Statistics Worksheet – 6

1. d) All of the mentioned
2. a) discrete
3. a) pdf
4. c) Mean
5. a) variance
6. b) standard deviation
7. c) 0 and 1
8. b) bootstrap
9. b) summarized

10. Histogram plot takes more space to represent a big dataset whereas boxplot requires less space to represent the big dataset. And histogram indicates the whole frequency distribution of a variable whereas boxplot summarizes its prominent features.

    In boxplot you cannot see the shape of distribution but in histogram we can clearly understand the shape of distribution, the variation & potential outliers.

11. We can select the metrics depending on what kind of problem we are trying to solve.

12. To assess the statistical significance, we have to do hypothesis testing. In that both null & alternate hypothesis would be stated first. Then secondly, we calculate the p-value which is the likelihood of getting the test's observed findings, if the null hypothesis is true. And now after that finally we select the threshold of the significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha. By this we can assess the statistical significance of an insight.

13. Any type of categorical data won't have the Gaussian distribution or log-normal distribution.

    Eg . The amount of time that a car battery lasts.

14. Examples where the median is better than the mean are:-
    a) Managing the home budget.
    b) In business.
    c) Buying a property.

15. Likelihood is the probability that a particular outcome is observed when the true value of the parameter is equivalent to the probability mass on.

SQL Worksheet -6

1. A), C), D)
2. A), C), D)
3. B) SELECT NAME FROM SALES
4. B) Insert, delete & update the records and values
5. B) Column alias
6. B) Commit
7. A) Parenthesis ()
8. C) TABLE
9. D) All of the mentioned.
10. A) ASC

11. Denormalization is an optimization technique in which we add the redundant data to one or more tables to improve read performance of the database.

12. Database cursor is the identifier associated with a group of rows. It is basically the sense or a pointer to the current row in the buffer.

13. Different types of queries are:-

a) SELECT query
b) Parameter query
c) Action query
d) Aggregate query

14. Constraints in SQL are used to specify rules for the data in the table. It limits the type of data that can go into the table & this ensures the accuracy & reliability of the data in the table.

15. Auto increment allows unique number to be generated automatically when a new record is inserted into a table.

Machine Learning Worksheet -6

1. B) Low R-squared value for train-set & high R-squared value for test – set.
2. C) Decision trees are not easy to interpret.
3. D) Decision tree.
4. A) Accuracy
5. B) Model B
6. A) and D) Ridge & Lasso
7. B) and C) Decision Tree and Random Forest
8. A) and C) Pruning and restricting the max depth of the tree.
9. A) and B)

10. The adjusted R-square increases only if the new predicator enhances the model above what would be obtained by probability and vice versa it would decreases when the predicator improves the model less than what is predicated by chance.

11. Differentiate between Ridge Regression and Lasso Regression :-

a) Ridge regression uses L2 regularization technique while Lasso uses L1 regularization technique.

b) In ridge regression the penalty is equal to the sum of the squares of the coefficients while in Lasso penalty is considered to be sum of the absolute values of the coefficients.

12. VIF is **Variance Inflation Factor** that measures the severity of multicollinearity in regression analysis. It is statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.
    The suitable value of VIF is 5; only variables with a VIF less than 5 are included in the model. However many sources says that VIF less than 10 are also acceptable.

13. Feature scaling is the important step of data pre-processing in machine learning. Scaling of data makes it easy for a model to learn and understand the problem.

14. Different metrics which are used to check the goodness of fit in linear regression are :-
    a) R-squared/Adjusted R-squared.
    b) Mean Square Error (MSE).
    c) Mean Absolute Error (MAE).
    d) Root Mean Square Error (RMSE).
15. Sensitivity/Recall – 1000/(1000+250) = 0.8
    Specificity – 1200/ (1200+50) = 0.96
    Precision – 1000/ (1000+50) = 0.95
    Accuracy – (1000+1200)/ (1000+1200+50+250) = 0.88