Name : Kaushik Kotian **Roll no.** 30 **Div :** D15B **Batch:**B

DMBI Lab

EXPERIMENT NO. 7

AIM : To implement a regression model using Rapid Miner and Python.

1. Preprocess data. Split data into train and test set
2. Build Regression model using inbuilt library function on training data
3. Calculate metrics based on test data using inbuilt function
4. Build a Regression model using a function defined by the student.
5. Calculate metrics based on test data using inbuilt function
6. Compare the results of all three ways of implementation.(Rapid Miner, Python Library)

Theory :

To implement a regression model using RapidMiner and Python, you can leverage both user-defined functions and built-in functions. Here's a general outline of how you can approach this:

Data Preparation:

- Load your dataset into RapidMiner for preprocessing. This may involve cleaning missing values, handling categorical variables, and scaling numeric features.
- Export the preprocessed data from RapidMiner to a format compatible with Python, such as CSV or Excel.

Regression Model Building in RapidMiner:

- Use RapidMiner's built-in operators for regression analysis, such as Linear Regression, Decision Tree Regression, or Support Vector Regression, depending on your data and problem.
- Configure the parameters of the regression model within RapidMiner, such as selecting input variables, setting regularization options, and specifying the target variable.

Exporting the Model:

- Once you have trained and validated your regression model in RapidMiner, export the model as a file (e.g., PMML format) that can be loaded into Python.

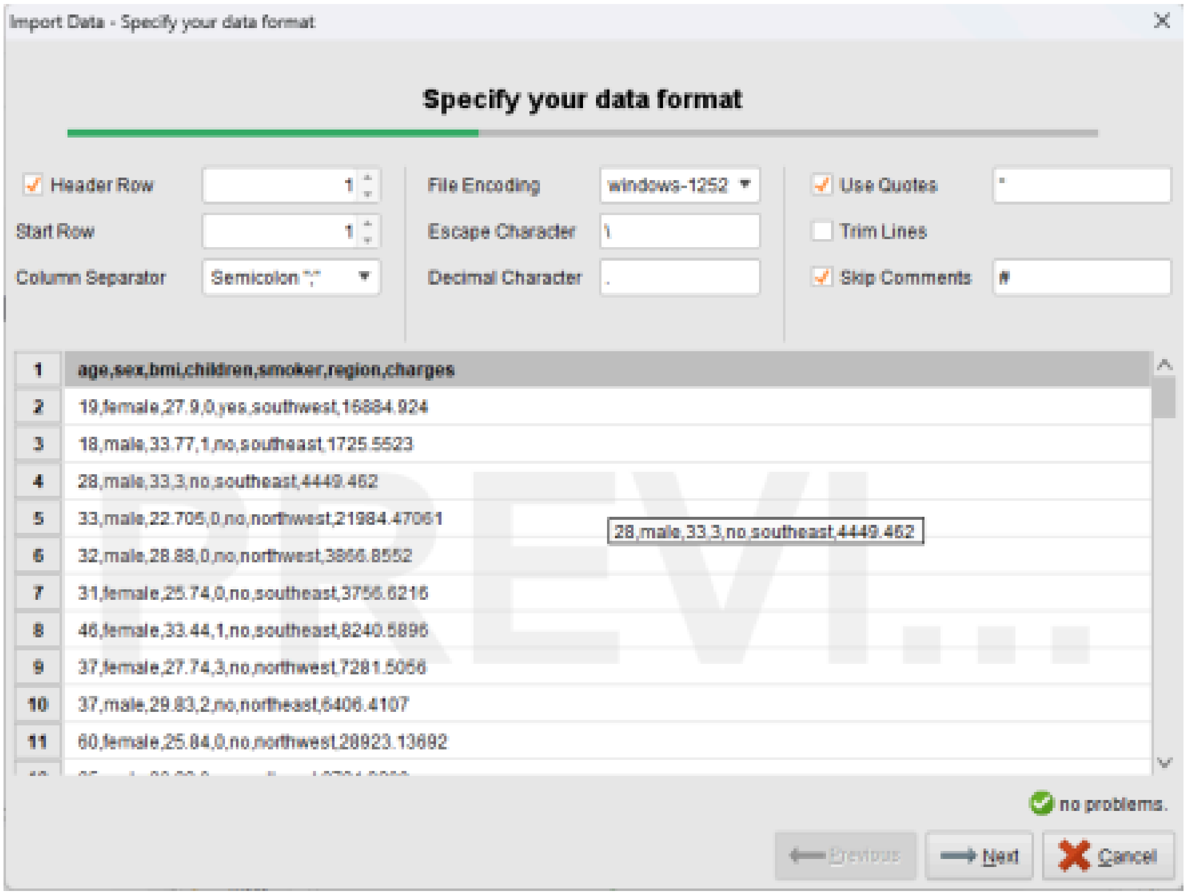
Loading the Model in Python:

- Use Python libraries such as pandas to load the preprocessed data and sklearn to load the exported regression model from RapidMiner.
- If needed, define custom functions in Python for any specific data transformations or model evaluation metrics that are not directly available in RapidMiner.

Prediction and Evaluation:

- Use the loaded regression model in Python to make predictions on new data or evaluate its performance on a test dataset.
- Implement evaluation metrics such as Mean Squared Error (MSE), R-squared, or others to assess the model's accuracy and reliability.

IMPLEMENTATION USING RAPID MINER

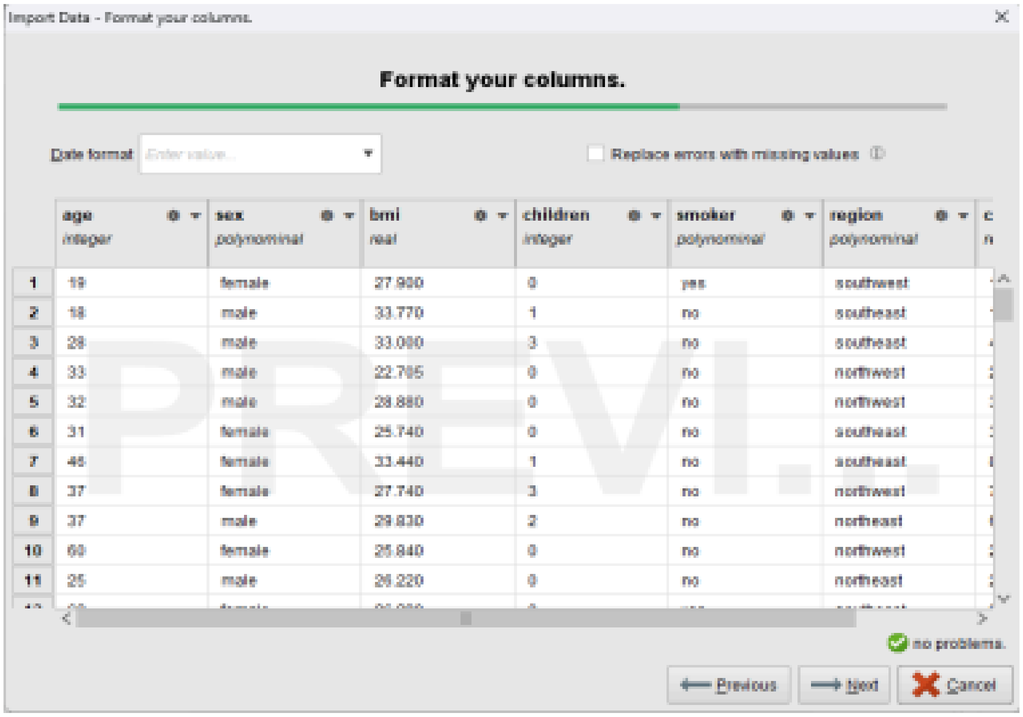


☑ Header Row: 1
Start Row: 1
Column Separator: Semicolon ";"
File Encoding: windows-1252
Escape Character: \
Decimal Character: .
☑ Use Quotes: "
□ Trim Lines
☑ Skip Comments: #

```
1   age,sex,bmi,children,smoker,region,charges
2   19,female,27.9,0,yes,southwest,16884.924
3   18,male,33.77,1,no,southeast,1725.5523
4   28,male,33,3,no,southeast,4449.462
5   33,male,22.705,0,no,northwest,21984.47061
6   32,male,28.88,0,no,northwest,3866.8552
7   31,female,25.74,0,no,southeast,3756.6216
8   46,female,33.44,1,no,southeast,8240.5896
9   37,female,27.74,3,no,northwest,7281.5056
10  37,male,29.83,2,no,northeast,6406.4107
11  60,female,25.84,0,no,northwest,28923.13692
```

28,male,33,3,no,southeast,4449.462

✓ no problems.



Import Data - Format your columns.

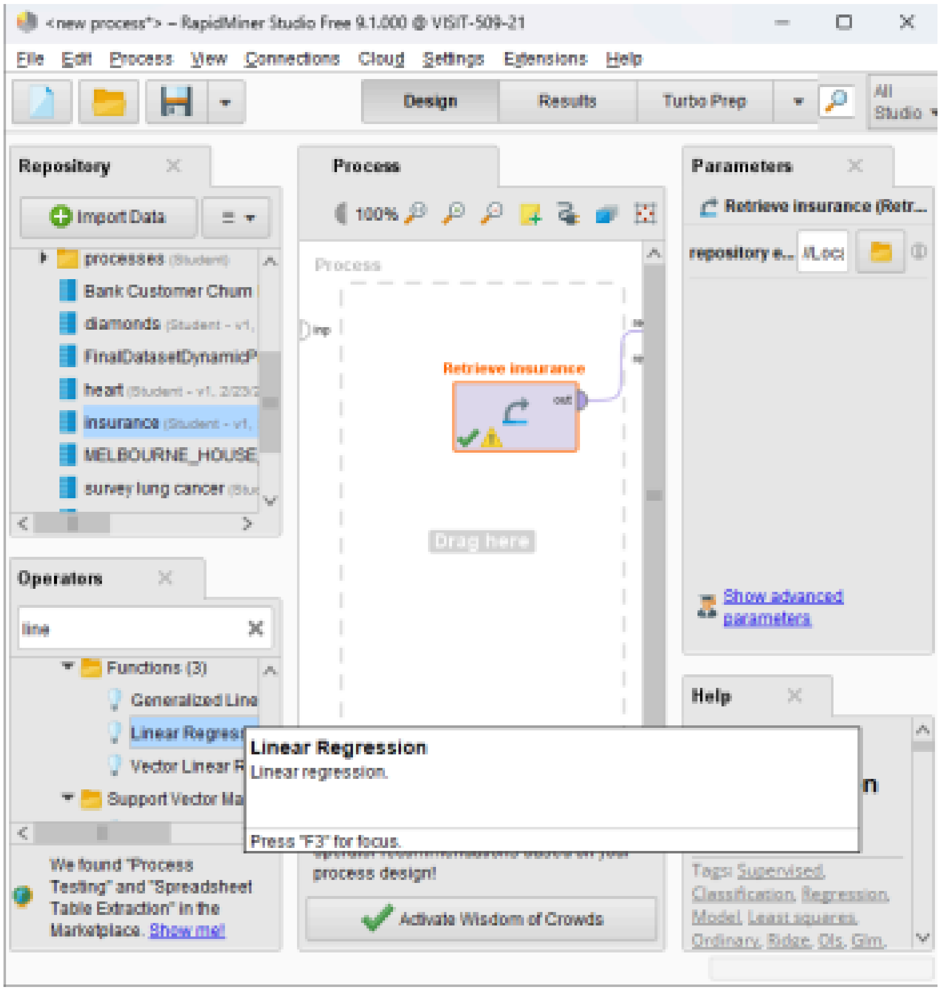
Format your columns.

Date format: Enter value...

☐ Replace errors with missing values

	age *integer*	sex *polynominal*	bmi *real*	children *integer*	smoker *polynominal*	region *polynominal*	c n
1	19	female	27.900	0	yes	southwest	
2	18	male	33.770	1	no	southeast	
3	28	male	33.000	3	no	southeast	
4	33	male	22.705	0	no	northwest	
5	32	male	28.880	0	no	northwest	
6	31	female	25.740	0	no	southeast	
7	46	female	33.440	1	no	southeast	
8	37	female	27.740	3	no	northwest	
9	37	male	29.830	2	no	northeast	
10	60	female	25.840	0	no	northwest	
11	25	male	26.220	0	no	northeast	

✓ no problems.



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21

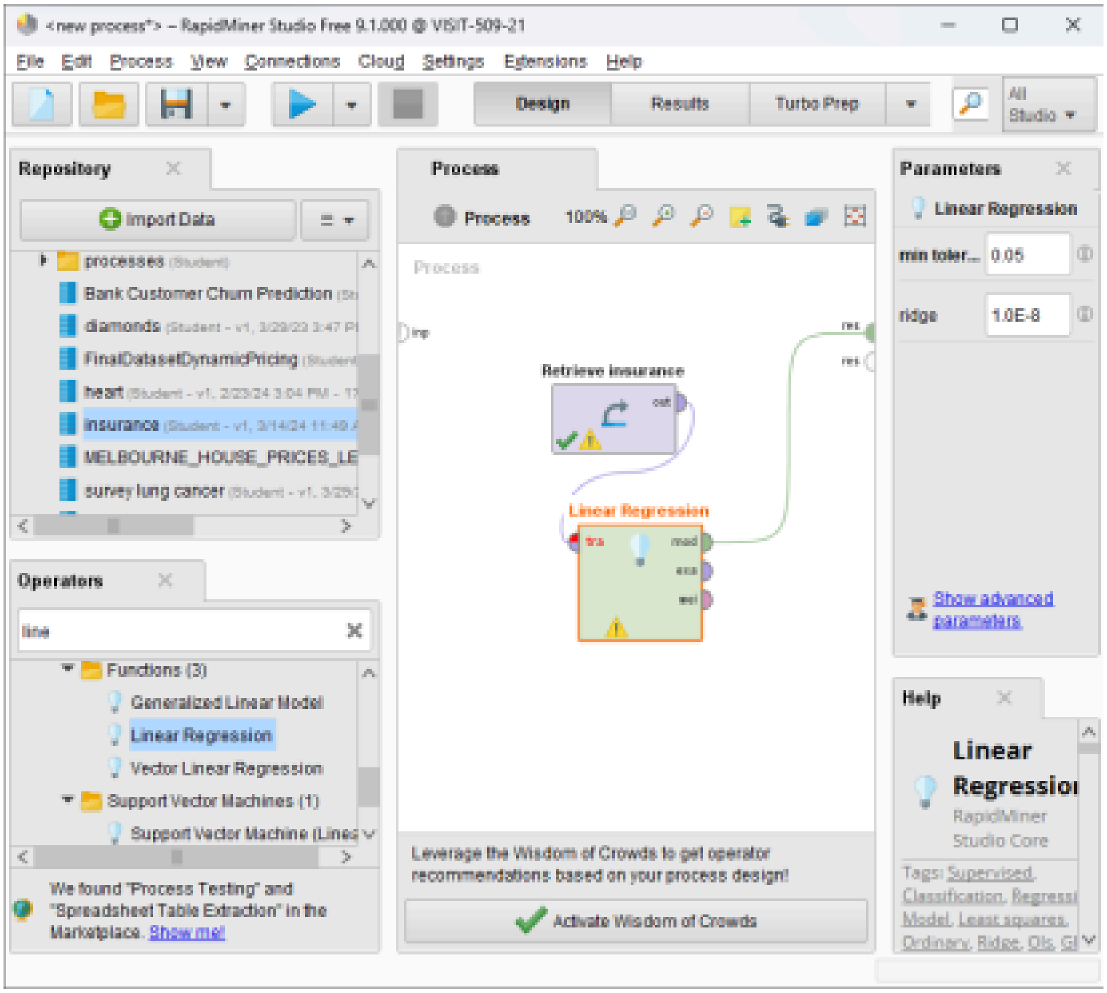
Linear Regression

Linear regression.

Press "F3" for focus.

process design!

We found "Process Testing" and "Spreadsheet Table Extraction" in the Marketplace. Show me!



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21

- processes (Student)
- Bank Customer Churn Prediction (St…
- diamonds (Student - v1, 3/29/23 3:47 P…
- FinalDatasetDynamicPricing (Student…
- heart (Student - v1, 2/23/24 3:04 PM - 17…
- insurance (Student - v1, 3/14/24 11:49 A…
- MELBOURNE_HOUSE_PRICES_LE…
- survey lung cancer (Student - v1, 3/29/…

line

- Functions (3)
  - Generalized Linear Model
  - Linear Regression
  - Vector Linear Regression
- Support Vector Machines (1)
  - Support Vector Machine (Linea…

We found "Process Testing" and "Spreadsheet Table Extraction" in the Marketplace. Show me!

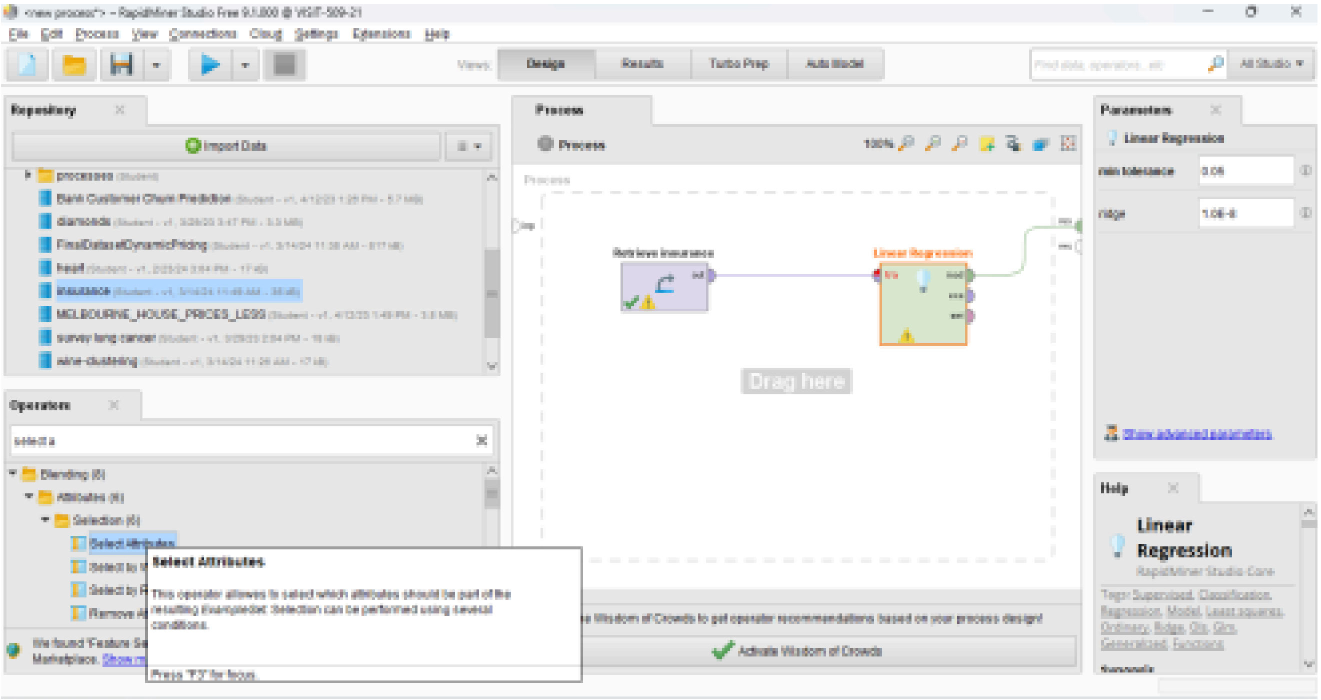
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Linear Regression

min toler…	0.05
ridge	1.0E-8

Linear Regression
RapidMiner Studio Core

Tags: Supervised, Classification, Regressi… Model, Least squares, Ordinary, Ridge, Ols, Gl…



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21

- processes (Student)
- Bank Customer Churn Prediction (Student - v1, 4/12/23 1:26 PM - 5.7 MB)
- diamonds (Student - v1, 3/29/23 3:47 PM - 3.3 MB)
- FinalDatasetDynamicPricing (Student - v1, 3/14/24 11:38 AM - 817 kB)
- heart (Student - v1, 2/23/24 3:04 PM - 17 kB)
- insurance (Student - v1, 3/14/24 11:49 AM - 38 kB)
- MELBOURNE_HOUSE_PRICES_LESS (Student - v1, 4/12/23 1:49 PM - 3.8 MB)
- survey lung cancer (Student - v1, 3/29/23 2:04 PM - 16 kB)
- wine-clustering (Student - v1, 3/14/24 11:26 AM - 17 kB)

select a

- Blending (6)
  - Attributes (6)
    - Selection (6)
      - Select Attributes
      - Select by W…
      - Select by R…
      - Remove A…

We found "Feature Se… Marketplace. Show m…

Select Attributes

This operator allows to select which attributes should be part of the resulting ExampleSet. Selection can be performed using several conditions.

Press 'F3' for focus.

…e Wisdom of Crowds to get operator recommendations based on your process design!

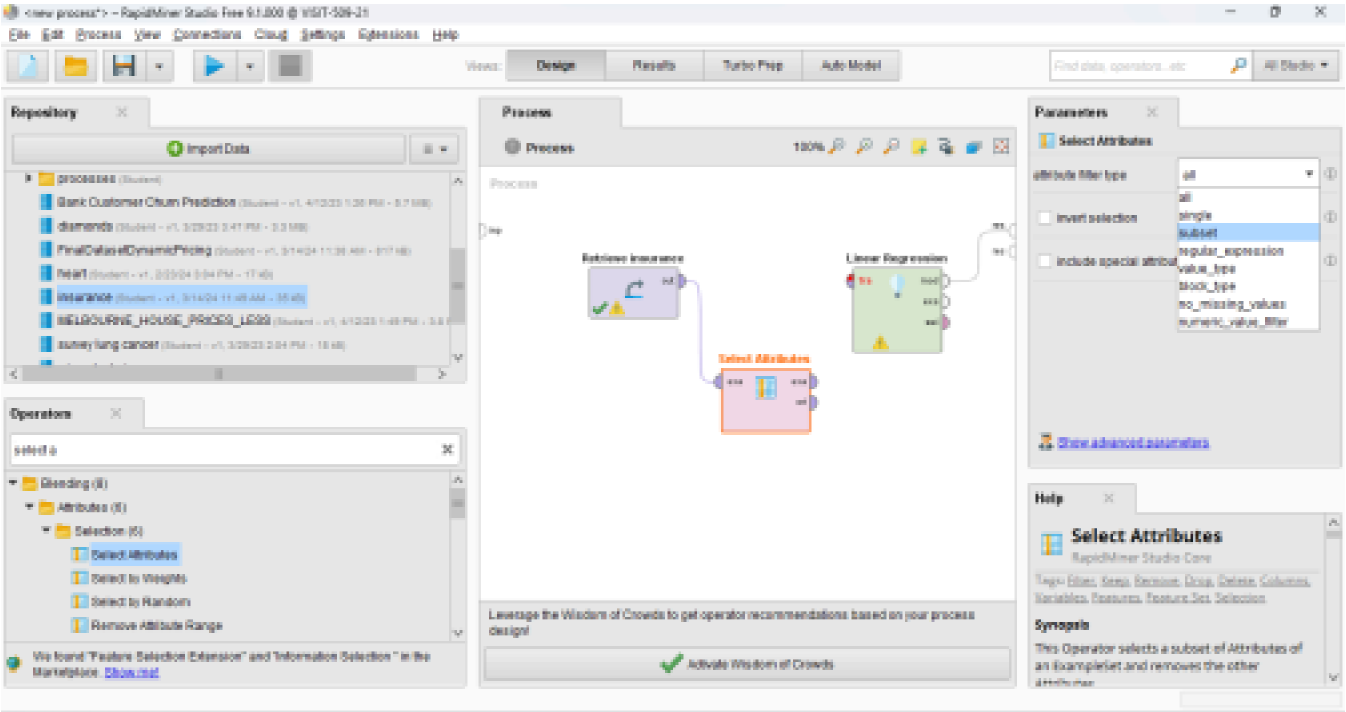
Linear Regression

min tolerance	0.05
ridge	1.0E-8

Linear Regression
RapidMiner Studio Core

Tags: Supervised, Classification, Regression, Model, Least squares, Ordinary, Ridge, Ols, Glm, Generalized, Functions

Synopsis



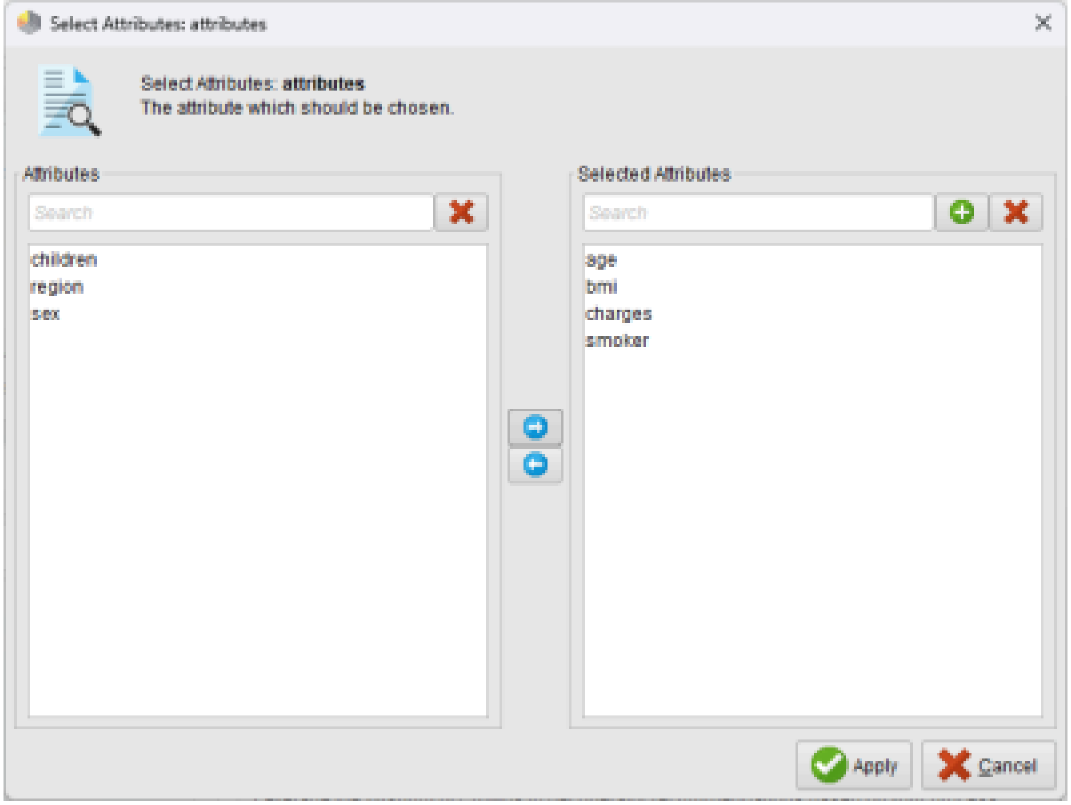
We found "Feature Selection Extension" and "Information Selection" in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Tags: Filter, Keep, Remove, Drop, Delete, Columns, Variables, Features, Feature Set, Selection

Synopsis

This Operator selects a subset of Attributes of an ExampleSet and removes the other attributes



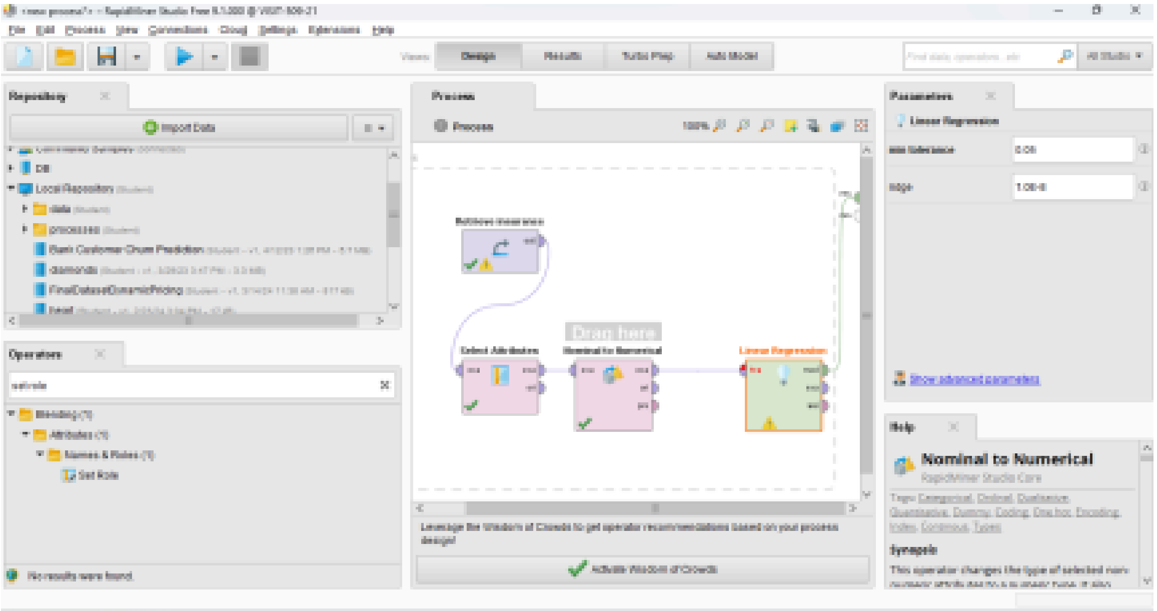
Select Attributes: **attributes**
The attribute which should be chosen.

Attributes

children
region
sex

Selected Attributes

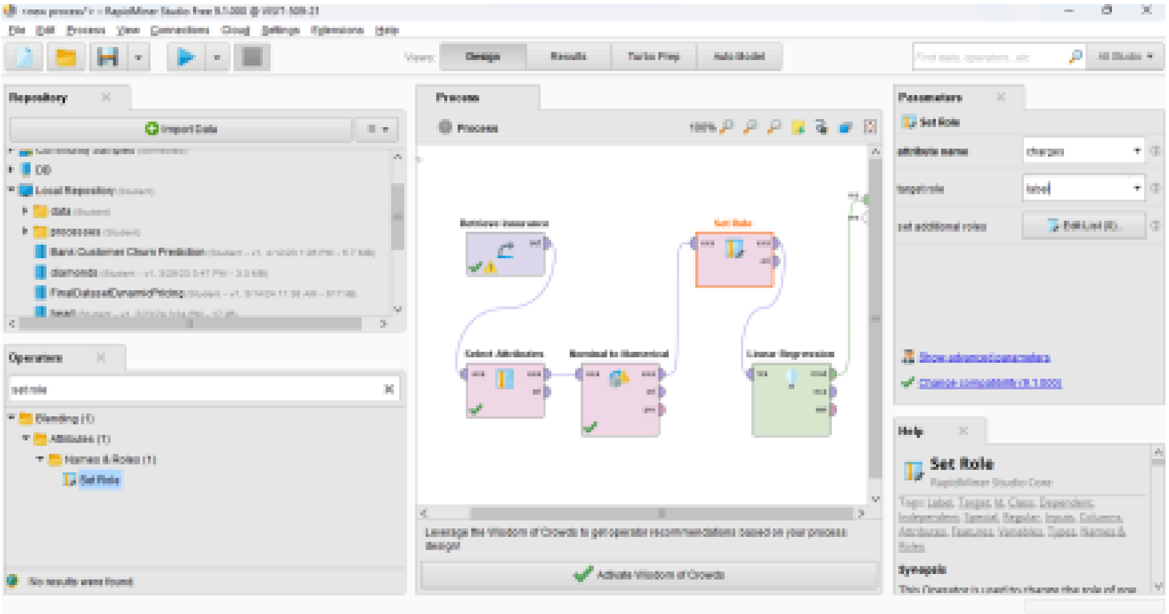
age
bmi
charges
smoker



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Community Samples (connected)
DB
Local Repository (Student)
data (Student)
processes (Student)
Bank Customer Churn Prediction
diamonds
FinalDatasetDynamicPricing
heart
Operators
setrole
Blending (1)
Attributes (1)
Names & Roles (1)
Set Role
No results were found.
Process
Retrieve insurance
Select Attributes
Nominal to Numerical
Linear Regression
Drag here
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Linear Regression
Show advanced parameters
Help
Nominal to Numerical
RapidMiner Studio Core
Tags: Categorical, Ordinal, Qualitative, Quantitative, Dummy, Coding, One hot, Encoding, Index, Continous, Types
Synopsis
This operator changes the type of selected non-numeric attributes to a numeric type. It also

Linear Regression

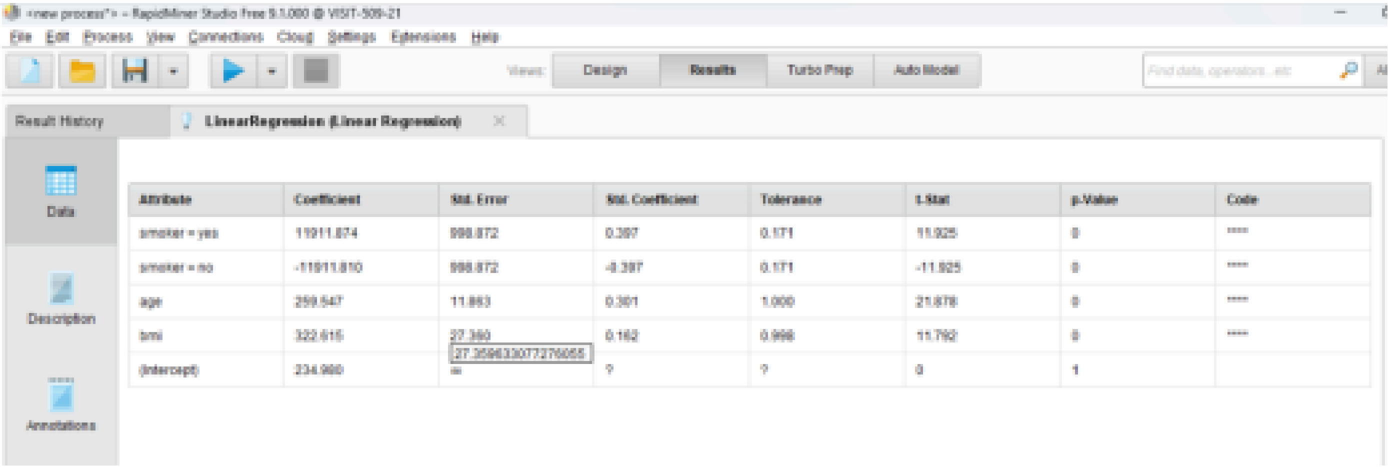
min tolerance	0.05
ridge	1.0E-8



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Community Samples (connected)
DB
Local Repository (Student)
data (Student)
processes (Student)
Bank Customer Churn Prediction
diamonds
FinalDatasetDynamicPricing
heart
Operators
set role
Blending (1)
Attributes (1)
Names & Roles (1)
Set Role
No results were found.
Process
Retrieve insurance
Select Attributes
Nominal to Numerical
Set Role
Linear Regression
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Set Role
Edit List (0)...
Show advanced parameters
Change compatibility (9.1.000)
Help
Set Role
RapidMiner Studio Core
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles
Synopsis
This Operator is used to change the role of one

Set Role

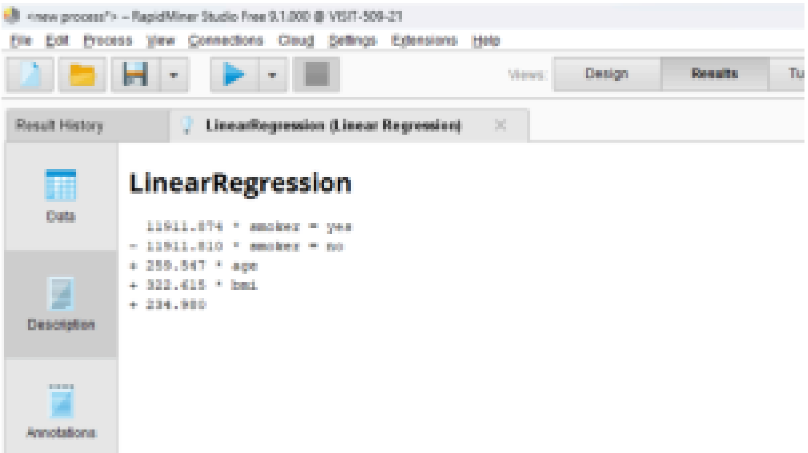
attribute name	charges
target role	label
set additional roles	Edit List (0)...



<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
Result History
LinearRegression (Linear Regression)
Data
Description
Annotations

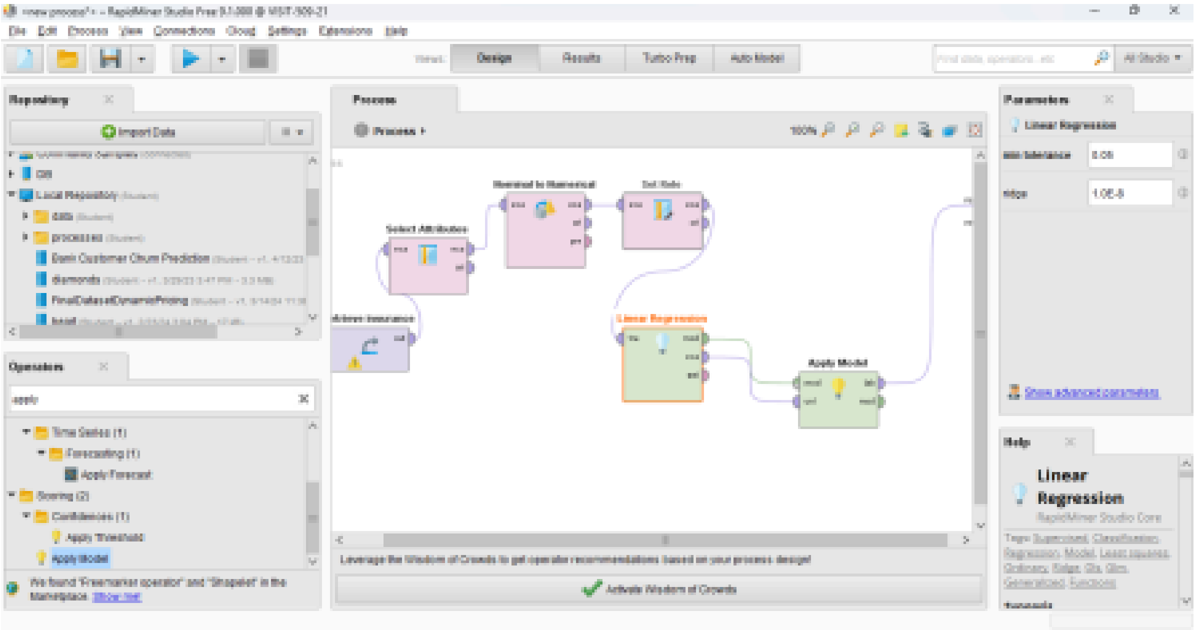
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
smoker = yes	11911.874	998.872	0.397	0.171	11.925	0	****
smoker = no	-11911.810	998.872	-0.397	0.171	-11.925	0	****
age	259.547	11.863	0.301	1.000	21.878	0	****
bmi	322.615	27.360	0.162	0.998	11.792	0	****
(intercept)	234.900	∞	?	?	0	1	

27.359633077276055


<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Tu
Result History
LinearRegression (Linear Regression)
Data
Description
Annotations

LinearRegression

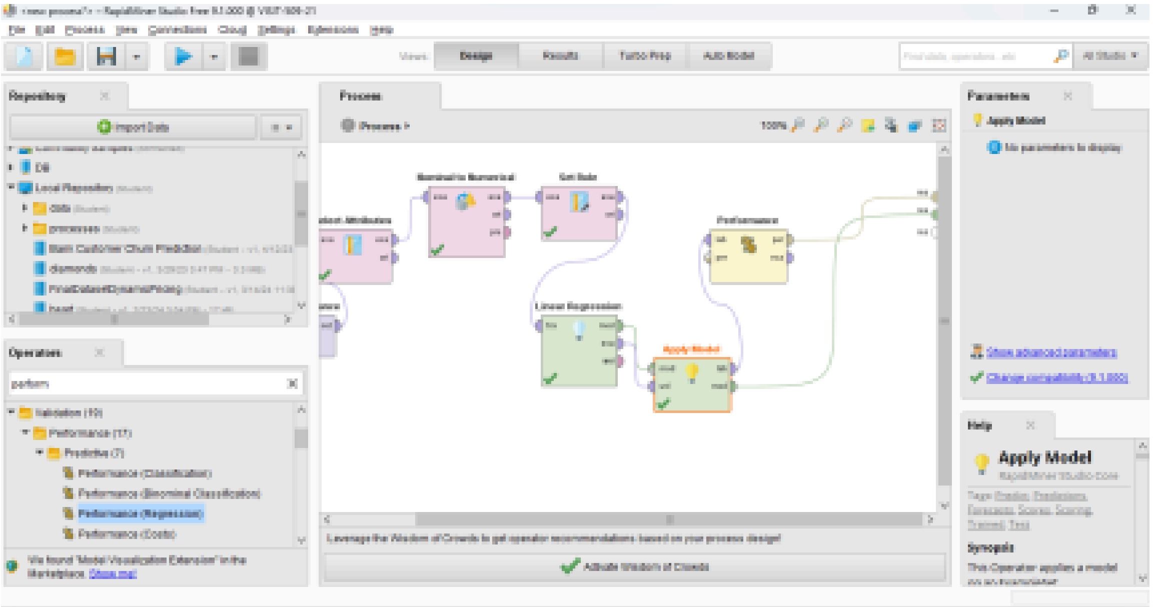
```
  11911.874 * smoker = yes
- 11911.810 * smoker = no
+ 259.547 * age
+ 322.615 * bmi
+ 234.980
```



min tolerance 0.05

ridge 1.0E-8

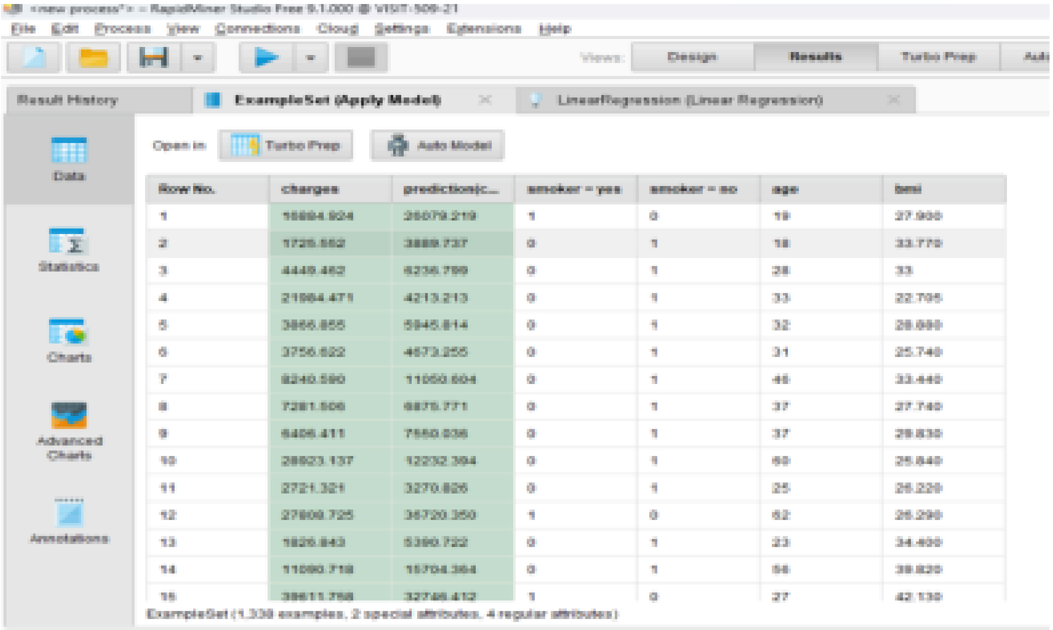
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!



No parameters to display

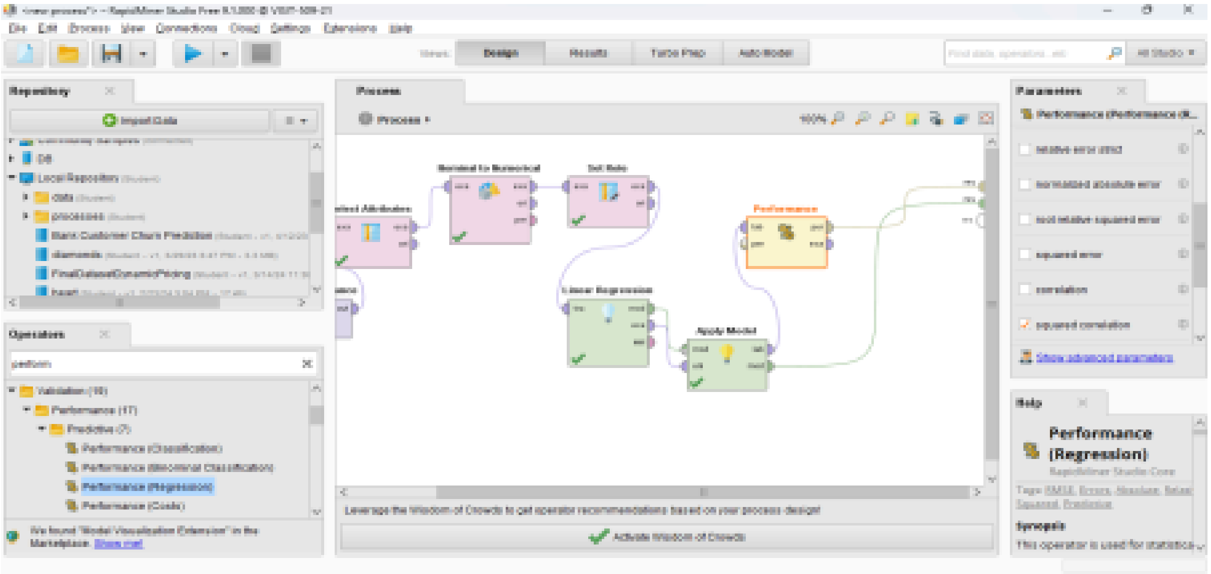
This Operator applies a model on an ExampleSet.

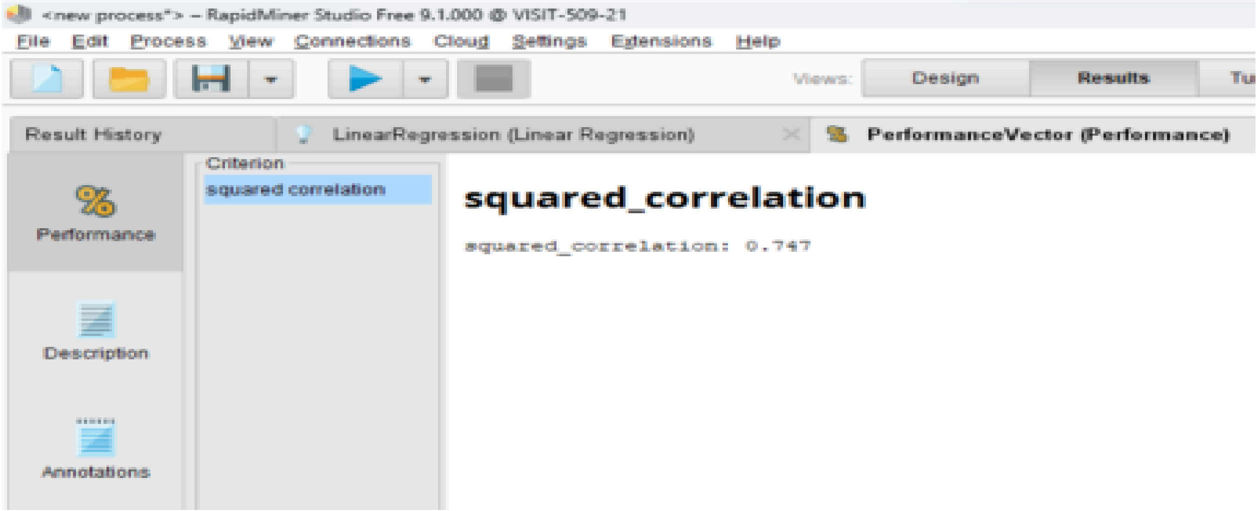
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!


<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto
Result History
ExampleSet (Apply Model)
LinearRegression (Linear Regression)
Data
Statistics
Charts
Advanced Charts
Annotations
Open in Turbo Prep Auto Model

Row No.	charges	prediction(c...	smoker = yes	smoker = no	age	bmi
1	16884.924	25079.219	1	0	19	27.900
2	1725.552	3889.737	0	1	18	33.770
3	4449.462	6236.799	0	1	28	33
4	21984.471	4213.213	0	1	33	22.705
5	3866.855	5945.814	0	1	32	28.880
6	3756.622	4673.255	0	1	31	25.740
7	8240.590	11050.604	0	1	46	33.440
8	7281.506	6875.771	0	1	37	27.740
9	6406.411	7550.036	0	1	37	29.830
10	28923.137	12232.394	0	1	60	25.840
11	2721.321	3270.826	0	1	25	26.220
12	27808.725	36720.350	1	0	62	26.290
13	1826.843	5390.722	0	1	23	34.400
14	11090.718	15704.364	0	1	56	39.820
15	39611.758	32746.412	1	0	27	42.130

ExampleSet (1,338 examples, 2 special attributes, 4 regular attributes)

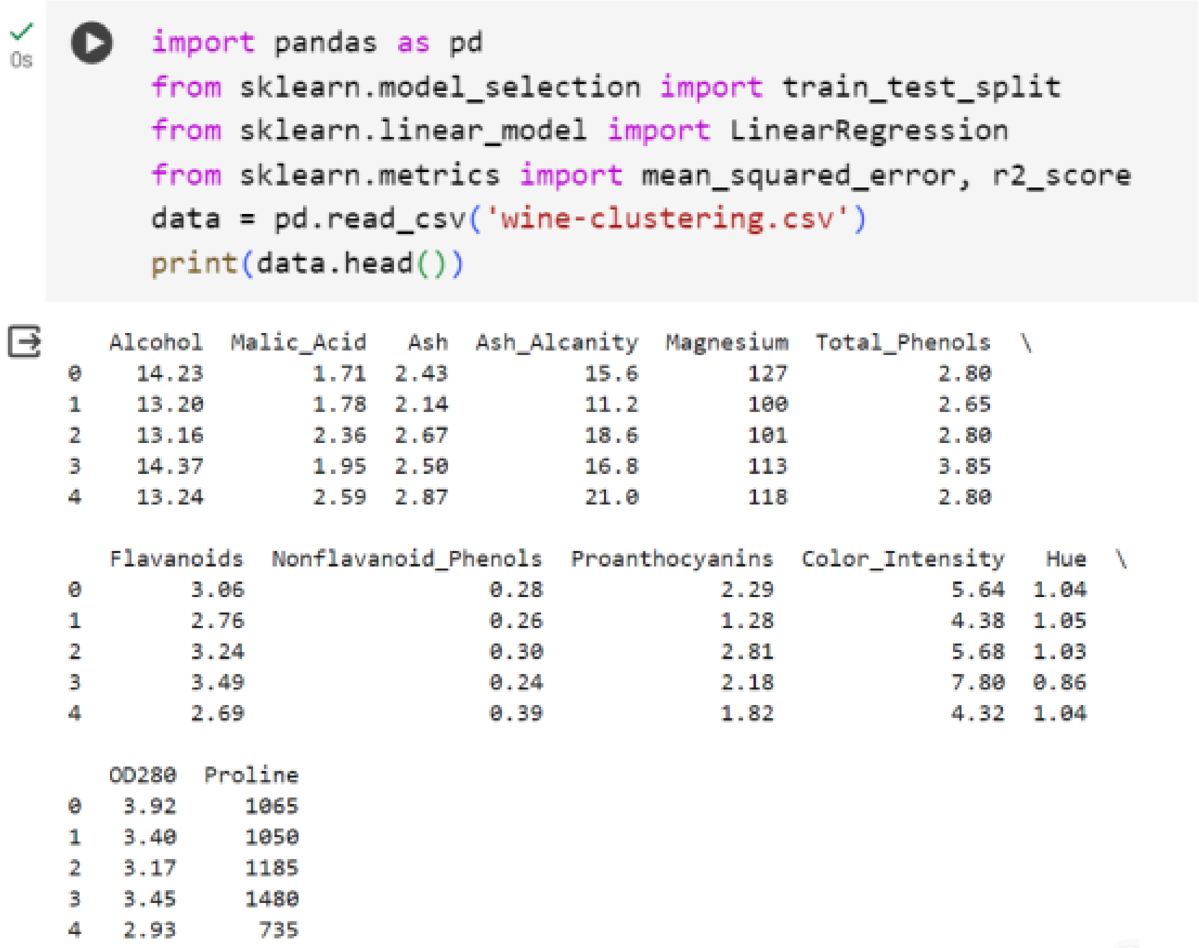

<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
DB
Local Repository
data
processes
Bank Customer Churn Prediction
diamonds
FinalDatasetDynamicPricing
heart
Operators
perform
Validation (18)
Performance (17)
Predictive (7)
Performance (Classification)
Performance (Binominal Classification)
Performance (Regression)
Performance (Costs)
We found 'Model Visualization Extension' in the Marketplace. Show me!
Process
Process ▸
100%
Select Attributes
Nominal to Numerical
Set Role
Performance
Linear Regression
Apply Model
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Performance (Performance (R...
relative error strict
normalized absolute error
root relative squared error
squared error
correlation
squared correlation
Show advanced parameters
Help
Performance (Regression)
RapidMiner Studio Core
Synopsis
This operator is used for statistica...


<new process*> – RapidMiner Studio Free 9.1.000 @ VISIT-509-21
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Tu
Result History
LinearRegression (Linear Regression)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
squared correlation

squared_correlation

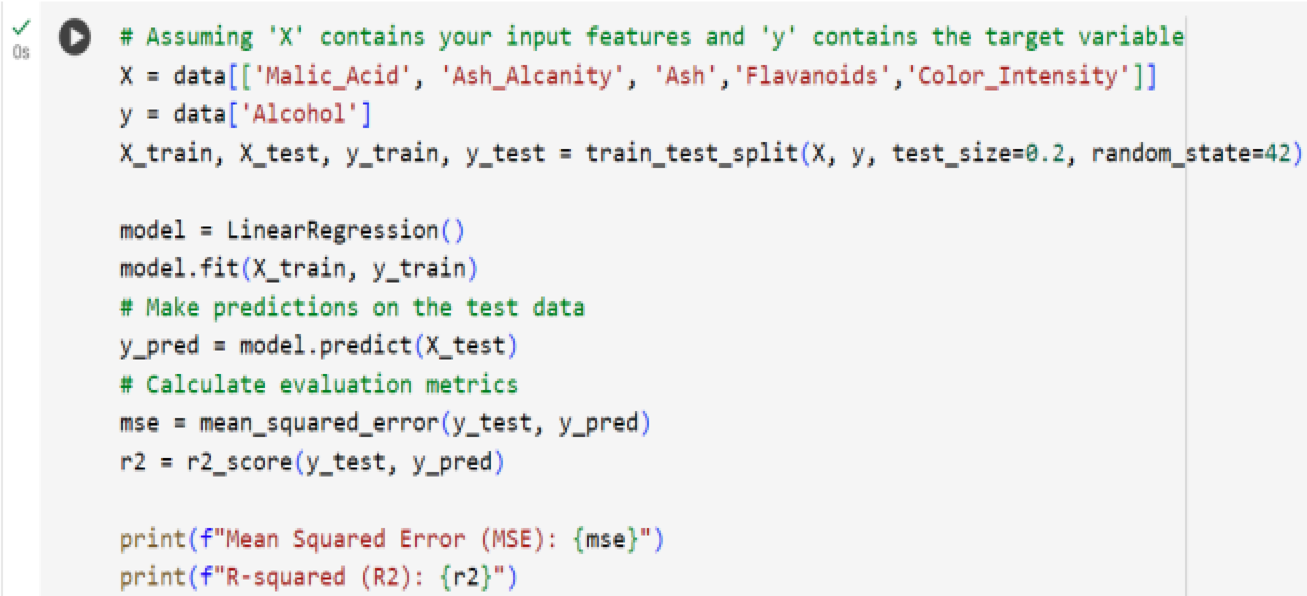
```
squared_correlation: 0.747
```

2. IMPLEMENTATION USING BUILT IN FUNCTION :



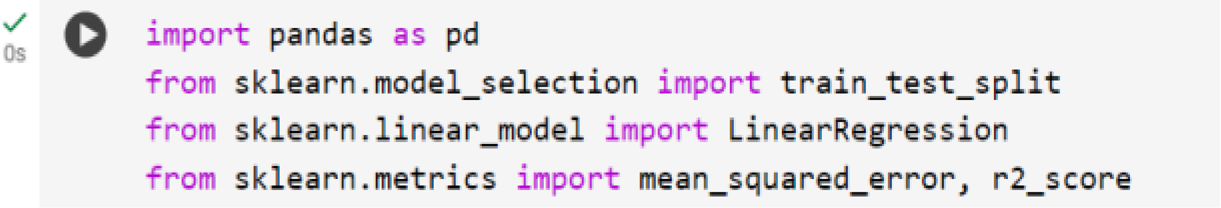
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
data = pd.read_csv('wine-clustering.csv')
print(data.head())
```

```
   Alcohol  Malic_Acid   Ash  Ash_Alcanity  Magnesium  Total_Phenols  \
0    14.23        1.71  2.43          15.6        127           2.80
1    13.20        1.78  2.14          11.2        100           2.65
2    13.16        2.36  2.67          18.6        101           2.80
3    14.37        1.95  2.50          16.8        113           3.85
4    13.24        2.59  2.87          21.0        118           2.80

   Flavanoids  Nonflavanoid_Phenols  Proanthocyanins  Color_Intensity   Hue  \
0        3.06                  0.28             2.29             5.64  1.04
1        2.76                  0.26             1.28             4.38  1.05
2        3.24                  0.30             2.81             5.68  1.03
3        3.49                  0.24             2.18             7.80  0.86
4        2.69                  0.39             1.82             4.32  1.04

   OD280  Proline
0   3.92     1065
1   3.40     1050
2   3.17     1185
3   3.45     1480
4   2.93      735
```

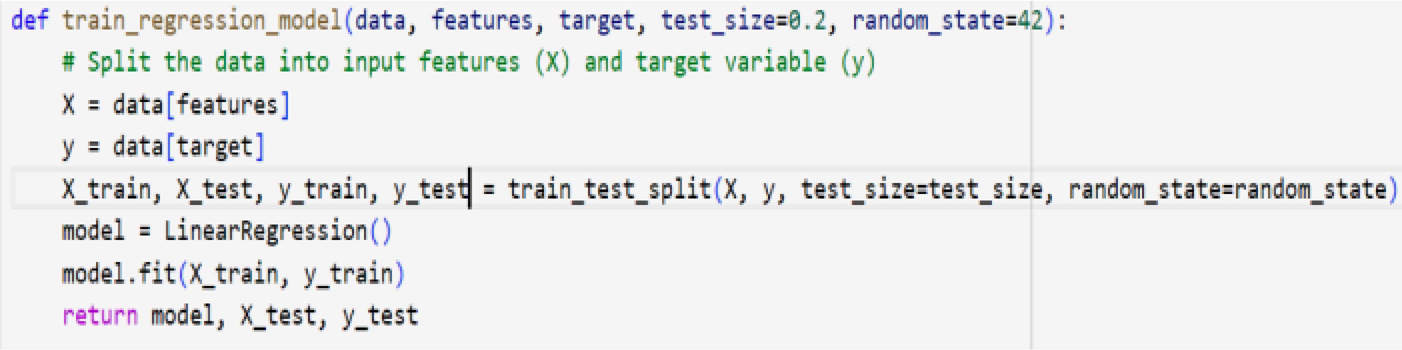


```
# Assuming 'X' contains your input features and 'y' contains the target variable
X = data[['Malic_Acid', 'Ash_Alcanity', 'Ash','Flavanoids','Color_Intensity']]
y = data['Alcohol']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
# Make predictions on the test data
y_pred = model.predict(X_test)
# Calculate evaluation metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error (MSE): {mse}")
print(f"R-squared (R2): {r2}")
```

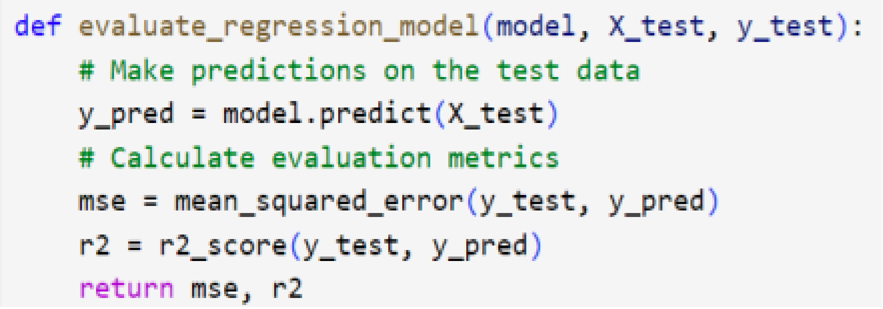
3. IMPLEMENTATION USING USER DEFINED FUNCTION :



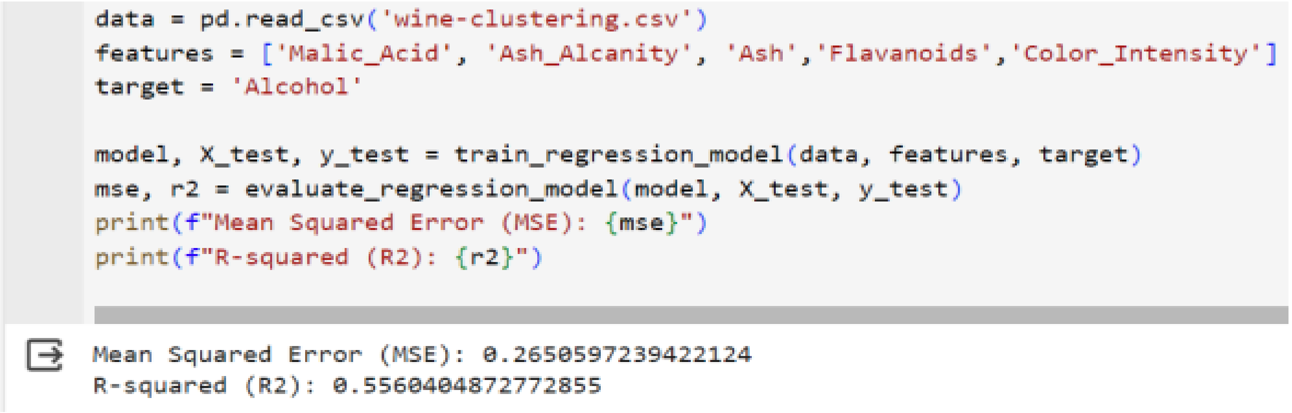
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```



```
def train_regression_model(data, features, target, test_size=0.2, random_state=42):
    # Split the data into input features (X) and target variable (y)
    X = data[features]
    y = data[target]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size, random_state=random_state)
    model = LinearRegression()
    model.fit(X_train, y_train)
    return model, X_test, y_test
```



```
def evaluate_regression_model(model, X_test, y_test):
    # Make predictions on the test data
    y_pred = model.predict(X_test)
    # Calculate evaluation metrics
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    return mse, r2
```



```
data = pd.read_csv('wine-clustering.csv')
features = ['Malic_Acid', 'Ash_Alcanity', 'Ash','Flavanoids','Color_Intensity']
target = 'Alcohol'

model, X_test, y_test = train_regression_model(data, features, target)
mse, r2 = evaluate_regression_model(model, X_test, y_test)
print(f"Mean Squared Error (MSE): {mse}")
print(f"R-squared (R2): {r2}")
```

```
Mean Squared Error (MSE): 0.2650597239422124
R-squared (R2): 0.5560404872772855
```

COMPARISON AND CONCLUSION : Comparing the Mean Squared Error (MSE) across Python's user-defined functions, Python's built-in functions, and RapidMiner reveals varying levels of flexibility, complexity, and performance. Python with user-defined functions allows for fine-tuning and optimization, potentially leading to lower MSE. Python's built-in functions offer a balance between simplicity and performance. RapidMiner's MSE depends on the efficiency of its built-in operators and workflow design. The Python implementation offers flexibility and control for customized data processing and model training using libraries like pandas and scikit-learn.