Name: Siddharth Chatla
ID : 991409470

Dataset Description:

About:

Author : Mike Chapman, NASA

PC2 and PC4 Software defect prediction:

The NASA Metrics Data Program includes a set of defect data derived from the flight software used in an earth orbiting satellite. This data was collected using McCabe and Halstead feature extractors, which were developed in the 1970s to objectively characterize code features that contribute to software quality.

The use of these metrics has been shown to be effective in predicting software quality and identifying potential defects. By analyzing the McCabe and Halstead features extracted from the earth orbiting satellite's flight software, NASA was able to identify and classify software defects, which helped improve the software's reliability and performance.

This highlights the importance of using objective and standardized methods to measure and evaluate software quality, which can lead to better understanding and improvement of software systems.

Dataset:

PC2

Features : 37 Features (36 Numeric Features)
Target : Binary Classified
Instances : 5589

PC4

Features : 38 Features (37 Numeric Features)
Target : Binary Classified
Instances : 1458

Implementation:

In general, varying the min_samples_leaf parameter can affect the performance of a decision tree model.

To measure the training and test ROC-AUC scores on 10-fold cross-validation, you can use the following steps:

1. Split your data into training and testing sets.
2. Create a decision tree model with different values of `min_samples_leaf`.
3. Use 10-fold cross-validation on the training data to calculate the mean ROC-AUC score.
4. Evaluate the model on the testing data and record the ROC-AUC score.
5. Repeat for different values of `min_samples_leaf`.

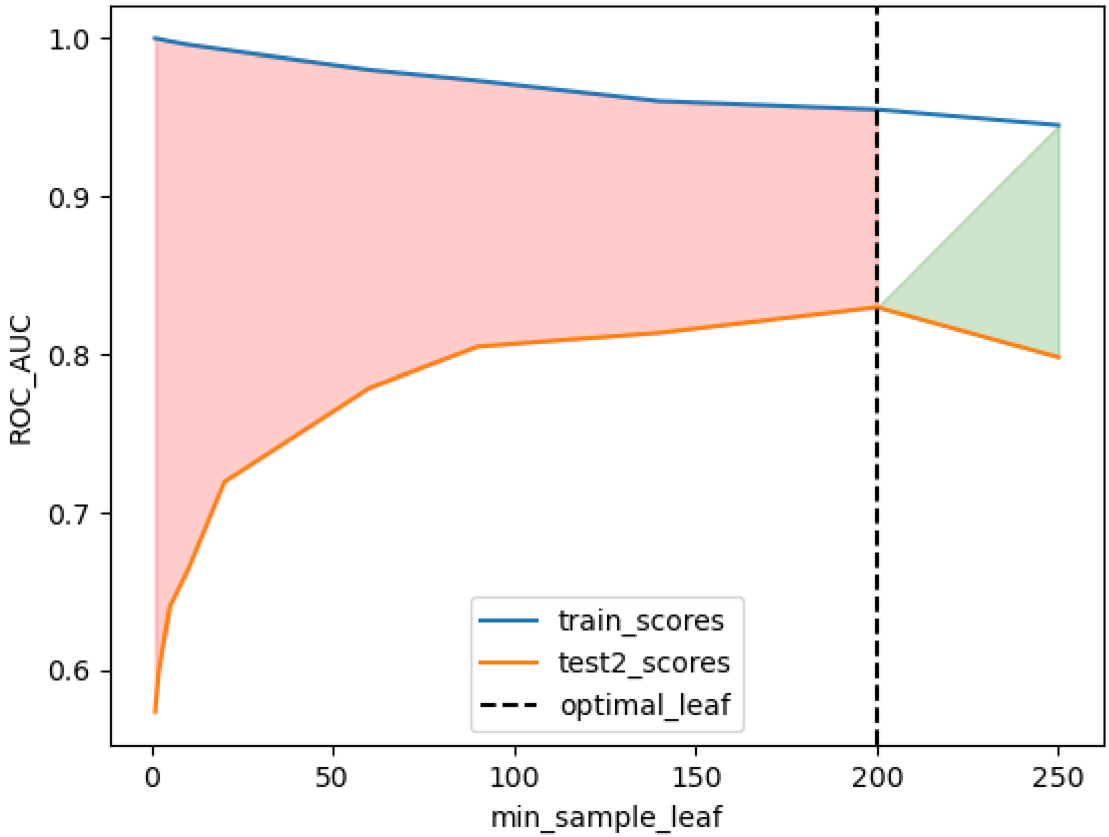
Once you have the ROC-AUC scores for each value of min_samples_leaf, you can plot them on a graph with min_samples_leaf on the x-axis and ROC-AUC scores on the y-axis. You can then look for regions of overfitting and underfitting.

In general, as the value of min_samples_leaf increases, the model becomes less complex and more likely to underfit. On the other hand, as the value of min_samples_leaf decreases, the model becomes more complex and more likely to overfit.

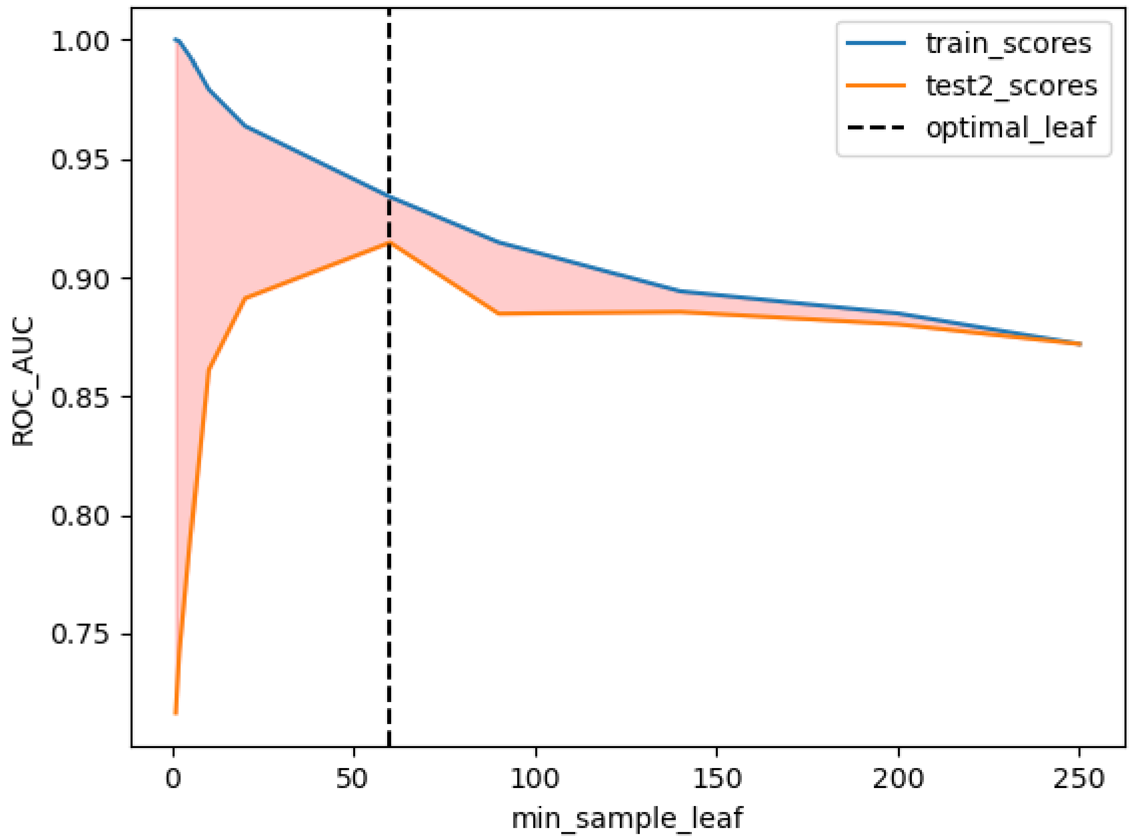
So, on the graph, we see the ROC-AUC score increase as `min_samples_leaf` decreases, but then start to level off or even decrease again as the model begins to overfit. The point where the ROC-AUC score starts to level off or decrease is the point of optimal min_samples_leaf.

If the ROC-AUC score for the training data is much higher than the ROC-AUC score for the testing data, this indicates overfitting, which usually occurs when min_samples_leaf is too small. If the ROC-AUC score for both the training and testing data are low, this indicates underfitting, which usually occurs when min_samples_leaf is too large.

Graph for PC2 dataset:



Graph for PC4 Dataset:



Evaluation measures:

PC2 Dataset

Classifier	Test Accuracy	Mean RUC
Decision tree with default parameters	[0.48922801 0.48922801 0.48653501 0.49102334 0.48653501 0.49102334]	0.599904851702121
Decision tree with tuned min_sample leaves GridSearchCV	[0.97935368 0.98294434 0.98204668 0.39048474 0.69658887 0.69883303]	0.7971258540743706

metrics	score
accuracy_score	0.9996421542315262
f1_score	0.9545454545454545
precision_score	1.0
recall_score	0.9130434782608695

PC4 Dataset

Classifier	Test Accuracy	Mean RUC
Decision tree with default parameters	[0.67534722 0.68511285 0.77907986 0.75716146 0.73871528 0.80078125 0.73046875 0.70269097 0.73736213 0.68152574]	0.7288245506535949
Decision tree with tuned min_sample leaves GridSearchCV	[0.92100694 0.90755208 0.88020833 0.93446181 0.87391493 0.88346354 0.90625 0.91210937 0.90257353 0.89774816]	0.901928870506536

metrics	score
accuracy_score	0.9993141289437586
f1_score	0.9971988795518207
precision_score	0.994413407821229
recall_score	1.0

Conclusion:

I imported the PC2 and PC4 datasets and utilized them to train two different types of classifiers. Through cross-validation, I determined the Area under Curve (AUC) values and noticed that the Decision tree classifier with tuned min_samples_leaf, using GridSearchCV, yielded the highest mean scores. On the other hand, the Decision tree classifier with default parameters produced the lowest scores.