**SQL WORKSHEET-5**

1. Write SQL query to show all the data in the Movie table.

**Answer**. select \* from movie;

2. Write SQL query to show the title of the longest runtime movie.

**Answer**. select title from movie order by runtime desc limit 1;

3. Write SQL query to show the highest revenue generating movie title.

**Answer**. select title from movie order by revenue desc limit 1;

4. Write SQL query to show the movie title with maximum value of revenue/budget.

**Answer**. select title from movie order by budget desc limit 1;

5. Write a SQL query to show the movie title and its cast details like name of the person, gender, character name, cast order.

**Answer**. select title, person\_name, gender, character\_name, cast\_order

-> from movie

->inner join movie\_cast on movie.movie\_id=movie\_cast.movie\_id

-> inner join gender on movie\_cast.gender\_id=gender.gender\_id

->inner join person on movie\_cast.person\_id=person.person\_id;

6. Write a SQL query to show the country name where maximum number of movies has been produced, along with the number of movies produced.

**Answer**. select country\_name, count(movie\_id) as count\_movie

-> from country

-> inner join production\_country on country.country\_id=production\_country.country\_id

-> group by country\_name order by count(movie\_id) desc limit 1;

7. Write a SQL query to show all the genre\_id in one column and genre\_name in second column.

**Answer**. select \* from genre;

8. Write a SQL query to show name of all the languages in one column and number of movies in that particular column in another column.

**Answer**. select language\_name, count(movie\_id) as movie\_count

-> from language

-> inner join movie\_language on language.language\_id=movie\_language.language\_id

-> group by language\_name;

9. Write a SQL query to show movie name in first column, no. of crew members in second column and number of cast members in third column.

**Answer**.

10. Write a SQL query to list top 10 movies title according to popularity column in decreasing order.

**Answer**. select title from movie order by popularity desc limit 10;

11. Write a SQL query to show the name of the 3rd most revenue generating movie and its revenue.

**Answer**. select title, revenue

-> from (select title, revenue from movies order by revenue desc limit 3) as comp

-> order by revenue limit 1;

12. Write a SQL query to show the names of all the movies which have “rumoured” movie status.

**Answer**. select title from movie where movie\_status=’rumoured’;

13. Write a SQL query to show the name of the “United States of America” produced movie which generated maximum revenue.

**Answer**. select title from movie

-> inner join production\_country on movie.movie\_id=production\_country.movie\_id

-> inner join country on production\_country.country\_id=country.country\_id

-> where country\_name='United States of America'

-> order by revenue desc limit 1;

14. Write a SQL query to print the movie\_id in one column and name of the production company in the second column for all the movies.

**Answer**. select movie\_id, company\_name

-> from movie\_company inner join production\_company

-> on movie\_company.company\_id=production\_company.company\_id

15. Write a SQL query to show the title of top 20 movies arranged in decreasing order of their budget.

**Answer**. select title from movie order by budget desc limit 20;

**STATISTICS WORKSHEET-5**

1. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies.

a) Mean b) Actual c) Predicted d) Expected

**Answer**. d) Expected

2. Chisquare is used to analyse

a) Score b) Rank c) Frequencies d) All of these

**Answer.** c) Frequencies

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

a) 4 b) 12 c) 6 d) 8

**Answer.** c) Mean

4. Which of these distributions is used for a goodness of fit testing?

a) Normal distribution

b) Chisqared distribution

c) Gamma distribution

d) Poission distribution

**Answer**. b) Chi-squared distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution

b) Hypergeometric Distribution

c) F Distribution

d) Poisson Distribution

**Answer**. c) F Distribution

6. A statement made about a population for testing purpose is called?

a) Statistic b) Hypothesis c) Level of Significance d) TestStatistic

**Answer**. b) Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis

**Answer**. a) Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

a) Two tailed b) One tailed c) Three tailed d) Zero tailed

**Answer**. a) Two-tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis b) Research Hypothesis c) Simple Hypothesis d) Null Hypothesis

**Answer**. b) Research Hypothesis

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

a) np b)n

**Answer.** a) np

**MACHINE LEARNING WORKSHEET 5**

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

**Answer.** RSS changes with a change in the units of actual y and predicted y. So, it is better to use R squared which is 1-RSS/TSS.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

**Answer.** ESS- Explained sum of squares also known as explained variation is the portion of total variation that measures how well the regression equation explains the relationship between independent and dependant variables. ESS measures how much variation is there in the modelled values. ESS is represented numerically as

ESS=

RSS- Residual sum of squares also called unexplained variation is the portion of total variation that measures errors between the actual values of y and the predicted values of y estimated by the regression model. It is a measure of the discrepancy between the data and an estimation model, such as a linear regression. Numerically RSs is given as

RSS=

The smaller the value of RSS relative to ESS, the better the regression line fits or explains the relationship between dependant variable and independent variable.

TSS- Also called total sum of squares is the sum of RSS and ESS. It measures the variation in the observed data. It is defined as the sum of squared difference between the observation and the overall mean and is given as

TSS=RSS+ESS=

3. What is the need of regularization in machine learning?

**Answer.** Regularization is a regression technique that constraints or shrinks the co-efficient estimates towards zero in order to discourage learning a complexed or more flexible model to reduce the risk of overfitting. Regularization reduces the error by fitting the function appropriately on the given training set to avoid over-fitting.

4. What is Gini–impurity index?

**Answer.** Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. It is the probability of misclassifying a randomly chosen element in the dataset if it were randomly labelled according to the class distribution in the dataset. Gini impurity of zero is the lowest and best possible impurity.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

**Answer.** Unregularized decision trees are prone to overfitting because decision trees are data intensive which means decision trees examine data in a lot of different ways. At each node, they look at every possible split of every feature available, and also multiple times using the same feature leading to unnecessary complexity and overfitting. We can regularize decision trees by- limiting max depth of the tree as deeper the tree more complex the result will be. We can also use ensemble technique to regularize the decision trees. Also, we can set stricter stopping criterion on when to split a node further.

6. What is an ensemble technique in machine learning?

**Answer.** Ensemble methods are techniques that combine several base models or weak learners to produce an optimal predictive strong learning model. Ensemble technique usually produce accurate solutions than a single model would.

7. What is the difference between Bagging and Boosting techniques?

**Answer.** Bagging also known as Bootstrap Aggregation is used to decrease variance in the prediction model. Bagging is a parallel method that fits different learners independently from each other and train them simultaneously. Each model receives an equal weight. Training data subsets are drawn randomly with replacement from the training dataset.

Boosting is a sequential ensemble method that iteratively adjusts the weight of observation as per the last classification. If an observation is incorrectly classified, it increases the weight of the observation. Boosting algorithm allocates weights to each resulting model during training based on their performance. A learner with good training data prediction will be assign a higher weight. Boosting decreases the bias error and builds a strong model. In Boosting, every model comprises of elements that were misclassified by the previous model.

8. What is out-of-bag error in random forests?

**Answer.** Out of bag error is a method of measuring error estimate of a machine learning model. Out of bag error is a method of measuring the prediction error of random forests. Random forests is trained using Bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observation ai = ( xi , yi ) . Out of bag error is the average error for each ai calculated using predictions from the trees which do not contain ai in their respective bootstrap sample.

9. What is K-fold cross-validation?

**Answer.** Cross validation is a model validation technique used to evaluate machine learning models on a limited data sample. Cross validation has a single parameter k that refers to the number of group that a given data sample is to be split into. For that reason, the technique is called k-fold cross validation. The goal of cross validation is to test model’s ability to predict the new data that was not used in estimating it, in order to overcome problems like over fitting or selection bias.

10. What is hyper parameter tuning in machine learning and why it is done?

**Answer.** Hyper parameter tuning is the method of optimizing the hyperparameters by choosing the optimal parameters for a machine learning algorithm. In simple terms the process of searching for the optimal parameters of a machine learning algorithm is called Hyper parameter tuning. Hyperparameters are the parameters used to control a machine learning process.

Hyperparameters are crucial in a machine learning algorithm as they control the overall behaviour of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a pre-defined loss function to give better results.

11. What issues can occur if we have a large learning rate in Gradient Descent?

**Answer**. When learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. As higher learning rate causes drastic updates which leads to divergent behaviours.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

**Answer.** Logistic regression cannot be used for classification of non-linear data as the decision boundary is linear.

13. Differentiate between Adaboost and Gradient Boosting.

**Answer**. Gradient boosting is based on minimising the loss function. Gradient boosting generates learners during the learning process. It builds first learner to predict the values and calculate the loss, it will build the second learner to predict the loss after the first step and so on.

Adaboost requires users to specify weak learners before the learning process. At each iteration adaboost changes the sample distribution by modifying the weights attached to each of the instances. It increases the weights of the wrongly predicted instances and decreases the weights of the correctly predicted instances. The weak learners thus focus on the difficult instances.

14. What is bias-variance trade off in machine learning?

**Answer.** A bias variance trade off is a way to diagnose the performance of an algorithm by breaking down its prediction error. Bias and variance are the prediction errors. If our model is too simple and has a few parameters then it may have high bias and low variance. On the other hand, if our model has large number of parameters then it will have high variance and low bias. We need to find a good balance between bias and variance without overfitting and underfitting the data. We can say an algorithm can’t be more complex or less complex at the same time.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

**Answer.** Linear kernel is used when the data is linearly separable i.e., it can be separated using a single line. It is used when there are a large number of features in the dataset. It is one of the most common used kernels.

RBF or radial basis function is used when the boundaries are hypothesized to be curve shaped. RBF kernel is a function whose values depends on the distance from the origin or from some point.

Polynomial kernel is a function used in SVM that represents the similarity of training samples in a feature space over polynomials of the original variables, allowing learning of non-linear models. In polynomial kernel we calculate the dot product by increasing the power of the kernel to be able to map our data into higher dimensions.