

Name: Nazmul Islam

ID: 232113

Ans. to the Q. no. 1

- a) Yes. This activity is a data mining task. It falls under the category of clustering.
- b) Yes. This activity is a data mining task. It falls under the category of classification.
- c) Yes. This activity is a data mining task. It falls under prediction task.
- d) No. This ~~act~~ activity is not a data mining task. Monitoring the heart rate of a patient for abnormalities is a real-time prediction or detection task. Usually data mining ~~work~~ involves with large datasets to find out pattern. So, This activity is not a data mining task.

Ans: to the Q.no. 2

a) Loan default prediction;

Goal: The goal of loan prediction is to identify potential customers who are likely to default on their loan payment.

Approach:

- 1) Analyze the historical loan data
- 2) Apply various classification algorithm on loan data
- 3) Build a model based on machine learning.
- 4) Consider the attributes of data: credit history, income, employment status, loan amount.
- 5) Preprocess the data with feature selection
- 6) Train the model with qualityful ~~data~~ and balanced data.

b) Credit card Fraud detection;

Goal: The goal of credit card fraud detection is to identify and prevent fraudulent transaction and protect both cardholders and financial institutions.

Approach:

- 1) Collect huge amount of credit card transactions.
- 2) ~~At preprocessing~~ We need to identify the anomaly to identify the unusual pattern or outliers in transaction data.
- 3) Train machine learning with important attributes data.
- 4) Transaction amount, merchant category, transaction date and time, location, cardholder information, transaction frequency these important attribute should be considered for this process.
- 5) Efficient classification algorithm should be chosen for this fraud detection.

Answer the Q. no. 3

The ~~the~~ Three data quality problem can be:

- 1) Noise
- 2) Duplicate data
- 3) Missing values

Noise: For objects noise is an extraneous object. Noise refers to modification of original values. It can misleading the data mining ~~task~~ task. ~~Train~~ Train with noise decrease the accuracy of the model. It also ~~create~~ ~~create~~ the raise the problem of overfitting, complexity, inaccurate decision making.

Duplicate data: A data set include data objects that are duplicates or almost duplicates, of one another. To detect and eliminate such duplicates, two main issues are addressed. ~~Similar data should be avoided and~~ Deduplication can help with this process.

Missing values: ~~Missing~~ Missing values also decrease the model performance. To handle this we can eliminate data object or estimate the missing values.

Ans. to the Q. no. 4

Temperature can be considered as either an interval or ratio, depending on the scale being used.

In Celsius ($^{\circ}\text{C}$) or Fahrenheit ($^{\circ}\text{F}$) temperature is measured on an interval scale.

On the other hand, in Kelvin (K) scale, temperature is measured on a ratio scale.

The distinction between interval and ratio scales lies in the presence or absence of a true zero point. In an interval scale, ratios between values are not meaningful since zero does not indicate the absence of the measured attributes. In a ratio scale, ratios have meaningful interpretation because zero represents the absence of the attributes.

That's why, ~~in~~ in Kelvin scale, temperature can be used as a ratio. On the other hand, in Celsius or Fahrenheit scales, temperature can be used as an interval.