DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY
UNIVERSITY EXAMINATIONS 2020/2021 ACADEMIC YEAR

SPECIAL/SUPPLEMENTARY EXAMINATION FOR THE DEGREE OF MASTER OF SCIENCE IN BUSINESS ANALYTICS

CIT 6114: MACHINE LEARNING

DATE: 19TH OCTOBER 2021 **TIME: 11:00AM-2:00PM**

Instructions: Answer Question 1 and Any Other Two

QUESTION 1: (30 MARKS)

a) Compare and contrast *Linear Regression* and *Logistic Regression* Machine Learning algorithms.

(4 marks)

b) Learning can be described as *committing to memory* or *receiving instructions*. However, these two aspects do not appropriately describe machine learning. Explain.

(3 marks)

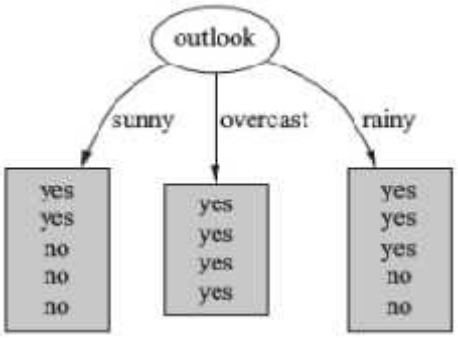
c) Distinguish between **hold-out** and **N-fold cross-validation** methods of evaluating Machine Learning models.

(4 marks)

d) Explain how the **Euclidean distance** between nominal attributes that involve color (e.g., the distance between the values red, green and blue) are determined in an instance based knowledge representation scheme.

(4 marks)

e) Calculate information values for all nodes of the outlook attribute in the following tree.



(3 marks)

f) Applying **1R algorithm** on the data in Table 1.2, determine:
   i) errors for the rules in the table **(4 marks)**
   ii) total error for each attribute shown in the table. **(4 marks)**

Table 1.2 Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

g) Distinguish between **Machine Learning** and **Data Mining**.

(4 marks)

QUESTION 2: (20 MARKS)

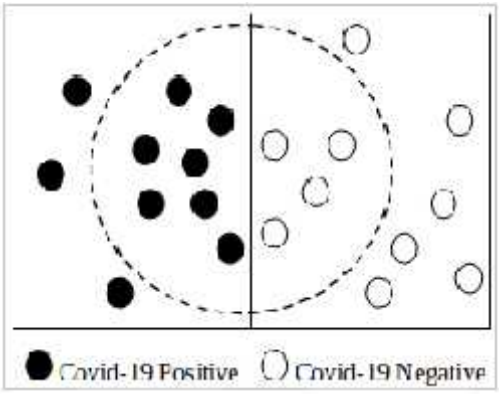
a) Demonstrate how **linear models** can be used in *binary classification* problems.

(4 marks)

b) Discuss **four** applications of data mining.

(8 marks)

c) An intelligent system was installed at DeKUT's main gate to detect Covid-19 status of people crossing the gate. Before deployment, it was tested on ten Covid–19 positive individuals and ten Covid-19 negative individuals. The results for the first round of testing are as shown in the figure below whereby the individuals in the dotted circle were categorized by the system as Covid–19 positive.



Subsequent runs of the testing yielded the following results:

Rounds of testing	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
No. of people correctly classified as positive	0	8	7	6	9	10
No. of people erroneously classified as positive	0	6	2	1	1	0
No. of people correctly classified as negative	10	4	8	9	9	10

Plot a Receiver Operator Characteristic (ROC) graph for the algorithm running this system.

(8 marks)

QUESTION 3: (20 MARKS)

a) With an illustration, briefly describe the *instance-based knowledge representation* i.e., the *nearest-neighbor* classification method.

(7 marks)

b) Explain **three** learning styles in data mining applications.

(6 marks)

c) Consider the following rules derived from contact lens data that tells the kind of contact lens to prescribe, given certain information about a patient.

```
If tear production rate = reduced then recommendation = none.
If age = young and astigmatic = no and tear production rate = normal
   then recommendation = soft
If age = pre-presbyopic and astigmatic = no and tear production
   rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope and
   astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no and
   tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = young and astigmatic = yes and tear production rate = normal
   then recommendation = hard
If age = pre-presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
```

Using these rules, construct a decision tree that describes the structure of the lens data.

(7 marks)

QUESTION 4: (20 MARKS)

a) Consider the data in table below that is associated with the type of lens that an optician recommended for several patients over a period of time.

Contact Lens Data

	Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
1	Pre-presbyopic	hypermetrope	yes	normal	none
2	pre-presbyopic	hypermetrope	yes	reduced	none
3	young	myope	yes	reduced	none
4	young	myope	no	normal	soft
5	presbyopic	myope	yes	reduced	none
6	young	myope	yes	normal	hard
7	young	myope	yes	reduced	none
8	young	hypermetrope	no	reduced	none
9	presbyopic	myope	yes	normal	hard
10	pre-presbyopic	myope	yes	normal	hard
11	pre-presbyopic	hypermetrope	no	reduced	none
12	young	hypermetrope	yes	reduced	none
13	young	hypermetrope	no	normal	soft
14	presbyopic	hypermetrope	no	normal	soft
15	pre-presbyopic	myope	no	reduced	none
16	pre-presbyopic	myope	no	normal	soft
17	young	hypermetrope	yes	normal	hard
18	presbyopic	hypermetrope	no	reduced	none

Consider a new patient with the attributes shown in the following table:

New_Patient

	Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
1	young	myope	yes	normal	?

i. Applying **Naïve Bayes** algorithm, determine the *likelihood* of the optician prescribing a hard lens for this patient.

(12 marks)

ii. Calculate the *probability* of the optician prescribing the hard contact lens.

(4 marks)

b) Discuss any two contemporary issues related to Data Mining and Big Data in Kenya.

(4 marks)

END.