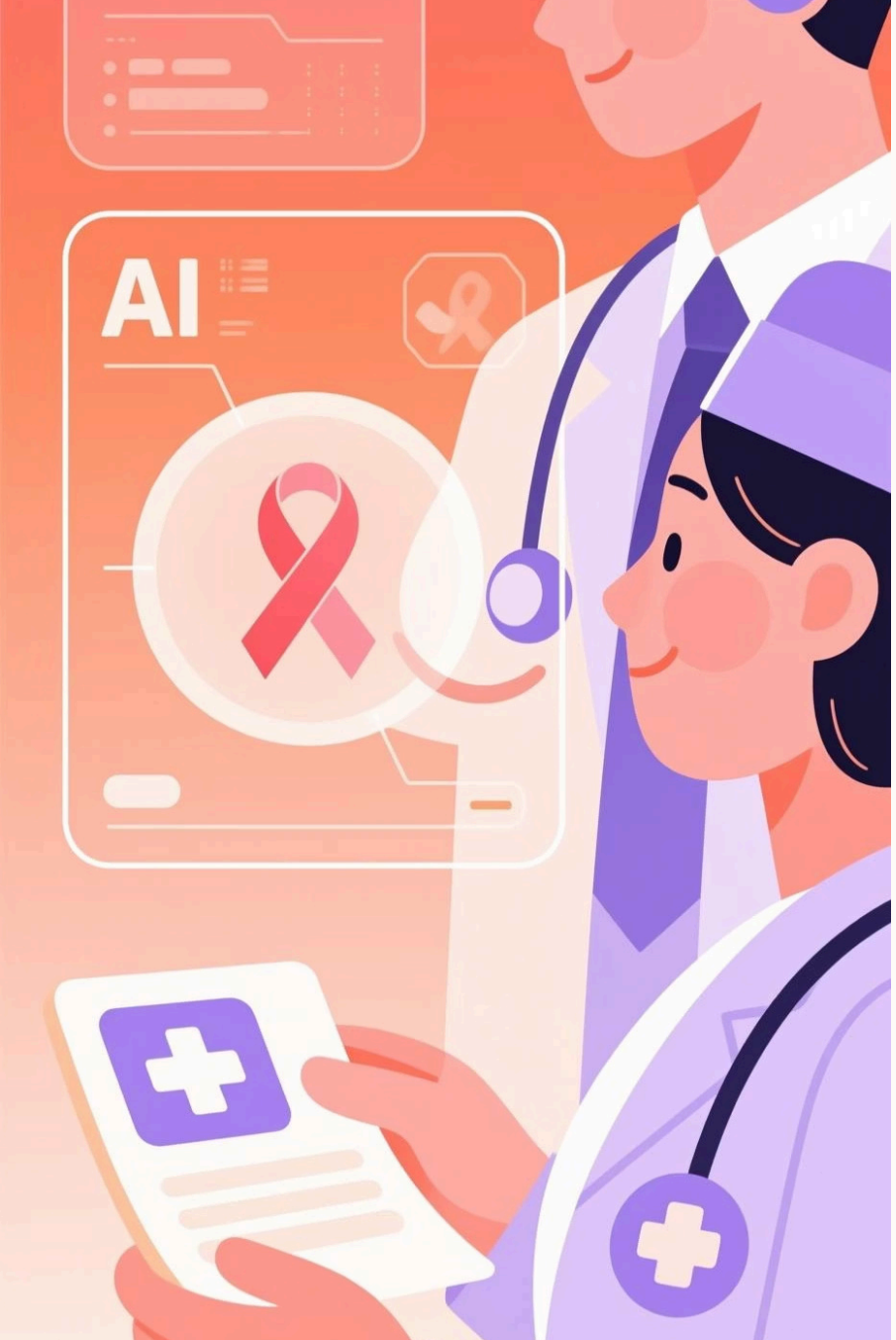


Breast Cancer Classification

Machine learning approach using Logistic Regression

CHAPTER 1

Essential Libraries

Pandas & NumPy

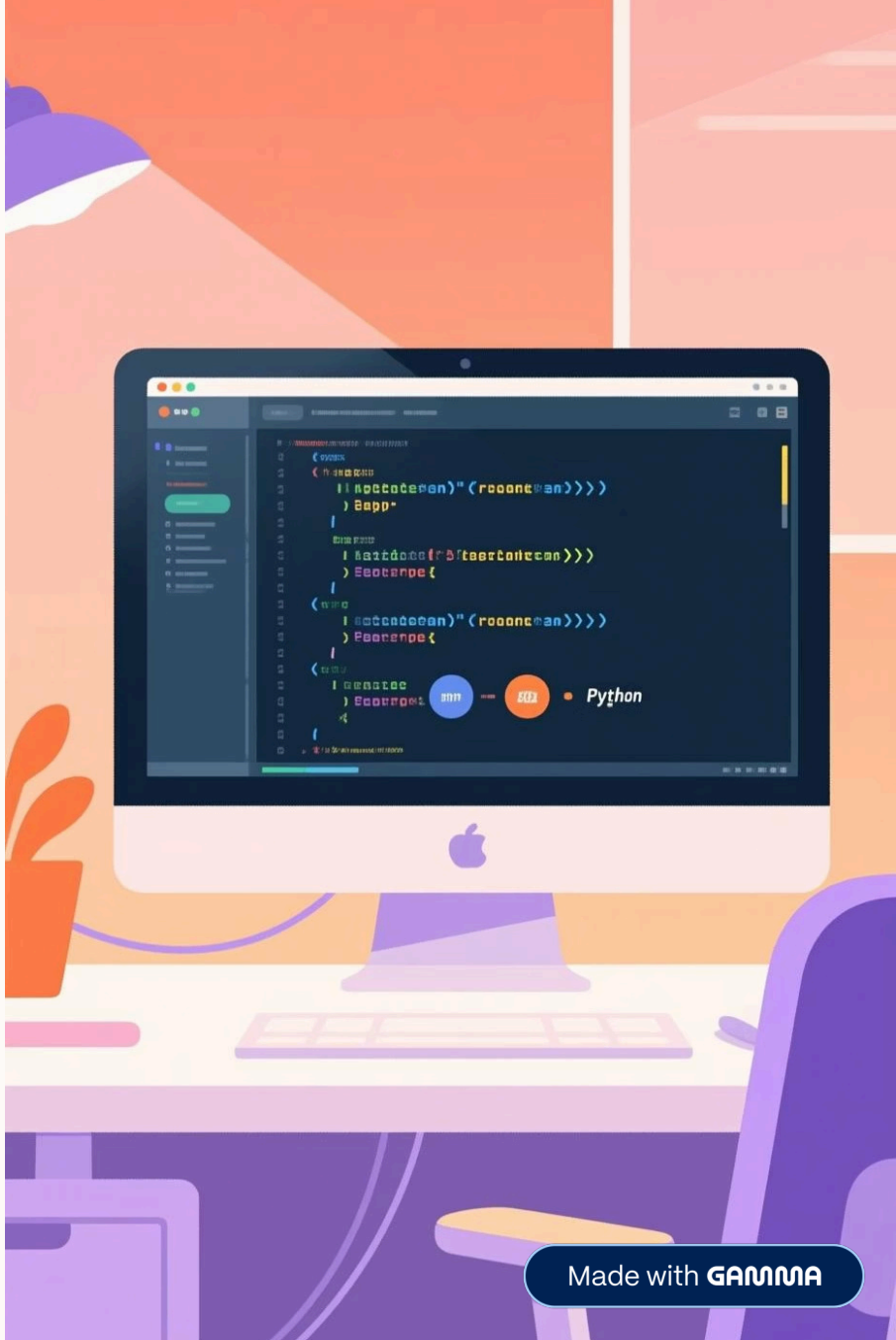
Data manipulation and numerical operations

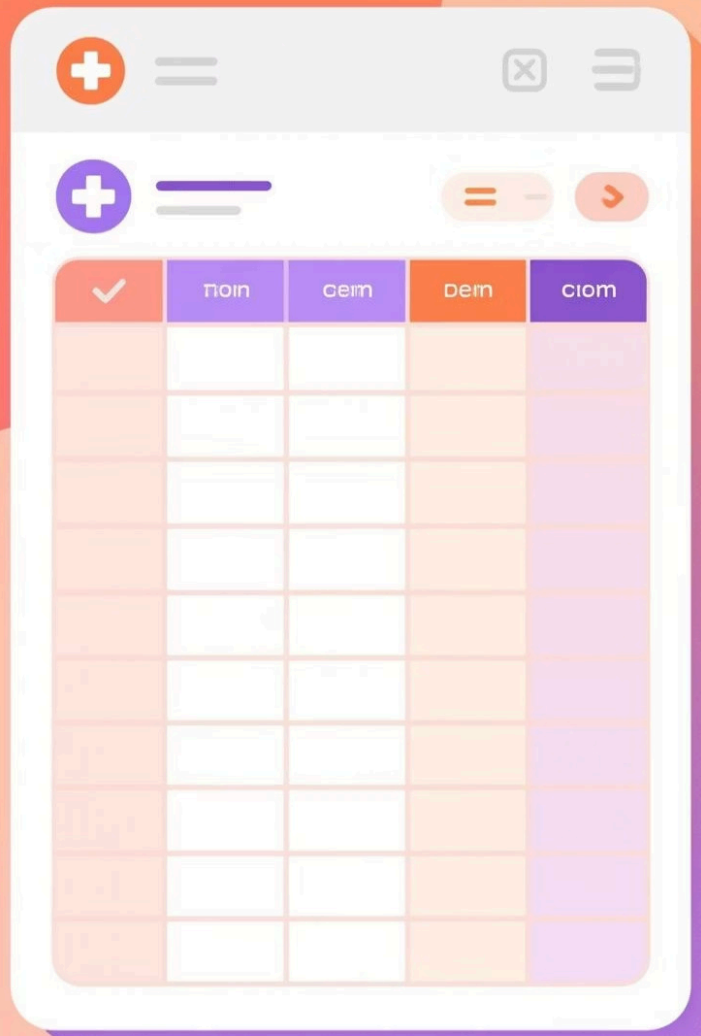
Matplotlib & Seaborn

Visualization and advanced plotting

Scikit-learn

Model training, scaling, and evaluation





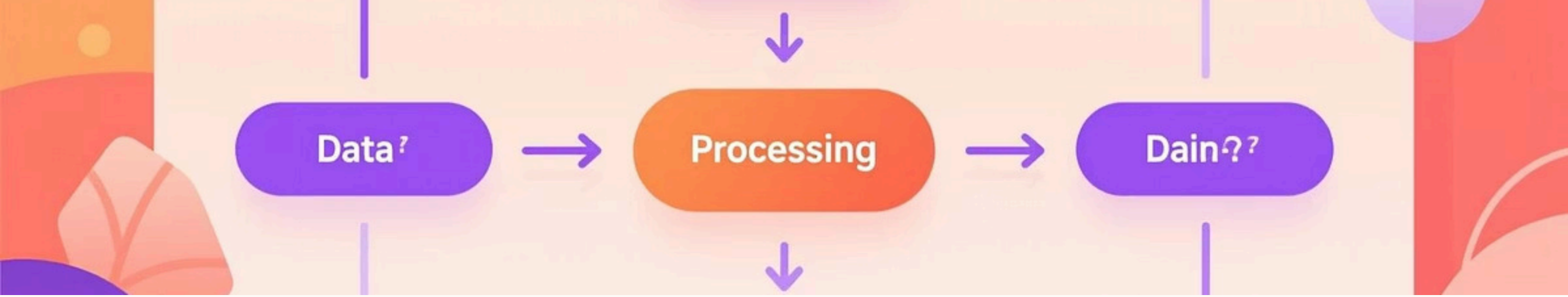
DATA EXPLORATION

Loading the Dataset

CSV file loaded into Pandas DataFrame

- Each row = one patient
- Columns = features + target variable
- Commands: df.head(), df.info(), df.describe()

Key insight: Understanding data structure is critical before modeling



CHAPTER 2

Data Preprocessing Pipeline

Separate Features

X = features, y = target (0 or 1)

Train-Test Split

80% training, 20% testing

Feature Scaling

Standardization: mean=0, std=1

> Feature scaling is critical for distance-based algorithms like Logistic Regression

Exploratory Data Analysis

Target Distribution

Class balance analysis using value_counts()

Feature Correlation

Identify features strongly influencing target

Correlation Heatmap

Detect multicollinearity among features

CHAPTER 3

Model Training

Logistic Regression

Binary classification algorithm

- Model learns feature-target relationships
- Predicts class labels (0 or 1)
- Random state ensures reproducibility

```
model = LogisticRegression(
 random_state=42
)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Core Evaluation Metrics

1

Accuracy Score

Overall correctness of predictions

Correct Predictions / Total Predictions

2

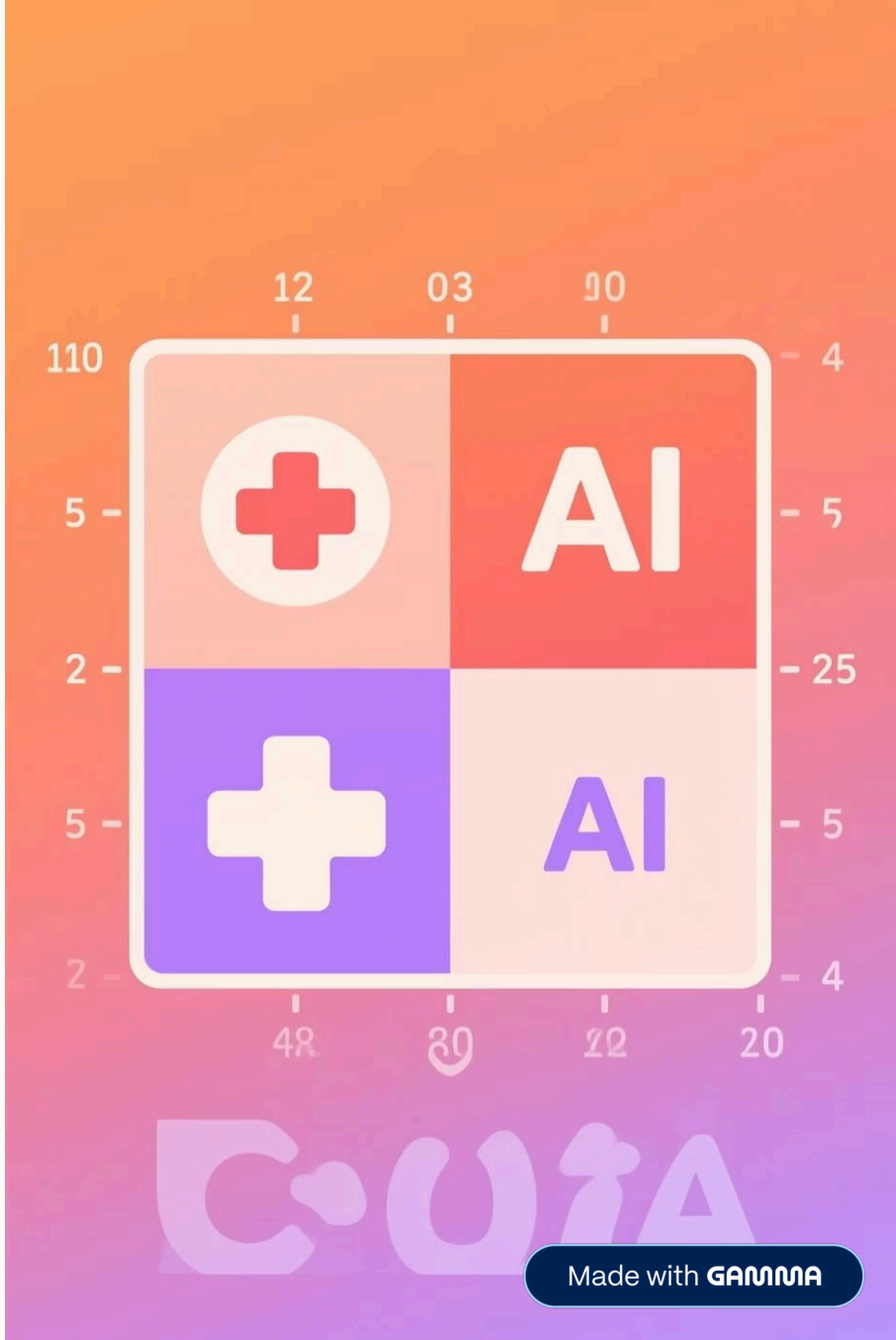
Confusion Matrix

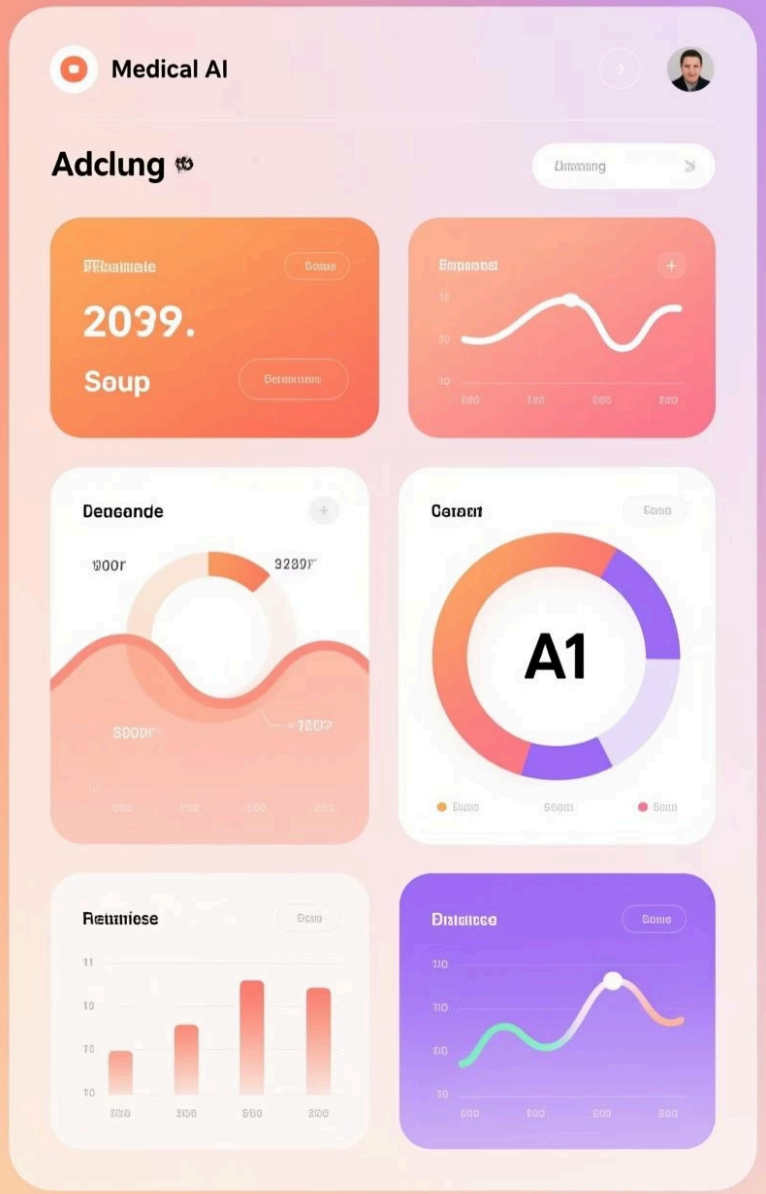
True Positives, True Negatives, False Positives, False Negatives

3

Classification Report

Precision, Recall, F1-Score, Support per class





ADVANCED METRICS

Beyond Basic Accuracy

F2 Score

Emphasizes recall over precision

Critical when false negatives are costly

F0.5 Score

Emphasizes precision over recall

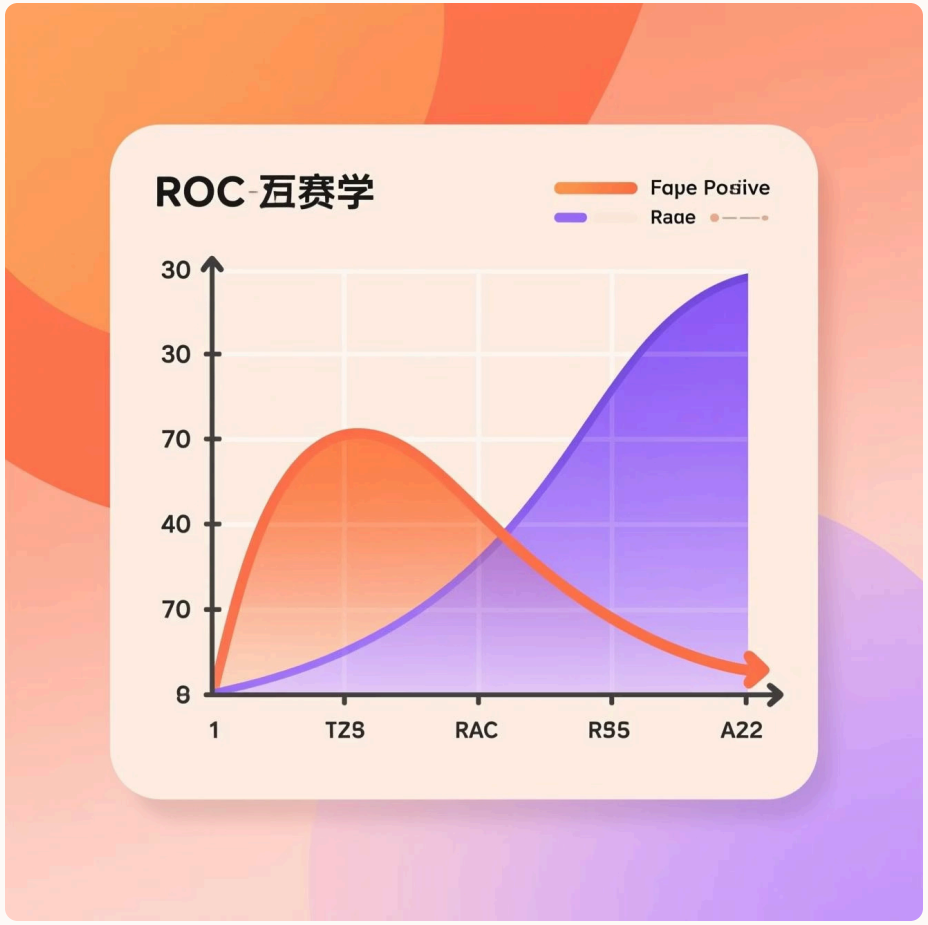
Important when false positives matter

Log Loss

Measures prediction confidence

Lower values indicate better probability estimates

ROC Curve & AUC Analysis



Model Performance

Probability predictions for positive class

0.5

Random Model

No predictive power

1.0

Perfect Classifier

Ideal performance

AUC (Area Under Curve) measures ability to distinguish between classes

Project Success

01

Data Preprocessing

Cleaned, scaled, and explored dataset

02

Binary Classification

Logistic Regression trained effectively

03

Comprehensive Evaluation

Multiple metrics beyond accuracy

04

ROC-AUC Validation

Confirmed classification strength