

# Introduction to Machine Learning with Python

David Schaupp | WS2025

# Content

## Introduction to Machine Learning:

- What is Machine Learning?
- Machine Learning Project Checklist

## Intro's:

- Colab
- Python
- Numpy
- Matplotlib

## Supervised Learning:

- Classification (Binary|Multiclass)
- Regression
- Support Vector Machines
- Decision Trees (Random Forest)

## Unsupervised Learning:

- Dimensionality Reduction
- Clustering (k-means, DBSCAN)

## Performances Measures:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC Curve
- Confusion Matrix



# Supervised Learning - Classification

## Classification:

- Binary
- Multiclass

## MNIST - Dataset:

- "Hello world" of ML
- Handwritten digits
- 70.000 images
- 28x28 pixels
- 0-9 digits



# Content

## Introduction to Machine Learning:

- What is Machine Learning?
- Machine Learning Project Checklist

## Intro's:

- Colab
- Python
- Numpy
- Matplotlib

## Supervised Learning:

- Classification (Binary|Multiclass)
- Regression
- Support Vector Machines
- Decision Trees (Random Forest)

## Unsupervised Learning:

- Dimensionality Reduction
- Clustering (k-means, DBSCAN)

## Performances Measures:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC Curve
- Confusion Matrix



# Performance Measures

Before we dive into formulas, let's understand *why* we need to measure performance.



## Model Selection

Choose the best algorithm for your specific problem.



## Optimization

Fine-tune your model's parameters for optimal results.



## Identifying Issues

Detect problems like overfitting and understand model weaknesses.

💡 \*\*Key Idea:\*\* Without metrics, we are flying blind. We have no objective way to know if our model is good, bad, or just lucky.



# Performance Measures

## Underfitting:

- The model is too simple to capture the underlying pattern in the data.
- Symptom:
  - Poor performance on both training and test data.
- Analogy:
  - Trying to fit a straight line to a wave. It misses the complexity entirely.

## Overfitting:

- The model learns the training data too well, including noise and random fluctuations.
- Symptom:
  - Great performance on training data, but poor performance on new, unseen data.
- Analogy:
  - Memorizing the answers to a test instead of learning the concepts.



# Performance Measures - CV

## Accuracy using Cross Validation:

- How do we get a reliable estimate of our model's performance on unseen data?
- Problem:
  - If we are "lucky" or "unlucky" with our split, our performance estimate might be misleadingly high or low.

## Different Variants of CV:

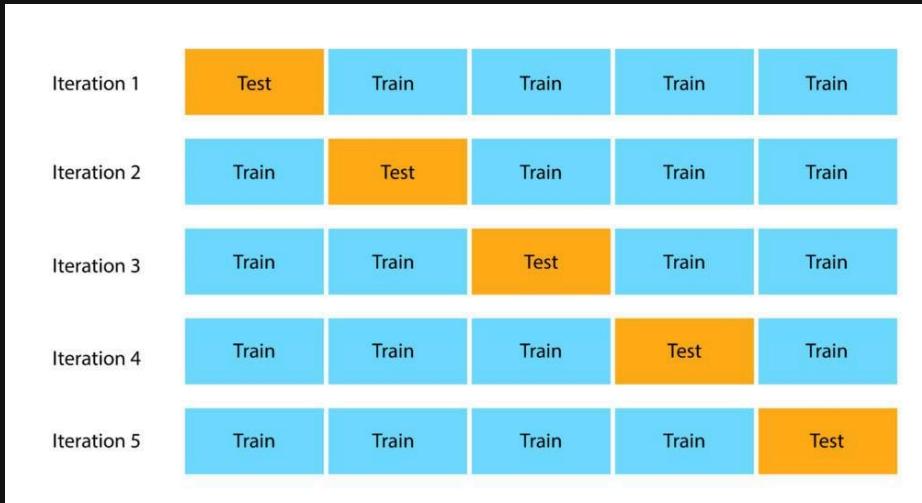
- Reapeated random sub-sampling validation
- **k-fold cross-validation**
- Leave-one-out cross-validation
- kx2 cross validation



# Performance Measures - CV

## Steps in k-fold CV:

- Shuffle the dataset randomly.
- Split the dataset into k groups
- For each unique group:
  - Take the group as a hold out or test data set
  - Take the remaining groups as a training data set
  - Fit a model on the training set and evaluate it on the test set
  - Retain the evaluation score and discard the model
- Summarize the skill of the model



# Performance Measures - CV

Working Example:

Dataset:

[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]

Choose k = 3 (most common k=5):

- Shuffle the data and split into 3 folds
- Fold 1: [0.5, 0.2]
- Fold 2: [0.1, 0.3]
- Fold 3: [0.4, 0.6]

For example:

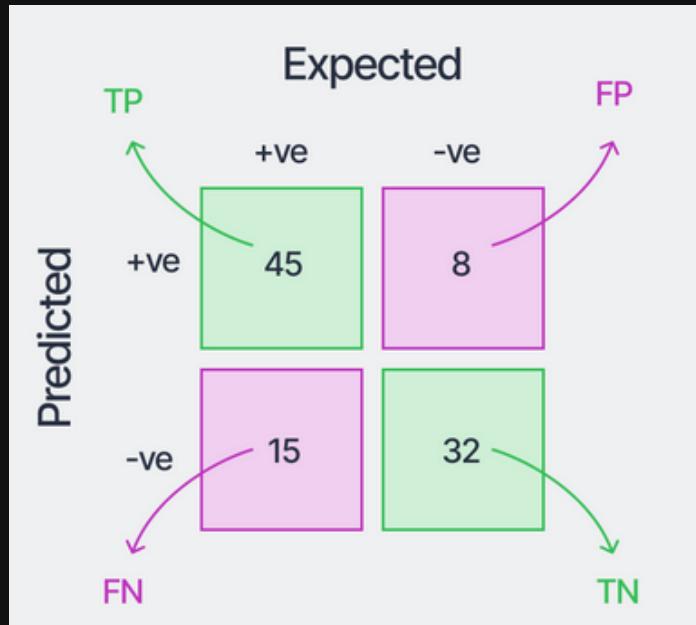
- Model1: Trained on Fold1 + Fold2, Tested on Fold3
- Model2: Trained on Fold2 + Fold3, Tested on Fold1
- Model3: Trained on Fold1 + Fold3, Tested on Fold2



# Performance Measures - CM

## Confusion Matrix:

- The Confusion Matrix is the foundation for most other classification metrics.
- It tells us where our model is getting confused.
- Tells us where the model gets "confused".
- The matrix is NxN, where N is the number of target values (classes).



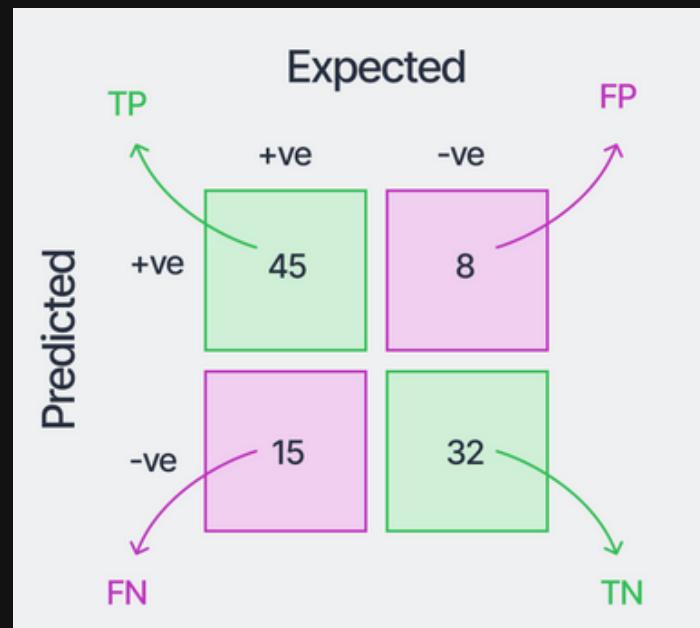
# Performance Measures - CM

## Anatomy of the Matrix:

- For a binary problem (e.g., "Disease" vs. "Healthy"), it's a 2x2 table
- True Positives (TP):
  - correctly predicted "Disease"
- True Negatives (TN):
  - correctly predicted "Healthy"
- False Positives (FP):
  - incorrectly predicted "Disease" (Type I Error)
- False Negatives (FN):
  - incorrectly predicted "Healthy" (Type II Error)

## Analogy:

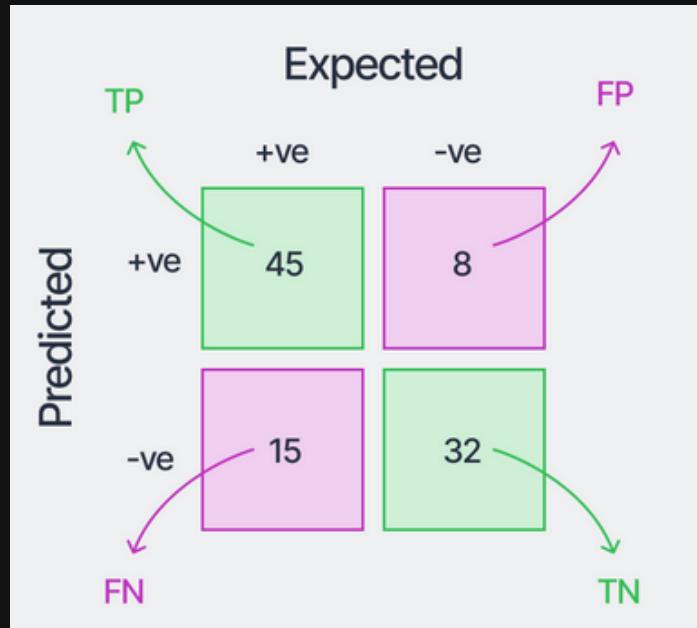
- Imagine a medical screening test.
- The confusion matrix shows you:
- How many sick people were correctly identified ( $TP = 45$ ).
  - How many healthy people were correctly identified ( $TN = 32$ ).
  - How many healthy people were wrongly told they are sick ( $FP = 8$ ).
  - How many sick people were wrongly told they are healthy ( $FN = 15$ ).



# Performance Measures - CM & Accuracy

## Accuracy Definition:

- The percentage of total predictions that were correct.



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{45 + 32}{45 + 32 + 8 + 15} = \frac{77}{100} = 0.77$$



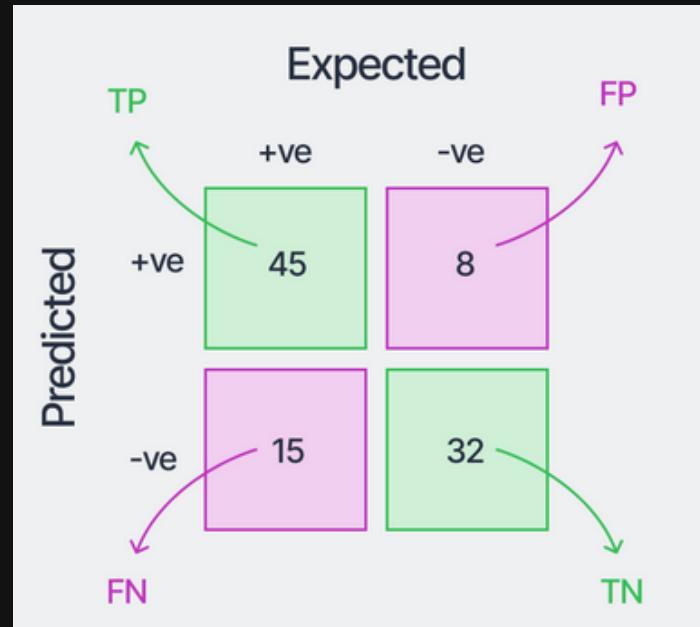
# Performance Measures - CM & Precision

## Precision Definition:

- Of all the times the model predicted "Positive", how many were actually correct?
- Question: "Are we sending innocent people to jail?"
- Focus: Minimizing False Positives (FP)

## Insight:

- Precision measures reliability of positive predictions.
- High precision reduces false alarms and unnecessary interventions.



$$\text{Precision} = \frac{TP}{TP + FP} = \frac{45}{45 + 8} = \frac{45}{53} \approx 0.849$$



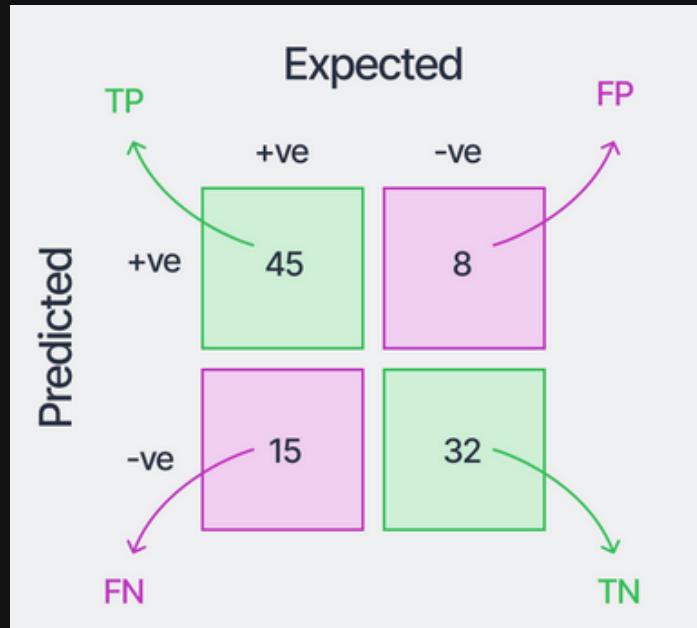
# Performance Measures - CM & Recall

## Recall Definition:

- Of all the actual "Positives" that exist, how many did the model find?
- Question: "Are we letting guilty people go free?"
- Focus: Minimizing False Negatives (FN)

## Insight:

- Recall measures completeness of positive detection.
- High recall means fewer missed true positives.



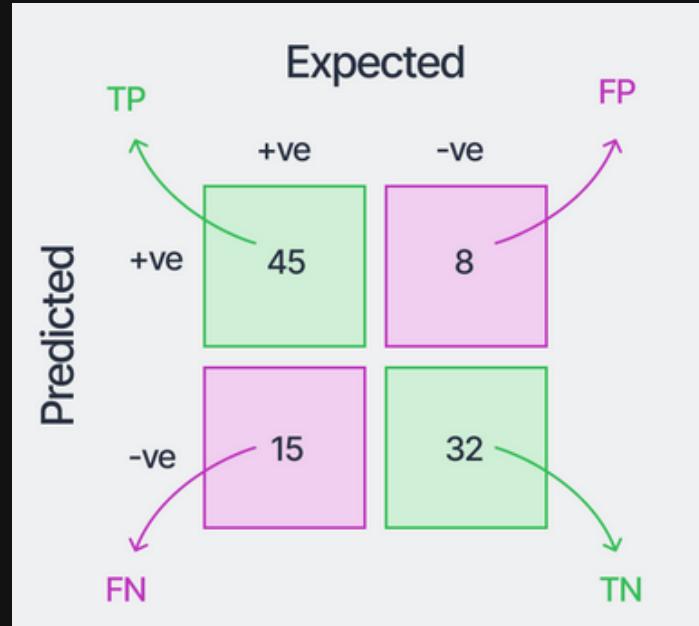
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{45}{45 + 15} = \frac{45}{60} = 0.75$$



# Performance Measures - CM & F1-Score

## F1-Score Definition:

- is the harmonic mean of precision and recall



$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.849 \times 0.75}{0.849 + 0.75} \approx 0.796$$



# Performance Measures - ROC/AUROC

What is it?

- ROC curve:  
A graph showing how well a binary classifier performs.
- AUC:  
A single number summarizing the ROC performance.

Game Show Analogy:

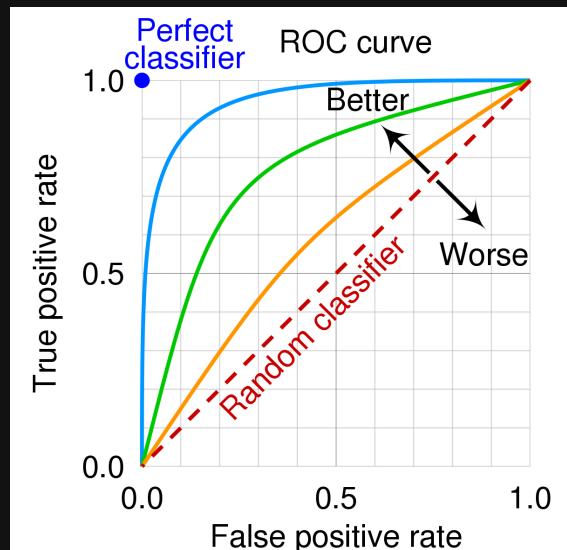
- Imagine a game show where you must decide if a contestant is a "Friend" (positive) or "Foe" (negative).
- True Positive Rate (Recall): how many "Friends" you correctly identify.
- False Positive Rate: how many "Foes" you incorrectly label as "Friends"
- The ROC curve tracks performance as you adjust suspicion (threshold).

AUC - The Overall Score:

- AUC = 1.0 → Perfect! You distinguish all Friends from Foes.
- AUC = 0.5 → Random guessing, like flipping a coin.
- AUC > 0.8 → Generally considered good performance.

Interpretation:

- AUC ranges from 0 to 1.
- Higher AUC = better model at separating positive and negative classes.
- Works well even when classes are imbalanced (e.g., rare diseases).



# Quick Reference Guide

Metric	Question it Answers	Formula	When to Use
Accuracy	Overall, how often is the model correct?	$(TP+TN)/Total$	When classes are balanced and all errors are equal.
Precision	When it predicts positive, how often is it right?	$TP/(TP+FP)$	When the cost of a False Positive is high.
Recall	Of all actual positives, how many did it find?	$TP/(TP+FN)$	When the cost of a False Negative is high.
F1-Score	What's the balance between Precision and Recall?	$2*(P*R)/(P+R)$	When you need a balance and classes are imbalanced.
AUC	How good is the model at separating classes?	$\sqrt{ }$	To compare models across all thresholds.



# Interactive Scenario: What Would You Do?

You are building a model for a bank to detect fraudulent credit card transactions.

The Situation:

- Only 0.1% of transactions are fraudulent (highly imbalanced).
- Cost of a False Negative (missing a fraud):  
The bank loses money, and the customer is unhappy. Let's say this costs €1000.
- Cost of a False Positive (flagging a valid transaction):  
The customer is inconvenienced by a blocked card and has to call the bank. Let's say this costs €10.

Which metric would you prioritize optimizing for? Why?



Accuracy



Precision



Recall

Answer: Recall



# Let's get our hands dirty

