ML worksheet Answers

1. b) 4
2. d) 1, 2 and 4
3. d)
4. a)
5. b)
6. d)
7. d)
8. b)
9. a)
10. a)
11. d)
12. a)
13. Cluster Analysis involves decision on number of clusters to be created, calculating the distances between centroid and data points. These centroid keeps on moving to the Average distance point to reach at the centre of the data to create best cluster which will further create best model for prediction.
14. silhouette coefficient value is used for measuring the quality of clusters
15. Cluster analysis is used in unsupervised learning where the data does not has its output or y label. It involves creation of cluster of data where the centroid is been created where most of the data is present and on the basis of which data analysis is done.

SQL worksheet Answers

1. A, c & d
2. A, b & d
3. B
4. B
5. A
6. C
7. B
8. B
9. D
10. C
11. Data warehouse is a storage of various data which are collected from varied sources. It is stored in an electronic storage from where it could get extracted and make it useful for business analysis which leads to business growth in long run. It's a process for collecting and managing the large amount of data.

12. 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.
13. The key characteristics of a data warehouse are as follows:

- Some data is denormalized for simplification and to improve performance
- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common

- The data load is controlled

In general, fast query performance with high data throughput is the key to a successful data warehouse.

14. Star schema is the basic and widely used schema which is used to create data warehouse. It includes one or more fact tables indexing any number of dimensional tables
15. The SET command is used with UPDATE to specify which columns and values that should be updated in a table in a SQL.

Statistics worksheet Answers

1. a)
2. a)
3. c)
4. d)
5. c)
6. b)
7. b)
8. a)
9. c)
10. Normal distribution is the term used when the data in the statistics is distributed around the mean of the data, they are not much deviated from the mean or standard deviation of the data. In this kind of distribution mean, median and mode all are almost equal. Half of the data is at the left of the mean and half at the left which mean its equally distributed. There are not much outliers in this kind of distribution. The curve is bell shaped.
11. Missing data could be handled by replacing the same to the mean, median or mode depending on the data type. This method should be used when there is minimal missing values in the data as it does not give the adequate result for better model building. Simple imputer method is best while filling the NaN values and giving accurate result for model building.
12. A/B testing is the type of hypothesis testing which involves comparison of two variants of single variable. It determines which variant is more effective and giving accurate output while model building or deciding the y variable in the statistics.
13. It is a non-standard practise but it can be used when we have lesser missing values in the data as it will not have much impact on model building.
14. Linear regression is a type of predictive analysis where we check the correlation between two variables or between the outcome and the variables of the data. Here, we check how the x variables are useful in predicting the y variables, how strong the relation between them. It can also be determined using mathematical formula : formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.
15. There are basically 02 types of statistics which are Descriptive and Inferential. Descriptive statistics further have 02 branches viz. Central tendency and Data dispersion. Central tendency includes mean, median and mode. Data dispersion includes range, variance, Standard deviation, percentile and skew. Inferential statistics includes z score, hypothesis testing, T test, F test, Anova test, chisquare test etc.