(**Answer only Q1 (b), Q2 (a), Q4 (A, B, C & D), Q5 (a, b), Q6 (a)**)

**Q1**. What is *data mining*? In your answer, address the following:

      (b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?

Ans:

Data mining is largely classified as the conjunction between statistics and AI/ML. As enormous amount of data needs to be filtered and processed to find out a pattern among the data for future predictions general statistics can’t handle this, which is why Data mining had come into the picture.

In my opinion I don’t think it’s a simple transformation of technology, it is a well-defined process that yields better pattern recognition in the data. It rather is an integration of multiple disciplines like machine learning, pattern recognition, data visualization and many more.

**Q2**. For each data set given below, give specific examples of classification,

clustering, association rule mining and anomaly detection tasks that can be performed on the data. For each task, state how the data matrix should be constructed (i.e., specify the rows and columns of the matrix).

(a) Ambulatory Medical Care data, which contains the demographic and medical visit information for each patient (e.g., gender, age, duration of visit, physician's diagnosis, symptoms, medication, etc).

Ans:

**Classification:**

* Task: Predicting whether patient need diagnosis based on demographic and medical visit information.
* Data Matrix:
  + Rows: Each row representing the patient visit.
  + Columns: Attributes are Gender, Age, Symptoms, Medication, Duration of visit, Physician’s diagnosis and last column should be variable indicating patient need diagnosis or not.

**Clustering:**

* Task: Grouping patients on similar symptoms, physician’s diagnosis.
* Data Matrix:
  + Rows: Each row representing the patient visit.
  + Columns: Similar to classification task, Attributes are Gender, Age, Symptoms, Medication, Duration of visit, Physician’s diagnosis

**Association Rule Mining:**

* Task: Identifying association between different symptoms and physician’s diagnosis.
* Data Matrix:
  + Rows: Each row representing the patient visit.
  + Columns: Attributes are medical conditions, symptoms, medications, etc. Each column would be binary, indicating the presence or absence of the corresponding attribute.

**Anomaly Detection:**

* Task: Identifying unusual or rare patterns in patient visits that may indicate errors, fraud, or outliers.
  + Data Matrix: Rows: Each row representing the patient visit.
  + Columns: Attributes are Gender, Age, Symptoms, Medication, Duration of visit, Physician’s diagnosis and attribute indicating any unusual patterns or outliers identified during preprocessing, such as exceptionally long visit durations.

**Q4**. For each attribute given, classify its type as:

* discrete or continuous AND
* qualitative or quantitative AND
* nominal, ordinal, interval, or ratio

Indicate your reasoning if you think there may be some ambiguity in some cases.

**Example**: Age in years.

**Answer**: Discrete, quantitative, ratio.

(a) Average number of hours a user spent on the Internet in a week.

Ans: Continuous, qualitative, ratio

(b) GPA of a student.

Ans: Continuous, qualitative

(c) Credit card number.

Ans: Discrete, Nominal

(d) Salary above the median salary of all employees in an organization.

Ans: Continuous, qualitative, ratio

1. **Q5**. Null values in data records may refer to missing or inapplicable values. Consider the following table of employees for a hypothetical organization:

|  |  |  |
| --- | --- | --- |
| **Name** | **Sales commission** | **Occupation** |
| John | 5000 | Sales |
| Mary | 1000 | Sales |
| Bob | *null* | Non-sales |
| Lisa | *null* | Non-sales |

The null values in the table refer to inapplicable values since sales commission are calculated for sales employees only. Suppose we are interested to calculate the similarity between users based on their sales commission.

**(a)**Explain what is the limitation of the approach to compute similarity if we replace the null values in sales commission by 0.

Ans:

When we replace the null values in sales commission with 0 for computing the similarity we will have the following limitations:

* **Information loss**: Because the null values is considered as inapplicable values or missing data which doesn’t mean 0 in sales commission.
* **Incorrect Similarity**: As null values are considered as inapplicable values replacing them with 0’s effect the calculated similarity.
* **Bias towards Non-Sales people**: Because Non-sales will have zero sales anytime when compared to sales people, which create bias between sales and non-sales people as we are not taking into account the different job role.

(b) Explain what is the limitation of the approach to compute similarity if we replace the null values in sales commission by the average value of sales commission (i.e., 3000).

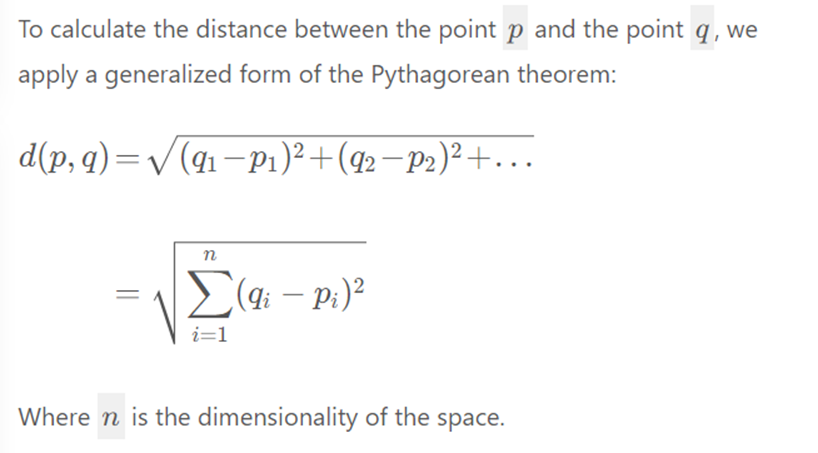
Ans:

When we replace the null values in sales commission with average value of sales commission (i.e., 3000) for computing the similarity we will have the following limitations:

* **Loss of Distinction**: By replacing null values with average values we will be losing the difference between employee who have made higher sales originally and who haven’t made.
* **Inflation of Similarity value**: By replacing the null values with average values will put people with null values making them similar to people who have worked originally. Which may result in incorrect similarity value with higher average then the original value, which opposite to the above case where we replaced null values with 0’s.
* **Bias towards sales person**: By replacing the null values with the average value of sales commission (i.e., 3000), we are ignoring the fact of different group of people that are sales and non-sales and only non-sales have null values in our data which puts sales people at disadvantage or biased has in fact non-sales have no sales commission to being with in first place.

**Q6**. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance*between the two objects.



Let’s assume p as (22, 1, 42, 10) and q as (20, 0, 36, 8) and by using the above formula we get the following has p= [p1=22,p2=1,p3=42,p4=10] and q= [q1=20,q2=0,q3=36,q4=8].

Using the above formula and substituting the values in it we get the following equation:

d(p,q)=Sqrt((20-22)2+(0-1)2+(36-42)2+(8-10)2)

= Sqrt((-2)2+(-1)2+(-6)2+(-2)2)

=Sqrt(4+1+36+4)

=Sqrt(45)

=6.708203932499369

