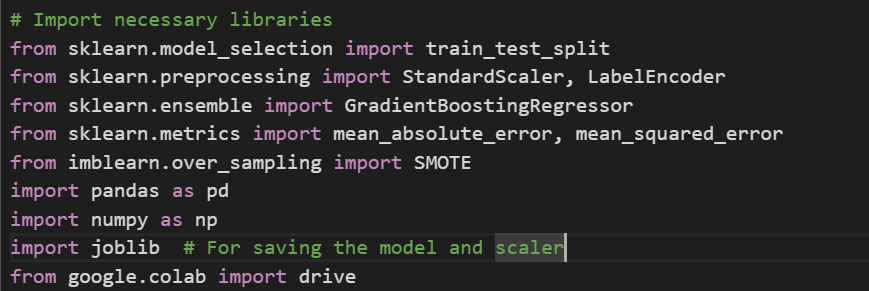
# ASSIGMENT -1 KRRISHIKA TANEJA 500120536 R2142230241 BATCH 12

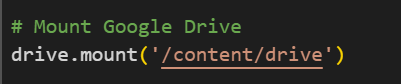
## Code with Explanations

### 1. Import Libraries

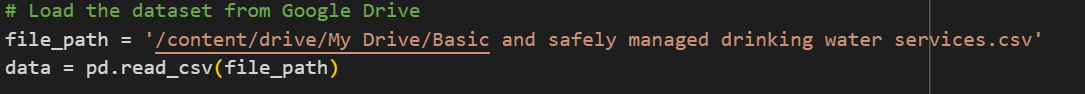
The necessary libraries are imported to handle data processing, machine learning, and model evaluation. These include

* pandas and numpy for data handling
* sklearn for modeling,
* imblearn for dealing with imbalanced data
* joblib for saving/loading models
* Google Colab Drive integration for accessing files in Google Drive.  
  

### 2. Mount Google Drive

This step connects Google Drive to the Colab environment, enabling the dataset to be loaded directly from Drive.  


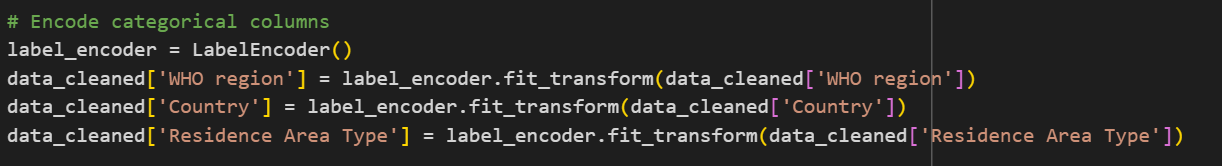
### 3. Load Dataset

Reads the CSV file containing data on drinking water services into a DataFrame for preprocessing and modeling.  


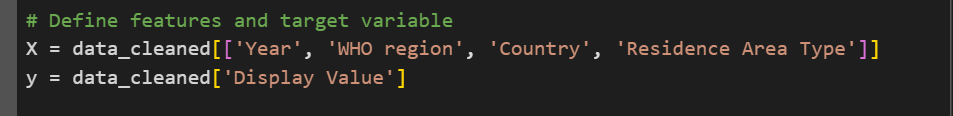
### 4. Preprocess the Data

Drops any rows with missing values in the 'Display Value' column, ensuring no missing data disrupts the modeling process.  

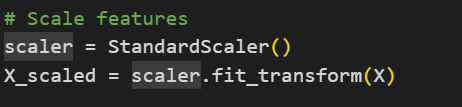

### 5. Encode Categorical Variables

Encodes categorical columns like 'WHO region', 'Country', and 'Residence Area Type' into numerical values, so they can be used effectively by the model.  


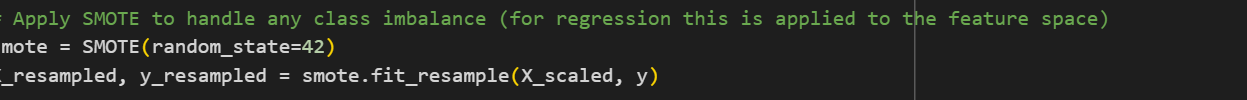
### 6. Define Features and Target

Splits the data into feature variables X and target variable y. Here, X includes 'Year', 'WHO region', 'Country', and 'Residence Area Type', while y contains 'Display Value' to predict.  


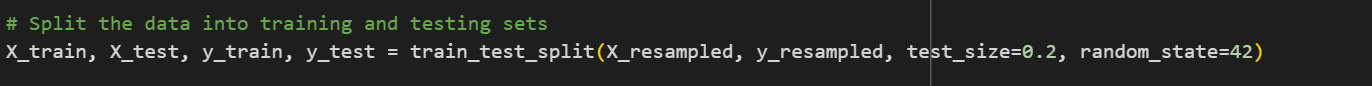
### 7. Scale Features

Uses StandardScaler to scale features to have mean 0 and variance 1, which improves the model's performance.

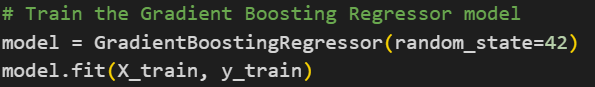
### 8. Apply SMOTE

SMOTE is used here to generate synthetic samples in feature space, helping to balance the data by oversampling minority classes, even though it’s typically for classification tasks.  


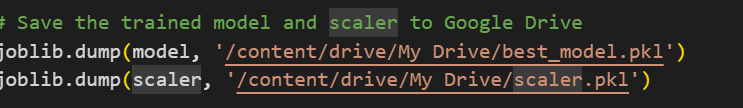
### 9. Split Data

Splits the resampled data into training and testing sets, ensuring the model is trained and evaluated on separate data.

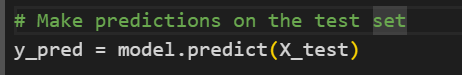
### 10. Train the Model

Initializes and trains a Gradient Boosting Regressor, which is suitable for regression tasks as it builds trees iteratively, correcting errors of previous iterations.

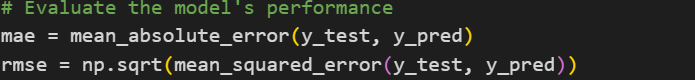
### 11. Save Model and Scaler

The trained model and scaler are saved to Google Drive using joblib. This step allows the model and scaler to be reused without retraining. 

### 12. Make Predictions

The trained model is used to predict values for the test data.  


### 13. Evaluate Model Performance

Evaluates the model using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which measure model performance. 

### 14. Display Evaluation Metrics and Predictions

Prints the evaluation metrics and compares the first 10 actual and predicted values in a Data Frame for visual inspection. 