Assignment 3 - Decision Trees

I. Choice of Data:

'Drugs A, B, C, X, Y for Decision Trees' imported from Kaggle website (IBM, 2021). The dataset belongs to lists of patients who received a treatment for the same disease. During their course of treatment, each patient responded to one of 5 medications, Drug A, Drug B, Drug c, Drug x and Drug y. The goal here is to build a model that can find out which drug might be appropriate for a future patient with the same illness. The features of this dataset are age, sex, blood pressure (refer to as BP), cholesterol and the sodium - potassium level (refer to as Na_to_K) of the patients, and the target is the drug that each patient responded to.

Reasons to use my DT implementation:

- Feature predicated have 5 unique categorical values. Therefore, the problem is building a multiclass classifier not a regression problem. Then, constructing a decision tree is a solution of such problem.
- Age and Sodium - Potassium level features are numerical (integer and float respectively). My DT implementation can handle features with these types.
- The other columns, sex, cholesterol, and blood pressure have categorical values which is perfectly fine for my implementation.

Database chosen suitable for a DT approach

- dataset is relatively small (only 200 lines) which is the kind of data where using a decision tree is recommended.
- Dataset has only 5 features. Using a DT will help provide better interpretability in the series of decision made to reach to the final classification.
- My DT doesn't require any prior data preparation, reducing the time and complexity it takes between data injection, model training then when using the model for prediction.

Data analysis:

observations about the dataset:

	Sex	BP	Cholesterol	Drug
count	200	200	200	200
unique	2	3	2	5
top	M	HIGH	HIGH	drugY
freq	104	77	103	91

	Age	Na_to_K
count	200.000000	200.000000
mean	44.315000	16.084485
std	16.544315	7.223956
min	15.000000	6.269000
25%	31.000000	10.445500
50%	45.000000	13.936500
75%	58.000000	19.380000
max	74.000000	38.247000

- Almost equal number of male and female patients.
- 91 out of 200 patients were cured using `drug y` which makes the decision more poised to this drug over the 4 other drug (imbalanced data).
- Almost equal number of patients has either high or low cholesterol level.
- Patients with normal cholesterol level did not use drug C.
- Average group age is 44 years.
- Average Sodium Potassium level is around 16.
- Drug A has been the cure for patients who are less than 51 years old whereas drug B was the cure for patients who are 51 years old and older.
- Patients with normal blood pressure were cured using either drug X or drug Y.
- Drug Y has been the cure for all patients who have na_to_k > 18.269. This makes na_to_k a key variable.

II. Decision Tree and accuracy:

Produced decision tree:

```
Decision Tree for drug dataset is as follows:
 Internal_Split_Node(attribute=('Na_to_K', 14.5175), subtrees={'larger_than_14.5175': Leaf(label='drugY', Samples=9
1), 'lower_or_equal_to_14.5175': Internal_Split_Node(attribute='BP', subtrees={'LOW': Internal_Split_Node(attribute
='Cholesterol', subtrees={'HIGH': Leaf(label='drugC', Samples=16), 'NORMAL': Leaf(label='drugX', Samples=17)}, sample
s=33, label_distribution={'drugX': 17, 'drugC': 16}, default_value=('drugX', 17)), 'NORMAL': Leaf(label='drugX', Samp
les=36), 'HIGH': Internal_Split_Node(attribute=('Age', 50.5), subtrees={'larger_than_50.5': Leaf(label='drugB', Sampl
es=16), 'lower_or_equal_to_50.5': Leaf(label='drugA', Samples=23)}, samples=39, label_distribution={'drugA': 23, 'dru
gB': 16}, default_value=('drugA', 23))}, samples=108, label_distribution={'drugX': 53, 'drugA': 23, 'drugC': 16, 'dru
gB': 16}, default_value=('drugX', 53))}, samples=200, label_distribution={'drugY': 91, 'drugX': 54, 'drugA': 23, 'dru
gC': 16, 'drugB': 16}, default_value=('drugY', 91))
This tree got an average accuarcy of 0.99 when running cross validation on 5 folds.
Accuracy this tree has on each fold is: [1.0, 1.0, 0.975, 0.975, 1.0]
```

Decision Tree Analysis (Hypotheses vs. produced DT)

- first branching of the resulting tree is per Sodium - Potassium (na_to_k). This split has been already observed in Figure 1. But instead of using 18.269 as a threshold, my tree uses 14.5175.
- Tree confirms my observation regarding patients having na_to_k > 14.5175. Tree is getting directly into a leaf node where patients are recommended to take drug Y.

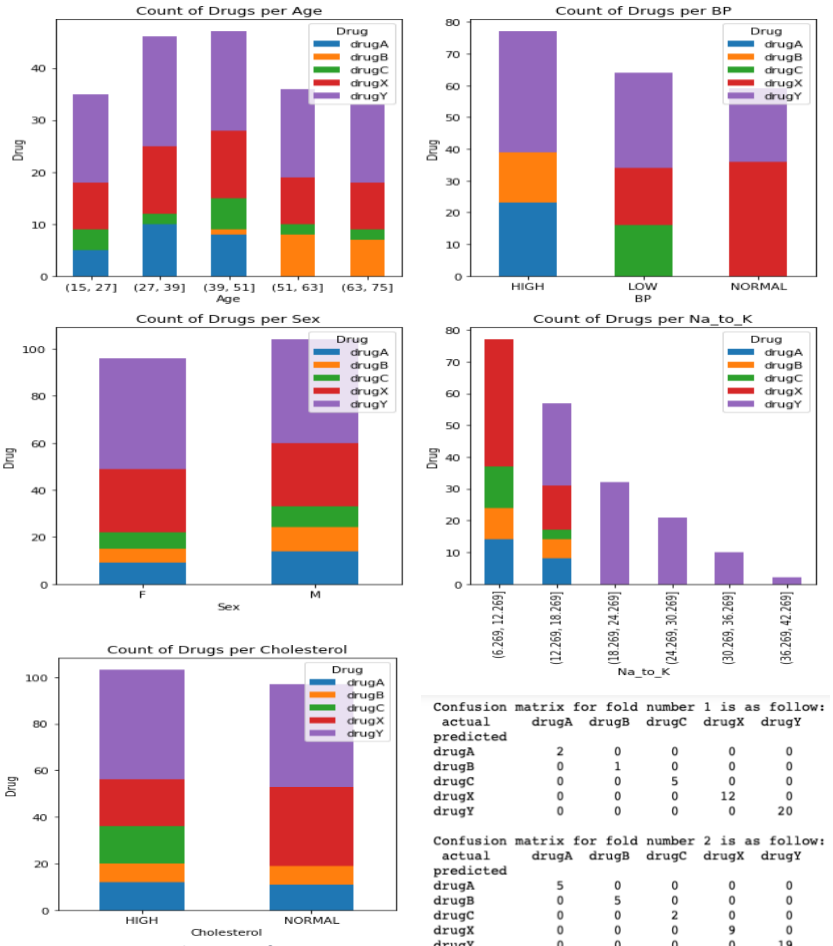


Figure 2 Bar chart per feature

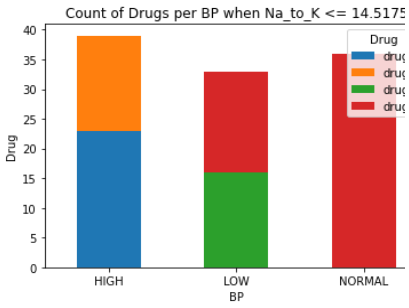


Figure 1

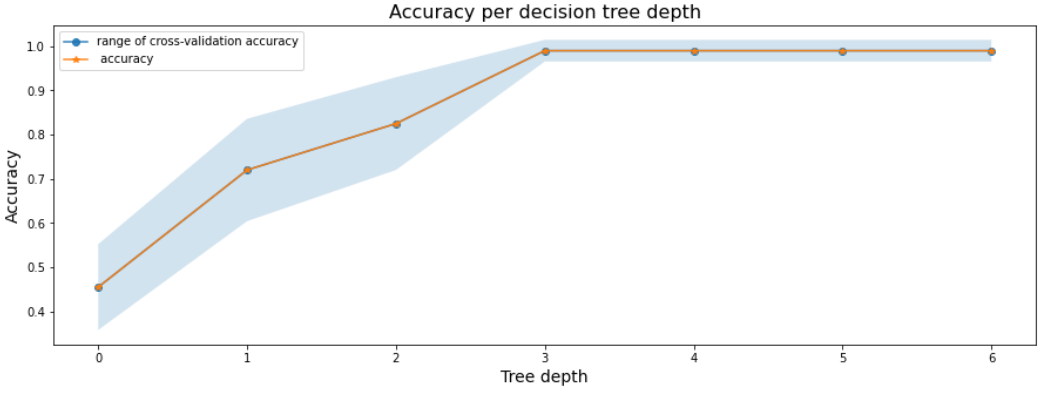
```
Confusion matrix for fold number 1 is as follow:
 actual     drugA  drugB  drugC  drugX  drugY
predicted
drugA           2      0      0      0      0
drugB           0      1      0      0      0
drugC           0      0      5      0      0
drugX           0      0      0     12      0
drugY           0      0      0      0     20

Confusion matrix for fold number 2 is as follow:
 actual     drugA  drugB  drugC  drugX  drugY
predicted
drugA           5      0      0      0      0
drugB           0      5      0      0      0
drugC           0      0      2      0      0
drugX           0      0      0      9      0
drugY           0      0      0      0     19

Confusion matrix for fold number 3 is as follow:
 actual     drugA  drugB  drugC  drugX  drugY
predicted
drugA           4      0      0      0      0
drugB           1      3      0      0      0
drugC           0      0      3      0      0
drugX           0      0      0     12      0
drugY           0      0      0      0     17

Confusion matrix for fold number 4 is as follow:
 actual     drugA  drugB  drugC  drugX  drugY
predicted
drugA           6      0      0      0      0
drugB           0      5      0      0      0
drugC           0      0      3      0      0
drugX           0      0      0     10      0
drugY           0      0      0      1     15

Confusion matrix for fold number 5 is as follow:
 actual     drugA  drugB  drugC  drugX  drugY
predicted
drugA           5      0      0      0      0
drugB           0      2      0      0      0
drugC           0      0      3      0      0
drugX           0      0      0     10      0
drugY           0      0      0      0     20
```

Figure 3

- When Sodium - Potassium level is lower or equal to 14.5175, drug X is suitable for patients that also have a normal BP, see Figure 2.
- For patients with na_to_k < 14.5 and low BP, cholesterol level is the determining factor. Drug C is suitable when cholesterol is High on the other hand drug X is recommended to those with normal cholesterol level.
- For patients with na_to_k < 14.5 and high BP, age is the determining factor. if patient is more than 50.5 years old, then drug B is recommended, otherwise, patient should get drug A. This confirm one of the observations I have.

Reflections and Takeaways:

This tree split almost evenly the drug data according to Sodium - Potassium level. This level is the most determining factor as it defines whether patient is more likely to be cured using drug Y. Because of this imbalance in data, I used a confusion matrix, see Figure 3, to check number of cases where drug y has been falsely recommended. Surprisingly, tree has only 1 misclassified case (out of 109 cases in all folds) where tree recommends that patients take drug Y and ground truth says that he took drug X. with Tree depth of 0, a leaf node will suggest a drug Y and we get a 0.45 accuracy. The accuracy increases with tree depth and hit a plateau (0.99) when tree depth is 3.



Bibliography:

IBM, 2021. *Drugs A, B, C, X, Y for Decision Trees*. *Kaggle* [Online]. Available from: https://www.kaggle.com/datasets/7f7bebbb020855aa344d19f2ce710e0128ec750263fc868840aa835eaf4cec05 [Accessed 11 December 2022].