How to build a ML model

Following are the steps to Build a ML model:-

Step1:- Data Ingestion

Data Ingestion is the process of importing library and dataset into our current program.

e.g.:-

```
import numpy as np
import pandas as pd
data = pd.read_csv(r"dataset1")
```

Step2:- EDA with visualization:-

EDA stands for Exploratory Data Analysis and it is defined as the process of understanding of dataset or get the insights of our dataset.

It includes following steps:-

Data Inspection: Review and understand the dataset by examining its structure, dimensions, and basic statistics.

--->

Handling Missing Values and duplicate values: Identify and address any missing data and duplicate points in the dataset.

--->.shape, .columns, .info(), .tail(), .head(), .unique().

Data Visualization: Create visual representations (plots, graphs) to explore relationships, distributions, and patterns in the data.

---> histograms, line-plot, bar, scatter, crosstab, etc

Feature Analysis: Assess individual features' characteristics and relationships with the target variable.

----> .describe(), .dtypes

Outlier Detection: Identify and handle outliers that may affect the model's performance.

---> Plotting Box plot

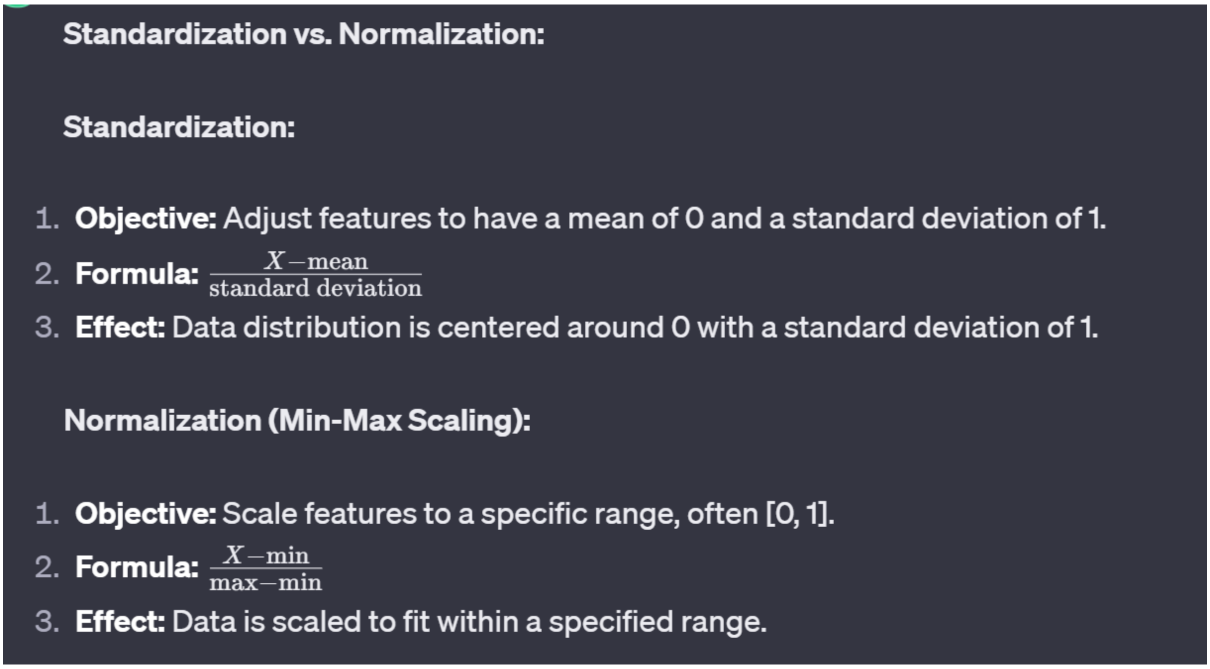
Correlation Analysis: Examine the correlation between features to understand interdependencies.

---> .corr

Standardization: Standardization is the process of rescaling features to have a mean of 0 and a standard deviation of 1.

This is particularly useful when features have different scales, as it ensures that each feature contributes equally to the model.

It's crucial to standardize training and testing sets separately to avoid data leakage. And done before training the model.



Standardization vs. Normalization:

Standardization:

1. **Objective:** Adjust features to have a mean of 0 and a standard deviation of 1.
2. **Formula:** $\frac{X - \text{mean}}{\text{standard deviation}}$
3. **Effect:** Data distribution is centered around 0 with a standard deviation of 1.

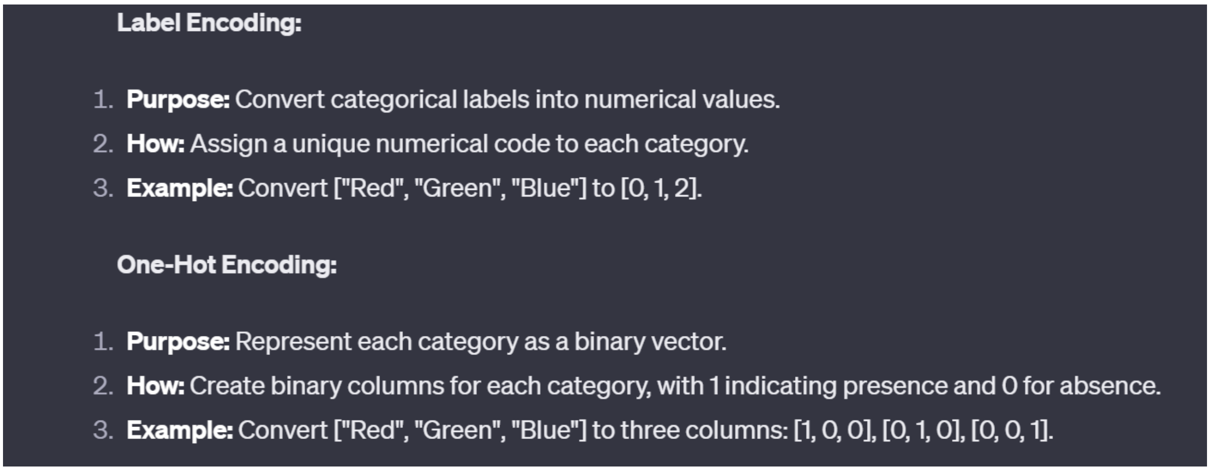
Normalization (Min-Max Scaling):

1. **Objective:** Scale features to a specific range, often [0, 1].
2. **Formula:** $\frac{X - \text{min}}{\text{max} - \text{min}}$
3. **Effect:** Data is scaled to fit within a specified range.

--->

Encoding: Encoding is the process of converting categorical data (text or labels) into a numerical format.

Common methods include Label Encoding and One-Hot Encoding.



Label Encoding:

1. **Purpose:** Convert categorical labels into numerical values.
2. **How:** Assign a unique numerical code to each category.
3. **Example:** Convert ["Red", "Green", "Blue"] to [0, 1, 2].

One-Hot Encoding:

1. **Purpose:** Represent each category as a binary vector.
2. **How:** Create binary columns for each category, with 1 indicating presence and 0 for absence.
3. **Example:** Convert ["Red", "Green", "Blue"] to three columns: [1, 0, 0], [0, 1, 0], [0, 0, 1].

Step3 Splitting of DataSet:-

Divide the data into training and testing sets.

Use functions like train_test_split of scikit-learn to randomly allocate data, e.g., 80% for training, 20% for testing.

Step4 Selecting a Model:-

1. **Understand Problem Type:** Identify whether your problem is of type classification or regression.

2. **Choose Model Type:** Select a suitable algorithm based on your problem type, such as

For Classification --> logistic regression, SVM, KNN, decision trees , random forest, etc.

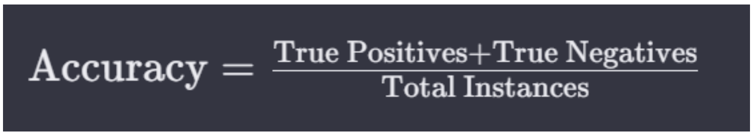
For Regression --> Linear Regression, Ridge Regression, Lasso Regression, etc

3. Model Fit: Model fit is the training phase where a model learns patterns from the provided data, adjusting its parameters to make accurate predictions.

4. **Evaluate Models:** Train multiple models and compare their performance to choose the best-performing model using metrics like :--

Confusion Matrix: A table that summarizes the performance of a classification algorithm by showing the counts of true positive, true negative, false positive, and false negative predictions.

Accuracy Metrics: A metric that measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. It is given by:



$$\text{Accuracy} = \frac{\text{True Positives}+\text{True Negatives}}{\text{Total Instances}}$$

Gini Score: A metric used in decision tree algorithms to evaluate the impurity of a set of categorical data. It ranges from 0 to 1, where 0 represents perfect purity and 1 represents maximum impurity.

Step5 Monitor and Update: Continuously assess model performance and update as needed for changing conditions or new data.