**WATER QUALITY ANALYSIS**

|  |  |
| --- | --- |
| **Date** | **28-10-2023** |
| **Team ID** | **714** |
| **Project Name** | **WATER QUALITY ANALYSIS** |

**Table of Contents**

|  |  |
| --- | --- |
| 1 | Introduction |
| 1.1 | Objectives |
| 1.2 | Scope and Methodology |
| 1.3 | Problem Statement |
| 2 | Design Thinking Process |
| 3 | Phases of Development |
| 4 | Data |
| 4.1 | Dataset description |
| 4.2 | Data collection |
| 4.3 | Data preprocessing |
| 5 | Analysis |
| 5.1 | Exploratory Data Analysis (EDA) |
| 5.1.1 | Detections and Visualization Outliers |
| 5.1.2 | Data Analysis in Jupyter Notebook |
| 6 | Dividing Dataset into Features and Target Variable |
| 7 | Model Building and Evaluation |
| 7.1 | Model 1 – Support Vector Machine |
| 7.2 | Model 2 – Decision Tree Classifier |
| 7.3 | Model 3 –Random Forest Classifier |
| 8 | Conclusion |
| 9 | References |
| 9.1 | Data Sources |
| 9.2 | Relevant Literature |

**1. Introduction:**

Water quality is a fundamental aspect of environmental health and public well-being. The availability of clean and safe water is essential for sustaining life and ecosystem integrity. Ensuring that our water resources meet acceptable quality standards is not only a regulatory requirement but also a moral and ecological imperative. The water quality analysis project was undertaken to assess and evaluate quality of water based on the parameter such as ph, hardness and so on.

**1.1 Objectives:**

The main objective of this project to assess the quality of water to define weather it’s portable for drinking water which ensuring the health of humans . Objective is to identify any potential issues or deviations from regulatory standards, and determine the potability of the water based on (PH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity) parameters.

**1.2 Scope and Methodology:**

The project encompassed a thorough examination of **restaurants and residential users**. We conducted field sampling at specific parameters which to ensure water analysis. The collected water samples were transported to the model which give more accuracy .

**1.3 Problem Statement:**

In this section, we introduce the primary problem addressed in our Water Quality analysis , emphasizing the significance of the analysis. We outline the specific objectives and define the scope of the project.

**2 Design Thinking Process:**

This part provides insights into the methodology and approach

we followed in this project. We describe the design thinking process we used to structure our analysis, including data collection strategies, data preprocessing steps, and the overall framework that guided our work using the IBM cognos .

**3 Phases of Development:**

We present a high-level overview of the different stages the project went through, including data collection, data preprocessing, analysis, and documentation, and how each phase contributed to achieving our objectives.

**4. Data**

**4.1 Dataset Description:**

In this segment, we introduce the dataset used for our analysis, including its origin, size, and the nature of the information it contains. We include a direct link to the dataset source: Water Quality Analysis

Dataset :

(<https://www.kaggle.com/datasets/adityakadiwal/water-potability>).

**Importance of Dependencies**

- NumPy: NumPy is a fundamental library for numerical operations in Python. It provides support for arrays and matrices, making it essential for data manipulation.

- Pandas: Pandas is a data manipulation library that simplifies data handling by providing data structures like DataFrames. It's indispensable for data exploration and preprocessing.

- Scikit-Learn: Scikit-Learn is a machine learning library that offers a wide range of algorithms for regression, classification, and more. It streamlines the implementation of machine learning models.

- Matplotlib and Seaborn and plotly: Data visualization is crucial for understanding the dataset and model results. Matplotlib, Seaborn and plotly are excellent for creating charts and graphs.

**4.2 Data Collection:**

We detail our data collection process, explaining where and how we sourced the water quality data from the [provided dataset (<https://www.kaggle.com/datasets/adityakadiwal/water-potability>) on Kaggle. We also clarify the data sources and their update frequencies, ensuring transparency and reliability.

**4.3 Data Preprocessing:**

This part outlines the steps taken to prepare the data for analysis. We describe data cleaning like addressing missing values, and transforming data , clearing the outliers to ensuring the data's suitability for clustering and analysis

**Importance of Dependencies**

- NumPy: NumPy is a fundamental library for numerical operations in Python. It provides support for arrays and matrices, making it essential for data manipulation.

- Pandas: Pandas is a data manipulation library that simplifies data handling by providing data structures like DataFrames. It's indispensable for data exploration and preprocessing.

- Scikit-Learn: Scikit-Learn is a machine learning library that offers a wide range of algorithms for regression, classification, and more. It streamlines the implementation of machine learning models.

- Matplotlib and Seaborn and plotly: Data visualization is crucial for understanding the dataset and model results. Matplotlib, Seaborn and plotly are excellent for creating charts and graphs.

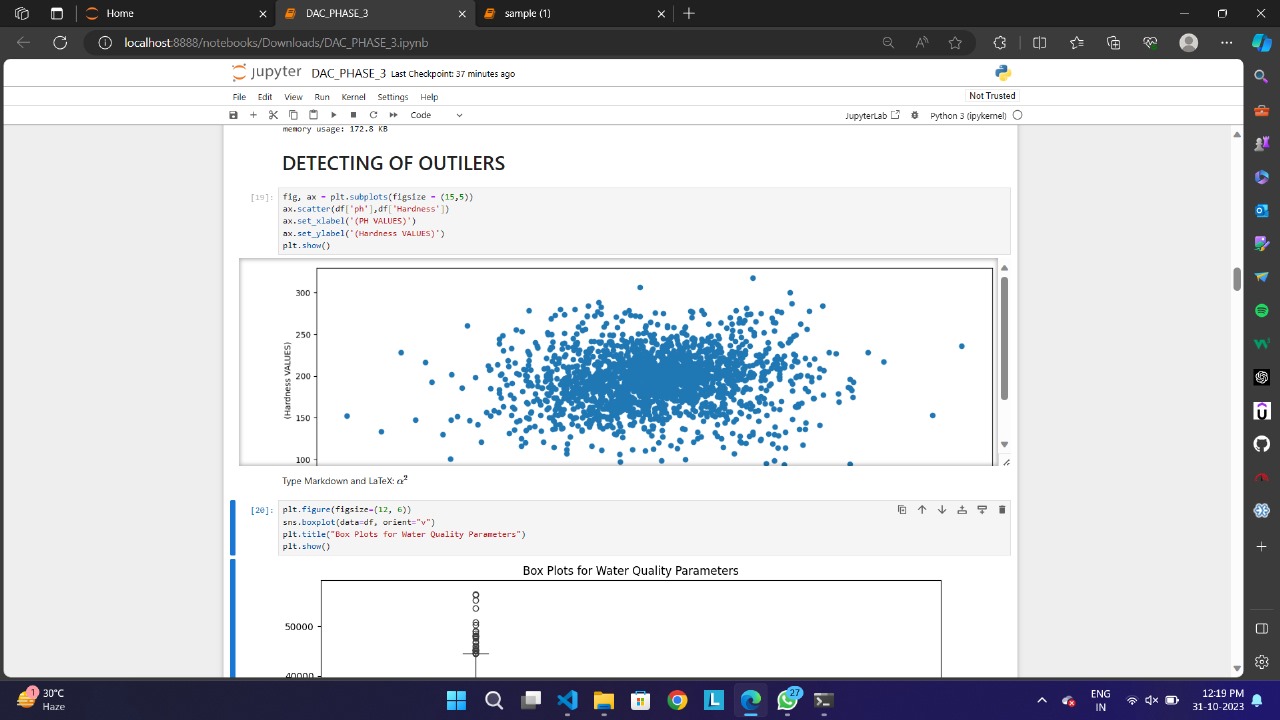
**5. Analysis**

**5.1 Exploratory Data Analysis (EDA):**

EDA is a critical phase where we delve into the dataset's intricacies. In this section, we present our findings, highlighting insights gained from visualizations and data exploration. We identify data characteristics, trends, and potential outliers, which laid the foundation for more advanced analysis.

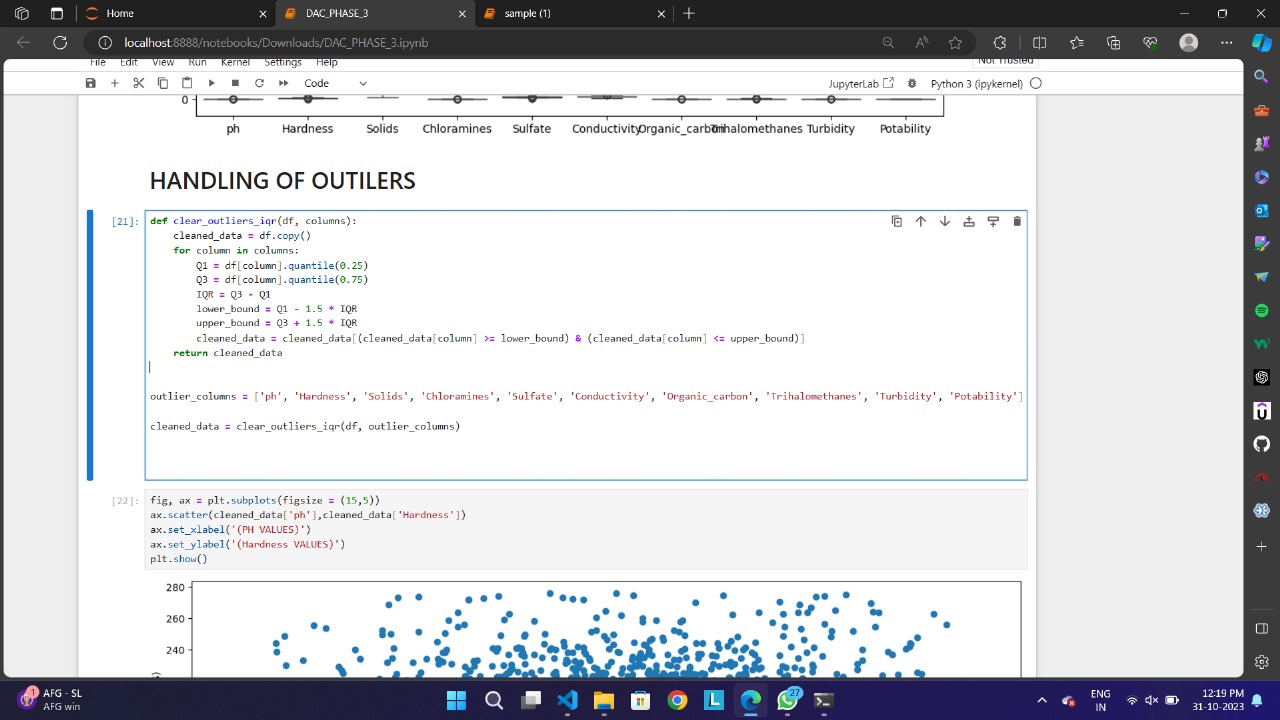
**5.1.1 Detections and Visualization Outliers:**

After we have identified outliers using statistical methods (e.g., IQR, or visualizations), we created visualizations to highlight and label these points in our dataset. This can be done using scatter plots , boxplot and other relevant charts.



**Handling outliers :**

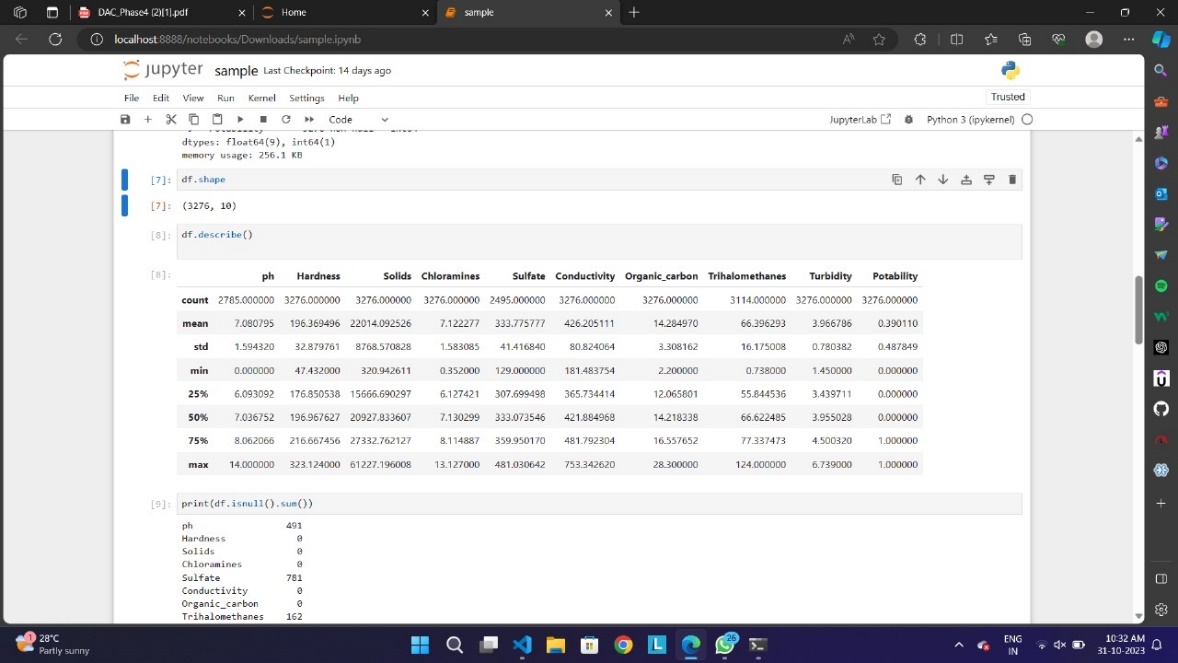
Outlier removal is a critical step in data cleaning to eliminate errors or anomalies in the data. In water quality analysis, identifying and addressing outliers is crucial for ensuring the accuracy and reliability of the analysis results. We have removed the extreme outliers which are likely to be data anomalies. We have used the Interquartile Range (IQR) to remove the outliers of the data .

****

**5.1.2 Data Analysis in Jupyter Notebook:**

In this step, we have used Jupyter Notebook to perform data analysis on your water quality dataset. we load the data using Python, typically with libraries like Pandas, Numpy and conduct various analyses, including:

**Descriptive statistics**: We have calculate summary statistics like mean, median, standard deviation, and percentiles for each water quality parameter (e.g., ph, Hardness, Solids, Chloramines, Sulfate ,Conductivity, Organic Carbon ,Trihalomethanes ,Turbidity).



**Data Visualization Techniques**

- Histograms: Used to visualize the distribution of numerical data, histograms provide insights into data spread and skewness.

- Box Plots: Box plots offer information about data distribution, central tendency, and potential outliers.

- Scatter Plots: Scatter plots help visualize relationships between two numerical variables.

- Bar plot : It representation of categorical data, used to compare values across different categories or groups. It's effective for making side-by-side comparisons.

-Count plot : It helping visualize the distribution of categories, A count plot is a type of bar plot that displays the frequency or count of categorical data

**Data visualization**: Create plots and charts to visualize trends, distributions, and relationships within the data. For example, you can use Matplotlib, Seaborn, or Plotly to generate line charts, scatter plots, histograms, count plots and more.

The goal of this step is to gain insights into the data, identify patterns, and prepare the results for presentation. we also save the analysis results in a format suitable for importing into IBM Cognos, such as a CSV file.

**6. Dividing Dataset into Features and Target Variable**

**Introduction**

To apply machine learning algorithms, we need to divide the dataset into two key components: features and the target variable.

**Features and Target Variable**

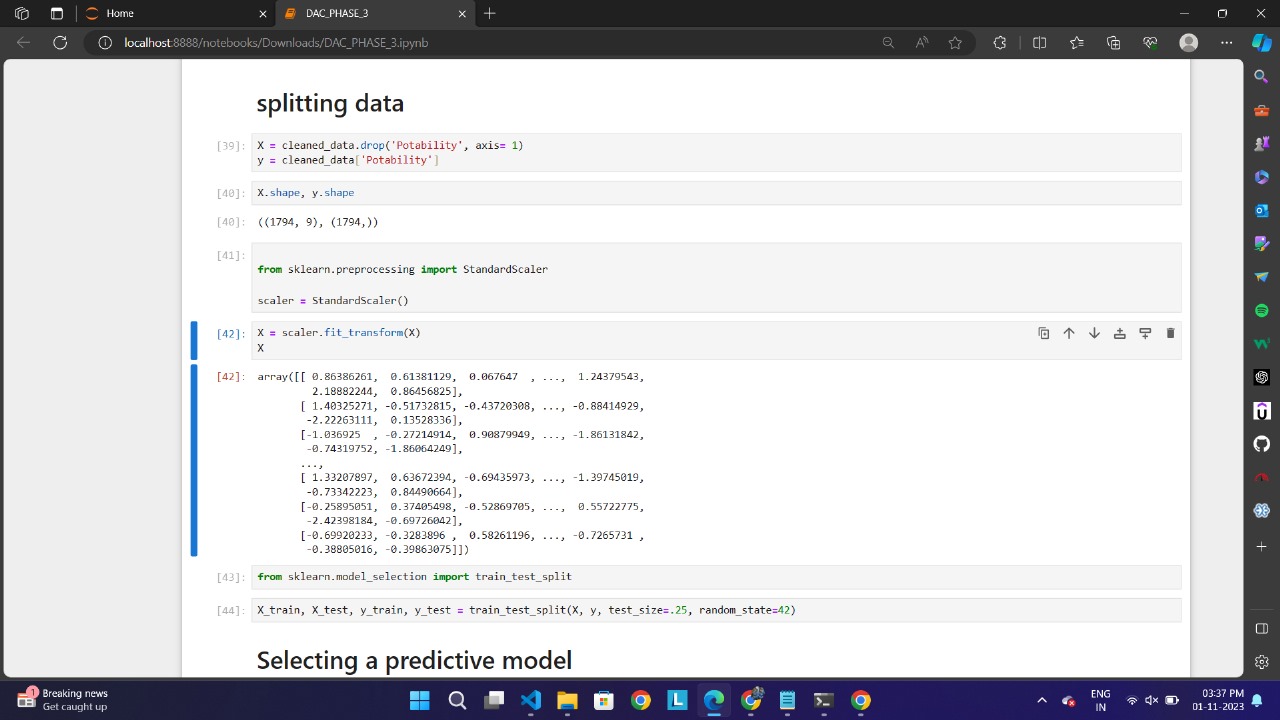
- Features (Independent Variables): These are the attributes used to make predictions. For water quality analysis, features could include (Ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity .)

- Target Variable (Dependent Variable): This is what we aim to predict – in this case, the Potability

**Data splitting :**

- Training Set: This is the portion of the data used for model training.

- Testing Set: The testing set is reserved for evaluating the model's performance.

****

**7. Model Building and Evaluation**

**Introduction**

In this pivotal section, we embark on the process of building and evaluating machine learning models for water quality analysis. We will explore various algorithms and assess their performance. **Model Building Steps**

- Data Splitting: Separating the data into training and testing sets. - Model Selection: Choosing the appropriate model for the task.

- Model Training: Fitting the model to the training data. - Hyper parameter Tuning: Optimizing model parameters for better performance.

- Model Evaluation Metrics: Using appropriate metrics to assess the model's performance

**7.1 Model 1 – Support Vector Machine :**

**SVM Classifier:**

A Support Vector Machine (SVM) classifier is a machine learning algorithm used for binary and multiclass classification tasks. It works by finding the best possible decision boundary (or hyperplane) that separates data points of different classes with the maximum margin, while also minimizing classification errors.

**Kernel Functions**: SVM can use various kernel functions, such as the linear, polynomial, or radial basis function (RBF) kernels, to handle different types of data and achieve non-linear classification.

**Support Vectors:** SVM identifies critical data points (support vectors) to define decision boundarie

**7.2 Model 2 – Decision Tree Classifier:**

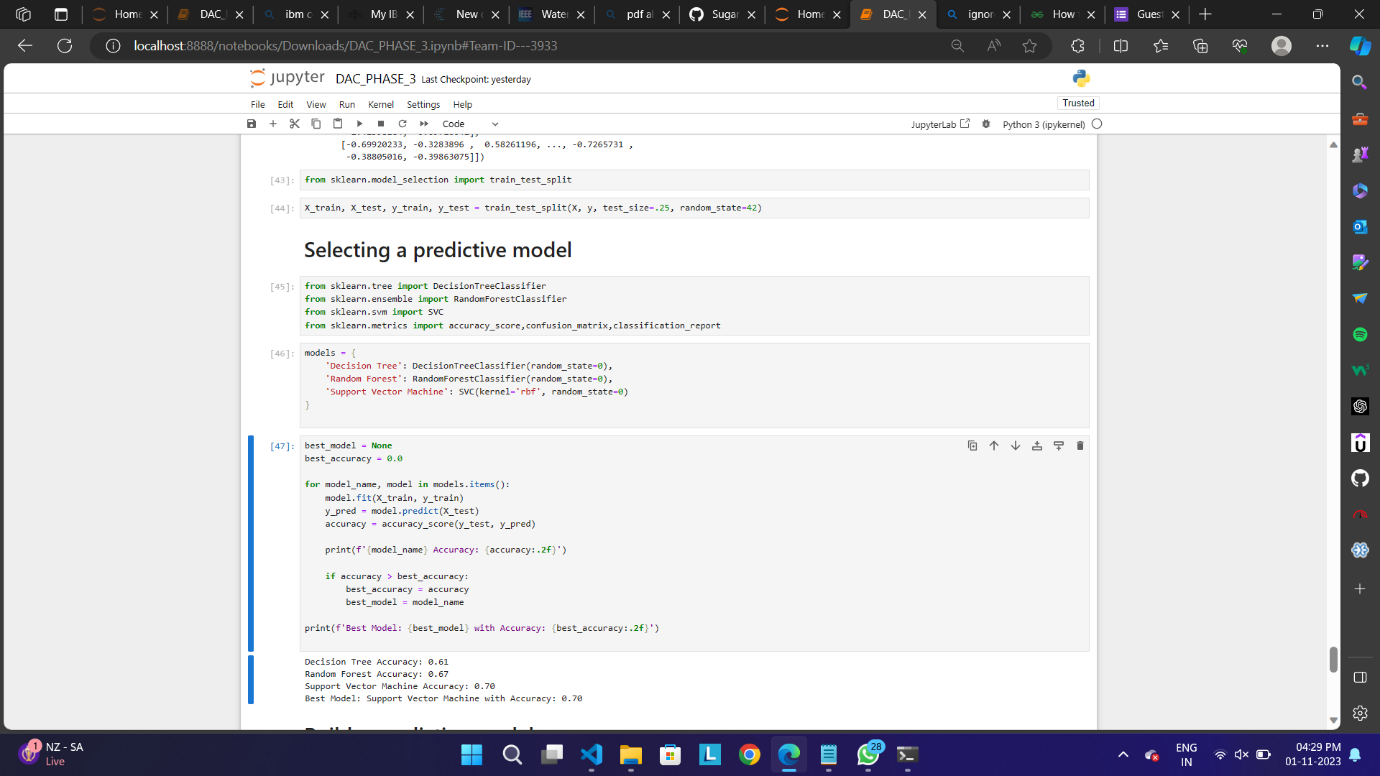
A Decision Tree Classifier is a machine learning algorithm used for classification tasks. It makes decisions by recursively splitting the dataset into subsets based on the most significant attributes

**Tree Structure:** A decision tree is a tree-like structure, where each internal node represents a feature or attribute, each branch represents a decision rule based on that feature, and each leaf node represents a class label or outcome.

**Decision Making:** To make a prediction, the algorithm starts at the root node and follows a path down the tree, making decisions at each node until it reaches a leaf node, which provides the final classification.

**7.3 Model 3 –Random Forest Classifier :**

A Random Forest Classifier is a machine learning model used in water quality analysis to predict and classify water conditions. It's like a diverse team of decision trees, where each tree votes on the water quality status based on various factors like pH, Hardness, Conductivity, Organic carbon, Trihalomethanes, Turbidity .

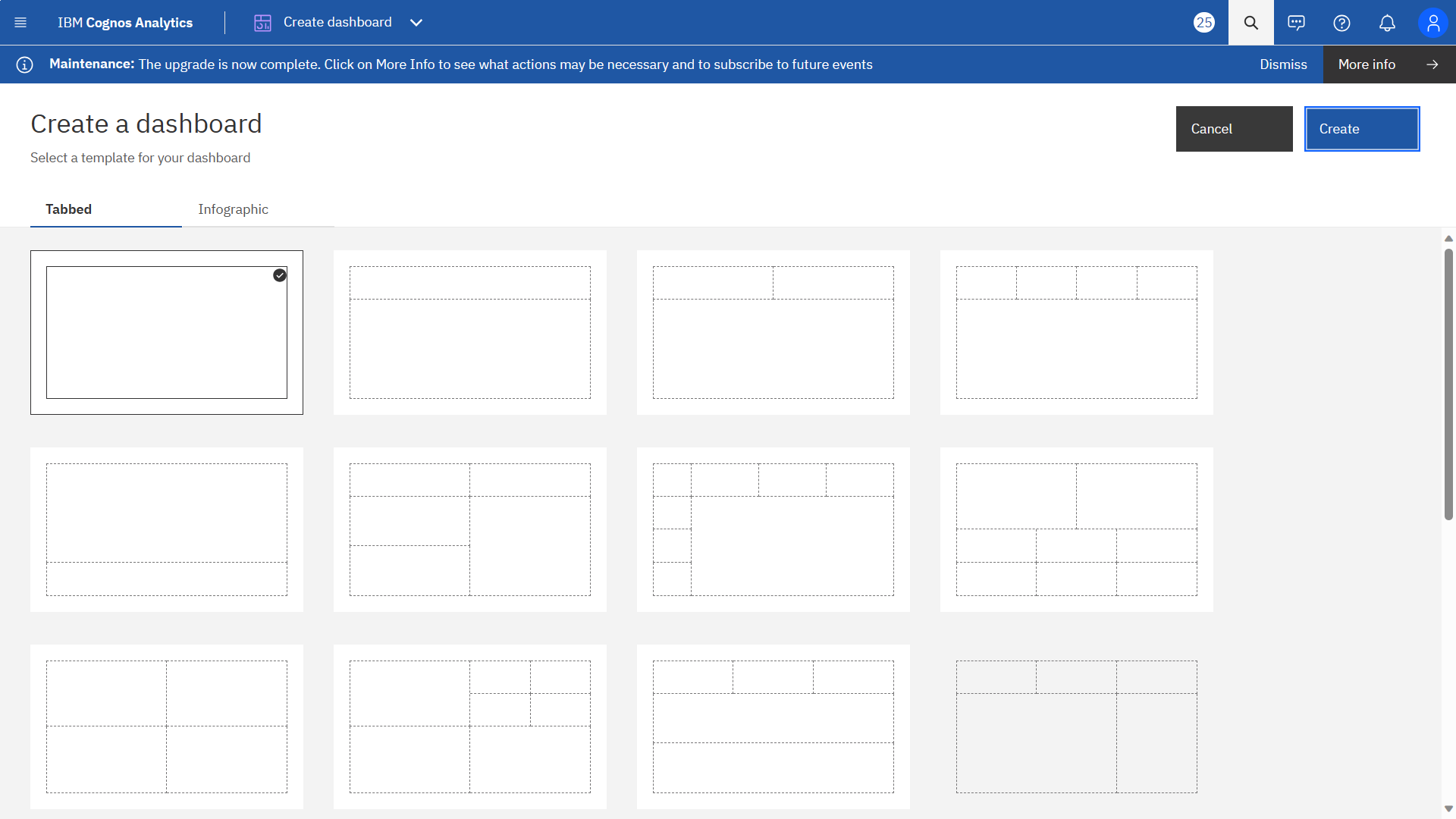
****

**Analysis using IBM Cognos:**

IBM Cognos allows us to create interactive dashboards that provide a consolidated view of the data and visualizations. Dashboards are a powerful way to present and share data summaries and insights with stakeholders. Here's a description of our dashboards in IBM Cognos

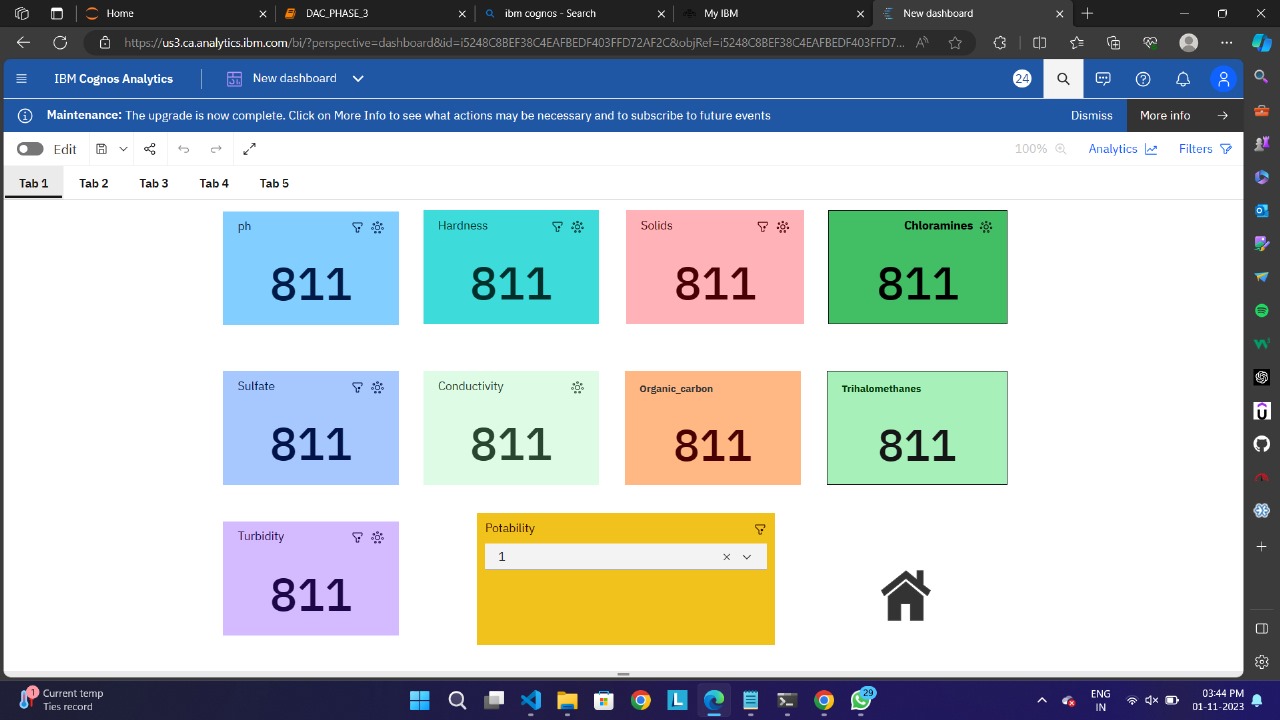
**Creating a Dashboard and Design Layout:**

In Cognos, we have created a new dashboard after uploading the dataset, which was cleaned and preprocessed using Jupyter Notebook in previous steps. We Customize the layout of new dashboard. we can arrange widgets and adjust their sizes and control their placement. This allows you to create a visually appealing and informative dashboard. We selected a blank layout of the dashboard which is used for visualizations and analysis of data.



**Interactivity Data Visualization:**

Dashboards typically use charts, graphs, tables, and other visual elements to represent data, making it easier for users to understand complex information quickly. Interactive dashboards allow users to explore data further, drill down into details, and perform actions like filtering or data entry.

****

**8. Conclusion**

In conclusion, anomaly detection techniques are indispensable tools across a wide range of industries and applications. These methods, whether based on statistics, machine learning, or domain-specific knowledge, play a vital role in identifying rare and unexpected patterns within data. By enabling the early detection of anomalies, these techniques help mitigate risks, enhance security, and improve the quality and reliability of systems. As data continues to grow in complexity and volume, the importance of effective anomaly detection only continues to rise, making it an essential component of modern data-driven decision-making processes.

**9. References**

**9.1 Data Sources:**

In this segment, we introduce the dataset used for our analysis, including its origin, size, and the nature of the information it contains. We include a direct link to the dataset source: Water Quality Analysis

Dataset :

(<https://www.kaggle.com/datasets/adityakadiwal/water-potability>).

The project references of water quality analysis which we have done using jupyter notebook and IBM Cognos was added below

Github:(https://github.com/Shibie23/Water-QualityAnalysis.git)

**9.2 Relevant Literature:**

We reference research papers, studies, or scholarly work that informed our analysis. These citations offer credibility to the methods and insights presented in the analysis.

The reference paper that we went through for our analysis are “Predictive Models for River Water and using Machine Learning and Big Data Quality” (<https://ieeexplore.ieee.org/document/9395832> ) , “WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes Techniques – A Survey “ ([https://ieeexplore.ieee.org/document/9766319](https://ieeexplore.ieee.org/document/9766319%20) )