RITWIK SRIVASTAVA - I6303400

MACHINE LEARNING
LAB 1 - DECISION TREES

Introduction

In this lab, our task involved tackling the Diabetes Classification Problem and the Glass Classification Problem. We were tasked with training, testing, and conducting experiments on the data provided. Our primary reference for understanding **`sklearn`** was the appendix accompanying the lab assignment. Our objectives included training both single-level and multi-level decision trees to assess their accuracy and subsequently compare the results obtained.

Moreover, the experimentation phase involved testing the accuracy of multi-level decision trees across different step sizes. Our goal was to identify regions indicating overfitting, optimality, and underfitting while providing sound reasoning behind these determinations.

LAB TASK C:

Conversion from One-Level Decision Trees to Multi-Level Decision Trees:

To implement the change, we had to change the parameter **`max_depth.`** The value when set to '1' gives us a one-level decision tree, while when set to 'None' gives us the multi-level decision tree.

```
clfGlass = tree.DecisionTreeClassifier(criterion='entropy', max_depth=None)
```

```
clfGlass = tree.DecisionTreeClassifier(criterion='entropy', max_depth=1)
```

The difference between the two decision trees is that one-level trees have a single split in their data while multi-level trees are more cmplicated, as they implement a multi split model which allows for deeper analysis. This should lead to more accurate results.

The parameters with which this data has been analysed:

```
test_size=0.2, random_state=42
```

The **`test_size = 0.2`** means that 20% of the data was used to test while 80% was used to train. Meanwhile, the **`random_state = 42`** was to ensure the reproducibility of results given the randomisation. The results clearly demonstrate that multi-level trees are more accurate for both the classification problems.

Diabetes Classification Problem

One-Level Decision Tree

```
Data Accuracy: 0.7402597402597403
```

Multi-Level Decision Tree

```
Data Accuracy: 0.7467532467532467
```

Glass Classification Problem

One-Level Decision Tree

```
Data Accuracy: 0.3488372093023256
```

Multi-Level Decision Tree

```
Data Accuracy: 0.7674418604651163
```

I tried to calculate these values but with a different **`test_size`**. I kept the **`random_state`** standard to make a comparison. This was done for my own understanding of these trees and the accuracy.

```
test_size=0.6, random_state=42
```

The **`test_size = 0.6`** means that 60% of the data was used to test while 40% was used to train. Meanwhile, the **`random_state = 42`** was to ensure the reproducibility of results given the randomisation. In this situation, there was half the data to train and we have obtained differing results where the accuracy advantage is lower for the multi-level graphs. This is because less data is being utilised to train complex models. Here, simpler models have the upper hand as undertrained complex models can overfit/underfit the data.

Diabetes Classification Problem

One-Level Decision Tree

```
Data Accuracy: 0.7266811279826464
```

Multi-Level Decision Tree

```
Data Accuracy: 0.6941431670281996
```

Glass Classification Problem

One-Level Decision Tree

```
Data Accuracy: 0.4263565891472868
```

Multi-Level Decision Tree

```
Data Accuracy: 0.5193798449612403
```

Difference in Accuracy Rate:

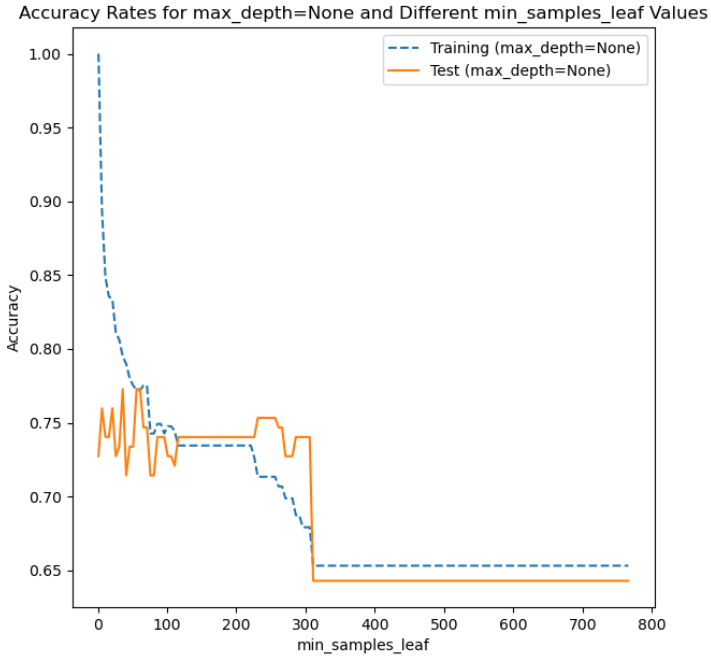
- The difference in accuracy rates between one-level and multi-level decision trees can stem for a multiude of factors. It is more likely that one-level trees underfit the data and multi-level trees over fit the data (when not pruned properly). This can result in the multi-level decision tree being more accurate for testing and training data.
- Since one-level trees have only a singular split in data, they are not able to identify patterns as well as multi-level trees can. Multi-level trees can be especially useful for huge datasets which cannot be efficiently calculated by hand. However, there is always an accuracy vs explainability trade-off that one needs to consider.
- In terms of explainability, it must be considered that one-level decision trees are more explananble than accurate than multi-level trees. This is due to the inherent simplicity of the former or the blatant complexity of the latter. This affects the interpretability of these trees, making one-level trees easier to interpret, visualise and work with.
- The lack of complexity from one-level trees allows for concise explanation of decision rules. This also prevents one-level trees from overfitting data.

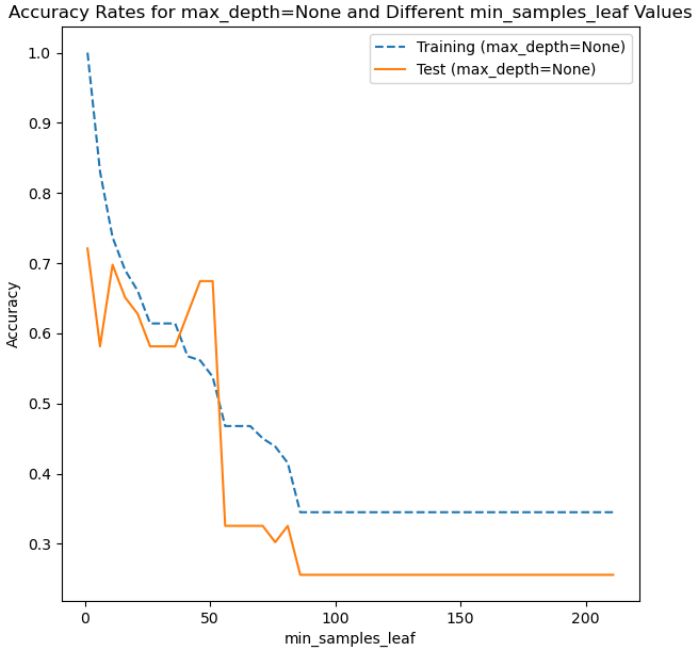
LAB TASK D:

The parameters with which this data has been analysed:

```
test_size=0.2, random_state=42
```

The **`test_size = 0.2`** means that 20% of the data was used to test while 80% was used to train. Meanwhile, the **`random_state = 42`** was to ensure the reproducibility of results given the randomisation. The first set of figures are representing the accuracy of the multi-level decision trees by error pre-pruning. The step size here is 5. For the experimental sets, the step size changes. With the change in step size, we can see that the estimations for overfitting, optimal fitting and underfitting differ. This is because of the step size, as with a step size of 20, we see that the pruning is more substantial and with the step size of 1, it is not pruned.





Here, we can see the two figures. The first figure is for the Diabetes Classification Problem and the second figure is for the Glass Classifcation Problem.

Diabetes Classification Problem

Overfitting: The accuracy of testing vs training drops just before the first 100 samples. Until this point the graph was overfit for the training data, which is why the accuracy was much higher than the testing data.

Optimality: The accuracy of testing vs training converges before 100 and until 300 samples. This is when it is optimally fitted for the data.

Underfitting: The accuracy of testing vs training is consistently lower beyond the 300 sample mark. This is when the data can be classified as underfit.

Glass Classification Problem

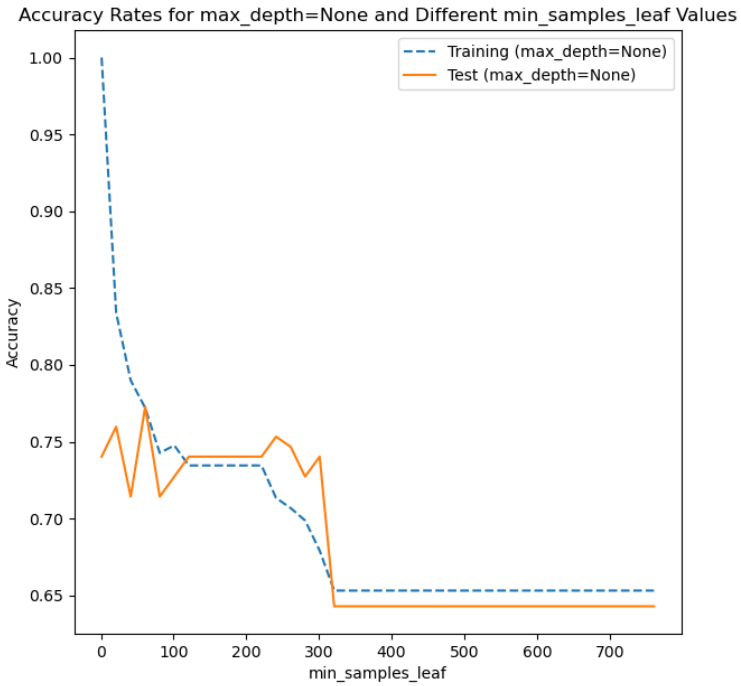
Overfitting: The accuracy of testing vs training drops just before the first 20 samples. Until this point the graph was overfit for the training data, which is why the accuracy was much higher than the testing data.

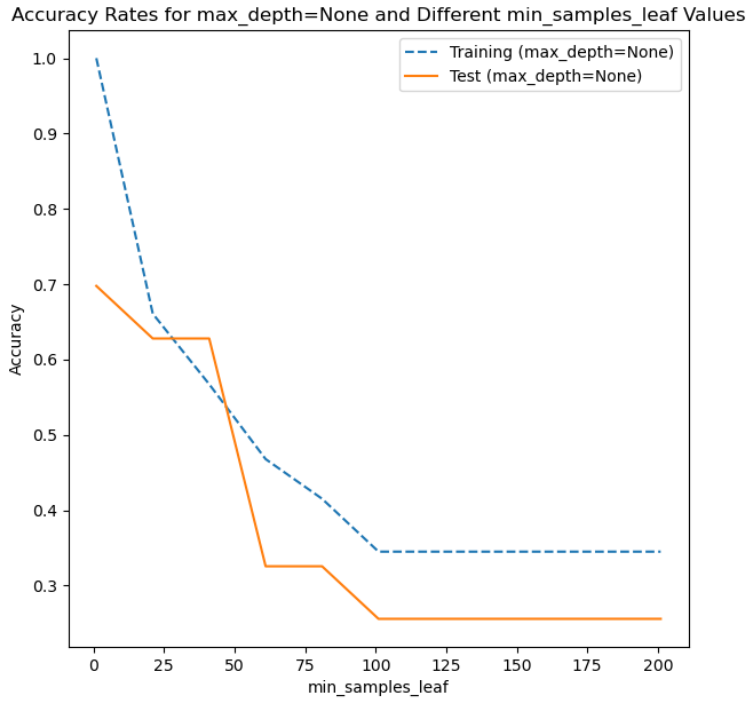
Optimality: The accuracy of testing vs training converges at 20 and until 40 samples. This is when it is optimally fitted for the data.

Underfitting :The accuracy of testing vs training is consistently lower beyond the 50 sample mark. This is when the data can be classified as underfit.

This was determined by comparing the testing and training data for the two classification problems. The plot figures help us estimate the points at which the trend changes from being overfit, optimally fit and underfit. The accuracy of the data is the point of representation for the same.

For the experiment (i), I turned up the step size from 5 to 20 and here are my results:





For the experiment (ii), I turned down the step size from 5 to 1 and here are my results: