**Statistics**

1. **D**
2. **A**
3. **A**
4. **C**
5. **C**
6. **A**
7. **C**
8. **B**
9. **b**
10. Box plot displays 25th , 50 and 75th percentile in the figure along with the fence. Histogram is used to check the frequency distribution of numerical data.
11. Selection of metrics depends upon the use case under consideration and the requirement of the use case. Commonly used metrices for continuous data is mean squared error , root mean squared error etc and metrices for discrete data are confusion\_matrix , accuracy , f1 score etc.

Any custom metric could also be used for evaluation of data.

1. Statistical significance can be accessed using hypothesis testing:  
   Stating a null hypothesis which is usually the opposite of what we wish to test and then using various hypothesis test to reject or fail to reject the null hypothsis.
2. Examples of non – normal and non -lognormal data – data of number of accidents took place per day since 2020 , amount of rainfall received in a region given a time period.
3. Median is considered better than mean in case of skewed data. The greater the skewness the larger will be the difference between mean and median. Hence any skewed distribution is an example of such case is data of per day covid-19 cases in India since last 60 days.
4. Likelihood is the chance of any event to happen i.e probability.

**SQL**

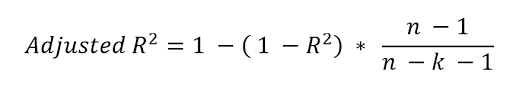
1. **A,C,D**
2. **A,C,D**
3. **B**
4. **C**
5. **D**
6. **B**
7. **A**
8. **C**
9. **D**
10. **A**
11. Denormalization in database is an optimization technique where redundant data is added to database tables so that the records can easily be fetched without writing complex joins.

It also improves the time duration of the records being fetched.

1. Database cursor is a identifier to the group of rows specifically it is a pointer to a row inside the row buffer.
2. There are five types of queries in database : DDL , DML , DCL , TCL , DQL
3. Constraint is nothing but rule specifier for the data in the table. They are used to limit the type of data that can go in table. This ensures the accuracy and reliability of the data in the table.
4. Auto increment is a unique number which is generated automatically upon the insertion of new record in a table. This is often the primary key which needs to be created on insertion of every new record.

**Machine Learning**

1. **C**
2. **B**
3. **C**
4. **B**
5. **B**
6. **A,C**
7. **B,C**
8. **A,C**
9. **A,B**
10. Formula for R squared is given as :

****

Here , n = the total sample size

k = the number of independent features (not correlated with target)

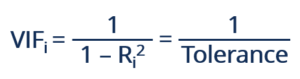
Now for a given problem if k increases than n-k-1 will decrease and hence (n-1)/(n-k-1) will increase . Again it gets multiplied by a constant (1-R2) and values gets more increased and finally when subtracted from 1 will result in a smaller value than the one which have been received if k=0.

1. Ridge regression aims at penalizing the steep sloped features by adding a value  λ\*(slope)2 to solve overfitting problem.

Here lamda is learning rate and m is the slope of regression line. At each iteration lambda \* slope2 is chosen such that the new slope is less than the previous slope m.

Lasso regression aims at removing the features of low slope with a small modification in the value λ\*(slope)2. Instead of choosing slope2 if takes only magnitude of slope and hence while penalizing the slope the features which already had very less slope gets removed.

1. The Variance Inflation factor measures the severity of multicollinearity in the regression analysis. In other words it measures the extent of multicollinearity in the data.



R2 is the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

So if VIF = 0 for a variable ie tolerance =0 (r2=1) means that the variable is not correlated with the others.

A VIF of 4 or above indicates multicollinearity might exist. So while including a feature it is good to have VIF <=4.

1. Having features on a similar scale can help the gradient descent converge more quickly towards the minima by smoothening the transition towards the global minima.
2. While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit. Other than this MAE ,RMSE can also be used to check the goodness of fit.

**Actual**

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 1000 | 50 |
| 0 | 250 | 1200 |

**Predicted**

Precision= TP/TP+FP=1000/1050 = 0.95

Sensitivity= TP/TP+FN = 1000/1250 =0.8

Accuracy= TP+TN/TP+TN+FN+FP = 2200/2700 = 0.81

Specificity= TN/TN+FP =0.96