



РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Э М И Т
И Н С Т И Т У Т
ЭКОНОМИКИ, МАТЕМАТИКИ
И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ



Интерактивный DataMining

Цифровая трансформация

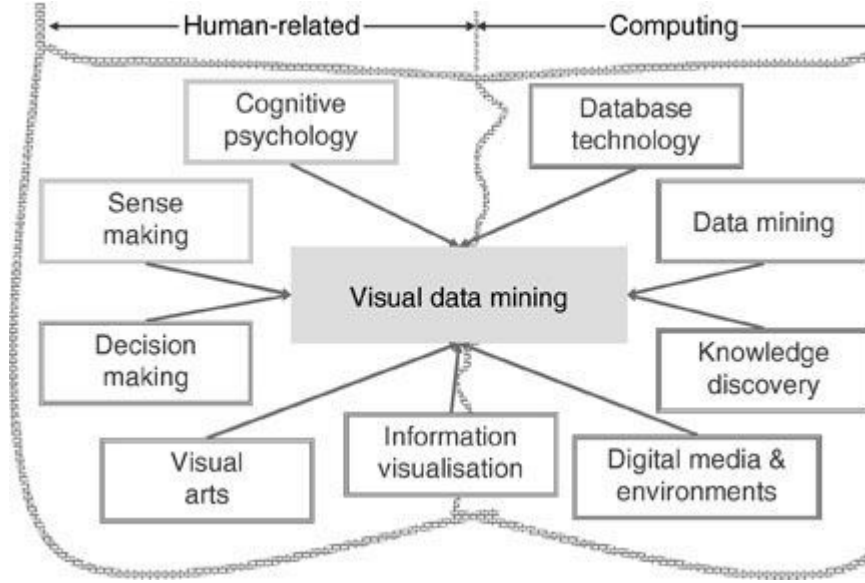
Почему мы используем инструменты визуального анализа данных?

Сегодня результаты анализа данных позволяют по новому взглянуть на процессы происходящие на предприятии. И для этого нужна модель для того или иного видения и последующего анализа этого видения

Нужен инструмент, который позволит вам быстро построить и проанализировать модель

Visual data mining (VDM)

Визуальный интеллектуальный анализ данных (VDM) - это процесс взаимодействия и аналитических рассуждений с одним или несколькими визуальными представлениями абстрактных данных. Этот процесс может привести к визуальному обнаружению устойчивых шаблонов в этих данных или предоставить некоторые рекомендации по применению других методов анализа данных и анализа данных.



Orange: Data Mining Fruitful and Fun

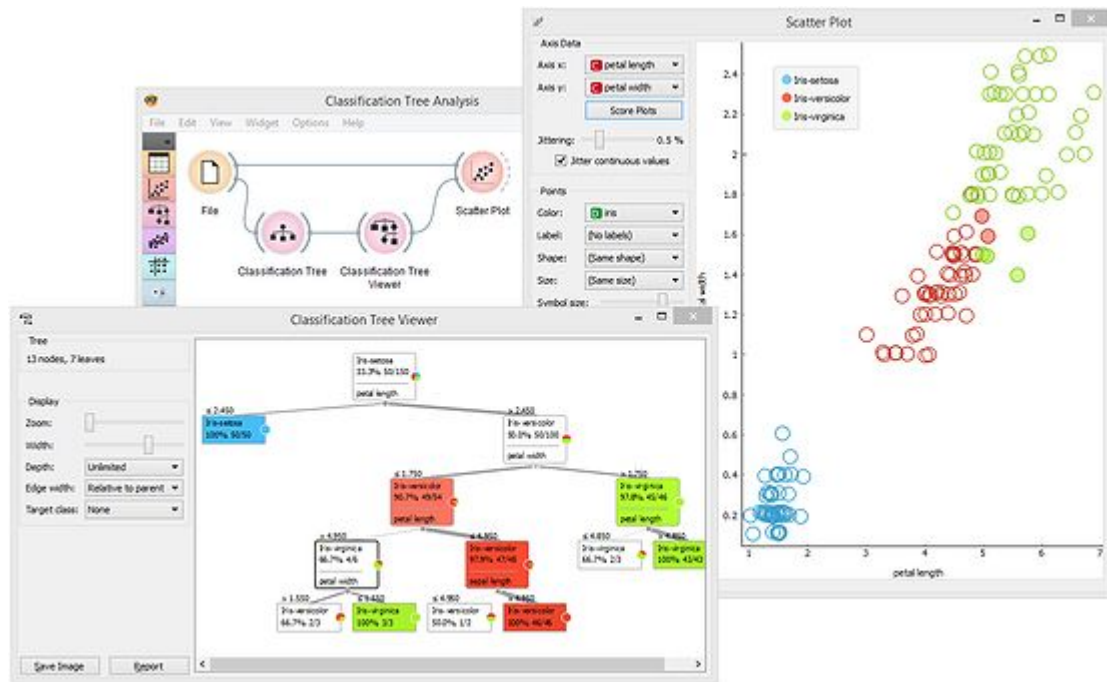
Данный продукт предлагает машинное обучение с открытым исходным кодом и визуализация данных для новичков и экспертов. Интерактивные рабочие процессы анализа данных с большим набором инструментов.

download [orange.biolab.si/download/](https://orangedatamining.com/download/)



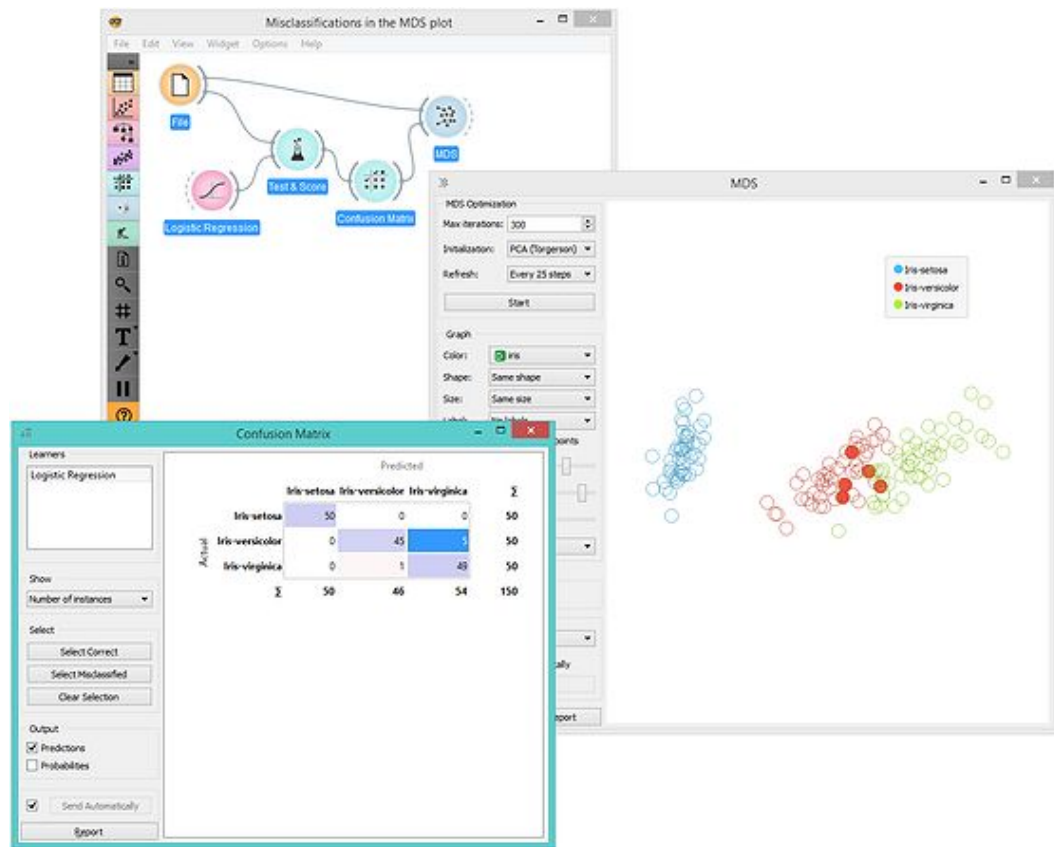
Orange: Интерактивный анализ данных

- Анализ данных с умной визуализацией данных.
- Традиционный статистический анализ
- Основные алгоритмы машинного обучения доступны в интерактивном использовании



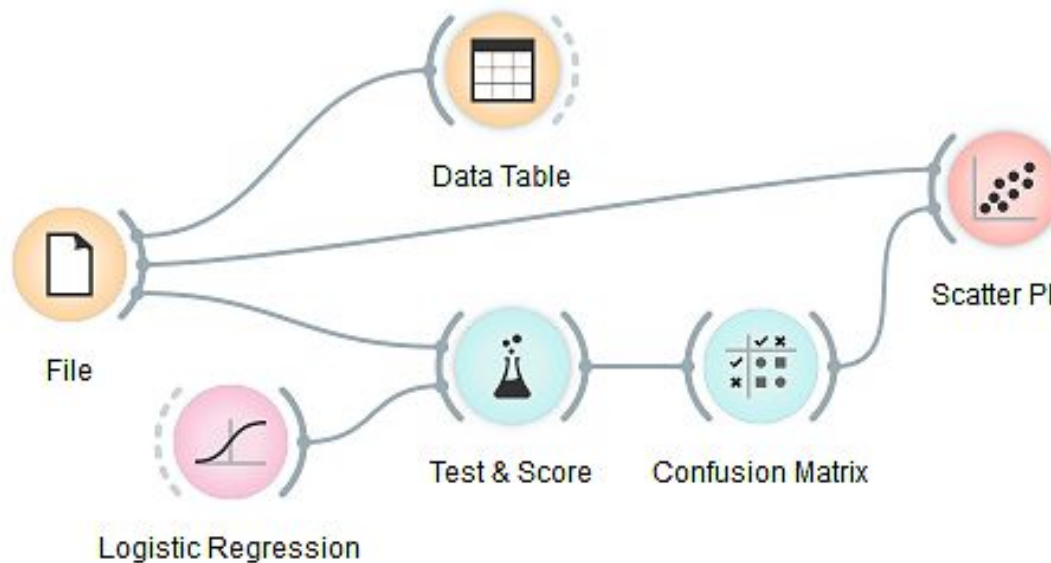
Для исследователей моделей

- Формирования представительных выборок
- Различные виды кроссвалидации
- Оценка качества предсказания
- Визуализация



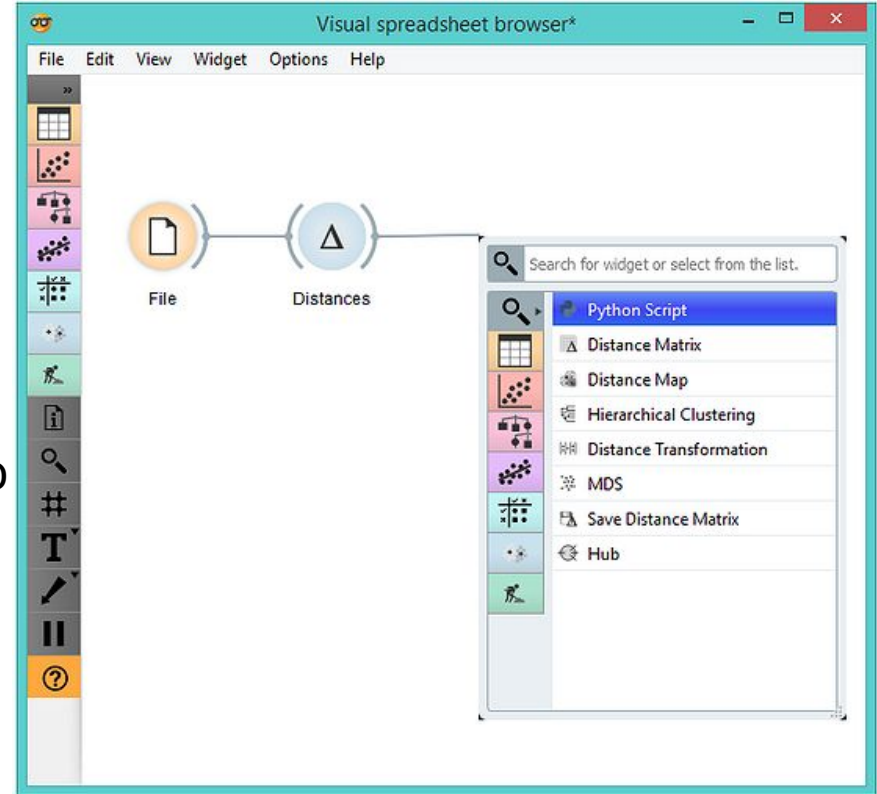
Component-Based Data Mining

В Orange, анализ данных выполняется путем объединения компонентов в рабочие процессы. Каждый компонент, называемый виджетом, встраивает некоторую задачу поиска, предварительной обработки, визуализации, моделирования или оценки данных.



Clever Workflow Design Interface

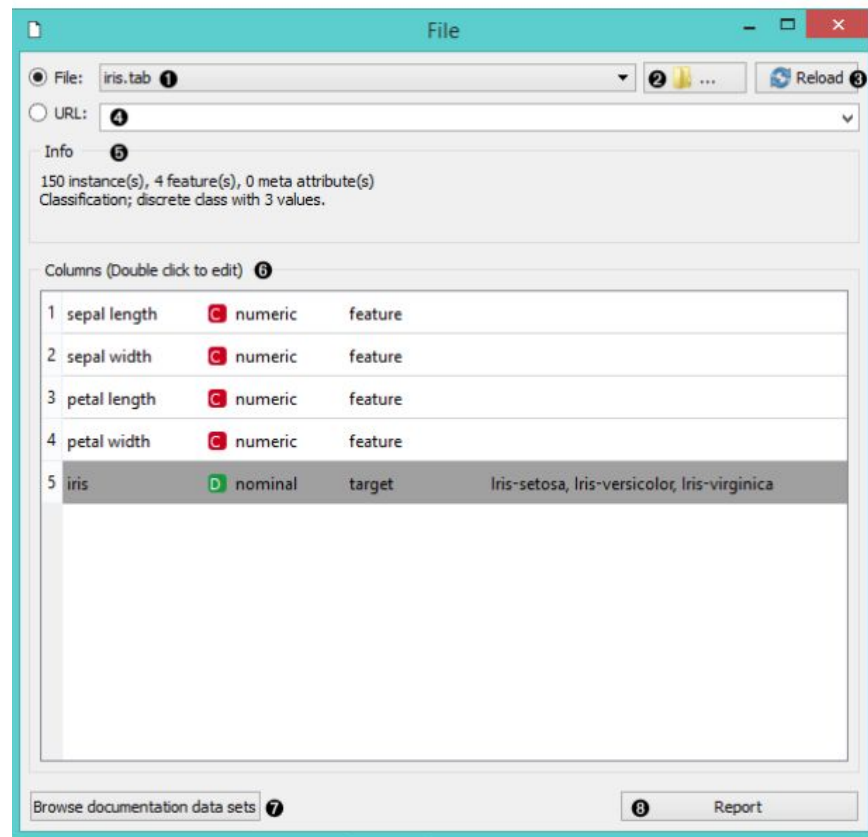
Orange растёт в использовании даже для начинающих. Начните с виджета Файл, и Orange автоматически предложит следующие виджеты, которые могут быть к нему подключены. Например, Orange знает, что вы, вероятно, захотите Иерархическую Кластеризацию после того, как вы настроили свой виджет Расстояний



Orange: File widget

Orange может импортировать файл с запятой или табуляцией, файл Excel или Google Sheet.

Необходимо пометить столбцы и выбрать целевую переменную, чтобы алгоритмы работали верно



Виджет viewer в Orange

Info

793 instances (no missing values)
2 features (no missing values)
Discrete class with 4 values (no missing values)
1 meta attribute (no missing values)

Variables

☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

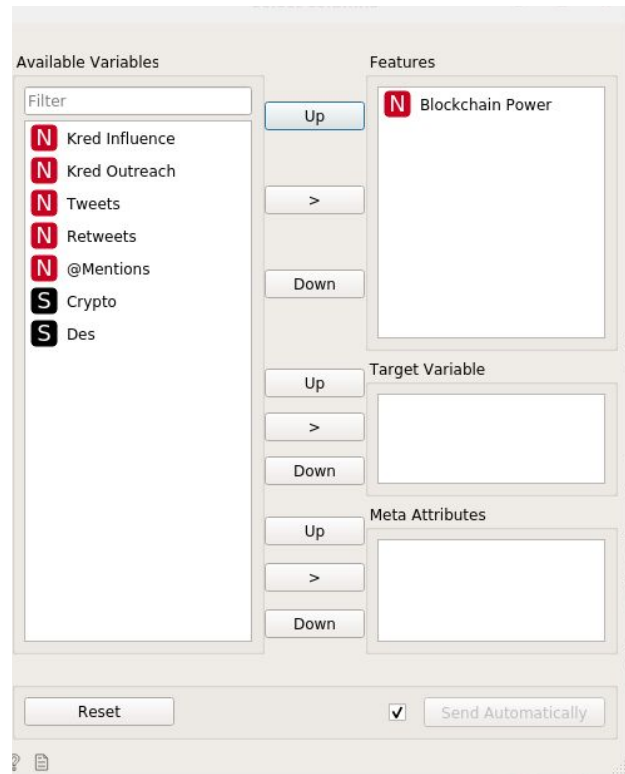
☒ Send Automatically

rfm	rfm	monetary_value	customer	recency	frequency
1	< 1142	Aaron Bergman	≥ 182	< 4	
9	< 1142	Adrian Shami	30 - 74	< 4	
10	< 1142	Aimee Bixby	30 - 74	4 - 6	
11	< 1142	Alan Barnes	< 30	6 - 8	
16	< 1142	Alan Shonely	≥ 182	6 - 8	
17	< 1142	Alejandro Ballentine	74 - 182	≥ 8	
20	< 1142	Aleksandra Gannaway	74 - 182	< 4	
22	< 1142	Alex Grayson	< 30	4 - 6	
23	< 1142	Alex Russell	74 - 182	< 4	
24	< 1142	Alice McCarthy	< 30	4 - 6	
25	< 1142	Allen Arnold	< 30	≥ 8	
26	< 1142	Allen Goldenen	74 - 182	4 - 6	
28	< 1142	Alyssa Crouse	≥ 182	< 4	
34	< 1142	Andrew Roberts	< 30	4 - 6	
37	< 1142	Andy Yotov	< 30	< 4	
38	< 1142	Anemone Ratner	≥ 182	< 4	
39	< 1142	Angele Hood	< 30	< 4	
42	< 1142	Ann Steele	74 - 182	6 - 8	
44	< 1142	Anna Chung	≥ 182	4 - 6	
51	< 1142	Anthony Garverick	74 - 182	< 4	
54	< 1142	Anthony O'Donnell	≥ 182	< 4	
56	< 1142	Anthony Witt	< 30	< 4	
59	< 1142	Art Foster	≥ 182	< 4	



Orange: Виджет выбора столбцов

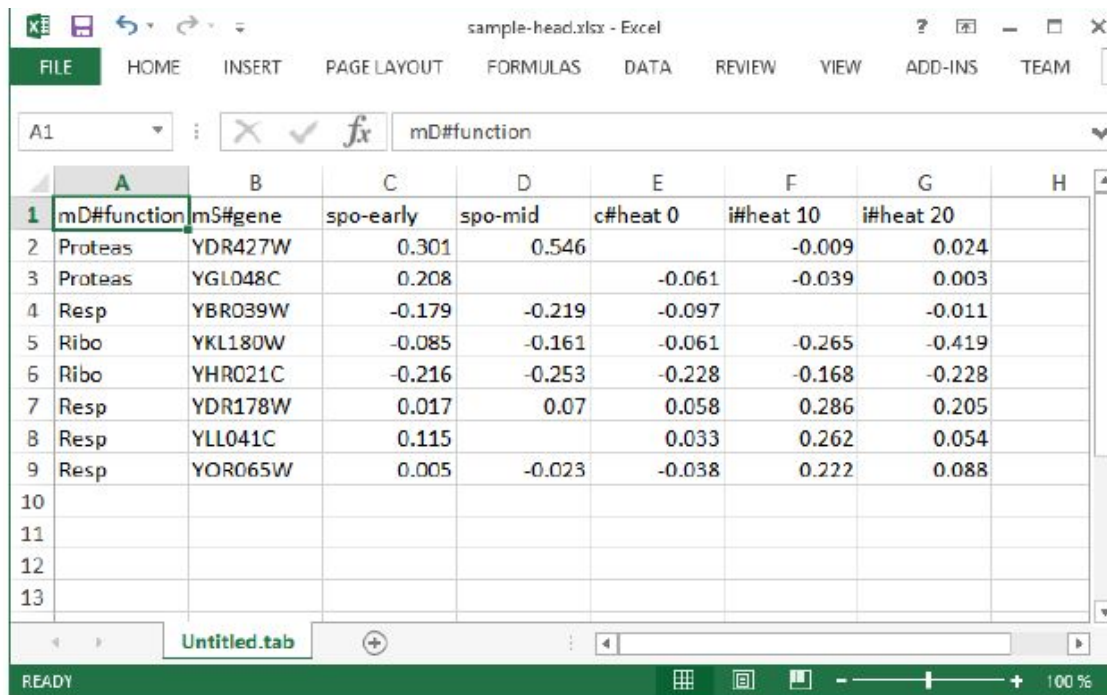
Изменение назначения атрибутов происходит в виджете «Выбрать столбцы» необходимо подтвердить, нажав кнопку «Применить». Данные из этого виджета поступают в таблицу данных, которая отображает данные.



Orange: Probably the best widget for mapping

Вы можете отобразить свои данные непосредственно в виджете MS Excel, используя знак хеша («#»). Префиксы для ролей атрибутов:

- c: категориальный
- m: мета
- i: пропустить
- w: вес экземпляра



sample-head.xlsx - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS TEAM

A1 : X ✓ fx mD#function

	A	B	C	D	E	F	G	H
1	mD#function	mS#gene	spo-early	spo-mid	c#heat 0	i#heat 10	i#heat 20	
2	Proteas	YDR427W	0.301	0.546		-0.009	0.024	
3	Proteas	YGL048C	0.208		-0.061	-0.039	0.003	
4	Resp	YBR039W	-0.179	-0.219	-0.097		-0.011	
5	Ribo	YKL180W	-0.085	-0.161	-0.061	-0.265	-0.419	
6	Ribo	YHR021C	-0.216	-0.253	-0.228	-0.168	-0.228	
7	Resp	YDR178W	0.017	0.07	0.058	0.286	0.205	
8	Resp	YLL041C	0.115		0.033	0.262	0.054	
9	Resp	YOR065W	0.005	-0.023	-0.038	0.222	0.088	
10								
11								
12								
13								

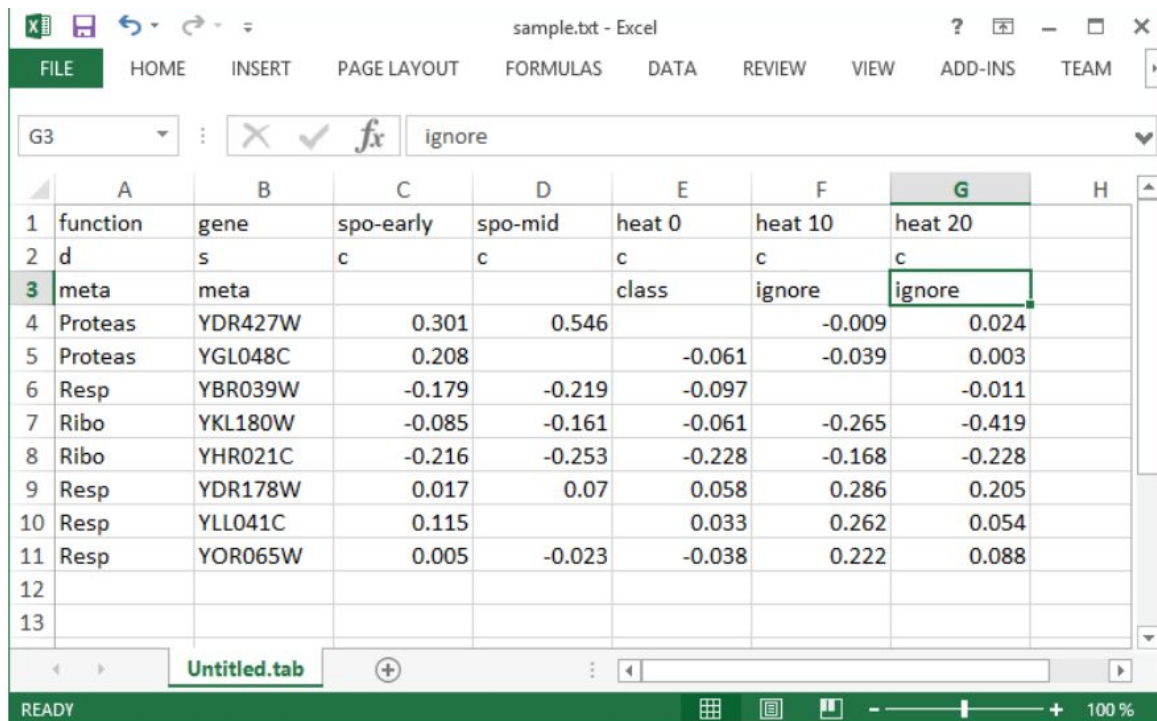
Untitled.tab

READY 100 %

Orange: формирование панелей данных

Типы переменных:

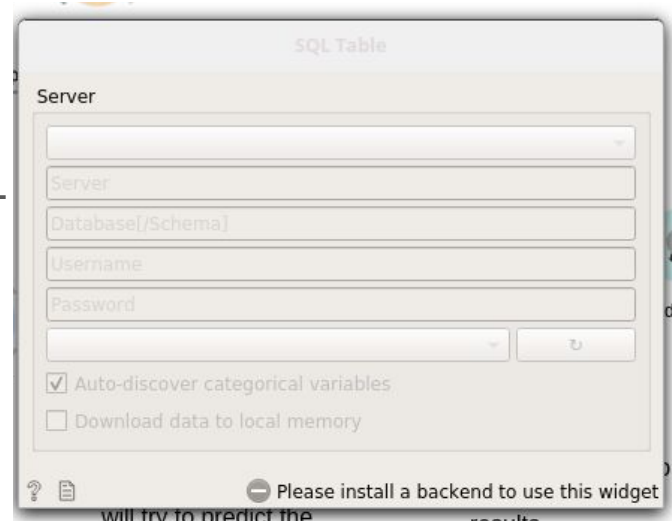
- C: Непрерывные
- D: Дискретные
- T: Временные
- S: тестовый



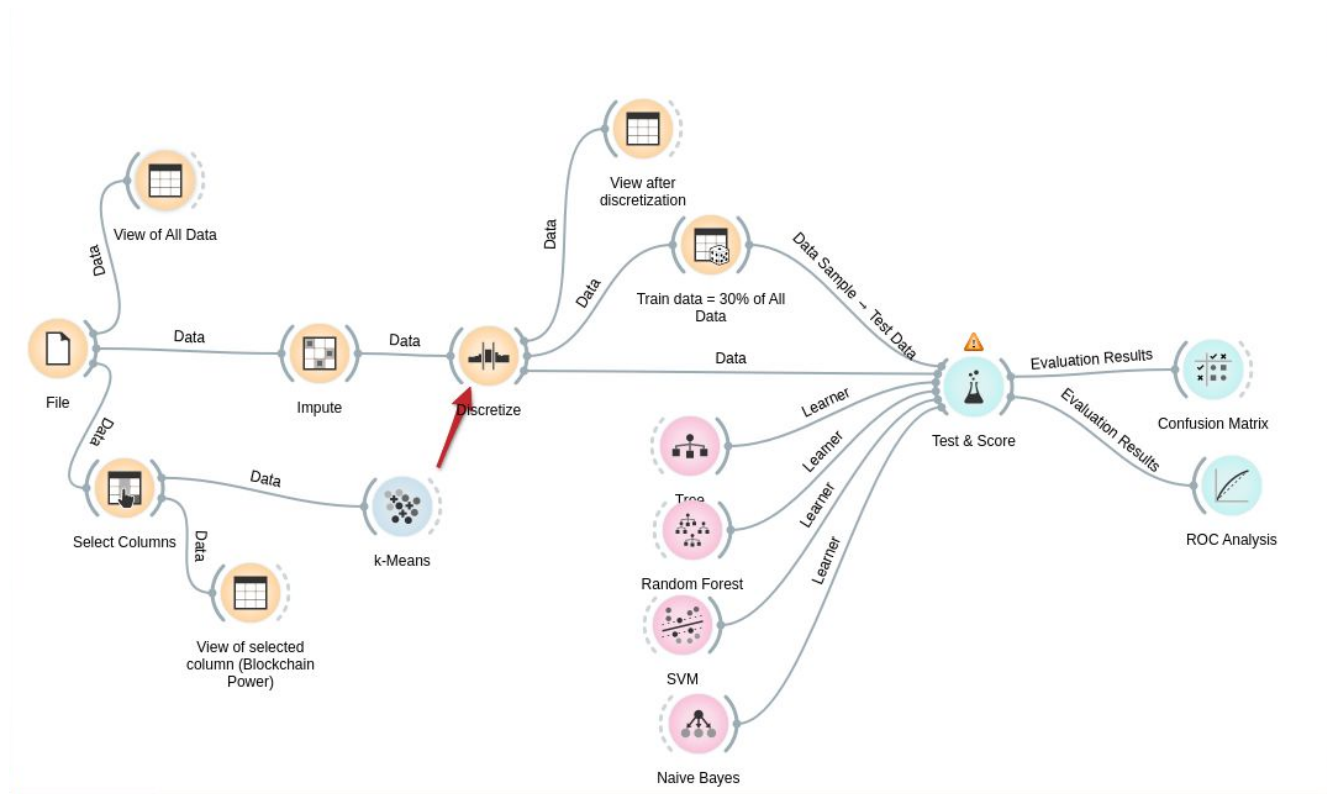
	A	B	C	D	E	F	G	H
1	function	gene	spo-early	spo-mid	heat 0	heat 10	heat 20	
2	d	s	c	c	c	c	c	
3	meta	meta			class	ignore	ignore	
4	Proteas	YDR427W	0.301	0.546		-0.009	0.024	
5	Proteas	YGL048C	0.208		-0.061	-0.039	0.003	
6	Resp	YBR039W	-0.179	-0.219	-0.097		-0.011	
7	Ribo	YKL180W	-0.085	-0.161	-0.061	-0.265	-0.419	
8	Ribo	YHR021C	-0.216	-0.253	-0.228	-0.168	-0.228	
9	Resp	YDR178W	0.017	0.07	0.058	0.286	0.205	
10	Resp	YLL041C	0.115		0.033	0.262	0.054	
11	Resp	YOR065W	0.005	-0.023	-0.038	0.222	0.088	
12								
13								

Orange: SQL = подключение

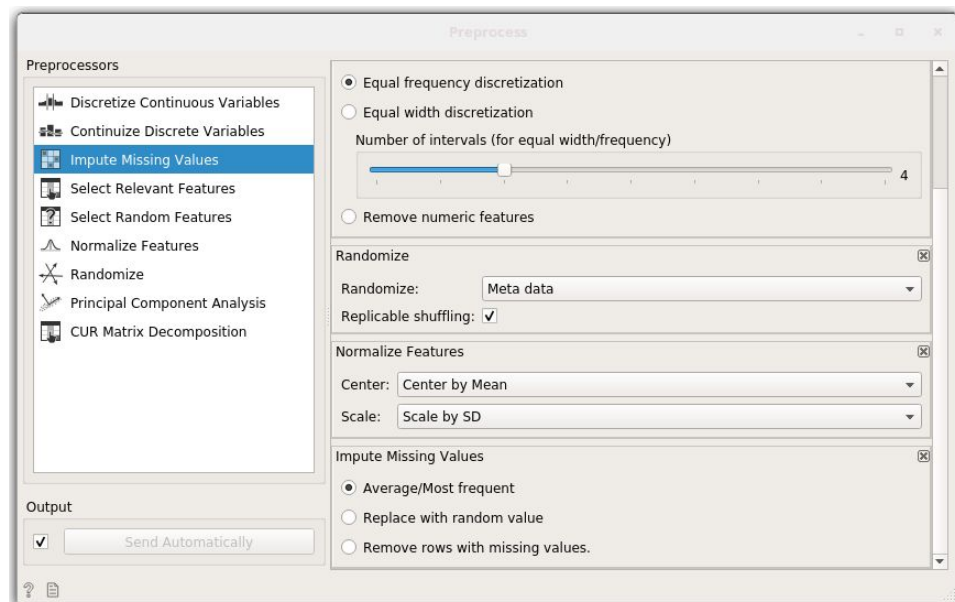
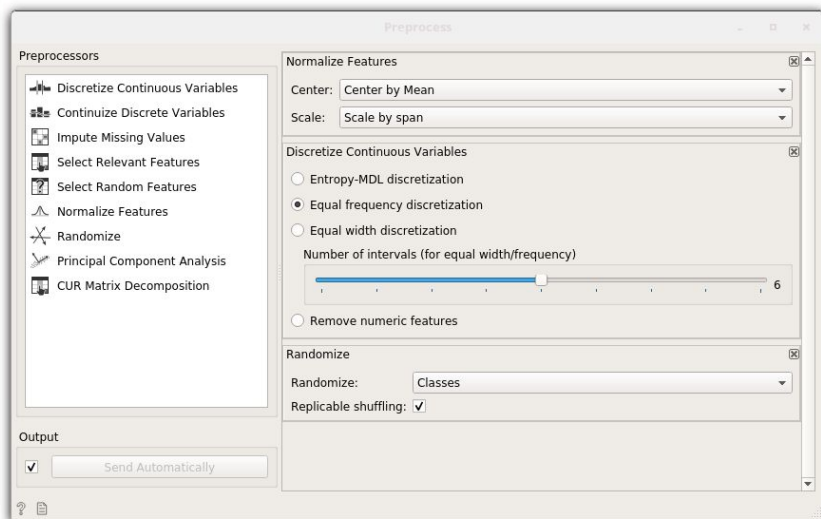
Виджет SQL обращается к данным, хранящимся в базе данных SQL. Он может подключаться к PostgreSQL (требуется модуль psycopg2) или к SQL Server (требуется модуль pumsql).



Как должно выглядеть и исследование (есть ошибки)

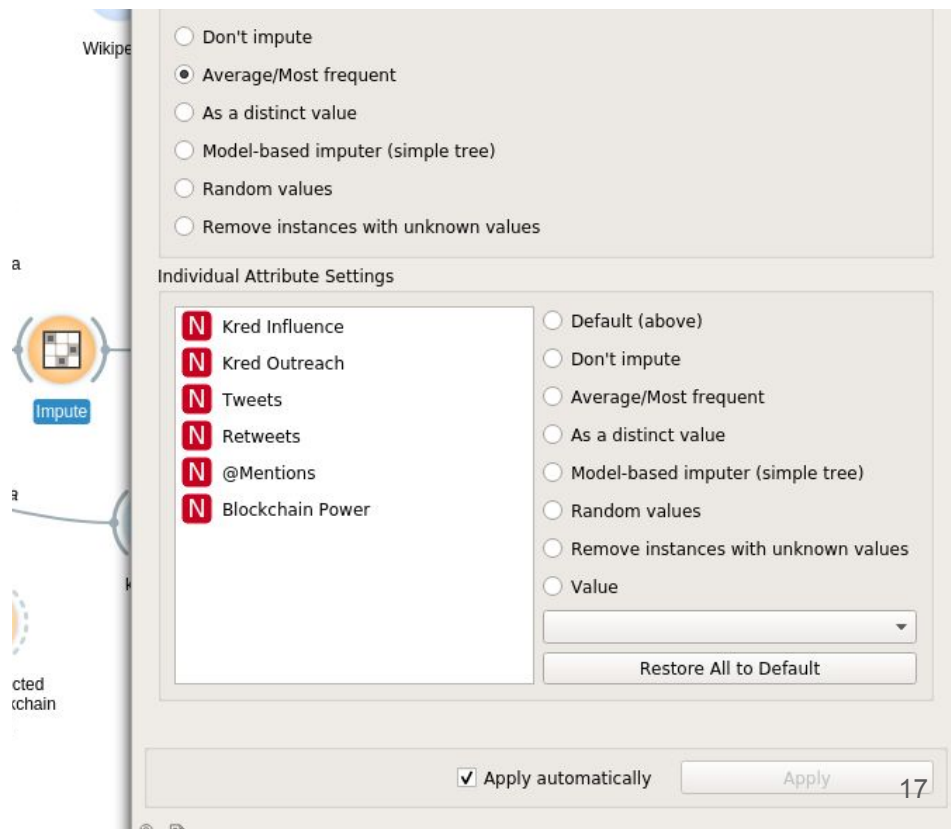


Виджет препроцессинга widget



Заполнение пропусков

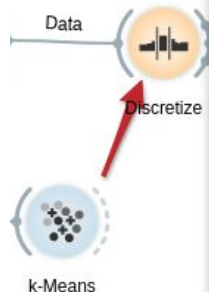
Например средним.



Дискретизация

- Entropy-MDL рекурсивно разделяет атрибут, максимизируя прирост информации, пока усиление не станет меньше минимальной длины описания разреза
- Equal-frequency распределяет на заданное количество интервалов
- Equal-width равномерно разбивает диапазон между наименьшим и наибольшим наблюдаемым значением

ikipedia



The screenshot shows a 'Discretize' window with the following settings:

- Default Discretization:**
 - ☐ Equal-frequency discretization
 - ☐ Equal-width discretization
 - ☒ Leave numeric
 - ☐ Entropy-MDL discretization
 - ☐ Remove numeric variables
- Num. of intervals:** 3
- Individual Attribute Settings:**
 - ☐ Default
 - ☐ Leave numeric
 - ☐ Entropy-MDL discretization
 - ☐ Equal-frequency discretization
 - ☒ Equal-width discretization
 - Num. of intervals:** 5
 - ☐ Remove attribute

The list of attributes includes: Kred Influence, Kred Outreach, Tweets, Retweets, @Mentions, and Blockchain Power: 198.80, 396.60, 5. The 'Blockchain Power' attribute is currently selected.

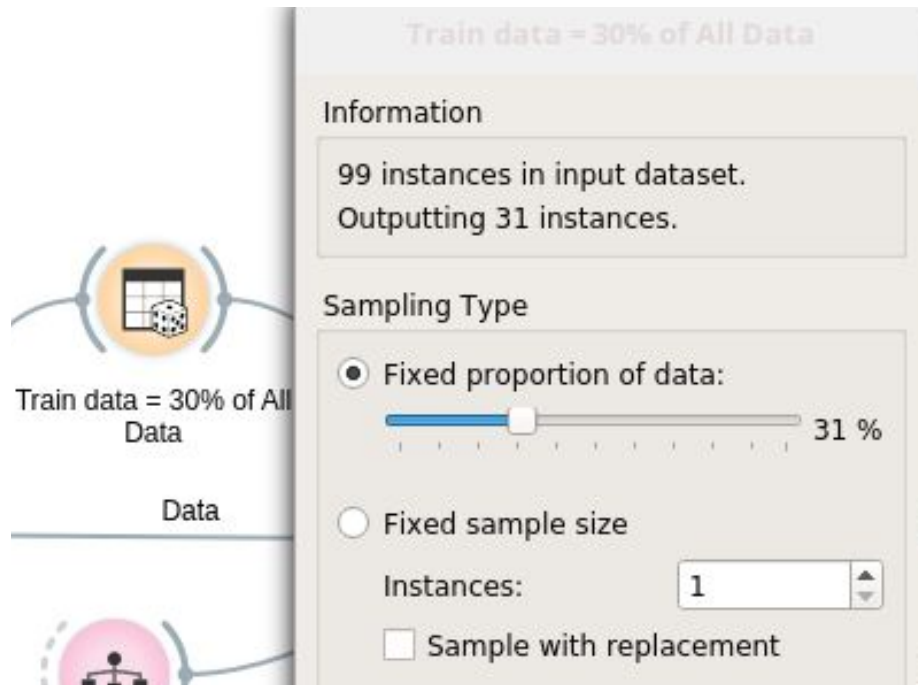
До дискретизации и после

	customer	recency	frequency	monetary_value
1	Aaron Bergman	415	3	887
2	Aaron Hawkins	12	7	1744
3	Aaron Smayling	88	7	3050
4	Adam Bellavance	54	8	7756
5	Adam Hart	34	10	3249
6	Adam Shillingsburg	28	9	3256
7	Adrian Barton	41	10	14474
8	Adrian Hane	60	7	1734
9	Adrian Shami	41	2	59
10	Aimee Bixby	41	5	968
11	Alan Barnes	25	8	1114
12	Alan Dominguez	29	8	6107
13	Alan Haines	55	4	1587
14	Alan Hwang	6	9	4806
15	Alan Schoenberger	22	5	4261
16	Alan Shonely	192	7	585
17	Alejandro Ballentine	166	9	914
18	Alejandro Grove	278	5	2584
19	Alejandro Savely	19	6	3214
20	Aleksandra Gannaway	118	4	368
21	Alex Avila	184	5	5564
22	Alex Grayson	8	5	662
23	Alex Russell	111	4	1056

rfm	rfm	
	monetary_value	customer
1	< 1142	Aaron Bergman
9	< 1142	Adrian Shami
10	< 1142	Aimee Bixby
11	< 1142	Alan Barnes
16	< 1142	Alan Shonely
17	< 1142	Alejandro Ballentine
20	< 1142	Aleksandra Gannaway
22	< 1142	Alex Grayson
23	< 1142	Alex Russell
24	< 1142	Alice McCarthy
25	< 1142	Allen Arnold
26	< 1142	Allen Goldenen
28	< 1142	Alyssa Crouse
34	< 1142	Andrew Roberts
37	< 1142	Andy Yotov
38	< 1142	Anemone Ratner
39	< 1142	Angele Hood
42	< 1142	Ann Steele
44	< 1142	Anna Chung
51	< 1142	Anthony Garverick
54	< 1142	Anthony O'Donnell
56	< 1142	Anthony Witt
59	< 1142	Art Foster

Выборки представительные

Data Sampler widget реализует несколько методов выборки данных. Он выводит выборочный и дополнительный набор данных (с экземплярами из входного набора, которые не включены в набор выборочных данных).



Test & Score:

Test & Score

Sampling

☐ Cross validation
Number of folds: 10
☒ Stratified

☐ Cross validation by feature

☐ Random sampling
Repeat train/test: 10
Training set size: 66 %
☒ Stratified

☐ Leave one out

☐ Test on train data

☒ Test on test data

Target Class

199 - 397

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree	0.967	0.968	0.923	1.000	0.857
SVM	0.750	0.581	0.480	0.333	0.857
Random Forest	1.000	1.000	1.000	1.000	1.000
Naive Bayes	1.000	1.000	1.000	1.000	1.000

Test & Score

Sampling

☐ Cross validation
Number of folds: 10
☒ Stratified

☐ Cross validation by feature

☐ Random sampling
Repeat train/test: 10
Training set size: 66 %
☒ Stratified

☐ Leave one out

☒ Test on train data

☐ Test on test data

Target Class

199 - 397

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree	0.987	0.970	0.933	0.955	0.913
SVM	0.720	0.657	0.541	0.392	0.870
Random Forest	1.000	1.000	1.000	1.000	1.000
Naive Bayes	0.960	0.919	0.846	0.759	0.957

Test data is present but unused. Select 'Test on test data' to use it.

Confusion Matrix или ошибок

		Predicted					
		< 199	199 - 397	397 - 594	594 - 792	≥ 792	Σ
Actual	< 199	35	0	0	0	0	35
	199 - 397	1	21	1	0	0	23
	397 - 594	0	0	19	0	0	19
	594 - 792	0	0	0	13	0	13
	≥ 792	0	0	0	3	6	9
Σ		36	21	20	16	6	99

Анализ текстов: Андерсен

