- **Study well what are the packages/Libraries/modules required for each program**
- **Matrix operation functions using NUMPY package**
- **Test Data & Training Data in Machine Learning**

Define Data Science.
The term “data science” combines two key elements: “data” and “science.”

1. **Data**: It refers to the raw information that is collected, stored, and processed. In today’s digital age, enormous amounts of data are generated from various sources such as sensors, social media, transactions, and more. This data can come in structured formats (e.g., databases) or unstructured formats (e.g., text, images, videos).
2. **Science**: It refers to the systematic study and investigation of phenomena using scientific methods and principles. Science involves forming hypotheses, conducting experiments, analyzing data, and drawing conclusions based on evidence.

Types of Data Visualization Techniques:

1. **Bar Charts:** Ideal for comparing categorical data or displaying frequencies, bar charts offer a clear visual representation of values.
2. **Line Charts:** Perfect for illustrating trends over time, line charts connect data points to reveal patterns and fluctuations.
3. **Pie Charts:** Efficient for displaying parts of a whole, pie charts offer a simple way to understand proportions and percentages.
4. **Scatter Plots:** Showcase relationships between two variables, identifying patterns and outliers through scattered data points.
5. **Histograms:** Depict the distribution of a continuous variable, providing insights into the underlying data patterns.

   Etc..

   - **Python Packages for Data Visualization – matplotlib, seaborn, etc..**

What is Pandas?

- Pandas is a Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis"

What are the Different Types of Data Structures in Pandas?
The two data structures that are supported by Pandas are **Series** and **DataFrames**.
Pandas Series is a one-dimensional labelled array that can hold data of any type. It is mostly used to represent a single column or row of data.
Pandas DataFrame is a two-dimensional heterogeneous data structure. It stores data in a tabular form. Its three main components are **data, rows,** and **columns**.

```
#To Find Confusion Matrix in Naïve Bayes use the below code
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:\n", cm)
```

What is the KNN Algorithm?

K-nearest neighbors algorithm (KNN)is a **supervised** learning algorithm that can be used to solve both classification and regression problem statements.

What is "K" in the K-nearest neighbor's algorithm?

K represents the number of nearest neighbours you want to select to predict the class of a given item, which is coming as an unseen dataset for the model.

Why is the odd value of "K" preferred over even values in the KNN Algorithm?

The odd value of K should be preferred over even values in order to ensure that there are no ties in the voting.

Why is the KNN Algorithm known as Lazy Learner?

The K-Nearest Neighbors (KNN) algorithm is called a *lazy learner* because it doesn't build an explicit model or perform any generalization during the training phase. Instead, it simply stores the training data and defers any computations until it receives a new input for classification or regression.

Why is it recommended not to use the KNN Algorithm for large datasets?

The Problem in processing the data:

KNN works well with smaller datasets because it is a lazy learner. It needs to store all the data and then make a decision only at run time. It includes the computation of distances for a given point with all other points. So if the dataset is large, there will be a lot of processing which may adversely impact the performance of the algorithm.

What is the difference between simple and multiple linear regression?

Simple linear regression models the relationship between one independent variable and one dependent variable, while multiple linear regression models the relationship between multiple independent variables and one dependent variable. The goal of both methods is to find a linear model that best fits the data and can be used to make predictions about the dependent variable based on the independent variables.

Aspect	**Simple Linear Regression**	**Multiple Linear Regression**
Definition	A statistical method for finding a linear relationship between two variables.	A statistical method for finding a linear relationship between more than two variables.
Number of independent variables	One.	More than one.
Number of dependent variables	One.	One.

Aspect	Simple Linear Regression	Multiple Linear Regression
Equation	$y = mx + b$	$y = b + m1x1 + m2x2 + ... + mnxn$
Purpose	Predict the value of the dependent variable based on the value of the independent variable.	Predict the value of the dependent variable based on the values of multiple independent variables.
Assumption	Assumes a linear relationship between the independent and dependent variables.	Assumes a linear relationship between the dependent variable and multiple independent variables.
Method	Uses a simple linear regression equation to estimate the regression line.	Uses multiple linear regression equations to estimate the regression plane or hyperplane.
Complexity	Less complex.	More complex.
Interpretation	Easy to interpret.	Complex to interpret.
Data requirement	Requires fewer data.	Requires more data.
Examples	Predicting the price of a house based on its size.	Predicting the performance of a student based on their age, gender, IQ, etc.