COMPUTER SCIENCE

Assignment 3

Lecturer: Meabh Loughman
Computer Science – Artificial Intelligence Basics
Group: BSC109224 A
Semester 2

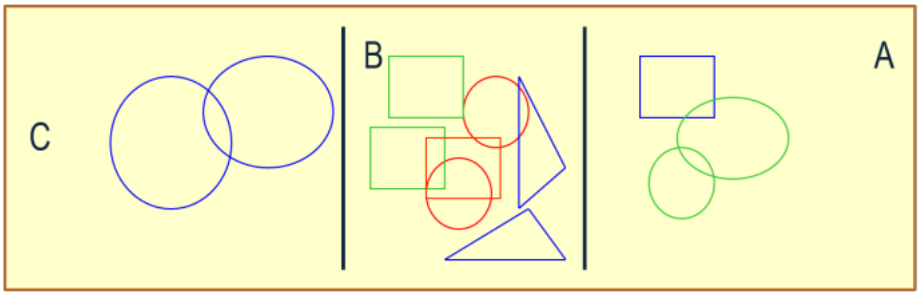
Lucas Madeira Maranho
76990@student.dorset-college.ie

SUMMARY

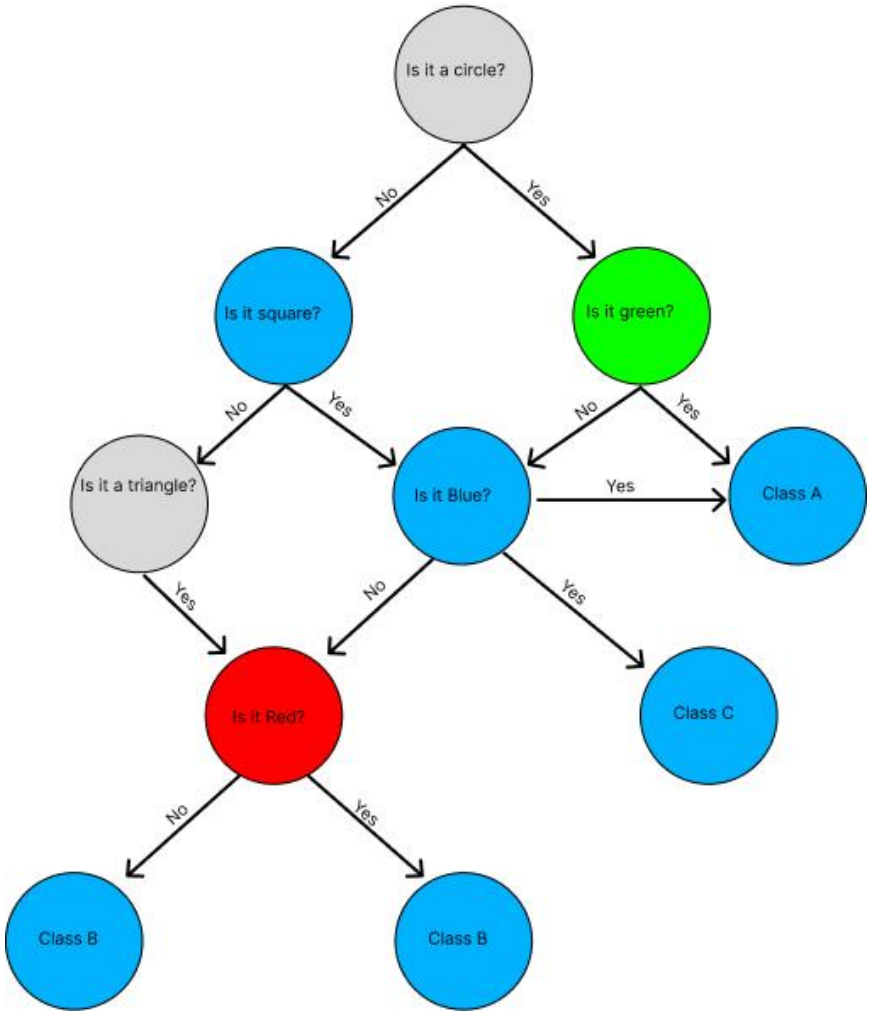
Assignment 3

Week beginning 10/02/25

Q1. Draw me a decision tree to classify these labelled data sets A, B and C



Tree for classification for A, B and C



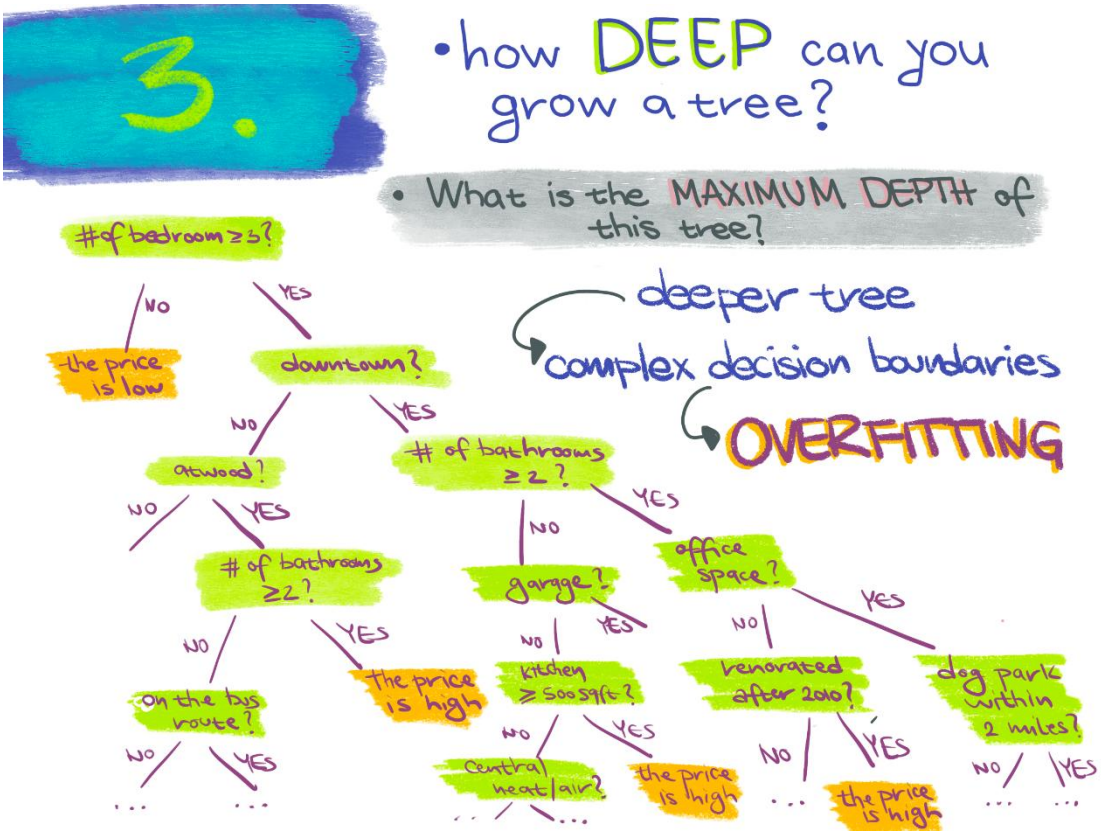
Author: Lucas Madeira Maranho 18/02/2025, Figma. (figure1)

As can you see in the image, decision tree is classified from top to the bottom creating branches, where the first question is about the format if it is a circle, the answer is yes or no, which split into two situations and the following step if it is green or not but if the answer is no, the next question is going to be about another format, which the question is if it is a square and so progressively, reaching the final classification, being A, B or C.

Q2. Are there any problems with using a decision tree algorithm.

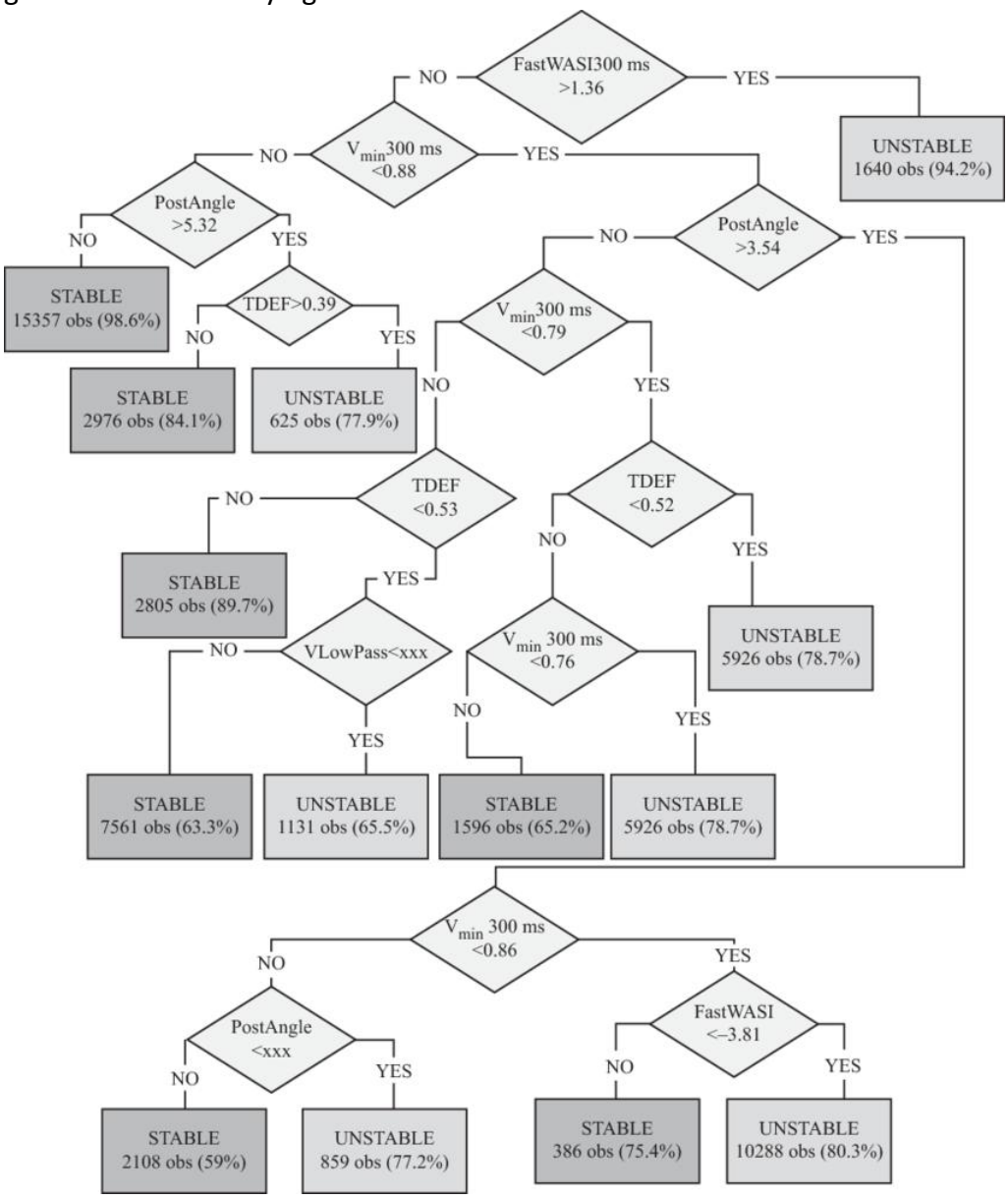
Yes, there are some disadvantages in using a decision tree algorithm:

1 – Overfitting - which means the data struggle in model complex and doesn't work well. Causing instability in small variations in data. As you can see in the image below figure (2). You just put information turning deeper and complex to understand.



Machine learning for biologists: Setup – Machine Learning for Biologists. Available at: https://carpentries-incubator.github.io/ml4bio-workshop/setup.html (Accessed: 18 February 2025). (Figure2)

2 – unstable - Decision tree can be unstable, because a small variation in the input can give a large change in output causing different results of the tree that was generated. For example, in the image below (Figure 3), it took 21 Decision tree generated for classifying stable OK = -1 and unstable cases ok = 1



Kamwa, Innocent. (2012). Wide-Area Early Warning Systems. (Figure 3)

3 – Not balanced – It is recommended to balance dataset prior to fitting in decision tree.

4 – Poor Performance – This kind of algorithm struggle when is related with linear relationship that are better constructed by algorithm like logistic regression or SVMs.

Q3. Identify a use of AI in a different industry, for example MYCIN blood infection system in medicine. Are there any ethical considerations for this specific use case?

Machine learning models based on fraud detection are trained in different data to identify anomalies, for example: unauthorized transactions.

Ethical Considerations:

1 – Bias in Ai Models – If the training data is biased, AI could target specific demographics disproportionately, leading to unfair discrimination.

2 – Lack of transparency – making it difficult to understand how they make decisions.

3 – Privacy Concerns – Systems require access to large amounts of personal data, creating doubt about data security and privacy.

Q4. Is there a better algorithm to use for classifying the data as in question 1?

There are some better algorithms to use for classifying a data instead of using decision tree algorithm.

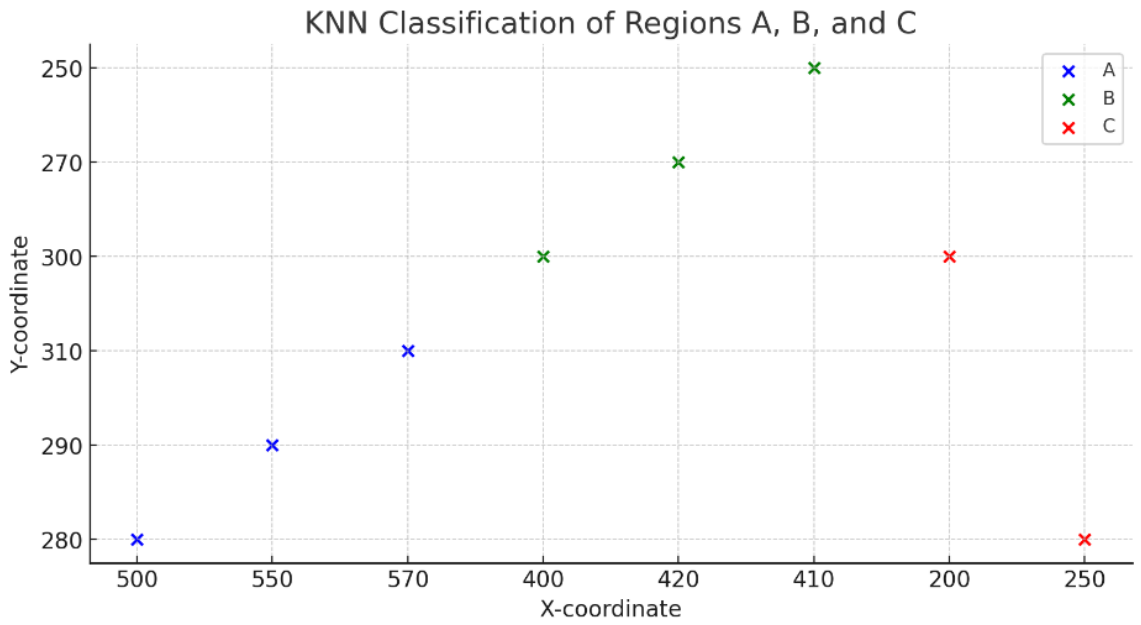
Decision tree algorithm is easy to interpret and understand, but most of the scientists use random forests, because only one tree is not enough to get an effective result.

Random forest is a version of decision tree that minimize overfitting.

Support Vector Machine (SVM) – Decisions non-linear, works very well with geometric classification.

Neural Networks (Deep learning) – If the data set becomes very complex.

K-nearest Neighbours (KNN) – Very useful if the classification depends on the local equality or similarity.



ChatGPT (2025) *K-Nearest Neighbours (KNN) Classification of Image Regions*. Generated using Python and Matplotlib. (Figure 4)

As you can see in the image above, I asked to ChatGPT create an image based on the image given from exercise 1, just to compare and show it.

In the K-Nearest Neighbours (KNN) classification by regions A, B and C focus in nearest data points.

KNN uses distance-based classification, changing shapes or small changes in position can turn another result while Decision tree create rules to separate the dataset depending on feature importance.

Decision tree would produce hard straight line and KNN would be more a fluid classification.

However, based on the question, my answer is yes, there are better algorithms to use for classifying the data, but in this case because the dataset is small and clearly structured, a decision tree is sufficient.

References:

https://scikit-learn.org/stable/modules/tree.html

https://www.scribbr.com/methodology/research-ethics/

https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=Building%20a%20Random%20Forest%20Model,-Let's%20see%20a&text=Here%2C%20we%20can%20see%20that,out%2Dof%2Dsample%20evaluation.