Lab2: Body Fat Prediction Dataset

For this lab, we use the **Body Fat Prediction** dataset, which contains anthropometric measurements collected from subjects (e.g., age, weight, height, and several body circumferences). The objective is to build a regression model that can **predict BodyFat (%)**, because body fat percentage is not always directly measurable in typical settings without specialized equipment.

The dataset also includes a **Density** attribute. However, Density is strongly related to body fat percentage because it is typically connected through established formulas, meaning it can behave like a "shortcut" feature that makes the prediction task unrealistically easy. In our use case, we assume that **Density is not available** (e.g., we do not have the appropriate sensor or measurement process in our lab). Therefore, in this lab we intentionally remove the Density column and focus on predicting **BodyFat (%) using only the measurements we can realistically obtain**.

Dataset Description

The **BodyFat** dataset contains **252 adult male subjects**. Each row is one subject. The goal is to predict **BodyFat (%)** from body measurements.

Target Variable

BodyFat: Body fat percentage (%) — this is the **target** to be predicted.

Features:

Age: Age (years)

Weight: Body weight (lbs)

Height: Height (inches)

Neck: Neck circumference (cm)

Chest: Chest circumference (cm)

Abdomen: Abdomen/waist circumference (cm)

Hip: Hip circumference (cm)

Thigh: Thigh circumference (cm)

Knee: Knee circumference (cm)

Ankle: Ankle circumference (cm)

Biceps: Biceps circumference (cm)

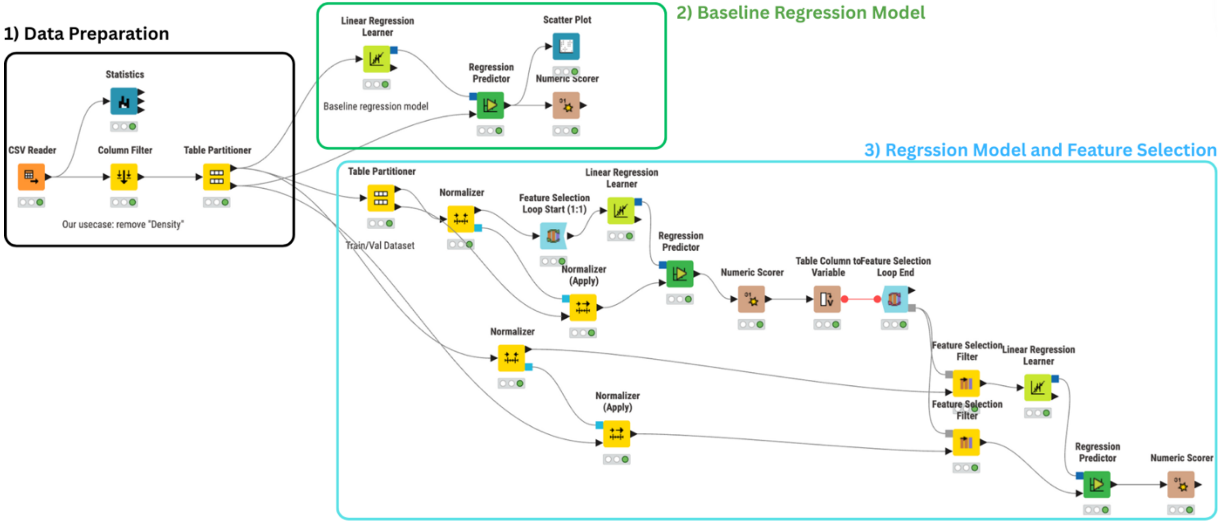
Forearm: Forearm circumference (cm)

Wrist: Wrist circumference (cm)

Density: Body density estimate (typically in g/cm³).

Lab note: We **remove Density** to simulate a realistic scenario where this measurement is **not** available.

KNIME Instructions



1. Data Preparation

1.1 CSV Reader

Load the BodyFat CSV file.

1.2 Statistics (optional)

Inspect distributions/summary statistics.

1.3 Column Filter

Remove **Density** (simulate a realistic setting where Density is not available).

Keep **BodyFat** and all other measurement columns.

1.4 Table Partitioner (Dev/Test split)

Partitioning method: **Random**

Split ratio: **80% / 20%**

Set your own fixed random seed for reproducibility.

Output 1 = **Dev (Train+Validation)**, Output 2 = **Test**

---

2. Baseline Regression Model (Linear Regression)

2.1 Linear Regression Learner

Input: **Dev (80%)** from Table Partitioner

Target/Response column: **BodyFat**

2.2 Regression Predictor

Model input: from Linear Regression Learner

Data input: **Test (20%)**

2.3 Numeric Scorer

Report at least **RMSE** (optionally R²).

2.4 Scatter Plot (optional)

Plot predicted vs actual BodyFat.

3. Regression Model + Feature Selection (Wrapper)

3.1 Table Partitioner (Train/Validation)

Input: **Dev (80%)**

Output: **75% Train / 25% Validation**

3.2 Normalizer (fit on Train only)

Input: **Train**

Method: standardization (z-score) is typical.

3.3 Normalizer (Apply)

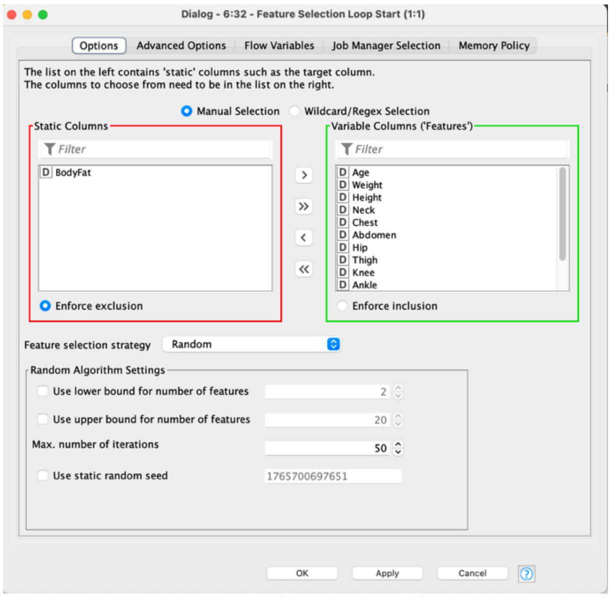
Apply the same normalization model to: **Validation set**

3.4 Feature Selection Loop Start (1:1)

Input: **normalized Train**

Ensure **BodyFat is the target**, and **BodyFat is NOT treated as a selectable feature**.

Double-clck the node to open configure view.



Dialog - 6:32 - Feature Selection Loop Start (1:1)

The list on the left contains 'static' columns such as the target column.
The columns to choose from need to be in the list on the right.

● Manual Selection ○ Wildcard/Regex Selection

Static Columns

D BodyFat

● Enforce exclusion

Variable Columns ('Features')

D Age
D Weight
D Height
D Neck
D Chest
D Abdomen
D Hip
D Thigh
D Knee
D Ankle

○ Enforce inclusion

Feature selection strategy: Random

Random Algorithm Settings

□ Use lower bound for number of features: 2

□ Use upper bound for number of features: 20

Max. number of iterations: 50

□ Use static random seed: 1765700697651

3.5 Linear Regression Learner (inside loop)

Target: **BodyFat**

3.6 Regression Predictor (inside loop)

Data input: **normalized Validation**

3.7 Numeric Scorer (inside loop)

Metric: **RMSE**

3.8 Table Column to Variable

Convert the RMSE output into a **Double flow variable**.

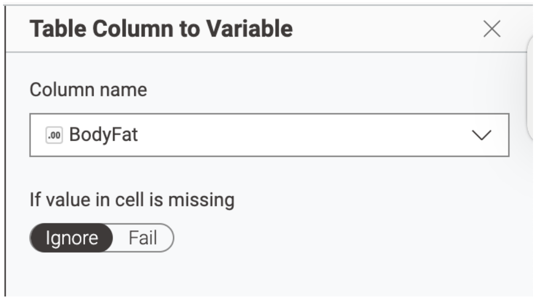


Table Column to Variable

Column name

.00 BodyFat

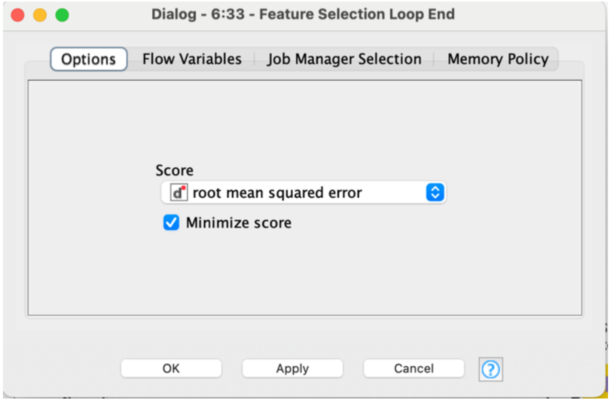
If value in cell is missing

[✓ Ignore] Fail

3.9 Feature Selection Loop End

Optimization: **minimize RMSE**

This selects the best feature subset based on Validation performance.



Score: root mean squared error

☑ Minimize score

3.10 Normalizer (fit on Train + Validation)

Input: **Train + Validation (From the first table partitioner) (80%)**

Method: standardization (z-score)

3.11 Normalizer (Apply)

Apply the same normalization model to: **Test set (20%)**

3.12 Feature Selection Filter (Dev)

Model input: from **Feature Selection Loop End**

Data input: **normalized Train + Validation (80%)**

Enable **Include static columns** so **BodyFat** remains available for training.

3.13 Linear Regression Learner (final)

Train on filtered **Train + Validation set**

Target: **BodyFat**

3.14 Feature Selection Filter (Test)

Model input: from **Feature Selection Loop End**

Data input: **normalized Test (20%)**

(Recommended to guarantee the test schema matches the selected feature set.)

3.15 Regression Predictor (final)

Model input: final learner output

Data input: filtered Test

3.16 Numeric Scorer (final)

Report final Test RMSE