MODULE 3

The Machine Learning Pipeline Architecture

Data Preparation & Preprocessing Strategy

PREPROCESSING WORKFLOW

Segregation

Separating the independent features (X) from the dependent target (y).

Integrity

Splitting data to prevent information leakage between training and testing sets.

Scaling

Normalizing numerical ranges to ensure equal weight during optimization.

FEATURE & TARGET SEPARATION

Feature Matrix (X)

The input variables. Usually a 2D matrix of shape (n_samples, n_features).

- Predictive measurements
- Explanatory variables

Target Vector (y)

The labels or values to predict. Usually a 1D vector of shape (n_samples,).

- Classification labels
- Regression continuous values

"X is what we know; y is what we want to find out."

THE SILENT MODEL KILLER

0

ALLOWED LEAKS

What is Data Leakage?

Leakage occurs when information from the testing set "leaks" into the training process. This leads to **overfitting** and unrealistic performance metrics.

> **Golden Rule:** Never fit your scaler or model on the test data. The test data should remain completely unseen until the very end.

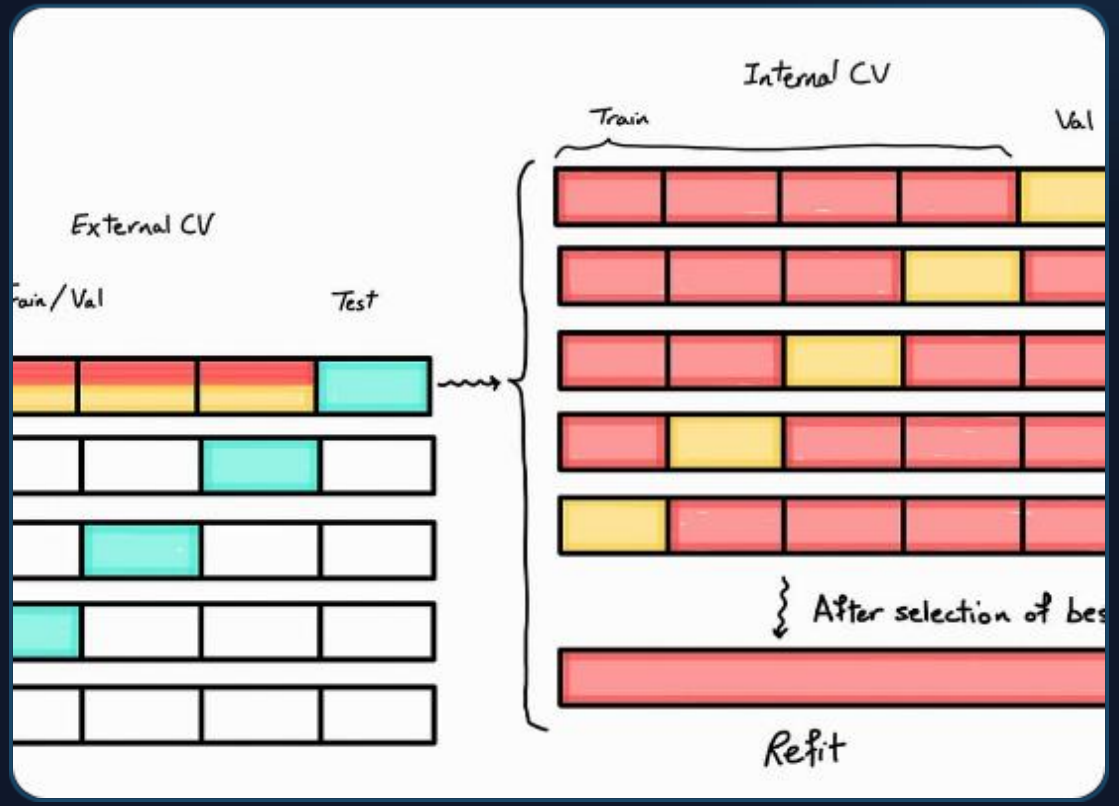
THE SPLIT STRATEGY

Training vs Testing

To evaluate a model fairly, we split the data matrix into two separate pools:

- **Training Set (80%):** Used by the algorithm to learn patterns.
- **Testing Set (20%):** Used as a proxy for "real world" unseen data.

Common split ratios range from 70/30 to 80/20 depending on dataset size.



IMPLEMENTING THE SPLIT

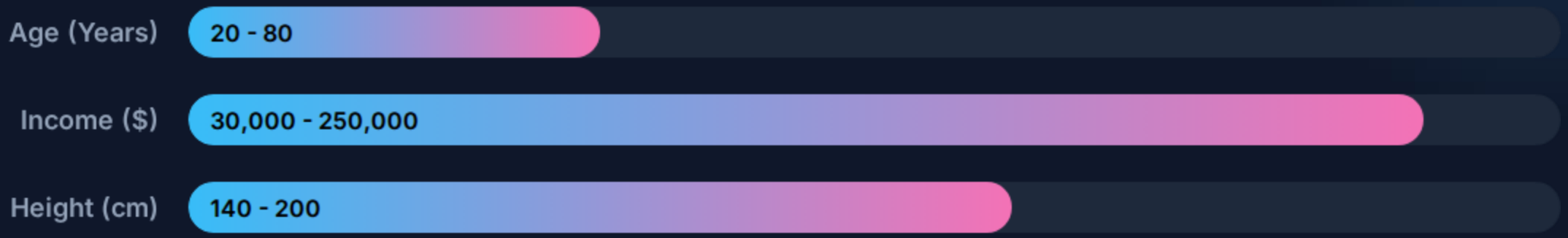

```
from sklearn.model_selection import train_test_split

# Perform the split
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42
)

# Output check
print(X_train.shape) # (120, 4)
print(X_test.shape)  # (30, 4)
```

Key Parameters

test_size: Fraction of data for the test set.

random_state: Seed for reproducibility.

WHY SCALE FEATURES?



Many algorithms (SVM, KNN, Neural Nets) use **Euclidean Distance**. If one feature has a massive range, it will dominate the distance calculation, making other features irrelevant.

STANDARDIZATION (Z-SCORE)

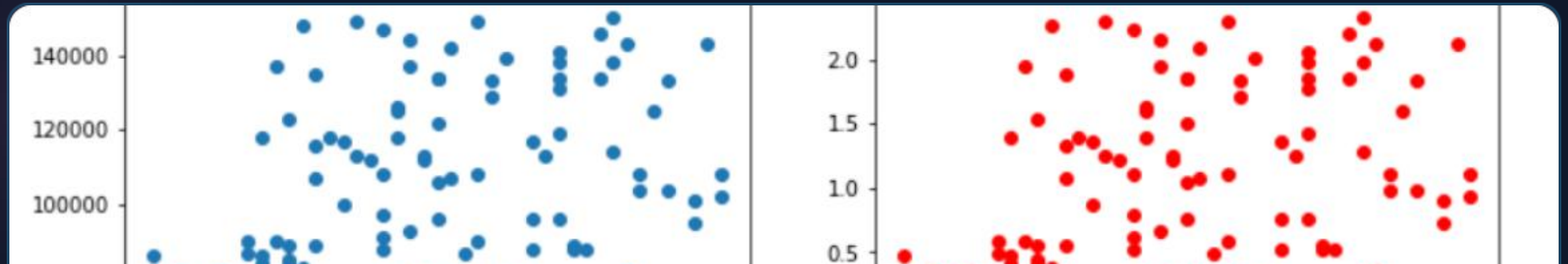
$$z = \frac{x - \mu}{\sigma}$$

Where μ = mean and σ = standard deviation

StandardScaler()

Transforms data to have a **Mean of 0** and a **Standard Deviation of 1.**

This centers the data around the origin and normalizes the spread of the data points.



THE FIT_TRANSFORM LOGIC

```
fro  sklearn.preprocessing impor  StandardScaler
m                          t
scaler = StandardScale ()
         r
# 1. Fit & Transform Training
X_train_scaled = scaler.fit_transfor (X_train)
Data                    m
# 2. Transform Test Data
X_test_scaled = scaler.transfor (X_test)
(ONLY)                  m
```

.fit_transform()

Calculates parameters (mean, std) AND applies transformation. Used on **Train** set.

.transform()

Applies previously learned parameters to new data. Used on **Test** set.

PIPELINE EXECUTION ORDER

Segregate

Drop y from X

Split

train_test_split()

Fit Scaler

Learn Train Stats

Scale Test

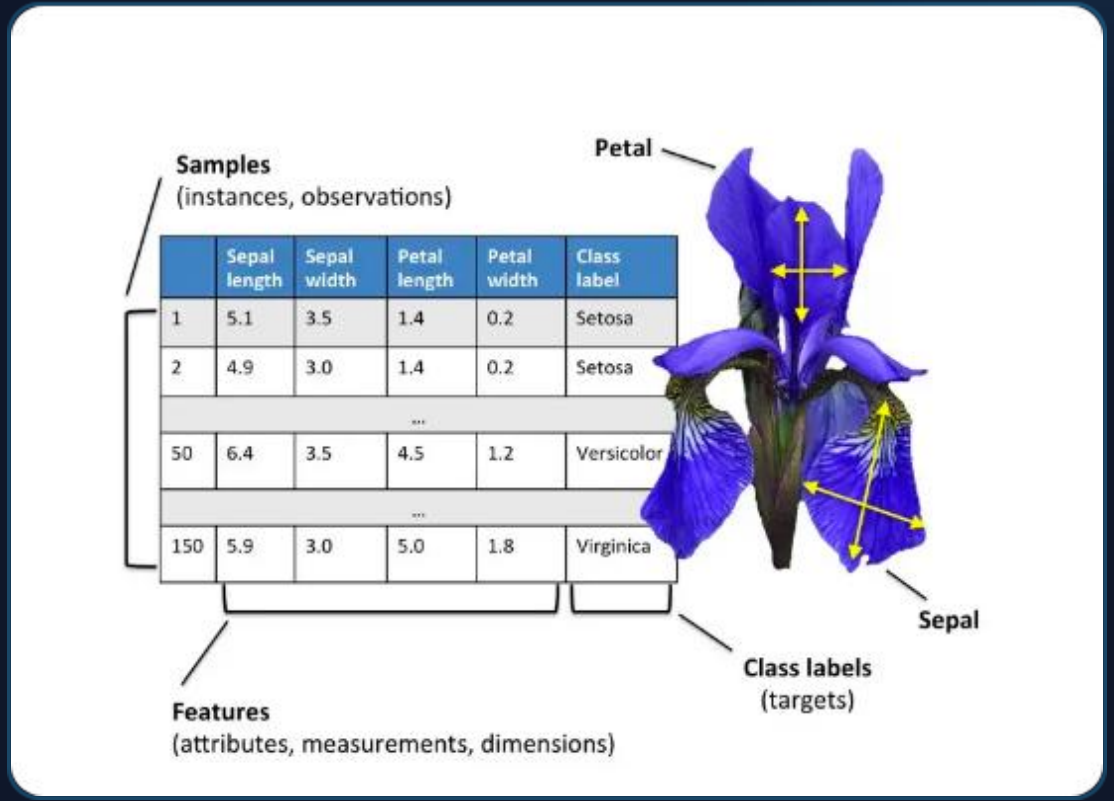
Transform Unseen

⚠ **Warning: Never swap Split and Fit! Scaler must not know Test stats.**

CASE STUDY: IRIS DATASET

The **Iris Dataset** is the standard "hello world" of ML. It contains measurements of 150 flowers.

Sample	Sepal L	Sepal W	Target
#1	5.1	3.5	Setosa
#2	4.9	3.0	Setosa
...	...	...	...
Scaled	-0.9	1.03	Setosa



MODULE DELIVERABLES

- **X_train / X_test:** Isolated feature matrices split at 80/20 or similar ratio.
- **y_train / y_test:** Corresponding target vectors for supervised training.
- **Standardized Arrays:** All numerical features transformed into NumPy arrays with 0 mean and 1 variance.
- **Leak-Free Preprocessing:** Verification that scaling parameters were derived *only* from the training subset.

Data is now ready for Mathematical Modeling.

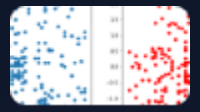
IMAGE SOURCES

https://miro.medium.com/v2/resize:fit:1400/1*NGEIpxYtvZJhqFz8hFGICg.jpeg

Source: medium.com

https://images.stockcake.com/public/7/c/9/7c908b17-8c83-4675-a270-3efd9c3f4ded_medium/code-on-screen-stockcake.jpg

Source: stockcake.com



https://editor.analyticsvidhya.com/uploads/157194.png

Source: www.analyticsvidhya.com

https://miro.medium.com/1*H2UmG5L1l5bzFCW006N5Ag.png

Source: eminebozkus.medium.com