## horizontal line



CSTP2301 Final Project V1

16.03.2024

**─**

ML

Jia Xi Lin

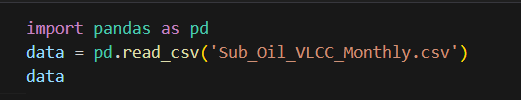
Muochu Hu

# Goals

1. Picking IDs 542236, 67321, 549295, 41108, 54982 to do the calculations.
2. Using a regression model to train and test the data set.
3. Comparing the actual value of y and predicted value of y and calculate the accuracy of them, fill the results in an Excel file.

#### **Step 1: Read the Dataset**

Begin by importing necessary libraries and loading the dataset. Use pandas to read the CSV file containing the data. This step is crucial as it sets up your data for preprocessing and analysis.

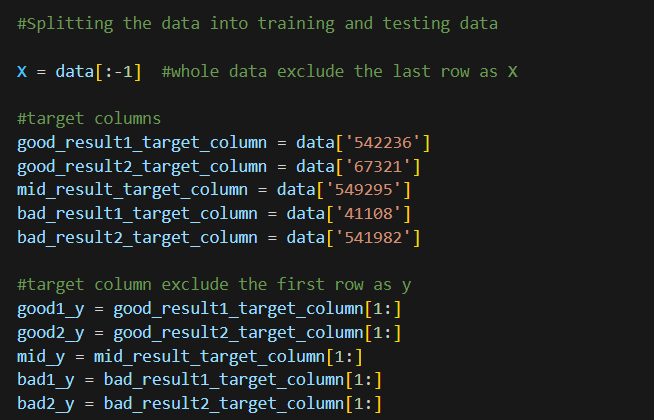


#### **Step 2: Data Preprocessing**

Prepare the dataset for the model. This involves splitting the dataset into features (X) and target variables (y). Since the model requires numerical inputs, ensure that the data is cleaned and appropriately formatted.

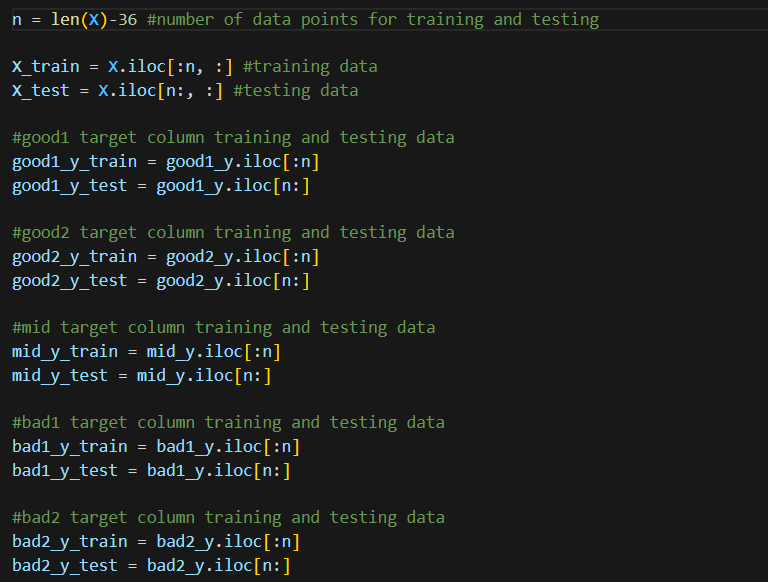
* Exclude the last row from the dataset for features (X).
* Define target columns for various scenarios (good, mid, and bad results).
* Exclude the first row from target columns to align with X.

To generate X and y, we can consider the entire table as X and copy the target column in a vector as y. Remember that the label of sample t in X is in row t+1 in y.



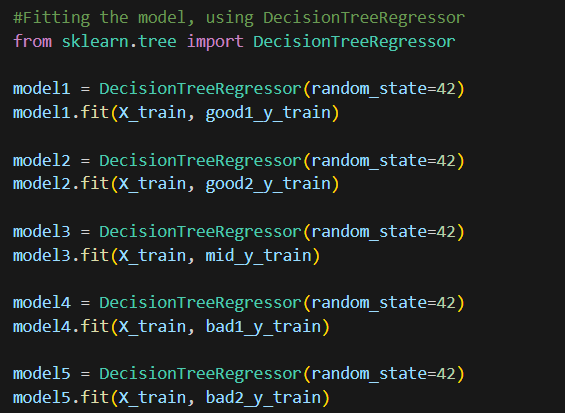
#### **Step 3: Splitting the Data into Training and Testing Sets**

It's essential to evaluate the model's performance on unseen data. Split the dataset into a training set used for learning and a testing set for evaluation. We aim to assess the model's accuracy over the last three years available in the dataset. To achieve this, consider training the model on all samples from the beginning up to n-36, and then test it on the last 36 samples (n is the total number of samples excluding the last one that doesn’t include a label)



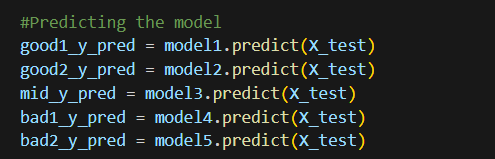
#### **Step 4: Model Selection and Training**

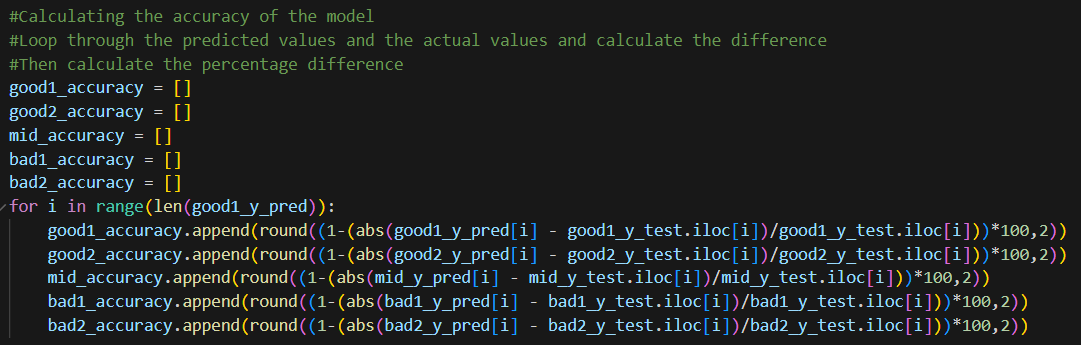
Choose an appropriate regression model based on your problem and dataset characteristics. Here, a DecisionTreeRegressor is used. Train separate models for each target variable using the training data.

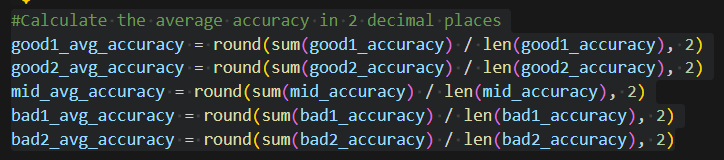


#### **Step 5: Model Prediction and Evaluation**

Use the trained models to predict outcomes on the testing set. Evaluate the models' performance by comparing the predicted values against the actual values in the testing set. Calculate accuracy as a measure of performance.



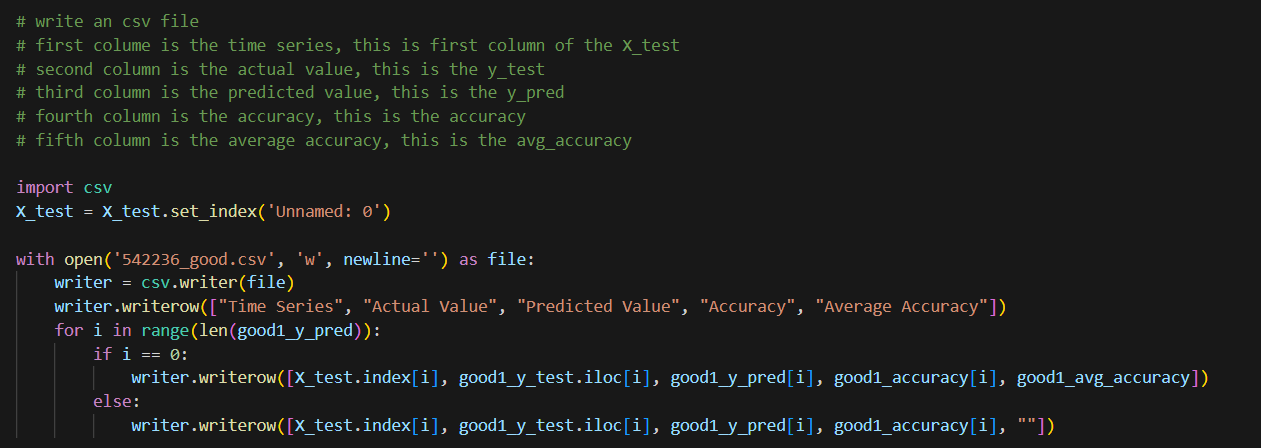




#### **Step 6: Exporting Results**

Organize the predictions, actual values, and accuracies into a structured format and export them as CSV files. This step is vital for reporting and further analysis.

Example of writing a CSV file for ID 542236 ‘s prediction:



### **Conclusion**

#### This project embarked on an exploration of regression models to predict various outcomes based on the dataset from Sub\_Oil\_VLCC\_Monthly.csv. By employing a DecisionTreeRegressor, we aimed to forecast outcomes under different conditions categorized as good, mid, and bad results.

#### Our methodology was grounded in a systematic approach, beginning with data preprocessing, followed by splitting the dataset into training and testing sets to ensure a robust evaluation of the model's predictive capabilities. The Decision Tree models were then trained on the respective subsets of the data, each targeting a different outcome variable.

#### The predictions made by these models on the testing set revealed insights into their performance. The accuracy calculations, derived from comparing predicted values against actual outcomes, served as a critical metric for assessing model efficacy. The exported results, encompassing time series, actual values, predicted values, and accuracies, provided a granular view of the model's performance across different scenarios.

Upon analyzing the calculated accuracies, it was discerned that the models predicting for the conditions labeled '542236' and '67321' demonstrated the highest accuracy, indicating their superior predictive performance. The model for '549295' yielded average accuracy, suggesting a moderate level of reliability in its predictions. Conversely, the models associated with '41108' and '541982' conditions exhibited the lowest accuracies, highlighting challenges in forecasting under these specific scenarios and underscoring areas for potential improvement in future iterations of the project.

#### 