CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 2 December 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions carrying forty (40) marks. SECTION II has three (3) practical questions carrying sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:

- **A computer**
- **Jupyter Notebook**
- **Pycharm IDE**
- **Pyspark library**
- **Python**
- **Java JDK**
- **Hadoop software**

SECTION I (40 MARKS)

1. With the correct techniques, computers can detect hidden patterns in data without explicit instructions. This subset of artificial intelligence is referred to as _____. (2 marks)

2. K-Means Clustering is an Unsupervised Learning algorithm that organises an unlabelled dataset into distinct clusters. What does “K” represent? (2 marks)

3. The Pandas Library functionality that allows concurrent processing of large tabular data on a single machine even if it exceeds its available memory is known as _____. (2 marks)

4. What method is used to calculate the correlation between two variables by ranking the data instead of using raw values when the data is not normally distributed and may contain outliers? (2 marks)

5. A non-parametric approach to estimating the probability density function of a random variable using kernels as weights is called _____. (2 marks)

6. What term refers to a situation in machine learning where a model is overly simplistic and unable to capture the underlying patterns in the training data, leading to poor performance on both the training and validation datasets? (2 marks)

7. The _____ Python package is used for the creation, manipulation and study of the structure, dynamics and functions of complex graphs. (2 marks)

8. What validation technique in machine learning involves splitting the dataset into two parts; one for training the model and the other for evaluating its performance? (2 marks)

9. The type of database engineering that involves refining raw data and identifying the most predictive attributes for use in modelling is known as _____. (2 marks)

10. What principle in data mining is employed to uncover relationships between variables in large datasets by calculating the frequency of item sets and establishing association rules? (2 marks)

11. Which method for managing outliers in machine learning entails standardising or normalising data so that it has a mean of 0 and a standard deviation of 1? (2 marks)

12. The special set of scalar values that is associated with the set of linear equations in the matrix equations is referred to as _____. (2 marks)

13. What is the process in data analysis that reduces the number of variables or features in a dataset while maintaining its essential information? (2 marks)

14. The supervised machine learning algorithm commonly employed for binary classification tasks such as determining whether an email is spam is known as _____. (2 marks)

15. What is the hierarchical clustering method that organises data points into clusters by successively merging the closest pairs of clusters until only one cluster remains? (2 marks)

16. Which ethical issue in big data analytics addresses the divide between those who control and own large-scale datasets and those who do not? (2 marks)

17. Which technique is employed to assess a dataset for its uniqueness, logic and consistency but is unable to detect inaccurate data values? (2 marks)

18. What method is used in data preparation to fill in the missing values of a dataset to ensure accurate and comprehensive analysis? (2 marks)

19. In big data analytics, data scientists introduce a problem where some labels in a labelled dataset are underrepresented during the training process and can lead to poor model performance. This type of problem is called _____. (2 marks)

20. What probability distribution is typically used to model the number of events occurring in a fixed interval of time or space, especially when these events occur independently of each other? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21". Use the word processor document to save your answers to questions (a) to (d).

    (a) Using two arrays a: [1, 2, 3] and b: [4, 5, 6], write Python code that will form a new array by stacking the given arrays using NumPy and display the output. (6 marks)

    (b) Write Python code to print four random integers between 1 and 15 using Numpy. (4 marks)

    (c) Given that an array has the following values: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9], write Python code to display the following values from the array: [1, 3, 5, 7, 9]. (4 marks)

    (d) Given below is an employees.csv file and the content inside it as shown. Write code to read the file into a dictionary using the CSV module in Python and the csv. DictReader class. (6 marks)

Name	Department	Birthday – month
John Smith	HR	July
Alice Johnson	IT	October
Bob Williams	Finance	January

Save and Upload "Question 21". **(Total: 20 marks)**

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) and (b).

(a) Write the Python code that will create the dataset given in the table below using a dictionary called "productdata". (4 marks)

	A	B	C
1	**Product**	**Quantity**	**Unit price**
2	Smartphone	10	500000
3	Router	50	25000
4	Tablet	30	90000
5	Monitor	20	300000
6	Smartphone	5	250000
7	Tablet	10	30000
8	Router	20	10000
9	Headphones	15	75000
10	Monitor	10	150000

(b) Using the dataset, write the Python code to perform the following tasks:

(i) Convert "Product" into numerical values and define the predictor and the response variables. (4 marks)

(ii) Split the data into training and testing sets and display the feature importance. (6 marks)

(iii) Plot the decision tree for predicting the unit price. (6 marks)

Save and Upload "Question 22". **(Total: 20 marks)**

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (e).

(a) Create the dataset given in the table below using Microsoft Excel and save it as "bag.csv". (3 marks)

BAG NUMBER	BAG WEIGHT	BAG COST
25	**35**	**4000**
32	**42**	**3000**
45	**21**	**3500**
22	**28**	**4300**
23	**34**	**3800**
28	**25**	**3500**
40	**20**	**3450**
30	**28**	**3600**
48	**27**	**3800**
29	**19**	**4500**

(b) Load the dataset and write the Python code to import the necessary libraries to extract data and perform logistic regression. (3 marks)

(c) Assuming the threshold for high cost to be Sh.3500, write the Python code that will convert the regression problem into a binary classification problem. (3 marks)

(d) Write the Python code to split the dataset into training and testing sets and standardise the features. (5 marks)

(e) Write the Python code to create and train the logistic regression model, make predictions, print the coefficients and capture the screenshot. (6 marks)

Save and Upload "Question 23". **(Total: 20 marks)**

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 19 August 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **A computer**
- **Python program**

SECTION I (40 MARKS)

1. In the context of data analytics, name the tool that helps users to make sense of complex economic data by producing interactive graphs and charts? (2 marks)

2. Write down the output of the code below. (2 marks)
```
print(type("Data,Analytics"))
```

3. Which ethical principle emphasises the algorithms' responsibility and transparency when utilised in big data analytics? (2 marks)

4. In machine learning, identify the optimisation algorithm used to discover coefficient values for a function that minimise a cost function to the greatest extent possible. (2 marks)

5. Name the graph theory analysing term that deals with the ratio of existing to potential edges. (2 marks)

6. Minimising noise in a dataset is critical for increasing data analysis and model performance. Which technique, in a broader sense, uses logarithmic, square root or power transformations to stabilise variance and reduce noise? (2 marks)

7. Which network data is represented as a two-column dataframe with a from and to column, each row representing one tie and the ability to add further information? (2 marks)

8. The automated process of transforming massive amounts of unstructured text into quantitative data to reveal insights, trends and patterns is known as _____. (2 marks)

9. The special set of scalar values that are associated with the set of linear equations in the matrix equations are referred to as _____. (2 marks)

10. What is the most effective tool that data scientists use to query, store, examine and improve the performance of their models by extracting context information from graph data? (2 marks)

11. Which supervised machine-learning algorithm is commonly utilised in missing value imputation when doing classification or regression tasks? (2 marks)

12. What is the name of a tree-like graph created using hierarchical clustering to represent the hierarchical links between groups? (2 marks)

13. The deep learning algorithm which consists of multiple layers and is mainly used for image processing and object detection is known as _____. (2 marks)

14. What is the name of the correlation matrix measure of statistical dependence between two variables based on the data ranks rather than the actual values of the data? (2 marks)

15. Specify the decision boundary that, in the intuition of support vector machine, divides a given set of data points with various class labels. (2 marks)

16. In big data preprocessing, the technique that preserves the mean and the sample size is referred to as _____. (2 marks)

17. What is the name of a non-parametric technique that uses kernels as weights to estimate the probability density function of a random variable? (2 marks)

18. Association rule learning works on the concept of If and Else Statement, such as: if A then B. What name is given to the "if" element? (2 marks)

19. The topic modelling technique that is used to analyse relationships between documents and the terms they contain is referred to as _____. (2 marks)

20. Linear discriminant analysis (LDA) is a dimensionality reduction technique primarily utilised in supervised classification problems. Give one assumption of LDA. (2 marks)

SECTION I1 (60 MARKS)

21. Create a word processing document named "Question 21". Capture screenshots where necessary and use the word processor document to save your answers to questions (a) to (b).

    (a) Using the data in the table below, write the python code that will find the correlation matrix between speed and time. (6 marks)

Speed(km/hr)	Time(min)
80	60
120	45
60	80
100	42
105	40
106	41
70	65
50	72
65	68
76	63
85	58

(b) Use the table below to answer the questions that follow.

Student number	Marks scored
NAN5467	80
NAN5468	70
NAN3456	65
NAN5469	87
NAN5466	56
NAN3459	76
NAN5367	34
NAN5268	76
NAN3486	56
NAN3467	65
NAN3468	76

(i) Write the python code that will create a data dictionary for the data then print the minimum mark, the maximum mark and the range mark. (8 marks)

(ii) Write the python code that will calculate and print the interquartile range. (3 marks)

(iii) Write the python code that will print the standard deviation and variance of the marks. (3 marks)

Save and upload "Question 21". **(Total: 20 marks)**

22. Create a word processing document named "Question 22". Capture screenshots where necessary and use the word processor document to save your answers to questions (a) to (d).

The table below illustrates the number of boys and girls present in different grades at Atela primary school. Use it to answer the questions that follow.

Grade	Boys	Girls
1	6	24
2	5	18
3	11	23
4	8	16
5	4	17
6	11	22
7	14	24
8	6	25
9	13	27
10	16	22
11	15	26

(a) Import the necessary python libraries that will support the use of K-means. (2 marks)

(b) Write the python code that will create arrays for the two variables boys and girls, in the dataset and turn the data into a set of data points. (6 marks)

(c) Find the best value of K by running K-means across the data for the range of values. (6 marks)

(d) Fit the K-means algorithm and plot the different clusters assigned to the data. (6 marks)

Save and upload "Question 22". **(Total: 20 marks)**

23. Create a word processing document named "Question 23" Capture screenshots where necessary and use the word processor document to save your answers to questions (a) to (b).

(a) Use the data below to produce a linear regression between the two variables, length and width. (10 marks)

Width = 6,9,8,7,3,18,3,7,5,14,13,9,7
Length = 79,96,77,83,101,86,105,85,92,78,75,82,88

(b) Write the python code that will create and display an array that resembles two variables given by number of cars and distance covered by the car in km in a data set and convert the data into a set of points. (6 marks)

No of cars = 7, 14, 10, 5, 8, 15, 13, 7, 9, 12
Distance in Km = 32, 31, 33, 26, 27, 32, 34, 35, 33, 37

(c) Write a python program that will run k-means across the data in part (a) for a range of 15 possible values. For each value of K in the range the K-means model should be trained before plotting the distance covered against the number of clusters. (4 marks)

Save and upload "Question 23". **(Total: 20 marks)**

……

kasneb

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 22 April 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:
- **A computer**
- **Python program**

SECTION I (40 MARKS)

1. In big data analytics, what term is used to describe the rate at which data is generated, collected, consumed and processed? (2 marks)

2. What statistical approach is utilised in big data analytics to determine the best fit for a set of data points? (2 marks)

3. Which fundamental component of probability theory assigns a numerical value between 0 and 1 to each event, reflecting the probability of that event occurring? (2 marks)

4. A data matrix's rows and columns have been switched. What is the generated matrix name after this procedure? (2 marks)

5. Give the name of the big data technique that involves inspecting data for errors, duplicates, inconsistencies, redundancies and inappropriate formats. (2 marks)

6. In relation to linear regression analysis, state the name of noise emanating from the independent variables being highly correlated with each other than other variables. (2 marks)

7. The method for preserving the sample size and mean in big data preprocessing is called: (2 marks)

8. Name the type of graph that has edges but no directionality best suited for modeling objects that have reciprocal relationships. (2 marks)

9. What is the name given to a top-down data clustering technique that divides data points into individual clusters until each data point is unique? (2 marks)

10. The linear model for classification and dimensionality reduction that is commonly used for feature extraction in pattern classification problems is called: (2 marks)

11. Name the Python Matplotlib module that provides simple functions for adding plot elements such as lines, images and text to the axes in a plot. (2 marks)

12. What is the name of data mining and machine learning approach that is used to find frequent item groups in a dataset that appears together in transactions? (2 marks)

13. What is the name of deep learning model that has been trained to interpret and transform a certain sequence of data input into a sequential output? (2 marks)

14. What is the name of a Python activation function that translates any input to the 0–1 range and returns a probability? (2 marks)

15. What is the name of non-parametric approach for estimating the probability density function of a random variable using kernels as weights? (2 marks)

16. The Naive Bayes algorithm which is based on Bayes theorem and predicts the tag of a text such as an email or a newspaper article is called: (2 marks)

17. What is the name of low-level Python graph plotting package used as a visualisation utility? (2 marks)

18. Once linear transformations are applied, what is the name of non-zero vector that can only be changed by its scalar factor? (2 marks)

19. Name the legislative framework that establishes guidelines for the exchange of personal data between nations outside of the European Union. (2 marks)

20. Which ethical principle highlights the need to provide clear and comprehensible information regarding data processing activities? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21". Use the word processor document to save your answers to questions (a) to (e) below:

(a) Create a folder on the desktop and name it "CISSE". Create an Excel document shown below and save it as a comma separated version (CSV) file named "**Car data**". (3 marks)

Car Number	Model	Buying price	Selling price
KDD007H	Toyota	500000	750000
KCD675T	Nissan	650000	890000
KBT584R	Mazda	760000	988000
KDA987E	Toyota	456980	540900
KBD745F	Toyota	567800	768000
KBS346Y	Mazda	456700	567000
KBD349R	Nissan	904000	986000
KAH875J	Mitsubishi	657000	790000
KBX975W	Nissan	567000	775000
KCG885D	Toyota	573000	674000
KCT567P	Toyota	678000	890000
KCR764Q	Mitsubishi	573000	609000

(b) Write python programming code that will retrieve data from the .csv file and import the relevant libraries to perform the hierarchical agglomerative clustering. (4 marks)

(c) Write python programming code that will use the buying price and selling price to display the data on a graph at a later point. (2 marks)

(d) Write python code to draw a dendrogram that will find the optimal number of clusters and the highest vertical distance that does not intersect with any clusters. (4 marks)

(e) (i) Write Python code to create an instance of Agglomerative Clustering using the Euclidean distance as the measure of distance between points and ward linkage to calculate the proximity of clusters. (4 marks)

(ii) Use a shorthand notation to display all the samples belonging to a category as a specific colour. (3 marks)

Capture a screenshot to demonstrate how you have performed the above tasks.

Save and upload "Question 21".

(Total: 20 marks)

22. Create a word processing document named "Question 22". Use the word processor document to save your answers to questions (a) and (b) below:

(a) The age and speed of 13 cars as they passed a toll booth were recorded where the x-axis represents age (x = 5, 7, 8, 7, 2, 17, 2, 9, 4, 11, 12, 9, 6) and the y-axis represents speed (y = 99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86).

Write python code that implements the following and displays the output:

(i) Draw a scatter plot. (4 marks)

(ii) Draw the line of linear regression. (6 marks)

(b) Create the arrays that represent the values of the x and y axis as shown below:
x = [5, 7, 8, 7, 2, 17, 2, 9, 4, 11, 12, 9, 6]
y = [99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86]

Write python code that will perform the following tasks.

(i) Write a method that returns slope, intercept, r, p and standard error of linear regression of x and y. (3 marks)

(ii) Create a function that uses the slope and intercept values to return a new value that represents where on the y-axis the corresponding x value will be placed. (4 marks)

(iii) A function that runs through x array generating a new array with new values for the y-axis. (3 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Save and upload "Question 22".

(Total: 20 marks)

23. Create a word processing document named "Question 23". Use the word processor document to save your answers to questions (a) and (b) below:

ITEM NAME	UNITS SOLD
Apple	500
Kent	120
Ngowe	800
Lemon	150
Lime	90
Dent corn	100
Amylomaize	250

(a) Using the data given in the above table:

(i) Write a Python code to display the item name with the units sold. (6 marks)

(ii) Write the Python code that will display the pie chart of the units sold in percentage. (4 marks)

(b) The table below shows a report of marks obtained by different students in a test:

Student Name	Marks obtained (%)
Oscar	75
Joseph	62
Alice	81
Anne	59
Vincent	64
Abraham	76

Using Python:

(i) Write code that displays students name and marks obtained. (4 marks)

(ii) Write the code that will support the visualisation of their performance by using a donut chart. (6 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Save and upload "Question 23". **(Total: 20 marks)**

.....

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 4 December 2023. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **A computer**
- **Python program**

SECTION I (40 MARKS)

1. The element of big data ethics which relates to being open about the sources of data and how that data is applied are referred to as: (2 marks)

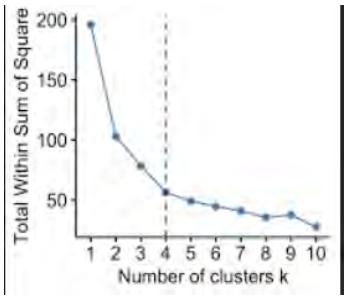
2. The hierarchical clustering algorithm that takes a "top-down" approach to clustering, which starts with all data points in one cluster and recursively divides them into smaller clusters which results in a hierarchical tree-like structure known as a dendrogram is called: (2 marks)

3. The name of the technique used to transform data in a way that it has a mean of 0 and a standard deviation of 1 is referred to as: (2 marks)

4. When carrying out text analysis in social media, which is the **MOST** popular library used to perform text analytics in Python? (2 marks)

5. The data mining principle which states that if an itemset is frequent, then all of its subsets must also be frequent is referred to as: (2 marks)

6. The method used to determine the optimal number of clusters in k-means clustering as shown in the figure below is called: (2 marks)



7. A clustering data mining method which is characterised by a centroid-based algorithm that minimises the variance of data points within a cluster is known as: (2 marks)

8. The contextual mining of text which identifies and extracts subjective information in source material and helps a business to understand the social opinion of their brand, product or service while monitoring online conversations is referred to as: (2 marks)

9. The type of relationship which exists between two variables such that as the value of one variable increases the other decreases is referred to as: (2 marks)

10. The type of data analytics that looks at past data to give an account of what has happened before is called: (2 marks)

11. In big data analytics, the divergence between a model's predictions and the real-world results is often quantified and analysed through a metric known as: (2 marks)

12. The supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event or observation is referred to as: (2 marks)

13. Which layer in Convolutional Neural Network (CNN) is used to reduce the spatial dimensions of the feature maps, retaining the most important information: (2 marks)

14. The machine learning library in Python that provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction is referred to as: (2 marks)

15. In today's fast-paced digital era, the abundance of continuously generated real-time data from various sources, including social media updates, sensor data, and financial market data, has presented a unique challenge and opportunity for businesses and organisations. The process where data that is generated continuously and in real-time is called: (2 marks)

16. The machine learning pipeline activity where the data is processed into a well-organised format is called: (2 marks)

17. The process of selecting, manipulating and transforming raw data into features that can be used in supervised learning is called: (2 marks)

18. The vector that is associated with a set of linear equations is referred to as: (2 marks)

19. With reference to NetworkX python library, state the function that will initialise an empty graph to facilitate addition of nodes and edges. (2 marks)

20. The type of Naive Bayes algorithm which assumes that all the variables have a normal distribution is referred to as: (2 marks)

SECTION II (60 MARKS)

21. Create a word document called "Question 21" to save and capture the screenshots for Questions 21. Create Python script called "BigDAnalytics" to perform the tasks in questions (a) to (g) below.

Use the CSV data provided below to answer the questions that follow:

Customer ID	Gender	Age	Purchase Amount	Town
1	Male	35	10,000	Nairobi
2	Female	28	7,500	Mombasa
3	Male	42	12,500	Nairobi
4	Female	30	8,560	Kisii
5	Male	22	9,575	Mombasa
6	Female	38	11,080	Nakuru
7	Female	45	13,520	Nakuru
8	Male	29	7,050	Kisii
9	Female	27	8,000	Nairobi
10	Male	31	10,575	Nairobi

(a) Using the data provided in the above table, create a CSV dataset called "salesInsights.csv" using Excel spreadsheet and save it in the appropriate location for data analytics (4 marks)

(b) Write Python statements to load the panda and matplotlib libraries assigning them relevant pseudonyms on the script "BigDAnalytics" (2 marks)

(c) Load the dataset "salesInsights" into a Python object called "data" on the script "BigDAnalytics" and display the first few rows of the dataset. (2 marks)

(d) Get and display the summary of the dataset to understand its basic statistics, structure and the data type. (2 marks)

(e) Visualise the dataset using an histogram with 20 bins, clearly showing the title, the X and Y axis. (4 marks)

(f) Calculate and display the aggregated total for the purchase amount by gender. (3 marks)

(g) Get the insights from the data for the highest purchase amount and the age of the youngest customer. (3 marks)

Save Question 21 document and upload.

(Total: 20 marks)

22. Create a word processing document named "Question 22". Use the word processor document to save your answers to questions (a) to (d).

(a) Given the data set X= 2, 5.5, 1.55, 6.1, 5.75, 6.8, 4.7, write the python codes to find the following the mode, variance and standard deviation (6 marks)

(b) Create an Excel document shown below and save it as a comma separated version (CSV) file named "Items". **(**3 marks**)**

Wine	Chips	Bread	Butter	Milk	Apple
Wine		Bread	Butter	Milk	
		Bread	Butter	Milk	
	Chips				Apple
Wine	Chips	Bread	Butter	Milk	Apple
Wine	Chips			Milk	
Wine	Chips	Bread	Butter		Apple
Wine	Chips			Milk	
Wine		Bread			Apple
Wine		Bread	Butter	Milk	
	Chips	Bread	Butter		Apple
Wine			Butter	Milk	Apple
Wine	Chips	Bread	Butter	Milk	
Wine		Bread		Milk	Apple
Wine		Bread	Butter	Milk	Apple
Wine	Chips	Bread	Butter	Milk	Apple
	Chips	Bread	Butter	Milk	Apple
	Chips		Butter	Milk	Apple
Wine	Chips	Bread	Butter	Milk	Apple
Wine		Bread	Butter	Milk	Apple

(c) Write python programming code that will retrieve data from the .csv file and import the relevant libraries to implement Apriori algorithm. (4 marks)

(d) Write python programming code that will convert pandas DataFrame into a list of lists, build the Apriori model and print out the rule. (7 marks)

Save Question 22 document and upload.

(Total: 20 marks)

23. Create a word processing document named “Question 23” use the word processor document to save your answers to questions (a) and (b).

(a) Given the two variables below, write the python code that will import the necessary libraries then find the correlation between the variables. The title should be “Correlation between DDMA and DCNA”.
(12 marks)

DDMA=1,5,4,3,9,7,8,2
DCNA=2,4,8,11,12,14,10.9

(b) You are given a list of integer values representing the ages of five persons 34, 23,45,67,43 and their associated weights 56, 48,89,57,73 respectively.

Required:

(i) Create two list data structures “lstage” and “lstwght” to store the persons’ ages and their weights respectively. (3 marks)

(ii) Create and display a well labeled scatter plot to show the relationship between the two lists, with the x-axis representing the age and the y-axis representing the weights. (5 marks)

Save Question 23 document and upload.

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 21 August 2023. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **A computer**
- **Python program**

SECTION I (40 MARKS)

1. The excel function that helps rotate (switch) the values from rows to columns and vice versa is called: (2 marks)

2. The term that describes the diversity and range of formats and types of big data available is referred as: (2 marks)

3. The type of data transfer where a user is informed of data transfer in a timely and straight forward manner is referred to as: (2 marks)

4. The big data ethics that involves concepts such as liberty, autonomy, security, data protection and data exposure is referred to as: (2 marks)

5. The type of visualisation that displays data as an interactive map allowing users to have a quick overview of the data sets is referred to as: (2 marks)

6. The statistical technique that is used to evaluate the relationship between two variables in a data set is referred to as: (2 marks)

7. State the term that best describes the type of statistical modelling that uses unsupervised Machine Learning to identify clusters or groups of similar words within a body of text. (2 marks)

8. All subsets of a frequent itemset must be frequent. If an itemset is infrequent, all its supersets will be infrequent. State the name of the algorithm depicts this scenario. (2 marks)

9. The special set of scalar values that are associated with the set of linear equations in the matrix equations is referred to as_____. (2 marks)

10. The big data mining technique that is used to identify critical abnormalities in data that could be indicative of a deeper issue is known as: (2 marks)

11. The Python's cross-platform, data visualisation and graphical plotting library for plotting histograms, scatter plots and bar charts is called: (2 marks)

12. State the term that best defines a NoSQL graph database that can be used to persist data in Python web applications and data projects. (2 marks)

13. Write a simple python expression that will print a data type of the variable g. (2 marks)

14. In data analytics, the clustering approach that involves grouping data into a tree of clusters is referred to as: (2 marks)

15. The type of supervised machine learning algorithm which assumes that the effect of a particular feature in a class is independent of other features is called: (2 marks)

16. The data mining technique that supports the identification of underlying relations between different items is referred to as: (2 marks)

17. The type of network that is modelled on the human brain and consists of connected nodes depicting the flow of information from one layer to the next is referred to as: (2 marks)

18. The statistical model is often used for classification and predictive analytics. It estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. This model is referred to as: (2 marks)

19. The machine learning technique that is capable of performing both regression and classification tasks with the use of multiple decision trees is referred to as: (2 marks)

20. The big data dimension reduction technique used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller one that still contains most of the information in the large set is referred to as: (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question Twenty One" and use the word processor document to save your answers to questions (a) to (e) below.

(a) Create a folder on the desktop called "DDMA". Create an excel document shown below and save it as a comma separated values (CSV) file named "**Customer**". (3 marks)

CustomerID	Gender	Age	Annual income (Ksh)	Spending score (1-100)
1	Male	33	190	39
2	Male	15	210	89
3	Female	62	200	6
4	Female	55	230	77
5	Male	45	180	40
6	Male	15	180	76
7	Female	62	190	6
8	Female	31	190	94
9	Male	23	190	3
10	Female	12	170	72
11	Male	35	170	54
12	Female	65	160	6
13	Female	20	160	63
14	Male	53	200	71
15	Male	24	210	89

(b) Write python programming code that will retrieve data from the customer .csv file and import the relevant libraries to perform the hierarchical agglomerative clustering. (4 marks)

(c) Write python programming code that will use the annual income and spending score to display the data on a graph at a later point. (2 marks)

(d) Write python code to draw a dendrogram that will find the optimal number of clusters and the highest vertical distance that does not intersect with any clusters. (4 marks)

(e) Write python code to create an instance of Agglomerative Clustering using the Euclidean distance as the measure of distance between points and ward linkage to calculate the proximity of clusters. Use a shorthand notation to display all the samples belonging to a category as a specific colour. (7 marks)

Save Question Twenty One document and upload.

(Total: 20 marks)

22. Create a word processing document named "Question Twenty Two". Use the word processor document to save your answers to questions (a) to (d) below:

(a) Write python code to create the two matrices (X1 and X2) below using NumPy array and add them. (6 marks)

X1 3, 6, 9
7, 8, -11
9, 5, 17

X2 10, -12, 6
13, 5, -12
14, 7, -16

(b) Write python code to transpose matrix X2. (4 marks)

(c) Write python code to print the first three rows and first 2 columns of matrix X1. (4 marks)

(d) Describe the three techniques for handling outliers in a dataset. (6 marks)

Save Question Twenty Two document and upload.

(Total: 20 marks)

23. Create a word processing document named "Question Twenty Three". Use the word processor document to save your answers to questions (a) to (c) below:

(a) Create a folder on the desktop and call it "DDMA". Create an Excel worksheet shown below, save it as a comma separated values (CSV) file named "**Edgelist**". (3 marks)

REGION	TENURE	AGE	MARITAL	ADDRESS	INCOME	EMPLOY	GENDER	CUSTCAT
2	13	33	1	11	99	5	0	1
3	12	44	1	9	89	29	0	4
3	13	54	0	8	112	5	1	3
2	67	32	1	6	45	0	1	1
2	54	23	1	12	87	8	0	3

(b) Write python programming code that will retrieve data from the Edgelist .csv file and import the relevant libraries for K-nearest neighbour clustering. (4 marks)

(c) Using the target column, 'custcat' that categorises the customers into four groups, write python code that will train and predict the model. Use K-Nearest neighbor. (7 marks)

(d) Write python code that will improve the model and find out the optimal k value. (6 marks)

Save Question Twenty Three document and upload.

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 24 April 2023. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions carrying forty (40) marks. SECTION II has three (3) practical questions carrying sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **A computer**
- **Python program**

SECTION I (40 MARKS)

1. Based on the BODMAS rule of operator precedence, what will be the output of the following code snippet when used in python? (2 marks)
   print (2**3) +(5+6)**(1+1))

2. The class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input is referred to as: (2 marks)

3. Multiple types of analytics provide organisations and people with information that can drive innovation, improve efficiency and mitigate risk. Which type of analytics is associated with following tasks: (2 marks)
   - Simulation
   - Forecasting

4. The type of analytics that relate to automated process of translating large volumes of unstructured text into quantitative data to uncover insights, trends and patterns is known as: (2 marks)

5. _____data analysis is a set of procedures designed to produce descriptive and graphical summaries of data with the notion that the results may reveal interesting patterns. (2 marks)

6. The big data mining technique that is used to identify critical abnormalities in data that could be indicative of a deeper issue is referred to as: (2 marks)

7. Big Data is characterised by volume, velocity, variety and veracity. The difference between actual value and the measured value of an observation is known as: (2 marks)

8. In graph theory the connections between the nodes that might hold properties are referred to as: (2 marks)

9. What is the name given to a collection of raster or vector data that can be used to display a map on mobile devices and within a browser? (2 marks)

10. What is the name given to the technique that improves the efficiency of the apriori algorithm by creating a dictionary that stores the candidate item sets as keys, and the number of appearances as the value? (2 marks)

11. The type of unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science is referred to as: (2 marks)

12. Which mechanism is used to refer to visual representation of a text and is increasingly being employed as a simple tool to identify the focus of written material? (2 marks)

13. The process of preparing the raw data and making it suitable for a machine learning model is referred to as: (2 marks)

14. Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. Which is the third step in the big data analytics process? (2 marks)

15. According to Bayesian statistics, what is the name given to the revised or updated probability of an event occurring after taking into consideration new information? (2 marks)

16. With reference to NetworkX python library, state the function that will initialise an empty graph to facilitate addition of nodes and edges. (2 marks)

17. The Naive Bayes algorithm that assumes all the variables have a normal distribution is called_____? (2 marks)

18. The term _____used to describe the type of analytics that provides marketers with a real-time snapshot of online conversations so they can understand their customers. (2 marks)

19. Association rule mining finds interesting associations and relationships among large sets of data items. The rule also shows how frequently an itemset occurs in a transaction. State two parts of an association rule. (2 marks)

20. The linear model for classification and dimensionality reduction that is commonly used for feature extraction in pattern classification problems is referred to as: (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" use the word processor document to save your answers to questions (a) and (b) below:

(a) Using Appropriate IDE and python perform the following:

(i) Define four bins of 0-25,26-50,51-75,76-100 (6 marks)

(ii) Plot a histogram of marks obtained by students. (4 marks)

(b) Write the python using appropriate IDE and python perform the following:

(i) Use the diagram below to develop appropriate code that displays the pie chart of the list of students enrolled with appropriate subjects: (8 marks)

Subject	**Number of enrolled students**
English	23
Mathematics	17
Biology	35
Chemistry	29
Physics	12

(ii) Generate pie chart display. **(2 marks)**

Capture screenshots to demonstrate how you have performed the above task.

Upload Question 21

(Total: 20 marks)

22. Create a word processing document named "Question 22" use the word processor document to save your answers to questions (a) to (c) below:

(a) Create the excel document shown below and save them as a comma separated version (CSV) file named **physicsmarks** and **mathematicsmarks** respectively. (5 marks)

mathematicsmarks

	A	B	C
1			
2	**StudentID**	**Maths_marks**	
3	**SD001**	**78**	
4	**SD002**	**45**	
5	**SD003**	**56**	
6	**SD004**	**65**	
7	**SD005**	**76**	
8	**SD006**	**69**	
9			
10			

physicsmarks

	A	B	C
1			
2	**StudentID**	**Physics_marks**	
3	**SD001**	**61**	
4	**SD002**	**72**	
5	**SD003**	**71**	
6	**SD004**	**54**	
7	**SD005**	**57**	
8	**SD006**	**51**	
9			

(b) Write python programming codes that will load the two data sets, convert both data sets into data frames and merge them. (7 marks)

(c) The speed of 15 cars is given by Speed = [99,86,87,88,111,86,103,87,94,78,77,85,86,85,97]. Write the python code to find the following:

(i) The median.

(ii) The mean

(iii) The mode (8 marks)

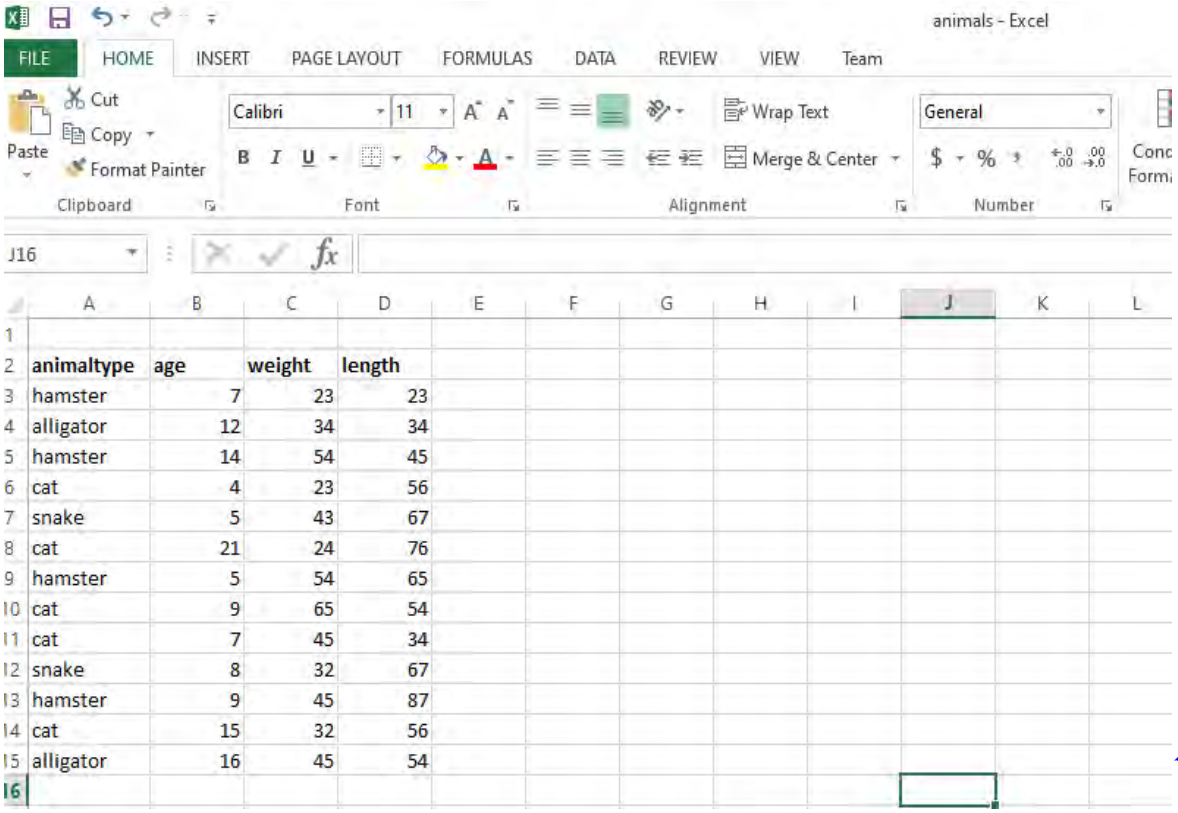
Capture the screenshots to demonstrate how you have performed the above task.

Upload Question 22

(Total: 20 marks)

23. Create a word processing document named “Question 23”. Use the word processor document to save your answers to questions (a) to (c) below:

(a) The spreadsheet below shows the types of animals and their respective age, weight and length.



	A	B	C	D
1				
2	**animaltype**	**age**	**weight**	**length**
3	hamster	7	23	23
4	alligator	12	34	34
5	hamster	14	54	45
6	cat	4	23	56
7	snake	5	43	67
8	cat	21	24	76
9	hamster	5	54	65
10	cat	9	65	54
11	cat	7	45	34
12	snake	8	32	67
13	hamster	9	45	87
14	cat	15	32	56
15	alligator	16	45	54
16				

Required:

(i) Create a folder named Animal. Enter the above data and save the file as animals in csv format within the Animal folder. (3 marks)

(ii) Write python programming code that will find the average weight of all the snakes, cats, hamsters, and alligators. (7 marks)

(b) Write python programming code that will generate the numbers for "actual" and "predicted" values in a dataset. (10 marks)

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 5 December 2022. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:
- **A computer**
- **Python program**

SECTION I

1. The category of numerical data in machine learning that varies over time and can have separate values at any given point is referred to as: (2 marks)

2. State the name given to a technique that converts higher dimensions dataset into lesser dimensions dataset to ensure provision of similar information. (2 marks)

3. State the name of the regression that describes data and explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables? (2 marks)

4. Identify the algorithm used for frequent itemset mining and association rule learning over relational databases. (2 marks)

5. The type of Naïve Bayes model that is used in classification and assumes that features follow a normal distribution is known as _____. (2 marks)

6. State the first step when explaining the working of K-Nearest Neighbor algorithm of machine learning. (2 marks)

7. Identify the learning approach with reference to problem solving features that takes input for a given problem then produce the end result (2 marks)

8. The type of linear regression in machine learning where more than one independent variable is used to predict the value of a numerical dependent variable is referred to as_____. (2 marks)

9. Identify the python package that can be used with natural language processing. (2 marks)

10. State the unsupervised machine learning technique that is capable of scanning documents and detecting phrase patterns within them. (2 marks)

11. Suggest an area of concern in big data ethics that outline the potential for immoral use of data. (2 marks)

12. State the term in big data that refers to the identification and removal of meaningless information present in data? (2 marks)

13. The python library that supports working with arrays and matrices is referred to as _____. (2 marks)

14. The data science concept that can be used to analyze a variety of network information and has various applications is referred to as _____. (2 marks)

15. State the term that describes the feature that will allow a data scientist to reduce a linear operation to much simpler problems? (2 marks)

16. A type of neural network that supports the modelling of time dependent and sequential data problems is referred to as_____. (2 marks)

17. State the term that can be used to describe the data visualization tool that is laid out on a map or table and uses different nuances and intensities of colors to represent its data? (2 marks)

18. The big data technology that is used to query continuous data flow and detect conditions quickly within a small time period is referred to as _____. (2 marks)

19. The unsupervised machine learning approach that is capable of scanning a set of documents, detecting phrase patterns within them and automatically clustering word groups is referred to as _____. (2 marks)

20. Identify the method that is used to predict the behavior of dependent variables in regression analysis. (2 marks)

SECTION II

21. Create a word processor document named "Question 21" and use the word processor document to save your answers to questions (a) to (b) below:

(a) Write a python code to draw a scatter plot with x values as 1,2,3,5,6,7,8,9,10, 12,13,14,15,16,18,19,21,22 and y values as 100,90,80,60,60,55,60,65,70,70,75,76,78,79,90,99,99,100.

Plot the line of linear regression on the scatter plot and display the output. (10 marks)

(b) Using appropriate python functions, draw the line of polynomial expression. (10 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Upload Question 21 document.

(Total: 20 marks)

22. Create a word processor document named "Question 22" and use the word processor document to save your answers to questions (a) to (c) below:

(a) Write a python code that will create and display an array that resembles two variables given by x (number of clusters) and y (inertia) in a data set and convert the data into a set of points.

X= 3, 4, 11, 4, 3, 10, 13, 7, 9, 13
Y= 20, 18, 23, 16, 17, 26, 24, 23, 22, 21 (9 marks)

(b) Write a python code that will run k-means across the data in part (a) for a range of 10 possible values. For each value of K in the range, the K-means model should be trained before plotting the inertia against the number of clusters. (10 marks)

(c) Write a python code that will fit the K-means algorithm at the point where the inertia becomes more linear then the code should plot the different clusters assigned to the data. (6 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Upload Question 22 document.

(Total: 25 marks)

23. Create a word processor document named "Question 23" and use the word processor document to save your answers to questions (a) and (b) below:

(a) Create a folder on the desktop and call it "BIG DATA". Create an excel document shown below, save it as a comma separated version (CSV) file named "**Analytics**". (3 marks)

	A	B	C	D
1				
2	SerialNo	StudyHours	MarksScored	
3	1	15	56	
4	2	25	93	
5	3	14	61	
6	4	10	50	
7	5	18	75	
8	6	0	32	
9	7	16	85	
10	8	5	42	
11	9	19	70	
12	10	16	66	
13	11	20	80	
14				

(b) Write python programming codes that will perform the following tasks.

(i) Retrieve data from analytics.csv file in question (a) above and print a summary of its description. (2 marks)

(ii) Split the data into training and testing sets. (3 marks)

(iii) Split each set into input and output attributes data. (3 marks)

(iv) Visualise the relationship between attributes in training and test data sets using scatter graph. (4 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Upload Question 23 document.

(Total: 15 marks)

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA ANALYTICS

MONDAY: 1 August 2022. Afternoon paper. **Time Allowed: 3 hours.**

This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. All questions are compulsory. Marks allocated to each question are shown at the end of the question.

SECTION I

1. _____ is a big data dimension reduction technique used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller set that still contains most of the information in the large set. (2 marks)

2. The _____ learning, is an area of machine learning concerned with how agents ought to take actions in an environment to maximise some notion of cumulative reward. (2 marks)

3. A deep learning neural network designed for processing structured arrays of data such as images is called _____. (2 marks)

4. _____is an umbrella term used to describe datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. (2 marks)

5. The data pre-processing technique that is used for handling missing values in dataset by use of statistical metrics is known as data _____. (2 marks)

6. You have been tasked as the data scientists to develop a machine learning model for a supermarket that determine the two products customers purchase together. The _____ algorithm is the best algorithm to carry out this task. (2 marks)

7. In social network analysis, _____ is the metric used to measure how quickly an entity can access more entities in a network (2 marks)

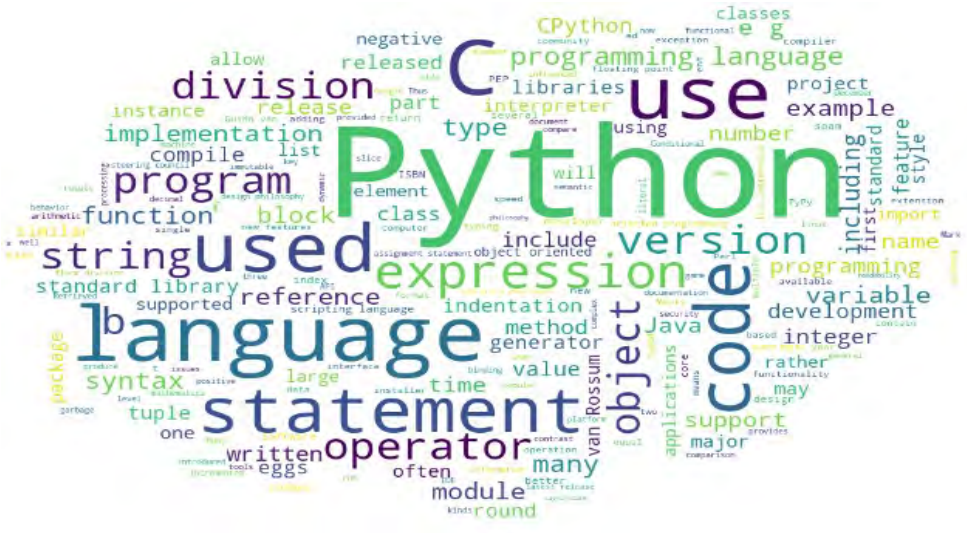
8. In a decision tree classification algorithm, all nodes except the root node are called leaf nodes. True or False? (2 marks)

9. When collecting data for analytics from participants, it is significant to give participants a reasonable and accurate understanding of how their data will be used. This is called _____. (2 marks)

10. In Natural Language Processing (NLP), eliminating affixes from a word such as “eating to eat” is called _____, as used in machine learning tasks focused on human languages. (2 marks)

11. The _____ network can be simply expressed as an information processing system designed to imitate the human brain structure and functions based on its source, features, and explanations. (2 marks)

12. Identify the name given to the data visualisation technique shown below? (2 marks)



13. The data structure made up of connection of nodes and edges, used data analytics for in fraud detection is called a _____? (2 marks)

14. The big data process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency is known as? (2 marks)

15. The _____regression model is used to solve classification problems in machine learning. (2 marks)

16. The activation function used in machine learning to transform the output of fully connected layer into a probability distribution of values which adds up to one is known as _____. (2 marks)

17. Given, import numpy as np, type a statement to Convert the python list, my_list = [1,2,3], into a single dimensional numpy array, without creating any spaces. (2 marks)

18. The artificial intelligence technique where large amount of data from an unlabeled dataset is divided into multiple categories according to internal similarity of the data and data in the same category is more similar than that in different categories is called? _____. (2 marks)

19. The clusters obtained in k-means meet the following conditions: (1) Objects in the same cluster are dissimilar (2) The similarity of objects in different clusters is high. True or False? (2 marks)

20. Consolidation of data from different sources into one unified hub, such as a data warehouse, so that users have centralised access to all the information they need for data mining, business intelligence reporting, and operational purposes is known as data _____ (2 marks)

SECTION II

21. Create a word processing document named “Question 21” use the word processor document to save your answers to questions (a) to (h) in form of screenshots.

Type the following dataset into excel document. Create a folder in drive C: and name it **AnalyticsAug**.

First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
Second	UON	Good	Yes	N	0
Second	MOI	Bad	Yes	Y	0
Second	JKUAT	Good	Yes	N	1
Second	JKUAT	Good	Yes	N	1
Second	JKUAT	Good	No	Y	1
First	UON	Good	No	Y	2
Second	UON	Bad	No	N	2

Required:

(a) Save the excel document in **AnalyticsAug** folder as a comma separated version (CSV) file and name it ***Colleges.csv.*** (3 marks)

(b) State python code to import the necessary python libraries. (3 marks)

(c) Key in python code to extract data from the source and print the output. (3 marks)

(d) State python code to transform the data into a numeric array and print the output. (3 marks)

(e) Type python code to separate independent variables from dependent variable and print the output. (3 marks)

(f) Explain the python code you would use to normalise the data and print the output. (3 marks)

(g) Explain how you would apply K-means algorithm to create clusters. (3 marks)

(h) State python code to visualise the generated clusters. (4 marks)

Upload Question 21 document.

(Total: 25 marks)

22. Create a word processing document named "Question 22" use the word processor document to save your answers to questions (a) to (e) in form of screenshots.

(a) Type the following dataset into excel document. Create a folder in drive C: and name it **Studentsdata**.

Save the excel document in **Studentsdata** folder as a comma separated version (CSV) file and name it ***Assignmentmarks.csv.*** (3 marks)

3		
4		**Marks**
5	**0**	**4**
6	**1**	**28**
7	**2**	**59**
8	**3**	**59**
9	**4**	**64**
10		
11		
12		
13		

(b) Write python code that loads data from the .csv file, prints the entire data, then computes and print the measures of central tendency. (5 marks)

(c) Write python code that will print the minimum mark, the maximum mark then, calculate and print the range. (3 marks)

(d) Write python code that will calculate and print the interquartile range. (5 marks)

(e) Write python code that will calculate and print the standard deviation and variance of the marks. (4 marks)

Upload Question 22 document.

(Total: 20 marks)

23. Create a word processing document named "Question 23" use the word processor document to save your answers to questions (a) to (d) in form of screenshots.

(a) Type the following dataset into excel document. Create a folder in drive C: and name it **Patientsdata**.

Save the excel document in **Patientsdata** folder as a comma separated version (CSV) file and name it ***Patientsweight.csv***. (3 marks)

	A	B	C	D	E	F
1						
2	**Patienttype**	**FirstName**	**Age (Years)**	**Weight (Kgs)**	**Height (meters)**	
3	Child	Joseph	8	15	2.3	
4	Teeneger	Millicent	15	16	3	
5	Youth	Ezra	28	43	4	
6	Old age	Walter	71	51	4	
7	Teeneger	Daniel	13	61	4	
8	Child	Grace	9	28	2.2	
9	Child	Vincent	4	26	1.8	
10	Teeneger	Alice	16	53	3.1	
11	Youth	Patricia	23	48	5	
12	Youth	Elizabeth	30	54	4.2	
13	Old age	Fidelis	67	46	5.2	
14	Old age	Dorcas	76	54	4.8	
15						
16						
17						

(b) Write python code that will find and print the average weight of all the patients by patient type. (4 marks)

(c) Write python code that will group the unique values from the patient column. (4 marks)

(d) Write python code that will find the mean of the height column for each patient group. (4 marks)

Upload Question 22 document.

(Total: 15 marks)

……..