Hello there :)

I am Emanuele and I am a biomedical engineer who got curious and just simply wanted to learn and break down into the ML and AI field. This is my first real GitHub project ever, feel free to help me out!

I will guide you through this, while I grasp new concepts and apply them directly on the python script.

I will try to cover as much as possible to grasp all the basics and more of the machine learning.

I will answer to 100 questions, covering these 15 macro-topics:

1. Dataset Basics
2. Data Exploration
3. Distribution Analysis
4. Relationship Analysis
5. Pattern Recognition
6. Data Preprocessing
7. Model Selection
8. Algorithm Understanding
9. Training Process
10. Performance Metrics
11. Model Comparison
12. Error Analysis
13. Feature Importance
14. Cross-Validation
15. Overfitting and Generalization
16. *Critical Thinking Questions*

That being said, lets start this journey tog “)

--------------------------------------------------------------------------------------------------------------------------------

**TASK num. 1: Dataset Basics**

1. What is the Iris dataset and why is it famous in machine learning?

2. How many samples, features, and classes are in the dataset?

3. What do the four features measure? (sepal length, sepal width, petal length, petal width)

4. What are the three species of iris flowers in the dataset?

5. Is the dataset balanced? (equal number of samples per class)

6. Are there any missing values or data quality issues?

The Iris dataset is a well-known and widely used dataset in the field of machine learning. It consists of 150 samples of iris flowers from three different species: Setosa, Versicolor, and Virginica. Each sample includes four features: sepal length, sepal width, petal length, and petal width.

The Iris dataset is often used as a beginner's dataset to understand classification and clustering algorithms in machine learning. By using the features of the iris flowers, researchers and data scientists can classify each sample into one of the three species. The dataset plays a crucial role in machine learning as a standard benchmark for testing classification algorithms. The dataset is balanced because we have an equal number of samples per class. There aren’t any missing values or issues.

**TASK num. 2: Data Exploration**

7. What are the mean, median, and standard deviation for each feature?

8. Which feature has the highest variability across all samples?

9. Which species tends to have the largest/smallest measurements for each feature?

10. Are there any obvious outliers in the data?

The data exploration is the process of understanding the characteristics of datasets and uncover patterns and relationships between the features and more. The mean is calculated by summing all the values in a dataset and dividing by the number of values. The median, on the other hand, is the middle value when the data is sorted in ascending order. It basically provides a better measure of central tendency for skewed data.

A math equations on a white background

AI-generated content may be incorrect.

To determine which feature has the highest variability across all samples, we can refer to the statistical summary of the dataset. According to the statistical summary, the feature with the highest variability is petal length, as it has the highest standard deviation among the four features.

Outliers in the Iris dataset analysis refer to data points that significantly deviate from the rest of the observations. In our case the dataset is well-balanced, therefore there aren’t any outliers.

**TASK num. 3: Distribution Analysis**

11. How are the values distributed for each feature? (normal, skewed, bimodal?)

12. Which features show clear separation between species in box plots?

13. Do any features appear to follow a normal distribution?

14. Which species has the most/least variation in measurements?

The distribution of values for each feature in the Iris dataset can be characterized in terms of **normality**, **skewness**, and **modality** (e.g., unimodal, bimodal). Based on exploratory data analysis and visual inspection (e.g., histograms, density plots, and Q-Q plots), here's a general breakdown of the distribution of each feature:

|  |  |  |
| --- | --- | --- |
| Feature | Distribution Type | Notes |
| Sepal Length | Near normal | Slightly right skewed |
| Sepal Width | Slightly skewed | May show mild bimodality |
| Petal Length | Bimodal | Clear separation between species |
| Petal Width | Bimodal | Distinct groups based on species (setosa vs others) |

Visual Tools Used:

* Histograms and density plots to assess modality.
* Boxplots to detect skewness and outliers.
* Q-Q plots to evaluate normality.

In box plots of the Iris dataset, some features show clear separation between species, while others exhibit overlap.

X = petal length, petal width, sepal length, sepal width # features

Y = [“setosa”, “versicolor”, “verginica”] # species = classi