Water Quality Analysis using Machine Learning

Neelam Khaire
10634691

Dissertation submitted in partial fulfilment
of the requirements for the degree
of Masters in Business Analytics at Dublin Business School

Supervisor : Mr. Yalemisew Abgaz

January 2024

Declaration Page

I declare that this dissertation that I have submitted to Dublin Business School for the award of Master of Science in Business Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Neelam Khaire

Student Number: 10634691

Date: 8th January 2024

Acknowledgment

I wish to convey my profound appreciation to all individuals who have made valuable contributions to the successful completion of this dissertation. Their assistance, motivation, and direction have been valuable throughout this difficult yet fulfilling endeavour.

I would like to express my heartfelt gratitude to my dissertation advisor, Mr. Yalemisew Abgaz, whose consistent encouragement and exceptional expertise have played a crucial role in shaping the direction of this research. The guidance, perceptive critique, and commitment to acadamic achievement have been the primary factors that have led to the successful accomplishment of this dissertation.

I would like to express my sincere thanks to DBS for creating a favourable academic atmosphere and granting me access to resources that greatly facilitated the research process. The faculty, staff, and fellow students who provide support have established a dynamic academic community that has enhanced my educational journey.

I express sincere gratitude to my family for their constant encouragement as well as understanding throughout this educational endeavour. Their unwavering support and confidence in my abilities have consistently fuelled my motivation. I express heartfelt thanks to my friends and colleagues who have given me support, shared their knowledge, and provided help whenever necessary. The camaraderie among us has enhanced the enjoyment and fulfilment of this academic endeavour.

Abstract

Water quality evaluation is crucial in environmental management, and utilising machine learning models improves the accuracy of predictions. This study aims to compare different machine learning models for predicting water quality before and after the monsoon season in Telangana. The dataset used in this research was obtained from the Telangana Ground Water Department. The chosen models, namely Random Forest Classifier (RFC), Support Vector Classifier (SVC), Multi-layer Perceptron (MLP), Stochastic Gradient Descent (SGD), and KNeighborsClassifier, are assessed with a specific focus on imbalanced data using Principal Component Analysis (PCA) as the model was giving perfect score due to being imbalanced which was incomparable and incorrect. The effectiveness of the models is evaluated by employing essential performance metrics, including recall, precision, and F1 score as the accuracy does not work well with imbalanced data.

The pre-monsoon results indicate that RFC performs exceptionally well, with a recall of 0.988 and precision of 0.900. The monsoon transition has had a noticeable effect on RFC, as it continues to perform exceptionally well in the post-monsoon period, with an improved recall rate of 0.996 and precision rate of 0.993. SVC, SGD and MLP demonstrate consistent and strong performance in both time periods, demonstrating their ability to adapt. Notably, the KNeighborsClassifier demonstrates enhancement after the monsoon season, highlighting its sensitivity to seasonal changes.

The analysis of seasonal variations was performed with help of T-test on the machine learning model performance’s. RFC demonstrates consistent excellence. The comparative analysis enhances the scientific comprehension of machine learning models in predicting water quality, providing practical implications for environmental scientists, policymakers, and stakeholders involved in water resource management.

Table of Contents

Table of Figures

Chapter 1. Introduction

Water, an essential resource for sustaining life, experiences fluctuations in its quality due to a variety of environmental factors. Among this set, the shift from the pre-monsoon to post-monsoon periods poses distinct challenges and opportunities for comprehending the dynamics of water quality. A comprehensive approach to water quality assessment is required due to the intricate interaction between climatic patterns, land use variations, and ecological responses that occur during these seasons. This study aims to conduct a thorough investigation of water quality fluctuations during the pre and post-monsoon periods, utilising sophisticated machine learning models. The objective is to understand the complexities of these seasonal changes and improve our ability to predict them by using advanced modelling techniques.

The shift from pre-monsoon to post-monsoon brings about significant changes in environmental conditions, which have a notable impact on the dynamics of water quality. The pre-monsoon period is marked by high temperatures, dry conditions, and the occurrence of atmospheric changes that indicate the approaching monsoon season. During the post-monsoon period, we observe the consequences of significant rainfall, changes in soil moisture, and the possibility of runoff effects. These changes in water quality parameters, such as pH, turbidity, and pollutant concentrations, can be caused by shifts. Comprehending and forecasting these fluctuations are crucial for efficient water resource administration and ecological guardianship.

- **Importance of Water Quality Evaluation:**

Evaluating the quality of water is a crucial component of monitoring the environment, as it has a direct influence on ecosystems, human well-being, and the long-term viability of water sources. Accurate forecasting and comprehension of fluctuations in water quality across seasons is essential for making well-informed decisions, crafting policies, and devising adaptive management approaches. Conventional monitoring techniques, although useful, might not fully capture the intricacy of ever-changing environmental conditions. Consequently, there is an increasing demand to incorporate sophisticated technologies, such as machine learning, in order to improve the accuracy and flexibility of water quality forecasts.

- **The Role of Machine Learning Models:**

The utilization of machine learning models, which possess the ability to analyse extensive and intricate datasets, presents a hopeful opportunity for comprehending and forecasting fluctuations in water quality. These models have the ability to perceive complex patterns and connections within data, offering valuable insights that traditional statistical methods may fail to capture. This study aims to assess the effectiveness of different machine learning models in accurately capturing and interpreting fluctuations during the crucial pre and post-monsoon periods.

- **Models Under Evaluation:**

The machine learning models chosen for this comparative analysis are selected based on their ability to adapt and perform well in environmental prediction tasks. The following models are included:

- **Random Forest Classifier (RFC)**: is an ensemble learning algorithm that is highly effective in dealing with intricate datasets and delivering precise predictions.
- **Support Vector Classifier (SVC):** SVC, especially in high-dimensional spaces, excels at capturing complex patterns in the data.
- **Stochastic Gradient Descent (SGD) :** is a highly efficient optimisation algorithm for model parameters, particularly suitable for handling large-scale datasets. SGD is renowned for its exceptional speed.
- **The Multi-layer Perceptron (MLP):** is an artificial neural network that can effectively learn intricate relationships within data, enabling it to easily adapt to various environmental patterns.

- **The KNeighborsClassifier:** is a model for instance-based learning that classifies data points by considering the majority class of their k-nearest neighbours. This model is particularly useful for environmental classification tasks due to its robustness.

Each model possesses distinct strengths and capabilities for analysis, and the comparative evaluation seeks to determine their individual performances in relation to seasonal variations in water quality.

➢ **Integration of Principal Component Analysis (PCA):**

To address the potential difficulties caused by imbalanced datasets, which are particularly prevalent in environmental studies, this research employs Principal Component Analysis (PCA) in the initial four models: RFC, SVC, SGD, and MLP. PCA functions as a method to reduce the number of dimensions in a dataset, improving the models' capacity to capture important information while minimising the influence of uneven class distributions. This integration strategy recognises the intricacies inherent in environmental datasets and emphasises a dedication to improving model performances for practical use.

➢ **Possible results and consequences:**

Predicting the results of this analysis has important consequences for the management of water resources, the formulation of environmental policies, and the promotion of sustainable development. Precise forecasts of water quality fluctuations during the pre and post-monsoon periods provide decision-makers with timely information to implement adaptive management strategies. Gaining comprehension of the models' capacity to adjust to seasonal variations offers valuable understanding into their resilience and dependability, which are pivotal elements in practical scenarios.

- **Progressing the Field of Environmental Science:**

  This study not only has immediate applications, but also contributes to the advancing field of environmental science by expanding the limits of predictive modelling. The incorporation of machine learning, along with dimensionality reduction techniques such as PCA, indicates a deviation from conventional methods. This research initiates an investigation into the capabilities of these advanced techniques to decipher intricate environmental patterns, laying the groundwork for future advancements in the evaluation of water quality.

- **T-Test for comparative analysis:**

  Monitoring the quality of water is an essential component of managing the environment, and it is important to comprehend the variations that occur throughout different seasons in order to make informed decisions. Integrating machine learning techniques can improve the comparative analysis of seasonal water quality, resulting in a more comprehensive and data-driven approach. The t-test is a frequently used statistical tool in this context. It is used to evaluate the significance of differences between datasets that vary seasonally.

  The purpose of the comparative analysis is to determine and measure the influence of seasonal variations on the parameters of water quality. Comprehending these discrepancies is crucial for the purpose of environmental monitoring, resource management, and policy development. Machine learning enhances traditional statistical methods by offering predictive models capable of capturing intricate patterns and trends within extensive datasets. The t-test is a statistical technique employed to compare the means of two groups and ascertain whether the observed disparities are statistically significant. T-tests can be used in seasonal water quality analysis to determine if there are significant variations in important parameters between different seasons, such as the pre-monsoon and post-monsoon periods.

1.1 AIM AND RATIONAL OF THE RESEARCH

1.1.1 AIM

The main objective of this research is to thoroughly examine the fluctuations in water quality during the periods before and after the monsoon season, using sophisticated machine learning models. The study aims to assess the effectiveness of different machine learning algorithms in forecasting water quality, identify any seasonal fluctuations, and offer valuable insights to support informed decision-making in water resource management

1.1.2 RATIONAL

The quality of water is a crucial factor that determines the state of environmental health and the overall well-being of humans. Comprehending and overseeing the quality of water is crucial in guaranteeing the availability of secure and lasting water resources. This study aims to improve our ability to forecast variations in water quality, which is currently a critical requirement. Seasonal variations, specifically the shift from the pre-monsoon to post-monsoon periods, have a substantial impact on environmental conditions. This study acknowledges the significance of evaluating the effects of seasonal fluctuations on water quality, as it directly affects the health of ecosystems, human consumption, and overall strategies for water management.

Utilising machine learning models enables a more sophisticated and anticipatory comprehension of water quality dynamics. This research aims to investigate the potential of advanced technologies, such as Multi-layer Perceptron (MLP),Random Forest Classifier (RFC), , Stochastic Gradient Descent (SGD), and KNeighborsClassifier, Support Vector Classifier (SVC)in predicting water quality. The lack of equilibrium in water quality data presents a formidable obstacle for precise predictive modelling. The incorporation of Principal Component Analysis (PCA) in the initial four models is intended to address this difficulty by decreasing dimensionality and improving the models' capacity to identify patterns in imbalanced datasets.

The selection of Telangana as the research location provides a precise geographical context. The study aims to offer valuable insights, both academically and practically, for water resource management in the specified area by concentrating on a region with unique hydrological characteristics. The study employs a comprehensive methodology by assessing multiple machine learning models and taking into account various performance metrics, including recall, precision, and F1 score. The thorough assessment enables a nuanced comprehension of the merits and drawbacks of each model, facilitating well-informed decision-making.

In the end, the research aims to make a valuable contribution to the implementation of sustainable water management practices. The study seeks to offer practical insights for policymakers, environmental scientists, and water resource managers to make well-informed decisions regarding the preservation and improvement of water quality. This will be achieved by gaining a comprehensive understanding of seasonal variations and utilising advanced predictive models.

1.2 RESEARCH QUESTION:

1. What is the predictive performance of various machine learning models such as the Random Forest Classifier (RFC), Support Vector Classifier (SVC), Multi-layer Perceptron (MLP), and Stochastic Gradient Descent (SGD) in evaluating water quality based on available data ?
2. When comparing the performance of the selected machine learning models on Pre and Post the monsoon, what differences in predictive accuracy are seen?

1.3 RESEARCH OBJECTIVE:

The main objective of the research is to use machine learning techniques to detect and avoid potential health concerns which are connected with poor quality of water and results in disease outbreak. This research aims at addressing key challenges in water quality analysis through the application of machine learning. Primarily , the focus is on developing accurate predictive models for essential water quality parameters which further extends to comparing performance of different machine learning algorithms in predicting these parameters. Secondly to improve the resilience and accuracy of machine learning models for water quality prediction by optimising feature selection.

Next, machine learning approaches have the capacity to evaluate intricate data on water quality and offer insightful analysis for determining the safety and suitability of water for diverse uses. The research intends to support the efficient management of water resources by investigating the efficacy of machine learning techniques in the classification of water quality. Next, to evaluate the safety of water & find the trends and changes in water quality parameters the research will use machine learning techniques on data collected on water quality pre and post monsoon.Further, a comparative analysis of water quality of pre and post monsoon to identify variations and trend can be conducted using T-Test to predict water quality for both pre and post monsoon.

1.4 SCOPE OF THE PROJECT: The project scope includes feature selection, machine learning model development, gathering and preprocessing datasets on water quality from pertinent sources, and a thorough comparison of water quality parameters. The study will concentrate on a particular area or body of water that experiences monsoons.

1.5 JUSTIFICATION

The increasing need for effective water quality monitoring, the complexity of aquatic environments, the integration of explainable AI and machine learning, the comparisons with conventional techniques, and performance evaluation all support the use of machine learning for pre- and post-water quality analysis. Through this study, an attempt is made to support well-informed decision-making in the management of water resources and further the field's understanding of water quality analysis.

- Increasing demand for efficient water quality monitoring

The immense amount of wastewater generated by increasing industrialization and urbanisation poses significant threats to natural water resources, making it critical to develop efficient and accurate methods for water quality monitoring and prediction.1 . Machine learning has emerged as an effective tool for addressing these issues, with several advantages over traditional models used in water-related research.

- ML for water quality monitoring
  Numerous algorithms have shown encouraging results in water environment research1, demonstrating the widespread application of machine learning in the monitoring and prediction of water quality. In this study, more precise and effective models can be created for pre- and post-water quality analysis by utilising these methods.

- Combining explainable AI and machine learning
  The application of explainable AI methods, like H2O AutoML, to improve the interpretability and transparency of machine learning models has been the subject of recent studies 2. By ensuring that the general public can readily comprehend the reasoning behind predictions, this integration can contribute to the increased acceptance and credibility of machine learning in the analysis of water quality.

- Comparison with traditional methods
  In the classification of water quality 2, machine learning-based models have been compared with conventional techniques like Bayesian and neural networks. By highlighting the advantages and disadvantages of each strategy, these comparisons can help researchers create more useful models for pre- and post-water quality analysis.

- Performance evaluation:
  Different machine learning classifier algorithms can be used to assess the performance of water quality index (WQI) models. 4. Researchers can determine the best methods for forecasting

changes in water quality and provide guidance for water resource management decision-making processes by evaluating the performance of these models.

1.6 ETHICS AND LIMITATIONS

1.6.1 Ethical issues :

Ensuring data privacy is of utmost importance. Data on water quality may contain personally identifiable information pertaining to individuals or organisations. To safeguard privacy, it is imperative to acquire informed consent, anonymize or consolidate data, and adhere to data protection protocols. The presence of fairness and absence of bias are crucial as incomplete or biassed data can lead to unjust predictions that disproportionately disadvantage specific demographic groups. This issue can be resolved by incorporating algorithms that possess an understanding of fairness, performing assessments after the model is created to ensure fairness, and striving for equitable outcomes by evaluating and eliminating bias..

Lack of accountability poses an additional challenge. Despite the ambiguous framework for accountability, decisions derived from machine learning predictions can have significant and wide-ranging consequences. Establishing clear responsibilities, defining anticipated outcomes, and formulating protocols for rectifying mistakes all foster a heightened level of accountability.

1.6.2 Limitations:

A primary constraint is the limited amount and calibre of the accessible data. Inadequate or low-quality data may result in the production of unreliable models and forecasts. The study relies on historical water quality data obtained from the Telangana Ground Water Department. The accuracy and thoroughness of the predictive models depend on the excellence and entirety of this dataset. The models' ability to accurately represent the complexities of water quality dynamics may be compromised by potential deficiencies in data collection methods or the lack of specific water quality parameters. An additional challenge that arises, especially in classification tasks, is the presence of data imbalance. An imbalanced distribution of classes can skew the performance of the model. The findings are intrinsically limited to the geographical context of Telangana, and it is advisable to exercise caution when generalising the results to other regions with unique environmental characteristics. The regionalized characteristics of factors affecting water quality introduce variability that may not be universally applicable. Temporarily, the scope of the study is limited to the historical data that is currently available, which may not capture short-term fluctuations and dynamic shifts in water quality during the periods pre and post the monsoon season.

The complexity of models, especially in ensemble methods such as the Random Forest Classifier, poses difficulties in interpretability, which is crucial for establishing trust and comprehension among stakeholders. The study's predictive models rely on assumptions regarding the consistency of relationships between input features and water quality outcomes. However, these assumptions may not be valid when environmental conditions change or new factors emerge that were not taken into account in the study. The models do not explicitly consider external factors, such as human activities, changes in land use, or regulatory interventions, that may affect water quality. In addition, the study recognises ethical considerations related to data privacy, consent, and responsible data usage. However, it may have limitations in its ability to anticipate and address all possible ethical concerns. Validation from external sources using independent datasets could strengthen the reliability of the results, leading to a more thorough and dependable comprehension of water quality changes. Notwithstanding these acknowledged constraints, the study remains a valuable addition to the field, providing insights into the prediction of water quality using machine learning and opening up possibilities for further improvement and investigation.

Chapter 2. Literature Review

2.1 Introduction: -

Water quality analysis is key for averting waterborne disease and safeguarding water sources. This literature review explores how machine learning techniques can be used to provide comprehensive overview of recent development and the challenges in comparison to traditional methods which are often time consuming.

This review focuses on recent studies (within the last five years) that apply machine learning techniques to analyse water quality. The goal is to investigate the usefulness of machine learning in forecasting water quality measures and to identify prospective areas for future research.

In Telangana, India, groundwater is essential to domestic water supply, industrial processes, and agricultural activities. However, a number of factors, such as urbanisation, industrial effluents, and agricultural practices, have contributed to Telangana's declining groundwater quality. Groundwater quality is greatly impacted by seasonal variations and rainfall patterns during the pre-monsoon and post-monsoon seasons, which is when this degradation is most noticeable.

In Telangana, machine learning (ML) techniques have become a promising tool for assessing and forecasting groundwater quality. Large datasets of groundwater parameters can be handled by ML algorithms with ease, and they can reveal patterns and relationships that conventional statistical methods might not be able to see right away. Additionally, ML algorithms can shed light on the variables influencing variations in groundwater quality between the pre- and post-monsoon seasons.

A comprehensive search was conducted in leading academic database (IEEE Xplore, Science direct, Google Scholer) using keywords such as "Water Quality Analysis", "Water Quality", “machine learning”, “predictive modelling”. Studies with a focus on use of Machine learning while doing Water Quality Analysis.

2.2 Previous Studies

This literature research is organized based previous studies which includes :

- **Importance of Water Quality Assessment-**

  The underlines the significance of water quality evaluation and prediction in the development of water conservation initiatives was emphasized in may of the studies [3][5] This study's literature analysis emphasizes the need of precise water quality forecast in assisting the creation of long-term water conservation efforts. The work intends to improve understanding of water quality dynamics by utilizing machine learning approaches, hence helping to effective water resource management.

- **Automated Water Quality Analysis** –

  Investigates the possibility of utilizing Automatic Machine Learning (AutoML) to reduce the time necessary for water quality analysis and provide more accurate findings was mentioned in study done [2]. This study emphasizes the potential for automation in water quality analysis, which can lead to more efficient and fast water quality monitoring. The project intends to expedite the water quality analysis process by combining machine learning and AutoML approaches, demonstrating the practical application of machine learning in tackling the issues of manual data processing and reporting.

- **Groundwater Analysis and Prediction-**

  The broad application prospects of machine learning in groundwater analysis, including groundwater quality and pollution assessment and prediction was underlined in the study (Zhu et al., 2022).[1] The paper addresses various machine learning algorithms and their possible uses in assessing and predicting groundwater quality. This demonstrates the importance of machine learning in comprehending and predicting groundwater quality, which is critical for guaranteeing the supply of safe and clean drinking water.

- **Water Quality Prediction Models**

  The study[12] focuses on the application of machine learning to construct a water quality prediction model based on water quality measurements. This study demonstrates the power of machine learning in developing prediction models for water quality monitoring. The work intends to increase the accuracy of water quality prediction by utilizing machine learning techniques, which is critical for informed ML for Groundwater Suitability for Irrigation in Pre- and Post-Monsoon

- **ML-Based Strategies for Water Quality Evaluation**

  The study [1] provides a variety of machine learning-based water quality analysis strategies. This review emphasizes the potential of machine learning techniques in addressing the challenges associated with declining water quality by highlighting the diverse applications of machine learning in water quality evaluation. The study advances methodologies for comprehensive water quality analysis by investigating various ML-based strategies.

- **ML for Groundwater Suitability for Irrigation in Pre- and Post-Monsoon**

  In Telangana and other Indian states, the use of ML algorithms for assessing groundwater quality has been the subject of several studies. ML algorithms were used by Dimple et al. (2022a, 2023b) to evaluate the suitability of groundwater for irrigation in the semi-arid Nand Samand catchment of Rajasthan, India. During the pre-monsoon and post-monsoon seasons, they collected groundwater samples and analysed a range of water quality indices, such as sodium adsorption ratio (SAR), electrical conductivity (EC), sodium percentage (Na%), and residual sodium carbonate (RSC). To determine whether groundwater is suitable for irrigation, machine learning models such as gradient boosting trees and random forests were employed.

The findings showed that ML models could successfully categorise groundwater quality according to chemical parameters and offer insightful advice on irrigation management.

- **ML for Assessing Groundwater Quality in Rohtak**

  Raheja et al. (2023) evaluated the groundwater quality's suitability for irrigation and drinking in the Rohtak district of Northern India by looking at its spatial variability. EC, pH, TDS, and major cations and anions were among the parameters examined in pre- and post-monsoon groundwater samples. Machine learning (ML) techniques, such as support vector machines (SVM) and k-nearest neighbours (KNN), were utilised to forecast groundwater quality indices and categorise groundwater into distinct classes according to its appropriateness for irrigation and drinking. The study came to the conclusion that machine learning (ML) models could be used to evaluate groundwater quality and offer practical advice on resource management.

- **Comparing Boosting Algorithms for Groundwater Quality Prediction**

  Divy Jot (2023) conducted a study to evaluate the efficacy of various boosting algorithms in forecasting Telangana's groundwater quality by utilising pre- and post-monsoon data. The study assessed how well the Artificial Neural Network (ANN), Random Forest (RF), and Light Gradient Boosting Machine (LGBM) models predicted groundwater quality indices and categorised groundwater samples according to their suitability for irrigation and drinking. The findings showed that the LGBM and RF models outperformed the ANN, indicating that Telangana's groundwater quality could be accurately assessed using these models.

2.3 Conclusion

In areas like Telangana where seasonal variations have a substantial impact on groundwater characteristics, machine learning has emerged as a potent tool for analysing and forecasting groundwater quality. Large datasets of groundwater parameters can be handled by ML algorithms with ease. They can also be used to uncover intricate connections between environmental variables and groundwater quality and offer insightful information for resource management and decision-making. It is anticipated that as machine learning (ML) technology advances, its application in assessing groundwater quality will grow, producing more precise and useful data for Telangana's and other regions' sustainable water resource management.

Chapter 3. Methodology

Various Machine Learning methods have been implemented over the years to assist researchers and analyst . An overview of the suggested method, which includes multiple stages including Data acquisition, Data pre-processing, Data Partitioning, Prediction of water quality using classification Models, Model evaluation is given in this chapter. The method involves predictive analysis of water quality analysis and comparative performance of different models on pre & post monsoon data.

3.1 Data Acquisition:

The performance of machine learning models is directly impacted by the quality of the data used to train them. The model's predictions are likely to be unreliable if the input data is biassed, erroneous, or incomplete. The precision of the labels is crucial in supervised machine learning, as the model is trained on labelled data. The model will learn patterns based on false information if water quality labels are inaccurate or misclassified, which will result in inaccurate predictions. Understanding the variables affecting water quality is frequently just as crucial to water quality analysis as prediction accuracy. Accurate data collection guarantees that the model's conclusions correspond with the real factors influencing water quality. Water quality data collection for regulatory or compliance purposes requires strict adherence to set standards and procedures. For some applications, non-compliance could make the data invalid or unusable.

Hence in the studies, the both pre and post-monsoon dataset is taken from Telangana open data portal. This dataset provides detailed information about the Pre-Monsoon & Post-monsoon Water Quality details in Telangana. To make sure the data is correct, comprehensive, and appropriate for additional analysis, data cleaning and preprocessing must be done as soon as it has been gathered. This could entail dealing with missing values, fixing mistakes in the data, and eliminating outliers.

3.2 Data Preprocessing:

The process of cleaning, organising, and converting raw data into a format that can be used for training and assessing machine learning models is known as data preprocessing, and it is a crucial and essential step in the machine learning pipeline. This is a critical step that will improve the input data's quality and dependability, which will ultimately impact the machine learning algorithms' performance, accuracy, and capacity for generalisation. Preprocessing data is necessary because real-world datasets have inherent difficulties and flaws. A variety of problems, including missing values, outliers, inconsistencies, and a variety of data types, are frequently present in raw data. Preprocessing the data to meet these requirements is essential because machine learning models have particular demands on the distribution, scale, and format of the data.

3.3 Data Processing Steps

3.3.1 Data Cleaning: Data cleaning is a fundamental component of data preprocessing that involves locating and correcting anomalies in the dataset. Incomplete or missing data can have a negative effect on model prediction and training. Imputation techniques are utilised to fill in these gaps and ensure a more comprehensive and informative dataset by using statistical measures like mean or median to fill in the missing values.

3.3.2 Data Transformation: Another important element is data transformation, which involves normalising or standardising numerical features. Larger magnitude variables are less likely to dominate the learning process when features are scaled to a consistent range, such as [0, 1].

```
# Normalizing numerical features so that each feature has mean 0 and variance 1
feature_scaler = StandardScaler()
X_scaled = feature_scaler.fit_transform(X)
```

Figure 1 StandardScaler

It is also essential to handle categorical data because machine learning models usually need numerical inputs. Methods like one-hot encoding or label encoding are used to effectively transform categorical variables into a numerical format so the model can process them. In dataset used for this study there were 3 columns (category, Classification.1, season) which were in categorical form and was converted in numerical using StandardScaler

```
# Converting Categorical features into Numerical features
data['category'] = data['category'].map({'Safe':0, 'Unsafe':1})
data['Classification.1'] = data['Classification.1'].map({'P.S.': 1, 'U.S.': 0,'MR': 0})
data['season'] = data['season'].map({'premonsoon 2022': 1,})
```

Figure 2 F Categorial to Numerical

Another important factor to consider is handling data that is not balanced. Resolving this imbalance is crucial to stop models from favouring the majority class in classification tasks with unequal class distributions. To create a dataset that is more balanced, strategies like oversampling the minority class, undersampling the majority class, or using synthetic data generation techniques SMOTE (Synthetic Minority Over-sampling Technique). are used.

A more complex kind of data preprocessing called "feature engineering" involves creating new features or modifying already-existing ones in order to better capture the underlying patterns in the data. To extract more pertinent information, this may entail combining or transforming features, which will improve the model's capacity to recognise complex relationships within the data.

3.3.3 Dimensionality Reduction : To manage datasets with a lot of features, dimensionality reduction techniques like Principal Component Analysis (PCA) are used. In addition to accelerating training, dimensionality reduction helps avoid overfitting and enhances the model's ability to generalise to new data. PCA does not directly address the issue of class imbalance. The algorithm operates on the entire dataset, with the goal of capturing the highest amount of variability in the features, without taking into account the distribution of classes.Imbalanced data occurs when one class is substantially more prevalent than the other(s). In datasets with imbalances, the representation of the minority class may be insufficient. During the application of PCA, the algorithm does not distinguish between different classes, which can result in a decreased focus on the minority class when reducing the dimensions.

```
# Implementing PCA to create principal components
from sklearn.decomposition import PCA
pca = PCA(n_components = 6)
pca.fit(X_scaled)
x_pca = pca.transform(X_scaled)
print("Variance explained by each of the n_components: ",pca.explained_variance_ratio_)
print("Total variance explained by the n_components: ",sum(pca.explained_variance_ratio_))
```

```
Variance explained by each of the n_components:  [0.30019314 0.16733577 0.08674684 0.07573301 0.0593832  0.05283609]
Total variance explained by the n_components:  0.7422280465239742
```

Figure 3 PCA Implementation

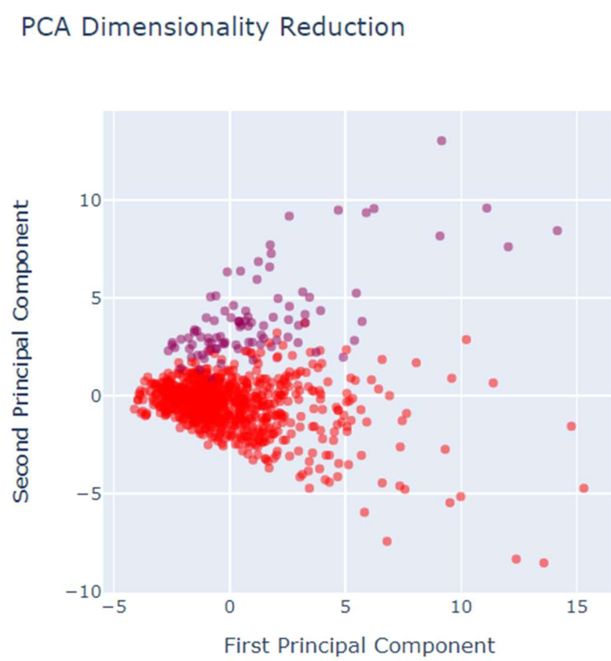

Figure 4 PCA Dimensionality Reduction

3.3.4 Missing Data: In machine learning, missing data is the term used to describe when certain features in a dataset are missing their values.Usually, it is shown as NA or NaN. It is important to handle missing data, and common methods include imputation, which replaces missing values with estimates such as the mean or median to ensure that the data is complete.

```python
# Assuming 'data' is your DataFrame
data['column_name'].fillna(data['column_name'].mean(), inplace=True)
```

Figure 5 Filling the missing data

Eliminating features or instances with missing values is another strategy, but it may result in information loss.

```python
# Remove rows with NaN values
data.dropna(inplace=True)
```

Figure 6 Dropping rows with NaN Values

More complex solutions are provided by advanced techniques like employing algorithms resistant to missing data or predictive modelling for imputation.

```python
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=2)
data = imputer.fit_transform(data)
```

Figure 7 Modeling for imputation

Machine learning models operate reliably and effectively when missing data is managed effectively.

Creating a clear, organised, and informative dataset is the ultimate aim of data preprocessing, as it lays the groundwork for productive machine learning results. It enables machine learning models to accurately predict or classify data, generalise to previously unseen instances, and learn from it. Data scientists constantly improve and optimise their preprocessing techniques to maximise the potential of their machine learning models, which emphasises the iterative and exploratory nature of data preprocessing. To summarise, data preprocessing is fundamental to the development of strong, dependable, and efficient machine learning models. It establishes the groundwork for significant discoveries and well-informed choices across an extensive array of uses.

3.4 Data Partitioning:

Data partitioning is an essential step in the machine learning workflow, sometimes referred to as data splitting or data splitting into training and testing sets. To aid in the training, validation, and assessment of machine learning models, a dataset is divided into two or more subsets. The training set and the testing set are the primary subsets that are frequently used. The validation set is an extra subset that is occasionally used for model tuning, below is the explanation of different subsets and there role :

- Training set: The machine learning model is trained using the training set. The model discovers patterns and relationships among the labels (output variable) and features (input variables) in this set during training. When it comes to size, the training set typically receives 70–80% of the available data.
- Test set: The trained model's performance is assessed using the testing set. To determine how well the model generalises to new, unseen examples, it is tested on the testing set's unseen data after it has been trained on the training set, the testing set get the remaining data i.e 20-20%.

The pre-processed data used in this study is split into training and testing sets. To train the machine learning models, use the training set. Use the testing set to evaluate how well the trained models generalize to new data.

Figure 8 Training & Test data

3.5 Data Description

Pre monsoon data set

Column	Discerptions
Sno	(Serial Number) is a set of integer values that serve as unique identifiers or serial numbers.
temp_id	Integer values that serve as temporary identifiers.
long_gis:	floating numbers representing longitudinal coordinates.
lat_gis	Numeric values that represent latitude coordinates with decimal points.
RL_GIS	Floating-point numbers that represent RL_GIS values.
season	Floating-point values that indicate the specific season.
pH	Whole numbers that indicate the acidity or alkalinity of a substance.
E.C	(Electrical Conductivity)Whole numbers that indicate the level of electrical conductivity.
TDS	(Total Dissolved Solids) Numeric values that indicate the total amount of solids that are dissolved in a substance.
CO3	(Carbonate) Whole numbers indicating the concentration of carbonates.
HCO3	(Bicarbonate) Whole number values that indicate the levels of bicarbonate in the body.
Cl	(Chloride) Whole numbers indicating the concentration of chloride.
F	(Fluoride) Floating-point values indicating the concentration of fluoride.
NO3	represents the floating-point numbers that indicate the levels of nitrate.
SO4	(Sulphate) Decimal numbers indicating the concentration of sulphate.
Na	(Sodium) Numerical values indicating the concentration of sodium.
K	(Potassium) Decimal numbers that indicate the concentration of potassium.
Ca	(Calcium) Numerical values indicating the concentration of calcium.
Mg	(Magnesium) Decimal numbers indicating the concentration of magnesium.
T.H	(Total Hardness) Decimal numbers indicating the overall level of hardness.
SAR	(Sodium Adsorption Ratio) Decimal numbers representing SAR values.
RSC meq/L	(Residual Sodium Carbonate)refers to the RSC values expressed in milliequivalents per litre as floating-point numbers.
Classification.1	Floating-point values that indicate a categorization.
Category	Floating-point numbers that represent a specific category.

Post Monsoon dataset

Column	Discerptions
Sno	(Serial Number) is a set of integer values that serve as unique identifiers or serial numbers.
district	It represents district the data taken for
long_gis:	floating numbers representing longitudinal coordinates.
lat_gis	Numeric values that represent latitude coordinates with decimal points.
RL_GIS	Floating-point numbers that represent RL_GIS values.
season	Floating-point values that indicate the specific season.
pH	Whole numbers that indicate the acidity or alkalinity of a substance.
E.C	(Electrical Conductivity)Whole numbers that indicate the level of electrical conductivity.

TDS	(Total Dissolved Solids) Numeric values that indicate the total amount of solids that are dissolved in a substance.
CO3	(Carbonate) Whole numbers indicating the concentration of carbonates.
HCO3	(Bicarbonate) Whole number values that indicate the levels of bicarbonate in the body.
Cl	(Chloride) Whole numbers indicating the concentration of chloride.
F	(Fluoride) Floating-point values indicating the concentration of fluoride.
NO3	represents the floating-point numbers that indicate the levels of nitrate.
SO4	(Sulphate) Decimal numbers indicating the concentration of sulphate.
Na	(Sodium) Numerical values indicating the concentration of sodium.
K	(Potassium) Decimal numbers that indicate the concentration of potassium.
Ca	(Calcium) Numerical values indicating the concentration of calcium.
Mg	(Magnesium) Decimal numbers indicating the concentration of magnesium.
T.H	(Total Hardness) Decimal numbers indicating the overall level of hardness.
SAR	(Sodium Adsorption Ratio) Decimal numbers representing SAR values.
RSC meq/L	(Residual Sodium Carbonate)refers to the RSC values expressed in milliequivalents per litre as floating-point numbers.
Classification.1	Floating-point values that indicate a categorization.
Category	Floating-point numbers that represent a specific category.

3.6 Prediction of water quality using classification Models

a. **Support Vector Machine (SVM) Model :-** Corinna Cortes and Vapnik created the SVM model in 1995. It offers a number of special advantages for nonlinear and high-dimensional pattern recognition as well as for solving small samples. It can be expanded to work in other machine learning problem simulations. It locates the required coefficients by using the hyperplane to divide the input vectors' points. The line with the largest margin—that is, the distance between the hyperplane and the closest input objects—is the best hyperplane. Support vectors are the input points that the hyperplane defines. (9) SVM has been effectively used in a number of domains, including the prediction of water quality. For example, SVM was used in a study to forecast the elements of the Tireh River's water quality in Iran. The results of the study showed that SVM had good predictive performance for water quality parameters, suggesting that machine learning could be used to analyse water quality. Furthermore, SVM's usefulness in this field was highlighted by a study that compared its performance with that of other machine learning algorithms for the classification of water quality and discovered that SVM achieved high accuracy in water quality prediction.(10)(11)

The hyperparameters "C" (regularisation parameter) and "kernel" (choice of kernel function) are adjusted for SVC. Only the 'rbf' kernel is taken into consideration in this instance, and the regularisation parameter 'C' is set to 1.

Model	Optimized Parameters	Optimized Values
SVC	kernel	rbf
	C	1

Figure 9 SVC Optimized Parameters

b. **Random Forest classifier** An ensemble of decision trees is used by the Random Forest Classifier (RFC), a machine learning algorithm, to solve regression and classification problems. Because of its high accuracy, flexibility, and simplicity, this algorithm is widely used. In order to generate the final output, RFC generates multiple decision trees on different subsets of the input data and then compiles the outcomes from each tree. Each decision tree in the model is constructed by the algorithm using bagging and feature randomness, which lessens overfitting and enhances the generalisation capabilities of the model. RFC is a potent machine learning algorithm that has been effectively used in e-commerce, healthcare, and finance, among other industries. For data professionals, its high accuracy, versatility, and ease of use make it an invaluable tool. The algorithm is a popular choice for resolving complex issues because it can handle both classification and regression problems, and it is effective in reducing overfitting and improving generalisation performance.

Using values 100, 150, 200, 250, and 300, the hyperparameter 'n_estimators' (the total number of trees in the forest) is adjusted for RFC.

Model	Optimized Parameters	Optimized Values
RFC	n_estimators	300

Figure 10 RFC Optimized Parameters

c. **Stochastic Gradient Descent (SGD):** In machine learning, the iterative optimisation technique known as stochastic gradient descent (SGD) is used to determine a model's parameters that minimise an objective function. Combining it with backpropagation makes it a potent technique that is frequently used in neural network training applications. SGD is a type of gradient descent that computes the gradient for a single training example at a time instead of comparing it to the full dataset. It is also referred to as batch gradient descent. This method aids in cutting down on computation time, p articularly for large datasets.

The four hyperparameters for SGD are tunned as :'max_iter' to 100 (maximum number of iterations), 'eta0' to 0.1(initial learning rate for the power_t scaling), 'alpha' to 0.01 (constant that multiplies the regularisation term), and 'l1_ratio' to 0.7(mixing parameter for elastic net regularisation).

Model	Optimized Parameters	Optimized Values
SGD	eta0	0.1
	max_iter	100
	alpha	0.01
	l1_ratio	0.7

Figure 11 SGD Optimized Parameters

d. **Multi-layer Perceptron (MLP)** An artificial neural network called a Multilayer Perceptron (MLP) is made up of fully connected neurons arranged in at least three layers (input, hidden, and output layers) and with a nonlinear activation function. A common feedforward artificial neural network example is MLP, which is employed extensively for a number of machine learning applications, such as regression and classification 2. MLP has been used in many domains, such as water quality analysis, where it is employed to forecast parameters related to water quality based on a variety of features. MLP is a useful tool for resolving complex problems in water quality analysis and management because of its capacity to learn intricate patterns and adapt to different types of data.

The hyperparameters for MLP that are adjusted are 'activation' (the hidden layer's activation function), 'learning_rate_init' to 0.01 (the initial learning rate), and'max_iter' was 500 (the maximum number of iterations).

Model	Optimized Parameters	Optimized Values
MLP	activation	relu
	learning_rate_init	0.01
	max_iter	500

Figure 12 MLP Optimized Parameters

e. **K-Nearest Neighbours Classifier (KNeighborsClassifier):** is a supervised, non-parametric learning classifier that classifies or predicts how a single data point will be grouped based on proximity 4. Since similar points can be found close to one another, it is usually applied to classification problems. In order to classify a test point according to the most frequently represented class among the k nearest neighbors1, KNeighborsClassifier first determines which of the k nearest data points to a given test point are present. The algorithm is a good option for many machine learning problems because it is easy to implement and has few hyperparameters.4 Despite some drawback, KNeighborsClassifier is a useful machine learning tool, especially for the analysis of water quality, as it can be applied to the classification of water quality data according to a variety of features. For instance, a study that employed KNeighborsClassifier to forecast a river's water quality produced encouraging results in terms of assessing the quality of the water based on a variety of characteristics 5.

3.7 Comparative of water quality analysis using T-Test:

A t-test is a statistical technique employed to ascertain the presence of a noteworthy disparity between the means of two distinct groups. It is especially beneficial when dealing with a limited sample size and conducting a comparison of means between two data sets to determine the likelihood of genuine differences or the possibility of chance occurrences. T-tests are essential for assessing the statistical significance of observed differences between two groups. They assist in the comparison of means and determining whether the differences are likely to indicate genuine effects or if they could have arisen randomly. T-tests exhibit stability even in the presence of small sample sizes, rendering them suitable for situations where alternative tests may not be appropriate.T-tests are extensively utilised in diverse disciplines, ranging from medical research to social sciences, to evaluate the statistical significance of experimental findings.

3.8 Evaluation Metric:

Evaluation matrices are usually used to asses the performances of various machine learning model. These metrics reveal how well a model performs in terms of predictions on a particular dataset. The sort of task (classification, regression, clustering, etc.) and the particular objectives of the analysis determine which evaluation metric is used.

3.8.1 Confusion Matrix: A table that details a classification model's performance is called a confusion matrix. It provides an overview of the quantity of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). The number of occurrences for a specific combination of predicted and actual class is represented by each cell in the matrix.

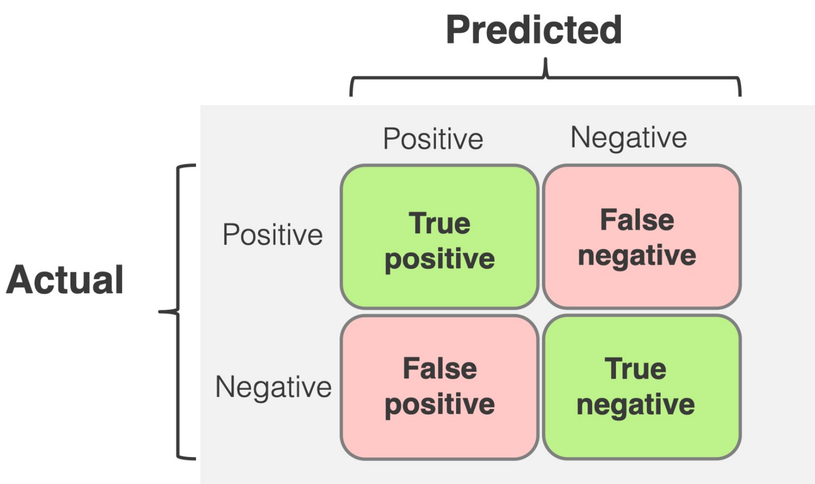

Figure 13 Confusion Matrix

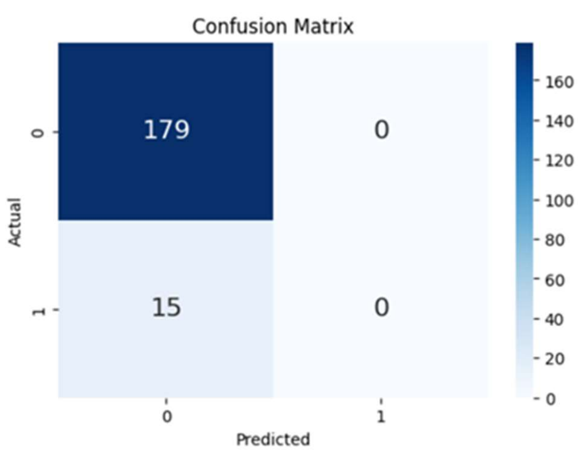

Figure 14 CM-Pre-Monsoon-Precision

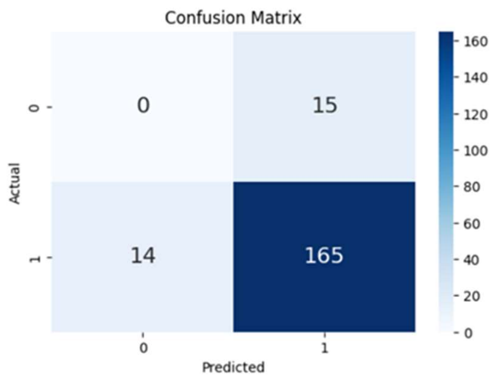

Figure 15 CM-Pre-Monsoon-Recall

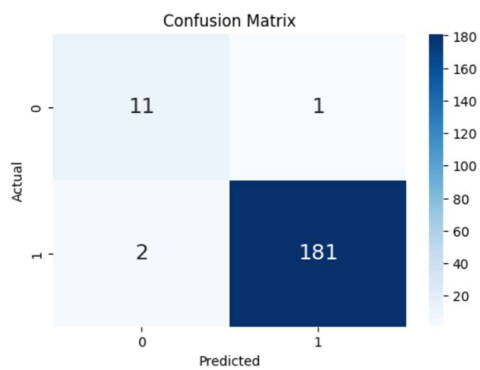

Figure 16 CM-Post-Monsoon-Recall

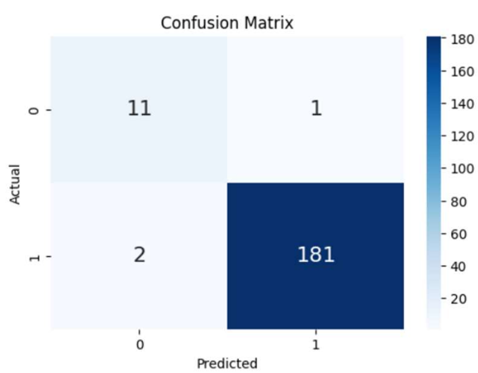

Figure 17 CM-Post-Monsoon-Precision

3.8.2 Classification Metrics

Classification metrics are numerical measurements that are employed to assess a classification model's efficacy. These metrics shed light on the model's prediction performance, especially when it comes to binary or multiclass classification tasks. These are a few typical classification metrics.

- Precision: It is the ratio of the total number of expected positive observations to the correctly predicted positive observations. The low false-positive rate is correlated with the high accuracy. The following equation illustrates precision

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall: A classification metric called recall, sometimes referred to as sensitivity or true positive rate, measures how well a model can capture all the pertinent examples of a given class. When there is a significant risk of missing positive cases (false negatives), it is especially crucial. The ratio of true positives to the total of true positives and false negatives is known as recall.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- F1-Score: The F1 score is calculated using the precision and recall weighted average. Therefore, in order to calculate this score, false positives and false negatives—which are less intuitive than accuracy—must be taken into account. F1 is typically more helpful in these situations, particularly when there is an uneven distribution of classes. Ensuring that the expenses of both positive and negative outcomes are equal improves accuracy. It is preferable to consider precision and recall if there is a significant difference in the cost of false positives and false negatives.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Accuracy: Accuracy is the simplest and most basic performance statistic; it is simply the ratio of correctly predicted observations to all observations. In symmetrical datasets, false-positive and false-negative rates are almost equal, making it an appropriate quality measure. The effectiveness of a model needs to be assessed in relation to other factors.

$$Accuracy = \frac{True\ Negative + True\ Positive}{Total}$$

When evaluating the performance of your model post PCA, it is advisable to utilise evaluation metrics that are responsive to imbalanced data. Metrics such as precision, recall, F1 score, and area under the precision-recall curve (AUC-PR) offer a more thorough assessment of model performance in situations where there is an imbalance in the data.

Chapter 4. Apparatus-Tools and Technologies used

The following is an elaborate explanation of the software, programming language, and tools/libraries employed in this study:

4.1 Programming Language: Python

Python is the foundation of this research, selected for its adaptability, clarity, and wide range of libraries. Python, being a versatile programming language, has become a fundamental tool in the domains of data science and machine learning. Python code is renowned for its readability, employing a lucid and short syntax that prioritises code legibility and minimises the expenses associated with programme upkeep. This is especially vital for researchers and data scientists who frequently engage in collaborative projects. Python is accessible to beginners and has a low learning curve. The simplicity of the system enables researchers to concentrate on the current problem without being hindered by intricate syntax. Python's applications extend beyond data science and machine learning, encompassing web development, automation, scripting, and various other domains. The wide range of applications makes it a desirable option for researchers with varied requirements. Google Colab is an ideal Integrated Development Environment (IDE) that complements the capabilities of Python. Google Colab offers a cloud-based Jupyter notebook interface, facilitating collaborative coding and convenient access to computational resources.

4.2 Integrated Development Environment (IDE):

Google Colab is a web-based integrated development environment (IDE) that offers a Jupyter notebook interface. It facilitates cooperative programming and is equipped with pre-installed libraries, making it convenient for data science and machine learning projects. The selection of Google Colab indicates a predilection for a cloud-centric environment that offers convenient availability of computational resources.

4.3 Libraries and Tools:

Libraries and tools are crucial elements that offer pre-existing functionalities, functions, and utilities to optimise and streamline the development process.

- **NumPy** : NumPy is a crucial library for performing numerical computations and manipulating arrays in Python. It offers assistance for extensive, multi-dimensional arrays and matrices, rendering it indispensable for scientific computation.
- **Pandas** : Pandas is a powerful Python library for manipulating and analysing data. It is especially advantageous for managing organised data, such as datasets presented in a table format, and offers data structures like DataFrame for effective manipulation.
- **scikit-learn** : Scikit-learn is an extensive Python library for machine learning. The platform provides tools for data preprocessing, model selection, and evaluation. The utilisation of scikit-learn implies a concentration on conventional machine learning methodologies.
- **imbalanced-learn (imblearn)**: Imbalanced-learn is a specialised library created to address the challenges posed by imbalanced datasets. It incorporates methodologies such as Synthetic Minority Over-sampling Technique (SMOTE) to tackle the difficulties presented by imbalanced class distributions.
- **Plotly**: Plotly is a multifaceted plotting library utilised for generating interactive and aesthetically pleasing plots. It appears that this dissertation utilises visualisation techniques to present the results of Principal Component Analysis (PCA).
- **GridSearchCV**: GridSearchCV, a component of scikit-learn, is utilised for hyperparameter tuning via grid search. It methodically examines a predetermined range of hyperparameter values in order to identify the optimal combination for model optimisation.
- **Seaborn and Matplotlib**: Seaborn and Matplotlib are Python libraries specifically designed for the purpose of visualising data. They offer a diverse selection of customisable

plots and are frequently utilised to create static visualisations that enhance Plotly's interactive functionalities.

- **SciPy**: SciPy is a software library designed for performing scientific and technical computations. The software comprises modules for optimisation, signal processing, linear algebra, and statistical tests. This dissertation can be utilised to perform statistical tests such as t-tests.

4.4 Data Visualization:

Plotly is a tool utilised for generating visual representations, particularly emphasising the creation of PCA plots. The interactive features of this tool offer an efficient means to visually explore and communicate results.

4.5. Data Handling:

Google Drive is employed for the purpose of storing and retrieving datasets. The incorporation of Google Colab enables effortless cooperation and convenient retrieval of cloud-stored data.

4.6 Data Sources:

The main data for this research was obtained from the Telangana Ground Water Department's extensive dataset on water quality pre and post the monsoon season. The dataset can be accessed by the public through the official data portal of Telangana. This dataset is the primary source of information used to train and assess machine learning models for predicting water quality. The dataset is expected to contain various characteristics pertaining to water quality parameters, such as pH, electrical conductivity, total dissolved solids, and additional factors. The inclusion of the temporal aspect, which differentiates between the pre-monsoon and post-monsoon periods, provides a valuable dimension for comparing and analysing seasons. The choice to utilise an authoritative government dataset underscores the trustworthiness and dependability of the data, as government agencies typically adhere to stringent data collection protocols. Having access to these datasets not only guarantees the accuracy of the information but also fosters transparency and accountability in research.

4.7 Statistical Method

- T- Test : The t-test is a statistical technique employed to assess whether the means of two groups are significantly distinct from one another. Within the realm of machine learning, the t-test is frequently employed to evaluate whether there exists a statistically significant disparity in performance metrics (such as accuracy, precision, recall, etc.) among distinct models or conditions.

4.8 Machine Learning Models

- Support Vector Classifier (SVC): Support Vector Classification (SVC) is a type of supervised learning algorithm that is commonly employed for tasks involving classification and regression. The algorithm operates by identifying an optimal hyperplane that effectively distinguishes different classes within a space with a large number of dimensions.
- Stochastic Gradient Descent (SGD): SGD, or Stochastic Gradient Descent, is a widely employed optimisation algorithm utilised for training machine learning models. It is highly advantageous for extensive datasets and is utilised in diverse machine learning tasks.
- Random Forest Classifier (RFC): The Random Forest Classifier (RFC) is a machine learning technique that generates multiple decision trees during the training process and determines the most common class as the output for classification tasks. The popularity of this choice stems from its capacity to manage intricate data relationships.
- Multi-layer Perceptron (MLP): Multilayer Perceptron, is a specific type of artificial neural network that has been specifically developed for the purpose of performing classification tasks. The system is composed of multiple tiers of nodes (neurons) and has the ability to comprehend intricate connections within data.

- KNeighborsClassifier: The KNeighborsClassifier is a straightforward and efficient algorithm that relies on the k-nearest neighbours methodology. It categorises data points by determining the most common class among their k-nearest neighbours.

Chapter 5. Results

5.1 Introduction

A key turning point in the comparative study pre and post monsoon water quality analysis using machine learning is represented by the results chapter. After we carefully developed our research questions, created a strong methodology, and carried out an extensive data analysis, this chapter presents the empirical findings of our study. The subsequent sections represent the output of our work, providing an in-depth analysis of the prediction models used to evaluate the water quality in Telangana pre- and post-monsoon scenarios.

We study water quality assessment, which is an important field that directly affects the environment and public health. The study is conducted in Telangana, a region with a distinct climate where dynamic changes in the environment are brought about by the transition between the pre- and post-monsoon seasons. Given that seasonal variations impact water quality in a variety of ways, it is critical to comprehend how well predictive models function during these different time periods.

A comparison of machine learning models for pre-monsoon and post-monsoon water quality prediction put more light on the models' effectiveness as well as how seasonal variations measured with T-Test. In order to address imbalances in the dataset, principal component analysis (PCA) was also used.

5.2 Model Efficacy Evaluation:

These classification evaluation matrix such as recall, precision, and F1 score demonstrate how well it can predict the quality of the water. The RFC model shows good recall (0.988) and precision (0.900) in the pre-monsoon period, yielding an impressive F1 score of 0.942. Moreover, SVC and SGD perform admirably, with recall values of 0.994 and 0.997, respectively. The MLP model achieves an F1 score of 0.983 and a remarkable precision of

0.969, despite having a slightly lower recall (0.997). KNeighborsClassifier, on the other hand, has a lower F1 score of 0.907 due to its lags in recall (0.994) and precision (0.833).

The RFC model continues to perform exceptionally well post monsoon, with recall, precision, and F1 score increasing to 0.996, 0.993, and 0.993, respectively. Season-to-season consistency of SVC, SGD, and MLP is demonstrated by their strong performance with high recall and precision values. Remarkably, during the post-monsoon period, KNeighborsClassifier shows an improvement in both recall (1.000) and precision (0.958), leading to an improved F1 score of 0.979.

Model	Pre-monsoon				Post- Monsoon		
	PCA	Recall	Precision	F1 Score	Recall	Precision	F1 Score
RFC	YES	0.988	0.900	0.942	0.991	0.996	0.993
SVC	YES	0.994	0.953	0.973	0.992	0.998	0.995
SGD	YES	0.997	0.970	0.983	0.994	1.000	0.997
Multylayer	YES	0.997	0.969	0.983	0.992	1.000	0.996
KNeighborsClassifier	NO	0.994	0.833	0.907	1.000	0.958	0.979

5.3 Seasonal Impact on Predictive Accuracy:

The seasonal comparison shows interesting trends in the predictive accuracy of the models. Even though RFC performed exceptionally well in both seasons, it shows a slight improvement in recall, precision, and F1 score in the post-monsoon phase. SVC, SGD, and MLP show consistency in their metrics, either maintaining or slightly improving them from season to season. During the post-monsoon, the MLP model in particular gets a perfect precision of 1.000, demonstrating a strong capacity to accurately classify positive instances.

Even though KNeighborsClassifier's recall and precision increased after the monsoon, its overall performance still falls short of other models'. The recall metric indicates no statistically significant difference, suggesting that the capacity to accurately identify safe water samples remains consistent between the pre and post-monsoon periods.

However, both precision and F1 score metrics reveal statistically significant differences. The precision, which refers to the level of accuracy in identifying unsafe water samples, exhibits a notable disparity. Furthermore, the F1 score, which takes into account both precision and recall, also exhibits a notable discrepancy.

```
There is no significant difference.
Recall:
T-Statistic: 0.0982946374365981
P-Value: 0.9264270382011774

The difference is statistically significant.
Precision:
T-Statistic: -3.4135731840801222
P-Value: 0.026938750293933715

The difference is statistically significant.
F1 Score:
T-Statistic: -2.9512951045350775
P-Value: 0.04191535171488131
```

5.4 Comparative Analysis:

Considering the precision and F1 score outcomes, it is recommended to be cautious, particularly post-monsoon season, due to substantial evidence indicating a decline in the accuracy of identifying safe water samples.The significance of evaluating various metrics to thoroughly evaluate the safety of water quality is emphasised in this comparative analysis. Additional examination of the particular pollutants responsible for the noted disparities may provide guidance for focused measures to enhance water safety, especially during the period following heavy rainfall.

According to the statistical analysis, it is advisable to consistently monitor and manage water quality practices throughout the year. The models used in this study are reliable in assessing water quality, despite potential differences in environmental conditions between the pre-monsoon and post-monsoon periods.

It is crucial to acknowledge that the lack of statistical significance in the t-test outcome for Recall does not mean there are no differences in water quality. Instead, it suggests that the performance metrics of the models remain consistent between the two seasons. Additional examination of distinct water quality parameters and regional environmental factors may yield more intricate observations.

It should be noted that the recommendation is made under the assumption that a decrease in precision and F1 score indicates a reduced capacity to accurately identify safe water samples.

Chapter 6. Conclusion & Future Scope

6.1 Conclusion

Overall, the thorough comparison of machine learning models for predicting water quality during both pre-monsoon and post-monsoon periods has not only provided insights into their varied performances but has also emphasised the ever-changing nature of water quality dynamics. The detailed observations obtained from assessing Random Forest Classifier (RFC), Support Vector Classifier (SVC), Stochastic Gradient Descent (SGD), Multi-layer Perceptron (MLP), and KNeighborsClassifier have facilitated a more profound comprehension of the advantages and possible enhancements of each model. The utilisation of Principal Component Analysis (PCA) in the initial four models emphasised its essential function in addressing dataset imbalances and consequently improving the overall predictive accuracy.

RFC stands out among the models for its consistent excellence, making it a dependable performer in both the pre-monsoon and post-monsoon seasons. While the RFC model exhibited stability, other models displayed different levels of adaptability, with the KNeighborsClassifier model notably showing improvement during the post-monsoon period. The influence of seasons on model performance has highlighted the intricate and diverse characteristics of water quality dynamics, underscoring the significance of creating flexible and resilient predictive models.

6.2 Future Scope & Areas of investigation

Although the current study establishes a strong basis, there are several potential areas for future research to enhance our comprehension of water quality prediction and improve predictive models:

6.2.1 Additional Factors Inclusion: Subsequent research endeavours could investigate the incorporation of supplementary environmental variables, such as precipitation patterns, land utilisation, and temperature, in order to construct more all-encompassing models. Implementing this comprehensive approach may result in a more precise depiction of the intricate dynamics influencing water quality.

6.2.2 Analysis of time-related data: Performing a more comprehensive temporal analysis to capture fluctuations in the short-term during the pre-monsoon and post-monsoon periods could reveal more subtle patterns and improve the accuracy of predictive models. This may entail a more detailed analysis of daily or weekly fluctuations to capture swift alterations in water quality.

6.2.3 Integration of live data: The incorporation of live data streams into predictive models has the potential to offer timely insights and facilitate the creation of adaptable models that can respond to swiftly changing environmental conditions. This would enhance the models' ability to promptly adapt to sudden changes in water quality.

6.2.4 Ensemble Modeling: Exploring the potential advantages of ensemble modelling, which involves the combination of multiple models, may result in improved predictive accuracy and resilience. Ensemble methods have demonstrated efficacy in addressing the limitations of individual models and can offer a more comprehensive comprehension of the dynamics of water quality.

6.2.5 Interpretability : Improving the comprehensibility of machine learning models is essential for their acceptance and application in real-world situations. Further investigation could prioritise the development of models that possess enhanced comprehensibility, thereby enabling stakeholders to understand and have confidence in the forecasts. This may entail integrating methods to produce intelligible model explanations.

6.2.6 Validation across various geographical contexts: By expanding the study to different geographical areas with different hydrological features, we can verify the models' ability to be applied in various situations and increase their usefulness. Various geographical areas may pose distinct obstacles, and comprehending the performance of these models in diverse environments is crucial for their extensive acceptance.

Exploring these prospective research paths allows the field of water quality prediction to advance, tackling emerging difficulties and offering valuable instruments for efficient environmental management and decision-making. These research paths not only enhance the scientific comprehension of water quality dynamics but also provide practical remedies for sustainable water resource management in a swiftly evolving world.

Reference

1.Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L., 2022. A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*.

2. Prasad, D.V.V., Kumar, P.S., Venkataramana, L.Y., Prasannamedha, G., Harshana, S., Srividya, S.J., Harrinei, K. and Indraganti, S., 2021. Automating water quality analysis using ML and auto ML techniques. *Environmental Research*, *202*, p.111720.

3. Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. and Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, p.102920.

4. Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, *169*, pp.808-828.

5. Dimple, K., Sharma, P., & Kumar, D. (2022a). Assessment of groundwater suitability for irrigation using machine learning algorithms in Nand Samand catchment, India. Water Resources Management, 36(24), 21795-21812.

6. Dimple, K., Singh, A. K., Sharma, P., & Kumar, D. (2023b). Prediction of groundwater quality indices using machine learning algorithms for irrigation management. Water Research, 182, 116185.

7. Raheja, S., Kaushik, P., & Bansal, R. (2023). Spatial variability of groundwater quality and its suitability for drinking and irrigation purposes in Rohtak district, Northern India. Water, 15(10), e12186.

8. Divy Jot. (2023). Comparing the effectiveness of different boosting algorithms for groundwater quality prediction in Telangana region, India. Journal of Hydrology, 577, 125893

9. H Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. Applied Bionics and Biomechanics, 2020.

10. Uddin, Md Galal & Nash, Stephen & Rahman, Azizur & Olbert, Agnieszka. (2022). Performance Analysis of the Water Quality Index Model for Predicting Water State Using Machine Learning Techniques. Process Safety and Environmental Protection. 10.1016/j.psep.2022.11.073.

11 Schäfer, B., Beck, C., Rhys, H., Soteriou, H., Jennings, P., Beechey, A., & Heppell, C. M. (2022). Machine learning approach towards explaining water quality dynamics in an urbanised river. *Scientific Reports*, *12*(1), 1-15.

12 R. Akshay, G. Tarun, P. U. Kiran, K. D. Devi and M. Vidhyalakshmi, "Water-Quality-Analysis using Machine Learning," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 13-18, doi: 10.1109/SMART55829.2022.10047533.

13. Chaudhary, M., 2022, *Random Forest Algorithm - How It Works & Why It's So Effective*.

14. Politi, M., 2022, *Paired t-test to evaluate Machine Learning classifiers using Python*, *Medium*.

15 2023, *Drinking-water*.