Biomechanical Health Assessment

Hsin-Ting Chou
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
and
Department of Computer Science
University of Cincinnati
Ohio, United State
justin0010523@gmail.com

***Abstract*—The data of biomechanical features are taken from the UCI Machine Learning Repository and were collected from some patients. The goal of this assignment is to predict if the patient is normal or not(abnormal). Using the six features in the excel file, I trained the model with decision tree. Finally, the health conditions were able to separate into two and three different conditions, corresponding to the two different files by decision tree.**

Index Terms—biomechanical, UCI Machine Learning, decision tree

1. INTRODUCTION

The first task of this assignment is to analyze the biomechanical health condition and train a model with decision trees by 230 training instances which are randomly split and selected in Data2. Then, predict the health condition by the rest of 80 testing instances. The minimum records per leaf node in decision trees should be 5, 15, 25, 40 and 50. So there will have 5 trees in this task. Compute and report the accuracy, precision (for each class), and recall (for each class) values for each of the five decision trees. Next, compare and give comment on these value and show these values on a plot. Explain for the observed trends or differences.

The second task is to repeat the first task with Data3 (There are three classes to deal with now). Additionally, reporting results for first task and comment on the comparison of results obtained for first task and this task.

The third task is also same as the first task, however, in Data2, “Abnormal” and “Normal” should be denoted as “1” and “0” class label. Then, due to pelvic incidence, pelvic tilt numeric, lumbar lordosis angle, sacral slope, pelvic radius, and degree spondylolisthesis which are record in the excel file, each feature has distinct correlation to the health condition. Calculate and report the correlations between each feature and the class label column. Drop the feature which correlated to the class label the most. Finally, repeat first task’s steps.

This report is organized as follows. Section 2 mentioned the first task. Section 3 mentioned the second task. Section 4 mentioned the third task. Section 5 summarized the conclusions

2. THE FIRST TASK

2.1. First task a.

According to different number of minimum records per leaf node, the corresponding diagrams are totally distinct with each other. The number of leaf nodes, depth of trees decreased sharply while the minimum records increased. In my opinion, due to the fixed number of all data, once each of the leaf node can record more information, decision trees do not need that much node to make decisions. That why the number of leaf nodes and depth of trees greatly reduce.

Base on different conditions, if now dataset have lots of features to analyze, a better choice is to decrease the number of minimum records to let decision tree make more decision. More decision can split data into more nodes which can easily figure out why data were separated to each node which means we have more recognize if each feature is useful or not.

On the contrary, if now the goal is to analyze data faster, a better choice is to increase the number of minimum records, thus, the number of leaf nodes, depth of trees, and processing time would reduce. Yet, to maintain the accuracy at the same time, minimum records should not be increased too much, so the speed and accuracy can both taken into consideration. The diagram of each minimum records per leaf node values is shown in Fig 1. to Fig 5.

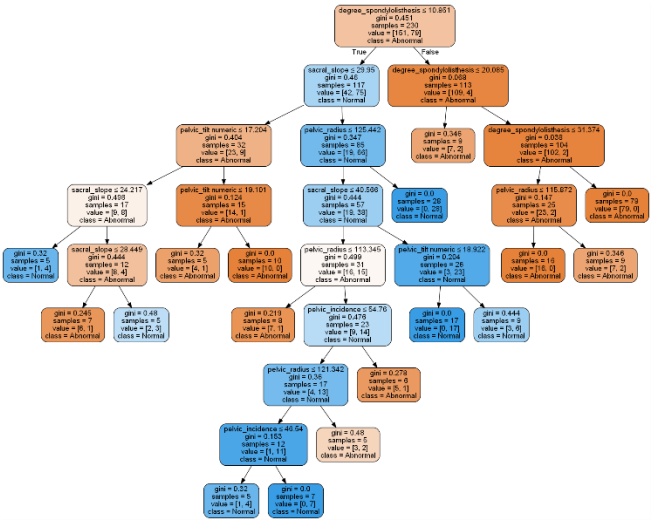

Fig 1. The diagram of “minimum records per leaf node” value of 5

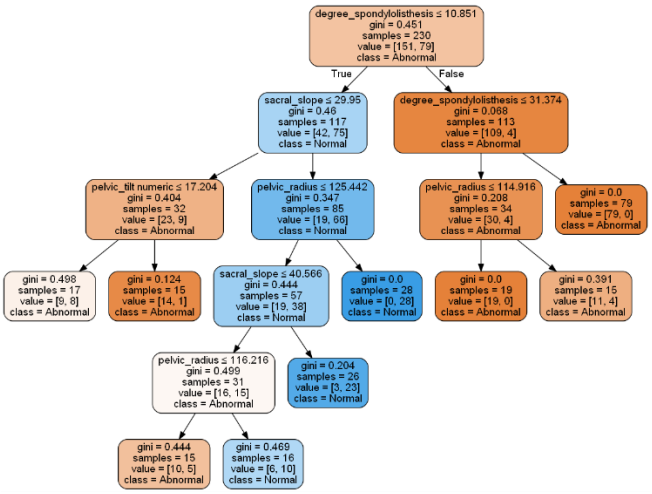

Fig 2. The diagram of “minimum records per leaf node” value of 15

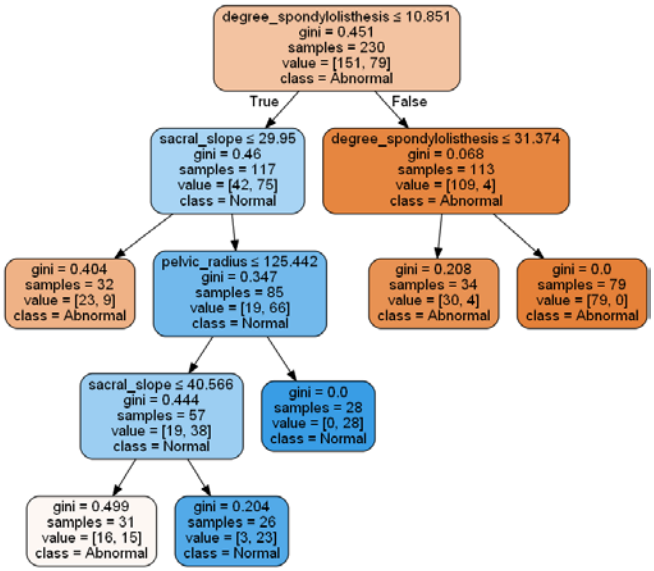

Fig 3. The diagram of “minimum records per leaf node” value of 25

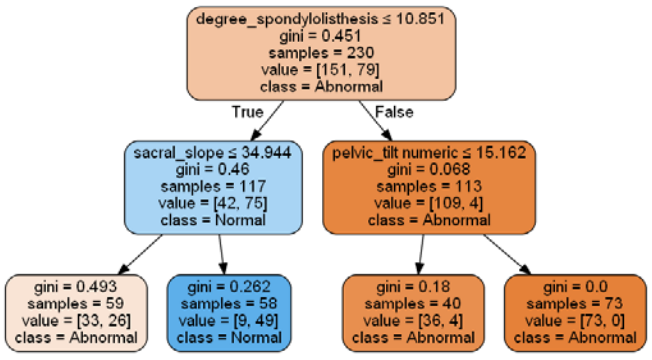

Fig 4. The diagram of “minimum records per leaf node” value of 40

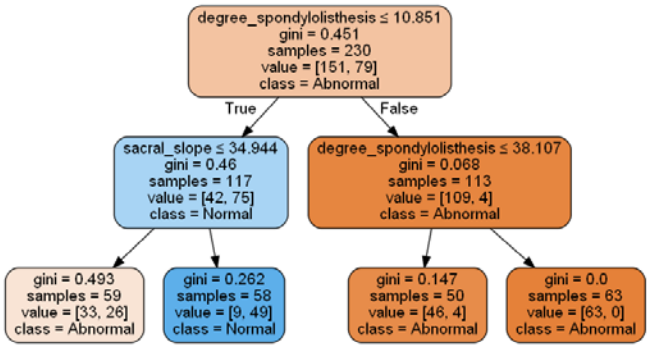

Fig 5. The diagram of "minimum records per leaf node" value of 50

2.2. First task b.

There are four types of conditions in decision tree which includes "True Positive", "True Negative", "False Positive", and "False Negative". They corresponding to the prediction which is true while the actuality is really true, the prediction which is false while the actuality is really false, the prediction which is true while the actuality is actually false, and the prediction which is false while the actuality is actually true. The formula of precision is shown in (1).

Precision means the percentage of the predicted results which are relevant[1].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Recall is the percentage of total relevant correct results predicted by the algorithm[1]. The formula of recall is shown in (2).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

Accuracy refers to the percentage of total correctness of value. The formula of accuracy is shown in (3).

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \quad (3)$$

The precision, recall, and accuracy of each decision tree are shown as follows.

Minimum records per leaf node values of 5	
Precision	Abnormal = 0.875
	Normal = 0.583
Recall	Abnormal = 0.831
	Normal = 0.667
Accuracy	Accuracy = 0.787

Minimum records per leaf node values of 15	
Precision	Abnormal = 0.895
	Normal = 0.652
Recall	Abnormal = 0.864
	Normal = 0.714
Accuracy	Accuracy = 0.825

Minimum records per leaf node values of 25	
Precision	Abnormal = 0.851
	Normal = 0.846
Recall	Abnormal = 0.966
	Normal = 0.524
Accuracy	Accuracy = 0.85

Minimum records per leaf node values of 40	
Precision	Abnormal = 0.815
	Normal = 0.6
Recall	Abnormal = 0.898
	Normal = 0.429
Accuracy	Accuracy = 0.775

Minimum records per leaf node values of 50	
Precision	Abnormal = 0.815
	Normal = 0.6
Recall	Abnormal = 0.898
	Normal = 0.429
Accuracy	Accuracy = 0.775

As shown in Fig 6, blue bar represents the value of abnormal precision while orange bar is the value of abnormal recall, green bar stands for the value of normal precision while red bar is the value of normal recall, and purple bar shows the value of accuracy.

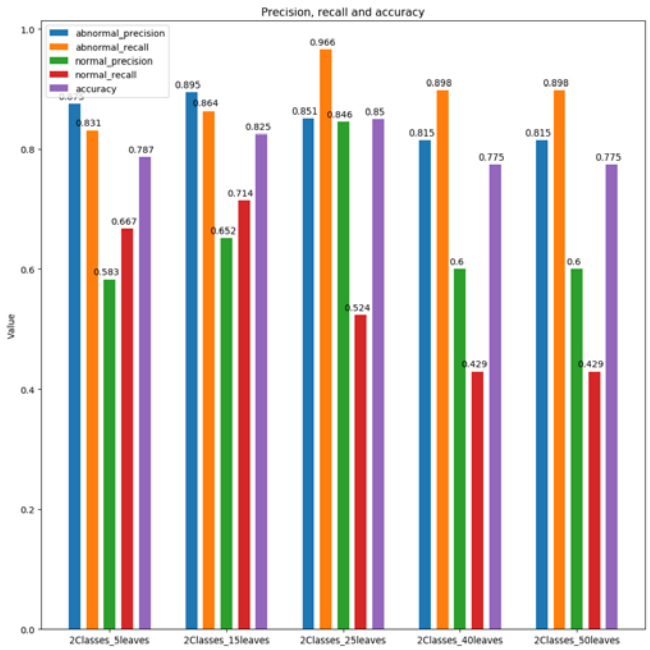

Fig 6. The diagram of each minimum record performance comparison

The confusion matrix of every minimum records is shown as below.

	Predicted abnormal	Predicted normal
Abnormal	49	10
Normal	7	14

Fig 7. The confusion matrix which “minimum records per leaf node” value of 5

	Predicted abnormal	Predicted normal
Abnormal	51	8
Normal	6	15

Fig 8. The confusion matrix which “minimum records per leaf node” value of 15

	Predicted abnormal	Predicted normal
Abnormal	57	2
Normal	10	11

Fig 9. The confusion matrix which “minimum records per leaf node” value of 25

	Predicted abnormal	Predicted normal
Abnormal	53	6
Normal	12	9

Fig 10. The confusion matrix which “minimum records per leaf node” value of 40

	Predicted abnormal	Predicted normal
Abnormal	53	6
Normal	12	9

Fig 11. The confusion matrix which “minimum records per leaf node” value of 50

The value of abnormal precision goes down which represents when the value of “True Positive” is fixed, the value of “True Positive” plus “False Positive” might increase. In contrast, if this value is fixed, the value of “True Positive” is decreasing.

In this case, abnormal refers to “True” and normal represent “False”. According to the above conditions, “True Positive” is the number which is predicted as abnormal and the actuality is abnormal. “True Negative” is the number which is predicted as normal and the actuality is normal. “False Positive” is the number which is predicted as abnormal and the actuality is normal. “False Negative” is the number which is predicted as normal and the actuality is abnormal. Due to these conditions, the value of which is predicted as abnormal and the actuality is abnormal plus which is predicted as abnormal and the actuality is normal might increase. Furthermore, the value really increased.

Next, the value of normal recall goes down which means when the value of “True Positive” is fixed, the value of “True Positive” plus “False Negative” might increase. On the contrary, if this value is fixed, the value of “True Positive” is decreasing.

In this case, normal refers to “True” and abnormal represent “False”. According to the above conditions, “True Positive” is the number which is predicted as normal and the actuality is normal. “True Negative” is the number which is predicted as abnormal and the actuality is abnormal. “False Positive” is the number which is predicted as normal and the actuality is abnormal. “False Negative” is the number which is predicted as abnormal and the actuality is normal. Due to the above conditions, the value of which is predicted as normal and the actuality is normal plus which is predicted as abnormal and the actuality is normal

might increase. However, this value is fixed, in the other words, the value of "True Positive" should decreased. In fact, the value of which is predicted as normal and the actuality is normal is decreasing.

Then, when the minimum records per leaf node value comes to 25, decision tree has the highest accuracy and balanced precision of each class relative to others. If now goes back to ***First task a***, I would choose this decision tree.

3. THE SECOND TASK

3.1. Second task a.

The difference of this task is almost same as ***First task a.*** Yet, there still have something new. Decision trees becoming balance while the minimum records increased in ***First task a.*** However, decision trees are still unbalance in this task. In my opinion, this is because the attribute of class increased to three. The "Abnormal" attribute in the first task is separated into "Hernia" and "Spondylolisthesis". Nodes which predicted as "Abnormal" might also be separated into "Hernia" and "Spondylolisthesis". That might be the reason why decision trees will not be balance even if the minimum records go up.

Due to the previous task, if now I have to choose one decision tree, I might choose the one which minimum records per leaf node value is 25 because the value of accuracy is possibly the highest and the value of precision of each class are probably the most balanced one.

The diagram of each minimum records per leaf node values is shown in Fig 12. to Fig 16.

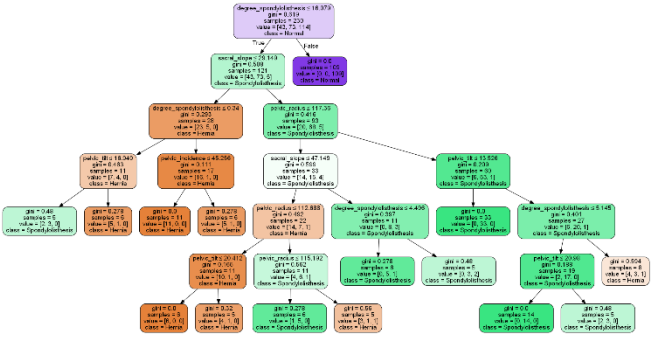

Fig 12. The diagram of "minimum records per leaf node" value of 5

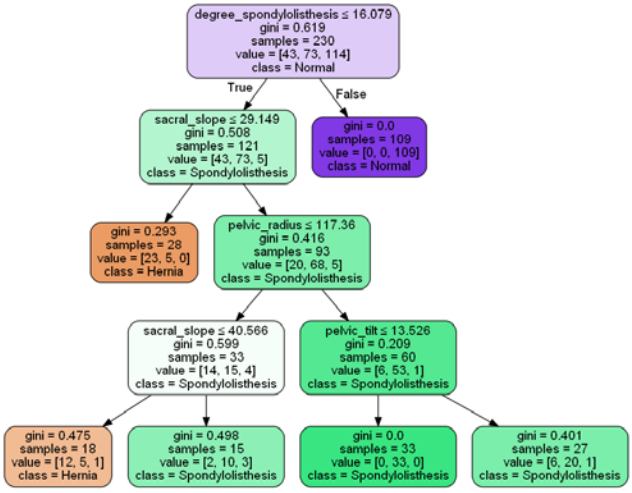

Fig 13. The diagram of "minimum records per leaf node" value of 15

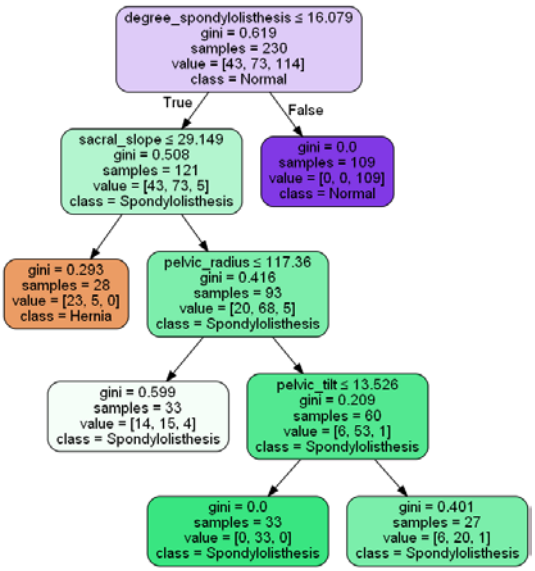

Fig 14. The diagram of "minimum records per leaf node" value of 25

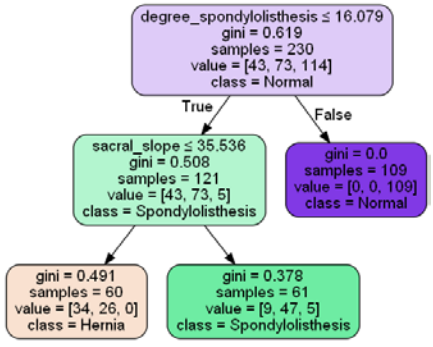

Fig 15. The diagram of "minimum records per leaf node" value of 40

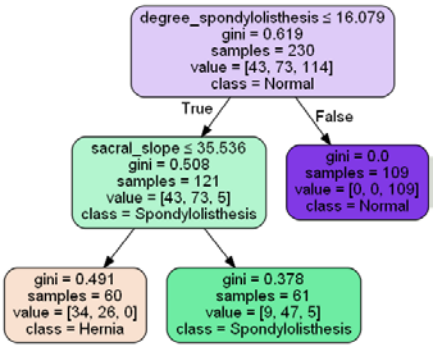

Fig 16. The diagram of "minimum records per leaf node" value of 50

3.2. Second task b.

The precision, recall, and accuracy of each decision tree are shown as follows.

Minimum records per leaf node values of 5	
Precision	Hernia = 0.571
	Spondylolisthesis = 0.923
	Normal = 0.75
Recall	Hernia = 0.706
	Spondylolisthesis = 1.0
	Normal = 0.556
Accuracy	Accuracy = 0.787

Minimum records per leaf node values of 15	
Precision	Hernia = 0.6
	Spondylolisthesis = 0.923
	Normal = 0.762
Recall	Hernia = 0.706
	Spondylolisthesis = 1.0
	Normal = 0.593
Accuracy	Accuracy = 0.8

Minimum records per leaf node values of 25	
Precision	Hernia = 0.417
	Spondylolisthesis = 0.923
	Normal = 0.586
Recall	Hernia = 0.294
	Spondylolisthesis = 1.0
	Normal = 0.63
Accuracy	Accuracy = 0.725

Minimum records per leaf node values of 40	
Precision	Hernia = 0.5
	Spondylolisthesis = 0.923
	Normal = 0.733
Recall	Hernia = 0.765
	Spondylolisthesis = 1.0
	Normal = 0.407
Accuracy	Accuracy = 0.75

Minimum records per leaf node values of 50	
Precision	Hernia = 0.5
	Spondylolisthesis = 0.923
	Normal = 0.733
Recall	Hernia = 0.765
	Spondylolisthesis = 1.0
	Normal = 0.407
Accuracy	Accuracy = 0.75

As shown in Fig 17, blue bar represents the value of hernia precision while orange bar is the value of hernia recall, green bar stands for the value of spondylolisthesis precision while red bar is the value of spondylolisthesis recall, and purple bar shows the value of normal precision while brown bar is the value of normal recall, and pink bar is the value of accuracy.

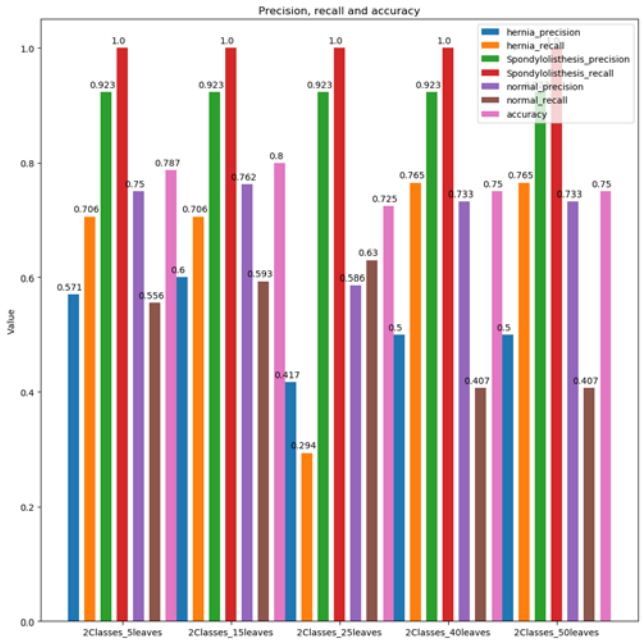

Fig 17. The diagram of each minimum record performance comparison

The value of spondylolisthesis precision and spondylolisthesis recall are fixed. Moreover, the value of spondylolisthesis recall is always one, which means the prediction to spondylolisthesis will never have "False Negative".

In this case, note that the decision tree which minimum records per leaf node value is 25, the

value of accuracy is not the highest and the value of precision of each class are not the most balanced as well. Instead, the one which minimum records per leaf node value is 15 performs the best. It is totally different with ***First task b.*** If now goes back to ***Second task a***, I would choose this decision tree.

4. THE THIRD TASK

4.1. Third task a.

In this task, first denote “Abnormal” and “Normal” as “1” and “0” class labels. Then, to compute the correlations between each feature and the class label column, correlation matrix has been used to achieve the goal. In correlation matrix, degree of spondylolisthesis has the highest magnitude of the correlation value to the labels. Consequently, degree of spondylolisthesis should be dropped from the dataset. The diagram of correlation matrix is shown in Fig 18.

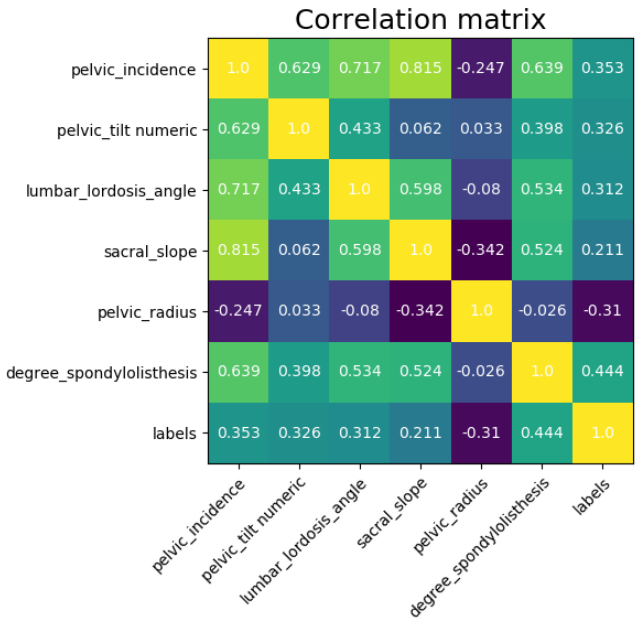

Fig 18. The diagram of the correlation matrix

Due to the effect of dropping a column of feature, the number of leaf nodes and depth of trees are much more than the previous two tasks. This is because decision tree has less data to decide which node should the data goes. It needs to make more decisions to separate data into different groups.

The diagram of each minimum records per leaf node values is shown in Fig 19. to Fig 23.

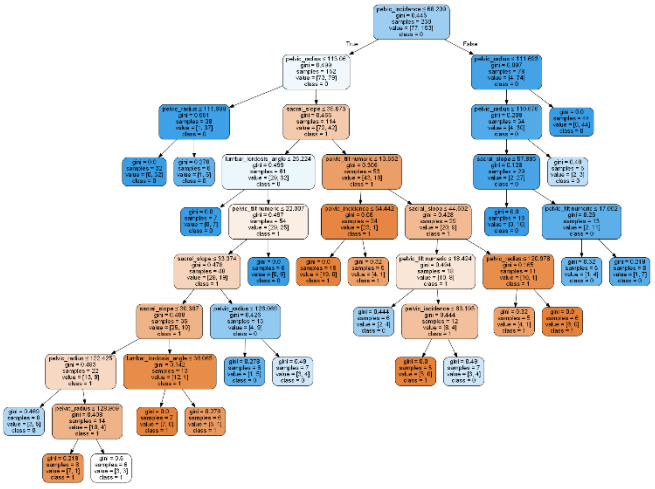

Fig 19. The diagram of “minimum records per leaf node” value of 5

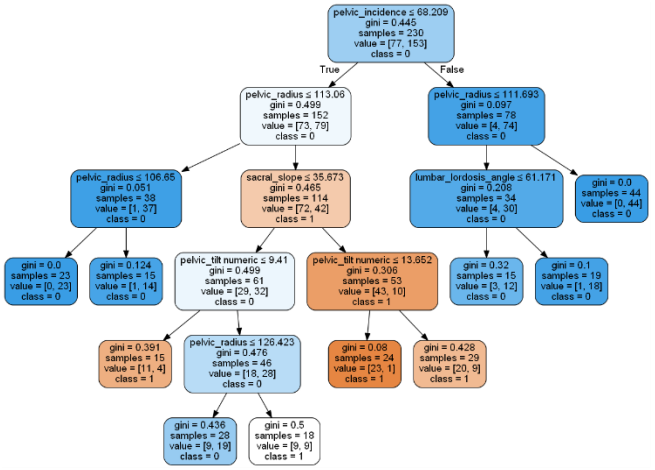

Fig 20. The diagram of “minimum records per leaf node” value of 15

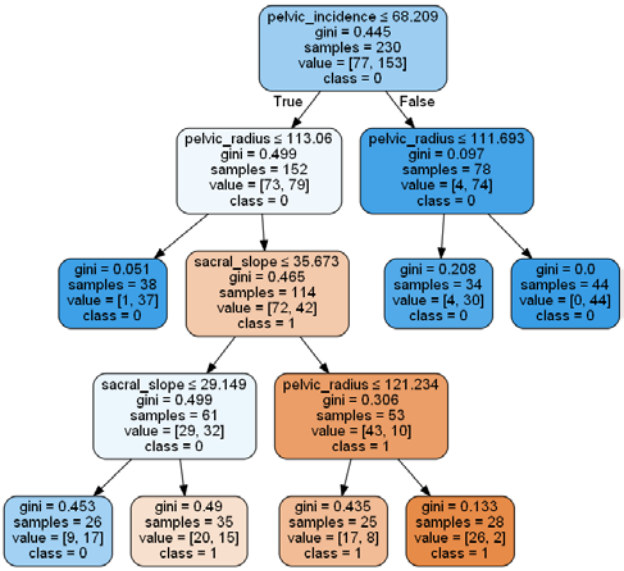

Fig 21. The diagram of "minimum records per leaf node" value of 25

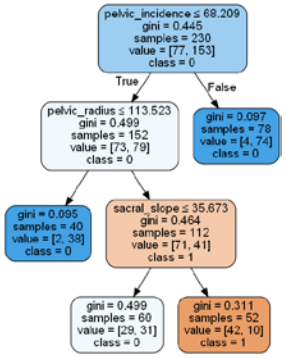

Fig 22. The diagram of "minimum records per leaf node" value of 40

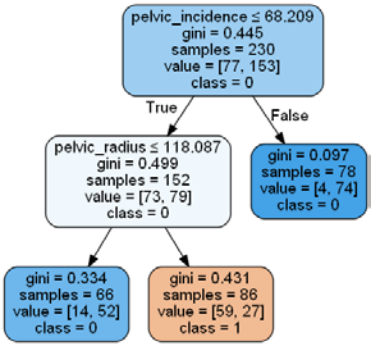

Fig 23. The diagram of "minimum records per leaf node" value of 50

4.2. Third task b.

The precision, recall, and accuracy of each decision tree are shown as follows.

Minimum records per leaf node values of 5	
Precision	Class 1 = 0.865
	Class 2 = 0.571
Recall	Class 1 = 0.789
	Class 2 = 0.696
Accuracy	Accuracy = 0.762

Minimum records per leaf node values of 15	
Precision	Class 1 = 0.865
	Class 2 = 0.571
Recall	Class 1 = 0.789
	Class 2 = 0.696
Accuracy	Accuracy = 0.762

Minimum records per leaf node values of 25	
Precision	Class 1 = 0.865
	Class 2 = 0.571
Recall	Class 1 = 0.789
	Class 2 = 0.696
Accuracy	Accuracy = 0.762

Minimum records per leaf node values of 40	
Precision	Class 1 = 0.865
	Class 2 = 0.571
Recall	Class 1 = 0.789
	Class 2 = 0.696
Accuracy	Accuracy = 0.762

Minimum records per leaf node values of 50	
Precision	Class 1 = 0.865
	Class 2 = 0.571
Recall	Class 1 = 0.789
	Class 2 = 0.696
Accuracy	Accuracy = 0.762

As shown in Fig 24, blue bar represents the value of abnormal precision while orange bar is the value of abnormal recall, green bar stands for the value of normal precision while red bar is the value of normal recall, and purple bar shows the value of accuracy.

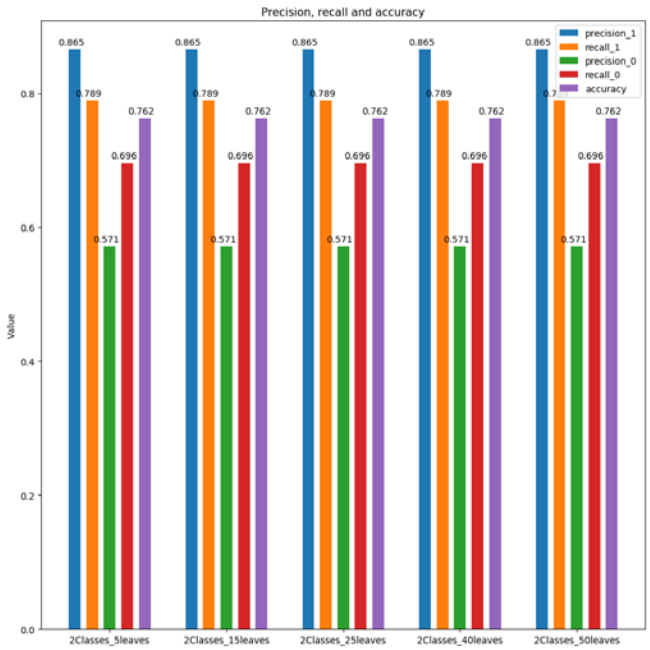

Fig 24. The diagram of each minimum record performance comparison

The confusion matrix of every minimum records is shown as below.

	Predicted 1	Predicted 0
1	13	10
0	10	47

Fig 25. The confusion matrix which “minimum records per leaf node” value of 5

	Predicted 1	Predicted 0
1	15	8
0	10	47

Fig 26. The confusion matrix which “minimum records per leaf node” value of 15

	Predicted 1	Predicted 0
1	15	8
0	14	43

Fig 27. The confusion matrix which “minimum records per leaf node” value of 25

	Predicted 1	Predicted 0
1	7	16
0	7	50

Fig 28. The confusion matrix which “minimum records per leaf node” value of 40

	Predicted 1	Predicted 0
1	16	7
0	12	45

Fig 29. The confusion matrix which “minimum records per leaf node” value of 50

The value of every precision, recall, and accuracy are the same, however, the value of each confusion matrix is not the same. This result represent no matter how minimum records changed, the value of “True Positive”, “True Negative”, “False Positive”, “False Negative” keep changing in proportion. Compared with the previous two tasks, though the accuracy is lower, all results are more stable than before.

5. CONCLUSION

In conclusion, the value of minimum records per leaf node is not the higher the better. There is a threshold of value in different conditions. If the value above the threshold, the performance might drop. The best way to choose the value is to visualize and plot the performance as this report. It could be easier to figure out which decision tree is better.

REFERENCES

[1] Shruti Saxena, *Precision vs Recall*, Medium, May 12, 2018. Accessed on: Aug. 14, 2019. [Online]. Available: https://towardsdatascience.com/precision-vs-recall-386cf9f89488

[2] Gaurav Singh, *Plot correlation matrix using pandas*, Stack Overflow, Apr. 3, 2015. [Online]. Available: https://stackoverflow.com/questions/29432629/plot-correlation-matrix-using-pandas

[3] Ben Keen, *Correlation in Python*, Ben Alex Keen, May 1, 2017. [Online]. Available: http://benalexkeen.com/correlation-in-python/

[4] Drazen Zaric, *Better Heatmaps and Correlation Matrix Plots in Python*, Medium, Apr. 16, 2019. [Online]. Available: https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

[5] Srishti Saha, *Baffled by Covariance and Correlation??? Get the Math and the Application in Analytics for both the terms*, Medium, Oct. 5, 2018. [Online]. Available: https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22

[6] Madhu Sanjeevi, *Chapter 4: Decision Trees Algorithms*, Medium, Oct. 7, 2017. [Online]. Available: https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

[7] Savan Patel, *Chapter 3 : Decision Tree Classifier — Theory*, Medium, May 11, 2017. [Online]. Available: https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

[8] Will Koehrsen, *How to Visualize a Decision Tree from a Random Forest in Python using Scikit-Learn*, Medium, Aug. 19, 2018. [Online]. Available: https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c

[9] Prateek Karkare, *Decision Trees — An Intuitive Introduction*, Medium, Jan. 18, 2019. [Online]. Available: https://medium.com/x8-the-ai-community/decision-trees-an-intuitive-introduction-86c2b39c1a6c

[10] Yury Kashnitsky, *Open Machine Learning Course. Topic 3. Classification, Decision Trees and k Nearest Neighbors*, Medium, Feb. 11, 2018. [Online]. Available: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-3-classification-decision-trees-and-k-nearest-neighbors-8613c6b6d2cd