DI501 INTRODUCTION TO DATA INFORMATICS
FINAL EXAM
PART 2
31 JANUARY 2022 (Duration: 90 minutes)

IMPORTANT:

In the final exam, you are not allowed to communicate with your classroom mates for any reason. You cannot discuss any matter with your friends during the exam. It is also strictly forbidden that you seek advice or feedback from any professionals in the domain, previous year students or any other. You are required to answer the questions alone. You are not allowed to give or receive any form of exam aid to any other student in this course during the exam. Any questions should be directed to the exam proctor via e-mail or Microsoft Teams course section.

If you need to scan your results on the paper, please use a professional application such as Office Lens with a good resolution. The image should only show your paper (not your desk, arm, fingers, pencils etc.) without any distortion. If your image does not satisfy these criteria and the document is not readable, you will get zero points. You need to crop the image if it's necessary.

Question 1 (25 pts):

Figure (a)

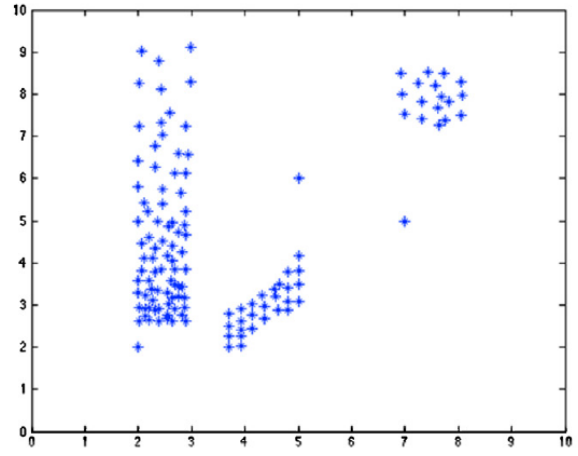
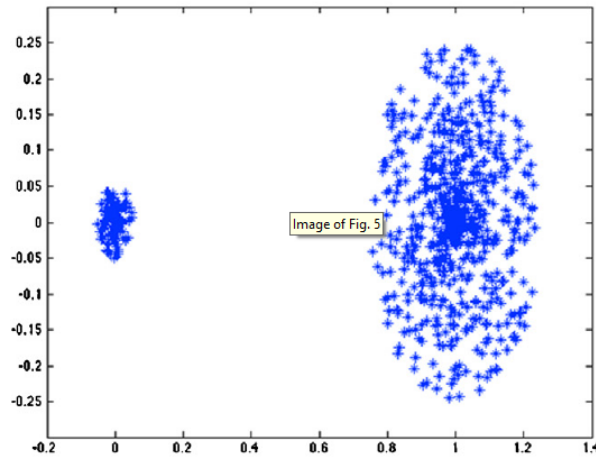


Figure (b)



Consider the two datasets shown in Figures while answering the following two (a and b) questions:

a) **(8 pts)** Compare approximately the average silhouette scores for k=3 and k=6 respectively using k-means algorithm (which one will most likely to be less than the other, you don't need to compute them) in Figure (a). Give your reasons. Show how the clusters might be formed on the images. Ensure to have two Figure (a) (copy and paste it) images to show clusters separately.

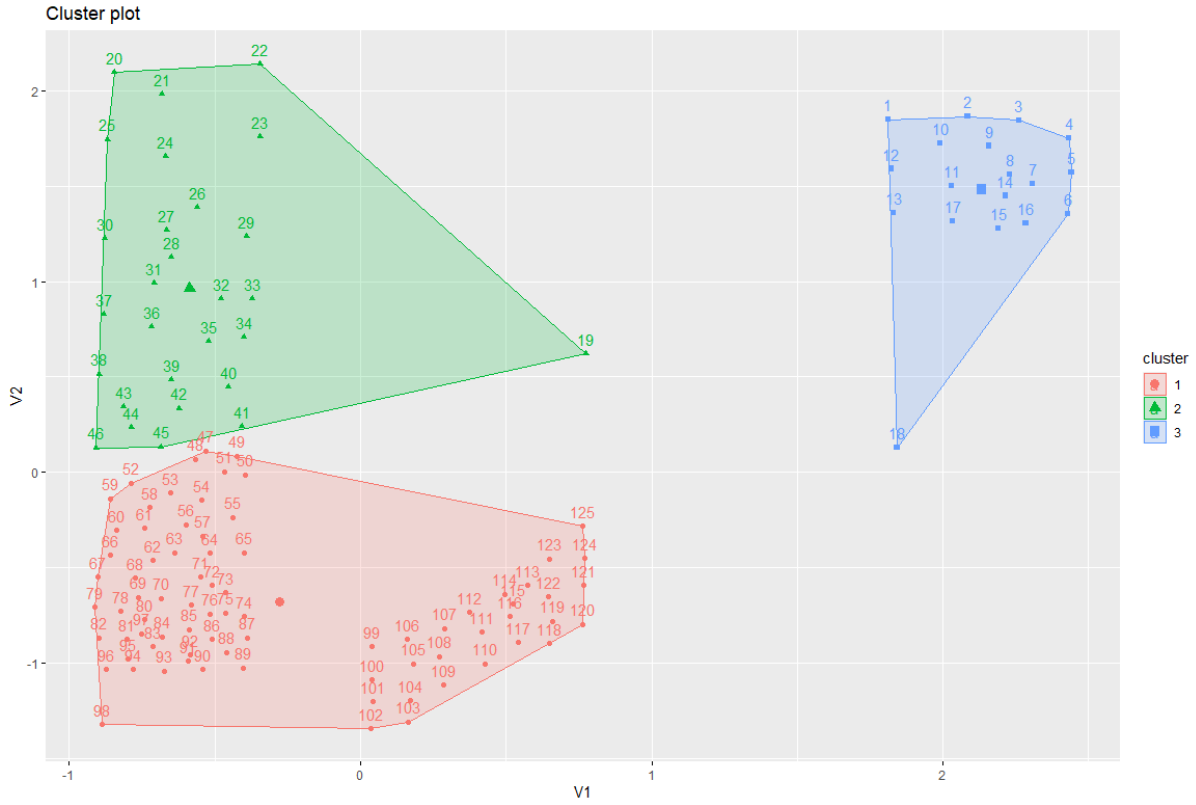
Note: Think whether outliers should be included in the clusters with the standard (vanilla) k-means algorithm. Moreover, think whether a cluster with one point cluster can be formed with this algorithm.

b) **(8 pts)** How many clusters approximately do you anticipate to have when eps= 0.2 and minpts=3 are chosen using density-based clustering in Figure (b)? What might happen when eps = 0.05 and minpts=3 are chosen?

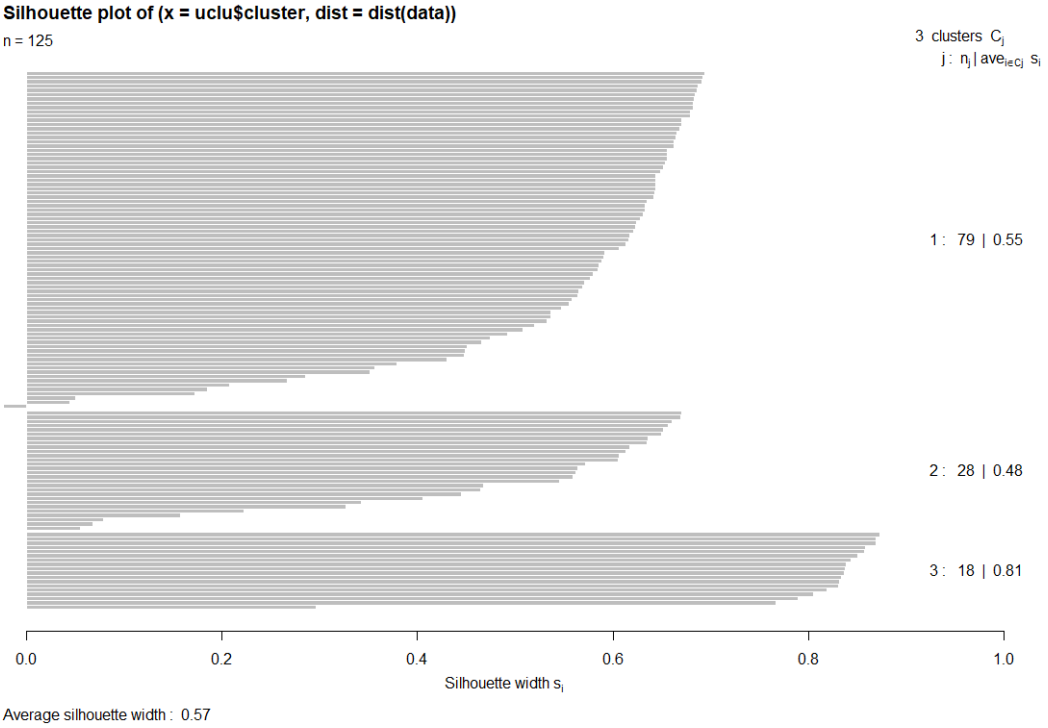
c) **(9 pts)** Does silhouette analysis work successfully for all types of cluster shapes? Justify your answer. Give an example for demonstration.

Answer 1:

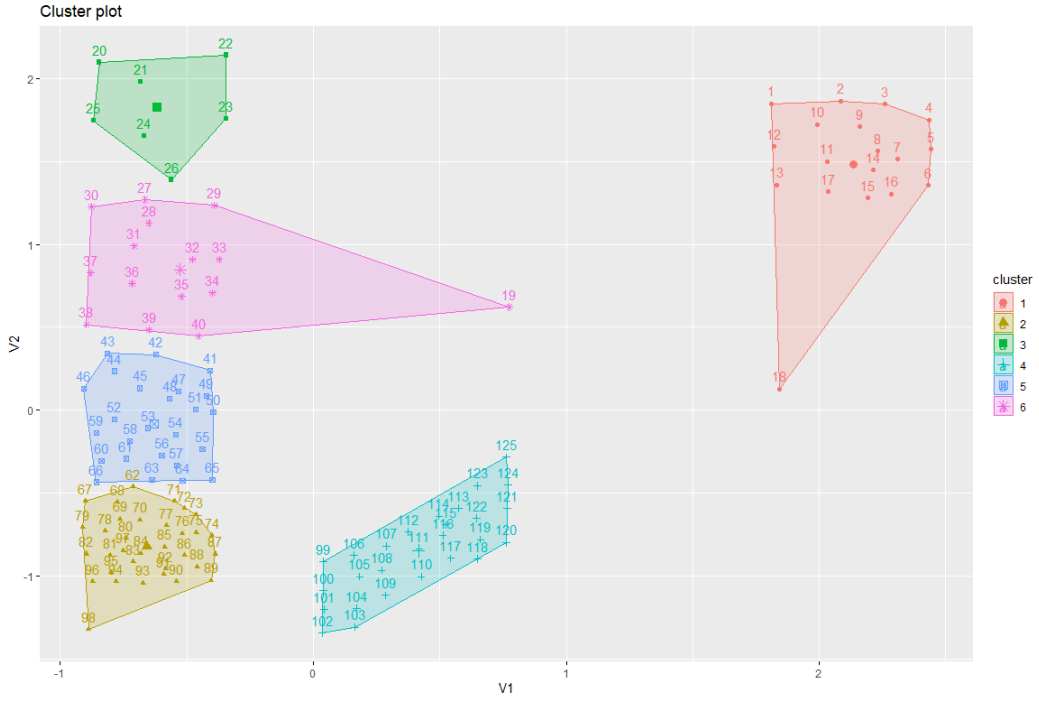
a) If we select k=3, it will look like this:



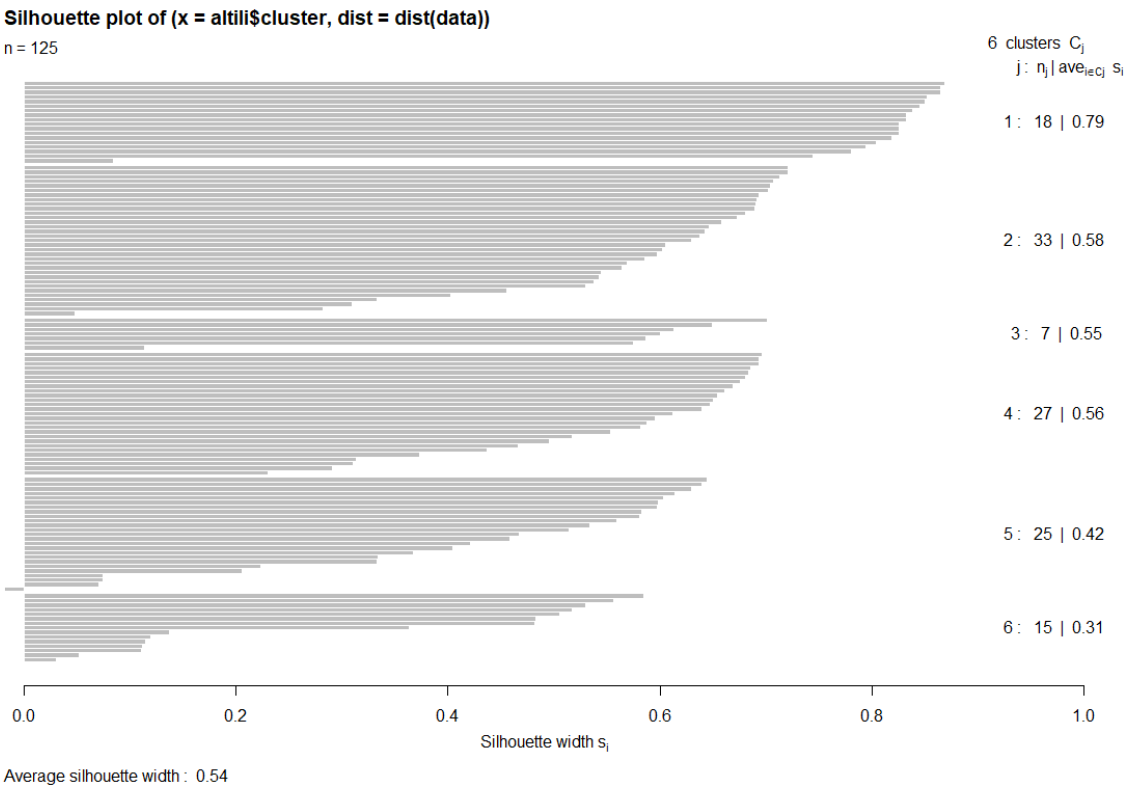
The silhouette score results:



All the outliers will be assigned to one of the clusters. If we select k=6, it will look like this:

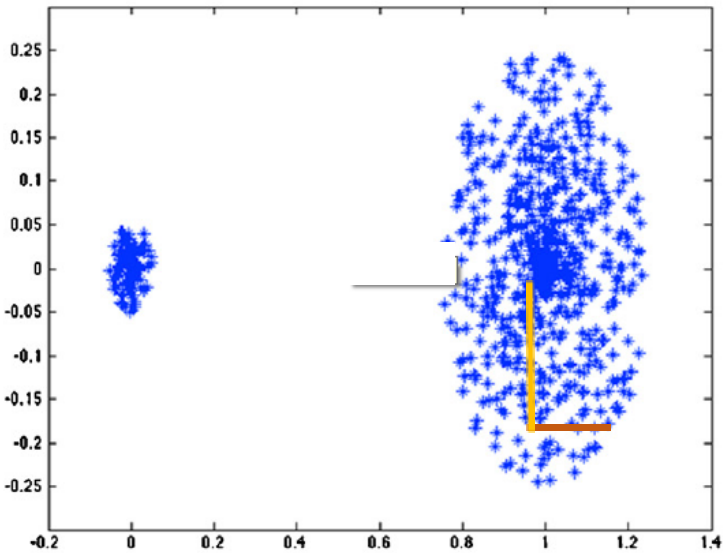


It is probable to end up with slightly different clusters such as single points assigned to different clusters. However, I wanted to ensure to end up with the best configurations. Therefore, I tried 200 different initializations in order to find the best clusters. You do not have this chance in this exam but it is obvious that the data points lying within x=2-3 range and y=2-9 range, will definitely split up more when k=6 is chosen instead of k=3 thus affecting the silhouette scores negatively. As confirmed in the below figure, many points appear to be close to the other clusters (with negative values). As a result, the average silhouette value will be a little less than the first one. Note that few outliers will not significantly affect the scores in the first case (when k=3).

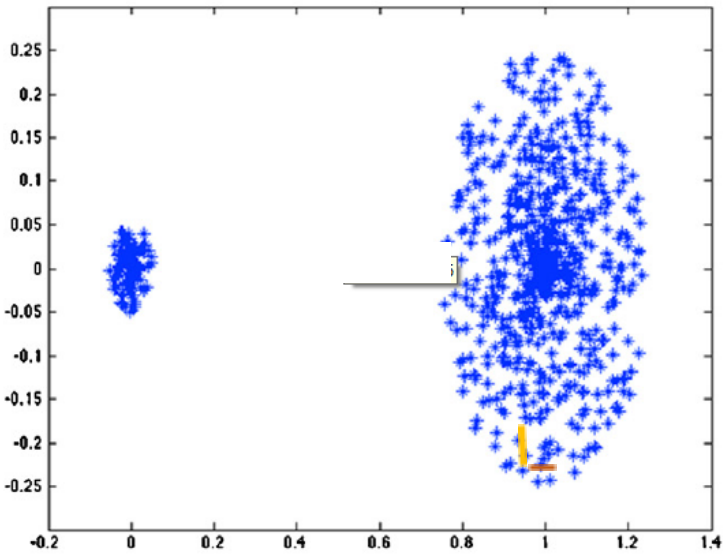


Outliers can be formed into a separate cluster.

b) When eps= 0.2 and minpts=3 are chosen using density-based clustering, there will be two clusters. Both axes are not shown as equally spaced visually but you should take into consideration the values on the axes (see yellow and orange lines). It shows that radius is quite large to include many data points.



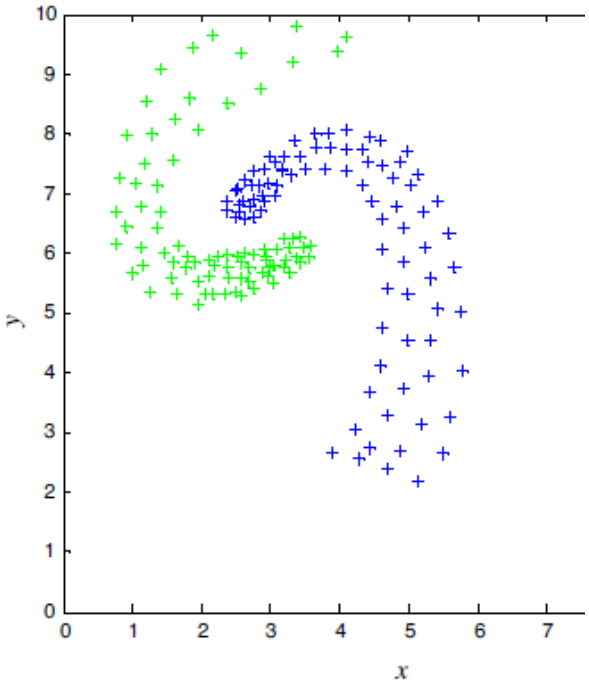
If we use eps = 0.05 and minpts=3, the yellow and orange lines implying the radius is shown on the Figure.



The chosen epsilon value still covers a lot of instances in the radius. Therefore, we expect no change. If we have reduced the eps significantly, then we would have seen that these clusters will be split into smaller meaningless clusters.

If you state that the number of clusters might probably increase due to eps decrease, I gave 4 pts.

c) Silhouette analysis cannot work successfully for all types of cluster shapes particularly with non-convex clusters. Because it considers the inter and intra cluster distances. For instance, consider the following dataset:



Instances locating in the denser part of each cluster are closer to each other but far from the centroid. Hence, (b-a) will result in a negative value.

Question 2 (20 pts):

a) Explain whether overfitting or underfitting is possible with cross-validation technique. Give your reasons.

b) Assume that you would like to predict the popularity of users given that you have mixed type variables in your dataset. The target variable is highly skewed continuous variable.

   i. Discuss the consequences of model fitting using 90% uniformly random selected training dataset and the remaining 10% for your testing dataset. Give at least two consequences as examples.

   j. Discuss the consequences of model fitting using nested cross validation based on this dataset. Give at least one negative consequence as an example. Propose a solution for this problem.

Answer 2:

a) **(10 pts)** Underfitting is possible with cross-validation. For instance, if the dataset is significantly imbalanced, certain instances might only appear in certain folds (maybe appearing in testing dataset only). Then, the model will not have a chance to learn these infrequent data points. In addition, if the hyperparameters are not tuned/determined very well, then you might have a underfitting problem.

Overfitting can be prevented with k-fold cross validation to some degree. However, it is still possible to observe it in your model. For instance, if your model is very complex and you use a static fold (not like in nested cross validation) which is very similar to your training dataset, you can still overfit to the training data. But this final model will fail to generalize to testing dataset.

If you said yes for both, I gave you 2 pts. But if you failed to give a correct reasoning, you get no points.

b) **(10 pts)** i) It is possible not to have sufficient data points representing the minority classes such as very popular users with millions of followers in your validation and/or testing dataset when they appear very few times. The model might not learn this data very well. In addition, the performance results might be biased towards the instances belonging to the majority instances such as unpopular users with few followers (overfitting to these instances).

ii) Still with nested cross validation, the same problem described above (i) might be observed. As a solution, the dataset might be (1) turned into a balanced dataset with sampling (2) split with stratified sampling to ensure instances belonging to minority class to be present in each fold (3) enriched with syntactic instances that are generated with an approach such as SMOTE.

Don't forget that although nested cross validation is computationally expensive, it is not affected by outliers or noise compared to standard k-fold scheme. Because you make use of different splits of validation dataset.

You can transform your target variable using log-transformation. It might help to a degree and still the instances appearing less frequently than others will not change. More specifically the input variable did not change. You can resemble this problem to anomaly detection.

Don't forget that we use nested cross validation to reduce bias while tuning hyperparameters across the folds. This ensures a better generalization compared to the other techniques.

Question 3 (10 pts):

Consider that our dataset includes the following input variables: V1 (binary), V2 (categorical with 3 possible values), V3 (binary). We have our Target (binary) variable as well.
We would like to run a naive Bayes classifier on this dataset.

a) How many parameters do we need to estimate to train our classifier?
b) How many parameters do we need if we don't have the naive Bayes conditional independence assumption?

Answer 3

a) **(5 pts)** For a naive Bayes classifier, we need to estimate P(Y=1),
P(V1 = 1|y = 0); P(V2 = C1|y = 0), P(V2 = C2|y = 0), P(V3 = 1|y = 0),
P(V1 = 1|y = 1); P(V2 = C1|y = 1), P(V2 = C2|y = 1), P(V3 = 1|y = 1).
Other probabilities can be obtained with the constraint that the probabilities sum up to 1.
So, we need to estimate 9 parameters (2*((2-1)+(3-1)+(2-1)))+1=9.

b) **(5 pts)** Without the conditional independence assumption, we need to estimate P(Y=1),
$P(Y = 1), P(Y = 1), P(Y = 1), P(Y = 1),$
$P(Y = 1), P(Y = 1),$
$P(Y = 1), P(Y = 1), P(Y = 1), P(Y = 1),$
$P(Y = 1), P(Y = 1),$

$P(Y = 0), P(Y = 0), P(Y = 0), P(Y = 0),$
$P(Y = 0), P(Y = 0),$
$P(Y = 0), P(Y = 0), P(Y = 0), P(Y = 0),$
$P(Y = 0), P(Y = 0),$

We need to estimate 23 parameters (2*3*2=12 of possible (V1,V2,V3)). Consider the constraint that the probabilities sum up to 1 (The ones with the yellow ones are not necessary for this purpose so they will be removed), we need to estimate 12- 1 =11 parameters for Y=1. Hence, 2 * (12-1) + 1=23 variables should be estimated.

Question 4 (10 pts):

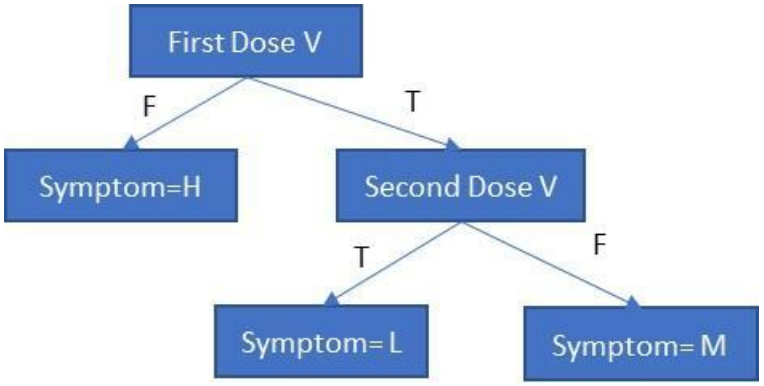
Consider the following dataset where we classify the symptoms of Covid as (low (L), medium (M) or high (H)) when people received their first dose of vaccination and the second (True (T) or False (F)).

Symptom	First Dose V	Second Dose V
L	T	T
M	T	F
L	T	F
H	F	F
H	T	F
H	F	F
M	T	F
M	T	T
L	T	T

a) Draw a decision tree according to ID3 algorithm using the above dataset. You are not supposed to show the calculations.
b) Discuss whether strongly calculated variables and redundant attributes affect the model performance (such as accuracy) of decision trees. Give your reasons.

Answer 4:

(5 pts) We draw the tree according to information gain which results in highest gain in the split (with purer nodes).



It is important to draw the tree correctly. For instance, showing the nodes and the labels over the splits should be correctly shown (in case of missing -2).
Also, in the algorithm, we repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes. Therefore, we are not allowed to use the same feature while constructing the decision tree (in case of not applying it -2).
You don't need to do the calculations since you can easily calculate how pure each split will be roughly. For instance, when First Dose V=False, all the leaves will end up with High symptoms. But this is not valid for the other feature.
While constructing the tree, I chose the label which happened to appear the most frequent in that specific branch.

(5 pts) I mistakenly wrote "strongly calculated variables". If you did not change it to "correlated" and use the term itself, there is nothing called strongly calculated variables in the literature😊 You could have said it so. It doesn't indicate the nodes resulting in higher impurity scores.
If you answer the question from the correlation perspective, the answer is as follow: The strongly correlated variables and redundant attributes do not affect the performance of the tree since it takes into consideration the best split. It will only affect the performance.