WORKSHEET – 4

WORKSHEET = SQL

1. Write a SQL query to show average number of orders shipped in a day (use Orders table).
ANS. SELECT AVG(orderNumber) from orders group by shippedDate;

2. Write a SQL query to show average number of orders placed in a day.
ANS.

3. Write a SQL query to show the product name with minimum MSRP (use Products table).
ANS. SELECT productName, MSRP FROM products WHERE MSRP = (SELECT MIN(MSRP) FROM products);

4. Write a SQL query to show the product name with maximum value of stockQuantity.
ANS. SELECT productname,quantityinstock FROM products ORDER BY QuantityInStock desc limit 1;

5. Write a query to show the most ordered product Name (the product with maximum number of orders).
ANS. SELECT productName FROM products WHERE productCode = (SELECT productCode
FROM orderdetails ORDER BY quantityOrdered DESC LIMIT 1);

6. Write a SQL query to show the highest paying customer Name.
ANS. SELECT customerName FROM customers WHERE customerNumber =
(SELECT customerNumber FROM payments ODER BY amount DESC LIMIT 1);

7. Write a SQL query to show cutomerNumber, customerName of all the customers who are from Melbourne city.
ANS. SELECT customerNumber,customerName FROM customers WHERE city IN ('Melbourne city');

8. Write a SQL query to show name of all the customers whose name start with “N”.
ANS. SELECT customerName FROM customers WHERE customerName LIKE 'N%';

9. Write a SQL query to show name of all the customers whose phone start with ‘7’ and are from city ‘LasVegas’.
ANS. SELECT customerNumber, customerName FROM customers WHERE city IN ('Melbourne city');

10. Write a SQL query to show name of all the customers whose creditLimit < 1000 and city is either “Las Vegas” or ”Nantes” or “Stavern”.
ANS. SELECT customerName,city from customers where creditLimit < 1000 AND city IN ('Las vegas','Nantes','Stavern');

11. Write a SQL query to show all the orderNumber in which quantity ordered <10.
ANS. SELECT orderNumber FROM orders WHERE orderNumber IN (SELECT orderNumber
FROM orderdetails where quantityOrdered < 10);

12. Write a SQL query to show all the orderNumber whose customer Name start with letter ‘N’.
ANS. SELECT orderNumber from orders WHERE customerNumber IN (SELECT
customerNumber FROM customers WHERE customerName like 'N%');

13. Write a SQL query to show all the customerName whose orders are “Disputed” in status.
ANS. SELECT customerName FROM customers where customerNumber IN (SELECT customerNumber
FROM orders WHERE status IN ('Disputed'));

14. Write a SQL query to show the customerName who made payment through cheque with checkNumber startingwith H and made payment on “2004-10-19”.
ANS. SELECT customerName FROM customers WHERE customerNumber = (SELECT customerNumber FROM
payments WHERE checkNumber LIKE 'c%' AND paymentDate='2020-12-21');

15. Write a SQL query to show all the checkNumber whose amount > 1000
ANS. SELECT checkNumber FROM payments where amount > 1000;

MACHINE LEARNING

In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:
Ans. C) between -1 and 1
A) between 0 and 1 B) greater than -1 C) between -1 and 1 D) between 0 and -1

2. Which of the following cannot be used for dimensionality reduction?
Ans. D) Ridge Regularisation
A) Lasso Regularisation B) PCA C) Recursive feature elimination D) Ridge Regularisation

3. Which of the following is not a kernel in Support Vector Machines?
Ans. C) hyperplane
A) linear B) Radial Basis Function C) hyperplane D) polynomial

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
Ans. A) Logistic Regression
A) Logistic Regression B) Naïve Bayes Classifier C) Decision Tree Classifier D) Support Vector Classifier

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be? (1 kilogram = 2.205 pounds)
Ans. C) old coefficient of 'X' ÷ 2.205
A) 2.205 × old coefficient of 'X' B) same as old coefficient of 'X' C) old coefficient of 'X' ÷ 2.205 D) Cannot be determined

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
Ans. B) increase
A) remains same B) increases C) decreases D) none of the above

7. Which of the following is not an advantage of using random forest instead of decision trees?
Ans. C) Random Forests are easy to interpret
A) Random Forests reduce overfitting B) Random Forests explains more variance in data then decision trees C) Random Forests are easy to interpret D) Random Forests provide a reliable feature importance estimate

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. Which of the following are correct about Principal Components?
Ans. D) All of the above
A) Principal Components are calculated using supervised learning techniques B) Principal Components are calculated using unsupervised learning techniques C) Principal Components are linear combinations of Linear Variables. D) All of the above

9. Which of the following are applications of clustering?
Ans. A, B, C and D
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts. C) Identifying spam or ham emails D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?
Ans= max_depth, max_features, min_samples_leaf
A) max_depth B) max_features C) n_estimators D) min_samples_leaf

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
Ans. An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data. Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset

IQR method= In IQR we take three quartile refer to Q1, Q2 and Q3. First quartile Q1 cover 25% data, second quartile Q2 cover 50% data, third quartile Q3 cover 75% data from whole data. Inter quartile range (IQR) = Q3-Q1 from this we removing outliers. After this we will calculate upper and lower fence value, Upper fence= Q3+(1.5*IQR) and
Lower fence= Q1-(1.5*IQR). Your any values greater than your upper fence or less than your lower fence so those value we said outliers

12. What is the primary difference between bagging and boosting algorithms?
Ans. BAGGING = Every model receives an equal weight. Objective to decrease variance, not to bias.
Every model is constructed independently. In Bagging weak learner Trained in parallel
BOOSTING= Model are weighted by their performance. Objective to decrease bias, not to variance.
In Boosting weak learners are Trained in sequentially

13. What is adjusted R2 in linear regression. How is it calculated?
Ans. Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.
Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2

14. What is the difference between Standardisation and Normalisation?
Ans. 1 =We do Normalisation on any dataset then a Normalised dataset will always have values range between 0 to 1.
2= we do Standardization on any dataset then a Standardized dataset will have a mean of 0 and std of 1, but there is no specific upper or lower bound for the maximum and minimum values

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation
Ans. Cross-validation= Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.
Advantage = In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities
Disadvantage = Increase Training time Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.

STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?
Ans. Suppose that we are interested in estimating the average height among all people. Collecting data are impossible of every person. we need to take random samples from population that are representative of the population. In this case Central Limit Theorem addresses this exactly.

2. What is sampling? How many sampling methods do you know?
Ans. When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.
Probability Sampling= involves random selection, allowing you to make strong statistical inferences about the whole group.
Non-Probability Sampling= involves non-random selection based on convenience or other criteria, allowing you to easily collect data.
3. What is the difference between type1 and type2 error?
Type1 error= The null hypothesis (Ho) is rejected when it is actually True, then the kind of error is known as type2 error. It is also called False Positive
Type2 error= The null hypothesis (Ho) is accepted when it is actually False, then the kind of error is known as type2 error. It is also called False Negative

4. What do you understand by the term Normal distribution?
Ans. In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region.
The Mean, Median and std are exactly the same.
The distribution can be described by two values, the mean and the standard deviation.
The normal distribution is often called the bell curve because the graph of its probability density looks like a bell.
5. What is correlation and covariance in statistics?
Ans. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.
Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

6. Differentiate between univariate , Biavariate, and multivariate analysis.
Ans. Univariate analysis is a basic kind of analysis technique for statistical data. Here the data contains just one variable and does not have to deal with the relationship of a cause and effect
The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot.
The multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.
7. What do you understand by sensitivity and how would you calculate it?
Ans. Sensitivity is a measure of how well a machine learning model can detect positive instances. It is also known as the true positive rate (TPR) or recall. Sensitivity is used to evaluate model performance because it allows us to see how many positive instances the model was able to correctly identify. A model with high sensitivity will have few false negatives
Sensitivity = (True Positive)/(True Positive + False Negative)

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
Ans. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.
The null hypothesis (H0) is the claim that there's no effect in the population.
The alternative hypothesis (H1). It claims that there's an effect in the population.

9. What is quantitative data and qualitative data?
Ans. Quantitative= Quantitative data are measures of values or counts and are expressed as number. Quantitative data are data about numeric variables.
Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

10. How to calculate range and interquartile range?
Ans. the range is the difference between the maximum and the minimum observation of the distribution.
Range = Xmax - Xmin
The difference between the upper and lower quartile is known as the interquartile range.
Interquartile range (IQR)= upper quartile (Q3) - lower quartile (Q1)

11. What do you understand by bell curve distribution?
Ans. A bell curve is a common type of distribution for a variable, also known as the normal distribution. The highest point on the curve, or the top of the bell, represents the most probable event in a series of data. The width of the bell curve is described by its standard deviation. which has a shape reminiscent of a bell.
The top of the curve shows the mean, mode, and median of the data collected.

12. Mention one method to find outliers.
Ans. By using Z-score we can get outliers of the data. Z-score tells you how many standard deviations away they are from the mean. If a value has high enough or low enough z-score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

13. What is p-value in hypothesis testing?

Ans. When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis. Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis.

14. What is the Binomial Probability Formula?

Ans. The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events

The binomial distribution formula is:$b(x; n, P) = nCx * Px * (1 – P)n – x$

b = binomial probability, x = total number of "successes" (pass or fail, heads or tails)

P = probability of success on an individual trial, n = number of trials

15. Explain ANOVA and it's applications

Ans. Analysis of variance, or ANOVA is a strong statistical technique that is used to show the difference between two or more means or components through significance tests

The ANOVA test is performed by comparing two types of variation, the variation between the sample mean, as well as variation within each of the samples.

the one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables

ANOVA is used in a business context to help manage budgets by comparing your budget to costs to help manage revenue and inventory, for example. ANOVA can also be used to forecast trends by analyzing patterns in data to better understand the future performance of sales. It's also a widely used statistical technique for comparing the relationship between factors that cause a rise in sales, such as how improving the features of a product resulted in a sales increase, which helps businesses tailor their product development and take measures to improve products in the future.