Task 3

Discuss what will happen if you decide to change the splitting criterion. Explain the new splitting criterion and how it might change your decision tree.

Changing the splitting criterion to Gini impurity, the decision tree can massively change.

Information gain is biased towards attributes with many distinct values. Gini impurity on the other hand, is biased towards attributes with fewer values. This will result in the root of the decision tree changing (as it has a lot of distinct values) and hence the whole decision tree would change as we would be biased towards features with the fewest distinct values.

Explain whether your evaluation method can indicate whether your tree is over- or underfitting.

Cross validation can help us determine if our tree is overfit. Our tree is trained on k-1 folds(from the training dataset) and then tested on the validation fold. Then we perform the same step but using folds from our full dataset. If the validation set result in a high level of accuracy but the validation set results on the full dataset perform poorly, then we are able to deduce that our tree is over fit. If it performs poorly on both validation tests then we know that it is underfit.