ambiguity, inconsistency etc.

# CHAPTER 6 CONCLUSION

For Analyzing the water quality of the actual Ganga river data collected from 5 points for 2001-2015, we classified the data into the categorical values on the basis of the biochemical parameters (Dissolved Oxygen, PH, Temperature, Total Coliform, Biochemical Oxygen Demand) on which we applied supervised machine learning algorithms for analyzing the accuracy of the model . And we found that the maximum accuracy on the dataset we formed has attained the maximum accuracy on using the Support Vector Classifier algorithm. With the help of the values predicted by our data we were able to classify the data into usable and waste water.

## REFERENCES

The following books were very helpful during the completion of project:

* Introduction to Machine Learning with Python: A Guide for Data Scientists

By: Müller-Cyran and Sarah Guido

* Python Machine Learning

By: Sebastian Raschka

Following websites that are very useful for the completion of the project

* Machine learning and Data Visualization using Python course – Pierian Data

# PERSONAL DETAILS

**Name:** Nihal Asati 

**Enr No.:** 151330

**Branch:** CSE

**Email Id:** [asatinihal2009@gmail.com](mailto:asatinihal2009@gmail.com)

**Phone No**.: 8462028346