MACHINE LEARNING

Q1. a)

Q2. d)

Q3. a)

Q4. a)

Q5. b)

Q6. d)

Q7. d)

Q8. b)

Q9. d)

Q10. a)

Q11. d)

Q12. a)

Q13. Clustering is an unsupervised approach to find out similarities and dissimilarities between different objects and then make a group of similar objects.

Q14. Cluster quality is measured through similar instance and keeping them together.

Q15. Cluster analysis is to classify the data into structures that are more easily to understood and manipulated or modified. Hierarchical Clustering and Kmeans Clustering are used most as a type of cluster analysis

WORKSHEET 1 SQL

Q1. A) , C) & D)

Q2. A) & B)

Q3. B)

Q4. B)

Q5.A)

Q6. C)

Q7. B)

Q8. B)

Q9. B)

Q10. A)

Q11. Data warehouse is a database that store information oriented to satisfy decision making request.

Q12. Online Analytical Processing (OLAP) is a category of software tools that analyze data stored in a database, whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture. OLAP is characterized by a large volume of data, while OLTP is characterized by large numbers of short online transactions.

Q13. Characteristics of Data Warehouse: It is a Subject Oriented, Time variant, Non-Volatile, integrated

Q14. Star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables.

Q15. It is a high level programming language of mathematical theory of sets.

STATISTICS WORKSHEET-1

Q1.A)

Q2.A)

Q3.B)

Q4.D)

Q5.C)

Q6.B)

Q7.B)

Q8.A)

Q9.C)

Q10. Normal Distribution means the data is symmetrically distributed i.e. the variables appears at regular frequency with no skewness, the mean, mode and median appear at distinct points. Height, birth, weight, job description is few example of such variables.

Q11. Missing value can be handle by deleting,by mean and mode, by fillna option.Missing value can be managed by Linear Regression Algorithm

Q12.A/B testing, also known as split testing, is a marketing experiment wherein you split your audience to test a number of variations of a campaign and determine which performs better. Like you can show version A of a piece of marketing content to one half of audience, and B to another.

Q13. Mean imputation is typically unethical practice because it removes the feature correlation. Suppose in a table with age and fitness, if 9 year old has a missing fitness score. And if we average the fitness score of 15 to 80 years, no doubt we will get the average fitness score but significantly get highest fitness level than he actually has.

Another one reason is mean imputation decreases variance of data while increase bias. So the consequence of reduced variance, the model is less accurate and the confidence interval is narrower.

Q14.

Linear Regression is one of the most fundamental algorithm in Machine Learning. It is the process of predicting a label based in features at hand. It is used for time series modeling an finding the casual effect relationship between variables and forecasting. Ex. Relationship between the stock prices of the company and various factors like customer reputation & company annual performance etc. can be studied using linear regression.

Q15. There are two types of statistics:

1.Descriptive statistics : It summarizes (and organize) characteristics of a data set. A data set is a collection of observation (or response) from a sample or entire population and

2.Inferential Statistics: It helps to decide whether your data confirms or refuse your hypothesis and whether it is generalize to a large population