The dataset:

In this report we used the dataset about weather in the north

Link of the dataset : https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?select=metrics.py

The goal of this project is to perform data preprocessing and model evaluation on a dataset with a target variable representing temperature, specifically "TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)." The project includes several tasks:

Data cleaning and preprocessing, including handling missing values, outlier removal, and normalization

Model selection based on the nature of the target variable (regression or classification)

Hyperparameter tuning and evaluation of machine learning models (SVR/SVC) for accurate predictions.

Reporting the results of model performance, including the evaluation of various metrics such as accuracy, precision, recall, F1 score, and regression errors.

the dataset is split into two parts for parallel processing, and each part is used to train and evaluate separate models

Data Preprocessing Steps:

1)Filling The Missing Values

2) Remove The Outlier Detection

3) Data Normalization

4) Categorical Data Encoding

5) Data Splitting , into tow equal subset

Model Selection:

The choice between a regression or classification model is determined based on the target variable. If the target variable has more than 10 unique values, a regression model is chosen (SVR - Support Vector Regression). Otherwise, a classification model (SVC - Support Vector Classifier) is selected

Tuning:

Tuning is performed using RandomizedSearchCV. This approach randomly samples from a distribution of hyperparameters and evaluates each combination to find the best configuration for the model.

Model Evaluation:

The model's performance is evaluated based on the type of task:

1 Classification

2 Regression

Results:

```
Performing Hyperparameter Tuning...

Best Parameters: {'C': 1.2973169683940733, 'epsilon':
0.025601864044243652, 'gamma': 0.025599452033620268}

Mean Squared Error (MSE): 0.0009
Root Mean Squared Error (RMSE): 0.0307
```

Challenges Faced During the Project:

1 It takes too much time to train the model

2 Outliers: Removing outliers based on the IQR method can lead to the loss of valuable information, especially when extreme values are valid observations

3 Model Selection: The process of determining whether the task should be treated as classification or regression was essential for choosing the right model. It required a careful inspection of the target variable's distribution.

4 Parallel Data Processing: Splitting the data into two parts and training models on each part independently added complexity in terms of managing multiple models and comparing their results.

Conclusion:

The project involved significant data preprocessing, model training, and evaluation. The results indicate that the models can provide useful predictions