ASSIGNMENT-4

MACHINE LEARNING

1. C
2. C
3. B
4. C
5. B
6. B
7. C
8. A,C
9. A,C,D
10. A,B,D
11. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers
12. Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification.
13. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

14. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
15. Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

    The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.

    Process parameters and controls are determined during the validation of any process or system.

SQL –

1. Select shippedDate from orders where av(count(shippedDate));
2. Select * from orders where avg(count(orderDate));
3. Select productName form products where min(MSRP);
4. Select productName fron products where max(quantityInStock);
5. Select max(count(productName)) from products;
6. Select customers.customerNumber,payments.customerNumber from customers,payments where max(amount) ;
7. Select customerNumber,customerName from customer where city='melbourne';
8. Select customerName from customers where name like'N%';
9. Select customerName from customer where phone like '7%' and city = 'las vegas';
10. Select customerName from customers where creditlimit <1000 and city = 'las vegas', ,'Nantes','staven';
11. Select orderNumber from orderdetail where quantityordered < 10;
12. Select orders.customerNumber,customers.customerNumber from orders,customers where customersName like 'N%';
13. NAN
14. Select customers.customerNumber ,payments.customerNumber from customers,payments where checkNumber like 'M%' and paymentDate ="2004-10-19";
15. Select checkNumber from payments where amount > 1000;

Statistics-

1. The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. it allows us to safely assume that the sampling distribution of the mean will be normal in most cases
2. Sampling means selecting the group that you will actually collect data from in your research. There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.
3. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.
4. A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.
5. Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.
6. Univariate analysis looks at one variable, Bivariate analysis looks at two variables and their relationship. Multivariate analysis looks at more than two variables and their relationship.

7. Sensitivity analysis is used to identify how much variations in the input values for a given variable impact the results for a mathematical model. Sensitivity analysis can identify the best data to be collected for analyses to evaluate a project's return on investment (ROI). The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.
8. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.
9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
10. The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
11. A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified

group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side

12. Data visualization method.
13. The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis
14. Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$.
15. Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.
16.

17.