Bagging with Decision Tree

**Ex No**: 6  
**Date**: 14/08/2024  
**Name**: Ruban s - 71762233039

**Aim:**

To build a Bagging ensemble model using a Decision Tree to predict employee attrition based on various features in the provided dataset.

**What is Bagging?**

Bagging, or Bootstrap Aggregating, is an ensemble machine learning technique used to improve the stability and accuracy of machine learning models. It is particularly effective for algorithms that have high variance, such as decision trees. Bagging works by creating multiple versions of a dataset using bootstrapping (sampling with replacement), training a model on each dataset, and then aggregating the predictions.

**Procedure:**

1. **Import Necessary Libraries and Load the Dataset:**
   * Import libraries such as pandas, numpy, scikit-learn, matplotlib, and seaborn.
   * Load the employee attrition dataset into a DataFrame.
2. **Encode Categorical Data:**
   * Use LabelEncoder or mapping to transform categorical columns (such as Gender, Department, Job\_Title) into numerical values.
3. **Prepare the Data:**
   * Select features like Age, Gender\_encode, Department\_encode, Job\_Title\_encode, Years\_at\_Company, Satisfaction\_Level, Average\_Monthly\_Hours, Promotion\_Last\_5Years, and Salary, and define the target variable (Attrition).
   * Split the data into training and testing sets using train\_test\_split.
4. **Define the Base Model:**
   * Use the DecisionTreeClassifier as the base estimator for bagging.
5. **Implement Bagging:**
   * Use BaggingClassifier with the DecisionTreeClassifier as the base estimator.
   * Define the number of base models (estimators) and other parameters such as max\_samples and max\_features.
   * Train the bagging model on the training data.
6. **Evaluate the Model:**
   * Make predictions on the test data.
   * Evaluate the model using metrics such as accuracy score, classification report, and confusion matrix.
7. **Plot Feature Distributions:**
   * Use seaborn to plot the distributions of features such as Age, Gender\_encode, Department\_encode, Job\_Title\_encode, and Satisfaction\_Level.
8. **Plot Confusion Matrix:**
   * Plot the confusion matrix for the Bagging model to visualize the classification performance.

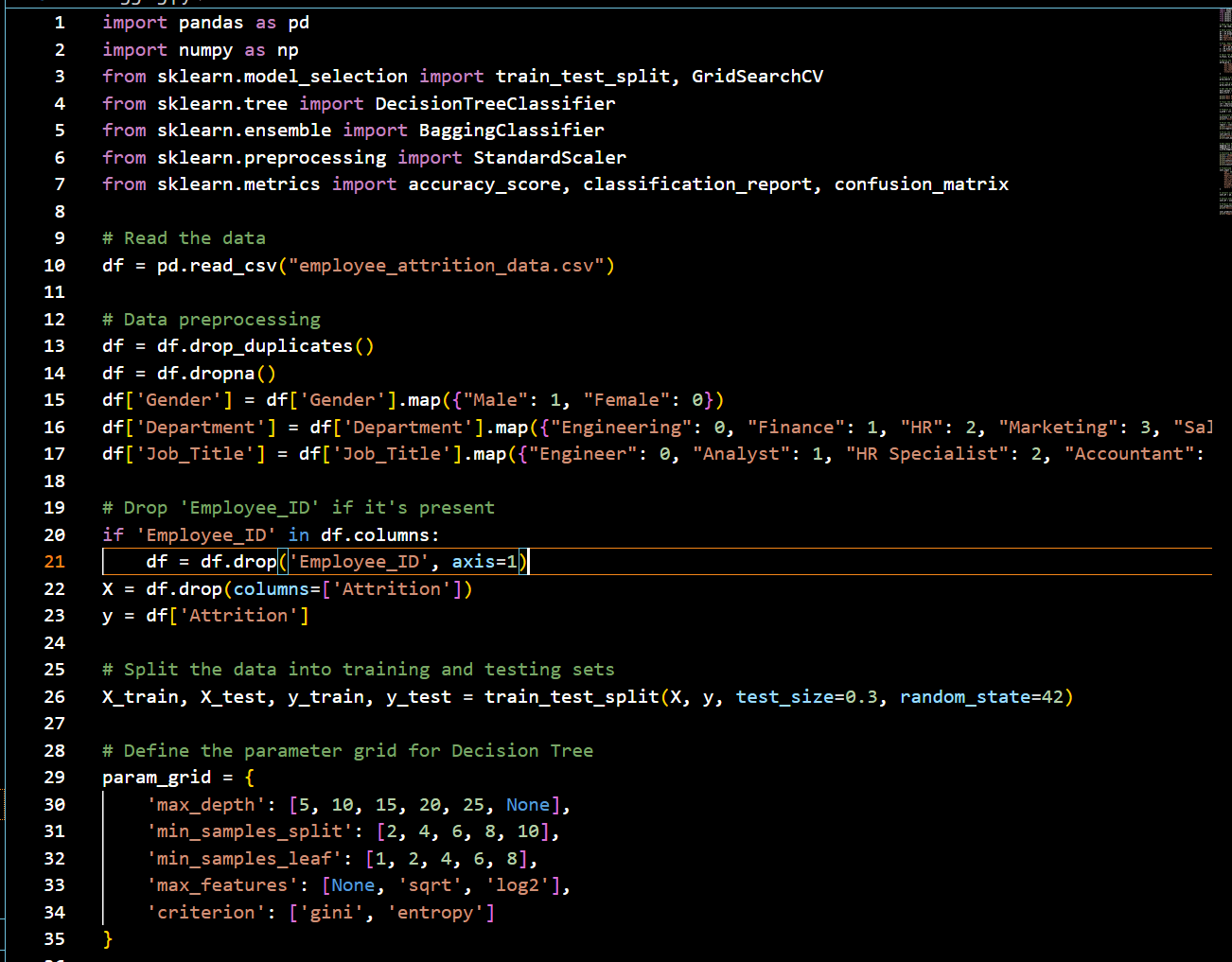
**Choosing Features in Bagging:**

Features are selected based on their relevance to the target variable. In this exercise, features like Age, Gender\_encode, Department\_encode, Job\_Title\_encode, Years\_at\_Company, Satisfaction\_Level, Average\_Monthly\_Hours, Promotion\_Last\_5Years, and Salary were chosen based on their correlation with the target variable Attrition.

**When Can Overfitting Occur?**

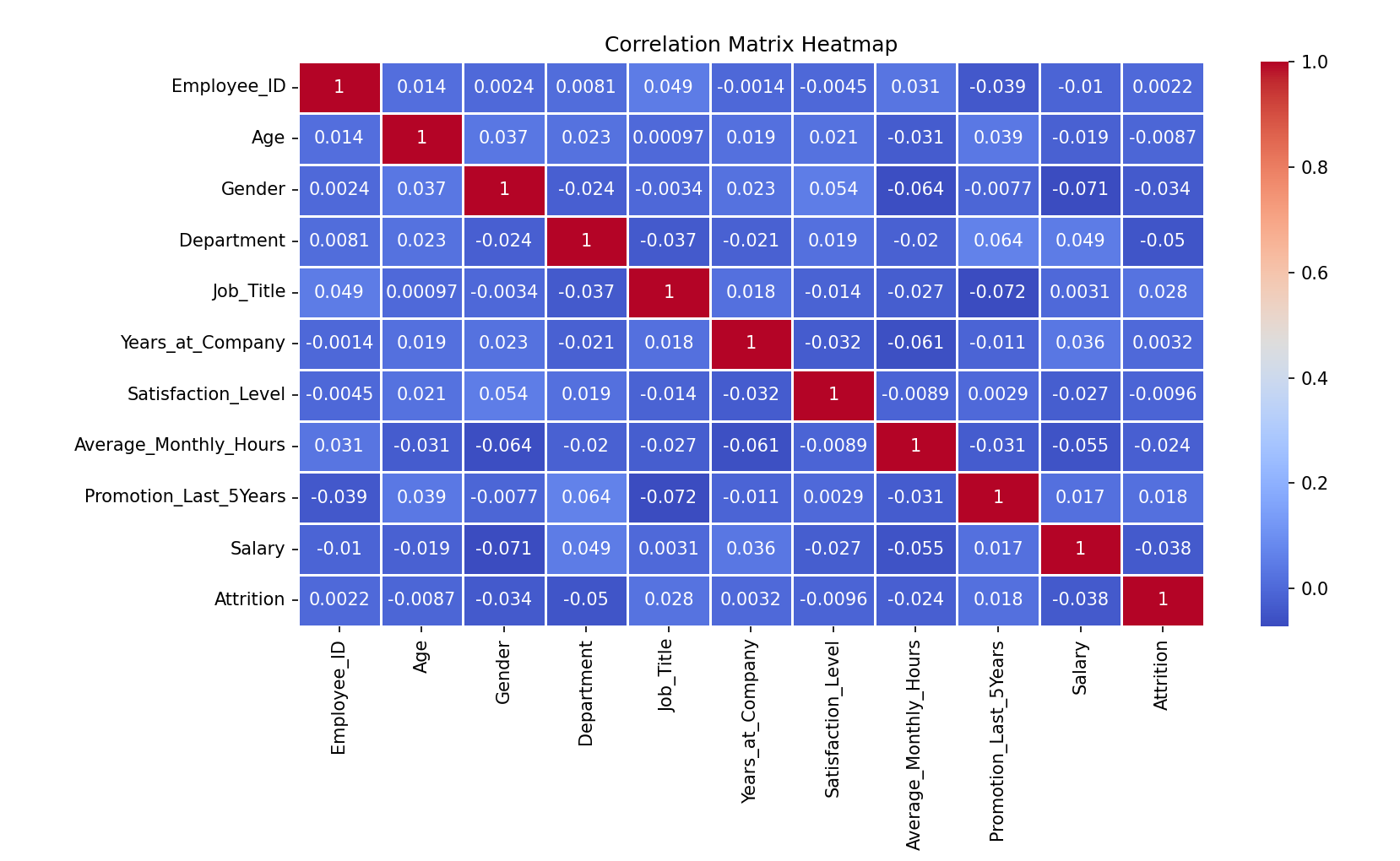
Overfitting can occur when a model is too complex and captures noise in the training data. In the context of Bagging, overfitting is less likely because the aggregation of multiple models reduces variance. However, it can still occur if individual decision trees are too deep or if the number of estimators in the Bagging ensemble is too small.

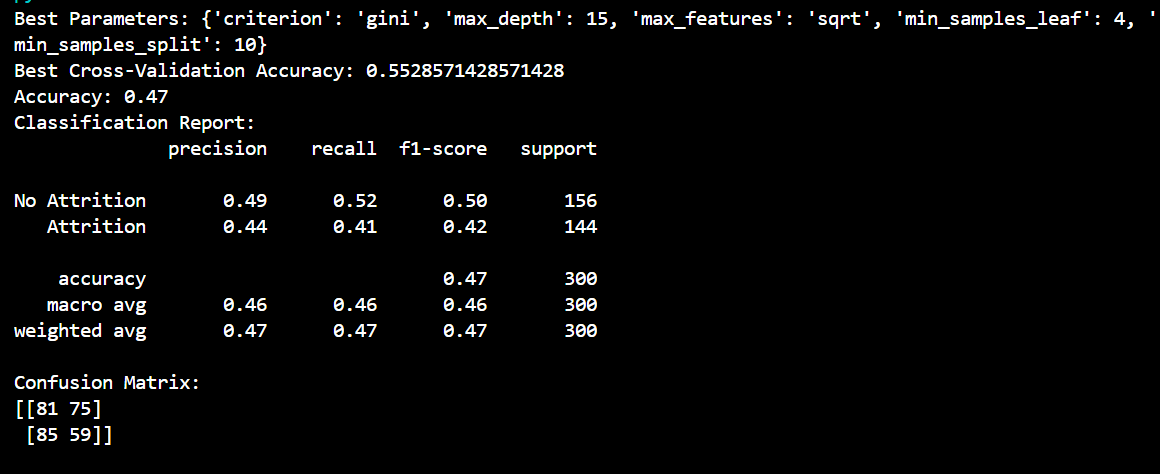
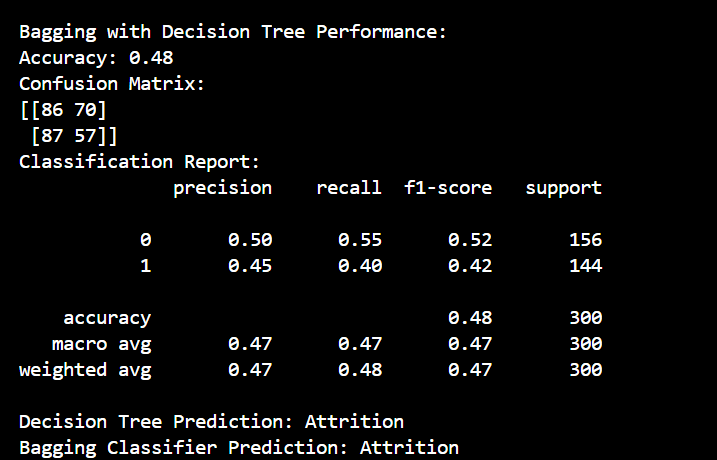
CODE:





**OUTPUT:**



**Conclusion:**

Bagging with Decision Trees is a powerful ensemble technique that improves the accuracy and robustness of predictions. By aggregating the outputs of multiple models, Bagging reduces variance and helps prevent overfitting. This approach is particularly effective for decision trees, which can be prone to high variance.