COMP9417 HomeWork 2

Ping GAO z5163482

March 29, 2020

1 Q1

1.1 Part A

DecisionTreeClassifier

Dataset	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
australian	72.61%	74.35%	75.36%	77.39%	77.83%	79.71%	83.77%	81.16%	80.72%	83.48%
balance-scale	69.92%	75.04%	69.12%	74.24%	74.40%	75.52%	78.08%	75.68%	77.92%	76.64%
hypothyroid	94.94%	96.31%	97.77%	99.18%	99.20%	99.42%	99.42%	99.52%	99.34%	99.20%

BernoulliNB with priors

Dataset	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
australian	73.48%	79.86%	81.45%	80.43%	79.71%	79.86%	79.86%	81.16%	82.17%	81.88%
balance-scale	46.08%	46.08%	46.08%	46.08%	46.24%	46.08%	46.08%	46.24%	46.24%	46.08%
hypothyroid	91.38%	91.81%	92.23%	92.23%	92.23%	92.26%	92.23%	92.23%	92.23%	92.23%

1.2 Part B

I think (3), (5) statements are true.

1.3 Part C

I choose (1).

2 Q2

2.1 Part A

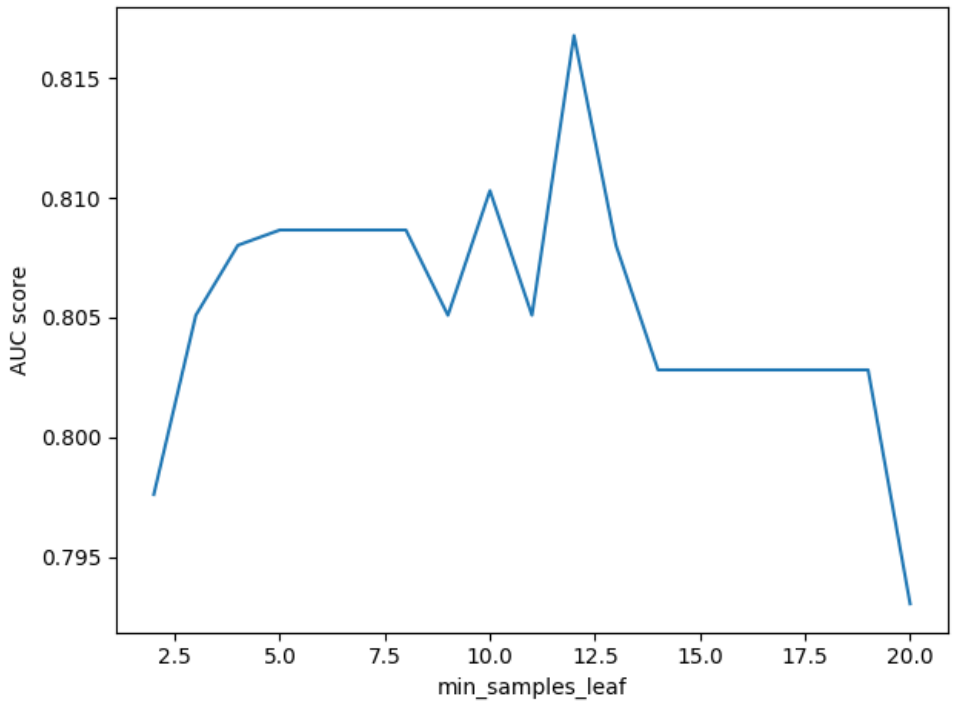
My accuracy score for the test dataset is 82.77%.
My accuracy score for the training dataset is 85.65%.

2.2 Part B

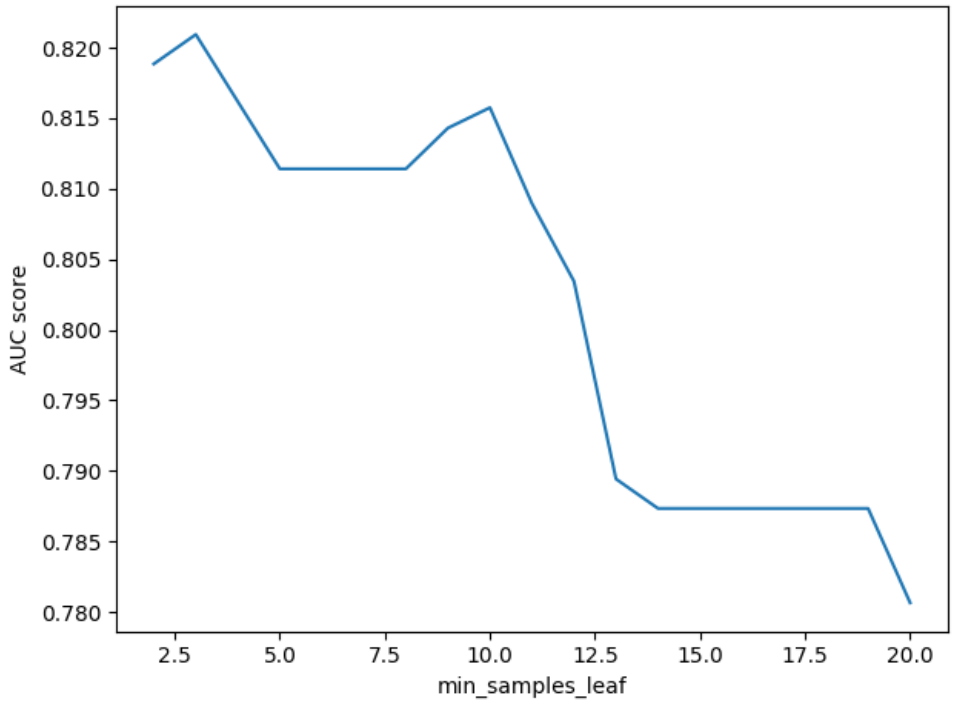
The min_samples_leaf number of 5 to 7 give the optimal result, this can be observed by compare the auc score in the part c's plots, pick the maximum score with the lowest varience.

2.3 Part C

Plot For Test Dataset



Plot For Train Dataset



2.4 Part D

We make the asumption that 'Sex' and 'Pclass' are independent feature.
Thus, (S=true $|G = female, C = 1) = P(S = true|G = female) * P(S = true|C = 1)$
$P(S = true|G = female) = P(G = female) \cap P(S = true)/P(G = female)$
$= 109/573$
$P(S = true|C = 1) = 136/216$
$P(S = true|G = female, C = 1) = 11.98\%.$