## CHAPTER – 1

## INTRODUCTION

Deep learning is a collection of statistical techniques of machine learning for learning feature hierarchies that are based on artificial neural networks. As neural network is going to mimic the human brain so deep learning is also a kind of mimic to human brain. The idea of deep learning is to build such algorithm that can mimic the brain. In deep learning, nothing is programmed explicitly. Deep learning algorithms are used in analyzing and interpreting large amounts of data. .

Aquaculture is a vital component for ensuring global food security. Water quality is affected by many factors, such as fish density, quality of the feed, feeding interval, climate, and more. The change in water quality will upset the balance of the system with algae bloom, bacterial growth. This can lead to severe problems, such as trigger stress, lack of food intake, vulnerability to diseases, and increased mortality rate of fish. With the help of deep learning algorithms aquaculture systems are upgraded to smart aquaculture systems for improving performance and efficiency.

This project focuses on predicting the quality of the water by applying various deep learning techniques. The outcome of these techniques are compared on the basis of mean absolute error, mean square error and root mean square error. Therefore, by predicting water quality will drastically improve productivity to make aquaculture more profitable and sustainable.

**1.1 Deep learning**

Deep learning can be considered as a subset of [machine learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning). It is a field that is based on learning and improving on its own by examining computer algorithms. While machine learning uses simpler concepts, deep learning works with artificial neural networks, which are designed to imitate how humans think and learn. In deep learning, a computer model learns to perform classification tasks directly from images, text and sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers. There are mainly seven types of deeps learning models. They are Feedforward Networks, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory Networks, Generative Adversarial Networks, Radial Basis Function Network, Multilayer Perceptrons and Deep Belief Networks. These algorithms can ingest and process unstructured data like images and automates feature extraction.

**Advantages of deep learning**

The following are the advantages of deep learning:

* **Feature generation automation**

Deep learning algorithms can generate new features from among a limited number located in the training dataset without additional human intervention. This means deep learning can perform complex tasks that often require extensive feature engineering.

* **Works well with unstructured data**

One of the biggest advantage of deep learning is its ability to work with unstructured data. Text, images, and voice are some of the most common data formats that are used. Classical machine learning algorithms are limited in their ability to analyze unstructured data, where deep learning algorithms will make the most impact on unstructured data. Training deep learning networks with unstructured data and appropriate labeling can help to get better results.

* **Better self-learning capabilities**

The multiple layers in deep neural networks allow models to become more efficient at learning complex features and performing more intensive computational tasks, i.e., execute many complex operations simultaneously. Deep learning algorithms have the ability to learn from its own errors. It can verify the accuracy of its predictions or outputs and make necessary adjustments.

* **Supports parallel and distributed algorithms**

A typical neural network or deep learning model takes days to learn the parameters that define the model. Parallel and distributed algorithms address this point by allowing deep learning models to be trained much faster. Models can be trained using local training (use one machine to train the model), with GPUs, or a combination of both. With data or the model itself being distributed across multiple machines, training is more effective.

* **Cost effectiveness**

While training deep learning models can be cost-intensive, once trained, it can help businesses cut down on unnecessary expenditure. In industries such as manufacturing, consulting, or even retail, the cost of an inaccurate prediction or product defect is massive. It often outweighs the costs of training deep learning models.

* **Advanced analytics**

Deep learning, when applied to data science, can offer better and more effective processing models. Its ability to learn unsupervised drives continuous improvement in accuracy and outcomes. It also offers data scientists with more reliable and concise analysis results.

* **Scalability**

Deep learning is highly scalable due to its ability to process massive amounts of data and perform a lot of computations in a cost- and time-effective manner. This directly impacts productivity and modularity and portability.

Apart from the above mentioned advantages, deep learning algorithms are also applied to customer data in CRM systems, social media and other online data to better segment clients, predict churn and detect fraud.

**1.2 Purpose** Water quality is one of the most important factors in maintaining and managing smart aquaculture. Clean water supports a diversity of plants and wildlife. This project predicts the quality of water for maintaining and managing smart aquaculture. The project uses advanced hybrid deep learning models such as cnn with lstm and cnn with gru and uses the concept of logistic regression algorithm to predict the accuracy of the models.

**1.3 Objective**

Deep learning is the most popular learning algorithm for predicting the future. This project focuses on predicting the quality of the water by applying various deep learning techniques. The outcomes of these techniques are compared on the basis of mean absolute error, mean square error, root mean absolute error.

The prediction made by these hybrid deep learning algorithms will help the one to decide whether to maintain aquaculture or not. CNN, LSTM and GRU are some of the algorithms in deep learning that are used in classifying whether the water is safe or not for aquaculture. By using these algorithms the risk associated with making a prediction is low. The training time of the models is reduced. After applying the algorithms on the data, logistic regression is used to evaluate the accuracy of the model. The main objective of this project is to predict quality of water for aquaculture.

**1.4 Scope**  Aquaculture is a type of farming that involves the process of cultivating aquatic organisms and aquatic plants in water under certain controlled parameters. Water quality is one of the most important factors in a healthy ecosystem. Clean water supports a diversity of plants and wildlife. Though it may seem unrelated at first, actions on land affect the quality of water so there is a need to maintain quality of water for maintaining or managing aquaculture. Aquaculture plays an important role in maintaining the world global security.

**1.5 Motivation**  During the last years, water quality has been threatened by various pollutants. Therefore, modeling and predicting water quality have become very important in controlling water pollution. This work focuses on finding the best feature selection technique and classifier for predicting the water quality to maintain aquaculture. Therefore, there is a need to execute feature selection with a classifier that can be applied to the water quality dataset. A comparative analysis of different methods is carried out and the performance evaluated. Using these techniques, farmers can easily identify the water quality is safe or not for aquaculture.

**1.6 Organization of the project**

The above is the introduction of the project and rest of the project is organized as follows:

**Chapter 2:** This chapter presents the system analysis of the project. The system analysis is to understand the module description and domain of the new system. Both the activities are equally important but the first activity serves as a basis of the functional specifications and then successful design of the proposed system.

**Chapter 3:** This chapter describes the system environment. It gives the usage of the software and source code. The source code contains the python libraries.

**Chapter 4:** This chapter describes the design and implementation details of proposed scheme. The design consists of UML diagrams such as data flow diagram, interaction diagram such as sequence diagram.

**Chapter 5:** This chapter presents the software testing of proposed system. The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product.

**Chapter 6:** This chapter includes information about screenshots of the proposed system. It explains how to upload the CSV file, pre processing the data, data analysis, implementation and prediction of results in terms of precision are in the form of a screenshot.

**Chapter 7:** This chapter presents conclusions of the proposed system of the project.

\*\*\*\*\*

## CHAPTER - 2

## SYSTEM ANALYSIS

Systems analysis is a problem solving method that involves looking at the wider system, breaking apart the parts and figuring out how it works to achieve a particular goal. It is a process of collecting and interpreting facts, identifying the problems and decomposition of a system into its components. System analysis is conducted to study a system or its parts to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

Analysis specifies what the system should do. A system is a general set of parts, step or components that are connected to form a more complex. The first step in solving a problem that involves a system is analyzing that system.

**2.1 System**

A system is “an orderly grouping of interdependent components linked together according to a plan to achieve a specific goal”. A system must have three basic constraints:

* A system must have some structure and behavior which is designed to achieve a predefined objective.
* Interconnectivity and inter dependence must exist among the system components.
* The system is effectively used to provide a large variety of information to the customer.

System analysis is the first step in any kind of perfect development. This is step if not done properly could lead to a faulty system. This step takes the maximum percentage of the total time allotted to the project.

**System perspective** Before loading data to the model, the user needs to preprocess the data then apply the recursive feature elimination algorithm to find out feature attributes to split the data accurately. The data is divided into training and testing data sets, by using a training set build a model and validate the model through the training set.

**2.2 Existing system**

Existing system is implemented by the hybrid Deep Learning (DL) models, Convolutional Neural Network (CNN) with the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) for aquaculture water quality prediction. Then, the system compared CNN with LSTM and CNN with GRU. The process is implemented without removing unwanted data so this system fails to predict the actual quality of water and techniques used in the existing system is not efficient for large amount of data.

**Drawbacks**

* The process is implemented without removing unwanted data.
* It is not efficient for large amount of data.
* Training time is high.

The following are the disadvantages of an existing system. Now a days aquaculture is the vital component of global food security. Water exhibits non linear nature so it is hard to determine the quality by considering few parameters.

**2.3 Proposed system**

In proposed system, the dataset of water is taken as input. To maintain and manage aquaculture, water quality monitoring and Water Quality Prediction (WQP) are essential. The system develops different deep and machine learning algorithm for predicting the water quality either the water is safe or not for aquaculture. To overcome the disadvantage of data duplicates here feature extraction techniques are used. The proposed system uses the hybrid models such as CNN with LSTM, CNN with GRU. Logistic Regression is performed on the result obtained by the CNN with LSTM and CNN with GRU. Then, finally calculating some performance metrics such as MAE, MSE, RMSE and MAPE the water quality is predicted. The main objective of this work is to predict water quality for aquaculture.

**Advantages**

* The process is implemented by replacing null values in the dataset with the value ‘0’.
* Training time is low by implementing the feature extraction techniques such as PCA (Principle Component Analysis).
* It is efficient for large amount of data.
* The risk associated with making a prediction is low.
* Accurate

The deep learning models like LSTM, GRU and CNN have the flexibility in capturing the non linear nature of the water. These models make time series prediction more precise and efficient in terms of training time and required processing power.

**2.4 Feasibility study**

The feasibility of the project is analyzed in this phase and a business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis, the feasibility study of the proposed system is to be carried out.

This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are as follows.

**Economical feasibility**

This study is carried out to check the economic impact that the system will have on the organization. The amount of funds that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well as within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

**Technical feasibility** This study is carried out to check the technical feasibility that is the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resource and being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**Social feasibility**  The aspect of the study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it.

After requirement gathering, the team comes up with a rough plan of the software process. At this step, the team analyzes if the software can be designed to fulfill all requirements of the user and if there is any possibility of software being no more useful. It is also analyzed if the project is economical, practical and technological feasible for the organization take up. The above constraints are considered, which help the developers to conclude the feasibility of software in the project.

**2.5 Software requirements**

A software requirement describes the intended purpose, requirements and nature of software to be developed. It also includes the yield and cost of the software. Software requirements specification plays an important role in creating quality software solutions. The specification is a representation process. Requirements are represented in a manner that ultimately leads to successful software implementation.

**Hardware requirements**

System requirements specify that the minimum requirements that are needed to execute the proposed system. The following are the hardware requirements, these are necessary to execute the project. The processor is an i3 Intel processor or above. Standard windows keyboard and two or three button mouse are used. Ram must be 4gb or above.

**Software requirements**  System requirements specify that the minimum requirements that are needed to execute the proposed system. The following are the software requirements, these are necessary to execute the project. Windows 10 operating system is used. Server side script python language and integrated development environment anaconda sypder are used to build the project.

\*\*\*\*\*\*

## CHAPTER – 3

## SYSTEM ENVIRONMENT

This chapter describes the system environment of the proposed system. The system environment provides support for the development and enhancement of software for management and control. The system environment consists of operating system and software.

**3.1 Operating system and environment**

Operating system is an interaction between user and hardware. Operating system creates user friendly environment it is the main function of the operating system. Another important function is resource management. The operating system controls and coordinates the execution of the programs. So it can able to control the execution of a program because it provides interface to various hardware components such as printer, monitor, keyboard, mouse etc.

**Windows 10**

Windows 10 is the most recent version of the operating system from microsoft. Officially it was released in and was initially offered free of charge to legitimate users of windows 10 and windows 8. This new version combines features from those two previous installments to suit the users in a better way for both desktop/laptop computers as well as mobile devices.

**Spyder**

Spyder is an open-source cross-platform IDE. The python spyder IDE is written completely in python. It is designed by scientists and is exclusively for scientists, data analysts, and engineers. It is also known as the Scientific Python Development IDE and has a huge set of remarkable features which are discussed below.

* Customizable syntax highlighting
* Availability of breakpoints (debugging and conditional breakpoints)
* Interactive execution which allows you to run line, file, cell, etc.
* Run configurations for working directory selections, command-line options, current/ dedicated/ external console, etc
* Can clear variables automatically ( or enter debugging )
* Navigation through cells, functions, blocks, etc can be achieved through the Outline Explorer
* It provides real-time code introspection (the ability to examine what functions, keywords, and classes are, what they are doing and what information they contain)

Apart from the above mentioned features some of the features of spyder ide are: automatic colon insertion after statements like if, while etc. It supports all the python commands inline display for graphics produced using [matplotlib](https://www.edureka.co/blog/python-matplotlib-tutorial/)

**3.2 Python**

Python is an [interpreted](https://en.wikipedia.org/wiki/Interpreted_language), [high level](https://en.wikipedia.org/wiki/High-level_programming_language) and [general-purpose programming](https://en.wikipedia.org/wiki/General-purpose_programming_language) [language.](https://en.wikipedia.org/wiki/General-purpose_programming_language) Created by [guido van rossum](https://en.wikipedia.org/wiki/Guido_van_Rossum) and first released in 1991, python design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with its notable use of [significant white space.](https://en.wikipedia.org/wiki/Off-side_rule) Python [language constructs](https://en.wikipedia.org/wiki/Language_construct) and [object oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aim to help [programmers](https://en.wikipedia.org/wiki/Programmers) write clear, logical code for small and large scale projects. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses english keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

**Python concepts**

Python is perhaps the best language accessible and it's an incredible one to begin programming with. Python is an open source general purpose language it supports the more concepts like object oriented, procedural, functional programming languages.

* Python is interpreted − Python is processed at runtime by the interpreter. User do not need to compile the program before executing it. This is similar to perl and php.
* Python is interactive – User can sit at a python prompt and interact with the interpreter directly to write the programs.
* Easy to interface with c/obj c/java/fortarn.
* Python is object-oriented −Python supports object-oriented style or technique of programming that encapsulates code within objects.
* Python is a beginner's language − Python is a great language for beginner-level programmers and supports the development of a wide range of applications from simple text processing to www browsers to games.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech giant companies like – google, amazon, facebook, instagram, dropbox, uber… etc.

**History of python**

Python was created by guido van rossum in the late eighties and mid nineties at the national exploration organization for arithmetic and software engineering in the nether lands.

Python is developed from numerous different dialects, including abc, modula-3, c, c++, algol-68, small talk and unix shell and other scripting dialects. Python is copyrighted. Like perl, python source code is currently accessible under the GNU. Python is presently kept up by a center advancement group at the organization, although guido van rossum still holds an essential job in coordinating its encouragement.

User created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks and developed a small number of powerful data types.

**3.3 Python features**

Python provides many useful features which make it popular and valuable from the other programming languages. It supports object oriented programming, procedural programming approaches and provides dynamic memory allocation. A few essential features have listed below.

* Easy-to-learn − Python has few keywords, a simple structure, and a clearly defined syntax. This allows the user to pick up the language quickly.
* Easy-to-read − Python code is more clearly defined and visible to the eyes.
* Easy-to-maintain − Python source code is fairly easy to maintaining.
* A broad standard library − Python bulk of the library is very portable and cross platform compatible on unix, windows and macintosh.
* Interactive mode − Python has support for an interactive mode that allows interactive testing and debugging of snippets of code.
* Portable − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* Extendable − User can add low-level modules to the python interpreter. These modules enable programmer to add or customize their tools to be more efficient.

Apart from the above mentioned features, python has a big list of good features, few supports functional and structured programming methods as well as object oriented programming methods. It can be used as a scripting language or can be compiled to byte code for building large applications. It provides very high level dynamic data types and supports dynamic type checking.

**3.4 Standard data types**

The data stored in memory can be of many types. For example, a person age is stored as a numeric value and his or her address is stored as alpha numeric characters. Python has various standard data types that are used to define the operations.

Python has six standard data types:

* Numbers
* Variables
* String
* List
* Tuple
* Dictionary

**Python numbers**

Number data types store numeric values. Number objects are created when you assign a value to them. Number objects are created when some value is assigned to a variable. The various types of numbers like int, float, complex is also as the functions in python.

**Python variables**

Variables are nothing but reserved memory locations to store values. This means that when user creates a variable system reserve some space in memory. Based on the data type of a variable, the interpreter allocates memory and decides what can be stored in the reserved memory.

**Python strings**

Strings in python are identified as a contiguous set of characters represented in the quotation marks. Python allows for either pair of single or double quotes. Subsets of strings can be taken using the slice operator ([ ] and [:]) with indexes starting at 0 at the beginning of the string and working their way from 1 at the end.

**Python lists**

Lists are the most versatile of python compound data types. A list contains items separated by commas and enclosed within square brackets. The values stored in a list can be accessed using the slice operator is represented as ( [ ] and [:] ) with indexes starting at 0 at the beginning of the list and working their way to end -1. The plus (+) sign is the list concatenation operator and the asterisk (\*) is the repetition operator.

**Python tuples**  A tuple is another sequence data type that is similar to the list. A tuple consists of several values separated by commas. Unlike lists however tuples are enclosed within parentheses. The main differences between lists and tuples are lists are enclosed in brackets while tuples are enclosed in parentheses and cannot be updated as lists.

**Python dictionary**

Python dictionaries are kind of hash table types. They work like associative arrays or hashes found in Perl and consist of key value pairs. A dictionary key can be almost any python type, but are usually numbers or strings. Values on the other hand, can be any arbitrary python object. Dictionaries are enclosed by curly braces and values can be assigned and accessed using square braces.

**3.5 Python libraries for deep learning**

[The python language](https://docs.python.org/3/reference/index.html#reference-index) describes the exact syntax and semantics of the python, this library reference manual describes the standard library that is distributed with python. The library contains built-in modules that provide access to system functionality such as file I/O that would be in accessible to python programmers if it is absent. As well as libraries written in python that provide standardized solutions for many problems that occurs in everyday programming. Some of these libraries are explicitly designed to encourage and enhance the portability of python programs by abstracting away platform-specifics into platform-neutral APIs.

In addition to the standard library, there is a growing collection of several thousand components from individual programs and modules to packages and entire application development frameworks, available from the python.

**NumPy**

Numpy is a python package. It stands for ‘numerical python’. It is a library consisting of multidimensional array objects and a collection of routines for processing of array numeric, the ancestor of numpy, was developed by jim hugunin. In 2005, travis oliphant created the numpy package by incorporating the features of numerary into the numeric package. There are many contributors to this open-source project.

With the revolution of data science, data analysis libraries like numpy, scipy, pandas etc. have seen a lot of growth.. Numpy provides a convenient and efficient way to handle the vast amount of data. Numpy is also very convenient with matrix multiplication

and data reshaping. Numpy is fast which makes it reasonable to work with a large set of data.

**Scikit-learn**

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in python. It is licensed under a permissive simplified bsd license and is distributed under many linux distributions, encouraging academic and commercial use. Extensions or modules for scipy care are conventionally named [scikits.](http://scikits.appspot.com/scikits) As such, the module provides learning algorithms and is named scikit-learn. The vision for the library is a level of robustness and support required for use in production systems. **Pandas**

Pandas is an open-source python library providing high performance data manipulation and analysis tools using its powerful data structures. Before pandas python was majorly used for data managing and preparation. It had very little contribution to data analysis. Pandas solved this problem. Using pandas, five typical steps in the processing and analysis of data was accomplished, regard less of the origin of data, prepare, manipulate, model and analyze. Python with pandas is used in a wide range of fields including academic and commercial domains including finance, economics, statistics, analytics etc.

**Matplotlib**

Matplotlib is one of the most popular python packages used for data visualization. It is a cross platform library for making 2D plots from data in arrays. Matplotlib is written in python and makes use of numpy the numerical mathematics extension of python. It can be used in python and Jupiter notebook and web application servers also. Matplotlib has a procedural interface named the pylab, which is designed to resemble matlab, a proprietary programming language developed by mathworks. Matplotlib was originally written by john d. hunter in 2003. The current stable version is 2.2.0 released in January 2018.

\*\*\*\*\*\*

## CHAPTER - 4

## UML DIAGRAMS

This chapter describes the design and implementation details of the proposed system. The design consists of UML diagrams such as a class diagram, collaboration diagram, use case diagram and sequence diagram.

UML stands for Unified Modeling Language. UML is a standardized general- purpose modeling language in the field of object oriented software engineering. The goal is for UML to become a common language for creating models of object- oriented computer software. In its current form, UML is comprised of two major components: A meta model and a notation. In the future, some form of method or process may also be added or associated with UML.

The Unified Modeling Language is a standard language for specifying, visualization, constructing and documenting the artifacts of the software system, as well as for other or business modeling non software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

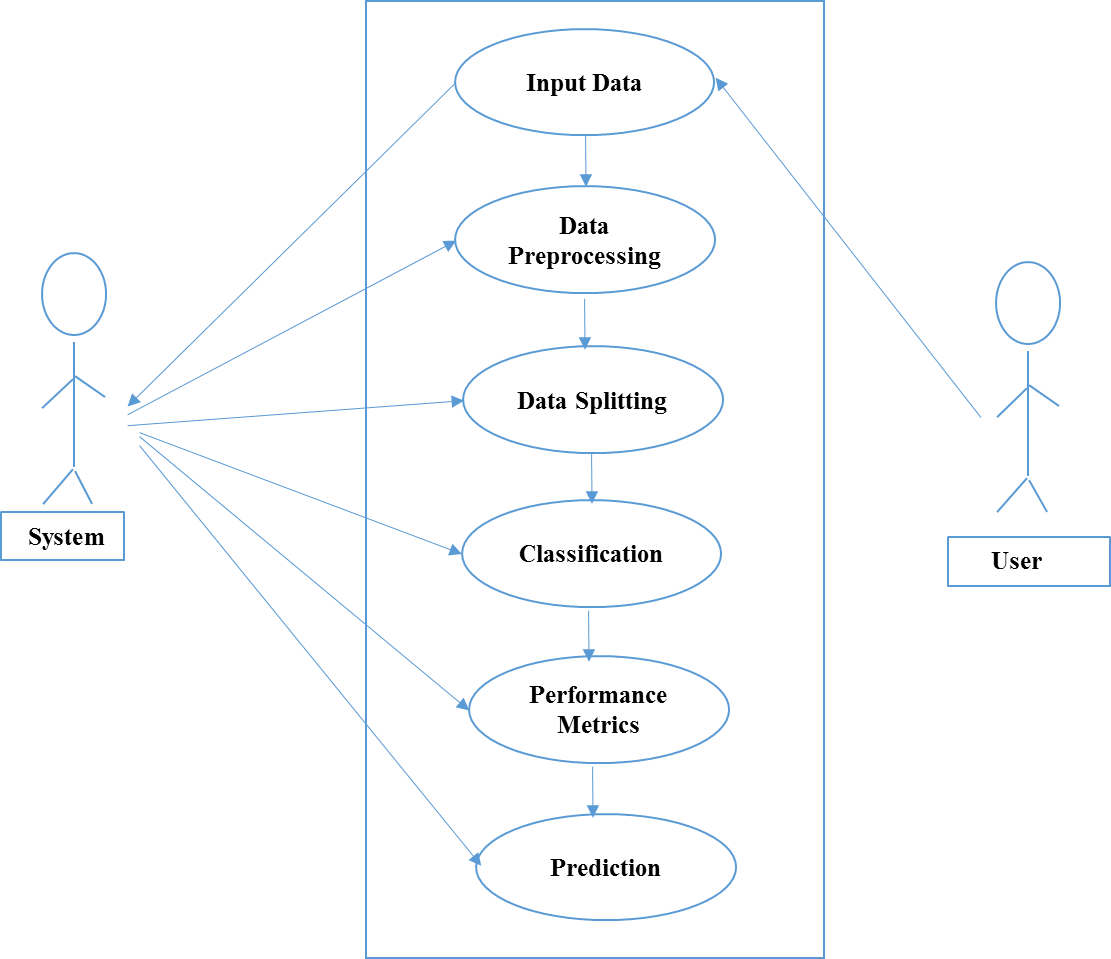
The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

* The primary goals in the design of the UML are as follows:
* Provide users a ready to use, expressive visual modeling language so that they can develop and exchange meaningful models.
* Provide extensibility and specialization mechanism to extend the core concepts.

The Unified modeling language allows the software engineer to express an analysis model using the modeling notation that is governed by set syntactic and semantic rules. A UML system is represented using five different views that described the system from a distinctly perspective. In general, the proto type of the project can be designed with the help of UML diagrams.

**4.1 Use case diagram**

A use case diagram in the UML is a type of diagram defined by and created from a use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals represented as use cases, and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



**Figure 4.1 Use case diagram of water quality prediction.**

The use case diagram which is represented in the figure 4.1 explains set of use cases their relationships. The flow of project is considered the following steps: first of all user upload test data and predict quality of water, after that it will process by using logistic regression algorithm to analyze the efficiency of the models and gives suitable results.

**4.2 Sequence diagram**

A sequence diagram in UML is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a message sequence chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

**Input Data**

**Preprocessing**

**Data splitting**

**Classification**

Select data

Missing value

DL and ML

Load data

Test and Train

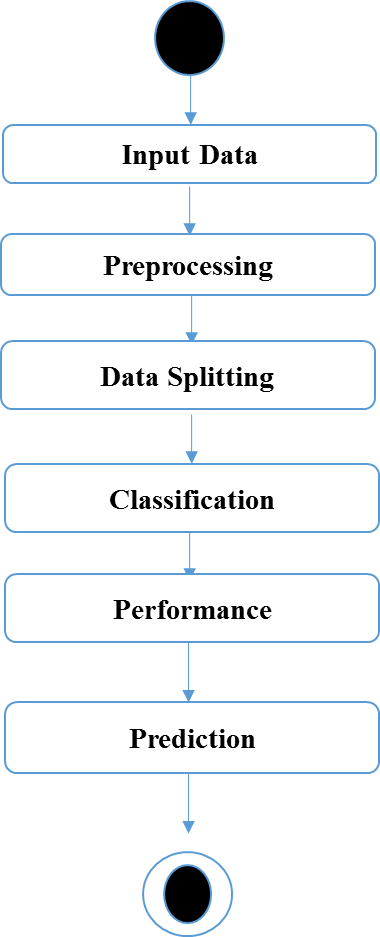
Performance

**Figure 4.2 Sequence diagram of water quality prediction**

The sequence diagram which is represented in the figure 4.2 explains set of sequences and their events in time sequence. The flow of project is considered the following steps: first of all user upload test data and predict quality of water, after that it will process by using logistic regression algorithm to analyze the efficiency of the models and gives suitable results.

**4.3 Activity diagram**

Activity diagrams are graphical representations of work flows of step wise activities and actions with support for choice, iteration and concurrency. In the UML activity diagram can be used to describe the business and operational behavior in step- by-step work flows of components in a system. An activity diagram shows the overall flow of control.



**Figure 4.3 Activity diagram of water quality prediction**

The activity diagram which is represented in the figure 4.3 explains set of activities and their actions in overall flow of control. The flow of project is considered the following steps: first of all user upload test data and predict quality of water, after that it will process by using logistic regression algorithm to analyze the efficiency of the models and gives suitable results.

**4.4 System design and implementation**

In the implementation phase, the result is evaluated according to the list of requirements that was created in the definition phase. It is also evaluated according to the designs. The purpose of the implementation phase is to deploy and enable operations of the new information system in the production environment.

**Data selection**

The input data was collected from dataset repository kaggle. The dataset is taken which contains the parameters of the water and the dataset contains the information about the water such as ph, Hardness, Solids. The dataset is in the format of .csv is fed into classification and regression algorithms.

**Data preprocessing**

Data pre-processing is the process of removing the unwanted data from the dataset. Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning. This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient. In this process, the null values such as missing values and nan values are replaced by 0 and data was cleaned of any abnormalities.

**Data splitting**

During the process, data is needed so that learning can take place. In addition to the data required for training, test data are needed to evaluate the performance of the algorithm in order to see how well it works. In this process, 80% of the dataset is to be the training data and the remaining 20% to be the testing data. Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.

One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, the data set is separated into a training set and testing set, most of the data is used for training the machine.

**Classification**

Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Classification is the task of predicting a discrete class label whereas regression is the task of predicting a continuous quantity. Before classification, the data is split into the test and train dataset.

* Most of data’s are used for training and smaller portion of the data’s are used for testing the model output.
* Training data is used for evaluate the model and testing data is used for predictive the model.
* After data splitting, classification algorithm is to be implemented. CNN, LSTM and GRU algorithms are implemented.

In this process, different classification algorithms such as CNN+LSTM, CNN+GRU are applied on the dataset and logistic regression is applied on the results obtained from the models.

**4.5 Model –1 CNN-LSTM**

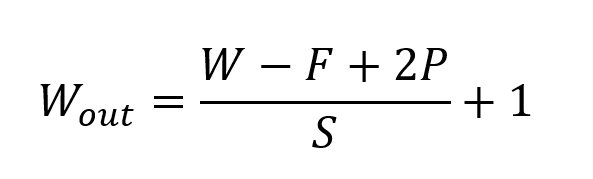
**CNN**

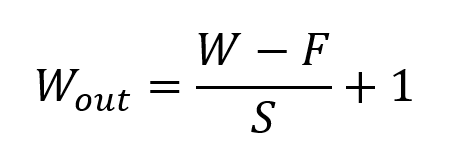
A convolutional neural network (CNN) is a subset of [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML). It is one of the various types of artificial [neural networks](https://www.techtarget.com/searchenterpriseai/definition/neural-network) which are used for different applications and data types. A CNN is a kind of network architecture for [deep learning](https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network) algorithms and is specifically used for classification and tasks that involve the processing of [pixel](https://www.techtarget.com/whatis/definition/pixel) data. CNN is used for identifying and recognizing objects. A CNN typically has three layers: a convolutional layer, a pooling layer, and a fully connected layer.

**Convolution layer**

The convolution layer is the core building block of the CNN. It carries the main portion of the network’s computational load. This layer performs a dot product between matrices, where one matrix is the set of learnable parameters otherwise known as a kernel, and the other matrix is the restricted portion of the receptive field

During the forward pass, the kernel slides across the height and width of the image-producing the image representation of that receptive region. The sliding size of the kernel is called a stride. The size of output volume can be determined by the following formula:



**Pooling layer**  The pooling layer replaces the output of the network at certain locations by deriving a summary statistic of the nearby outputs. This helps in reducing the spatial size of the representation, which decreases the required amount of computation and weights. The output volume can be determined by the following formula: 

There are several pooling functions such as the average of the rectangular neighborhood, l2 norm of the rectangular neighborhood, and a weighted average based on the distance from the central pixel. However, the most popular process is max pooling, which reports the maximum output from the neighborhood.

**Fully connected layer**

Neurons in this layer have full connectivity with all neurons in the preceding and succeeding layer as seen in regular FCNN. This is why it can be computed as usual by a matrix multiplication followed by a bias effect. The FC layer helps to map the representation between the input and the output.

**LSTM** Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

**Structure of LSTM**

LSTM has a chain structure that contains four neural networks and different memory blocks called **cells.** Information is retained by the cells and the memory manipulations are done by the**gates.** There are three gates: forget gate, input gate, output gate.

**Forget gate**

The information that is no longer useful in the cell state is removed with the forget gate. Two inputs x\_t (input at the particular time) and h\_t-1 (previous cell output) are fed to the gate and multiplied with weight matrices followed by the addition of bias. The resultant is passed through an activation function which gives a binary output. If for a particular cell state the output is 0, the piece of information is forgotten and for output 1, the information is retained for future use.

**Input gate**

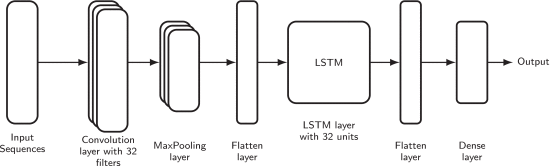
The addition of useful information to the cell state is done by the input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate using inputsh\_t-1 and x\_t. Then, a vector is created usingtanhfunction that gives an output from -1 to +1, which contains all the possible values from h\_t-1 and x\_t. At last, the values of the vector and the regulated values are multiplied to obtain the useful information.

**Output gate**

The task of extracting useful information from the current cell state to be presented as output is done by the output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter by the values to be remembered using inputsh\_t-1 and x\_t. At last, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.

**CNN–LSTM model**

In CNN–LSTM model the first part is CNN to which the data is fed, and it extracts the features there is a dropout layer after the convolutional layer(conv1d) and pooling layer( maxpooling 1d). In this model we use conv1d with 32 filters and a kernel size of 3n and ReLU is used as the activation function. A pooling layer maxpooling1d follows this with a pooling size of 2. CNN with maxpooling is used to reduce the length of the input sequence which is fed into the LSTM.



**Figure 4.5 CNN-LSTM model diagram**

The CNN layer extracts all the features and then feeds them to the LSTM layer. The output of the CNN layer is fed to the LSTM layer after pooling through the flatten layer. Then, the LSTM layer outputs its output to the dense layer through the flatten layer and finally the prediction is output at the dense layer.

**4.6 Model – 2 CNN–GRU**

**GRU**

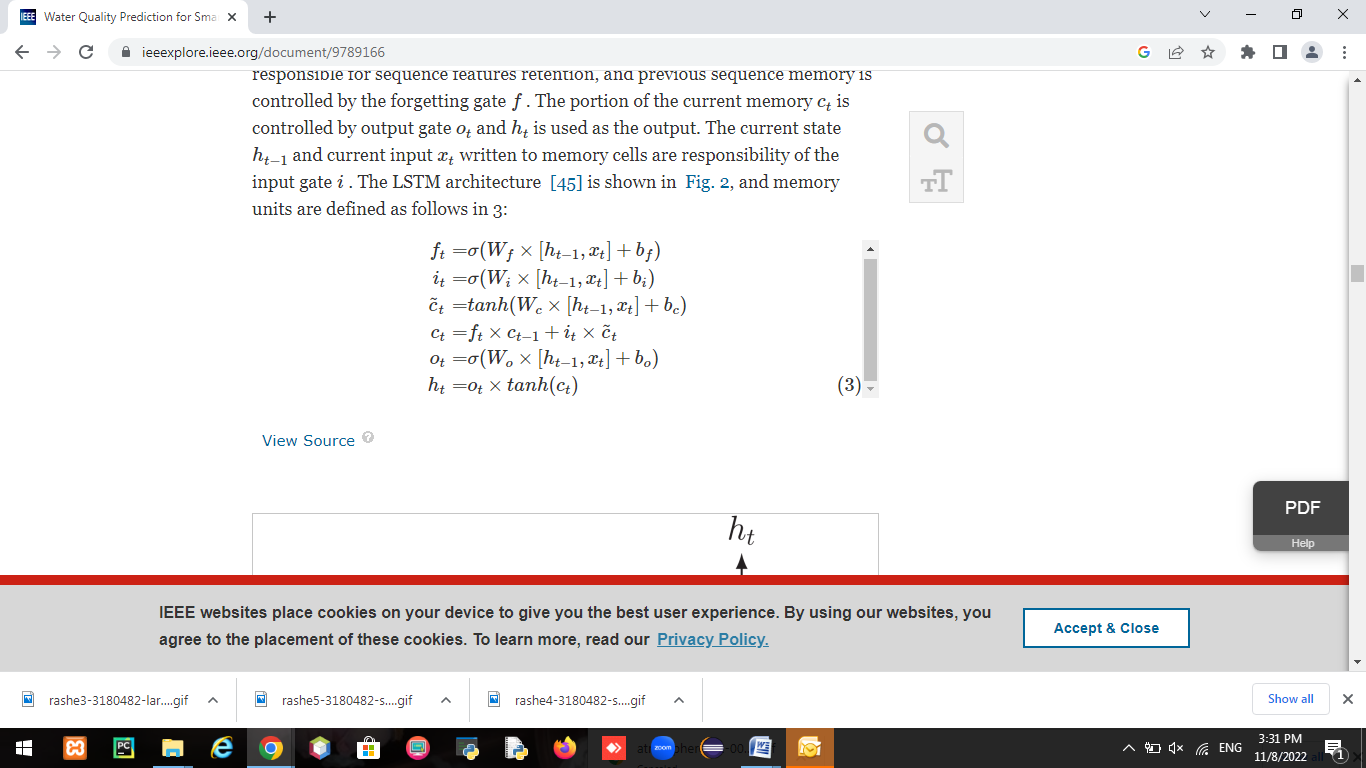
To solve the Vanishing-Exploding gradients problem often encountered during the operation of a basic Recurrent Neural Network, many variations were developed. One of the most famous variations is the Long Short Term Memory Network (LSTM). One of the lesser-known but equally effective variations is the Gated Recurrent Unit Network (GRU).

Unlike LSTM, it consists of only three gates and does not maintain an Internal Cell State. The information which is stored in the Internal Cell State in an LSTM recurrent unit is incorporated into the hidden state of the Gated Recurrent Unit. This collective information is passed onto the next Gated Recurrent Unit. The different gates of a GRU are as described below:

**Update gate (z):**  It determines how much of the past knowledge needs to be passed along into the future. It is analogous to the Output Gate in an LSTM recurrent unit.

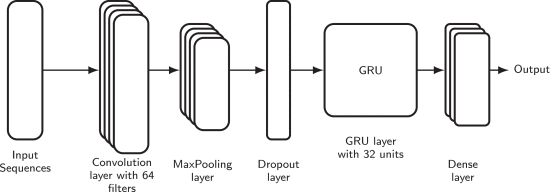
**Reset gate (r):**  It determines how much of the past knowledge to forget. It is analogous to the combination of the Input Gate and the Forget Gate in an LSTM recurrent unit.

**Current memory gate ( ):**  It is often over looked during a typical discussion on gated recurrent unit network. It is incorporated into the reset gate just like the input modulation gate is a sub-part of the input gate and is used to introduce some non-linearity into the input and to also make the input zero-mean. Another reason to make it a sub-part of the reset gate is to reduce the effect that previous information has on the current information that is being passed into the future. The memory units are defined as follows:



There are no additional memory cells to store information; GRU can control information inside the unit. The update gate decides whether to pass the previous output ht−1 to the next cell. The reset gate reads the input sequences when the gate is set to zero and forgets the previously calculated state. As a result, GRU has fewer tensor operations than LSTM and runs typically faster than LSTM.

**CNN–GRU model**  The CNN-GRU models use Conv1D with 64 filters, and a kernel size of 5 and ReLU is used as the activation function. A pooling layer MaxPooling1D follows this with a pooling size of 4.



**Figure 4.6 CNN-GRU model diagram**

The input data is given to the CNN layer and CNN layer extracts all the features by using max pooling layer and then feds to the GRU layer, result is produced at the dense layer.

**4.7 Framework design**

Frame design is transition from a user oriented document to programmers or data base personnel. The planning may be a solution, the way to approach the creation of a replacement system. This is often composed of several steps. It provides the understanding and procedural details necessary for implementing the system recommended within the feasibility study.

Designing goes through logical and physical stages of development, logical design reviews this physical system prepares input and output specification details of implementation plan and prepares a logical design rehearse.

The input design is that the link between the knowledge system and therefore the user. It comprises the developing specification and procedures for data preparation and people steps are necessary to place transaction data in to a usable form for processing are often achieved. The planning of input focuses on controlling the quantity of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the method simple. The input is meant in such how in order that it provides security and simple use with retaining the privacy.

Input design considered the subsequent things:

* What data should tend as input?
* How the information should be arranged or should be coded?

Dialog to guide the operating personnel in providing input and methods for preparing input validations and steps to follow when error occurs.

A quality output is one that which meets the wants of the top user and presents the knowledge clearly in any system of results of processing are communicated to the users and to other system through outputs. In output design it's determine how the knowledge is to be displaced for immediate needs and also for text output. The direct source information is given to the user. The efficient and the intelligent output design which improves the system relationship to assist user for decision making. Select methods for presenting information.

* The creation of document, report and other formats that contain information produced about the system.
* The output sort of a data system should accomplish one or more of the subsequent objectives.
* Convey information about past activities to current status and projections of the future.
* Signal important events, opportunities, problems or warnings.
* Trigger an action
* Confirm an action.

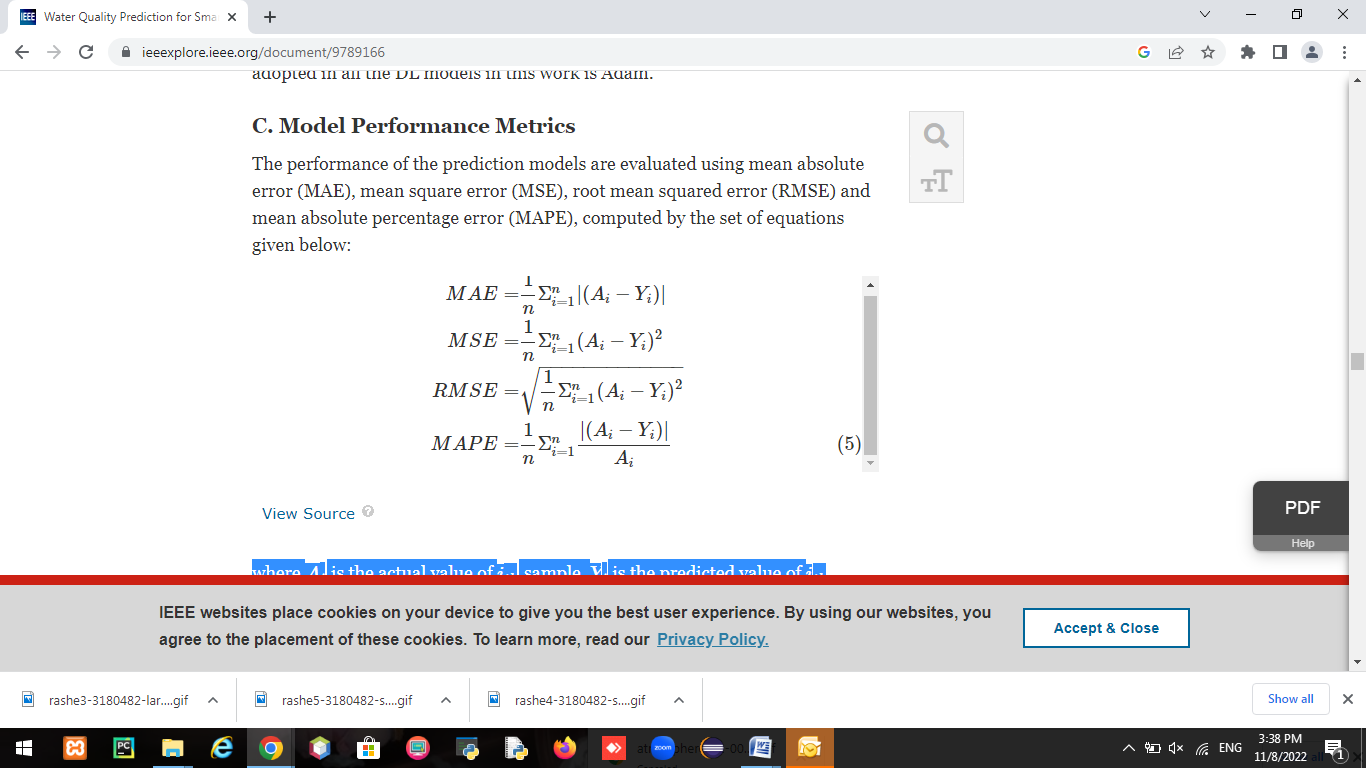
Designing computer output should proceed in an organized, well thought out manner the proper output must be developed while ensuring that every output element is meant in order that people to find the system which can be used easily and effectively. When analysis designs computer output, they ought to identify the precise output that's needed to satisfy the wants.

**Methodology**  Architecture diagram may be a diagram of a system during which the principle parts or functions are represented by blocks connected by lines that shows the relationships of the blocks. The diagram is usually used for a better level less detailed description aimed more at understanding the general concepts and fewer at understanding small print of the implementation. Predictive modeling is to built a model that capable of creating predictions. The method includes a machine learning algorithm that learns certain properties from a training dataset so as to form those predictions.

The dataset is taken as input for sample data which consists of attributes. Within the data preprocessing step the null values are removed then feature normalization which is completed as a neighbor hood of knowledge preparation. The info is converted in to supporting format and classification takes place. The specified results of the proposed system are generated.

**Model performance metrics**

The performance of the prediction models are evaluated using mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE) and mean absolute percentage error (MAPE), computed by the set of equations given below:



where Ai is the actual value of ith sample, Yi is the predicted value of ith sample and n is the number of samples.

**Logistic regression** Logistic regression is one of the most popular machine learning algorithms, which comes under the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either yes or no, 0 or 1, true or false, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic regression is much similar to the linear regression except that how they are used. Linear regression is used for solving regression problems, whereas logistic regression is used for solving the classification problems.Logistic regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.Logistic regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

**Accuracy check**

Accuracy is the number of correctly predicted data points out of all the data points. More formally, the most important thing in the implementation is to understand the problem and to know the purpose of the problem.

Therefore before starting the implementation user need to understand the problem because the good result depends on the better understanding of the problem. User creates a machine learning system called model and this model is created by providing training. To train a model, user need data, hence implementation starts by collecting data. After collecting data is pre processed and model is implementing to evaluate and check the accuracy.

\*\*\*\*

## CHAPTER – 5

## SYSTEM TESTING

Software testing is the process used to help identify the correctness, completeness, security, and quality of developed computer software. Testing is a process of technical investigation, performed on behalf of stake holders, that is intended to reveal quality related information about the product concerning the context in which it is intended to operate. This includes but is not limited to, the process of executing a program or application with the intent of finding errors. Quality is not absolute it is value to some person. With that in mind, testing can never completely establish the correctness of arbitrary computer software. Testing furnishes a criticism or comparison that compares the state and behavior of the product against a specification. An important point is that software testing should be distinguished from the separate discipline of Software Quality Assurance (SQA), which encompasses all business process areas, not just testing.

The software system needs to be checked for its intended behavior and direction of progress at each development stage to avoid duplication of efforts, time and cost over runs and to assure completion of the system within a stipulated time. System testing and quality assurance come to aid for checking the system. It includes product level quality (testing) and process level quality.

Testing is the process or activity that checks the functionality and correctness of software according to specified user requirements in order to improve the quality and reliability of system. This includes but is not limited to, the process of executing a program or application with the intent of finding errors. It is an expensive time consuming and critical approach in system development which requires per planning of overall testing process. A successful test is one that finds the errors. It executes the program with explicit intention of finding error i.e., making the program fail. The data fields predefined processes and successive processes must be considered for testing.

**5.1 Types of testing** Testing is the process of executing a program with the aim of finding errors. To make software perform well it should be error free. If testing is done successfully it will remove all the errors from the software. There are different types of testing presents according to the flow of execution. The following is some different types of testing as mentioned below.

**Unit testing**

The unit testing is the operation in which each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

Each module can be tested using the following two strategies:

* Black box testing
* White box testing

**Black box testing**

Black box testing treats the software as a "black box," examining functionality without any knowledge of internal implementation, without seeing the source code. The testers are only aware of what the software is supposed to do, not how it does it. Black box testing methods include: [equivalence partitioning](https://en.wikipedia.org/wiki/Equivalence_partitioning), [boundary value analysis](https://en.wikipedia.org/wiki/Boundary_value_analysis), [all-pairs testing](https://en.wikipedia.org/wiki/All-pairs_testing), [state transition](https://en.wikipedia.org/wiki/State_transition_table) [tables,](https://en.wikipedia.org/wiki/State_transition_table) table testing, [fuzz testing](https://en.wikipedia.org/wiki/Fuzz_testing), [model-based testing](https://en.wikipedia.org/wiki/Model-based_testing), [use case](https://en.wikipedia.org/wiki/Use_case) testing, [exploratory testing](https://en.wikipedia.org/wiki/Exploratory_testing), and specification-based testing. Specification-based testing aims to test the functionality of software according to the applicable requirements. This level of testing usually requires thorough [test cases](https://en.wikipedia.org/wiki/Test_case) to be provided to the tester, who then can simply verify that for a given input, the output value either "is" or "is not" the same as the expected value specified in the test case. Test cases are built around specifications and requirements, i.e., what the application is supposed to do. It uses external descriptions of the software, including specifications, requirements, and designs to derive test cases. These tests can be [functional](https://en.wikipedia.org/wiki/Functional_testing) or [non-functional,](https://en.wikipedia.org/wiki/Non-functional_testing) though usually functional. Specification-based testing may be necessary to assure correct functionality, but it is insufficient to guard against complex or high-risk situations.

One advantage of the black box technique is that no programming knowledge is required. Whatever biases the programmers may have had, the tester likely has a different set and may emphasize different areas of functionality. On the other hand, black-box testing has been said to be "like a walk in a dark labyrinth without a flashlight. Because they do not examine the source code, there are situations when a tester writes many test cases to check something that could have been tested by only one test case or leaves some parts of the program untested.

**Types of black box testing**

There are different types of black box testing:

**Functional testing** – This black box testing type is related to functional requirements of a system; it is done by software testers.

**Non-Functional testing** – This type of black box testing is not related to testing of a specific functionality, but non-functional requirements such as performance, scalability, usability.

**Regression testing** – Regression testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

**White box testing**

White box testing verifies the internal structures or workings of a program, as opposed to the functionality exposed to the end-user. In white-box testing, an internal perspective of the system as well as programming skills, are used to design test cases. The tester chooses inputs to exercise paths through the code and determine the appropriate outputs.

While white-box testing can be applied at the [unit](https://en.wikipedia.org/wiki/Unit_testing), [integration](https://en.wikipedia.org/wiki/Integration_testing), and [system](https://en.wikipedia.org/wiki/System_testing) levels of the software testing process, it is usually done at the unit level. It can test paths within a unit, paths between units during integration, and between subsystems during a system–level test. Though this method of test design can uncover many errors or problems, it might not detect unimplemented parts of the specification or missing requirements.

**5.2 Integration testing**

Integration testing is defined as the testing of combined parts of an application to determine if they function correctly. Integration testing can be done in two ways: bottom-up integration testing and top down integration testing. Bottom up integration testing begins with unit testing, followed by tests of progressively higher level combinations of units called modules or builds. In top down integration testing, the highest level modules are tested first and p regressively, lower level modules are tested after.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check the components or software applications.

**Test strategy and approach**

Field testing will be performed manually and functional tests will be written in detail.

**Test objectives**

These are the test objectives

* All field entries must work properly.
* Pages must be activated from the identified link.
* The entry screen, messages and responses must not be delayed.

**Features to be tested**

* Verify that the entries are of the correct format.
* No duplicate entries should be allowed.
* All links should take the user to the correct page.

**Validation testing**

After the culmination of black box testing, software is completed assembly as a package, interfacing errors have been uncovered and corrected and final series of software validation tests begin validation testing can be defined as many, but a single definition is that validation succeeds when the software functions in a manner that can be reasonably expected by the customer

**5.3 Acceptance testing**

Acceptance testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements. As it is conducted by the quality assurance team who will gauge whether the application meets the intended specifications and satisfies the client requirement. The quality assurance team will have a set of pre written scenarios and test cases that will be used to test the application.

Acceptance testing is the process of trying to discover every conceivable fault or weakness in a work product. It tests a configuration to ensure known and predictable results. Acceptance testing process is used to discover whether the proposed system meets the requirements as expected before or not.

Acceptance testing are not only intended to point out simple spelling mistakes, cosmetic errors, or interface gaps, but also to point out any bugs in the application that will result in system crashes or major errors in the application. By performing acceptance tests on an application, the testing team will reduce how the application will perform in production. There are also legal and contractual requirements for acceptance of the system.

\*\*\*\*\*\*

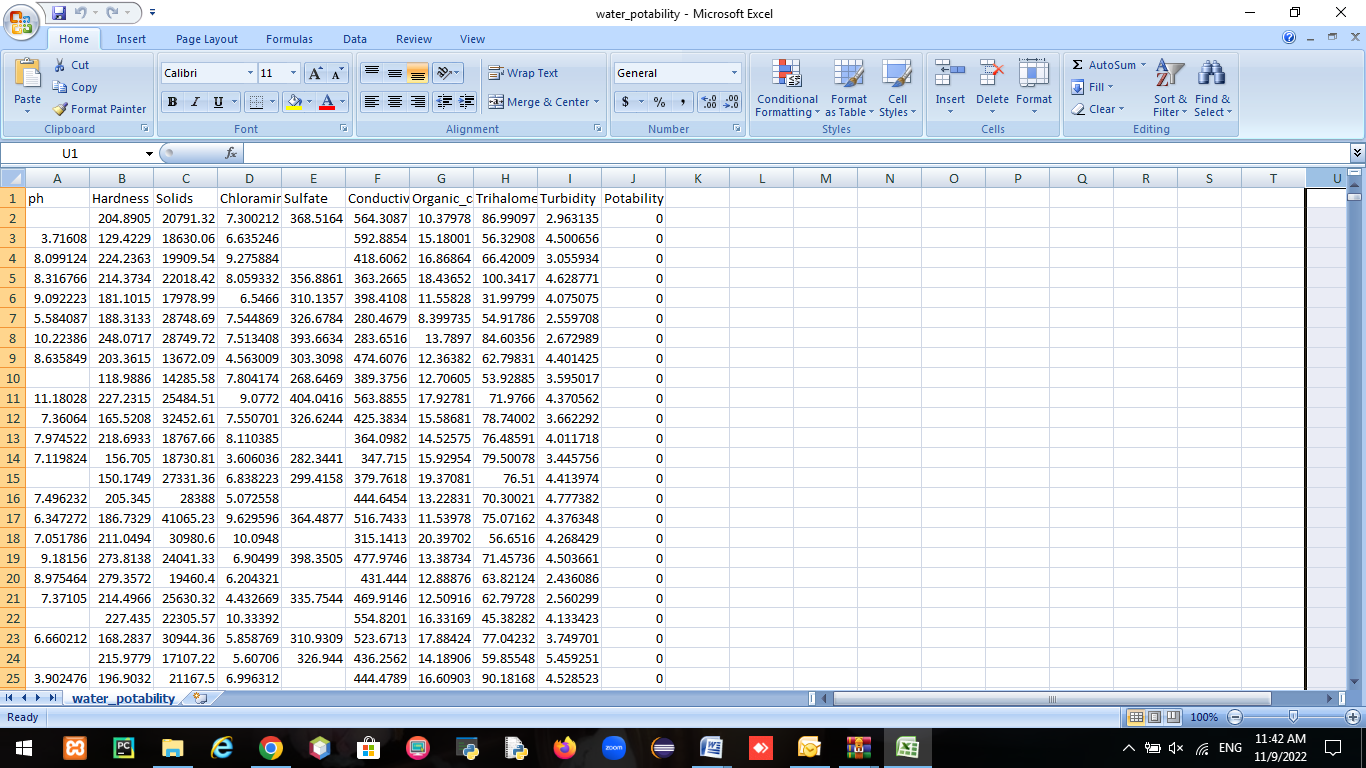
## CHAPTER – 6

## RESULT ANALYSIS

The results of the proposed system are explained below. This contains all the results of the proposed system in screenshot format. It explains how to upload the CSV file pre processing the data, data analysis, model implementation and prediction the results in terms of precision accuracy is in the form of a screenshot.

**6.1 CSV file**

CSV file (comma-separated values) is a text file that stores data in the form of columns, separated by commas and rows are distinguished by line breaks. A CSV file is a special type of file that can created or edited in excel. Instead of storing information in columns, CSV files store data separated by commas. When text and numbers are saved in a CSV file, it is easy to move them from one program to another.



**Figure 6.1 CSV file**

The data as shown in figure 6.1 CSV file was downloaded from kaggle. Each row represents a parameter values, each column contains parameter attributes.

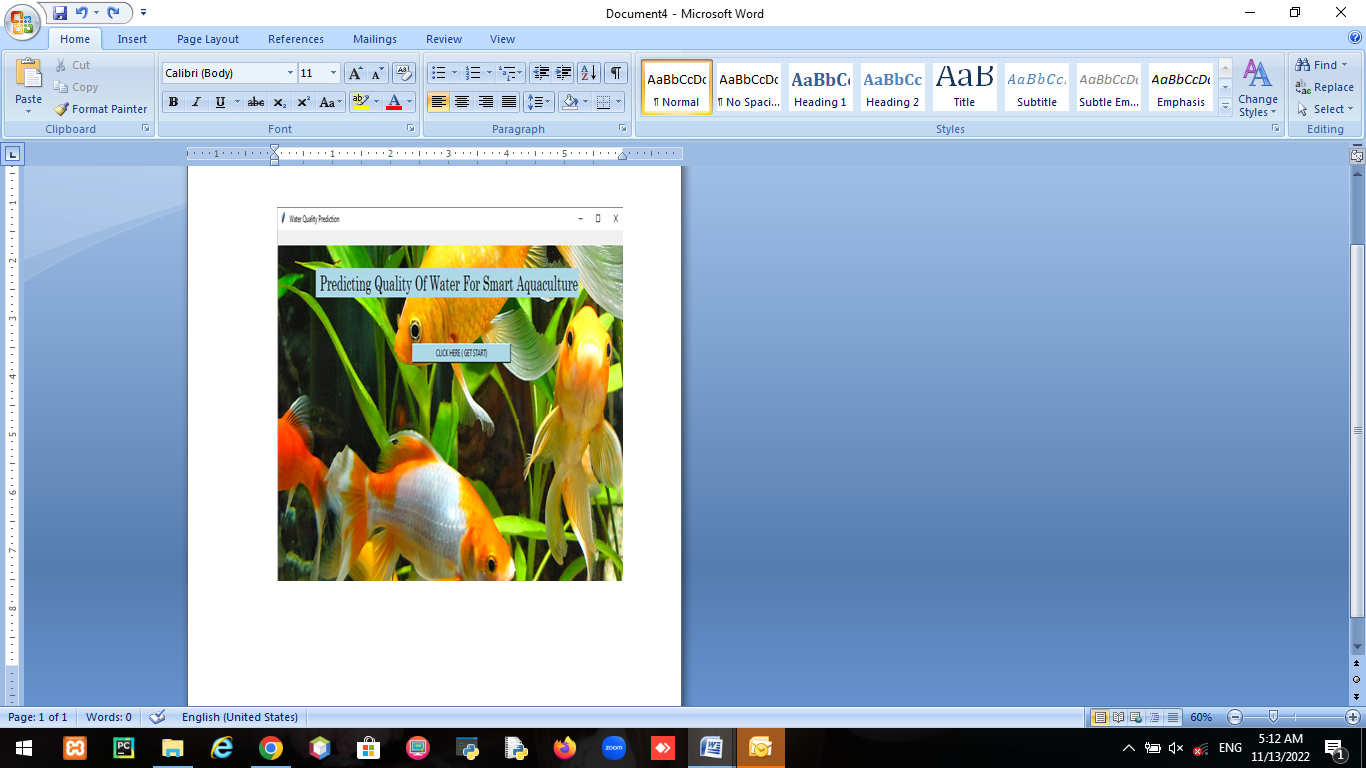
The collected CSV file has 10 columns and 3277 rows of attributes data. The attributes are explained below:

* **PH value –** PH is an important parameter in evaluating the acid- base balance of water. It is the indication of acidic or alkaline condition of the water.
* **Hardness –** Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.
* **Solids –** Water has a ability to dissolve a wide range of inorganic and some organic minerals or salts such as calcium, sulfates etc. The water with high TDS value indicates that water is highly mineralized.
* **Chloramines –** Chlorine and chloramines are the major disinfectants present in water. Chloramines are mostly formed when ammonia is added to chlorine.
* **Sulfate –** Sulfates are naturally occurring substances that are found in minerals and soil.
* **Conductivity –** Pure water is not a good conductor of electric current. Generally the amount of dissolved solids in water determines the electrical conductivity of water.
* **Organic carbon –** Total organic carbon is the measure of the total amount of carbon in organic compounds in water.
* **Trihalomethanes –** THMs are chemicals which may be found in water treated with chlorine.
* **Turbidity –** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is the measure of light emitting properties of water.
* **Potability –** States whether the water is safe or not. It contains only two values 0 or 1. 0 indicates not safe and 1 indicates safe nature of water.
* **Temperature -** Water temperature is **a** measure of the kinetic energy ofwater and is expressed in degrees Fahrenheit (F) or Celsius (C). Water temperature varies according to season, depth and, in some cases, time of day.

These parameters play an important role in predicting the quality of water to maintain and manage aquaculture. Water quality plays an important role to have profitable aquaculture.

## 6.2 Home page

This is the home page of the project that shows the prediction of water quality for aquaculture based on the parameters in the dataset.

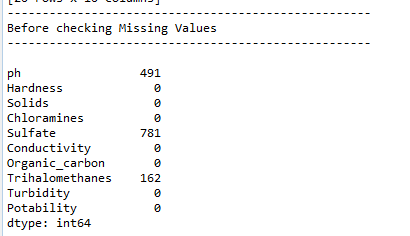


**Figure 6.2 Home page**

The above figure 6.2 shows the home page of the project. This page provides an option to start prediction of water quality for aquaculture. It performs the prediction of water quality based on the parameters in the dataset by applying the deep learning algorithms.

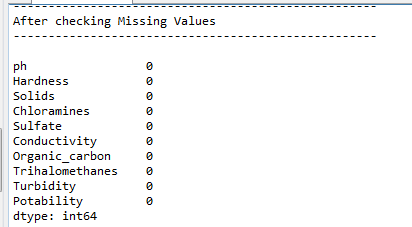
**6.3 Processing**

Before processing the input data was collected from kaggle website. The system checks for the null values in the rows and those values are replaced by ‘0’, so that the models results accurately.



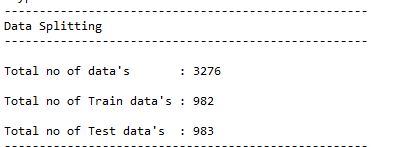
**Figure 6.3.1 Processing before checking null values**

The above figure 6.3.1 shows the number of null values in the dataset and those null values are replaced with ‘0’.



**Figure 6.3.2 Processing after checking null values**

Here the above figure 6.3.2 represents the analysis of data after checking for null values those null values are replaced with 0 so that there are no null value rows in dataset. Further it processes the data set into two that is train and test data sets.

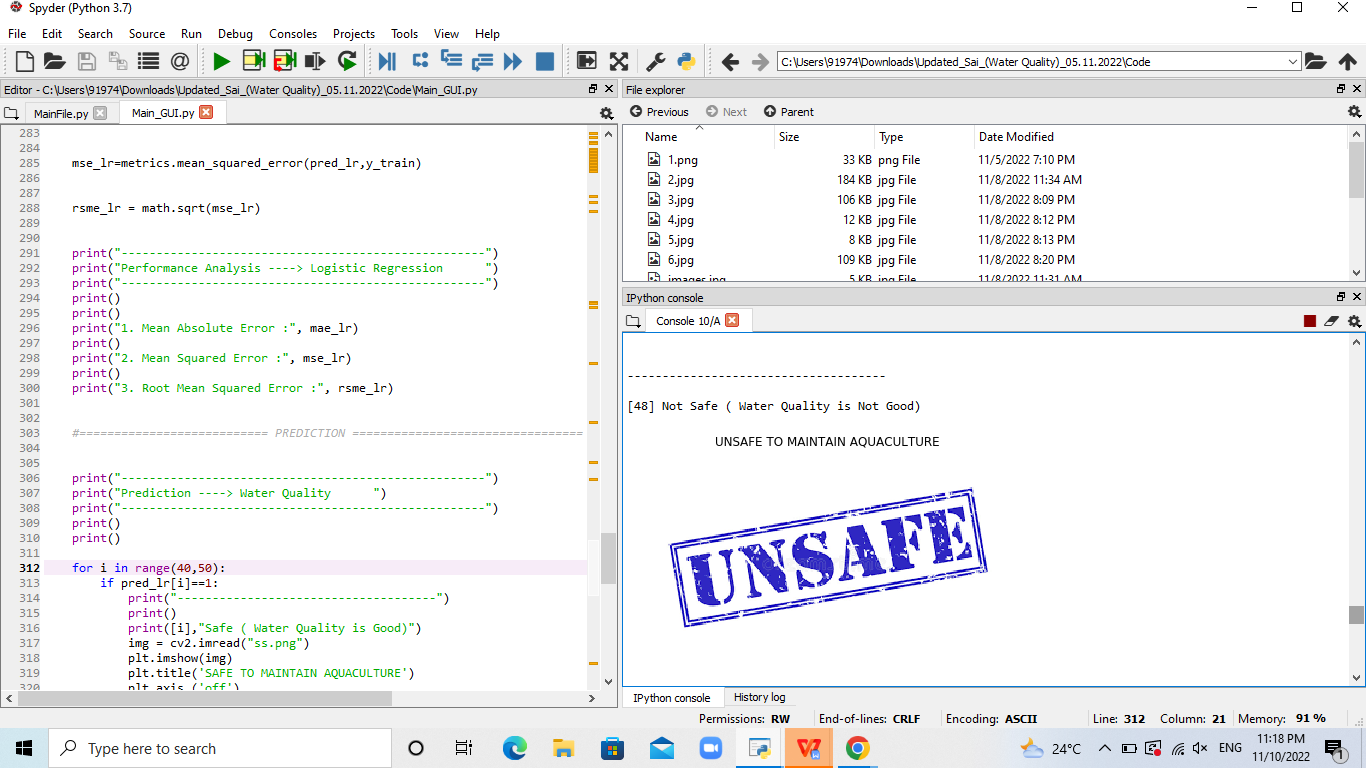


**Figure 6.3.3 Process of data splitting**

Here the above figure 6.3.3 represents the splitting of data into two that is for test data and train data. The train data is given to the system to train the model. Once the model is trained the test data is used to test the model performance.

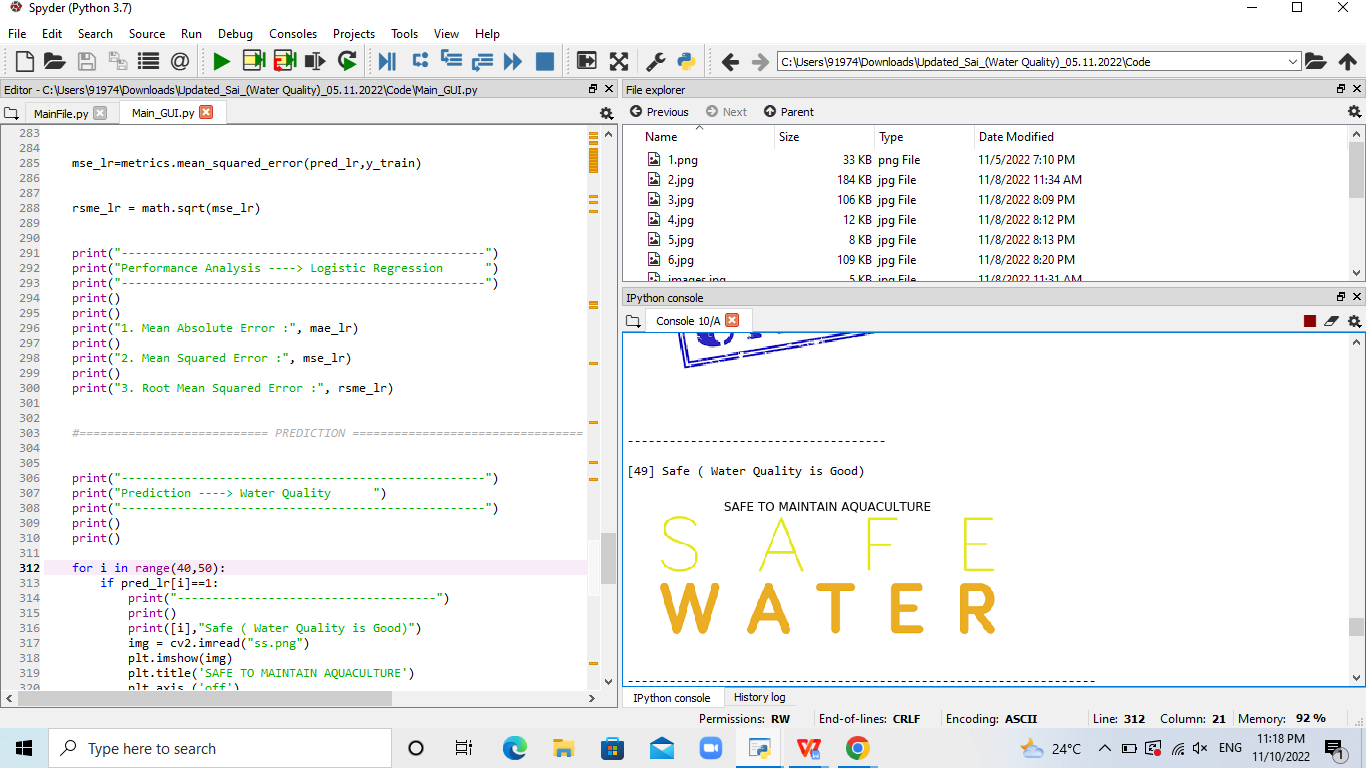
**6.4 Prediction**

The prediction page shows the result of dataset that is uploaded for predicting the quality of water to maintain and manage aquaculture, after prediction logistic regression algorithm provides the better performance of the model.

****

**Figure 6.4.1 Prediction of water quality for aquaculture as unsafe**

Here the above figure 6.4.1 represents that the water quality is not good to maintain aquaculture. Based upon the parameters the quality of the water is predicted so that it helps in maintaining and managing aquaculture.



**Figure 6.4.2 Prediction of water quality for aquaculture as safe.**

Here the above figure 6.4.2 represents that the water is safe to maintain and manage aquaculture.

From the analysis it is concluded that the parameters like ph, solids, hardness, chloramines, sulfate and conductivity of water plays an important role in predicting the water quality.

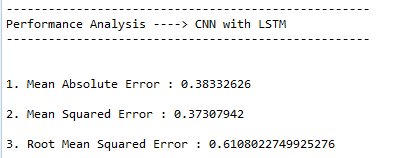
Table 6.4 Maximum minimum values of the parameters.

|  |  |
| --- | --- |
| **Parameters** | **Min- Max Value** |
| PH | 7.0 – 9.0 |
| Hardness  (mg/l) | 100 - 400 |
| Temperature  (°F) | 50 - 90 |
| Solids  (mg/l) | 700 – 1000 |
| Chloramines (mg/l) | 2.5- 4 |
| Potability | 0 - 1 |

The above table determines the minimum and maximum values of the parameters that are to be noted for predicting the quality of water for aquaculture.

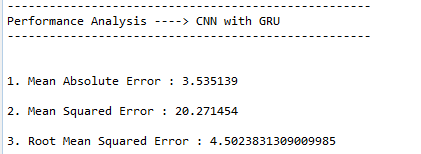
**6.5 Performance**

To predict the quality of water two models are used. The first one is CNN with LSTM and another one CNN with GRU. The performance of these models is calculated by mean absolute error, mean square error and root mean squared error.



**Figure 6.5.1 Performance analysis of CNN with LSTM**

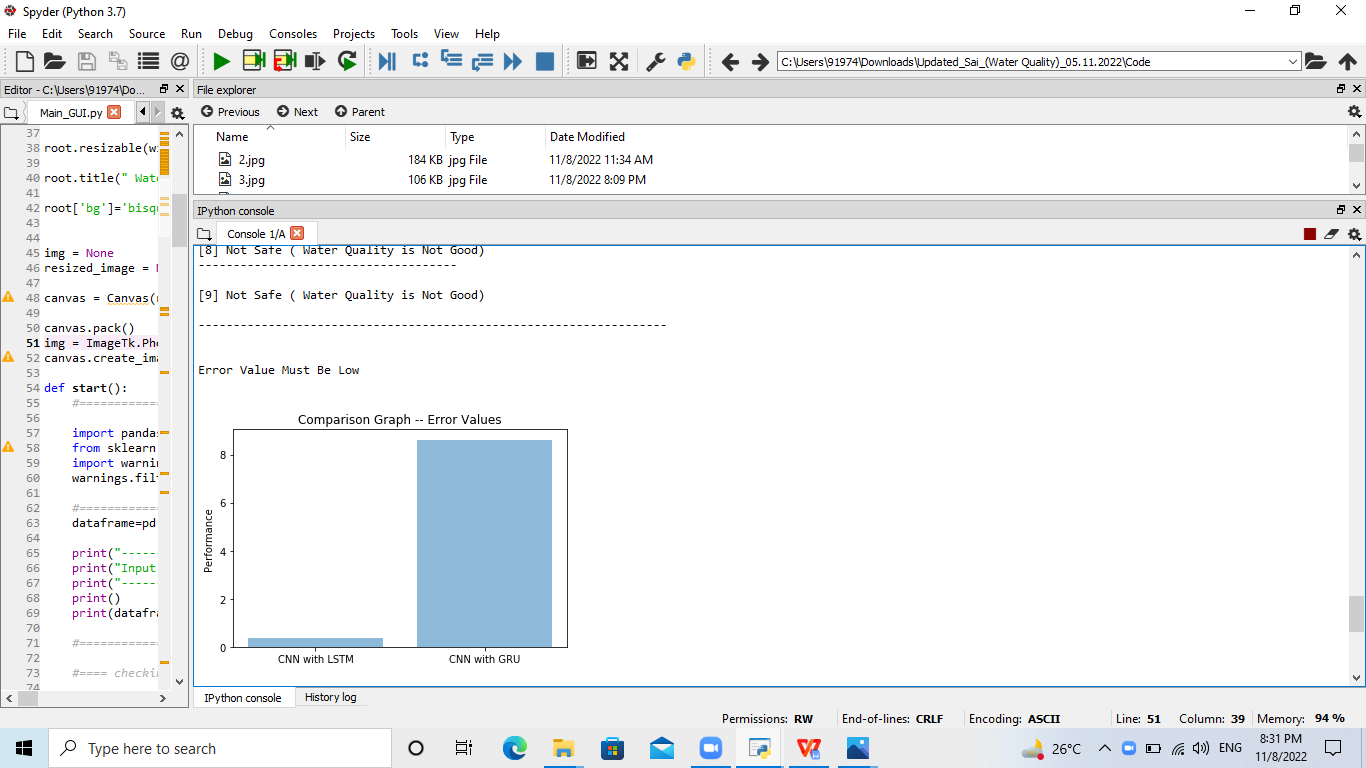
Here the above figure 6.5.1 represents the performance of the model CNN with LSTM. The mean absolute error, mean square error and root mean squared error are calculated on the CNN with LSTM model, to determine the accuracy of the model.

 **Figure 6.5.2 Performance analysis of CNN with GRU**

Here the above figure 6.5.2 represents the performance of the model CNN with GRU. The performance is calculated by handling the metric errors.

**6.6 Accuracy**

Accuracy is measured from training data and testing data. Based on the accuracy the performance of algorithms can be determined. Accuracy is one metric for evaluating classification models. Informally accuracy is the fraction of predictions our model got right. Formally accuracy is the number of correctly predicted data points out of all the data points.



**Figure 6.6 Accuracy of the models.**

Here the above figure 6.6 accuracy of the models, comparison graph represents the accuracy of the models. By using logistic regression algorithm the comparison of CNN-LSTM and CNN-GRU was done, CNN-LSTM model has low error rate than CNN-GRU so it is concluded that the model CNN-LSTM is more accurate.

\*\*\*\*\*

## CHAPTER - 7

## CONCLUSIONS & FUTURE WORK

Deep learning is the most popular learning algorithm for predicting the future. The paper presented the various deep learning algorithms for predicting the quality of the water for smart aquaculture on the basis of ph, hardness, solids, chloramines, sulfate, conductivity and turbidity. Experiments were conducted on dataset and it has been collected and it has been established then the CNN-LSTM model gives highest prediction accuracy.

The process of prediction starts from collecting data and processing of data, imputation of missing values, experimental analysis of data set and then model building to an evaluation of model and testing on test data. By combining ph, sulfate along with other parameters like solids and hardness, water quality prediction for a certain considered data set can be made. Here the two models CNN-LSTM and CNN-GRU were implemented to predict the quality of water. Logistic regression is applied on the results obtained by the models. Results reveal that CNN-LSTM model is the best classifier when all parameters are combined. This will not only help farmers in maintaining the aquaculture but also bridge the gap between technology and farming.

When apply logistic regression, the accuracy of models was determined. The output which has been shown in figure is currently an application but future work would be implementing different hybrid models, building an application where the user can calculate the parameters of water and can upload it in app, so that they can predict the results and converting the whole system in their regional language.

\*\*\*\*\*

## REFERENCES

[1] World Population Projected to Reach 9.8 Billion in 2050, and 11.2 Billion in 2100, United Nations, New York, NY, USA, Jun. 2017. [Online]. Available: https://www.un.org/development/desa/en/news/population/worldpopulation-prospects-2017.html.

[2] D. Li and C. Li, ‘‘Intelligent aquaculture,’’ J. World Aquaculture Soc., vol. 51, no. 4, pp. 808–814, Aug. 2020.

[3] Z. Hu, R. Li, X. Xia, C. Yu, X. Fan, and Y. Zhao, ‘‘A method overview in smart aquaculture,’’ Environ. Monitor. Assessment, vol. 192, no. 8, pp. 1–25, Jul. 2020.

[4] S. S. Ahuja, ‘‘Monitoring water quality, pollution assessment, and remediation to assure sustainability,’’ in Monitoring Water Quality, S. Ahuja, Ed. Amsterdam, The Netherlands: Elsevier, 2013, pp. 1–18.

[5] C. E. Boyd and C. S. Tucker, Pond Aquaculture Water Quality Management. New York, NY, USA: Springer, 1998.

[6] J. A. Martos-Sitcha, J. M. Mancera, P. Prunet, and L. J. Magnoni, ‘‘Editorial: Welfare and stressors in fish: Challenges facing aquaculture,’’ Frontiers Physiol., vol. 11, p. 162, Feb. 2020.

[7] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, ‘‘ARIMA models to predict next-day electricity prices,’’ IEEE Trans. Power Syst., vol. 18, no. 3, pp. 1014–1020, Aug. 2003.

[8] T. Arcomano, I. Szunyogh, J. Pathak, A. Wikner, B. R. Hunt, and E. Ott, ‘‘A machine learning-based global atmospheric forecast model,’’ Geophys. Res. Lett., vol. 47, no. 9, 2020, Art. no. e2020GL087776.

\*\*\*\*\*\*