Group 2 (DA-02)Theory Assignment

Concept questions

1. C, D, K, I

2.
   a. Data analytics techniques could be used to set credit limits for customers and predict default rates.
   b. Classifying customer dissatisfaction to locate areas with problems.
   c. Measuring and predicting traffic and weather conditions.
   d. Finding areas of interest of users for showing better advertisements

3.
   a. Yes. Dividing the customers of a company according to their gender is a form of classification of data. Therefore, it is a data analytics task.
   b. Yes. Dividing the customers of a company according to their profitability is also a form of classification of data. Therefore, it is a data analytics task.
   c. Yes. Computing the total sales of a company makes the sales data more understandable by making it smaller. Therefore, it is a data analytics task.
   d. No. Sorting a student database based on student identification number is not making data smaller or more understandable. Therefore, it is not a data analytics task.
   e. No. Predicting the outcomes of tossing a (fair) pair of dice does not involve analysing any data. Therefore, it is not a data analytics task.
   f. Yes. Predicting the future stock price of a company using historical records involves analysing past data and predicting which is a data analytics task.
   g. Yes. Monitoring the heart rates of patients for abnormalities involves analysing the heart rate data of patients which is a data analytics task.
   h. Yes. Monitoring seismic waves for earthquake activities consists of analysing seismic data and predicting future earthquakes which is a data analytics task.
   i. No. Extracting the frequencies of a sound wave is a scientific experiment and does not involve analysing data rather collection of data. Therefore, it is not a data analytics task.

4. Data analytics can help the Internet search engine company in many ways particularly using methods of search analytics. Some of them are as follows:
   a. Clustering techniques could be used to group a particular type of user based on the search patterns.
   b. Regression techniques could be used to predict search requests and optimise the results based user preference.
   c. Classification could be used to analyse the various groups of users to deliver advertisements customised to the users.
   d. Anomaly detection could be used to find users who have unusual amounts of requests and find programs using the search engine.

e. Association rule mining can be used to find relationships between search patterns and users’ online purchases for better ad placements.

5.

a. This is not a privacy issue since the data collected does not reveal personal details of the people involved.
b. This is an intrusion of privacy since the users’ personal information is being used.
c. This is not a data privacy issue if the images do not reveal personal information or any sensitive information.
d. This is not a data privacy issue since the data used is divulged by the people.
e. This may be a data privacy issue since the data used may or may not be used with the consent of the people involved.
f. This is not a data privacy issue if the messages posted are viewed by the intended audience.

6.

a. Measures of central tendencies (Mean, Median, Mode)
b. Standard deviation
c. Variance
d. Skewness
e. Kurtosis
f. Frequency distribution

7.

a. Sheer size of data: Considering the rate at which data is being generated by in the current fields and the rate at which it is increasing, processing algorithms would have to improve significantly to keep up.
b. Data quality: As size of data increases, accuracy of the information decreases. To address this issue, data procurement will have to be overtaken to ensure clean data.
c. Outliers: Outliers which could safely be ignored in small data can no longer be ignored in big data. They cause an impact in the analysis and understanding of the data which would otherwise have been ignored.
d. Privacy: When personal information is used in big data, it can infer conclusions about people that were unintended to be made public.
e. Data visualization: Most visualisation techniques like graphs and charts can only show a simplistic view of data. This view does not give the complete information in case of big data as the source is much more complex information.

8.

a. Atmospheric data and weather reports
b. Default rates of customers
c. Character samples of various fonts and scripts

d. Basic information about a person that differentiates the person such as name, ethnic group, gender, nationality, vernaculars, occupation, region

e. Information about outliers and deviations from the consensus

9. The 3 V values of streaming data in case of a cricket broadcast are:

a. Volume: The processing station receives massive amounts of footage from live feed from several cameras throughout the game along with all the numbers of the game.

b. Variety: The variety of information received such as video, text and numbers which include feed from the various cameras in the event, the scores of the game and the statistics associated with the various players and teams.

c. Velocity: All the data being received must be processed into understandable formats and broadcast to the public. In live events the speed of processing this information is much higher.

Objective questions

1. B
2. D
3. D
4. B
5. A
6. D
7. B, D
8. B
9. B