MACHINE LEARNING

1. D
2. B
3. A
4. A
5. B
6. A & D
7. B&C
8. A &B
9. **Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?**
   The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.
10. **Differentiate between Ridge and Lasso Regression.**
    The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.
11. **What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?**
    Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.
    A rule of thumb commonly used in practice is if a **VIF** is > 10, you have high multicollinearity
12. **Why do we need to scale the data before feeding it to the train the model?**
    Feature scaling is essential for machine learning algorithms that calculate distances between data. ... Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
13. **What are the different metrics which are used to check the goodness of fit in linear regression?**
    Five metrics give us some hints about the goodness-of-fit of our model. The first two metrics, the Mean Absolute Error and the Root Mean Squared Error (also called Standard Error of the Regression), have the same unit as the original data.
14. **From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy**
    Sensitivity=0.45
    Specificity= 0.96

Precision=0.95
Recall=0.45
Accuracy=0.88

SQL

1. A,C,D
2. A,C,D
3. B
4. C
5. B
6. B
7. A
8. C
9. A
10. D
11. **What is denormalization?**
    Denormalization is a database optimization technique in which we add redundant data to one or more tables. ... For example, in a normalized database, we might have a Courses table and a Teachers table. Each entry in Courses would store the teacherID for a Course but not the teacherName.
12. **What is a database cursor?**
    A database cursor is an identifier associated with a group of rows. It is, in a sense, a pointer to the current row in a buffer. You must use a cursor in the following cases: Statements that return more than one row of data from the database server: A SELECT statement requires a select cursor.
13. **What are the different types of the queries?**
    Five types of SQL queries are 1) Data Definition Language (DDL) 2) Data Manipulation Language (DML) 3) Data Control Language(DCL) 4) Transaction Control Language(TCL) and, 5) Data Query Language (DQL)
14. **Define constraint?**
    SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table.
15. **What is auto increment?**
    Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table. Often this is the primary key field that we would like to be created automatically every time a new record is inserted.

STATISTICS

1. D
2. A
3. A
4. C
5. A
6. A
7. C
8. B
9. **What is the difference between a boxplot and histogram?**
   Histograms and box plots are graphical representations for the frequency of numeric data values. ... Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.
10. **How to select metrics?**

    Good metrics are important to your company growth and objectives. Your key metrics should always be closely tied to your primary objective.
    Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement.
    Good metrics inspire action.
11. **How do you assess the statistical significance of an insight?**

Calculating the statistical significance is rather extensive if you calculate it by hand and this is why it's typically calculated using a calculator. When you calculate it by hand, however, it will help you more fully understand the concept. Here are the steps for calculating statistical significance:

Create a null hypothesis.
Create an alternative hypothesis.
Determine the significance level.
Decide on the type of test you'll use.
Perform a power analysis to find out your sample size.
Calculate the standard deviation.
Use the standard error formula.
Determine the t-score.
Find the degrees of freedom.
Use a t-table.

12. **Give examples of data that doesnot have a Gaussian distribution, nor log-normal**

There are many data types that follow a non-normal distribution by nature. Examples include: Weibull distribution, found with life data such as survival times of a product. Log-normal distribution, found with length data such as heights.

13. **Give an example where the median is a better measure than the mean.**
    In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency. When you have ordinal data, the median or mode is usually the best choice.
14. **What is the Likelihood?**
    In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. But in both frequentist and Bayesian statistics, the likelihood function plays a fundamental role.