Name: Sambhav Jain

Roll No.: – 24CS60R39

Topic - Decision Tree implementation on given Diabetes dataset.

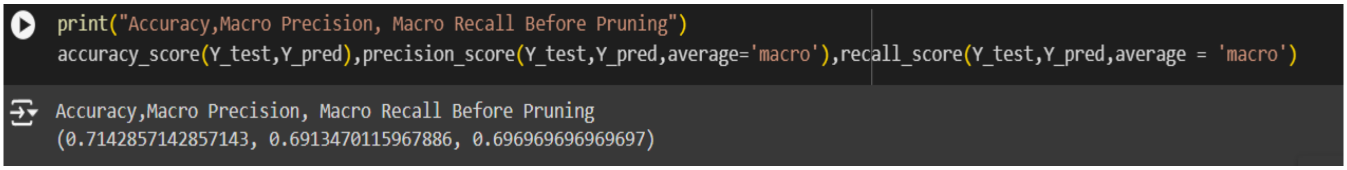
Dataset –

- Noiseless Data: diabetes.csv
- Noisy Data: diabetes_noise.csv

1. Diabetes.csv (Noiseless data)

(Before Pruning)

- Accuracy - 0.7142857142857143
- Macro Precision - 0.6913470115967886
- Macro Recall - 0.696969696969697

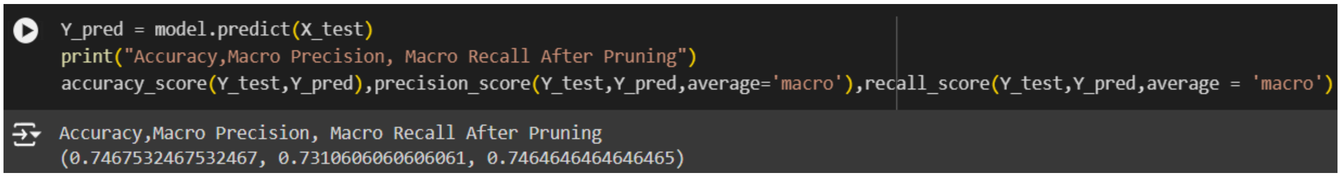


```
print("Accuracy,Macro Precision, Macro Recall Before Pruning")
accuracy_score(Y_test,Y_pred),precision_score(Y_test,Y_pred,average='macro'),recall_score(Y_test,Y_pred,average = 'macro')
```

```
Accuracy,Macro Precision, Macro Recall Before Pruning
(0.7142857142857143, 0.6913470115967886, 0.696969696969697)
```

(After Pruning)

- Accuracy - 0.7467532467532467
- Macro Precision - 0.7310606060606061
- Macro Recall - 0.7464646464646465



```
Y_pred = model.predict(X_test)
print("Accuracy,Macro Precision, Macro Recall After Pruning")
accuracy_score(Y_test,Y_pred),precision_score(Y_test,Y_pred,average='macro'),recall_score(Y_test,Y_pred,average = 'macro')
```

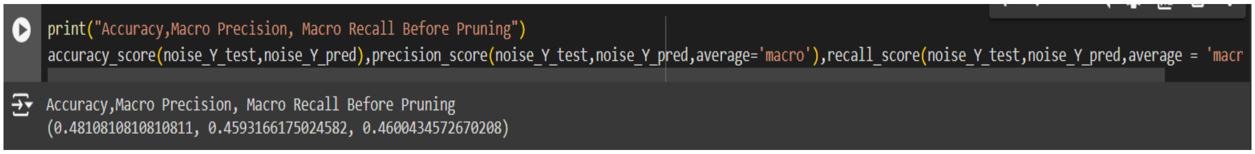
```
Accuracy,Macro Precision, Macro Recall After Pruning
(0.7467532467532467, 0.7310606060606061, 0.7464646464646465)
```

2. Diabetes_noise.csv (Noisy data)

(Before Pruning)

- Accuracy - 0.4810810810810811
- Macro Precision - 0.4593166175024582

- Macro Recall - 0.4600434572670208

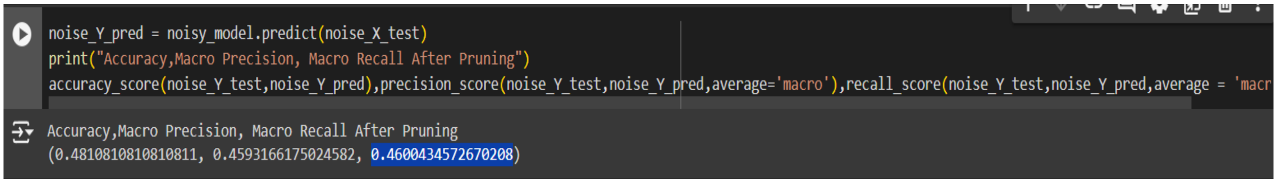


```
print("Accuracy,Macro Precision, Macro Recall Before Pruning")
accuracy_score(noise_Y_test,noise_Y_pred),precision_score(noise_Y_test,noise_Y_pred,average='macro'),recall_score(noise_Y_test,noise_Y_pred,average = 'macr
```

```
Accuracy,Macro Precision, Macro Recall Before Pruning
(0.4810810810810811, 0.4593166175024582, 0.4600434572670208)
```

(After Pruning)

- Accuracy - 0.4810810810810811
- Macro Precision - 0.4593166175024582
- Macro Recall - 0.4600434572670208



```
noise_Y_pred = noisy_model.predict(noise_X_test)
print("Accuracy,Macro Precision, Macro Recall After Pruning")
accuracy_score(noise_Y_test,noise_Y_pred),precision_score(noise_Y_test,noise_Y_pred,average='macro'),recall_score(noise_Y_test,noise_Y_pred,average = 'macr
```

```
Accuracy,Macro Precision, Macro Recall After Pruning
(0.4810810810810811, 0.4593166175024582, 0.4600434572670208)
```

Conclusion: In the first dataset we are given noiseless data, this means the data is consistent and there are no outliers. On data when I apply Decision tree Algorithm without pruning, then it gives accuracy of 71.43%. After that we are asked to apply pruning on the same dataset, then model gives Accuracy of 74.68%. Without pruning the model is trying to fit each feature even though some features are useless and unnecessarily contribute to reducing the overall accuracy. After that we do pruning then the model ignores the useless features that increases the overall accuracy.

On the noisy dataset which includes outliers, we already expected the model to perform worse than noiseless. It gives the accuracy of 48.11% before pruning. Since noise can confuse the model during learning. However, after pruning, the accuracy did not change. This shows that even after pruning, the accuracy may not improve, which also depends on the type of pruning, post-pruning, or pre-pruning. The post-pruning technique is more effective in increasing the accuracy.