

## Lab3: H2O AutoML in KNIME

This lab demonstrates how to train and evaluate a classification model using **H2O AutoML** inside KNIME.

### Learning goals

Prepare a dataset for AutoML.

Train H2O AutoML models with a reproducible train/test split.

Evaluate predictions using KNIME's Scorer and interpret the confusion matrix and key metrics.

### Dataset Description

The **Mall Customer** dataset contains **200 customers**. Each row is one subject. The goal is to predict the **Spending Class (High or Low Spender)** of the customer.

#### Target Variable

**High Spender:** “yes” if spending score  $\geq 60$ ; else “no”.

#### Features:

**CutomerID**

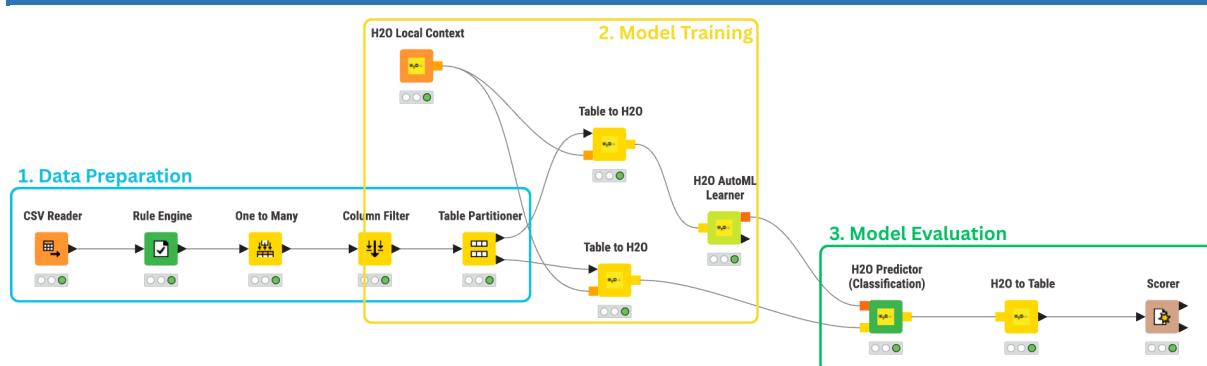
**Gender:** Male/Female

**Age:** (years)

**Annual Income:** (k\$)

**Spending Score:** Derived Metric (0-100)

## KNIME Instructions



#### 1) Data Preparation (blue block)

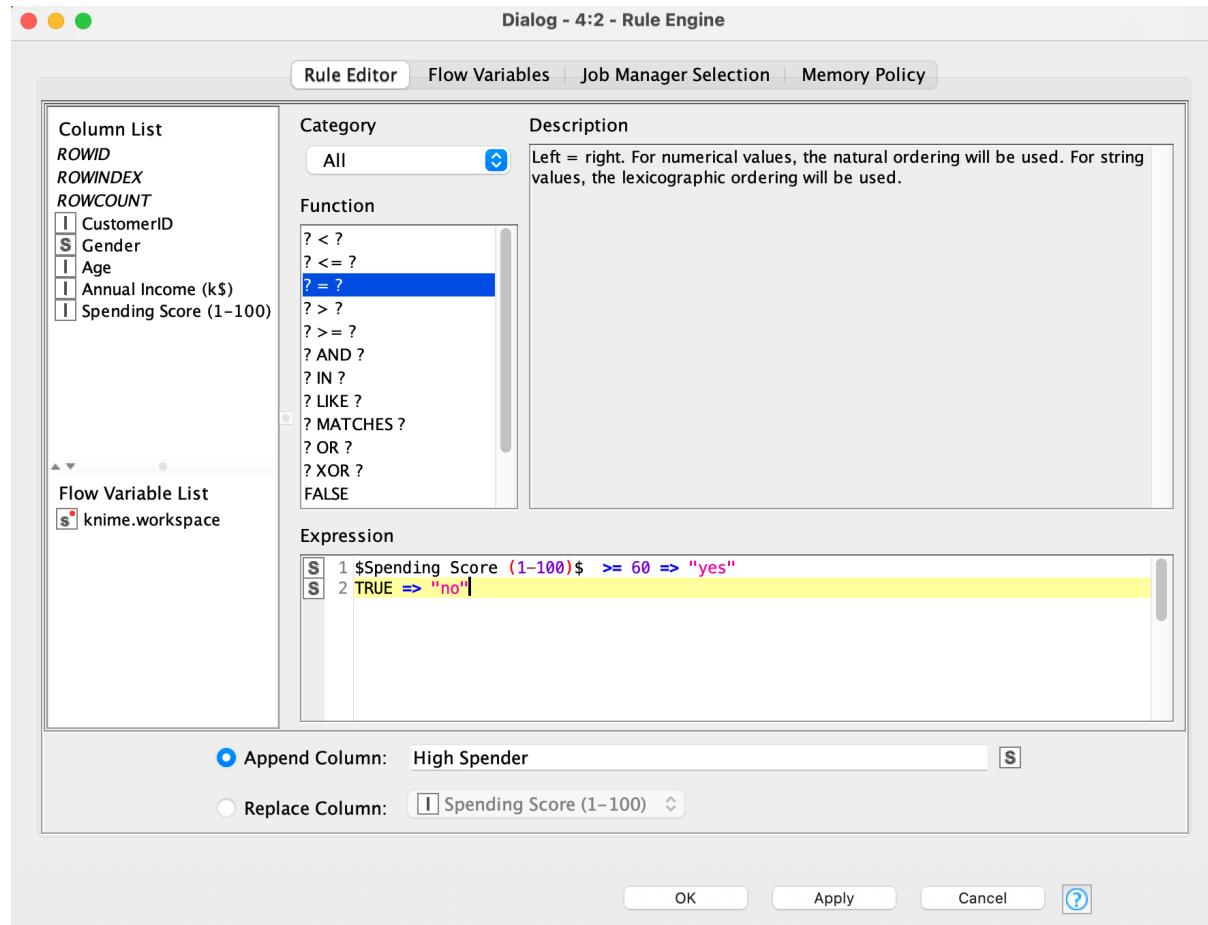
##### 1.1 CSV Reader

Load your lab CSV file.

- Configure the file path.

## 1.2 Rule Engine

Create/clean the target label (class column).



## 1.3 One to Many

Encode categorical predictors into dummy/indicator columns.

**One to Many** X

Columns to transform

Manual     Wildcard     Regex

Search Aa

Available columns

High Spender

Columns to transform

Gender

>    >>    <    <<

*Any unknown column*

Remove included columns from output

#### 1.4 Column Filter

Select the final modeling schema.

The screenshot shows a 'Column Filter' interface with the following components:

- Title:** Column Filter
- Search Bar:** A search bar with a magnifying glass icon and the placeholder text "Search". To its right is a dropdown menu with the letters "Aa".
- Filter Types:** A horizontal row of four buttons: "Manual" (selected), "Wildcard", "Regex", and "Type".
- Exclude List:** A list titled "Excludes" containing three items:
  - CustomerID
  - Spending Score (1-1...
  - Female
- Includes List:** A list titled "Includes" containing four items:
  - Age
  - Annual Income (k\$)
  - High Spender
  - Male
- Navigation:** Between the two lists are four small arrows: a single greater than sign (>), a double greater than sign (»), a single less than sign (<), and a double less than sign («).
- Text at Bottom:** The text "Any unknown column" is centered at the bottom of the includes list.

## 1.5 Table Partitioner

Split into training and test sets.

The screenshot shows the configuration interface for the 'Table Partitioner' node in KNIME. The interface includes the following sections:

- First partition type:** A dropdown menu with options 'Relative (%)' (selected) and 'Absolute'.
- Relative size:** An input field containing '80' with up and down arrow buttons for adjustment.
- Sampling strategy:** A dropdown menu with options 'Random' (selected), 'Stratified' (highlighted in dark grey), 'Linear', and 'First rows'.
- Group column:** An input field containing 'High Spender' with a dropdown arrow button.
- Fixed random seed:** A checked checkbox with the value '2026' in an input field with up and down arrow buttons.
- If input table is empty:** A dropdown menu with options 'Fail' (selected) and 'Output empty table(s)'.

## 2) Model Training (yellow block)

### 2.1 H2O Local Context

Initialize the local H2O backend.

### 2.2 Table to H2O (Train)

Convert the KNIME training table into an H2OFrame.

Connect: Table Partitioner (Train) → Table to H2O.

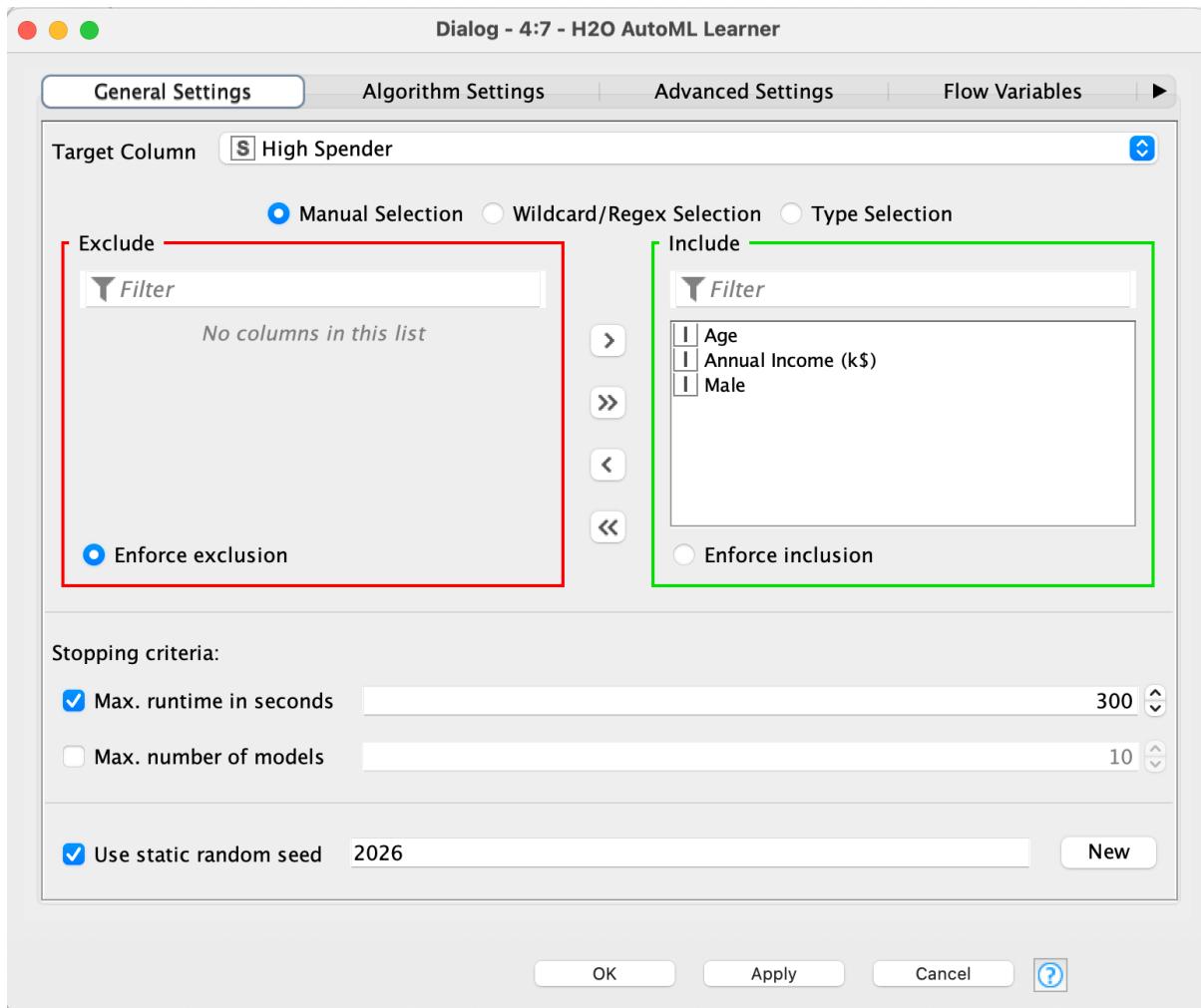
### 2.3 Table to H2O (Test)

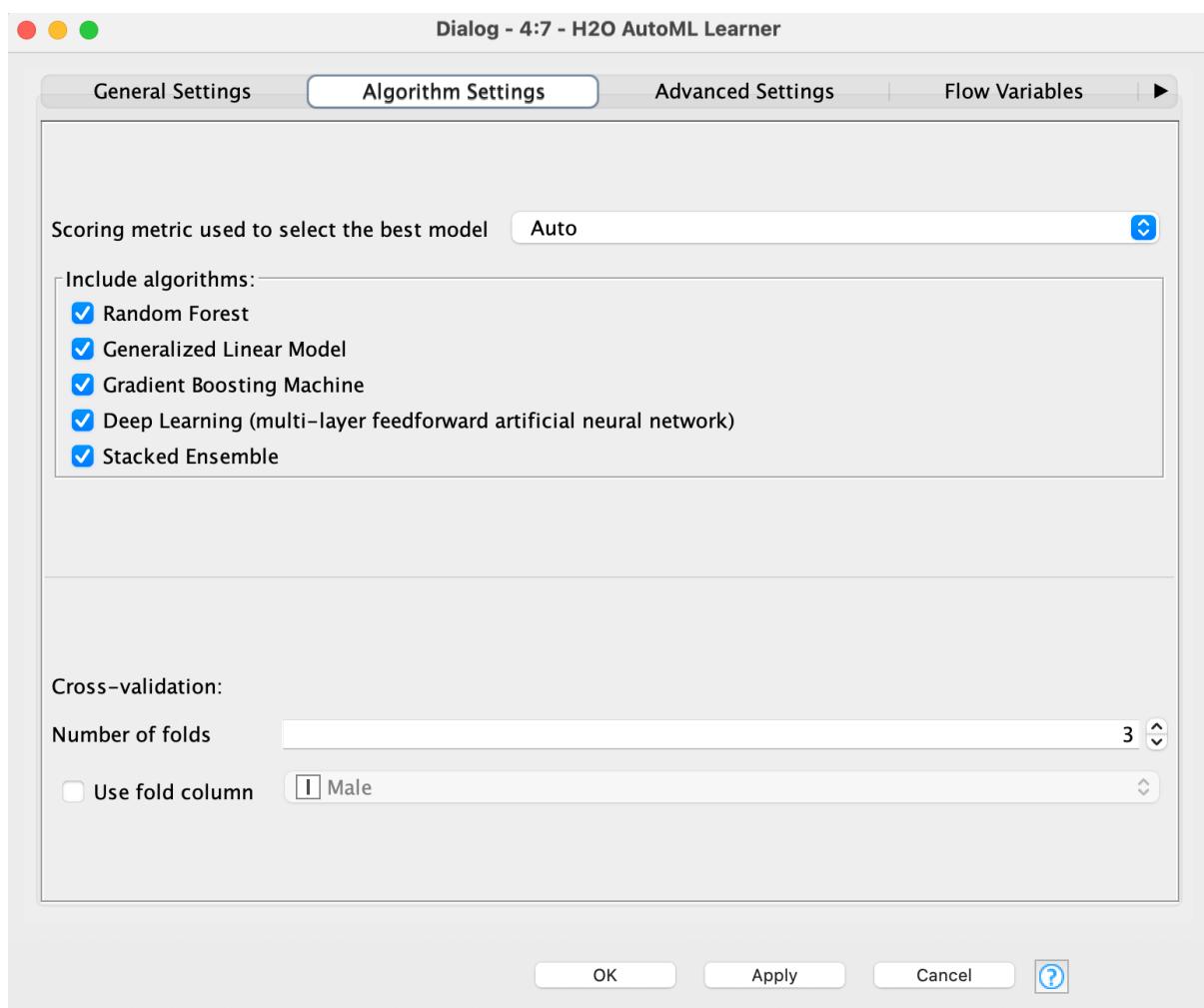
Convert the KNIME test table into an H2OFrame.

Connect: Table Partitioner (Test) → Table to H2O.

## 2.4 H2O AutoML Learner

Train a leaderboard of models automatically.





### 3) Model Evaluation (green block)

#### 3.1 H2O Predictor (Classification)

Score the test H2OFrame using the trained AutoML model.

Connect:

- Model input: H2O AutoML Learner output
- Data input: Table to H2O (Test) output

Output: H2OFrame with predicted class and class probabilities.

#### 3.2 H2O to Table

Convert predictions back to a KNIME table so you can use standard KNIME scoring and visualization nodes.

#### 3.3 Scorer

Evaluate classification performance.

**Scorer** X

First column

▼

Second column

▼

Sorting strategy

▼

Reverse order

Use name prefix

Missing values

Ignore  Fail