CA03 **Max Kaiser**
Decision Tree

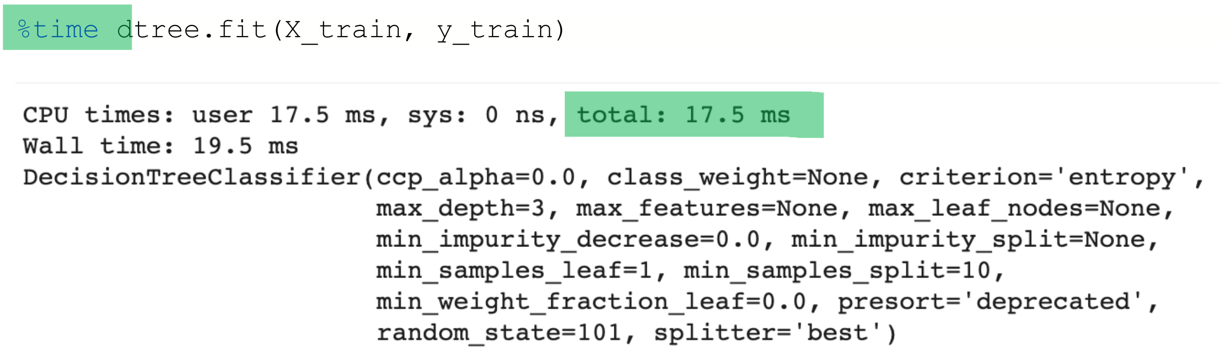
- My GitHub link: Assignment Folder CA03 https://github.com/mkaiser6/Intro_to_ML/tree/main/CA03

Q.1.1 Why does it makes sense to discretize columns for this problem? AND What might be the issues (if any) if we DID NOT discretize the columns.

It makes sense to discretize columns (numerical values --> discrete categories) for this problem because some columns like e.g., age contain too many values so that the algorithm struggles to identify meaningful/intersting patterns in the data.

Putting the values in ordered and discrete buckets/bins helps the algorithm to deal with outliers, skewness and to decrease entropy (degree of information disorder) -- observations in bins are more similar.

Q.8.1 How long was your total run time to train the model?



```
%time dtree.fit(X_train, y_train)
```

```
CPU times: user 17.5 ms, sys: 0 ns, total: 17.5 ms
Wall time: 19.5 ms
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=3, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=10,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=101, splitter='best')
```

Q.8.2 Did you find the BEST TREE?

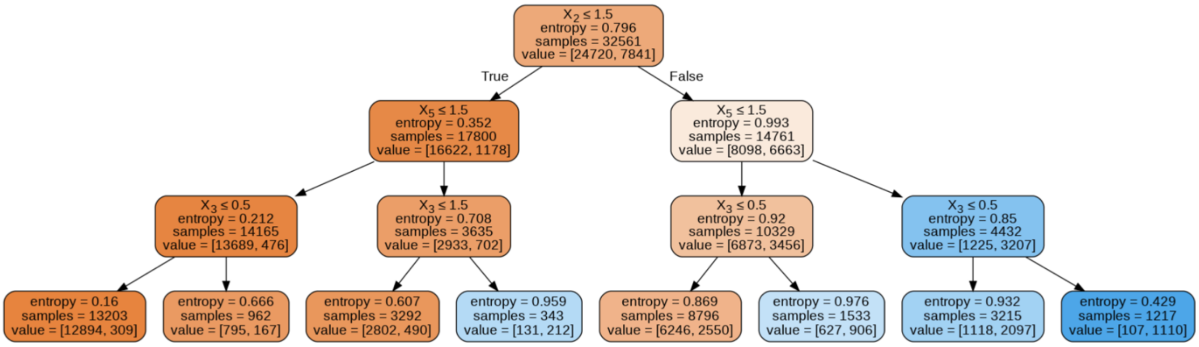
Parameters for the best performing tree in terms of (accuracy, precision, Recall, F-1 Score -balance)

```
dtree = 
DecisionTreeClassifier(min_samples_split=10,min_samples_leaf=1,max_dept
h=8,random_state=101,max_features=None,criterion="entropy")
```

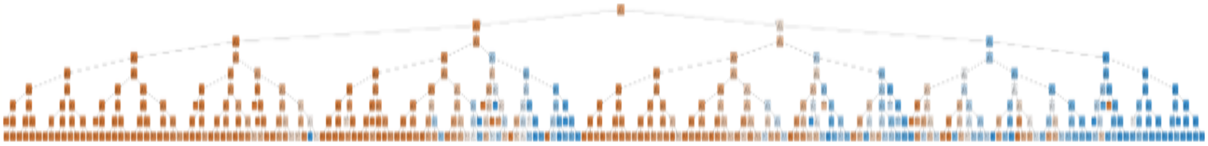
BUT in order to reduce complexity and size of the tree, the pre-pruned model with max_depth of 3 is less complex, explainable, and easy to understand than the previous decision tree model plot.

Q.8.3 Draw the Graph of the BEST TREE Using GraphViz

max_depth = 3



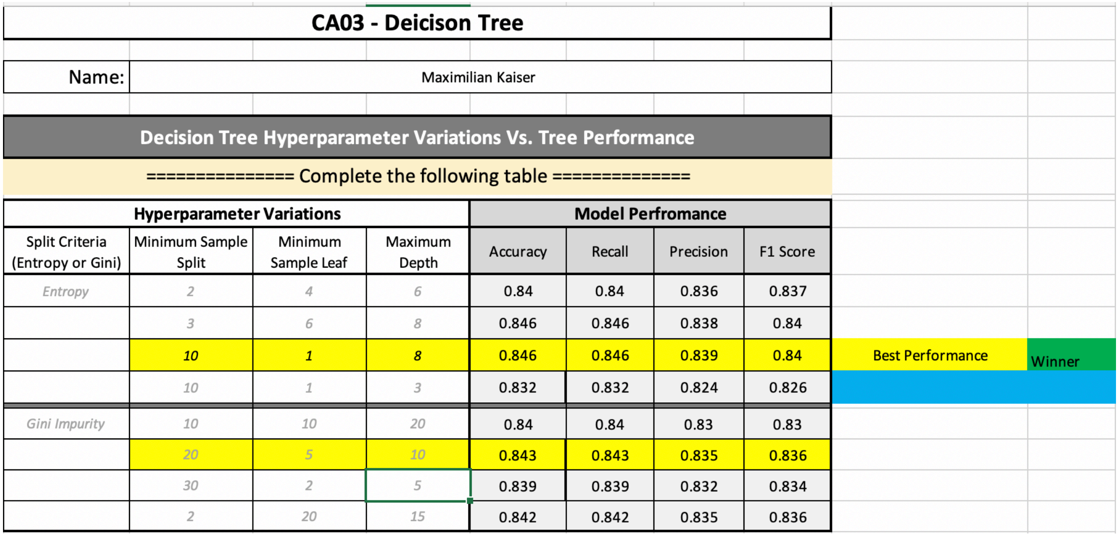
Compared to max_depth = 8



- Too complex
- This unpruned tree is unexplainable and not easy to understand
- higher value of maximum depth → overfitting

Q.8.4 What makes it the best tree?

The F-1 score which represents how precision and recall are balanced is pretty high with 0.84 (better model). Usually, anything better than 0.8 is considered good. Accuracy is also pretty high with 84,6% correct prediction. We don’t want 100% accuracy (over-fitting) on the training dataset.



CA03 - Deicison Tree

Name: Maximilian Kaiser

Decision Tree Hyperparameter Variations Vs. Tree Performance

=============== Complete the following table ==============

Hyperparameter Variations				Model Perfromance					
Split Criteria (Entropy or Gini)	Minimum Sample Split	Minimum Sample Leaf	Maximum Depth	Accuracy	Recall	Precision	F1 Score		
Entropy	2	4	6	0.84	0.84	0.836	0.837		
	3	6	8	0.846	0.846	0.838	0.84		
	10	1	8	0.846	0.846	0.839	0.84	Best Performance	Winner
	10	1	3	0.832	0.832	0.824	0.826		
Gini Impurity	10	10	20	0.84	0.84	0.83	0.83		
	20	5	10	0.843	0.843	0.835	0.836		
	30	2	5	0.839	0.839	0.832	0.834		
	2	20	15	0.842	0.842	0.835	0.836		

Q.10.1 What is the probability that your prediction for this person is accurate?

The prediction for the new person is 84% accurate. (see Google Colab)