Summary of project

Datasets:

1. Primary dataset - train.csv and test.csv for Mercedes-Benz-Greener-Manufacturing

Tasks:

1. If for any column(s), the variance is equal to zero, then you need to remove those variable(s).
2. Check for null and unique values for test and train sets.
3. Apply label encoder.
4. Perform dimensionality reduction.
5. Predict your test_df values using XGBoost.

Steps involved:

1. Pandas library has been used to import data from datafile and doing exploratory data analysis, removing columns with zero variance, removing rows with null values.

2. Since unique values in test data is not a subset of that in train data column wise Unicode label encoding has been applied.

3. PCA has been done on data and using variance plot , n_components = 4 has been selected for dimensionality reduction.

4. XGBRegressor has been used to fit the model and predict output for test data. Scoring has also been done on validation data.