# Machine Learning Assignment-4

1. C
2. B
3. C
4. A
5. C
6. B
7. C
8. D
9. D
10. A,C,D
11. Outliers are anomalies which are extreme values falling outside 3 standard deviations in a normal distribution. Interquartile Range is difference between the third quartile (Q3) and the first quartile (Q1) = Q3-Q1
12. Bagging decreases variance, not bias while Boosting decreases bias, not variance.
13. The formula for adjusted R2 score is equal to 1−[(1−R2)(n−1)n−k−1] where R2 is r2 score while n is number of points in data sample and k is the number of independent variables in the model.
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 resampling procedure used to evaluate machine learning models on a limited data sample. The main advantage of using cross-validation is to detect overfitting.

# SQL Worksheet-4

1. SELECT AVG(ORDERNUMBER) FROM ORDERS GROUP BY SHIPPEDDATE;
2. SELECT AVG(ORDERNUMBER) FROM ORDERS GROUP BY ORDERDATE;
3. SELECT MIN(MSRP) FROM PRODUCTS;
4. SELECT MAX(QUANTITYINSTOCK) FROM PRODUCTS GROUP BY PRODUCTNAME;
5. SELECT P.PRODUCTNAME, SUM(S.QUANTITYORDERED) FROM PRODUCTS P, ORDERDETAILS S WHERE S.PRODUCTCODE=P.PRODUCTCODE GROUP BY S.QUANTITYORDERED,P.PRODUCTNAME;
6. SELECT C.CUSTOMERNAME, MAX(P.AMOUNT) FROM CUSTOMERS C, PAYMENTS P WHERE C.CUSTOMERNUMBER=P.CUSTOMERNUMBER GROUP BY P.AMOUNT,C.CUSTOMERNAME;
7. SELECT CUSTOMERNUMBER, CUSTOMERNAME FROM CUSTOMERS WHERE CITY LIKE ‘MELBOURNE’;
8. SELECT CUSTOMERNAME FROM CUSTOMERS WHERE CUSTOMERNAME LIKE ‘N%’;
9. SELECT CUSTOMERNAME FROM CUSTOMERS WHERE TO\_CHAR(PHONE) LIKE TO\_CHAR(7%) AND UPPER(CITY) LIKE ‘LAS VEGAS’;
10. SELECT CUSTOMERNAME FROM CUSTOMERS WHERE CREDITLIMIT<1000 AND UPPER(CITY) IN (‘LAS VEGAS’,’NANTES’,’STAVERN’);
11. SELECT ORDERNUMBER FROM ORDERDETAILS WHERE QUANITYORDERED < 10;
12. SELECT O.ORDERNUMBER FROM ORDERS O, CUSTOMERS C WHERE C.CUSTOMERNUMBER=O.CUSTOMERNUMBER AND C.CUSTOMERNAME LIKE ‘N%’;
13. SELECT C.CUSTOMERNAME FROM ORDERS O, CUSTOMERS C WHERE C.CUSTOMERNUMBER=O.CUSTOMERNUMBER AND O.STATUS LIKE ‘DISPUTE’;
14. SELECT C.CUSTOMERNAME FROM CUSTOMERS C, PAYMENTS P WHERE C.CUSTOMERNUMBER=P.CUSTOMERNUMBER AND P.CHECKNUMBER LIKE ‘H%’ AND TO\_CHAR(P.PAYMENTDATE,’YYYY-MM-DD’) LIKE ‘2004-10-19’;
15. SELECT CHECKNUMBER FROM PAYMENTS WHERE AMOUNT>1000;

# Statistics Worksheet-4

1. The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger. The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases.
2. Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.
3. Type II error is the error that occurs when the null hypothesis is accepted when it is not true. Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.
4. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
5. Covariance is a measure to indicate the extent to which two random variables change. Correlation is a measure used to represent how strongly two random variables are related to each other.
6. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.
7. Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall.

Sensitivity = (True Positive)/(True Positive + False Negative)

1. Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You’re basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won’t be repeatable and so has little use. Null Hypothesis(H0): H0 always assume there is no significant effect/difference within the specified population. Hypothesis testing is a statistical test based on two hypothesis: the null hypothesis(H0), and the alternative hypothesis(H1).
2. Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language.
3. The difference of third quartile(Q3) and first quartile(Q1) is known as interquartile range(IQR).
4. The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean.
5. Z-score is one method to find outliers. To visualize the outliers we can also use box-plots.
6. The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested. P is also described in terms of rejecting H0 when it is actually true, however, it is not a direct probability of this state.
7. 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).
8. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.