Random Forest Documentation

1. Introduction

Random Forest is a supervised machine learning algorithm that uses an ensemble of decision trees to perform classification tasks. It improves prediction accuracy by reducing overfitting through the aggregation of multiple tree predictions.

2. Objective

- Develop an interactive web application using Streamlit for Random Forest classification.
- Provide users with the ability to upload a dataset, select features, train the model, and evaluate its performance.

3. Dataset Description

- Dataset: User-provided CSV file.
- Features: User-selected input variables.
- Target: User-selected variable to predict.
- Number of records and attributes depend on the uploaded dataset.

4. Implementation Details

- **Frontend**: Developed using Streamlit for user interaction.
- **Backend**: Random Forest implemented using `scikit-learn`.
- **Steps**:
  1. Upload a CSV dataset.
  2. Select features and target variable.
  3. Configure hyperparameters (e.g., number of estimators, max depth).
  4. Train the model using Random Forest.
  5. Evaluate the model using accuracy, confusion matrix, and classification report.

5. Results and Analysis

- Model accuracy and evaluation metrics are displayed.
- Feature importance can be visualized using bar charts.

- The impact of different hyperparameters on model performance can be analyzed.

6. Challenges and Solutions

- Managed large datasets efficiently using parallel processing.

- Prevented overfitting through hyperparameter tuning.

- Addressed class imbalance using appropriate techniques.

7. Conclusion

Random Forest is a robust algorithm that delivers high accuracy and handles large datasets effectively. The web application allows users to explore model behavior and analyze feature importance.

8. References

- Scikit-learn Documentation

- Streamlit Documentation

- Dataset Source