WOMEN'S UNIVERSITY IN AFRICA

Addressing gender disparity and fostering equity in University Education

FACULTY OF MANAGEMENT AND ENTREPRENEURIAL SCIENCES

BSc HONOURS DEGREE IN COMPUTER SCIENCE

MAIN PAPER

HCS322:   MACHINE LEARNING

INTAKE 1: THIRD YEAR SECOND SEMESTER

TIME: 3 HOURS MORNING

INSTRUCTIONS TO CANDIDATES

Answer any **four** questions.

QUESTION 1

a) Write short notes on the following concepts as they are used in machine learning:

     i)   Internet of things; [3]

     ii)   Artificial intelligence; [3]

     iii)   Machine learning; and [3]

     iv)   Natural language processing [3]

b) Wakanda Studios have recently established a music promotion studio in Mbare – Harare. It is the desire of the Wakanda Management to classify the music likes of the people in Harare and segment the market according to the potential insights that can be produced from the questionnaire they distributed online. As modern as the management of Wakanda is, they have seen it fit to analyse the responses which have since been captured in a csv file. As a newly appointed data scientist of Wakanda Studios, you have been invited by the Management to provide them with answers to the following questions:

     (i)   Which machine learning model can we build to understand the consumer music likes? [1]

     (ii)   With the aid of a diagram, explain how you will build the model using the machine learning approach. [12]

QUESTION 2

You have been asked to present to a panel of business professionals who wish to transform their business approach techniques by introducing machine learning techniques in their business intelligence functions. The professionals have expressed interest in understanding the types of machine learning that are available for consideration despite variations in business requirements.

a) Explain to the panel with the aid of examples, the following types of machine learning:

     i. Supervised learning; [4]

     ii. Semi Supervised learning; [4]

     iii. Unsupervised learning; and [4]

     iv. Reinforcement learning [4]

b) Supervised machine learning is associated with two types of data spaces. Identify and explain the two types of data spaces in supervised machine learning. [9]

QUESTION 3

Figure 1 is an extract of a model to classify transactions as either fraudulent or genuine. Use Figure 1 to answer questions a, b, c, d and e.

```
In [14]: data.describe()
```

Out[14]:

	CountryCode	Amount	Value	PricingStrategy	FraudResult
count	95662.0	9.566200e+04	9.566200e+04	95662.000000	95662.000000
mean	263.0	6.717846e+03	9.900584e+03	2.255974	0.002018
std	0.0	1.233068e+05	1.231221e+05	0.732924	0.044872
min	263.0	-1.000000e+06	2.000000e+00	0.000000	0.000000
25%	263.0	-5.000000e+01	2.750000e+02	2.000000	0.000000
50%	263.0	1.000000e+03	1.000000e+03	2.000000	0.000000
75%	263.0	2.800000e+03	5.000000e+03	2.000000	0.000000
max	263.0	9.880000e+06	9.880000e+06	4.000000	1.000000

a) Identify the name of the dataset used in the model. [1]

b) List the features of the dataset [5]

c) Given that classification models thrive on labeled data, list the features of the X input space and the Y output feature. [5]

d) Do you think the feature 'CountryCode' can be used in the classification model given that all the transactions were recorded with the CountryCode 263.0? Justify your answer. [4]

e) Using the feature 'Value', account for the statistical insights that can be discovered using the '**describe()**' function in python. [10]

QUESTION 4

The string below shows a line of code to fit a model in machine learning.

history=model.fit(X_train,Y_train,epochs=200,batch_size=10,verbose=1,validation_split=0.2
,callbacks=[checkpoint, earlystop])

With reference to the code above, answer questions

a) For how many iterations is the model expected to run? [2]

b) Briefly comment on the effect of increasing the batch size as well as reducing it on the performance of the model. [5]

c) What is the purpose of introducing the 'earlystop' checkpoint in the model? [2]

d) Give one monitor that can be used as a checkpoint for earlstopping and briefly explain how it works. [4]

Figure 2 is an extract of the history captured during the training of a model.

```
Train on 61223 samples, validate on 15306 samples
Epoch 1/200
61020/61223 [=============================>.] - ETA: 0s - loss: 0.0018 - mean_squared_error: 0.0018
Epoch 00001: val_mean_squared_error improved from inf to 0.00082, saving model to Transaction Classifier.h5
61223/61223 [==============================] - 7s 120us/sample - loss: 0.0017 - mean_squared_error: 0.0017 - val_loss: 8.1517e-
04 - val_mean_squared_error: 8.1517e-04
Epoch 2/200
60840/61223 [=============================>.] - ETA: 0s - loss: 0.0010 - mean_squared_error: 0.0010
Epoch 00002: val_mean_squared_error did not improve from 0.00082
61223/61223 [==============================] - 5s 86us/sample - loss: 0.0010 - mean_squared_error: 0.0010 - val_loss: 0.0010 -
val_mean_squared_error: 0.0010
Epoch 3/200
60820/61223 [=============================>.] - ETA: 0s - loss: 0.0011 - mean_squared_error: 0.0011
Epoch 00003: val_mean_squared_error did not improve from 0.00082
61223/61223 [==============================] - 5s 88us/sample - loss: 0.0011 - mean_squared_error: 0.0011 - val_loss: 0.0010 -
val_mean_squared_error: 0.0010
Epoch 4/200
60940/61223 [=============================>.] - ETA: 0s - loss: 0.0014 - mean_squared_error: 0.0014
Epoch 00004: val_mean_squared_error did not improve from 0.00082
61223/61223 [==============================] - 5s 88us/sample - loss: 0.0014 - mean_squared_error: 0.0014 - val_loss: 0.0011 -
val_mean_squared_error: 0.0011
Epoch 5/200
60790/61223 [=============================>.] - ETA: 0s - loss: 0.0010 - mean_squared_error: 0.0010
Epoch 00005: val_mean_squared_error did not improve from 0.00082
61223/61223 [==============================] - 5s 88us/sample - loss: 0.0010 - mean_squared_error: 0.0010 - val_loss: 9.0346e-0
4 - val_mean_squared_error: 9.0346e-04
Epoch 6/200
60960/61223 [=============================>.] - ETA: 0s - loss: 9.6951e-04 - mean_squared_error: 9.6951e-04
Epoch 00006: val_mean_squared_error did not improve from 0.00082
```

Figure 2: Snippet from model training

Using Figure 2, answer questions below

e) What is the train sample size? [2]

f) What is the validate sample size? [2]

g) What is the name of the saved model? [2]

h) What is the val_mean_squared_error of the model at the fifth iteration? [2]

i) Identify the model loss at the fourth iteration of the model. [2]

j) What is the ideal condition of the val_mean_squared_error with reference to model performance as the model training progresses? [2]

QUESTION 5

a) Explain the basic problem in machine learning. [5]

b) What is the purpose of Exploratory Data Analysis? [10]

c) Explain any five 'Vs' of big data. [10]

QUESTION 6

a) What is overfitting and underfitting and why is it undesirable for your machine learning algorithm. [8]

b) How do you overcome overfitting in machine learning? [5]

c) Describe in detail four metrics that are used for measuring the performance of machine learning algorithms. [12]

END