Algorithm for ML Models for Normality

Algorithm 1 Data Generation Algorithm

Require: Sample size n, number of samples N, distribution type `dist`, label `label`

Ensure: Dataset with statistical features and labels

for $i = 1$ to N **do**
    Generate n samples from the specified distribution `dist`
    Calculate statistical features:
        Skewness, kurtosis, Jarque-Bera statistic, Anderson-Darling statistic
        Zero-crossing rate, Gini coefficient, number of outliers
        Shapiro-Wilk statistic, Liliefors statistic, Cramer-von Mises statistic
        Sample size, range, coefficient of variation, and energy
    Assign the label `label` to the data
end for
Combine all generated data into a single dataset
return Dataset containing statistical features and labels

2. Data Preparation for Machine Learning Models

Algorithm 2 Data Preparation Algorithm

Require: Dataset

Ensure: Training and testing datasets with normalized features

Shuffle the dataset to randomize sample order
Split the data into training (70%) and testing (30%) sets
for each numeric feature in the dataset **do**
    Apply min-max normalization: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
end for
Ensure the `Label` column is treated as a categorical variable
return Training and testing datasets

3. Model Building and Testing Step

Algorithm 3 Model Building and Testing Algorithm

Require: Training and testing datasets
Ensure: Trained models and performance metrics
1: Define models: Logistic Regression, Random Forest, ANN, GBM, SVM, KNN
2: **for** each model **do**
3: Train the model using the training dataset
4: Predict labels for the test dataset
5: Evaluate the model using a confusion matrix
6: **end for**
7: Calculate performance metrics: accuracy, precision, recall, F1-score
8: **return** Trained models, predictions, and performance metrics

4. Validation Step

Algorithm 4 Validation Algorithm

Require: Validation dataset, trained models
Ensure: Performance metrics on the validation dataset
1: Generate validation dataset with distributions not used in training
2: Calculate statistical features as in the data generation step
3: Shuffle the validation dataset
4: Normalize the validation dataset using min-max normalization
5: **for** each trained model **do**
6: Predict labels for the validation dataset
7: Evaluate predictions using a confusion matrix
8: **end for**
9: Calculate performance metrics: accuracy, precision, recall, F1-score
10: **return** Performance metrics on the validation dataset

5. Plotting and Visualization

Algorithm 5 Plotting and Visualization Algorithm

Require: Trained models, test data
Ensure: ROC curves and variable importance plots
1: Plot ROC curves for all models:
2: Compute true positive rate (TPR) and false positive rate (FPR)
3: Plot the ROC curve and compute AUC
4: For Random Forest model:
5: Calculate variable importance scores
6: Plot variable importance to visualize feature contributions
7: **return** ROC curves and variable importance plots