Data Flow & Feature Engineering Pipeline

RAW DATA

CSV File

4,424 rows

37 columns

DATA CLEANING

✓ Remove duplicates

✓ Handle missing (KNN)

✓ Drop invalid rows

ENCODING

Categorical → Numeric

One-Hot: gender, marital

Label: daytime/evening

FEATURE ENG

+ Success rate

+ Average grade

+ Academic progression

ACADEMIC FEATURES

THEORY

Tinto

68%

Bean

32%

Weight

• Grades (sem 1, 2)

• Success rate

• Enrolled units

• Approved units

• Evaluations

FINANCIAL FEATURES

• Tuition status

• Scholarship

• Debtor status

• Payment history

DEMOGRAPHIC FEATURES

• Age at enrollment

• Gender

• Marital status

• Parental education

• Previous qualification

NORMALIZATION (Z-Score)

x_norm = (x - μ) / σ   [Applied to all 46 features]

PROCESSED DATASET: 46 Features × 4,424 Samples

Train: 3,096 (70%) | Validation: 664 (15%) | Test: 664 (15%)