S1	S2	S3	S4	Total

Student Name: ______________________

Number: ________________

Yıldız Technical University

BLM4800–Introduction to Data Mining

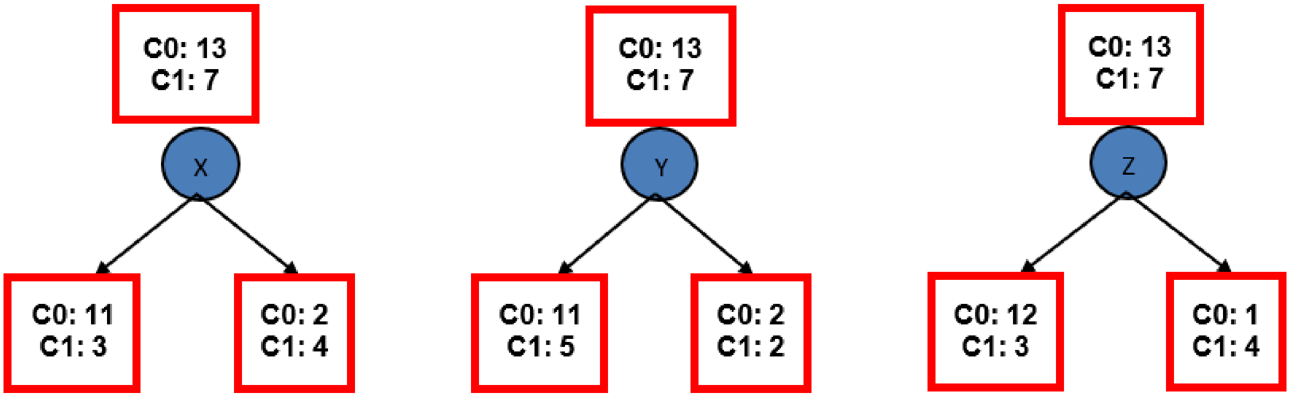
Midterm - Spring 2022-2023

- **Duration:** 90 minutes
- **Exam information:**
  - Attempts to cheat in the exam will not be tolerated. If an attempt to cheat is discovered, it will be severely punished.
  - Read all the questions carefully before you start answering them.
  - The point value of each question is indicated next to the question.

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

1. [25 points].

The example diagram below shows the candidate values to be used to construct the decision tree and the class distribution after the split. Accordingly, which of the attributes X, Y, and Z splits the dataset most efficiently? Explain the reason by using the Gini index formula. Show your calculations in details to get full credit.



p -> parent   l -> left
r -> right

S1: $x_p \to 1-\left(\frac{13}{20}\right)^2-\left(\frac{7}{20}\right)^2=0,46$

$x_l \to 1-\left(\frac{11}{14}\right)^2-\left(\frac{3}{14}\right)^2=0,34$

$x_r \to 1-\left(\frac{2}{6}\right)^2-\left(\frac{4}{6}\right)^2=0,44$

$\to 0,34\cdot\frac{14}{20}+0,44\cdot\frac{6}{20}=0,37$

for x tree $0.46>0.37$

$0.37.$

$y_p \to 0,46$

$y_l \to 1-\left(\frac{11}{16}\right)^2-\left(\frac{5}{16}\right)^2=0,43$

$y_r \to 1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=0,5$

$\to 0,43\cdot\frac{16}{20}+0,5\cdot\frac{4}{20}=0.44.$

$z_p \to 0,46$

$z_l \to 1-\left(\frac{12}{15}\right)^2-\left(\frac{3}{15}\right)^2=0,32$

$z_r \to 1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2=0,32$

$\to 0,32\cdot\frac{15}{20}+0,32\cdot\frac{5}{20}=0,32$

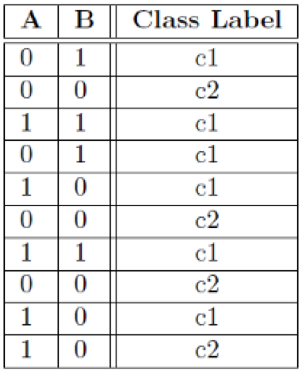
Result is 0.32 because the GINI index measures the impurity of a dataset, and a lower GINI index indicates less impurity and more homogeneous dataset. Therefore, a lower GINI index shows that an attribute divides the dataset better and enables the creation of a more efficient decision tree.

M. Kagan Bulut

20011901

2. [30 points].

Consider the training data set shown in the table below and answer the following questions.



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. (8 points).

Calculate the following conditional probabilities.

$P(A = 1|C = c1)$:

$P(A = 0|C = c1)$:

$P(B = 1|C = c1)$:

$P(B = 0|C = c1)$:

$P(A = 1|C = c2)$:

$P(A = 0|C = c2)$:

$P(B = 1|C = c2)$:

$P(B =0|C =c2)$:

2-) a-) $P(A=1 / C=c_1) = \frac{4}{6}$ $P(B=1 / C=c_1) = \frac{4}{6}$

$P(A=1 / C=c_2) = \frac{1}{4}$ $P(B=1 / C=c_2) = \frac{0}{4}$

$P(A=0 / C=c_1) = \frac{2}{6}$ $P(B=0 / C=c_1) = \frac{2}{6}$

$P(A=0 / C=c_2) = \frac{3}{4}$ $P(B=0 / C=c_2) = \frac{4}{4}$

b-) $P(\overbrace{A=1, B=0 \mid C=C_1}^{4/6 \,\cdot\, 2/6}) \cdot P(\overbrace{C=c_1}^{\frac{6}{10}}) = \frac{\overset{2}{\cancel{4}}}{6} \cdot \frac{\cancel{2}}{\underset{3}{\cancel{6}}} \cdot \frac{\cancel{6}}{\underset{5}{\cancel{10}}} = \frac{2}{15}$

$P(\overbrace{A=1, B=0 \mid C=C_2}^{1/4}) \cdot P(\overbrace{C=C_2}^{4/10}) = \frac{1}{4} \cdot \frac{4}{10} = \frac{1}{10}$

c-) A=1 B=0 → c1

A=1 B=0 → c1

A=1 B=0 → C_2

Since there are two instance of C1 and 1 instance of C_2 for A=1 and B=0, we determine the Class 1

b. (10 points).

Using the conditional probabilities, you have calculated, estimate the class label for a test sample x=(A =1, B=0) given by the Naive Bayes method. Show your calculation method in detail.

c. (12 points).

In the same question, estimate the class label for x=(A=1, B=0) using the K-Nearest Neighbor algorithm for the Euclidean distance for k=3. Show your calculation method in detail.

3. **[25 points].**

Answer the following questions.

a. (8 points).

Calculate the Cosine and Euclidean distances between the defined vectors of x=(2,-1,0,2,-3), y=(-1,1,-1,0,0,-1).

b. (8 points).

In a dataset, feature A has the following data: A=[200,400,800,1000,2000,2200] Scale this property with min-max normalization to be between min=1 and max=11.

3-)

a-) $\left(\overbrace{(2-(-1))}^{9}{}^2 + \overbrace{(-1-1)}^{4}{}^2 + \overbrace{(0-(-1))}^{1}{}^2 + \overbrace{(2-0)}^{4}{}^2 + \overbrace{(-3-(-1))}^{4}{}^2 \right)^{\frac{1}{2}}$

$= (22)^{\frac{1}{2}} \rightarrow$ result $= \sqrt{22}$

cosine $= \dfrac{0+0+3-2-1}{|y| * |x|} = 0$

b-) $\dfrac{\text{number min}}{\text{max} - \text{min}} * 10 + 1$

$10 * \dfrac{200 - 200}{2200 - 200} + 1 = 1$

$10 * \dfrac{800 - 200}{2000} + 1 = 4$

$10 * \dfrac{2000 - 200}{2000} + 1 = 10$

$10 * \dfrac{400 - 200}{2000} + 1 = 2$

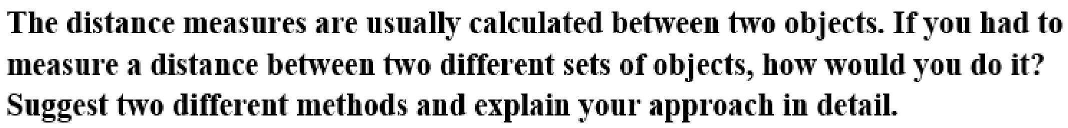
$10 * \dfrac{1000 - 200}{2000} + 1 = 5$

$10 * \dfrac{2200 - 200}{2000} + 1 = 11$

1, 2, 4, 5, 10, 11

c-)

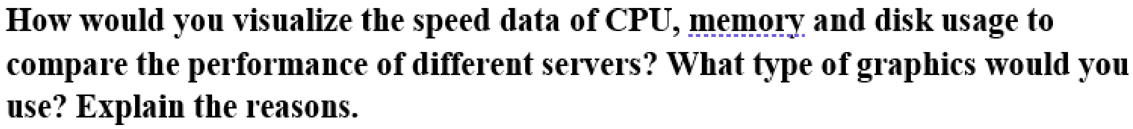
c. (4 points).



The distance measures are usually calculated between two objects. If you had to measure a distance between two different sets of objects, how would you do it? Suggest two different methods and explain your approach in detail.

Two methods for measuring the distance between two different sets of objects are the Jaccard distance and the Hamming distance. The Jaccard distance measures the dissimilarity between two sets by calculating the size of the intersection divided by the size of the union of the two sets. The Hamming distance calculates the number of positions where the corresponding elements of two sets are different.

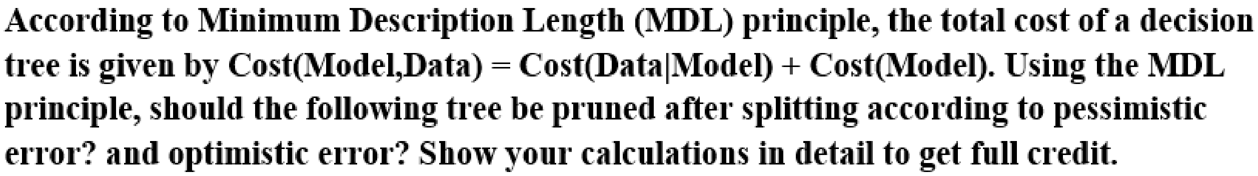
d. (5 points).



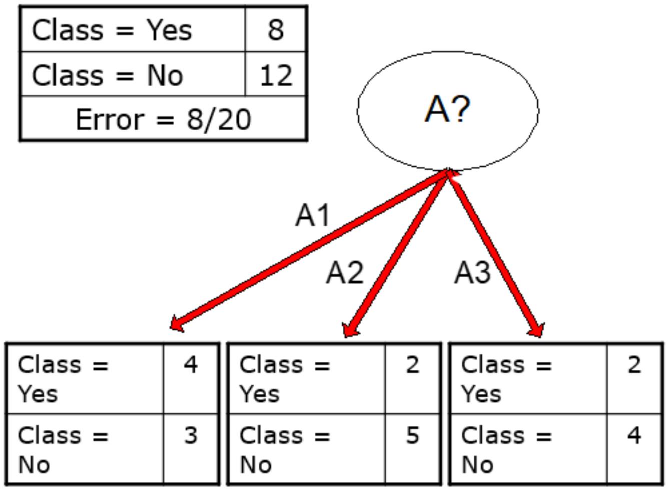
How would you visualize the speed data of CPU, memory and disk usage to compare the performance of different servers? What type of graphics would you use? Explain the reasons.

We can use line charts or heat maps to compare the performance of different servers based on CPU, memory, and disk usage data. Line charts are suitable for visualizing the changes in time series data, and we can easily compare the performance of different servers by using different lines or colors for each server. Heat maps, on the other hand, display the performance of different servers over time using color-coded cells, which can be useful for examining the overall trends across all servers.

4. [20 points].



According to Minimum Description Length (MDL) principle, the total cost of a decision tree is given by Cost(Model,Data) = Cost(Data|Model) + Cost(Model). Using the MDL principle, should the following tree be pruned after splitting according to pessimistic error? and optimistic error? Show your calculations in detail to get full credit.



Class = Yes	8
Class = No	12
Error = 8/20	

Class = Yes	4
Class = No	3

Class = Yes	2
Class = No	5

Class = Yes	2
Class = No	4

) Training error $\rightarrow \frac{8}{20}$

Pestimistic error $\rightarrow \frac{8+(0,5).1}{20}$

Before SPLit

Train Error $= \frac{3+2+2}{20} = \frac{7}{20}$

Pessimist Error $= \frac{3.(0,5)+7}{20} = \frac{8,5}{20}$

AFter SPLit

for possimistic error $\rightarrow \frac{8,5}{20} = \frac{8,5}{20} \rightarrow$ there no need to split dataset because result is equal

for optimistic error $\rightarrow \frac{8}{20} > \frac{7}{20} \rightarrow$ we need to split dataset because before split greater than after split