ASSIGNMENT 1

1. (30) Take Data2 and split it into randomly selected 210 training instances and remaining 100 as test instance. Create decision trees using the training set and the “minimum records per leaf node” values of 3, 8, 12, 30, and 50.

a. Show the trees for all the five cases of min record values. Comment on what you see in a comparative analysis of the five trees. Just reporting the numbers is not enough; you must try to give an explanation of the changes observed. Which of these five trees would you prefer to use and why?

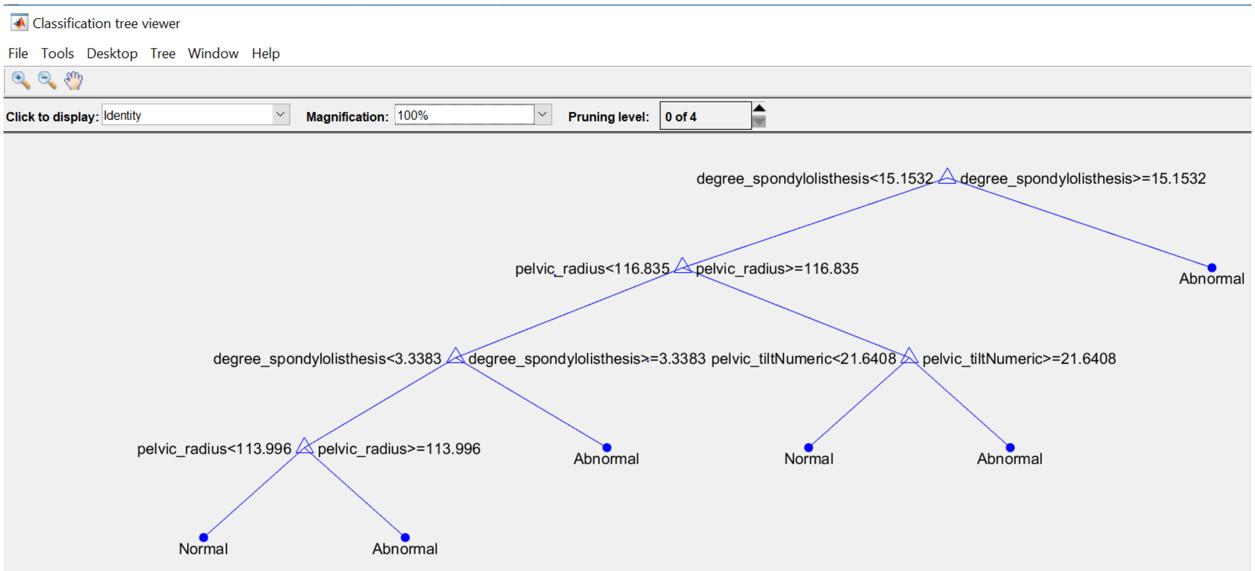
b. For each of the five decision trees compute and report the accuracy, precision, and recall values. Comment on the comparison of these values and show these values on a plot. Give your reasons for the observed trends/differences.

SOLUTION:

Matlab Code for Plotting Decision Trees:

```
DataTable=readtable('Biomechanical_Data_column_2C_weka.csv');

DataTableTrain=DataTable(51:260,:);

DataTableTest=DataTable([1:50 261:310],:);


Tree3=fitctree(DataTableTrain,'class','MinLeafSize',3);

view(Tree3,'Mode','graph')


Tree8=fitctree(DataTableTrain,'class','MinLeafSize',8);

view(Tree8,'Mode','graph')


Tree12=fitctree(DataTableTrain,'class','MinLeafSize',12);

view(Tree12,'Mode','graph')


Tree30=fitctree(DataTableTrain,'class','MinLeafSize',30);

view(Tree30,'Mode','graph')


Tree50=fitctree(DataTableTrain,'class','MinLeafSize',50);

view(Tree50,'Mode','graph')
```

DECISION TREE FOR MIN RECORD VALUE=3:


Classification tree viewer

Click to display: Identity

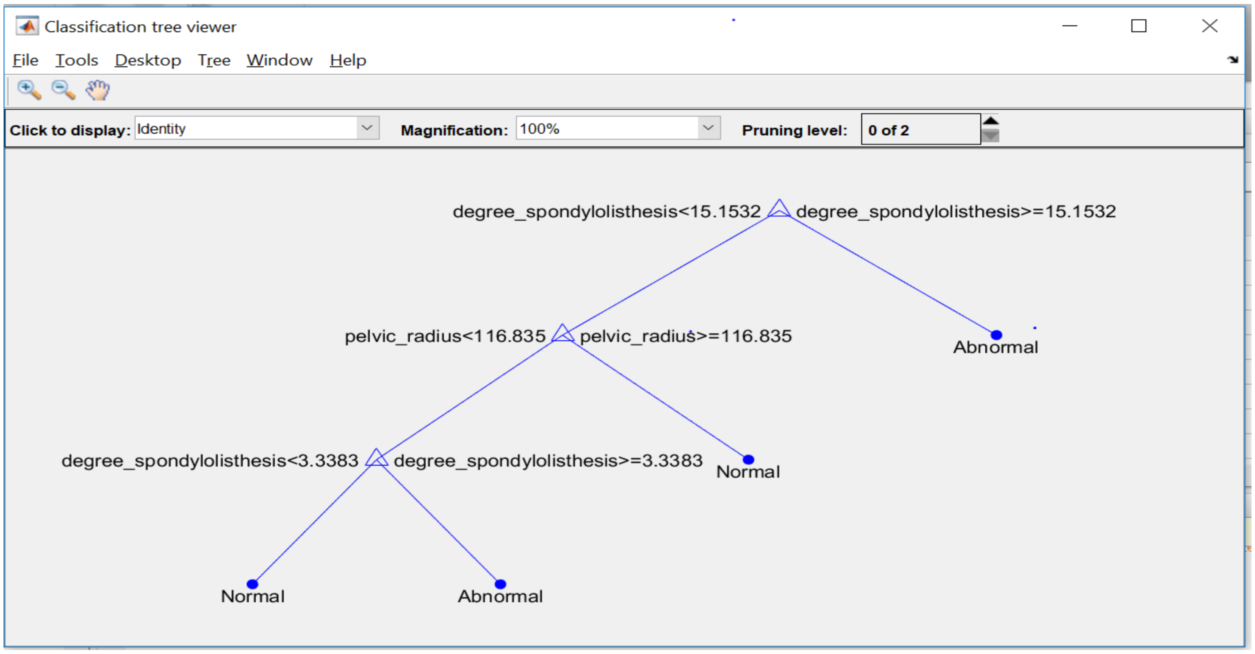
Magnification: 100%

Pruning level: 0 of 4

File
Tools
Desktop
Tree
Window
Help
degree_spondylolisthesis<15.1532
degree_spondylolisthesis>=15.1532
pelvic_radius<116.835
pelvic_radius>=116.835
Abnormal
degree_spondylolisthesis<3.3383
degree_spondylolisthesis>=3.3383
pelvic_tiltNumeric<21.6408
pelvic_tiltNumeric>=21.6408
pelvic_radius<113.996
pelvic_radius>=113.996
Abnormal
Normal
Abnormal
Normal
Abnormal

As the min Record Value=3, the decision tree takes into consideration all the conditions until the next split min Record value becomes lesser than 3. Hence the tree looks big.

DECISION TREE FOR MIN RECORD VALUE=8:


Classification tree viewer

Click to display: Identity

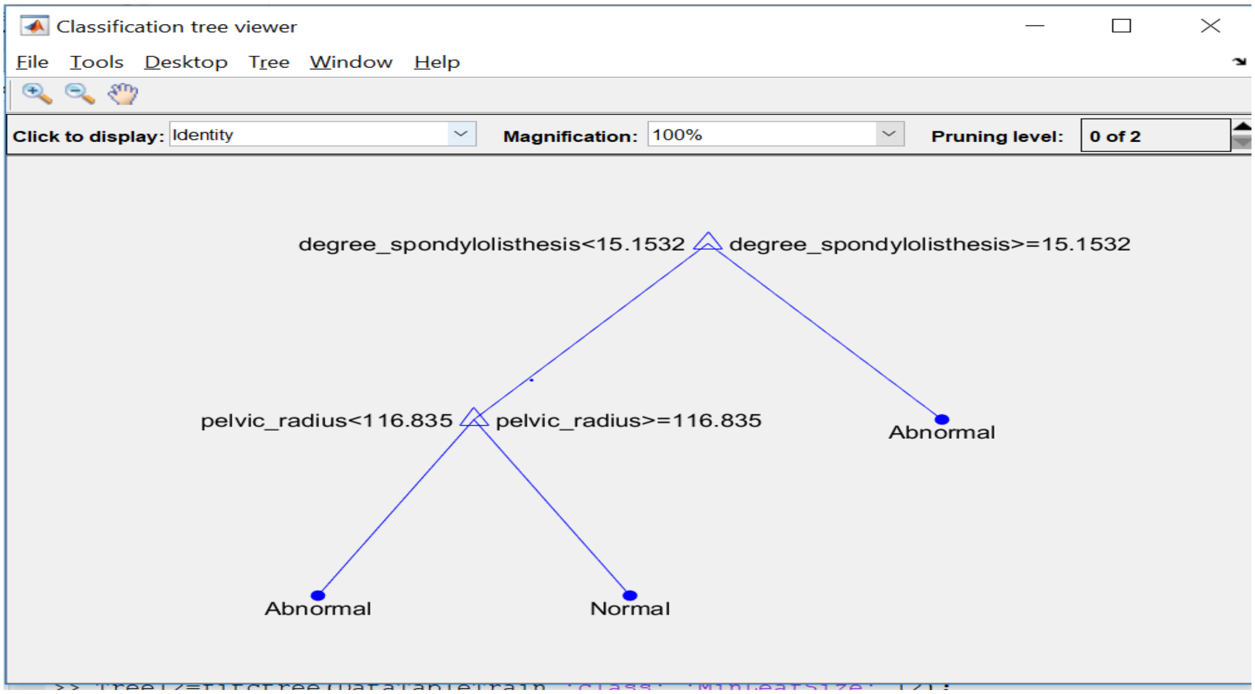
Magnification: 100%

Pruning level: 0 of 2

File
Tools
Desktop
Tree
Window
Help
degree_spondylolisthesis<15.1532
degree_spondylolisthesis>=15.1532
pelvic_radius<116.835
pelvic_radius>=116.835
Abnormal
degree_spondylolisthesis<3.3383
degree_spondylolisthesis>=3.3383
Normal
Normal
Abnormal

As the min record value has been increased to 8, the tree size reduced by a couple of branches. The Leaf node size lesser than 8 has diminished.

DECISION TREE FOR MIN RECORD VALUE=12:



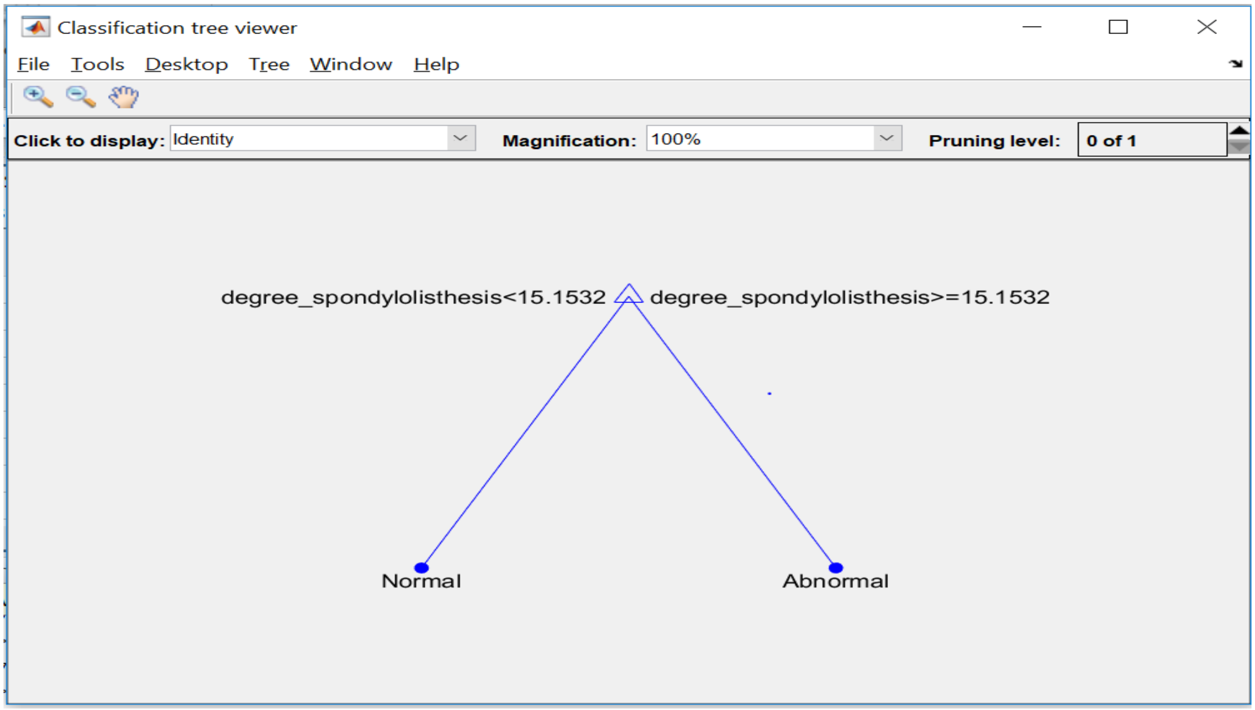
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 2

On Increasing the min leaf node value, the Decision Tree Size has decreased more.

DECISION TREE FOR MIN RECORD VALUE=30:


Classification tree viewer
File
Tools
Desktop
Tree
Window
Help
degree_spondylolisthesis<15.1532
degree_spondylolisthesis>=15.1532
Normal
Abnormal

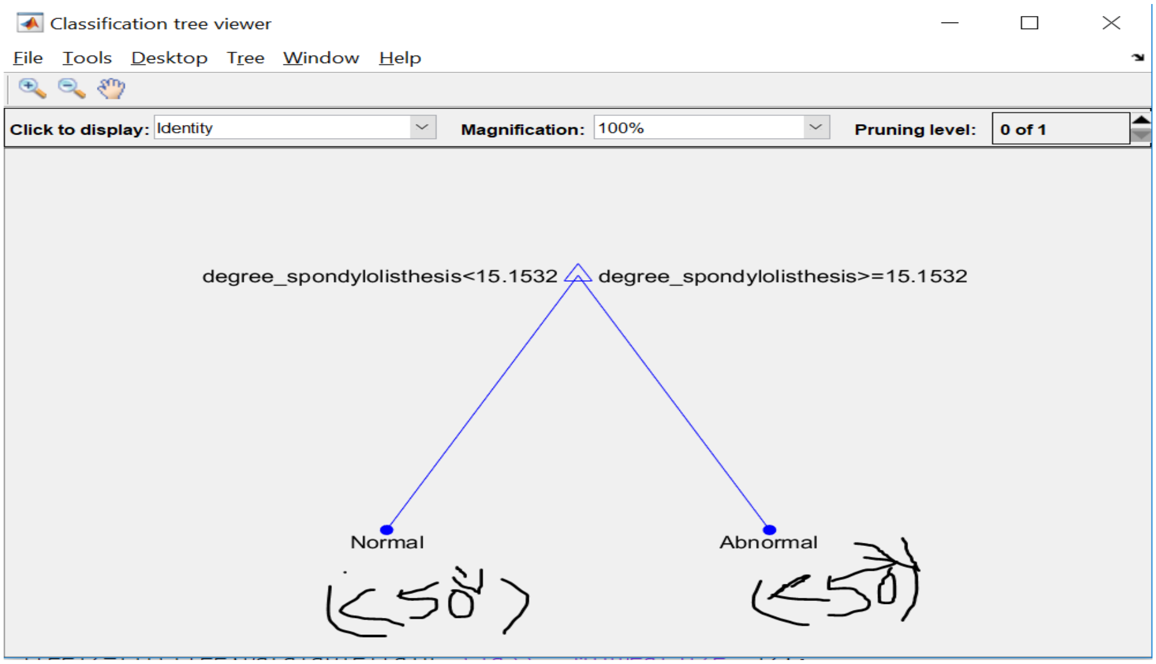
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 1

When the min Leaf Node Size Variable is Increased to 30, the Tree Reduces to a single condition Tree.

DECISION TREE FOR MIN RECORD VALUE=50:


Classification tree viewer
File Tools Desktop Tree Window Help
degree_spondylolisthesis<15.1532
degree_spondylolisthesis>=15.1532
Normal
Abnormal
(<50)
(<50)

Click to display: Identity Magnification: 100% Pruning level: 0 of 1

This Tree Looks similar compared to the Tree with min Leaf Node size equals 30.

COMPARISON OF FIVE DECISION TREES:

From the above 5 Trees, as the Min Leaf Node Size Increases, the Tree Size decreases because of the limit in splitting leaf nodes are increasing.

COMMENTS ON TREE PREFERENCE AND EXPLANATION:

Based on the Above Plotting of the 5 Trees with the Min Record Values 3,8,12,30,50, the Trees have been classified into 6,4,3,2,2 Number of Final Child Nodes respectively. On comparing the Number of Nodes following observations can be Made:

1.The Trees with min. Leaf Sizes 30 and 50 are not preferred since the No. of Nodes is only 2 and by the concept of Underfitting comes into Account, the Efficiency of the Tree is very less.

2. The Tree with min Leaf Size of 3 is not preferred since the No. of Nodes is 6 and by the concept of Overfitting, Noise may get interfered, so the Efficiency would be less

3.Now the Trees with min Leaf Sizes of 8 and 12 are preferred. But on assuming the generalization Errors of the Trees as same, By Occam's Razor Principle Tree with Node size 12 is considered as the Best. Further Analysis could be made using Accuracy, Precision and Recall Values in the Next Pages.

1.b. Matlab Code for Calculation of Accuracy, Precision and Recall Values:

```
prediction=predict(Tree3,DataTableTest);
DataTableTestcell=table2cell(DataTableTest);
cell=DataTableTestcell(:,[7]);
C=confusionmat(cell,prediction);

prediction8=predict(Tree8,DataTableTest);
C8=confusionmat(cell,prediction8);
prediction12=predict(Tree12,DataTableTest);
C12=confusionmat(cell,prediction12);
prediction30=predict(Tree30,DataTableTest);
C30=confusionmat(cell,prediction30);
prediction50=predict(Tree50,DataTableTest);
C50=confusionmat(cell,prediction50);

Accuracyc3=(C(1,1)+C(2,2))/(C(1,1)+C(1,2)+C(2,1)+C(2,2));
Accuracyc8=(C8(1,1)+C8(2,2))/(C8(1,1)+C8(1,2)+C8(2,1)+C8(2,2));
Accuracyc12=(C12(1,1)+C12(2,2))/(C12(1,1)+C12(1,2)+C12(2,1)+C12(2,2));
Accuracyc30=(C30(1,1)+C30(2,2))/(C30(1,1)+C30(1,2)+C30(2,1)+C30(2,2));
Accuracyc50=(C50(1,1)+C50(2,2))/(C50(1,1)+C50(1,2)+C50(2,1)+C50(2,2));
Precision3=(C(1,1))/(C(1,1)+C(2,1));
Precision8=(C8(1,1))/(C8(1,1)+C8(2,1));
Precision12=(C12(1,1))/(C12(1,1)+C12(2,1));
Precision30=(C30(1,1))/(C30(1,1)+C30(2,1));
Precision50=(C50(1,1))/(C50(1,1)+C50(2,1));
Precision3normal=(C(2,2))/(C(2,2)+C(1,2));
Precision8normal=(C8(2,2))/(C8(2,2)+C8(1,2));
Precision12normal=(C12(2,2))/(C12(2,2)+C12(1,2));
Precision30normal=(C30(2,2))/(C30(2,2)+C30(1,2));
Precision50normal=(C50(2,2))/(C50(2,2)+C50(1,2));
Recall3=(C(1,1))/(C(1,1)+C(1,2));
Recall8=(C8(1,1))/(C8(1,1)+C8(1,2));
Recall12=(C12(1,1))/(C12(1,1)+C12(1,2));
Recall30=(C30(1,1))/(C30(1,1)+C30(1,2));
Recall50=(C50(1,1))/(C50(1,1)+C50(1,2));
Recall3normal=(C(2,2))/(C(2,2)+C(2,1));
Recall8normal=(C8(2,2))/(C8(2,2)+C8(2,1));
Recall12normal=(C12(2,2))/(C12(2,2)+C12(2,1));
Recall30normal=(C30(2,2))/(C30(2,2)+C30(2,1));
Recall5normal0=(C50(2,2))/(C50(2,2)+C50(2,2));

AccuracyPlot=plot([3,8,12,30,50],[Accuracyc3 Accuracyc8 Accuracyc12 Accuracyc30 Accuracyc50])
PrecisionPlot=plot([3,8,12,30,50],[Precision3 Precision8 Precision12 Precision30 Precision50])
RecallPlot=plot([3,8,12,30,50],[Recall3 Recall8 Recall12 Recall30 Recall50])
PrecisionnormalPlot=plot([3,8,12,30,50],[Precision3normal Precision8normal Precision12normal
Precision30normal Precision50normal])
```

RecallnormalPlot=plot([3,8,12,30,50],[Recall3normal Recall8normal Recall12normal Recall30normal Recall5normal0])

ACCURACY VALUES FOR FIVE DECISION TREES:

1.Accuracy for Decision Tree with min Leaf Node Size=3 is 0.6100

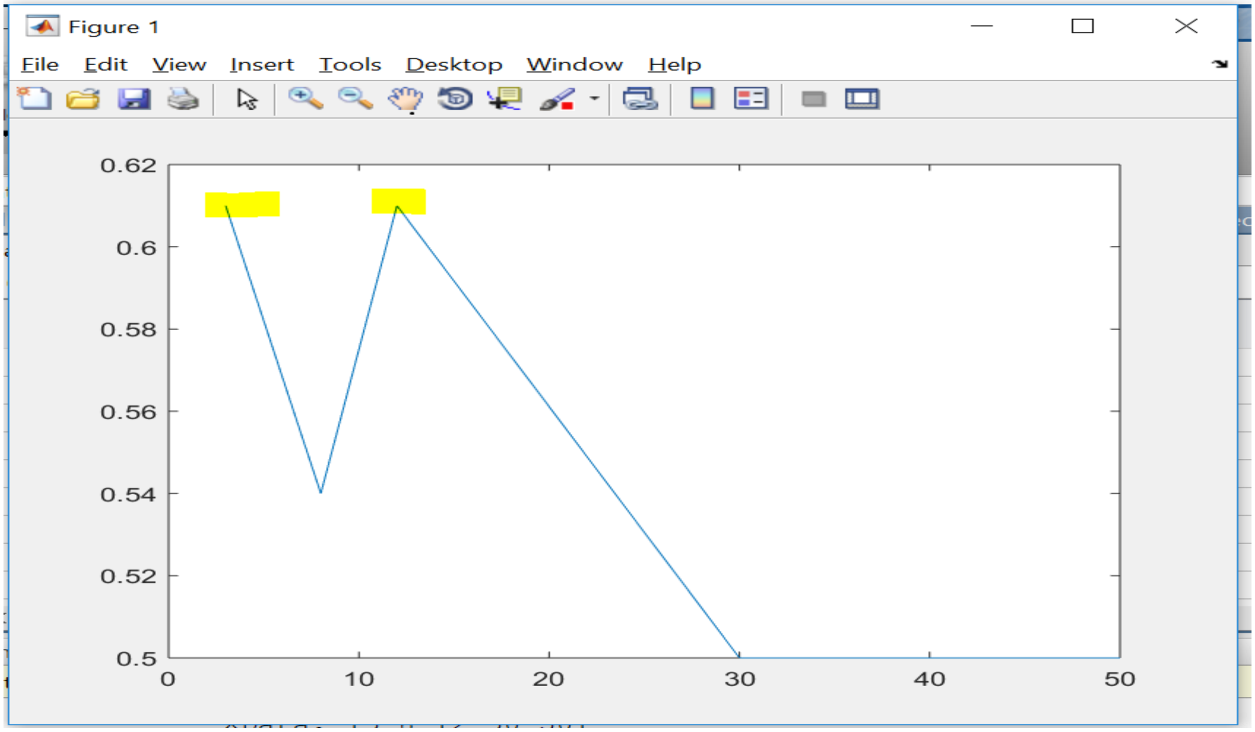
2. Accuracy for Decision Tree with min Leaf Node Size=8 is 0.5400

3. Accuracy for Decision Tree with min Leaf Node Size=12 is 0.6100

4. Accuracy for Decision Tree with min Leaf Node Size=30 is 0.5000

5. Accuracy for Decision Tree with min Leaf Node Size=50 is 0.5000

ACCURACY PLOT:



From Accuracy Plot, Plots with Leaf Node Values 3,12 are having the Highest Accuracy. Hence these 2 are Accurate Trees. We will analyze more through Precision and Recall Values as well.

PRECISION VALUES FOR FIVE DECISION TREES(ABNORMAL CLASS):

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.6774

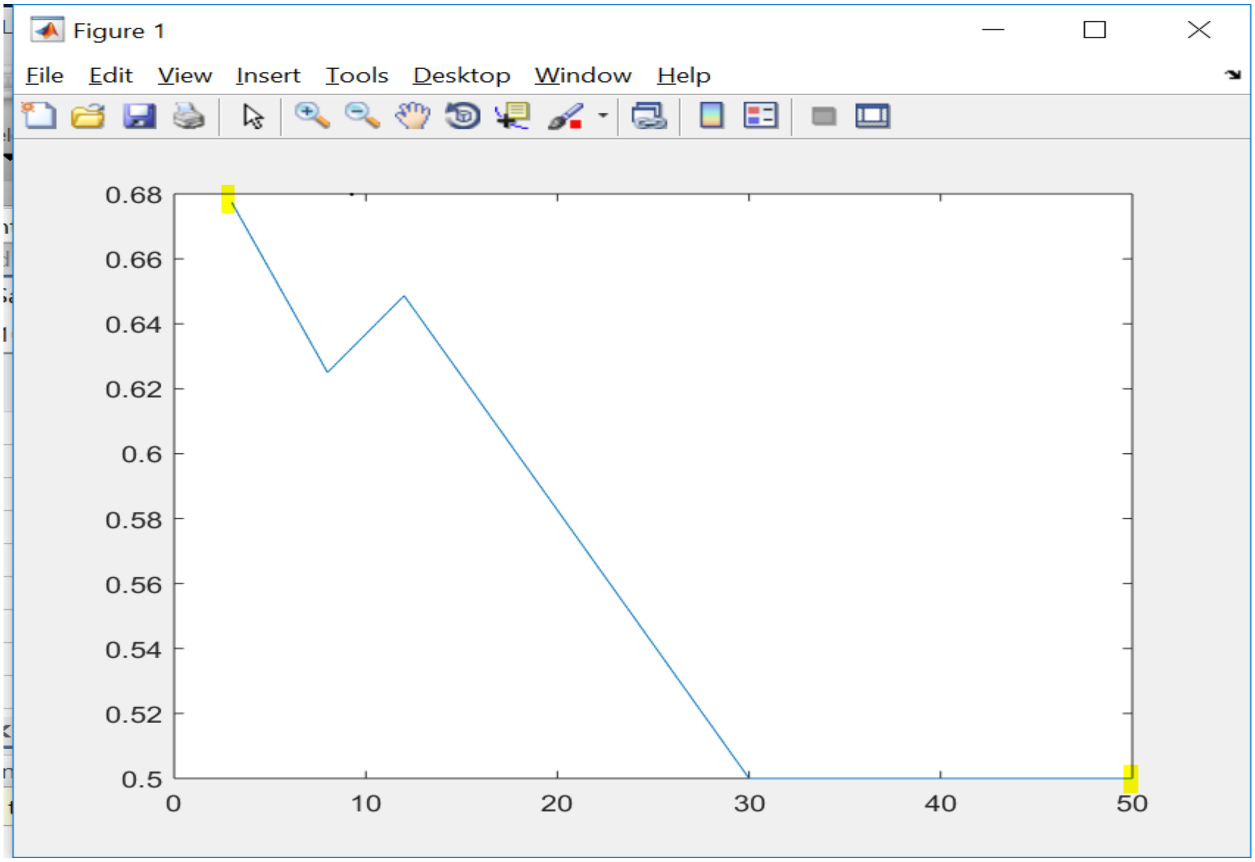
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.6250

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.6486

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.5000

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.5000

PRECISION PLOT:



The Number of True Positives Predicted per Total Predicted Positives is Precision which is Highest for Decision Tree with Min Leaf Node size 3 and Least for 50. Let's Check the Recall Values Now.

PRECISION VALUES FOR FIVE DECISION TREES(NORMAL CLASS):

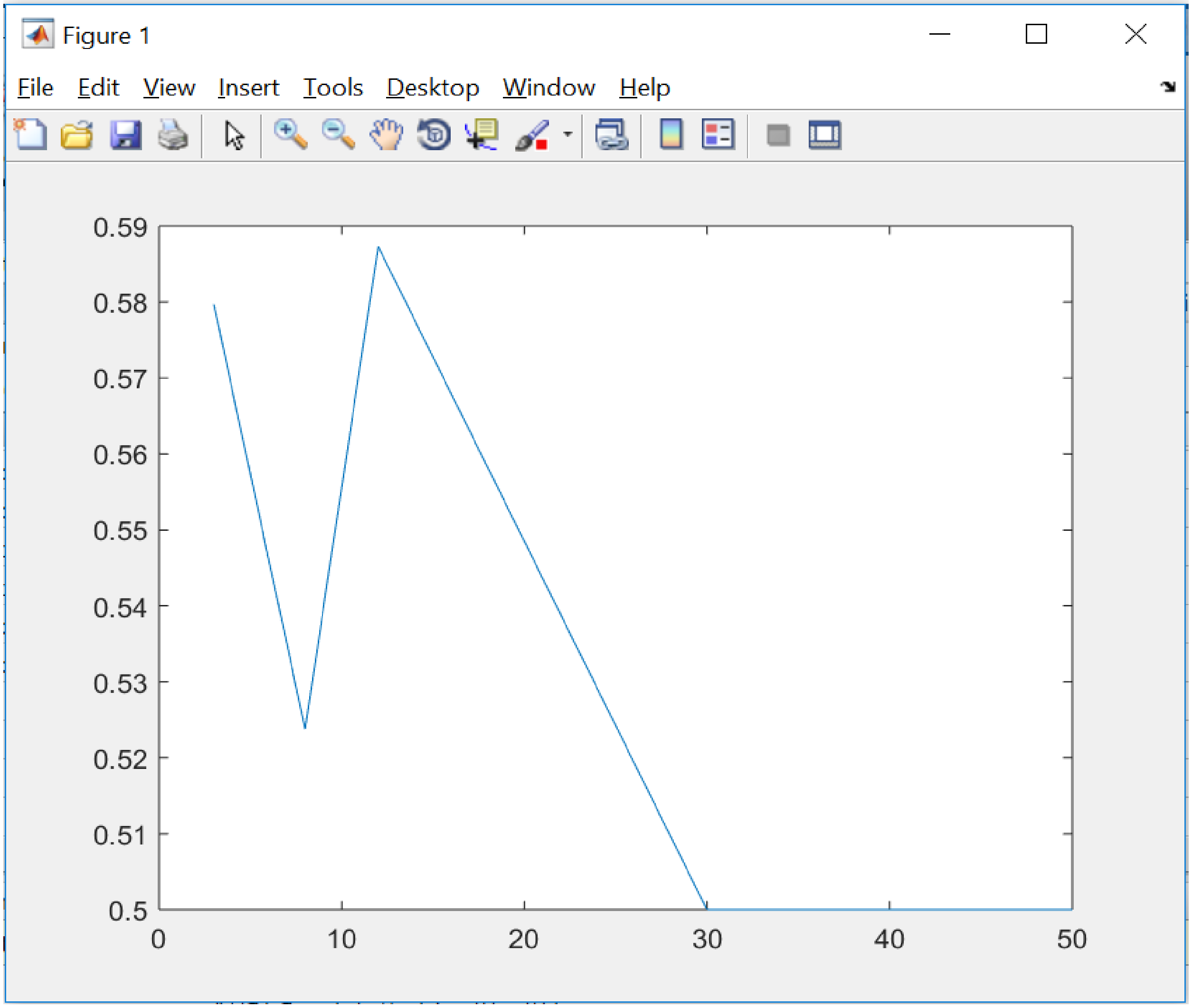
1.Precision for Decision Tree with min Leaf Node Size=3 is 0.5797

2. Precision for Decision Tree with min Leaf Node Size=8 is 0.5238

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.5873

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.5000

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.5000



The Precision graph for the Class Attribute ’Normal’ is as shown which gives the Number of True ‘Normals’ predicted out of Total ‘Normals’ Predicted. This Value is Highest for the Tree with min Leaf Size Value=12 and minimum for min Leaf Size Value=30 and 50

RECALL VALUES FOR FIVE DECISION TREES(ABNORMAL CLASS):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.4200

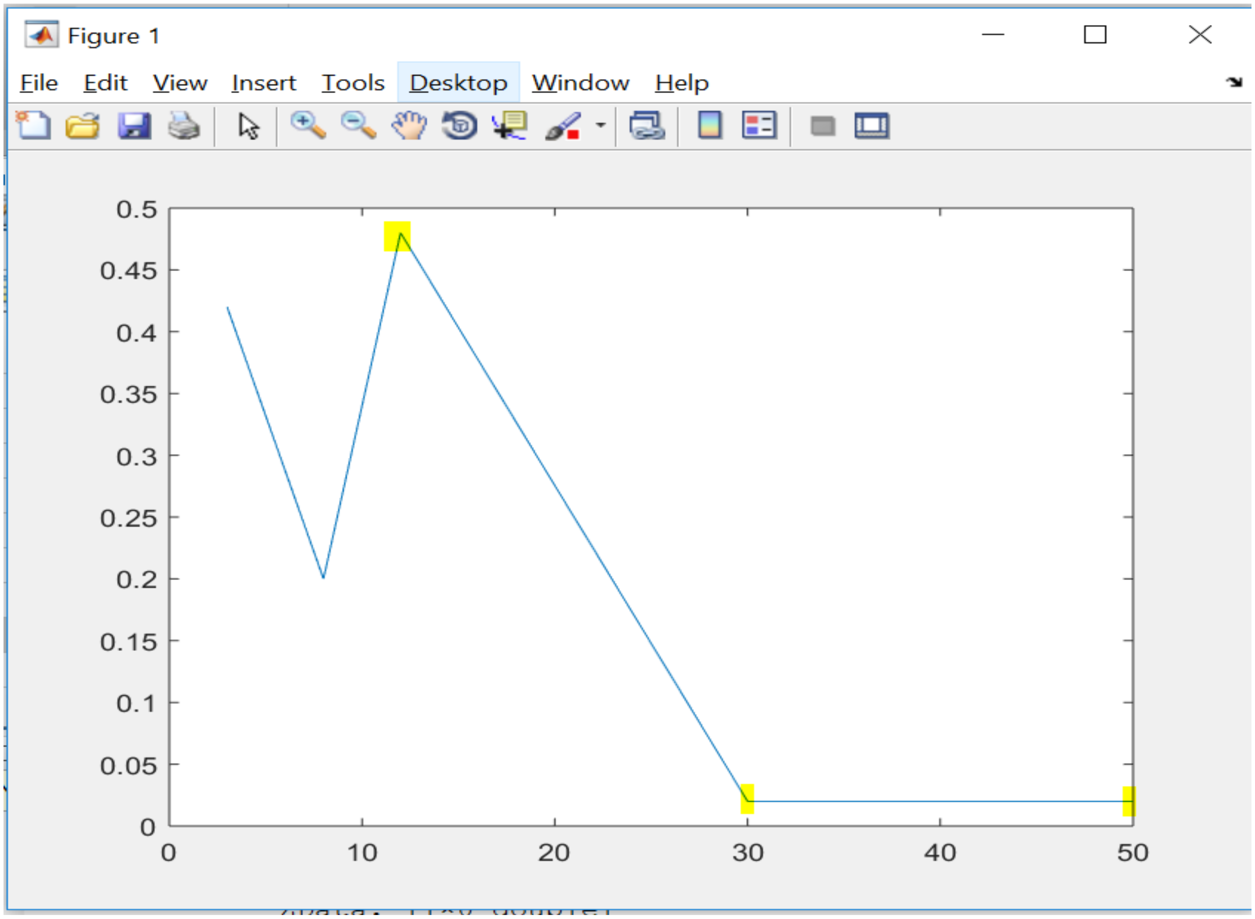
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.2000

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.4800

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.0200

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.0200

RECALL PLOT:



From the Recall Plot ,the Recall Value is Maximum for Leaf Node Min Size=12 and Maximum for Min Leaf Node Size=30,50

RECALL VALUES FOR FIVE DECISION TREES(NORMAL CLASS):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.8000

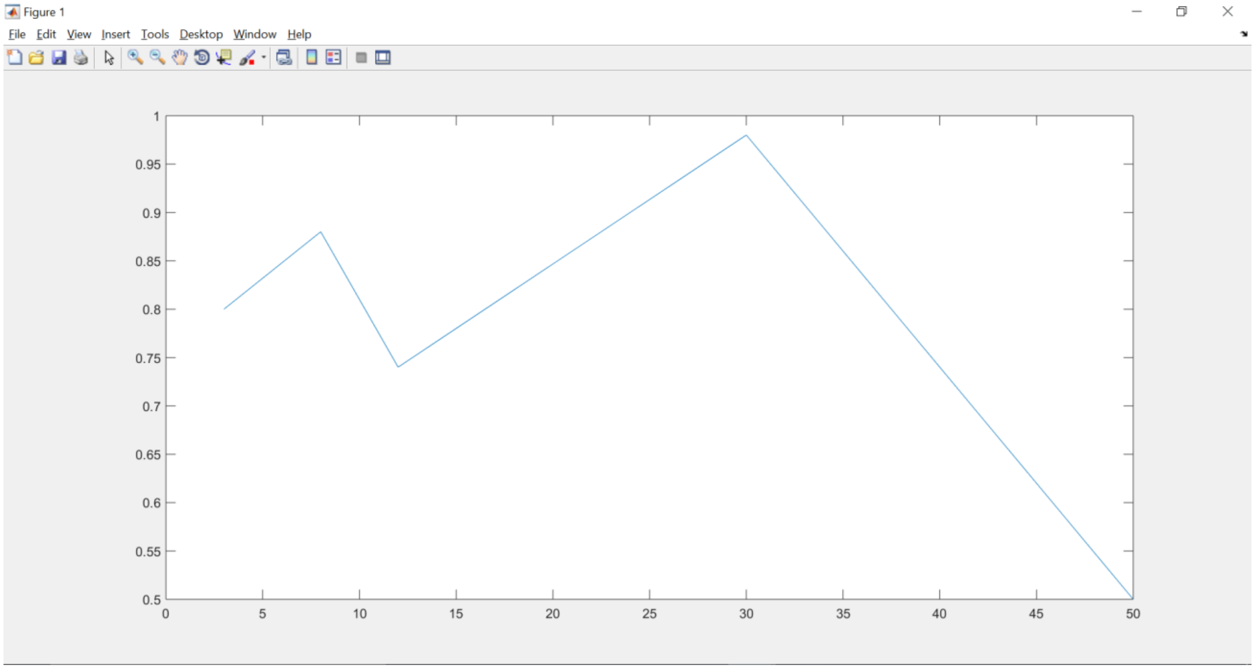
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.8800

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.7400

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.9800

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.5000

RECALL PLOT(NORMAL CLASS):



The Recall Plot for Class Attribute ’Normal’ is as shown. The Maximum Recall for this Tree is observed for values 8,30 due to Lesser Noise and Minimum Recall for this Tree is observed for the Tree with min Leaf Size50 due to Insufficient Examples.

COMMENTS ON ACCURACY, PRECISION AND RECALL VALUES:

1.The Accuracy Values are Highest for the Trees with min Leaf Node Sizes 3 and 12, this is because the Tree with min Leaf Size 3 is having a number of Nodes and the Noise interfering with the Classification would be very less. For Class 12, the Accuracy is the Best because of the Number of Nodes is only 3(which is Intermediate) and the Noise would be Negligible. Hence the trees with min Leaf Size 3 and 12 are having Highest Accuracy. The Trees with min Leaf Node Size values 30 and 50 are having very less Accuracy clearly due to Under Fitting Concept

> Based on the Accuracy Values, the Trees with min Leaf Size 3 and 12 are selected. Let's Investigate the parameters Precision and Recall.

2. The Precision Values are Highest for Tree with min Leaf Size 3 and then the Tree with min Leaf size 12. Precision is the Fraction of Retrieved Documents that are Relevant and based on the Plot trees with Leaf Node sizes 3 and 12 are Predicting are more Relevant compared to the Other. The Trees with min Leaf Node Size 30 and 50 are predicting least Relevant Values mostly due to the Number of Conditions of Split are very less to predict the Relevant Values.

>Based on the Precision Values, the Trees with min Leaf Size 3 and 12 are selected. Let us Investigate the Recall Values before selecting our Best Tree.

3. The recall Values for the Trees give the fraction of Relevant Documents that are successfully Predicted. Clearly, from the plot the tree with min Leaf Node Size 12 is having the Highest Recall Values for Abnormal Class and 30 is having the Highest Recall for Normal Class, this could be because the Number of Nodes is Intermediate for the Tree and the Noise could be very Less. The Trees with leaf Node Size 50 has lower Recall again due to Underfitting Concept the Number of Examples is not sufficient .

>Based on the Recall Values, the Tree with min Leaf Size 12 and 30 is having the Highest Recall Value.

PREFERENCE:

>Based on the Accuracy, Precision and recall Value and taking the concepts Underfitting and Overfitting into Consideration based on the Number of Nodes per Tree, the Tree with min Leaf Node Size value 12 is the Preferred Tree.

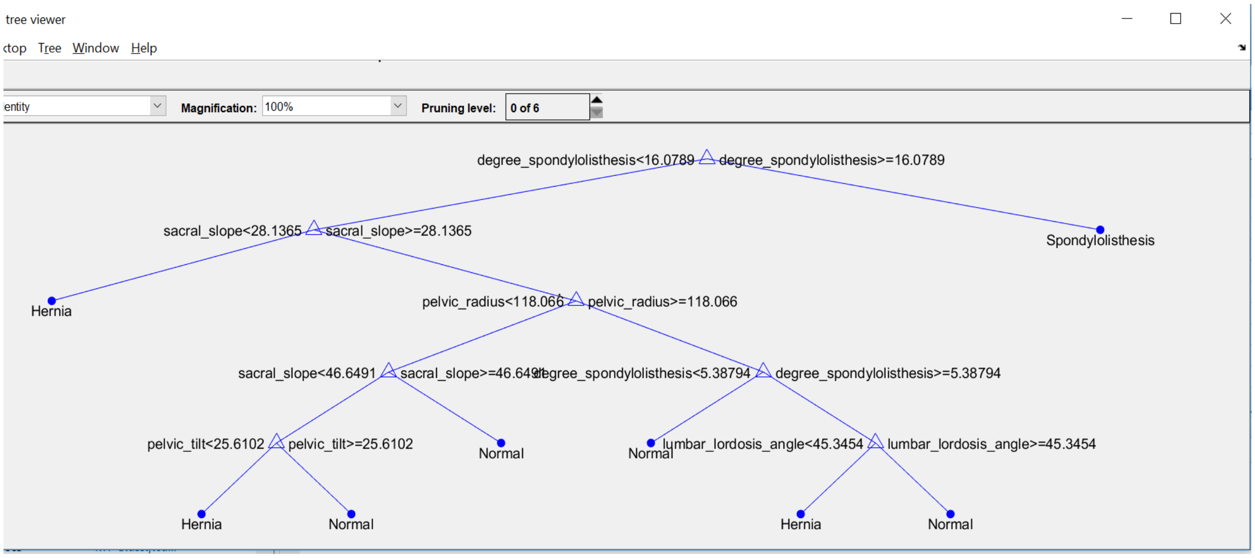
2. (30) Repeat the same tasks as done in Question-1 above for Data3 (Now the decision tree has three classes to work with). In addition to reporting results for parts (a) and (b) comment on the comparison of results obtained for (1a) and (2a) and also for (1b) and (2b). Give your analysis for the differences in results. Label this answer as 2c in your submission.

Solution:

Matlab Commands for Plotting Decision Trees:

```
ThreecTable=readtable('BiomechanicalData_column_3C_weka.csv');

ThreecTableTrain=ThreecTable([1:41 62:170 212:271],:);

ThreecTableTest=ThreecTable([42:61 171:211 272:310],:);


Treethreec3=fitctree(ThreecTableTrain,'class','MinLeafSize',3);

view(Treethreec3,'Mode','graph')


Treethreec8=fitctree(ThreecTableTrain,'class','MinLeafSize',8);

view(Treethreec8,'Mode','graph')


Treethreec12=fitctree(ThreecTableTrain,'class','MinLeafSize',12);

view(Treethreec12,'Mode','graph')


Treethreec30=fitctree(ThreecTableTrain,'class','MinLeafSize',30);

view(Treethreec30,'Mode','graph')


Treethreec50=fitctree(ThreecTableTrain,'class','MinLeafSize',50);

view(Treethreec50,'Mode','graph')
```

DECISION TREE FOR MIN RECORD VALUE=3:



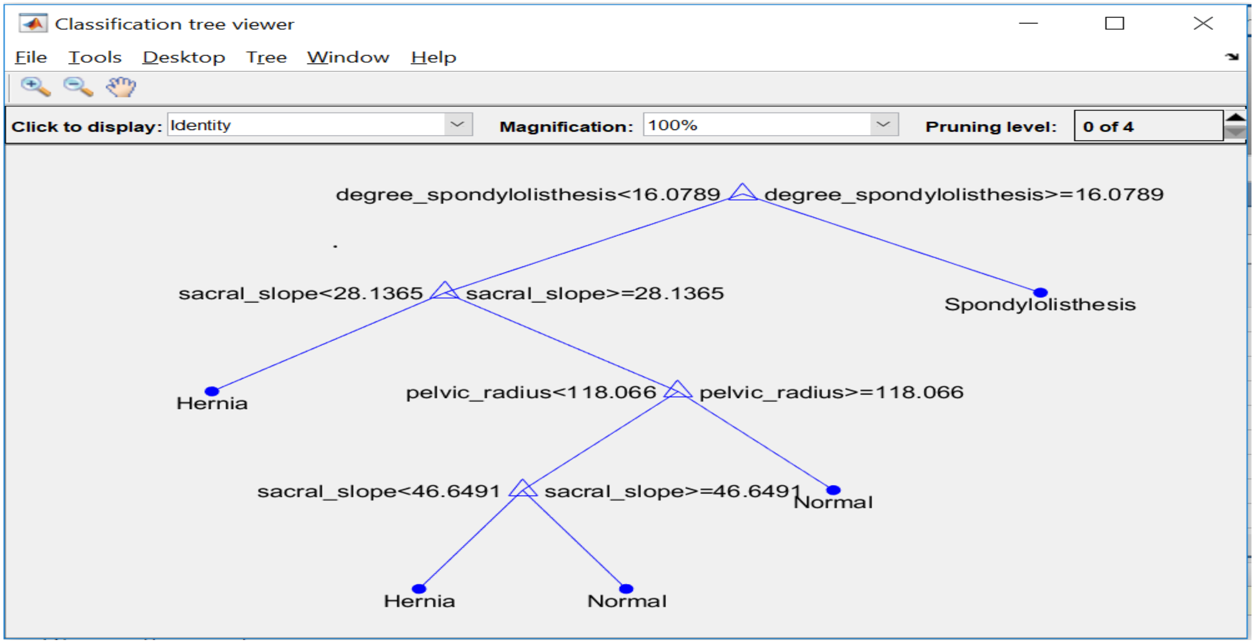
entity

Magnification: 100%

Pruning level: 0 of 6

Since the Min Leaf Node Value=3, the Decision Tree looks Bigger with a maximum Number of Conditions involved compared to the remaining.

DECISION TREE FOR MIN RECORD VALUE=8:



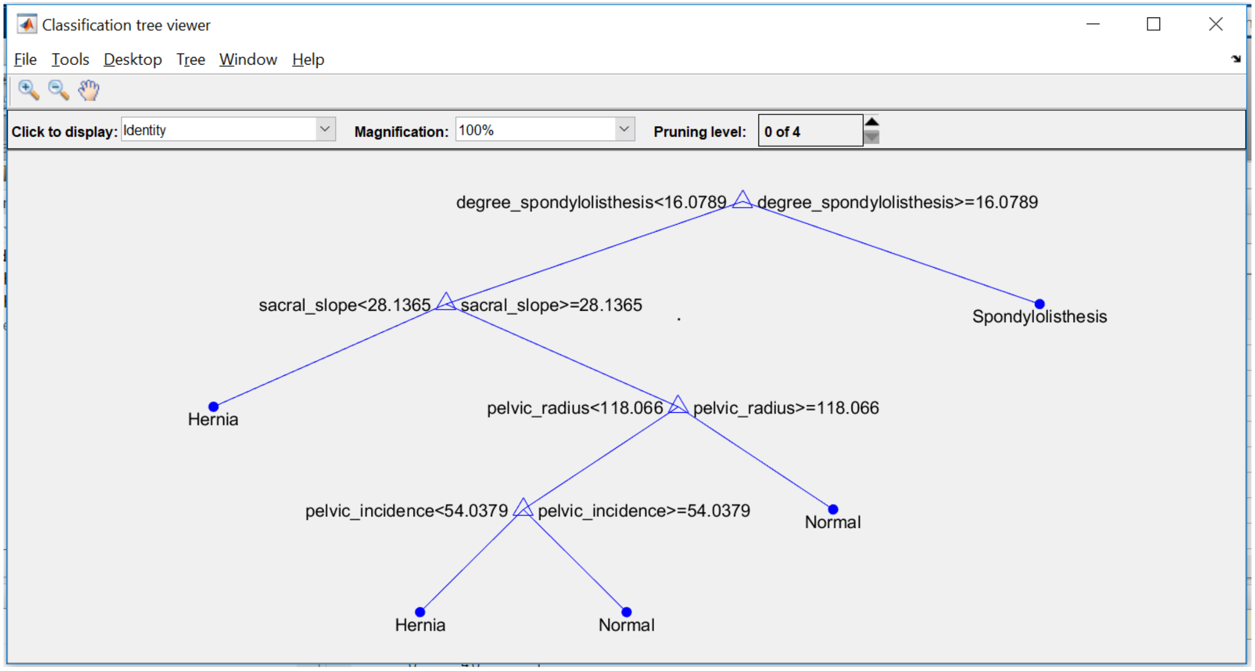
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 4

The Tree Size reduced as the Min Leaf Node Value is Increased to 8 as shown due to some of the Splitting Conditions are not satisfied after reducing the Leaf Size after 8.

DECISION TREE FOR MIN RECORD VALUE=12:


Classification tree viewer
File Tools Desktop Tree Window Help
degree_spondylolisthesis<16.0789
degree_spondylolisthesis>=16.0789
sacral_slope<28.1365
sacral_slope>=28.1365
Spondylolisthesis
Hernia
pelvic_radius<118.066
pelvic_radius>=118.066
pelvic_incidence<54.0379
pelvic_incidence>=54.0379
Normal
Hernia
Normal

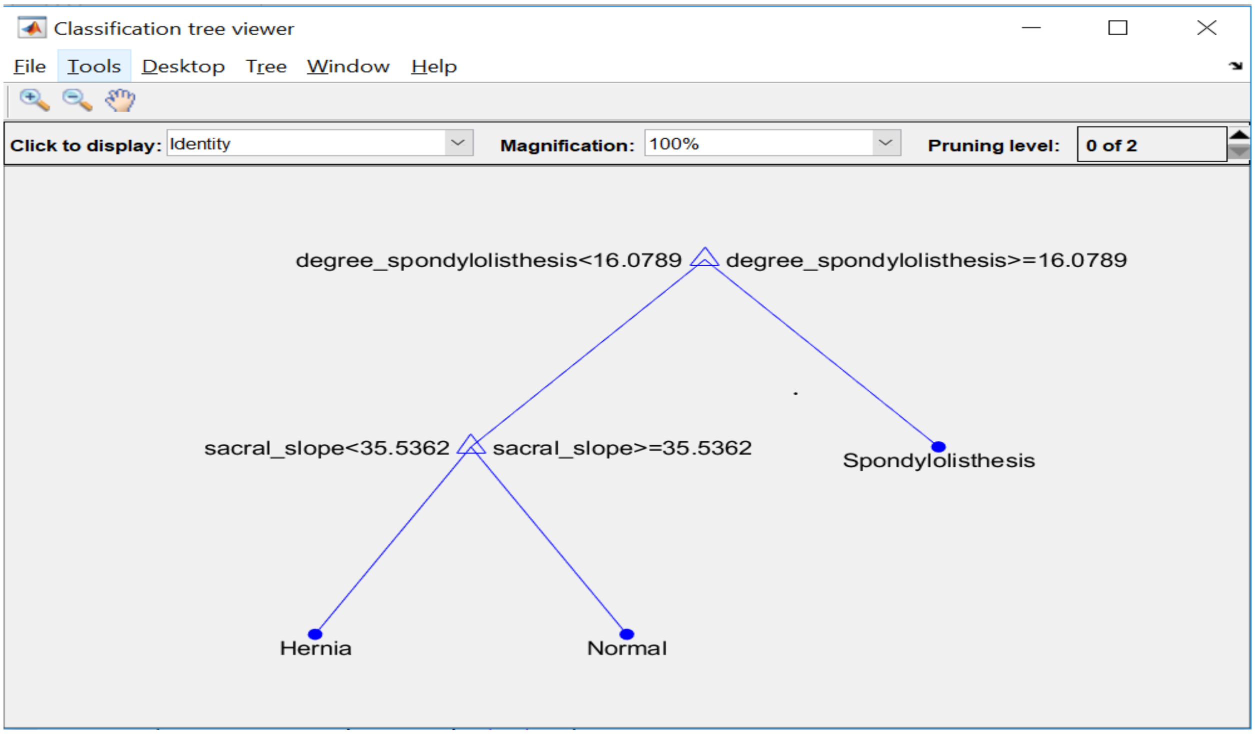
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 4

For Min record value=12, the Decision Tree looks similar to the Value at 8 with the Given Data.

DECISION TREE FOR MIN RECORD VALUE=30:


Classification tree viewer
File Tools Desktop Tree Window Help
degree_spondylolisthesis<16.0789
degree_spondylolisthesis>=16.0789
sacral_slope<35.5362
sacral_slope>=35.5362
Spondylolisthesis
Hernia
Normal

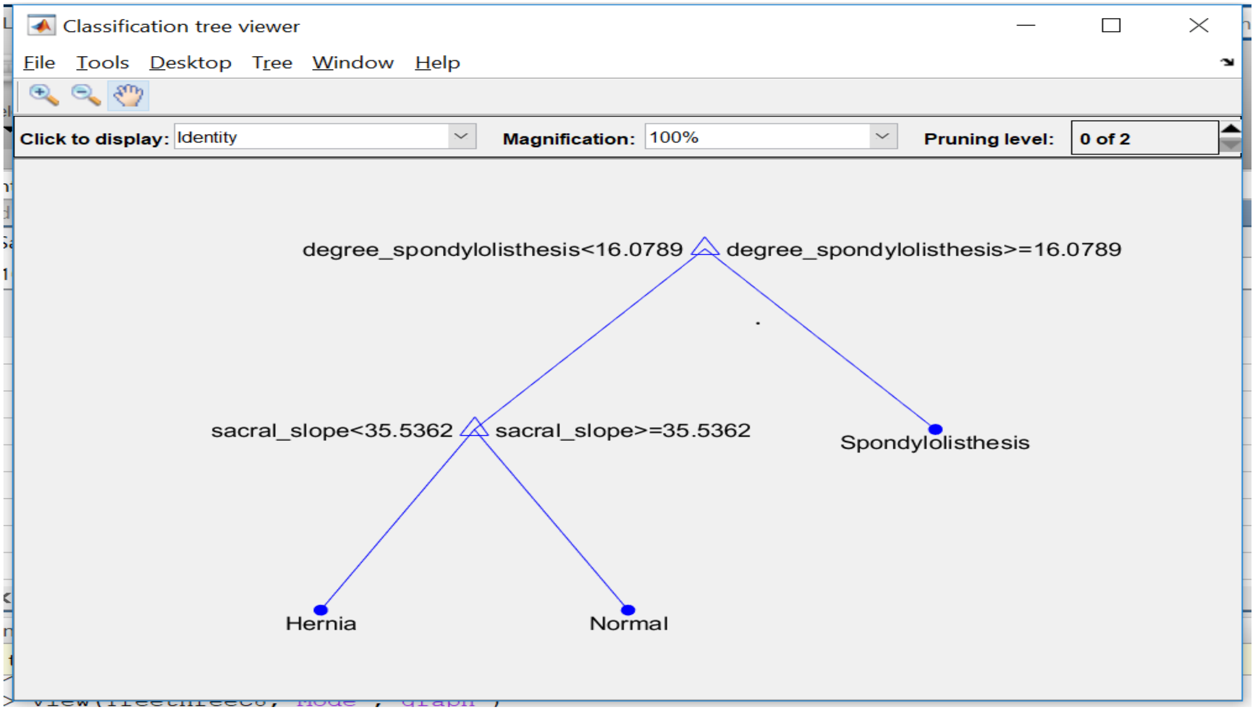
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 2

With the Increase of min Record Value, the Decision Tree Size Reduces as shown and let's check the Tree at value 50.

DECISION TREE FOR MIN RECORD VALUE=50:



Click to display: Identity

Magnification: 100%

Pruning level: 0 of 2

Classification tree viewer
File Tools Desktop Tree Window Help
degree_spondylolisthesis<16.0789
degree_spondylolisthesis>=16.0789
sacral_slope<35.5362
sacral_slope>=35.5362
Spondylolisthesis
Hernia
Normal

For Min record Value-50, the Decision Tree looks as shown with only 2 splits.

<u>COMPARISON OF FIVE DECISION TREES:</u>

From the above 5 Trees, it is clear that, as the Min Leaf Node Size Increases, the Tree Size decreases because of the limit in splitting leaf nodes are increasing.

<u>ESTIMATION OF THE BEST TREE:</u>

- By having a look at the Decision Trees for the 5 min Leaf Sizes 3,8,12,30,50 the Trees have 8,5,5,3,3 Nodes Respectively.

1. Based on the Number of Nodes, the Trees with min Leaf Size Values 30 and 50 are not Preferred due to Least Number of Splitting Parameters and due to Underfitting Concept.
2. The Tree with min Leaf Size 3 is also not preferred due to Number of Nodes being 8 by which the Noise might have an Interference.
3. The Trees with min Leaf Size values 8 and 12 are preferred. By looking at the Trees the best of 2 cannot be predicted as both are exactly Same. But by looking at the Accuracy, Precision, Recall Values, clearly all these are Highest for min Leaf Node with value 8. Hence preferred Tree is the Tree with min Leaf Node Value 8. Furthermore, analysis of the Decision based on Accuracy, Precision and recall plots is mentioned below.

2.b. MATLAB CODE FOR CALCULATING ACCURACY, PRECISION AND RECALL VALUES:

```
predictionthree=predict(Treethreec3,ThreecTableTest);
DataTableTestcellthree=table2cell(ThreecTableTest);
cellthree=DataTableTestcellthree(:,[7]);
Cthree=confusionmat(cellthree,predictionthree);

predictionthree8=predict(Treethreec8,ThreecTableTest);
Cthree8=confusionmat(cellthree,predictionthree8);
predictionthree12=predict(Treethreec12,ThreecTableTest);
Cthree12=confusionmat(cellthree,predictionthree12);
predictionthree30=predict(Treethreec30,ThreecTableTest);
Cthree30=confusionmat(cellthree,predictionthree30);
predictionthree50=predict(Treethreec50,ThreecTableTest);
Cthree50=confusionmat(cellthree,predictionthree50);

Accuracycthree3=(Cthree(1,1)+Cthree(2,2)+Cthree(3,3))/(Cthree(1,1)+Cthree(1,2)+Cthree(1,3)+Cthree(2,1)+Cthree(2,2)+Cthree(2,3)+Cthree(3,1)+Cthree(3,2)+Cthree(3,3));
Accuracycthree8=(Cthree8(1,1)+Cthree8(2,2)+Cthree8(3,3))/(Cthree8(1,1)+Cthree8(1,2)+Cthree8(1,3)+Cthree8(2,1)+Cthree8(2,2)+Cthree8(2,3)+Cthree8(3,1)+Cthree8(3,2)+Cthree8(3,3));
Accuracycthree12=(Cthree12(1,1)+Cthree12(2,2)+Cthree12(3,3))/(Cthree12(1,1)+Cthree12(1,2)+Cthree12(1,3)+Cthree12(2,1)+Cthree12(2,2)+Cthree12(2,3)+Cthree12(3,1)+Cthree12(3,2)+Cthree12(3,3));
Accuracycthree30=(Cthree30(1,1)+Cthree30(2,2)+Cthree30(3,3))/(Cthree30(1,1)+Cthree30(1,2)+Cthree30(1,3)+Cthree30(2,1)+Cthree30(2,2)+Cthree30(2,3)+Cthree30(3,1)+Cthree30(3,2)+Cthree30(3,3));
Accuracycthree50=(Cthree50(1,1)+Cthree50(2,2)+Cthree50(3,3))/(Cthree50(1,1)+Cthree50(1,2)+Cthree50(1,3)+Cthree50(2,1)+Cthree50(2,2)+Cthree50(2,3)+Cthree50(3,1)+Cthree50(3,2)+Cthree50(3,3));

Precisionthree3=(Cthree(1,1))/(Cthree(1,1)+Cthree(2,1)+Cthree(3,1));
Precisionthree8=(Cthree8(1,1))/(Cthree8(1,1)+Cthree8(2,1)+Cthree8(3,1));
Precisionthree12=(Cthree12(1,1))/(Cthree12(1,1)+Cthree12(2,1)+Cthree12(3,1));
Precisionthree30=(Cthree30(1,1))/(Cthree30(1,1)+Cthree30(2,1)+Cthree30(3,1));
Precisionthree50=(Cthree50(1,1))/(Cthree50(1,1)+Cthree50(2,1)+Cthree50(3,1));

Precisionthree3spondy=(Cthree(2,2))/(Cthree(1,2)+Cthree(2,2)+Cthree(3,2));
Precisionthree8spondy=(Cthree8(2,2))/(Cthree8(1,2)+Cthree8(2,2)+Cthree8(3,2));
Precisionthree12spondy=(Cthree12(2,2))/(Cthree12(1,2)+Cthree12(2,2)+Cthree12(3,2));
Precisionthree30spondy=(Cthree30(2,2))/(Cthree30(1,2)+Cthree30(2,2)+Cthree30(3,2));
Precisionthree50spondy=(Cthree50(2,2))/(Cthree50(1,2)+Cthree50(2,2)+Cthree50(3,2));

Precisionthree3normal=(Cthree(3,3))/(Cthree(1,3)+Cthree(2,3)+Cthree(3,3));
Precisionthree8normal=(Cthree8(3,3))/(Cthree8(1,3)+Cthree8(2,3)+Cthree8(3,3));
Precisionthree12normal=(Cthree12(3,3))/(Cthree12(1,3)+Cthree12(2,3)+Cthree12(3,3));
Precisionthree30normal=(Cthree30(3,3))/(Cthree30(1,3)+Cthree30(2,3)+Cthree30(3,3));
Precisionthree50normal=(Cthree50(3,3))/(Cthree50(1,3)+Cthree50(2,3)+Cthree50(3,3));

Recallthree3=(Cthree(1,1))/(Cthree(1,1)+Cthree(1,2)+Cthree(1,3));
Recallthree8=(Cthree8(1,1))/(Cthree8(1,1)+Cthree8(1,2)+Cthree8(1,3));
```

```
Recallthree12=(Cthree12(1,1))/(Cthree12(1,1)+Cthree12(1,2)+Cthree12(1,3));
Recallthree30=(Cthree30(1,1))/(Cthree30(1,1)+Cthree30(1,2)+Cthree30(1,3));
Recallthree50=(Cthree50(1,1))/(Cthree50(1,1)+Cthree50(1,2)+Cthree50(1,3));

Recallthree3spondy=(Cthree(2,2))/(Cthree(2,1)+Cthree(2,2)+Cthree(2,3));
Recallthree8spondy=(Cthree8(2,2))/(Cthree8(2,1)+Cthree8(2,2)+Cthree8(2,3));
Recallthree12spondy=(Cthree12(2,2))/(Cthree12(2,1)+Cthree12(2,2)+Cthree12(2,3));
Recallthree30spondy=(Cthree30(2,2))/(Cthree30(2,1)+Cthree30(2,2)+Cthree30(2,3));
Recallthree50spondy=(Cthree50(2,2))/(Cthree50(2,1)+Cthree50(2,2)+Cthree50(2,3));

Recallthree3normal=(Cthree(3,3))/(Cthree(3,1)+Cthree(3,2)+Cthree(3,3));
Recallthree8normal=(Cthree8(3,3))/(Cthree8(3,1)+Cthree8(3,2)+Cthree8(3,3));
Recallthree12normal=(Cthree12(3,3))/(Cthree12(3,1)+Cthree12(3,2)+Cthree12(3,3));
Recallthree30normal=(Cthree30(3,3))/(Cthree30(3,1)+Cthree30(3,2)+Cthree30(3,3));
Recallthree50normal=(Cthree50(3,3))/(Cthree50(3,1)+Cthree50(3,2)+Cthree50(3,3));

AccuracyPlotthree=plot([3,8,12,30,50],[Accuracycthree3 Accuracycthree8 Accuracycthree12
Accuracycthree30 Accuracycthree50])
PrecisionPlotthree=plot([3,8,12,30,50],[Precisionthree3 Precisionthree8 Precisionthree12
Precisionthree30 Precisionthree50])
RecallPlotthree=plot([3,8,12,30,50],[Recallthree3 Recallthree8 Recallthree12 Recallthree30
Recallthree50])

PrecisionPlotthreespondy=plot([3,8,12,30,50],[Precisionthree3spondy Precisionthree8spondy
Precisionthree12spondy Precisionthree30spondy Precisionthree50spondy])
PrecisionPlotthreenormal=plot([3,8,12,30,50],[Precisionthree3normal Precisionthree8normal
Precisionthree12normal Precisionthree30normal Precisionthree50normal])

RecallPlotthreespondy=plot([3,8,12,30,50],[Recallthree3spondy Recallthree8spondy
Recallthree12spondy Recallthree30spondy Recallthree50spondy])
RecallPlotthreenormal=plot([3,8,12,30,50],[Recallthree3normal Recallthree8normal
Recallthree12normal Recallthree30normal Recallthree50normal])
```

2.b. **ACCURACY VALUES FOR FIVE DECISION TREES:**

1.Accuracy for Decision Tree with min Leaf Node Size=3 is 0.8300

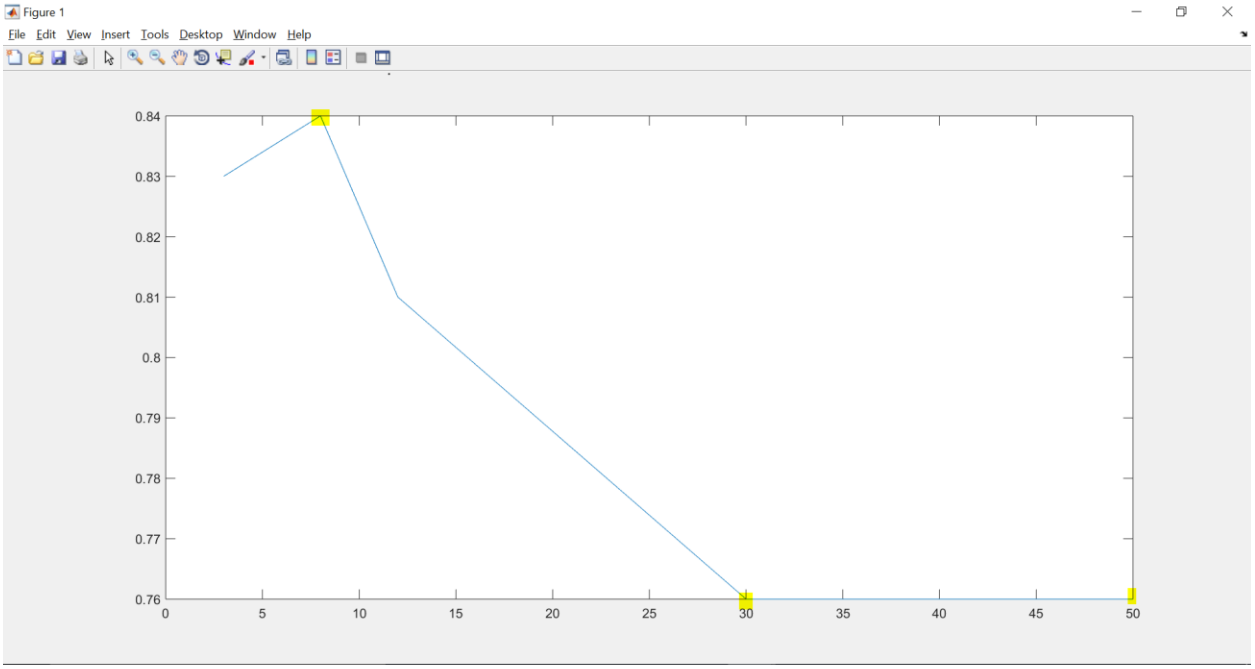
2. Accuracy for Decision Tree with min Leaf Node Size=8 is 0.8400

3. Accuracy for Decision Tree with min Leaf Node Size=12 is 0.8100

4. Accuracy for Decision Tree with min Leaf Node Size=30 is 0.7600

5. Accuracy for Decision Tree with min Leaf Node Size=50 is 0.7600

ACCURACY PLOT:



The Accuracy is Highest for Min Leaf Size=8 and Lowest for Min Node Sizes 30 and 50.The Reason for High Accuracy is due to more Number of Training Examples and Lesser Noise.

PRECISION VALUES FOR FIVE DECISION TREES('Abnormal' Class):

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.5714

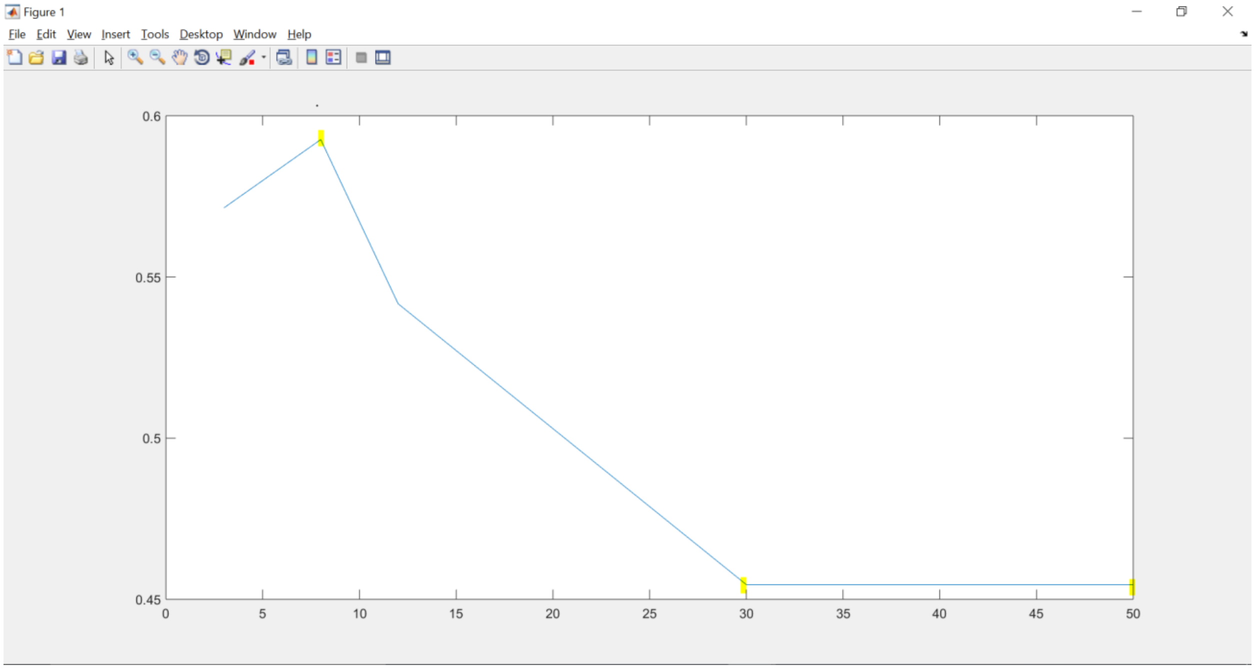
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.5926

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.5417

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.4545

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.4545

PRECISON PLOT:



The Number of True Positives Predicted from the Total Positives is Highest for Min Leaf Value=8 and is minimum for min Leaf Value =30 and 50.For min Leaf Sizes 30 and 50 the Tree Size is very Less, and all the Conditions haven't been Covered. For min Leaf Size Value 8,the Tree size is medium and also the Noise could be less ,Hence, it has the Highest Precision.

PRECISION VALUES FOR FIVE DECISION TREES('Spondylolisthesis' Class):

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.9756

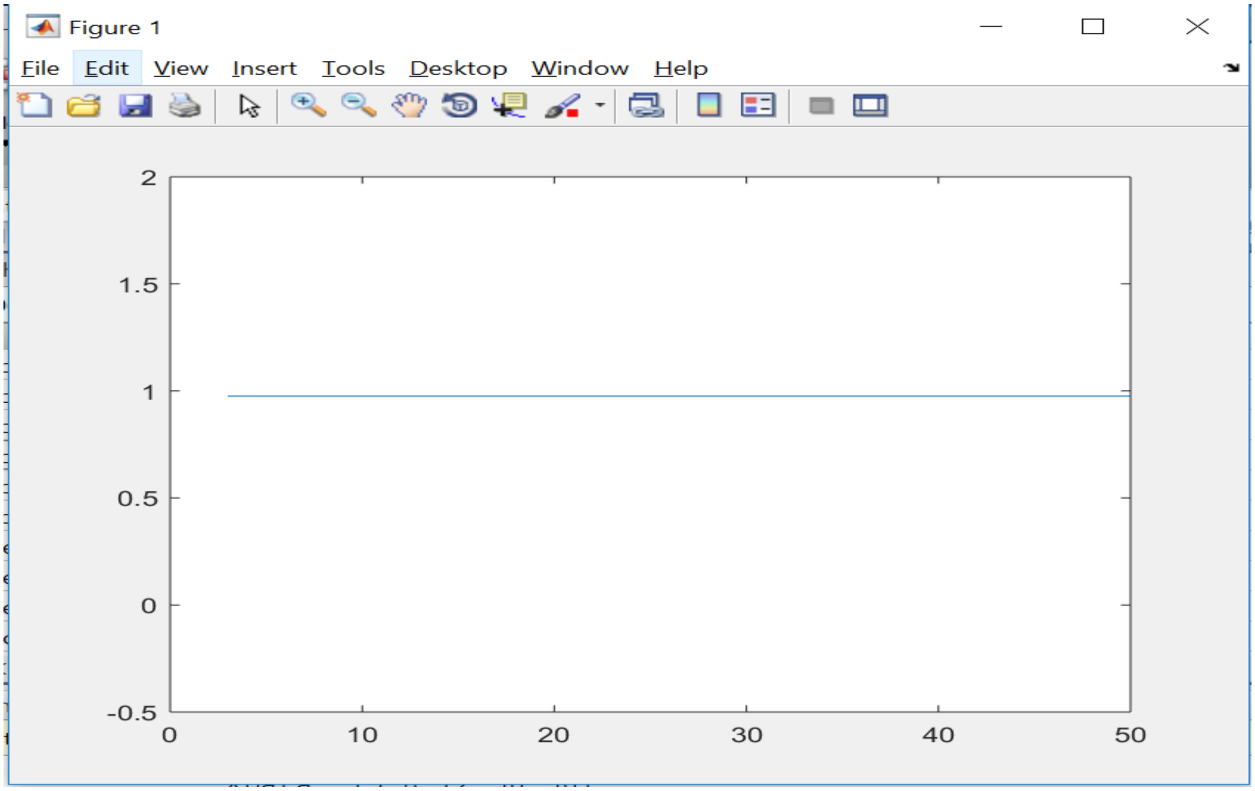
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.9756

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.9756

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.9756

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.9756

PRECISON PLOT:



For all the min Leaf Size values, the Decision Trees have 0.9756 Precision for Spondy Class which are the True Spondys out of the Predicted Spondys. We cannot guess the Better Tree based on this Precision Plot.

PRECISION VALUES FOR FIVE DECISION TREES(‘Normal’ Class):

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.8710

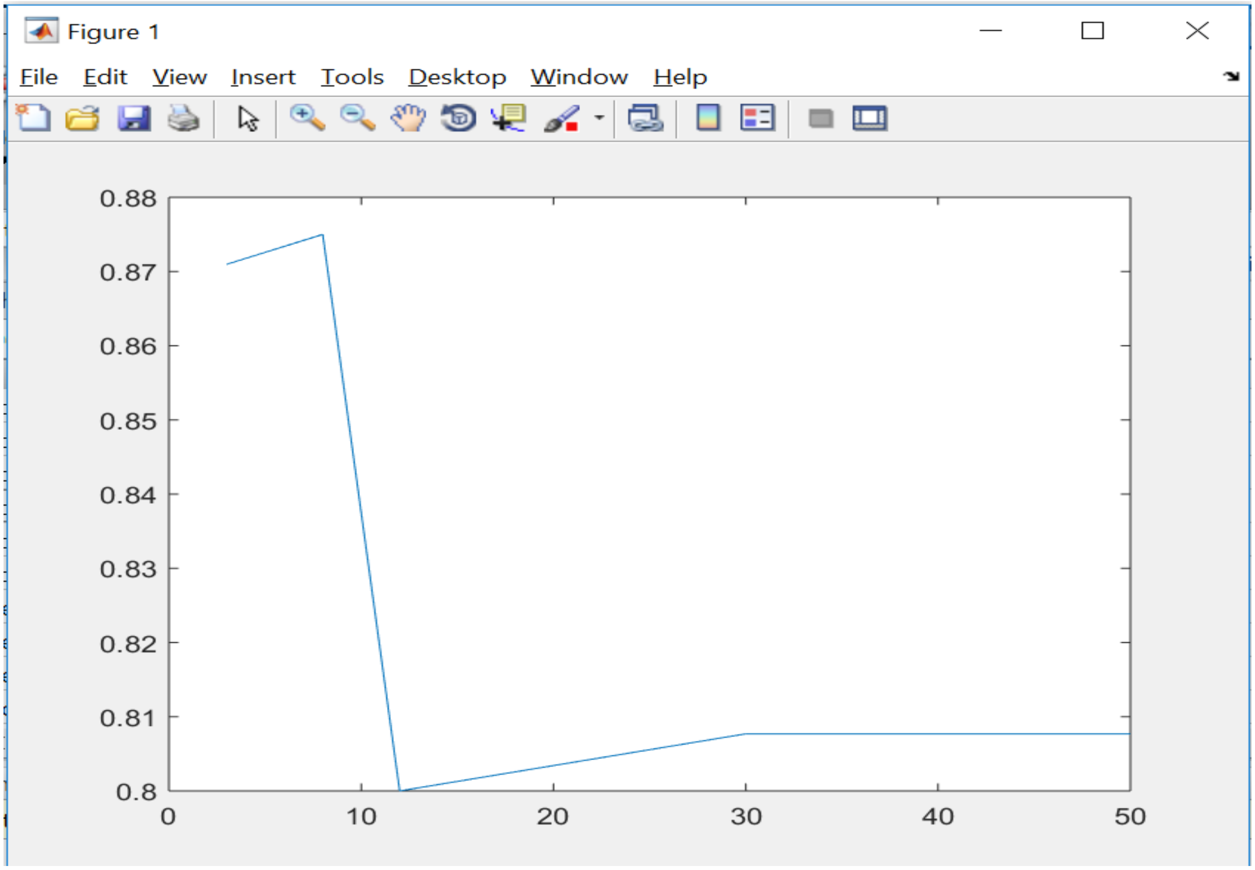
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.8750

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.8000

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.8077

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.8077

PRECISON PLOT:



The Precision Plot is Higher for max Leaf Node Value=8 for the Class Attribute ‘Normal’ and is more Likely because it has more Sufficient Examples and Lesser Noise.

RECALL VALUES FOR DECISION TREES('Abnormal' Class):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.8421

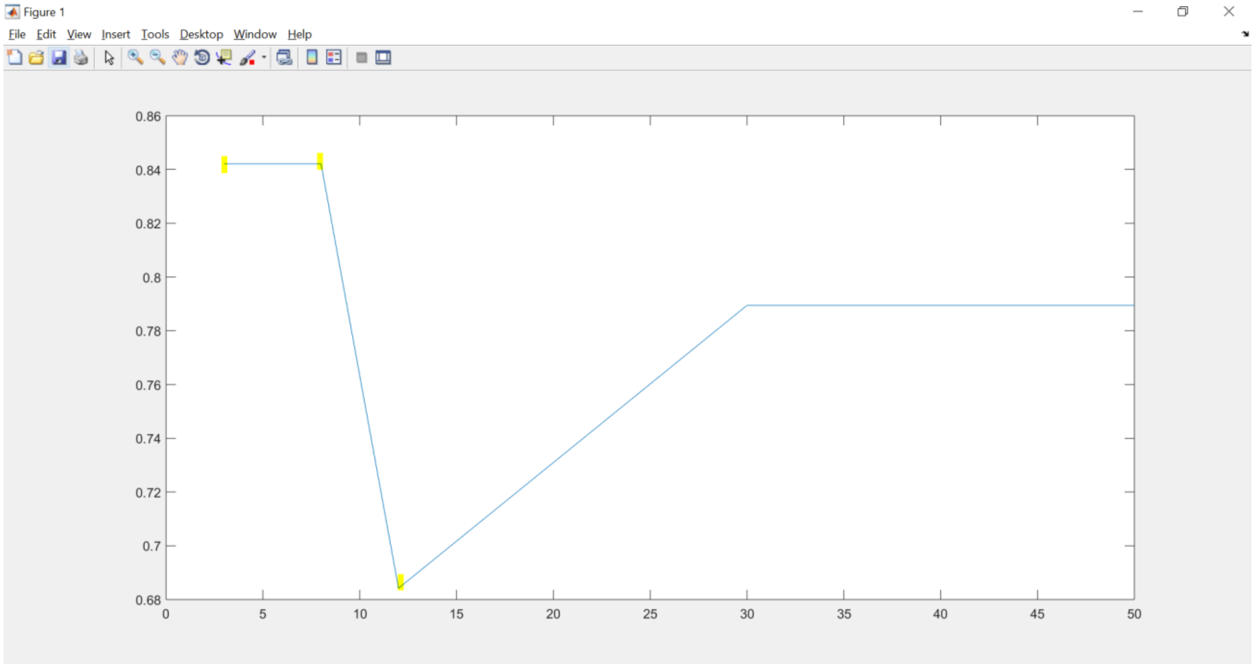
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.8421

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.6842

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.7895

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.7895

<u>RECALL PLOT</u>:



The Recall Values are Highest for min Leaf Sizes 3 and 8 and are least for Leaf Size 12,mostly due to the Tree Size and more training Examples have been covered.

RECALL VALUES FOR DECISION TREES(FOR CLASS 'Spondylolisthesis'):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.9756

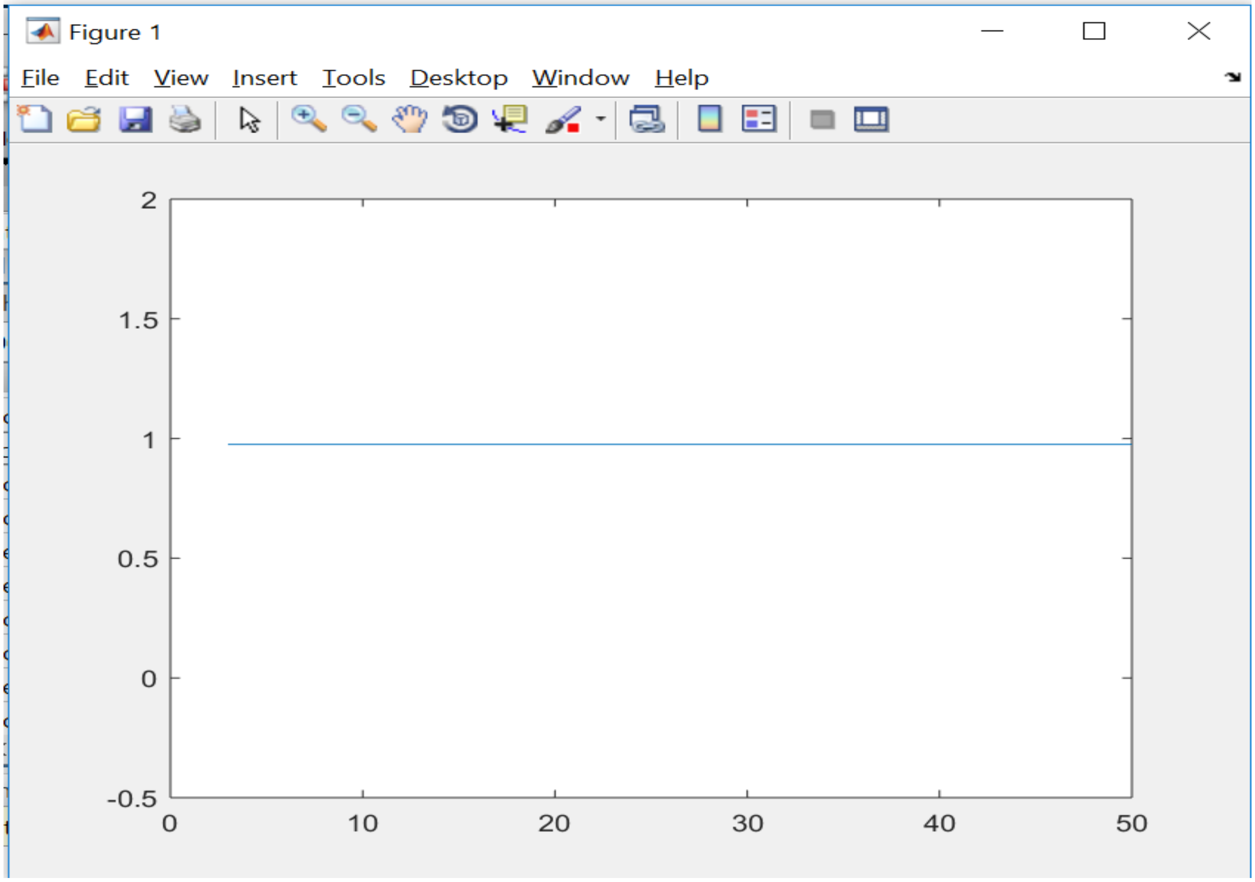
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.9756

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.9756

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.9756

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.9756

RECALL PLOT:



The Recall Plot is constant for Class Attribute ‘Spondy’ Hence the best Class prediction cannot be made using recall for this Attribute.

RECALL VALUES FOR DECISION TREES(FOR CLASS Normal):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.6750

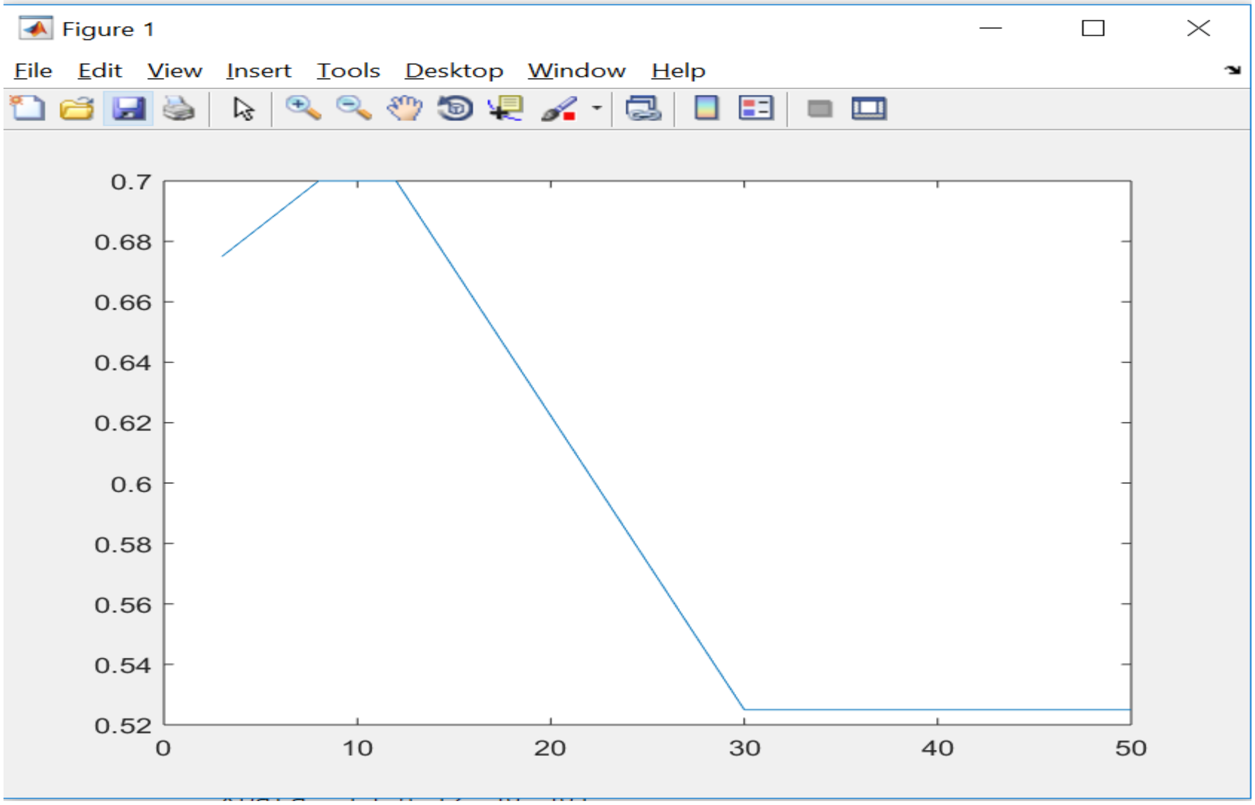
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.7000

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.7000

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.5250

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.5250

RECALL PLOT:



The Recall Plot for Class Attribute ’Normal’ is Highest at min Leaf Size Values 8 and 12 due to the Tree Sizes being Intermediate and having Lesser Noise or Wrong assumptions and is Highest at min Leaf Node Sizes 30 and 50 due to Fewer Examples.

COMMENTS ON ACCURACY, PRECISION AND RECALL VALUES:

1.Based on the Accuracy Plot, the Accuracy is Highest for the Min Leaf Node value with Value 8. The Reason is because the Tree being having intermediate (Not more or less) Nodes, it is exempted from the Concept of Underfitting and Overfitting). The Trees with min Leaf Node Sizes 30 and 50 are having very few Nodes and the Training Conditions are not sufficient for a Good Prediction.

- Based on Accuracy, the Tree with min Leaf Size 8 is preferred, let's Examine the Precision and recall plots as well.

2. The Precision plot for the 3 class is similar to the Accuracy Plot and is Highest when Min Leaf Node Value is 8 and Least when Min Leaf Node Value is 30 and 50. The Higher Precision could be because of the Tree being exempted from Under fitting and Overfitting. Lesser Precision is because of Lesser Training Examples in order to Predict the True Positives out of all Predicted Positives.

> Based on Precision, the Tree with min Leaf Size 8 is preferred, Now let's look at the Recall Values.

3.The Recall plot for 3 class is highest for Tree with Min Leaf Node Value 3 and 8 as the Number of Training Examples are more. The Recall Plot is lowest for min Leaf Node Value 12 .The Reason could be due to Incorrect Examples used in Training or Noise.

PREFERENCE:

1.Based on the Accuracy, Precision and Recall Values and taking the Concepts of Underfitting and Overfitting into consideration, the Tree with min Leaf Size 8 is preferred for the 3 class Model.

2.C. COMPARISON OF RESULTS IN TWO CLASS DECISION TREE AND THREE CLASS DECISION TREES:

For a 3 class Problem, the Accuracy has been Increased to 0.8400 clearly compared to 0.6100 for a 2 class Problem. By increasing the class Attributes, the Tree is Accurately able to Predict .

When comparing the Precision for a 2 class and 3 class, the Precision Values for 3 class also have Improved for couple of class Attributes .For the third Attribute the Value looks like 2 class Precision Value. Overall Precision looks Improved for a 3 class and we can Infer that the Prediction Capacity for a 3 class has been Improved compared to 2 class Problem .The Recall Values also are higher for 3 class Problem.

Hence the 3 class has Improved Performance compared to 2 class ,the reason could be because the Tree Formation is accurate when distinguishing a greater Number of Attributes .for this given Data.

3. (30) Take Data2 for this question. Partition each column into four sets of equal width of values. Assign these intervals as values 0, 1, 2, and 3 and replace each value in the original data by its corresponding interval number. a. Show the boundaries for each interval for each attribute. b. Learn a decision tree with this transformed data and compute performance parameters in the same way as done for 1b and 2b. c. Compare the performance metric as obtained in 1b with those obtained here in 3b. Explain the differences in performance and give your intuitive reasons why these differences are observed.

Solution:

MATLAB CODE FOR MAKING SAMPLED DATA:

```
binranges2=[-6.55,7.445,21.44,35.39,49.43]
i2=table2array(DataTable(:,2))
[bincounts2,ind2]=histc(i2,binranges2)
ind2=ind2-1
binranges3=[14,41.935,69.87,97.805,125.745]
i3=table2array(DataTable(:,3))
[bincounts3,ind3]=histc(i3,binranges3)
ind3=ind3-1
binranges4=[13.36,40.375,67.39,94.405,121.43]
i4=table2array(DataTable(:,4))
[bincounts4,ind4]=histc(i4,binranges4)
ind4=ind4-1
binranges5=[70.08,93.32,116.57,139.82,163.08]
i5=table2array(DataTable(:,5))
[bincounts5,ind5]=histc(i5,binranges5)
ind5=ind5-1
binranges6=[-11.05,96.34,203.73,311.12,418.55]
i6=table2array(DataTable(:,6))
[bincounts6,ind6]=histc(i6,binranges6)
ind6=ind6-1

ind=[ind1 ind2 ind3 ind4 ind5 ind6 ]
ind=array2table(ind)
SampledData=horzcat(ind(:,:),DataTable(:,7))

SampledData.Properties.VariableNames={'pelvic_incidence','pelvic_tilt_numeric','lumbar_lordosis_angle','sacral_slope','pelvic_radius','degree_spondylolisthesis','class'}
```

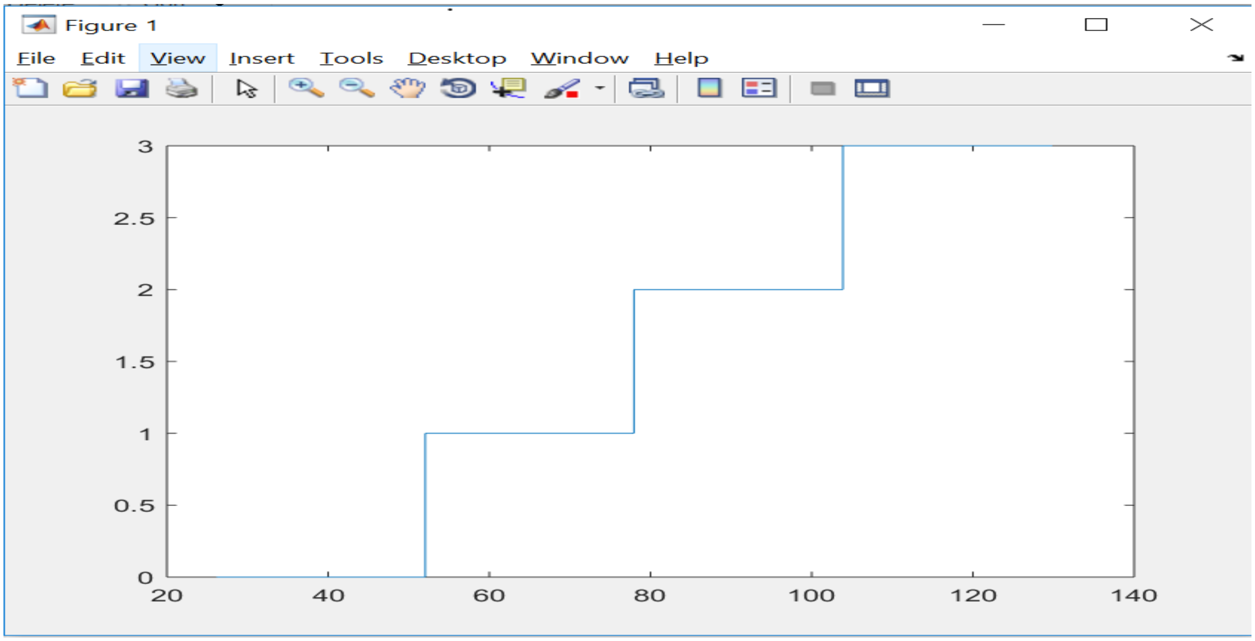
3.a. BOUNDARIES FOR EACH INTERVAL FOR EACH ATTRIBUTES:

1.Pelvic Incidence Boundaries:

binranges1 = [26.147,52.06,77.96,103.88,129.84]

Samples= {0,1,2,3}

Plotting the Values

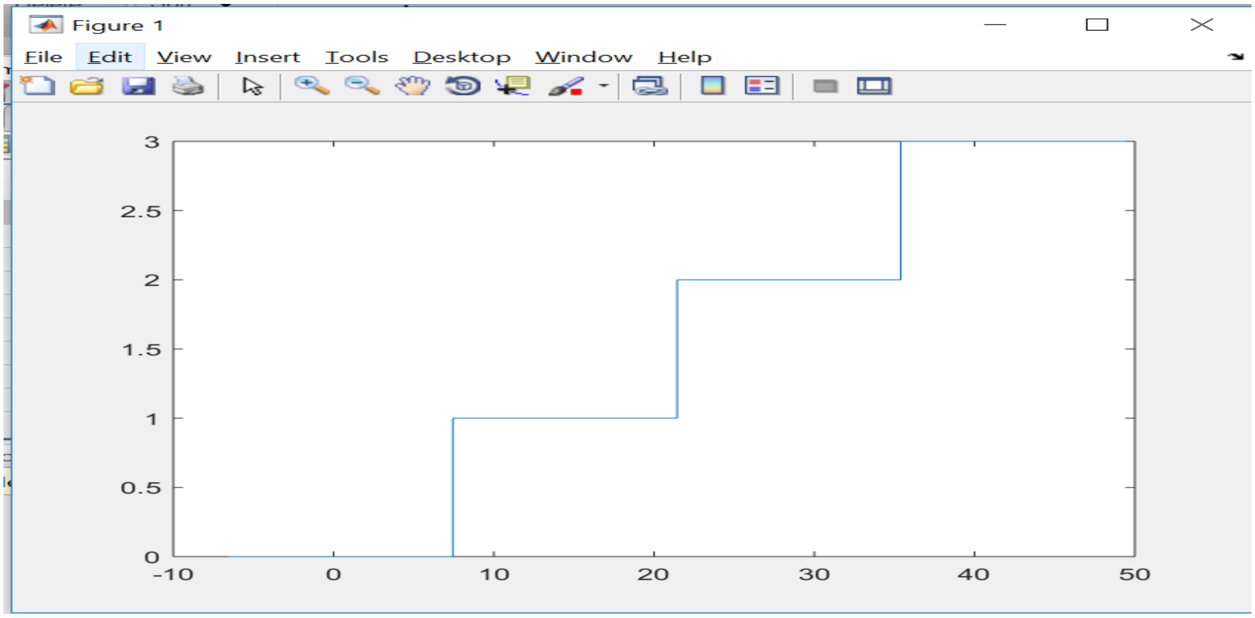


2. ‘pelvic_tilt’ numeric Boundaries:

binranges2=[-6.55,7.445,21.44,35.39,49.43]

Samples= {0,1,2,3}

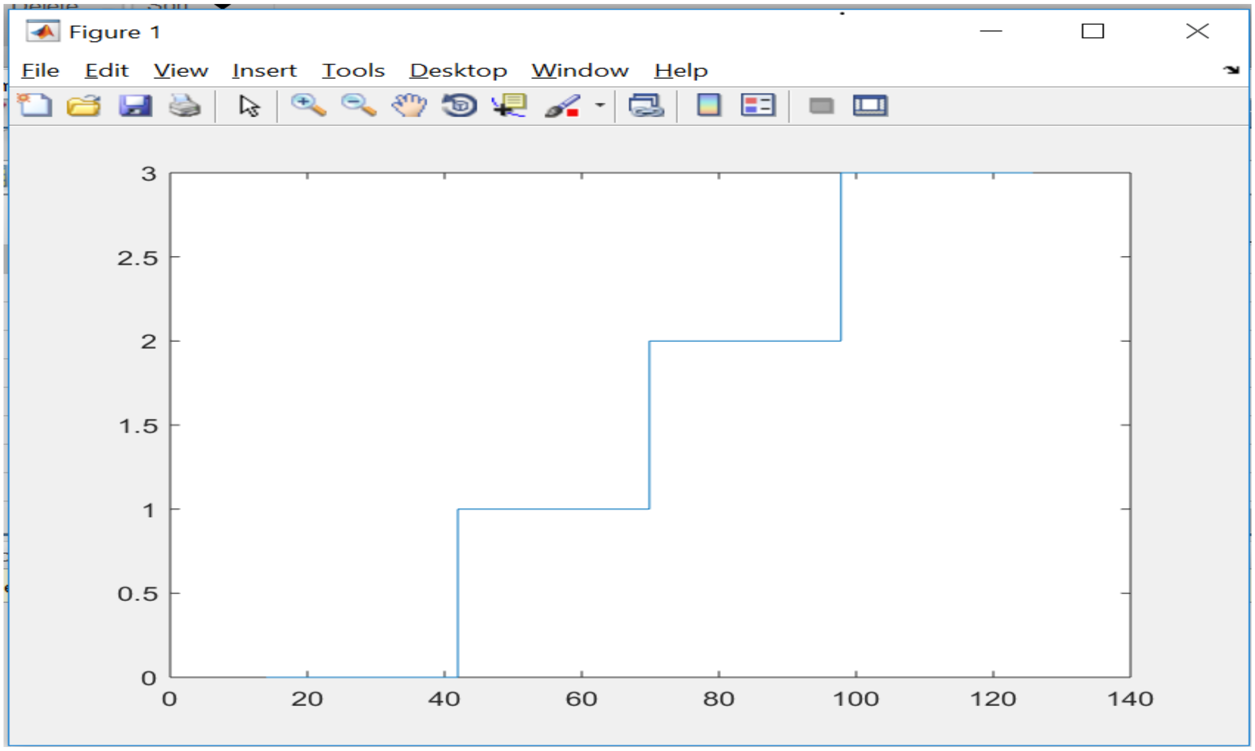
Plotting the Values



3. ‘lumbar_lordosis_angle’ Boundaries:

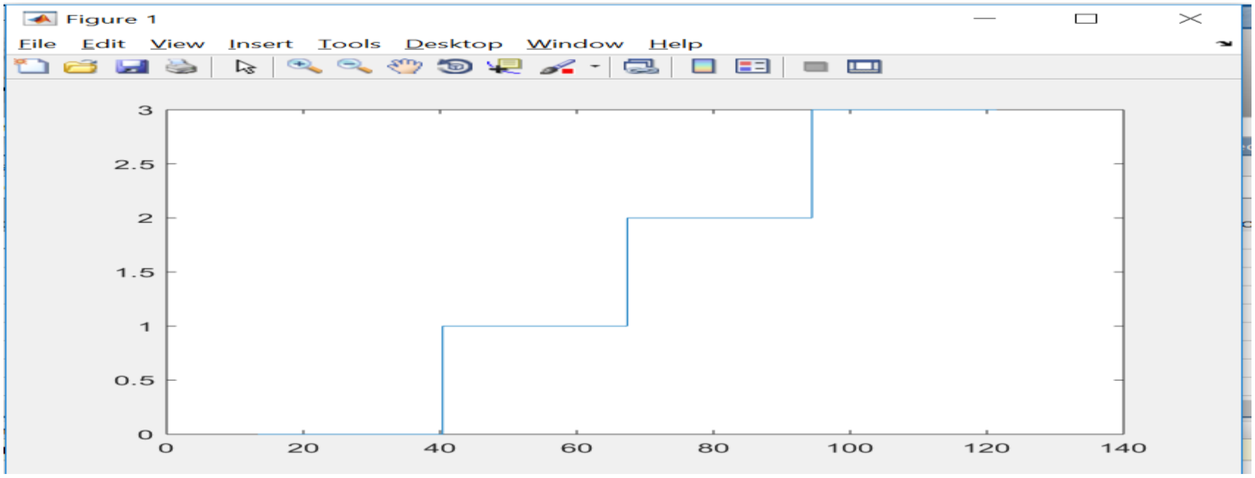
binranges3=[14,41.935,69.87,97.805,125.745]

Samples= {0,1,2,3}
Plotting the Values:



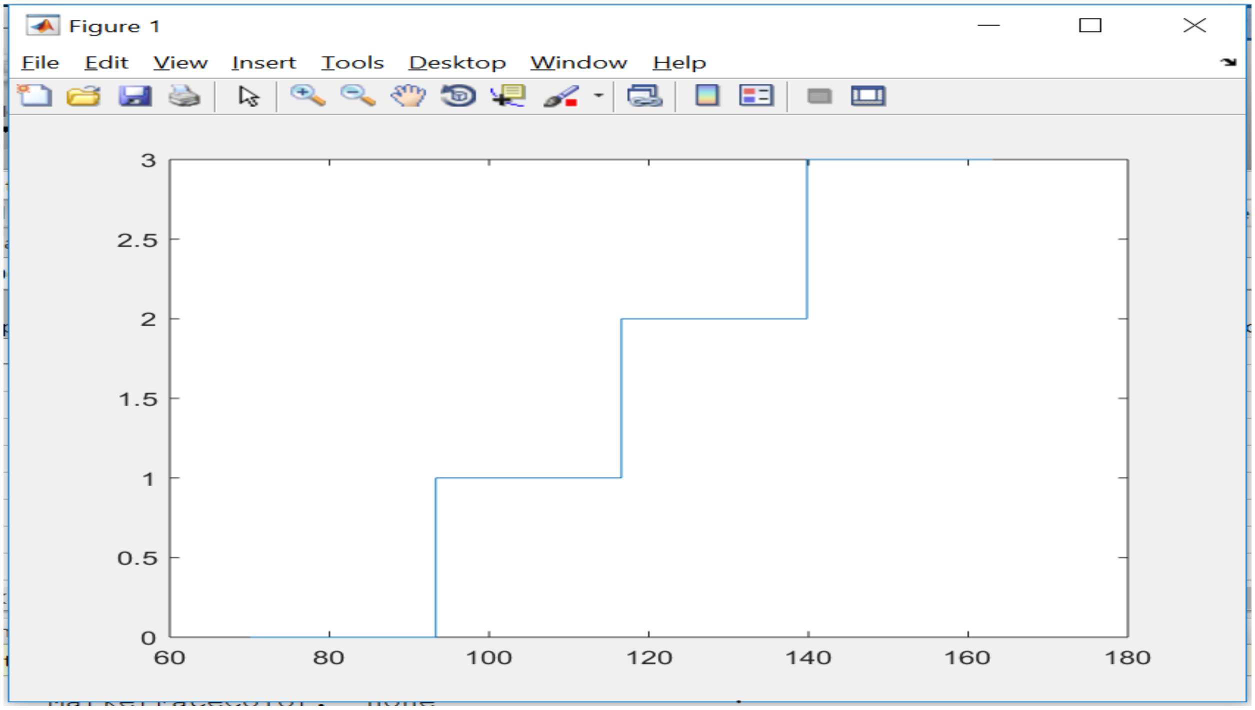
4. ’sacral_slope’ Boundaries:

binranges4=[13.36,40.375,67.39,94.405,121.43]
Samples= {0,1,2,3}
Plotting the Values:



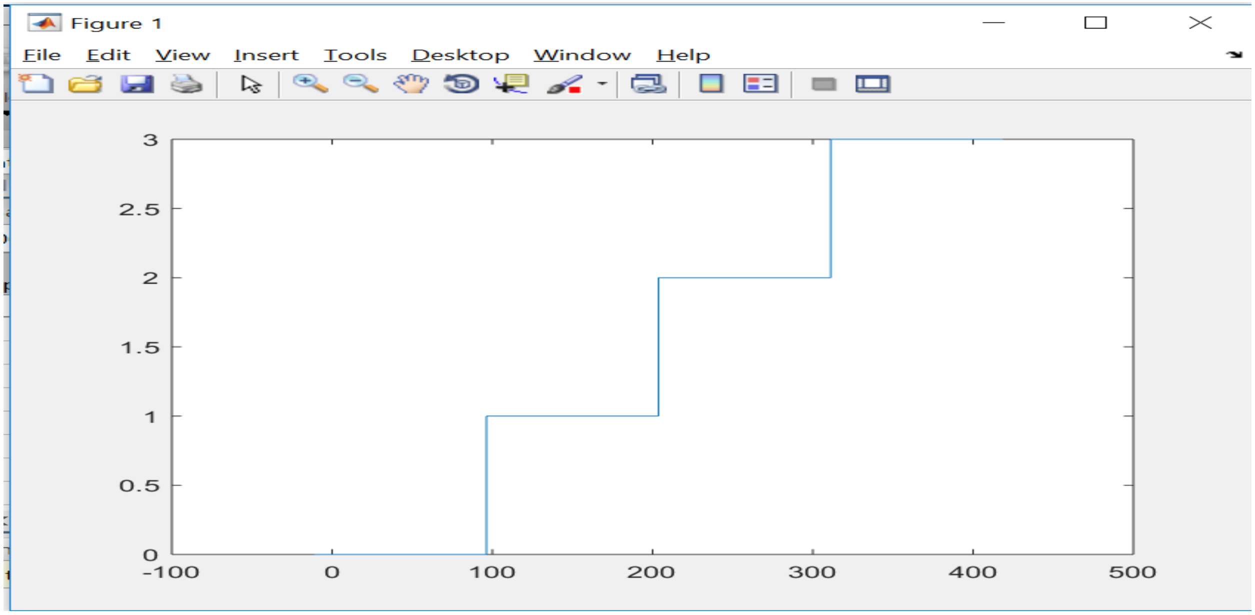
5. ‘pelvic_radius’ Boundaries:

binranges5=[70.08,93.32,116.57,139.82,163.08]
Samples= {0,1,2,3}
Plotting the Values:



6. 'degree_spondylolisthesis' Boundaries:

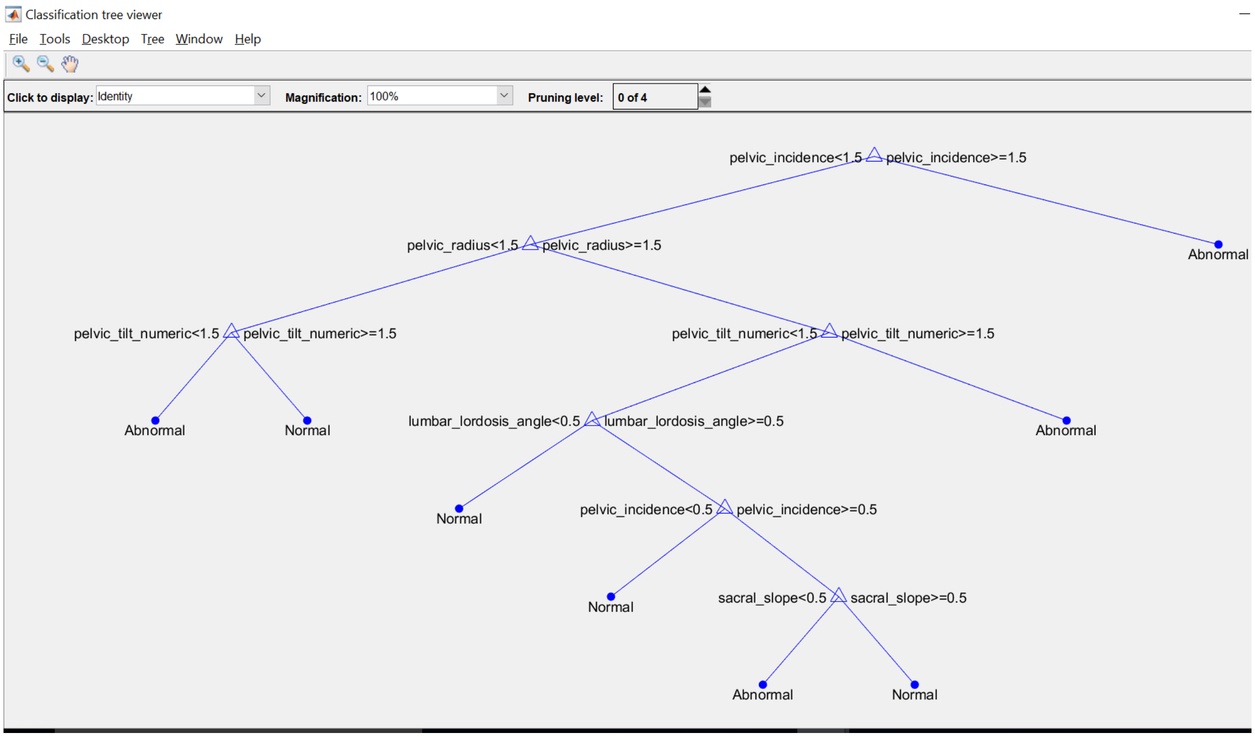
binranges6=[-11.05,96.34,203.73,311.12,418.55]
Samples= {0,1,2,3}
Plotting the Values:



3.b. MATLAB CODE FOR LEARNING AND PLOTTING DECISION TREES:

```
SampledDataTrain=SampledData(51:260,:);
SampledDataTest=SampledData([1:50 261:310],:);

SampledTree3=fitctree(SampledDataTrain,'class','MinLeafSize',3);
view(SampledTree3,'Mode','graph')

SampledTree8=fitctree(SampledDataTrain,'class','MinLeafSize',8);
view(SampledTree8,'Mode','graph')

SampledTree12=fitctree(SampledDataTrain,'class','MinLeafSize',12);
view(SampledTree12,'Mode','graph')

SampledTree30=fitctree(SampledDataTrain,'class','MinLeafSize',30);
view(SampledTree30,'Mode','graph')

SampledTree50=fitctree(SampledDataTrain,'class','MinLeafSize',50);
view(SampledTree50,'Mode','graph')
```

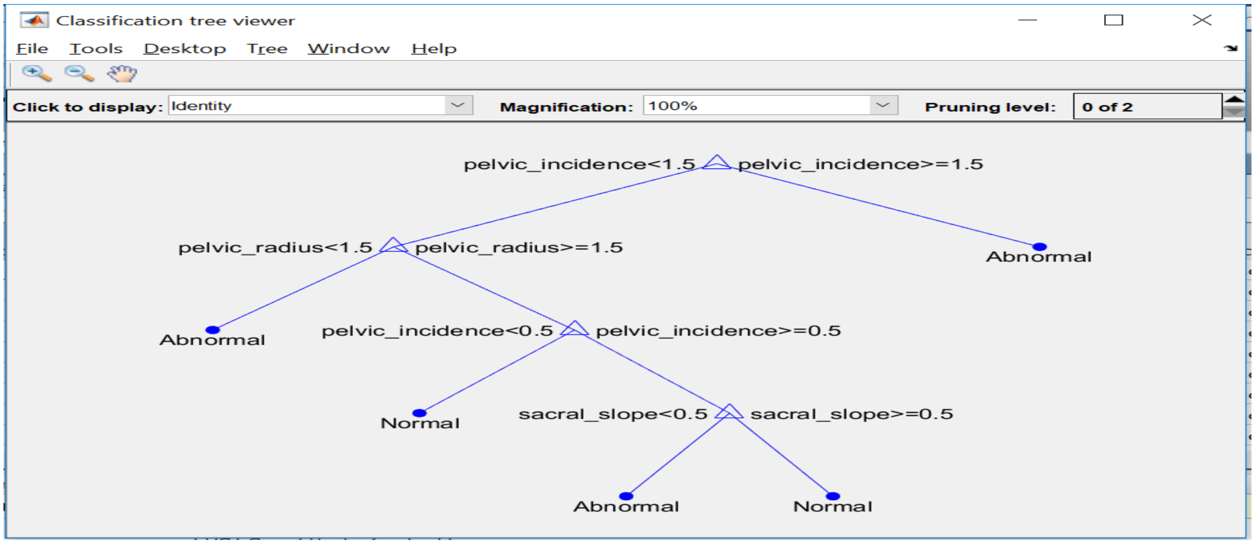
DECISION TREE FOR MIN RECORD VALUE=3:



Click to display: Identity
Magnification: 100%
Pruning level: 0 of 4

The Decision Tree for the Sampled data for min Records per leaf is 3 is big, Since the min Records per leaf are lower the splitting Conditions are Higher .

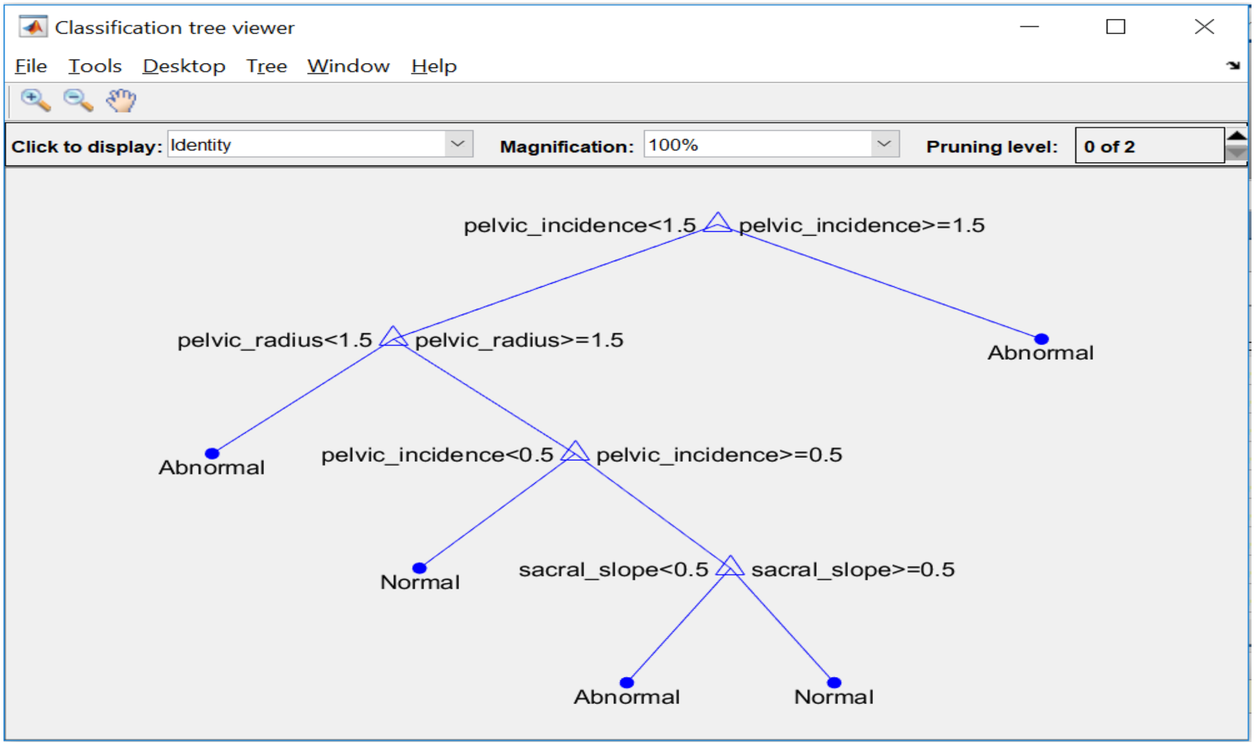
DECISION TREE FOR MIN RECORD VALUE=8:



Click to display: Identity
Magnification: 100%
Pruning level: 0 of 2

The Decision Tree for Tree with min Records per Tree equals 8 has reduced Nodes compared to the one with 3 as the splitting Stopped when the Records are going less than 8

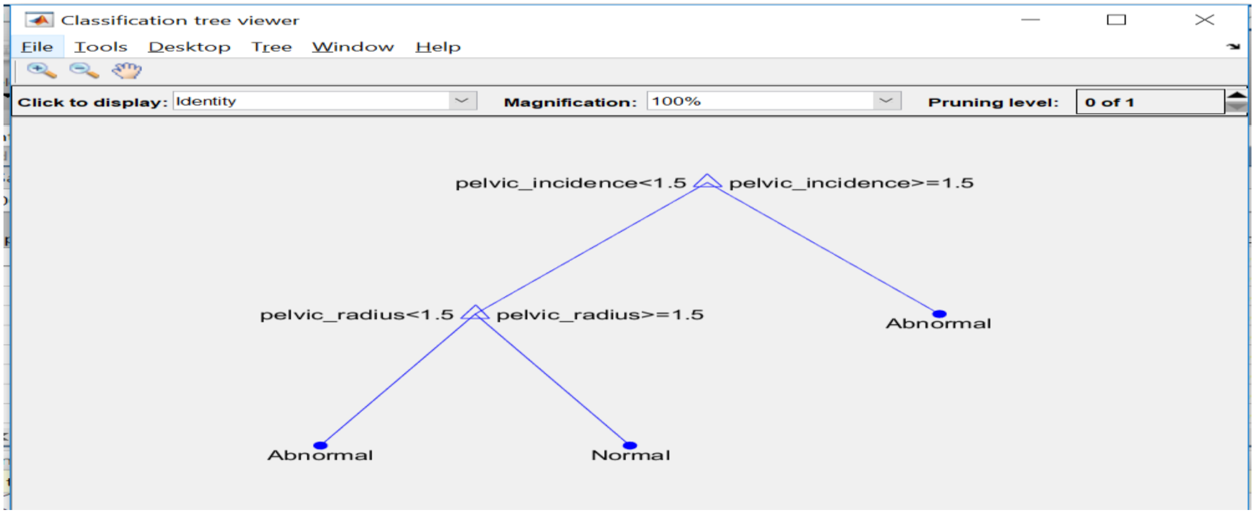
DECISION TREE FOR MIN RECORD VALUE=12:



Click to display: Identity
Magnification: 100%
Pruning level: 0 of 2

The Decision Tree for the Tree with min Records per Leaf are similar to the one with men Records per Leaf=8 based on the Sampled Data.

DECISION TREE FOR MIN RECORD VALUE=30:



Classification tree viewer

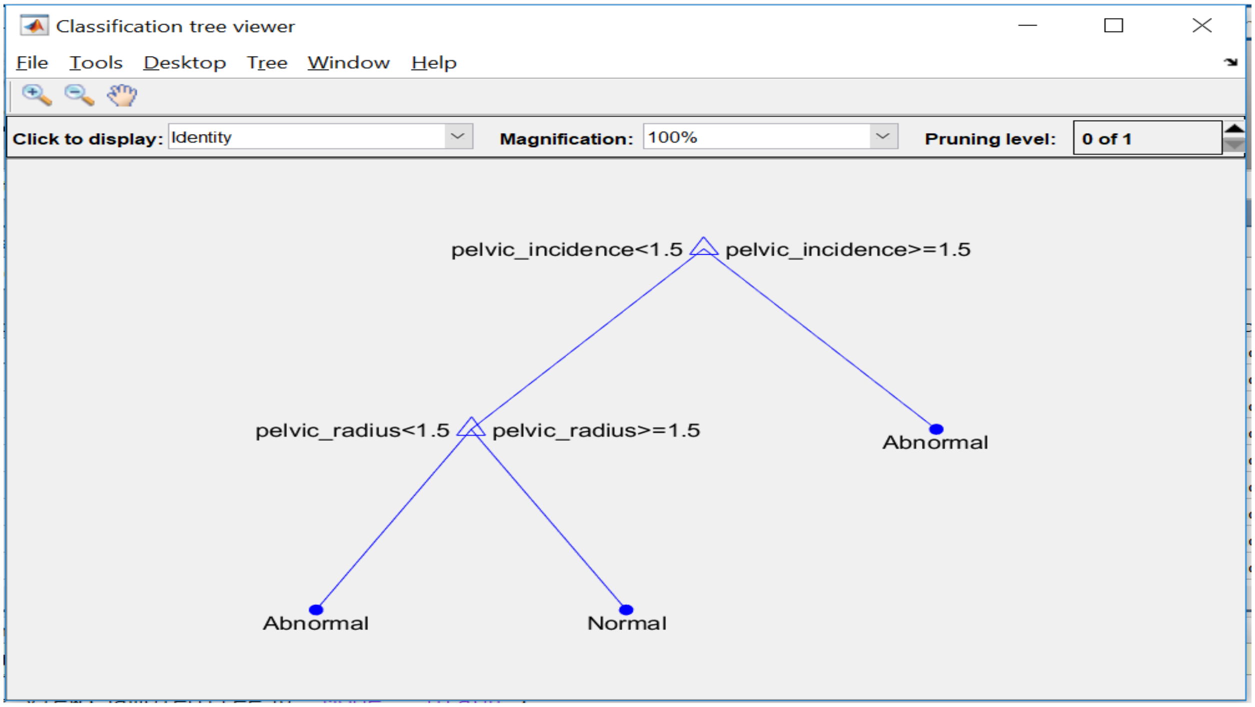
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 1

The Decision Tree for min Record Value 30 has Deduced to 2 condition Tree based on the Sampled Data.

DECISION TREE FOR MIN RECORD VALUE=50:



Classification tree viewer

Click to display: Identity

Magnification: 100%

Pruning level: 0 of 1

The Decision Tree for a Tree with min Records per Leaf Value being 50 is like the one with 30 and is a 2 conditioned Tree with Underfitting Properties.

COMPARISON OF FIVE DECISION TREES:

From the above 5 Trees, as the Min Leaf Node Size Increases, the Tree Size decreases because of the limit in splitting leaf nodes are increasing.

COMMENTS ON TREE PREFERENCE AND EXPLANATION:

The Trees with min Records per Leaf 3,8,12,30 and 50 Records per Leaf have the Child Nodes as 8,5,5,3,3 respectively and we observe the Following Parameters from the Decision Trees.

1. The Trees with min Records per Leaf 30 and 50 cannot be selected as the Trees have a very minimum Number of Training Conditions and might have Insufficient Example to predict the Training Model.
2. The Trees with min Records per Leaf 3 cannot be Used as the Number of Child Nodes being 8 which is Higher and may contain Noise.
3. Hence the Trees with min Records per Leaf as 8 and 12 can be preferred. But when selecting the Best Tree from both these Trees is not possible by looking at the trees, as both the Trees are exactly similar. Hence, we look at Accuracy, Precision and Recall Values.
4. By Looking at the Accuracy, Precision and Recall Values, the Tree 8 has Higher Accuracy, Precision and Recall Values. Hence the Tree 8 is Preferred of the 5 Trees.

- **Matlab Code for Calculation of Accuracy, Precision and Recall Values:**

```
predictionsample3=predict(SampledTree3,SampledDataTest);
SampledDataTestcell=table2cell(SampledDataTest);
Sampledcell=SampledDataTestcell(:,[7]);
Csample3=confusionmat(Sampledcell,predictionsample3);
predictionsample8=predict(SampledTree8,SampledDataTest);
Csample8=confusionmat(Sampledcell,predictionsample8);
predictionsample12=predict(SampledTree12,SampledDataTest);
Csample12=confusionmat(Sampledcell,predictionsample12);
predictionsample30=predict(SampledTree30,SampledDataTest);
Csample30=confusionmat(Sampledcell,predictionsample30);
predictionsample50=predict(SampledTree50,SampledDataTest);
Csample50=confusionmat(Sampledcell,predictionsample50);

Accuracysample3=(Csample3(1,1)+Csample3(2,2))/(Csample3(1,1)+Csample3(1,2)+Csample3(2,1)+Csample3(2,2));
Accuracysample8=(Csample8(1,1)+Csample8(2,2))/(Csample8(1,1)+Csample8(1,2)+Csample8(2,1)+Csample8(2,2));
Accuracysample12=(Csample12(1,1)+Csample12(2,2))/(Csample12(1,1)+Csample12(1,2)+Csample12(2,1)+Csample12(2,2));
Accuracysample30=(Csample30(1,1)+Csample30(2,2))/(Csample30(1,1)+Csample30(1,2)+Csample30(2,1)+Csample30(2,2));
Accuracysample50=(Csample50(1,1)+Csample50(2,2))/(Csample50(1,1)+Csample50(1,2)+Csample50(2,1)+Csample50(2,2));

Precisionsample3=(Csample3(1,1))/(Csample3(1,1)+Csample3(2,1));
Precisionsample8=(Csample8(1,1))/(Csample8(1,1)+Csample8(2,1));
Precisionsample12=(Csample12(1,1))/(Csample12(1,1)+Csample12(2,1));
Precisionsample30=(Csample30(1,1))/(Csample30(1,1)+Csample30(2,1));
```

```
Precisionsample50=(Csample50(1,1))/(Csample50(1,1)+Csample50(2,1));

Precisionsample3normal=(Csample3(2,2))/(Csample3(2,2)+Csample3(1,2));
Precisionsample8normal=(Csample8(2,2))/(Csample8(2,2)+Csample8(1,2));
Precisionsample12normal=(Csample12(2,2))/(Csample12(2,2)+Csample12(1,2));
Precisionsample30normal=(Csample30(2,2))/(Csample30(2,2)+Csample30(1,2));
Precisionsample50normal=(Csample50(2,2))/(Csample50(2,2)+Csample50(1,2));

Recallsample3=(Csample3(1,1))/(Csample3(1,1)+Csample3(1,2));
Recallsample8=(Csample8(1,1))/(Csample8(1,1)+Csample8(1,2));
Recallsample12=(Csample12(1,1))/(Csample12(1,1)+Csample12(1,2));
Recallsample30=(Csample30(1,1))/(Csample30(1,1)+Csample30(1,2));
Recallsample50=(Csample50(1,1))/(Csample50(1,1)+Csample50(1,2));

Recallsample3normal=(Csample3(2,2))/(Csample3(2,2)+Csample3(2,1));
Recallsample8normal=(Csample8(2,2))/(Csample8(2,2)+Csample8(2,1));
Recallsample12normal=(Csample12(2,2))/(Csample12(2,2)+Csample12(2,1));
Recallsample30normal=(Csample30(2,2))/(Csample30(2,2)+Csample30(2,1));
Recallsample50normal=(Csample50(2,2))/(Csample50(2,2)+Csample50(2,1));

AccuracyPlotsample=plot([3,8,12,30,50],[Accuracysample3 Accuracysample8 Accuracysample12
Accuracysample30 Accuracysample50])
PrecisionPlotsample=plot([3,8,12,30,50],[Precisionsample3 Precisionsample8 Precisionsample12
Precisionsample30 Precisionsample50])
RecallPlotsample=plot([3,8,12,30,50],[Recallsample3 Recallsample8 Recallsample12 Recallsample30
Recallsample50])
PrecisionPlotsamplenormal=plot([3,8,12,30,50],[Precisionsample3normal Precisionsample8normal
Precisionsample12normal Precisionsample30normal Precisionsample50normal])
RecallPlotsamplenormal=plot([3,8,12,30,50],[Recallsample3normal Recallsample8normal
Recallsample12normal Recallsample30normal Recallsample50normal])
```

ACCURACY VALUES FOR FIVE DECISION TREES:

1.Accuracy for Decision Tree with min Leaf Node Size=3 is 0.6100

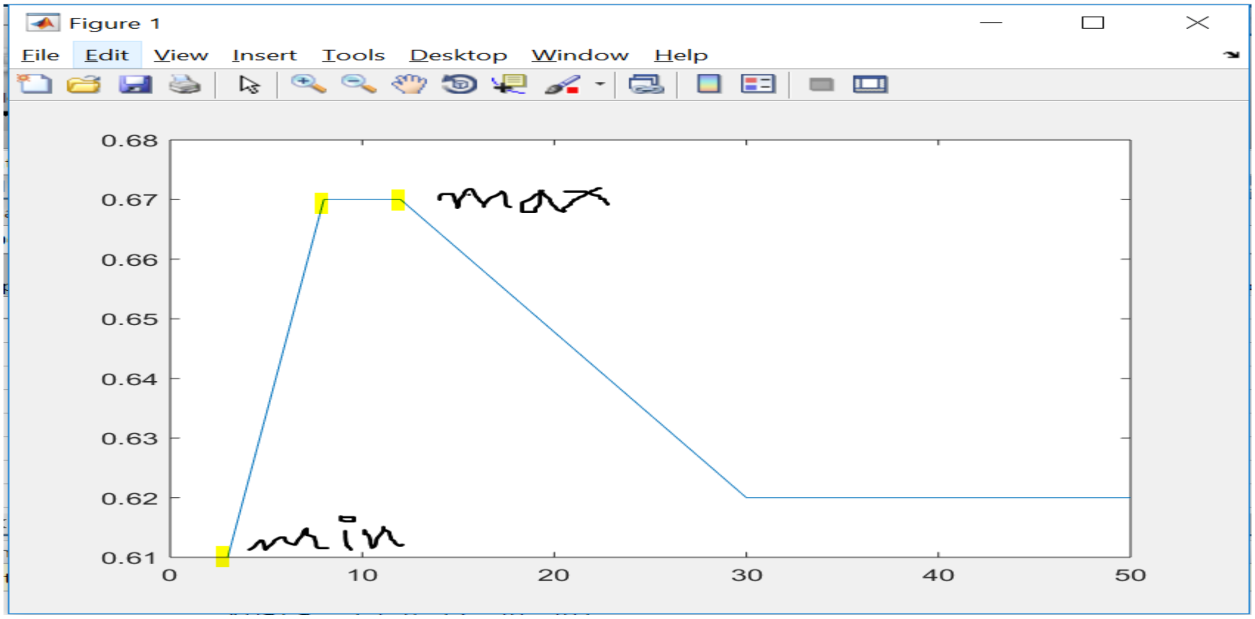
2. Accuracy for Decision Tree with min Leaf Node Size=8 is 0.6700

3. Accuracy for Decision Tree with min Leaf Node Size=12 is 0.6700

4. Accuracy for Decision Tree with min Leaf Node Size=30 is 0.6200

5. Accuracy for Decision Tree with min Leaf Node Size=50 is 0.6200

ACCURACY PLOT:



The Accuracy Plot for Sampled Data is Highest at min Leaf Node Size Values 8 and 12,The Reason could be because the Tree Size is more compared to min Leaf Sizes 30 and 50 and also the Noise is Lesser compared to the Tree Size with min Leaf Size 3.

PRECISION VALUES FOR FIVE DECISION TREES:

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.6571

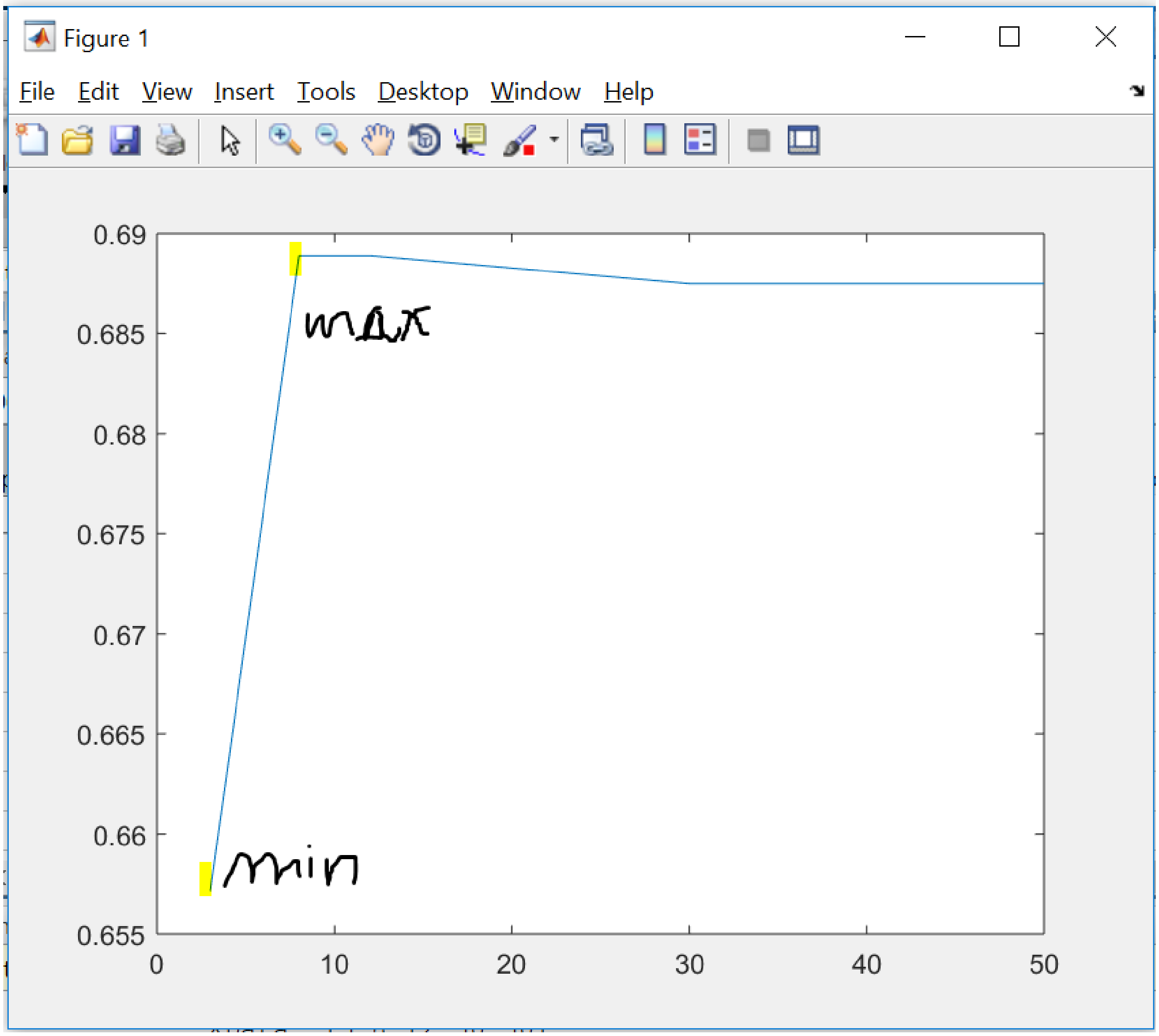
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.6889

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.6889

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.6875

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.6875

PRECISION PLOT:



The Precision Value is Highest at min Leaf Size Value 8,because of Higher Tree Size and more Training Examples used and lesser Noise.

PRECISION VALUES FOR FIVE DECISION TREES(CLASS='Normal'):

1.Precision for Decision Tree with min Leaf Node Size=3 is 0.5846

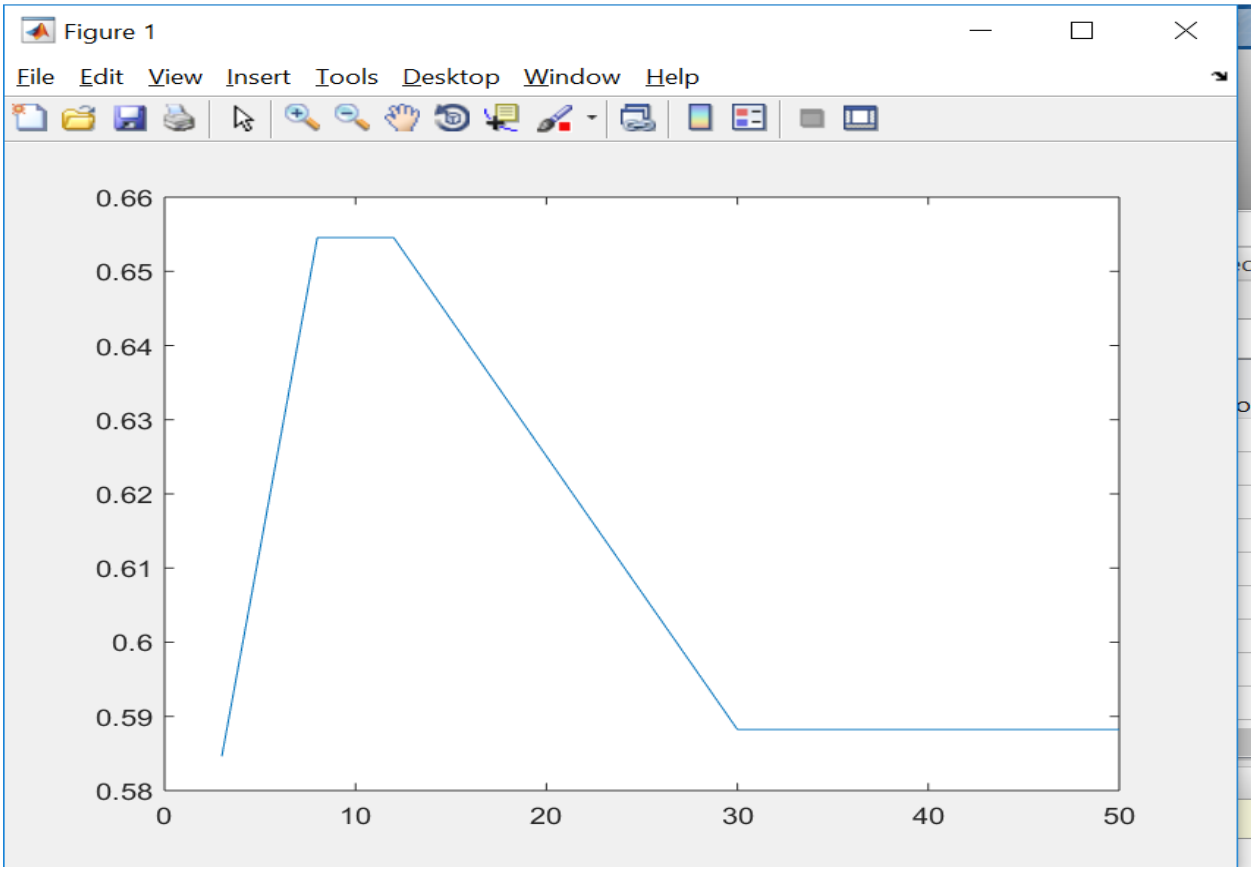
2. Precision for Decision Tree with min Leaf Node Size=8 is 0.6545

3. Precision for Decision Tree with min Leaf Node Size=12 is 0.6545

4. Precision for Decision Tree with min Leaf Node Size=30 is 0.5882

5. Precision for Decision Tree with min Leaf Node Size=50 is 0.5882

PRECISION PLOT:



The Precision Plot is maximum at min Leaf Size Values 8,12 because of Intermediate Tree Sizes and Lesser Noise Parameter. Minimum at Leaf Sizes 30,50 due to very Lower Tree Sizes.

RECALL VALUES FOR FIVE DECISION TREES:

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.4600

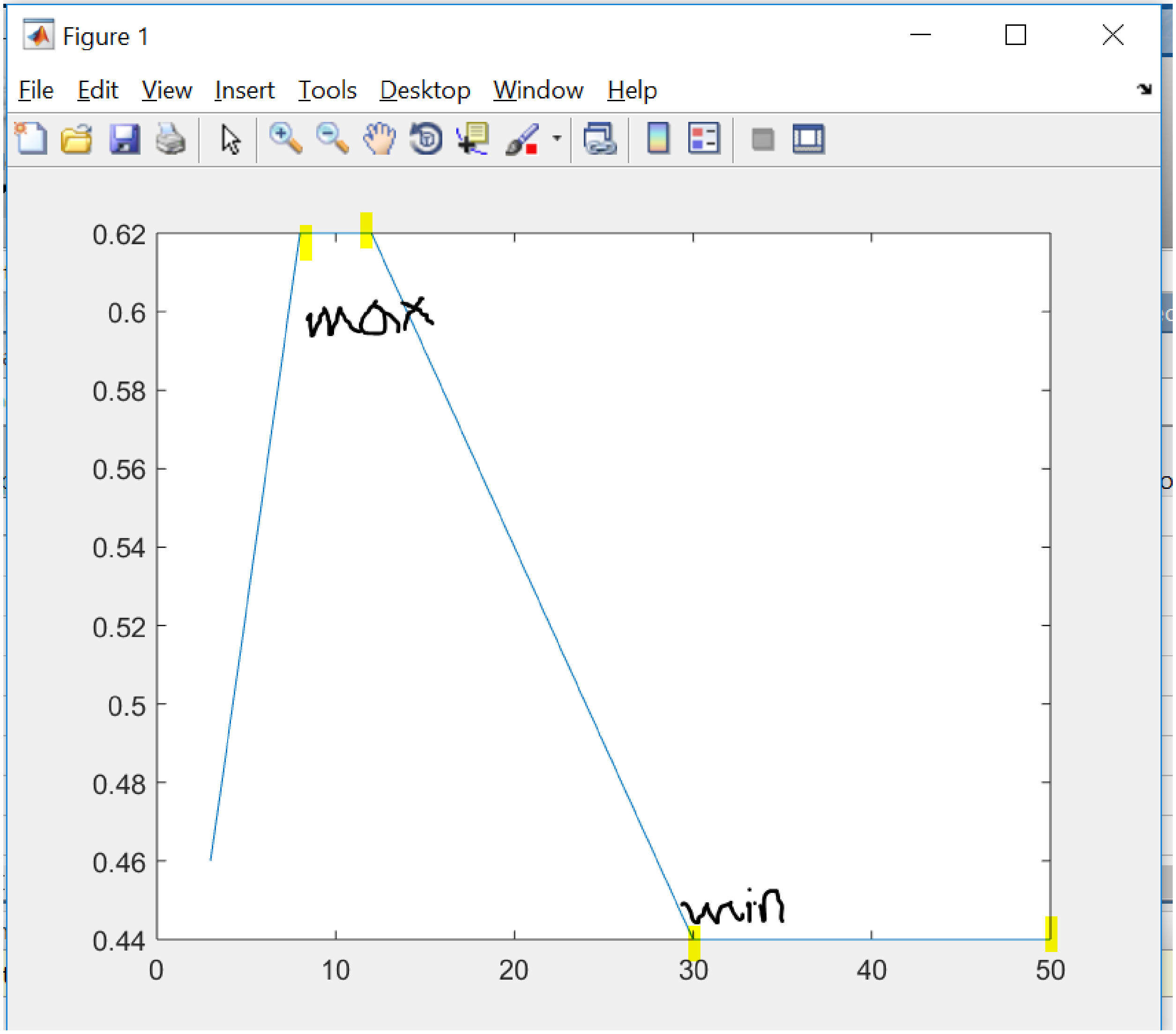
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.6200

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.6200

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.4400

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.4400

RECALL PLOT:



The Recall Values are higher at min Leaf Sizes 8 and 12 due to the Training Tree has sufficient Examples to train and also the Size is not too High and hence the Noise could be lesser.

RECALL VALUES FOR FIVE DECISION TREES(Class='Normal'):

1.Recall for Decision Tree with min Leaf Node Size=3 is 0.7600

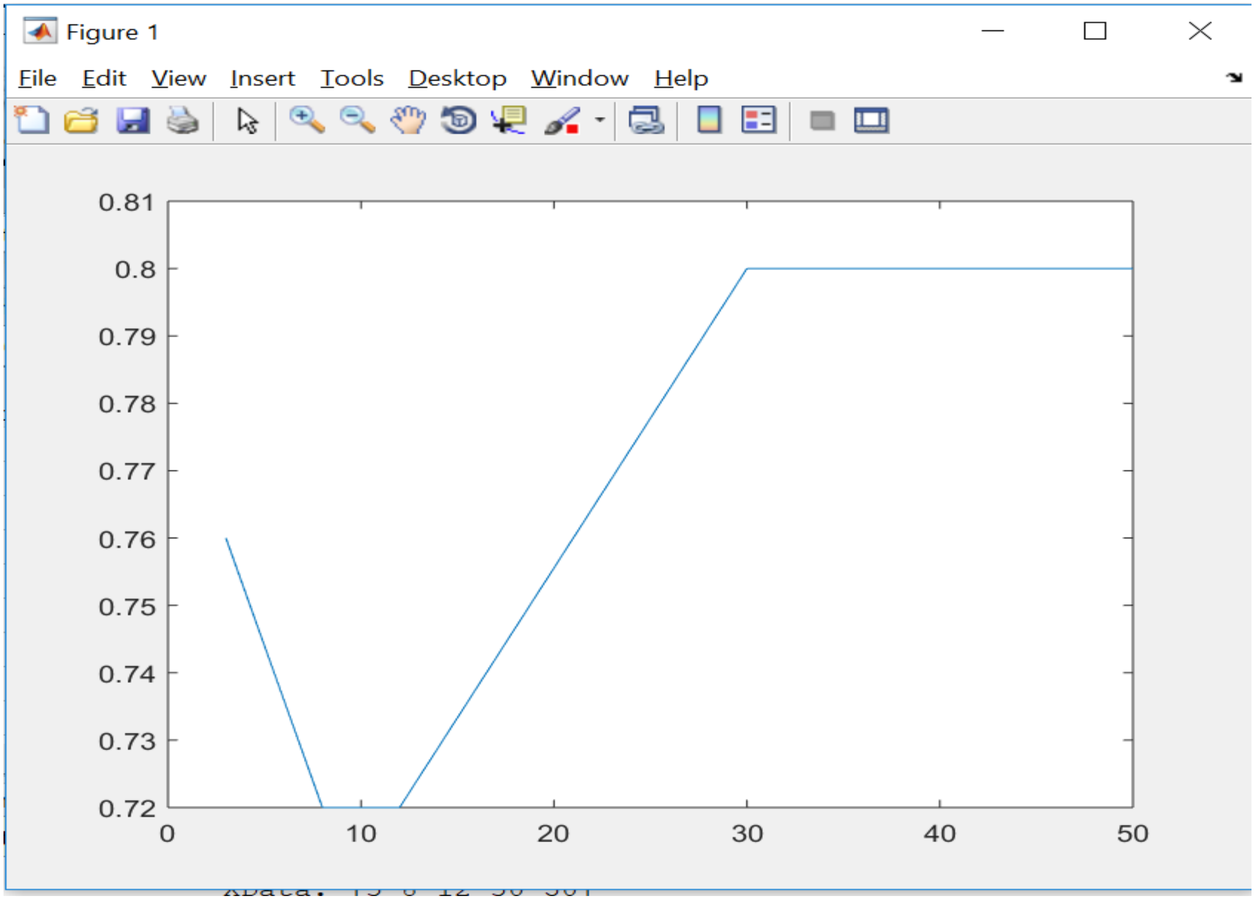
2. Recall for Decision Tree with min Leaf Node Size=8 is 0.7200

3. Recall for Decision Tree with min Leaf Node Size=12 is 0.7200

4. Recall for Decision Tree with min Leaf Node Size=30 is 0.8000

5. Recall for Decision Tree with min Leaf Node Size=50 is 0.8000

RECALL PLOT:



The Recall Plot for Decision Tree with Class Attribute ’Normal’ is High at Lesser Tree Sizes. This could be because of the Sampling Issues and Errors that we Expect in the Data converted to bins.

COMMENTS ON ACCURACY, PRECISION AND RECALL VALUES:

1. From the Accuracy Plot, the Trees with min records per Leaf 8 and 12 are having the Highest Accuracy which is 0.67 and lowest being 0.62 for the Trees with min Records per Leaf as 30 and 50. The Reason for this is clear. Trees with min Records per Leaf 30 and 50 are affected by Underfitting and the Training Examples are not enough. But the Trees with Records 8 and 12 have intermediate Node Values and Conditions to make the Accurate Prediction.

- ➢ From the Accuracy Plot, the Tree with min Records per Leaf 8 and 12 are preferred, they are more Accurate compared to the other.

2. From the Precision Plot, the Tree with min Records per Leaf 8 is Having the Highest Precision compared to the Other. Precision is the number of True Predicted Values from all the Positives Predicted. Hence, the Tree with min Records per Leaf as 8 is having the best Predictions compared to the others.

- ➢ From the Precision Plot, the Tree with min Records per Leaf as 8 is Preferred among the others.

3.From the Recall Plot, the Tree with min Records per Leaf 8 and 12 are having the Highest Recall which is 0.62 which means that From the Parent Node ,62% of the Records are Truly from the Parent Node. The Recall is Lesser for the Nodes with min Records per Leaf 30 and 50.

> From the Recall Plot, the Preference of Tree would be given to the Trees with minimum Records per Leaf as 8 and 12.

PREFERENCE:

Based on Accuracy, Precision and Recall Values, the Tree with min Records per Leaf 8 is having the Highest Values .Also considering Number of Nodes per Tree and Underfitting, Overfitting concepts ,the Tree with min Records per Leaf 0f 8 is the Preferable Tree.

3.C.COMPARISON OF DECISION TREES BUILT USING REAL AND SAMPLE DATA:

> The Accuracy, Precision and Recall Values for the Real Data for the Preferred Tree(8) are 0.6100,0.6486 and 0.4800.

> But the Accuracy, Precision and Recall Values for Sampled Data are 0.6700,0.6889,0.6200 .

> All the 3 parameters have been Enhanced by Sampling the Data. Hence the Performance of the Sampled Decision is as good as Normal Data.

> Hence Bin Data can be Used Effectively for Data Computation since it is Effective and Easy to compute. The Time Taken for Computation for Bin Data is less compared to Real Data.