Strathmore UNIVERSITY

School of Computing and Engineering Sciences

Master of Science in Sustainable Energy Transitions

End of Semester Examination

MSET 8103: Data Science Concepts

Date: 16th August 2024 **Time 18:00-20:30 Hours**

Instructions: Answer Question **ONE** and any other **TWO** Questions

Question ONE (20 MARKS)-Compulsory

a) In a large city, the energy utility company supplies electricity to residential and commercial buildings. To manage peak demand periods, the utility company uses a system called "demand response," where they encourage consumers to reduce their electricity usage during peak hours in exchange for financial incentives. However, many consumers agree to participate in demand response programs but do not reduce their usage as expected. Research suggests that about 15% of consumers who agree to reduce their electricity consumption during peak periods fail to do so, leading to potential overloads in the electricity grid.

The utility company can predict this behavior based on historical data and typically enrolls more consumers in the demand response program. However, this is a risky strategy because, on some occasions, fewer consumers than expected might fail to reduce their usage, leading to insufficient load reduction during peak times and potential grid overload. The company wants to assess the probability that the grid will overload by enrolling 200 consumers, where each has an 85% probability of successfully reducing their usage as agreed.

i. Using binomial distribution model in R-programming, determine the probability that fewer than 170 consumers will reduce their electricity usage during a peak demand period. [4 marks]

ii. To determine the pattern of consumer behavior in the demand response program, the energy utility company wants a machine learning model that can predict whether a consumer will successfully reduce their electricity usage during peak periods. The predictions can help the company better manage the enrollment process and avoid grid overloads.

   a. With reasons, what machine learning category does this problem fall? [2 marks]

b. Discuss the steps that will be used to develop the machine learning model for predicting consumer energy consumption behavior. [5 marks]
c. Using the steps in (ii), write a code segment in R-programming that can be used to predict the consumer energy consumption behavior. [4 marks]

b) Data science plays a crucial role in many fields like energy transitions. Discuss contextual the relevance of data science concepts to an expert in energy transitions. [5 marks]

Question TWO (15 Marks)

a. An Electricity utility company wants to classify new residential consumers into one of three categories based on their expected electricity consumption patterns: Low, Medium, or High. The company has historical data on previous consumers, including various features such as the size of the household, income level, average monthly electricity usage, and whether they have energy-efficient appliances. The company wants to use the K-Nearest Neighbors (KNN) algorithm to classify new consumers based on these features.

Assume you are given the dataset called **energy_consumption.csv**, which contains the following columns: household_size, income_level, avg_monthly_usage, energy_efficient_appliances, and consumption_category, detail the tasks you will perform to classify the new consumers using the KNN algorithm. [7 marks]

a) Suppose a dataset consisting energy consumption in a neighbourhood of Nairobi were recorded as follows 567,1823, 517, 583, 317, 367, 250, 503, 317, 567, 583, 517, 650, 567, 450 and 350. Using r-software syntax give the code segments that can be used to;

i. Compute the mean energy consumption [2 marks]
ii. Generate a tabular presentation of the data set [2 marks]
iii. Identify the value that occurs most often [2 marks]
iv. Compute the standard deviation of the energy consumption [2 marks]

Question THREE (15 Marks)

a. Exploratory Data Analysis (EDA) encompasses various techniques designed to summarize the main characteristics of a dataset. These techniques can be broadly classified into several forms, each with specific objectives and methods.
i. Discuss aspects where exploratory data analysis techniques can be used by the energy utility company in Kenya [5 marks]
ii. Explain the various forms of exploratory data analysis. [5 marks]

b. You are working as a data analyst for an energy company that has provided you with a dataset energy_usage.csv. This dataset contains information on the electricity consumption of various households over the past year. The dataset includes the following columns: Household_ID, Region, Household_Size, Monthly_Consumption_kWh, Annual_Income, Energy_Efficient_Appliances, Peak_Hour_Usage, and Renewable_Energy_Usage.

Explain to the management how you will proceed to perform an Exploratory Data Analysis (EDA) on the provided dataset to uncover insights into household electricity consumption patterns. [5 marks]

Question Four (15 Marks)

a. Discuss the relevance of clustering techniques to an operations manager in a power utility company. [5 marks]
b. The sample dataset below relates to solar energy production. The goal is to segment these solar power plants into clusters based on similar characteristics. This can help in identifying groups of plants that might benefit from similar operational strategies, maintenance schedules, or investment in technology upgrades.

Plant_id	Average_Solar_Irradiance	Panel_Efficiency	Average_Daily_Output	Maintenance_Cost	Region
1	5.8	17%	450	3000	North
2	6.2	18%	480	3200	East
3	5.0	15%	400	2800	South
4	6.5	19%	510	3400	West
5	5.5	16%	420	2900	North
6	6.0	18%	460	3100	East
7	5.2	15%	405	2850	South
8	6.3	19%	500	3300	West
9	5.7	17%	435	2950	North
10	6.1	18%	475	3150	East

With explanation and code segment, demonstrate how this goal can be achieved using the K-Means clustering technique. [10 marks]