DUT

DURBAN UNIVERSITY OF TECHNOLOGY
INYUVESI YASETHEKWINI YEZOBUCHWEPHESHE

FACULTY OF ACCOUNTING AND INFORMATICS

DEPARTMENT OF INFORMATION TECHNOLOGY

2023
YEAR END MAIN EXAMINATION

INSTRUCTIONAL PROGRAMME : Bachelor of ICT/Advanced Diploma in ICT
INSTRUCTIONAL OFFERING : Machine Intelligence 3
PAPER NUMBER : 1
SUBJECT CODE/S : MAIN301/MCHI301
DATE : 22 NOVEMBER 2023
DURATION : 3 HOURS
TIME : 14H00 - 17H00
TOTAL MARKS : 100
NUMBER OF PAGES : 11 (including cover sheet)
EXAMINER/S : MRS S HOOSEN
MODERATOR/S : MR L VORSTER

INSTRUCTIONS/REQUIREMENTS:-

1. Answer all questions in the answer book

2. Calculators are permitted

3. No notes will be permitted.

Do not turn the page until permission is given

QUESTION 1 [15 marks]

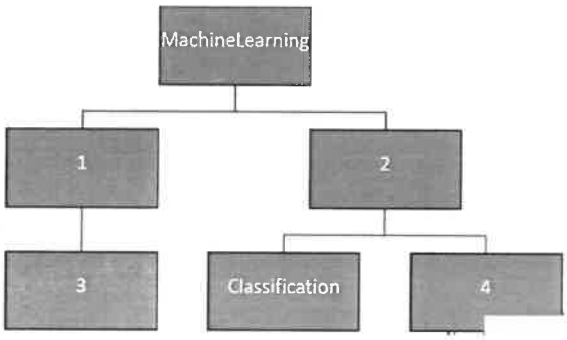
Write down the numbers 1.1 to 1.15 and next to each write down only the alphabet of the choice you have selected.

1.1 What type of learning needs to be applied to the following problem statement?

"You want to train a machine to help you predict how long it will take you to drive home from your workplace."

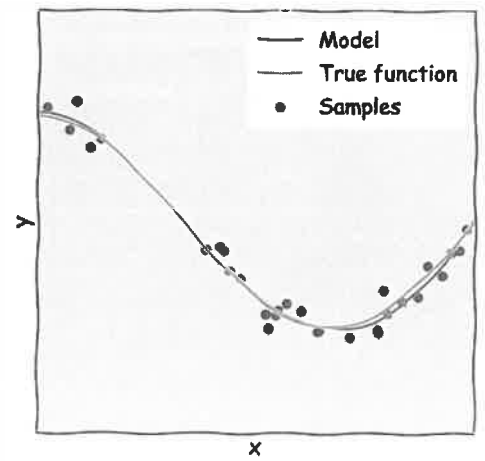
A) Unsupervised Learning with Clustering
B) Supervised Learning with Regression
C) Supervised Learning with Classification
D) Reinforcement Learning with Agents

1.2 The diagram below represents an overview of the types of machine learning. Study this diagram and then select the correct option by matching the numbers in the diagram with the corresponding type of learning.



A) 1: Supervised; 2: Unsupervised; 3: Regression; 4: Clustering
B) 1: Unsupervised; 2: Supervised; 3: Clustering; 4: Regression
C) 1: Supervised; 2: Unsupervised; 3: Clustering; 4: Regression
D) 1: Unsupervised; 2: Supervised; 3: Regression; 4: Clustering

1.3 What best describes the model function illustrated below?



A. This is an example of a model with high bias
B. The model function has the right complexity to fit the true function
C. The model function does not have enough complexity to fit the true function and underfits the model
D. The model function has too many complexities and over fits the model

1.4 In the snippet of code below which method compares features in a dataset and generates a threshold value for the comparisons, and which method locates individual elements in the dataset, respectively.

```
In [134]: def get_correlation(data,threshold):
              corr_column=set()
              corrmatrix=data.corr()
              for i in range(len(corrmatrix.columns)):
                  for j in range(i):
                      if abs(corrmatrix.iloc[i,j])>threshold:
                          colname=corrmatrix.columns[i]
                          corr_column.add(colname)
              return corr_column
```

A) corrmatrix, i
B) .corr(), set()
C) .corr(), .iloc[]
D) set(), iloc[]

1.5 This type of learning is used to identify different segments of customers and group them into categories like gender, age and location.

A) Reinforcement learning with agents
B) Unsupervised learning with clustering
C) Supervised learning with regression
D) Supervised learning with classification

1.6 The algorithm that finds the best fit model by automatically selecting the best features in a dataset and handles missing data is _____?

A) Neural Networks
B) SVM
C) Random Forest
D) Decision Trees

1.7

	A	B	C	D	E	F	G	H	I
	Type	Flour	Milk	Sugar	Butter	Egg	Baking Pov	Vanilla	Salt
	Muffin	55	28	3	7	5	2	0	0
	Muffin	47	24	12	6	9	1	0	0
	Muffin	47	23	18	6	4	1	0	0
	Muffin	45	11	17	17	8	1	0	0
	Muffin	50	25	12	6	5	2	1	0
	Muffin	55	27	3	7	5	2	1	0
	Muffin	54	27	7	5	5	2	0	0
	Muffin	47	26	10	10	4	1	0	0
	Muffin	50	17	17	8	6	1	0	0
	Muffin	50	17	17	11	4	1	0	0
	Cupcake	39	0	26	19	14	1	1	0
	Cupcake	42	21	16	10	8	3	0	0
	Cupcake	34	17	20	20	5	2	1	0
	Cupcake	39	13	17	19	10	1	1	0
	Cupcake	38	15	23	15	8	0	1	0
	Cupcake	42	18	25	9	5	1	0	0
	Cupcake	36	14	21	14	11	2	1	0
	Cupcake	38	15	31	8	6	1	1	0
	Cupcake	36	16	24	12	9	1	1	0
	Cupcake	34	17	23	11	13	0	1	0

x_train,x_test,y_train,y_test=train_test_split(X, y, test_size=0.3)

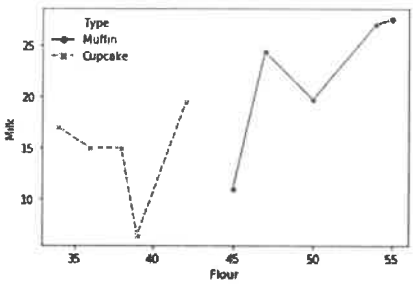
Given the dataset and the code above, what will be the output for x_train.shape, x_test.shape ?

A) (12, 8), (8,)
B) (14,8) (14,)
C) (14,) (6, 8)
D) (14, 8), (6, 8)

1.8 This module implements several loss, score, and utility functions to measure classification performance of algorithms when applied to data.
A) sklearn. scores
B) sklearn. metrics
C) sklearn.models
D) sklearn.accuracy

1.9 These types of machine learning algorithms are based on supervised learning and they model target prediction values based on independent variables. They mainly used for finding out the relationship between variables and forecasting.
A) Decision Trees
B) Random Forest
C) Linear Regression
D) SVM

1.10 What will be the correct coding with parameter listings for the plots below:



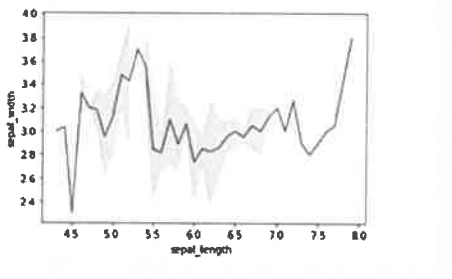
A) ax=sns.lineplot(x='Flour',y='Milk',data=raw_data,hue='Type', ci=False, markers=True, style='Type')
B) ax=sns.line(x='Flour',y='Milk',data=raw_data,hue='Type', ci=True, markers=True, style='Type')
C) ax=sns.lineplot(x='Flour',y='Milk',data=raw_data,hue='Type', ci='False', markers=True, 'style'='Type')
D) ax=sns.lineplot(x=Flour,y=Milk,data=raw_data,hue='Type', ci=False, markers=True, style='Type')

1.11 Given the following code:

```
import seaborn as sns

data = sns.load_dataset("iris")
```

Select the line of code that will plot the graph below:



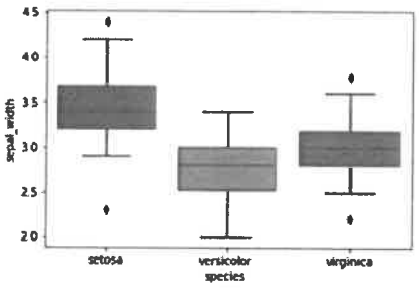
A. sns.lineplot(x="sepal_length", y="sepal_width", data=data, hue='Type')
B. sns.lineplot(x="sepal_length", y="sepal_width", data=data, ci="True")
C. sns.lineplot(x="sepal_width", y="sepal_length", data=data)
D. sns.lineplot(x="sepal_length", y="sepal_width", data=data)

1.12 Given the following code:

```
import seaborn as sns
import matplotlib.pyplot as plt
data = sns.load_dataset("iris")
```

Select the code that will display the plot below:

Output:



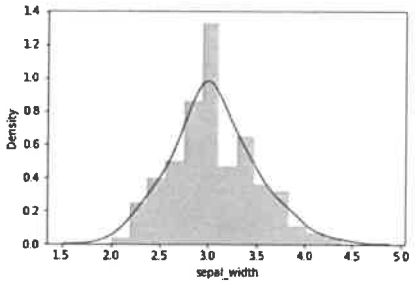
A) plt.boxplot(x='species', y='sepal_width', data=data)
plt.show()
B) sns.barplot(x=species, y=sepal_width, data=data)
plt.show()
C) sns.boxplot(x='species', y='sepal_width', data='data')
plt.show()
D) sns.boxplot(x='species', y='sepal_width', data=data)
plt.show()

1.13 Given the code below:

```
import seaborn as sns
import matplotlib.pyplot as plt
data = sns.load_dataset("iris")
```

Select the code that will display the plot below:

Output:



A) sns.histplot(data['sepal_width'])
B) sns.distplot(data['Density'])
C) sns.distplot(['sepal_width'])
D) sns.distplot(data['sepal_width'])

1.14 This is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency.

A) Histogram
B) Line Graph
C) Scatter Plot
D) Bar chart

1.15 Colouring different points of a scatter plot can be done through the use of:

A) ci
B) markers
C) Hue
D) Matplotlib inline

(15)

QUESTION 2[20 marks]

2.1. Study the diagrams A and B below and and give a clear and detailed explanation of what each illustrates with regards to Machine Learning.

Diagram A

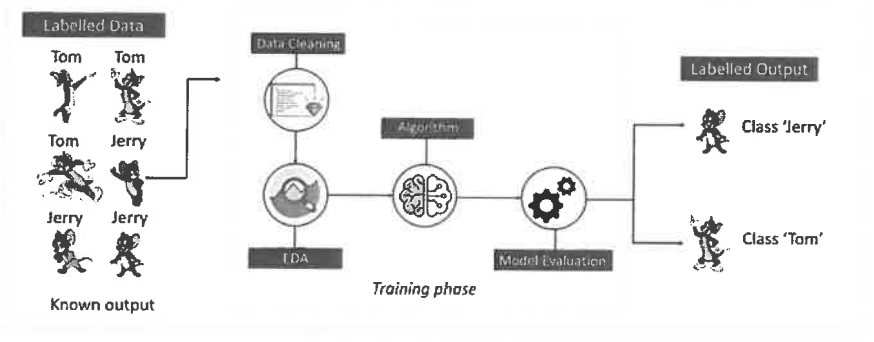
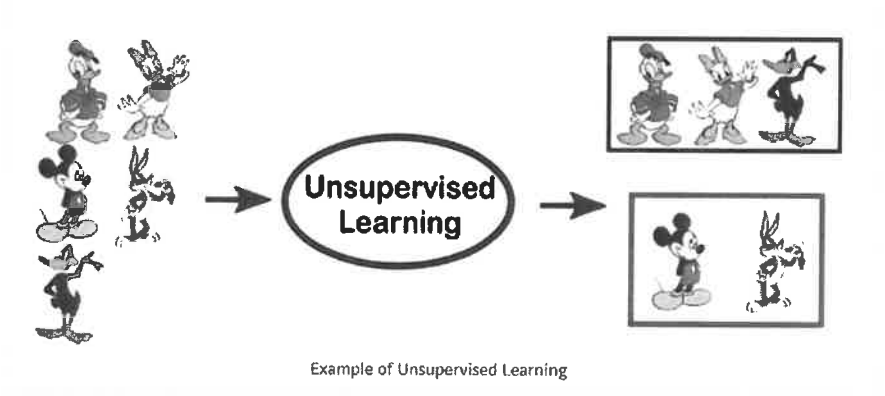


Diagram B



(6)

2.2 **"Ever since the technical revolution, we've been generating an immeasurable amount of data. As per research, we generate around 2.5 quintillion bytes of data every single day! It is estimated that by 2020, 1.7MB of data will be created every second for every person on earth."**

With regards to the statement above, discuss the need and importance of Machine Learning as the most in demand technology in today's market. (8)

2.3 Data Collection is regarded as a very important stage in the ML workflow process. From your understanding of Machine Learning why does data play such a vital role. (3)

2.4 "Data collected from the real world is transformed to a clean dataset."

List 3 ways on how one could clean the raw data in a dataset. (3)

QUESTION 3 [15 marks]

Study the dataset given below and answer the questions that follow:

Age	Height (in foot)	Weight (in kgs)
5	3.5	20
7	3.11	25
9	4.1	26
11	4.7	32
13	4.11	35
15	5.1	40
17	5.2	45
19	5.3	48
21	5.5	50
23	5.55	51
25	5.55	55

3.1 What can you conclude about the relationship between the 3 feature attributes given and explain the impact this may have on the ML model during the training phase of the ML workflow process. (5)

3.2 What statistical method can we implement in ML to help measure the strength of the relationship between 2 variables. (2)

3.3 With regards to ML, list 4 methods of cleaning a dataset and for each method explain what is the reason for removing these feature attributes. (8)

QUESTION 4 [10 marks]

Explain the working of each of the following ML algorithms using a diagram as well as the general Python code for implementation.

4.1 Linear Regression Algorithm (5)

4.2 Support Vector Machine (5)

QUESTION 5 [10 marks]

5.1 The code below implements a Decision Tree classifier to train and test a ML model using the muffins vs cupcakes dataset. Complete the code by filling in the missing information.

```
from __5.1.1___ import ____5.1.2____
model=____5.1.3_____
__5.1.4____(ingredients,type_label)
new =[[47, 26, 10, 10,  4,  1,  0,  0]]
predict=___5.1.5____(new)
if(predict)==0:
    print("muffin")
else:
    print("cupcake")
```

(5)

5.2 What is your understanding of the algorithm below. Exactly what is it used for in machine learning and how does it work. (5)

```
def get_correlation(data,threshold):
    corr_column=set()
    corrmatrix=data.corr()
    for i in range(len(corrmatrix.columns)):
        for j in range(i):
            if abs(corrmatrix.iloc[i,j])>threshold:
                colname=corrmatrix.columns[i]
                corr_column.add(colname)
    return corr_column
```

QUESTION 6 [15 marks]

6.1 Splitting your dataset is essential for an unbiased evaluation of prediction performance. The code below illustrates the application of the train, test, split method, with 60% of data used for training and 40% for testing. The program also uses evaluation measures of precision, accuracy and recall.

Study the incomplete program below and fill in the parts that are missing.

```
from sklearn.model_selection import ___6.1.1_____
from ___6.1.2_____ import accuracy_score,precision_score,___6.1.3____
from sklearn.ensemble import _____6.1.4_____
x_train,x_test,y_train,y_test=train_test_split(ingredients,type_label,
___6.1.5_=0.4)

model=RandomForestClassifier()
model.fit(x_train,____6.1.6____)
predict=model.predict(__6.1.7_____)
accuracy=____6.1.8_____(y_test,predict)
precision=____6.1.9_____(y_test,predict)
recall=   6.1.10   (y_test,predict)
accuracy,precision,recall
```

(10)

6.2 Splitting a dataset might also be important for detecting if your model suffers from one of two very common problems, called under fitting and overfitting. What is your understanding of each problem and explain how it affects the created model? (5)

QUESTION 7 [15 marks]

Study the following segment of Python code on data visualisations and answer the questions that follow.

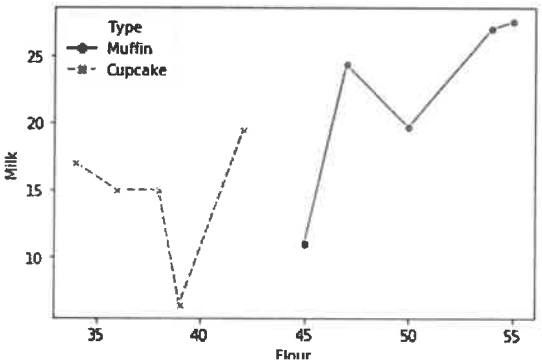
```
In [1]: #Plots and visualisations in ML
        import numpy as np
        import pandas as pd
        from matplotlib import pyplot as plt
        import seaborn as sns
        %matplotlib inline
```

```
In [3]: raw_data=pd.read_csv('muffins_cupcakes.csv')
        raw_data.head()
```

Out[3]:

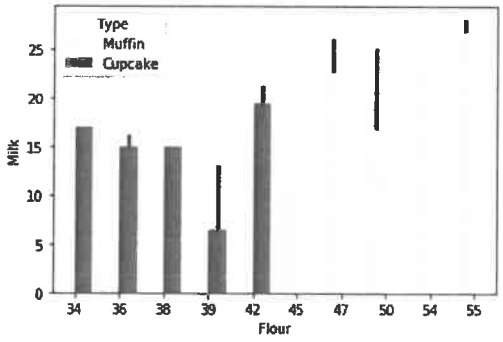
	Type	Flour	Milk	Sugar	Butter	Egg	Baking Powder	Vanilla	Salt
0	Muffin	55	28	3	7	5	2	0	0
1	Muffin	47	24	12	6	9	1	0	0
2	Muffin	47	23	18	6	4	1	0	0
3	Muffin	45	11	17	17	8	1	0	0
4	Muffin	50	25	12	6	5	2	1	0

7.1 Write the code that will display a lineplot as follows:



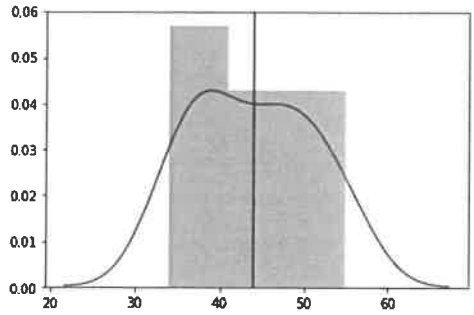
(5)

7.2 Write the code that will display a bar plot, in red, as follows:



(5)

7.3 Write the code that will display a histogram as follows. Note that the distribution plot is in red and the vertical mean axes line is in blue.



THE END

TOTAL=100