STAT 6650 Homework 2

María Victoria Liendro
Master's in Data Engineering
mzl0201@auburn.edu

Spring 2025

1) c) In Figure 1 we can see the accuracy for different values of maximum depth (from 1 to 6) and minimum samples (from 1 to 15). In blue we can see the accuracy in the training set and in orange the accuracy in the testing set The accuracies obtained go from 0.93 to 0.98.

1) d) For my train function, if we consider the worst-case scenario where all samples have a different value for every feature, the complexity is $O(n \log n)$ per feature per split. For each split I am using recursion, so in the scenario where my tree is full the complexity will be $O(2^p)$. Finally, it will be:

$$O(d \cdot n \log n \cdot 2^p)$$

For my predict function the complexity is $O(m * p)$, where m is the number of samples to test and p is the path from the root to the leaf, worst-case it will be the maximum length. Then:

$$O(m * p)$$

2) d) Table 1 shows the values for the AUC in the training and test set and the computation time, across the three methods. kfoldCV got the highest testAuc but it was the slowest one, being four times slower that holdout and two times MonteCarloCV. The parameters used were:

- `max_depth` = 4
- `min_samples` = 5
- `testSize` = 0.2
- `s` = 5
- `k` = 5

3) d) Table 2 shows that KNN was stable through all the training sizes, however Decision Tree, performed better at every size. Surprisingly, the best performance was for the Tree using 0.90 of the training set. Probably using the whole dataset the model was overfitting.

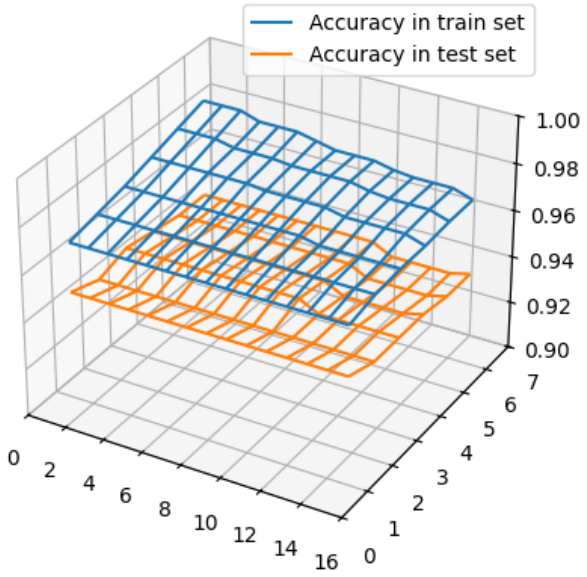

Figure 1: Figure 1

	trainAuc	**testAuc**	**timeElapsed**
holdout	0.6485	0.5195	0.0724
kfoldCV	0.6348	0.5364	0.2980
MonteCarloCV	0.6485	0.5195	0.1685

Table 1: Performance metrics for different methods.

Model	**Training Set Size**	**AUC**	**Accuracy**
KNN	100%	0.5308	0.9354
KNN	99%	0.5308	0.9354
KNN	95%	0.5308	0.9354
KNN	90%	0.5308	0.9354
Decision Tree	100%	0.5341	0.9417
Decision Tree	99%	0.5341	0.9417
Decision Tree	95%	0.5341	0.9417
Decision Tree	90%	0.5736	0.9479

Table 2: Performance of k-NN and Decision Tree models with different training set sizes.