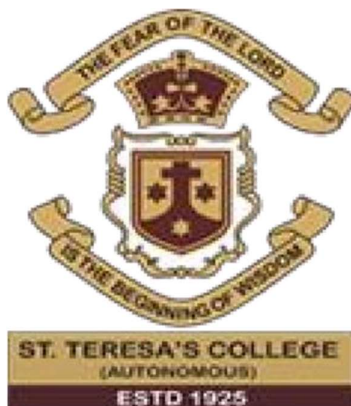


ST. TERESA'S COLLEGE (AUTONOMOUS)
AFFILIATED TO MAHATMA GANDHI UNIVERSITY,
KOTTAYAM



WATER QUALITY ANALYSIS
PROJECT REPORT

In partial fulfilment of the requirements for the award of the degree of
BACHELOR OF SCIENCE IN
COMPUTER APPLICATIONS
[TRIPLE MAIN]

Submitted By

Khadeeja Nazreen

III B.Sc. Computer Applications [Triple Main]

Register No: SB21CA014

Under the guidance of

Ms. Arunima P S

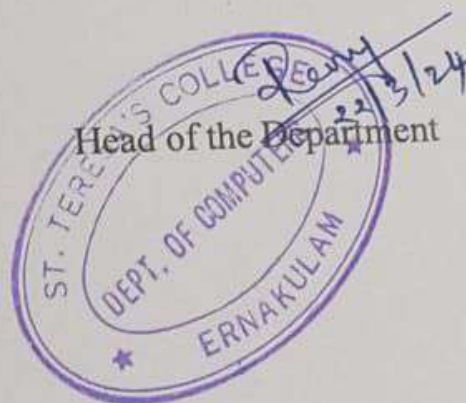
DEPARTMENT OF COMPUTER APPLICATIONS
2021 – 2024

ST. TERESA'S COLLEGE (AUTONOMOUS)
AFFILIATED TO MAHATMA GANDHI UNIVERSITY,
KOTTAYAM



CERTIFICATE

This is to certify that the project report entitled "*Water Quality Analysis*" is a bona-fide record of the work done by **KHADEEJA NAZREEN (SB21CA014)** during the year 2021 – 2024 and submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Applications (Triple Main) under Mahatma Gandhi University, Kottayam.



Internal Examiner

Date: 21-03-2024

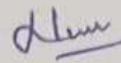
External Examiner

DECLARATION

I, **KHADEEJA NAZREEN** (Register no: **SB21CA014**), B.Sc. Computer Applications [Triple Main] final year student of St. Teresa's College (Autonomous), Ernakulam, hereby declare that the project submitted named "Water Quality Analysis" for the Bachelor's Degree in Computer Applications [Triple Main] is my original work. I further declare that the said work has not previously been submitted to any other university or academic body.

Date: 21-03-2024

Place: Ernakulam



KHADEEJA NAZREEN

ACKNOWLEDGEMENT

I would like to convey my heartfelt gratitude to **Rev. Dr. Sr. Vinitha** (CSST) Manager, Director **Rev. Sr. Emeline** (CSST) and Principal **Dr. Alphonsa Vijaya Joseph** for providing me with this wonderful opportunity to work on a project with the topic ‘Water Quality Analysis’.

I would like to express my profound gratitude to the Head of the Department of Computer Applications **Ms. Remya C J.** , my project guide **Ms. Arunima P S** and all other faculty of the department for their contributions to the completion of my project. The completion of the project would not have been possible without their help and insights.

Finally, I take this opportunity to thank all them who has directly or indirectly helped me with my project.

KHADEEJA NAZREEN

ABSTRACT

Water quality analysis plays a pivotal role in environmental monitoring and management, facilitating timely interventions to safeguard ecosystems and public health. Leveraging machine learning algorithms, this research aims to compare and evaluate the predictive performance of five prominent models—Linear Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN)—in forecasting water quality parameters. A comprehensive dataset encompassing diverse physicochemical and biological variables from various water bodies forms the basis of analysis. Through rigorous evaluation of predictive accuracy, computational efficiency, and robustness, this study sheds light on the strengths and limitations of each algorithm in the context of water quality prediction. The findings offer valuable insights for policymakers, water resource managers, and environmental scientists, guiding the development of effective decision support systems for sustainable water resource management and conservation efforts.

CONTENTS

1.Introduction.....	1
1.1 About Project.....	1
1.2 Objective of Project.....	1
2 Methodology	2
2.1 Data Collection	2
2.2 Method used.....	2
2.2.1 Logistic Regression.....	2
2.2.2 Decision Tree	2
2.2.3 Random Forest.....	3
2.2.4 Support Vector Machine (SVM)	3
2.2.5 K Neighbors	3
2.3Data Wrangling.....	4
3. Source Code	5
4. Results	6
5. Conclusion	10
6. Literature Review.....	11
7. Reference.....	14

1 .INTRODUCTION

1.1 About project

Water quality is a vital aspect of environmental health and sustainability, directly impacting ecosystems, human health, and economic activities. With increasing anthropogenic activities and environmental stressors, the need for accurate and timely water quality prediction has become paramount for effective management and conservation efforts. Traditional methods of water quality assessment often rely on costly and time-consuming laboratory analyses, limiting their practicality for real-time monitoring and decision-making.

This study focuses on comparing the performance of five commonly used ML algorithms—Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and K Neighbors—in predicting water quality parameters. By leveraging a comprehensive dataset encompassing various environmental variables and water quality indicators, we aim to assess the predictive accuracy, computational efficiency, and robustness of each algorithm. Such comparative analysis is essential for identifying the most suitable ML model for water quality prediction tasks across diverse environmental settings.

1.2 Objective of project

Evaluate the performance of five machine learning algorithms—Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and K Neighbor—in predicting water quality parameters.

Compare the predictive accuracy, computational efficiency, and robustness of the selected machine learning algorithms using a comprehensive dataset containing physicochemical and biological parameters from diverse water bodies

2. METHODOLOGY

2.1 Data collection

Dataset I have used is from kaggle and in that I have many factors that contribute to the potability of water. There are factors like pH, hardness , solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity and atleast potability.

2.2 Methods used

Here for comparing the accuracy of the machine learning algorithms, I have choosen 5 algorithms , they are; Logistic Regression, Decision Tree, Random Forest, SVM , an K- Neighbours.

2.2.1 Logistic Regression

Logistic Regression is a statistical technique used for binary classification tasks, where the outcome variable is categorical and has two possible outcomes. Despite its name, logistic regression is primarily used for classification rather than regression. It models the probability of the occurrence of a certain event by fitting a logistic function to the observed data.

In the context of water quality prediction, logistic regression can be applied to classify water samples as either meeting or failing certain quality standards based on a set of predictor variables such as pH, temperature, dissolved oxygen, and pollutant concentrations. By estimating the probability of water samples belonging to a specific class (e.g., safe or polluted), logistic regression enables decision-makers to assess the likelihood of water quality exceedances and prioritize management actions accordingly.

2.2.2 Decision Tree

Decision Trees are versatile and intuitive machine learning algorithms used for both classification and regression tasks. They model decision rules by recursively partitioning the feature space into subsets, aiming to minimize impurity or maximize information gain at each split. In the context of water quality prediction, decision trees can effectively capture complex relationships between environmental variables and water quality indicators.

One of the primary advantages of decision trees is their interpretability. The resulting tree structure provides clear insights into the decision-making process, making it easy to understand and interpret the factors driving water quality variations. Decision trees are particularly useful for identifying important features and their interactions, which can inform environmental management strategies and guide targeted monitoring efforts.

2.2.3 Random Forest

Random Forest is a powerful ensemble learning algorithm that leverages the collective wisdom of multiple decision trees to make predictions. It operates by constructing a multitude of decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees for classification and regression tasks, respectively.

In the realm of water quality prediction, Random Forest offers several advantages over standalone decision trees. Firstly, it mitigates the issue of overfitting commonly associated with individual decision trees by aggregating predictions from multiple trees. This ensemble approach results in improved generalization performance and robustness, especially when dealing with noisy or high-dimensional datasets.

2.2.4 Support Vector Machine (SVM)

Support Vector Machines (SVM) are a class of supervised learning algorithms used for classification and regression tasks. SVM works by finding the optimal

hyperplane that separates data points into different classes while maximizing the margin between the classes. In the context of water quality prediction, SVM can effectively handle both linear and nonlinear relationships between predictor variables and water quality parameters.

One of the key advantages of SVM is its ability to capture complex decision boundaries in high-dimensional feature spaces. By using kernel functions, SVM can map input data into higher-dimensional spaces where linear separation becomes possible, allowing for more flexible and accurate modeling of nonlinear relationships in the data.

2.2.5 K Neighbors

K-Nearest Neighbors (KNN) is a simple yet effective non-parametric algorithm used for both classification and regression tasks. In KNN, the prediction for a new data point is determined by the majority vote (for classification) or averaging (for regression) of its k nearest neighbors in the feature space. KNN does not explicitly learn a model during the training phase; instead, it memorizes the entire training dataset, making it particularly suitable for small to moderate-sized datasets.

In the context of water quality prediction, KNN offers several advantages. Firstly, it is easy to understand and implement, requiring minimal parameter tuning and preprocessing. KNN does not make any assumptions about the underlying data distribution, making it robust to outliers and noisy data. Additionally, KNN naturally handles multiclass classification tasks and can accommodate both numerical and categorical features.

2.3 Data Wrangling

Data wrangling, also known as data munging or data preprocessing, refers to the process of cleaning, transforming, and enriching raw data into a format suitable for analysis or modeling. It is a critical step in the data science workflow, as the quality and structure of the data greatly influence the accuracy and reliability of

subsequent analyses or models. Below are the key steps involved in data wrangling:

Data Collection: The first step is to gather data which was already available in kaggle.

Data Cleaning: The code addresses missing values (NaNs) in the DataFrame by replacing them with the mean value of each respective column. This ensures that the dataset is complete and ready for analysis. Filling missing values with the mean is a common approach in data cleaning to preserve the overall distribution of the data while minimizing the impact of missing data.

Data Transformation: By filling missing values with the mean, the code transforms the dataset by replacing missing values with meaningful estimates. This allows for a more comprehensive analysis of the dataset, as missing values no longer hinder computations or visualizations. Additionally, the transformation ensures that the dataset is in a suitable format for further analysis or modeling, as missing values are handled appropriately.

3.SOURCE CODE

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
main_df = pd.read_csv("/content/water_potability.csv")
df = main_df.copy()
```

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

```
ax = sns.countplot(x = "Potability",data= df, saturation=0.8)
plt.xticks(ticks=[0, 1], labels = ["Not Potable", "Potable"])
plt.show()
```

```
df['ph'] = df['ph'].fillna(df['ph'].mean())
df['Sulfate'] = df['Sulfate'].fillna(df['Sulfate'].mean())
df['Trihalomethanes'] = df['Trihalomethanes'].fillna(df['Trihalomethanes'].mean())
df.head()
```

4.RESULTS

4.1 Logistic Regression

```
X = df.drop('Potability', axis=1)
y = df['Potability']
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X = scaler.fit_transform(X)
X
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report

model_lg = LogisticRegression(max_iter=120, random_state=0, n_jobs=20)
# Training Model
model_lg.fit(X_train, y_train)
```

```
pred_lg = model_lg.predict(X_test)
# Calculating Accuracy Score
lg = accuracy_score(y_test, pred_lg)
print(lg)
```

```
0.6284658040665434
```

The accuracy point is 0.6284658040665434

4.2 Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
# Creating model object
model_dt = DecisionTreeClassifier( max_depth=4, random_state=42)
# Training Model
model_dt.fit(X_train,y_train)
pred_dt = model_dt.predict(X_test)
# Calculating Accuracy Score
dt = accuracy_score(y_test, pred_dt)
print(dt)
```

```
0.6451016635859519
```

The accuracy point is 0.6451016635859519

4.3 Random Forest

```
from sklearn.ensemble import RandomForestClassifier
# Creating model object
model_rf = RandomForestClassifier(n_estimators=300,min_samples_leaf=0.16, random_state=42)
# Training Model
model_rf.fit(X_train, y_train)
# Making Prediction
pred_rf = model_rf.predict(X_test)
# Calculating Accuracy Score
rf = accuracy_score(y_test, pred_rf)
print(rf)
```

```
0.6284658040665434
```

The accuracy point is 0.6284658040665434

4.4 K Neighbors

```
from sklearn.neighbors import KNeighborsClassifier
# Creating model object
model_kn = KNeighborsClassifier(n_neighbors=9, leaf_size=20)
# Training Model
model_kn.fit(X_train, y_train)
# Making Prediction
pred_kn = model_kn.predict(X_test)
# Calculating Accuracy Score
kn = accuracy_score(y_test, pred_kn)
print(kn)
```

```
0.6534195933456562
```

The accuracy point is 0.6534195933456562

4.5 Support Vector Machine (SVM)

```
from sklearn.svm import SVC, LinearSVC
model_svm = SVC(kernel='rbf', random_state = 42)
model_svm.fit(X_train, y_train)
# Making Prediction
pred_svm = model_svm.predict(X_test)
# Calculating Accuracy Score
sv = accuracy_score(y_test, pred_svm)
print(sv)
```

```
0.6885397412199631
```

The accuracy point is 0.6885397412199631

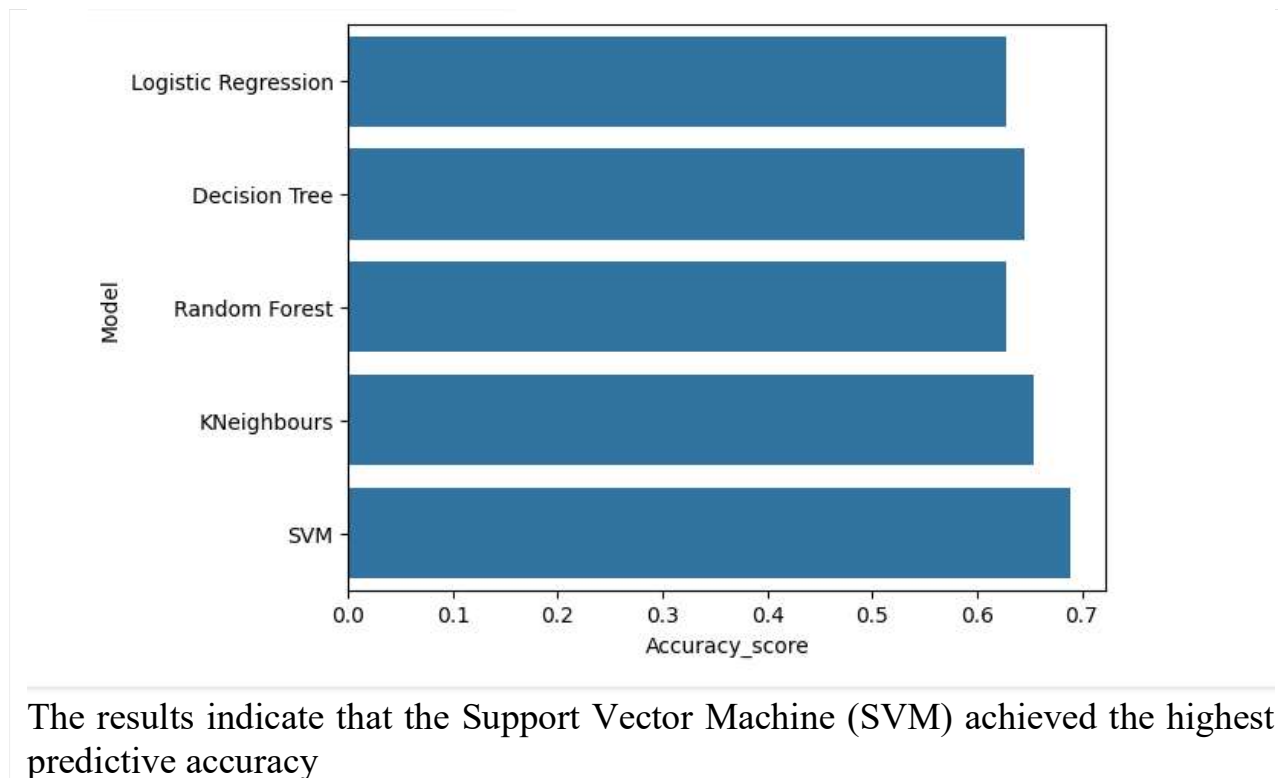
Water Quality Analysis

```
models = pd.DataFrame({
    'Model': ['Logistic Regression', 'Decision Tree', 'Random Forest', 'KNeighbours', 'SVM'],
    'Accuracy_score': [lg, dt, rf, kn, sv]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

models.sort_values(by='Accuracy_score', ascending=False)
```

	Model	Accuracy_score
4	SVM	0.688540
3	KNeighbours	0.653420
1	Decision Tree	0.645102
0	Logistic Regression	0.628466
2	Random Forest	0.628466

Water Quality Analysis



5. CONCLUSION

Based on the comparative analysis of five machine learning algorithms for water quality prediction, the results indicate that Support Vector Machines (SVM) achieved the highest predictive accuracy, followed by K-Nearest Neighbors (KNN) and Decision Trees. Conversely, Logistic Regression and Random Forest exhibited lower predictive performance in this context.

The superiority of SVM in accuracy suggests its effectiveness in capturing complex relationships within water quality datasets, particularly in scenarios where the decision boundaries are nonlinear and high-dimensional. KNN, despite its simplicity, also demonstrated competitive performance, highlighting its ability to adapt to local data structures and patterns. Decision Trees, while performing reasonably well, may have limitations in capturing intricate relationships compared to SVM and KNN.

The lower accuracy observed with Logistic Regression and Random Forest could be attributed to their inherent assumptions and limitations in handling nonlinear relationships or capturing complex interactions among variables in the dataset.

In conclusion, the findings emphasize the importance of selecting appropriate machine learning algorithms tailored to the characteristics of the dataset and the complexity of the problem at hand. While SVM emerges as the top performer in this study, KNN and Decision Trees offer viable alternatives, depending on the specific requirements and constraints of the application. Nonetheless, further exploration and fine-tuning of model parameters may be warranted to enhance the predictive performance of all algorithms, particularly those exhibiting lower accuracy. Overall, this research contributes valuable insights into the selection and application of machine learning techniques for water quality prediction, paving the way for improved decision-making and management of water resources in diverse environmental settings.

6. LITERATURE REVIEW

SI. NO	TITLE	AUTHOR	YEAR OF PUBLICATIONS	SUMMARY
1.	Comparison of Water Quality Classification Models using Machine Learning	Radhakrishnan , Neha, and Anju S. Pillai.	2020	Water quality classification models were compared using machine learning algorithms, with the Decision Tree algorithm achieving the highest accuracy of 98.50%. Another study compared five classification algorithms and found that the Lazy model using the K Star algorithm had an accuracy of 86.67% in classifying water quality. Based on the analysis of water quality detection using machine learning models, the Decision Tree algorithm was found to be the most suitable classification model for labeling the quality class of water

2.	Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: A case study in the Tri An Reservoir, Vietnam	Nguyen, Hao Quang, et al.	2021	<p>The study tested the performance of different machine learning models for estimating chlorophyll-a (Chl-a) concentration in the Tri An Reservoir, Vietnam, using Sentinel-2 MSI data and in situ water quality measurements</p> <p>Support vector machine (SVM) is a commonly used supervised kernel-based algorithm for various learning problems, including ecology and environmental management</p>
3.	Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models	Shah, Muhammad Izhar, et al.	2021	<p>The study investigated the predictive performance of gene expression programming (GEP), artificial neural network (ANN), and linear regression model (LRM) for modeling monthly total dissolved solids (TDS) and specific conductivity (EC) in the upper Indus River at two outlet stations</p>
4.	Water Quality Evaluation of the Yangtze River in China Using Machine Learning Techniques and Data Monitoring on Different Time Scales	Zhenzhen Di, Miao Chang and Peikun Guo	2019	<p>Machine learning techniques were used to assess water quality in the Yangtze River, China, based on weekly and real-time data. The study revealed variable spatiotemporal distribution characteristics for water quality and pollutants, providing insights for environmental decision support systems and watershed management</p>

5.	Machine Learning Methods for Better Water Quality Prediction	Ahmed, Ali Najah, et al,	2019	Machine learning methods such as Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN), and Multi-Layer Perceptron Neural Networks (MLP-ANN) have been implemented for water quality prediction. Artificial Neural Network (ANN) is used to evaluate and assess the correlation among water quality parameters based on experimental data.
----	--	--------------------------	------	---

7. REFERENCE

1. Radhakrishnan, Neha, and Anju S. Pillai. "Comparison of water quality classification models using machine learning." *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020.
2. Shah, Muhammad Izhar, et al. "Predictive modeling approach for surface water quality: development and comparison of machine learning models." *Sustainability* 13.14 (2021): 7515.
3. Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." *Journal of Hydrology* 578 (2019): 124084.
4. Nguyen, Hao Quang, et al. "Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: A case study in the Tri An Reservoir, Vietnam." *Water Environment Research* 93.12 (2021): 2941-2957.
5. Di, Zhenzhen, Miao Chang, and Peikun Guo. "Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales." *Water* 11.2 (2019): 339.