

Drinking Water Quality Detection
Using Machine Learning Techniques

Major Project Report Submitted in partial fulfilment

of the requirement for undergraduate degree of

**BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING**
By

K. B. Indraneel (221710301030)

D. Akhil Kumar (221710301012)

O. Sai Ram Reddy (221710301046)

A. Kartheek (221710301004)

Under the Guidance of

Ms. G. Mounika

Assistant Professor



Department Of Computer Science and Engineering

GITAM School of Technology

GITAM (Deemed to be University)

Hyderabad-502329

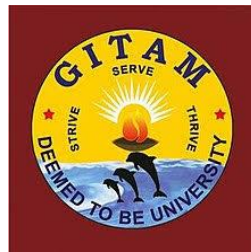
July 2020

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF TECHNOLOGY**

GITAM

(Deemed-to-be-University u/s 3 of UGC Act 1956)

HYDERABAD CAMPUS



DECLARATION

We submit this major project work entitled **“Drinking water quality detection using machine learning techniques”** to GITAM (Deemed to be University), Hyderabad, in partial fulfilment of the requirements for the award of the degree of **“Bachelor of Technology”** in **“Computer Science and Engineering”**. We declare that it was carried out independently by us under the guidance of **(Ms. G. Mounika)**, Asst. Professor, GITAM (Deemed to be University), Hyderabad, India.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Place: Hyderabad

Date:

Name and Signature of Candidate

K. B. Indraneel (221710301030)

D. Akhil Kumar (221710301012)

O. Sai Ram Reddy (221710301046)

A. Kartheek (221710301004)

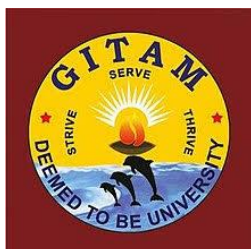
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM SCHOOL OF TECHNOLOGY

GITAM

(Deemed-to-be-University u/s 3 of UGC Act 1956)

HYDERABAD CAMPUS



CERTIFICATE

This is to certify that the Major Project Report entitled - “**Drinking water quality detection using machine learning techniques**” is being submitted by **K. B. Indraneel (221710301030)**, **D. Akhil Kumar (221710301012)**, **O. Sai Ram Reddy (221710301046)**, **A. Kartheek (221710301004)** submitted in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering**.

Guided by

Ms. G. Mounika

Head of the Department

Dr. Phani Kumar

Professor & HOD

ACKNOWLEDGEMENT

Our project would not have been successful without the help of several people. We would like to thank the personalities who were part of our project in numerous ways, those who gave us outstanding support from the birth of the project.

We would like to thank our honorable Pro-Vice-Chancellor, **Prof. N. Siva Prasad**, for providing the necessary infrastructure and resources for the accomplishment of our project.

We would like to thank respected **Prof. N. Seetharamaiah**, Principal, School of Technology, for his support during the tenure of the project.

We would like to thank respected **Prof. S. Phani Kumar**, Head of the Department of Computer Science & Engineering, for providing the opportunity to undertake this project and encouragement in the completion of this project.

We would like to thank respected **Ms. G. Mounika**, Assistant Professor in the Computer Science Department, for the esteemed guidance, moral support and invaluable advice provided by her for the success of the project.

We would like to thank our parents and friends who extended their help, encouragement and moral support directly or indirectly in our project work.

Sincerely,

K. B. Indraneel (221710301030)

D. Akhil Kumar (221710301012)

O. Sai Ram Reddy (221710301046)

A. Kartheek (221710301004)

ABSTRACT

The quality of water plays an essential role in a healthy ecosystem. Drinking water is one of the most critical and essential need of human beings. Its quality changes from one spot to another, contingent upon the state of the water source from which it is taken and the treatment it gets. Drinking water that is not adequately treated or that travels through an improperly maintained distribution system may increase the risk of contamination, leading to several health issues and causes diseases like cholera, diarrhea and typhoid. In order to test the quality of the water sample, it usually takes 24-48 hours in a lab which is a time-consuming process and involves many resources. Hence the project's aim is to detect the quality of the drinking water using various machine learning techniques.

The dataset used is taken from “Kaggle” website. The attributes of the dataset such as temperature, ph value, conductivity, nitrate content, fecal coliform etc are considered in order to detect the quality of drinking water. The algorithms used in this process are random forest and support vector machine. The dataset is tested against the two algorithms and the one which gives the highest accuracy is considered.

TABLE OF CONTENTS

1. Machine learning.....	1-9
1.1 Introduction.....	1
1.2 Importance of machine learning.....	2
1.3 Applications of machine learning.....	2
1.4 Classification of learning models.....	3
1.4.1 Supervised learning.....	4
1.4.2 Unsupervised learning.....	5
1.4.3 Reinforced learning.....	5
1.5 Types of supervised learning.....	6
1.5.1 Regression.....	6
1.5.2 Classification.....	7
2. Python.....	10-12
2.1 Introduction.....	10
2.2 History of python.....	10
2.3 Python variables.....	10
2.4 Strings in python.....	11
2.5 Python lists.....	11
2.6 Python tuples.....	11
2.7 Python sets.....	12
2.8 Python dictionaries.....	12
3. Literature survey.....	13
4. Case study.....	14-15
4.1 Problem statement.....	14
4.2 Data set.....	14
4.3 Objective of case study.....	15
5. Design.....	16-19
5.1 Architecture.....	16
5.2 Process flow.....	16
5.3 Data flow.....	17
5.4 Flow chart.....	17
5.5 UML diagram.....	18
5.6 Sequence diagram.....	18

5.7 Activity diagram.....	19
6. Implementation.....	20-26
6.1 Importing libraries.....	20
6.2 User interface.....	20
6.3 Data splitting.....	25
6.4 Support vector machine.....	25
6.5 Random forest.....	26
7. Result analysis.....	27-28
7.1 Random forest.....	27
7.2 Support vector machine.....	28
8. Conclusion and future scope.....	29
References.....	30

LIST OF FIGURES

Fig 1.1 Introduction to machine learning	1
Fig 1.2 The process flow.....	2
Fig 1.3 Classification of learning models.....	4
Fig 1.4 Supervised learning.....	4
Fig 1.5 Unsupervised learning.....	5
Fig 1.6 Reinforced learning.....	6
Fig 1.7 Regression model.....	7
Fig 1.8 Classification model.....	8
Fig 1.9 Decision Tree for Playing Tennis.....	8
Fig 1.10 Random Forest.....	9
Fig 4.1 CSV file of water.....	14
Fig 5.1 Architecture diagram.....	16
Fig 5.2 Process flow diagram.....	16
Fig 5.3 Data flow diagram.....	17
Fig 5.4 Flow chart	17
Fig 5.5 UML diagram.....	18
Fig 5.6 Sequence diagram.....	18
Fig 5.7 Activity diagram.....	19
Fig 6.1 Importing libraries.....	20
Fig 6.2 Implementing user interface.....	20
Fig 6.3 Water extraction points.....	21
Fig 6.4 Creation of rfacts table.....	21
Fig 6.5 Rfacts table.....	22
Fig 6.6 Rfacts data entry.....	22
Fig 6.7 Updated rfacts table.....	23
Fig 6.8 Quality factors.....	23
Fig 6.9 Creation of qfacts table.....	23
Fig 6.10 Qfacts table.....	24
Fig 6.11 Qfacts data entry.....	24
Fig 6.12 Updated qfacts table.....	25
Fig 6.13 Support vector machine.....	25
Fig 6.14 Random forest.....	26

Fig 7.1 Random forest UI.....	27
Fig 7.2 Accuracy of random forest.....	27
Fig 7.3 Support vector machine UI.....	28
Fig 7.4 Accuracy of support vector machine.....	28

CHAPTER-1

MACHINE LEARNING

1.1 INTRODUCTION

Machine learning (ML) is that the study of algorithms that improves through its experience. It's defined as the science of getting computers act like humans do and improve their learning over time in self determining fashion by supplying them data. It is seen as a subset of AI.

These algorithms develop a model that validate sample data referred to as training data set to perform predictions without being expressly customized to attempt and do so. It involves computers learning from data provided so they perform certain tasks. A subset of ML is connected with computational insights, which make predictions using computers.

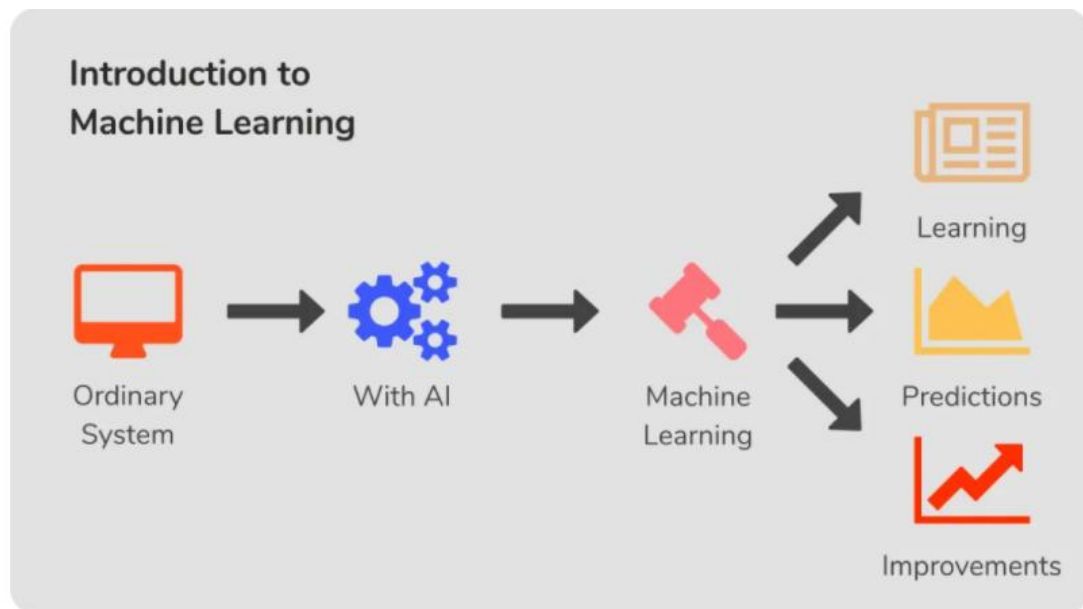


Fig 1.1 Introduction to Machine Learning

1.2 IMPORTANCE OF MACHINE LEARNING

In future, automation process will replace most of the human-work. To meet human capabilities, computers must be intelligent and this can be accomplished using machine learning. Now-a-days with the help of machine learning companies are replacing human work such as responding to customer calls, maintaining the financial transactions, etc Machine

learning can handle major problems like image detection in self-driving cars, prediction of natural disaster locations and helps to understand the interaction of medication with medical conditions before clinical trials. Hence machine learning plays an important role.

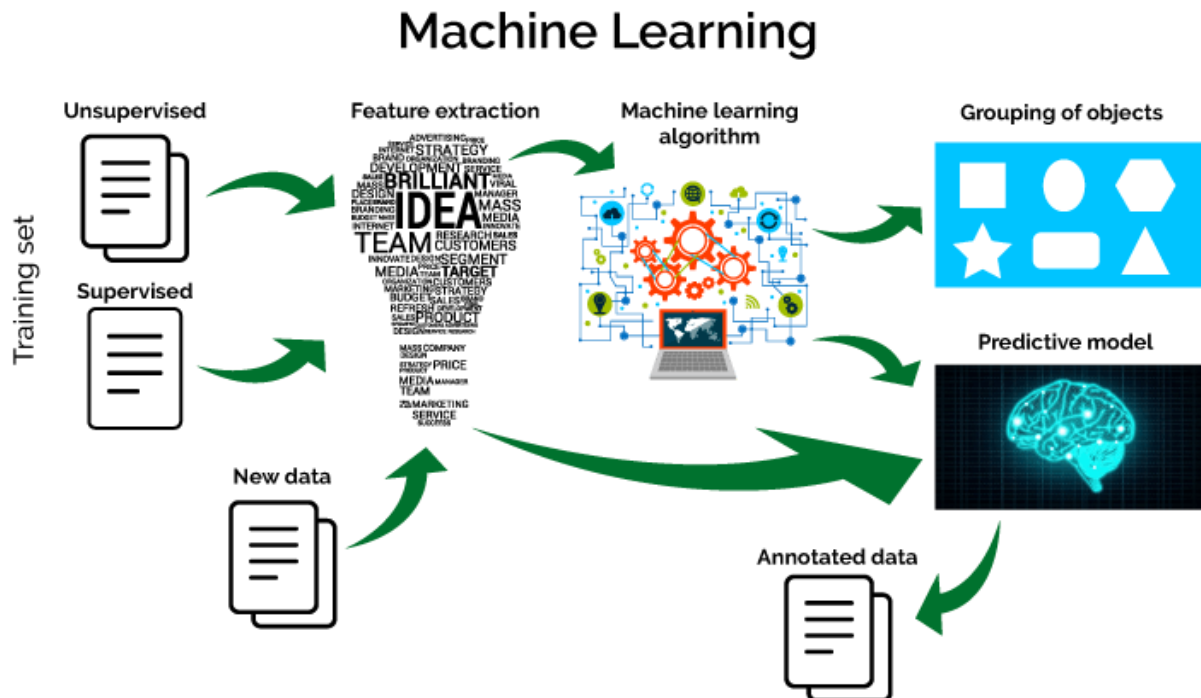


Fig 1.2 The Process Flow

1.3 APPLICATIONS OF MACHINE LEARNING

1. Healthcare: The sensors that screen everything like pulse rates, number of steps walked, sleeping patterns, oxygen and sugar levels have generated a enormous amounts of knowledge. It allows doctors to treat the patient's health in real-time. ML algorithms can detect cancer tumors, identifies carcinoma and also can analyze images related to eye diseases.

2. Government: Machine learning helps governing body to predict future scenarios and adjust to quickly changing situations. ML can help to boost cyber security and cyber intelligence, operational preparedness, helps in counterterrorism efforts, predictive maintenance and reduce failure rates.

3. Marketing and sales: With the help of Machine learning marketing sector is being revolutionized as many companies are using computer science and machine learning to improve customer satisfaction. As per Forbes, fifty percent of enterprise executives believe that AI and ML play an important role in improving customer support and experiences.

4. E-commerce: Social media uses machine learning to research customer's search history and make suggestions on similar according to the past habits. Many experts believe that the long term of retail is going to be dominated by AI and ML as deep learning allows business applications to become even better at capturing and analyzing data to customize one's shopping experience and develop personalized marketing campaigns.

5. Transportation: Accuracy and efficiency play an important role in earning profits within this sector, so it gives the ability to predict the potential problems. Machine Learning's analysis and modeling functions helps the businesses to link public transportation and cargo transport sectors perfectly. Machine learning acts as a crucial component within supply chain management since it uses algorithms to search the parameters that shows an impact on supply chain's success.

6. Financial services: The information provided by ML in the industry made investors to spot new opportunities and trade. Data processing highlights high-risk clients and helps cyber surveillance to investigate the signs of fraud. ML helps to regulate financial portfolios, assess risk for loans and insurance underwriting.

7. Oil and gas: With the use of ML and AI new energy sources and mineral deposits have been found within the ground. It helps in predicting the refinery sensor failure, and streamline oil distribution to improve efficiency and reduce shrink costs. ML is changing the industry with its case-based reasoning and reservoir modeling. Machine learning helps to make this dangerous industry more secure.

8. Manufacturing: It is not strange to know that machine learning is being used in this industry. Machine learning applications in this sector are about to achieve the goal of improving operations from gathering raw materials to final delivery, significantly reducing error rates, improve in predictive maintenance and increase in inventory turn.

1.4 CLASSIFICATION OF LEARNING MODELS

These algorithms can be divided into three groups based on the essence of learning.

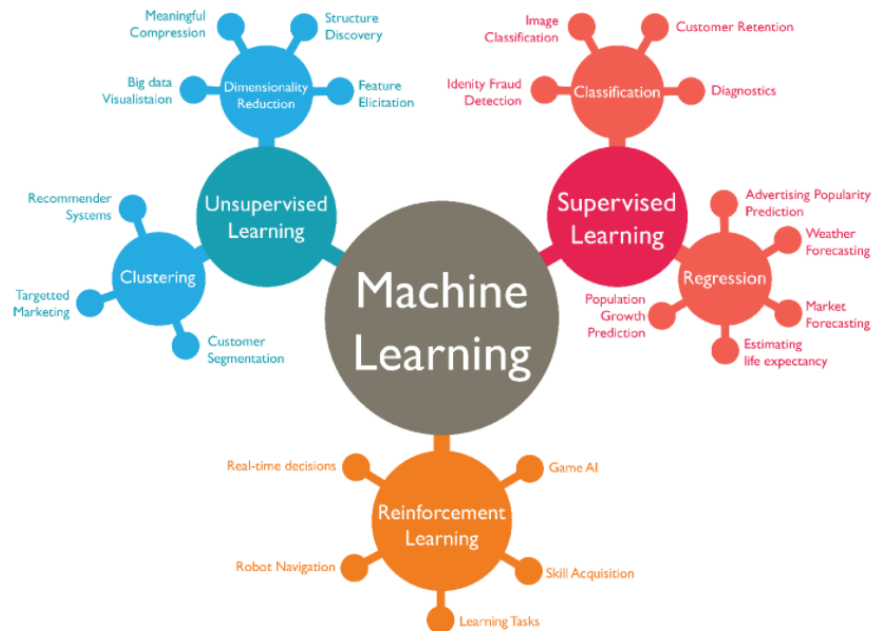


Fig 1.3 Classification of Learning Models

1.4.1 Supervised learning: It is defined as the task of constructing an input-output pair by learning a function that maps an input to a corresponding output. It employs a feature derived from labeled training data, which consists of a set of tested examples. Supervised learning occurs if an algorithm learns from sample data and related responses of target data, which may include labels or numerical values, such as classes, in order to predict the correct response when presented with new examples.

This method is like a student learning under the supervision of a lecturer. It is like tutor providing suitable examples for the scholar to understand and remember and As a result, the student is able to derive abstract rules from the examples.

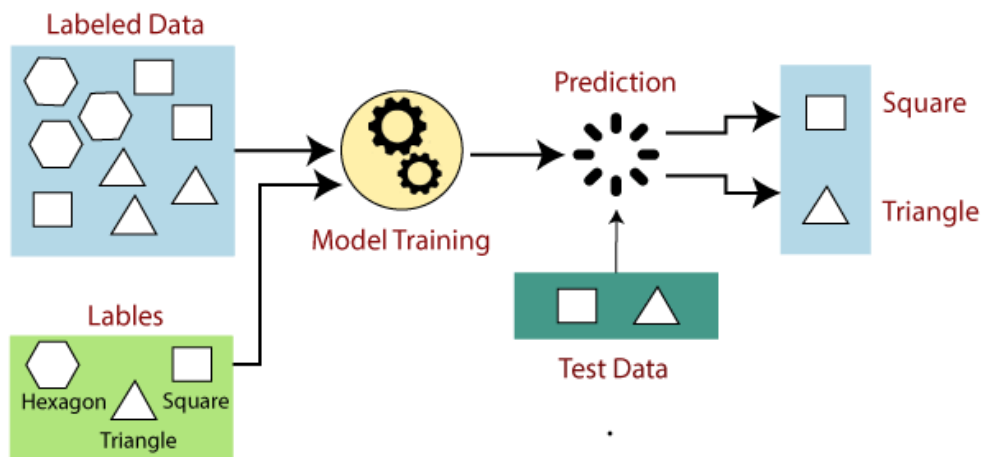


Fig 1.4 Supervised Learning

1.4.2 Unsupervised learning: It is a form of algorithm that is used to make inferences from datasets that contain computer files with no labeled responses. Cluster analysis is the most popular unsupervised learning technique, which is used for data exploration to uncover hidden correlations or data grouping. Unsupervised learning occurs when an algorithm learns from simple examples with no related answer, allowing the algorithm to see knowledge patterns on its own.

This form of algorithm restructures data as well as new features that are representative of a category or a replacement set of unrelated values. They're great for giving humans new inputs for supervised machine learning algorithms, as well as insights into the sense of intelligence.

As a means of learning, it is similar to the methods humans use to decide if certain objects or events belong to the same class, such as evaluating the degree of similarity between objects. This type of learning is assisted by several recommendation systems that can be found on the internet as part of marketing automation.

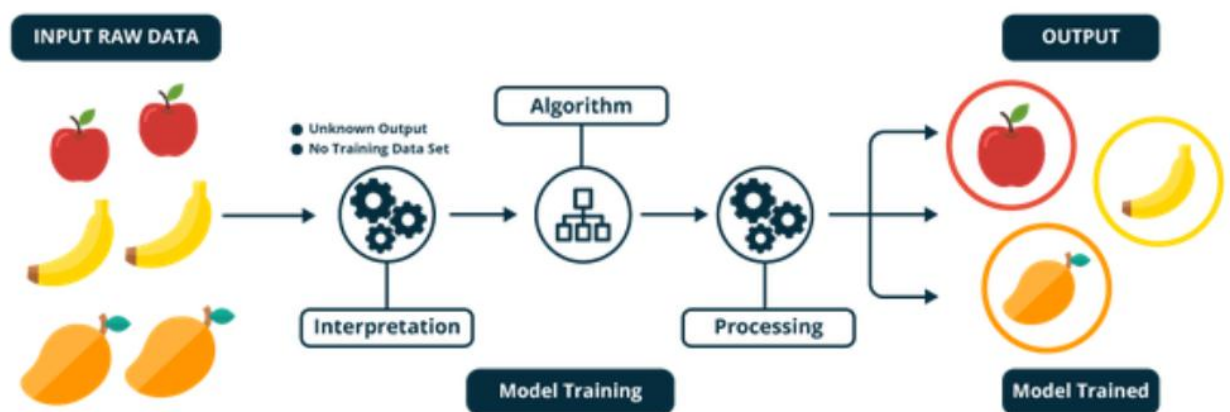


Fig 1.5 Unsupervised Learning

1.4.3 Reinforcement learning: The algorithm is given examples without labels. However, an example that is followed by positive or valid feedback that is compatible with the algorithm's proposed answer falls into this category, which is relevant to applications in which the algorithm must make decisions with consequences in the human environment. (As a consequence, unlike unsupervised instruction, the product is prescriptive instead of descriptive)

Errors aids in learning because they require a penalty (cost, loss of time, regret, discomfort, and so on), demonstrating that one plan of action is less able to succeed over others. This is an exciting example of reinforcement learning as computers learn to play video games on their own.

In this case, an application provides the algorithm with examples of specific situations, such as trapping the gamer in a maze while avoiding an enemy. The applying informs the algorithm of the outcome of its decisions, and learning occurs as it attempts to escape and survive what it learns to be dangerous.

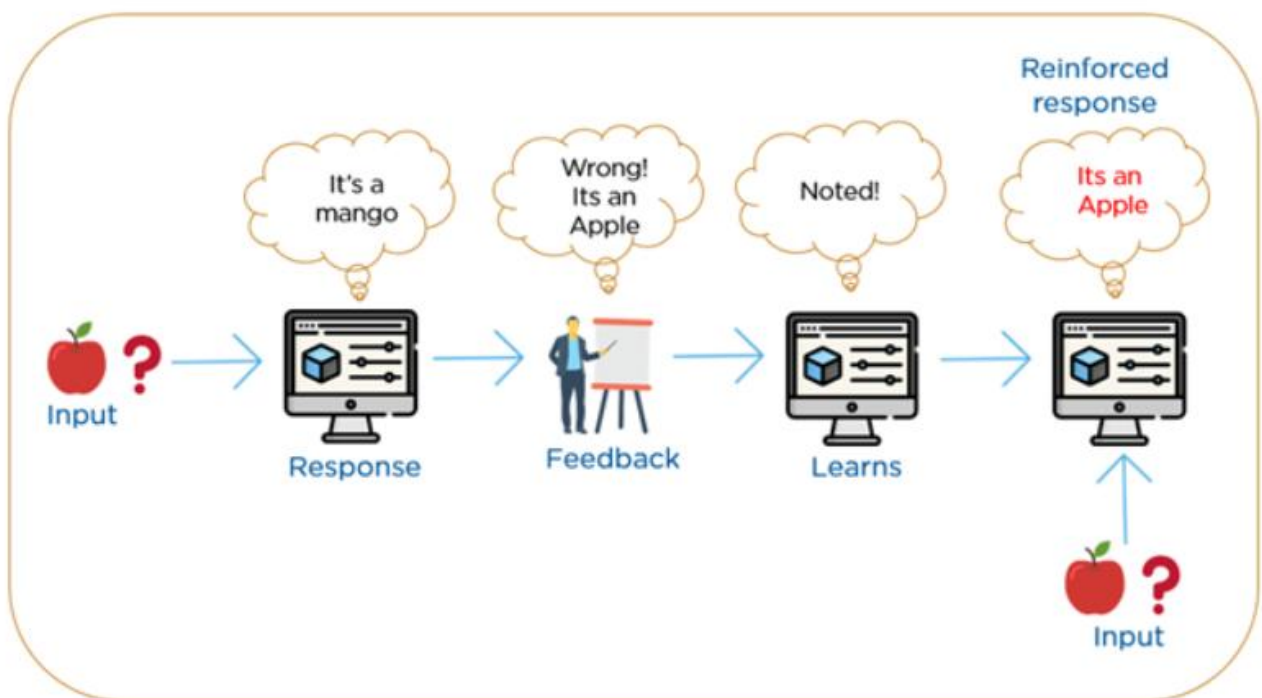


Fig 1.6 Reinforced Learning

1.5 TYPES OF SUPERVISED LEARNING

It can be further divided into two categories:

1.5.1 Regression

Regression algorithms predict never-ending values which supports the input data. The regression problems mainly estimate a mapping function supporting the input and output data. The regression model is used when the target variable is a quantity like scores, income, weight or height, or the probability of a binary category like the chances of having rain some regions.

The various types of this algorithm include:

A. Simple linear regression: With this method, a connection can be estimated between two variables employing a line when both the variables are quantitative.

B. Multiple linear regression: It's a step beyond linear regression. It can estimate the values of a variable which supports two or more independent variables.

C. Polynomial regression: It's main aim is to find a nonlinear relationship between dependent and independent variables.

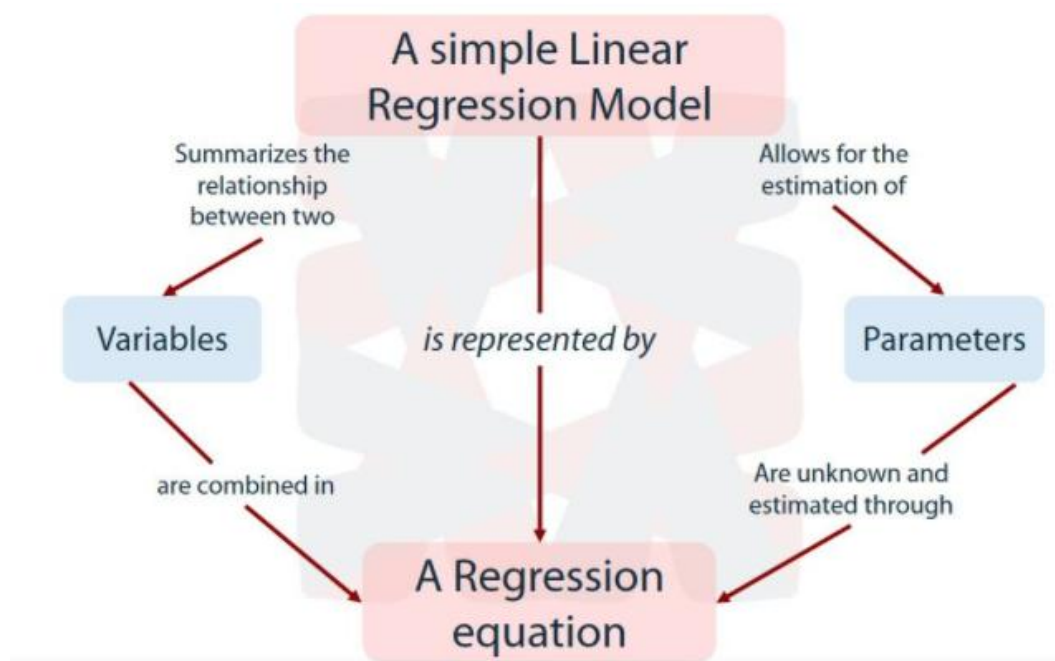


Fig 1.7 Regression Model

1.5.2 CLASSIFICATION

It is a predictive model that estimates a linking function that finds distinct output variables from input variables which might be categories or labels. The prediction of category or label of input variables is done by the mapping function of classification algorithm. Both discrete and real-valued variables can be used in a classification algorithm, but the examples must be categorized into one of two or more groups.

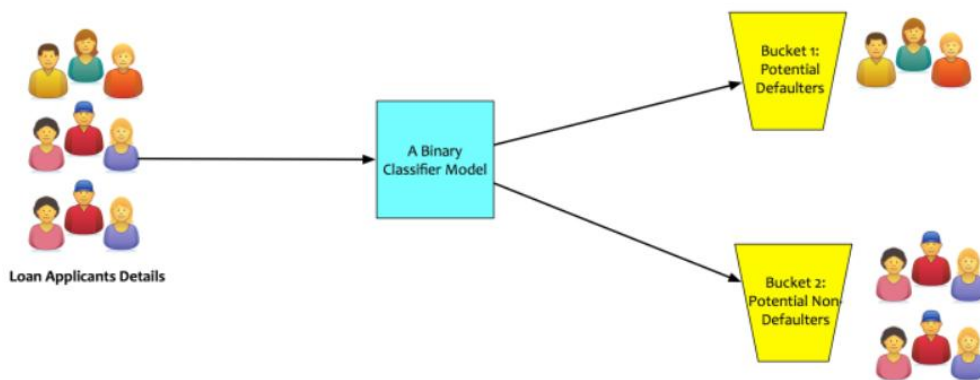


Fig 1.8 Classification Model

Classification algorithms can be further divided into following types:

1. Decision tree classification

In this algorithm by using a decision tree, a classification model is built. Here every node indicates a test case for an attribute and every branch of the node is a possible value for that attribute.

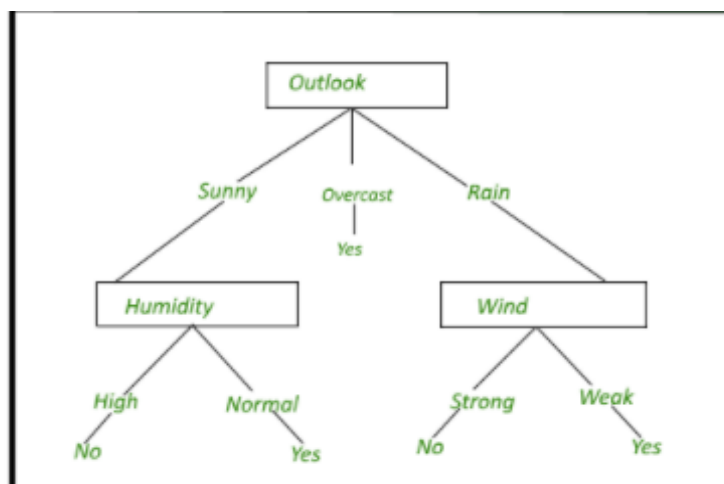


Fig 1.9 Decision Tree for Playing Tennis

2. Random forest classification

In this algorithm a collection of decision trees are chosen randomly from the subset of training set. The random forest classification algorithm combines the outputs of all the different decision trees to arrive at a final output estimate, that is more accurate than any of the individual trees.

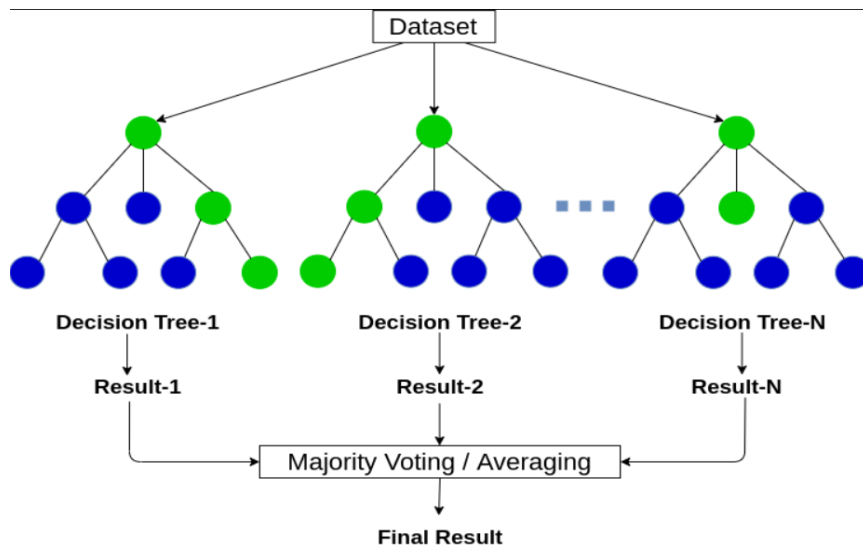


Fig 1.10 Random Forest

CHAPTER-2

PYTHON

2.1 INTRODUCTION

Python is a free language that's used in data science, web development, computing, and lots of scientific applications. Learning it allows the programmer to specialize in a problem solving, instead of that specialize in syntax. Its relative size and simplified syntax provides it a position over languages like Java and C++, yet the abundance of libraries gives it the facility needed to accomplish great things.

Following are the features of python:

1. It's a scripting language that's high-level, interpreted, interactive, and object-oriented.
2. It is interpreted since the interpreter processes it at runtime.

It is Object-Oriented in the sense that it promotes Object-Oriented programming, which encapsulates code inside objects.

2.2 HISTORY OF PYTHON

1. GUIDO VAN ROSSUM is the creator of Python. During the early 1990s
2. Python version 2.0 was released on October 16, 2000, with a slew of new features, including Unicode support.
3. On December 3, 2008, Python 3.0 was announced. Many of its features were reverted to Python version 2.6 and 2.7.

2.3 PYTHON VARIABLES

1. Variables are storage locations set aside for storing values. This indicates that whenever a variable is created some memory is allocated for it.
2. Declaring a variable in Python does not have a particular instruction. When a value is allocated to a variable, it becomes a variable.
3. Python variables don't explicitly reserve space in memory. The declaration happens automatically when a value is assigned to variable.

4. Python has the following data types :

- Strings
- Lists
- Tuples
- Sets
- Dictionary

2.4 STRINGS IN PYTHON

1. Since Python lacks a char data type, a single character is simply a one-length string.
2. Strings are enclosed by single or double quotation marks, for example, `a = "Hello"` or `a='Hello'`.
3. The variable name and string are followed by the equal to sign when assigning a string to a variable.
4. Square brackets are commonly used to access string components.

2.5 PYTHON LISTS

1. It is formed with square brackets and is used to store several items in a single variable regardless of the data type.
2. List items may be modified, arranged, and repeat values are allowed.
3. List items are numbered, with index [0] for the first item, index [1] for the second, and so on.
4. They're adaptable, which means you can modify, add, or remove things from a list after it's been created.
5. Examples of list:

```
list1 = ["apple", "banana", "cherry"]
```

```
list2 = [1, 19, 6, 3, 4]
```

```
list3 = [False, True, True, False]
```

2.6 PYTHON TUPLES

1. Tuples, which are represented with round brackets, are used to store more than one item in a single variable.
2. They're ordered, unchangeable and repeated values are allowed.

3. They're immutable, which means you can't add, delete, or alter objects after they've been formed.
4. These objects are indexed, with the first item having an index of [0], the second having an index of [1], and so on.
5. Examples of tuple

```
Tuple = ("apple", "banana", "cherry")
```

2.7 PYTHON SETS

1. Sets are used to storing a large number of items in a single variable.
2. It's a set that's both unindexed and unordered, and it's denoted by curly brackets.
3. Set objects are unchangeable, unordered, and don't allow duplicate values.
4. Unordered means that when it's used, the set items will appear in any order and aren't marked by a key or index. They're unchangeable, which means we can't change them once they've been formed.
5. Example of set

```
Set = {"apple", "banana", "cherry", "apple"}
```

2.8 PYTHON DICTIONARIES

1. Data is usually stored in key: value pairs in dictionaries.
2. It's an unstructured, modifiable array that doesn't allow duplicate values.
3. Keys and values are expressed by curly brackets in dictionaries.
4. Dictionary objects are organized into key: value pairs and can be found by looking up the key name.
5. Dictionaries are modifiable, which means that things can be added, removed, or changed after they have been produced.
6. Example: Dict = { "brand": "Ford", "model": "Mustang", "year": 1964, "year": 2020 }

CHAPTER-3

LITERATURE SURVEY

The project entitled “Drinking Water Quality Detection Using Machine Learning Techniques” is based on the research paper published on IEEE of DOI 10.1109/CAS47993.2019.9075730. The algorithms used in this paper were Decision tree classifier and Neural Networks (ANN) is used to predict the quality of drinking water. After the implementation it was found that ANN was the best among the two in predicting the quality of drinking water with an accuracy of 89.25% followed by Decision tree classifier with 81.25%.

SVM and random forest are the algorithms used in this project to predict the quality of drinking water. Random Forest is better suited to multiclass problems than SVM, which is better suited to two-class classifications. To use SVM on a multiclass problem, it must be broken down into multiple binary classification problems. Similarly, Decision Tree is a tree-based approach that is called non-parametric because it makes no assumptions about the distribution of data or the true model's structure, and it needs less data cleaning than other approaches. Hence, considering all the pros and cons of the above algorithms one can predict that Random forest or Support Vector Machine would predict the quality of the water in better way with higher accuracy.

CHAPTER-4

CASE STUDY

4.1 PROBLEM STATEMENT

Drinking water quality detection: The main goal of this problem is to test the quality of the drinking water with the help of its attributes like acidity, ph value, density, salt content. Since water is a basic necessity and there is always a scope of finding pollutants in it which shows negative impact on health of the one who consumes it. Generally it takes few days in order to test the quality of drinking water in laboratory and involves huge cost overheads in order to maintain the laboratory. So with the help of this model time and money can be saved in testing the quality of drinking water.

4.2 DATA SET

The data set used in this project has been acquired from “Kaggle” website. The data set consists of 2200 rows and 19 columns or attributes out of which first 13 columns represent river attributes and the last 6 columns represent quality attributes of water.

1367.81	1	1.53	1	0.2079	1	578.23	22	1	90.3827	13	55.62	6.75	8.9	13.1	3.4	2	1	1
2003.15	0	2.59	0	0.0137	0	1433.05	12	0	65.8742	6	5	1.31	7.2	14.2	1.4	0	0	0
2176.88	0	2.7	0	0.0213	0	1402.55	3	0	3.2274	2	13.06	2.33	7.1	14.5	1.4	1	0	0
125.07	2	0.25	2	0.4993	2	202.54	49	2	214.1973	18	146.2	20.25	5.2	11.6	5.7	3	1	2
1764.45	1	1.66	1	0.0562	1	755.98	26	1	109.8859	10	47.29	5.46	8.7	12.4	3.2	2	1	1
150.25	2	0.29	2	0.3491	2	214.07	50	2	234.5082	17	200.39	11.92	5.7	11.7	5.8	3	1	2
2960.06	0	3.37	0	0.0368	0	1944.65	10	0	11.8541	0	5.76	1.87	7.3	14.1	1.1	0	0	0
532.33	2	0.29	2	0.3801	2	405.73	48	2	240.8542	17	197.59	11.94	5.3	12	4.3	3	1	2
1457.9	1	2.1	1	0.083	1	544.19	35	1	170.7999	15	54.35	7.14	7.9	13.3	3.7	2	1	1
140.25	2	0.52	2	0.4984	2	82.42	42	2	279.7352	20	169.47	11.14	6.2	10.5	4.9	3	1	2
872.72	2	0.66	2	0.3487	2	264.54	44	2	297.3227	19	143.05	11.3	6.2	11.2	4.2	3	1	2
2914.14	0	3.02	0	0.0199	0	1687.92	3	0	5.8341	3	6.22	1.44	6.7	15.7	0	1	0	0
973.85	2	0.58	2	0.3054	2	346.29	49	2	271.8584	16	237.47	14.89	5	11.1	5.4	3	1	2
965.97	2	0.04	2	0.3951	2	445.62	45	2	205.4787	18	229.15	29.05	5.2	11.1	4.8	3	1	2
2178.18	0	3.49	0	0.0018	0	1718.72	18	0	56.404	4	7.41	1.03	7	15.6	0.4	1	0	0
262.35	2	0.12	2	0.4957	2	137.83	47	2	276.9208	18	205.32	28.42	6.3	10.8	5.1	3	1	2
2198.91	0	3.54	0	0.034	0	1526.67	16	0	47.5677	2	16.67	3.62	7.3	14.6	0.8	0	0	0
976.78	2	0.36	2	0.4275	2	123.14	46	2	292.5373	17	173.17	23.38	5.3	11.7	5	3	1	2
496.75	2	0.87	2	0.4486	2	471.74	49	2	219.5059	18	179	10.99	5.3	12	4.5	3	1	2
2102.06	0	3.78	0	0.0059	0	1921.77	3	0	14.9947	3	15.8	1.87	7	15.3	1	1	0	0
756.08	2	0.16	2	0.3062	2	370.06	50	2	219.0918	21	248.08	13.57	5.9	10.8	5.3	3	1	2
2929.26	0	2.94	0	0.0216	0	1603.98	4	0	21.1178	0	19.86	3.17	7.3	14.4	0.4	0	0	0
25.38	2	0.55	2	0.4265	2	164.81	39	2	223.8352	21	220.59	13.51	5.5	11.3	4.9	3	1	2
1921.89	1	1.37	1	0.1513	1	673.34	21	1	82.0145	12	49.78	8.94	7.6	12.7	3.3	2	1	1
665.19	2	0.67	2	0.2698	2	330.64	41	2	219.0719	20	174.96	28.29	6.3	10.6	4.4	3	1	2
2372.26	0	3.24	0	0.0001	0	1183.42	6	0	11.3374	7	15.87	4.55	6.7	14.5	0.1	1	0	0
2222.58	0	3.83	0	0.0321	0	1564.5	5	0	24.6157	1	12.15	1.35	7.5	14	0.2	0	0	0
1458.99	1	1.89	1	0.0593	1	979.72	34	1	179.948	14	129.33	6.36	7.9	13	3.8	2	1	1
2397.05	0	3.31	0	0.0157	0	1633.7	11	0	33.993	8	1.55	1.98	6.7	15.5	1	0	0	0
2903.78	0	3.33	0	0.0499	0	1262.8	4	0	36.8437	4	6.75	3.73	6.8	16	1	0	0	0
1722.3	1	1.46	1	0.1676	1	770.51	22	1	108.2088	11	78.61	7.23	8	13.9	3.6	2	1	1

Fig 4.1 CSV File of Water

4.3 OBJECTIVE OF CASE STUDY

The main objective of the drinking water quality dataset is to predict which of the physiochemical features make good drinking water. This model examines the dataset and predicts the quality of water based on the 7 attributes of dataset. Random forest and SVM are used to estimate drinking water quality. The data set is trained and tested against each algorithm. Results of the algorithms are compared and the one which gives the highest accuracy is considered best to predict the quality of drinking water.

CHAPTER-5

DESIGN

5.1 ARCHITECTURE

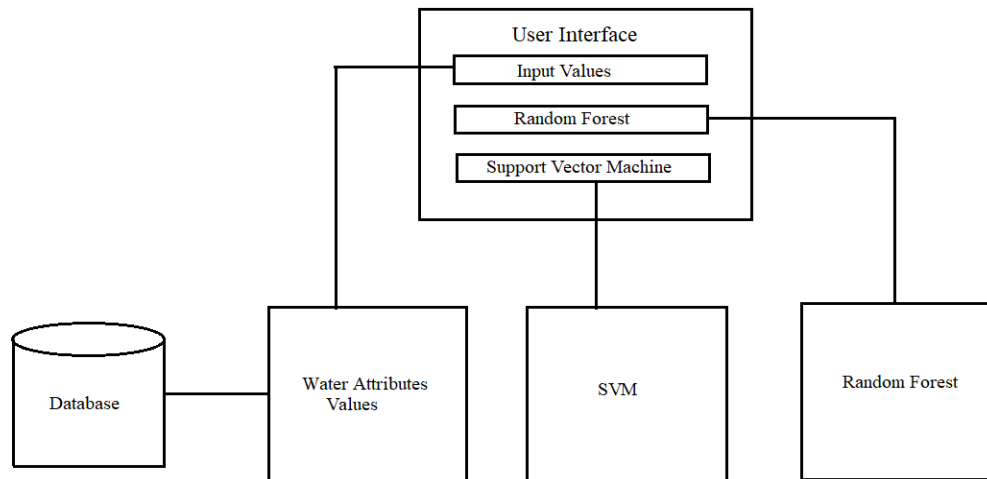


Fig 5.1 Architecture Diagram

The Architecture contains 3 modules:

1. User Interface
2. Connecting to database
3. Implementing algorithms

5.2 PROCESS FLOW

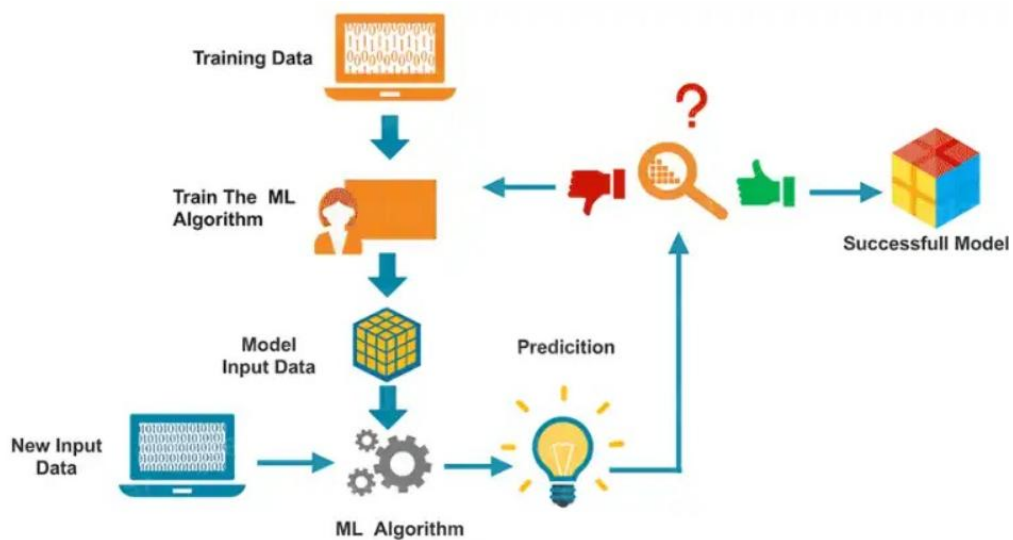


Fig 5.2 Process Flow Diagram

5.3 DATA FLOW

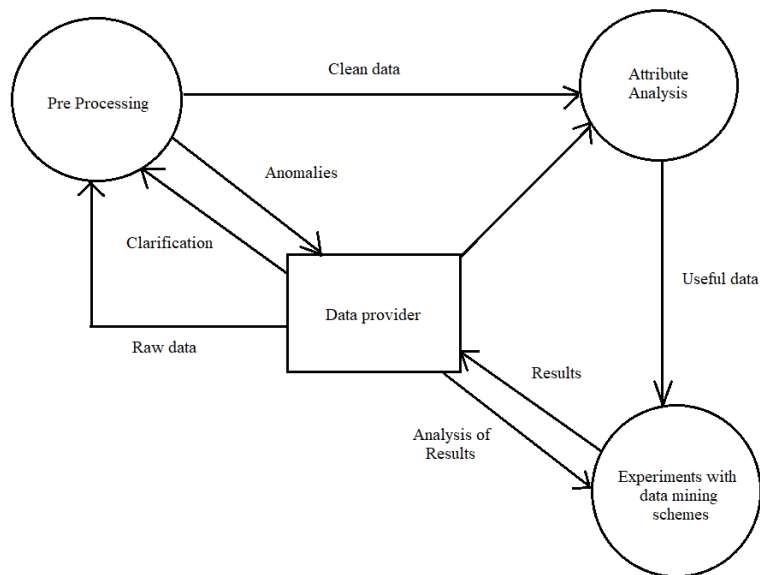


Fig 5.3 Data Flow Diagram

Data flow diagrams generally depict the flow of information in a system graphically. The flow starts from gathering the dataset, pre-processing the data, visualizing the data and ends at estimating the accuracies of the model.

5.4 FLOW CHART

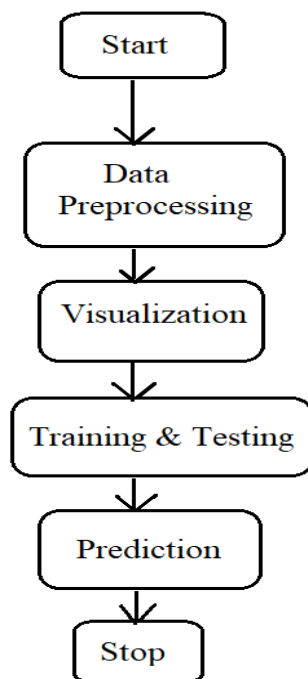


Fig 5.4 Flow Chart

5.5 UML DIAGRAM

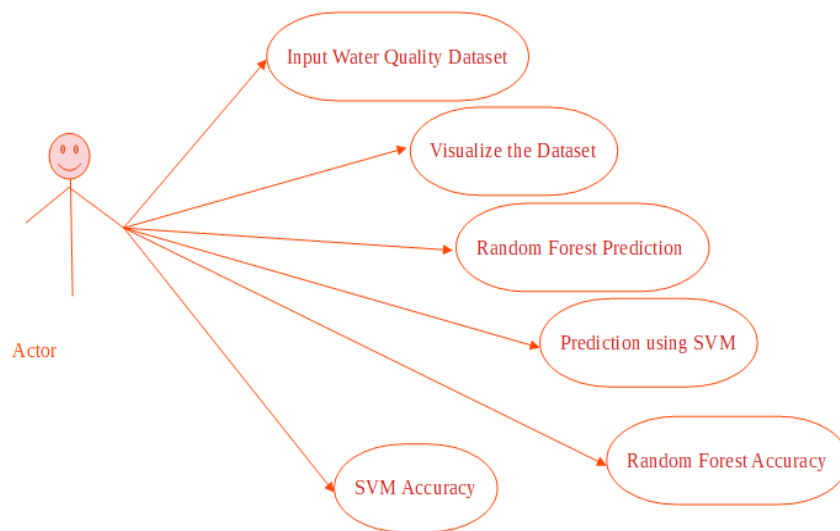


Fig 5.5 UML Diagram

The user interacts with the application through a UI, which allows the customer to perform the specified operations, such as supplying the Training Dataset and Test Data, and then measuring the SVM and RF accuracies, as well as visualization and predictions.

5.6 SEQUENCE DIAGRAM

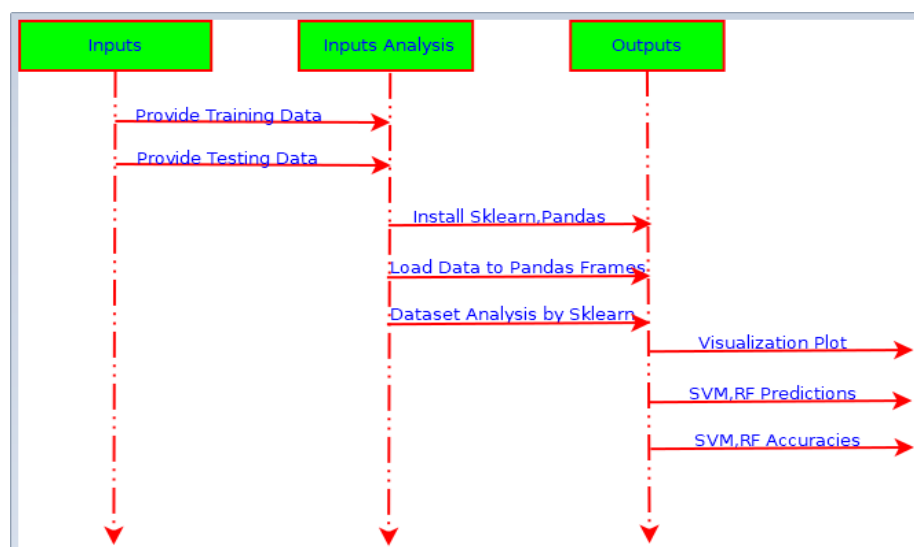


Fig 5.6 Sequence Diagram

From the above-mentioned sequence diagram, one must proceed in the following order: Filling the necessary information as seen in the diagram above. Provide the Training Dataset and the Test Data to be evaluated, followed by the calculation of SVM and RF accuracies, as well as visualization and forecasts.

5.7 ACTIVITY DIAGRAM

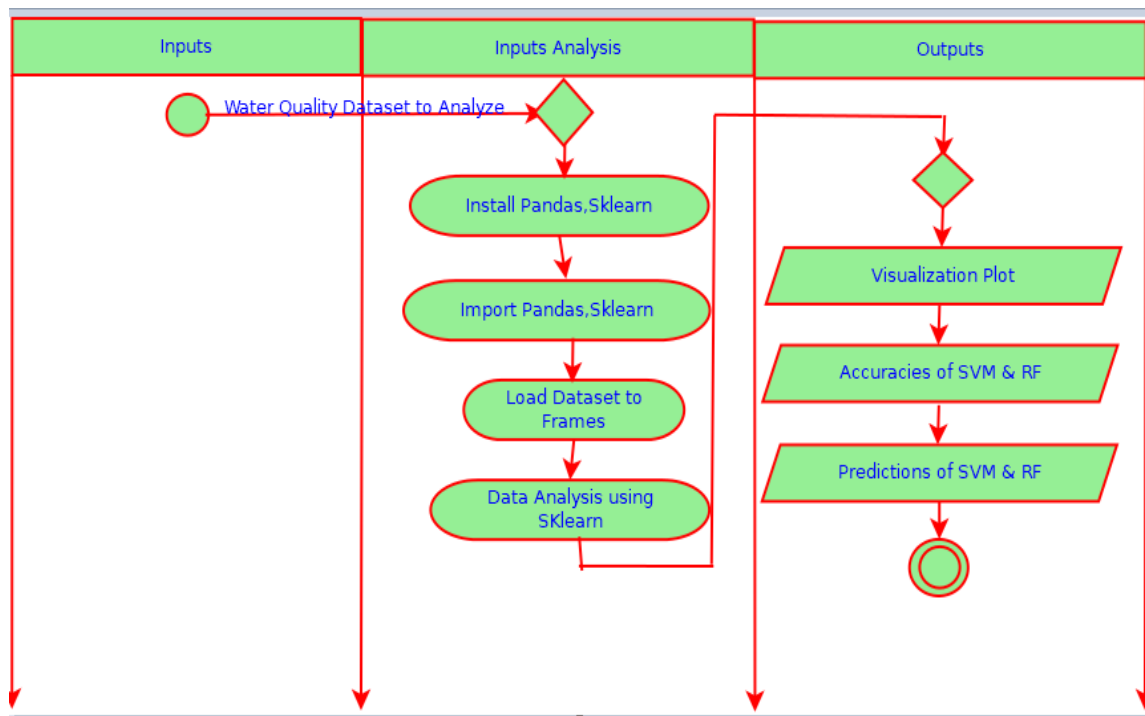


Fig 5.7 Activity Diagram

An activity diagram is essentially a flow chart that depicts the flow of information from one activity to the next. In the above diagram, one can see that the first step is to provide the Test Data as well as the Training Dataset, which consists of details, and then train the model, test the model, and measure the accuracies as well as the SVM and RF predictions.

CHAPTER-6

IMPLEMENTATION

6.1 IMPORTING LIBRARIES

```
import numpy as np
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn import *
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

Fig 6.1 Importing Libraries

- The Numpy includes data structures for multi-dimensional arrays and matrices that can be used to perform mathematical operations.
- Pandas package provide fast, flexible, and expressive data structures which are used to process data frames.
- Sklearn is the main package which is used for machine learning and contains all the necessary modules for implementing algorithms.
-

6.2 USER INTERFACE

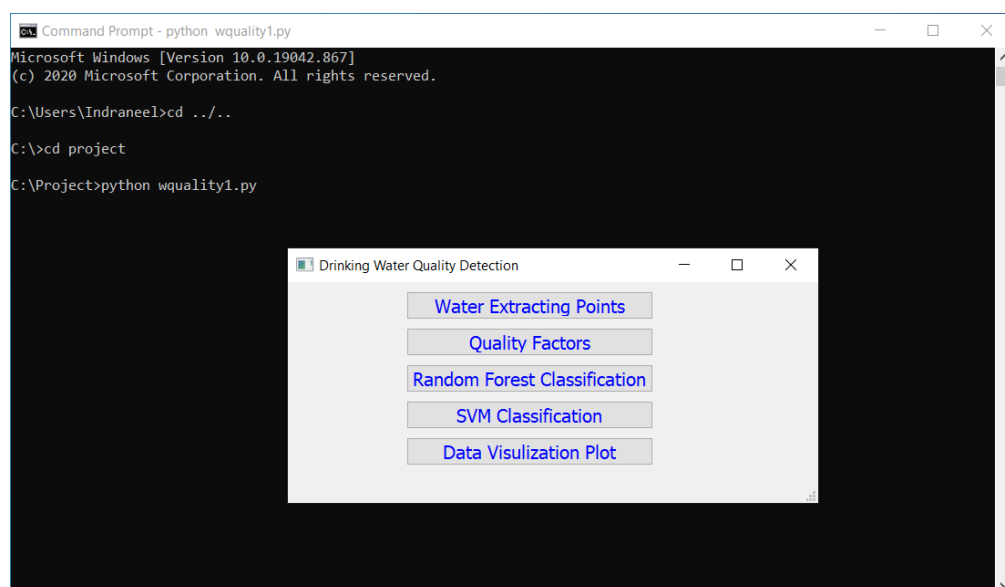


Fig 6.2 Implementing User Interface

The above User Interface contains five buttons. First two buttons are used to enter the input data, third button is used for implementing random forest, fourth button is used for implementing support vector machine and the last button is used for displaying a visual plot. When user clicks on first button below window is displayed and the user needs to enter the data.

Fig 6.3 Water Extraction Points

SQLite is the software used for connecting to a database. A database has been created which consists of two tables “qfacts” and “rfacts” where q and r stands for quality and river. Initially the rfacts table contains only 1 row of data

```
CREATE TABLE rfacts
(
pid varchar(6) NOT NULL,
s1 char(8),s2 char(8),s3 char(8),s4 char(8),s5 char(8),s6 char(8),s7 char(8),s8 char(8),s9 char(8),s10 char(8),s11 char(8),s12 char(8)
);
```

Fig 6.4 Creation of Rfacts Table

A table named rfacts is created and is used to store attributes of river. The table consists of 14, except the first column all the remaining columns are of char data type of length 8. First column is used to determine the sample number hence it is assigned with a constraint NOT NULL which means the field should not be left empty while entering the data.

```

C:\Project\sqlite3.exe
SQLite version 3.34.1 2021-01-20 14:10:07
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .open wquall1
sqlite> .tables
rfacts rfacts
sqlite> select * from rfacts;
001|1000|1|2.00|1|0.01|1|100.00|21|1|10.00|5|100
sqlite>

```

Fig 6.5 Rfacts Table

Extraction point No	2
River width [21 - 3000]m	400
Nearby Industrial Areas [0 - low, 1 - medium, 2 - high]	1
River slantness [0.00 - 4.00]	0.45
Ghat Area [0 plain area, 1 area with curves, 2 Ghat area]	0
Salty content in the river soil Bed [0.00 - 0.50]	0.35
No.River flow diverters [0 - 2]	1
Straight river length before the point [1.00 - 2000.00]m	500
No.of streams mixing in last five Kms [0 - 50]	25
Deep bends of river [0 - No, 1 - slight, 2 - Deep bend]	1
Presence of clay soil in kg/cubicmeter [0.00 - 300.00]	200
No.of extracting points in last 5kms [0 - 20]	10
River Curvature at the spot [0.00 - 250.00]	150
<input type="button" value="Store Details in Data base"/>	

Fig 6.6 Rfacts Data Entry

In the database a new row has been added to the rfacts table as the user has clicked on the store button at the bottom of the above window.

```

C:\Project\sqlite3.exe
SQLite version 3.34.1 2021-01-20 14:10:07
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .open wqual1
sqlite> .tables
qfacts  rfacts
sqlite> select * from rfacts;
001|1000|1|2.00|1|0.01|1|100.00|21|1|10.00|5|100
sqlite> select * from rfacts;
001|1000|1|2.00|1|0.01|1|100.00|21|1|10.00|5|100
2|400|1|0.45|0|0.35|1|500|25|1|200|10|150
sqlite>

```

Fig 6.7 Updated Rfacts Table

When the user clicks on second button in the user interface the below window is displayed which consists of 6 attributes excluding the sample number and data related to quality factors is to be entered in the window by the user.

Fig 6.8 Quality Factors

```

CREATE TABLE qfacts
(
sid varchar(6) NOT NULL,
s1 char(6),s2 char(6),s3 char(6),s4 char(6),s5 char(6),s6 char(6)
);

```

Fig 6.9 Creation of Qfacts Table

A table named as qfacts is created and is used to store attributes of water. The table consists of 7, except the first column all the remaining columns are of char data type of length 6. First column is used to determine the sample number hence it is assigned with a constraint NOT NULL which means the field should not be left empty while entering the data. The qfacts table contains only 2 rows of data.

```

C:\Project\sqlite3.exe
SQLite version 3.34.1 2021-01-20 14:10:07
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .open wqual1
sqlite> .tables
qfacts  rfacts
sqlite> select * from qfacts;
0001|3.45|6|12.6|4.6|1|0
2|15.5|7.0|12.8|5.4|2|1
sqlite>

```

Fig 6.10 Qfacts Table

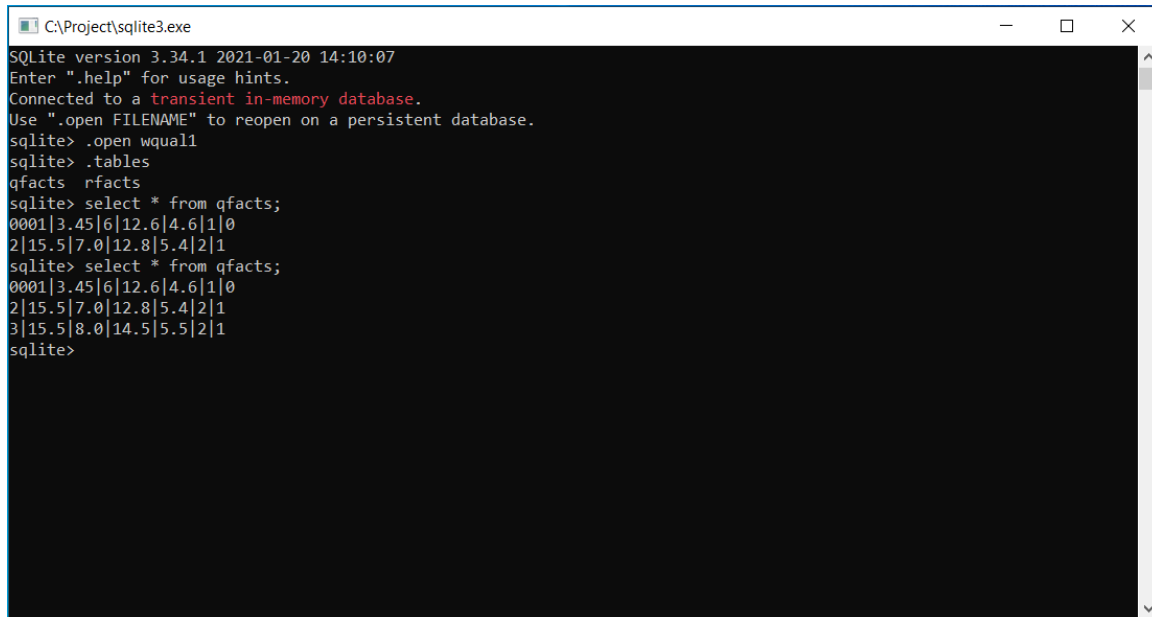
After entering the data user needs to click on the button at the bottom of the window with the name store in order to store the data in the database.

Sample No [0001 - 9999]	3
Salinity Level [0.00 - 30.00]	15.5
PH level [5.0 - 9.0]	8.0
Dissolved Oxygen [10.1 - 16.0]	14.5
Zinc Level [0.1 - 6.0]	5.5
Normaized Carbon level [0-3]	2
Flourides Presence [0 for No, 1 for Yes]	1

Store Quality Factors in Data base

Fig 6.11 Qfacts Data Entry

In the database a new row has been added to the qfacts table and now it consists of 3 rows.



```
C:\Project\sqlite3.exe
SQLite version 3.34.1 2021-01-20 14:10:07
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .open wqual1
sqlite> .tables
qfacts  rfacts
sqlite> select * from qfacts;
0001|3.45|6|12.6|4.6|1|0
2|15.5|7.0|12.8|5.4|2|1
sqlite> select * from qfacts;
0001|3.45|6|12.6|4.6|1|0
2|15.5|7.0|12.8|5.4|2|1
3|15.5|8.0|14.5|5.5|2|1
sqlite>
```

Fig 6.12 Updated Qfacts Table

6.3 DATA SPLITTING

In machine learning in order to access the performance of the classifier one need to train the classifier using training set and then test the performance of the classifier on unseen testing set. The whole dataset is generally divided into two parts i.e. training set and testing set. In this project the data set has been divided in the ratio of 8:2. Eighty percent of the data set is used for training , while the remaining twenty percent is used for testing.

6.4 SUPPORT VECTOR MACHINE

```
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
training_data = np.genfromtxt('waterset.csv', delimiter=',')
inputs = training_data[:, :-1]
outputs = training_data[:, -1]
training_inputs = inputs[:1200]
training_outputs = outputs[:1200]
testing_inputs = inputs[1200:]
testing_outputs = outputs[1200:]
classifier = SVC()
classifier.fit(training_inputs, training_outputs)
predictions = classifier.predict(testing_inputs)
accuracy = 100.0 * accuracy_score(testing_outputs, predictions)
print ("The accuracy of SVM Classifier on testing data is: " + str(accuracy))
```

Fig 6.13 Support Vector Machine

They are just a group of supervised learning methods for classification, regression, and detecting outliers.

The advantages of SVM are:

1. Only accurate when the number of samples are less than the number of dimensions.
2. It uses a subset of coaching points (called support vectors) within the decision function, making it memory efficient.
3. Versatile: Different Kernel functions can be specified for the option function. Custom kernels may be defined in addition to the standard kernels.

The disadvantages of SVM are:

1. If the number of features exceeds the number of samples, over fitting in chosen kernel function will be prevented.
2. Probability estimates are determined using a rich five-fold cross-validation method rather than directly by SVMs.

6.5 RANDOM FOREST

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
training_data = np.genfromtxt('waterset.csv', delimiter=',')
inputs = training_data[:, :-1]
outputs = training_data[:, -1]
training_inputs = inputs[:1200]
training_outputs = outputs[:1200]
testing_inputs = inputs[1200:]
testing_outputs = outputs[1200:]
classifier = RandomForestClassifier()
classifier.fit(training_inputs, training_outputs)
predictions = classifier.predict(testing_inputs)
accuracy = 100.0 * accuracy_score(testing_outputs, predictions)
print ("The accuracy of RF Classifier on testing data is: " + str(accuracy))
```

Fig 6.14 Random Forest

A Random Forest is a technique that performs both regression and classification tasks by combining multiple decision trees and a technique known as Bootstrap and Aggregation, also known as bagging. Rather than relying on individual decision trees, the basic concept is to combine several decision trees to determine the final production. It's a form of learning in which you combine different algorithms or use the same algorithm multiple times to create a more efficient prediction model. The Random Forest algorithm constructs decision trees from data samples, and then receives predictions from each one before voting on the best solution. Since it combines the outcomes to reduce over-fitting, it's a better ensemble solution than a single decision tree.

CHAPTER 7

RESULT ANALYSIS

7.1 RANDOM FOREST

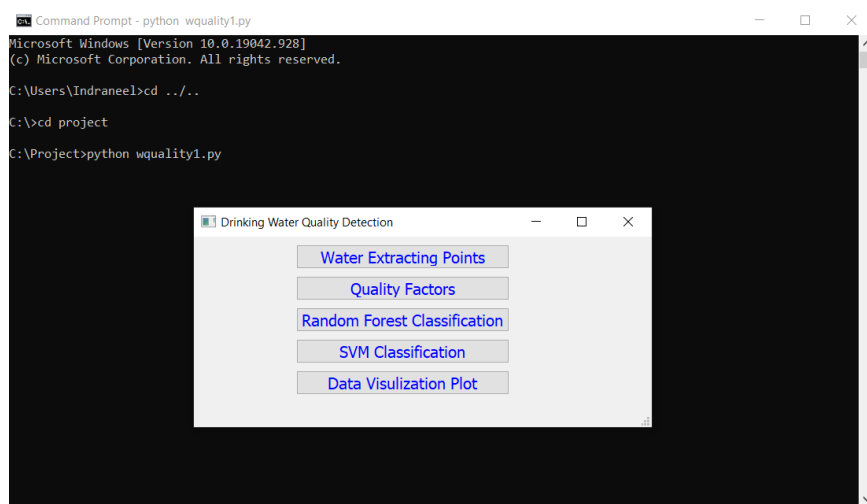


Fig 7.1 Random Forest UI

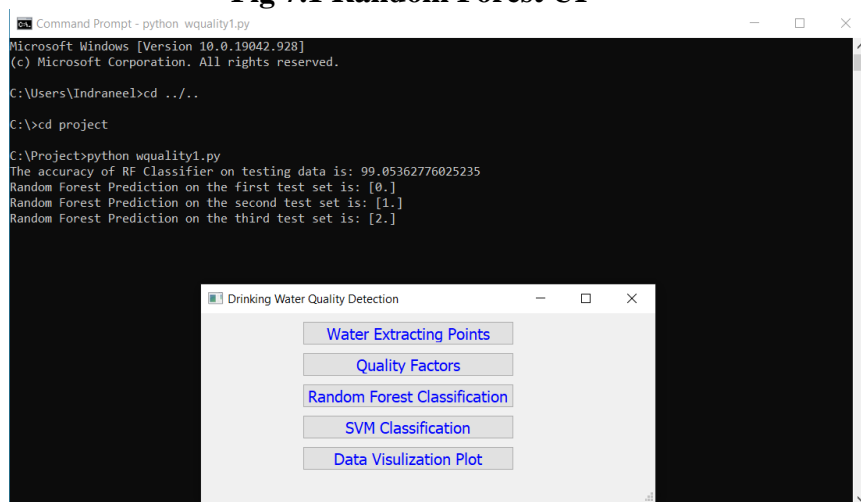


Fig 7.2 Accuracy of Random Forest

It is found that random forest has achieved an accuracy of 99. Three different test sets are used in order to show all the possible results. First test set has displayed the number 0 which means that the water has good quality, similarly for second and third the numbers 1 and 2 is displayed which refers to poor quality and very poor quality of water.

7.2 SUPPORT VECTOR MACHINE

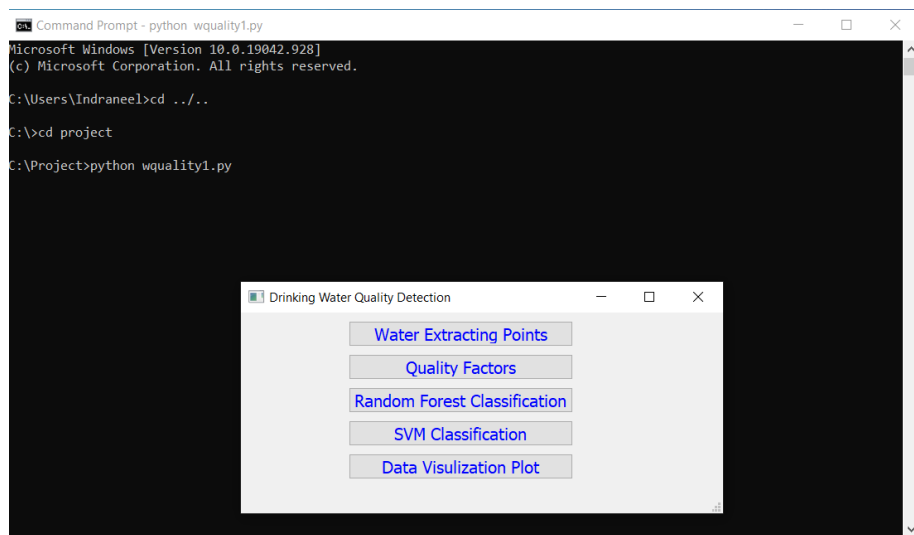


Fig 7.3 Support Vector Machine UI

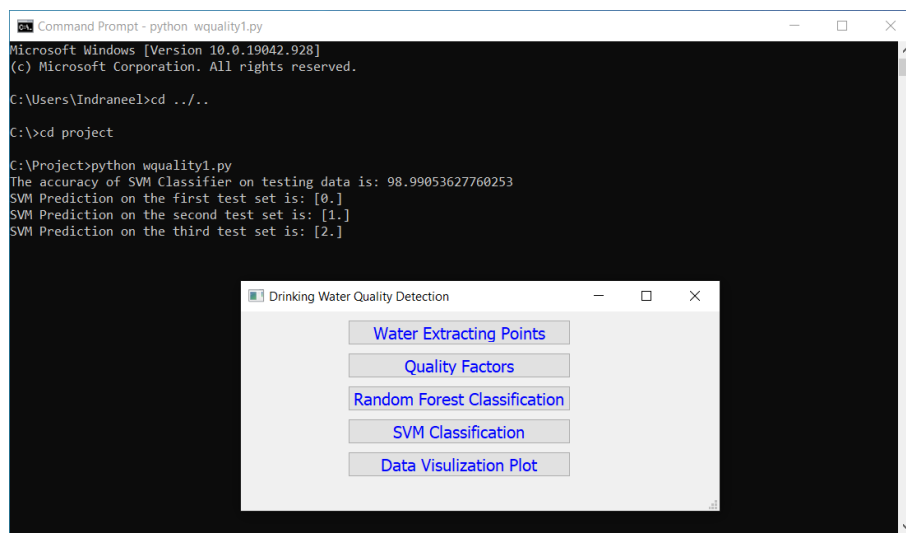


Fig 7.4 Accuracy of Support Vector Machine

It is found that SVM has achieved an accuracy of 98. Three different test sets are used in order to show all the results. First test set has displayed the number 0 which means that the water has good quality, similarly for second and third the number 1 and 2 are displayed which refers to poor quality and very poor quality of water

CHAPTER 8

CONCLUSION & FUTURE SCOPE

Firstly, the data set used for this project has been acquired from “kaggle” website. Secondly, several packages like numpy, pandas, seaborn, matplotlib have been used in order to perform the statistical analysis, visualization and data splitting. Thirdly, the data set has been split in the ratio of 8:2 i.e training and testing set using which Random Forest and SVM have been trained and tested. Based on the results it is concluded that Random Forest is the most suitable algorithm for predicting the quality of drinking water with an accuracy of “0.99”.

The project entitled “Drinking Water Quality Detection Using Machine Learning Techniques.” is very useful to the Waterworks department to take measures to improve the water quality, when it is predicted to be not good. The project is useful to save people against the diseases that can occur due to poor water quality. This project finally leads to the improvement of people life span.

REFERENCES

1. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
2. <https://www.geeksforgeeks.org/decision-tree/>
3. [https://searchenterpriseai.techtarget.com/definition/machine-learning-ML#:~:text=Machine%20learning%20\(ML\)%20is%20a,to%20predict%20new%20output%20values.](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.)
4. <https://www.python.org/>
5. <https://www.numpy.org/>