United International University
B.Sc. in Data Science

DS 1101: Fundamentals of Data Science

Final Exam: Summer 2024 Time: 2 Hours Marks: 50

Any evidence of plagiarism or copy will be punishable according to the proctorial rules of UIU

Answer all of the following questions.

1. You work at a tourism company that offers special discount cards to its customers. Last year, they called many customers and a fraction of the customers accepted the offer. Here is the data that was collected:

Serial No.	Job Type	Income Level	Likes to Hangout	Tours per year	Offer Taken
1	Engineer	High	Yes	2	Yes
2	Doctor	High	Yes	1	No
3	Engineer	Medium	No	3	Yes
4	Teacher	Medium	No	2	Yes
5	Doctor	High	Yes	3	Yes
6	Engineer	Medium	No	2	Yes
7	Teacher	High	Yes	1	No
8	Doctor	High	No	1	No
9	Teacher	High	No	2	Yes
10	Teacher	Medium	Yes	3	Yes
11	Engineer	High	No	1	No
12	Engineer	High	No	2	No

(a) Your task is to learn a Decision tree based on this data to predict whether a particular customer will take the offer. Which attribute should be the root of the decision tree? Show detailed calculations. [5]

(b) If any node in the tree has a majority vote of 80% or more, it will be converted into a decision node that aligns with the majority vote. Given this information, do you think a second level is needed? Explain why/why not. [3]

(c) A new high-earning engineer comes to your company to check out offers. You got to know that he likes to hangout but does not like to spend on more than 1 trip each year. Based on your decision tree, will this new customer accept the offer? [2]

2. (a) For each of the following scenarios, identify the task T, the experience E, and the performance measure P: [1x5=5]

i. A music app learns your preferences based on the songs you listen to and the ones you skip and recommends songs that match your taste.

ii. A voice assistant improves its ability to understand and respond to your commands as you interact with it more frequently.

iii. A self-driving car learns to navigate through heavy traffic based on its experiences driving in different conditions.

iv. A smart thermostat learns your temperature preferences throughout the day and automatically adjusts to optimize comfort.

v. A facial recognition system is trained on a large dataset of images, learning to match faces with high accuracy.

(b) Answer the following questions:

i. What are the key differences between Conventional Programming and Machine Learning? Draw diagrams to explain your answer. [2]

ii. Differentiate between Supervised, Semi-Supervised, and Unsupervised Learning. Your comparison should at least contain a definition, example cases, and popular algorithm names for each type of learning. [3]

3. (a) A global e-commerce company collects data from millions of users daily, including: [6]

- **Transaction data** from purchases
- **User reviews and feedback** in text, images, and videos
- **IoT sensors** monitoring warehouse inventory

The company aims to provide real-time product recommendations, manage inventory dynamically, and analyze customer feedback.

Based on this scenario, identify three potential challenges the company may encounter due to the nature of their data. For each challenge, recommend one appropriate technology that could be used to address it effectively with proper reasoning.

(b) A national weather service needs to process real-time sensor data from thousands of monitoring stations to provide immediate weather alerts for severe conditions (e.g., storms, floods).

Is Apache Spark or MapReduce more suitable for this task? Justify your choice based on factors such as processing speed and the ability to handle real-time data. [4]

4. You are given a dataset, *customer_satisfaction.csv*, that contains information about customers and their satisfaction scores based on several features. Your task is to clean, normalize, and standardize the dataset to prepare it for use in a machine learning model. The following contains the description of some of the columns:

   i **income**: Annual income of the customer in USD (float, contains missing values).

   ii **purchase_history**: Total number of items purchased in the last year (integer).

   iii **satisfaction_score**: Customer satisfaction score (target variable, integer, ranges from 1 to 10).

customer_id	age	income	purchase_history	satisfaction_score	signup_date	loyalty_tier
C001	25	55000	5	8	2015-06-12	Gold
C002	45	NaN	15	6	2016-04-22	Silver
C003	35	72000	10	NaN	2017-09-15	Bronze
C004	23	45000	NaN	7	2018-12-11	G0ld
C005	NaN	50000	8	9	2019-07-30	Bronz
C006	120	85000	20	10	2015-11-01	Silver
C007	42	62000	NaN	5	2016-08-14	Bronze
C008	32	-30000	7	8	2018-01-25	Silver
C009	NaN	NaN	NaN	NaN	2017-05-16	Silver
C010	27	58000	12	6	2016-02-20	Silver

Table 1: Customer Satisfaction Dataset

Now try to answer the questions below:

(a) Identify three potential techniques that will help to clean this dataset. [3]

(b) Correct the *loyalty_tier* column. [2]

(c) Handle all the visible missing values using pandas methods. [3]

(d) Find out the columns that contains outliers and correct them [2]

5. You are a data scientist working for a healthcare analytics company that is developing à model to predict patient health outcomes using various medical and lifestyle factors. The dataset provided includes several numerical features and one target variable indicating whether the patient has experienced a major health event.

Id	age	blood_prossure	cholesterol_level	BMI	exercise_hours	income	outcome
P001	45	135	210	28	5	50000	yes
P002	72	160	250	32	2	100000	yes
P003	30	120	180	24	10	75000	no
P004	65	145	220	29	3	25000	yes
P005	25	110	160	21	15	30000	no
P006	55	180	275	36	0	150000	yes
P007	40	140	200	27	8	120000	no
P008	60	170	265	31	4	90000	yes

Table 2: Sample of the patient health dataset

The dataset *patient_health_data.csv* contains some of the following features:

i. **blood_pressure**: The patient's systolic blood pressure in mmHg (values range from 90 to 200).

ii. **cholesterol_level**: The cholesterol level of the patient in mg/dL (values range from 100 to 300).

iii. **BMI**: The patient's BMI, calculated from weight and height (values range from 15 to 50).

iv. **exercise_hours**: The number of hours the patient exercises in a typical week (values range from 0 to 20).

v. **income**: The patient's annual income in USD (values range from $10,000 to $500,000).

vi. **outcome**: The **target variable** indicating whether the patient has experienced a major health event (yes or no).

Now try to answer the questions below:

(a) For Column *BMI & income*, decide whether **data normalization** or **data standardization** is more appropriate. Justify your choice based on the nature of the feature in the context of healthcare. [4]

(b) How would you detect and handle outliers in the features like **cholesterol_level** and **exercise_hours** . [4]

(c) Discuss **Z-Score** briefly. Explain how this technique will help you to detect outliers in your given context? [2]