CA03 – Decision Tree

Jamie Meugniot

**Q.1.1.**

Discretizing columns for this problem make sense given the continuous data that would exist in these columns. Values can be different for each record yet vary slightly. Discretizing columns creates clear levels within the nodes and thus we can build the decision tree and split based on these clearly identified levels.

**Q.1.2.**

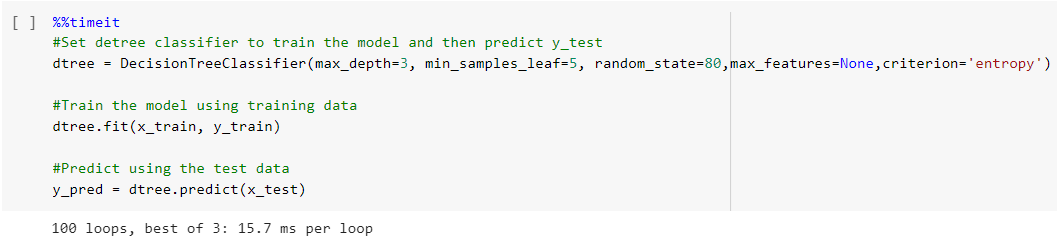
An issue that can arise from not discretizing the columns would be any present outliers and highly skewed variables. Decision trees using entropy as the criterion will also be unable to use continuous data and another model will need to be selected.

**Q.7.1.**



**Q.8.1.**

Using %%timeit, it took 15.7ms to train the model.



**Q.8.2.**

After completing the hyperparameter variations and reviewing the corresponding model performance, I found that the best tree yields an F1 Score of 64% and an AUC of 76% using ‘gini’ as the criterion.

Of all the variations of hyperparameter testing, I would say that this is the best tree.

**Q.8.3.**

In Colab Document: <https://colab.research.google.com/drive/1DlJaIGOIsOP4bg4djhoSrYUeaOnyoRNd?usp=sharing>

**Q.8.4.**

I know this is the best model because the AUC was highest at those given parameters meaning it performs well across various threshold values.

**Q.10.1.**