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

SELECT \*

FROM movie;

1. Write SQL query to show the title of the longest runtime movie
2. Write SQL query to show the highest revenue generating movie title

SELECT title,MAX(revenue)

FROM movie;

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

SELECT title,MAX(budget)

FROM movie;

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

SELECT title,person,gender,character\_name,cast\_order

FROM movie,movie\_cast,person,gender

JOIN gender

ON gender.gender\_id=movie\_cast.gender\_id

JOIN person

ON person.person\_id= movie\_cast.person\_id

JOIN movie

ON movie.movie\_id=movie\_cast.movie\_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

SELECT country

FROM movie,production\_country,country

JOIN country

ON country.country\_id=production\_country.country\_id

JOIN movie

ON movie.movie\_id=production\_country.country\_id

ORDER BY COUNT(movie) DESC LIMIT 1;

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

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.

SELECT language\_name

FROM language,movie\_languages,movie

JOIN language

ON language.language\_id=movie\_languages.languages\_id

JOIN movie

ON movie.movie\_id= movie\_languages.languages\_id

ORDER BY COUNT(movie) DESC LIMIT 1;

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.

SELECT title, movie\_crew.COUNT(person\_id), movie\_cast.COUNT (person\_id)

FROM movie ,movie\_cast,movie\_crew

ORDER BY movie;

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

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

SELECT title

FROM movie

ORDER BY revenue DESC LIMIT 1 OFFSET 2;

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

SELECT titles from movie

WHERE movie.movie\_status like '% rumoured %';

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

SELECT MA(revenue)

WHERE movie.movie\_id=production\_country.movie\_id

And country.country\_id=production\_country.country\_id

And country.country\_name like '% United States of America %';

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

SELECT movie\_id,company\_name

FROM movie\_company,production\_company

WHERE production\_company.company\_id=movie\_company.company\_id;

MACHINE LEARNING

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?

RSS  for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

R-squared statistic or coefficient of determination is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

RSS is better as it gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is not explained by our model.

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.

RSS is defined as between the *observed* value and the *predicted* value.

TSS  is the squared differences between the observed *dependent variable* and its **mean.**

 ESS is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*.

The equation the three are related:

**TSS** = **RSS** + **ESS**.

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

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to **properly fit a model onto our test set**. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

4. What is Gini–impurity index?

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

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

6. What is an ensemble technique in machine learning?

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

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

|  |  |
| --- | --- |
| Bagging | Boosting |
| It is a method of merging the same type of predictions. | It is a method of merging different type f predictions. |
| It decreases variance, not bias, and solves over-fitting issues in a model | It decreases bias, not variance. |
| Model receives an equal weight. | Model is weighed based on their performance. |

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

The **out**-**of**-**bag** **error** is an **error** estimation technique often used to evaluate the accuracy of a **random** **forest** and to select appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split, referred to as mtry.

9. What is K-fold cross-validation?

 It is a **data partitioning strategy**so that you can effectively use your dataset to build **a more generalized model**. The main intention of doing any kind of machine learning is to develop a more generalized model which can perform well on **unseen data**.

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

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model

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

 If we choose a learning rate that is too large, it might overshoot the minimum value of the error function, and may even never reach the optimal solution.

13. Differentiate between Adaboost and Gradient Boosting.

AdaBoost or Adaptive Boosting is the first [Boosting ensemble model](https://analyticsindiamag.com/ensemble-modeling-explained-through-music/). The method automatically adjusts its parameters to the data based on the actual performance in the current iteration. Meaning, both the weights for re-weighting the data and the weights for the final aggregation are re-computed iteratively. In practice, this boosting technique is used with [simple classification trees](https://arxiv.org/pdf/1403.1452.pdf) or stumps as base-learners, which resulted in improved performance compared to the classification by one tree or other single base-learner.

Gradient Boost is a robust [machine learning algorithm](https://analyticsindiamag.com/pseudo-labelling-a-guide-to-semi-supervised-learning/) made up of Gradient descent and Boosting. The word ‘gradient’ implies that you can have two or more derivatives of the same function. Gradient Boosting has three main components: additive model, loss function and a weak learner. The technique yields a direct interpretation of boosting methods from the perspective of numerical optimisation in a function space and generalises them by allowing optimisation of an arbitrary loss function.

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

Bias is the gap between the actual value and the predicted value. Variance means how much-predicted values are scattered in relation to each other

The bias–variance tradeoff is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters. It’s a way to diagnose the performance of an algorithm by breaking down its prediction error.

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

Linear

It is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used

Equatin is:

F(x, xj) = sum( x.xj)

(Here, x, xj represents the data you’re trying to classify.)

RBF

It is a general-purpose kernel; used when there is no prior knowledge about the data.  
Equation is:

[Gaussian radial basis function (RBF)](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2017/08/gaussian-radial-basis-function-RBF.png)

Gaussian radial basis function (RBF)

Polynomial:

It is popular in image processing.  
Equation is:

[Polynomial kernel equation](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2017/08/polynomial-kernel.png)

Polynomial kernel equation

where d is the degree of the polynomial.