SENG 474 Assignment 1

Experiments and Analysis

Francis German

V00893968

Second classification problem

The second classification problem selected for this assignment was predicting house prices. The dataset was taken from Zillow's home value prediction Kaggle competition data. The number of input features has been modified and the task was changed into predicting whether the house price is above or below median value[5]. it contains a large number of instances 1460 compared to the 297 heart disease instances but contains around the same and or less number of features (10) to (14), compared to the heart disease dataset. I wanted to know if the larger samples would produce more accurate result than a dataset with less samples.

Background information

the data was split 80% for training and 20% for testing for both the heart disease and house price data sets and had default max_depth of 0. The best attributes was found for each classifier: decision trees, random forests, and neural networks. These attributes where then used to analyze the effects of changing the split between training and test data.

Performance and analysis

The following sections analyses the performance and results of the heart disease and house price datasets being used to train decision trees, neural networks and random forest.

DECISION TREE

The DecisionTreeClassifier from scikit learn was used to create decision tree classifiers for the data sets [4]. DecisionTreeClassifier uses cost complexity pruning to avoid over-fitting the decision tree [3] this pruning technique is parameterized by the cost complexity parameter . The complexity parameter was compared against the accuracy of the tree as it was pruned. This was done using both Gini index and entropy as split criterion. The heart disease results with gini split criterion is shown in figure 1 and result with entropy In figure 2.

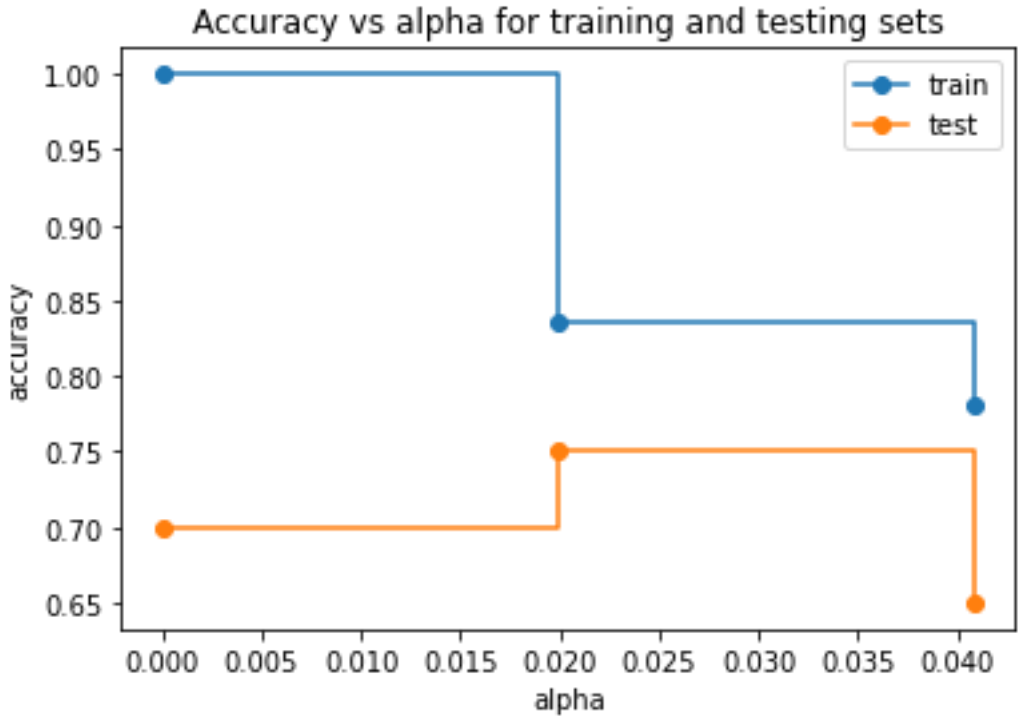

FIGURE 1- TRAINING AND TEST RESULT HEART DISEASE DATASET (SPLIT CITERION- GINI)

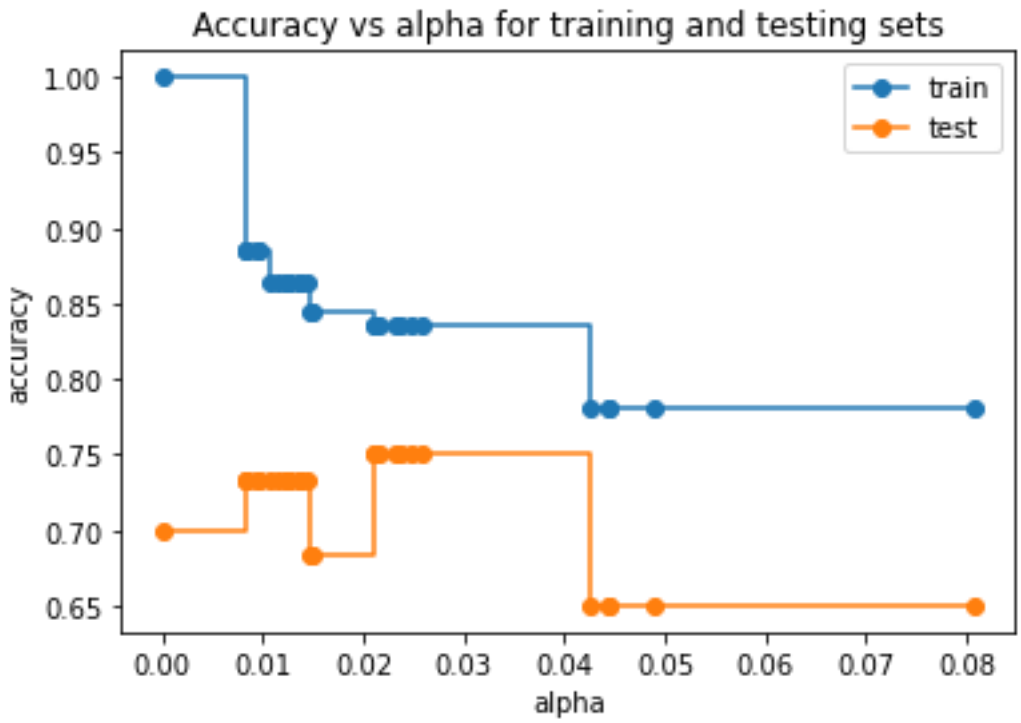

FIGURE 2- TRAINING AND TEST RESULT HEART DISEASE DATASET (SPLIT CITERION-ENTROPY)

For the heart disease dataset using the gini index as the spilt criterion yielded the best results of approximately 75% accuracy on the test set which was slightly better than the 73% accuracy achieved using entropy as the split criterion. The best preforming trees gotten from the gridsearchCV plugin showed the hyperparameter for this case was a split criterion of gini and max depth=3 and random _state= 0, and it gave an accuracy prediction of 75%.

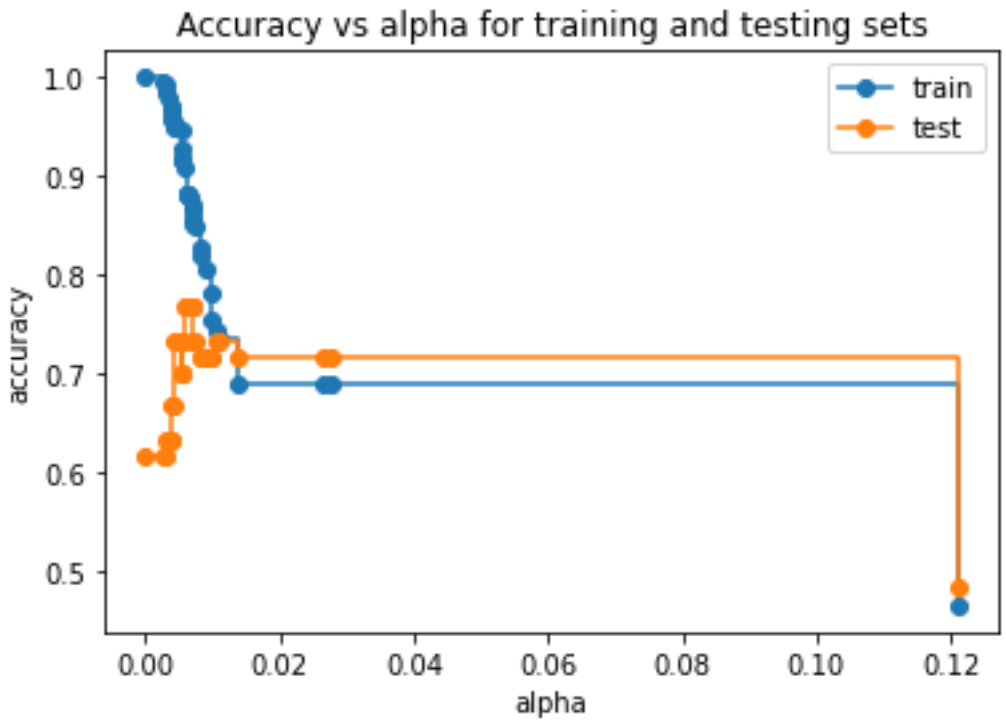

FIGURE 3- TRAINING AND TEST RESULT HOSUE PRICE DATASET (SPLIT CITERION- GINI)

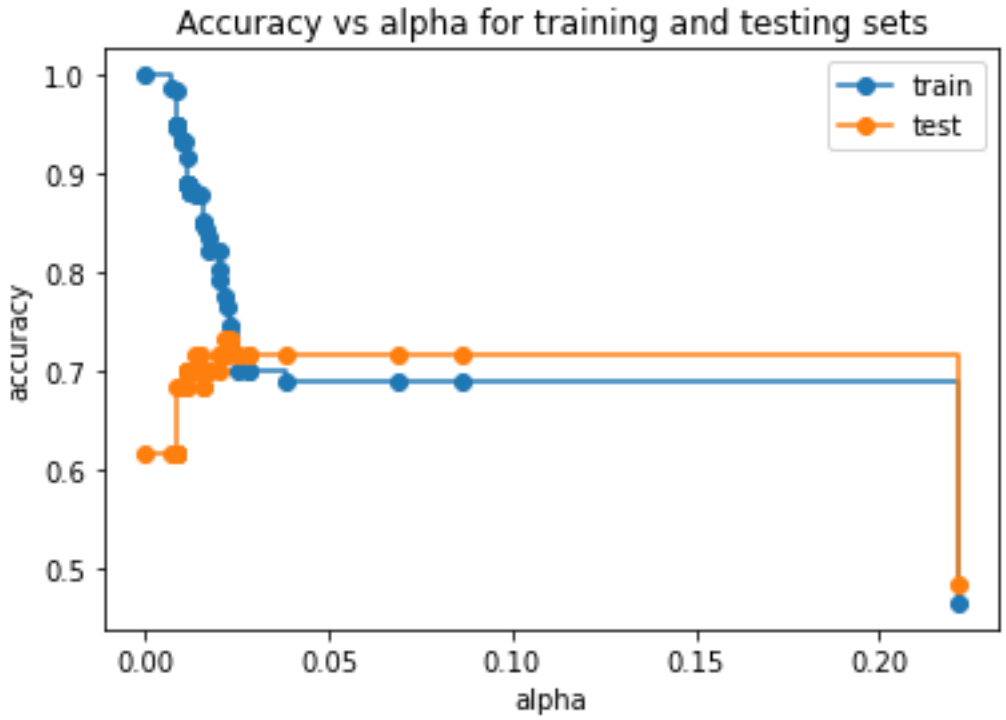

FIGURE 4- TRAINING AND TEST RESULT HOUSE PRICE DATASET(SPLIT CITERION- ENTROPY)

For the house price dataset gini split criterion were able to achieve approximately 73% accuracy on the test dataset while the entropy split criterion also achieved a 73%. The accuracy is represented in figure 3 and 4.

For both datasets using gini as a split criterion produced a smaller tree after pruning. The gini index preformed slightly better on the heart disease data while both split criteria preformed equally well on the larger house price dataset.

NEURAL NETWORK

The MLPClassifier from scikit learn was used to create neural network classifiers for the data sets [6]. The networks were created using the stochastic gradient descent solver and adam, I tested the neural network for the heart disease and house price dataset with 3 hidden layers (the input layer, 1 hidden layer, and the output layer).

For the first test conducted for the heart disease dataset, I started with 12 nodes in the hidden layer and 100 iterations or epochs which gave a result of 100% accuracy then I increased the number of nodes in the hidden layer to 62 and achieved a higher accuracy prediction of 91% as can see model accuracy in figure 5 shows the accuracy with the number of epoch compared to figure 6.

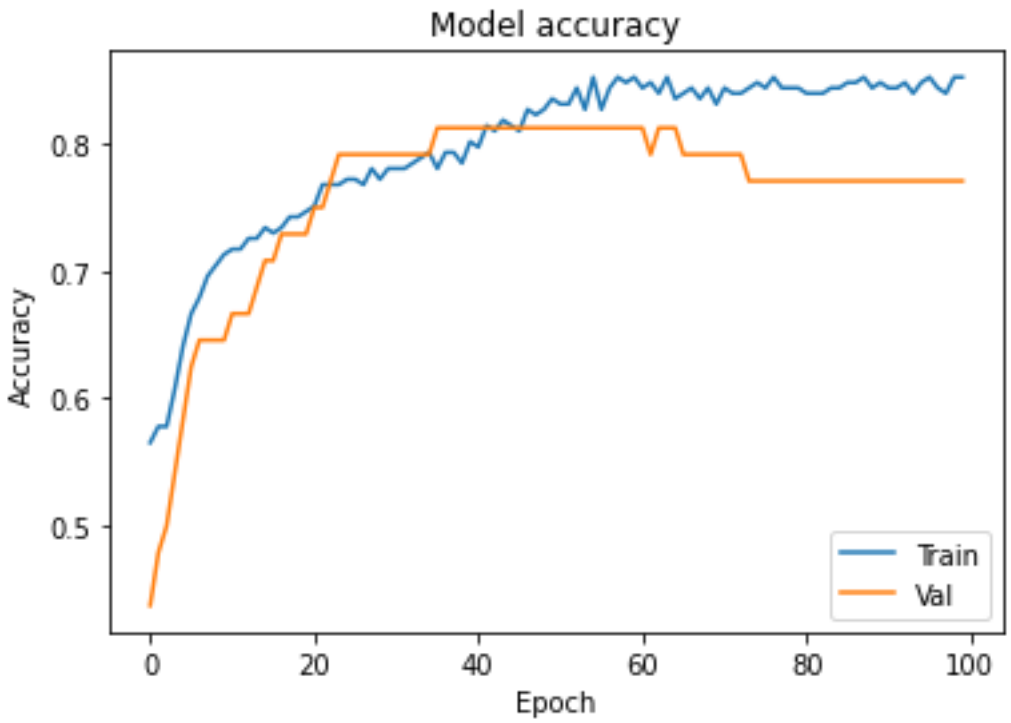

FIGURE 5. 12 NODES IN HIDDEN LAYERS- HEART DISEASE DATASET

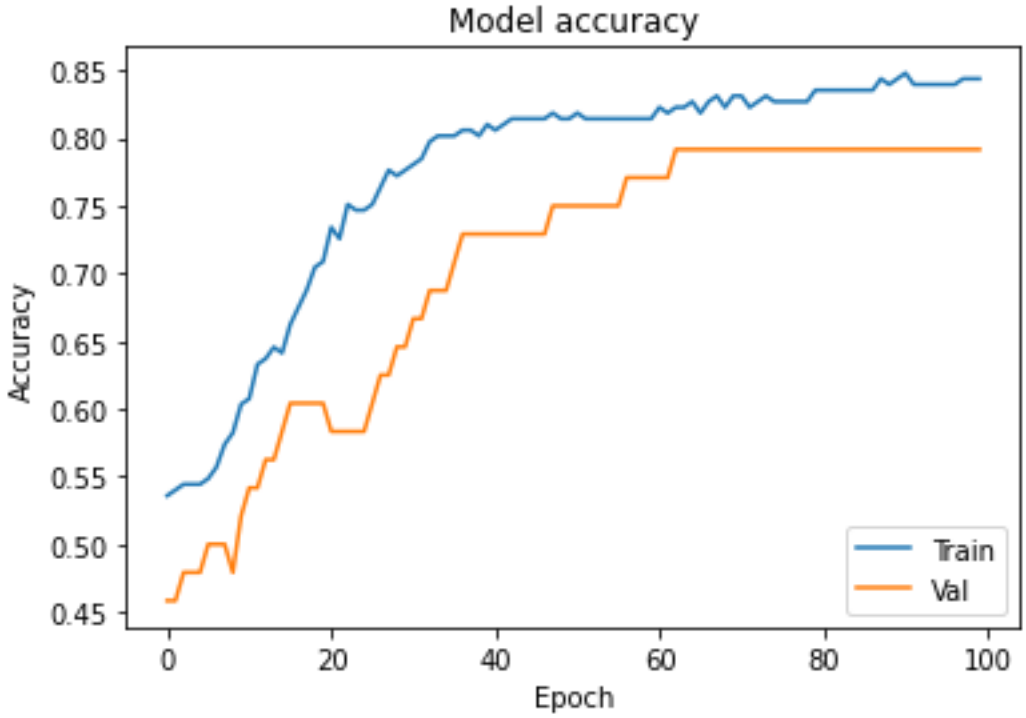

FIGURE 6. 62 NODES IN HIDDEN LAYERS- HEART DISEASE DATASET

Then for the house price dataset, I carried out the same test and based on the previous test conducted I was expecting similar results. For the test with 12 nodes in the hidden layer It achieved 86% accuracy while with the test with 62 nodes in the hidden layer got a 81% as seen in both figures 7 and 8.

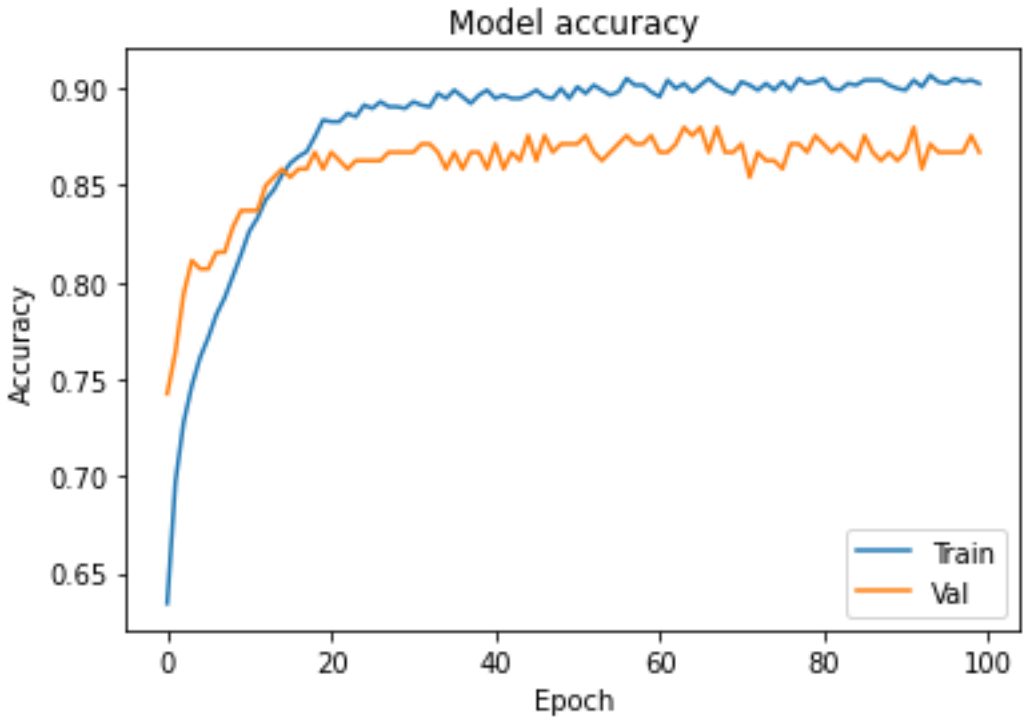

FIGURE 7. 12 NODES IN HIDDEN LAYER- HOUSE PRICE DATASET

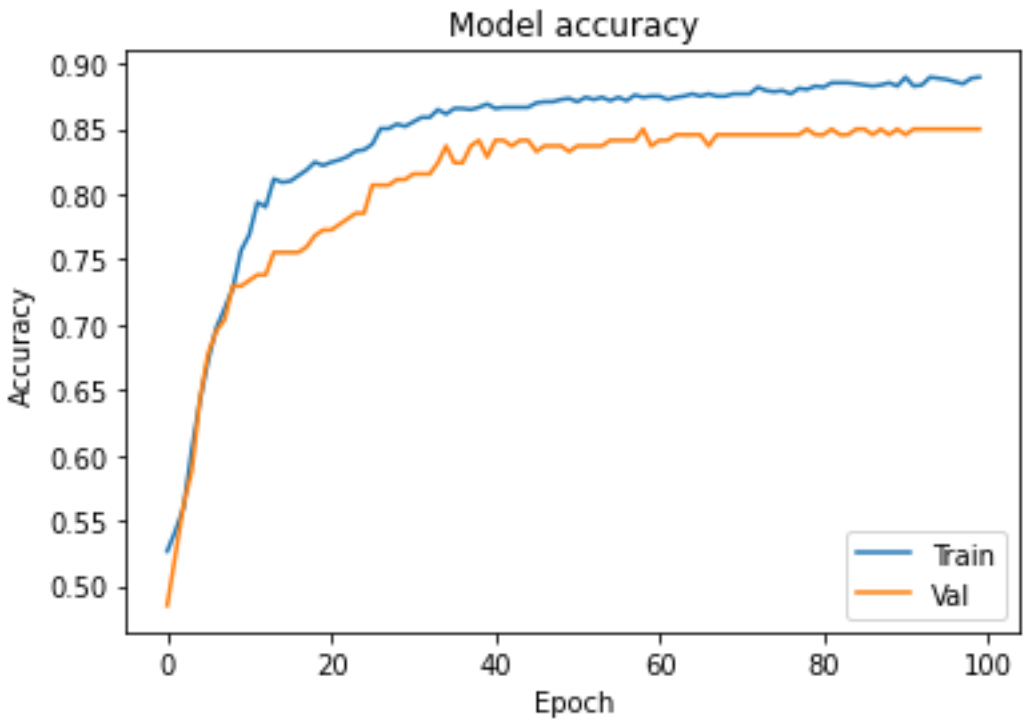

FIGURE 8. 62 NODES IN HIDDEN LAYERS- HOUSE PRICE DATASET

For both datasets using more nodes in the hidden layers increased the accuracy prediction and using less nodes decreased the accuracy prediction. The SGD yielded the best accuracy results after conducting a test with both sgd and adam, the test conducted with the adam solver had 12 nodes in the hidden layers and 100 iterations or epoch and produce an accuracy prediction of 75% for the heart disease dataset then I increased the number of nodes from 12 to 62 but still got the same accuracy prediction.

RANDOM FOREST

The RandomForestClassifier from scikit learn was used to create decision tree classifiers for the data sets [2]. Several forests were generated by combining different variations of maximum size of the forest, then decided to go with 500

trees and 50 trees for test conducted, for all forests gini was used for split criterion. For the heart disease training dataset random forests were able to achieve 76% accuracy with a forest size 500 trees compared to the test data with smaller forest size of 50 which gave a 80% accuracy prediction while the test for the house price dataset with 500 trees gave an accuracy prediction of 93% and the test conducted with 50 trees gave a 92% accuracy. The forests that preformed the best on the test set were smaller forests of 50 trees with default depth of 0. The smaller forests were able to achieve 80% and 92% accuracy on the test data.

The random forest classifier was able to achieve a much higher accuracy on the house price dataset compared to the heart disease dataset. I found out that the less amount of trees in the forest the more accurate the prediction with a default depth of 0. A small forest with trees of depth 0 performs just as well as or if not better than a larger forest of similar depth.

In conclusion, For both the heart disease and the house price dataset the random forest, and neural network produced the best test accuracies compared to the decision tree classifier. On the heart disease dataset the neural network classifier had the highest accuracy prediction of 100% based on best case parameters while the decision tree classifier had the lowest accuracy prediction of 75%, the random forest came out on top for the house price dataset with 93% for best case parameter. After all the tests conducted I learned that the highest test accuracy came from best case parameter and not the classifier being used.

References

1] UCI, “Neural Network Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.[Accessed February 4, 2021].

2] UCI, “Random Forest Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html . [Accessed February 4, 2021].
3] UCI, “Cost-Complexity Pruning” [Online]. Available: https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html. [Accessed February 4, 2020].
4] UCI, “Decision Tree Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Accessed February 4, 2020].
5] UCI, “House price dataset” [Online]. Available: https://medium.com/intuitive-deep-learning/build-your-first-neural-network-to-predict-house-prices-with-keras-eb5db60232c .[Accessed February 4, 2020].

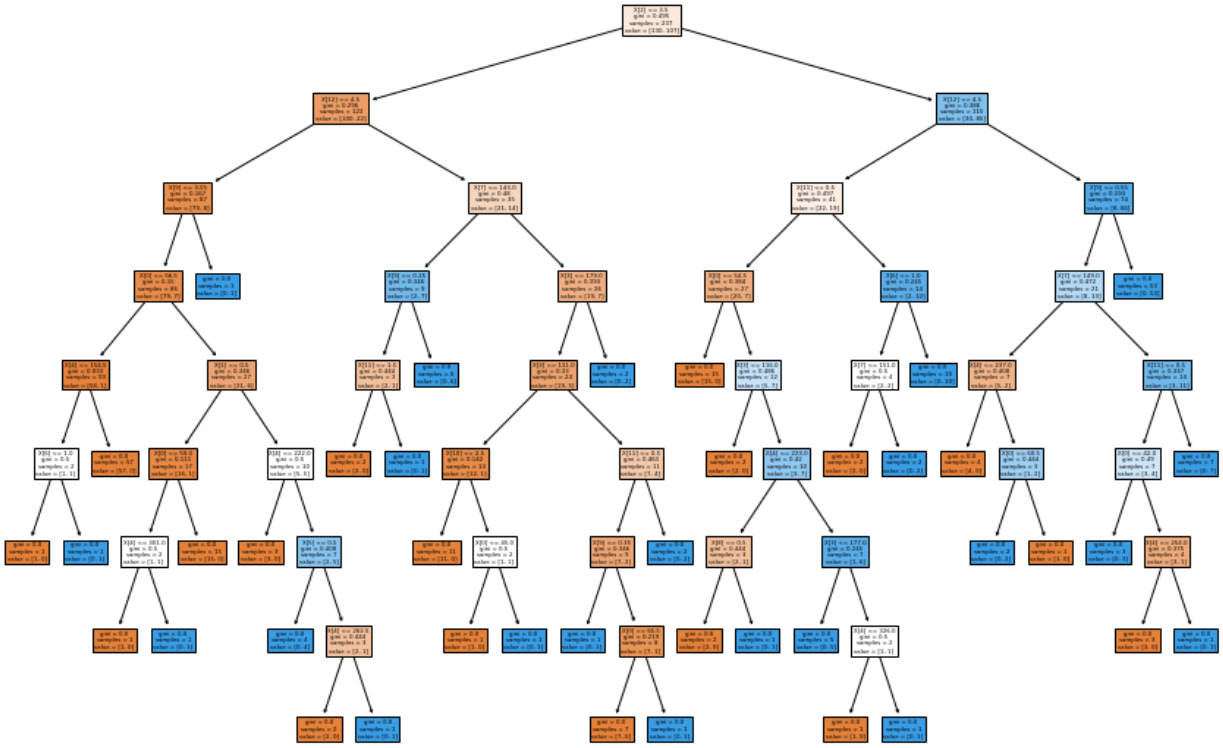

FIGURE 9. DECISION TREE SPLIT CITERION (GINI)(HEART DISEASE)

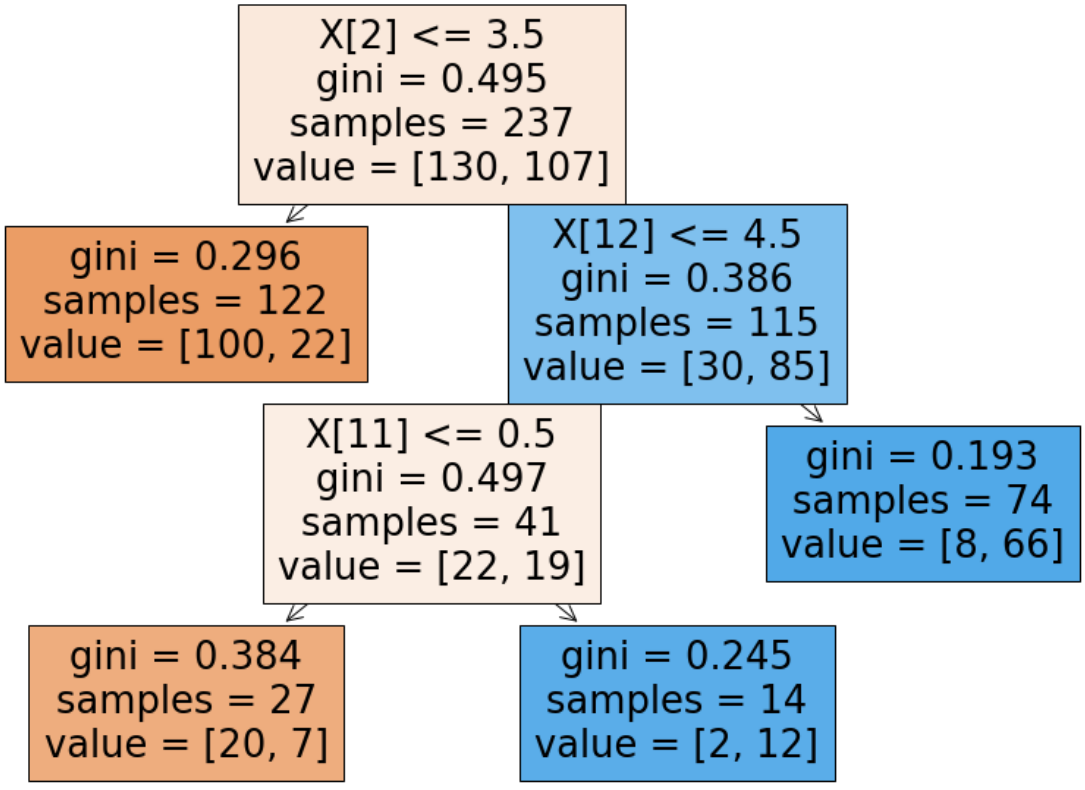

FIGURE 10. DECISION TREE SPLIT CITERION (GINI) POST PRUNNING (HEART DISEASE)

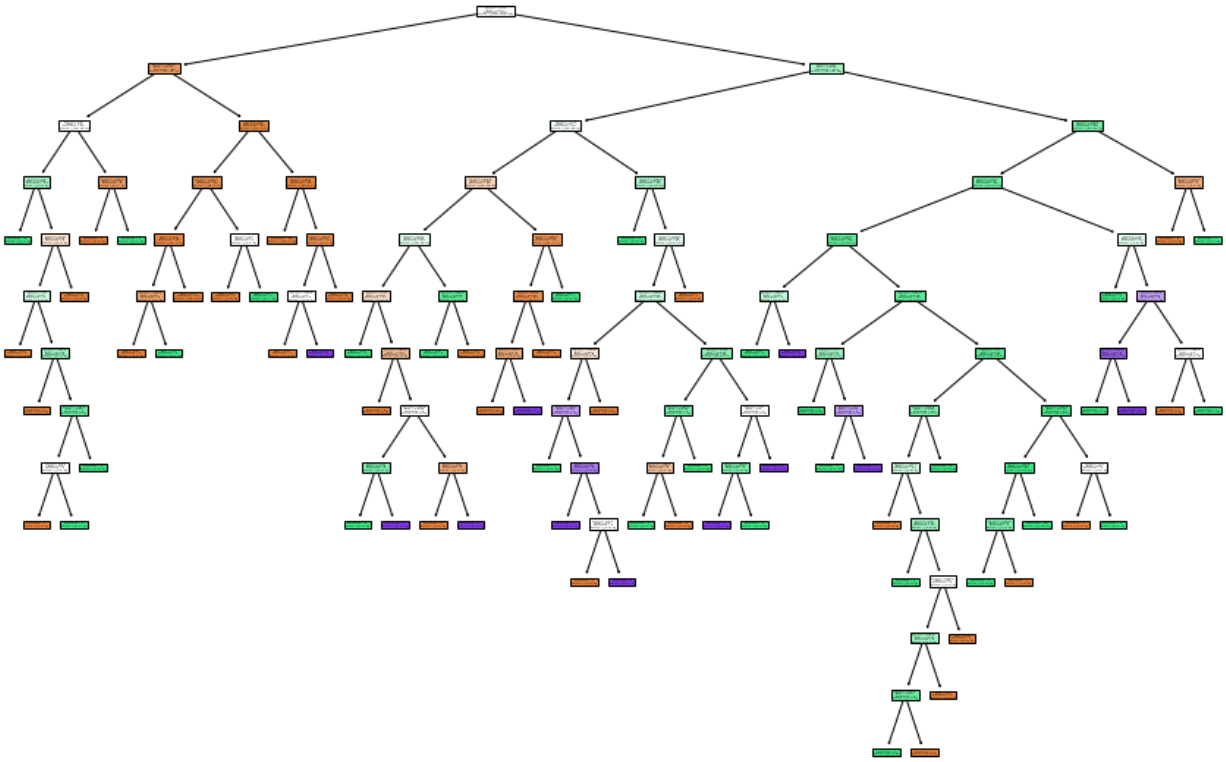

FIGURE 11. DECISION TREE SPLIT CITERION (GINI) HOUSE PRICE

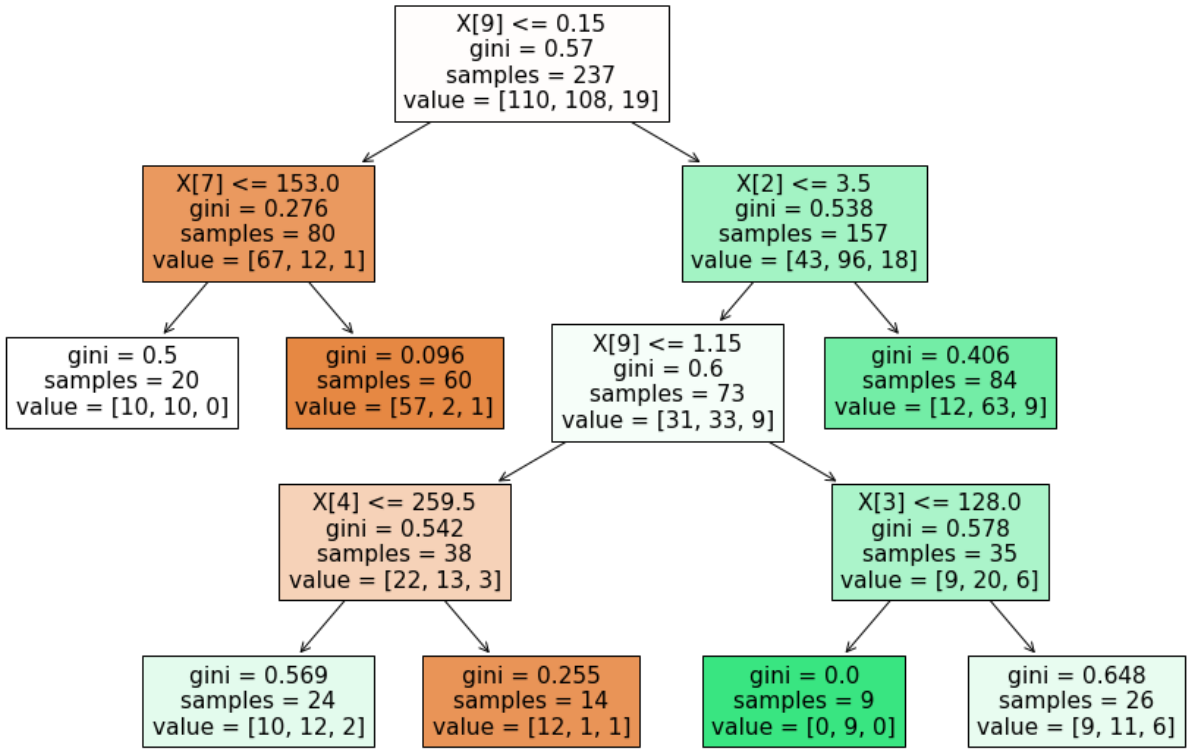

FIGURE 12. DECISION TREE SPLIT CITERION (GINI) POST PRUNNING HOUSE PRICE