**SQL subjective question**

**11. What are joins in SQL?**

A SQL Join statement is used to combine data or rows from two or more tables based on a common field between them

**12. What are the different types of joins in SQL?**

There are four different types of the JOINs in SQL:

* **(INNER) JOIN**: Returns records that have matching values in both tables
* **LEFT (OUTER) JOIN:** Returns all records from the left table, and the matched records from the right table
* **RIGHT (OUTER) JOIN:** Returns all records from the right table, and the matched records from the left table
* **FULL (OUTER) JOIN:** Returns all records when there is a match in either left or right table

**13. What is SQL Server?**

**SQL SERVER** is a relational database management system (RDBMS) developed by Microsoft. It is primarily designed and developed to compete with MySQL and Oracle database.

**14. What is primary key in SQL?**

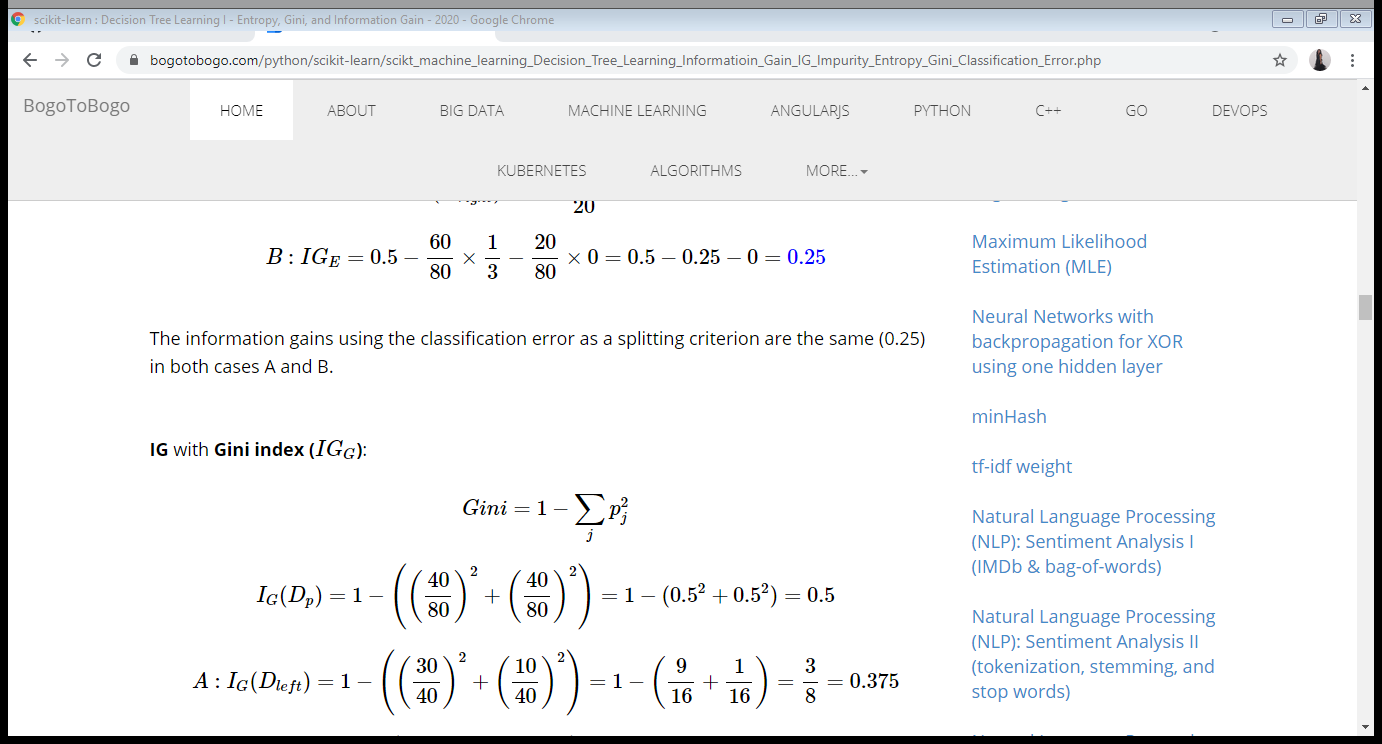
The primary key constraint uniquely identifies each record in a table. Primary keys must contain UNIQUE values, and cannot contain NULL values. A table can have only ONE primary key; and in the table, this primary key can consist of single or multiple columns (fields).

**15. What is ETL in SQL?**

**ETL** stands for Extract, Transform and Load, which is a process used to collect data from various sources, transform the data depending on business rules/needs and load the data into a destination database.

**Machine learning subjective questions:**

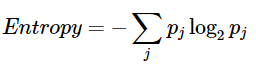
**9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.**



Gini = 1 - (0.40^2 + 0.60^2)

= 1 - (0.16+0.36)

= 1 – (0.52) = 0.48



Entropy = - [0.4 \* log2(0.4) + 0.6 \* log2(0.6)]

= - [0.4 \* -1.32192809489 + 0.6 \* -0.736965594166 ]

= 0.97

**10. What are the advantages of Random Forests over Decision Tree?**

Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees.

Although decision trees have a low bias, they suffer from a high variance which makes them less useful for most practical applications.

Advantages of RandomForest are:

* Robust to outliers.
* Works well with non-linear data.
* Lower risk of overfitting.
* Can handle several features at once
* Runs efficiently on a large dataset.

**11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.**

**Scaling is required for correct prediction and results**. In case when the values of one of the column is very high as compare to others, the impact of the column with higher value will be much higher as compared to the impact of other low valued columns**. The feature with high magnitude will weigh lot more than features having low magnitude even if they are more crucial in determining the output**. Hence the prediction might not give the expected results and may fail to satisfy the business use case.

* MIN-MAX SCALAR
* STANDARD SCALAR

**12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.**

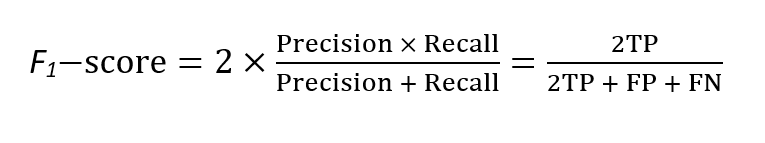
Feature scaling aims to speed up the process of convergence of gradient descent.

**13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?**

In a dataset with highly unbalanced classes, the classifier will always “predicts” the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

**14. What is “f-score" metric? Write its mathematical formula.**

The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.



**15. What is the difference between fit(), transform() and fit\_transform()?**

In the fit() method, where we use the required formula and perform the calculation on the feature values of input data and fit this calculation to the transformer.

In the transform () method, where we apply the calculations that we have calculated in fit() to every data point in feature F. We have to use .transform().

This fit\_transform() method is basically the combination of fit method and transform method, it is equivalent to fit().transform(). This method performs fit and transform on the input data at a single time and converts the data points.