DSBDAL Practical Sample Oral Questions—-

1. DSBDAL Practical —Follow the instructions given by Nalini Mam and Tejashwini Mam.

   Following are Some sample question related to DSBDAL Practical also read the concepts of Group A,B and C Dsbdal practical.

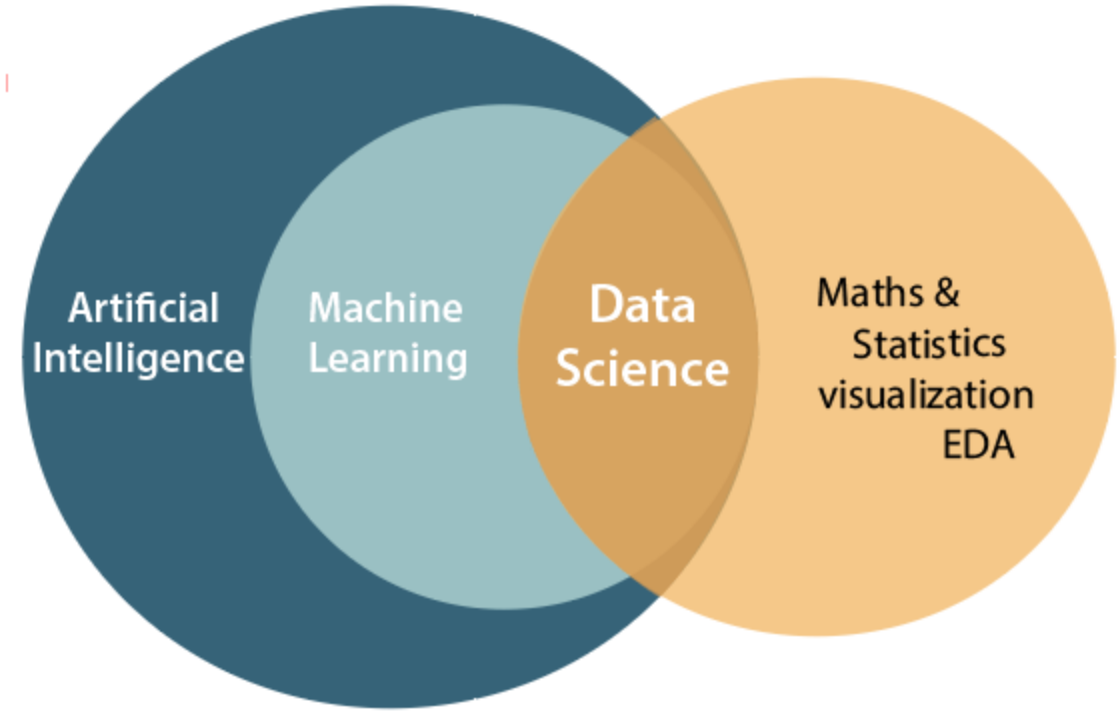
1) What do you understand by the term Data Science?

- Data science is a multidisciplinary field that combines **statistics, data analysis, machine learning, Mathematics, computer science**, and related methods, to understand the data and to solve complex problems.
- Data Science is a deep study of the massive amount of data, and finding useful information from raw, structured, and unstructured data.
- Data science is similar to data mining or big data techniques, which deals with a huge amount of data and extract insights from data.
- It uses various tools, powerful programming, scientific methods, and algorithms to solve the data-related problems.

---

2) What are the differences between Data Science, Machine Learning, and Artificial intelligence?

Data science, Machine learning, and Artificial Intelligence are the three related and most confusing concepts of computer science. Below diagram is showing the relation between AI, ML, and Data Science.



Following are some main points to differentiate between these three terms:

Data Science	Artificial Intelligence	Machine Learning
Data science is a multidisciplinary field that is used for deep study of data and finding useful insights from it.	Artificial Intelligence is a branch of computer science that build intelligent machines which can mimic the human brain.	Machine learning is a branch of computer science which enables machines to learn from the data automatically.
Data Science is not exactly a subset of artificial intelligence and machine learning, but it uses ML	Artificial Intelligence is a wide field which ranges from natural language processing to deep learning.	Machine learning is a subset of Artificial Intelligence and a part of data science.

algorithms for data analysis and future prediction.		
The goal of Data science is to find hidden patterns from the raw data.	The goal of artificial intelligence is to make intelligent machines.	The goal of machine learning is to allow a machine to learn from data automatically.
Data science finds meaningful insights from data to solve complex problems.	Artificial intelligence creates intelligent machines to solve complex problems.	Machine learning uses data and train models to solve some specific problems.

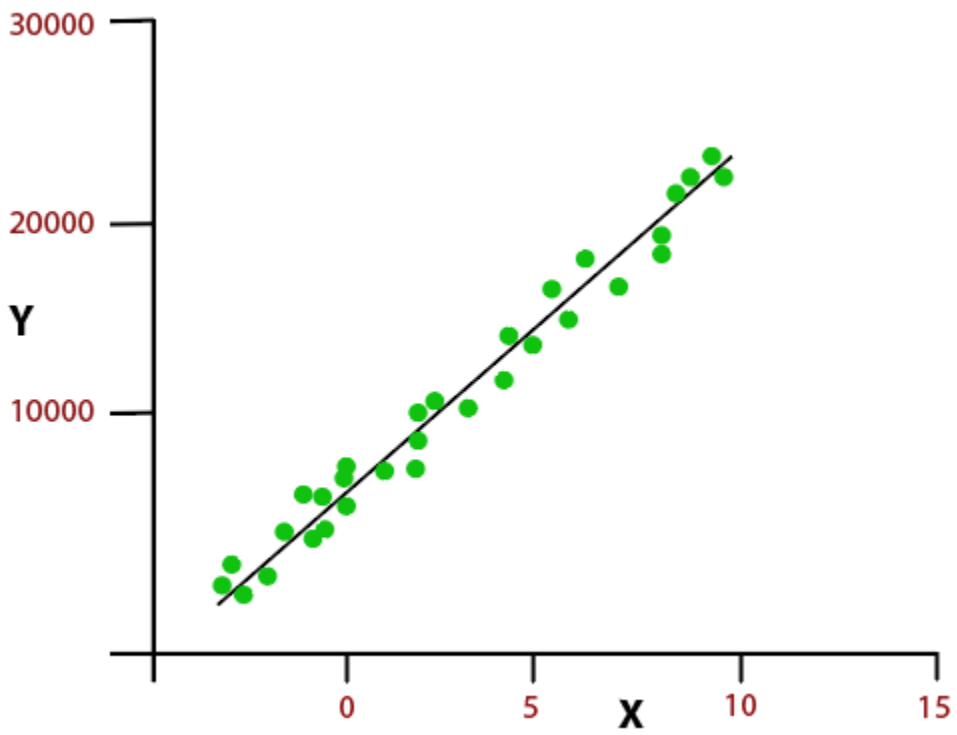
---

3) Discuss Linear Regression?

- Linear Regression is one of the popular machine learning algorithms based on supervised learning, which is used for understanding the relationship between input and output numerical variables.
- It applies **regression analysis**, a predictive modeling technique that finds a relationship between the dependent and independent variables.
- It shows the linear relationship between **independent** and **dependent variables**, hence it is called a linear regression algorithm.
- Linear Regression is used for prediction of continuous numerical variables such as sales/day, temperature, etc.
- It can be divided into two categories:
    a. **Simple Linear Regression**
    b. **Multiple Linear Regression**

If we talk about simple linear regression algorithm, then it shows a linear relationship between the variables, which can be understood using the below equation, and graph plot.

1. y=mx + c



---

4) Differentiate between Supervised and Unsupervised Learning?

Supervised and Unsupervised learning are types of Machine learning.

Supervised Learning:

Supervised learning is based on the supervision concept. In supervised learning, we train our machine learning model using sample data, and on the basis of that training data, the model predicts the output.

Unsupervised learning:

Unsupervised learning does not have any supervision concept. Hence, in unsupervised learning machine learns without any supervision. In unsupervised learning, we provide data which is not labeled, classified, or categorized.

Below are some main differences between supervised and unsupervised learning:

Sr. No.	Supervised Learning	Unsupervised learning
1.	In supervised learning, the machine learns in supervision using training data.	In unsupervised learning, the machine learns without any supervision.
2.	Supervised learning uses labeled data to train the model.	Unsupervised learning uses unlabeled data to train the model.
3.	It uses known input data with the corresponding output.	It uses unknown data without any corresponding output.
4.	It can be grouped into **Classification** and **Regression** algorithms.	It can be grouped into **Clustering** and **Association** algorithms.
5.	It has more complex computation than Unsupervised learning.	It has less complex computation than supervised learning.
6.	It provides more accurate and reliable output.	It provides less reliable and less accurate output.
7.	It can also use Off-line data analysis.	It uses real-time data analysis.

5) What do you understand by bias, variance trade-off?

When we work with a supervised machine learning algorithm, the model learns from the training data. The model always tries to best estimate the mapping function between the output variable(Y) and the input variable(X). The estimation for target function may generate the prediction error, which can be divided mainly into **Bias error**, and **Variance error**. These errors can be explained as:

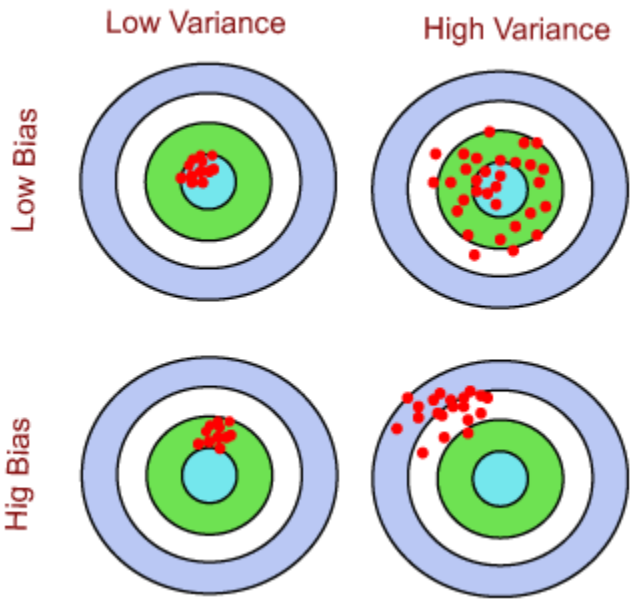
- **Bias Error:** Bias is a prediction error which is introduced in the model due to oversimplifying the machine learning algorithms. It is the difference of predicted output and actual output. There are two types of bias:
    - **High Bias:** If the suggested predicted values are much different from actual value, then it is called as high bias. Due to high bias, an algorithm may miss the relevant relationships between the input features and target output, which is called **underfitting**.
    - **Low Bias:** If the suggested predicted values are less different from actual value, then it is called as **low bias**.
- **Variance Error:** If the machine learning model performs well with training dataset, but does not perform well with test dataset, then variance occurs. It can also be defined as **an error caused by the model's sensitivity to small fluctuation in training dataset**. The high variance would cause Overfitting in machine learning model, which means an algorithm introduce noise along with the underlying pattern in data to the model.

Bias Variance tradeoff:

In the machine learning model, we always try to have low bias and low variance, and

- If we try to increase the bias, the variance decreases
- If we try to increase the variance, the bias decreases.

Hence, trying to get an optimal bias and variance is called **bias-variance trade-off**. We can define it using the Bull eye diagram given below. There are four cases of bias and variances:



- If there is low bias and low variance, the predicted output is mostly close to the desired output.
- If there is low bias and high variance, the model is not consistent.
- If there is high variance and low bias, the model is consistent but predicted results are far away from the actual output.
- If there is high bias and high variance, then the model is inconsistent, and also predictions are much different with actual value. It is the worst case of bias and variance.

---

6) Define Naive Bayes?

Naive Bayes is a popular classification algorithm used for predictive modeling. It is a supervised machine learning algorithm which is based on Bayes theorem.

It is easy to build a model using Naive Bayes algorithm when working with a large dataset. It is comprised of two words, Naive and Bayes, where Naive means features are unrelated to each other.

In simple words, we can say that "**Naive Bayes classifier assumes that the features present in a class are statistically independent to the other features**."

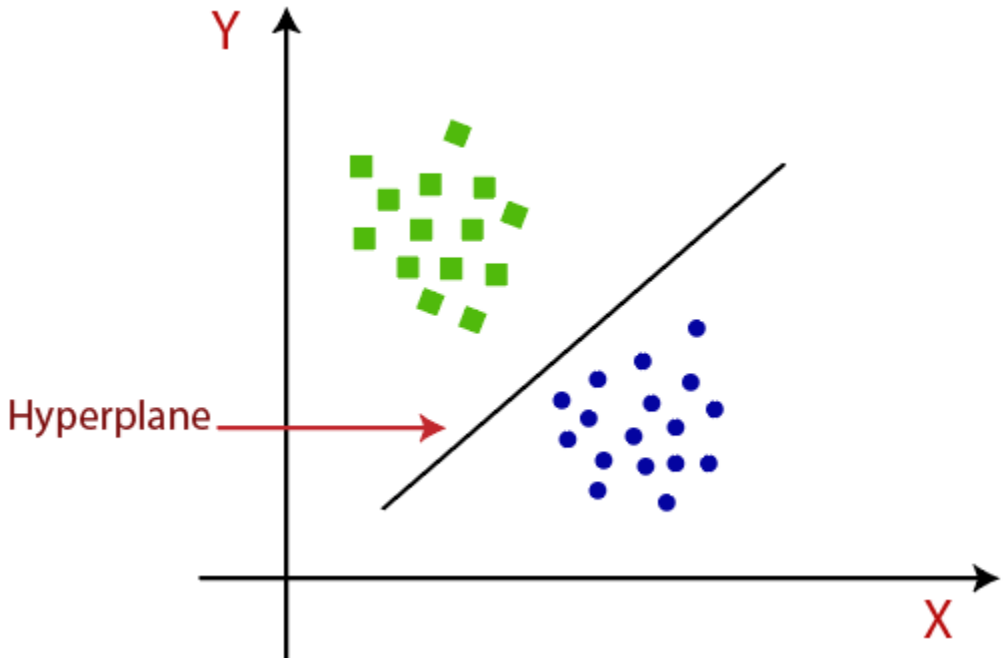
---

7) What is the SVM algorithm?

SVM stands for **Support Vector Machine**. It is a supervised machine learning algorithm which is used for classification and regression analysis.

It works with labeled data as it is a part of supervised learning. The goal of support vector machine algorithm is to construct a hyperplane in an N-dimensional space. The **hyperplane** is a dividing line which distinct the objects of two different classes, it is also known as a **decision boundary**.

If there are only two distinct classes, then it is called as **Binary SVM classifier**. A schematic example of binary SVM classifier is given below.



The data point of a class which is nearest to the other class is called a support vector.

There are two types of SVM classifier:

- **Linear SVM classifier:** A classifier by which we can separate the set of objects into their respective group by drawing a single line, i.e., hyperplane, called as linear SVM classifier.
- **Non-Linear SVM classifier:** Non-linear SVM classifier applies on those objects which cannot be classified into two groups by a single line.

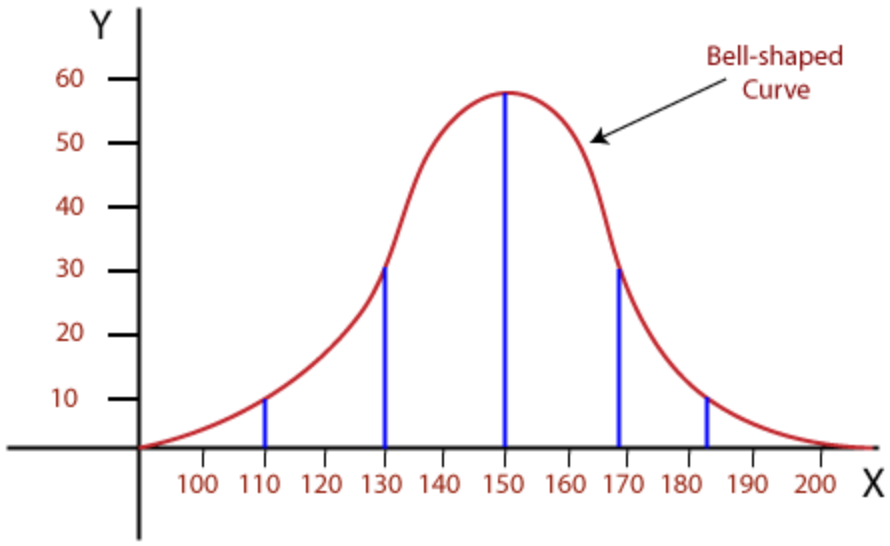
On the basis of error function, we can divide a SVM model into four categories:

- **Classification SVM Type1**
- **Classification SVM Type2**
- **Regression SVM Type1**
- **Regression SVM Type1**

---

8) What do you understand by Normal distribution?

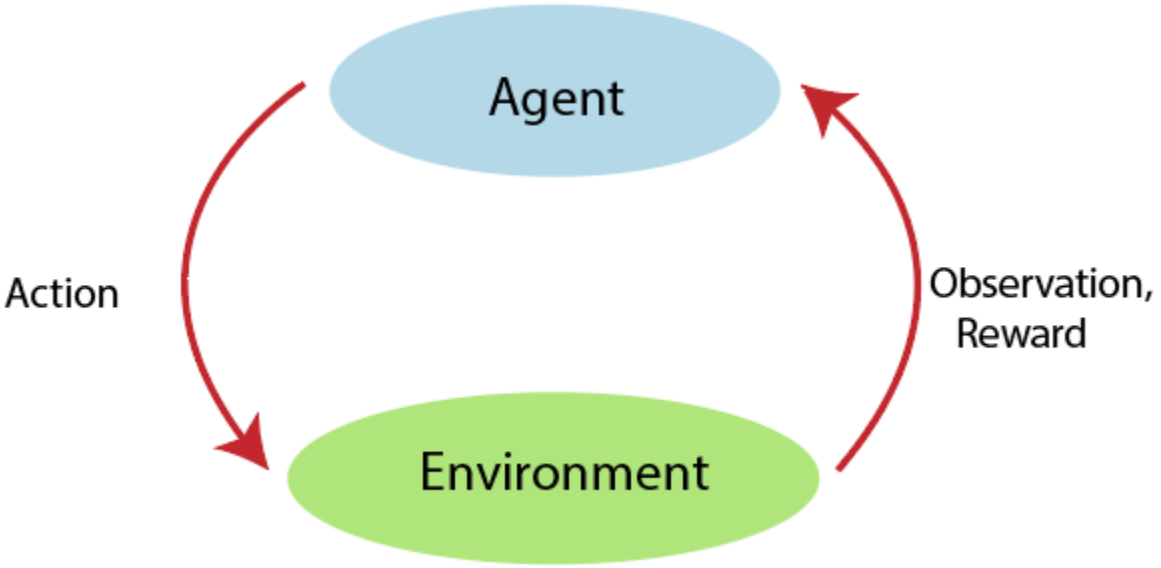
- If the given data is distributed around a central value in the bell-shaped curve without any left or right bias, then it is called **Normal distribution**. It is also called a **Bell Curve** because it looks like a bell?shaped curve.
- The normal distribution has a mean value, half of the data lies to the left of the curve, and half of the data lies right of the curve.
- In probability theory, the normal distribution is also called a **Gaussian distribution**, which is used for the probability distribution.
- It is a probability distribution function used to see the distribution of data over the given range.
- Normal distribution has two important parameters: **mean(μ)** and **standard deviation(σ)**.



---

9) Explain Reinforcement learning.

- Reinforcement learning is a type of machine learning where an agent interacts with the environment and learns by his actions and outcomes. On each good action, he gets a positive reward, and for each bad action, he gets a negative reward. Consider the below image:



- The goal of an agent in reinforcement learning is to maximize positive rewards.
- In reinforcement learning, algorithms are not explicitly programmed for tasks but learns with experiences without any human intervention.
- The reinforcement learning algorithms is different from supervised learning algorithms as there is no any training dataset is provided to the algorithm. Hence the algorithm automatically learns from experiences.

---

10) What do you mean by p-value?

- The p-value is the probability value which is used to determine the statistical significance in a hypothesis test.
- Hypothesis tests are used to check the validity of the null hypothesis (claim).
- P-values can be calculated using p-value tables or statistical software.
- The p-values lies between 0 and 1. It can have mainly two cases:
    - (p-value<0.05): A small p-value indicates strong evidence against the null hypothesis, so we can reject the null hypothesis.
    - (p-value>0.05): A large p-value indicates weak evidence against the null hypothesis, so we consider the null hypothesis as true.

---

11) Differentiate between Regression and Classification algorithms?

Classification and Regression both are the supervised learning algorithms in machine learning, and uses the same concept of training datasets for making predictions. The main difference between both the algorithms is that the output variable in regression algorithms is **Numerical** or **continuous**, whereas in Classification algorithm output variables are **Categorical** or **discrete**.

Regression Algorithm: A regression algorithm is about mapping the input variable x to some real numbers such as percentage, age, etc. Or we can say regression algorithms are used if the required output is continuous. **Linear regression is a famous example of the regression algorithm**.

Regression Algorithms are used in **weather forecasting, population growth prediction, market forecasting, etc**.

Classification Algorithm: A classification algorithm is about mapping the input variable x with a discrete number of labels such as true or false, yes or no, male-female, etc. Or we can say Classification algorithm is used if the required output is a discrete label. **Logistic regression** and **decision trees** are popular examples of a classification algorithm. The classification algorithm is used for **image classification, spam detection, identity fraud detection, etc**.

---

12) Which is the best suitable language among Python and R for text analytics?

Both R and Python are the suitable language for text analytics, but the preferred language is Python, because:

- Python has Pandas library, by which we can easily use data structure and data analysis tools.
- Python performs fast execution for all types of text analytics.

---

13) What do you understand by L1 and L2 regularization methods?

Regularization is a technique to reduce the complexity of the model. It helps to solve the over-fitting problem in a model when we have a large number of features in a dataset. Regularization controls the model complexity by adding a penalty term to the objective function.

There are two main regularization methods:

L1 Regularization:

- L1 regularization method is also known as Lasso Regularization. L1 regularization adds a penalty term to the error function, where penalty term is the sum of the absolute values of weights.
- It performs feature selection by providing 0 weight to unimportant features and non-zero weight to important features.
- It is given below:

$$L(x, y)= \text{Min}(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} |\mathbf{w_i}|)$$

  - Here $\sum_{i=1}^{n}(y_i - w_i x_i)^2$ is the sum of the squared difference between the actual value and the predicted value.
  - $\lambda \sum_{i=1}^{n} |\mathbf{w_i}$ is **regularization term**, and λ is penalty parameter which determines how much to penalize the weights.

L2 Regularization:

- L2 regularization method is also known as Ridge Regularization. L2 regularization does the same as L1 regularization except that penalty term in L2 regularization is the sum of the squared values of weights.
- It performs well if all the input features affect the output and all weights are of approximately equal size.
- It is given as:

$$L(x, y) = \text{Min}(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}(w_i)^2)$$

- Here, $\sum_{i=1}^{n}(y_i - w_i x_i)^2$ is the sum of the squared difference between actual value and predicted value.
- $\lambda \sum_{i=1}^{n}(w_i)^2$ is the regularization term, and λ is the penalty parameter which determines how much to penalize the weights.

---

14) What is the 80/20 rule? Explain its importance in model validation?

In machine learning, we usually split the dataset into two parts:

- **Training set:** Part of the dataset used to train the model.
- **Test set:** Part of the dataset used to test the performance of the model.

The best ratio to split the dataset is 80-20%, to create the validation set for machine learning model. Here, 80% is assigned for the training dataset, and 20% is for the test dataset. This ratio maybe 90-20%, 70-30%, 60-40%, but these ratios would not be preferable.

Importance of 80/20 rule in model validation:

The process of evaluating a trained model on the test dataset is called as **model validation** in machine learning. In model validation, the ratio of splitting dataset is important to avoid Overfitting problem. The best preferable ration is 80-20%, which is also known as 80/20 rule, but it also depends upon the amount of data in a dataset.

---

15) What do you understand by confusion matrix?

- Confusion matrix is a unique concept of the statistical classification problem.
- Confusion matrix is a type of table which is used for describing or measuring the performance of Binary classification model in machine learning.

- The confusion matrix is itself easy to understand, but the terminologies used in the matrix can be confusing. It is also known as **Error matrix**.
- It is used in statistics, data mining, machine learning, and different Artificial Intelligence applications.
- It is a table with two dimensions, "actual and predicted" and identical set of classes in both dimensions of the table.
- The confusion matrix has four following cases:
    - **True Positive(TP):** The predictions is positive and its actually true.
    - **False Positive(FP):** The prediction is positive but its actually false.
    - **True Negative(TN):** The prediction is negative but its actually true.
    - **False Negative(FN):** The prediction is negative and its false.

The classification accuracy can be obtained by the below formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

---

16) What is the ROC curve?

ROC curve stands for **Receiver Operating Characteristics** curve, which graphically represents the performance of a binary classifier model at all classification threshold. The curve is a plot of true positive rate (TPR) against false positive rate (FPR) for different threshold points.

---

17) Explain the Decision Tree algorithm, and how is it different from the random forest algorithm?

- Decision tree algorithm belongs to supervised learning which solves both classifications and Regression problems in machine learning.

- Decision tree solves problems using a tree-type structure which has leaves, decision nodes, and links between nodes. Each node represents an attribute or feature, each branch of the tree represent the decision, and each leaf represents the outcomes.
- Decision tree algorithm often mimic human thinking hence, it can be easily understood as compared to other classifications algorithm.

Difference between Decision Tree and Random Forest algorithm:

Decision Tree Algorithm	Random Forest Algorithm
Decision tree algorithm is a tree-like structure to solve classification and regression problems.	Random forest algorithm is a combination of various decision trees which gives the final output based on the average of each tree output.
Decision tree may have a chance of Overfitting problem.	Random Forest reduces the chance of Overfitting problem by averaging out several trees predictions.
Simpler to understand as it is based on human thinking.	This algorithm is comparatively complex.
It gives less accurate result as compared to the random forest algorithm.	It gives a more accurate result.

---

18) Explain the term "Data warehouse".

The data warehouse is a system which is used for analysis and reporting of data collected from operational systems and different data sources. Data warehouse plays an important role in Business Intelligence.

In a data warehouse, data is extracted from various sources, transformed (cleaned and integrated) according to decision support system needs, and stored into a data warehouse.

The data present in the data warehouse after analysis does not change, and it is directly used by end-users or for data visualization.

Advantages of Data Warehouse:

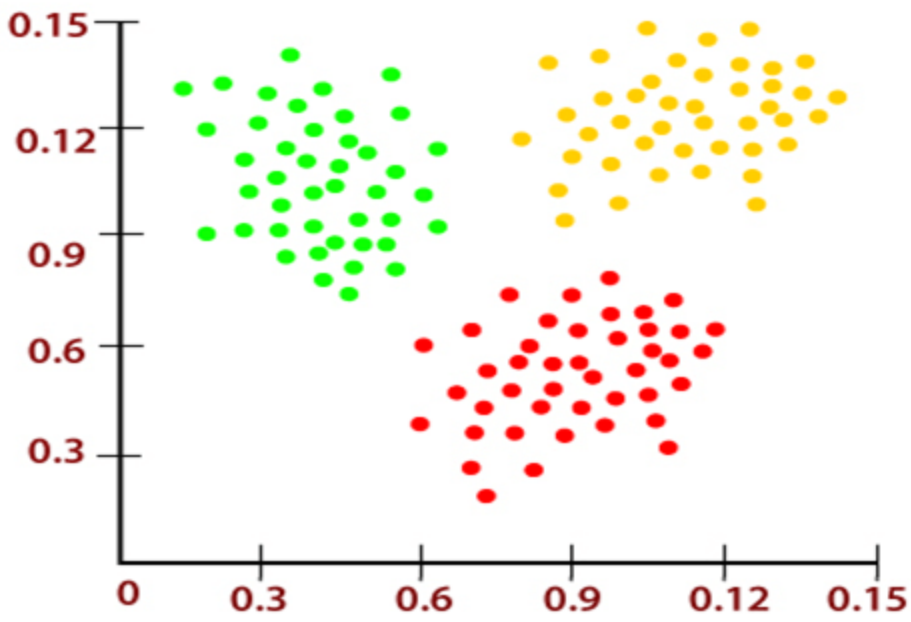
- Data Warehouse makes data more readable, hence, strategic questions can be easily answered using various graphs, trends, plots, etc.
- Data warehouse makes data analysis and operation faster and more accurate.

---

19) What do you understand by clustering?

Clustering is a way of dividing the data points into a number of groups such that data points within a group are more similar to each other than data points of other groups. These groups are called clusters, and hence, the similarities within the clusters is high, and similarities between the clusters is less.

The clustering techniques are used in various fields such as **machine learning, data mining, image analysis, pattern recognition, etc**.



Clustering is a type of supervised learning problems in machine learning. It can be divided into two types:

- **Hard Clustering**
- **Soft Clustering**

---

20) How to determine the number of clusters in k-means clustering algorithm?

In k-means clustering algorithm, the number of clusters depends on the value of k.

---

21) Differentiate between K-means clustering and hierarchical clustering?

The K-means clustering and Hierarchical Clustering both are the machine learning algorithms. Below are some main differences between both the clustering:

K-means clustering	Hierarchal Clustering
K-means clustering is a simple clustering algorithm in which objects are divided into clusters.	Hierarchal clustering shows the hierarchal or parent-child relationship between the clusters.
In k-means clustering, we need prior knowledge of k to define the number of clusters which sometimes may be difficult.	In hierarchal clustering, we don't need prior knowledge of the number of clusters, and we can choose as per our requirement.
K-means clustering can handle big data better than hierarchal clustering.	Hierarchal clustering cannot handle big data in a better way.
Time complexity of K-means is O(n) (Linear).	Time complexity of hierarchal clustering is $O(n^2)$(Clustering).

---

22) How is Data Science different from Data Analytics?

When we deal with data science, there are various other terms also which can be used as data science. Data Analytics is one of those terms. The data science and data analytics both deal with the data, but the difference is how they deal with it. So to clear the confusion between data science and data analytics, there are some differences given:

Data Science:

Data Science is a broad term which deals with structured, unstructured, and raw data. It includes everything related to data such as data analysis, data preparation, data cleansing, etc.

Data science is not focused on answering particular queries. Instead, it focuses on exploring a massive amount of data, sometimes in an unstructured way.

Data Analytics:

Data analytics is a process of analysis of raw data to draw conclusions and meaningful insights from the data. To draw insights from data, data analytics involves the application of algorithms and mechanical process.

Data analytics basically focus on inference which is a process of deriving conclusions from the observations.

Data Analytics mainly focuses on answering particular queries and also perform better when it is focused.

What are outliers and how to handle them?

Outliers are referred to the anomalies or slight variances in your data. It can happen during the data collection. There are 4 ways in which we can detect an outlier in the data set. These methods are as follows: Boxplot is a method of detecting an outlier where we segregate the data through their quartiles. A scatter plot displays the data of 2 variables in the form of a collection of points marked on the cartesian plane. The value of one variable represents the horizontal axis (x-axis) and the value of the other variable represents the vertical axis (y-axis). While calculating the Z-score, we look for the points that are far away from the center and consider them as outliers.

2) What type of big data problems Apache Spark can solve?

As we know that Apache Spark is an open-source big data framework. It provides an expressive APIs to facilitate big data professionals to execute streaming and batching efficiently. It is designed for fast computation and also provides a faster and more general data processing platform engine.

Apache Spark was developed at UC Berkeley in 2009 as an Apache project called "lighting fast cluster computing". It can distribute data in a file system across the cluster and processes that data in parallel.

Using Spark, we can write an application in Java, Python, Scala or R language.

---

3) What was the need for Apache Spark?

Many general-purpose cluster computing tools in the market, such as Hadoop MapReduce, Apache Storm, Apache Impala, Apache Giraph and many more. But each one has some limitations in its functionalities.

We can see the limitations as:

- Hadoop MapReduce can only allow for batch processing.
- If we talk about stream processing, then only Apache Storm / S4 can perform it.
- If we need interactive processing, then only Apache Impala / Apache Tezcan perform it.
- If we need to perform graph processing, then only Neo4j / Apache Giraph can do it.

Here, we can see that no single engine can perform all the tasks together. So, there was a requirement of a powerful engine that can process the data in real-time (streaming) and batch mode and respond to sub-second and perform in-memory processing.

This is how Apache Spark comes into existence. It is a powerful open-source engine that offers interactive processing, real-time stream processing, graph processing, in-memory processing and batch processing. It provides a very fast speed, ease of use, and a standard interface simultaneously.

---

4) Which limitations of MapReduce Apache Spark can remove?

Apache Spark was developed to overcome the limitations of the MapReduce cluster computing paradigm. Apache Spark saves things in memory, whereas MapReduce keeps shuffling things in and out of disk.

Following is a list of few things which are better in Apache Spark:

- Apache Spark keeps the cache data in memory, which is beneficial in iterative algorithms and can easily be used in machine learning.
- Apache Spark is easy to use as it knows how to operate on data. It supports SQL queries, streaming data as well as graph data processing.
- Spark doesn't need Hadoop to run. It can run on its own using other storages like Cassandra, S3, from which Spark can read and write.
- Apache Spark's speed is very high as it can run programs up to 100 times faster in-memory or ten times faster on disk than MapReduce.

5) Which languages Apache Spark supports? / Which are the languages supported by Apache Spark?

Apache Spark is written in Scala language. It provides an API in Scala, Python, Java, and R languages to interact with Spark.

---

6) What is the key difference between Apache Spark and MapReduce?

Following is the list of main differences between Apache Spark and MapReduce:

Comparison Parameter	Apache Spark	MapReduce
Data processing:	Apache Spark can process data in batches as well as in real-time.	MapReduce can process data in batches only.

Speed:	The processing speed of Apache Spark is extremely high. It runs almost 100 times faster than Hadoop MapReduce.	Hadoop MapReduce is slower than Apache Spark in the case of large scale data processing.
Data Storage:	Apache Spark stores data in the RAM, i.e., in-memory. It is easier to retrieve it, and that's why it is best to use in Artificial Intelligence.	Hadoop MapReduce stores data in HDFS. So, it takes a long time to retrieve the data from there.
Caching:	Apache Spark provides caching and in-memory data storage.	Hadoop MapReduce is highly disk-dependent.

---

7) What are the most important categories of the Apache Spark that comprise its ecosystem?

Following are the three important categories in Apache Spark that comprise its ecosystem:

- **Core Components:** Apache Spark supports five main core components. These are Spark Core, Spark SQL, Spark Streaming, Spark MLlib, and GraphX.
- **Cluster Management:** Apache Spark can be in the following three environments. These are the Standalone cluster, Apache Mesos, and YARN.
- **Language support:** We can integrate Apache Spark with some different languages to make applications and perform analytics. These languages are Java, Python, Scala, and R.

---

8) What is the difference between Apache Spark and Hadoop?

The key differences between Apache Spark and Hadoop are specified below:

- Apache Spark is designed to efficiently handle real-time data, whereas Hadoop is designed to efficiently handle batch processing.
- Apache Spark is a low latency computing and can process data interactively, whereas Hadoop is a high latency computing framework, which does not have an interactive mode.

Let's compare Hadoop and Spark-based on the following aspects:

Feature Criteria	Apache Spark	Hadoop

Speed:	Apache Spark is 100 times faster than Hadoop.	It is also very fast but not as much as Apache Spark.
Processing:	It is used for Real-time & Batch processing.	This is used for Batch processing only.
Learning Difficulty:	It is easy to learn because of high-level modules.	It is tough to learn.
Interactivity:	It has interactive modes.	It doesn't have interactive modes except for Pig & Hive.
Recovery:	Allows recovery of partitions	Fault-tolerant

---

9) What are some key features of Apache Spark?

Following is the list of some key features of Apache Spark:

Polyglot: Spark provides high-level APIs in Java, Scala, Python and R. We can write Spark code in any of these four languages. It provides a shell in Scala and Python. The Scala shell can be accessed through ./bin/spark-shell and Python shell through ./bin/pyspark from the installed directory.

Speed: Apache Spark provides an amazing speed upto 100 times faster than Hadoop MapReduce for large-scale data processing. We get this speed in Spark through controlled partitioning.

Multiple Formats: Apache Spark supports multiple data sources like Parquet, JSON, Hive and Cassandra. These data sources can be more than just simple pipes that convert data, pull it into Spark, and provide a pluggable mechanism to access structured data though Spark SQL.

Evaluation is lazy: Apache Spark doesn't evaluate itself until it is necessary. That's why it attains an amazing speed. Spark adds them to a DAG of computation for transformations, and they are executed only when the driver requests some data.

Real-Time Computation: The computation in Apache Spark is done in real-time and has less latency because of its in-memory computation. Spark provides massive scalability, and the Spark team has documented users of the system running production clusters with thousands of nodes and supports several computational models.

Hadoop Integration: Apache Spark is smoothly compatible with Hadoop. This is great for all the Big Data engineers who work with Hadoop. Spark is a potential replacement for the MapReduce functions of Hadoop, while Spark can run on top of an existing Hadoop cluster using YARN for resource scheduling.

Machine Learning: The MLlib of Apache Spark is used as a component of machine learning, which is very useful for big data processing. Using this, you don't need to use multiple tools, one for processing and one for machine learning. Apache Spark is great for data engineers and data scientists because it is a powerful, unified engine that is both fast and easy to use.

What is Scala?

Scala is a general-purpose programming language. It supports object-oriented, functional and imperative programming approaches. It is a strong static type language. In Scala, everything is an object whether it is a function or a number. It was designed by Martin Odersky in 2004.

Scala Program Example

```
1. object MainObject{
2.     def main(args:Array[String]){
3.         print("Hello Scala")
4.     }
5. }
```

---

2) What are the features of Scala?

There are following features in Scala:

- **Type inference:** In Scala, you don't require to mention data type and function return type explicitly.
- **Singleton object:** Scala uses a singleton object, which is essentially class with only one object in the source file.
- **Immutability:** Scala uses immutability concept. Immutable data helps to manage concurrency control which requires managing data.
- **Lazy computation:** In Scala, computation is lazy by default. You can declare a lazy variable by using the lazy keyword. It is used to increase performance.
- **Case classes and Pattern matching:** In Scala, case classes support pattern matching. So, you can write more logical code.
- **Concurrency control:** Scala provides a standard library which includes the actor model. You can write concurrency code by using the actor.
- **String interpolation:** In Scala, string interpolation allows users to embed variable references directly in processed string literals.
- **Higher order function:** In Scala, higher order function allows you to create function composition, lambda function or anonymous function, etc.
- **Traits:** A trait is like an interface with partial implementation. In Scala, the trait is a collection of abstract and non-abstract methods.
- **Rich set of collection:** Scala provides a rich set of collection library. It contains classes and traits to collect data. These collections can be mutable or immutable.

References:

1. https://www.javatpoint.com/data-science-interview-questions
2. https://www.javatpoint.com/apache-spark-interview-questions
3. https://www.javatpoint.com/scala-interview-questions