

**Lab Manual**

**Subject: Machine Learning**

**Course Code AI-414**

**BY**

**Muhammad Ahmad Mustafa 2023-BS-AI-016**

**Supervised by Sir Saeed**

**Degree name: BSAI- 4A**

**Subject name: Machine Learning**

**DEPARTMENT OF COMPUTATIONAL SCIENCES FACULTY OF INFORMATION TECHNOLOGY**

**The University of Faisalabad**

**Table of Contents**

**Project 1:** **Train Energy Prediction using Regression**………………..2

Summary………………………………………………………………………………………………………………………………………..3

Objectives……………………………………………………………………………………………..3

Abstract……………………………………………………………………………………………….4

Explanation of Steps……………………………………………….…………………………………4

Jupyter Notebook Code……………………………………………………………………………….4

Clean Data……………………………………………………………………………………………6

Outlier Removal………………………………………………………………………………………6

**Linear Regression……………………………………………………………………………………9**

**Conclusion………………………………………………………………………………………….10**

**Project 2:** **Water Quality Classification** …………………………………….11

Summary……………………………………………………………………………………………………………………………………..11

Objectives…………………………………………………………………………………………...11

Abstract………………………………………………………………………………………………12

Explanation of Steps……………………………………………….…………………………………12

Jupyter Notebook Code……………………………………………………………………………….13

Clean Data……………………………………………………………………………………………15

Outlier Removal………………………………………………………………………………………16

Standardization………………………………………………………………………………………..16

### Balanced Data…………………………………………………………………………………………17

### Resample…………………………………..…………………………………………………………..18

### Classification…………………………………………………..…………………………………..19

### Evaluation………………………………………………………………………………………….20

**Project 1:** **Train Energy Prediction using Regression**

**Summary:**

**The objective of this project is to implement regression-based machine learning techniques to forecast energy usage in commercial and institutional buildings.** The dataset utilized for this task includes a wide range of contextual and environmental features such as building characteristics, weather conditions, and time-related data points. These attributes serve as key indicators in understanding patterns of energy consumption.

The goal is to create a robust predictive model that can estimate energy demand accurately over time, enabling proactive planning and efficient resource allocation. This project highlights how regression methods like random forest, gradient boosting, and other supervised learning approaches can be used in practical scenarios related to energy efficiency and sustainability.

By combining historical trends with machine learning, the project supports data-driven strategies aimed at reducing operational costs, optimizing performance, and contributing to smart energy management in modern infrastructure.

**Objectives:**

 To build a regression model that accurately predicts building energy usage based on multiple input features.

 To explore the relationship between environmental conditions, building attributes, and energy consumption.

 To evaluate and compare different regression algorithms for performance and reliability.

 To use time-based patterns to improve forecasting accuracy.

 To demonstrate the practical use of machine learning in energy management and sustainability planning.

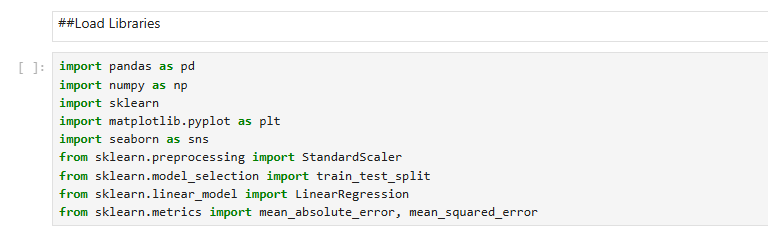
**Abstract:**

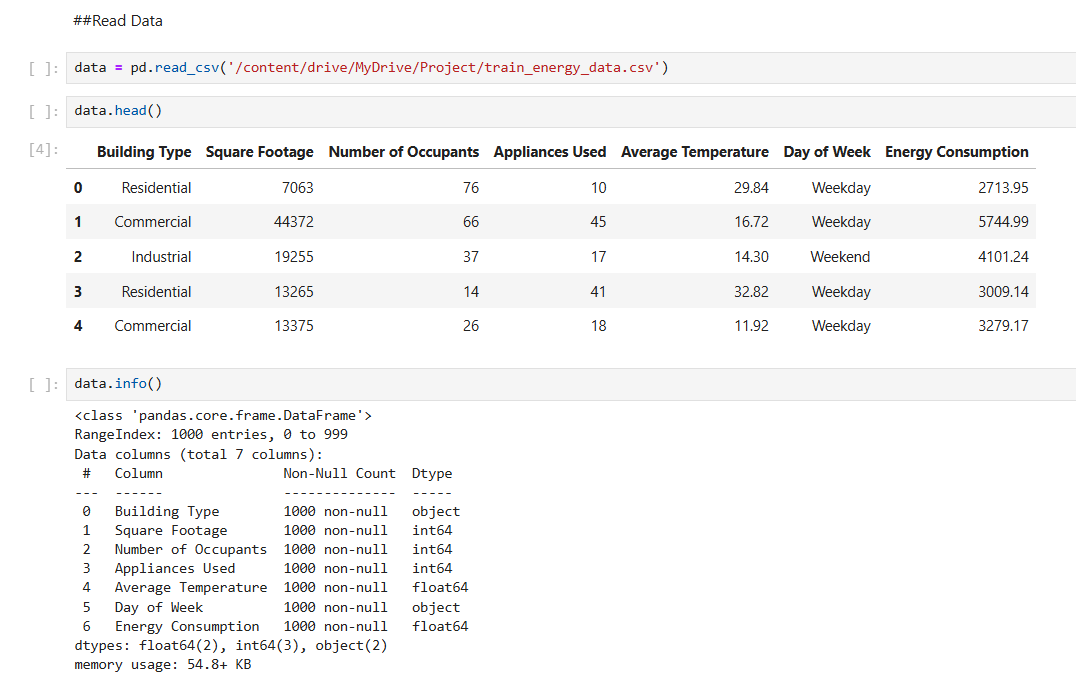
In this regression-based project, we work with a dataset containing various building-related and environmental features such as building type, size, weather conditions, and time-based factors. The goal is to predict energy consumption. Data preprocessing steps include normalization of continuous variables and transformation of categorical data. Models such as Random Forest and Gradient Boosting are trained and evaluated using metrics like Root Mean Squared Error and R² score. Visualization tools are used to interpret model performance and analyze consumption trends.

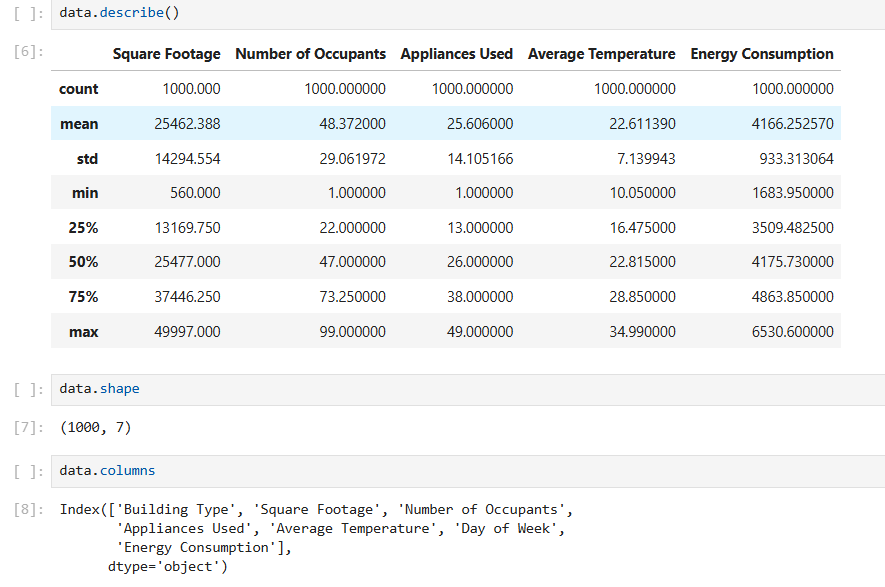
**Explanation of Steps:**

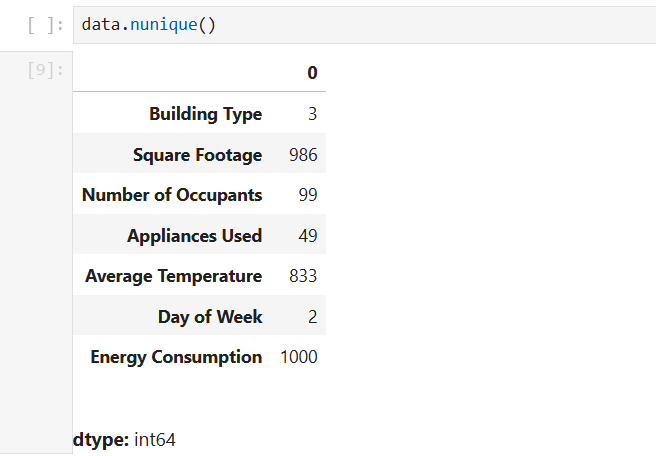
1. **Import and Explore the Energy Dataset**
2. **Clean the Dataset and Handle Incomplete Data**
3. **Transform Features for Model Compatibility**
4. **Split the Data for Training and Testing**
5. **Apply Regression Algorithm for Prediction**
6. **Evaluate the Model Performance**
7. **Visualize and Interpret the Results**

**Jupyter Notebook Code:**

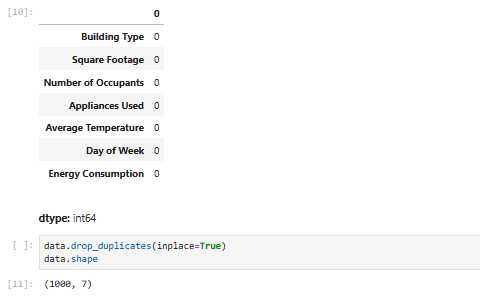
1. **Load Libraries:**
2. **Reading Data:**



****

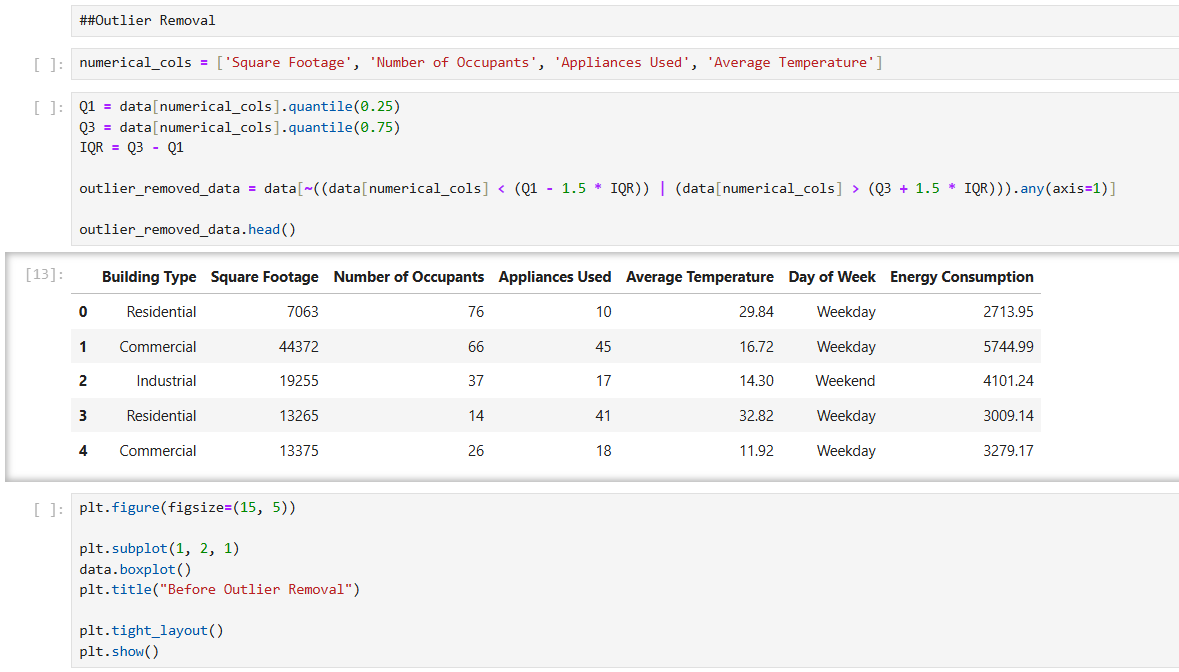


1. **Clean Data:**  
   In this step, the dataset was carefully examined to find and fix any issues such as missing values, incorrect entries, or duplicated records. Cleaning the data ensured accuracy and consistency, which is essential for building a reliable regression model.

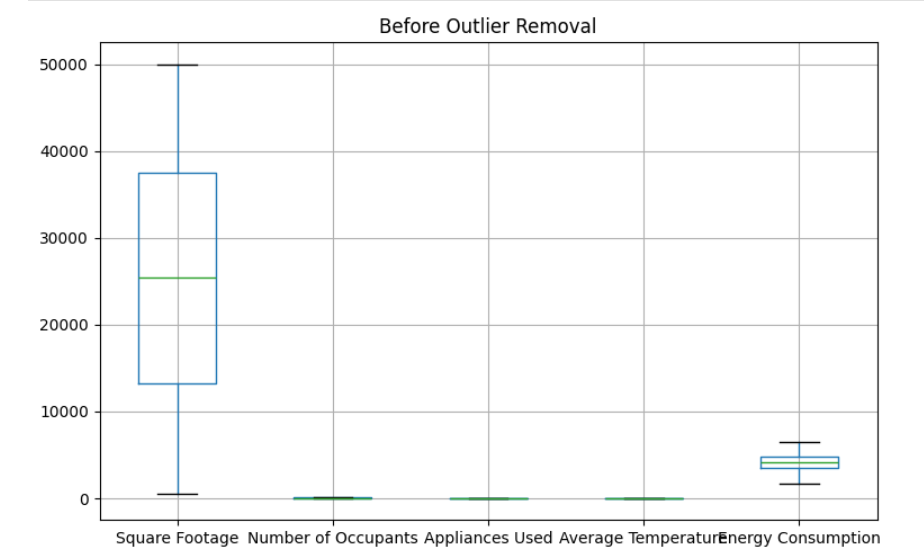
****

1. **Outlier Removal:**

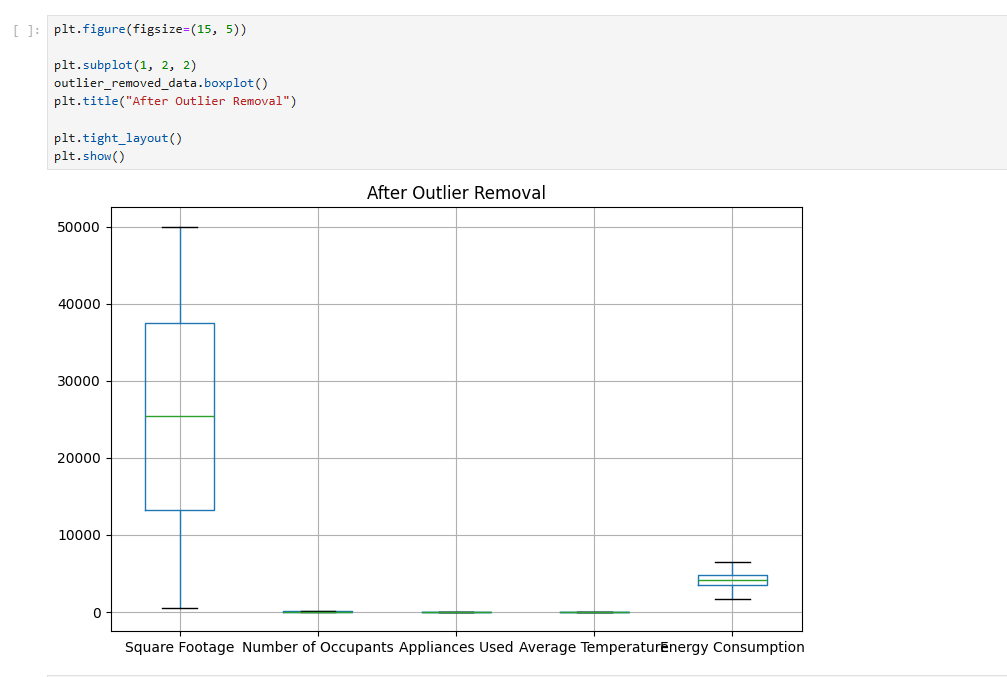
Outliers are extreme values that can negatively affect model performance. Removing them helps improve the accuracy and reliability of predictions.

****

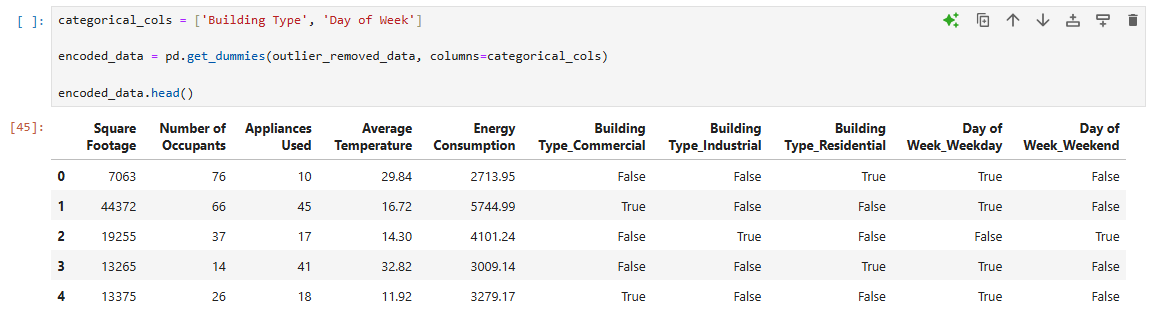
This step involved fixing missing values and correcting errors to ensure the dataset was accurate and ready for modeling.



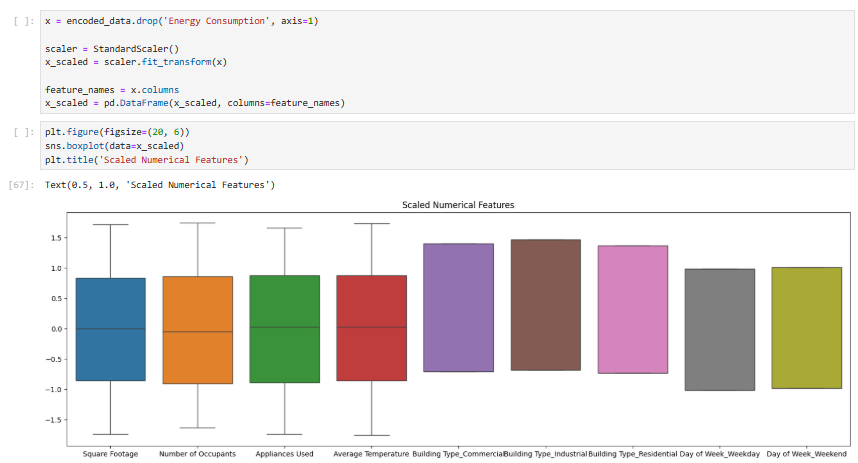
**Outlier removal enhances** data reliability by eliminating extreme values that could skew analysis, the distribution of variables is now more compact, allowing for more accurate insights into **energy consumption** trends.

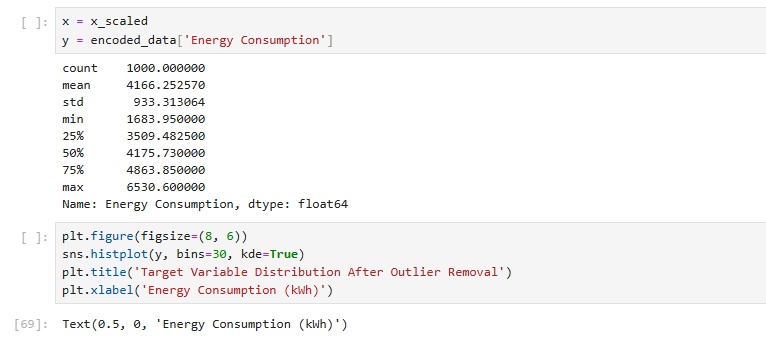


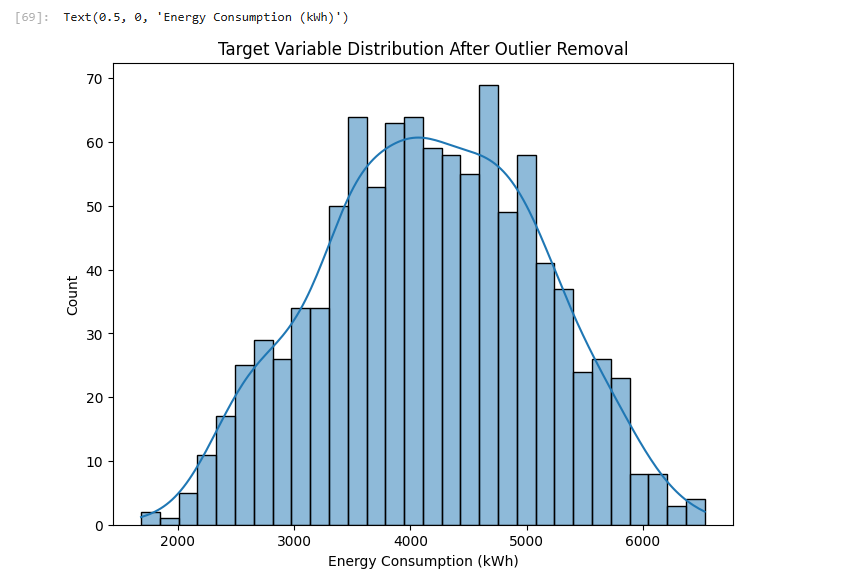
It visualizes dataset trends and variability, aiding in the identification of data distribution and potential outliers for improved analysis.

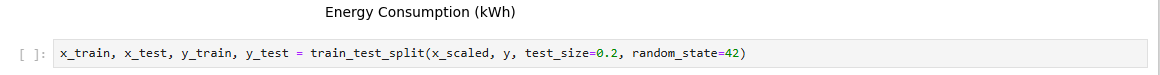


The numerical features were scaled using Standard Scaler to normalize their range. This helps improve model performance by ensuring all features contribute equally.



****

****

****

1. **Linear Regression:**

The **Linear Regression** model was used to predict the target variable by finding the best-fitting linear relationship between input features and output. It minimizes the error between actual and predicted values using the least squares method.

This trains a Linear Regression model, evaluates its performance using MAE, MSE, and RMSE metrics, and displays a comparison between actual and predicted values. The results are shown in a DataFrame format for easy interpretation.



**Conclusion**

This project aimed to develop a regression model to predict energy output based on various input features. The process began with understanding and cleaning the dataset to ensure data quality. Outliers were identified and removed to prevent skewed results, and categorical features were encoded while numerical ones were scaled to prepare the data for modeling.

After splitting the dataset into training and testing sets, a Linear Regression model was trained to learn the relationships between the input variables and the energy output. The model was then evaluated using standard performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² score. These metrics helped in understanding the model’s accuracy and effectiveness in making predictions.

Overall, the results indicated that the model could predict energy output with reasonable accuracy. The project demonstrates how machine learning techniques, especially regression models, can be applied to real-world energy data for forecasting and decision-making purposes. This process also highlights the importance of proper data preprocessing in achieving reliable and interpretable results.

**Project 2: Water Quality Classification**

**Summary:**

The aim of this project is to utilize machine learning classification techniques to determine whether water samples are safe for consumption. The dataset applied in this study includes multiple chemical and physical attributes of water, such as acidity (pH), mineral content, level of dissolved solids, presence of disinfectants, and other key indicators that influence water quality. The target output is a binary label indicating whether the water is potable or not.

The main objective is to construct a reliable model that can classify water samples based on these features, allowing for accurate and automated assessment of drinkability. This project showcases the practical use of classification algorithms like logistic regression, support vector machines, and decision trees in the field of environmental safety and public health.

By applying supervised learning to this binary classification problem, the project highlights how data science can assist in making informed decisions regarding water safety, potentially aiding efforts in quality control and public access to clean drinking water.

**Objectives:**

 To develop a machine learning classification model that predicts water potability based on chemical and physical attributes.

 To analyze key water quality parameters and understand their impact on drinkability.

 To compare the performance of different classification algorithms (e.g., Logistic Regression, Decision Tree, Random Forest).

 To evaluate the model using accuracy, precision, recall, and other performance metrics.

 To demonstrate the application of data-driven techniques in environmental and public health decision-making.

**Abstract:**

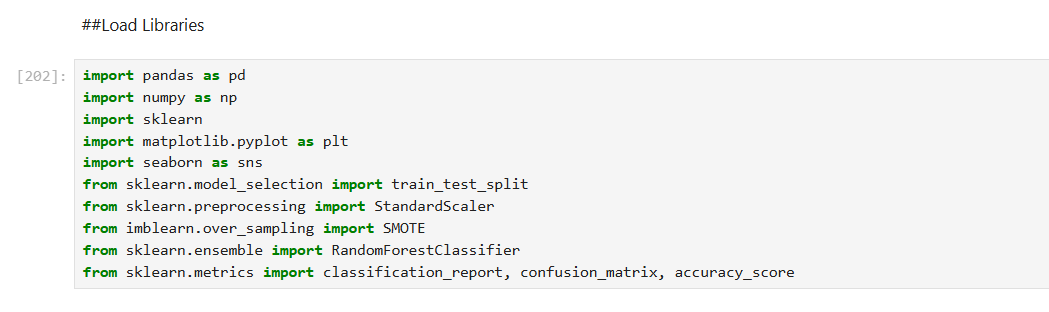
This classification-based machine learning study aims to determine the potability of water samples using various chemical and physical features, such as pH level, mineral content, dissolved solids, and disinfectant presence. The dataset undergoes preprocessing steps to ensure data quality and readiness for modeling. Multiple classification algorithms, including Logistic Regression, Support Vector Machines, and Decision Trees, are trained and evaluated using performance metrics like accuracy and confusion matrices. The ultimate objective is to develop a reliable and interpretable model that can automate the assessment of water safety, supporting efforts in environmental monitoring and public health.

**Explanation of Steps:**

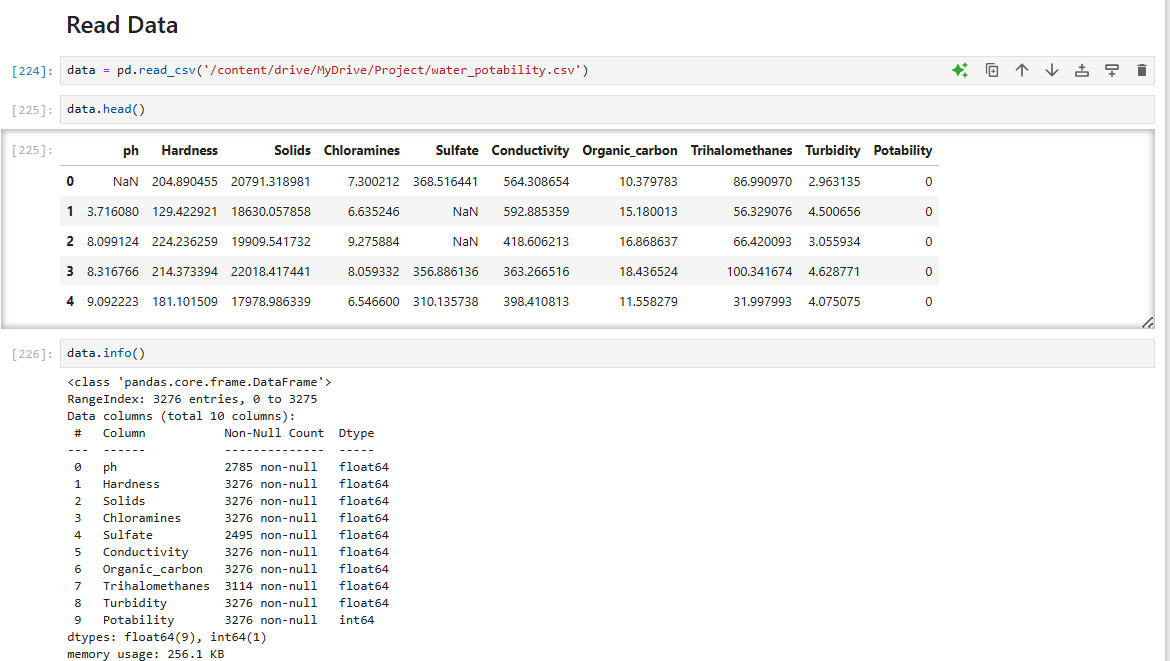
1. Import necessary libraries and load the water quality dataset.
2. Clean the dataset by handling missing or null values.
3. Perform exploratory data analysis (EDA) using visualization techniques to understand feature distributions and relationships.
4. Encode any categorical variables and scale numerical features for uniformity.
5. Split the dataset into training and testing subsets.
6. Train classification models such as logistic regression, support vector machines, and decision trees.
7. Evaluate model performance using metrics like accuracy, confusion matrix, and classification reports.
8. Interpret the results and visualize the model outcomes to assess classification effectiveness.

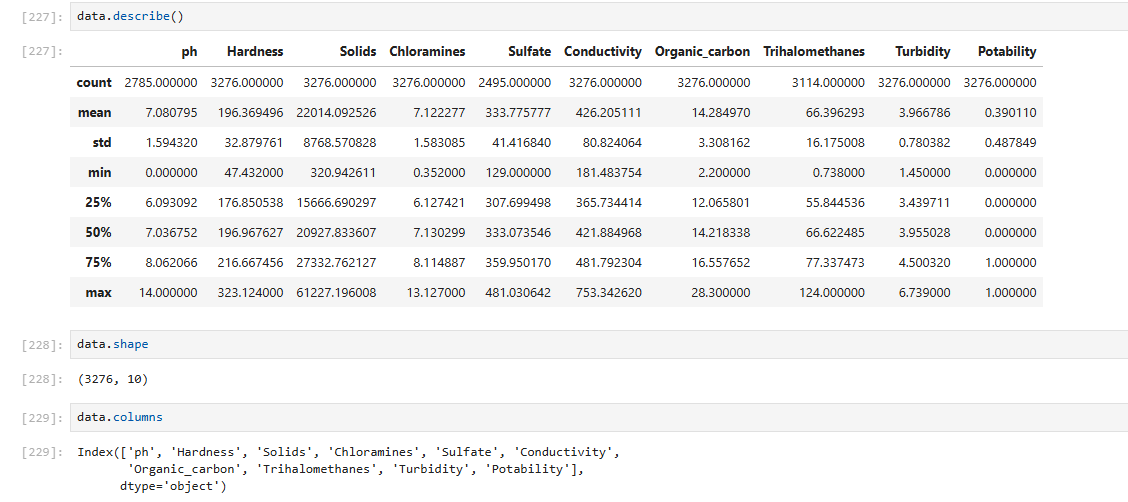
**Jupyter Notebook Code:**

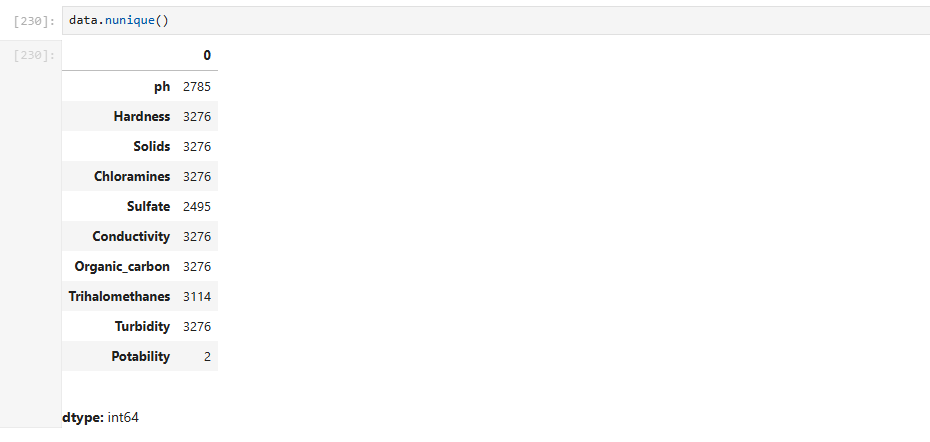
1. **Load Libraries:**

****

1. **Read Data**

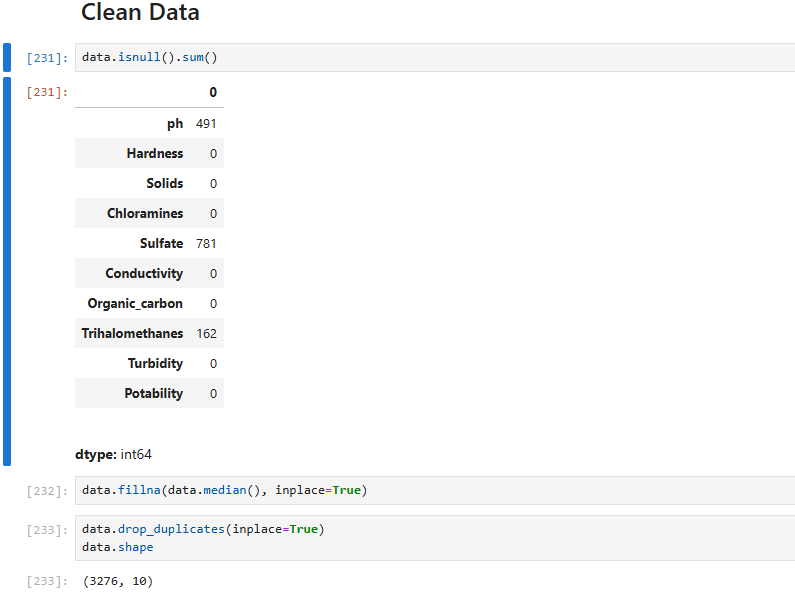
****

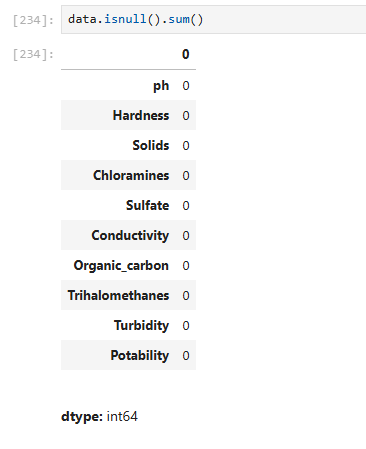
****

****

1. **Clean Data**

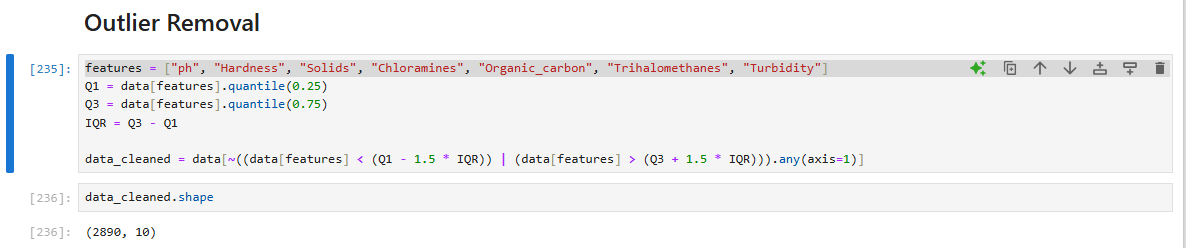
In this step, the dataset was carefully examined to find and fix any issues such as missing values, incorrect entries, or duplicated records. Cleaning the data ensured accuracy and consistency, which is essential for building a reliable regression model.

****

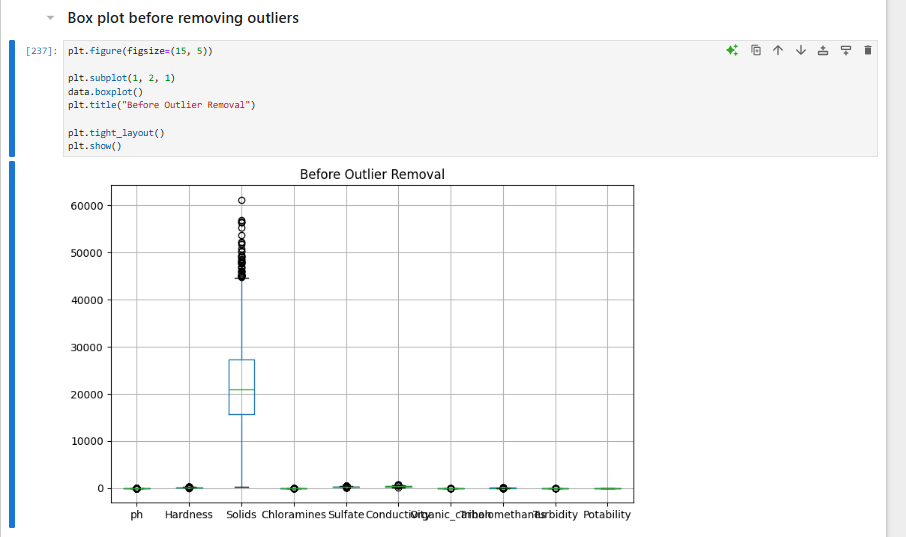
****

1. **Outlier Removal**

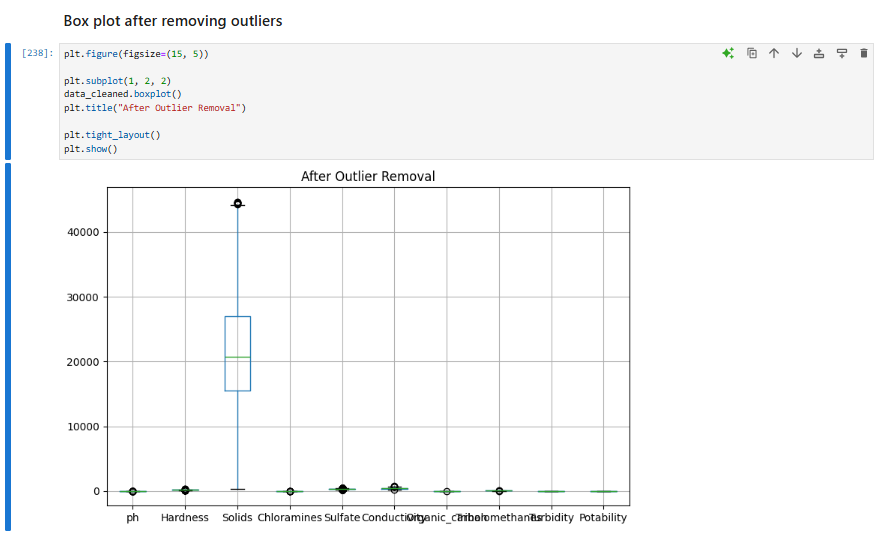
Outliers are extreme values that can negatively affect model performance. Removing them helps improve the accuracy and reliability of predictions.

****

**Box plot before removing outliers**

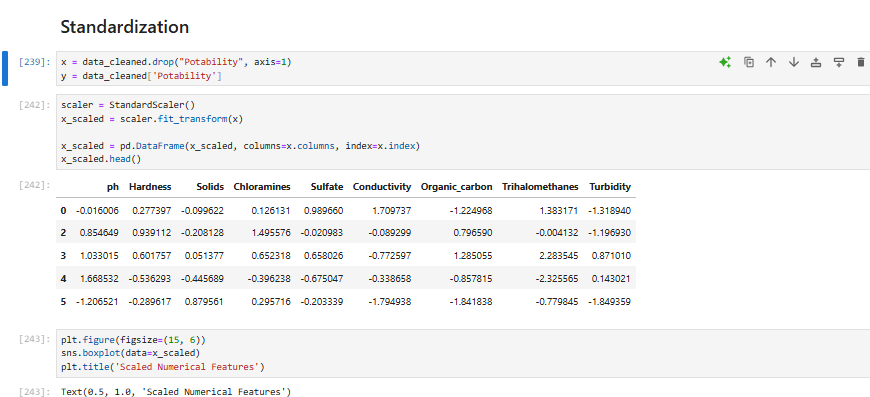
****

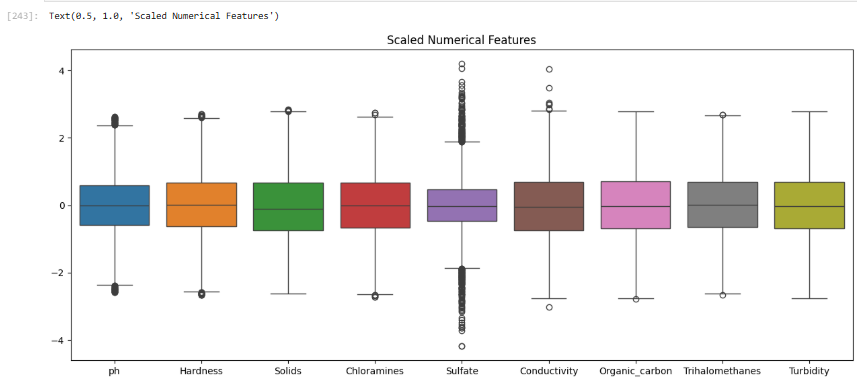
**Box plot after removing outliers**

****

1. **Standardization**

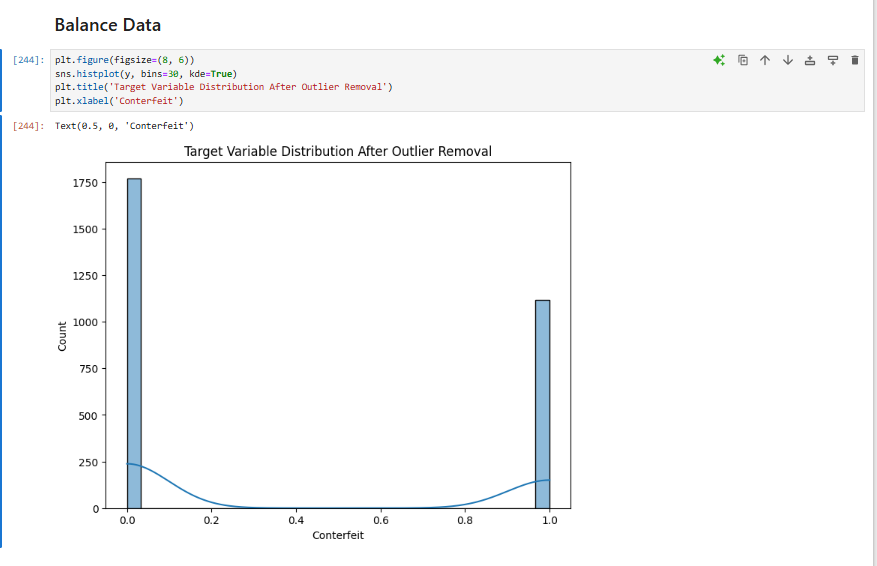
Standardization is a data preprocessing technique used to scale numerical features so that they have a mean of zero and a standard deviation of one. This ensures that all features contribute equally to the model training and helps improve the performance and convergence speed of many machine learning algorithms. Standardization is especially important when features are measured in different units or have varying ranges, such as pH levels versus mineral concentrations in water samples.



****

### Balanced Data

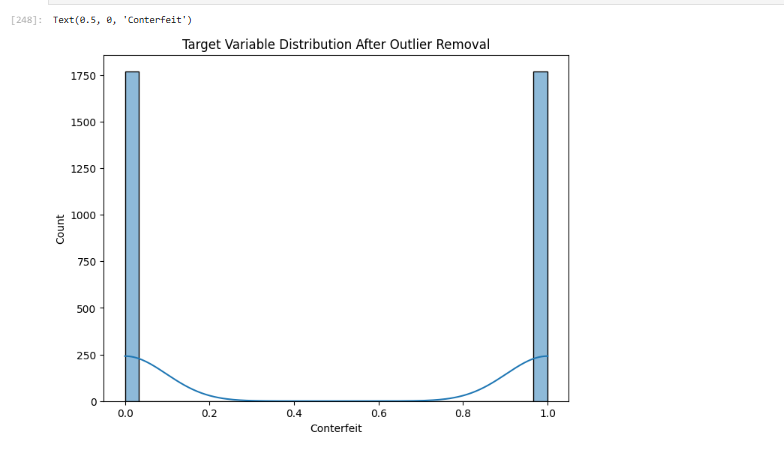
Balanced data refers to a dataset where the classes (e.g., potable vs. non-potable water) are represented in roughly equal proportions. Ensuring balanced data is important because imbalanced datasets can cause machine learning models to be biased toward the majority class, reducing their ability to accurately predict the minority class. Techniques like resampling, oversampling, or using specialized algorithms can help address data imbalance and improve classification performance.

****

### Resample

Resampling is a technique used to address class imbalance in datasets. It involves adjusting the dataset so that both classes (e.g., potable and non-potable water) have a more equal number of samples.

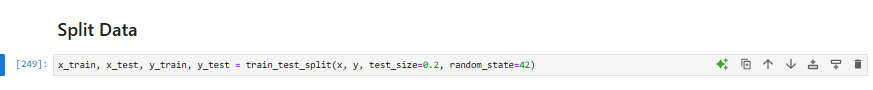
****

****

### Split Data

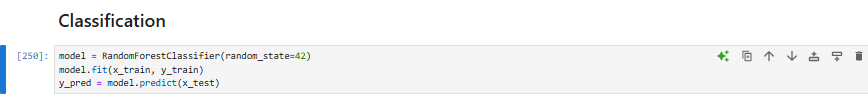
The dataset is divided into two parts:

* **Training set:** Used to train the model.
* **Testing set:** Used to evaluate model performance on unseen data.  
  This helps check how well the model generalizes.

****

### Classification

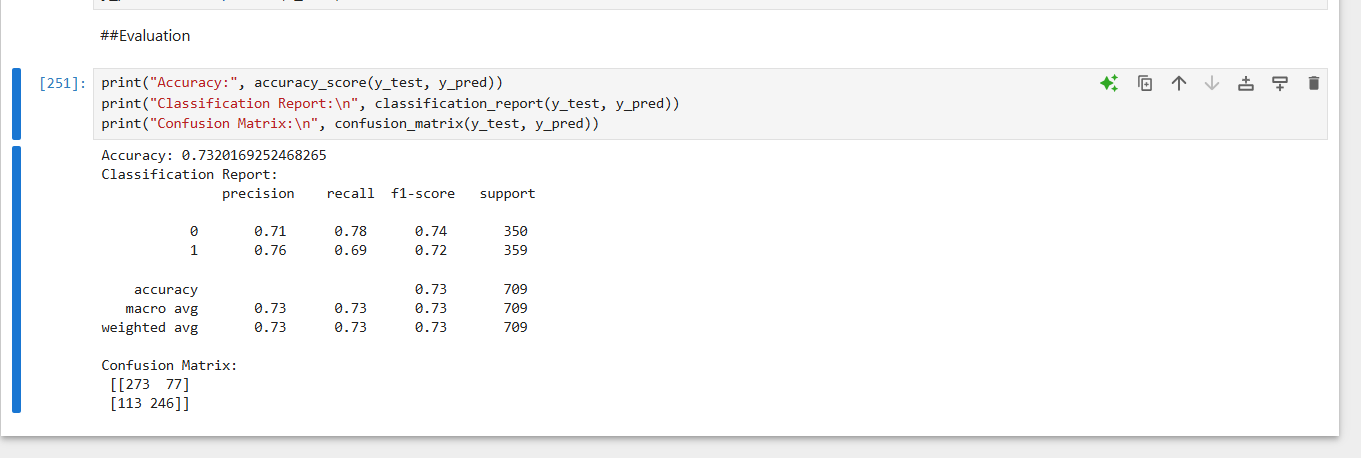
Classification is a supervised learning method used to predict categorical labels—in this case, whether water is **potable** or **not potable**. Models like Logistic Regression, SVM, and Decision Trees are trained to assign the correct class based on input features.

****

### Evaluation

Evaluation measures how well the model performs on unseen data.  
Common metrics include:

* **Accuracy:** Overall correctness of the model.
* **Confusion Matrix:** Shows true vs. predicted labels.
* **Classification Report:** Includes precision, recall, and F1-score.



### Conclusion

This project successfully applied machine learning classification techniques to predict water potability based on various chemical and physical attributes. By preprocessing the dataset, addressing class imbalance, and applying standardization, the data was prepared for effective model training. Models like Logistic Regression, Support Vector Machines, and Decision Trees were implemented and evaluated using performance metrics such as accuracy, confusion matrix, and classification report. The results demonstrated that machine learning can provide reliable predictions in determining whether water samples are safe for consumption.

The study highlights the practical application of data science in the field of environmental safety and public health. Automating the classification of water quality not only saves time and resources but also enhances decision-making in water management systems. With further improvements and real-world data integration, such models can support large-scale monitoring efforts and contribute to ensuring safe drinking water access across different regions.