CS 5990 (Advanced Data Mining) - Assignment #1

Bronco ID: |0|1|7|3|1|6|7 |5|4|

Last Name: ELAVAZHAGAN

First Name: RASHMI

1.

a. **Dividing the customers of a company according to their gender.**

   No, this activity does not involve discovering patterns or extracting information from data. It is a simple data organization or categorization task based on existing attributes.

b. **Monitoring seismic waves for earthquake activities.**

   Yes, this task involves analyzing and detecting patterns in seismic data to predict potential earthquake activities. It falls under data mining as it involves extracting valuable insights from large datasets to make predictions.

c. **Computing the total sales of a company**.

   No, this task involves aggregating numerical data (sales figures) to derive a single value (total sales). It is not data mining as it doesn't involve extracting patterns or making predictions from data.

d. **Predicting the outcomes of tossing a (fair) pair of dice.**

   No, predicting the outcomes of tossing dice relies on probability theory rather than data mining.

e. **Predicting the future stock price of a company using historical records.**

   Yes, this task falls under data mining. It involves analyzing historical stock price data to identify patterns and trends that can be used to make predictions about future stock prices. Techniques such as time series analysis and machine learning are often employed for this purpose.

f. **Monitoring the heart rate of a patient for abnormalities**.

Yes, Monitoring heart rate for abnormalities involves analyzing real-time physiological data to detect irregular patterns that may indicate health issues. While it involves analyzing data, it is more closely related to real-time monitoring and anomaly detection rather than traditional data mining, which typically deals with historical or batch data analysis**.**

2.

a) Brightness as measured by a light meter. **Continuous and Ratio**
b) Brightness as measured by people's judgments. **Discrete and ordinal**
c) Density of a substance in grams per cubic meter. **Continuous and Ratio**
d) Time of each day in the meaning of a 12-hour clock. **Continuous and Interval**
e) CPP bronco IDs. **Discrete and Nominal**

3.

Data Preprocessing:

- Dimensionality Reduction: In the data preprocessing phase, dimensionality reduction techniques are primarily applied. The main objective here is to reduce the complexity of the dataset by eliminating redundant features or reducing the number of dimensions while preserving important information. Techniques like Principal Component Analysis (PCA) or feature selection methods help improve computational efficiency and mitigate the curse of dimensionality.

Data Mining:

- Machine Learning Techniques: During the data mining phase, the primary focus is on applying machine learning algorithms to discover patterns, associations, classifications, and other insights from the data. Supervised learning algorithms (e.g., decision trees, support vector machines) and unsupervised learning algorithms (e.g., k-means clustering, association rule mining) are commonly used for this purpose. These algorithms are essential for extracting useful knowledge from large datasets.

Postprocessing:

- Visualization: In the postprocessing phase, the primary technique employed is visualization. Once the results of data mining algorithms are obtained, visualization techniques are crucial for presenting and communicating these results to stakeholders. Visualizations help convey insights, trends, and predictions derived from the data in an intuitive and interpretable manner, enabling informed decision-making and facilitating further analysis or action.

4.

a) **Association rule mining** - Its main objective is to identify the relationships and connections between the items in the dataset. The output is predicted by the data that we will provide as input.In the example input that is provided: Messi -> Football.
b)**Anomaly Detection** - Detect notable deviations from the usual pattern. Outliers are indicated in the above diagram, which deviates from the average of all the clusters.
c) **Classification** - Labeling or classifying instances according to their characteristics. Football is categorized as a sport, and voting as politics.
d) **Clustering**- Placing similar items in one group is known as clustering.
Elections, voter registration, ballots, and turnout are grouped into a single cluster vote. The football cluster is no different.

5.

a. What is the **most likely task** that data scientists are trying to accomplish?

   The most likely task is cheat class, specifically to predict whether an individual is likely to cheat on their taxes based on attributes like refund status, marital status, and taxable income.

b. **In general**, what is a feature and how would you **exemplify** it with **this data**?

   A feature is an individual measurable property or characteristic of a phenomenon being observed. In this dataset, examples of features include Refund, Marital Status, and Taxable Income. Each feature provides data that can be used to understand patterns or make predictions.

c. **In general**, what is a feature value and how would you **exemplify** it with **this data**?

   A feature value is the actual data or measurement associated with a feature for a particular instance. For example, for the feature Marital Status, a feature value could be "Single", "Married", or "Divorced". For Taxable Income, a feature value could be "125k", indicating the amount of income.

d. **In general**, what is dimensionality and how would you **exemplify** it with **this data**?

   Dimensionality refers to the number of features (attributes) that are present in the dataset. For this dataset, the dimensionality is 5, considering the features Tid, Refund, Marital Status, Taxable Income, and Cheat.

e. **In general**, what is an instance and how would you **exemplify** it with **this data**?

   An instance (or record) is a single, complete set of values for all the features that describe one observation. In this dataset, an instance could be the data for Tid 1: {Refund: Yes, Marital Status: Single, Taxable Income: 125k, Cheat: No}.

f. **In general**, what is a class and how would you **exemplify** it with **this data**?

   A class is the outcome or label that a classification task aims to predict based on

the features. In this dataset, Cheat is the class with possible values "Yes" or "No", indicating whether someone is likely to cheat on their taxes.

6. Ratio=Value in Field 2 Value in Field

   Calculating the ratios for each row:

   For Row 1:

   - Ratio = 233.8/33.4 ≈ 7

   For Row 2:

   - Ratio = 119.7/17.1≈ 7

   For Row 3:

   - Ratio = 168.0/24.0≈ 7

   We can see that the ratios are approximately the same for each row, indicating a consistent relationship between the values in Field 2 and Field 3 across the dataset.

7.

(a) x = (1 1 0 0 0), y = (0 0 0 1 1). Jaccard, Cosine, Euclidean, Correlation.

J(X,Y) = $\frac{a_{11}}{a_{11}+b_{10}+c_{01}}$

= $\frac{0}{0+2+2}$=0

Cosine Similarity = $\frac{x \cdot y}{\|x\| \cdot \|y\|}$

$x \cdot y=(1)(0)+(1)(0)+(0)(0)+(0)(1)+(0)(1)=0$

$\| x \| =\sqrt{1^2 + 1^2 + 0^2 + 0^2 + 0^2}=\sqrt{2}$

$\| y \| =\sqrt{0^2 + 0^2 + 0^2 + 1^2 + 1^2}=\sqrt{2}$

Cosine Similarity = $\frac{0}{\sqrt{2} \cdot \sqrt{2}}$= 0

Euclidean Distance: $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

$$= \sqrt{(1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2}$$

$$= \sqrt{1 + 1 + 0 + 1 + 1} = \sqrt{4} = 2$$

$$\text{Correlation} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} - \sqrt{n\sum y^2 - (\sum y)^2}}$$

n=5

$$\sum xy = (0)(1)+(1)(0)+(0)(1)+(1)(0)+(1)(0)=0$$

$$\sum x = 2 \ , \sum y = 2, \sum x^2 = 2, \sum y^2 = 2$$

$$\text{Correlation} = \frac{(5\cdot 0) - (2\cdot 2)}{\sqrt{(5\cdot 2) - (2)^2} \cdot \sqrt{(5\cdot 2) - (2)^2}} = \frac{-4}{6} = \frac{-2}{3} = -0.66$$

(b) x = (0 1 0 1 1), y = (1 0 1 0 0). Jaccard, Cosine, Euclidean, Correlation.

$$J(X,Y) = \frac{a_{11}}{a_{11} + b_{10} + c_{01}}$$

$$= \frac{0}{0+3+2} = 0$$

$$\text{Cosine Similarity} = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

$$x \cdot y = (0)(1)+(1)(0)+(0)(1)+(1)(0)+(1)(0)=0$$

$$\| x \| = \sqrt{0^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{3}$$

$$\| y \| = \sqrt{1^2 + 0^2 + 1^2 + 0^2 + 0^2} = \sqrt{2}$$

$$\text{Cosine Similarity} = \frac{0}{\sqrt{3} \cdot \sqrt{2}} = 0$$

$$\text{Euclidean Distance: } \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$= \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2}$$

$$= \sqrt{5} = 2.23$$

$$\text{Correlation} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} - \sqrt{n\sum y^2 - (\sum y)^2}}$$

n=5

$$\sum xy = (0)(1)+(1)(0)+(0)(1)+(1)(0)+(1)(0)=0$$

$$\sum x = 3, \sum y = 2, \sum x^2 = 3, \sum y^2 = 2$$

$$\text{Correlation} = \frac{(5\cdot 0) - (3\cdot 2)}{\sqrt{(5\cdot 3)-(3)^2} \cdot \sqrt{(5\cdot 2)-(2)^2}} = \frac{-6}{6} = -1$$

8.

a)

To find when the vectors are perpendicular, their dot product should be zero.

The dot product of two vectors u and v is given by:

$u \cdot v = u1 \cdot v1 + u2 \cdot v2$

Given u=(2,k) and v=(3,−2), their dot product is: $u \cdot v = (2)(3)+(k)(-2)=6-2k$

$u \cdot v = 0$.

$6-2k=0$

$2k=6$

$\boxed{k=3}$

b)

To find when the vectors are parallel, one must be a scalar multiple of the other.

If u is a scalar multiple of v, then u=cv, where c is a scalar.

$u = cv$

$(2, k) = c(3,-2)$

$(2, k) = 3c,-2c$

$2=3c$

$c= 2/3$

$k = -2c$

$c = 2/3$

$k = -2(2/3)$

$k = -4/3$

9.

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	$10
2	Shoes	0	09/10/22	$15
3	TV	1	09/09/22	$20

a)

ID	Item	Downtown	Date	Price
1	Electronics	2	09/09/22	$30
2	Shoes	0	09/10/22	$15

b)

Item	Date	Price
Watch	09/09/22	$10
Shoes	09/10/22	$15
TV	09/09/22	$20

c)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	0.2
2	Shoes	0	09/10/22	0.5
3	TV	1	09/09/22	0.04

d)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	$10
2	Shoes	0	09/10/22	$15

a) Aggregation.

The processed dataset seems to aggregate data based on the date, possibly summing up prices for items sold on the same date. This is evident from the fact that the original dataset has multiple entries for the same date, but the processed dataset has only one

entry per date with the total price.

b) Feature Selection

The processed dataset appears to select specific features or columns from the original dataset. In this case, only the "Item", "Date", and "Price" columns are retained, while the "ID" and "Downtown" columns are removed. Feature selection involves selecting a subset of relevant features from the original dataset for further analysis.

c) Normalization

The processed dataset seems to normalize the price values, converting them into a normalized scale between 0 and 1. This is evident from the fact that the original price values have been replaced with normalized values.

d) Sampling

Sampling involves selecting a subset of data from a larger dataset. If the processed dataset contains only a subset of the original raw dataset's entries, then this is an example of **sampling**.

10.

https://github.com/RashmiElavazhagan/Smilarity.py/blob/main/Similarity.py