EXERCISES

You've just finished training a random forest for spam classification, and it is getting abnormally bad performance on your validation set, but good performance on your training set. Your implementation has no bugs. What could be causing the problem?

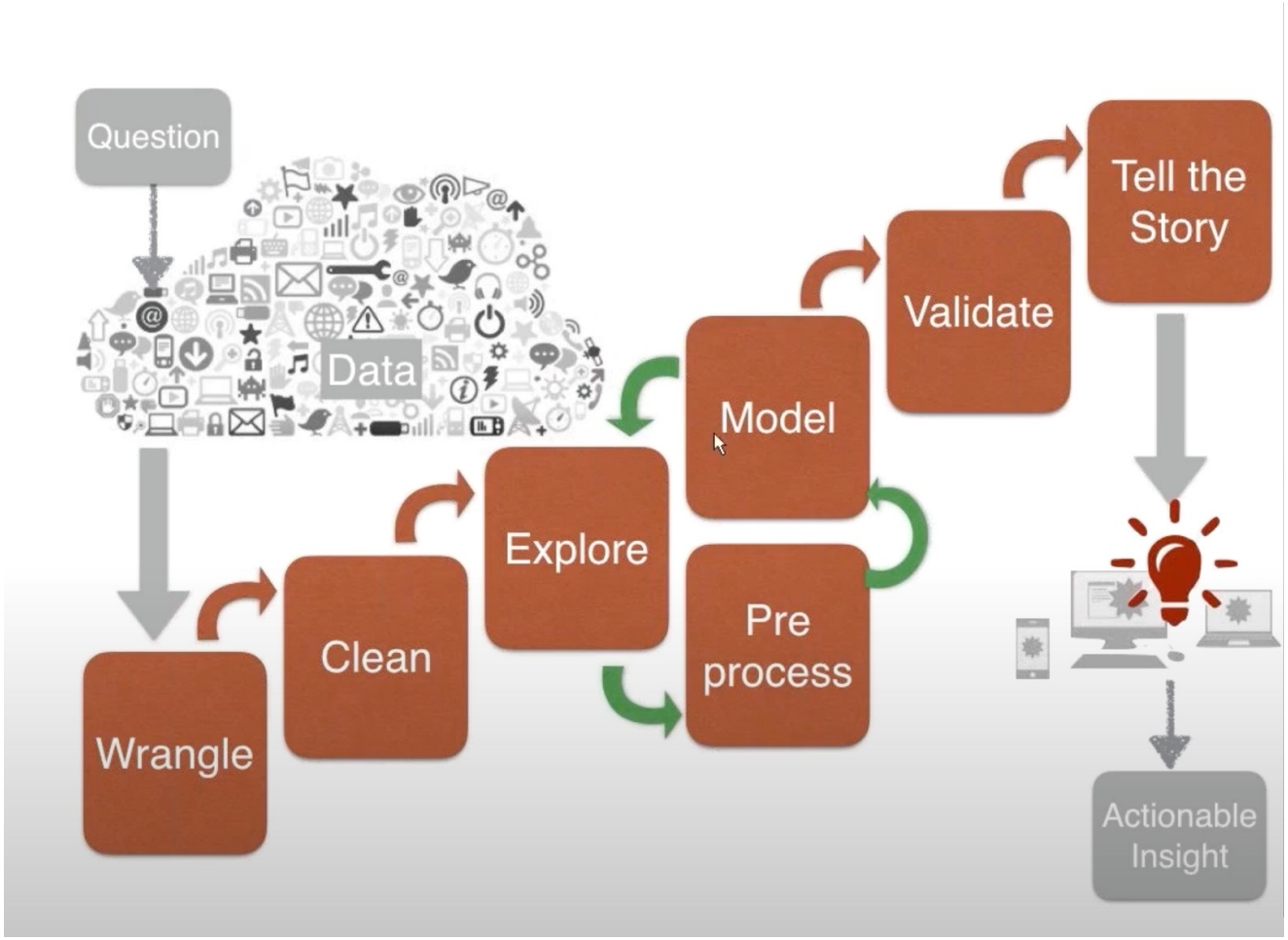
- Your decision trees are too deep
- You are randomly sampling too many features when you choose a split
- You have too few trees in your ensemble
- Your bagging implementation is randomly sampling sample points without replacement

[3 pts] In terms of the bias-variance decomposition, a 1-nearest neighbor classifier has _____ than a 3-nearest neighbor classifier.

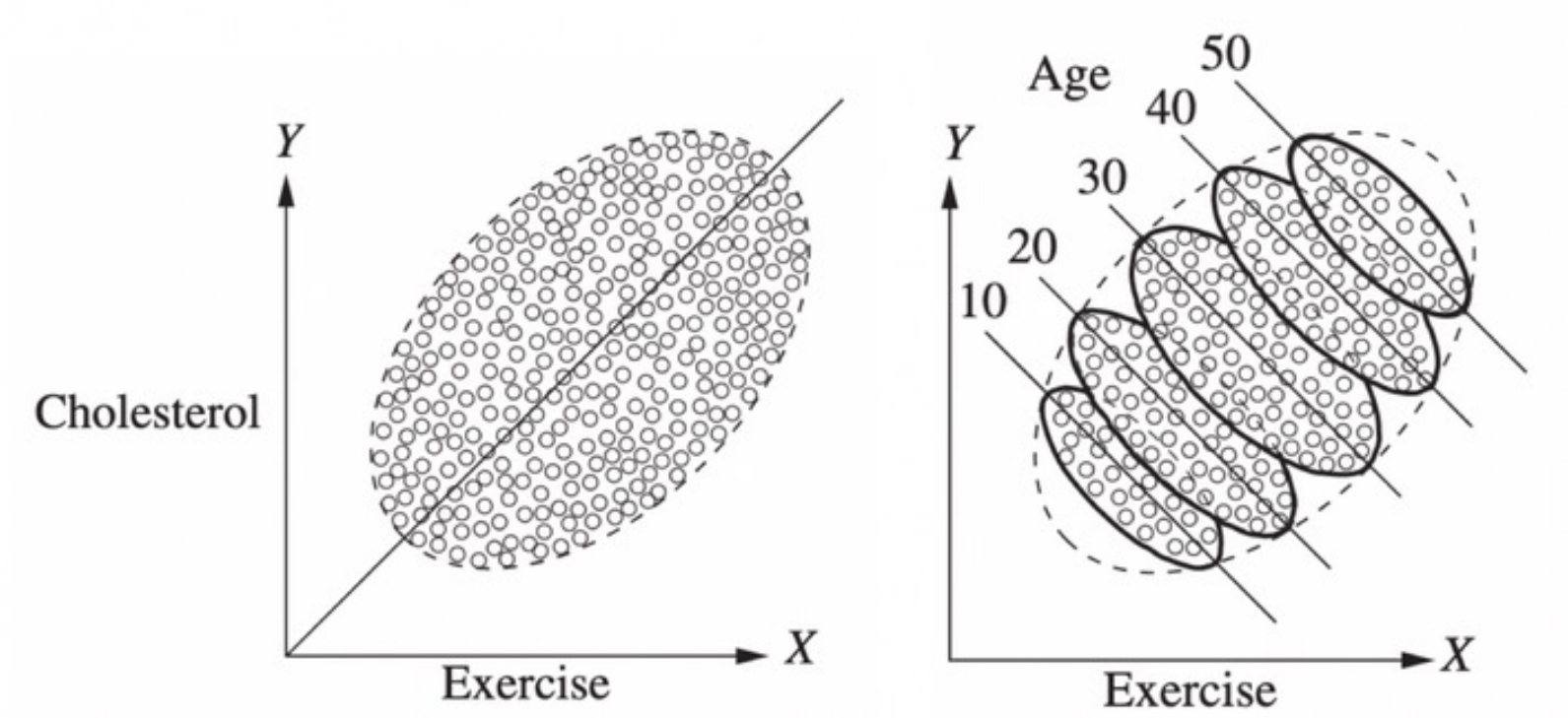
- ☐ higher variance
- ☐ higher bias
- ☐ lower variance
- ☐ lower bias

Describe a data science pipeline

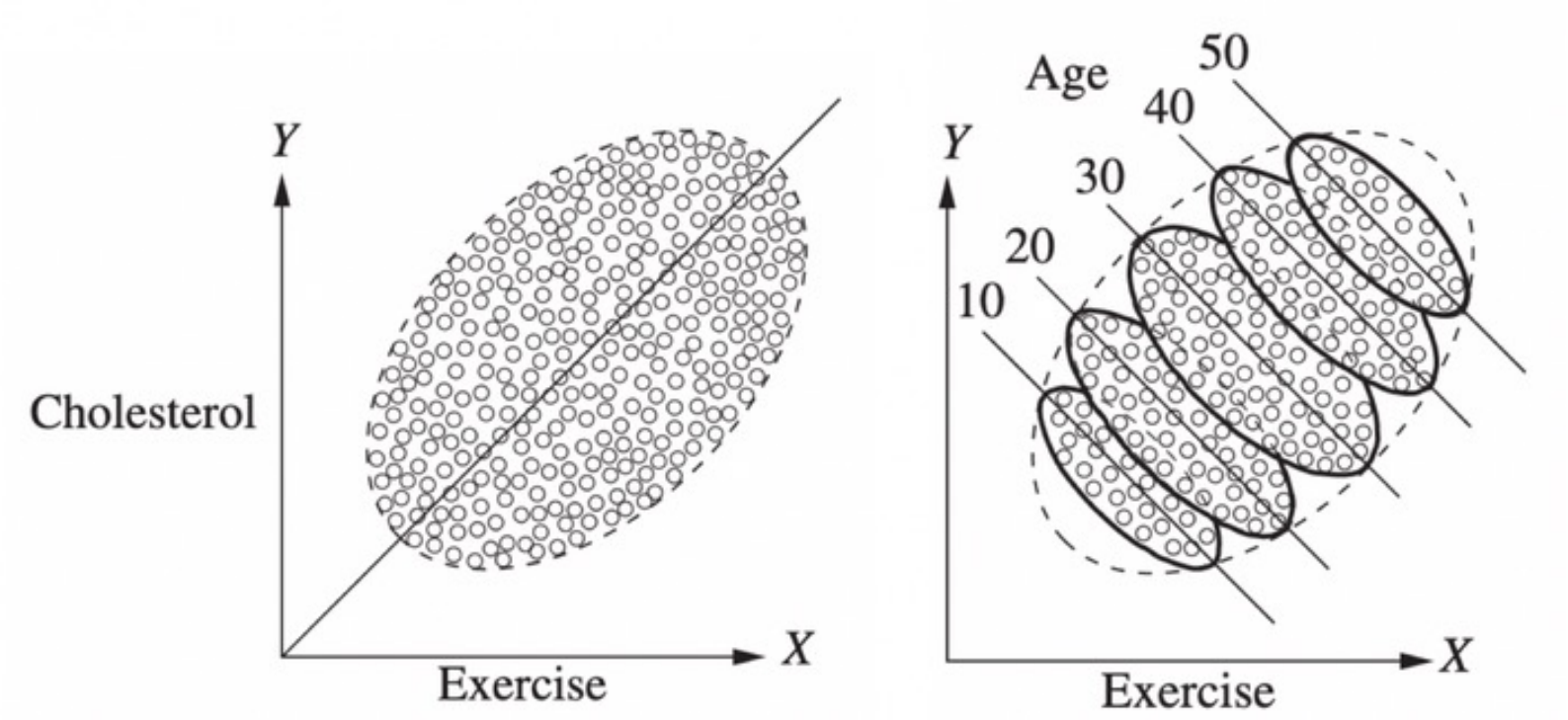
The Data Science pipeline



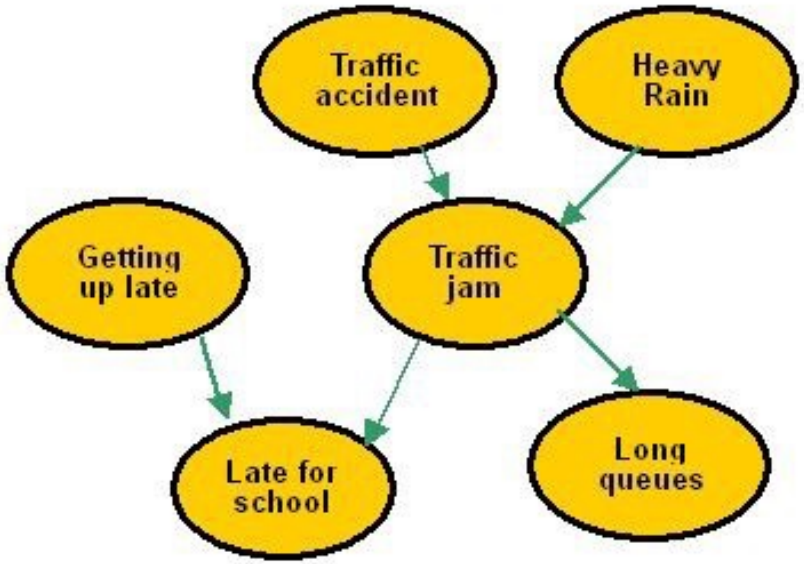
Taking into account the following corellation diagrams, is there a causality between cholesterol and exercise?



This is a graph showing the correlative relationship between Exercise and Cholesterol (which *looks like a causal relationship but is not).* If we just look at the correlative relationship between cholesterol and exercise, it looks like there's a causal relationship between the two. But this correlation actually happens because both cholesterol and exercise share a common cause or confounder: age.



Is the following a causal diagram?



Can decision trees be used for performing clustering?

A. True
B. False

For two runs of K-Mean clustering is it expected to get same clustering results?

A. Yes

B. No

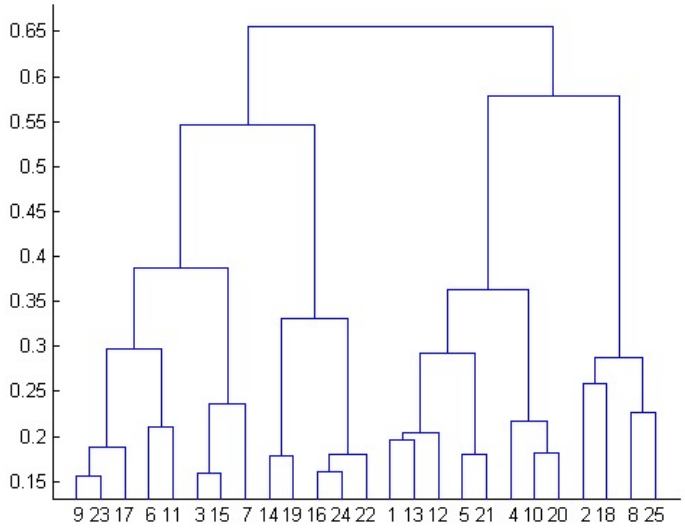
After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?

A. There were 28 data points in clustering analysis

B. The best no. of clusters for the analyzed data points is 4

C. The proximity function used is Average-link clustering

D. The above dendrogram interpretation is not possible for K-Means clustering analysis



In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Options:

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

E. 1, 2, 3 and 4

If two variables V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?

1. If V1 and V2 has a correlation of 1, the cluster centroids will be in a straight line
2. If V1 and V2 has a correlation of 0, the cluster centroids will be in straight line

Options:

A. 1 only

B. 2 only

C. 1 and 2

D. None of the above

What should be the best choice for number of clusters based on the following results: