CSCI4150U: Data Mining
Lab 03

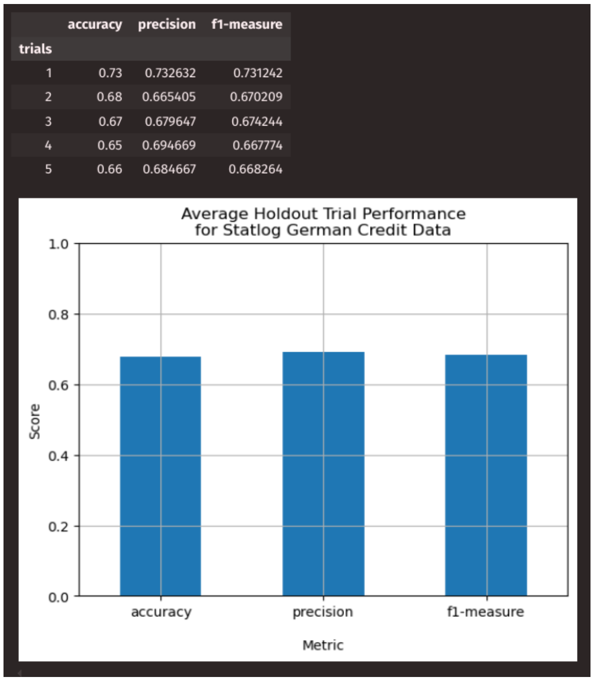
Syed Naqvi
100590852

October 9, 2024

Part I:

1.

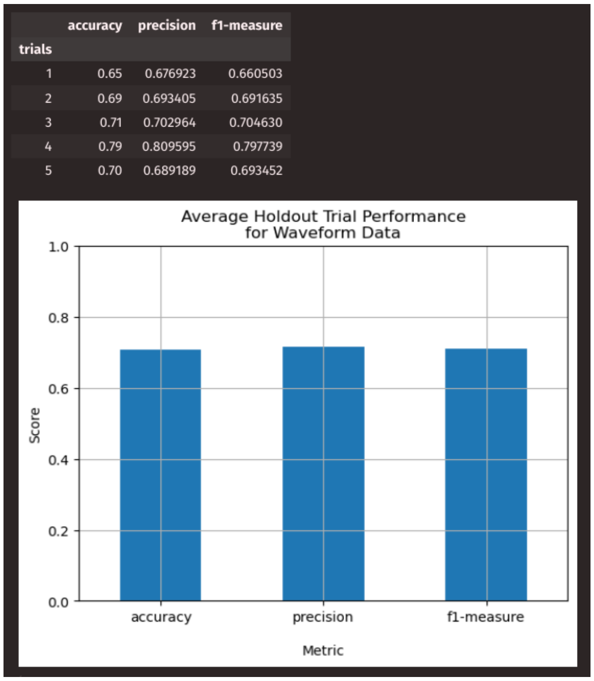
Average and individual trial results across 5 randomized trials using accuracy, precision and f1 scores as performance metrics based on a (90%-train and 10%-test) **holdout** validation method, decision tree classifiers and Gini impurity.



	accuracy	precision	f1-measure
trials			
1	0.73	0.732632	0.731242
2	0.68	0.665405	0.670209
3	0.67	0.679647	0.674244
4	0.65	0.694669	0.667774
5	0.66	0.684667	0.668264

Average Holdout Trial Performance
for Statlog German Credit Data
1.0
0.8
0.6
0.4
0.2
0.0
Score
accuracy
precision
f1-measure
Metric

Figure 1: Decision tree performance based on holdout method and Gini impurity using Statlog German Credit Dataset

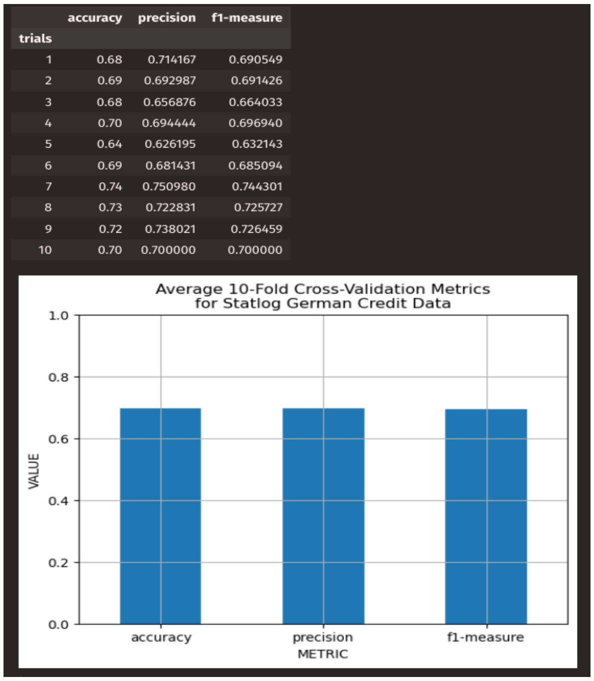


	accuracy	precision	f1-measure
trials			
1	0.65	0.676923	0.660503
2	0.69	0.693405	0.691635
3	0.71	0.702964	0.704630
4	0.79	0.809595	0.797739
5	0.70	0.689189	0.693452

Figure 2: Decision tree performance basd on holdout method and Gini impurity using Waveform Dataset

2.

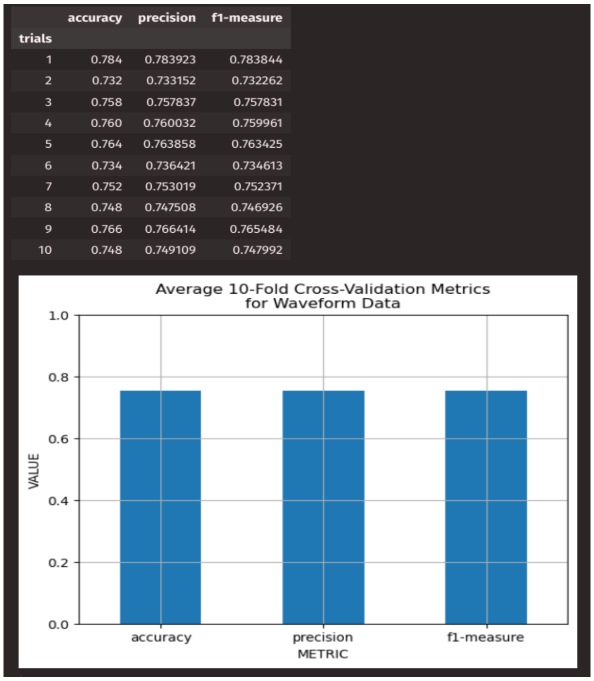
Average and individual trial results using accuracy, precision and f1 scores based on **10-fold cross-validation**, decision tree classifiers and Gini impurity.



	accuracy	precision	f1-measure
trials			
1	0.68	0.714167	0.690549
2	0.69	0.692987	0.691426
3	0.68	0.656876	0.664033
4	0.70	0.694444	0.696940
5	0.64	0.626195	0.632143
6	0.69	0.681431	0.685094
7	0.74	0.750980	0.744301
8	0.73	0.722831	0.725727
9	0.72	0.738021	0.726459
10	0.70	0.700000	0.700000

Average 10-Fold Cross-Validation Metrics
for Statlog German Credit Data
1.0
0.8
0.6
0.4
0.2
0.0
VALUE
accuracy
precision
f1-measure
METRIC

Figure 3: Decision tree performance based on 10-fold cross-validation and Gini impurity using Statlog German Credit Dataset



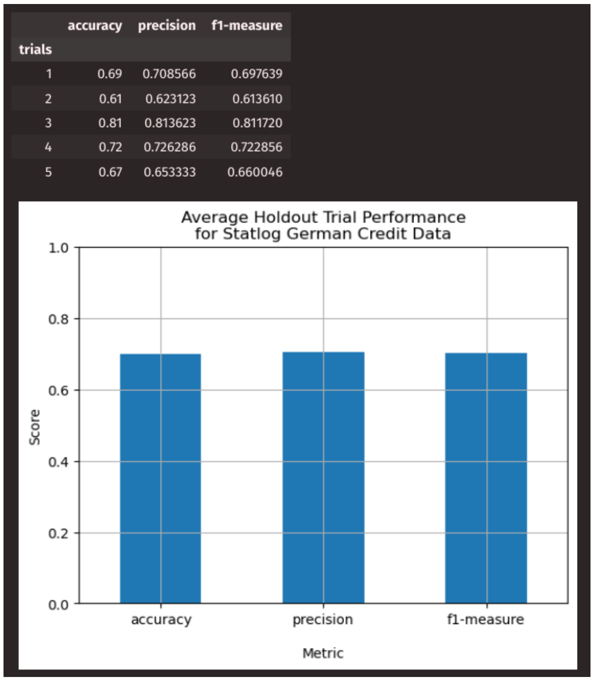
	accuracy	precision	f1-measure
trials			
1	0.784	0.783923	0.783844
2	0.732	0.733152	0.732262
3	0.758	0.757837	0.757831
4	0.760	0.760032	0.759961
5	0.764	0.763858	0.763425
6	0.734	0.736421	0.734613
7	0.752	0.753019	0.752371
8	0.748	0.747508	0.746926
9	0.766	0.766414	0.765484
10	0.748	0.749109	0.747992

Figure 4: Decision tree performance based on 10-fold cross-validation and Gini impurity using Waveform Dataset

Part II:

1.

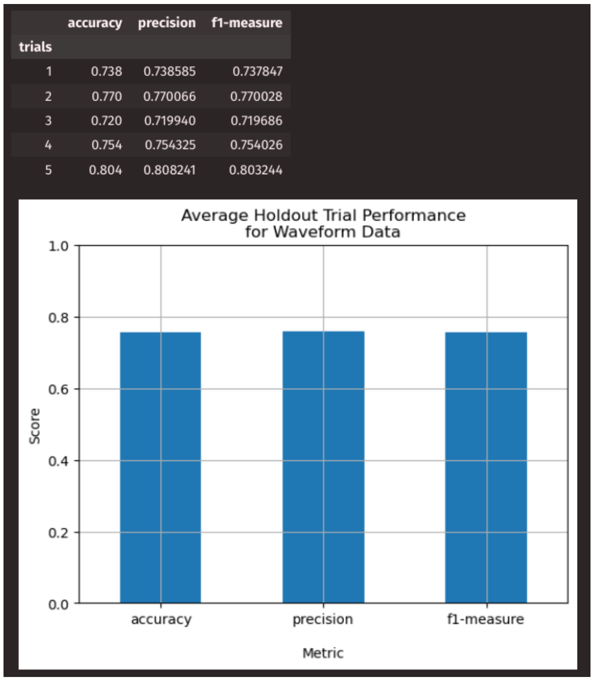
Repetition of **Part I**, this time using **’Entropy’** instead of **’Gini;** as the decision-tree impurity measure.



	accuracy	precision	f1-measure
trials			
1	0.69	0.708566	0.697639
2	0.61	0.623123	0.613610
3	0.81	0.813623	0.811720
4	0.72	0.726286	0.722856
5	0.67	0.653333	0.660046

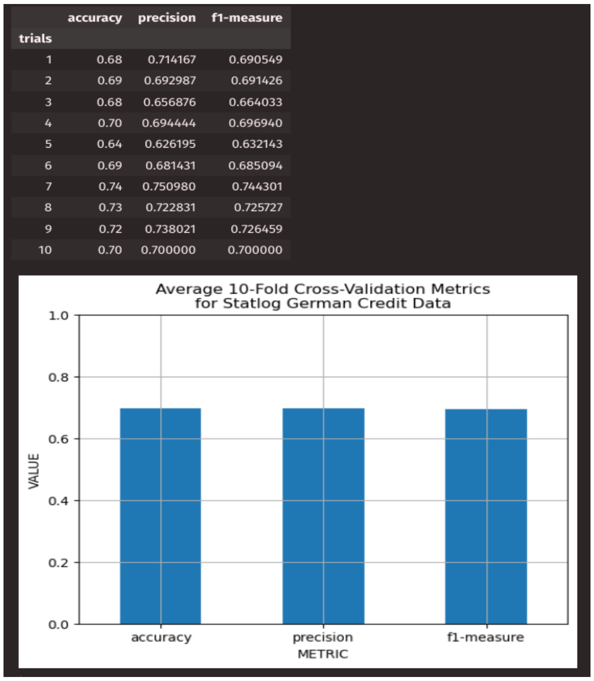
Average Holdout Trial Performance
for Statlog German Credit Data
1.0
0.8
0.6
0.4
0.2
0.0
Score
accuracy
precision
f1-measure
Metric

Figure 5: Decision tree performance based on holdout method and Entropy impurity using Statlog German Credit Dataset



	accuracy	precision	f1-measure
trials			
1	0.738	0.738585	0.737847
2	0.770	0.770066	0.770028
3	0.720	0.719940	0.719686
4	0.754	0.754325	0.754026
5	0.804	0.808241	0.803244

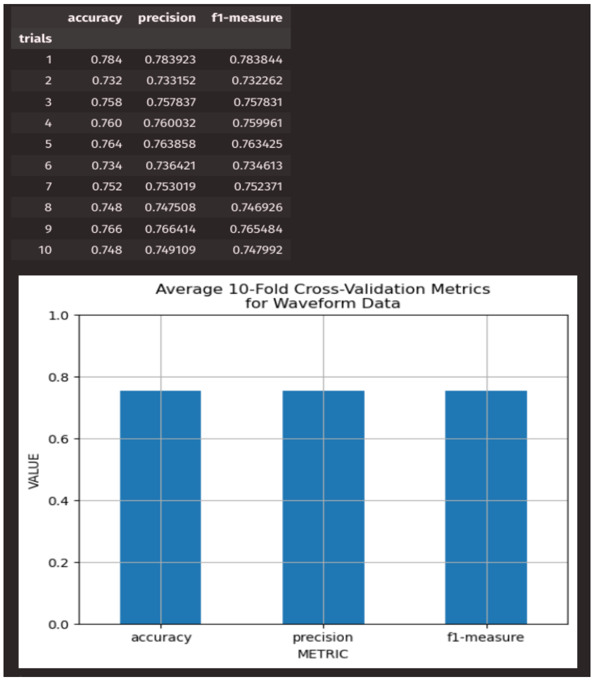
Figure 6: Decision tree performance based on holdout method and Entropy impurity using Waveform Dataset



	accuracy	precision	f1-measure
trials			
1	0.68	0.714167	0.690549
2	0.69	0.692987	0.691426
3	0.68	0.656876	0.664033
4	0.70	0.694444	0.696940
5	0.64	0.626195	0.632143
6	0.69	0.681431	0.685094
7	0.74	0.750980	0.744301
8	0.73	0.722831	0.725727
9	0.72	0.738021	0.726459
10	0.70	0.700000	0.700000

Average 10-Fold Cross-Validation Metrics
for Statlog German Credit Data
1.0
0.8
0.6
0.4
0.2
0.0
VALUE
accuracy
precision
f1-measure
METRIC

Figure 7: Decision tree performance based on 10-fold cross-validation and Entropy impurity using Statlog German Credit Dataset

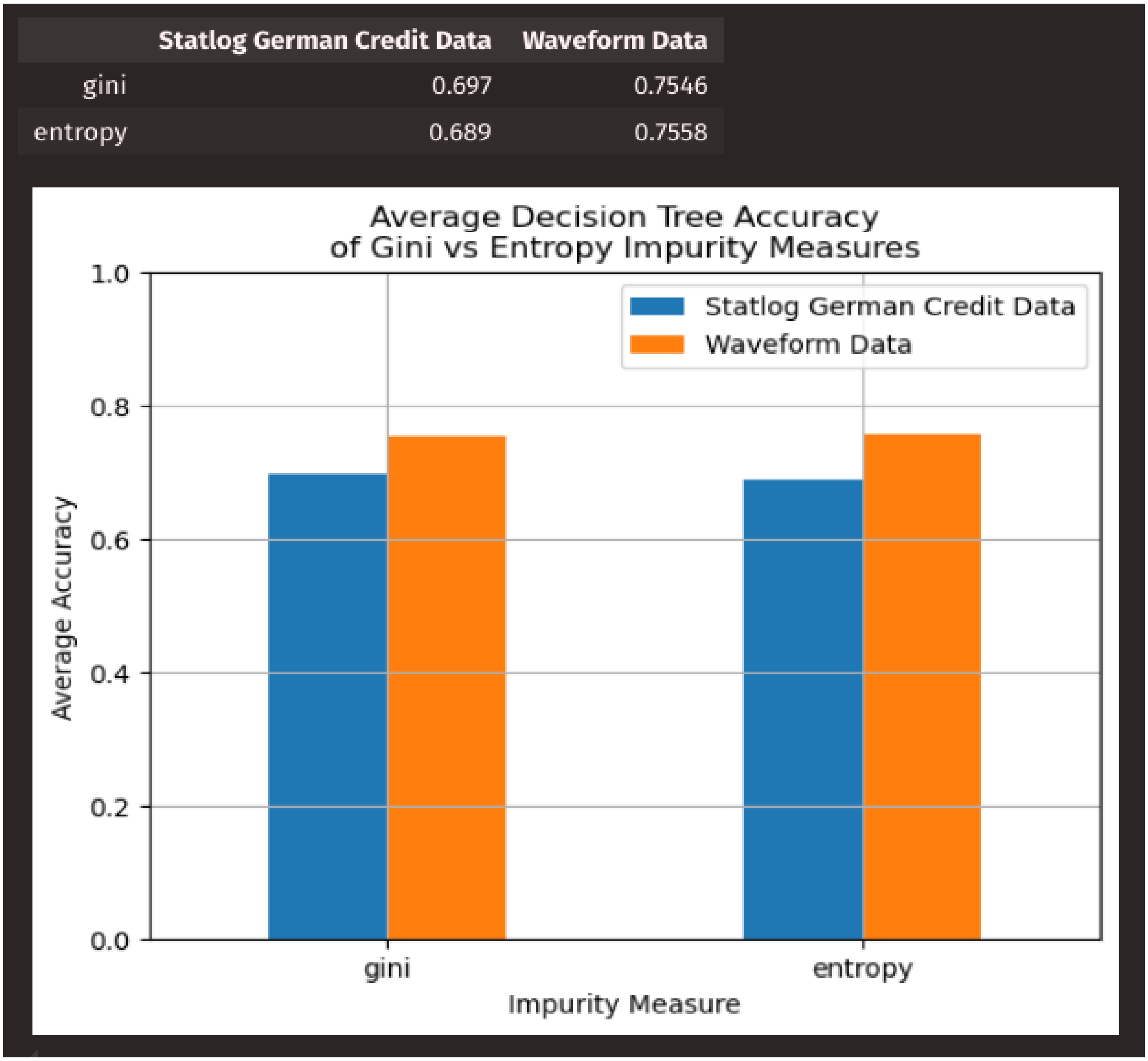


	accuracy	precision	f1-measure
trials			
1	0.784	0.783923	0.783844
2	0.732	0.733152	0.732262
3	0.758	0.757837	0.757831
4	0.760	0.760032	0.759961
5	0.764	0.763858	0.763425
6	0.734	0.736421	0.734613
7	0.752	0.753019	0.752371
8	0.748	0.747508	0.746926
9	0.766	0.766414	0.765484
10	0.748	0.749109	0.747992

Figure 8: Decision tree performance based on 10-fold cross-validation and Entropy impurity using Waveform Dataset

2.

Comparing the average **10-fold cross-validation** based accuracy of decision-tree classifiers using **’Gini’** vs **’Entropy’** as impurity measures for both **Statlog German Credit** and **Waveform Datasets**.



	Statlog German Credit Data	Waveform Data
gini	0.697	0.7546
entropy	0.689	0.7558

Average Decision Tree Accuracy
of Gini vs Entropy Impurity Measures
1.0
0.8
0.6
0.4
0.2
0.0
Average Accuracy
Statlog German Credit Data
Waveform Data
gini
entropy
Impurity Measure

Figure 9: Accuracy comparison of decision tree classifiers using Entropy vs Gini impurity measures

Part III:

1.

Plotting decision tree accuracies based on the (90%-train and 10%-test) **holdout** validation method and varying max tree depths.

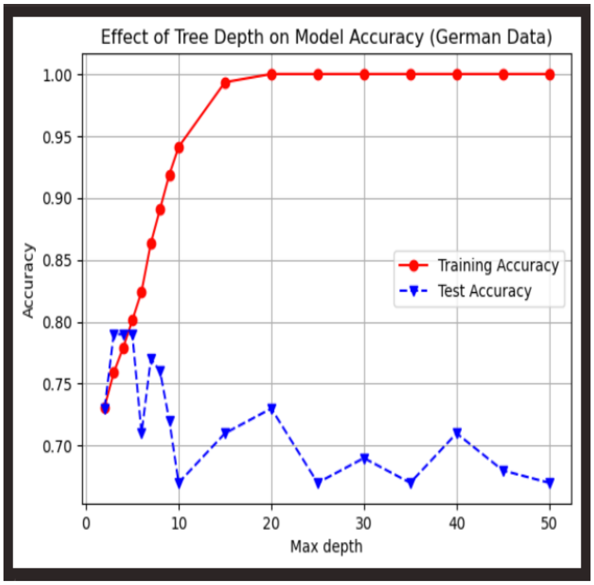


Figure 10: Decision tree accuracies for Statlog German Credit Dataset

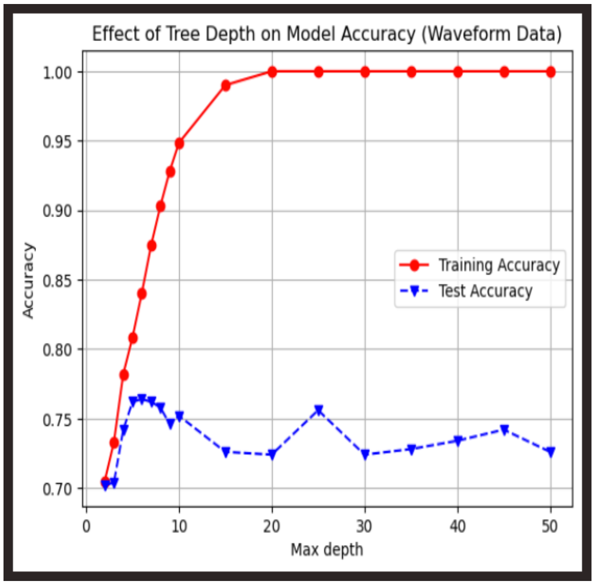


Figure 11: Decision tree accuracies for Waveform Dataset

2.

We can see clearly that with increasing tree depth there is an overfitting phenomenon observable across both datasets. The model achieves an increasingly better fit to its training data until it can perfectly classify all records. However, we see consistently that after a short period of increasing test accuracy, the model experiences a net decrease in test accuracy with increasing tree depth. This indicates that after a certain tree depth, the model starts to become highly specialized and overfits to its training data. This results in a model that poorly generalizes to unseen test data.