

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | DECISION TREE ASSIGNMENT | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | Siddharth Mehrotra |  | | | |
|  | | | | SID- 55354954CS3481 Fundamentals of Data Science—Dr. WONG, Hau San Raymond |  | | | |
|  | | |  | | |  | | |

**Multiple decision trees based on different partitions of the dataset into a training set and a test set, and the structures and classification performances of these different trees.**

**Code:**

**Text

Description automatically generated**

The parameters were selected from the following code with different values:

**Graphical user interface, text, application

Description automatically generated**

**Description:**

The following trees have been generated based on different parameters such as criterion, test size, random\_state, max\_depth, min\_samples\_leaf, min\_samples\_split, max\_leaf\_nodes and the stratify function. Some of the best trees with highest accuracy will be chosen to study in more detail in the second part of this report.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Tree | | | | |
| **Impurity Measure** | **Parameters** | **Test Accuracy** | **No. of leaf nodes**  **Length of the longest path** | **Balanced**  **(T/F)** |
|  | #1 | | | | |
| Gini | test\_size=0.2, random\_state=1,  max\_depth = 2  stratify=target\_class | 0.8064516129032258  ~ 80% | 4  2 | T |
|  |  | | | | |
| Entropy | test\_size=0.2, random\_state=1,  max\_depth = 2,  stratify=target\_class | 0.7903225806451613  ~79% | 4  2 | F |
|  |  | | | | |
| Gini | test\_size = 0.22, random\_state = 27,  max\_depth = 3  min\_samples\_leaf =20  max\_leaf\_nodes=5  stratify = target\_class | 0.9130434782608695  ~91% | 5  3 | F |
|  |  | | | | |
| Entropy | test\_size = 0.22, random\_state = 27,  max\_depth = 3  min\_samples\_leaf = 20  stratify = target\_class | 0.855072463768116  ~85% | 5  3 | T |
|  |  | | | | |
| Entropy | test\_size = 0.35, random\_state = 27,  max\_depth = 3  min\_samples\_leaf = 20  stratify = target\_class | 0.8256880733944955  ~82% | 5  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.35, random\_state = 27,  max\_depth = 3  min\_samples\_leaf = 20  stratify = target\_class | 0.8165137614678899  ~81% | 5  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.3 random\_state = 164,  max\_depth = 3  min\_samples\_leaf = Default  stratify = target\_class | 0.8709677419354839  ~87% | 7  3 | T |
|  |  | | | | |
| Entropy | test\_size = 0.3, random\_state = 164,  max\_depth = 3  min\_samples\_leaf = Default  stratify = target\_class | 0.8494623655913979  ~85% | 7  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.4, random\_state = 20,  max\_depth = 3  min\_samples\_leaf = 20  stratify = target\_class | 0.8548387096774194  ~85% | 6  3 | T |
|  |  | | | | |
| Entropy | test\_size = 0.4, random\_state = 20,  max\_depth = 3  min\_samples\_leaf = 20  stratify = target\_class | 0.8256880733944955  ~82% | 5  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.2, random\_state = 14,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8225806451612904  ~82% | 7  4 | F |
|  |  | | | | |
| Entropy | test\_size = 0.2, random\_state = 14,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8387096774193549  ~84% | 7  4 | F |
|  |  | | | | |
| Gini | test\_size = 0.3, random\_state = 14,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8172043010752689  ~81% | 6  3 | T |
|  |  | | | | |
| Entropy | test\_size = 0.3, random\_state = 14,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8256880733944955  ~82% | 6  4 | F |
|  |  | | | | |
| Gini | test\_size = 0.35, random\_state = 84,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8532110091743119  ~85% | 6  3 | F |
|  |  | | | | |
| Entropy | test\_size = 0.35, random\_state = 84,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8623853211009175  ~86% | 6  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.37, random\_state = 189,  max\_depth = 4  min\_samples\_leaf = 20  stratify = target\_class | 0.8869565217391304  ~88% | 11  4 | T |
|  |  | | | | |
| Entropy | test\_size = 0.4, random\_state = 146,  max\_depth = 4  min\_samples\_leaf = 1  stratify = target\_class | 0.7983870967741935  ~79% | 10  3 | T |
|  |  | | | | |
| Gini | test\_size = 0.3, random\_state = 164,  max\_depth = 3  min\_samples\_leaf = 20  max\_leaf\_nodes=4  stratify = target\_class  splitter='best' | 0.8709677419354839  ~87% | 4  3 | F |
|  |  | | | | |
| Entropy | test\_size = 0.3, random\_state = 164,  max\_depth = 3  min\_samples\_leaf = 20  max\_leaf\_nodes=4  stratify = target\_class  splitter='best' | 0.8494623655913979  ~84% | 4  3 | F |

**Analysing selected trees for classification performance associated with the different classes and the sequence of decisions that lead to the misclassification**

Chart, scatter chart

Description automatically generated

*Plot 1. Pelvic Radius vs Grade of Spondylolisthesis*

Scatter chart

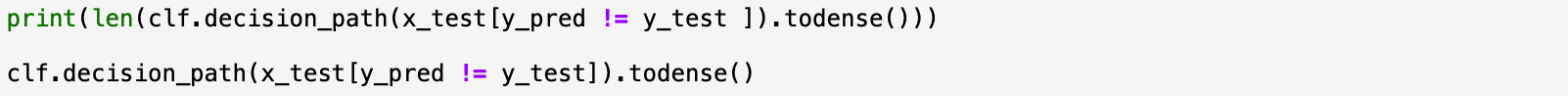
Description automatically generated

*Plot 2. Sacral Slope vs Grade of Spondylolisthesis*

**Code:**

Text, letter

Description automatically generated



1. Tree #2

Diagram

Description automatically generated

Impurity= Entropy

Test\_size= 0.2

Random\_state = 1

Max\_depth = 2

Accuracy = 79%

No. of leaf nodes = 2

Length of longest path= 2

Balanced

A picture containing chart

Description automatically generated

Class accuracy:

DH: nan

NO: 0.6060606060606061

SL: 1.0

Decision Path:

Matrix

([[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0]])

It can be observed that the accuracy of DH class is 0, NO are classified correctly 60% of the times and

SL is classified correct 100% of the time. The model testing size is 80% but the maximum depth is set to 2. Therefore, this is an example of underfitting.

Using the decision matric, it can be observed that all the values end up at node 3. This can be observed using the scatter plot 2, where the values of sacral slope are less than 46. Most of the DH values can be misclassified as falling under normal. The training size is set to 80% but the maximum depth has been limited to 2, which leads to underfitting.

1. Tree #3

Diagram

Description automatically generated

Impurity= Gini

Test\_size= 0.22

Random\_state = 27

Max\_depth = 3

**Accuracy = 91%**

Max\_leaf\_nodes = 5

No. of leaf nodes = 5

Length of longest path= 3

Unbalanced

A picture containing graphical user interface

Description automatically generated

Class accuracy:

DH: 0.9

NO: 0.8333333333333334

SL: 0.9714285714285714

Decision Path:

matrix([[1, 1, 0, 0, 1, 0, 1, 0, 0],

[1, 1, 0, 1, 0, 0, 0, 0, 1],

[1, 1, 0, 1, 0, 0, 0, 0, 1],

[1, 1, 0, 0, 1, 0, 1, 0, 0],

[1, 0, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 1, 0, 0, 0, 1, 0]])

This is the most accurate tree structure, the structure was built using random seed of 27, test size 22%, max depth 3 and with Gini impurity. Along with that, we have limited the max leaf nodes to 5 and the minimum samples to split as 20. It can be observed in scatter plot 1 that all the values with SL grade more than 16.5 are classified as SL. This observation resonates with our tree structure and the decision path matrix. Observing the matrix we can see that there are only 2 values which are misclassified as SL whose grade values are higher than 16.5. Using the class accuracy we can observe 90% of DH instances are correctly classified, 83% of NO are classified accurately and almost all the SL instances are classified correctly with the accuracy of 97%.

1. Tree #6

Diagram

Description automatically generated

Impurity= Gini

Test\_size= 0.35

Random\_state = 27

Max\_depth = 3

Accuracy = 81%

Max\_leaf\_nodes = None

No. of leaf nodes = 5

Length of longest path= 3

Unbalanced

Graphical user interface

Description automatically generated with medium confidence

Class accuracy:

DH: 0.6071428571428571

NO: 0.8636363636363636

SL: 0.8983050847457628

Decision Path:

matrix([[1, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 1, 1, 0, 0, 0, 0],

[1, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 1, 0, 1, 0, 0, 0],

[1, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 1, 0, 1, 0, 1, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 1, 0, 1, 0, 0, 0],

[1, 1, 1, 0, 0, 0, 0, 0, 0],

[1, 0, 0, 0, 0, 0, 1, 1, 0]])

It can be observed that the accuracy of DH class is 60%, NO are classified correctly 86% of the times

and SL is classified correct 89% of the time. Although the maximum depth is 3, the training size is only 65% which leads to underfitting underfitting.

Using the decision matric, it can be observed that all the 9 values have been misclassified as DH under node 3. This can be observed using the scatter plot 1, where the values of pelvic radius is less than 117 and grade of SL is less than 11. Most of the NO values can be misclassified as falling under DH. The training size is set to 65% which leads to underfitting.

1. Tree #19

Diagram

Description automatically generated

Impurity= Gini

Test\_size= 0.3

Random\_state = 164

Max\_depth = 3

**Accuracy = 87%**

Max\_leaf\_nodes = 4

Min\_samples\_leaf=4

No. of leaf nodes = 5

Length of longest path= 3

Unbalanced

A picture containing graphical user interface

Description automatically generated

Class accuracy:

DH: 0.7

NO: 0.8214285714285714

SL: 0.9777777777777777

Decision Path:

matrix([[1, 1, 0, 1, 0, 1, 0],

[1, 1, 0, 0, 1, 0, 0],

[1, 1, 0, 1, 0, 1, 0],

[1, 1, 0, 0, 1, 0, 0],

[1, 1, 0, 1, 0, 0, 1],

[1, 1, 0, 1, 0, 0, 1],

[1, 1, 0, 1, 0, 1, 0],

[1, 1, 0, 1, 0, 0, 1],

[1, 1, 0, 1, 0, 1, 0],

[1, 1, 0, 1, 0, 1, 0],

[1, 0, 1, 0, 0, 0, 0],

[1, 1, 0, 1, 0, 1, 0]])

This tree structure uses test size of 30% and a training size of 70%. The random state is set to be 164, max leaf nodes to be 4. Choosing max leaf node forces the algorithm to choose the most optimal leaf nodes, which can be seen in our tree structure.

In the above decision path matric we can observe that 6 instances has been misclassified as normal when actually they were supposed to be DH. This can be observed in scatter plot 1 where the grade of SL is less than 16 but the pelvic radius is greater than 125.

1. Tree #11

Diagram

Description automatically generated

Impurity= Gini

Test\_size= 0.2

Random\_state = 14

Max\_depth = 4

**Accuracy = 82%**

Max\_leaf\_nodes = None

Min\_samples\_leaf=20

No. of leaf nodes = 7

Length of longest path= 4

Unbalanced

Graphical user interface

Description automatically generated

Class accuracy:

DH: 0.6666666666666666

NO: 0.7368421052631579

SL: 0.9354838709677419

Decision Path

matrix([[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1],

[1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0],

[1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0],

[1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0],

[1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0],

[1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0],

[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0],

[1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0]])

It can be observed that the accuracy of DH class is 60%, NO are classified correctly 73% of the times and SL is classified correct 93% of the time. The model testing size is 80% but the maximum depth is set to 4.

Using the decision matrix, it can be observed that there’s not a clear relation between misclassified samples. They are largely distributed. The training size is set to 80% but the maximum depth has been allowed to be 4, which leads to overfitting.

**Further Analysis**

**Code:**

**Text

Description automatically generated**

**Chart, line chart

Description automatically generated**Chart, line chart

Description automatically generated

*Test size 0.2 Test Size = 0.25*

**Chart, line chart

Description automatically generated**Chart, line chart

Description automatically generated

*Test Size = 0.3 Test size = 0.35*

**Chart, line chart

Description automatically generated**Chart, line chart

Description automatically generated

*Test size = 0.4 Test size = 0.45*

**Using the above graphs there have been several patterns detected in different test size and maximum depths. It can be observed that the maximum depth of a decision tree should be limited around 3 for maximum accuracy. Maximum depth of 3 and a test size of ~75% prevents the tree from underfitting and overfitting. Limiting the maximum leaf nodes and minimum samples split further promotes accuracy.**