MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
TERNOPIL NATIONAL ECONOMIC UNIVERSITY
FACULTY OF COMPUTER INFORMATION TECHNOLOGIES
DEPARTMENT OF INFORMATION COMPUTER SYSTEMS
AND CONTROL

GUIDELINES

for practical work

on discipline "Data Mining"

for students of specialty 122 "Computer Science"

Ternopil

2019

Guidelines for practical work on discipline "Data Mining" for students of specialty 122 "Computer Science" / P. Bykovyy, D. Zahorodnia, G. Hladiy / Ed. A. Sachenko. - Ternopil-2019. - 63 p.

Associate Professor of

Compiled by *Bykovyy P.,* Ph.D., Assistant Professor, Department of Information Computer Systems and Control
Zahorodnia D., Ph.D., Senior Lecturer, Department of Information Computer Systems and Control
Hladiy G. Ph.D., Associate Professor, Department of Information Computer Systems and Control

Reviewers: Trembach Rostyslav, Ph.D., Associate Professor of the Department of Automation of Technological Processes and Production of Ternopil Ivan Puluj National Technical University

Andriy Melnyk, Ph.D., Associate Professor, Computer Science Department, Ternopil National Economic University

Responsible for the issue: Anatoly Sachenko, Doctor of Technical Sciences, Professor, Head of the Department of Information Computer Systems and Control

Approved at a meeting of the Department of Information Computer Systems and Control, protocol No. 3 from 10.30.2019.

Reviewed and recommended for publication by the Academic Council of the Faculty of Computer Information Technologies, protocol No. 3 dated 10/31/2019.

TABLE OF CONTENTS

INTRODUCTION

Aim of discipline “Data Mining” is to study of modern methods of the data processing, and also discovery in the untilled arrays of data earlier unknown, practically useful knowledge and patterns, necessary for an acceptance decisions.

The competencies, forming of that provide the study of discipline:

- knowing of data warehousing concepts and their operative analytical processing and intellectual analysis;
- ability to discover in data earlier unknown knowledge necessary for making decision in the different spheres of professional activity.

At the end of the course, students will be able to:

- identify, describe, reproduce concept stage, objectives, standards Data Mining;
- know the methods of classification, prediction, cluster analysis, associative search rules;
- perform visual analysis – Visual Mining; analysis of textual information – Text Mining; obtaining knowledge from the Web – Web Mining;
- use the tools of analysis processes – Process Mining;
- create multidimensional data models.

PRACTICAL WORK #1
Correlation and Regression Analysis using MS Excel

To perform correlation and regression analysis of data using MS Excel, you can use the standard functions of this software, as well as the Add-In "Analysis ToolPak". This MS Excel add-in becomes available after you install MS Office. To use this add-in, you must first download it. For this:

1. Click the ***File*** tab and select ***Options***.
2. Click ***Add-Ins*** tab, and then in the ***Manage*** box, click ***Excel Add-Ins***.
3. Click ***Go***.

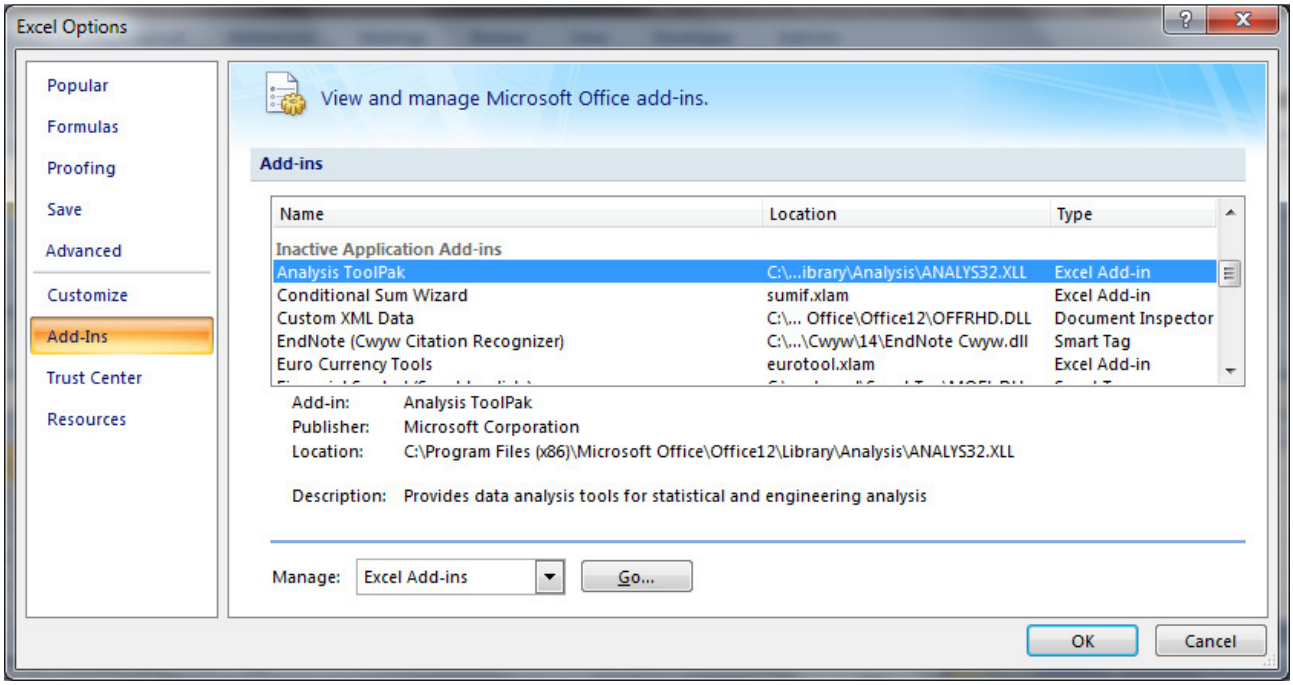


Fig.1. ***Add-Ins*** dialog box of ***Excel Options***

4. Select the ***Analysis ToolPak*** check box in the ***Add-Ins*** dialog box and then click ***OK***.

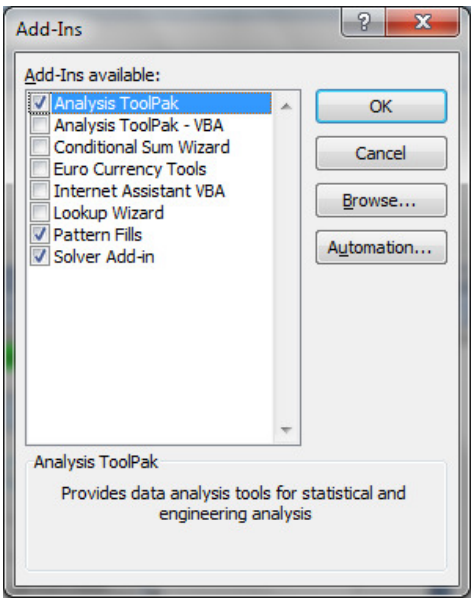
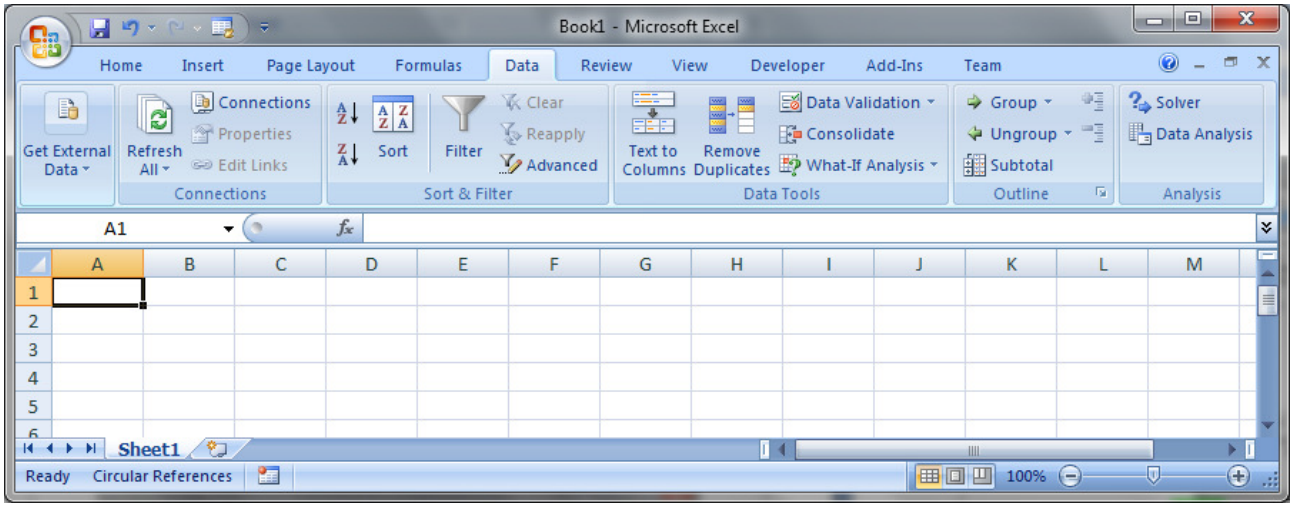


Fig.2. Choosing of Analysis Tool Pack at Add-Ins

5. After loading the Analysis ToolPak, the ***Data Analysis*** button will appear on the ***Data*** tab (***Analysis*** group).



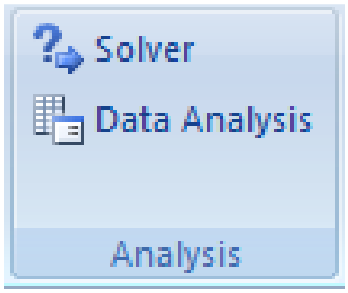


Fig.3. Command ***Data Analysis*** *in the analysis group*

6. Next, it is possible to choose one of 19 tools of Data Analysis package.

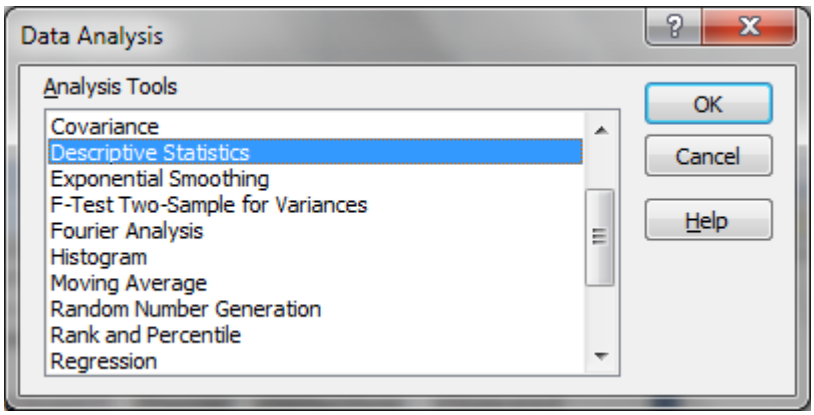


Fig.4. List of tools for data analysis

Before performing correlation or regression analysis, it is worth revising the descriptive statistics of the data set. **Descriptive statistics** - a quantitative description of the basic properties of a dataset. The purpose of descriptive statistics is to summarize the primary results obtained from observations and experiments, that is, it provides a brief summary of the sample and the observations made.

Selecting the **Descriptive Statistics** tool in the **Analysis Package**, we get a one-dimensional statistical report containing information about the central trend and the variability (variation) of the input data. This tool provides a complete set of indicators for descriptive statistics, which includes: average; standard error; median; fashion; standard deviation; sample variance; excess; asymmetry; interval; minimum; maximum; amount; score; reliability level.

To Calculate Excel Descriptive Statistics you should follow next steps

Step 1: **Type your data into Excel (you can use column from Table 1),** in a single column. For example, if you have ten items in your data set, type them into cells A1 through A10.

Step 2: **Click the "Data" tab** and then click "Data Analysis" in the Analysis group.

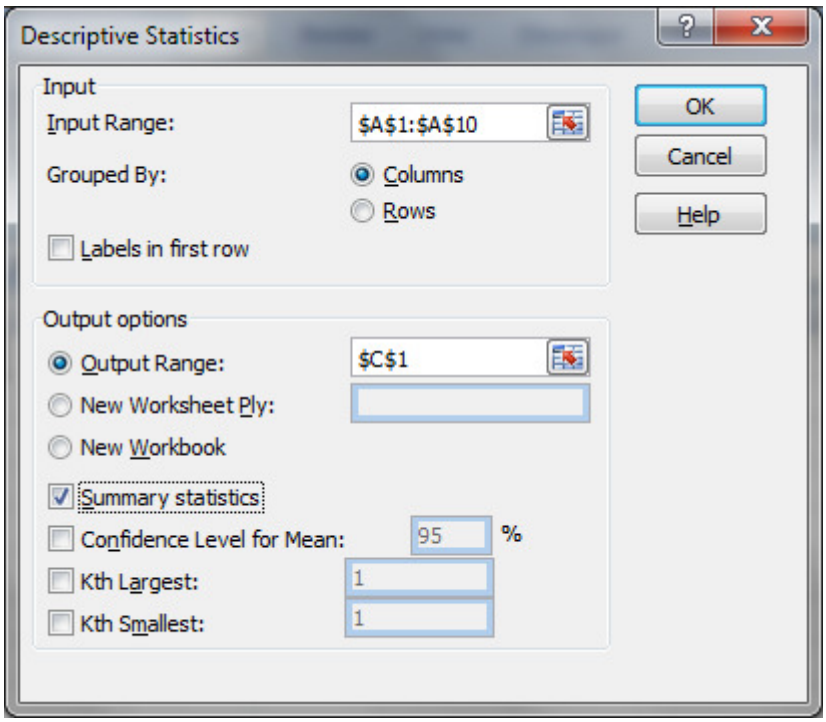
Step 3: **Highlight "Descriptive Statistics"** in the pop-up Data Analysis window.

Step 4: **Type an input range into the "Input Range" text box.** For this example, type "A1:A10" into the box.

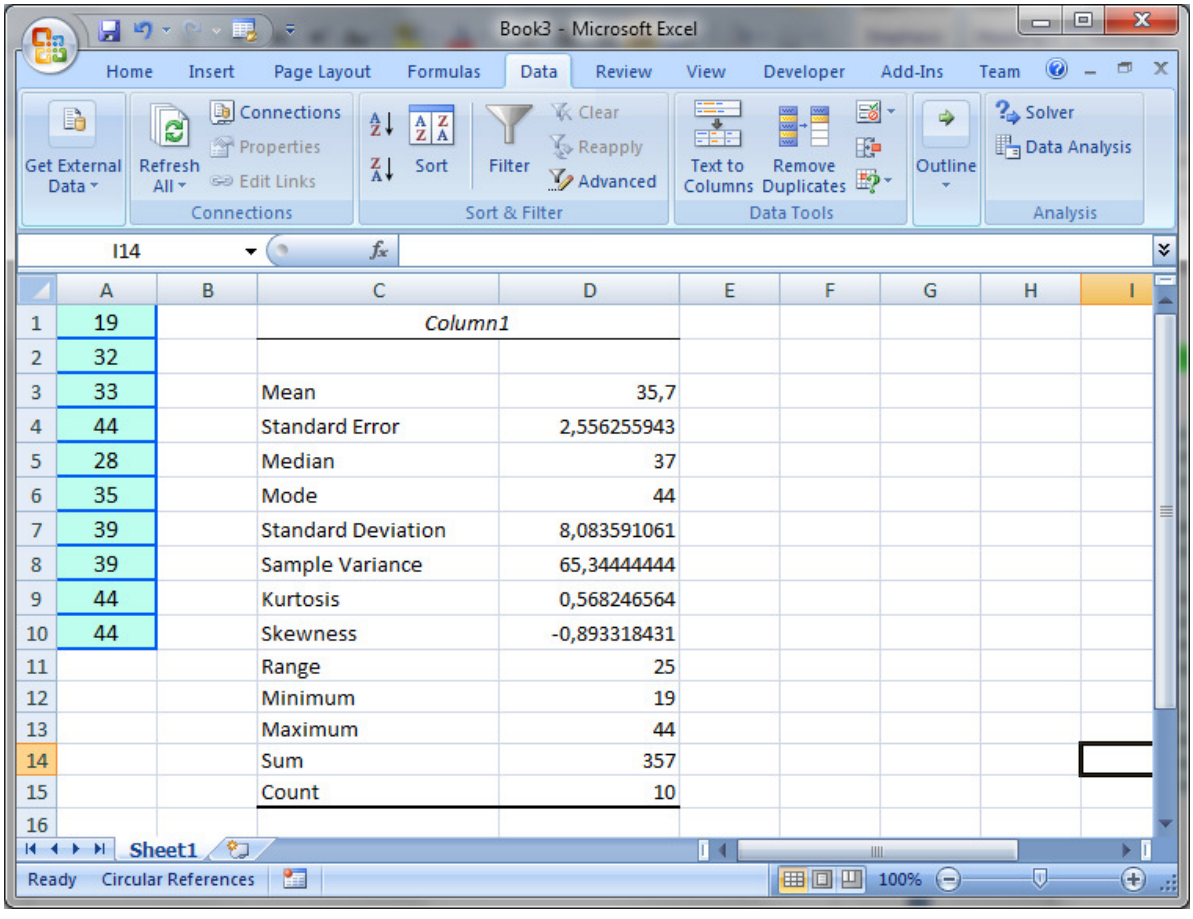
Step 5: **Check the "Labels in first row" check box** if you have titled the column in row 1, otherwise leave the box unchecked.

Step 6: **Type a cell location into the "Output Range" box.** For example, type "C1." Make sure that two adjacent columns do not have data in them.

Step 7: **Click the "Summary Statistics" check box and then click "OK"** to display Excel descriptive statistics. A list of descriptive statistics will be returned in the column you selected as the Output Range.



Report "Descriptive Statistics" for two variables ***x*** and ***y*** are given in the table below (Fig.5).



	A	B	C	D	E	F	G	H	I
1	19		*Column1*						
2	32								
3	33		Mean	35,7					
4	44		Standard Error	2,556255943					
5	28		Median	37					
6	35		Mode	44					
7	39		Standard Deviation	8,083591061					
8	39		Sample Variance	65,34444444					
9	44		Kurtosis	0,568246564					
10	44		Skewness	-0,893318431					
11			Range	25					
12			Minimum	19					
13			Maximum	44					
14			Sum	357					
15			Count	10					
16									

Fig.5. Descriptive statistics of the dataset

Correlation Analysis

First, we will show you how to calculate correlation using standard MS Excel functions.

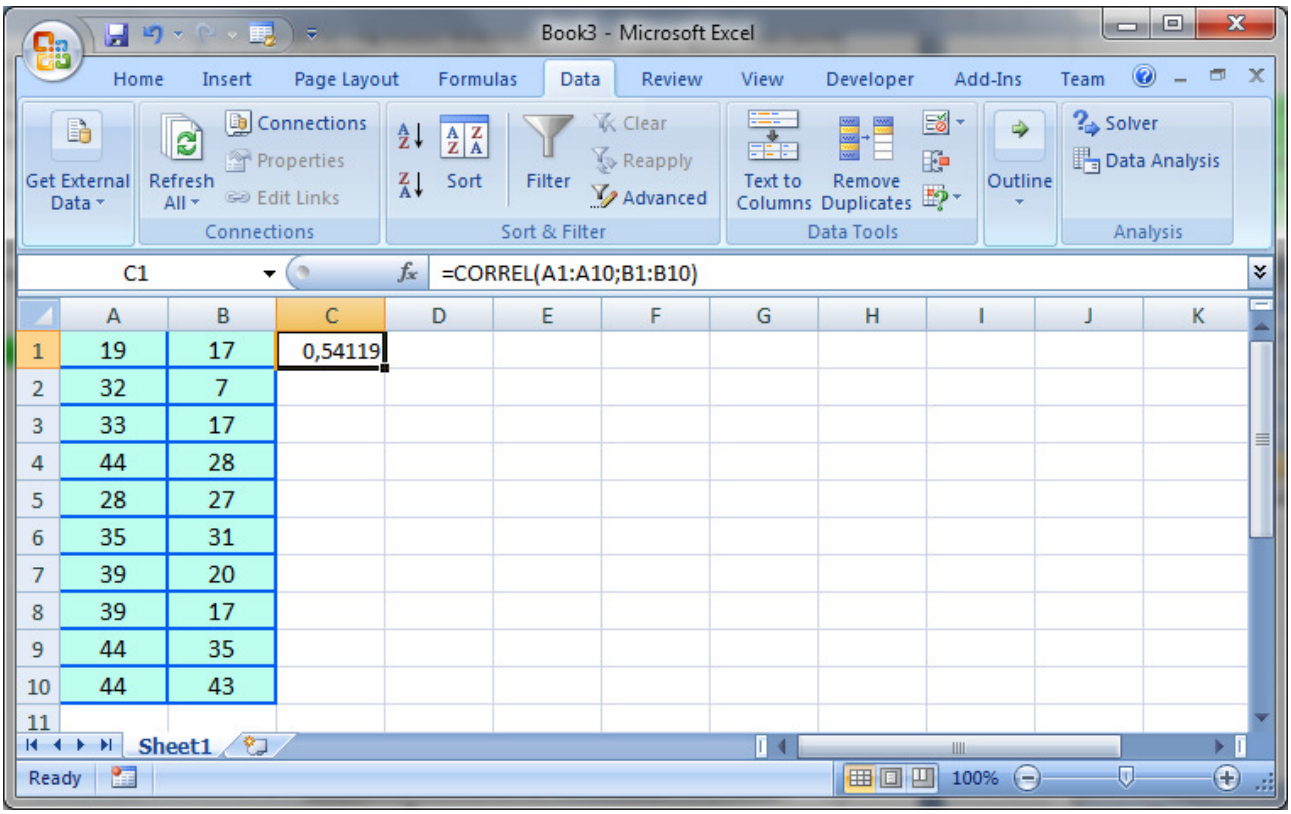
Example 1. 10 students were tested on verbal and imaginative thinking. Measured average time for solving the problems test in seconds (Table 1). Studied the question of the existence of the relationship between time solutions to both problems. The ***X*** variable – average time imaginative solution, and the ***Y*** variable – average time of test solution of verbal problems.

Table 1

Results of testing students

Student	X	Y
1	19	17
2	32	7
3	33	17
4	44	28
5	28	27
6	35	31
7	39	20
8	39	17
9	44	35
10	44	43

To identify the relationship, you must first enter the data (***X***, ***Y***) in the table. Then to calculate correlation coefficient it is necessary to set the cursor to cell **C1**, enter the formula **=CORREL(A1:A10;B1:B10)**. In Excel (except Analysis ToolPak) to calculate the linear correlation coefficients are used functions **CORREL (*array1*; *array2*)** and **PEARSON (*array1*; *array2*)**, where ***array1*** – range of cells for sample ***X***; ***array2*** – range of cells for sample ***Y***.



	A	B	C	D	E	F	G	H	I	J	K
1	19	17	0,54119								
2	32	7									
3	33	17									
4	44	28									
5	28	27									
6	35	31									
7	39	20									
8	39	17									
9	44	35									
10	44	43									
11											

Fig.6. Correlation calculation result using ***CORREL*** function

Example 2. Data were collected monthly observations of weather and attendance of museums and parks (Table 2). Necessary to determine whether there is a relationship between the weather and attendance of museums and parks.

Table 2

Statistics on sunny days and attendance at museums and parks

Sunny days	**Museum visitors**	**Park visitors**
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

To perform the correlation analysis type in the range **A1:G3** initial data. Then on ***Data*** tab click ***Data analysis*** button and select ***Correlation***.

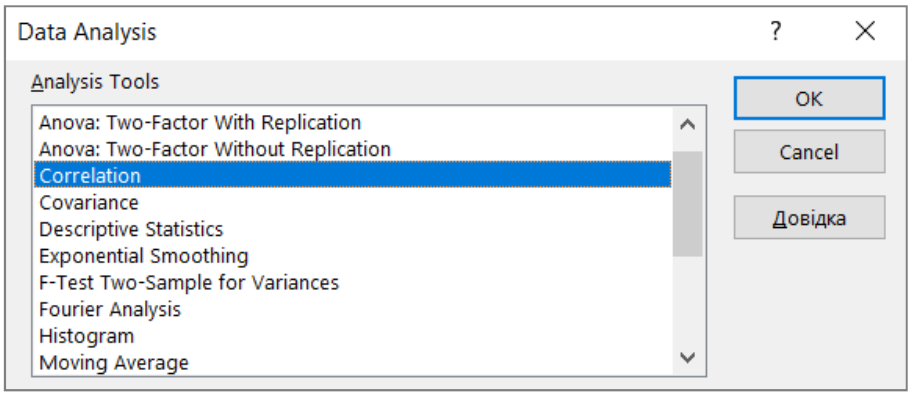


Fig.7. Choosing the ***Correlation*** tool

In the dialog box select ***Input Range***: **A2:C7,** and that the data reviewed by ***Columns***. Specify the ***Output Range*** (**E1**) and click ***OK***.

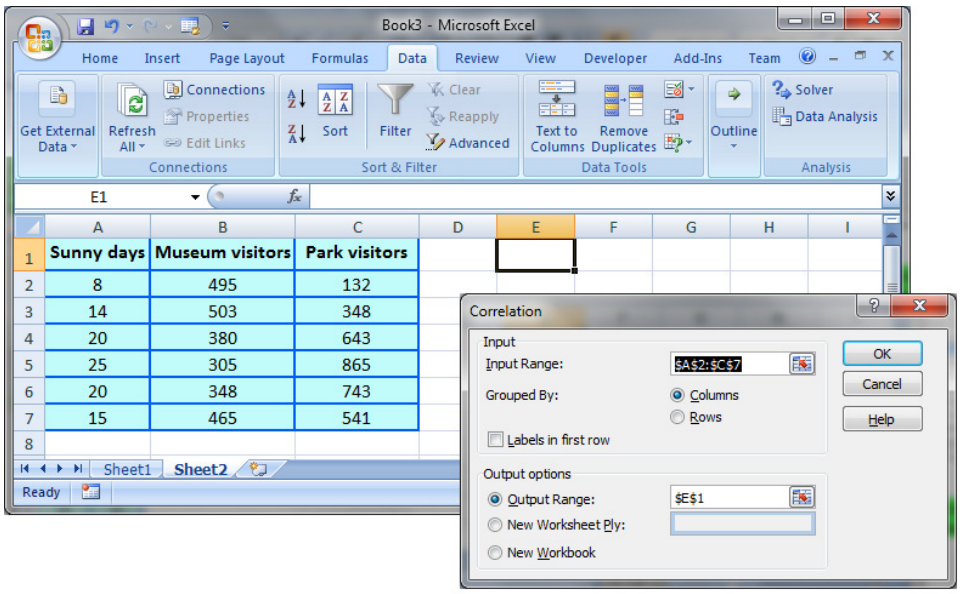


Fig.8. Input and output parameters

The results of implementation are presented in the right table of Fig.9.

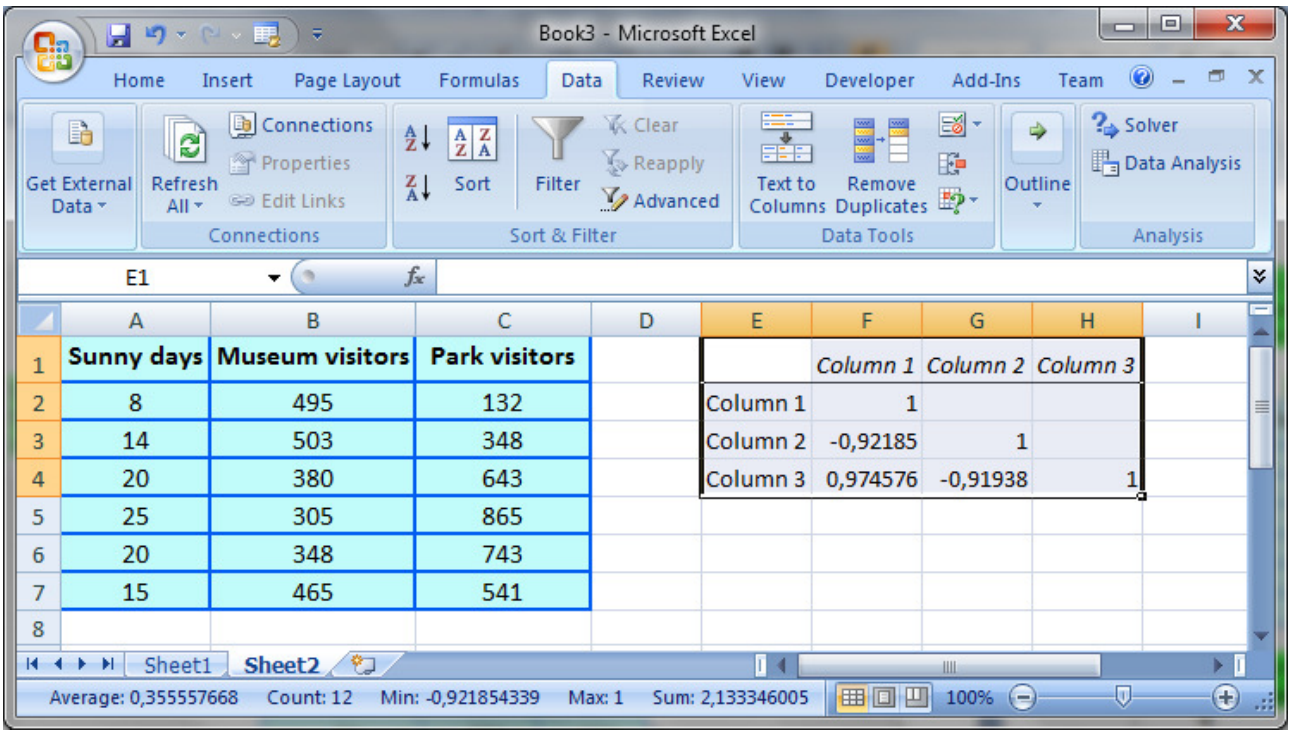


Fig.9. Results of correlation analysis

The correlation between the weather and visiting the museum is -0.92, and between the weather and visiting the park is 0.97, between visits to the park and museum is -0.92.

The analysis revealed dependence:

- a strong degree of inverse linear relationship between museums and the number of sunny days;
- almost linear (very strong direct) connection between a visit to the park and the weather.

Between visits to museums and parks, there is a strong inverse relationship.

Regression Analysis

Example 3. For 20 travel companies were collected advertising costs and the number of tourists who took advantage of the services of each firm after campaign (Fig.10). It is necessary to build a linear regression model on the condition that the variables have normal distribution using:

- Function **LINEST**.
- Graphical method
- Analysis ToolPak (**Regression** tool).

Approach 1.

1. Open a new workbook and create MS Excel spreadsheet according to conditions.

2. For the coefficients ***a*** and ***b*** of the linear regression equation ***y=b*x+a***, which describes the dependence of the number of tourists attracted by the cost of advertising, we use statistical function **LINEST**.

	A	B	C	D
1	**Company**	**Adverticement costs , $ (X)**	**Number of tourists, thousands (Y)**	
2	1	5	720	
3	2	5	750	
4	3	7	800	
5	4	8	820	
6	5	9	800	
7	6	10	880	
8	7	11	950	
9	8	12	820	
10	9	13	900	
11	10	14	950	
12	11	15	920	
13	12	15	980	
14	13	16	980	
15	14	17	970	
16	15	18	980	
17	16	18	1010	
18	17	19	1100	
19	18	20	1100	
20	19	20	1115	
21	20	21	1110	
22				
23	Equation Coefficients		b	a
24				
25	Regression equation			
26				

Fig.10. Entering tabular data

3. Select two cells **C24**:**D24** and insert the **LINEST** function with arguments. Here: **Known_y's** – range of values *Number of tourists*, **Known_x's** – range of values *Advertising Costs*. Press the key combination **SHIFT + CTRL + ENTER**.

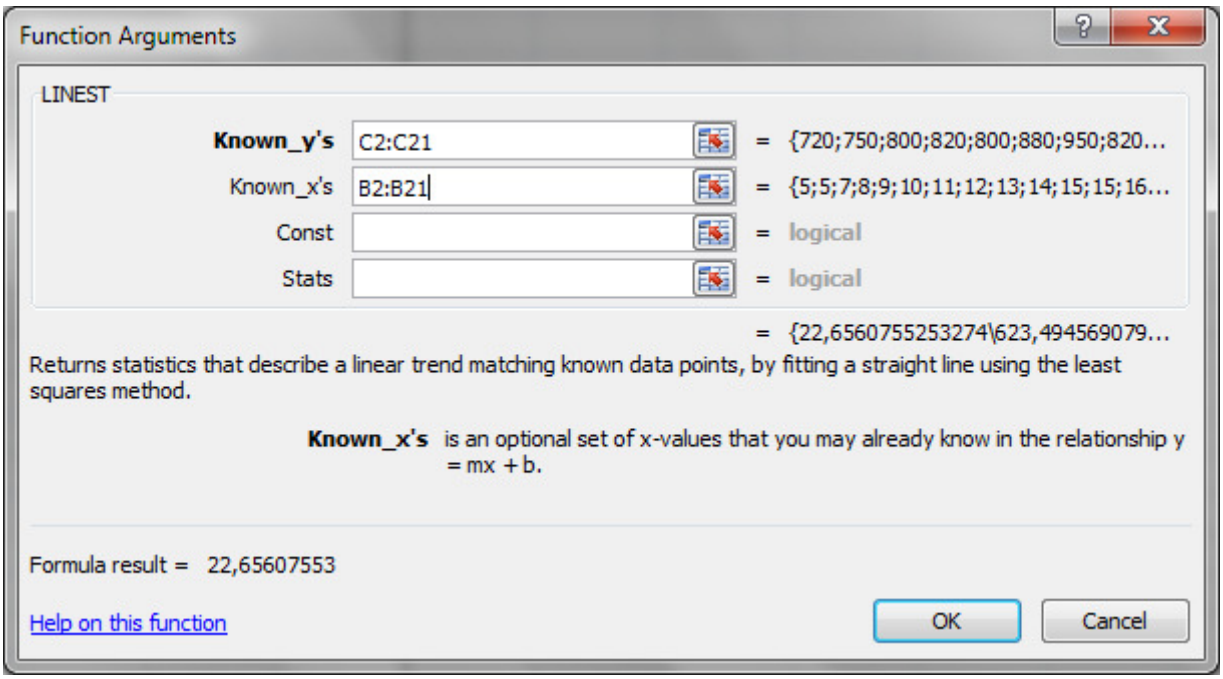


Fig.11. Arguments of **LINEST** function

4. In cell **C27** enter the regression equation ***y=b*x+a*** (substitute the calculated linear regression coefficients instead of ***b*** and ***a***).

	A	B	C	D
23	Equation Coefficients		b	a
24			22,65607553	623,4945691
25	Regression equation			
26			y=22,656*X+623,494	
27				

Fig.12. Calculation results

5. Calculate the standard error of the model by the formula

$$S = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - m}},$$

where y_i – are the actual values of the productive attribute obtained from observational data;

$\hat{y}_i$ – calculated values of the effective attribute;

n – the sample size;

m – the number of parameters in the regression equation.

For this:

- In the cell range **D2: D21**, calculate the value of the resultant characteristic $\hat{y}_i$ by substituting the values of the independent regressive signs x_i into the linear regression equation.
- In the cell range **E2:E21** calculate the deviations of the actual values of the resultant characteristic from the calculated values $y_i - \hat{y}_i$.
- In cell range **F2:F21** calculate the squared deviations $(y_i - \hat{y}_i)^2$.
- In cell **G2** calculate the standard error of the model using the specified formula.

Approach 2.

1. To obtain the regression equation, we construct the correlation field of the variables **X** (advertising costs) and **Y** (number of tourists).

2. Select the range of cells **B2: C21**, on the **Insert** tab in the **Charts** group, select the chart type - **Scatter**. Name the chart "**Correlation field**", for the X axis - **Adverticement costs**, $, for the Y axis - **Number of tourists, thousands** (on the **Layout contextual tab**). Indicate the location - a separate sheet (in the context menu the command **Move Chart**).

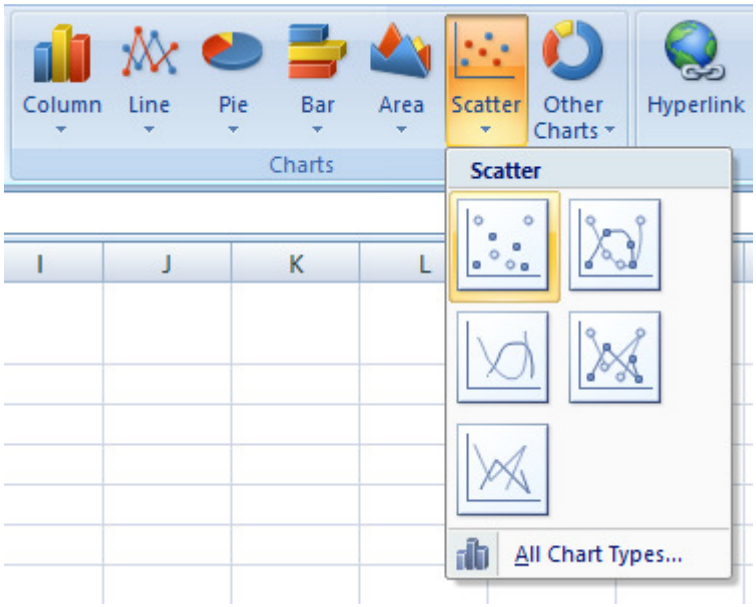
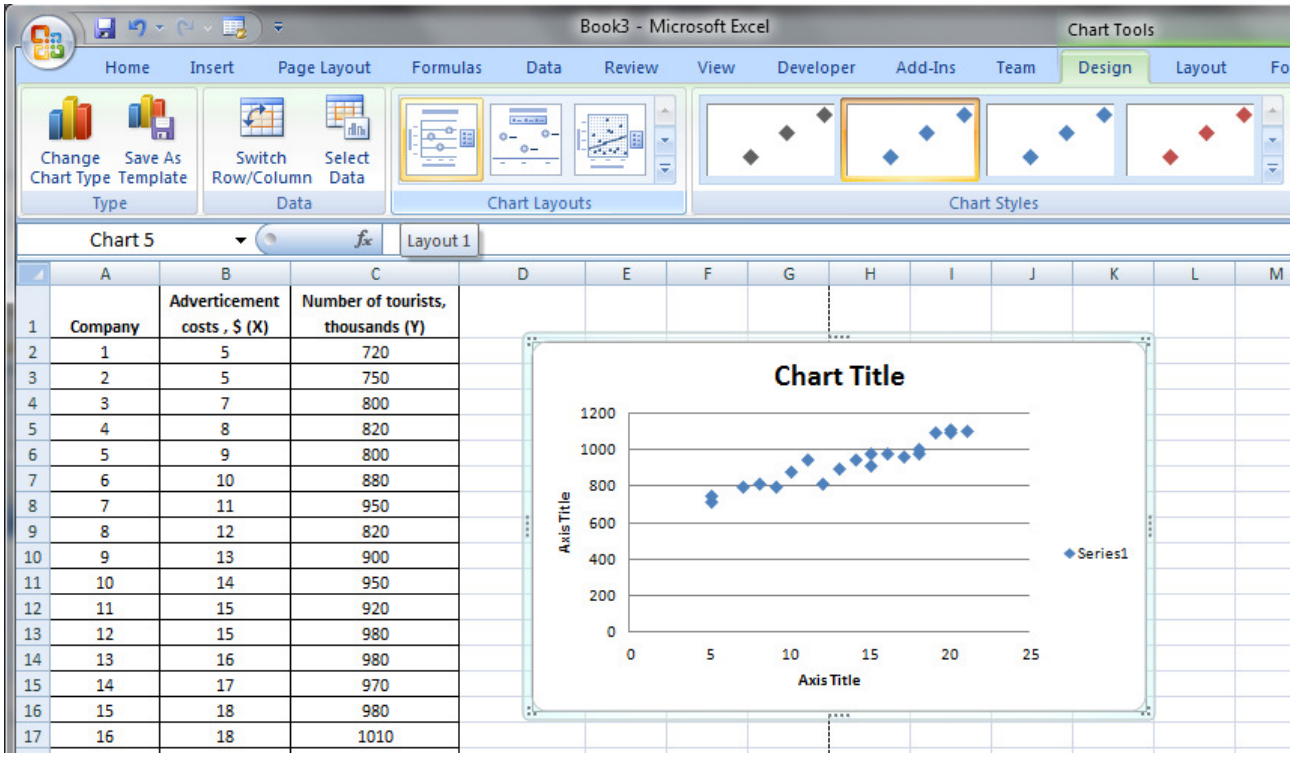


Fig.13. Select a Scatter from the Insert tab



	Company	Adverticement costs , $ (X)	Number of tourists, thousands (Y)
2	1	5	720
3	2	5	750
4	3	7	800
5	4	8	820
6	5	9	800
7	6	10	880
8	7	11	950
9	8	12	820
10	9	13	900
11	10	14	950
12	11	15	920
13	12	15	980
14	13	16	980
15	14	17	970
16	15	18	980
17	16	18	1010

Fig. 14. Insert title on layout

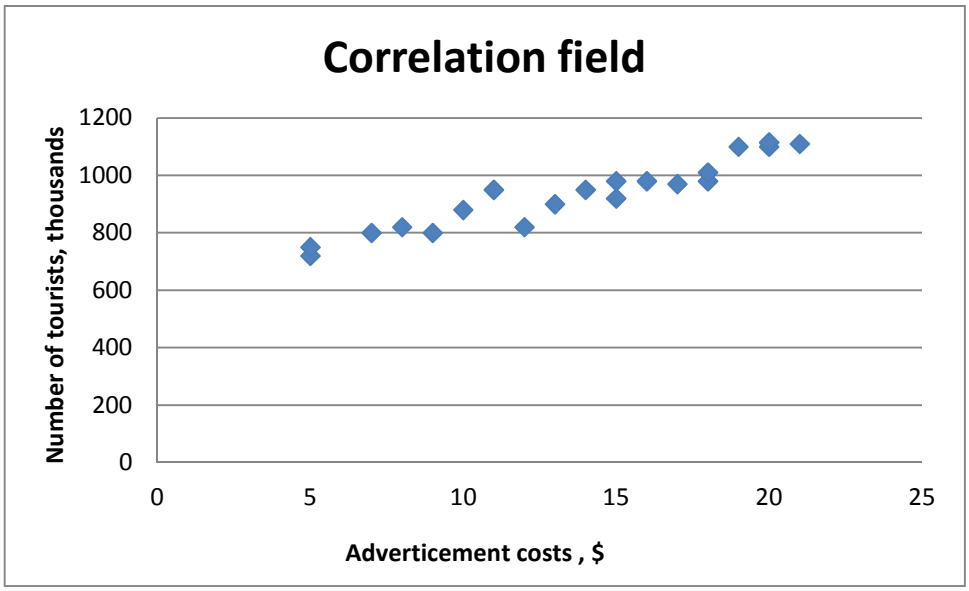


Fig.15. The results of the graphical method

3. Add a trend line to the exact schedule: look at the team and see the trends (**Layout** contribution) - div. Figure 14. **TrendLine** - A graphical representation strains changes in a series of data.

4. The selection of typical trends: **Linear TrendLine**, as well as testing methods for children by the method of hire squares **y = y * x + a** (Fig. 16)

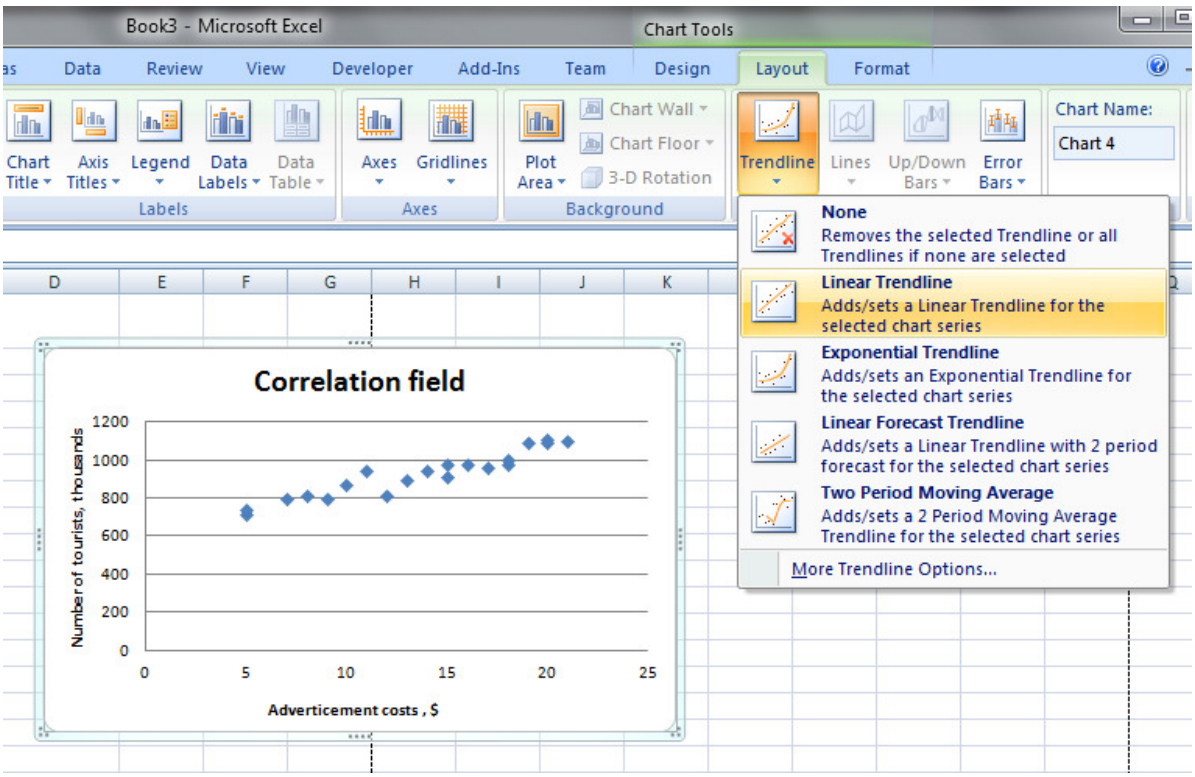


Fig.16. Choosing the type of trend line

5. On the **More TrendLine Options…** , check the **Display equation on chart** and **Display R-squared value on chart.** On chart R^2 - is a number from 0 to 1, which reflects the proximity of the trend line to the actual data. The trend line is most true when the value of $\mathbf{R^2}$ is close to **1**.

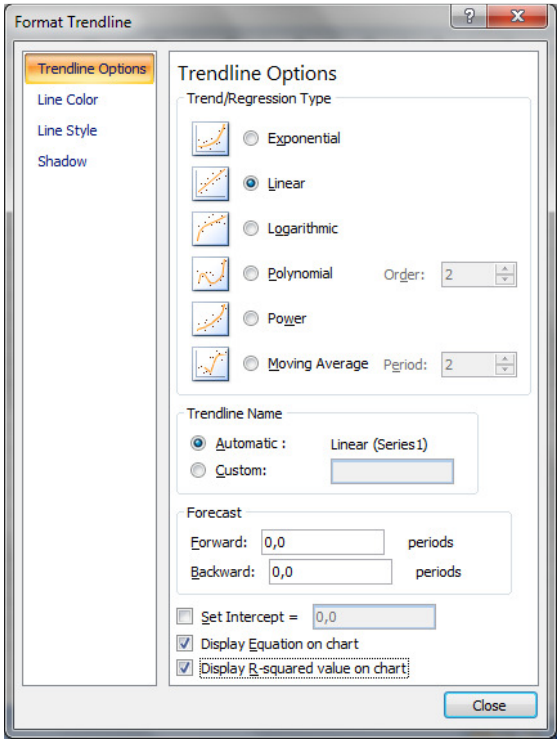


Fig.17. The choice of showing the equation and value of R^2 on the chart

Click the **OK** button. The final result is shown in Fig.18.

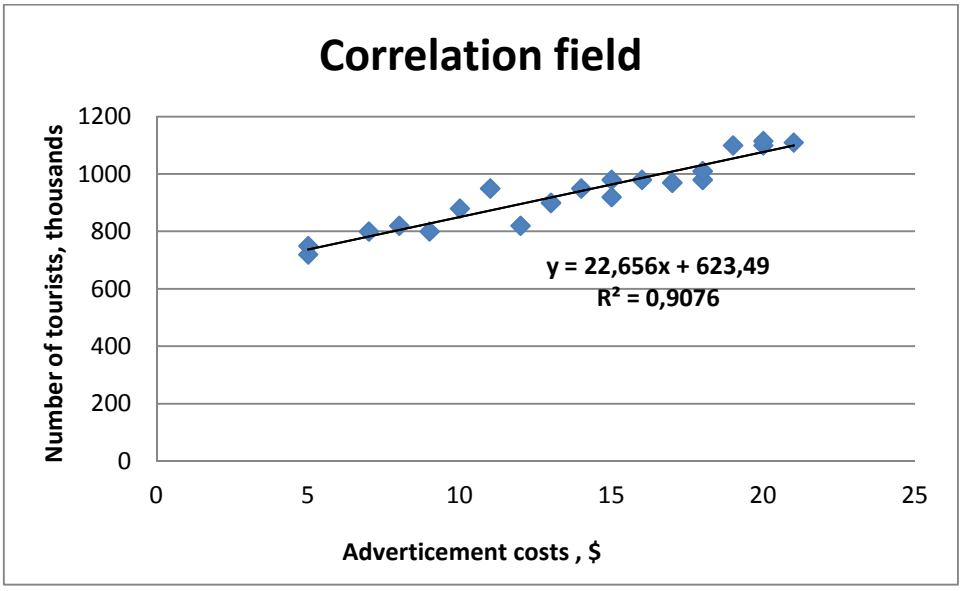


Fig.18. Trend line

Approach 3.

At first ascertain, that the package of analysis is activated.

1. Go to the **Data Analysis**. Select ***Regression*** tool in ***Analysis Tools*** list. Click on the ***OK*** button.

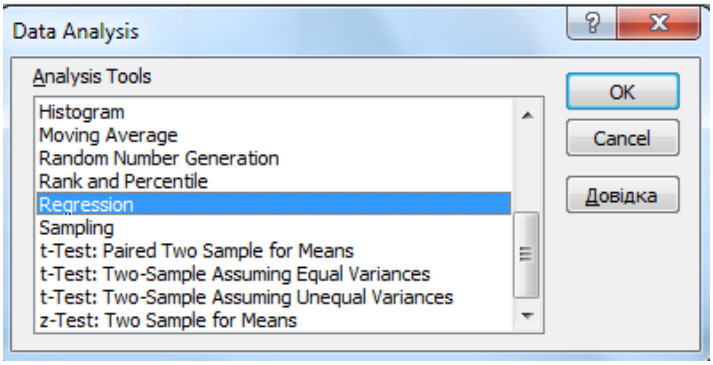


Fig.19. Choosing a ***Regression*** tool

2. You will see a dialog box ***Regression***.

- in the text box ***Input Y Range*** enter a range of values of the dependent variable **$C$2:$C$21**.
- in the text box ***Input X Range*** enter a range of values of the independent variables **$B$2:$B$21**.
- make sure, that in the field ***Confidence Level*** it is entered **95%** and in the ***Output Options*** block a switch is set in position of ***New Worksheet Ply***.
- Click on the button ***OK***.

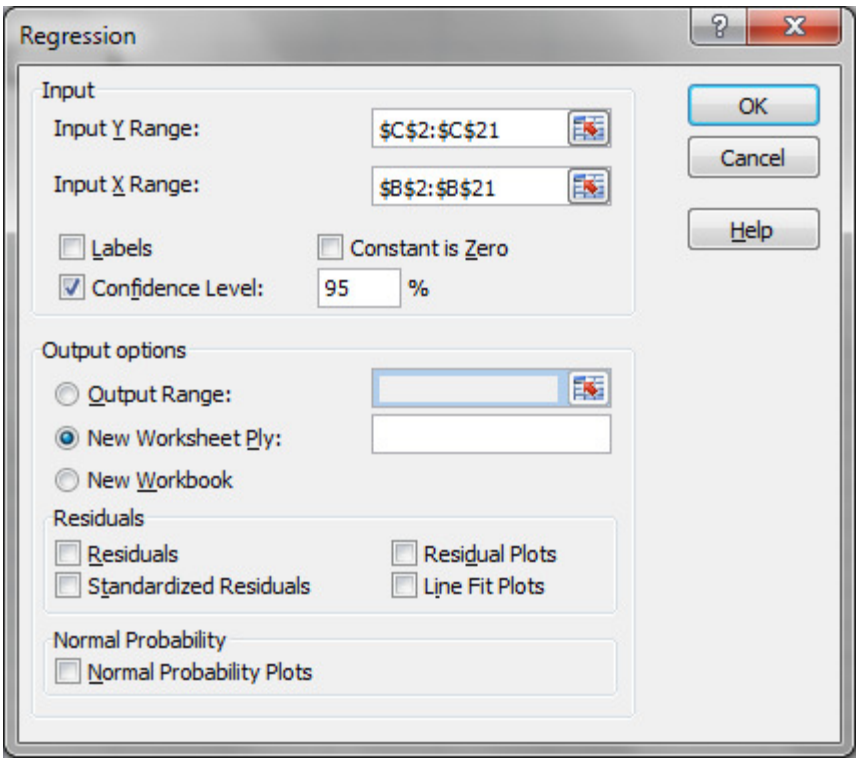


Fig.20. Setting regression model

3. As a result, a new worksheet will show the results of using the ***Regression*** tool.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,952663992							
5	R Square	0,907568681							
6	Adjusted R Square	0,902433608							
7	Standard Error	37,82190664							
8	Observations	20							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	252824,8108	252824,8108	176,7391876	9,53939E-11			
13	Residual	18	25748,93919	1430,496622					
14	Total	19	278573,75						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
17	Intercept	623,4945691	24,75187034	25,18979618	1,73703E-15	571,4928192	675,4963189	571,4928192	675,4963189
18	X Variable 1	22,65607553	1,704190957	13,29432915	9,53939E-11	19,07570319	26,23644786	19,07570319	26,23644786
19									

Fig.21. Summary of regression analysis

How to interpret the results? What do they mean?

❑ **Regression statistics**

- **Multiple R** = 0,952 (Pearson correlation coefficient) indicates the presence of a very strong relationship between the studied variables.

- **R Square** = 0,907 (determination coefficient) - characterizes the quality of the regression line, that is, the degree of correspondence between the regression model and the source data (always lies in the range from 0 to 1). R^2 = 0,907 - means that the regression model describes 90,7% of cases.
- **Adjusted R Square** gives the exact value (0,902).
- **Standard Error** = 37,73 - assessment of the adequacy of the linear model.

❑ **Analysis of Variance**

- Column **df** - the number of degrees of freedom (used when checking the adequacy of the model according to statistical tables).
- Column **SS** (sum of squares) - the given fate of the variance.
- Column **MS** - auxiliary values for calculating the Fisher criterion.
- Column **F** - Fisher criteria
- Column **Significance F** (of Fisher criteria) - assessment of the adequacy of the constructed model. If the significance is **F <0,05**, means that the model can be considered as adequate with a probability of **0,95**.

❑ **Odds**

- **Y-Intercept** - coefficient a = 625,031.
- **X Variable 1** - coefficient b = 22,507
- **Standard error, t-statistics -** auxiliary values used to check the significance of the model coefficients.
- **P-value -** an assessment of the significance of the model coefficients. If **P** is less than **0,05**, then with a probability of **0.95**, we can assume that the corresponding coefficient of the model is significant (that is, it cannot be considered equal to **0** and **Y** significantly depends on the corresponding **X**).
- **Lower** and **upper 95%** - confidence intervals for the model coefficients.

Conclusions

As a result of **Regression analysis** using MS Excel, we:

- constructed regression equations (Y = 22,507 * X + 625,031)
- established the form of dependence and the direction of the relationship between the variables (positive linear regression, which is expressed in the uniform growth of the function);
- evaluated the quality of the obtained regression line (R^2 = 0.907)
- were able to see deviations of the calculated data from the data of the initial set (on the graph)
- predicted future values of the dependent variable.

Task

Download data from the European Central Bank (www.ecb.int/stats/eurofxref/eurofxref-hist.zip) about the exchange rate of the euro against other foreign currencies. The downloaded file (in format *.csv) to convert in MS Excel format (*.xls or *.xlsx). In accordance with the individual variant leave in spreadsheet only two currencies (i.e. the X and Y variables), deleting the rest of the columns in the table. Using MS Excel to explore these data: 1) to get the descriptive statistics; 2) to perform correlation analysis; 3) to execute regression analysis.

Individual variants

1. USD-JPY
2. USD-DKK
3. GBP-JPY
4. HUF-GBP
5. PLN-SEK
6. TRY-HUF
7. IDR-DKK
8. INR-PLN
9. NZD-MXN
10. ILS-MYR
11. THB-HKD
12. CNY-BRL
13. PHP-TRY
14. MXN-INR
15. SEK-NZD
16. BRL-ILS
17. HKD-CNY
18. MYR-THB
19. MXN-CNY
20. TRY-IDR
21. GBP-CNY

PRACTICAL WORK #2
Correlation and Regression Analysis using RapidMiner

1. Linear Regression

In this tutorial, we will use linear regression, a statistical technique to perform estimation using RapidMiner tool. As you open the RapidMiner development environment, you will reach a welcome screen. Or, if you are using the RapidMiner for first time, you will be asked to define a repository, where your sources will be saved. To create new model, use ***File -> New***. You will end-up at process definition window. Here, you will be defining your model. For this tutorial, using linear regression.

For any data-mining task, the first step is about data. We need to read data. RapidMiner provides support for many file formats. For reading data, use Operator from ***Import->Data*** folder with in ***Operators*** window. For example, in our tutorial, we use ***Read CSV*** operator. Each operator in RapidMiner has input and output ports. For example, ***Read CSV*** has ***fil*** as input port and ***out*** as output port. As we drag and drop Operator on ***Process*** window, Operator output port automatically gets connected with the ***Process*** window ports named ***res***.

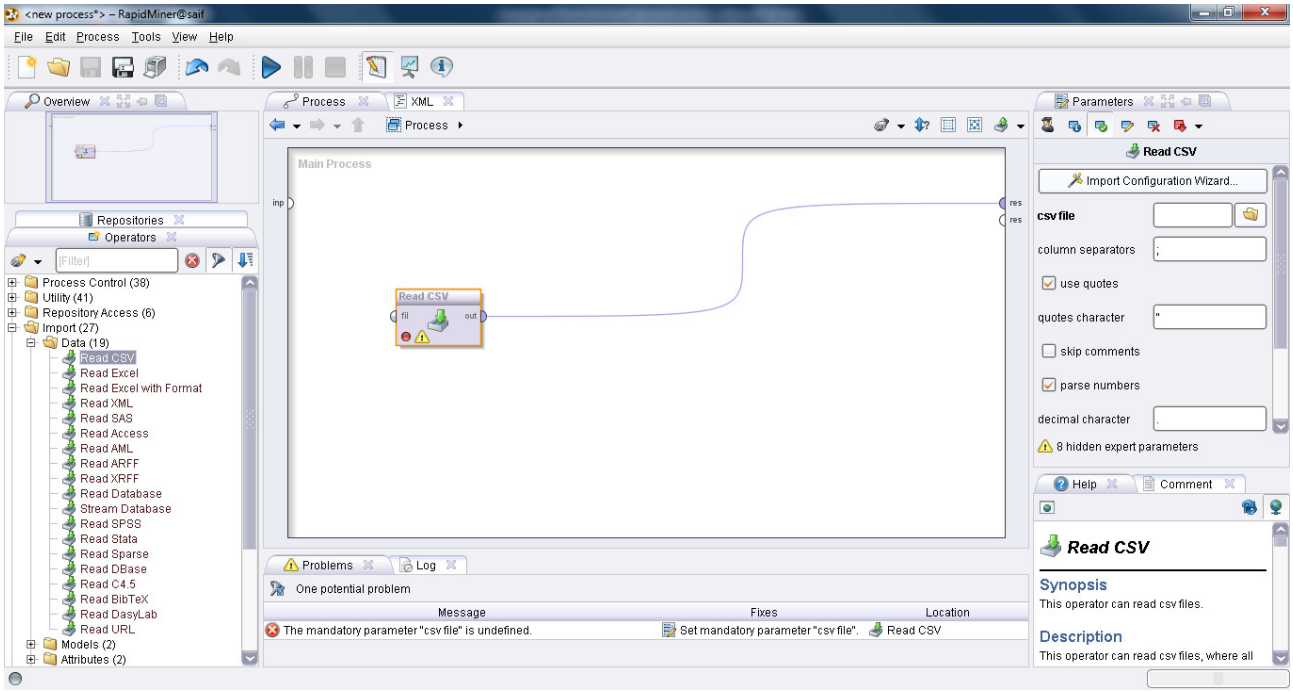


Fig.1 ***Read CSV*** operator

Make a click on ***Read CSV*** operator and check the parameters window. Using parameters window, you can set the Operator configuration values. We can use ***Import Configuration Wizard*** to import the data. ***Import Configuration Wizard*** asks you to select the data set in step 1. In this tutorial, we will be using European Central Bank Euro foreign exchange reference rates. The dataset can be found at

http://www.ecb.int/stats/eurofxref/eurofxref-hist.zip?a7b2dcf7c964da63a7e63ce6ae295613 . Once data set is selected, in step 2 it asks about how to parse the data file. It also gives visualization of how data will be read by the operator.

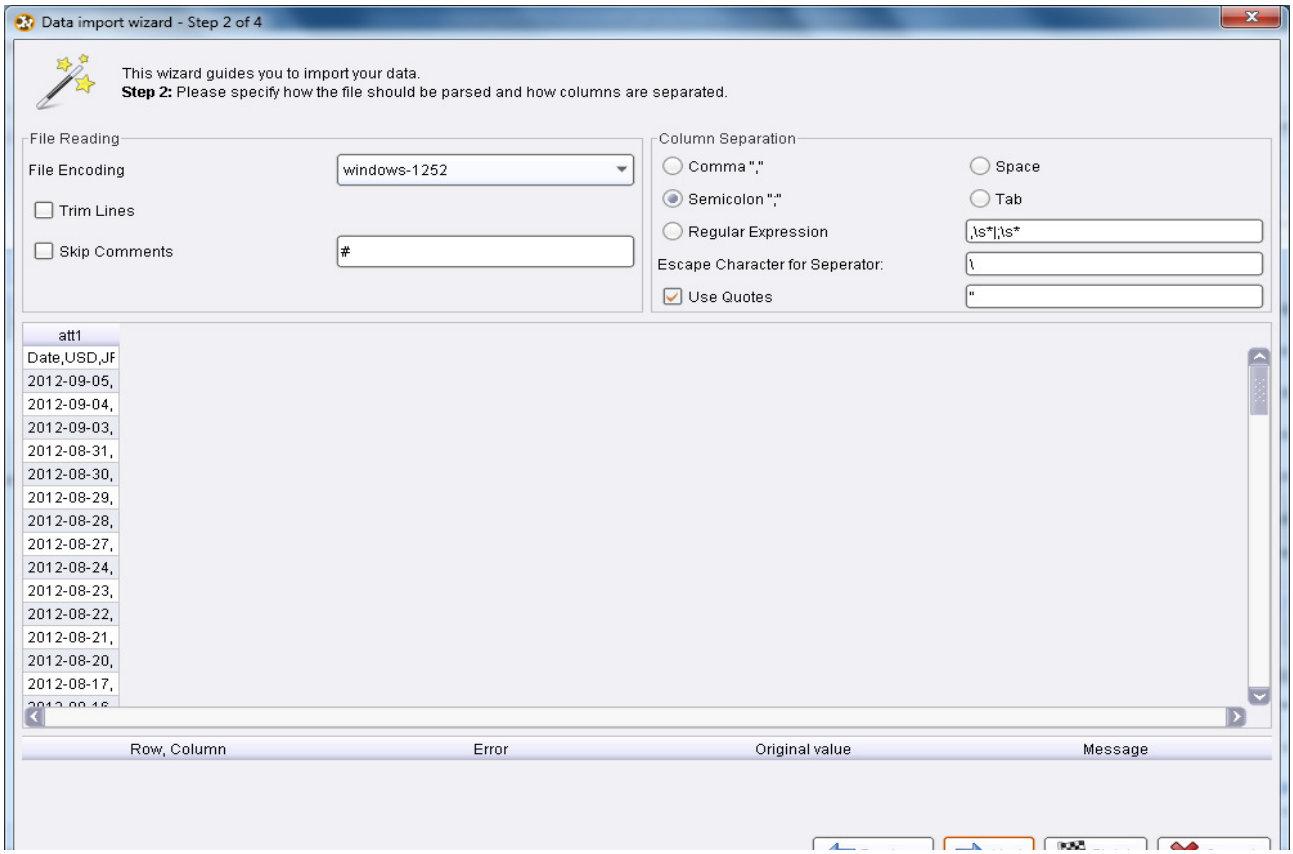


Fig.2. Visualization of how data will be read by the operator

As the file is comma separate, please select the ***Comma*** radio button from the top right ***Column Separation*** panel. In step 3, it show the data. In step 4, it allows you to select attributes, set attributes/columns names, identify their data types, and also allows us to identify their roles, i.e., either certain attribute is used as label for classification task or as a prediction column. For example in Figure below we have only selected USD and GBP, because we are only interested in estimating/predicting the value of GBP with change in USD. Furthermore, as linear regression is a bivariate method, it operate on two variables, i.e., provided the value of one it can predict/estimated another. As you can observe that we identify the role for GBP as ***label***, because we want to estimate/predict the value of GBP. ***Linear Regression*** operator requires at-least one attribute with ***label*** role. Another important point to consider here is that ***Linear Regression*** can only learn/estimate/predict numerical attributes, which means that none of selected attributes should be nominal.

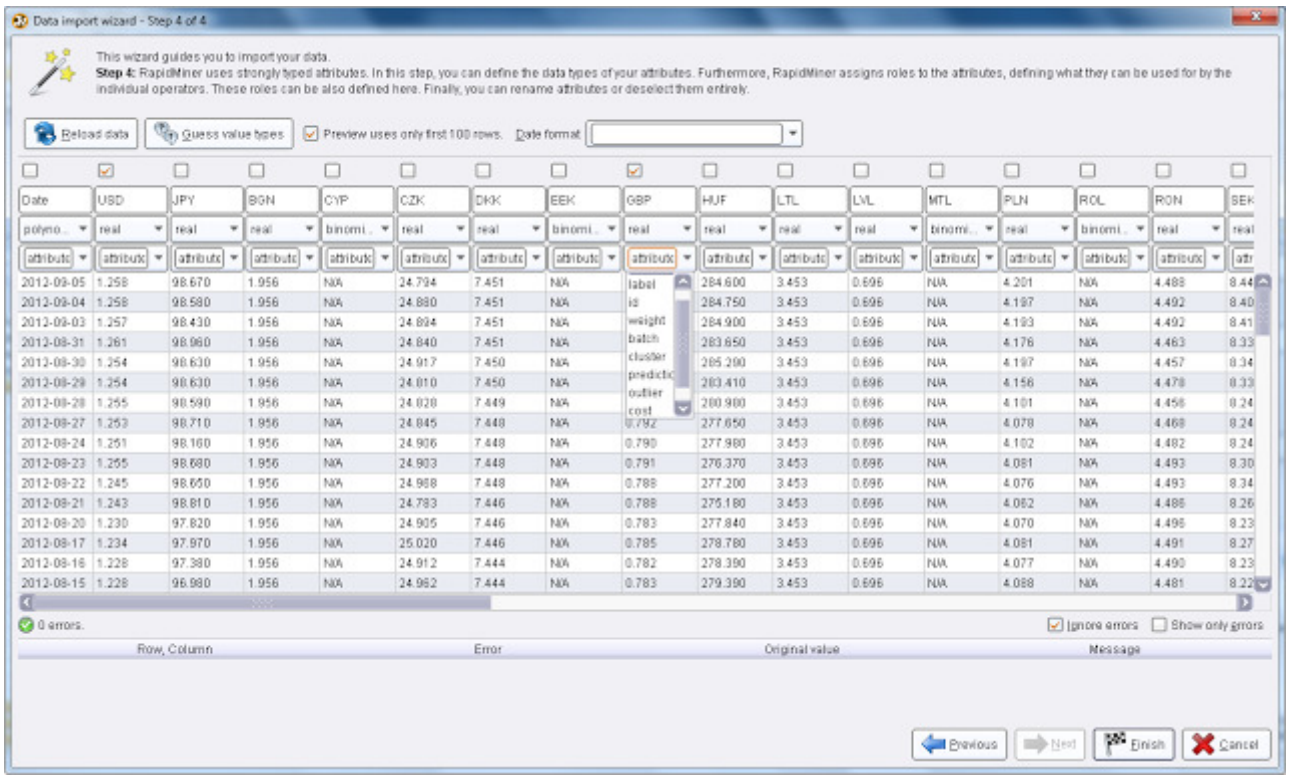


Fig. 3. Data Import Wizard

Once we finish the ***Import Configuration Wizard***, the red circle in ***Read CSV*** operator will change to yellow showing that configuration is complete.

Now in ***Operators*** window, select the ***Modeling -> Classification and Regression -> Function Fitting*** folder. This folder contains ***Linear Regression*** operator. Drag and drop ***Linear Regression*** operator on ***Process*** window and connect the output port ***out*** of ***Read CSV*** operator with input port ***tra*** of ***Linear Regression*** operator.

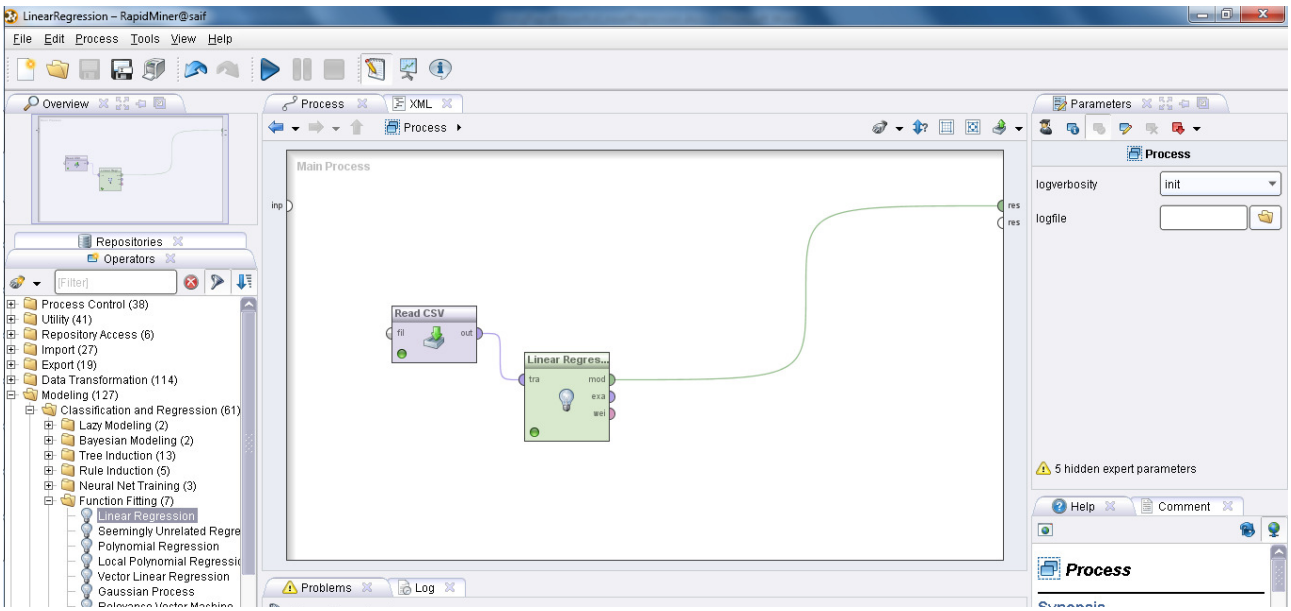


Fig. 4. ***Linear Regression*** operator.

Now we have completed the very basic step to create our model. As shown in figure-above, connect the output port named ***mod*** of ***Linear Regression*** operator with process window ***res*** port. To create the model, simple click the blue play button from tool bar or use ***Process -> Run*** menu or use ***F11***. A prompt will appear prompting "*Close old results before starting process*", better say '*Yes*'. Another prompt prompting to switch to result view will appear. Result can be viewed in tabular as well we text form as shown below:

Fig.5. ***Linear Regression*** results in text form

Or it can be viewed in tabular form as shown below:

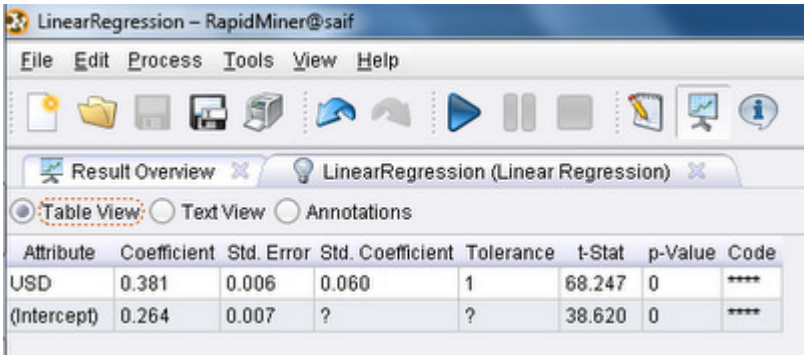


Fig.6. ***Linear Regression*** results in table form

Now as we have created a model, we can use it to estimate/predict the value of GBP using the value of USD. For this purpose, drop another ***Read CSV*** operator and name it ***New Data***. This will read the data values for which we want to perform prediction. Rename the old ***Read CSV*** operator to ***Historical Data***. Now to make prediction, we have to apply model. For this purpose, in RapidMiner, we have an ***Apply Model*** operator available in ***Modeling -> Model Application*** folder of ***Operators*** window. ***Apply Model*** operator has two input ports, i.e., ***mod*** to provide the model input and ***unl*** to provide the data to perform prediction. Connect the ***mod*** output port of ***Linear Regression*** operator with ***mod*** input port of ***Apply Model*** operator. Similarly connect new data at ***unl*** input port of ***Apply Model*** operator.

Connect ***lab*** output port of ***Apply Model*** operator with the ***res*** port of process window. In the end, you model should look like the one as shown in figure below:

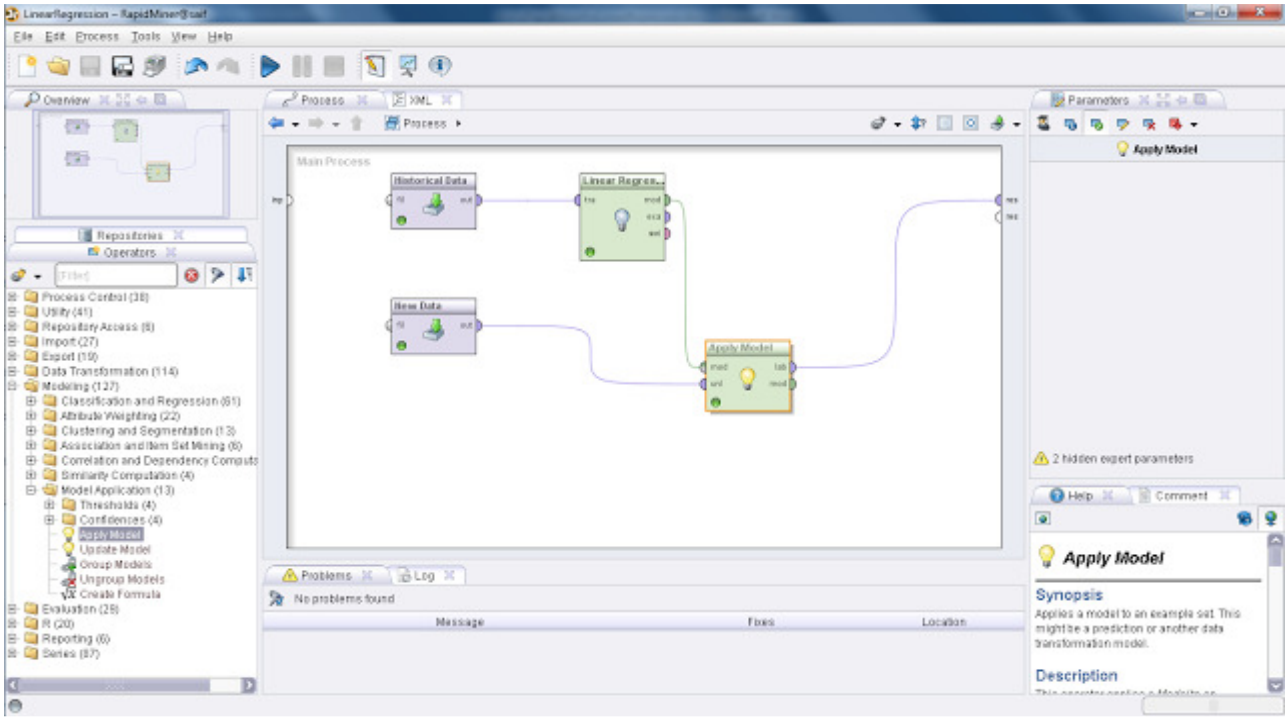
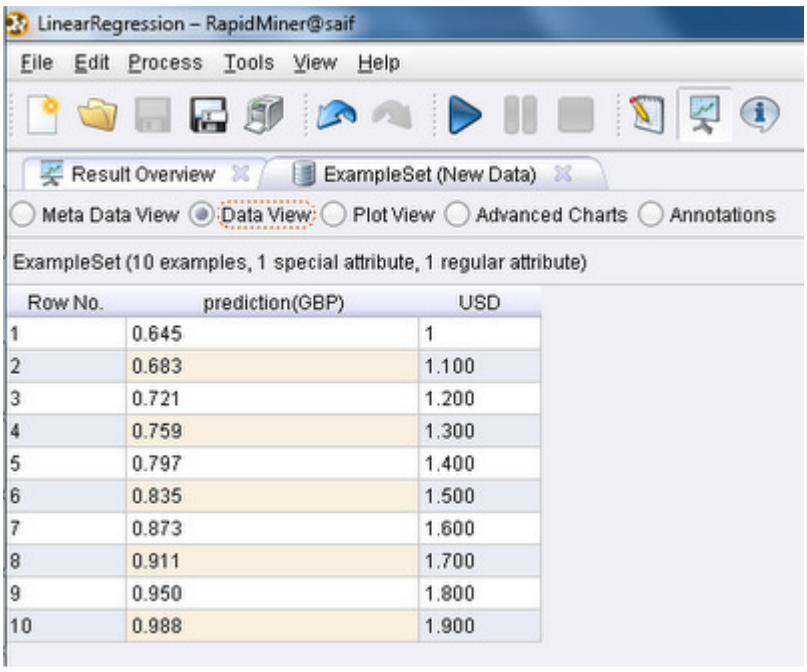


Fig.7. Model with ***mod*** operatotr

Now execute the process and you will get the predicted values, i.e., GBP values for your provided input, i.e., USD values. The output for our provided values is shown in figure below:



LinearRegression – RapidMiner@saif

File Edit Process Tools View Help

Result Overview | ExampleSet (New Data)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (10 examples, 1 special attribute, 1 regular attribute)

Row No.	prediction(GBP)	USD
1	0.645	1
2	0.683	1.100
3	0.721	1.200
4	0.759	1.300
5	0.797	1.400
6	0.835	1.500
7	0.873	1.600
8	0.911	1.700
9	0.950	1.800
10	0.988	1.900

Fig. 8. Predicted (GBP) values

2. Correlation

Realization of correlation analysis will consider on the example of the data placed to link: https://sites.google.com/site/dataminingforthemasses/home/Chapter04DataSet.csv. Information is taken from the database of company that supply with a fuel for heating of houses. It contains such attributes: *Insulation*, *Temperature*, *Heating_Oil*, *Num_Occupants*, *Avg_Age*, *Home_Size*. Want to explore how these factors affect fuel consumption. We choose correlation as a method of identifying the relationship between factors.

Import the *Chapter04DataSet.csv* into your RapidMiner data repository. Save it with the name ***Chapter4DataSet***. If you need a refresher on how to bring this data set into your RapidMiner repository, refer to steps from part I – Regression. The steps will be the same, with the exception of which file you select to import. Import all attributes, and accept the default data types. When you are finished, your repository should look similar to Figure:

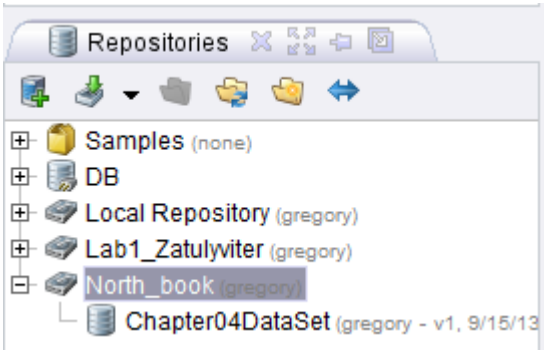


Fig.9. Repository view

If your RapidMiner application is not open to a new, blank process window, click the new process icon, or click ***File > New*** to create a new process. Drag your ***Chapter04DataSet*** into your main process window. Go ahead and click the run (play) button to examine the data set's meta data. If you are prompted, you may choose to save your new model. For this example, we will save the model as *Chapter04_Process*. Meta Data view is shown in the figure below:

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (1218 examples, 0 special attributes, 6 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	Insulation	integer	avg = 6.214 +/- 2.768	[2.000 ; 10.000]	0
regular	Temperature	integer	avg = 65.079 +/- 16.932	[38.000 ; 90.000]	0
regular	Heating_Oil	integer	avg = 197.394 +/- 56.248	[114.000 ; 301.000]	0
regular	Num_Occupants	integer	avg = 3.113 +/- 1.691	[1.000 ; 10.000]	0
regular	Avg_Age	real	avg = 42.706 +/- 15.051	[15.100 ; 72.200]	0
regular	Home_Size	integer	avg = 4.649 +/- 2.321	[1.000 ; 8.000]	0

Fig. 10. Meta Data view

We can see that our six attributes are shown. There are a total of 1,218 homes represented in the data set. Our data set appears to be very clean, with no missing values in any of the six attributes, and no inconsistent data apparent in our ranges or other descriptive statistics. If you wish, you can take a minute to switch to ***Data View*** to familiarize yourself with the data. It feels like these data are in good shape, and are in no further need of data preparation operators, so we are ready to move on to…

Switch back to design perspective. On the ***Operators*** tab in the lower left hand corner, use the search box and begin typing in the word "correlation". The tool we are looking for is called ***Correlation Matrix***. You may be able to find it before you even finish typing the full search term. Once you have located it, drag it over into your ***Process*** window and drop it into your stream. By default, the ***exa*** port will connect to the ***res*** port, but in this example, we are interested in creating a matrix of correlation coefficients that we can analyze. Thus, is it important for you to connect the ***mat*** (matrix) port to a ***res*** port, as illustrated in Figure.

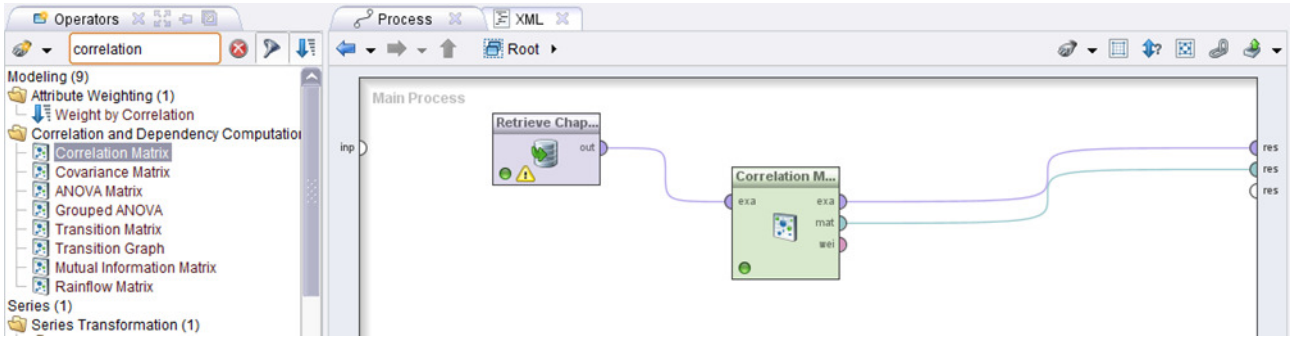
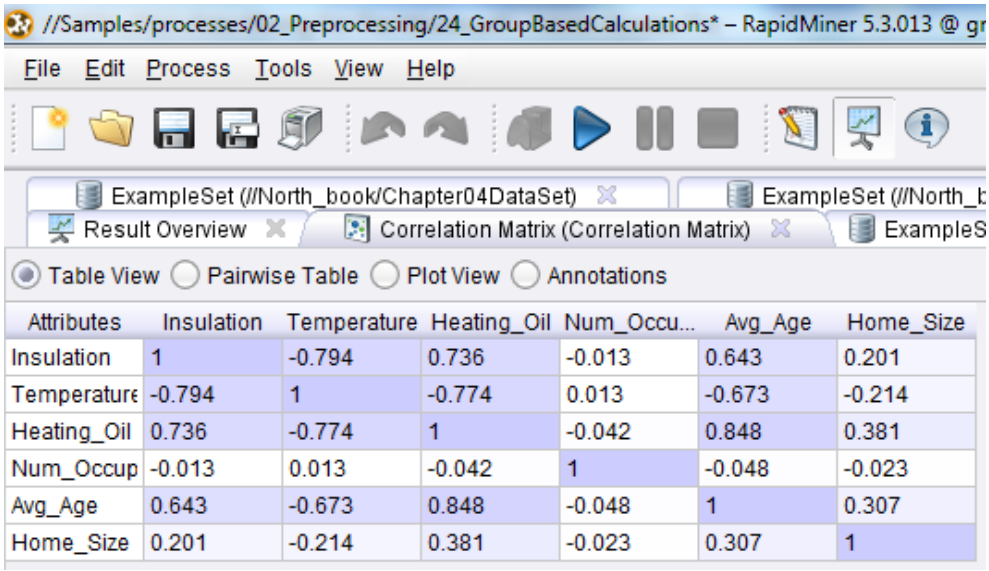


Fig. 11. Connection of the ***mat*** (matrix) port to a ***res*** port

Correlation is a relatively simple statistical analysis tool, so there are few parameters to modify. We will accept the defaults, and run the model. The results will be similar to Figure:



Attributes	Insulation	Temperature	Heating_Oil	Num_Occu...	Avg_Age	Home_Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num_Occup	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1

Fig. 12. Results table view

In Figure, we have our correlation coefficients in a matrix. Correlation coefficients are relatively easy to decipher. They are simply a measure of the strength of the relationship between each possible set of attributes in the data set. Because we have six attributes in this data set, our matrix is six columns wide by six rows tall. In the location where an attribute intersects with itself, the correlation coefficient is '1', because everything compared to itself has a perfectly matched relationship. All other pairs of attributes will have a correlation coefficient of less than one. To complicate matters a bit, correlation coefficients can actually be negative as well, so all correlation coefficients will fall somewhere between -1 and 1. We can see that this is the case in Figure, and so we can now move on to the step of evaluation.

All correlation coefficients between 0 and 1 represent positive correlations, while all coefficients between 0 and -1 are negative correlations. While this may seem straightforward, there is an important distinction to be made when interpreting the matrix's values. This distinction has to do with the direction of movement between the two attributes being analysed. Let's consider the relationship between the *Heating_Oil* consumption attribute, and the *Insulation* rating level attribute. The coefficient there, as seen in our matrix in Figure, is 0.736. This is a positive number, and therefore, a positive correlation. But what does that mean? Correlations that are positive mean that as one attribute's value rises, the other attribute's value also rises. However, a positive correlation also means that as one attribute's value falls, the other's also falls. Data analysts sometimes make the mistake in thinking that a negative correlation exists if an attribute's values are decreasing, but if its corresponding attribute's values are also decreasing, the correlation is still a positive one. This is illustrated in Figure.

Heating_Oil ↑	*Insulation* ↑
Heating_Oil ↓	*Insulation* ↓

Fig. 13. Mistakes in data analysts

Next, consider the relationship between the *Temperature* attribute and the *Insulation* rating attribute. In our matrix, we see that the coefficient there is -0.794. In this example, the correlation is negative, as illustrated in Figure

Temperature ↓	*Insulation* ↑
Temperature ↑	*Insulation* ↓

Fig 14. Mistakes in data analysts

So correlation coefficients tell us something about the relationship between attributes, and this is helpful, but they also tell us something about the strength of the correlation. As previously mentioned, all correlations will fall between 0 and 1 or 0 and -1. The closer a correlation coefficient is to 1 or to -1, the stronger it is. Figure illustrates the correlation strength along the continuum from -1 to 1.

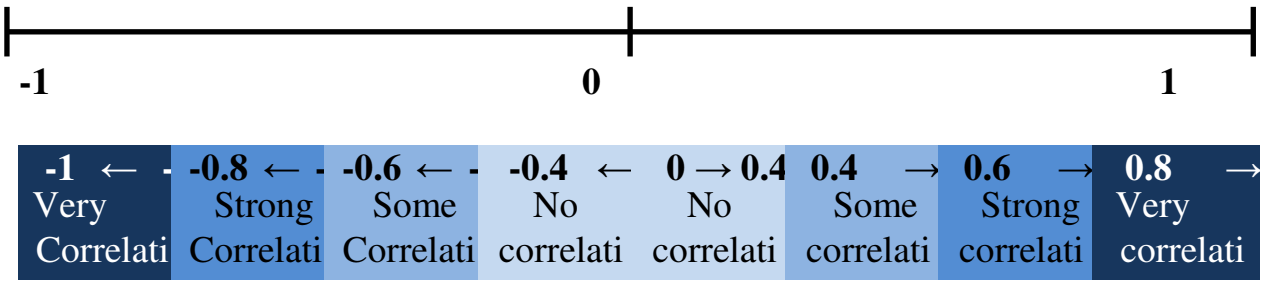


Fig .15. Correlation strength along the continuum from -1 to 1.

RapidMiner attempts to help us recognize correlation strengths through colour coding. In the Figure with matrix, we can see that some of the cells are tinted with shades of purple in graduated colours, in order to more strongly highlight those with stronger correlations. It is important to recognize that these are only general guidelines and not hard-and-fast rules.

RapidMiner provides the user with extensive visualization capabilities.

While still in results perspective, click on the ***ExampleSet*** tab (which exists assuming you left the ***exa*** port connected to a ***res*** port when you were in design perspective). Click on the ***Plot View*** radio button. Examine correlations that you found in your model visually by creating a scatter plot of your data. Choose one attribute for your ***x-Axis*** and a correlated one for your ***y-Axis***. Experiment with the ***Jitter*** slide bar.

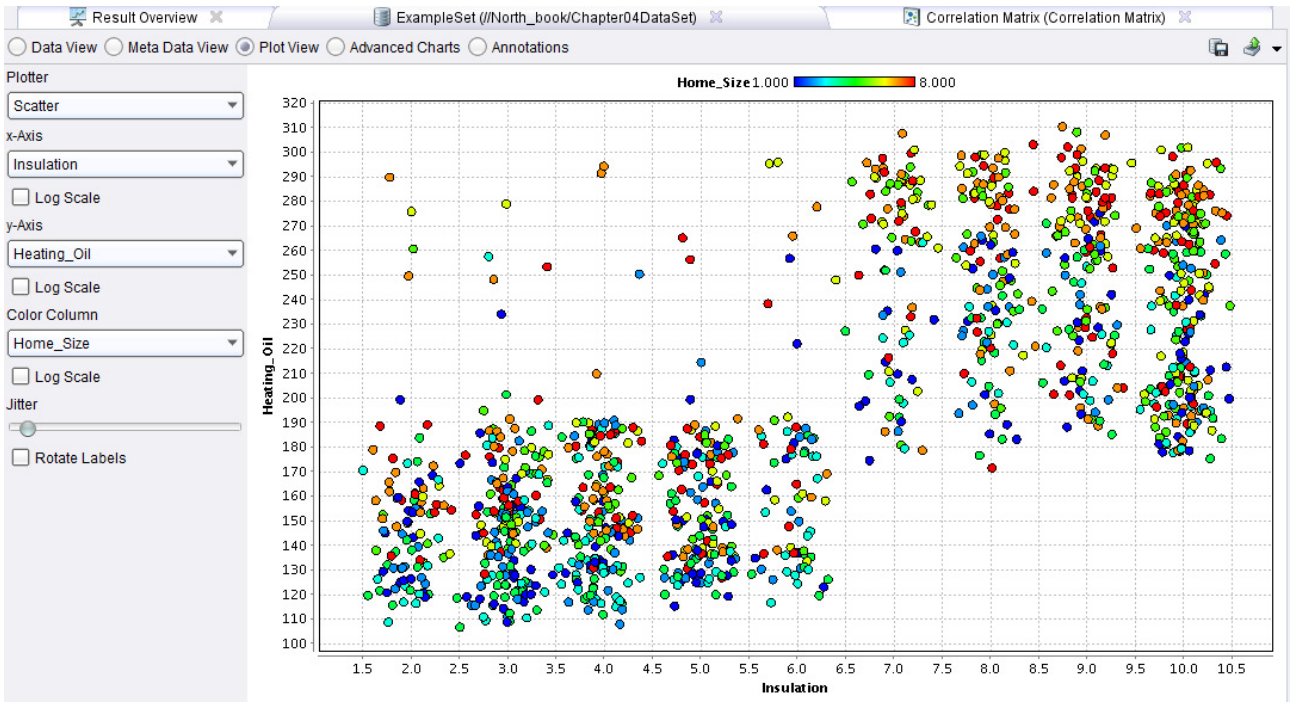


Fig. 16. Correlations plot view

For an additional visual experience, try a ***Scatter 3D*** or ***Scatter 3D Color*** plot. Note that with 3D plots in RapidMiner, you can click and hold to rotate your plot in order to better see the interactions between the data.

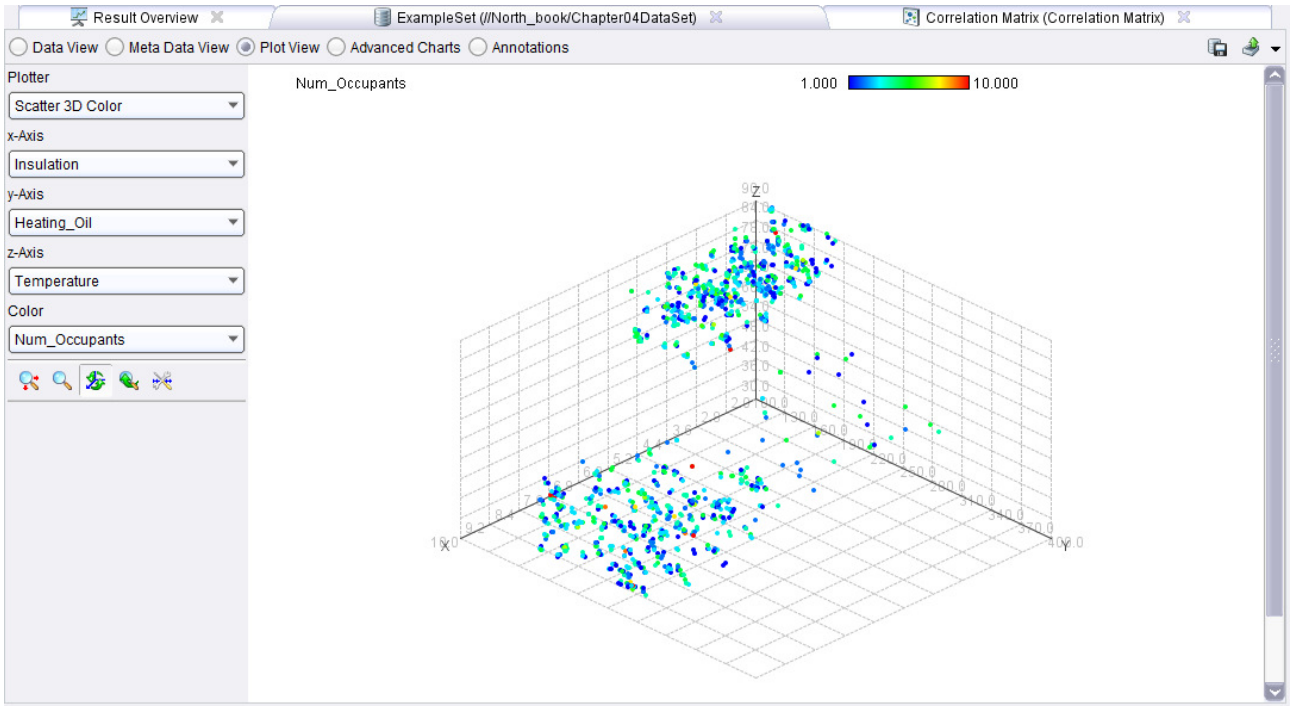


Fig. 17. Correlations ***Scatter 3D*** or ***Scatter 3D Color*** plots

Tasks

Download data from the European Central Bank (www.ecb.int/stats/eurofxref/eurofxref-hist.zip) about the exchange rate of the euro against other foreign currencies and place them in the repository of RapidMiner. In accordance with the individual variant leave in table only two currencies (i.e. the *X* and *Y* variables), deleting the rest of the columns in the table. Using RapidMiner to explore these data: 1) to get the descriptive statistics; 2) to perform correlation analysis; 3) to execute regression analysis. In a report necessarily to place a copy of the screenshots with the results and give their interpretation.

Individual variants

1. USD-JPY
2. USD-DKK
3. GBP-JPY
4. HUF-GBP
5. PLN-SEK
6. TRY-HUF

7. IDR-DKK
8. INR-PLN
9. NZD-MXN
10. ILS-MYR
11. THB-HKD
12. CNY-BRL
13. PHP-TRY
14. MXN-INR
15. SEK-NZD
16. BRL-ILS
17. HKD-CNY
18. MYR-THB
19. MXN-CNY
20. TRY-IDR
21. GBP-CNY

PRACTICAL WORK #3
Clustering Methods

Using the insurance company's claims database, we extract three attributes for 547 randomly selected individuals. The three attributes are the insured's weight in pounds as recorded on the person's most recent medical examination, their last cholesterol level determined by blood work in their doctor's lab, and their gender. As is typical in many data sets, the gender attribute uses 0 to indicate Female and 1 to indicate Male. We will use this sample data from database to build a cluster model and to help understand how company's clients, the health insurance policyholders, appear to group together based on their weights, genders and cholesterol levels. We should remember as we do this that means are particularly susceptible to undue influence by extreme outliers, so watching for inconsistent data when using the **k-Means clustering** data mining methodology is very important.

Data set for this example can be downloaded from the link https://sites.google.com/site/dataminingforthemasses/home/Chapter06DataSet.csv. Please note that this demo example. **Individual tasks (data sets) for each student listed below!**

If you would like to follow along with this example exercise, go ahead and download the data set now, and import it into your RapidMiner data repository. At this point, you are probably getting comfortable with importing CSV data sets into a RapidMiner repository, but remember that the steps are outlined in Chapter 3 if you need to review them. Be sure to designate the attribute names correctly and to check your data types as you import. Once you have imported the data set, drag it into a new, blank process window so that you can begin to set up your k-means clustering data mining model. Your process should look like Figure 1.

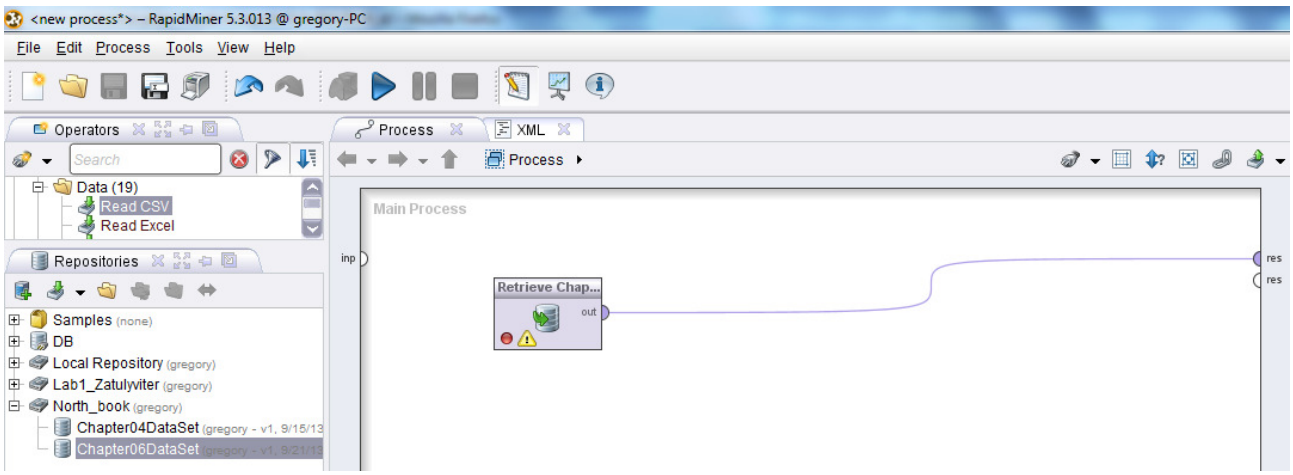


Fig.1. Data set *Chapter06DataSet* added to a new process

Go ahead and click the play button to run your model and examine the data set. In Figure 2 we can see that we have 547 observations across our three previously defined attributes. We can see the averages for each of the three attributes, along with their accompanying standard deviations and ranges. None of these values appear to be inconsistent (remember the earlier comments about using standard deviations to find statistical outliers). We have no missing values to handle, so our data appear to be very clean and ready to be mined.

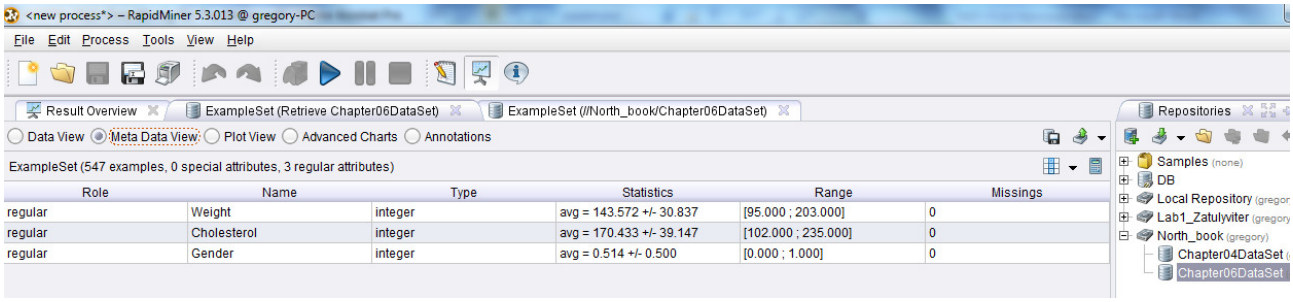


Fig.2. A view of data set's meta data

The '*k*' in *k-means* clustering stands for some number of groups, or clusters. The aim of this data mining methodology is to look at each observation's individual attribute values and compare them to the means, or in other words averages, of potential groups of other observations in order to find natural groups that are similar to one another. The *k-means* algorithm accomplishes this by sampling some set of observations in the data set, calculating the averages, or means, for each attribute for the observations in that sample, and then comparing the other attributes in the data set to that sample's means. The system does this repetitively in order to 'circle-in' on the best matches and then to formulate groups of observations which become the clusters. As the means calculated become more and more similar, clusters are formed, and each observation whose attributes values are most like the means of a cluster become members of that cluster. Using this process, *k-means* clustering models can sometimes take a long time to run, especially if you indicate a large number of "max runs" through the data, or if you seek for a large number of clusters (*k*). To build your *k-means* cluster model, complete the following steps:

1. Return to design view in RapidMiner if you have not done so already. In the operators search box, type "k-means" (be sure to include the hyphen). There are three operators that conduct *k-means* clustering work in RapidMiner. For this exercise, we will choose the first, which is simply named "***k-Means***". Drag this operator into your stream, and shown in Figure 3.

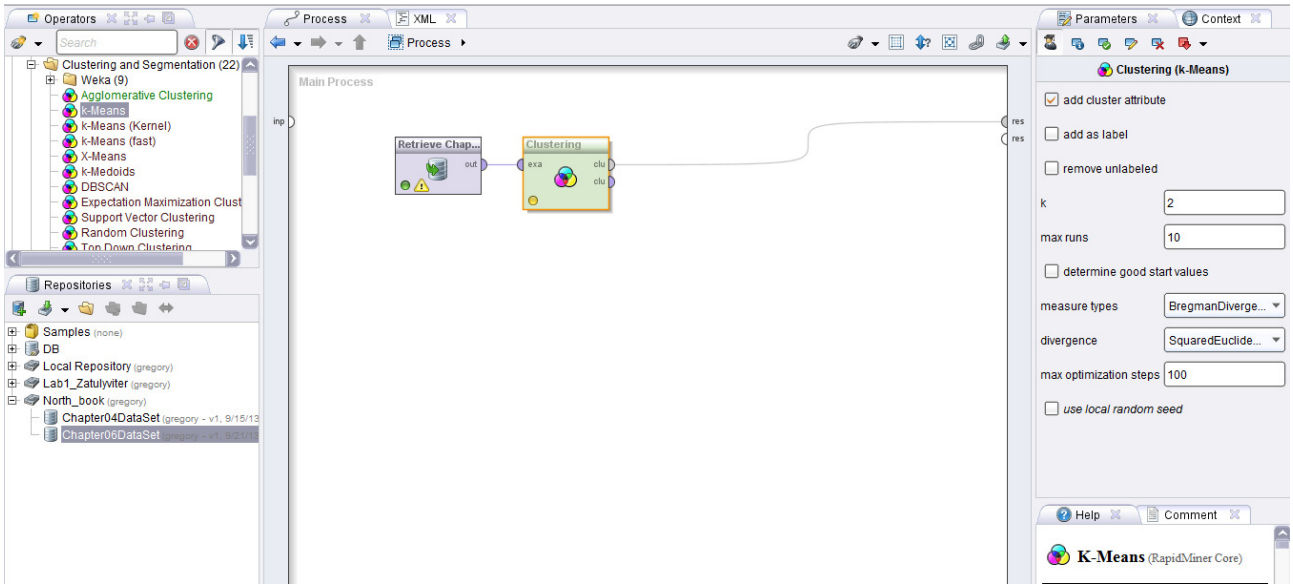


Fig.3. Adding the ***k-Means*** operator to model

2. Because we did not need to add any other operators in order to prepare our data for mining, our model in this exercise is very simple. We could, at this point, run our model and begin to interpret the results. This would not be very interesting however. This is because the default for our *k,* or our number of clusters, is 2, as indicated by the black arrow on the right hand side of Figure 3. This means we are asking RapidMiner to find only two clusters in our data. If we only wanted to find those with high and low levels of risk for coronary heart disease, two clusters would work. However, we have already recognized that there are likely a number of different types of groups to be considered. Simply splitting the data set into two clusters is probably not going to give us the level of detail. Because we felt that there were probably at least 4 potentially different groups, let's change the *k* value to four, as depicted in Figure 4. We could also increase of number of 'max runs', but for now, let's accept the default and run the model.

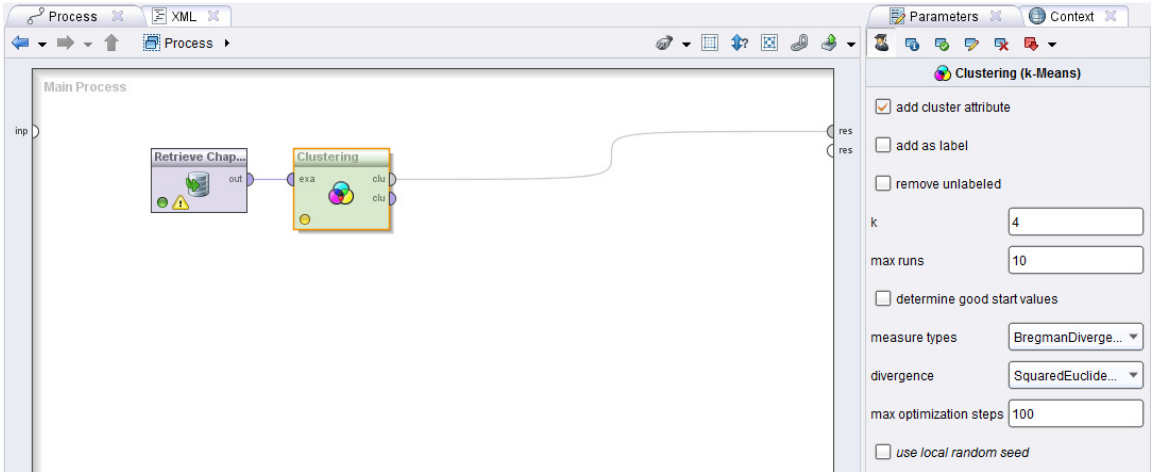


Fig.4. Setting the desired number of clusters for model

3. When the model is run, we find an initial report of the number of items that fell into each of our four clusters. (Note that the clustered are numbered starting from 0). In this particular model, our clusters are fairly well balanced.

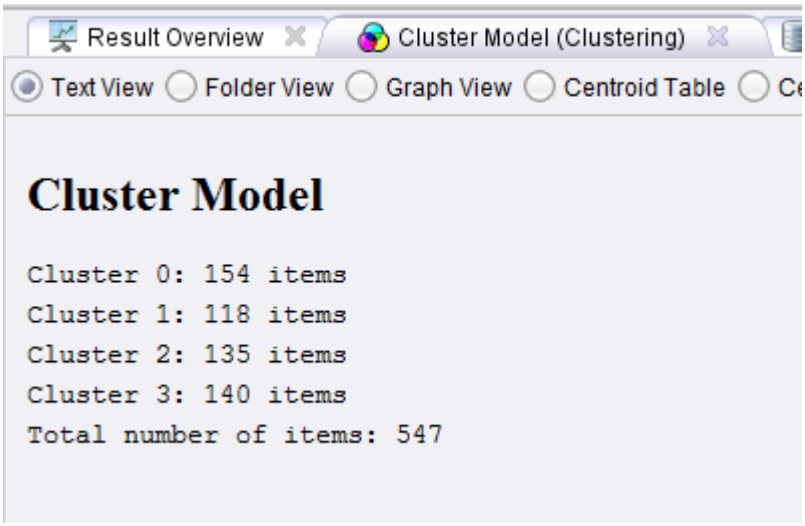


Fig.5. The distribution of observations across four clusters

We could go back at this point and adjust our number of clusters, our number of 'max runs', or even experiment with the other parameters offered by the ***k-Means*** operator. There are other options for measurement type or divergence algorithms. When you are satisfied with your model parameters, you can proceed to evaluation.

Refer back to Figure 5. There are a number of radio buttons, which allow us to select options for analyzing our clusters. We will start by looking at our Centroid Table. This view of our results, shown in Figure 6, gives the means for each attribute in each of the four clusters we created.

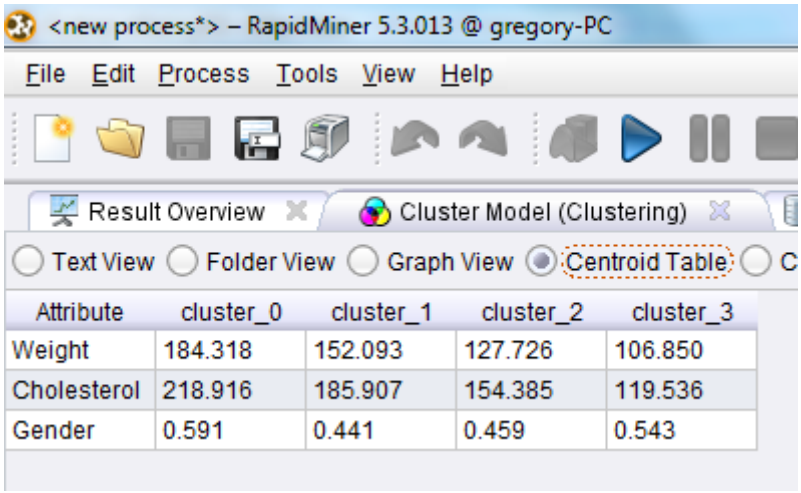


Fig.6. The means for each attribute in our four (*k*) clusters

We see in this view that cluster 0 has the highest average weight and cholesterol. With 0 representing Female and 1 representing Male, a mean of 0.591 indicates that we have more men than women represented in this cluster. Knowing that high cholesterol and weight are two key indicators of heart disease risk, policyholders need to pay attention to this. You should note that in this example, the clusters' numeric order (0, 1, 2, 3) corresponds to decreasing means for each cluster. This is coincidental. Sometimes, depending on your data set, cluster 0 might have the highest means, but cluster 2 might have then next highest, so it's important to pay close attention to your centroid values whenever you generate clusters.

Therefore, we know that cluster 0 is the group with highest risk. Who are the members of this highest risk cluster? We can find this information by selecting the ***Folder View*** radio button. ***Folder View*** is depicted in Figure 7.

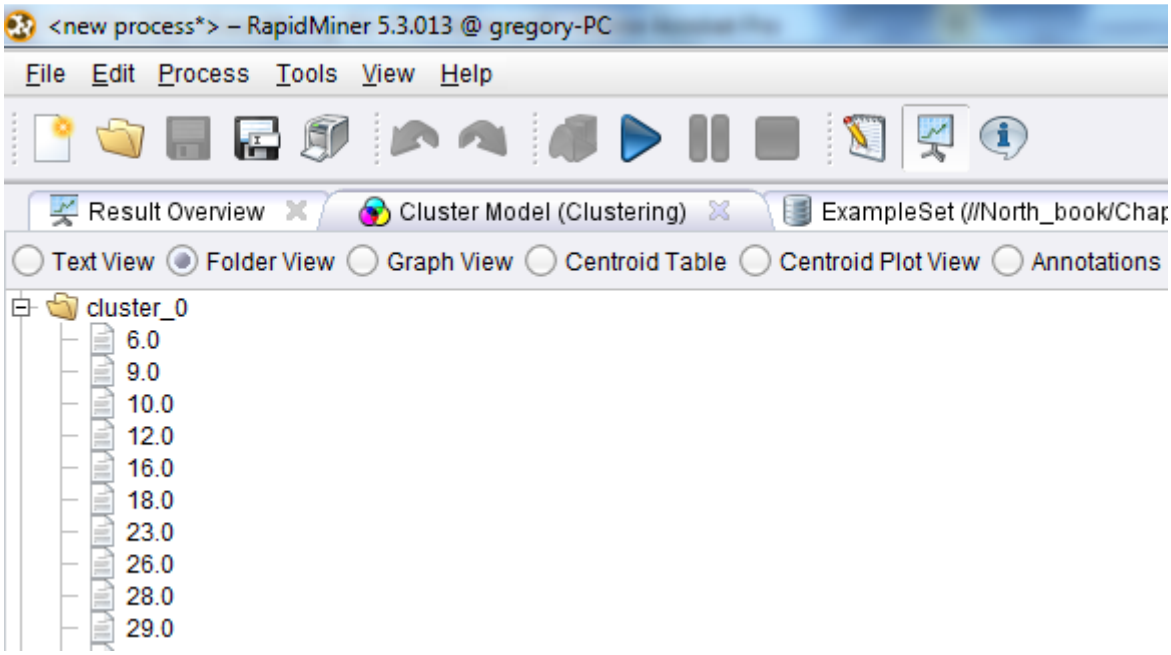


Fig.7. Folder view showing the observations included in Cluster 0

By clicking the small **+** sign next to **cluster_0** in ***Folder View***, we can see all of the observations that have means, which are similar to the mean for this cluster. Remember that these means are calculated for each attribute. You can see the details for any observation in the cluster by clicking on it. Figure 8 shows the results of clicking on observation 6 (6.0):

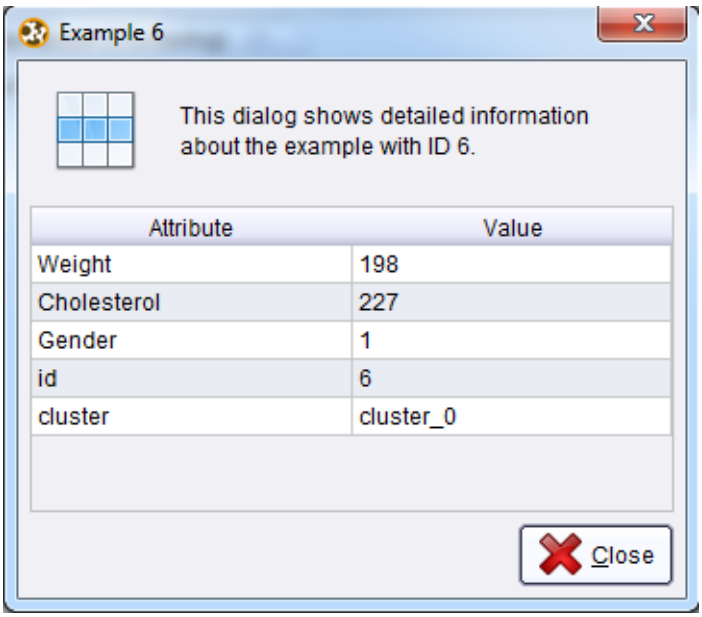


Fig.8. The details of an observation within cluster 0

Visual presentation of clustering results can be seen on the ***ExampleSet*** tab, choosing the ***Plot View*** button. The user has the possibility to choose the different types of diagrams. The scatter plots (2D and 3D) for our example shown on Figures 9 and 10.

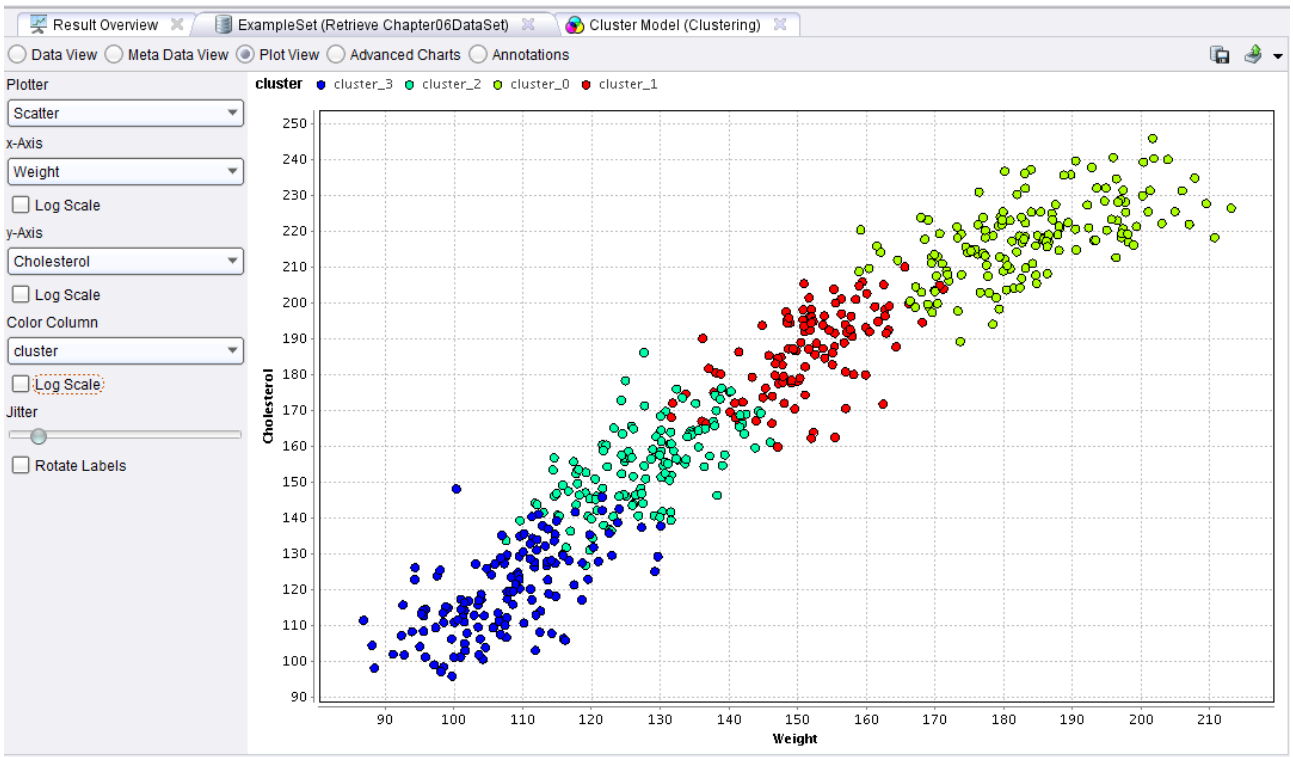


Fig.9. Scatter plot for clusters

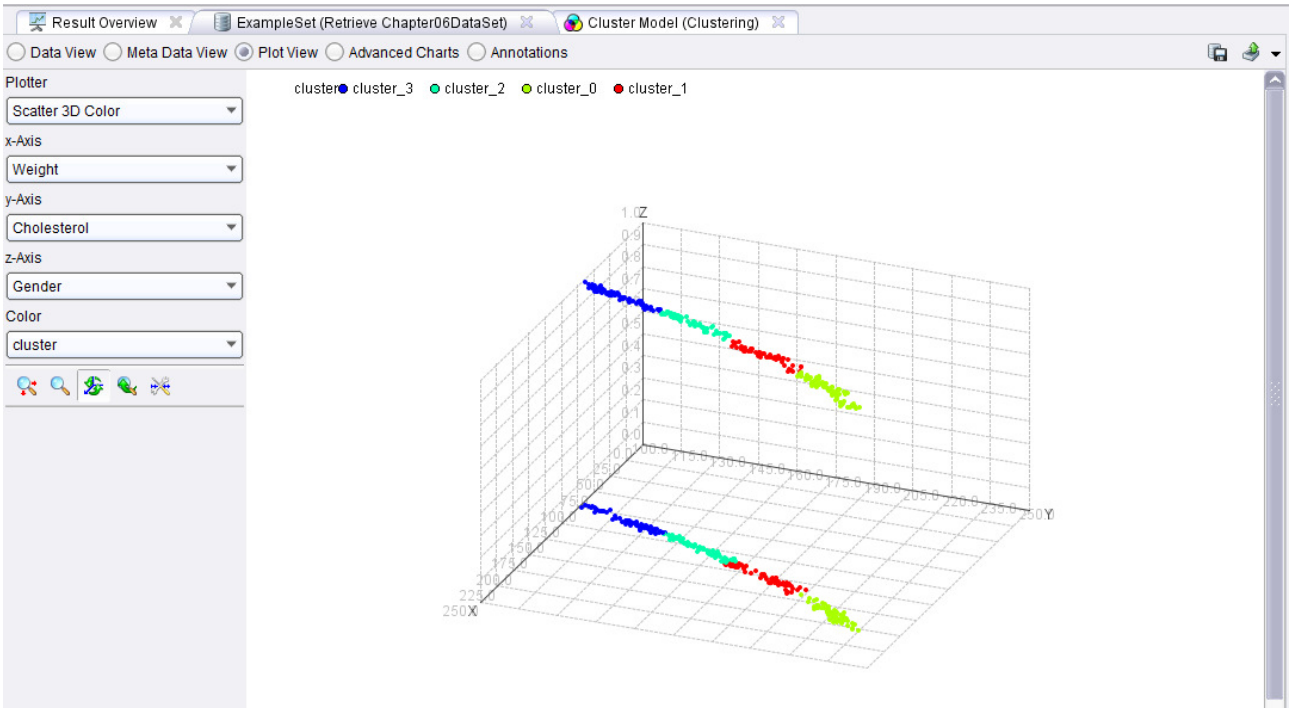


Fig.10. 3D Scatter plot for clusters

Tasks

Need to download a data set in RapidMiner repository according to the given link. Perform data clustering, preliminary setting number of clusters. Conduct experiments with a model, varying the number of clusters and other parameters. A graphical representation of the clustering results to place in the report.

Individual variants

1. http://archive.ics.uci.edu/ml/datasets/Sponge
2. http://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant
3. http://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
4. http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series
5. http://archive.ics.uci.edu/ml/datasets/Plants
6. http://archive.ics.uci.edu/ml/datasets/Amazon+Access+Samples
7. http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
8. http://archive.ics.uci.edu/ml/datasets/KEGG+Metabolic+Relation+Network+(Directed)
9. http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption
10. http://archive.ics.uci.edu/ml/datasets/seeds
11. http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

12. http://archive.ics.uci.edu/ml/datasets/3D+Road+Network+(North+Jutland,+Denmark)
13. http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities+Dataset
14. http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling
15. http://archive.ics.uci.edu/ml/datasets/Wine
16. http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
17. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
18. http://archive.ics.uci.edu/ml/datasets/Energy+efficiency
19. http://archive.ics.uci.edu/ml/datasets/Fertility

PRACTICAL WORK #4
Decision Trees

- Explain what decision trees are, how they are used and the benefits of using them.
- Recognize the necessary format for data in order to perform predictive decision tree mining.
- Develop a decision tree data mining model in RapidMiner using a training data set.
- Interpret the visual tree's nodes and leaves, and apply them to a scoring data set in order to deploy the model.
- Use different tree algorithms in order to increase the granularity of the tree's detail.

The construction of decision tree will consider on the example of data set of company that trades in eReaders. It is known that some customers want to buy gadgets immediately, other in the near time, yet other put aside a purchase on later. Marketers believe that they can categorize his company's customers into one of four groups that will eventually buy the new eReader: Innovators, Early Adopters, Early Majority or Late Majority. It helps to plan an advertising campaign for the respective consumer groups.

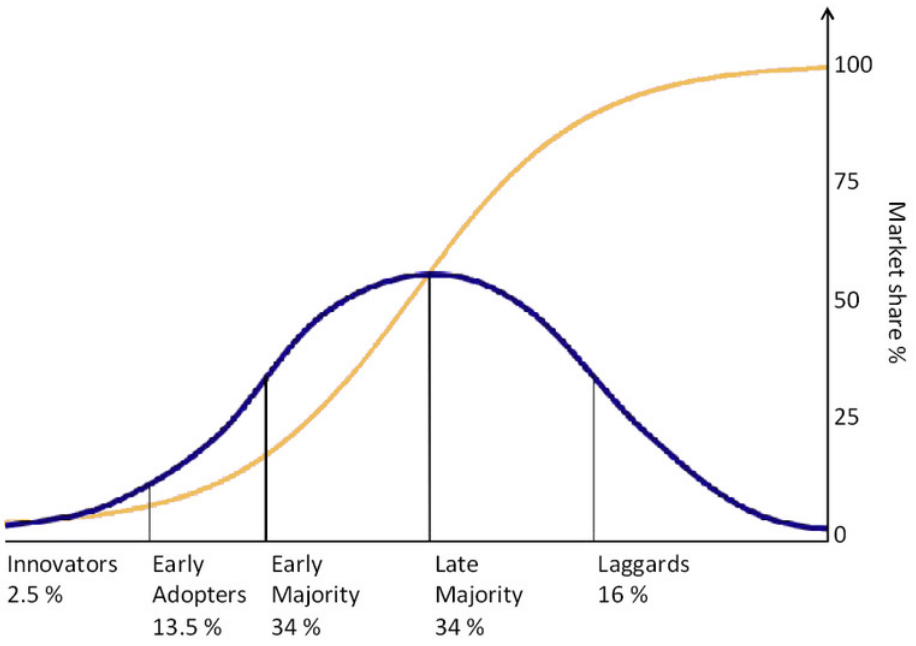


Fig.1. Sample of **decision tree** model

They have decided to use a **decision tree** model in order to find good early predictors of buying behavior. Because company does all of its business through its

web site, there is a rich data set of information for each customer, including items they have just browsed for, and those they have actually purchased. Marketers prepared two data sets. The training data set contains the web site activities of customers who bought the company's previous generation reader, and the timing with which they bought their reader. The second is comprised of attributes of current customers, which will buy the new eReader. They hope to figure out which category of adopter each person in the scoring data set will fall into based on the profiles and buying timing of those people in the training data set.

Data sets comprised of the following attributes: *User_ID*, *Gender*, *Age*, *Marital_Status*, *Website_Activity*, *Browsed_Electronics_12Mo*, *Payment_Method, Bought_Electronics_12Mo*, *Bought_Digital_Media_18Mo*, *Bought_Digital_Books*, *eReader_Adoption*.

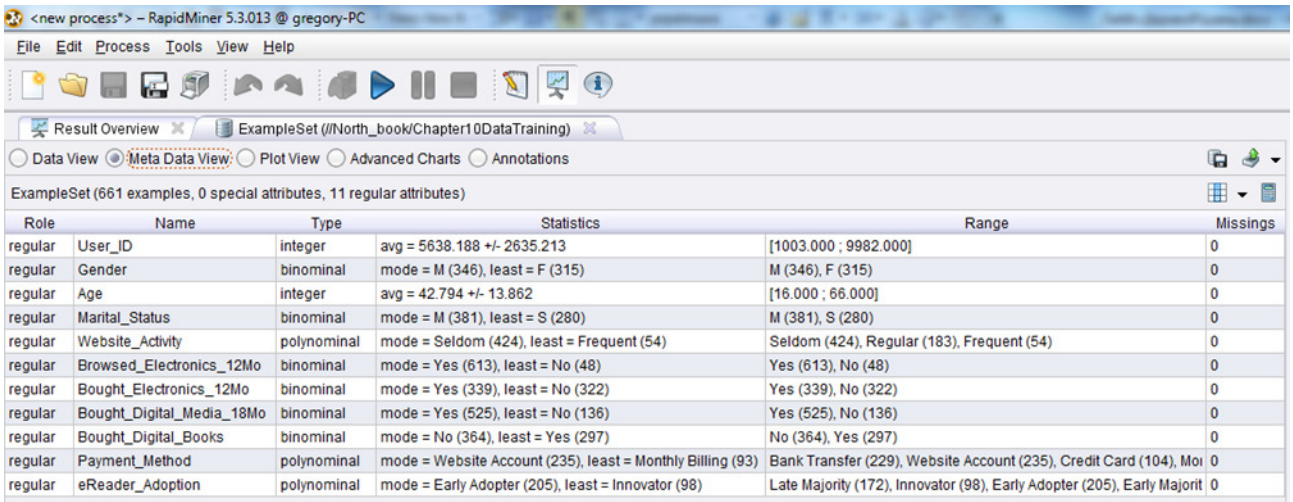
Our demo example consists of two data sets: *Chapter10DataSet_Training.csv* and *Chapter10DataSet_Scoring.csv*. Download these from

(https://sites.google.com/site/dataminingforthemasses/home/Chapter10DataSet_Scoring.csv

https://sites.google.com/site/dataminingforthemasses/home/Chapter10DataSet_Training.csv), then complete the following steps:

1. Import both data sets into your RapidMiner repository. You do not need to worry about attribute data types because the Decision Tree operator can handle all types of data. Be sure that you do designate the first row of each of the data sets as the attribute names as you import. Save them in the repository with descriptive names, so that you will be able to tell what they are.

2. Drag and drop both of the data sets into a new main process window. Rename the ***Retrieve*** objects as *Training* and *Scoring* respectively. Run your model to examine the data and familiarize yourself with the attributes.



Role	Name	Type	Statistics	Range	Missings
regular	User_ID	integer	avg = 5638.188 +/- 2635.213	[1003.000 ; 9982.000]	0
regular	Gender	binominal	mode = M (346), least = F (315)	M (346), F (315)	0
regular	Age	integer	avg = 42.794 +/- 13.862	[16.000 ; 66.000]	0
regular	Marital_Status	binominal	mode = M (381), least = S (280)	M (381), S (280)	0
regular	Website_Activity	polynominal	mode = Seldom (424), least = Frequent (54)	Seldom (424), Regular (183), Frequent (54)	0
regular	Browsed_Electronics_12Mo	binominal	mode = Yes (613), least = No (48)	Yes (613), No (48)	0
regular	Bought_Electronics_12Mo	binominal	mode = Yes (339), least = No (322)	Yes (339), No (322)	0
regular	Bought_Digital_Media_18Mo	binominal	mode = Yes (525), least = No (136)	Yes (525), No (136)	0
regular	Bought_Digital_Books	binominal	mode = No (364), least = Yes (297)	No (364), Yes (297)	0
regular	Payment_Method	polynominal	mode = Website Account (235), least = Monthly Billing (93)	Bank Transfer (229), Website Account (235), Credit Card (104), Mor	0
regular	eReader_Adoption	polynominal	mode = Early Adopter (205), least = Innovator (98)	Late Majority (172), Innovator (98), Early Adopter (205), Early Majorit	0

Fig.2. Metadata for *Scoring* set

3. Switch back to design perspective. While there are no missing or apparently inconsistent values in the data set, there is still some data preparation yet to do. First, the ***User_ID*** is an arbitrarily assigned value for each customer. The customer doesn't use this value for anything, it is simply a way to uniquely identify each customer in the data set. As such, it should not be included in the model as an independent variable.

We can handle this attribute in one of two ways. First, we can remove the attribute using a ***Select Attributes*** operator. Alternatively, we can try a new way of handling a non-predictive attribute. This is accomplished using the ***Set Role*** operator. Using the search field in the ***Operators*** tab, find and add ***Set Role*** operators to both your training and scoring streams. In the ***Parameters*** area on the right hand side of the screen, set the role of the ***User_ID*** attribute to '***id***'. This will leave the attribute in the data set throughout the model, but it won't consider the attribute as a predictor for the ***label*** attribute. Be sure to do this for both the training and scoring data sets, since the ***User_ID*** attribute is found in both of them (Figure 3).

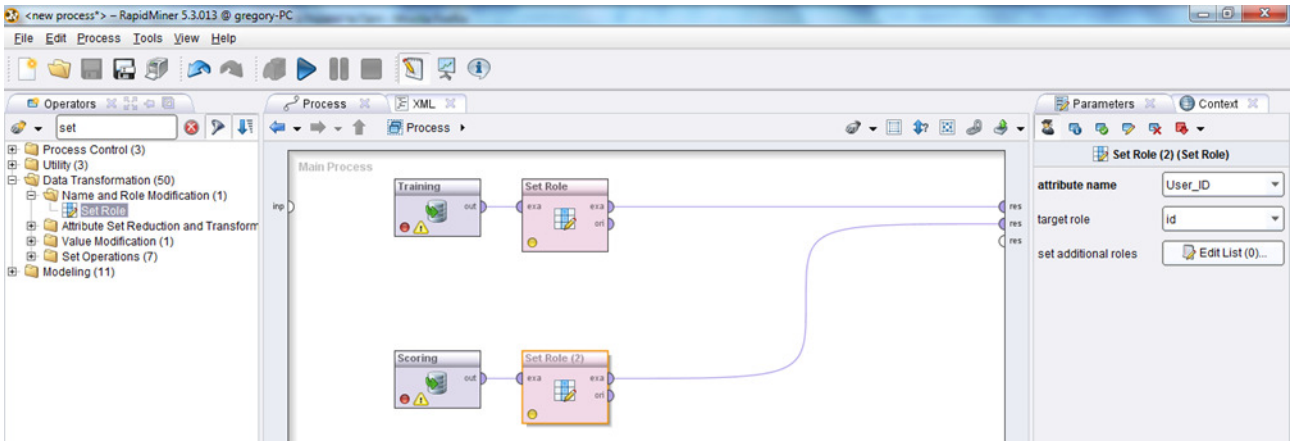


Fig.3. Setting the ***User_ID*** attribute to an '***id***' role

4. One of the nice side-effects of setting an attribute's role to '***id***' rather than removing it using a ***Select Attributes*** operator is that it makes each record easier to match back to individual people later, when viewing predictions in results perspective. You could use such an approach to leave in peoples' names or ID numbers so that you could easily know who to contact during the deployment phase of data mining project.

Before adding a ***Decision Tree*** operator, we still need to do another data preparation step. The ***Decision Tree*** operator expects the training stream to supply a '***label***' attribute. For this example, we want to predict which adopter group next-gen eReader customers are likely to be in. Therefore, our label will be ***eReader_Adoption*** (Figure 4).

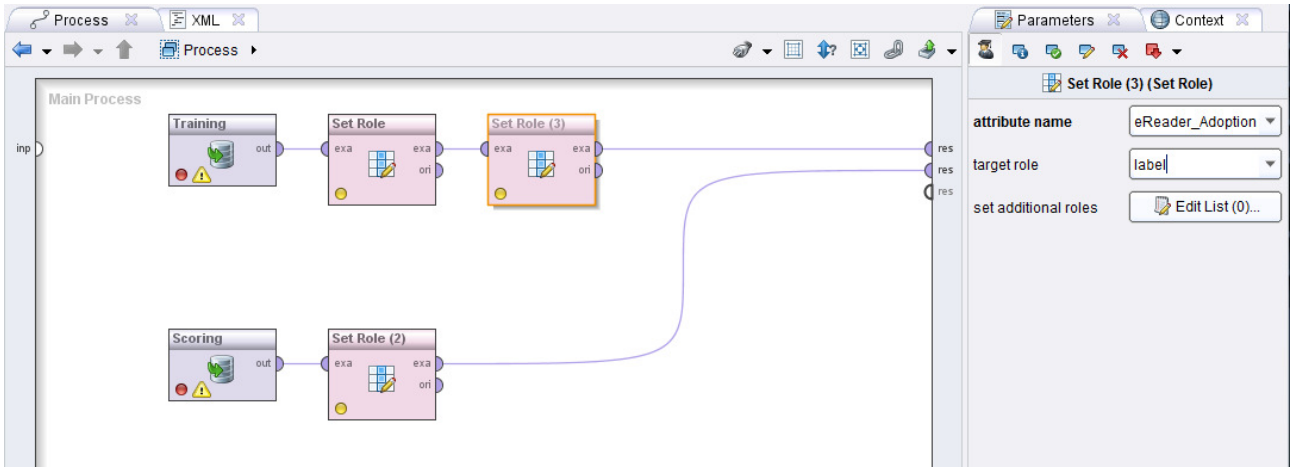


Fig.4. Setting the ***eReader_Adoption*** attribute as the ***label*** in our training stream

5. Next, search in the ***Operators*** tab for 'Decision Tree'. Select the basic ***Decision Tree*** operator and add it to your training stream as it is in Figure 5.

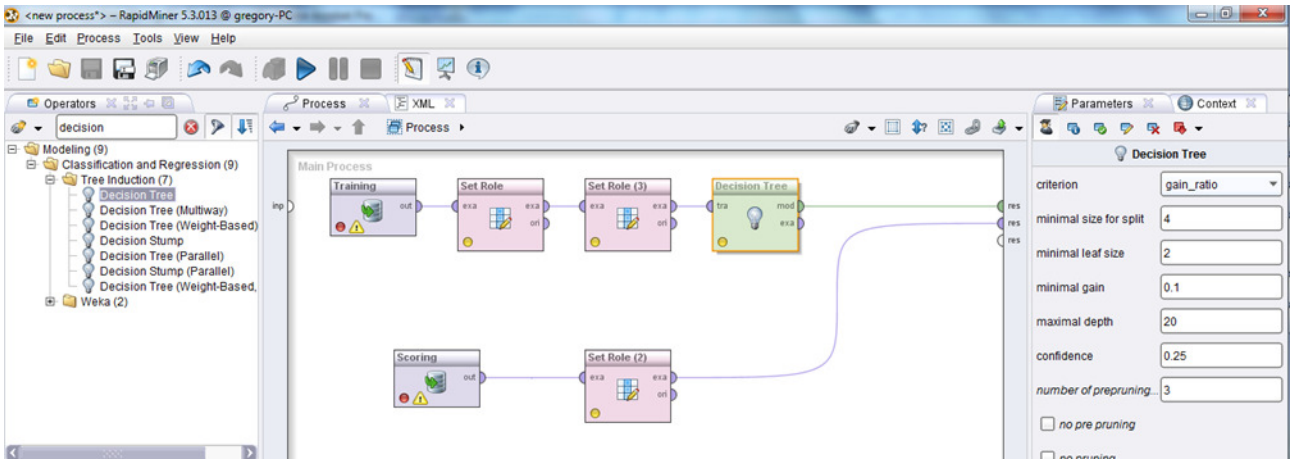


Fig.5. The ***Decision Tree*** operator added to model

6. Run the model and switch to the ***Tree (Decision Tree)*** tab in results perspective. You will see our preliminary tree (Figure 6).

In Figure 6, we can see what are referred to as nodes and leaves. The nodes are the gray oval shapes. They are attributes which serve as good predictors for our ***label*** attribute. The leaves are the multicolored end points that show us the distribution of categories from our ***label*** attribute that follow the branch of the tree to the point of that leaf. We can see in this tree that ***Website_Activity*** is our best predictor of whether or not a customer is going to adopt (buy) the company's new eReader. If the person's activity is *frequent* or *regular*, we see that they are likely to be an Innovator or Early Adopter, respectively. If however, they *seldom* use the web site, then whether or not they have *bought digital books* becomes the next best predictor of their *eReader adoption* category. If they have not *bought digital books* through the web site in the past, ***Age*** is another predictive attribute which forms a node, with younger folks

adopting sooner than older ones. This is seen on the branches for the two leaves coming from the ***Age*** node in Figure 6. Those who *seldom* use the company's website, have never *bought digital books* on the site, and are older than 25½ are most likely to land in the Late Majority category, while those with the same profile but are under 25½ are bumped to the Early Majority prediction. In this example, you can see how you read the nodes, leaves and branch labels as you move down through the tree.

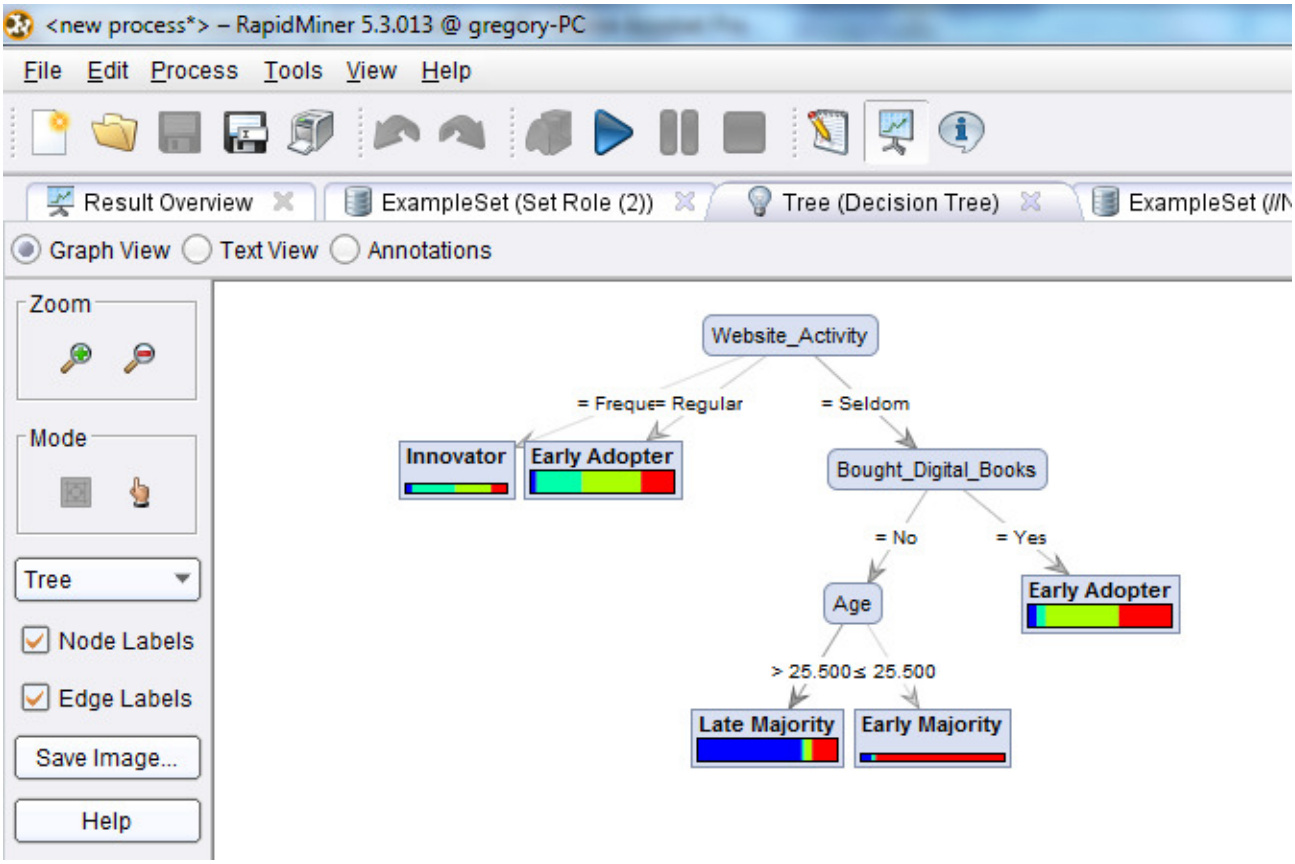


Fig.6. Decision tree results

Before returning to design perspective, take a minute to try some of the tools on the left hand side of the screen. The magnifying glasses can help you see your tree better, spreading out or compacting the nodes and leaves to enhance readability or to view more of a large tree at one time. Also, try using the 'hand' icon under Mode (see the arrow on Figure 6). This allows you to click and hold on individual leaves or nodes and drag them around to enhance your tree's readability. Finally, try hovering your mouse over one of the leaves in the tree. In Figure 7, we see a tool-tip hover box showing details of this leaf. Although our training data is going to predict that 'regular' web site users are going to be Early Adopters, the model is not 100% based on that prediction. In the hover, we read that in the training data set, 9 people who fit this profile are Late Adopters, 58 are Innovators, 75 are Early Adopters and 41 are Early Majority. When we get to Evaluation phase, we will see that this uncertainty in our data will translate into confidence percentages.

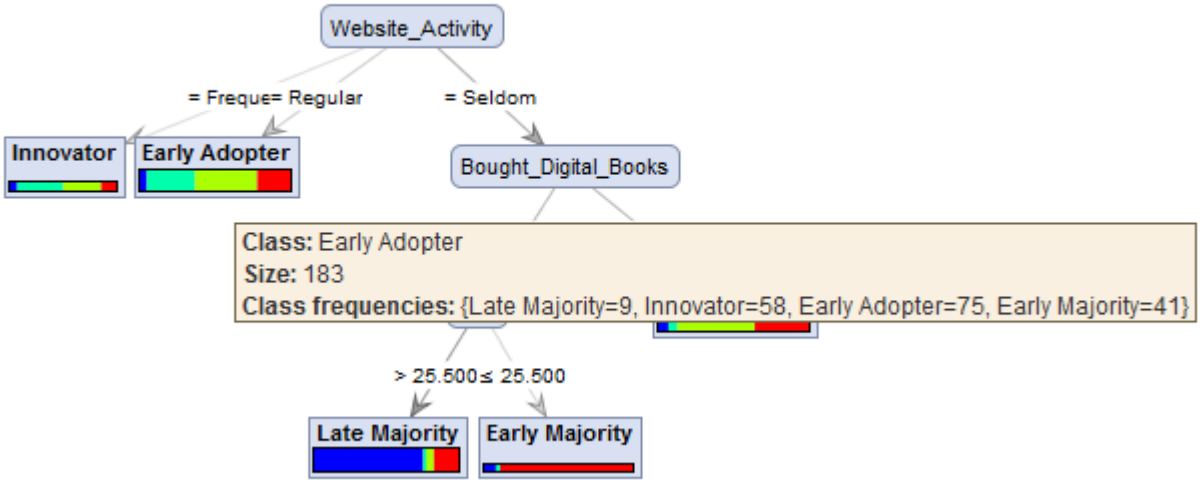


Fig.7. A tool-tip hover showing expanded leaf detail in our tree.

With our predictor attributes prepared, we are now ready to move on to modeling.

8. Return to design perspective. In the ***Operators*** tab search for and add an ***Apply Model*** operator, bringing your training and scoring streams together. Ensure that both the ***lab*** and ***mod*** ports are connected to ***res*** ports in order to generate our desired outputs (Figure 8).

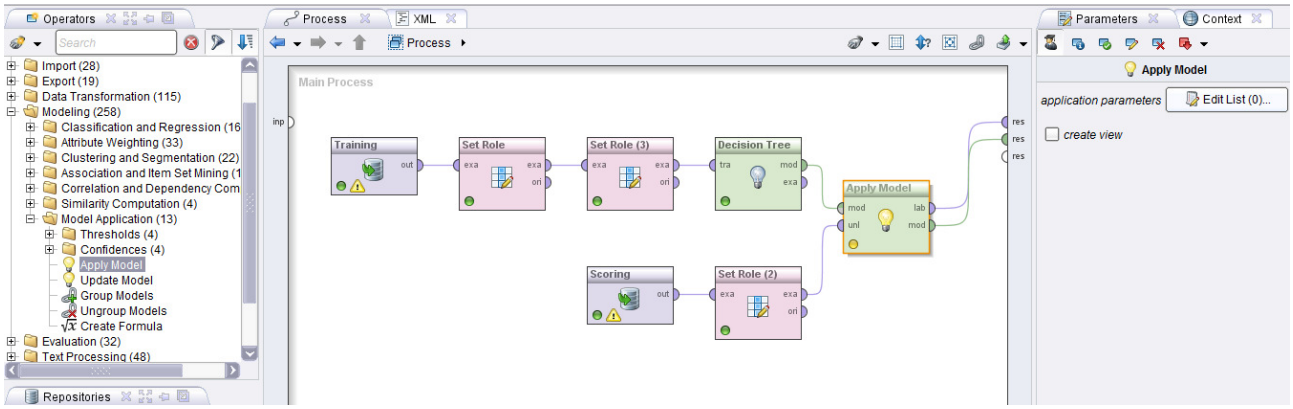
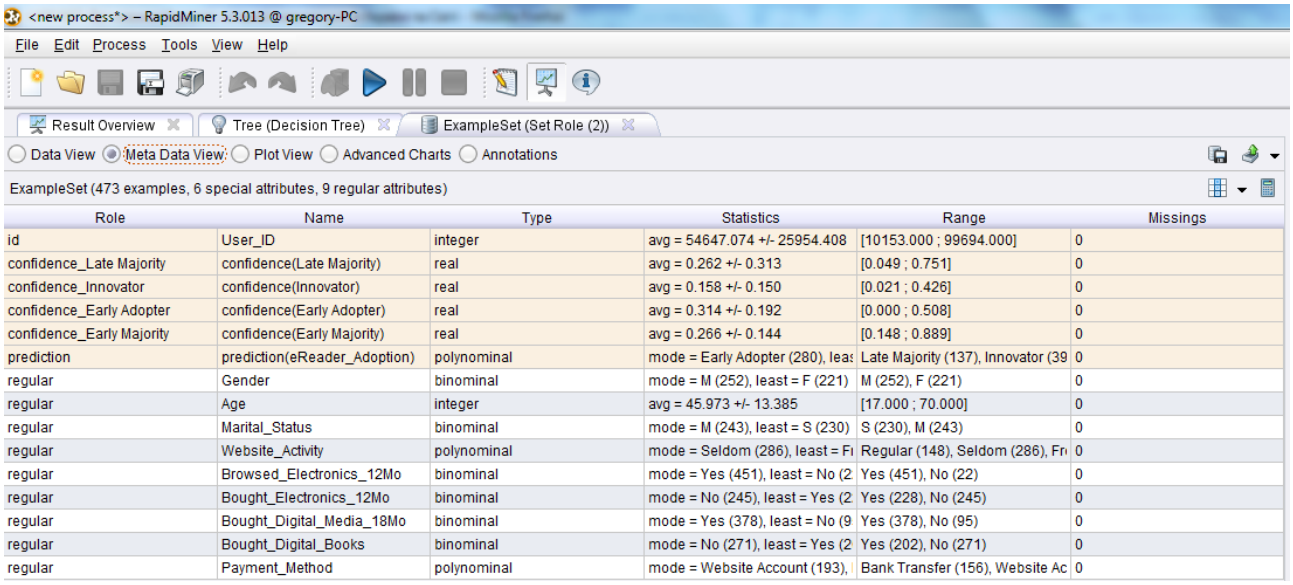


Fig.8. Applying the model to our scoring data, and outputting label predictions (*lab*) and a decision tree model (*mod*)

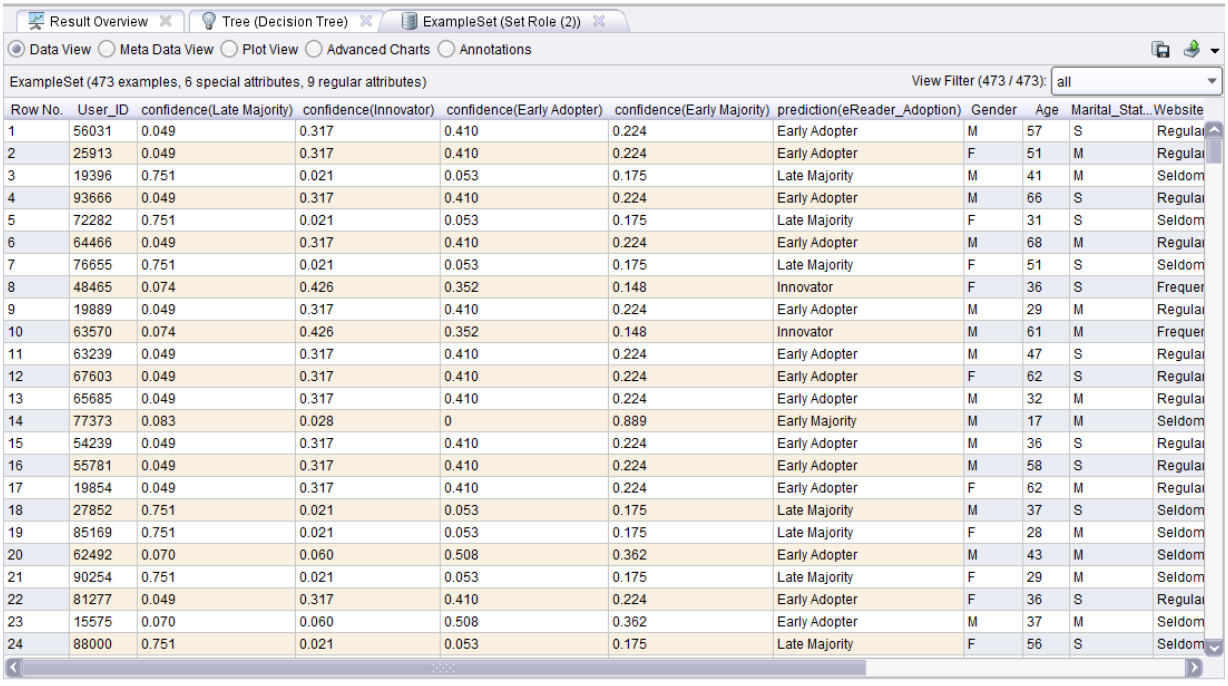
9. Run the model. You will see familiar results – the tree remains the same as it was in Figure 6, for now. Click on the ***ExampleSet*** tab next to the ***Tree*** tab. Our tree has been applied to our scoring data. RapidMiner has created confidence attributes, along with a prediction attribute.



Role	Name	Type	Statistics	Range	Missings
id	User_ID	integer	avg = 54647.074 +/- 25954.408	[10153.000 ; 99694.000]	0
confidence_Late Majority	confidence(Late Majority)	real	avg = 0.262 +/- 0.313	[0.049 ; 0.751]	0
confidence_Innovator	confidence(Innovator)	real	avg = 0.158 +/- 0.150	[0.021 ; 0.426]	0
confidence_Early Adopter	confidence(Early Adopter)	real	avg = 0.314 +/- 0.192	[0.000 ; 0.508]	0
confidence_Early Majority	confidence(Early Majority)	real	avg = 0.266 +/- 0.144	[0.148 ; 0.889]	0
prediction	prediction(eReader_Adoption)	polynominal	mode = Early Adopter (280), leas	Late Majority (137), Innovator (39	0
regular	Gender	binominal	mode = M (252), least = F (221)	M (252), F (221)	0
regular	Age	integer	avg = 45.973 +/- 13.385	[17.000 ; 70.000]	0
regular	Marital_Status	binominal	mode = M (243), least = S (230)	S (230), M (243)	0
regular	Website_Activity	polynominal	mode = Seldom (286), least = Fı	Regular (148), Seldom (286), Fr	0
regular	Browsed_Electronics_12Mo	binominal	mode = Yes (451), least = No (2	Yes (451), No (22)	0
regular	Bought_Electronics_12Mo	binominal	mode = No (245), least = Yes (2	Yes (228), No (245)	0
regular	Bought_Digital_Media_18Mo	binominal	mode = Yes (378), least = No (9	Yes (378), No (95)	0
regular	Bought_Digital_Books	binominal	mode = No (271), least = Yes (2	Yes (202), No (271)	0
regular	Payment_Method	polynominal	mode = Website Account (193),	Bank Transfer (156), Website Ac	0

Fig.9. Meta data for scoring data set predictions

10. Switch to ***Data View*** using the radio button. We see in Figure 10 the prediction for each customer's adoption group, along with confidence percentages for each prediction.



Row No.	User_ID	confidence(Late Majority)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)	prediction(eReader_Adoption)	Gender	Age	Marital_Stat...	Website
1	56031	0.049	0.317	0.410	0.224	Early Adopter	M	57	S	Regular
2	25913	0.049	0.317	0.410	0.224	Early Adopter	F	51	M	Regular
3	19396	0.751	0.021	0.053	0.175	Late Majority	M	41	M	Seldom
4	93666	0.049	0.317	0.410	0.224	Early Adopter	M	66	S	Regular
5	72282	0.751	0.021	0.053	0.175	Late Majority	F	31	S	Seldom
6	64466	0.049	0.317	0.410	0.224	Early Adopter	M	68	M	Regular
7	76655	0.751	0.021	0.053	0.175	Late Majority	F	51	S	Seldom
8	48465	0.074	0.426	0.352	0.148	Innovator	F	36	S	Frequer
9	19889	0.049	0.317	0.410	0.224	Early Adopter	M	29	M	Regular
10	63570	0.074	0.426	0.352	0.148	Innovator	M	61	M	Frequer
11	63239	0.049	0.317	0.410	0.224	Early Adopter	M	47	S	Regular
12	67603	0.049	0.317	0.410	0.224	Early Adopter	F	62	S	Regular
13	65685	0.049	0.317	0.410	0.224	Early Adopter	M	32	M	Regular
14	77373	0.083	0.028	0	0.889	Early Majority	M	17	M	Seldom
15	54239	0.049	0.317	0.410	0.224	Early Adopter	M	36	S	Regular
16	55781	0.049	0.317	0.410	0.224	Early Adopter	M	58	S	Regular
17	19854	0.049	0.317	0.410	0.224	Early Adopter	F	62	M	Regular
18	27852	0.751	0.021	0.053	0.175	Late Majority	M	37	S	Seldom
19	85169	0.751	0.021	0.053	0.175	Late Majority	F	28	M	Seldom
20	62492	0.070	0.060	0.508	0.362	Early Adopter	M	43	M	Seldom
21	90254	0.751	0.021	0.053	0.175	Late Majority	F	29	M	Seldom
22	81277	0.049	0.317	0.410	0.224	Early Adopter	F	36	S	Regular
23	15575	0.070	0.060	0.508	0.362	Early Adopter	M	37	M	Seldom
24	88000	0.751	0.021	0.053	0.175	Late Majority	F	56	S	Seldom

Fig.10. Predictions and their associated confidence percentages using our decision tree

Also, in Fig. 10 we see that there are four confidence attributes, corresponding to the four possible values in the label (***eReader_Adoption***). We interpret these the same way that we did with the other models though – the percentages add to 100%, and the prediction is whichever category yielded the highest confidence percentage.

RapidMiner is very (but not 100%) convinced that person 77373 (Row 14, Figure 10) is going to be a member of the early majority (88.9%). Despite some uncertainty, RapidMiner is completely sure that this person is not going to be an early adopter (0%).

11. We have already begun to evaluate our model's results, but what if we feel like we'd like to see greater detail, or granularity in our model. Surely, some of our other attributes are also predictive in nature. Remember that CRISP-DM is cyclical in nature, and that in some modeling techniques, especially those with less structured data, some back and forth trial-and-error can reveal more interesting patterns in data. Switch back to design perspective, click on the ***Decision Tree*** operator, and in the ***Parameters*** area, change the '***criterion***' parameter to '***gini_index'***, as shown in Figure 11.

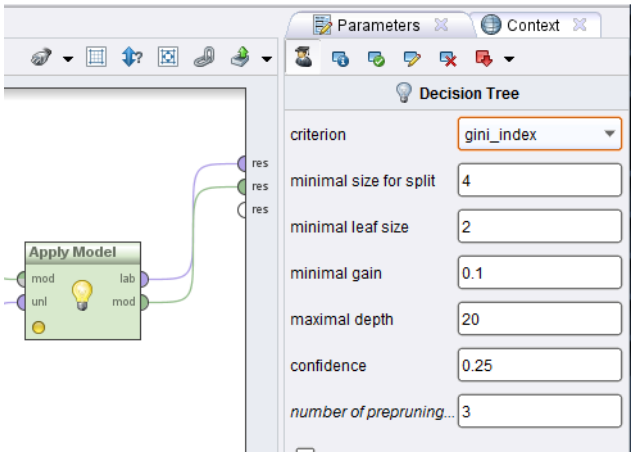


Fig.11. Constructing our decision tree model using the ***gini_index*** algorithm rather than the ***gain_ratio*** algorithm

Now, re-run the model and we will move on to evaluation. (Fig.12). We see in this tree that there is much more detail, more granularity in using the Gini algorithm as our parameter for our decision tree. We could further modify the tree by going back to design view and changing the minimum number of items to form a node (size for split) or the minimum size for a leaf. Even accepting the defaults for those parameters though, we can see that the Gini algorithm alone is much more sensitive than is the Gain Ratio algorithm in identifying nodes and leaves. Take a minute to explore around this new tree model. You will find that it is extensive, and that you will to use both the ***Zoom*** and ***Mode*** tools to see it all. You should find that most of our other independent variables (predictor attributes) are now being used, and the granularity with which we can identify each customer's likely adoption category is much greater. How active the person is on firm's employer's web site is still the single best predictor, but gender, and multiple levels of age have now also come into play. You will also find that a single attribute is sometimes used more than once in a

single branch of the tree. Decision trees are a lot of fun to experiment with, and with a sensitive algorithm like Gini generating them, they can be tremendously interesting as well.

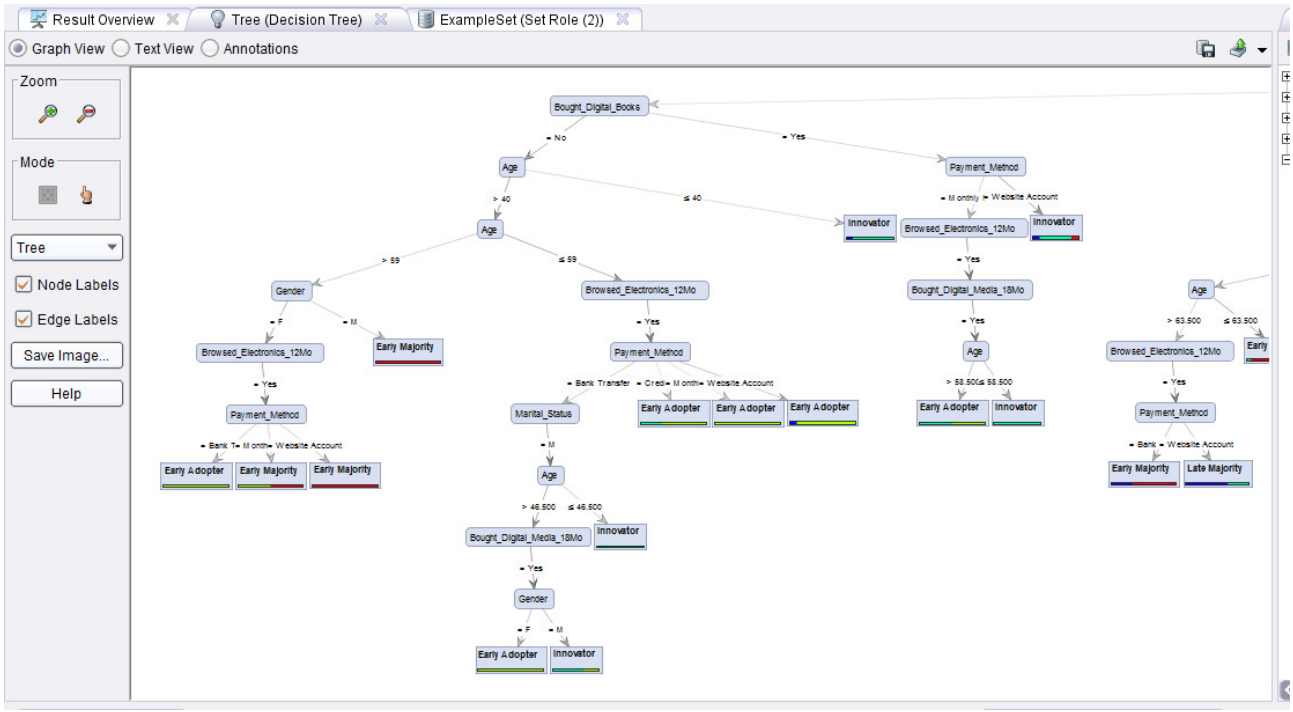
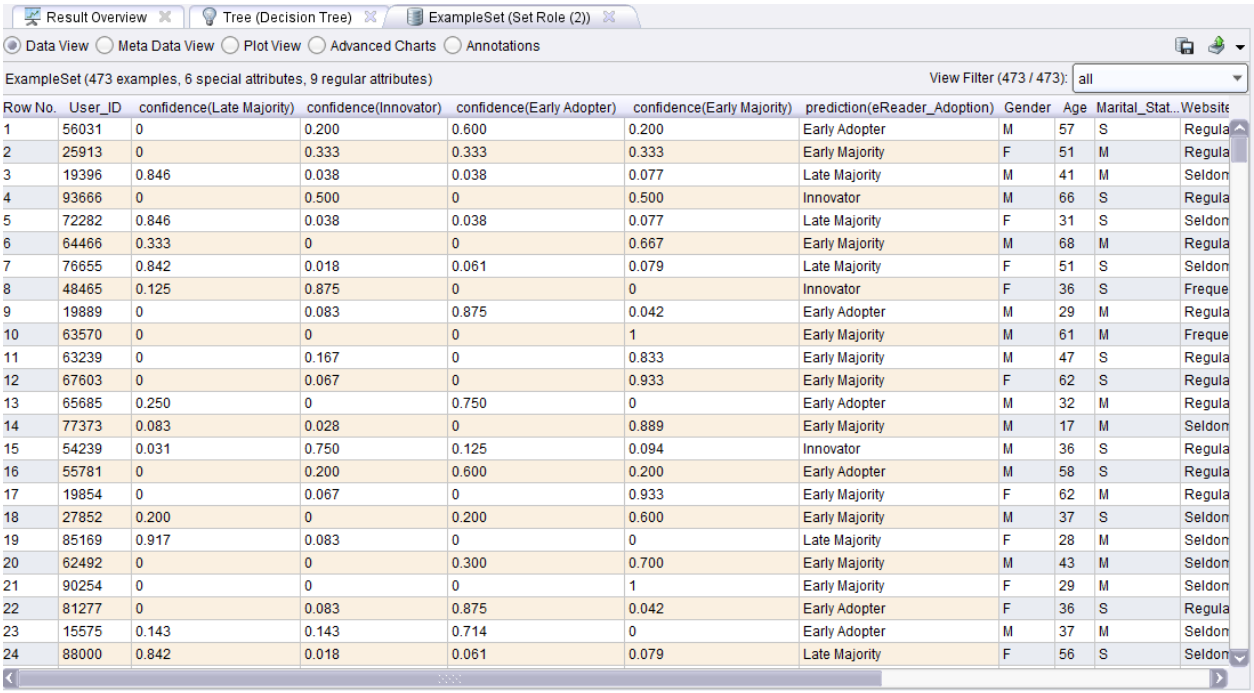


Fig.12. Tree resulting from a ***gini_index*** algorithm

Switch to the ***ExampleSet*** tab in ***Data View***. We see here (Figure 13) that changing our tree's underlying algorithm has, in some cases, also changed our confidence in the prediction.



Result Overview | Tree (Decision Tree) | ExampleSet (Set Role (2))

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (473 examples, 6 special attributes, 9 regular attributes) — View Filter (473 / 473): all

Row No.	User_ID	confidence(Late Majority)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)	prediction(eReader_Adoption)	Gender	Age	Marital_Stat...	Website
1	56031	0	0.200	0.600	0.200	Early Adopter	M	57	S	Regula
2	25913	0	0.333	0.333	0.333	Early Majority	F	51	M	Regula
3	19396	0.846	0.038	0.038	0.077	Late Majority	M	41	M	Seldon
4	93666	0	0.500	0	0.500	Innovator	M	66	S	Regula
5	72282	0.846	0.038	0.038	0.077	Late Majority	F	31	S	Seldon
6	64466	0.333	0	0	0.667	Early Majority	M	68	M	Regula
7	76655	0.842	0.018	0.061	0.079	Late Majority	F	51	S	Seldon
8	48465	0.125	0.875	0	0	Innovator	F	36	S	Freque
9	19889	0	0.083	0.875	0.042	Early Adopter	M	29	M	Regula
10	63570	0	0	0	1	Early Majority	M	61	M	Freque
11	63239	0	0.167	0	0.833	Early Majority	M	47	S	Regula
12	67603	0	0.067	0	0.933	Early Majority	F	62	S	Regula
13	65685	0.250	0	0.750	0	Early Adopter	M	32	M	Regula
14	77373	0.083	0.028	0	0.889	Early Majority	M	17	M	Seldon
15	54239	0.031	0.750	0.125	0.094	Innovator	M	36	S	Regula
16	55781	0	0.200	0.600	0.200	Early Adopter	M	58	S	Regula
17	19854	0	0.067	0	0.933	Early Majority	F	62	M	Regula
18	27852	0.200	0	0.200	0.600	Early Majority	M	37	S	Seldon
19	85169	0.917	0.083	0	0	Late Majority	F	28	M	Seldon
20	62492	0	0	0.300	0.700	Early Majority	M	43	M	Seldon
21	90254	0	0	0	1	Early Majority	F	29	M	Seldon
22	81277	0	0.083	0.875	0.042	Early Adopter	F	36	S	Regula
23	15575	0.143	0.143	0.714	0	Early Adopter	M	37	M	Seldon
24	88000	0.842	0.018	0.061	0.079	Late Majority	F	56	S	Seldon

Fig.13. New predictions and confidence percentages using Gini

Let's take the person on Row 1 (ID 56031) as an example. In Figure 10, this person was calculated as having at least some percentage chance of landing in any one of the four adopter categories. Under the Gain Ratio algorithm, we were 41% sure he'd be an early adopter, but almost 32% sure he might also turn out to be an innovator. In other words, we feel confident he'll buy the eReader early on, but we're not sure how early. Firm will have to decide during the deployment phase. But perhaps using Gini, we can help him decide. In Figure 13, this same man is now shown to have a 60% chance of being an early adopter and only a 20% chance of being an innovator. The odds of him becoming part of the late majority crowd under the Gini model have dropped to zero. We know he will adopt (or at least we are *predicting* with 100% confidence that he will adopt), and that he will adopt early. While he may not be at the top of firm's list when deployment rolls around, he'll probably be higher than he otherwise would have been under gain_ratio. Note that while Gini has changed some of our predictions, it hasn't affected all of them. Re-check person ID 77373 briefly. There is no difference in this person's predictions under either algorithm – RapidMiner is quite certain in its predictions for this young man. Sometimes the level of confidence in a prediction through a decision tree is so high that a more sensitive underlying algorithm won't alter an observation's prediction values at all.

Tasks

The data repositories are on a web site ***http://archive.ics.uci.edu/ml/datasets***. To choose one of data sets (see the Classification column) and download it. It mainly files of *txt* or *csv* format. To check, does a table from the data file have the column titles. If no, to add the names of attributes (list them is on a web page) as first line and save a file. To add a file to repository of RapidMiner. Farther to build a decision tree, choosing one of attributes as label. To build a tree for two criteria: ***gain_ratio*** and ***gini_index***.

PRACTICAL WORK #5
Text Mining

After completing the reading and exercises in this work, you should be able to:

- Explain what text mining is, how it is used and the benefits of using it.
- Recognize the various formats that text can be in, in order to perform text mining.
- Connect to and import text as a data source for a text mining model.
- Develop a text mining model in RapidMiner including common text-parsing operators such as **tokenization**, **stop word** filtering, **n-gram** construction, **stemming**, etc.
- Apply other data mining models to text mining results in order to predict or classify based on textual analysis.

The text mining module of RapidMiner is an optional add-in. When you installed RapidMiner, we mentioned that you might want to include the ***Text Processing*** component. Whether you did or did not at that time, we will need it for this work, so we can add it now. Even if you did add it earlier, it might be a good idea to complete all of the steps below to ensure your ***Text Processing*** add-in is up-to-date.

1. Open RapidMiner to a new, blank process. From the application menu, select ***Help*** > Update and Extensions (Marketplace)…

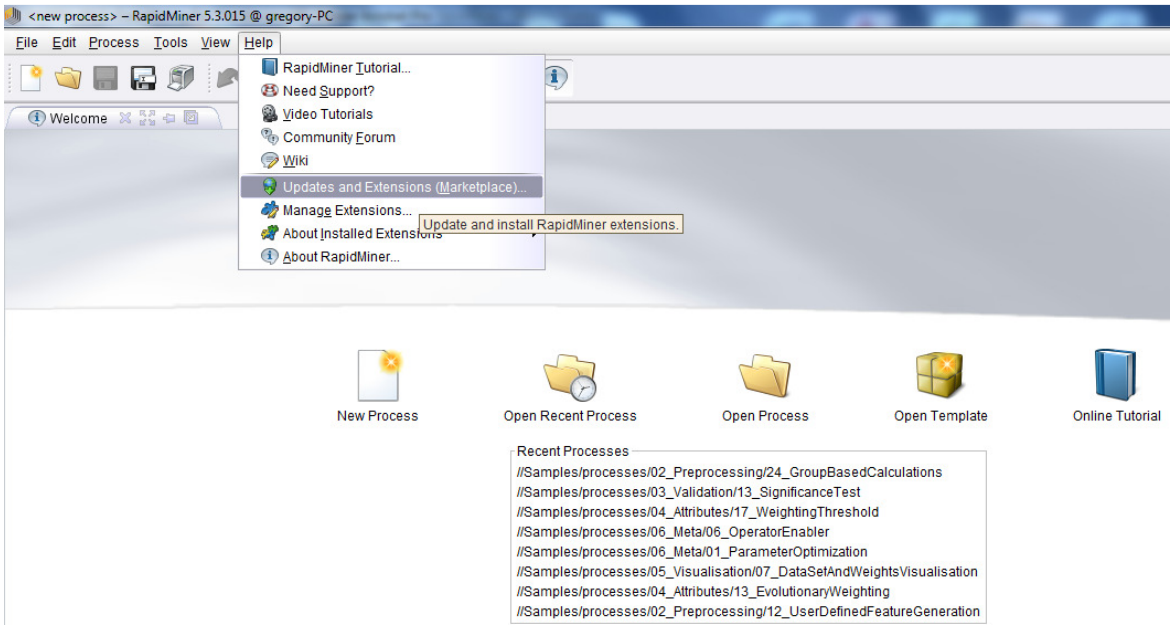


Fig.1. Updating RapidMiner add-ins

2. Your computer will need to be connected to the Internet, so that it can check Rapidminer's servers to see if any updates are available. Once the connection has

been established and the software has checked for available updates, you will see a window similar to Figure 2. Locate ***Text Processing*** in the list. If it is grayed out, that means that the add-in is installed and up-to-date on your computer. If it is not installed, or not up to the current version, it will be orange. You can double click the small square to the left of the ***Text Processing*** icon (the circle with 'ABC' in it). Then click the ***Install*** button to add or update the module. When it is finished, the window will disappear and you will be back to your main RapidMiner window.

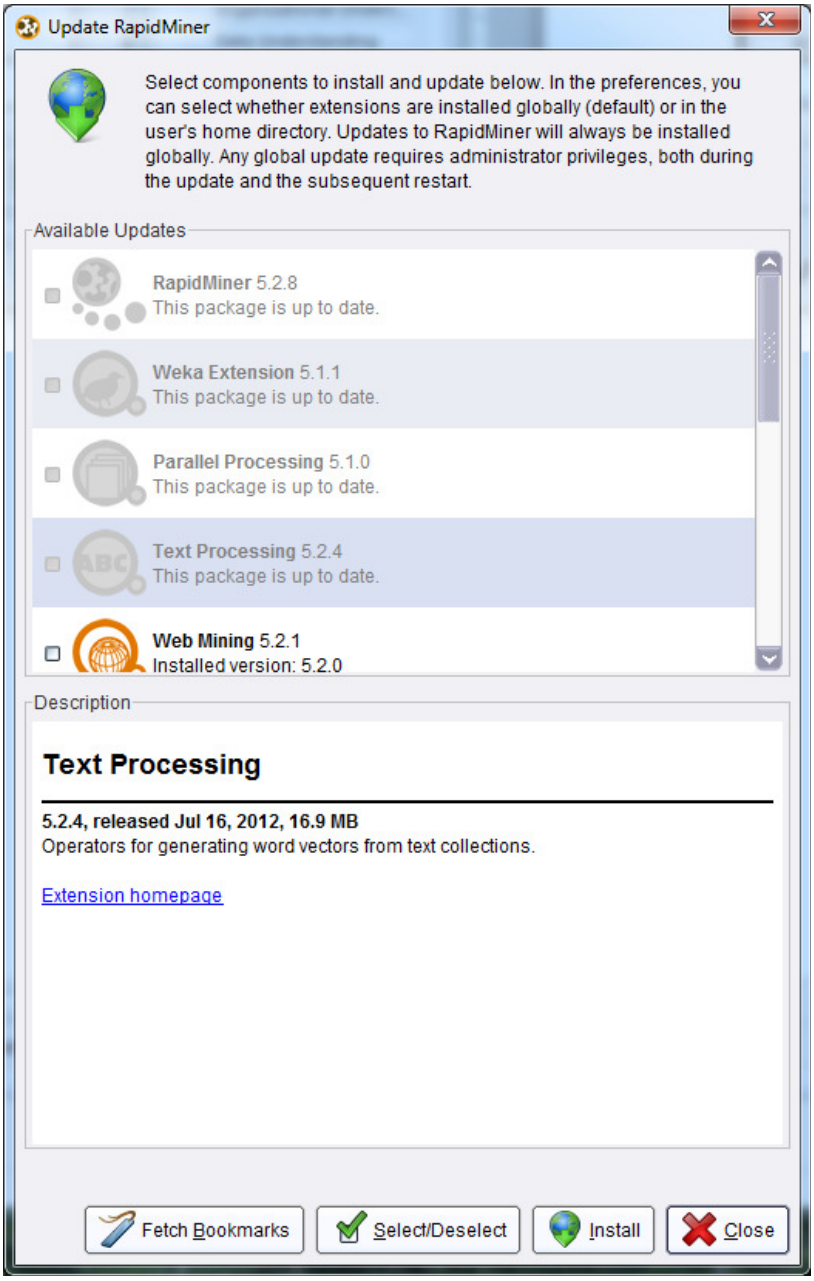


Fig.2. Adding/updating the RapidMiner ***Text Processing*** add-in

3. In the ***Operators*** tab in the lower left hand area of your RapidMiner window, locate and expand the ***Text Processing*** operators folder by clicking on the **+** sign next to it.

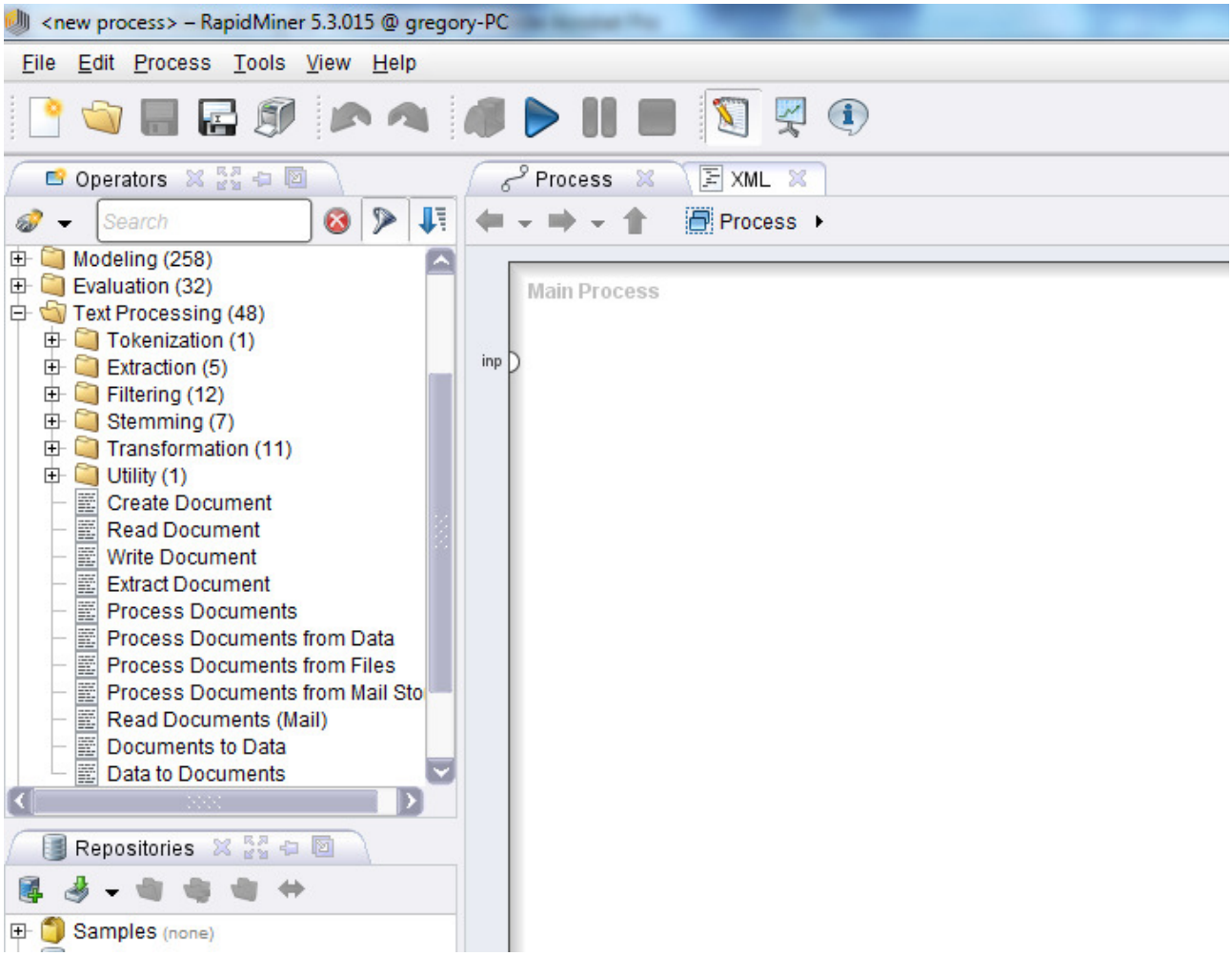


Fig.3. Finding tools in the ***Text Processing*** operator area

4. Download your document (for example, *The Entire Works of Mark Twain*) through Project Gutenberg's website as a text document. Save the document in a file on your computer.

5. Open RapidMiner and click ***New Process***. On the left hand pane of your screen, there should be the ***Operators*** tab – this is where you can search and find all of the operators for RapidMiner and its extensions. By searching the ***Operators*** tab for "read", you should get an output like Figure 4:

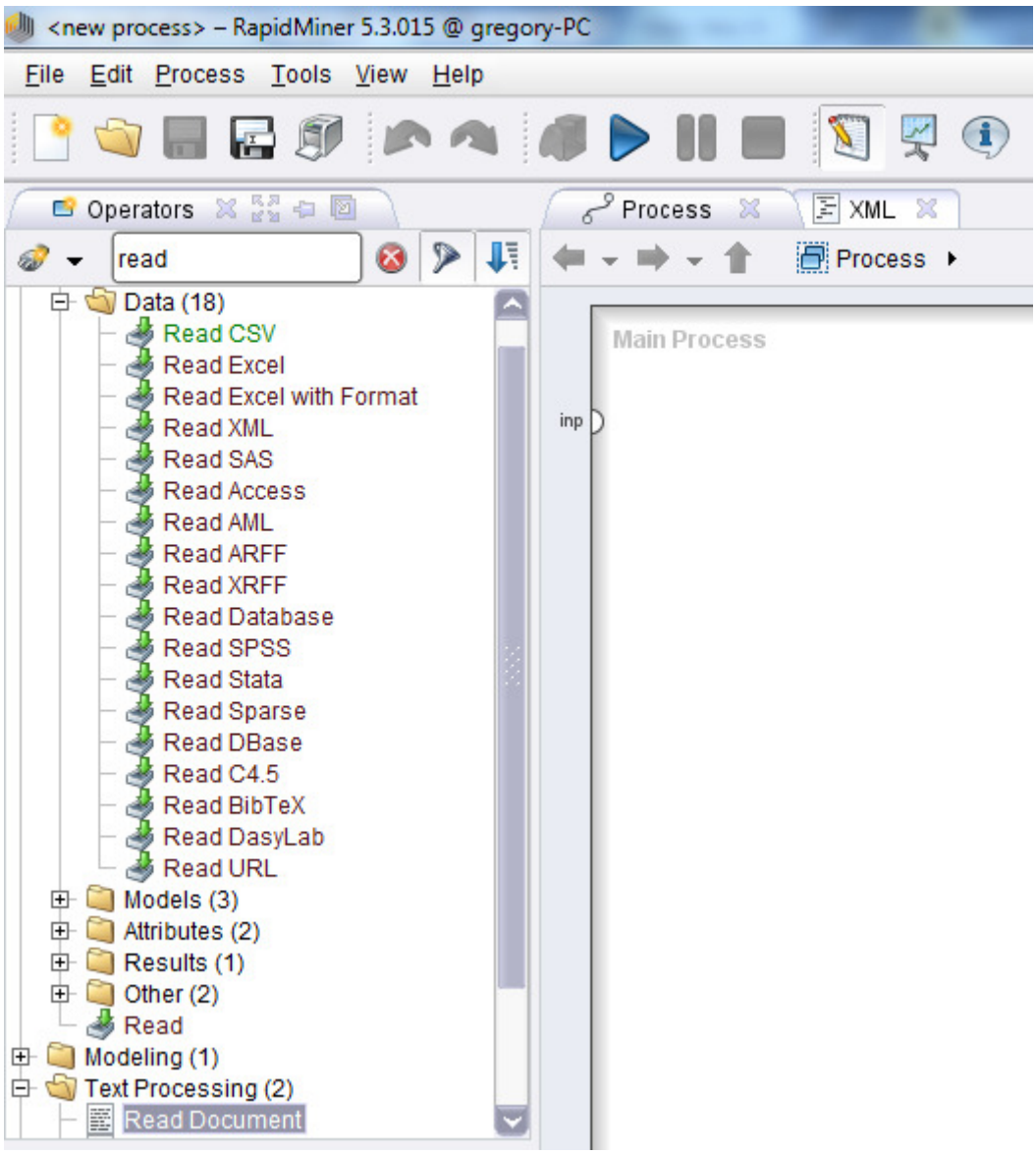


Fig.4. The read operators in the ***Operators*** tab

There are multiple read operators depending on which file you have, and most of them work the same way. If you scroll down, there is a ***Read Documents*** operator. Select this operator and enter it into your ***Main Process*** window by dragging it. When you select the ***Read Documents*** operator in the ***Main Process*** window, you should see a ***file*** uploader in the right-hand pane (Figure 5).

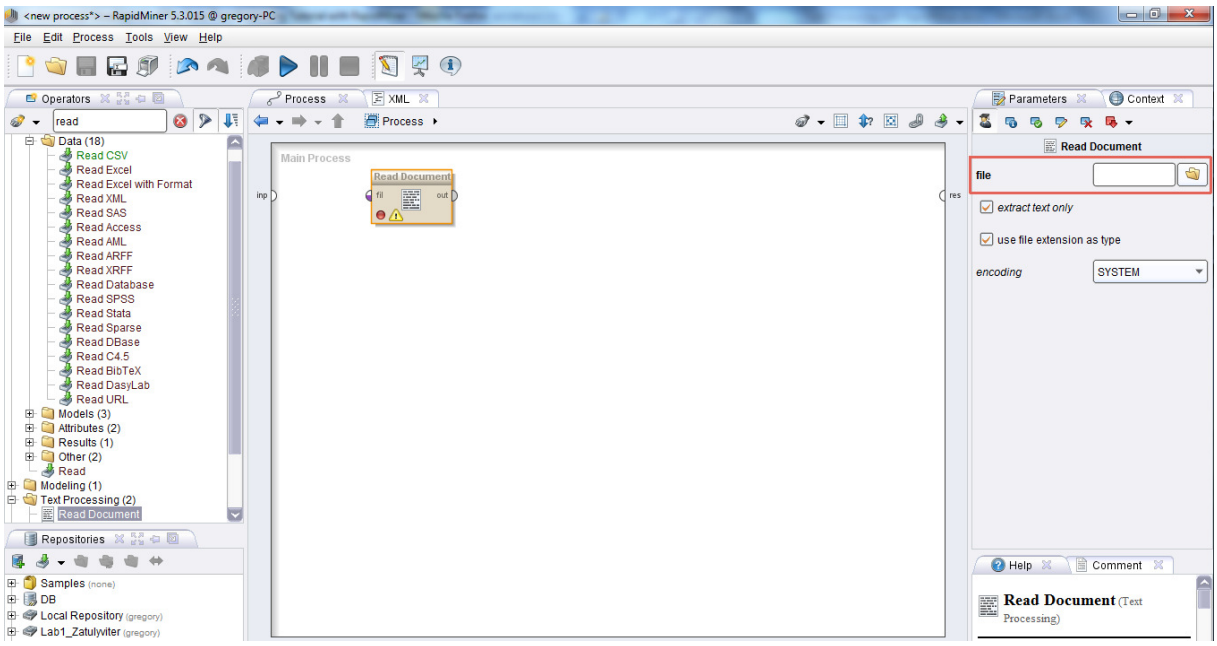


Fig.5. The ***Read Documents*** operator in the ***Main Process*** window

Select the text file you want to use.

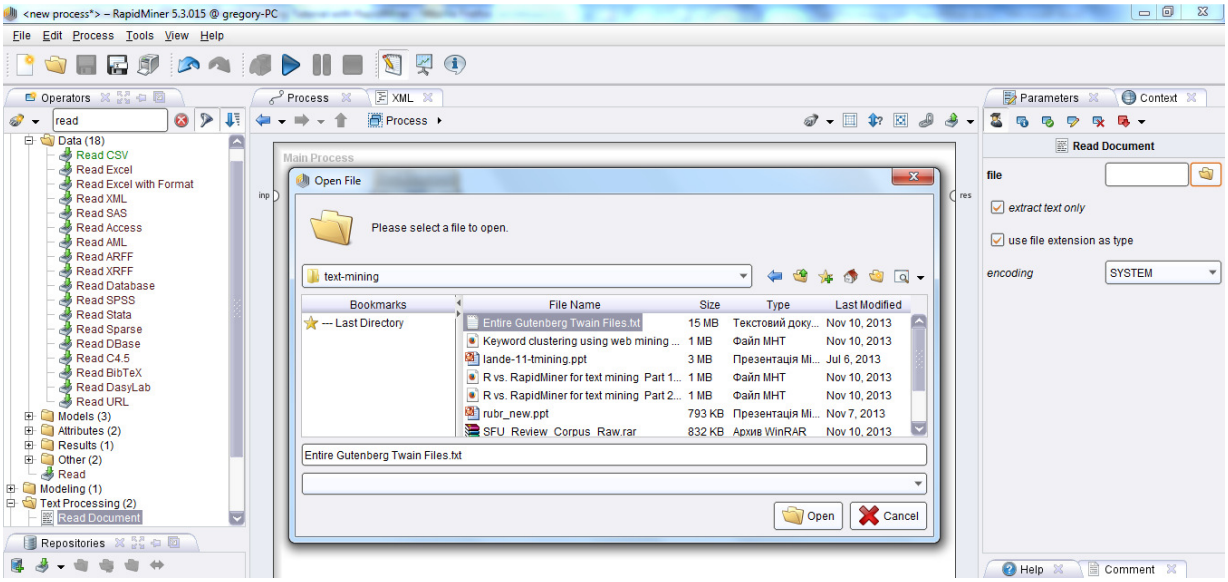


Fig.6. Select the desired text file

6. After you have chosen your file, make sure that the output port on the ***Read Documents*** operator is connected to the ***res*** node in your ***Main Process***. Click the ***play*** button to check that your file has been received correctly. Switch to the results perspective by clicking the icon that looks like a display chart above the ***Process*** tab at the top of the ***Main Process*** pane. Click the ***Document (Read Document)*** tab. Your output text should look something like this depending on the file you have chosen to process:

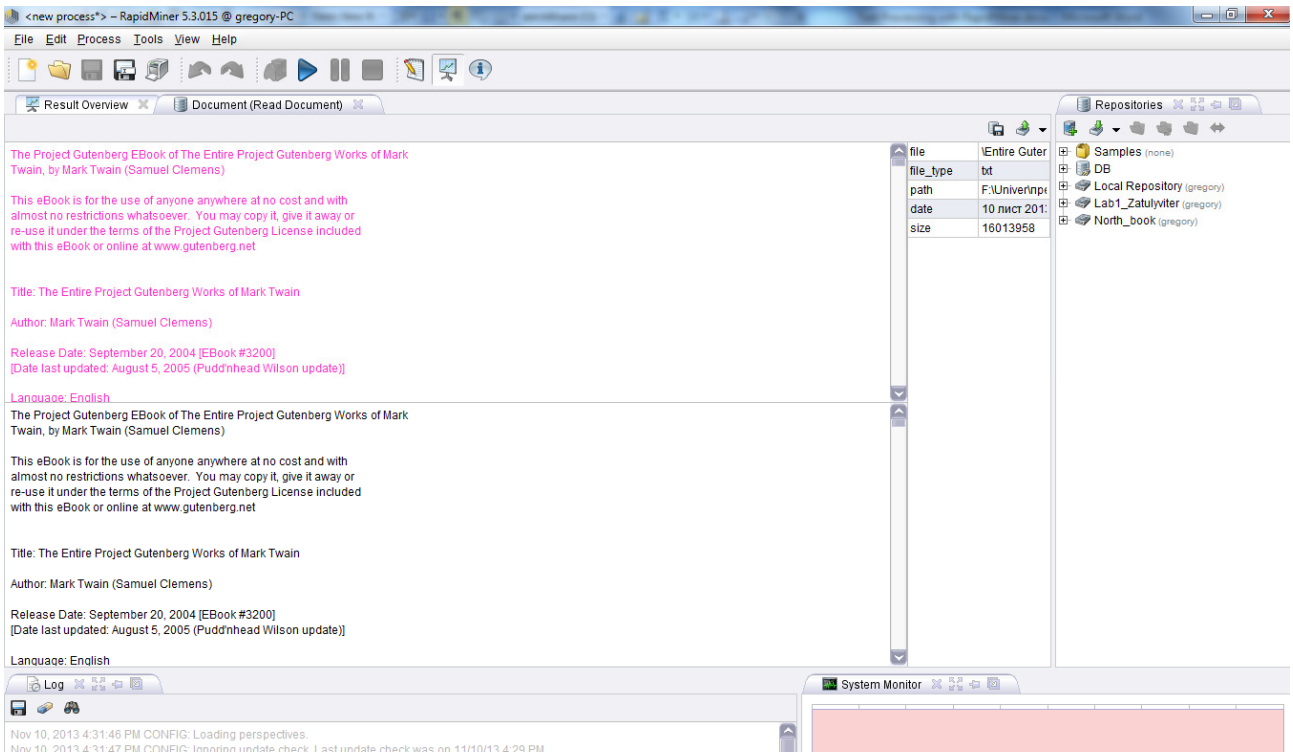


Fig.7. ***Document (Read Document)*** tab

7. Now we will move on to processing the document to get a list of its different words and their individual count. Search the ***Operators*** list for ***Process Documents***. Drag this operator the same way as you did for the ***Read Documents*** operator into the ***Main Process*** pane.

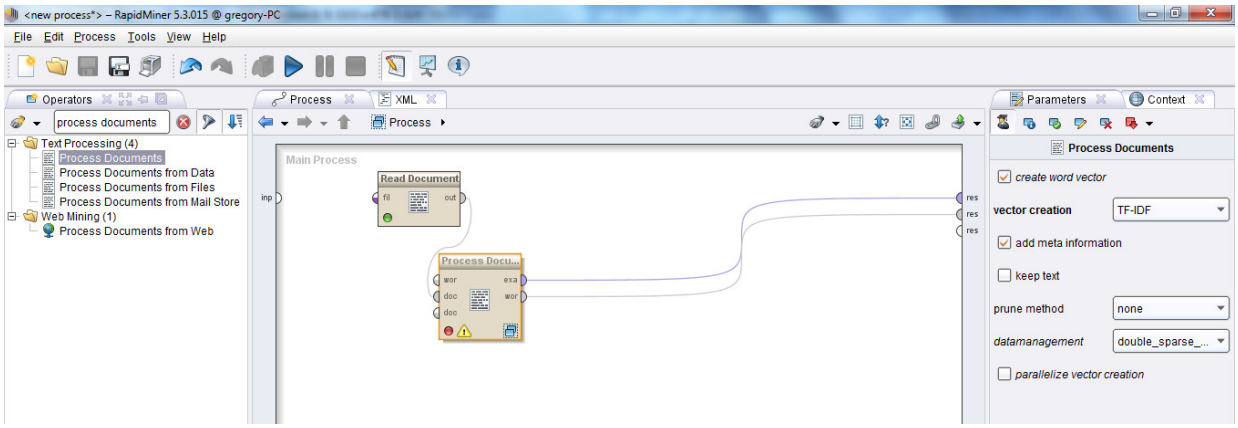


Fig.8. Added ***Read Documents*** operator into the ***Main Process*** pane

Double click the ***Process Documents*** operator to get inside the operator. This is where we will link operators together to take the entire text document and split it down into its word components. This consists of several operators that can be chosen by going into the ***Operator*** pane and looking at the ***Text Processing*** folder. You should see several more folders such as ***Tokenization***, ***Extraction***, ***Filtering***,

Stemming, ***Transformation***, and ***Utility***. These are some of the descriptions of what you can do to your document. The first thing that you would want to do to your document is to tokenize it. Tokenization creates a "bag of words" that are contained in your document. This allows you to do further filtering on your document. Search for the ***Tokenize*** operator and drag it into the ***Process Documents*** process.

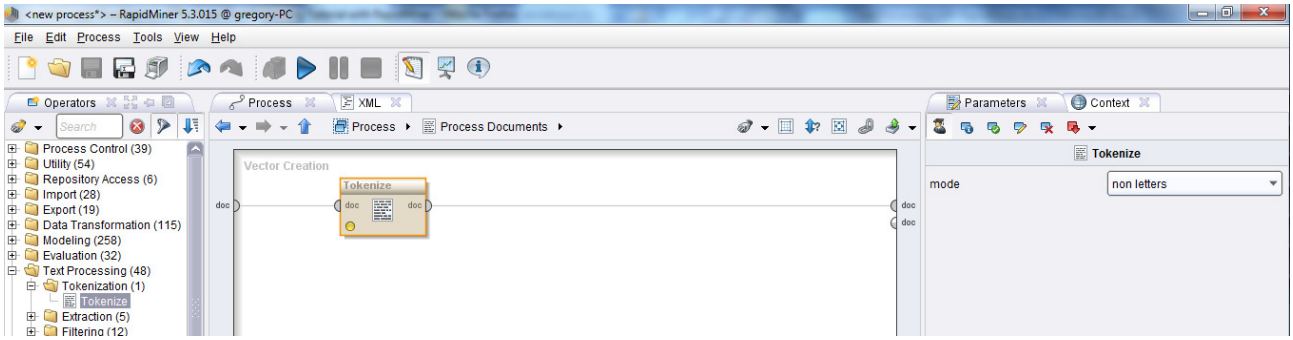


Fig.9. The ***Tokenize*** operator

Connect the ***doc*** node of the process to the ***doc*** input node of the operator if it has not automatically connected already. Now we are ready to filter the bag of words. In ***Filtering*** folder under the ***Text Processing*** operator folder, you can see the various filtering methods that you can apply to your process. For this example, we want to filter certain words out of our document that don't really have any meaning to the document itself (such as the words *a*, *and*, *the*, *as*, *of*, etc.); therefore, we will drag the ***Filter Stopwords (English)*** into my process because our document is in English. Also, we want to filter out any remaining words that are less than three characters. Select ***Filter Tokens by Length*** and set your parameters as desired (in this case, we want our min number of characters to be 3, and our max number of characters to be an arbitrarily large number since we don't care about an upper bound). Connect the nodes of each subsequent operator accordingly as in Figure 10.

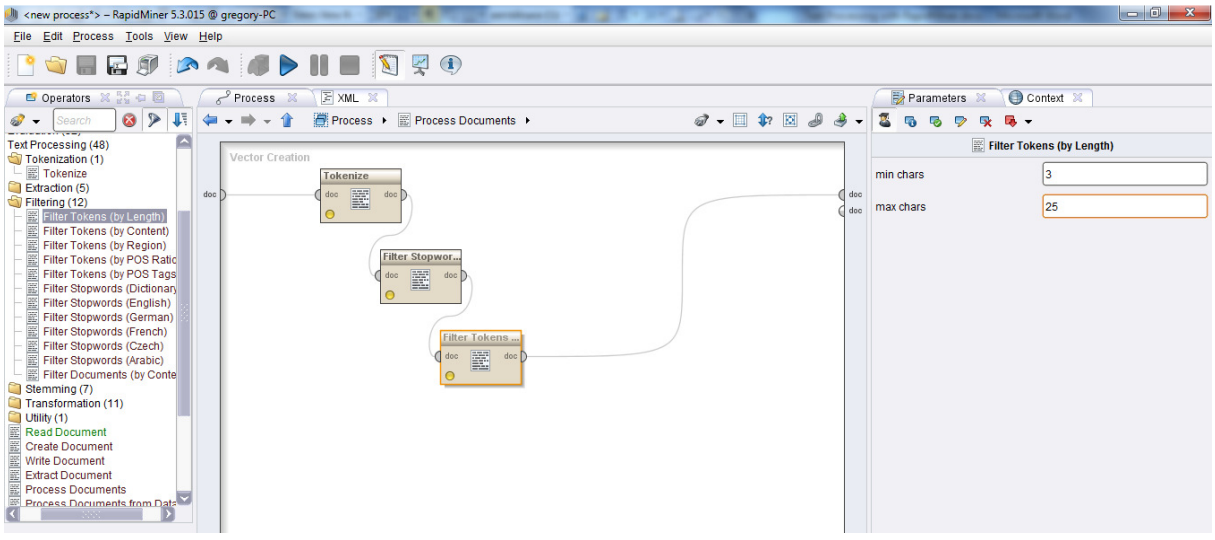


Fig.11. Added ***Filter Stopwords (English)*** and ***Filter Tokens by Length*** operators

After we filtered the bag of words by stopwords and length, we want to transform all of our words to lowercase since the same word would be counted differently if it was in uppercase vs. lowercase. Select the operator ***Transform Cases*** and drag it into the process.

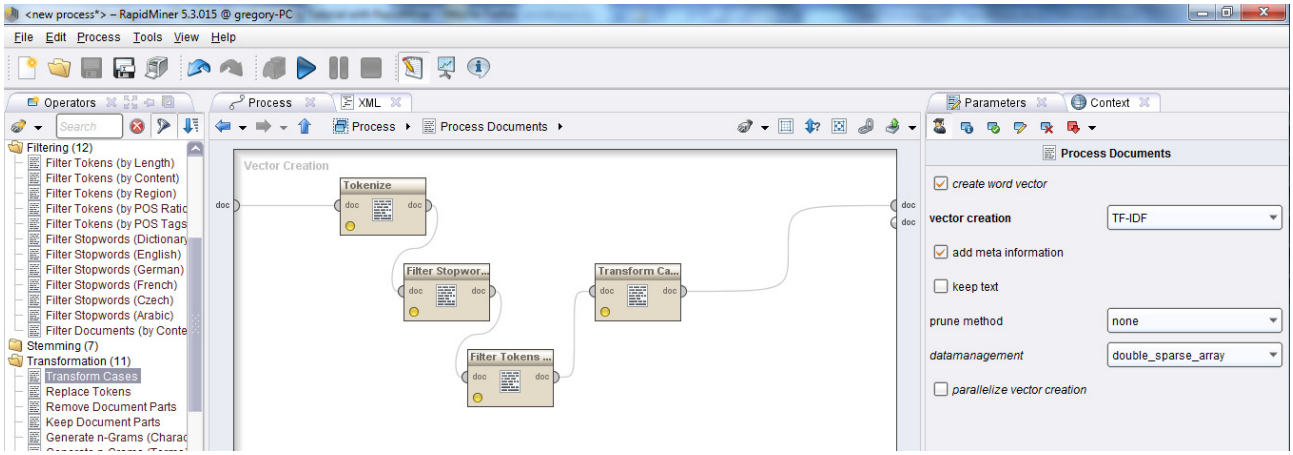
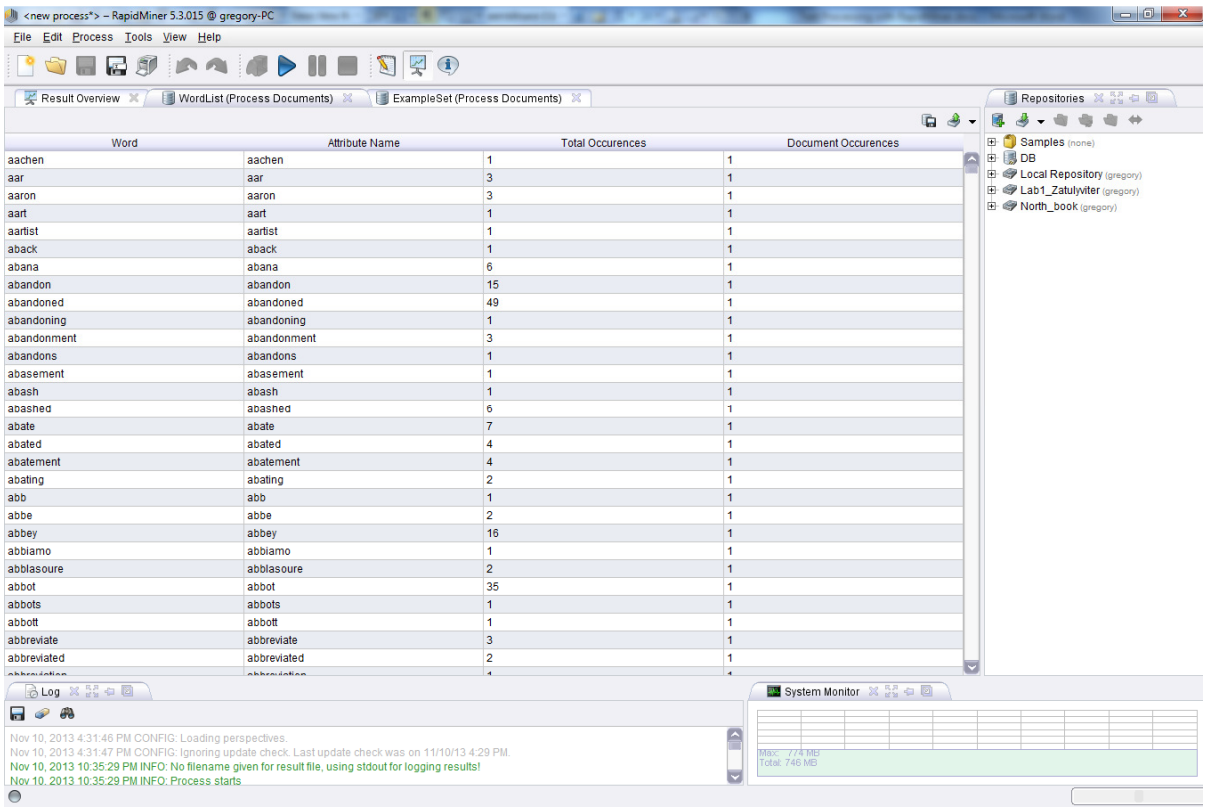


Fig.12. Added the ***Transform Cases*** operator

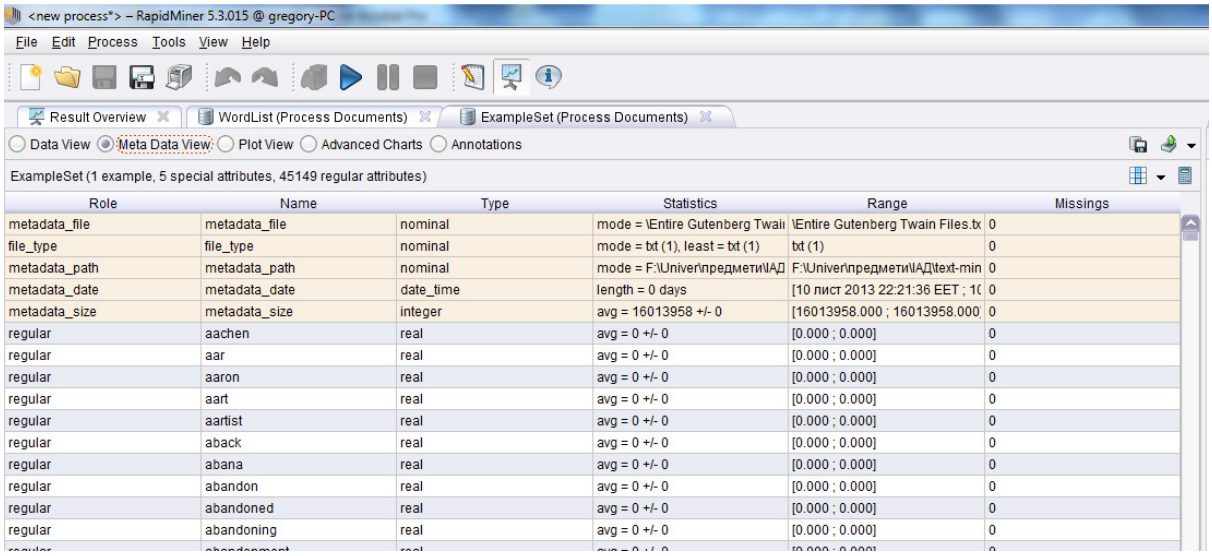
8. Now that we have the sufficient operators in our process for this example, we check all of our node connections and click the ***Play*** button to run our process. If all goes well, your output should look like this in Figure 13:



Word	Attribute Name	Total Occurences	Document Occurences
aachen	aachen	1	1
aar	aar	3	1
aaron	aaron	3	1
aart	aart	1	1
aartist	aartist	1	1
aback	aback	1	1
abana	abana	6	1
abandon	abandon	15	1
abandoned	abandoned	49	1
abandoning	abandoning	1	1
abandonment	abandonment	3	1
abandons	abandons	1	1
abasement	abasement	1	1
abash	abash	1	1
abashed	abashed	6	1
abate	abate	7	1
abated	abated	4	1
abatement	abatement	4	1
abating	abating	2	1
abb	abb	1	1
abbe	abbe	2	1
abbey	abbey	16	1
abbiamo	abbiamo	1	1
abblasoure	abblasoure	2	1
abbot	abbot	35	1
abbots	abbots	1	1
abbott	abbott	1	1
abbreviate	abbreviate	3	1
abbreviated	abbreviated	2	1

Fig.13. The result of tokenization

If you switch the ***ExampleSet (Process Documents)*** tab and choose the ***Meta Data View*** button, we see a few new special attributes, created by RapidMiner (Figure 14).



Role	Name	Type	Statistics	Range	Missings
metadata_file	metadata_file	nominal	mode = \Entire Gutenberg Twain	\Entire Gutenberg Twain Files.tx	0
file_type	file_type	nominal	mode = txt (1), least = txt (1)	txt (1)	0
metadata_path	metadata_path	nominal	mode = F:\Univer\предмети\АД	F:\Univer\предмети\АД\text-min	0
metadata_date	metadata_date	date_time	length = 0 days	[10 лист 2013 22:21:36 EET ; 10	0
metadata_size	metadata_size	integer	avg = 16013958 +/- 0	[16013958.000 ; 16013958.000	0
regular	aachen	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	aar	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	aaron	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	aart	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	aartist	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	aback	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	abana	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	abandon	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	abandoned	real	avg = 0 +/- 0	[0.000 ; 0.000]	0
regular	abandoning	real	avg = 0 +/- 0	[0.000 ; 0.000]	0

Fig.14. A view of the words from our input documents as tokens (attributes)

You are now able to see a word list containing all the different words in your document and their occurrence count next to it in the ***Total Occurences*** column. If you do not get this output, make sure that all of your nodes are connected correctly and also to the right *type*. Some errors are because your output at one node does not match the type expected at the input of the next node of an operator.

At this point, we have a model that is capable of mining and displaying to us the words that are most frequent in our text documents. This will be interesting for us to review, but there are a few more operators that you should know about in addition to the ones we are using here. These are highlighted by arrows in Figure 15, and discussed below.

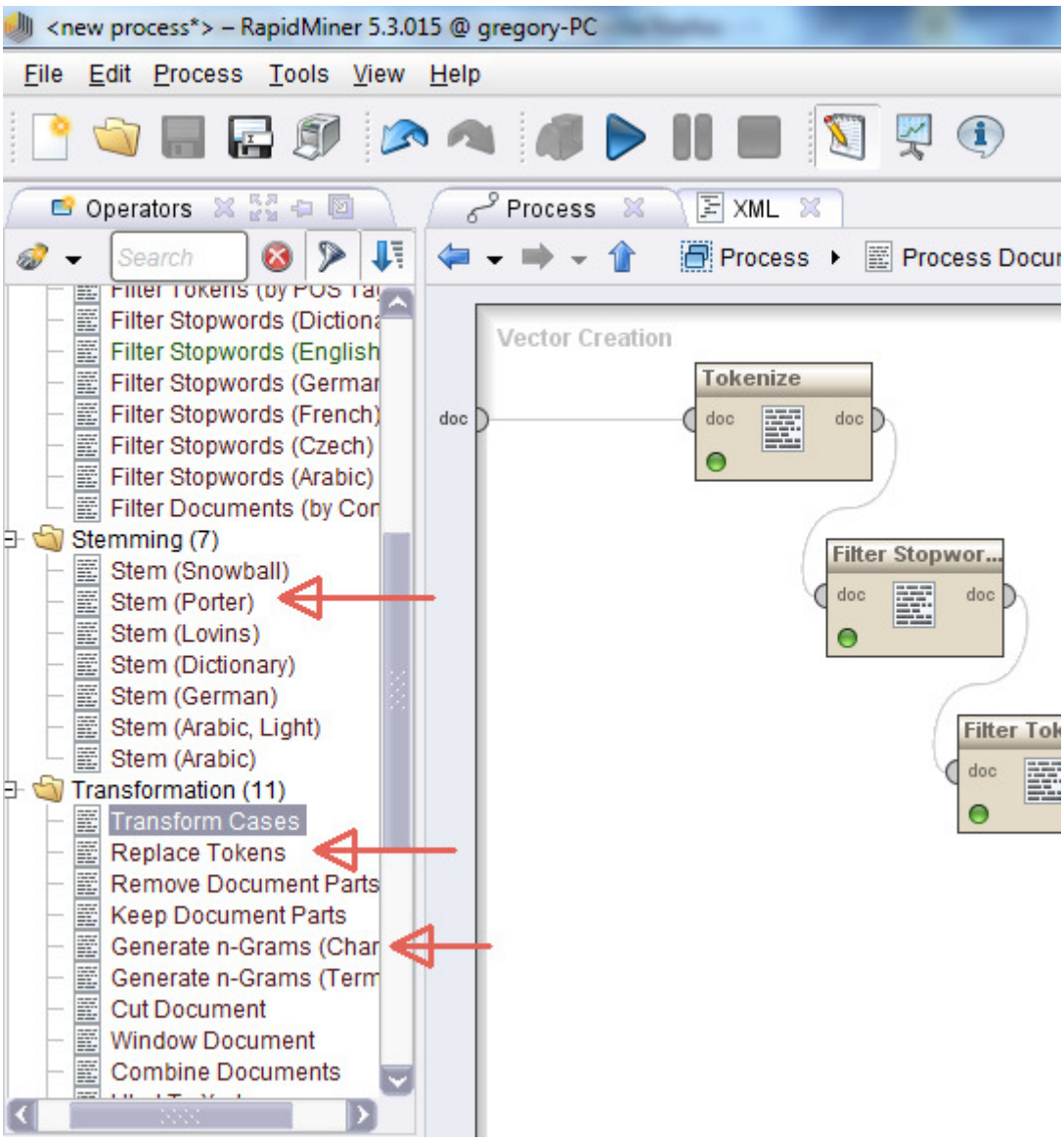


Fig.15. Additional text mining operators of interest

❖ **Stemming**: In text mining, *stemming* means finding terms that share a common root and combining them to mean essentially the same thing. For example, 'America', 'American', 'Americans', are all like terms and effectively refer to the same thing. By stemming (you can see there are a number of stemming operators using different algorithms for you to choose from), RapidMiner can reduce all instances of these word variations to a common form, such as 'Americ', or perhaps 'America', and have all instances represented in a single attribute.

❖ **Generate n-Grams**: In text mining, an *n-gram* is a phrase or combination of words that may take on meaning that is different from, or greater than the meaning of each word individually. When creating n-grams, the *n* is simply the maximum number of terms you want RapidMiner to consider grouping together. Take for example the token 'death'. This word by itself is strong, evoking strong emotion. But now consider the meaning, strength and emotion if you were to add a ***Generate n-Grams*** operator to your model with a size of 2 (this is set in the

parameters area of the n-gram operator). Depending on your input text, you might find the token 'death_penalty'. This certainly has a more specific meaning and evokes different and even stronger emotions than just the token 'death'. What if we increased the n-gram size to 3? We might find a token 'death_penalty_execution'. Again, more specific meaning and perhaps stronger emotion is attached. Understand that these example gram tokens would only be created by RapidMiner if the two or three words in each of them were found together, and in close proximity to one another in the input text. Generating grams can be an excellent way to bring a more granular analysis to your text mining activities.

❖ **Replace Tokens**: This is similar to replacing missing or inconsistent values in more structured data. This operator can come in handy once you've tokenized your text input. Suppose for example that you had the tokens 'nation', 'country', and 'homeland' in your data set but you wanted to treat all of them as one token. You could use this operator to change both 'country' and 'homeland' to 'nation', and all instances of any of the three terms (or their stems if you also use stemming) would subsequently be combined into a single token.

These are a just a few of the other operators in the ***Text Processing*** area that can be nice additions to a text mining model. There are many others, and you may experiment with these at your leisure. For now though, we will proceed to modelling.

Click the blue up arrow to move from your sub-process back to your main process window.

Tasks

1. http://www.gutenberg.org/ebooks/100 - The Complete Works of William Shakespeare
2. http://www.gutenberg.org/ebooks/31100 - The Complete Works of Jane Austen
3. http://www.gutenberg.org/ebooks/5400 - The Complete Works of Winston Churchill
4. http://www.gutenberg.org/ebooks/6300 - The Complete Works of Gilbert Parker
5. http://www.gutenberg.org/ebooks/4800 - The Complete Poetical Works of Percy Bysshe Shelley
6. http://www.gutenberg.org/ebooks/1365 - The Complete Poetical Works of Henry Wadsworth Longfellow
7. http://www.gutenberg.org/files/2981/ - The Complete Memoirs of Casanova
8. http://www.gutenberg.org/files/18500/ - The Complete Works of Robert Burns
9. http://www.gutenberg.org/ebooks/6049 - Works of John Bunyan
10. http://www.gutenberg.org/ebooks/3254 - Complete John Galsworthy Works
11. http://www.gutenberg.org/ebooks/5000 - The Notebooks of Leonardo Da Vinci

12. http://www.gutenberg.org/files/9600/ - The Complete Works of John Greenleaf Whittier
13. http://www.gutenberg.org/ebooks/29090 - The Complete Poetical Works of Samuel Taylor Coleridge
14. http://www.gutenberg.org/ebooks/3252 - The Complete Oliver Wendell Holmes, Sr. Works
15.

REFERENCES

1. North M. Data Mining for the Masses / Matthew North. – Global Text Project, 2012. – 264 p. – http://dl.dropboxusercontent.com/u/31779972/DataMiningForTheMasses.pdf
2. Han J. Data Mining: Concepts and Techniques: 3rd ed. / Jiawei Han, Micheline Kamber, Jian Pei. – NY: Elsevier, 2012. – 740 p.
3. Tan P. Introduction to Data Mining / P. Tan, M. Steinbach, V. Kumar. – Boston: Addison Wesley, 2006. – 769 p.
4. RapidMiner Studio Manual. – Rapid-I GmbH, 2014. – 116 p.
5. RapidMiner 8: Operator Reference Manual. – Rapid-I GmbH, 2018. – 988 p.
6. Bramer M. Principles of Data Mining / Max Bramer. – London: Springer, 2013. – 455 p.
7. Cios K. Data Mining: A Knowledge Discovery Approach / Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan. – NY: Springer, 2007. – 660 p.
8. Larose D. T. Data Mining Methods and Models / Daniel T. Larose. – NY: Wiley-Interscience, 2006. – 340 p.
9. Rajaraman A. Mining of Massive Datasets Stanford University / A. Rajaraman, J. Lescovec, J. Ullman. – The Stanford InfoLab, 2013. – http://i.stanford.edu/~ullman/mmds/book.pdf.
10. RapidMiner: Data Mining Use Cases and Business Analytics Applications / Edited by Markus Hofmann, Ralf Klinkenberg. – CRC Press, 2014. – 525 p.
11. Witten I. H. Data Mining: Practical Machine Learning Tools and Techniques: 3rd ed. / Ian H. Witten, Frank Eibe, Mark A. Hall. – NY: Elsevier, 2011. – 665 p.

Signed for publication on November 1, 2019
Format 60x84/16. Offset paper.
Print on duplicator. Deputy No. 6-1984
Cond. print. sheets 2.57. Acc. publ. sheets 2.62.
Amount 100 copies

Printed by Private Entrepreneur Shpak V.B.
Certificate of state registration: B02 No. 924434 dated 12/12/2006.
Certificate of tax payer Series E No. 897220
Ternopol, steet Prosvity 6
tel. 097 299 38 99, 063 300 86 72
Email: tooums@ukr.net