NOTE: DO Comments in coding questions .

Assignment 1: Study of Models

Q1.) Differentiate between Supervised & unsupervised learning.

Q2.) Define numpy array and how is it different from pandas Dataframe & pandas Series? Which python data structure does pandas dataframe represent?

Q3.) What are the significance of box plots, violin plots, histogram plots and scatter plots in data visualization?

Q4.) what will be output on screen (dont run the code do it with pen+paper)

(a)

```
def f(n):
    if n == 1:
        return "A"
    if n == 2:
        return "B"
    return f(n-1) + f(n-2)

print(f(5))
```

(b)

```
def show(n):
    if n == 0:
        return
    print(n, end="-")
    show(n-1)
    print(n, end="-")

show(3)
```

Q5.) What is hypothesis testing? When do we use t-distribution, F-distribution and X square distribution? (search youtube lectures or NPTEL for this ,you may study them in core AI/DS course)

Q6.)

The Space Farm Productivity Crisis

A Mars agriculture colony reports that their hydroponic farms are underperforming.

Tasks: Load the given dataset (`farm_metrics.csv`) using Pandas.

Compute Mean, median, and variance of plant height & Correlation between nutrient levels and yield.

Use NumPy to:
Generate **simulated yield values** from a normal distribution.
Compute matrix transpose + inverse of the nutrient interaction matrix.

Plot:
Scatter plot of **nutrient_level vs yield**.
Histogram of simulated yields.

Q7.)

The Lost Treasure Ship Analysis

An ocean research team scanned the ocean floor and received 3D coordinates of objects.They believe one cluster is a lost treasure ship.

Tasks: Read coordinates (`sonar_points.csv`) using Pandas.
Use NumPy to:
Calculate distances of each point from origin.
Normalize the coordinates (mean 0, variance 1).
Compute covariance matrix of points.

Plot:
2D scatter plot (x vs y).
Histogram of depth (z).

Use Pandas to find the top 5 deepest points.

Q8.)The Chocolate Factory Quality Control

A chocolate factory tracks sugar %, cocoa %, and weight of each chocolate bar.**Tasks:**Load (`choco_quality.csv`) and

Using Pandas:
Remove missing values.
Describe statistics.

Using NumPy:
Compute a "quality score" = 0.3*sugar + 0.7*cocoa.
Generate random noise using `np.random.normal`.

Plot:
Scatter plot **cocoa% vs sugar%**.
Histogram of weight distribution.

Q9.) The interstellar research team receives two encoded signals:

> **Signal A**: Generated using a **sinusoidal wave** sampled at random points
> **Signal B**: Generated using a **cosine wave** sampled at random points

Each incoming signal is reshaped into a **4×6 matrix**.Your task is to help decode the signals.Generate the signals of matrix 4X6 using sine and cosine of random numbers ,generated in range of 0 to 2pie.

Compute the dot product of the two matrices (A · B).
Ensure shapes are aligned properly (you may need transpose).

Compute the **cross product row-wise**, i.e., for each pair of corresponding rows in A and B, compute: cross(A[i],B[i])

Q10.) Brain round 🙂 just for brainstorming.

<u>Poisonous Wine</u>

So there's this king. Someone breaks into his wine cellar where he stores 1000 bottles of wine. This person proceeds to poison one of the 1000 bottles, but gets away too quickly for the king's guard to see which one he poisoned or to catch him.

The king needs the remaining 999 safe bottles for his party in 4 weeks. The king has 10 prisoners who deserve execution. The poison takes about 3 weeks to take effect, and any amount of it will kill whoever drinks it. How can he figure out which bottle was poisoned in time for the party?

You have a train to catch!

Spiderman has two close friends, Mary Jane & Gwen Stacy. After every mission, he rushes to the central subway. One line heads towards Mary's place, and another towards Stacy. Trains from each line leave every 10 minutes. Spiderman being impartial always boards the first train that leaves.

However, he observes that he ends up visiting Mary Jane nine times more often than Gwen Stacy. Can you decipher why?

Assignment 2: Study of Models

You are hired as Data analytics role in Real Estate firm in USA. You are given California housing data with columns:

1. longitude: A measure of how far west a house is; a higher value is farther west

2. latitude: A measure of how far north a house is; a higher value is farther north

3. housingMedianAge: Median age of a house within a block; a lower number is a newer building

4. totalRooms: Total number of rooms within a block

5. totalBedrooms: Total number of bedrooms within a block

6. population: Total number of people residing within a block

7. households: Total number of households, a group of people residing within a home unit, for a block

8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

9. medianHouseValue: Median house value for households within a block (measured in US Dollars)

10. oceanProximity: Location of the house w.r.t ocean/sea

Your tasks are to analys the data thoroughly and create Multiple linear regression model to predict housing prices of locality.

- Load the **California Housing dataset** from sklearn.datasets.Use pandas library.Deliverables:
  Shape of dataset
  Column names
  First 10 rows

- Use .describe() , .info() for understanding data.Which feature has largest variance?

Univariate Analysis (Histograms)

- Plot **histograms for all numeric columns with proper labelling of axis.**
  Try:
  different bins
  Explain most widely used methods to eliminate skewness of column features in comments.
- Use Box plot for each feature to detect outliers in MedInc AveRooms,Population
- **Correlation Heatmap:**Compute **correlation matrix** ,Plot a **heatmap**

Latitude/Longitude Visualization

- Scatter plot:
  Longitude vs Latitude
  Color (cmap) by **MedHouseVal**
  Point size based on **Population**

Explain why scaling is required before PCA.

- Apply PCA on features
  Plot explained variance ratio
  Choose top 2 principal components

  Scatter plot:PC1 vs PC2
  Color points by **MedHouseVal**

- **Create a scikit-learn Pipeline for multiple linear regression and print coefficients- intercept matrix (exclude latitude and longitude in this final model becoz that is not relevant data : use correlation matrix to prove this point also)**
- Evaluate model on metrics of MSE loss, MAE loss, R2 score , adjusted R2 score.

Is high R2 score always good? Is low training loss always preferred?

- Plot:Predicted vs Actual
  Residuals vs Predicted values

BONUS TASKS : Train Ridge & Lasso Regression and explain them with diagrams.