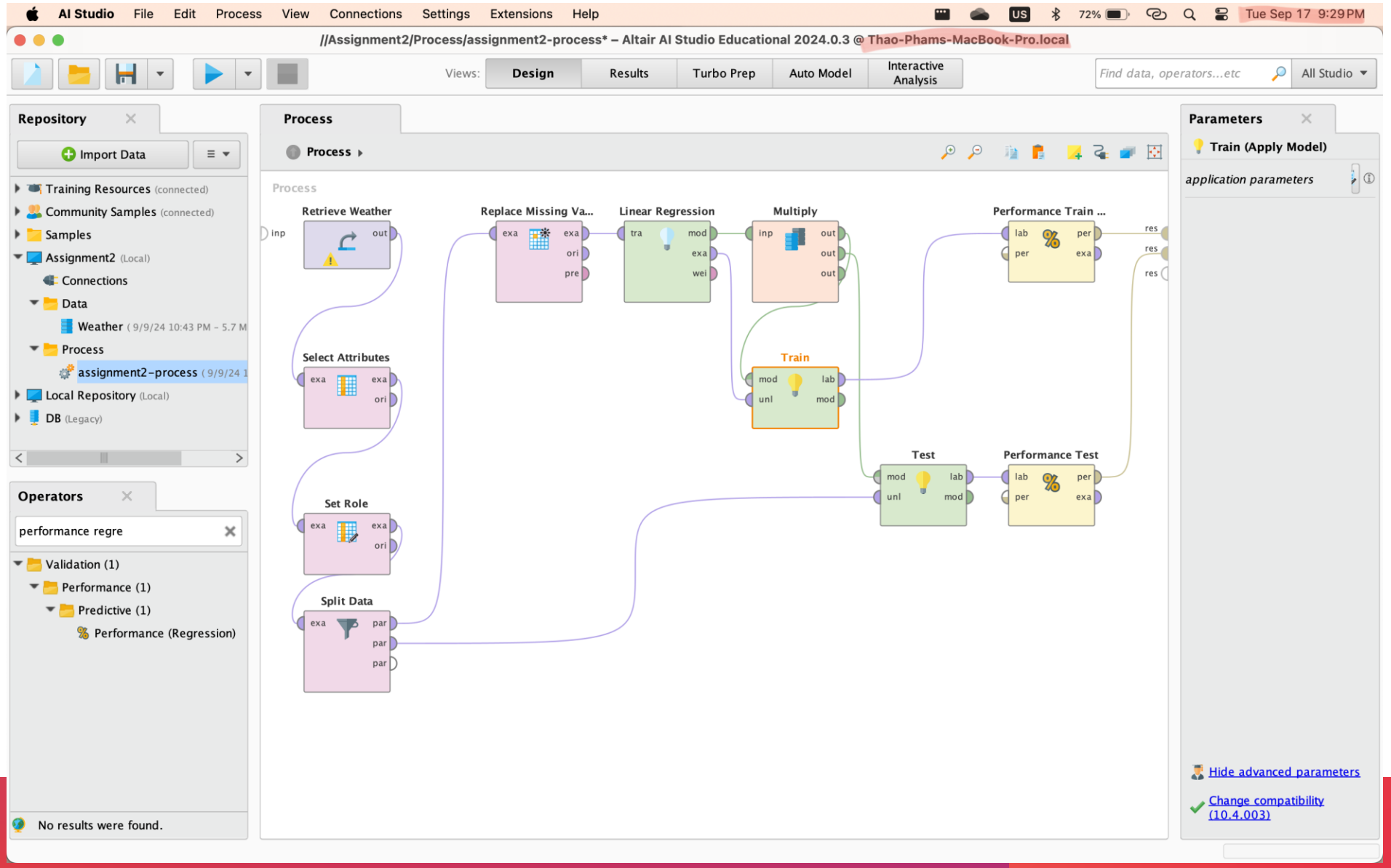
MSBA 265 – FALL 2024
ASSIGNMENT 2
LINEAR REGRESSION MODEL
WEATHER

PHUC MINH THAO PHAM - 989468542

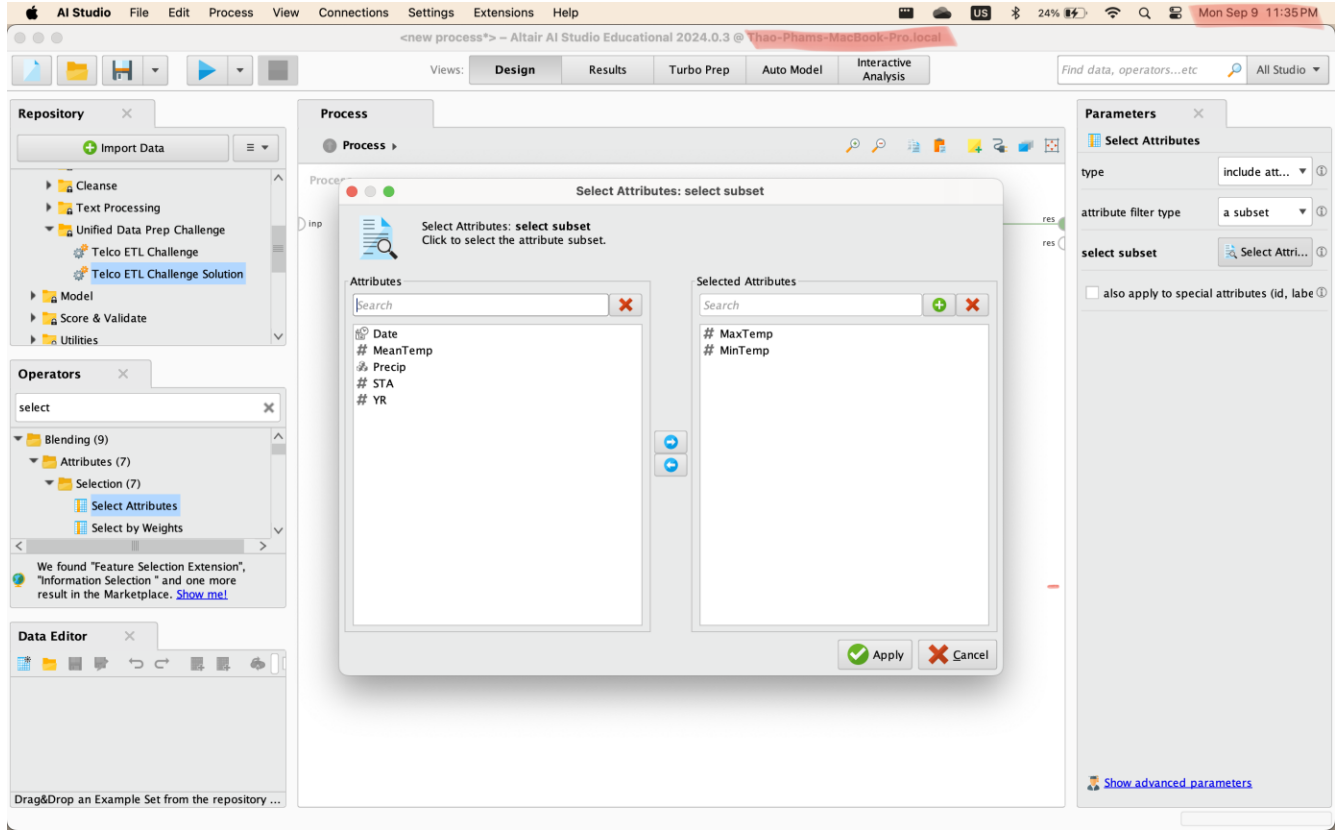
DATASET AND QUESTION?

- Weather Conditions in World War Two
- Is there a relationship between the daily minimum and maximum temperature?
- Can you predict the maximum temperature given the minimum temperature?


AI Studio File Edit Process View Connections Settings Extensions Help
Tue Sep 17 9:29 PM
//Assignment2/Process/assignment2-process* – Altair AI Studio Educational 2024.0.3 @ Thao-Phams-MacBook-Pro.local
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Assignment2 (Local)
Connections
Data
Weather (9/9/24 10:43 PM – 5.7 M
Process
assignment2-process (9/9/24 1
Local Repository (Local)
DB (Legacy)
Operators
performance regre
Validation (1)
Performance (1)
Predictive (1)
Performance (Regression)
No results were found.
Process
Process
Process
Retrieve Weather
Replace Missing Va...
Linear Regression
Multiply
Performance Train ...
Select Attributes
Train
Set Role
Test
Performance Test
Split Data
Parameters
Train (Apply Model)
application parameters
Hide advanced parameters
Change compatibility (10.4.003)

PROCESS

RETRIEVE WEATHER

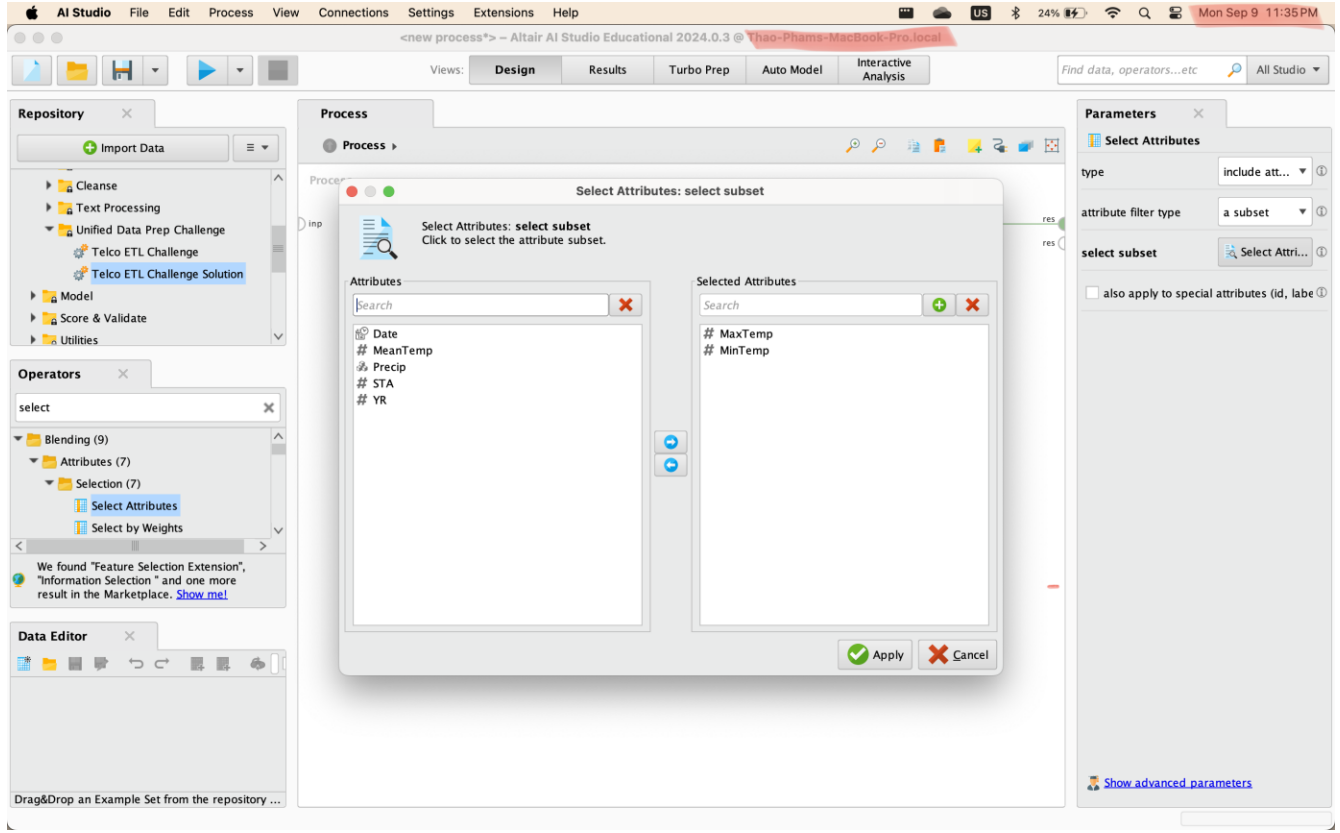


Operation tab – typing **Retrieve.**

Drag and drop to the process tab.

Parameters tab, selecting the correct dataset.

SELECT ATTRIBUTE



Operation tab – typing **Select Attribute.**

Drag and drop to the process tab.

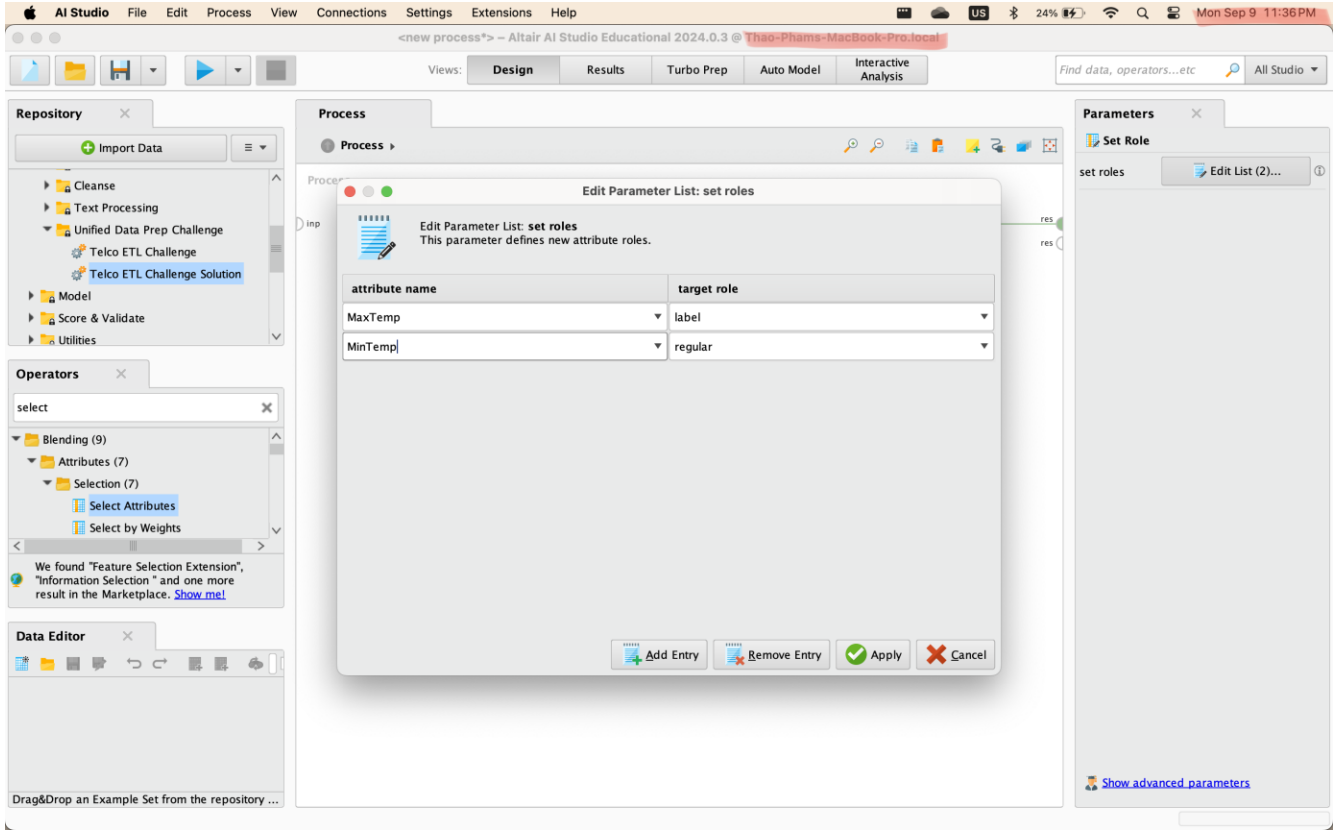
Parameters tab, then **attribute filter type** choose **a subset.**

Choose M**axTemp** and M**inTemp.**

Click **OK**.

Connecting **Retrieve Weather.output** to **Select Attributes.example set input**

SELECT ROLE



Operation tab, type **Set Role**.

Drag and drop to the process tab.

Parameter tab, click **Edit List**.

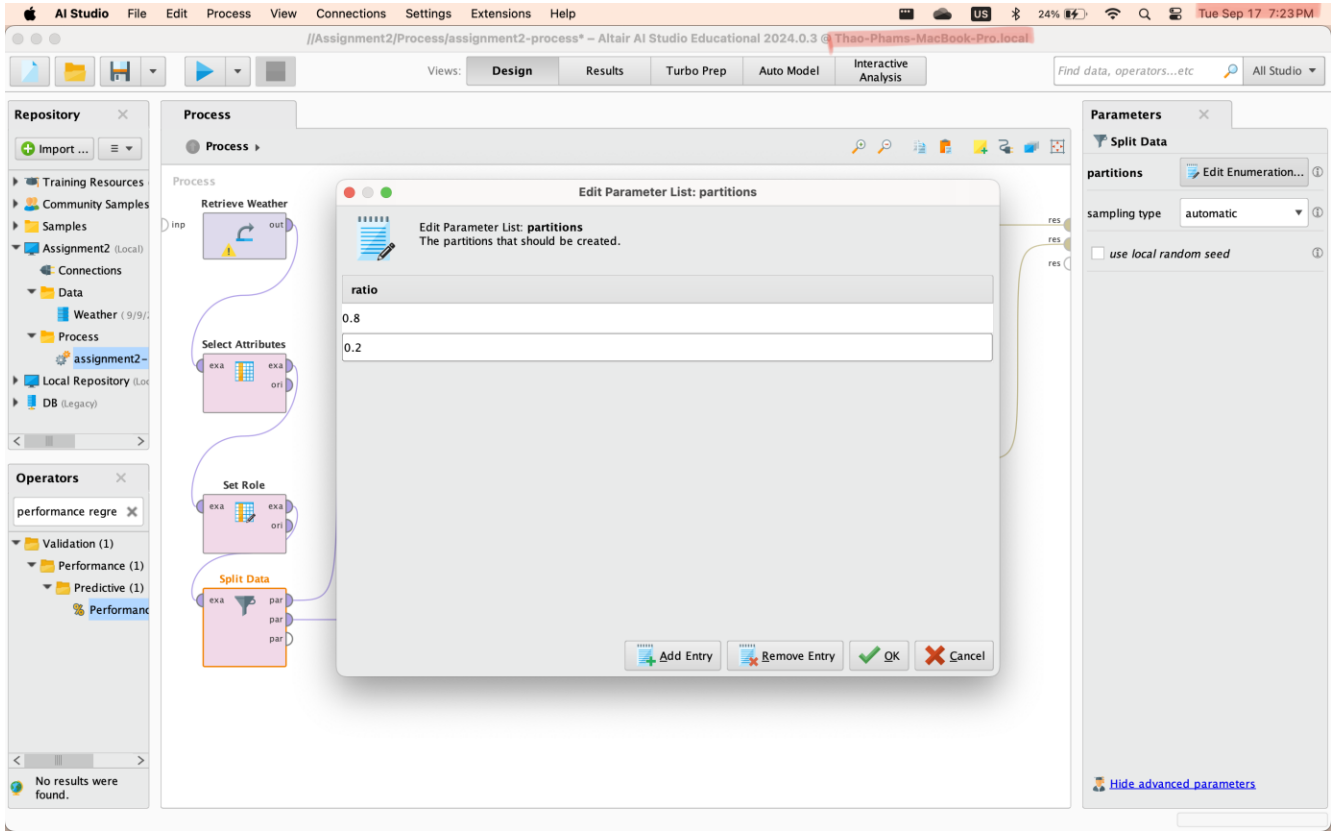
Set **MaxTemp** as **label** (dependent variable).

Set **MinTemp** as **regular** (independent variable).

Click **Apply.**

Connecting **Select Attributes.example set output** to **Set Role.example set input**

SPILT DATA



Operation tab, type **Split Data**.

Drag and drop to the process tab.

Parameter tab, click **Edit Enumeration List** in **partitions**.

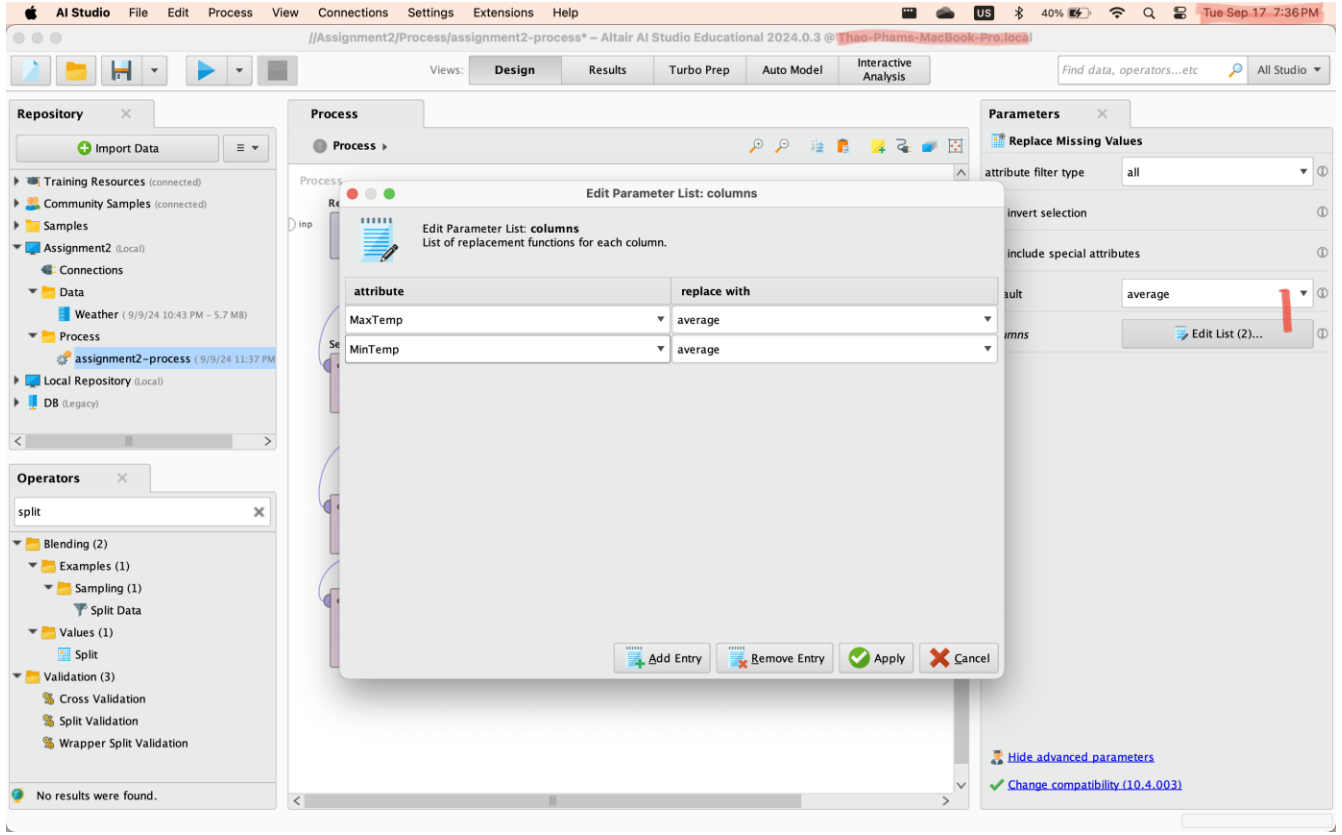
Add Entry – enter **0.8** for ratio.

Add Entry – enter **0.2** for ratio.

Click **OK**.

Connecting **Set Role.example set output** to **Split Data.example set**

REPLACE MISSING VALUES



Operation tab, type **Replace missing Values**.

Drag and drop to the process tab.

Parameter tab, click **Edit List** in ***column***.

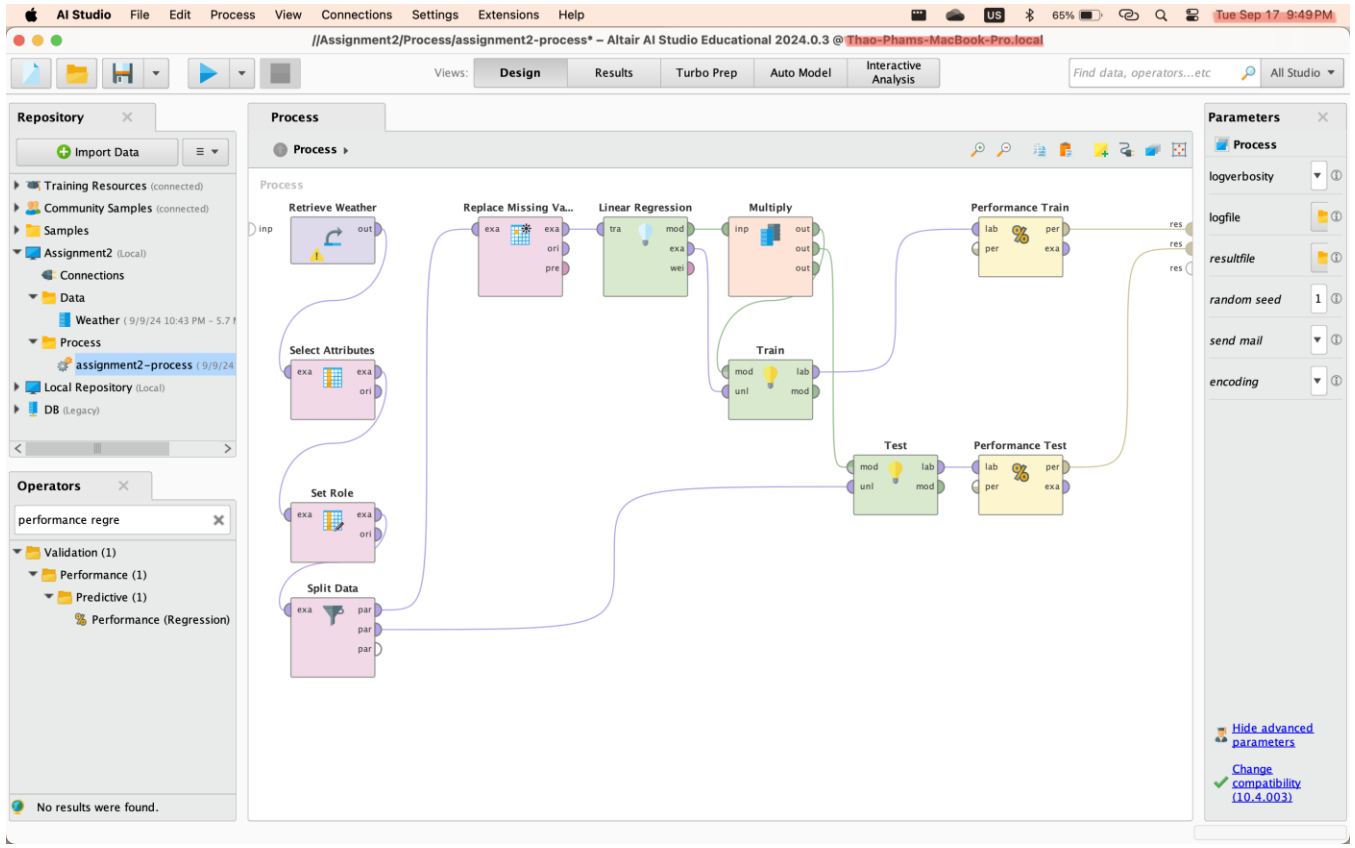
Add Entry – select **MaxTemp** as **attribute** and **average** in **replace with**.

Add Entry – select **MinTemp** as **attribute** and **average** in **replace with**.

Click **Apply**.

Connecting **Split Data.partition 1** to **Replace Missing Values.example set input**

LINEAR REGRESSION



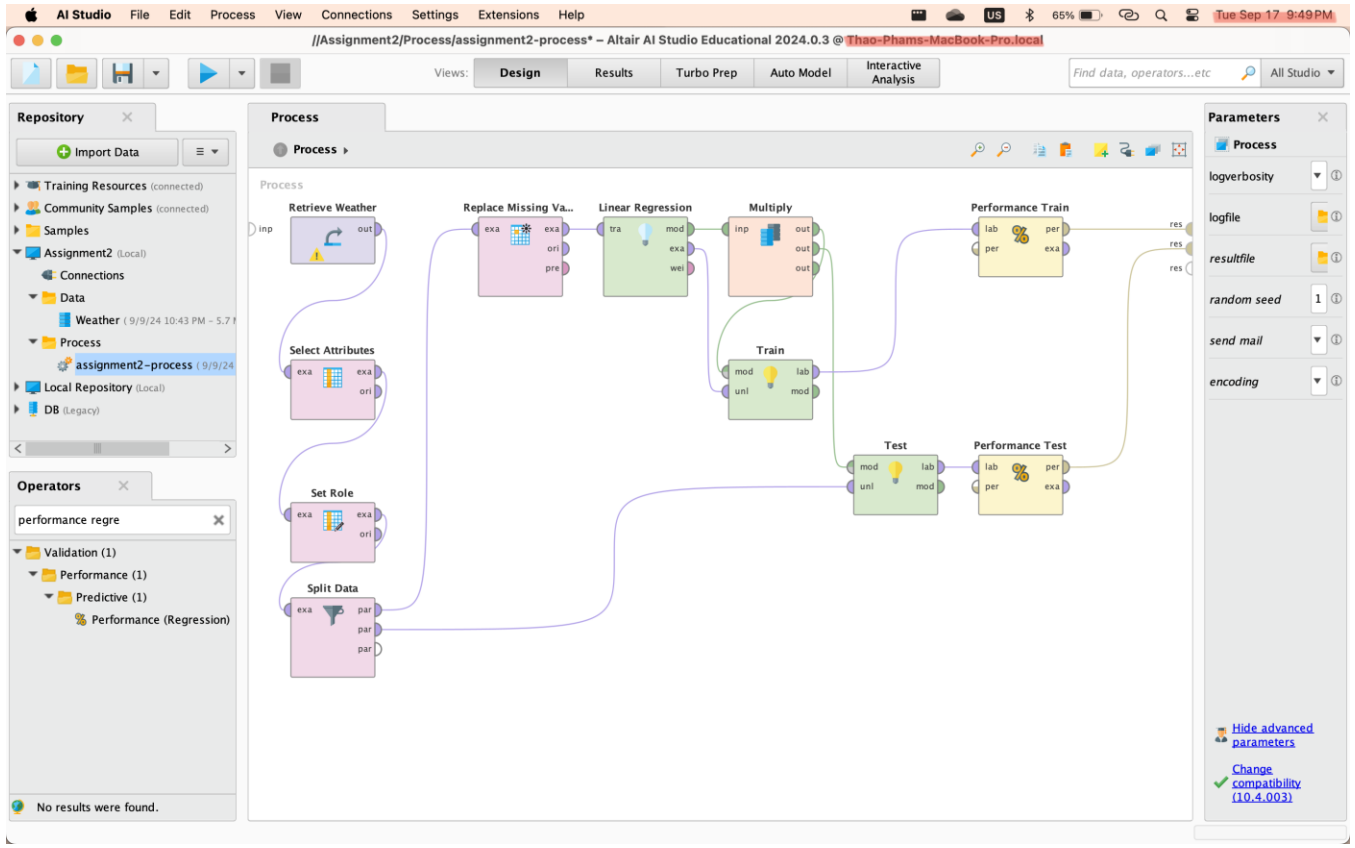
Operation tab, type **Linear Regression** .

Drag and drop to the process tab.

Connecting **Replace Missing Values.example set output** with **Linear Regression.training set**

Keep it as default in the **Parameter** tab.

MULTIPLY

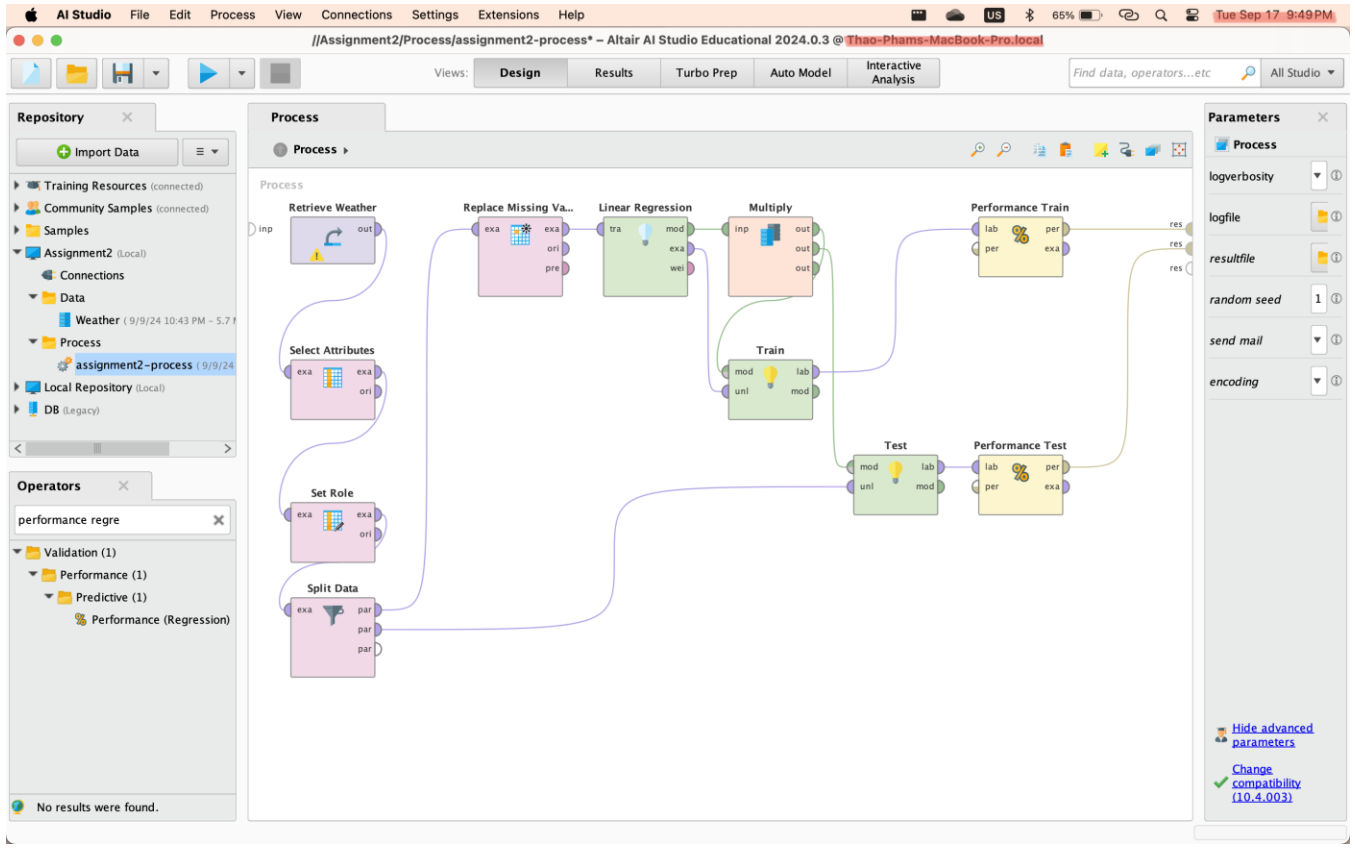


Operation tab, type **Multiply**.

Drag and drop to the process tab.

Connecting **Linear Regression.model** to **Multiply.input**

APPLY MODEL - TRAIN



Operation tab, type **Apply Model**.

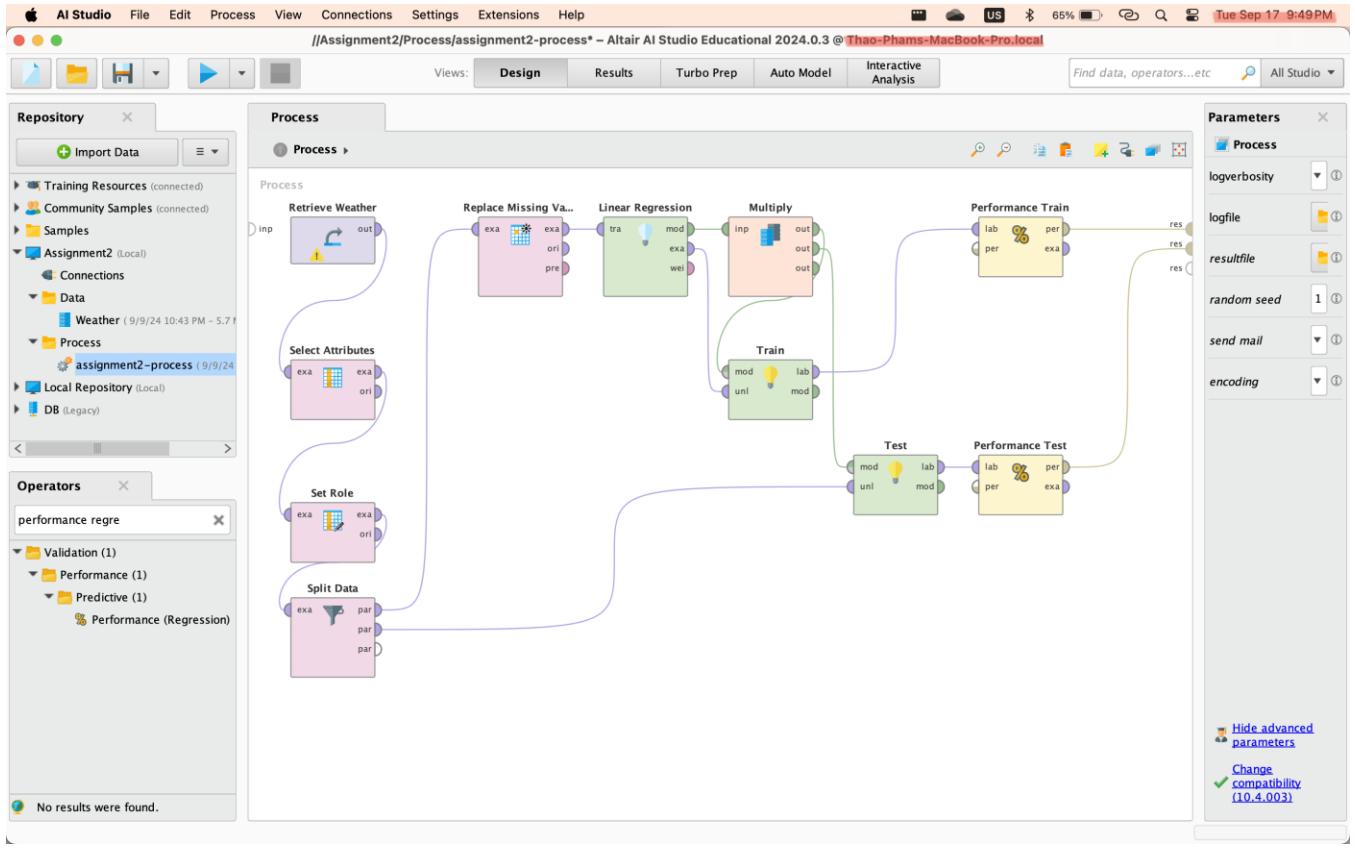
Drag and drop to the process tab.

Double click on its name to change to **Train**

Connecting **Linear Regression.exampleSet** to **Train.unlabelled data**

Connecting **Multiply.output 1** to **Train.model**

APPLY MODEL - TEST



Operation tab, type **Apply Model**.

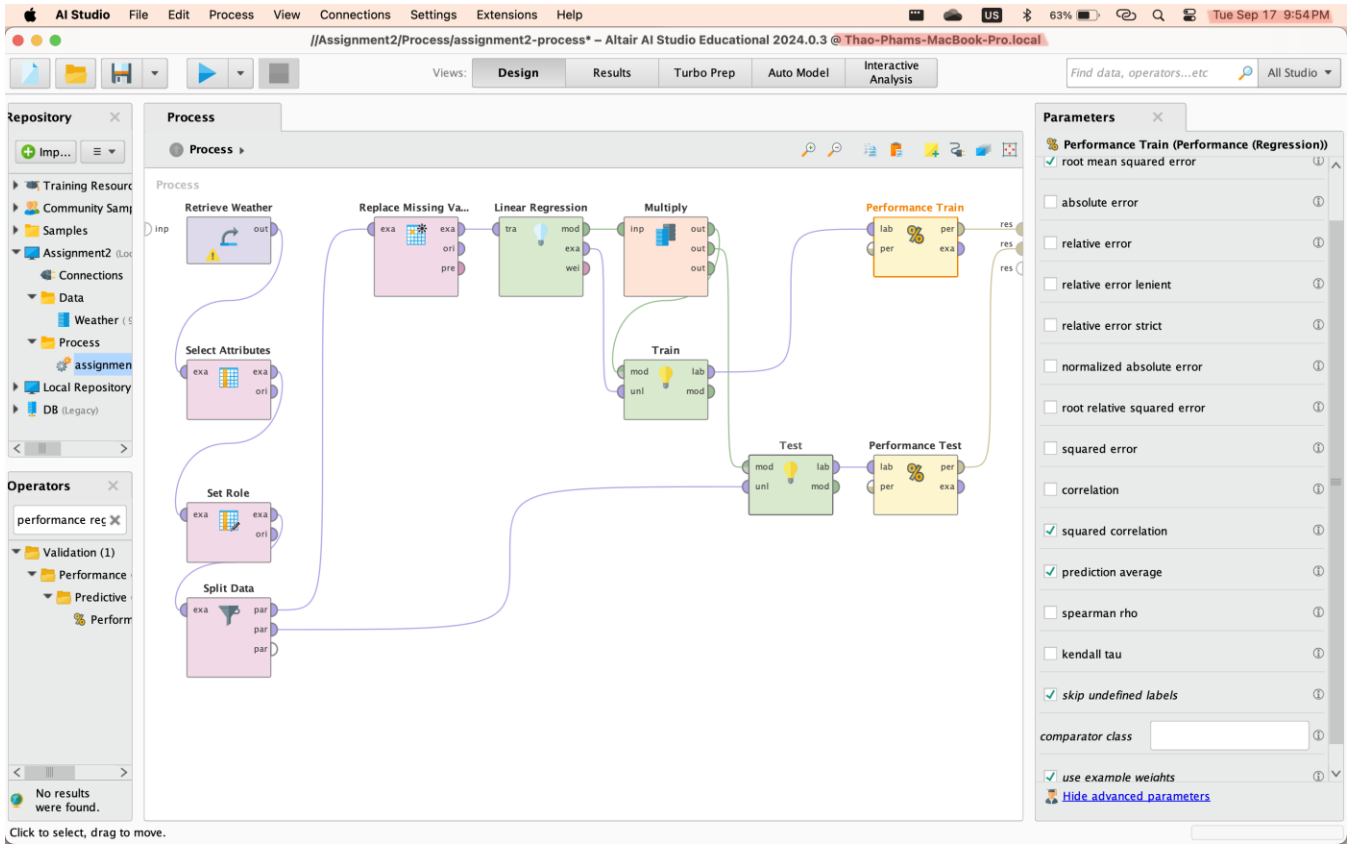
Drag and drop to the process tab.

Double click on its name to change to **Test**

Connecting **Multiply.output 2** to **Test.model**

Connecting **Split Data.partition 2** to **Test.unlabelled data**

PERFORMANCE-TRAIN



Operation tab, type **Performance (Regression)**.

Drag and drop to the process tab.

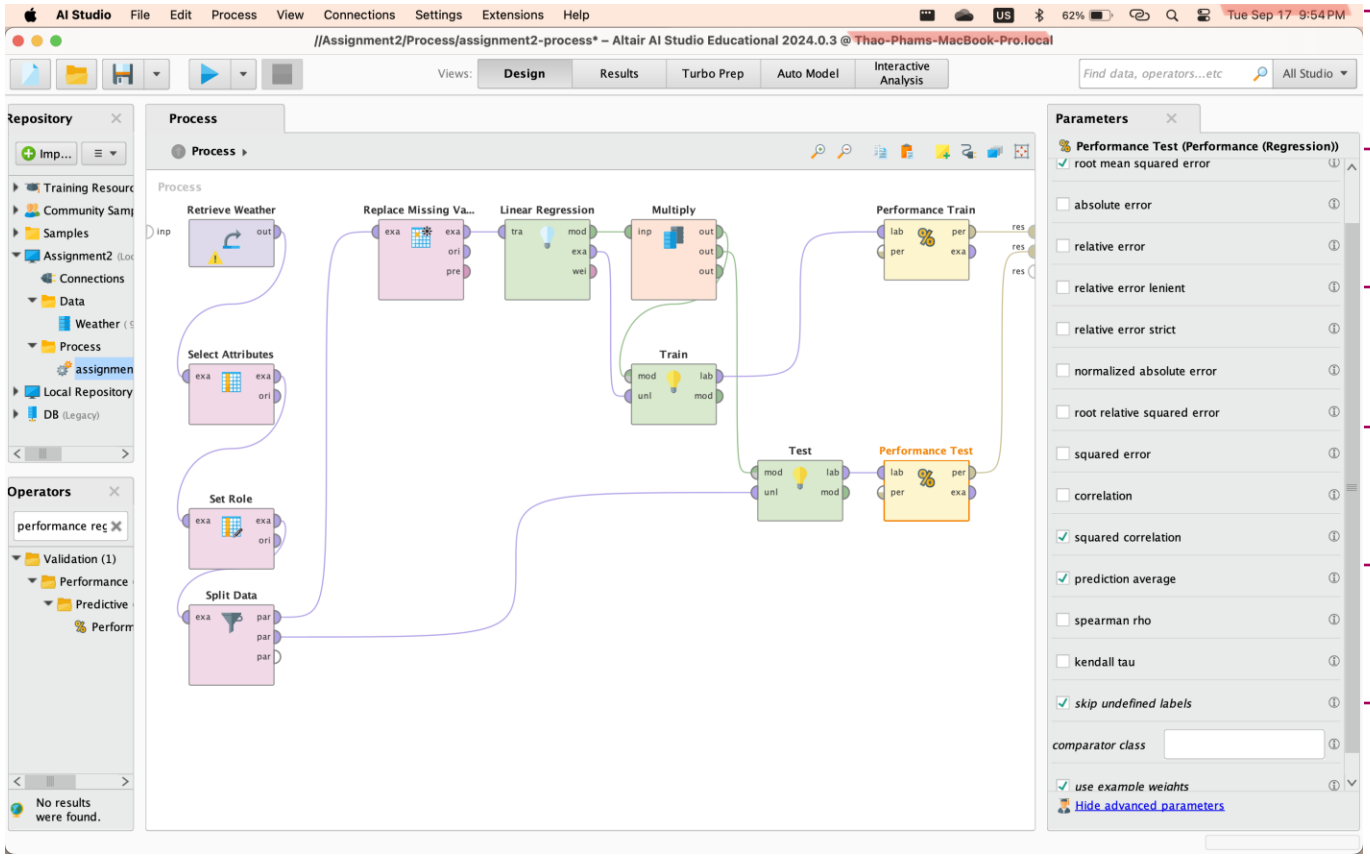
Double click on its name to change to **Performance Train**.

Connecting **Train.labelled data** to **Performance Train.labelled data**

Connecting **Performance Train.performance** to **Process.result 1**

Parameters tab, selecting **root mean squared errors, squared correlation, prediction average, skip undefined labels, use example weights.**

PERFORMANCE-TEST



Operation tab, type **Performance (Regression)**.

Drag and drop to the process tab.

Double click on its name to change to **Performance Test**.

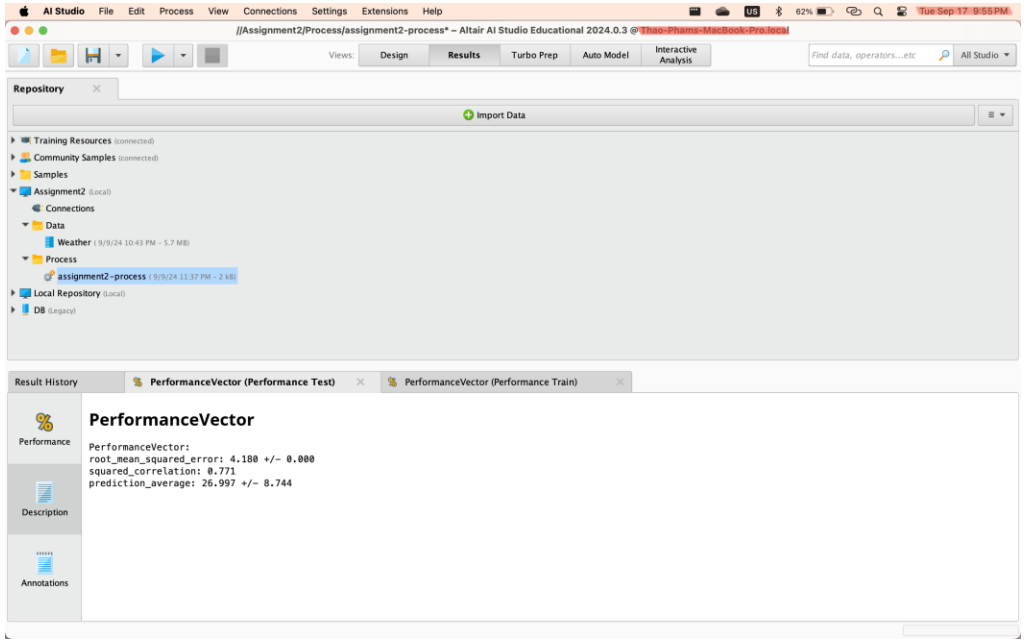
Connecting **Test.labelled data** to **Performance Test.labelled data**

Performance Test.performance to **Process.result 2**

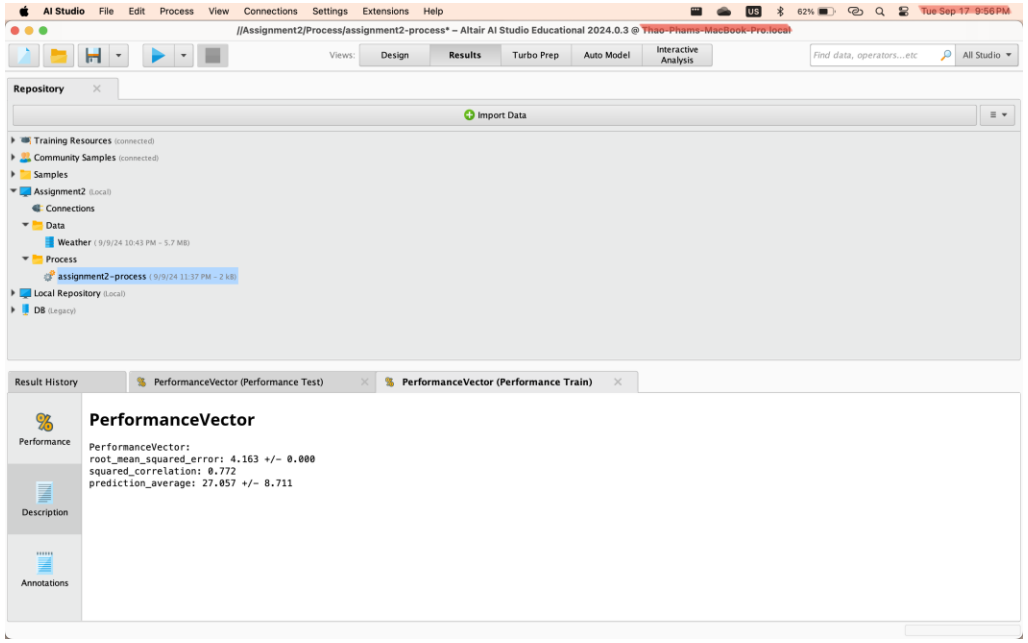
Parameters tab, selecting **root mean squared errors, squared correlation, prediction average, skip undefined labels, use example weights.**

PERFORMANCE RESULTS

Performance Test Results



Performance Train Results



INTERPRETING THE RESULTS

- The results of both models are very similar to each other:
  - Root Mean Square Error: 4.163 - 4.18, which means on average, the model's predictions for maximum temperature will be off by 4 degrees off from the actual maximum temperature.
  - Squared Correlation: 0.771 - 0.772, on the range 0 to 1, where a value closer to 1 means a stronger linear relationship between the max and min temperatures. This means around 77% of the variance in the maximum temperature can be predicted by the minimum temperature.
  - Prediction Average: 26.997 – 27.057, which are the average predicted maximum temperature.

CONCLUSION

- Based on the results, there is a strong and statistically significant relationship between the daily minimum and maximum temperatures.
- Yes, we can predict the maximum temperature given the minimum temperature.

RESEARCH ITEMS

- From Operator:
  - Retrieve:
    - Access data that is imported in the Repository and load them into the Process (GmbH, n.d.-f).
    - Parameters: repository entry which is the path to the object should be loaded (GmbH, n.d.-f).
    - Returns an object whose path was specified in repository entry parameter (GmbH, n.d.-f),
  - Select Attributes:
    - Selects a subset of Attributes of an ExampleSet and removes the other attributes (GmbH, n.d.-g).
    - Its inputs are the data was loaded (GmbH, n.d.-g).
    - Data that was selected is delivered to the output port (GmbH, n.d.-g).
    - Parameters: types, attribute filter type, select attribute, select subset, expression, exclude expression, type of value, special attributes (GmbH, n.d.-g).

RESEARCH ITEMS

- Select Role:
  - Changes the role of one or more attributes (GmbH, n.d.-h).
  - Parameters: attribute_name: the variable, target_role: the new role assigned to it (GmbH, n.d.-h).
  - Receiving inputs then modified role is output of this operator (GmbH, n.d.-h).
- Split Data:
  - this tool divides a dataset into a specified number of subsets with specific proportions (GmbH, n.d.-i).
  - Its input is the output of the previous operator, for example, output of the retrieve operator (GmbH, n.d.-i).
  - Can output multiple partition ports (GmbH, n.d.-i).
  - Parameters: partitions, sampling_type, use_local_random_seed, local_random_seed (GmbH, n.d.-i).

RESEARCH ITEMS

- Replace Missing Data:
  - learnt this tool replace missing data with a value and the replace type of numerical in the parameters (GmbH, n.d.-e).
  - Parameters allows the method you want to use for selecting attributes (GmbH, n.d.-e).
- Linear regression:
  - this operator calculates a linear regression model from the input ExampleSet (GmbH, n.d.-b).
  - Input only numeric attributes, so the Nominal to Numerical operator should be used before this operation if you've nominal attributes (GmbH, n.d.-b).
  - Outputs linear regression model (GmbH, n.d.-b).

RESEARCH ITEMS

- Multiply:
  - takes object from the input port and delivers copies of it to the output ports, while not effecting other copies if one is changed (GmbH, n.d.-c)
  - Inputs any Rapidminer Object that should be copies (GmbH, n.d.-c).
  - Outputs the copy of the input object. An output port is created for more copies when an output port is connected. Therefore, all ports don't change copies of the input objects (GmbH, n.d.-c).
- Apply Model:
  - the operator applies a model on a dataset (GmbH, n.d.-a).
  - Inputs a model and unlabeled data (GmbH, n.d.-a).
  - Outputs a model and labelled data (GmbH, n.d.-a).
  - Parameters: applications_parameter, create_view (GmbH, n.d.-a).

RESEARCH ITEMS

- Performance:
  - Assesses how well a regression model performs by providing various statistical metrics (GmbH, n.d.-c).
  - Can only be used with regression tasks only (GmbH, n.d.-c).
  - Inputs labeled data and performance (GmbH, n.d.-c).
  - Outputs performance and dataset (GmbH, n.d.-c).
  - Parameters: root mean squared error, absolute error, relative error, relative error lenient, relative error strict, correlation, squared error, squared correlation, and so on (GmbH, n.d.-c).

REFERENCE

GmbH, R. (n.d.-a). *Apply Model - RapidMiner Documentation*. https://docs.rapidminer.com/10.2/studio/operators/scoring/apply_model.html

GmbH, R. (n.d.-b). *Linear Regression - RapidMiner Documentation*.

https://docs.rapidminer.com/9.8/studio/operators/modeling/predictive/functions/linear_regression.html

GmbH, R. (n.d.-c). *Multiply - RapidMiner Documentation*.

https://docs.rapidminer.com/9.7/studio/operators/utility/multiply.html#:~:text=Description,no%20effect%20on%20other%20copies.

GmbH, R. (n.d.-d). *Performance (Regression) - Altair RapidMiner Documentation*.

https://docs.rapidminer.com/latest/studio/operators/validation/performance/predictive/performance_regression.html#:~:text=The%20Performance%20(Regression)%20operator%20is,common%20criteria%20for%20that%20type.

REFERENCE

GmbH, R. (n.d.-e). *Replace missing values - Altair RapidMiner Documentation*. https://docs.rapidminer.com/latest/studio/operators/cleansing/missing/replace_missing_values.html

GmbH, R. (n.d.-f). *Retrieve - Altair RapidMiner Documentation*. https://docs.rapidminer.com/latest/studio/operators/data_access/retrieve.html

GmbH, R. (n.d.-g). *Select Attributes - Altair RapidMiner Documentation*. https://docs.rapidminer.com/latest/studio/operators/blending/attributes/selection/select_attributes.html#:~:text=The%20first%20Select%20Attributes%20Operator,choosing%20binominal%20and%20non%2Dbinominal.

GmbH, R. (n.d.-h). *Set Role - RapidMiner Documentation*. https://docs.rapidminer.com/9.1/studio/operators/blending/attributes/names_and_roles/set_role.html

GmbH, R. (n.d.-i). *Split Data - RapidMiner Documentation*. https://docs.rapidminer.com/9.6/studio/operators/blending/examples/sampling/split_data.html

Weather conditions in World War Two. (2017, November 1). Kaggle. https://www.kaggle.com/datasets/smid80/weatherww2/data

THE END

THANK YOU FOR READING

THAO PHAM