Lebanese University Faculty of Sciences I	INFO 437-447 E+F Final	Date: 11/06/2024 Duration: 1h 30 min

Exercise I (DT and BN) (30 pts)

Consider the training dataset shown in the following Table.

A	B	Class Label
0	1	c1
0	0	c2
1	1	c1
0	1	c1
1	0	c1
0	0	c2
1	1	c1
0	0	c2
1	0	c1
1	0	c2

(a) Calculate the gain in the Gini index when splitting into attributes A and B, respectively. Show your calculation details. According to the gain, which will you choose as the first attribute to split in the decision tree induction?

(b) Calculate the gain in the misclassification error when splitting into attributes A and B, respectively. Show your calculation details. According to the gain, which one will you choose as the first attribute to split in the decision tree induction?

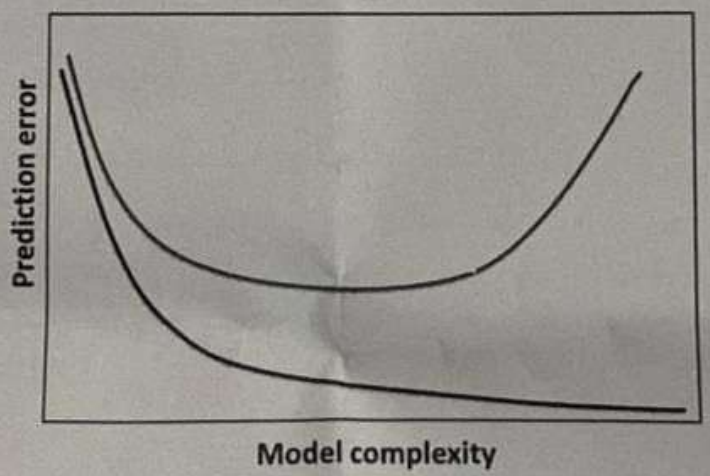
(c) Compute the conditional probabilities:
$P(A = 1|C = c1)$, $P(A = 0|C = c1)$, $P(B = 1|C = c1)$, $P(B = 0|C = c1)$, $P(A = 1|C = c2)$, $P(A = 0|C = c2)$, $P(B = 1|C = c2)$, and $P(B = 0|C = c2)$. [2 marks]

(d) Use the computed conditional probabilities to predict the class label for a test sample ($A = 1$; $B = 0$) using the naive Bayes approach. [2 marks]

Exercise II (20 pts)

A. Answer the following questions:
   a. What is overfitting in machine learning? And what is a common symptom of overfitting?
   b. What is the main assumption of a Naive Bayesian Network?

B. Consider the following graph



Draw the graph on your sheet and show your answers inside the graph

a. Which of the curves is more likely to be the training error and which is more likely to be the validation error?

b. In which regions of the graph are bias and variance low and high? Indicate clearly on the graph with four labels: "low variance", "high variance", "low bias", and "high bias".
c. In which regions does the model overfit or underfit? Indicate clearly on the graph by labeling "overfit" and "underfit".

Exercise III (K-Nearest Neighbor) (20 pts)

After an IT course, the exam results were recorded along with some data about the students. The results can be found in the table below. (GPA is the Grade Point Average.)

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
1	555 - 3452	Java	No	3.1	Yes
2	555 - 6294	Java	No	2.0	No
3	555 - 9385	C++	Yes	3.5	Yes
4	555 - 9387	Python	Yes	2.5	Yes
5	555 - 9284	Java	Yes	3.9	No
6	555 - 0293	C++	No	2.9	No
7	555 - 9237	Java	No	1.9	No
8	555 - 3737	Python	Yes	3.2	Yes

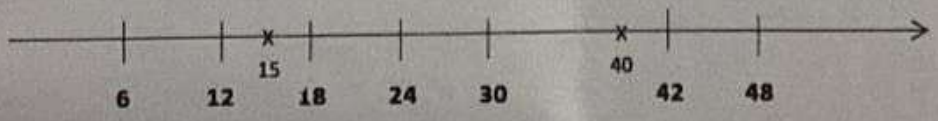
Suppose we build a K-NN classifier as follows. The ID and Phone number are unrelated to a student's capacity to pass the exam, so they are discarded. For nominal Attributes we will consider the distance between two items to be 1 if they are different and 0 if they are the same. Use this K-NN classifier with $K = 3$ to predict whether the following student (who overslept and missed the original exam) will pass the re-exam.

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
9	555 - 6295	C++	Yes	3.0	?

Exercise IV (Clustering) (30 pts)

Consider the following dataset of one-dimensional instances: {6, 12, 18, 24, 30, 42, 48}.

1. Using 15 and 40 as the initial centroids, follow the K-means algorithm to partition the dataset into 2 clusters. Show your work.



2. Calculate the SSE of the clustering you obtained above. Show your work.
3. Follow the single link version of hierarchical clustering to cluster this dataset. Show your work and the resulting dendrogram.