****

**ZU/WI/25/EXM/6**

**UNIVERSITY EXAMINATION 2023/2024**

**ORDINARY EXAMINATION FOR BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY / COMPUTER SCIENCE / SOFTWARE ENGINEERING.**

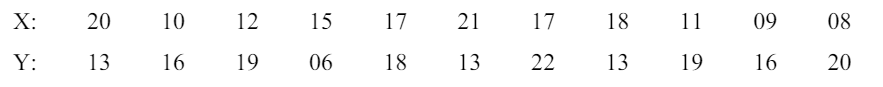
**BDM 221 DATA SCIENCE PROGRAMMING.**

**DATE: AUGUST 2024 TIME: 2 HOURS**

**Instructions: Answer Question ONE (Compulsory) and any other Two Questions**

**QUESTION ONE (30 MARKS) – COMPULSORY**

1. Imagine you're tasked with analyzing customer reviews for a new product your company has launched. To do this effectively, you'll need to use various data science techniques. Give the meaning of the following technologies in this context: **(5 Marks)**
2. Classification
3. Decision tree
4. Prediction
5. Data cleaning
6. Clustering
7. Given the data below, answer the questions that follows:



1. By showing all the steps, calculate the linear regression equation for the two sets of data. **(6 Marks)**
2. Using Python Scikit learn, write a program to show linear regression output clearly indicating dependent and independent variables. **(6 Marks)**
3. You have been hired by a healthcare organization to develop a predictive model to identify patients at risk of readmission within 30 days of discharge. The organization wants to use this model to allocate resources more effectively and improve patient outcomes.
   1. Outline the data science lifecycle you would follow to develop this predictive model. **(3 Marks)**
   2. Discuss how you would evaluate the effectiveness of your model and the metrics you would use. **(2 Marks)**
   3. Describe the ethical considerations and data privacy issues you need to address in this healthcare scenario. **(2 Marks)**
4. Suppose a certain bank wants to deploy a new system for assessing credit worthiness of its customers. The new system uses a feed forward network with a supervised learning algorithm. Suggest what the bank should have before the system is used. Discuss problems associated with this requirement. **(6 Marks)**

**QUESTION TWO (20 MARKS**

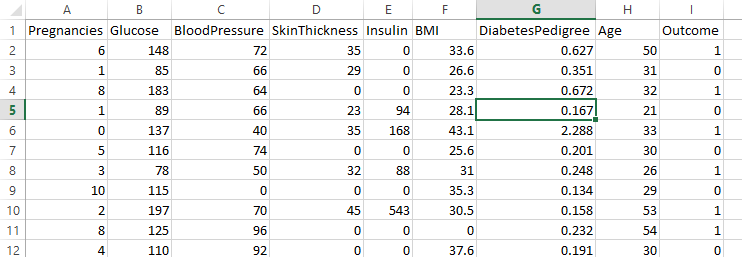
1. Data pre-processing and conditioning is one of the key factors that determine whether a data science project will be a success or not. For each of the following areas, describe the affect this issue can have on our data science session and what techniques can we use to counter this problem.
2. Noisy data **(2 Marks)**
3. Missing data **(2 Marks)**
4. Data normalization and scaling **(2 Marks)**
5. Data type conversion **(2 Marks)**
6. Attribute and instance selection **(2 Marks)**
7. You are a data scientist working for an e-commerce company. You have been given a dataset containing information about customer transactions, including the following columns: customer\_id, transaction\_id, transaction\_date, product\_id, quantity, and transaction\_amount.
8. Write a Python program to load the dataset, display the first 10 rows, and get summary statistics for the numerical columns. **(3 Marks)**
9. Give a program in Python to plot the distribution of transaction\_amount and identify any outliers using a boxplot. **(4 Marks)**
10. Write a program to calculate and plot the total transaction amount per day in Python language. **(3 Marks)**

**QUESTION THREE (20 MARKS)**

1. Describe the procedure for processing, analyzing and presenting data obtained from:
   1. A map **(2 Marks)**
   2. Semi-structured questionnaire **(2 Marks)**
   3. Key informant interview **(2 Marks)**
   4. Spatial imagery **(2 Marks)**
2. You are a data scientist working for a car dealership company. You have been given a dataset containing information about the vehicles in the dealership's inventory. The dataset includes columns such as vehicle\_id, brand, model, year, mileage, price, and color. Additional Information: The mileage column represents the number of miles the vehicle has been driven. The price column represents the selling price of the vehicle.
3. If you were required to change the data type of the year column from string to integer in the dataset, how would you do this programmatically in Python program. Give the source Code. **(3 Marks)**
4. Write a Python program that loads the dataset containing vehicle information, filters it to include only vehicles with mileage less than 50,000 miles, and displays the filtered dataset. **(4 Marks)**
5. As a Health Records and Information Officer, discuss how you would collect, analyze and present data in a health facility, tools you would use and the various measures you would use in your analysis. **(5 Marks)**

**QUESTION FOUR (20 MARKS)**

1. Using appropriate examples, elaborate the usage of the following Python libraries in data science programming. **(5 Marks)**
2. Pandas
3. Matplotlib
4. Scipy
5. Scikit-learn
6. Tensorflow
7. Use the Pima Indian Diabetes dataset shown below to perform the following tasks using Python programming language:



1. Import the required libraries for the decision tree analysis and load in the required data. **(3 Marks)**
2. Determine the target and feature variables **(3 Marks)**
3. Dividing the data into training and testing sets in the ratio of 70:30. **(3 Marks)**
4. Building the Decision Tree Model using scikit-learn. **(3 Marks)**
5. Determining how accurately the classifier predicts the outcome. The accuracy is computed by comparing actual test set values and predicted values.**(3 Marks)**