MODULE 12

SIMPLE LINEAR REGRESSION

Sort by :

Hamzah L200154013

\

Informatics Study Program

Faculty of Communication and Informatics

Muhammadiyah University of Surakarta

Praktikum Steps

Sample Case:

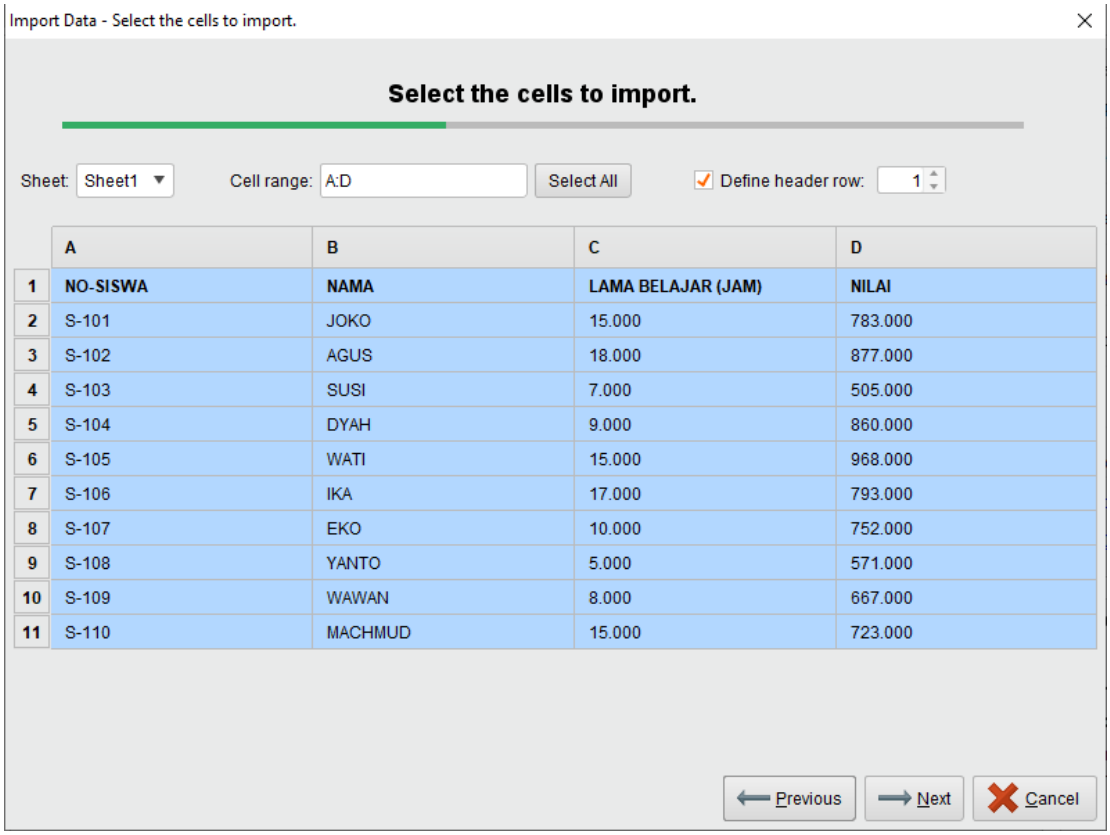
In a class that has 10 students a survey was conducted on the length of the student study and their test/exam result. The student data will be used as basic calculation to predict other student exam result based on how long they study.

Look for t-Stat value and Linear Regression Model

1. Open MS. Excel, and make table student like the picture below. Save with the name **Tabel_LamaBelajardanNilaiUjian.xls** (Excel format 2013 *xls).

NO-SISWA	NAMA	LAMA BELAJAR (JAM)	NILAI
S-101	JOKO	15	783
S-102	AGUS	18	877
S-103	SUSI	7	505
S-104	DYAH	9	860
S-105	WATI	15	968
S-106	IKA	17	793
S-107	EKO	10	752
S-108	YANTO	5	571
S-109	WAWAN	8	667
S-110	MACHMUD	15	723

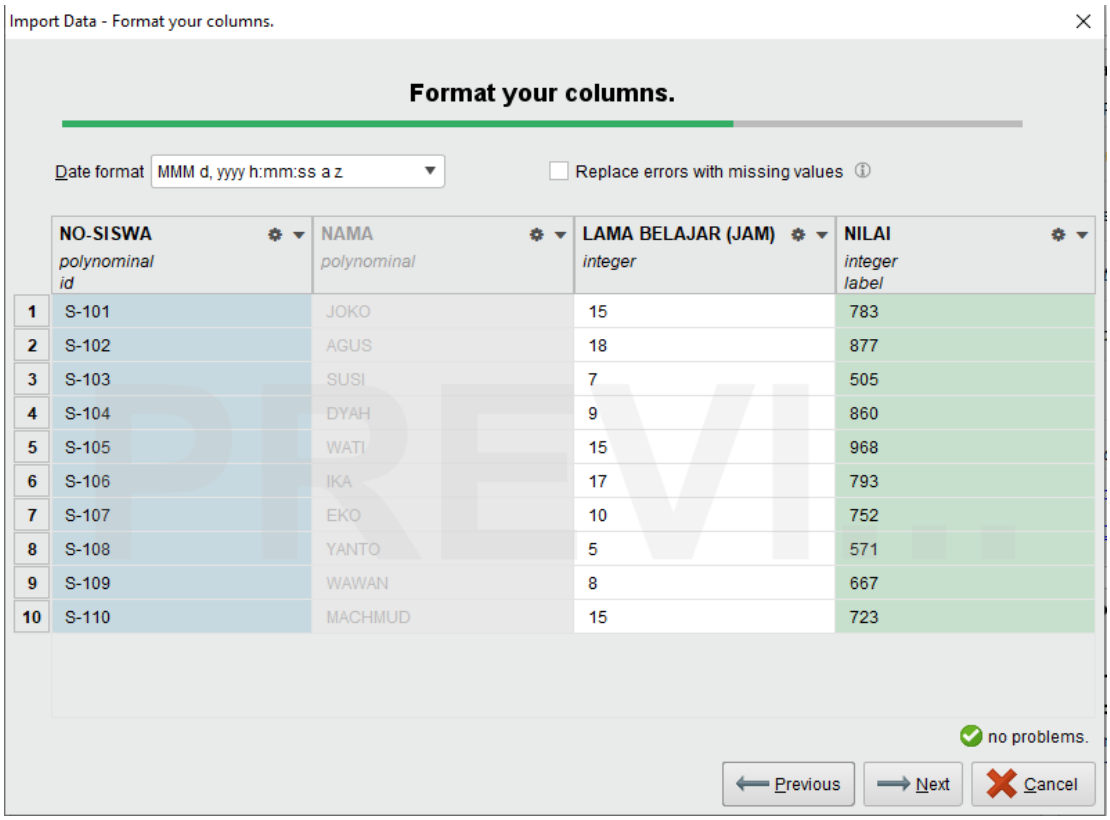
2. Open RapidMiner application and import the excel file, select all cells and click next.



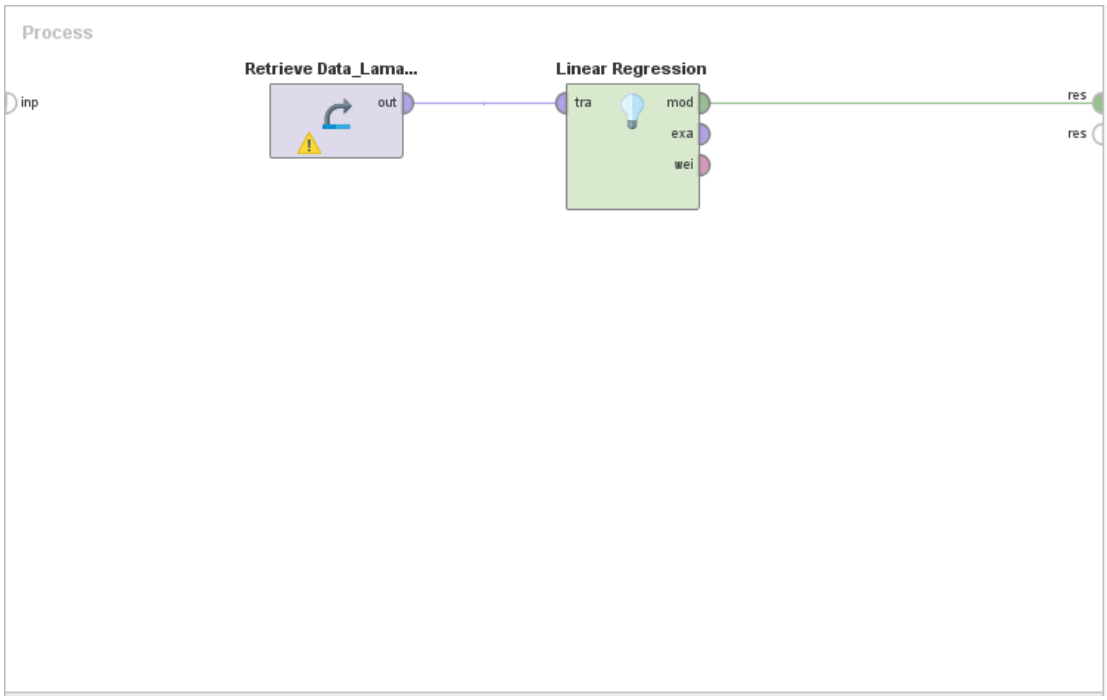
	A	B	C	D
1	NO-SISWA	NAMA	LAMA BELAJAR (JAM)	NILAI
2	S-101	JOKO	15.000	783.000
3	S-102	AGUS	18.000	877.000
4	S-103	SUSI	7.000	505.000
5	S-104	DYAH	9.000	860.000
6	S-105	WATI	15.000	968.000
7	S-106	IKA	17.000	793.000
8	S-107	EKO	10.000	752.000
9	S-108	YANTO	5.000	571.000
10	S-109	WAWAN	8.000	667.000
11	S-110	MACHMUD	15.000	723.000

3. Change data type and attribute type as follows:
   a. NO_SISWA: polynomial, id
   b. NAMA: choose exclude column
   c. LAMA JAMA BELAJAR: integer
   d. NILAI: integer, label

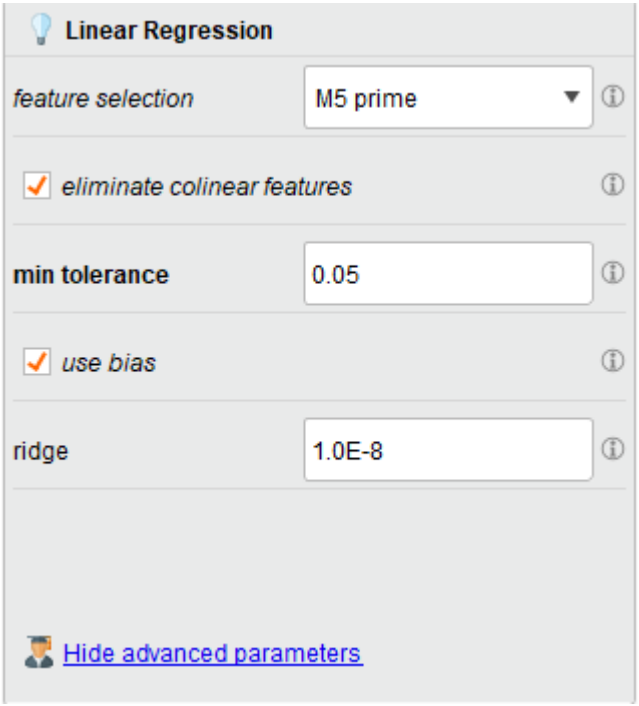
   Save with name **Data_LamaBelajardanNilaiUjian.**



4. Drag **Data_LamaBelajardanNilaiUjian** into the process area, then add **Linear Regression** operator. Connect each port like the picture below.



5. Double click on **Linear Regression** operator, in parameter box change **min tolerance** value to **0.05** (tolerance limit of 5%).



6. (Additional) If the input data is in nominal type or polynomial, add **Nominal to Numerical** operator between input data and linear regression operator. Set **coding type** parameter to **unique integer** (This step can be skipped in our praktikum).
7. Run the process and there is the result:
   a) Table View (looking for the value of t-hitung)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
LAMA BELAJAR (...	21.608	7.645	0.707	1	2.827	0.022	**
(Intercept)	492.769	96.909	?	?	5.085	0.001	****

From the table above can be seen that the value of t-Stat is 2.827, then we compared with the provided t-table. Because there is only one independent variable (Lama Jam Belajar) we use **one-tail**, **min tolerance 0.05** and **10** sample data and found the value is 1.812.

***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587

2.827 is higher that 1.812 (t-Stat > t-Table), so we can conclude that the independent variable (Lama Jam Belajar) is significantly affected **Nilai Ujian** variable (dependent variable.

   b) Text View (Model Regression)

LinearRegression

```
  21.608 * LAMA BELAJAR (JAM)
+ 492.769
```

Based on the text view result we got equation:
= 21.608 * LAMA BELAJAR (JAM) + 492.769

Then a simple linear regression equation model can be made to look for the value of variable Y (**Nilai Ujian**) based on variable X_1 (Lama Jam Belajar). The following regression models are formed:

Y = 21.608 * LAMA BELAJAR (JAM) + 492.769

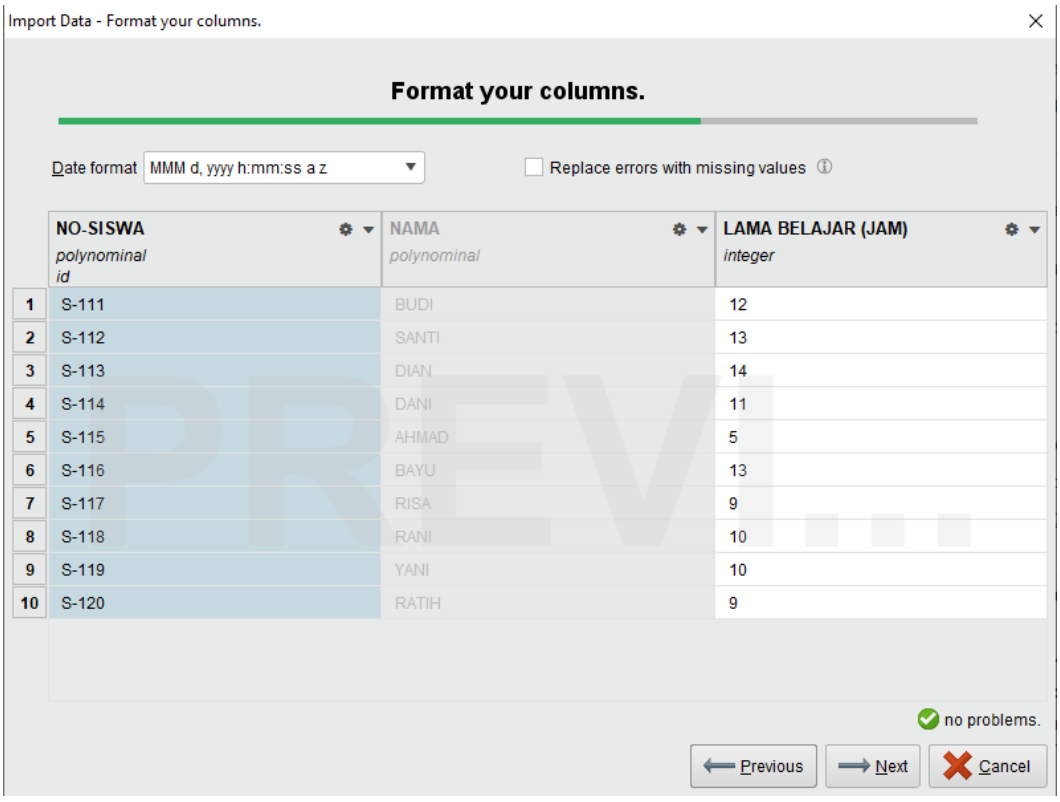
With this model, **Nilai Ujian** can be found by entering the value of **Lama Jam Belajar** in variable X_1.

Look for t value and Linear Regression Model using RapidMiner

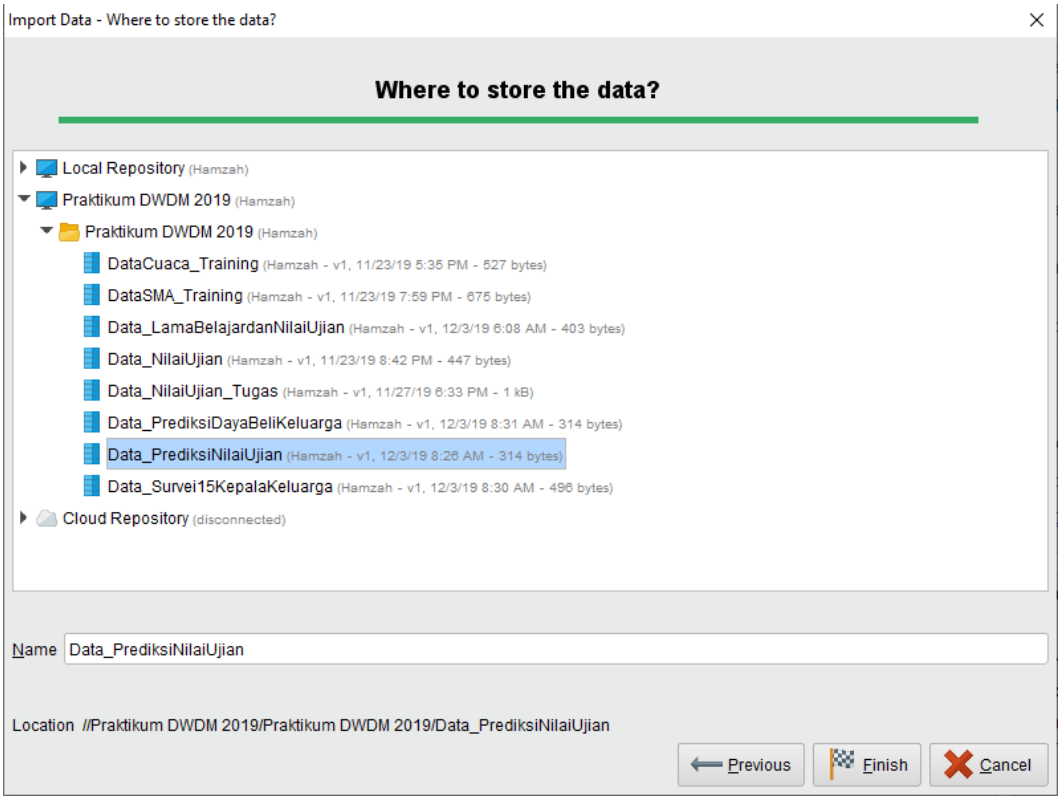
1. Open Ms. Excel make a table of student data like in the picture, and save it with name **Tabel_PrediksiNilaiUjian.xls** (Format Excel 2003*.xls).

NO-SISWA	NAMA	LAMA BELAJAR (JAM)
S-111	BUDI	12
S-112	SANTI	13
S-113	DIAN	14
S-114	DANI	11
S-115	AHMAD	5
S-116	BAYU	13
S-117	RISA	9
S-118	RANI	10
S-119	YANI	10
S-120	RATIH	9

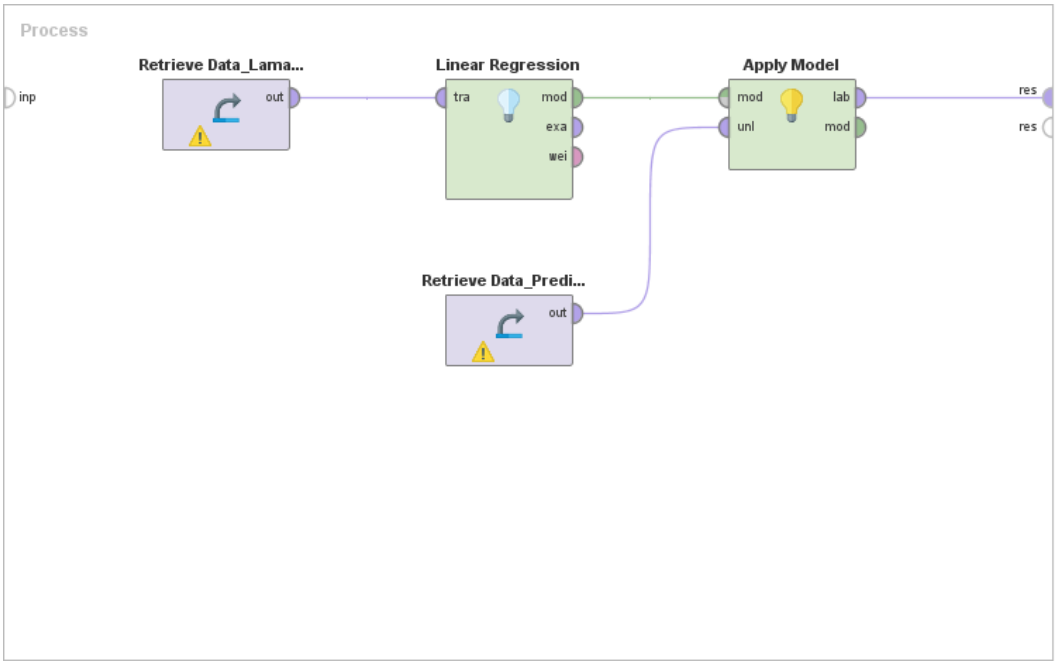
2. Open RapidMiner application, and import the file excel that we make earlier. Change the data type and their attribute into:
   a) NO_SISWA : polynomial, id
   b) NAMA : Exclude column
   c) LAMA JAM BELAJAR : integer



3. Save with name **Data_PrediksiNilaiUjian.**



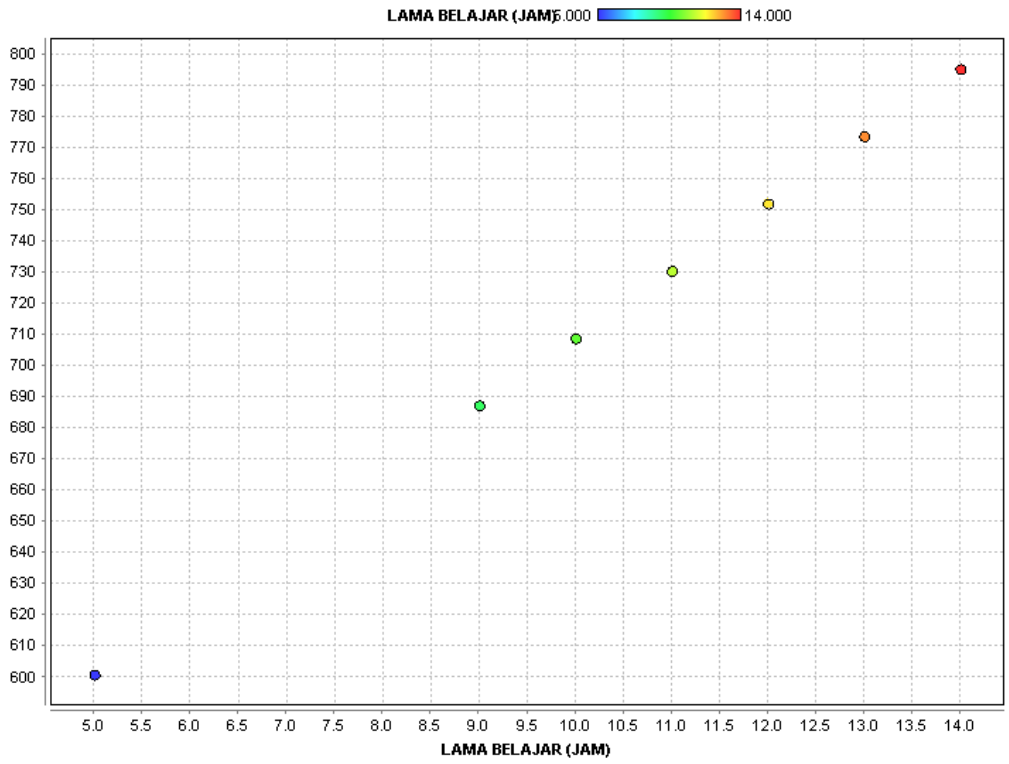
4. Still with the previous process we did in earlier activity, drag **Data_PrediksiNilaiUjian** in process area and add apply model operator then arrange like in the picture below after run the process.



5. The result of the process:
   a) Data View (Prediction of Nilai Ujian)

Row No.	NO-SISWA	prediction(N...	LAMA BELA...
1	S-111	752.061	12
2	S-112	773.668	13
3	S-113	795.276	14
4	S-114	730.453	11
5	S-115	600.807	5
6	S-116	773.668	13
7	S-117	687.238	9
8	S-118	708.845	10
9	S-119	708.845	10
10	S-120	687.238	9

   b) Charts View (Scatter Plot)



Proofing the Regression Mode

Y = 21.608 * LAMA BELAJAR (JAM) + 492.769

We will predict the **Nilai Ujian** using regression model above in excel file and compare it with result from RapidMiner, it is correct or wrong.

NO_SISWA	NAMA	LAMA BELAJAR (JAM)	Prediction (NILAI)	Prediction (NILAI)
			Tabel	Model Regresi
S-111	BUDI	12	752,0607648	752,065
S-112	SANTI	13	773,6684128	773,673
S-113	DIAN	14	795,2760608	795,281
S-114	DANI	11	730,4531168	730,457
S-115	AHMAD	5	600,8072289	600,809
S-116	BAYU	13	773,6684128	773,673
S-117	RISA	9	687,2378209	687,241
S-118	RANI	10	708,8454688	708,849
S-119	YANI	10	708,8454688	708,849
S-120	RATIH	9	687,2378209	687,241

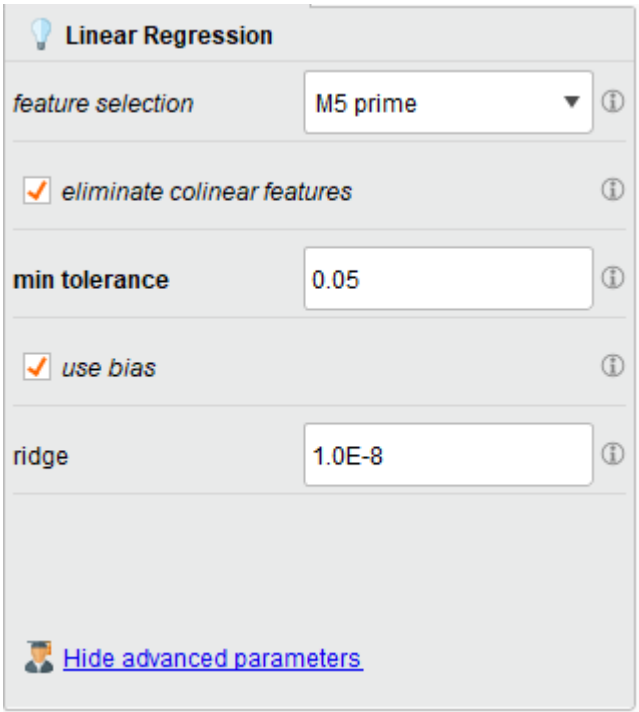
As we can see in table above, the result is the same with RapidMiner and Excel (RapidMiner left and Excel right). We can assume that the regression model is correct.

Task

1. Table Survey Family

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA	DAYA BELI (RUPIAH)
1	1000000	6	834000
2	1400000	7	1200000
3	200000	3	134000
4	1400000	6	1167000
5	500000	3	334000
6	1700000	5	1360000
7	400000	3	267000
8	1900000	5	1520000
9	300000	3	200000
10	500000	4	375000
11	700000	7	600000
12	1900000	3	1267000
13	800000	4	600000
14	1500000	4	1125000
15	1300000	7	1115000

2. Simple Linear Regression using RapidMiner with condition:
   a) Independent Variable (X) = Pendapatan (X_1), Jumlah Anggota Keluarga (X_2)
   b) Dependent Variable (Y) = Daya Beli
   c) Minimum Tolerance = 5%



3. Linear Regression Result:
   a) Table View

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
PENDAPATAN (...	0.739	0.021	0.924	0.857	35.037	0	****
JUMLAH ANGGO...	47807.624	7833.319	0.161	0.857	6.103	0.000	****
(Intercept)	-180222.487	36497.284	?	?	-4.938	0.000	****

t-Stat value is:

i. $X_1 = 35.037$

ii. $X_2 = 6.103$

t-Table for 2 independent variables with 15 sample data is: 2.131

***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073

Because X_1 (35.037) and X_2 (6.103) is higher than the t-Table (2.131). We can conclude that variable Y (Daya Beli) is significantly affected by both variable X_1 and X_2.

4. The Linear Regression Model is

LinearRegression

```
  0.739 * PENDAPATAN (RUPIAH)
+ 47807.624 * JUMLAH ANGGOTA KELUARGA
- 180222.487
```

5. With data testing provided below predict the value of Daya Beli with:

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA
1	900000	5
2	800000	3
3	500000	2
4	1900000	6
5	600000	2
6	800000	5
7	1000000	6
8	1100000	4
9	1000000	4
10	500000	3

a) Using Regression Model in steps 4 in excel

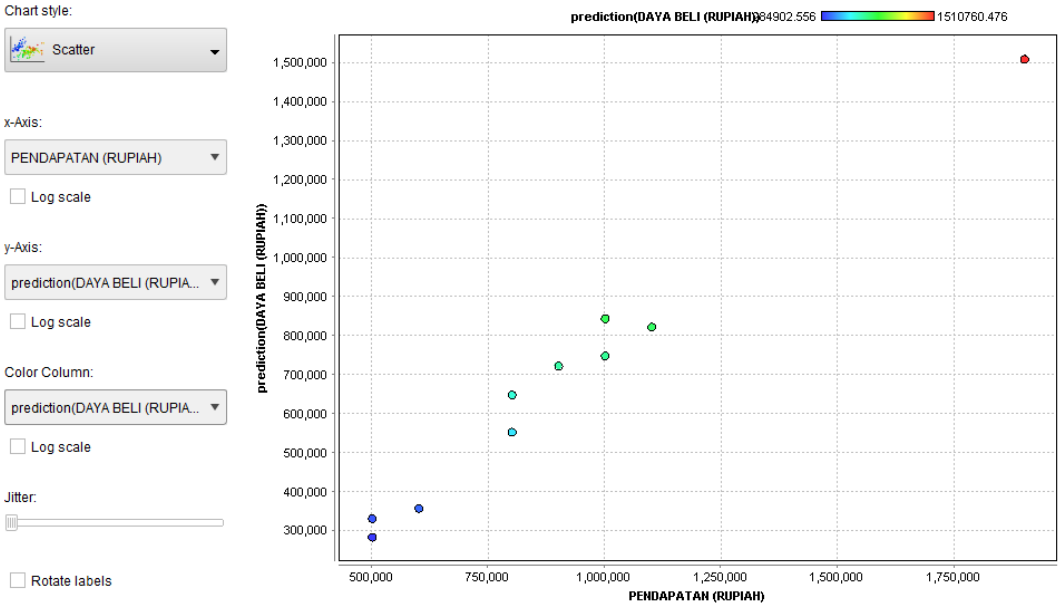
E3 | = 0,739 * B3 + 47807,624 * C3-180222,487

	A	B	C	D	E
1	NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA	Prediction (NILAI)	Prediction (NILAI)
2				Tabel	Model Regresi
3	1	900000	5	723933,2625	723915,633
4	2	800000	3	554416,0562	554400,385
5	3	500000	2	284902,5556	284892,761
6	4	1900000	6	1510760,476	1510723,257
7	5	600000	2	358804,5146	358792,761
8	6	800000	5	650031,3035	650015,633
9	7	1000000	6	845642,8452	845623,257
10	8	1100000	4	823929,5569	823908,009
11	9	1000000	4	750027,5979	750008,009
12	10	500000	3	332710,1792	332700,385

b) Using RapidMiner

Row No.	NO. RESPON...	prediction(D...	PENDAPATA...	JUMLAH AN...
1	1	723933.263	900000	5
2	2	554416.056	800000	3
3	3	284902.556	500000	2
4	4	1510760.476	1900000	6
5	5	358804.515	600000	2
6	6	650031.304	800000	5
7	7	845642.845	1000000	6
8	8	823929.557	1100000	4
9	9	750027.598	1000000	4
10	10	332710.179	500000	3

6. The distribution pattern uses plot view (Scatter) with the following sets:
   a) x-Axis = Pendapatan (Rupiah),
      y-Axis = Prediction (Daya Beli (Rupiah)),
      Color Column = Prediction (Daya Beli (Rupiah))



   b) x-Axis = Jumlah Anggota Keluarga,
      y-Axis = Prediction (Daya Beli (Rupiah)),
      Color Column = Prediction (Daya Beli (Rupiah))