CSE 520: Data Analysis and Visualization

Mid-semester examination: February 2026

Total marks: 50 (Overall weightage: 30%)

Section 1

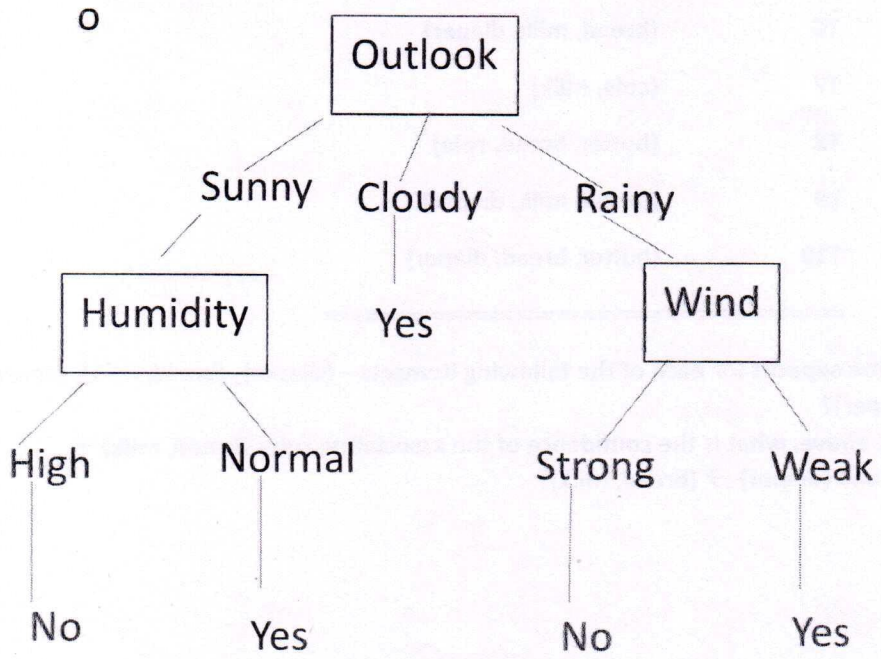
Answer any 6 of the following. Each question carries 2 marks.

a) Why is data preprocessing required, when analysing data? State at least 4 reasons.
b) What are the different types of qualitative attributes? Give one example for each.
c) What is an asymmetric binary attribute? Why are distance measures different for symmetric and asymmetric binary attributes?
d) What is over-fitting? What are two typical reasons for model overfitting?
e) What is the main principle on which the *Apriori* algorithm is based?
f) Why is conflict resolution needed in rule-based classifiers?
g) What is post-pruning?
h) In the Naïve Bayesian algorithm, why is the term 'naïve' used?

Section 2

Answer any 6 of the following. Each question carries 3 marks.

a) What are some techniques for handling missing data?
b) Define the terms drill-down, roll-up and star schema
c) What are some of the key challenges in data integration?
d) What is a data cube? Give an example of a 3-dimensional data cube, in tabular form.
e) State some key differences between OLTP and OLAP systems.
f) List down the main steps of the Apriori algorithm.
g) How does boosting work?
h) What are all the rules that can be listed down from the following decision tree with two class labels, viz., *PlayTennis=Yes* and *PlayTennis=No*?



Section 3

Answer any 4 of the following. Each question carries 5 marks. Please show the detailed steps in the sheets provided, and enter the final answer in LMS. Alternately, if you are comfortable, you may show the steps and final answer in LMS.

a) **I) Find the Euclidean distance and Manhattan distance between the variables P and Q, where P and Q are represented by the points (10,5) and (4,13) in 2-d space.**
2) Given a data set containing a range of numbers from 30,000 to 3,00,000, what is the normalised value of 63,000, under min-max normalization, with a min of 0 and a max of 1?

b) **Consider the following transaction dataset.**

TID	Items
T1	{butter, milk, diaper}
T2	{butter, bread, cola, diaper}
T3	{butter, bread, milk, diaper}
T4	{butter, cola, milk, diaper}
T5	{bread, cola, diaper}
T6	{bread, milk, diaper}
T7	{cola, milk}
T8	{butter, bread, cola}
T9	{butter, milk, diaper}
T10	{butter, bread, diaper}

1) **What is the support for each of the following itemsets – {diaper}, {bread, milk}, {bread, milk, diaper}?**
2) **From the above, what is the confidence of the association rules {bread, milk} → {diaper} and {diaper} → {bread, milk}?**

c) Consider the training examples shown below for a binary classification problem.

ID	Gender	CarType	ShirtSize	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra large	C0
6	M	Sports	Extra large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

1) Compute the Gini index for the overall collection of attributes.
2) Compute the overall Gini index for the carType attribute using multiway split.

d) Consider the following data set for a binary classification problem with 2 classes, 0 and 1. The three attributes A,B and C also take values 0 and 1.

Id	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Estimate the following conditional probabilities, using Bayes theorem.

1) P(A=1|+)
2) P(A=1|-)
3) P(B=0|+)

4) P(B=0|-)

e) Consider the confusion matrix for a classifier.

Actual \ predicted	isFraud = Y	isFraud = N
isFraud = Y	6000	1000
isFraud = N	1000	2000

Evaluate the following for the above:

1) Precision
2) Sensitivity
3) Specificity
4) Accuracy
5) Recall