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)

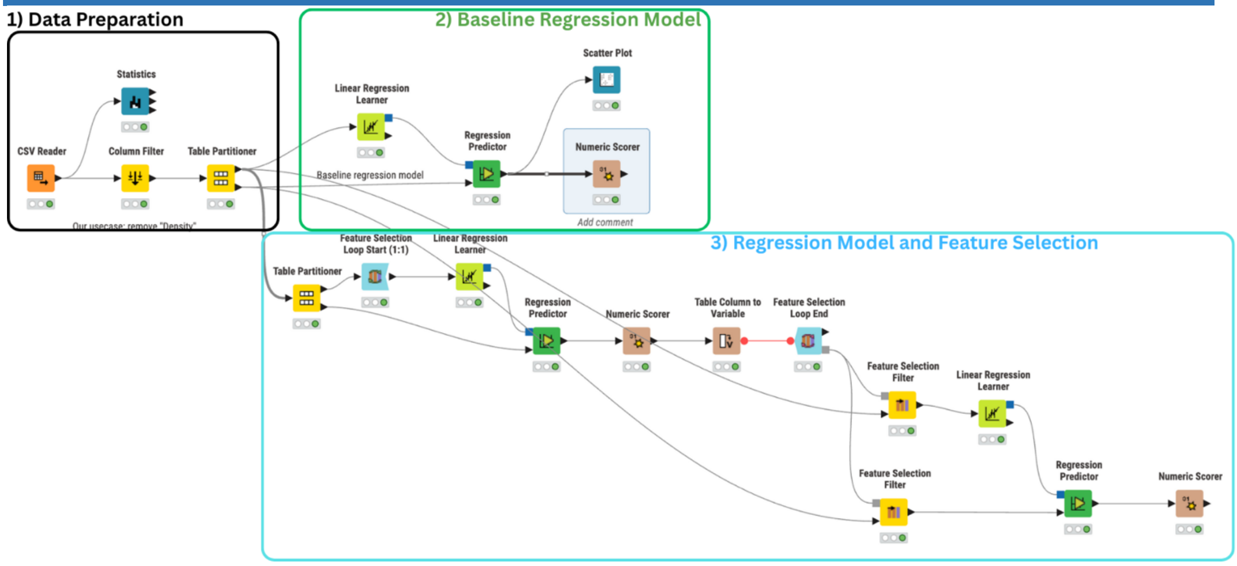
Forearm: Forearm circumference (cm)

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**

Also set its Fixed random seed = 2026.

2. Baseline Regression Model (Linear Regression)

2.1 Linear Regression Learner

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

Target/Response column: **BodyFat**

You can view the output of the learner by right-clicking then open view.

Linear Regression Result View -...

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Age	0.0707	0.036	1.9638	0.051
Weight	-0.076	0.059	-1.2885	0.1992
Height	-0.0553	0.1	-0.5531	0.5809
Neck	-0.3915	0.2581	-1.5169	0.131
Chest	-0.078	0.1124	-0.6941	0.4885
Abdomen	0.9838	0.0947	10.388	0.0
Hip	-0.2453	0.1651	-1.486	0.139
Thigh	0.2603	0.1603	1.6242	0.106
Knee	-0.1712	0.2957	-0.5791	0.5632
Ankle	0.1581	0.2345	0.674	0.5011
Biceps	0.2138	0.1887	1.1331	0.2586
Forearm	0.5594	0.2098	2.666	0.0083
Wrist	-1.6342	0.6103	-2.6777	0.0081
Intercept	-15.7686	19.0049	-0.8297	0.4078

R-Squared: 0.7579
Adjusted R-Squared: 0.7411

2.2 Regression Predictor

Model input: from Linear Regression Learner

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

2.3 Numeric Scorer

Report at least **RMSE** (optionally R^2).

RowID	Prediction (BodyFat) Number (Float)
R^2	0.681
mean absolute error	3.448
mean squared error	18.71
root mean squared error	4.326
mean signed difference	1.172
mean absolute percentage error	0.235
adjusted R^2	0.681

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**

Also set its Fixed random seed = 2026.

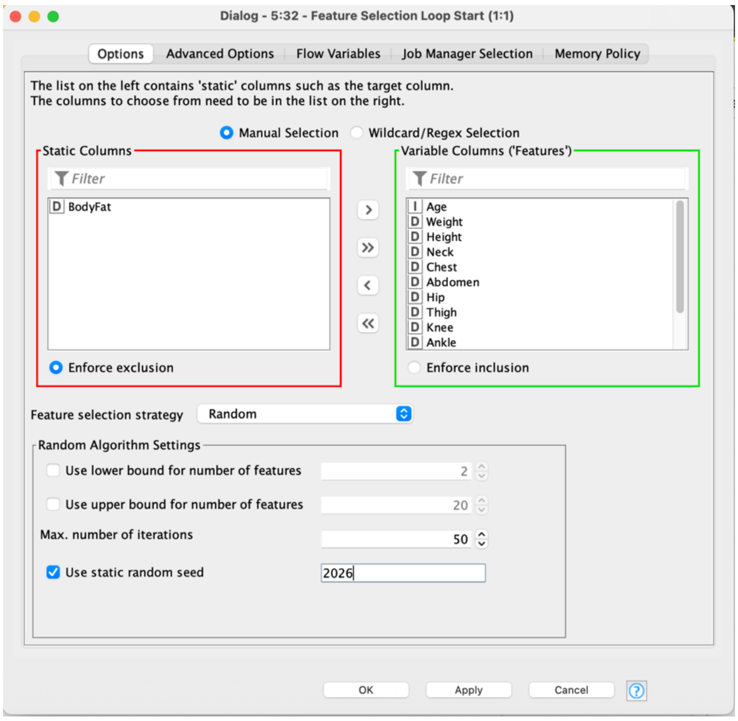
3.2 Feature Selection Loop Start (1:1)

Input: **Train dataset**

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

Double-clck the node to open configure view.

Also set static random seed = 2026



3.3 Linear Regression Learner (inside loop)

Target: **BodyFat**

3.4 Regression Predictor (inside loop)

Data input: **Validation Dataset**

3.5 Numeric Scorer (inside loop)

Output: A table of regression evaluation metrics (e.g., **RMSE**, **MAE**, and/or **R^2**) computed on the **Validation** predictions.

▶ 1: Statistics   Flow Variables

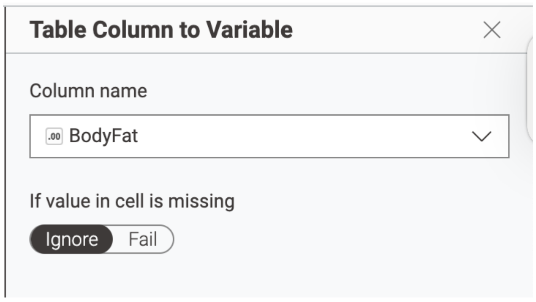
Rows: 7 | Columns: 1

#	RowID	BodyFat (Number (Float))
1	R^2	-0.743
2	mean absolute error	5.174
3	mean squared error	39.176
4	root mean squared error	6.259
5	mean signed difference	-1.121
6	mean absolute percentage error	0.299
7	adjusted R^2	-0.743

3.6 Table Column to Variable

Purpose: Convert the selected metric from the Numeric Scorer output (e.g., **RMSE**) into a **Double flow variable**.

Use: This flow variable is passed to **Feature Selection Loop End** so the loop can compare feature subsets and select the best one (e.g., **minimize RMSE** or **maximize R^2**).

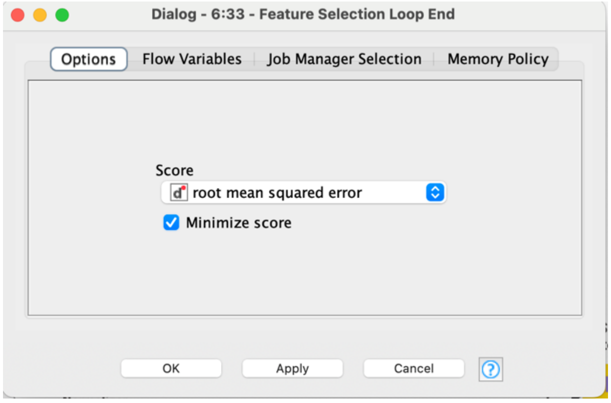


Owner ID	Data Type	Variable Name	Value
5:35	DoubleType	adjusted R^2	0.5140224576415714
5:35	DoubleType	mean absolute percentage error	0.2750082820930372
5:35	DoubleType	mean signed difference	-0.8720463004785128
5:35	DoubleType	root mean squared error	6.844443940206473
5:35	DoubleType	mean squared error	46.84641285062911
5:35	DoubleType	mean absolute error	4.917554272951872
5:35	DoubleType	R^2	0.5140224576415714

3.7 Feature Selection Loop End

Optimization: **minimize RMSE**

This selects the best feature subset based on Validation performance.



3.8 Normalizer (fit on Train + Validation)

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

Method: standardization (z-score)

3.9 Feature Selection Filter (Dev)

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

Data input: **Dev + Validation (80%)**

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

3.10 Linear Regression Learner (final)

Train on filtered **Train + Validation set**

Target: **BodyFat**

3.11 Feature Selection Filter (Test)

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

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

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

3.12 Regression Predictor (final)

Model input: final learner output

Data input: filtered Test

3.13 Numeric Scorer (final)

Report final Test metrics

RowID	BodyFat Number (Float)
R^2	0.595
mean absolute error	3.664
mean squared error	20.762
root mean squared error	4.557
mean signed difference	-1.615
mean absolute percentage error	0.207
adjusted R^2	0.595

How to Interpret Regression Performance (BodyFat %)

What each metric means (and what "better" looks like):

- **R^2 / Adjusted R^2 (0 to 1):** how much variance in *BodyFat* the model explains. **Higher is better.**
- **MAE (Mean Absolute Error):** average absolute prediction error in **percentage points**.
  Example: **MAE = 3.45** means the prediction is off by ~**3.45 BodyFat%** on average.
  Lower is better.
- **RMSE (Root Mean Squared Error):** like MAE but **penalizes large errors more**. **Lower is better.**
- **MSE:** squared version of error; mainly used because RMSE is derived from it. **Lower is better.**
- **Mean Signed Difference (Bias):** shows whether the model systematically over/under-predicts.
  Closer to 0 is better (positive = overpredict; negative = underpredict).
- **MAPE:** average percentage error (relative error). **Lower is better**, but can be sensitive when true values are small.

Comparing Your Two Results (Baseline vs Feature-Selected)

- **Baseline** performs better on the main accuracy metrics:
  - **R^2:** 0.681 (better than 0.595)
  - **MAE:** 3.448 (better than 3.664)
  - **RMSE:** 4.326 (better than 4.557)
- **Feature-selected model** has a slightly better **MAPE** (0.207 vs 0.235), but overall it **does not improve** the model on R^2/MAE/RMSE.