ФГБОУ ВО

Уфимский государственный авиационный технический университет

Кафедра ВМиК

Отчет

по лабораторной работе №2

Тема: «Система интеллектуального анализа данных RapidMiner»

Выполнили: ст. гр. ПРО-411

Букатина В.В. и Кабирова Л.Р.

Проверила: Харисова Э. А.

Уфа 2019

**Цель:** приобрести навыки корреляционного анализа набора данных с помощью системы интеллектуального анализа данных RapidMiner

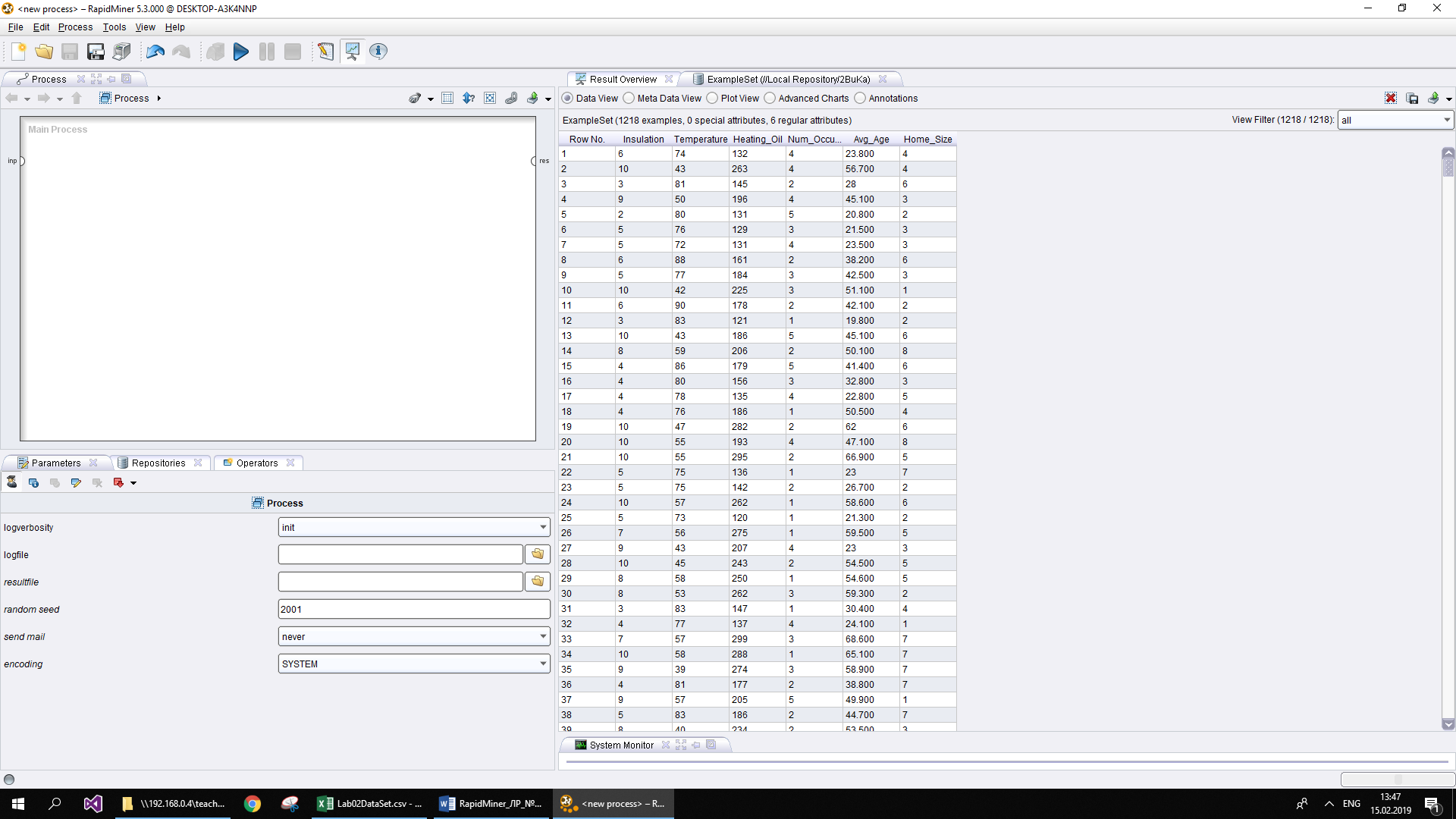
**Задачи:**

1. Выполнить обучающее задание

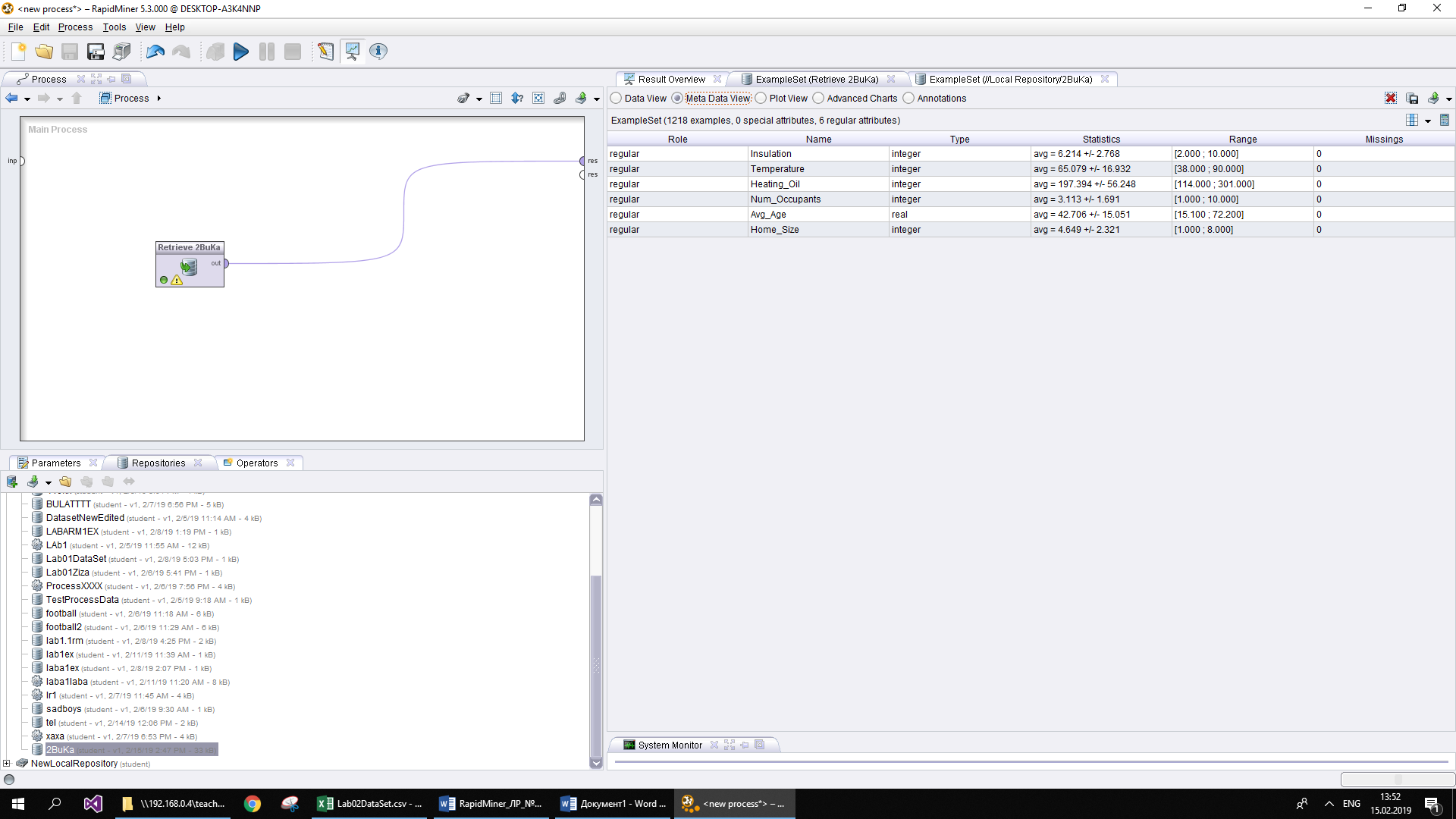
2. Выполнить задание для самостоятельной работы

**1. Обучающее задание**

Создадим новый проект и импортируем туда набор данных для обучающего задания ко второй лабораторной работе, затем добавим его на главный процесс (аналогично лабораторной работе 1).

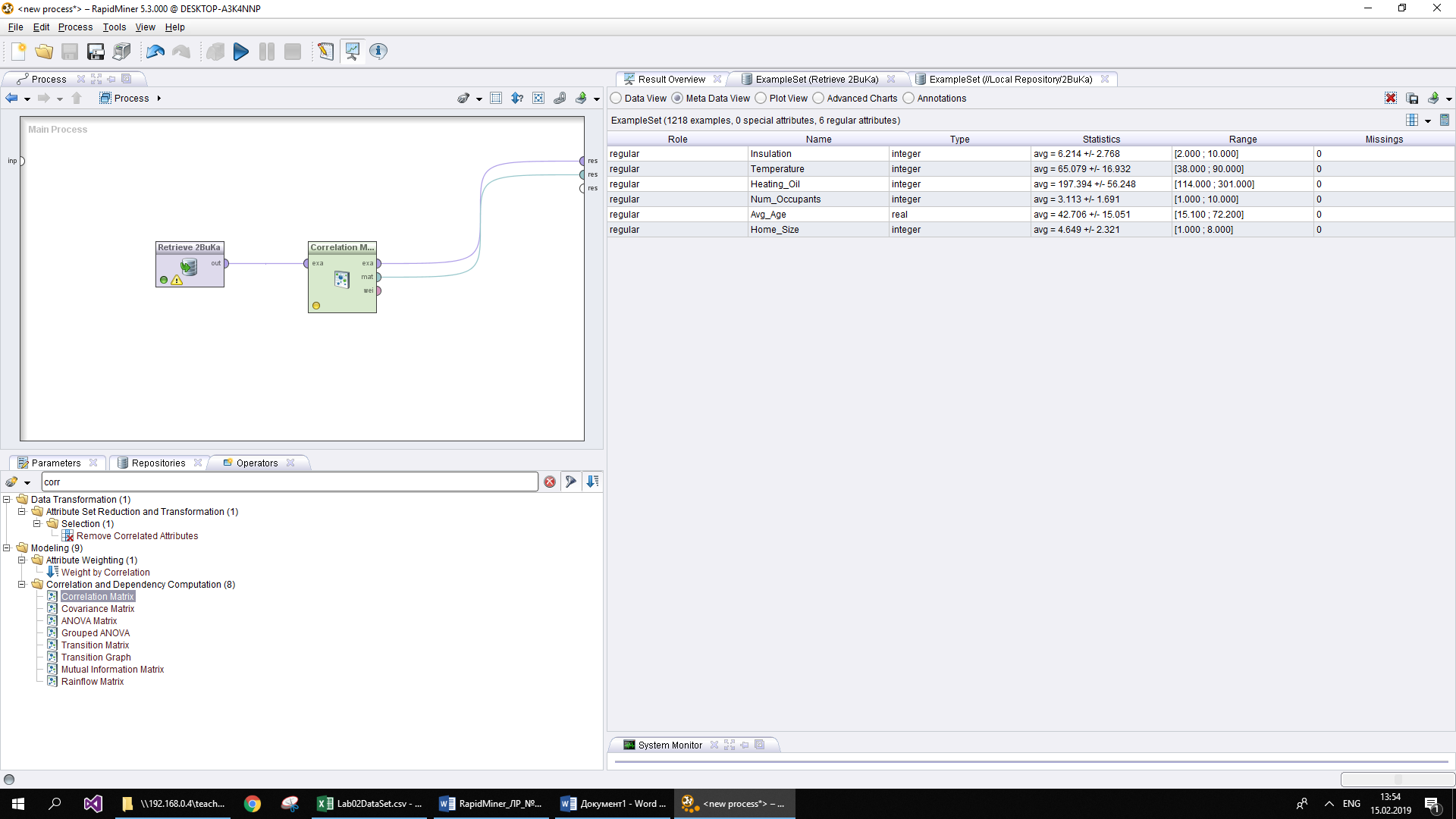
*Рисунок 1 –*  и*мпорт файла Lab02DataSet.csv в программу RapidMiner 5*

Соединим выходной и входной порты exa операторов Retrieve и Correlation Matrix соответственно, затем выходной Correlation Matrix mat к res.

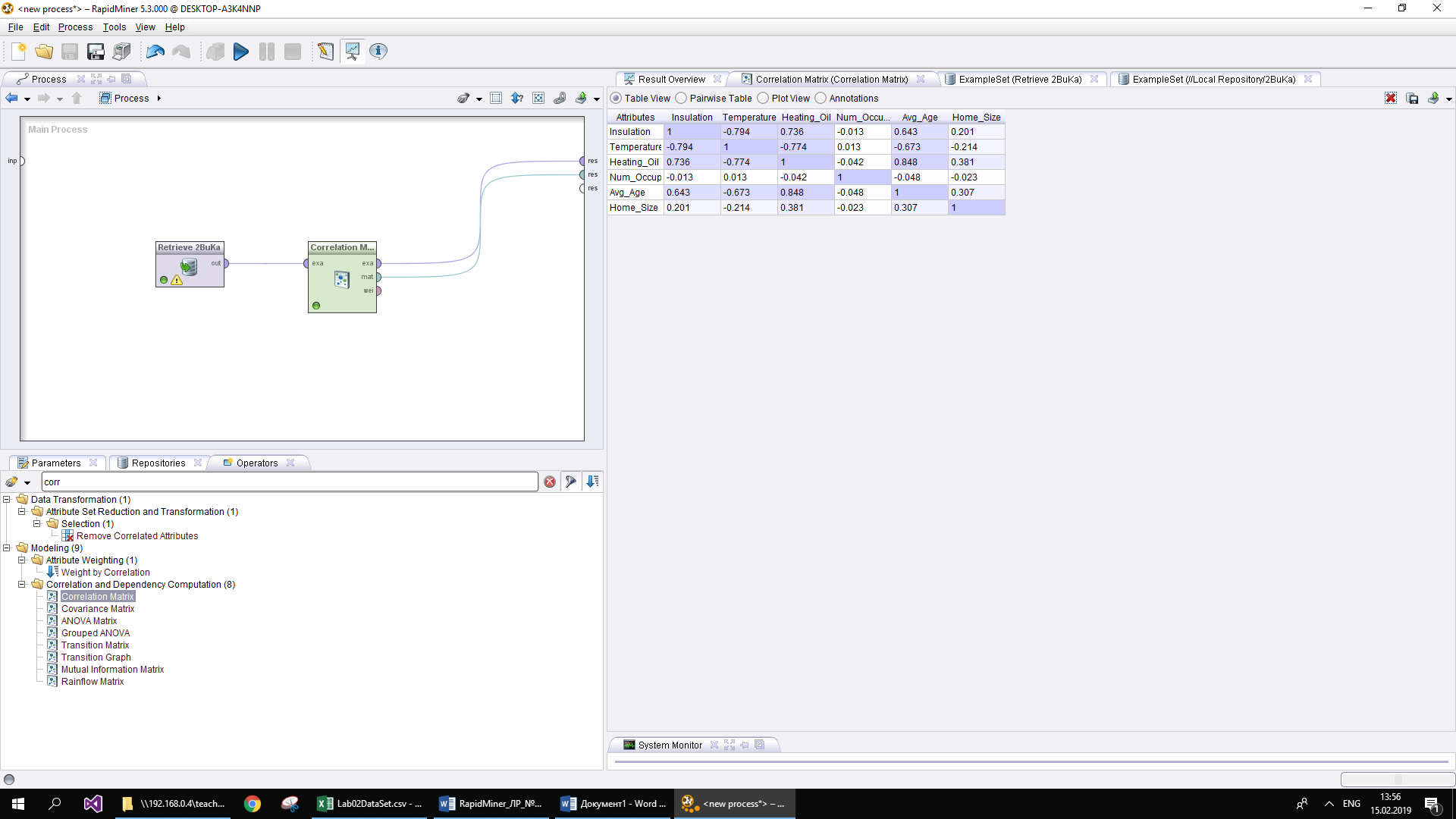


*Рисунок 2* – *Окно результата процесса*

Запустим процесс и *RapidMiner 5 автоматически переключит* рабочую область в Results, где отображается полученная корреляционная матрица.



*Рисунок 3 – Main Process после добавления оператора Correlation Matrix*



*Рисунок 4 – окно Correlation Matrix*

Корреляционная матрица – квадратная матрица *P* размера *n\*n*, где *n* – количество атрибутов. – корреляционная связь между *X*-м и *Y*-м атрибутом. Все диагональные элементы матрицы *P* равны 1.

Коэффициент корреляции рассчитывается по формуле:

где , – среднее значение выборок.

Коэффициент корреляции – показатель статистической связи двух атрибутов, изменяется в пределах от -1 до 1.

Доказательство:

Разделив обе части двойного неравенства на получим

Корреляцию считают:

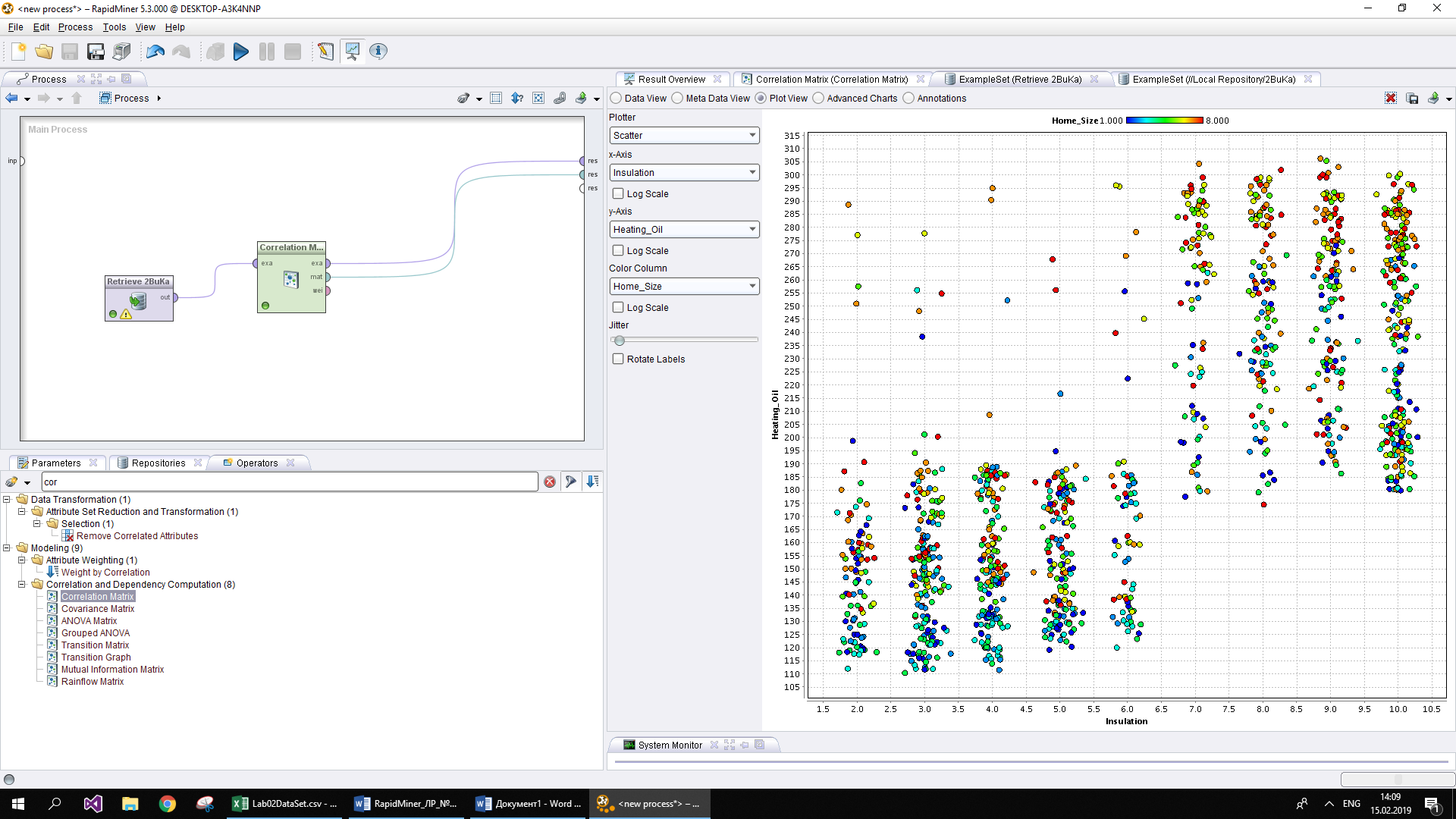
* Функциональной при
* Сильной, если коэффициент корреляции
* Средней при
* Умеренной при
* Слабой при
* Очень слабой при

Если , то корреляционная связь между атрибутами прямая (увеличение значения одного атрибута ведет к увеличению значения другого), при – обратная.

Визуализируем набор данных для подтверждения расчета корреляционной матрицы. Переключаемся во вкладку Plot View и настраиваем параметры.

Таблица 1

|  |  |  |  |
| --- | --- | --- | --- |
| *Первый атрибут* | *Второй атрибут* | *Зависимость* | *Описание* |
| Heating\_Oil | Insulation | Сильная, прямая зависимость  (0.736) | От количества блоков нагревательного материала зависит уровень теплоизоляции домов, чем больше материала – тем выше теплоизоляция |
| Home\_Size | Heating\_Oil | Умеренная, прямая зависимость  (0.381) | От размерности дома зависит количество нагревательных материалов, чем больше дом – тем больше материалов |
| Home\_Size | Insulation | Слабая, прямая зависимость  (0.201) | От размерности дома не значимо зависит уровень теплоизоляции |



*Рисунок 5 – Визуализация наборов данных*

Как видно, график представлен на Рис. 5. Визуализация Heating Oil, Insulation показывает, что увеличение одного признака Insulation приводит к увеличению другого Heating\_Oil (Heating\_Oil -> Insulation = 0.736) т.е. для увеличения теплоизоляции домов нужно увеличивать количество блоков нагревательного материала в течении года на один дом. Подтверждается сильная, прямая связь между атрибутами.

В связи с прямой связью между Avg\_age и Heating\_oil, а также между Heating\_Oil и Insulation, логично, что между Avg\_age и Insulation тоже имеется связь, но не сильная, а средняя. Аналогично связаны атрибуты Avg\_age – Heating\_Oil – Temperature, только связь обратная из-за того, что между Heating\_Oil и Temperature также связь обратная.

В среднем для домов большего размера требуется больше нагревательных блоков, поэтому связь между Home\_Size и Heating\_Oil прямая и умеренная.

Чем выше средний возраст жильцов дома, тем, вероятнее всего, у них больше семья и доход, поэтому они могут позволить себе дом большего размера. Отсюда следует умеренная прямая связь между Home\_Size и Avg\_age.

Остальные связи слабые или очень слабые, потому что эти атрибуты между собой практически не связаны.

**2. Задание для самостоятельного выполнения**

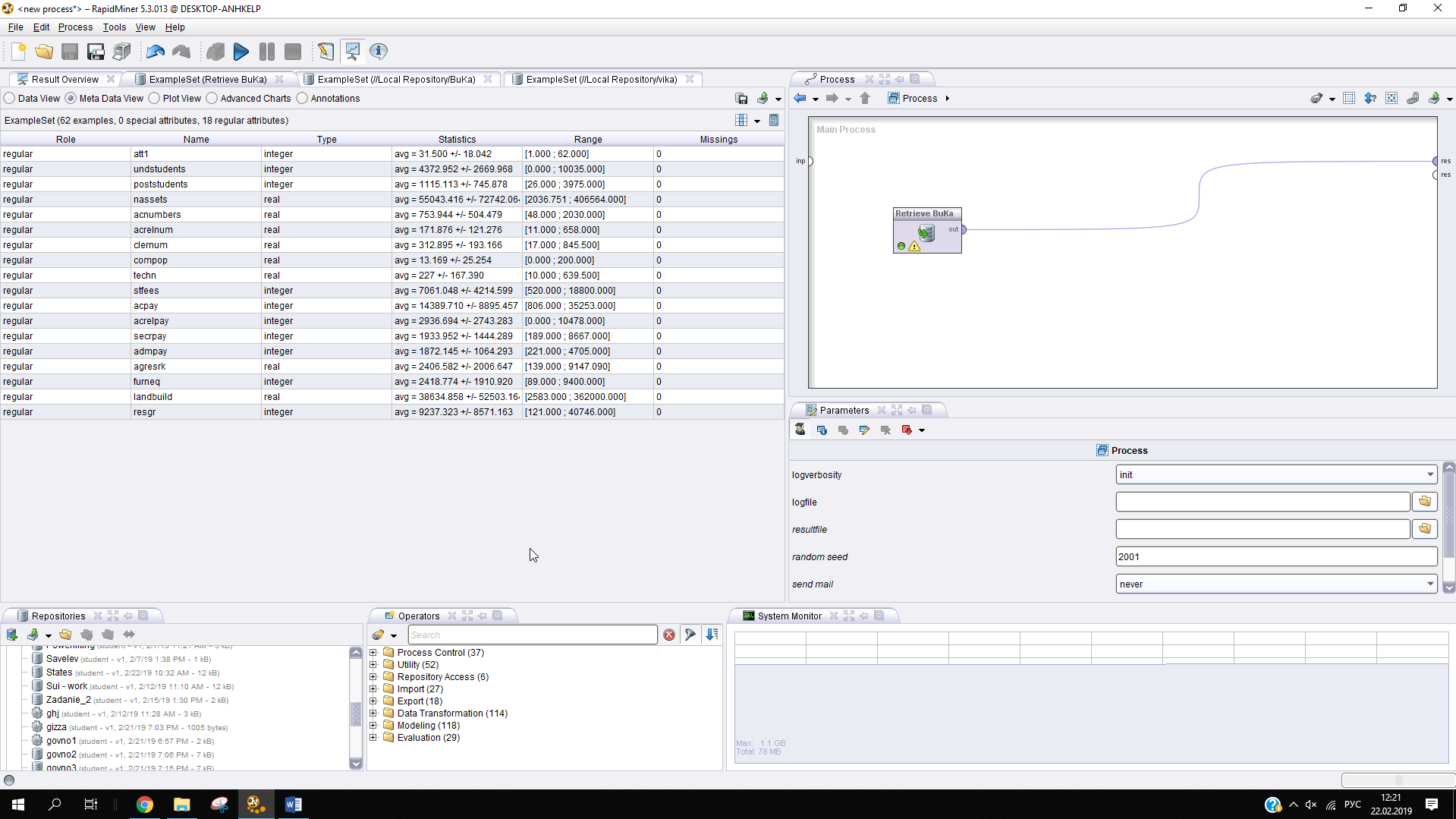
В качестве набора данных был выбран датасет University.csv с информацией о студентах, обучающихся в университете.

Источник: http://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/University.html.

Данный файл содержит следующие атрибуты:

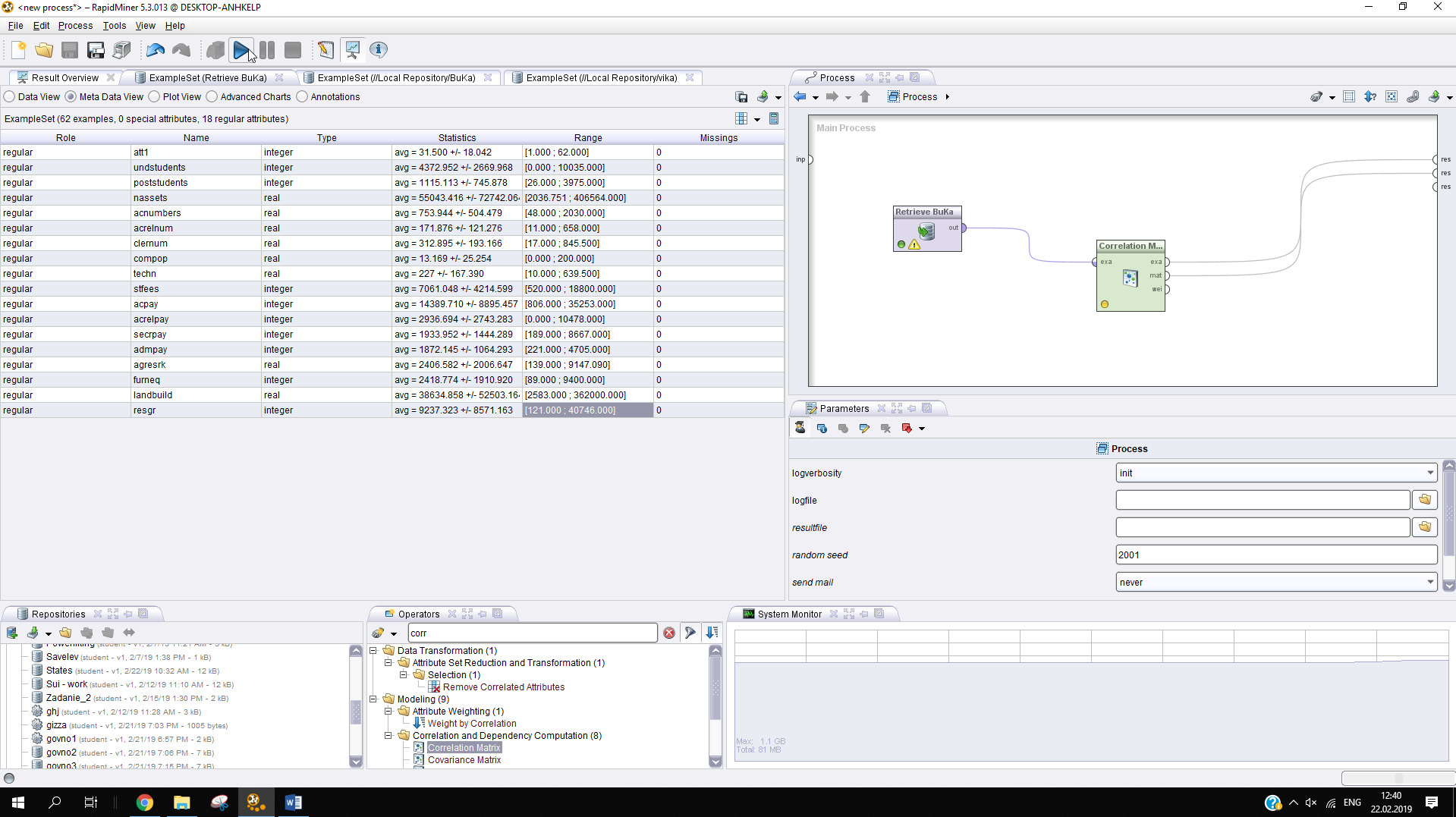
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *№* | *Название* | *Описание* | *Тип* | *Формат* |
| 1 | undstudents | студенты | Integer | Число [0.000 ; 10035.000] |
| 2 | poststudents | аспиранты | Integer | Число  [26.000 ; 3975.000] |
| 3 | nassets | чистые активы | Real | Число  [2036.751 ; 406564.000] |
| 4 | acnumbers | совокупный рейтинг успеваемости | Real | Число  [48.000 ; 2030.000] |
| 5 | acrelnum | академические номера | Real | Число  [11.000 ; 658.000] |
| 6 | clernum | канцелярские номера | Real | Число  [17.000 ; 845.500] |
| 7 | compop | операторы компьютеров | Real | Число  [0.000 ; 200.000] |
| 8 | techn | технический номер | Real | Число  [10.000 ; 639.500] |
| 9 | stfees | студенческие сборы | Integer | Число  [520.000 ; 18800.000] |
| 10 | acpay | академическая оплата | Integer | Число  [806.000 ; 35253.000] |
| 11 | acrelpay | оплата за научные труды | Integer | Число  [0.000 ; 10478.000] |
| 12 | secrpay | оплата труда секретаря | Integer | Число  [189.000 ; 8667.000] |
| 13 | admpay | административная плата | Integer | Число  [221.000 ; 4705.000] |
| 14 | agresrk | совокупный исследовательский рейтинг | Real | Число  [139.000 ; 9147.090] |
| 15 | furneq | мебель и оборудование | Integer | Число  [89.000 ; 9400.000] |
| 16 | landbuild | земля и здания | Real | Число  [2583.000 ; 362000.000] |
| 17 | resgr | исследовательские гранты | Integer | Число  [121.000 ; 40746.000] |

Импортируем набор данных в RM, оставив лишь вышеперечисленные атрибуты. Затем создаем новый проект и импортируем туда набор данных, затем добавим его на главный процесс.



*Рисунок 6 - Импорт данных*

Корреляционный анализ атрибутов набора данных. Для проведение корреляционного анализа атрибутов набора данных, необходимо найти корреляционную матрицу. Для этого во вкладке Operators найдем оператор Correlation Matrix и добавим его на главный процесс.



*Рисунок 7 – добавление операторов в поле Main Process*

Корреляционная матрица – квадратная матрица *P* размера *n\*n*, где *n* – количество атрибутов. – корреляционная связь между *X*-м и *Y*-м атрибутом. Все диагональные элементы матрицы *P* равны 1.

Коэффициент корреляции рассчитывается по формуле:

где , – среднее значение выборок.

Коэффициент корреляции – показатель статистической связи двух атрибутов, изменяется в пределах от -1 до 1.

Доказательство:

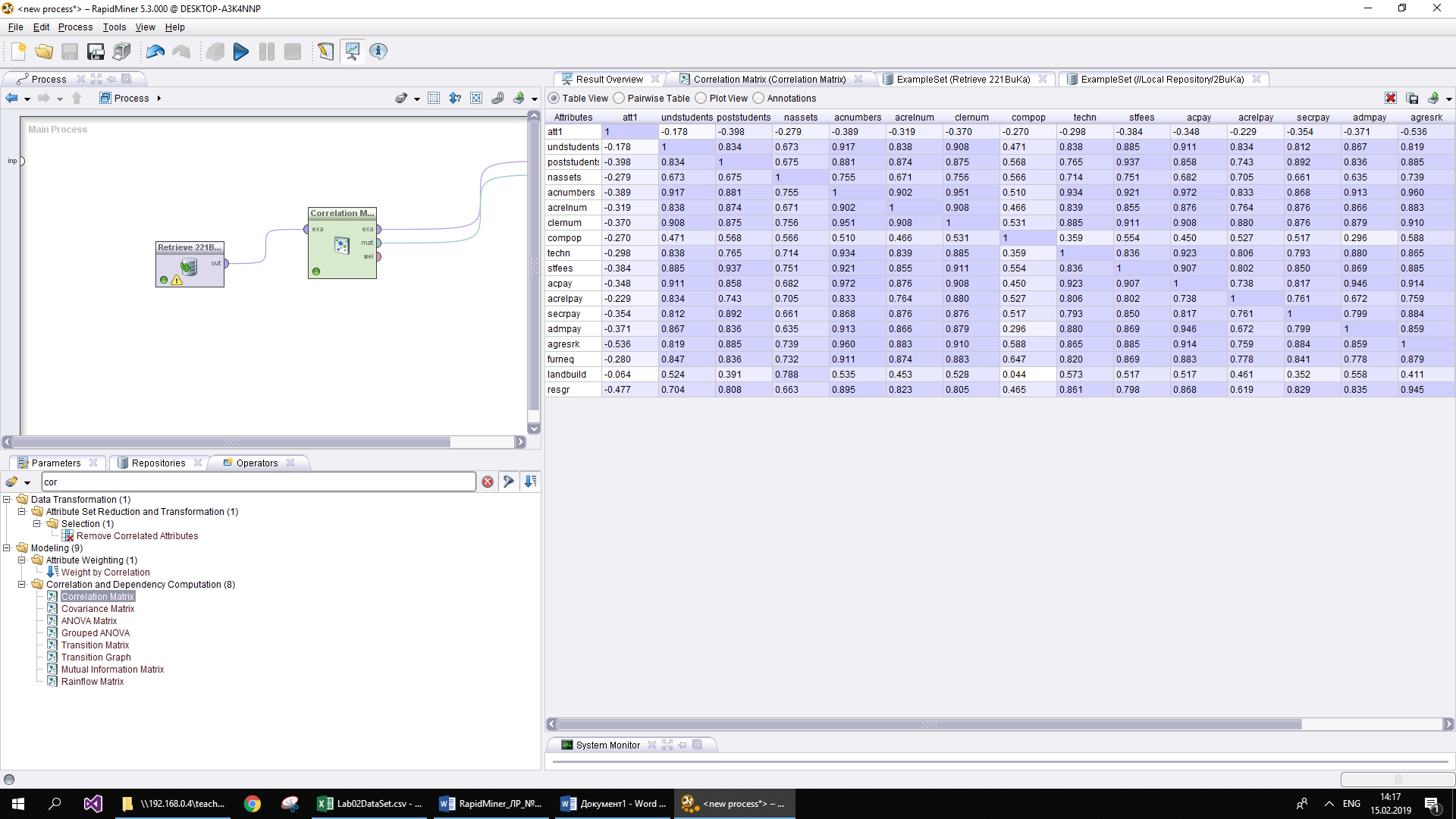
Разделив обе части двойного неравенства на получим

Корреляцию считают:

* Функциональной при
* Сильной, если коэффициент корреляции
* Средней при
* Умеренной при
* Слабой при
* Очень слабой при

Если , то корреляционная связь между атрибутами прямая (увеличение значения одного атрибута ведет к увеличению значения другого), при – обратная.

Запустим процесс и RapidMiner автоматически переключит рабочую область в Results, где отображается полученная корреляционная матрица.



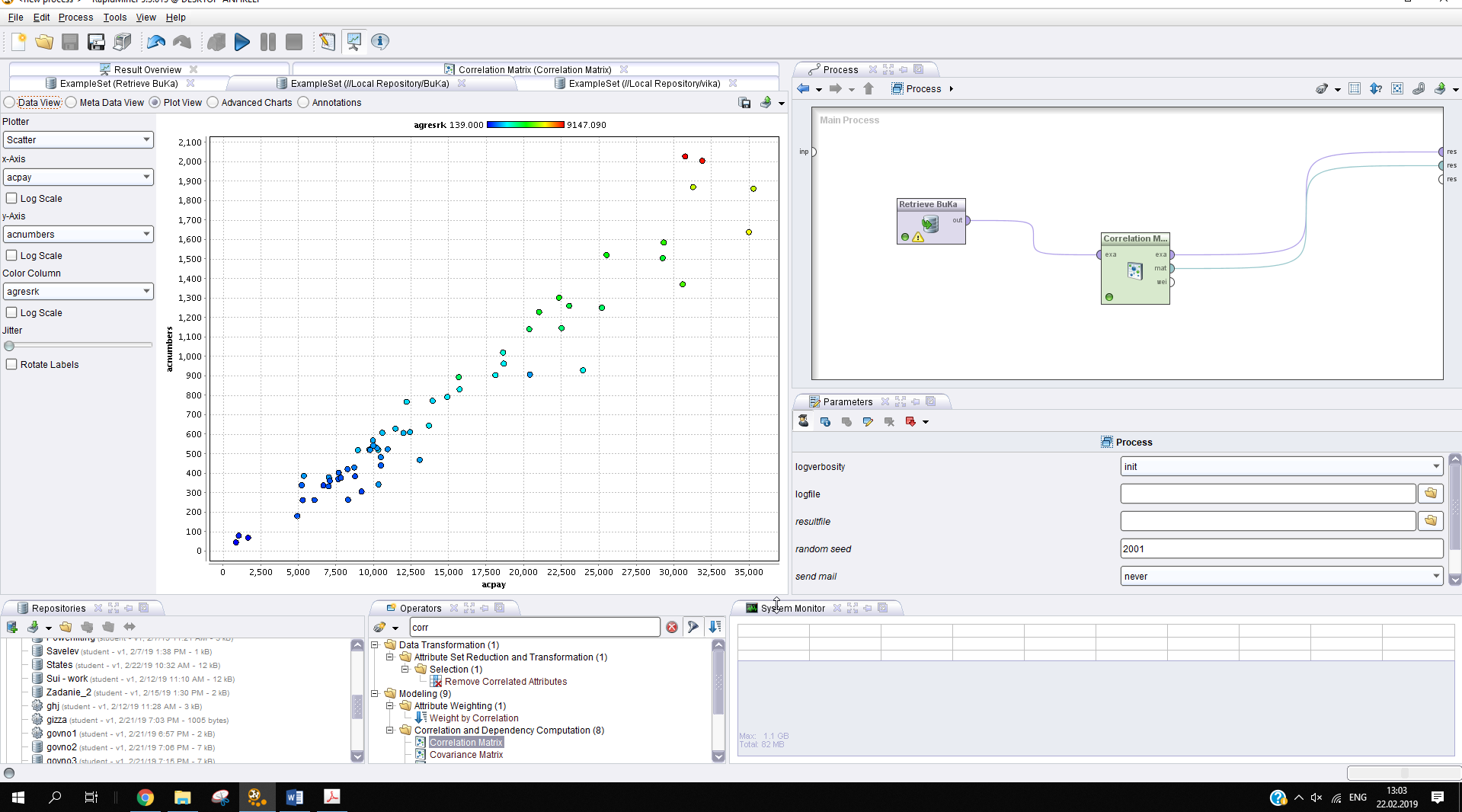
*Рисунок 8 – результат работы оператора Корреляционной матрицы*

Ввиду большого количества атрибутов рассмотрим только 8 из них: undstudents, poststudents, acnumbers, acrelnum, acpay, acrelpay, agresrk и resgr.

*Таблица 2*

|  |  |  |  |
| --- | --- | --- | --- |
| *Первый атрибут* | *Второй атрибут* | *Зависимость* | *Описание* |
| undstudents | poststudents | Сильная, прямая зависимость  (0.84) | Количество студентов зависит от количества аспирантов, существует норма по количеству мест для приема на обучение |
| undstudents | acnumbers | Сильная, прямая зависимость  (0.92) | Совокупный рейтинг успеваемости формируется с учетом количества студентов |
| undstudents | acrelnum | Сильная, прямая зависимость  (0.84) | Академический номер присваивается каждому студенту, с увеличением студентов увеличится количество академических номеров |
| undstudents | acpay | Сильная, прямая зависимость  (0.91) | Стипендия формируется по количеству студентов |
| undstudents | acrelpay | Сильная, прямая зависимость  (0.83) | Стипендия за научные труды формируется по количеству студентов |
| undstudents | agresrk | Сильная, прямая зависимость  (0.82) | Совокупный исследовательский рейтинг формируется по количеству студентов |
| undstudents | resgr | Сильная, прямая зависимость  (0.70) | Количество выдаваемых исследовательских грантов зависит от количества студентов |
| poststudents | acnumbers | Сильная, прямая зависимость  (0.89) | Совокупный рейтинг успеваемости формируется с учетом количества аспирантов |
| poststudents | acrelnum | Сильная, прямая зависимость  (0.87) | Академический номер присваивается каждому аспиранту, с увеличением аспирантов увеличится количество академических номеров |
| poststudents | acpay | Сильная, прямая зависимость  (0.86) | Стипендия формируется по количеству аспирантов |
| poststudents | acrelpay | Сильная, прямая зависимость  (0.74) | Стипендия за научные труды формируется по количеству аспирантов |
| poststudents | agresrk | Сильная, прямая зависимость  (0.89) | Совокупный исследовательский рейтинг формируется по количеству аспирантов |
| poststudents | resgr | Сильная, прямая зависимость  (0.81) | Количество выдаваемых исследовательских грантов зависит от количества аспирантов |
| acnumbers | acrelnum | Сильная, прямая зависимость  (0.90) | Cовокупный рейтинг успеваемости формируется по количеству академических номеров |
| acnumbers | acpay | Сильная, прямая зависимость  (0.97) | Большая стипендия начисляется студентам с большим совокупным рейтингом успеваемости |
| acnumbers | acrelpay | Сильная, прямая зависимость  (0.83) | Большая стипендия за научные труды начисляется студентам с большим совокупным рейтингом успеваемости |
| acnumbers | agresrk | Сильная, прямая зависимость  (0.96) | Высокие показатели в исследовательском рейтинге гарантируют высокие показатели и в рейтинге успеваемости |
| acnumbers | resgr | Сильная, прямая зависимость  (0.90) | Высокий рейтинг успеваемости увеличивает шансы студента на получение большого исследовательского гранта |
| acrelnum | acpay | Сильная, прямая зависимость  (0.88) | Студенты с большими академическими номерами получают большую стипендию |
| acrelnum | acrelpay | Сильная, прямая зависимость  (0.76) | Студенты с большими академическими номерами получают большую стипендию за научные труды |
| acrelnum | agresrk | Сильная, прямая зависимость  (0.88) | Студенты с большими академическими номерами занимают высокие места в рейтинге успеваемости |
| acrelnum | resgr | Сильная, прямая зависимость  (0.82) | Студенты с большими академическими номерами получают большие исследовательские гранты |
| acpay | acrelpay | Сильная, прямая зависимость  (0.74) | Если студент получает оплату за научные труды, значит он хорошо учится и получает большую стипендию |
| acpay | agresrk | Сильная, прямая зависимость  (0.91) | Студенты с большими стипендиями занимают высокие места в рейтинге успеваемости |
| acpay | resgr | Сильная, прямая зависимость  (0.87) | Студенты с большими стипендиями вследствие хорошей учебы получают гранты на исследования |
| acrelpay | agresrk | Сильная, прямая зависимость  (0.76) | Студенты, имеющие высокий исследовательский рейтинг, проводят много исследований и получают большую оплату за научные труды |
| acrelpay | resgr | Средняя, прямая зависимость  (0.62) | Студенты, имеющие исследовательские гранты, проводят исследования и получают большую оплату за научные труды |
| agresrk | resgr | Сильная, прямая зависимость  (0.95) | Студенты, имеющие исследовательские гранты, проводят исследования и оказываются на высоких строчках в совокупном исследовательском рейтинге |

Визуализируем полученную корреляционную матрицу с помощью Plot View.



*Рисунок 9 - Визуализация корреляционной матрицы*

Визуализируем атрибуты acpay, acnumbers и agresrk. Как видно, график представлен на Рис. 9.

Визуализация acpay, acnumbers показывает, что увеличение одного признака acpay приводит к увеличению другого acnumbers (acpay -> acnumbers = 0.96), большая стипендия начисляется студентам с большим совокупным рейтингом успеваемости, т.е. для увеличения стипендии необходимо увеличивать совокупный исследовательский рейтинг. Подтверждается сильная, прямая связь между атрибутами.

Также прямую и сильную связь имеют следующие атрибуты: acnumbers и agresrk (acnumbers -> agresrk = 0.91). Высокие показатели в исследовательском рейтинге гарантируют высокие показатели и в рейтинге успеваемости, т.е. от результатов совокупного исследовательского рейтинга зависят результаты и в совокупном рейтинге успеваемости, чем больше исследовательский рейтинг – тем выше рейтинг успеваемости.

Атрибуты acpay и agresrk (acpay -> agresrk = 0.97) имеют сильную и прямую связь. Студенты с большими стипендиями занимают высокие места в рейтинге успеваемости. Т.е. для того, чтобы иметь большую стипендию необходимо находится на верхних строчках совокупного рейтинга успеваемости.

**Вывод**

При выполнении лабораторной работы были приобрести навыки подготовки набора данных для его интеллектуального анализа с помощью системы RapidMiner, а также выполнены следующие задачи:

* Выполнено обучающее задание в соответствии с руководством по выполнению лабораторной работы;
* Выполнено задание для самостоятельной работы в соответствии с руководством по выполнению лабораторной работы;
* Оформлен отчет о выполнении лабораторной работы в соответствии с требованиями к его оформлению.