**Name: Aryan Girish Kasliwal**

**SID: 55972222**

**EID: akasliwal2**

**CS3481: Fundamentals of Data Science- Assignment 1**



Construct multiple decision trees based on the default training set/test set partition using different parameter settings. Compare the structures and classification performances of these decision trees. (25%)

**Decision tree - 1:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="best" |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:**

Diagram

Description automatically generated

Since the max\_depth parameter is set to 2, the classifier makes the tree reach a depth of 2 before evaluating the results. The attributes chosen are based on their entropy where the highest entropy was for b2, followed by pred\_minus\_obs\_H\_b1 then pred\_minus\_obs\_H\_b9. Based on the value of b2 being less than or equal to 35.5, the values were split in two. All forests of type ‘h’ and ‘s’ have a value of b2 less than 35.5 hence all forests of these types would lie on the left side of the decision tree.

On the left side, based on the value of pred\_minus\_obs\_H\_b1 being less than or equal to 41.39 the tree was further split into two nodes to form the leaf nodes. It was observed that a majority of forest type ‘h’ had pred\_minus\_obs\_H\_b1 less than or equal to 41.39 with just two outliers. It can also be observed that most forests of type ‘s’ formed the other lead node where the value of pred\_minus\_obs\_H\_b1 was greater than 41.39 with just one outlier.

On the right side, based on the value of pred\_minus\_obs\_H\_b9 being less than or equal to -7.525, the tree was further split into two nodes to form the leaf nodes. It was observed that all forest type ‘o’ have a value of pred\_minus\_obs\_H\_b9 less than or equal to -7.525 with zero outliers. It can also be observed that a majority of forests of type ‘d’ formed the other leaf node where the value of pred\_minus\_obs\_H\_b9 was greater than -7.525 with two outliers.

The accuracy of this decision tree was 80% when run on the test set.

**Decision tree - 2:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=3 | splitter="best" |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:**Diagram

Description automatically generated

Since the max\_depth parameter is set to 3, the classifier makes the tree reach a depth of 3 before evaluating the results. The attributes chosen are based on their entropy where the highest entropy was for b2, followed by pred\_minus\_obs\_H\_b1 then b9 then pred\_minus\_obs\_S\_b7 then pred\_minus\_obs\_H\_b7 then pred\_minus\_obs\_S\_b3. Based on the value of b2 being less than or equal to 35.5, the values were split in two. All forests of type ‘h’ and ‘s’ have a value of b2 less than 35.5 hence all forests of these types would lie on the left side of the decision tree whereas ‘o’ and ‘d’ would lie on the right side of this node.

On the left side, based on the value of pred\_minus\_obs\_H\_b1 being less than or equal to 41.39 the tree was further split into two nodes. It was observed that a majority of forest type ‘h’ had pred\_minus\_obs\_H\_b1 less than or equal to 41.39 with just two outliers. It can also be observed that most forests of type ‘s’ have a value of pred\_minus\_obs\_H\_b1 greater than 41.39 with just one outlier. On the side that has pred\_minus\_obs\_H\_b1 less than or equal to 41.39 (left side), the forests were further classified based on pred\_minus\_obs\_S\_b7 being less than or equal to 0.36. It is observed that all forests of type ‘h’ have a pred\_minus\_obs\_S\_b7 value less than 0.36 hence forming the leaf node for forest type ‘h’ whereas the outlying forest class ‘s’ has this value greater than 0.39 hence forming the leaf node ‘s’. On the side that has pred\_minus\_obs\_H\_b1 greater than 41.39 (right side), the forests were further classified based on pred\_minus\_obs\_H\_b7 being less than or equal to -17.815. It is observed that all forests of type ‘h’ have pred\_minus\_obs\_H\_b7 value less than -17.815 hence forming the leaf node for forest type ‘h’ whereas all forests of type ‘s’ have this value greater than -17.815 with just one outlying forest type ‘d’ forming the leaf node for forest type ‘s’.

On the right side, based on the value of b9 being less than or equal to 62.5, the tree was further split into two nodes to form the leaf nodes. It was observed that majority forest types ‘d’ have a value of b9 less than or equal to 62.5 with two outliers hence forming a leaf node for forest type ‘d’. It can also be observed that most forests of type ‘o’ have a value of b9 greater than 62.5 with one outlying type ‘d’ also having this value greater than 62.5. This was then classified based on the attribute pred\_minus\_obs\_S\_b3 being less than or equal to -1.83. It is observed that all forest types ‘o’ have this value less than -1.83 hence forming the leaf node for type ‘o’. Whereas the outlying type ‘d’ has this value greater than -1.83 hence forming the leaf node for the outlying ‘d’ node.

The accuracy of this decision tree was 78.1538 when run on the test set.

**Decision tree - 3:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="random " |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:**

**Diagram

Description automatically generated**

The depth of this decision tree was set to 2, but the splitter parameter was set to random. This made the tree split on the best random value on each node instead of the best value. The root node is pred\_minus\_obs\_H\_b5 being less than or equal to -41.662. It was observed a majority of class ‘o’ have this value less than the threshold with a few ‘d’ types as well. Whereas a majority of type ‘s’ have this value greater than -41.662 with a few type ‘d’ and a lot of outliers from classes ‘d’ and ‘o’.

On the left side, the node is split based on the value of b9 being less than or equal to 69.7. It is observed that 12 type ‘d’ forests have this value less than 69.7 along with 5 outlier from class ‘o’ hence forming the leaf node for ‘d’ whereas 26 type ‘o’ forests have this value greater than 69.7 hence forming leaf node for ‘o’.

On the right side, the node is split based on the value of b5 being less than or equal to 53.471. It is observed that 52 type ‘s’ forests have this value less than or equal to 53.471 along with 40 type ‘h’ and 1 each of ‘o’ and ‘d’ hence forming the leaf node for ‘s’ whereas 41 type ‘d’ along with ‘8’ type ‘h’, 5 type ‘o’ and 7 type ‘s’ forests hence forming the leaf node for type ‘d’.

It is noteworthy here that this tree does not have any leaf nodes for forest type ‘h’, which means it is incapable to classifying any forest as type ‘h’. This hinders its accuracy by a lot with a value of only 69.8461%.

**Decision tree - 4:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="best " |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=3 | class\_weight=None | ccp\_alpha=0 |

**Tree:**

A diagram of a flowchart

Description automatically generated with low confidence

The depth of this tree was set to 2, the splitter parameter was set to “best” hence choosing the best possible value to split at each node but the maximum leaf nodes was set to 3 only. This makes this tree interesting as it has to classify 4 types of forests but is only allowed to have 3 leaf nodes, hence having to miss one type of tree for sure.

The root node is b2 being less than or equal to 35.5. 53 tree types ‘d’ have this value greater, hence forming the leaf node on the right with 37 outliers from forest class ‘o’ which is never classified by this tree.

On the left side, the attribute pred\_minus\_obs\_H\_b1being less than or equal to 41.39 is the deciding factor to classify the forests into the two leaf nodes. Majority of type ‘h’ forests have this value less than 41.39 with just one outlying value of type ‘s’ hence forming the leaf node ‘h’. Whereas most type ‘s’ forests have this value greater than 41.39 with two outlying values from type ‘h’ and one outlying value from type ‘d’ hence forming the leaf node for type ‘s’.

The accuracy of this decision tree was 72.9230 when run on the test set.

**Tree comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Tree 1** | **Tree 2** | **Tree 3** | **Tree 4** |
| **Accuracy** | 80 | 78.1538 | 69.8461 | 72.9230 |
| **No. Leaf nodes** | 4 | 7 | 4 | 3 |
| **Depth** | 2 | 3 | 2 | 2 |
| **Splitter** | Best | Best | Random | Best |



Exchange the training and test set and repeat the tasks in (a). (25%)

**Decision tree - 1:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="best" |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:** Diagram

Description automatically generated

Since the max\_depth parameter is set to 2, the classifier makes the tree reach a depth of 2 before evaluating the results. The attributes chosen are based on their entropy where the highest entropy was for b2, followed by b1 then pred\_minus\_obs\_H\_b8. Based on the value of b2 being less than or equal to 34.5, the values were split in two. All forests of type ‘h’ and ‘s’ have a value of b2 less than 34.5 hence all forests of these types would lie on the left side of the decision tree.

On the left side, based on the value of b1 being less than or equal to 69.5 the tree was further split into two nodes to form the leaf nodes. It was observed that a majority of forest type ‘s’ had b1 less than or equal to 69.5 with 15 outliers from other types of leaf nodes. It can also be observed that most forests of type ‘h’ formed the other lead node where the value of b1 was greater than 69.5 with 5 outliers.

On the right side, based on the value of pred\_minus\_obs\_H\_b8 being less than or equal to -2.51, the tree was further split into two nodes to form the leaf nodes. It was observed that all forest type ‘o’ have a value of pred\_minus\_obs\_H\_b8 less than or equal to -2.51 with11 outliers. It can also be observed that a majority of forests of type ‘d’ formed the other leaf node where the value of pred\_minus\_obs\_H\_b8 was greater than -2.51 with 21 outliers.

The accuracy of this decision tree was 92.42% when run on the training set.

**Decision tree - 2:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=3 | splitter="best" |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:**

**Diagram

Description automatically generated with medium confidence**

Since the max\_depth parameter is set to 3, the classifier makes the tree reach a depth of 3 before evaluating the results. The attributes chosen are based on their entropy where the highest entropy was for b2, followed by b1 then pred\_minus\_obs\_H\_b8 then pred\_minus\_obs\_H\_b5 then b8 then b4 and then b1. Based on the value of b2 being less than or equal to 34.5, the values were split in two. All forests of type ‘h’ and ‘s’ have a value of b2 less than 34.5 hence all forests of these types would lie on the left side of the decision tree whereas ‘o’ and ‘d’ would lie on the right side of this node.

On the left side, based on the value of b1 being less than or equal to 69.5 the tree was further split into two nodes. It was observed that a majority of forest type ‘s’ had b1 less than or equal to 69.5. It can also be observed that most forests of type ‘h’ have a value of b1 greater than 69.5. On the side that has b1 less than or equal to 69.5 (left side), the forests were further classified based on b1 being less than or equal to 42.5. It is observed that majority forests that have b1 value less than 42.6 are of type ‘d’ hence forming the leaf node for forest type ‘d’ whereas the other forest class ‘s’ has this value greater than 42.5 hence forming the leaf node ‘s’. On the side that has b1 greater than 69.5 (right side), the forests were further classified based on b4 being less than or equal to 112.5. It is observed that all forests of type ‘h’ have b4 value less than 112.5 hence forming the leaf node for forest type ‘h’ with 9 outliers whereas all forests of type ‘h’ have this value greater than 112.5 with 0 outliers hence forming the leaf node for forest type ‘h’.

On the right side, based on the value of pred\_minus\_obs\_H\_b8 being less than or equal to -2.51, the tree was further split into two nodes. It was observed that majority forest types ‘o’ have this value less than or equal to -2.51. It can also be observed that most forests of type ‘d’ have a value of b9 greater than -2.51. The left side was then classified based on the attribute b8 being less than or equal to 37.5. It is observed that all forest types ‘o’ have this value less than 37.5s hence forming the leaf nodes for type ‘o’. On the right hand side the tree was classified based on the value of pred\_minus\_obs\_H\_b5 being less than or equal to -27.335. It was observed that majority of type ‘d’ forests had this value less than or equal to -27.335 whereas majority of forests that had this value greater than the threshold were of type ‘s’.

The accuracy of this decision tree was 92.42% when run on the training set.

**Decision tree - 3:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="random " |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=None | class\_weight=None | ccp\_alpha=0 |

**Tree:**

**Diagram

Description automatically generated**

The depth of this decision tree was set to 2, but the splitter parameter was set to random. This made the tree split on the best random value on each node instead of the best value. The root node is pred\_minus\_obs\_H\_b5 being less than or equal to -38.351. It was observed a majority of class ‘o’ have this value less than the threshold with a few ‘d’ types as well. Whereas a majority of type ‘s’ have this value greater than the threshold with a few type ‘d’ and a lot of outliers from classes ‘d’ and ‘o’.

On the left side, the node is split based on the value of pred\_minus\_obs\_H\_b6 being less than or equal to -53.116. It is observed that 31 type ‘o’ forests have this value less than -53.116 along with 4 outlier from class ‘d’ hence forming the leaf node for ‘o’ whereas 34 type ‘d’ forests have this value greater than 69.7 hence forming leaf node for ‘o’ along with 8 outlier from class ‘o’.

On the right side, the node is split based on the value of b5 being less than or equal to 55.172. It is observed that 121 type ‘s’ forests have this value less than or equal to the threshold along with 35 type ‘h’ and 1 of type ‘o’ and 20 of type ‘d’ hence forming the leaf node for ‘s’ whereas 47 type ‘d’ along with 3 type ‘h’, 6 type ‘o’ and 15 type ‘s’ forests hence forming the leaf node for type ‘d’.

It is noteworthy here that this tree does not have any leaf nodes for forest type ‘h’, which means it is incapable to classifying any forest as type ‘h’. This hinders its accuracy by a lot with a value of only 65.1515 %.

**Decision tree - 4:**

**Parameters:**

|  |  |  |
| --- | --- | --- |
| criterion='entropy' | max\_depth=2 | splitter="best " |
| min\_samples\_split=2 | min\_samples\_leaf=1 | min\_weight\_fraction\_leaf=0 |
| min\_impurity\_decrease=0 | max\_features=None | random\_state=None |
| max\_leaf\_nodes=3 | class\_weight=None | ccp\_alpha=0 |

**Tree:**

A picture containing text, businesscard

Description automatically generated

The depth of this tree was set to 2, the splitter parameter was set to “best” hence choosing the best possible value to split at each node but the maximum leaf nodes was set to 3 only. This makes this tree interesting as it has to classify 4 types of forests but is only allowed to have 3 leaf nodes, hence having to miss one type of tree for sure.

The root node is b2 being less than or equal to 34.5. 87 forest types ‘d’ have this value greater, hence forming the leaf node on the right with 44 outliers from forest class ‘o’ which is never classified by this tree and 6 outliers from class ‘s’.

On the left side, the attribute b1being less than or equal to 69.5 is the deciding factor to classify the forests into the two leaf nodes. Majority of type ‘s’ forests have this value less than 69.5 hence forming the leaf node ‘s’. Whereas most type ‘h’ forests have this value greater than 69.5 hence forming the leaf node for type ‘s’.

The accuracy of this decision was 75.7575% when run on the training set.



For selected trees in (a) and (b), observe the classification performance associated

with the different classes, and determine which pair(s) of classes are likely to be

confused with each other. (25%)

1. This part has the confusion matrices when the training set is used to train and testing set is used to test.

**Confusion matrix for tree 1:**

Chart

Description automatically generated

We can infer from the matrix which forest types can be confused with other forests respectively. For example, in row 1, forest ‘d’ matches with forest ‘d’ 52 times and matches with forest ‘h’ 0 times. This means the result of the decision trees prediction of forest ‘d’ matches with the correct answer ‘d’ 52 times and matches with correct answer ‘h’ 0 times. In other terms, ‘d’ is confused as ‘h’ 0 times.

For the above matrix, we can infer that when the true value is ‘d’, it is confused most to be ‘s’, 20 times.

**Confusion matrix for tree 2:**

Chart

Description automatically generated

Based on the above confusion matrix we can infer that just like the first tree, when the true value is ‘d’ its confused with (predicted to be) ‘s’ 20 times.

**Confusion matrix for tree 3:**

A picture containing chart

Description automatically generated

Based on the above confusion matrix we can infer that just like the first tree, when the true value is ‘d’ its confused with (predicted to be) ‘s’ 58 times.

**Confusion matrix for tree 4:**

A picture containing chart

Description automatically generated

Based on the matrix above we can infer that when the true value is ‘o’ it is confused to be value ‘d’ 44 times. This can be understood from the tree clearly because the tree does not have a leaf node ‘o’ hence all ‘o’ values are classified to be something else.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | D | H | O | S |
| D | 270 | 6 | 26 | 118 |
| H | 0 | 117 | 0 | 35 |
| O | 72 | 2 | 99 | 11 |
| S | 15 | 66 | 0 | 463 |

Overall it can be said that when the true value is ‘d’ it is most confused to be ‘s’ – 118 times.

1. This part has the confusion matrices when the testing set is used to train and training set is used to test.

**Confusion matrix for tree 1:**

Chart

Description automatically generated

In this matrix, the highest confusion is when the true value is ‘h’ and the decision tree predicts ‘s’.

**Confusion matrix for tree 2:**

Chart

Description automatically generated

Based on the above confusion matrix we can infer that just like the first tree, when the true value is ‘h’ its confused with (predicted to be) ‘s’ 6 times.

**Confusion matrix for tree 3:**

Chart

Description automatically generated with medium confidence

Based on the above confusion matrix we can infer that just like the first tree, when the true value is ‘h’ its confused with (predicted to be) ‘s’ 43 times.

**Confusion matrix for tree 4:**

A picture containing chart

Description automatically generated

Based on the matrix above we can infer that when the true value is ‘o’ it is confused to be value ‘d’ 37 times. This can be understood from the tree clearly because the tree does not have a leaf node ‘o’ hence all ‘o’ values are classified to be something else.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | D | H | O | S |
| D | 211 | 0 | 3 | 2 |
| H | 8 | 123 | 0 | 61 |
| O | 45 | 0 | 103 | 0 |
| S | 7 | 6 | 1 | 222 |

Overall it can be said that when the true value is ‘d’ it is most confused to be ‘s’ – 61 times.



For selected confused class pairs in (c), identify the corresponding leaf node(s) and

analyze the sequence of decisions that lead to the misclassification. (25%)

1. This part has the confusion matrices when the training set is used to train and testing set is used to test.

**Tree 1:**

Diagram

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘d’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there is one class ‘d’ value in the leaf node that classifies as ‘s’ and 0 ‘s’ nodes when the tree classifies ‘s’ as ‘d’. On running this training result on the test set, we get 20 wrong attributes.

**Tree 2:**

Diagram

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘d’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there is one class ‘d’ value in the leaf node that classifies as ‘s’. On running this training result on the test set, we get 20 wrong attributes.

**Tree 3:**

Diagram

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘d’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there are 7 class ‘s’ values in the leaf node that classifies as ‘d’. Likewise there is 1 class ‘d’ lassified as ‘s’. On running this training result on the test set, we get 58 wrong attributes.

**Tree 4:**

A diagram of a flowchart

Description automatically generated with low confidence

In this tree, we know from part (c) that the most confused type is true value ‘o’ with predicted value ‘d’. We can see from the arrows drawn above where the confusion arises. We can see that there are 37 class ‘o’ values in the leaf node that classifies as ‘d’. On running this training result on the test set, we get 44 wrong attributes.

1. This part has the confusion matrices when the testing set is used to train and training set is used to test.

**Tree 1:**

Diagram

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘h’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there are 5 class ‘h’ values in the leaf node that classifies as ‘s’. Likewise, there are 9 class ‘s’ values in the leaf node that classifies as ‘h’. On running this training result on the test set, we get 6 wrong attributes.

**Tree 2:**

**Diagram

Description automatically generated with medium confidence**

In this tree, we know from part (c) that the most confused type is true value ‘h’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there are 4 class ‘h’ values in the leaf node that classifies as ‘s’. Likewise, there are 9 ‘s’ values classified as ‘h’. On running this training result on the test set, we get 6 wrong attributes.

**Tree 3:**

Diagram

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘h’ with predicted value ‘s’. We can see from the arrows drawn above where the confusion arises. We can see that there are 35 class ‘h’ values in the leaf node that classifies as ‘s’. On running this training result on the test set, we get 43 wrong.

**Tree 4:**

A picture containing text, businesscard

Description automatically generated

In this tree, we know from part (c) that the most confused type is true value ‘o’ with predicted value ‘d’. We can see from the arrows drawn above where the confusion arises. We can see that there are 44 class ‘o’ values in the leaf node that classifies as ‘d’. On running this training result on the test set, we get 37 wrong.