A Project Report

On

Water Potability Detection

Submitted by

Ayush – 2115000248

Hitansh Mangla–2115000473

Achintya Gupta – 2115000043

Nikunj Maheshwari - 2115000672

Supervisor

Mr. Sanjay Madaan
Technical Trainer
CSE Department

Department of Computer Engineering &Application
G.L.A. UNIVERSITY

GLA University, Mathura - 281406

29/11/2023

BONAFIDE CERTIFICATE

Certified that this project report “**Water Potability Detection**” is the Bonafide work of
“Ayush – 2115000248

Hitansh Mangla **–** 2115000473

Achintya Gupta – 2115000043

Nikunj Maheshwari -2115000672”
who carried out the project work under mysupervision.

SIGNATURE (HOD)	**SIGNATURE (SUPERVISOR)**
HEAD OF THE DEPARTMENT	**Mr. Sanjay Madaan**
CSE Department	**(Technical Trainer)**
	CSE Department

Submitted for the project viva-voce examination held on 29 November 2023

ACKNOWLEDGEMENT

Presenting the ascribed project paper report in this very simple and official form, we would like to place my deep gratitude to GLA University for providing us with the instructor Mr. Sanjay Madaan , our technical trainer and supervisor.

He has been helping us since Day 1 of this project. He provided us with the roadmap, and the basic guidelines explaining on how to work on the project. He has been conducting regular meetings to check the progress of the project and providing us with the resources related to the project. Without his help, we wouldn’t have been able to complete this project.

And at last but not least we would like to thank our dear parents for helping us to grab this opportunity to get trained and also my colleagues who helped me find resources during the training.

Thanking You

Name of Candidate:Ayush

(2115000248)

Name of Candidate: Hitansh Mangla

(2115000473)

Name of Candidate:Achintya Gupta

(2115000043)

Name of Candidate: Nikunj Maheshwari

(2115000672)

CERTIFICATE

This is to certify that the above statement made by the students is correct to the best of my knowledge and belief.

Date:

Place: Mathura

Name and Signature with Affiliation of Supervisor

Mr. Sanjay Madaan

CONTENTS :-

Table of Contents

1. ABSTRACT

A Water Potability detection using Machine Learning aims to assess the safety of a specific water source for drinking by employing various ML algorithms to analyze the quality of water.

This project aims to automate the assessment process by developing a model that can predict water potability based on water quality data

KEY FEATURES:-

- **Data Collection**: Gather a dataset of water quality measurements, which should include various parameters such as pH, turbidity, hardness, chloride, sulfate, dissolved solids, and more.

- **Data Preprocessing**: Clean and pre-process the dataset, handling missing values, outliers, and data normalization

- **Feature Selection/Engineering:** Choose relevant features (parameters) that are most likely to influence water potability.

- **Data Splitting:** Split the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set to fine-tune hyper parameters, and the testing set to evaluate the model's performance.

- **Model Selection**: Select an appropriate machine learning algorithm for binary classification. Common choices include logistic regression, decision trees, random forests, support vector machines, or neural networks.

- **Model Training:** Train the selected model on the training data, using the water quality parameters as input and the potability labels as the target variable.

- **Model Evaluation:** Assess the model's performance on the validation set, considering metrics like accuracy, precision, recall, F1 score, and the receiver operating characteristic (ROC) curve.

- **Hyperparameter Tuning**: Optimize the model's hyperparameters to improve its performance. Techniques like grid search or random search can be used.

- **Model Testing**:- Evaluate the final model on the testing dataset to assess its real-world performanc

CHAPTER-1

INTRODUCTION

1.1 CONTEXT:-

The "Water Potability Detection" mini project addresses this critical issue by leveraging machine learning algorithms to assess the potability of water samples. The project utilizes data science techniques to analyze various water quality parameters and predict whether a given sample is suitable for consumption.

1.2 MOTIVATION:-

"In the pursuit of a safer and healthier future, our mini project leverages the power of Machine Learning to revolutionize water potability detection. By employing advanced algorithms, we aim to provide an efficient and reliable solution for identifying the safety of drinking water. This project is driven by a profound motivation to ensure access to clean and potable water, contributing to the well-being of communities and fostering a sustainable environment."

1.3 OBJECTIVE:-

The primary objective of this project is to utilize machine learning techniques to assess and confirm the potability of a specific water source, such as a local well, tap water, or a small scale water supply system. This project aims to automate the assessment process by developing a model that can predict water potability based on water quality data.

1.4 DATASET:-

Proposed system is implemented using the water potability dataset from Kaggle . The water_potability.csv file contains water quality metrics for 3276 dataset, 9 features and one class variable.

- Feature lists are pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity

- Potability is the class label

CHAPTER -2

ALGORITHM ANALYSIS

2.1 MODEL SELECTION:-

There are many algorithm present in the supervised machine learning such as Logistic Regression Support Vector Machines, Random Forests ,Decision Trees etc. Among these all we have selected three of the algorithm for the project and trained them and finally choosed the one of the best algorithm among all these algorithm.

2.2 PROBLEM STATEMENT:-

With the population and the ever increasing need for various resources we are faced with a dilemma in how to manage our lives. Sometimes we end up by utilizing a poor or contaminated source of water for our use.

The proposed system aims to provide the solution for the same by allowing users to monitor the water quality from a given sample of water and predict whether water is contaminated or not.

2.3 Supervised Learning Algorithms:-

- **Logistic Regression**
- **Support Vector Machine**
- **K-Nearest Neighbors**
- **Decision Trees**

Libraries and other Requirement

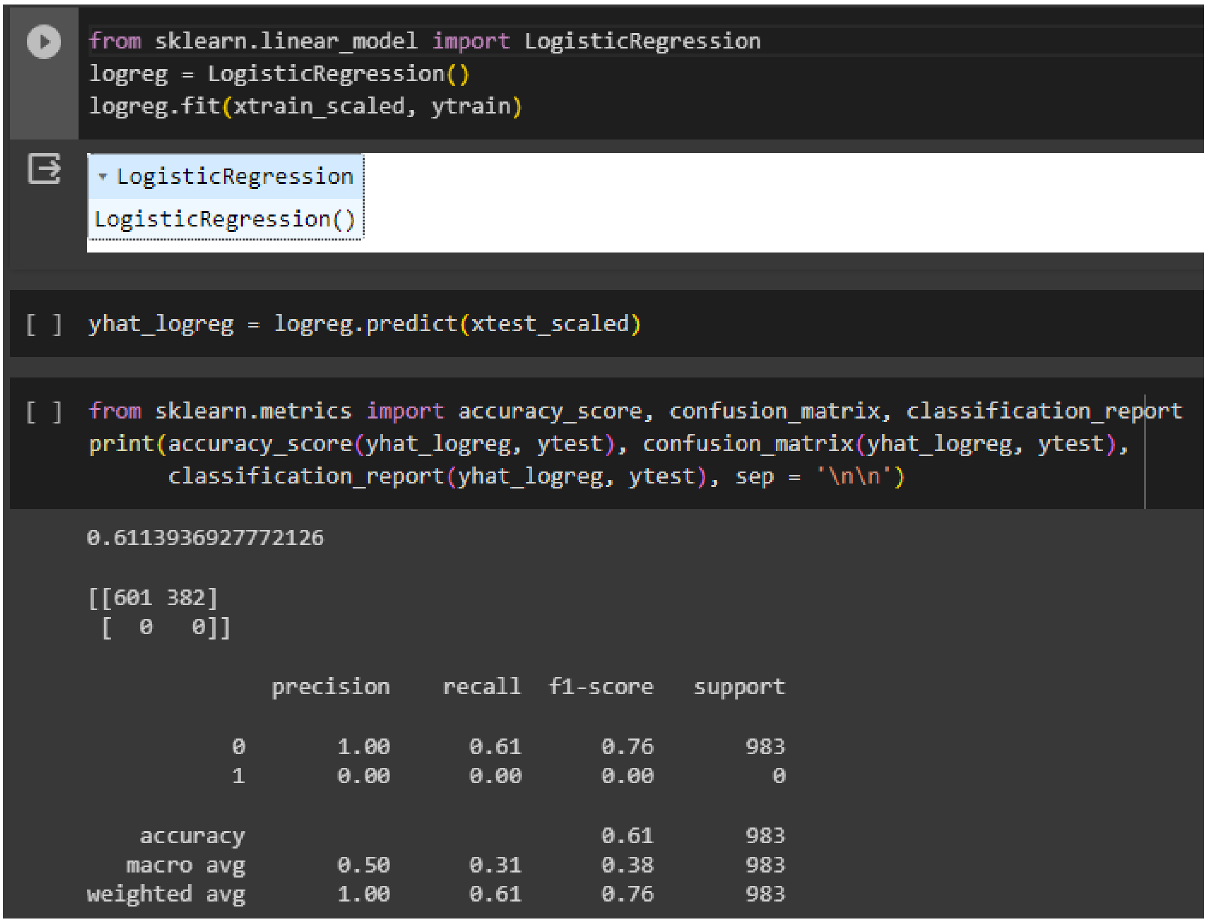
- Software used: Google Colab
- Language used: Python ,Pandas , Numpy, Matplotlib
- Machine learning Algorithm

CHAPTER – 3

IMPLEMENTATION AND MODEL SELECTION

❖ LOGISTIC REGRESSION:-

Logistic regression is one of the Machine Learning algorithms of Supervised Learning. 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.



```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(xtrain_scaled, ytrain)
```

```
▾ LogisticRegression
LogisticRegression()
```

```
[ ] yhat_logreg = logreg.predict(xtest_scaled)
```

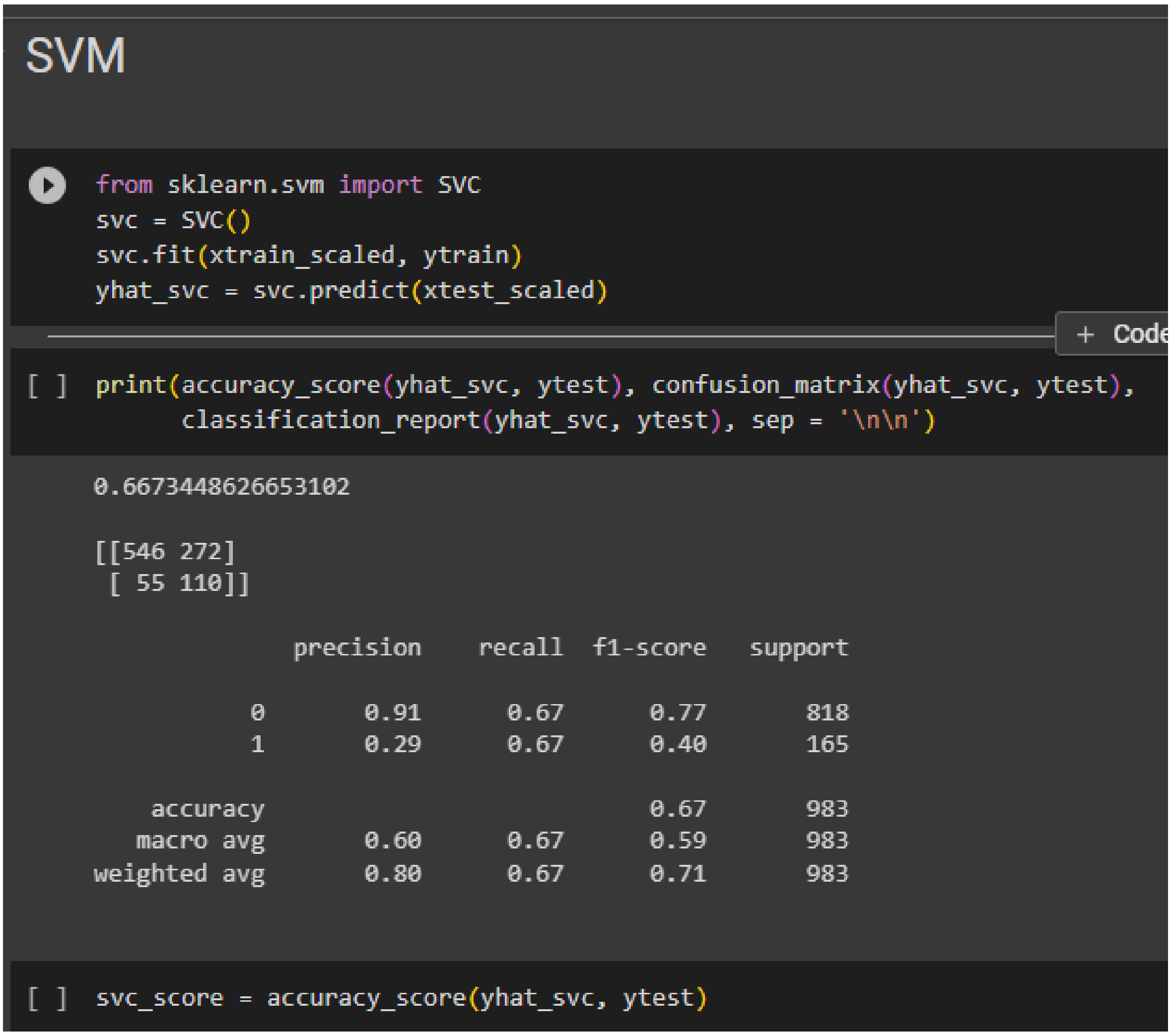
```
[ ] from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
    print(accuracy_score(yhat_logreg, ytest), confusion_matrix(yhat_logreg, ytest),
          classification_report(yhat_logreg, ytest), sep = '\n\n')
```

```
0.6113936927772126

[[601 382]
 [  0   0]]

              precision    recall  f1-score   support

           0       1.00      0.61      0.76       983
           1       0.00      0.00      0.00         0

    accuracy                           0.61       983
   macro avg       0.50      0.31      0.38       983
weighted avg       1.00      0.61      0.76       983
```

❖ SUPPORT VECTOR MACHINE:-

A support vector machine (SVM) is a type of supervised learning algorithm used in machine learning to solve classification and regression tasks .

SVMs are particularly good at solving binary classification problems, which require classifying the elements of a data set into two groups.



SVM

```
from sklearn.svm import SVC
svc = SVC()
svc.fit(xtrain_scaled, ytrain)
yhat_svc = svc.predict(xtest_scaled)
```

```
[ ] print(accuracy_score(yhat_svc, ytest), confusion_matrix(yhat_svc, ytest),
          classification_report(yhat_svc, ytest), sep = '\n\n')
```

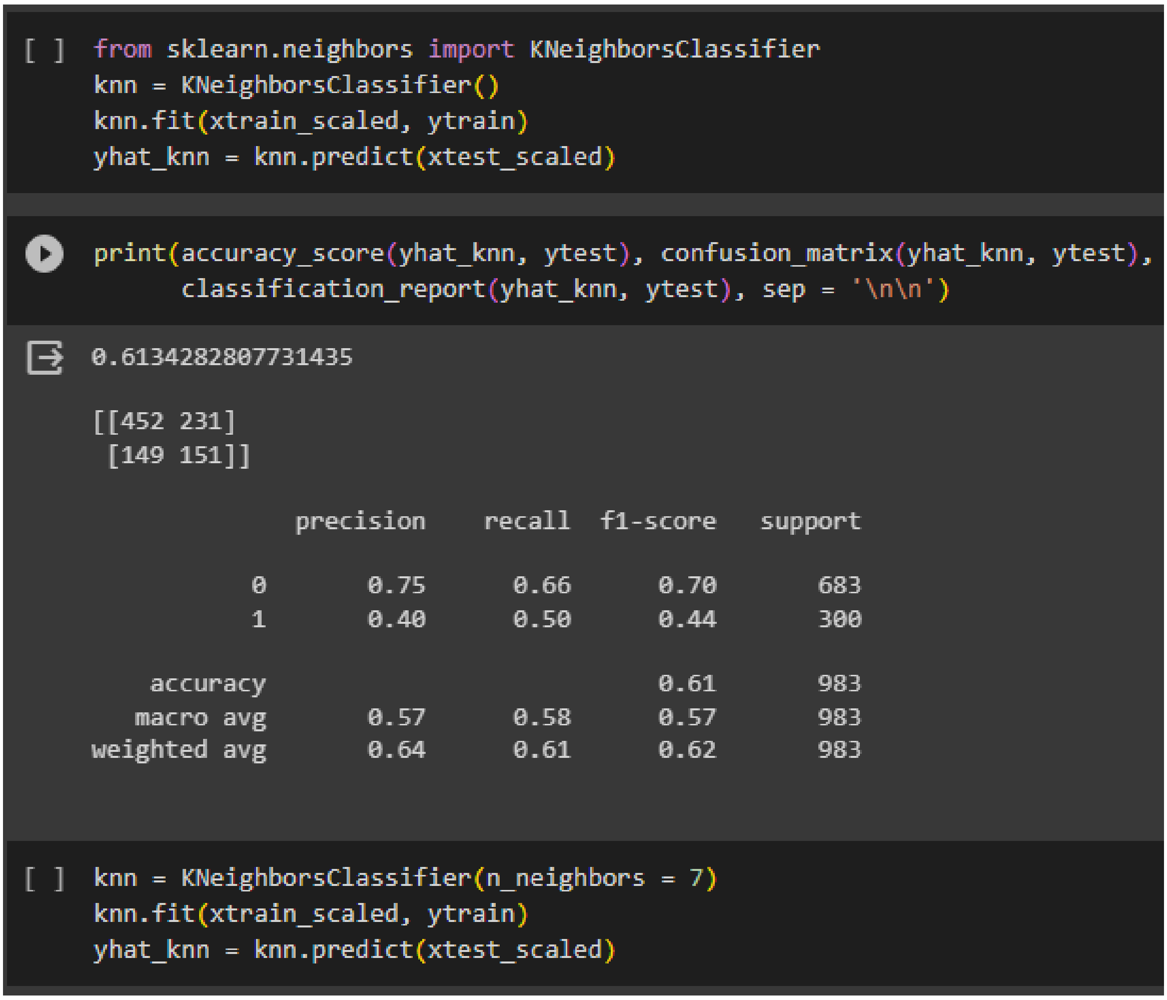
```
0.6673448626653102

[[546 272]
 [ 55 110]]

              precision    recall  f1-score   support

           0       0.91      0.67      0.77       818
           1       0.29      0.67      0.40       165

    accuracy                           0.67       983
   macro avg       0.60      0.67      0.59       983
weighted avg       0.80      0.67      0.71       983
```

```
[ ] svc_score = accuracy_score(yhat_svc, ytest)
```

K-NEAREST NEIGHBOR:-

KNN stands for “K-Nearest Neighbor”.It is a supervised machine learning algorithm.

The algorithm can be used to solve both classification and regression problem statement



```
[ ]  from sklearn.neighbors import KNeighborsClassifier
     knn = KNeighborsClassifier()
     knn.fit(xtrain_scaled, ytrain)
     yhat_knn = knn.predict(xtest_scaled)
```

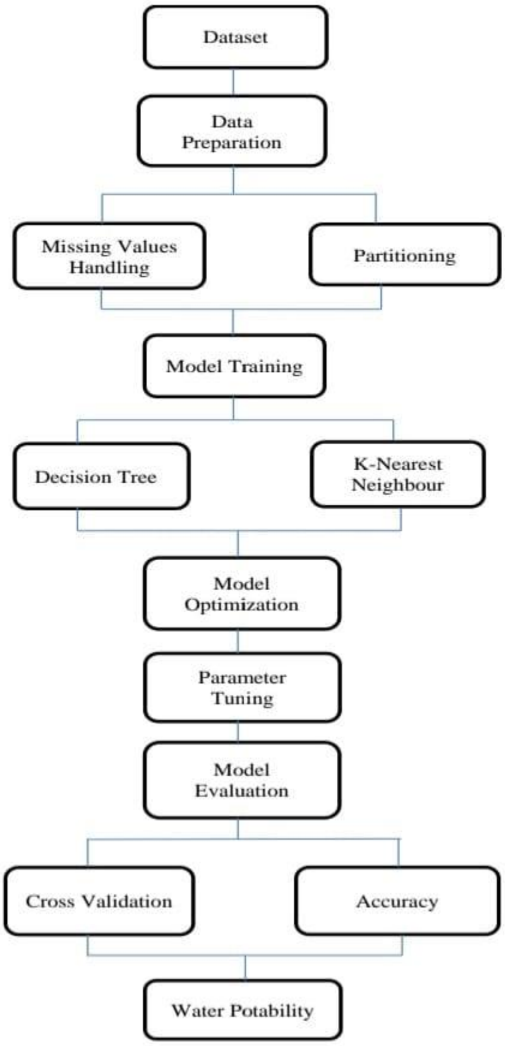
```
print(accuracy_score(yhat_knn, ytest), confusion_matrix(yhat_knn, ytest),
      classification_report(yhat_knn, ytest), sep = '\n\n')
```

```
0.6134282807731435

[[452 231]
 [149 151]]

              precision    recall  f1-score   support

           0       0.75      0.66      0.70       683
           1       0.40      0.50      0.44       300

    accuracy                           0.61       983
   macro avg       0.57      0.58      0.57       983
weighted avg       0.64      0.61      0.62       983
```

```
[ ]  knn = KNeighborsClassifier(n_neighbors = 7)
     knn.fit(xtrain_scaled, ytrain)
     yhat_knn = knn.predict(xtest_scaled)
```

FLOW CHART

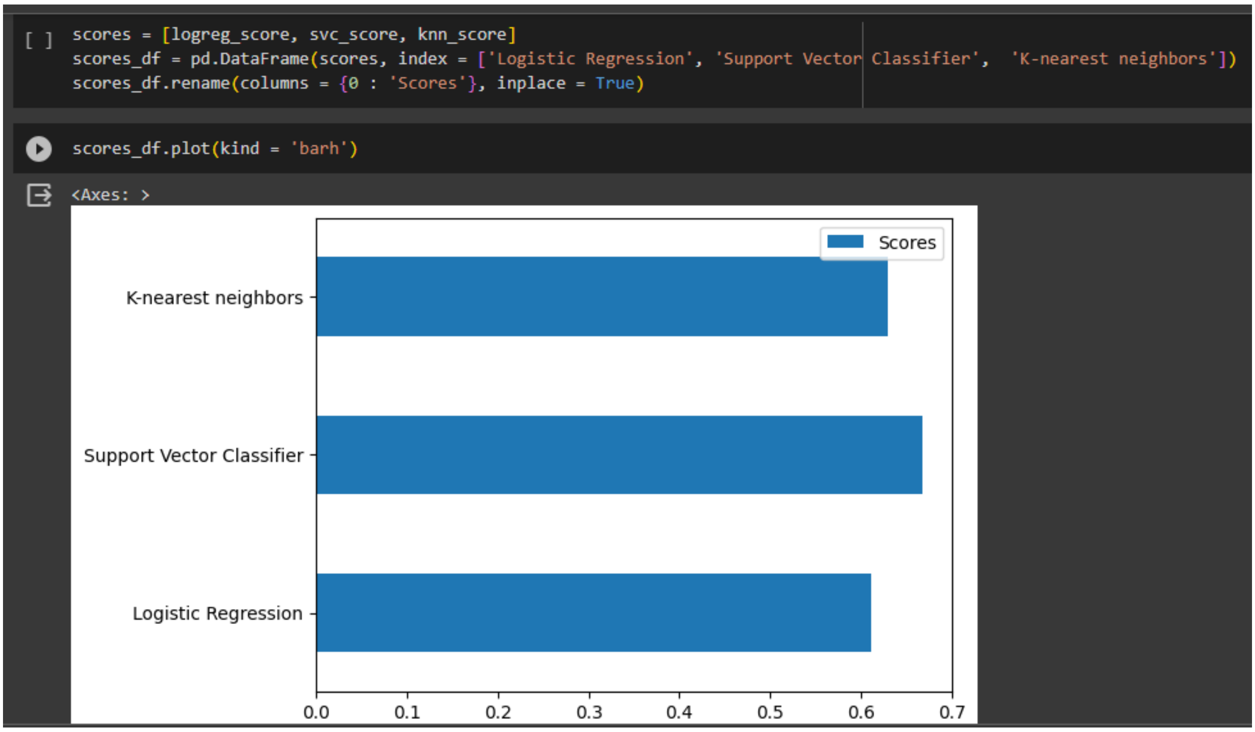
Fig: Scatter Plot of Hardness and Solids



CHAPTER -5

RESULTS

❖ In this we have a bar graph showing comparison between all the model's accuracy score and we have come to a result that **SVM** is the best model among all these models.



```
scores = [logreg_score, svc_score, knn_score]
scores_df = pd.DataFrame(scores, index = ['Logistic Regression', 'Support Vector Classifier',  'K-nearest neighbors'])
scores_df.rename(columns = {0 : 'Scores'}, inplace = True)
```

```
scores_df.plot(kind = 'barh')
```

```
<Axes: >
```

CHAPTER -5

CONCLUSION

In conclusion, the Water Potability Detection mini project leveraging Machine Learning (ML) has proven to be a significant stride towards ensuring safe and clean drinking water. Through the utilization of advanced algorithms, predictive models were developed to assess the potability of water based on various chemical and physical parameters. The accuracy and efficiency of these models underscore the potential of ML in addressing water quality concerns.

This project has broader implications for public health, enabling rapid and reliable identification of potable water sources. The successful implementation highlights the capability of ML to analyze complex datasets, providing a valuable tool for water quality monitoring. As we move forward, the integration of such technologies into real-world water management systems could revolutionize the way we safeguard public health and environmental sustainability. This mini project serves as a stepping stone towards creating smarter, data-driven solutions for ensuring access to clean and safe drinking water for communities around the globe.

REFERENCES

Dataset:-

URL:- https://www.kaggle.com/datasets/adityakadiwal/water-potability