WATER QUALITY PREDICTION USING MACHINE LEARNING

Mini Project Report

Submitted by

Vaisakh R Menon

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Computer Applications
Of
A P J Abdul Kalam Technological University

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®
ANGAMALY-683577, ERNAKULAM(DIST)
MARCH 2022

DECLARATION

I, **Vaisakh R Menon**, hereby declare that the report of this project work, submitted to the Department of Computer Applications, Federal Institute of Science and Technology (**FISAT**), Angamaly in partial fulfillment of the award of the degree of Master of Computer Application is an authentic record of our original work.

The report has not been submitted for the award of any degree of this university or any other university.

Date : 04-03-2022

Place: Angamaly

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®

ANGAMALY, ERNAKULAM-683577

DEPARTMENT OF COMPUTER APPLICATIONS

CERTIFICATE

This is to certify that the project report titled **"Water Quality Prediction Using Machine Learning"** submitted by **Vaisakh R Menon** towards partial fulfillment of the requirements for the award of the degree of Master of Computer Applications is a record of bonafide work carried out by them during the year 2022.

Project Guide **Head of the Department**

Submitted for the viva-voice held on ..... at .....

Examiner1 :

Examiner2 :

ACKNOWLEDGEMENT

Gratitude is a feeling which is more eloquent than words, more silent than silence. To complete this project work I needed the direction, assistance and cooperation of various individuals, which is received in abundance with the grace of God.

I hereby express our deep sense of gratitude to **Dr. Manoge George**, Principal of FISAT and **Dr. C Sheela**, Vice principal of FISAT, for allowing us to utilize all the facilities of the college.

My sincere thanks to **Dr. Deepa Mary Mathew**, Head of the department of Computer Applications FISAT and scrum master and our Internal guide for this project **Ms.Anju L and Dr. Sujesh P Lal** for giving valuable guidance, constructive suggestions and comment during my project work. I also express my boundless gratitude to all the lab faculty members for their guidance.

Finally I wish to express a whole heart-ed thanks to my parents, friends and well-wishers who extended their help in one way or other in preparation of my project. Besides all, I thank GOD for everything.

ABSTRACT

Water is used for various purposes and it has a strong impact on public health and the environment. Drinking contaminated water can cause many diseases. Even some of the packaged water that is available does not have the appropriate mineral content which in turn leads to an adverse health effect. Many people rely upon the water from river, lake which is much prone to pollution. They contain more nitrate contaminant which causes more BOD i.e. there will be less oxygen content in the water, which is not healthy for drinking. The proposed system is to check whether the given water sample is eligible for drinking by creating an application for which Machine Learning is used by taking some of the basic parameters of the water sample of tested data.

Contents

Chapter 1

INTRODUCTION

In India, many people depend on rivers and lake water for drinking and various purposes. Considering this condition, it is required to know whether the source of water is eligible for use or is contaminated. Due to many factors contaminated water causes many problems such as diseases like dysentery, cholera, diarrhoea, etc. In this paper, an old data set is taken to predict current water quality to check whether it is eligible for drinking and other purposes. Natural water resources like groundwater and surface water have always been the cheapest and most widely available resources of freshwater. These resources are also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. The effects of water quality deterioration are far-reaching, impacting health, environment and infrastructure in a very adverse manner. According to the United Nations (UN), waterborne diseases cause the death of more than 1.5 million people each year, much greater than deaths caused by accidents, crimes and terrorism. Therefore, it is very crucial to devise novel approaches and methodologies for analyzing water quality and to forecast future water quality trends. Different methodologies have been proposed and applied for analysis and monitoring of water quality as well as time series analysis. The methodologies range from statistical techniques, visual modelling, analysis algorithms and prediction algorithms and decision making.

Chapter 2

PROOF OF CONCEPT

2.1 Existing System

Early work considered using regression models to predict the trends and values of stock price for the next day. But using Linear Regression, the short term trends and price have high error rates.Therefore it might not be the ideal implementation. And also the attributes are not actually contributing enough for training the linear model.

2.2 Proposed System

Data in the real-time process: Real-time data processing involves continuous input, process and output of data which is processed in a short period. Data from batch processes: Batch process means the data is collected in a large volume all at once. It can consist of millions of records for a day and can be stored in a variety of ways. Database: It is a collection of information that is organised so that it can be easily accessed, managed and updated. Data cleaning: It is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated and improperly formatted. Feature generation: It is the process of taking raw, unstructured data and defining features in your anal-

ysis. Evaluation: It is the process of applying various algorithms on a dataset to find which gives the best accuracy for the taken dataset. Modelling: Splitting the data into training and testing. Model output: It is the process of displaying the accuracy of the applied algorithms..

2.3 Objectives

The aim of this study is the prediction of water quality components using artificial intelligence (AI) techniques including DEEP NEURAL NETWORK, SVM, and group method of data handling (GMDH). Therefore, in the first part of this section, the studied area is introduced and then ranges of measured water quality components are presented. Overviews on applied AI models are then presented.

Chapter 3

SCRUM MEETINGS

On 24-11-2021

Started searching the miniproject topic based on the new technology such as deep learning, IoT, machine learning, classification, prediction etc.

On 29-11-2021

The topic ”Water Quality Prediction” was selected and did the detail study of the topic, the required data set was selected. The data set was searched from the different site such as kaggle, NSE1 History etc.

On 06-12-2021

This day I submitted the synopsis and research paper to guide for the topic approval.

On 15-12-2021

After getting approval from the guide, the algorithm and model for the project were structured. Then the algorithm were chosen. Then quick study was done about which algorithms would be best for the project.

The selected ones were:
- Random Forest
- Decision Tree

On 18-12-2021

On this day guide took a detailed class on how to do the project, what IDEs to use, what paper are referred, what steps are follow to do the project and so on

On 06-01-2022

According to the project the required IDE such as Spyder, Colab were chosen .Even checked whether the system was efficient to train the model. Here colab to code the project,then started to deploying the model using the algorithm.Python language is used to code the project.

On 10-01-2022

After the project first review according to guide's opinion decided to concentrate on Water Quality Prediction based on relevant attributes.

On 13-01-2022

Used different algorithm/data model then choose the maximum accuracy one. The algorithm used are:-
Random Forest
Decision Tree

On 19-01-2022

Started to do project coding. Firstly study the data set and download the data set from NSE History Data. The data set contains data for water quality indices.

On 25-01-2022

Testing the data application

On 28-01-2022

The training done in two different data model then choose the maximum accuracy with regression for predicting the Water Quality.Random Forest is used for prediction.

On 02-02-2022

Created the git repository.

On 07-02-2022

Used flask for connection.

Chapter 4

IMPLEMENTATION

The aim of this study is the prediction of water quality components using artificial intelligence (AI) techniques including DEEP NEURAL NETWORK, SVM, and group method of data handling (GMDH). Therefore, in the first part of this section, the studied area is introduced and then ranges of measured water quality components are presented. Overviews on applied AI models are then presented.. A Machine learning algorithm is used here for predicting this Water Quality. The algorithm used here is Random Forest.

ALGORITHM

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems. 1. Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

2. Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.

3. Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

4. Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30

5. Stability- Stability arises because the result is based on majority voting/ averaging.

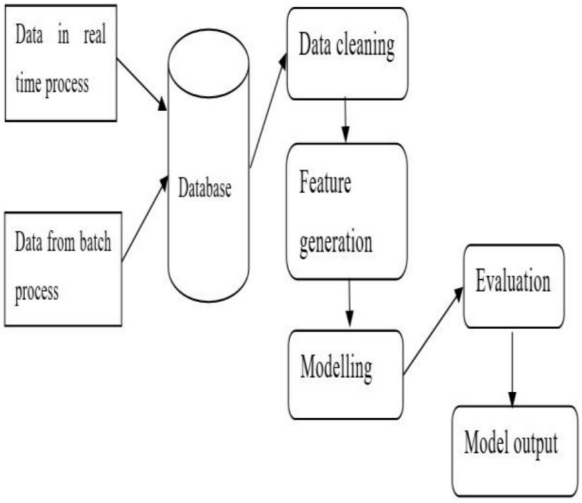
Decision Tree Algorithm

Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret..Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification prob-

lems too.The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

4.1 System Architecture

The use case diagram that describes the operation of the system.



4.2 Data set

The data requirements is very high for the project.We get a data set that contains the component like:-

• PH

• Hardness

- Solids
- Chlorimes
- Sulphate
- Conductivity
- Organic carbon
- Turbidity

4.3 Modules

4.3.1 Data Preprocessing

Explore the data set and analyse it. The new derived attributes are required for the proper analysis and prediction of variation of water quality.
The derived attributes are :

4.3.2 Data Splitting

For the proper implementation and testing the processed data is divided into training data (70%) and test data (30%).

4.3.3 Modelling

The training data is used to create the model for the application using Random Forest. And the change different parameters and tune to create the best model for the purpose.

4.3.4 Implementation

Model deployment is simply the engineering task of exposing an ML model to real use. The python web framework flask is used for the implementation.

Chapter 5

RESULT ANALYSIS

The system predicts the water quality based on the input wter quality index and gives the idea whether the water is safe to drink.
Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used: In most cases, high accuracy value represents a good model, but considering the fact that we are training a classification model in our case, an article that was predicted as true while it was actually false (false positive) can have negative consequences; similarly, if an article was predicted as false while it contained factual data, this can create trust issues.
Accuarcy = TP+TN/TP+TN+FP+FN

Chapter 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

To evaluate the conventional algorithm, a data set is built and studied a trend of price variation for the period of limited days. Machine Learning algorithms are applied on the data set to predict the water quality. This gives the predicted value for percentage change for expiry date. And then according to that the investor can make decision to minimize the loss. Data was collected from National Stock Exchange History. The model gives about 68% accuracy so that it would valuable for the consumer to decide the water is drinkable or not.

6.2 Future Scope

In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate. And further advanced analysis can find more dependent attributes and this can make the model more accurate.

Chapter 7

SOURCE CODE

7.0.1 model.py

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sb

from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report, accuracy_score, classification_report, precision_score,rec
from sklearn.model_selection import GridSearchCV

from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import seaborn as sns
import plotly.graph_objs as go

import pickle

df = pd.read_csv('C:/Users/vaisa/prj/water_potability.csv')
df.head(10)
df.info
df.isnull().sum()
df['ph'].fillna(value = df['ph'].median() , inplace = True)
df['Sulfate'].fillna(value = df['Sulfate'].median() , inplace = True)
df['Trihalomethanes'].fillna(value = df['Trihalomethanes'].median() , inplace = True)

df.isnull().sum()
#round off the pH value to one decimal place
df['ph'] = df['ph'].round(decimals = 1)
df["ph"].head()
```

```
df["Type of Water"] = ""
for i in range(0,len(df)):
    if (df["ph"][i] > 9):
        df["Type of Water"][i] = "Alkaline water"
    elif (df["ph"][i] <= 9 and df["ph"][i] > 8):
        df["Type of Water"][i] = "Bottled waters labeled as alkaline"
    elif (df["ph"][i] <= 8 and df["ph"][i] > 7.5 ):
        df["Type of Water"][i] = "Ocean water"
    elif(df["ph"][i] == 7.5 ):
        df["Type of Water"][i] = "Tap water"
    elif(df["ph"][i] < 7.5 and df["ph"][i] >=6.5):
        df["Type of Water"][i] = "Common bottled waters"
    elif(df["ph"][i] < 6.5 and df["ph"][i] >=5.5):
        df["Type of Water"][i] = "Distilled reverse osmosis water"
    else:
        df["Type of Water"][i] = "Acidic water"

df["Type of Hardness"] = ""
for i in range(0,len(df)):
    if (df["Hardness"][i] >=0  and df["Hardness"][i] < 17.1):
        df["Type of Hardness"][i] = "Soft"
    elif (df["Hardness"][i] >= 17.1 and df["Hardness"][i] < 60):
        df["Type of Hardness"][i] = "Slightly hard"
    elif (df["Hardness"][i] >= 60 and df["Hardness"][i] < 120 ):
        df["Type of Hardness"][i] = "Moderately hard"
    elif(df["Hardness"][i] >= 120 and df["Hardness"][i] < 180):
        df["Type of Hardness"][i] = "Hard"
    else:
        df["Type of Hardness"][i] = "Very Hard"
```

```
df.head()

plt.figure(figsize=(7,5))
colors = sb.color_palette('twilight')[0:6]

df['Potability'].value_counts().plot(kind='pie',labels = ['','','',''], autopct='%1.1f
plt.legend(labels=['Non-Potable', 'Potable'])
plt.show()


df.drop('Potability', axis=1).hist(figsize=(12,8));




index_vals = df['Potability'].astype('category').cat.codes
#df.value_counts(["Type of Hardness","Potability"],sort=True,ascending=True).plot(kind
pd.value_counts(df.values.flatten())
plt.figure(figsize=(12,7))
plt.xlabel("Type of Hardness")
plt.ylabel("Potability")
plt.title("Item_Weight and Item_Outlet_Sales_Analysis")
plt.plot(df["Type of Hardness"],df["Potability"],'.',alpha=0.3)


#one hot encoding
df = pd.get_dummies(df, columns = ['Type of Water', 'Type of Hardness'])
df.head()

df.shape
x = df.drop(['Potability'], axis = 1)
y = df['Potability']

x_train, x_test, y_train, y_test= train_test_split(x, y, test_size=0.3, random_state =
scores = {}
RF = RandomForestClassifier()
```

```
param_grid = {'max_depth': [80, 90, 100, 110]}
RF_GS = GridSearchCV(RF, param_grid,cv=5,return_train_score=True)

RF_GS.fit(x_train,y_train)

print(RF_GS.best_estimator_)
predicted_values = RF_GS.best_estimator_.predict(x_test)

x = metrics.accuracy_score(y_test, predicted_values)
y_pred_RF = RF_GS.best_estimator_.predict(x_test)
print("Random Forest Accuracy is: ", x*100)
print()
print(classification_report(y_test,predicted_values))

print('Accuracy: ', accuracy_score(y_pred_RF, y_test))
print('Precision: ', precision_score(y_pred_RF, y_test))
print('Recall: ', recall_score(y_pred_RF, y_test))
print('F1 score: ', f1_score(y_pred_RF, y_test))

scores['Random Forest'] = accuracy_score(y_pred_RF, y_test)*100



from sklearn.svm import SVC
model_linear = SVC(kernel = "linear")
model_linear.fit(x_train,y_train)
pred_test_linear = model_linear.predict(x_test)
acclinear=np.mean(pred_test_linear==y_test)
print(classification_report(y_test,pred_test_linear))
print('Accuracy: ', accuracy_score(pred_test_linear, y_test))
print('Precision: ', precision_score(pred_test_linear, y_test))
print('Recall: ', recall_score(pred_test_linear, y_test))
print('F1 score: ', f1_score(pred_test_linear, y_test))

scores['svm'] = accuracy_score(pred_test_linear, y_test)*100
```

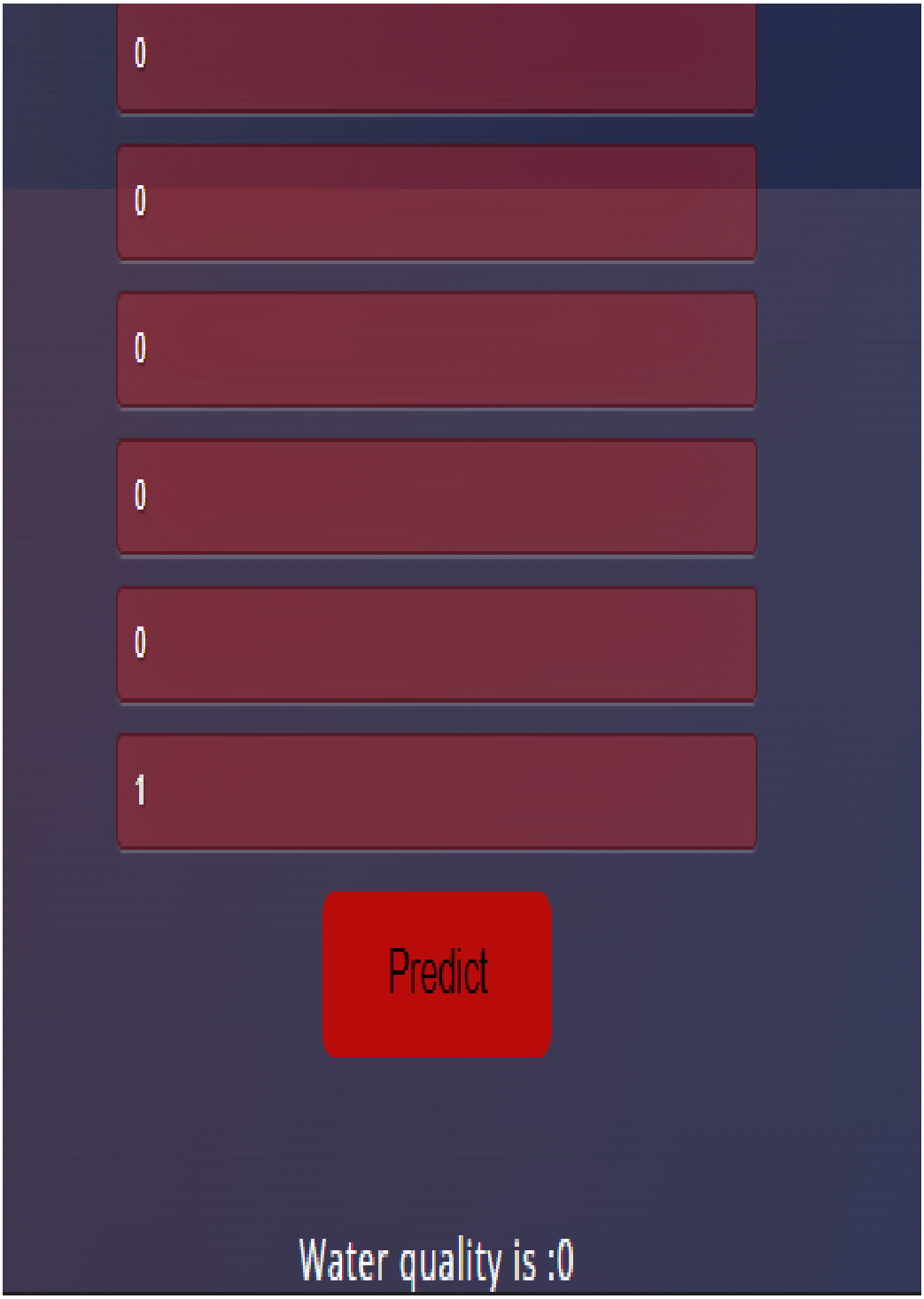
```
DT = DecisionTreeClassifier(criterion="entropy",random_state=150)

param_grid = {'criterion':['gini','entropy'],'max_depth':[4,5,6,7,8,9,10,11,12,15,20,3
DT_GS = GridSearchCV(DT, param_grid,cv=5,return_train_score=True)

DT_GS.fit(x_train,y_train)

print(DT_GS.best_estimator_)
predicted_values = DT_GS.best_estimator_.predict(x_test)

x = metrics.accuracy_score(y_test, predicted_values)
y_pred_DT = DT_GS.best_estimator_.predict(x_test)
print("Decision Tree Accuracy is: ", x*100)
print()
print(classification_report(y_test,predicted_values))

print('Accuracy: ', accuracy_score(y_pred_DT, y_test))
print('Precision: ', precision_score(y_pred_DT, y_test))
print('Recall: ', recall_score(y_pred_DT, y_test))
print('F1 score: ', f1_score(y_pred_DT, y_test))

scores['Decision Tree'] = accuracy_score(y_pred_DT, y_test)*100
scores


filename='C:/Users/vaisa/prj/water_potability.pkl'
pickle.dump(RF_GS, open(filename, 'wb'))
# load the model from disk
#loaded_model = pickle.load(open("C:/Users/admin/Documents/project-rijo/water quality
#result = loaded_model.score(x_train,y_train)
```

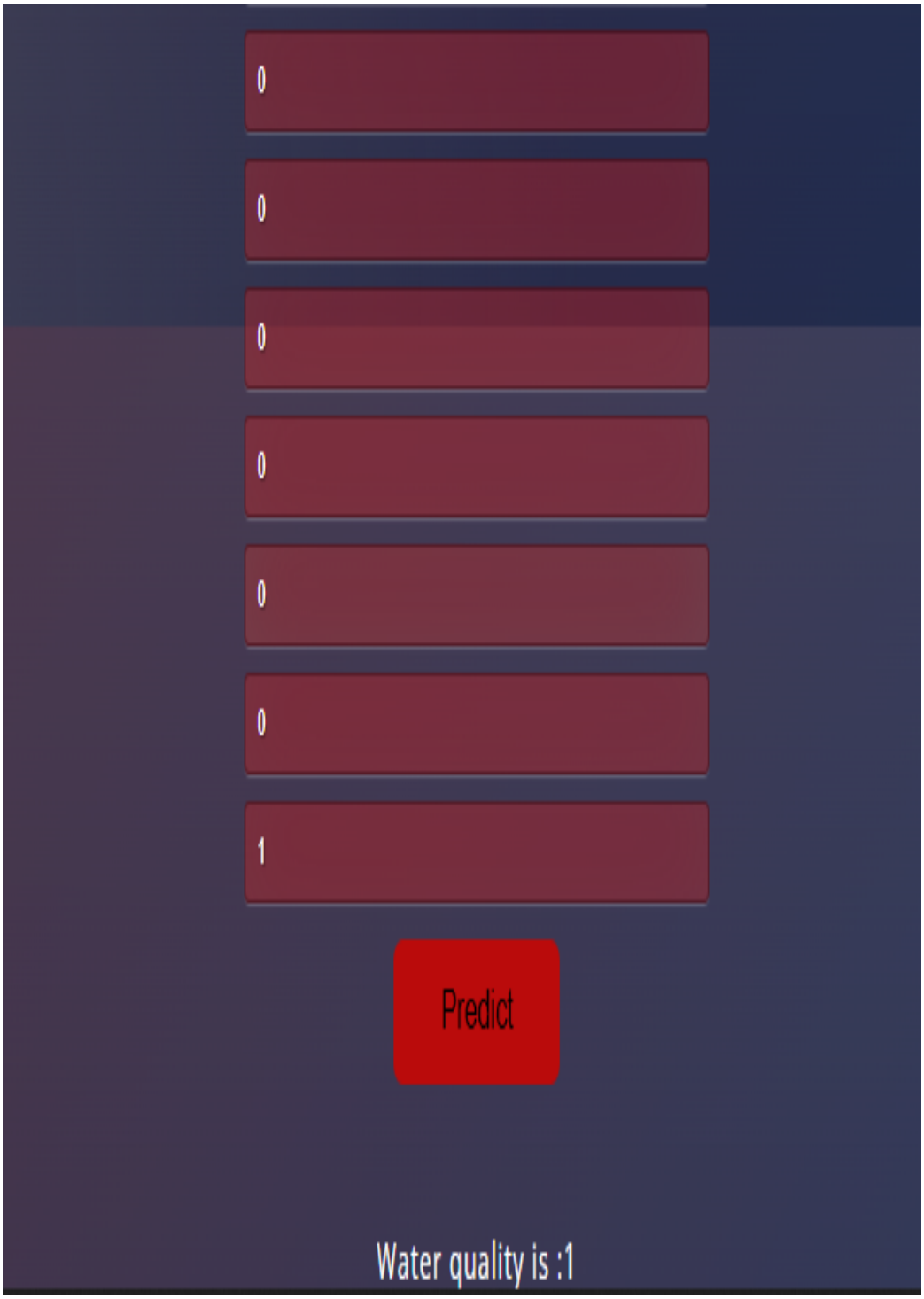
Chapter 8

SCREEN SHOTS


Water quality prediction
ph
Hardness
Solids
Please fill out this field
Chloramines
Sulfate
Conductivity
10
Trihalomethanes
Turbidity
Type of Water_Acidic water 0/1
Type of Water_Alkaline water 0/1
Type of Water_Bottled waters labeled as alkaline 0/1
Type of Water_Common bottled waters 0/1
Type of Water_Distilled reverse osmosis water 0/1
Type of Water_Ocean water 0/1
Type of Water_Tap water 0/1
Type of Hardness_Hard 0/1
Type of Hardness_Moderately hard 0/1
Type of Hardness_Slightly hard 0/1
Type of Hardness_Very Hard 0/1
Predict


Water quality prediction
7
204
20791
7.3
368.5
564
10
86
2.96
0
0
0
1
0
0
0
0
0
0
1
Predict


0
0
0
0
0
1
Predict
Water quality is :0


0
0
0
0
0
0
1
Predict
Water quality is :1

Chapter 9

REFERENCES

- www.youtube.com
- Aaditya Gupta, Chesta Bansal, and Agha Imran Husain, “Ground Water Quality Monitoring Using Wireless Sensors and Machine Learning”, IEEE, Vol., No., pp.121-125,2018.
- Touglas kwasi boah, Scethen Boakye twum, and Kenneth V. Pelig-ba “Mathematical Computation of Water Quality Index of via Dam in Upper East Region of Ghana”, Hikari Ltd, Vol:3, No.1, pp.11-16, 2015.
- F. J Thakor, D. K. Bhoi, H.R. Dabhi, S.N. Pandya and Nikitaraj B. Chauhan, “Water Quality Index (W.Q.I) of Pariyej Lake Dist. Kheda-Gujarat, Vol. 6, No.2, pp. 225- 231, 2011