НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

INTRODUCTION TO PRINCIPAL COMPONENTS ANALYSIS (PCA)

Анализ баз данных в публичном управлении

Кольцов С.Н.

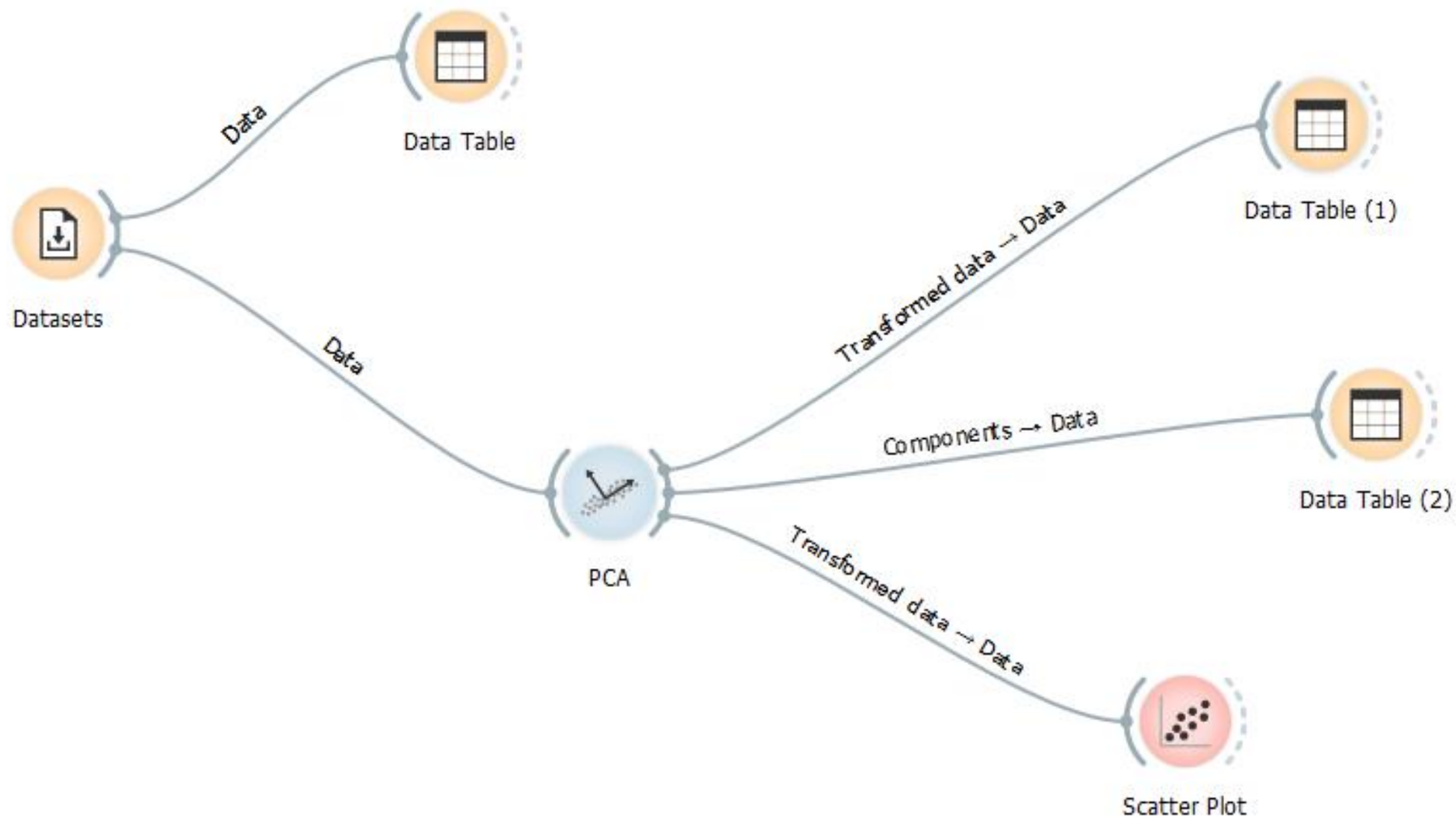
Saint Petersburg, 07.09.2018

PRINCIPAL COMPONENTS ANALYSIS ИЛИ МЕТОД ГЛАВНЫХ КОМПОНЕНТ

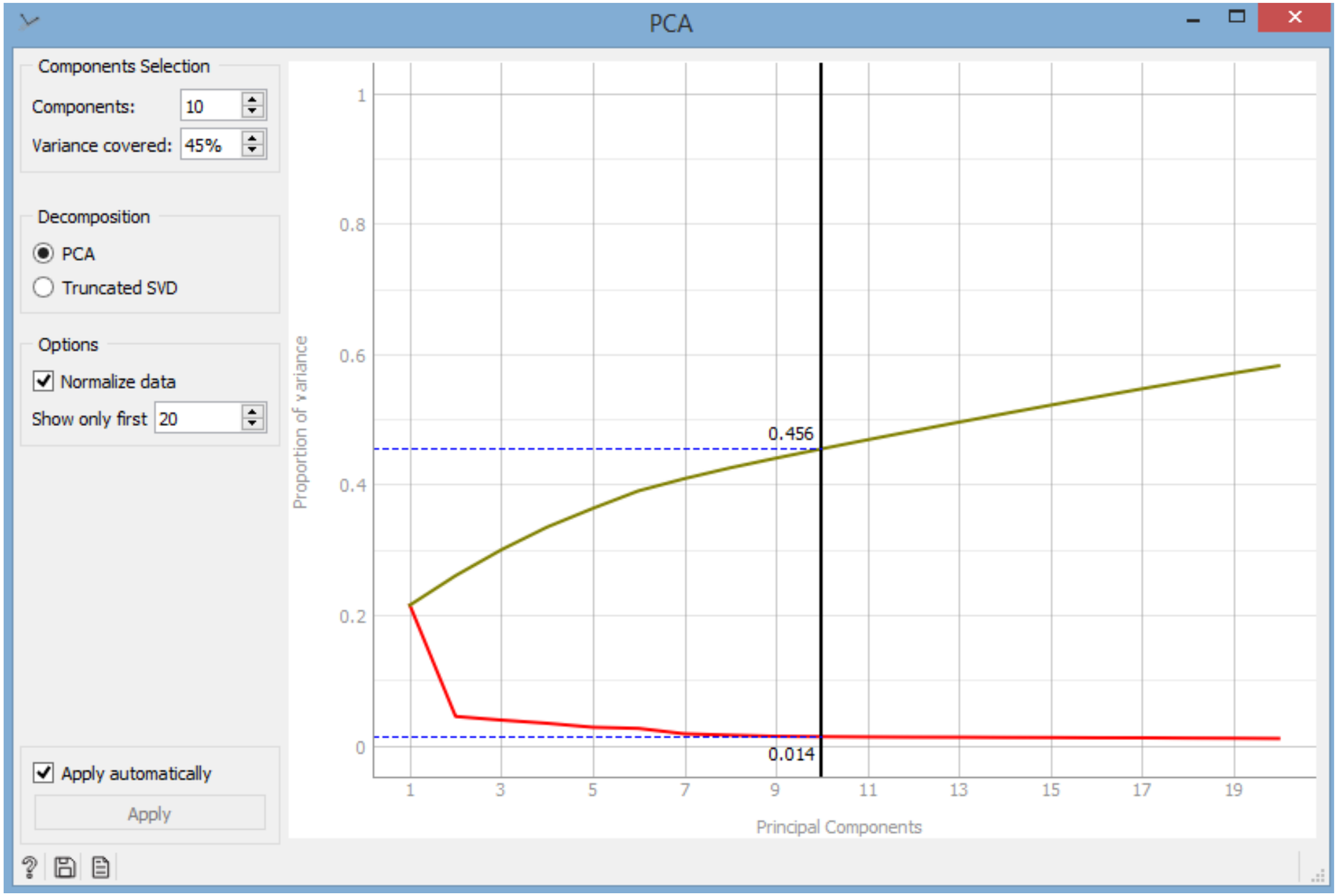
Метод главных компонент (англ. principal component analysis, PCA) — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Изобретён Карлом Пирсоном в 1901 году. Суть модели заключается в преобразовании исходных данных в систему новых координат, в которых новые оси (компоненты), во первых, перпендикулярны друг к другу (то есть нескоррелированы), во вторых, все компоненты ранжированно связаны с общим статистическим разбросом наблюдений, т.е. первая из них объясняет наибольшую долю вариации исходных данных, а в ряду последующих главных компонент доля объясненной дисперсии постоянно снижается. Соответственно можно отбросить часть компонент и таким образом уменьшить число признаков.

Результатом PCA-преобразования матрицы наблюдений **X** является матрица **T** *факторных значений* (или "счетов" scores) размерностью $n \times u$, содержащая проекции исходных точек выборки **X** в новом u-мерном базисе. Другая матрица **P** размерностью ($u \times m$) содержит *факторные нагрузки* (loadings) – коэффициенты пропорциональности, обеспечивающие пересчет данных из m-мерного пространства исходных переменных в u- мерное пространство главных компонент

PCA: ОБЩАЯ СХЕМА РАСЧЕТА



PCA: ЗАДАНИЕ ЧИСЛА КОМПОНЕНТ



PCA: ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ
МАТРИЦА ПРЕОБРАЗОВАННЫХ ДАННЫХ

Data Table (1)

Info

188 instances (no missing values)
10 features (no missing values)
Continuous target variable (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

	HDI	Country	PC1	PC2	PC3	PC4	PC5	PC6
1	0.949	Norway	-8.695	-0.323	-3.672	0.860	-0.031	-2.478
2	0.939	Australia	-6.964	-0.258	-1.927	1.134	-0.003	-1.363
3	0.939	Switzerland	-7.642	-0.483	-3.136	1.429	-0.371	-2.433
4	0.926	Germany	-8.039	0.592	-1.931	0.922	-1.540	-2.112
5	0.925	Denmark	-7.679	-0.666	-2.484	0.620	-1.960	-1.856
6	0.925	Singapore	-6.055	-0.237	-2.445	1.904	4.089	-2.006
7	0.924	Netherlands	-7.444	-0.181	-2.789	0.415	-0.984	-2.225
8	0.923	Ireland	-6.457	-0.723	-1.297	0.494	-0.444	-1.450
9	0.921	Iceland	-7.289	-0.118	-3.733	0.793	-0.005	-0.953
10	0.920	Canada	-7.116	-0.170	-2.300	1.580	-0.583	-1.552
11	0.920	United States	-6.706	3.658	0.212	2.837	0.139	-2.513
12	0.917	Hong Kong	-5.921	-0.353	-1.020	0.513	2.908	-2.197
13	0.915	New Zealand	-5.688	-0.476	-1.442	1.415	0.889	-1.114
14	0.913	Sweden	-8.244	-0.399	-2.592	0.524	-0.598	-1.576
15	0.912	Liechtenstein	-1.006	-0.216	-0.465	0.302	0.118	-0.443
16	0.909	United Kingdom	-7.166	0.192	-1.034	1.037	-1.355	-1.990
17	0.903	Japan	-7.312	1.820	-1.559	0.634	-0.763	-1.927
18	0.901	Korea	-5.922	0.760	-0.514	0.324	1.516	-0.024
19	0.899	Israel	-4.783	-0.842	-1.065	1.642	1.012	-1.526
20	0.898	Luxembourg	-7.579	-0.766	-2.118	0.820	0.379	-3.055
21	0.897	France	-7.438	0.308	-0.913	0.077	-0.559	-1.879
22	0.896	Belgium	-7.363	-0.571	-1.839	-0.039	-0.977	-2.098

PCA: ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ
МАТРИЦА ПЕРЕСЧЕТА

Data Table (2)

Info
10 instances (no missing values)
134 features (no missing values)
No target variable.
1 meta attribute (no missing values)

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

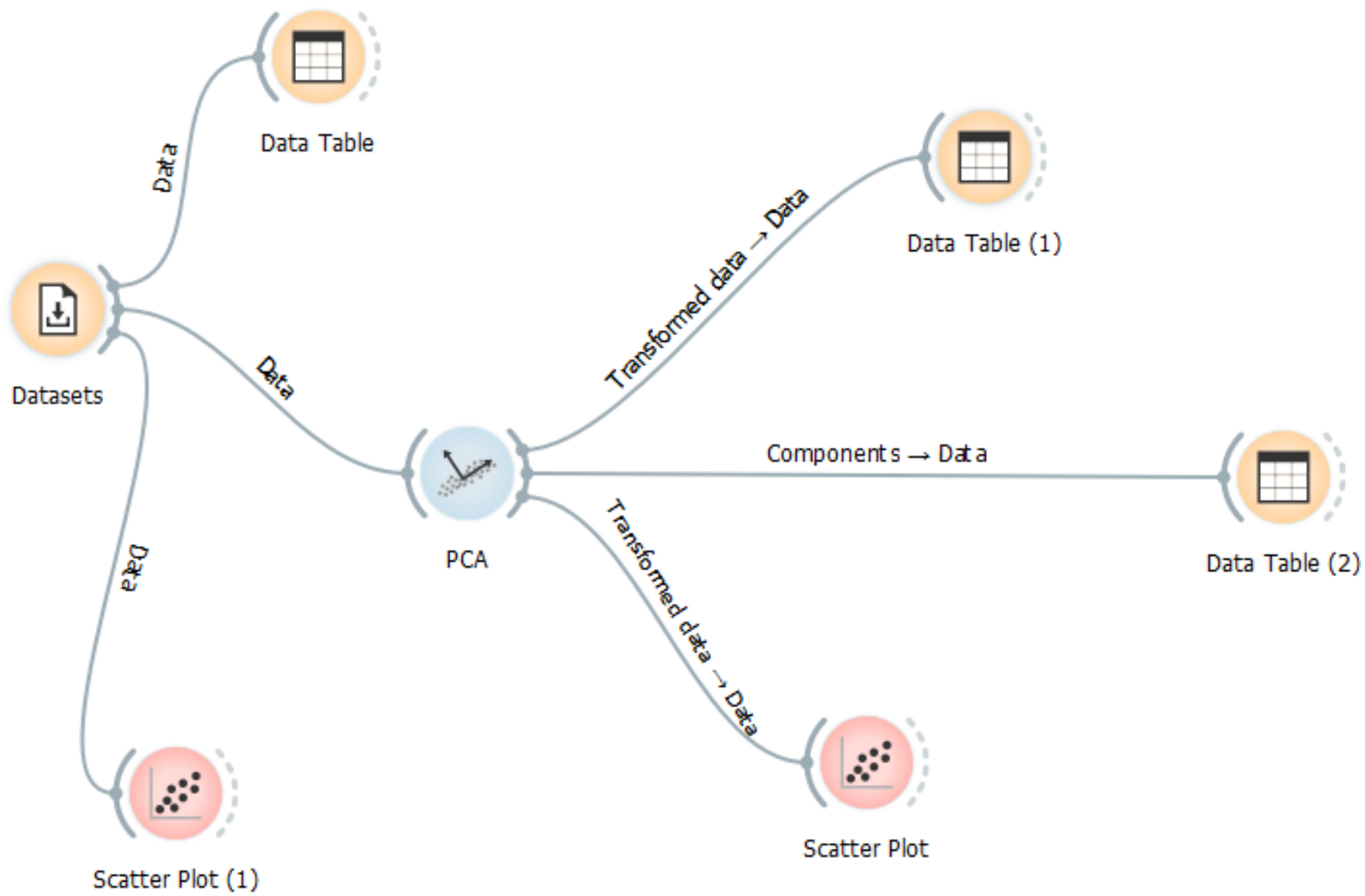
Selection
☐ Select full rows

Restore Original Order

☑ Send Automatically

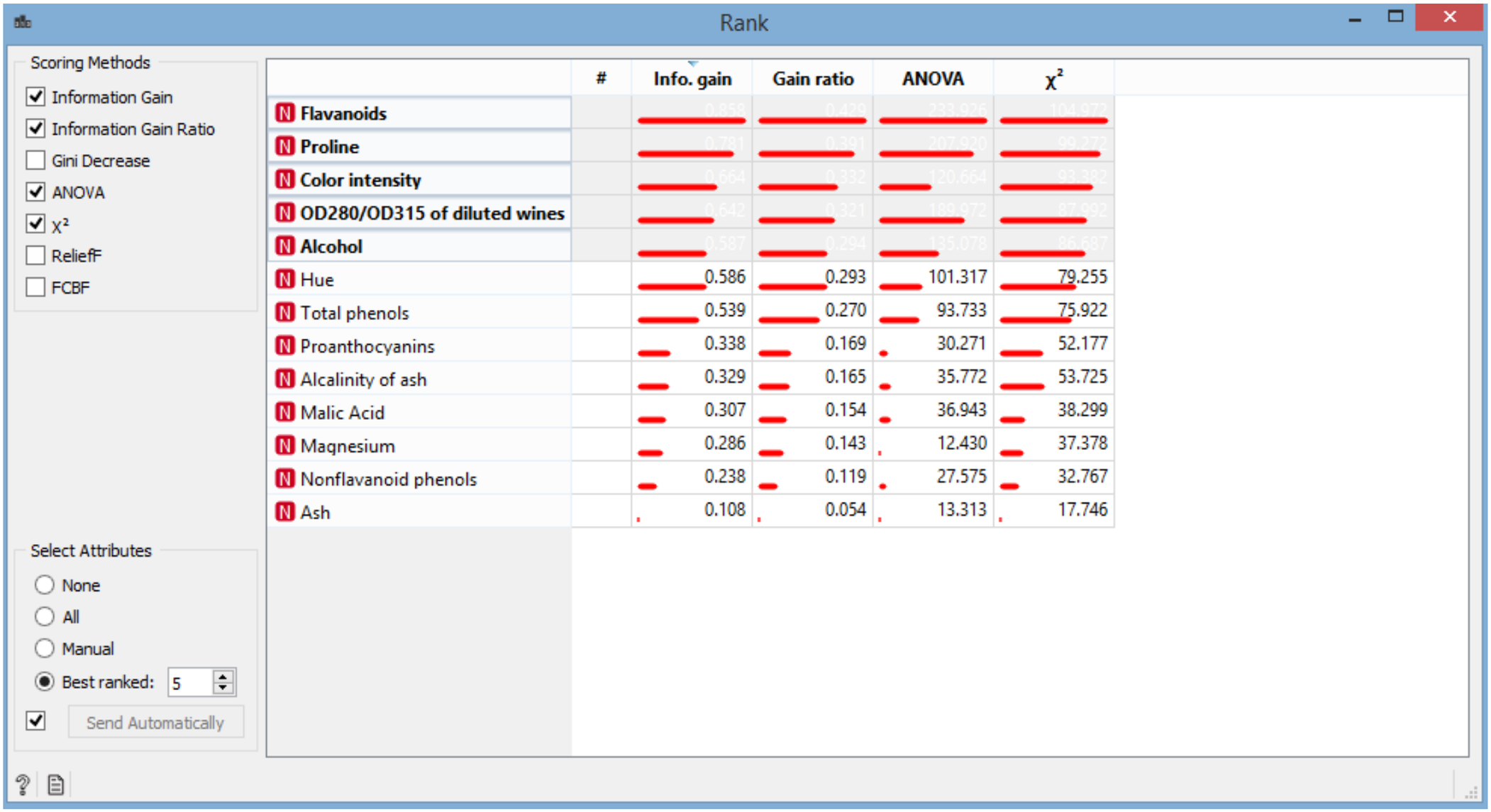
	component	Life expectancy	Mean years of schooling	Gross national income (GNI) per capita	Gender Development Index value	Developme
1	PC1	-0.169	-0.168	-0.121	-0.127	
2	PC2	0.039	-0.030	-0.014	-0.051	
3	PC3	-0.029	0.008	-0.111	-0.012	
4	PC4	-0.038	0.065	0.078	0.207	
5	PC5	0.140	-0.081	0.156	-0.033	
6	PC6	0.052	-0.025	-0.228	0.148	
7	PC7	-0.053	0.054	-0.075	0.066	
8	PC8	0.082	-0.077	-0.094	-0.037	
9	PC9	-0.003	0.054	-0.011	-0.041	
10	PC10	0.021	0.008	-0.052	-0.056	

PCA: EXAMPLE DATASET WINE AND CANSER



RANK – РАНЖИРОВАНИЕ ФИЧ

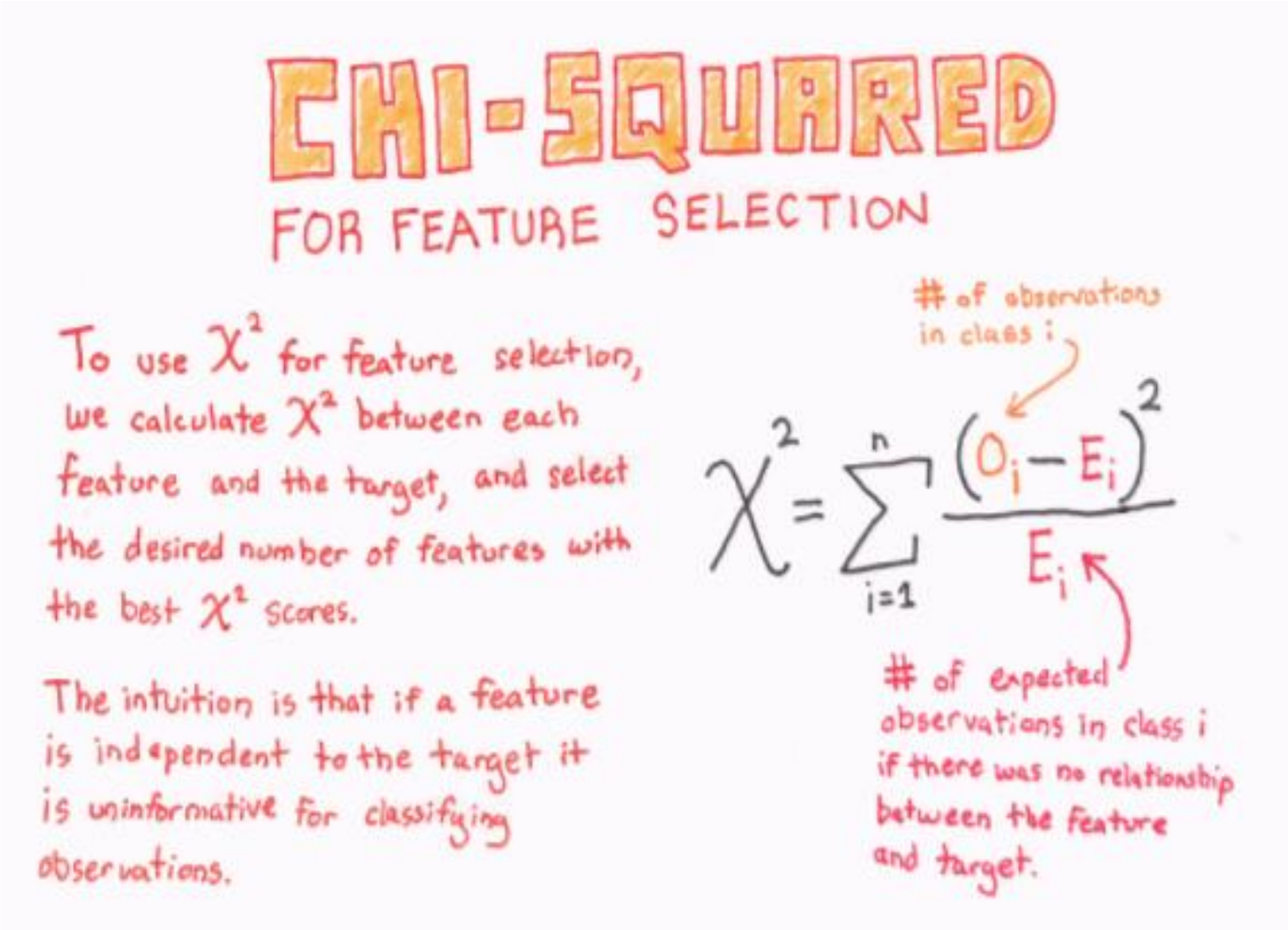
Rank - виджет, кооторый позволяет выбрать наиболее значимые фичи из списка в соответствии с корреляцией с классами (target value).
Выбор осуществляется при помощи нескольких метрик.



	#	Info. gain	Gain ratio	ANOVA	χ^2
Flavanoids		0.858	0.429	233.926	104.973
Proline		0.781	0.391	207.920	99.272
Color intensity		0.664	0.332	120.664	93.382
OD280/OD315 of diluted wines		0.642	0.321	189.972	87.992
Alcohol		0.587	0.294	135.078	86.687
Hue		0.586	0.293	101.317	79.255
Total phenols		0.539	0.270	93.733	75.922
Proanthocyanins		0.338	0.169	30.271	52.177
Alcalinity of ash		0.329	0.165	35.772	53.725
Malic Acid		0.307	0.154	36.943	38.299
Magnesium		0.286	0.143	12.430	37.378
Nonflavanoid phenols		0.238	0.119	27.575	32.767
Ash		0.108	0.054	13.313	17.746

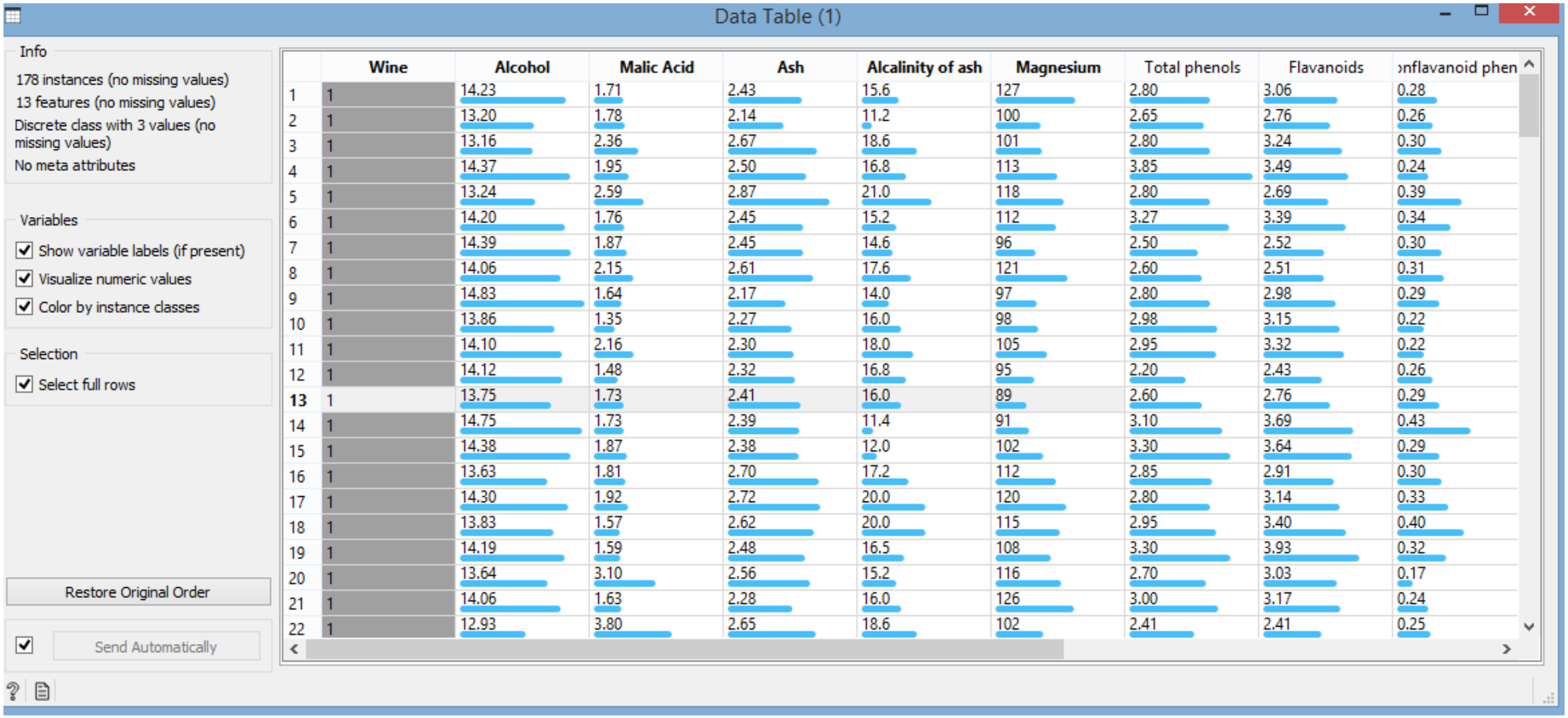
Пусть у нас есть датасет, где одна колонка это значения целевой переменной, а остальные колонки это признаки (фиич). Можно попытаться вычислить насколько целевая переменная зависит от того или иного признака. Как это можно сделать.

Можно рассчитать индивидуальную величину **chi-square statistic,** для разных фич. То есть берем наблюдаемую величину (например, тональную оценку текста или стоимость квартиры) и значение фичи для данной оценки. Считаем **chi-square.** Далее берем следующую фичу и снова считаем статистику.



В итоге перебираем все фичи и получаем набор score для каждой фичи. Сортируем и отбираем нужное количество фич из списка. Выбираем фичи с максимальными значениями Chi2. Основной недостаток заключается в том, что в алгоритме рассматривается каждая фича по отдельности.

RANK – EXAMPLE WINE



	Wine	Alcohol	Malic Acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	onflavanoid phen
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17
21	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24
22	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25

Датасет содержит колонку wine (перечень классов), остальные колонки фичи, характеризующие тип вина. Попробует выбрать какие из фич наиболее значимы в определении классов.

RANK – EXAMPLE WINE

Rank

Scoring Methods

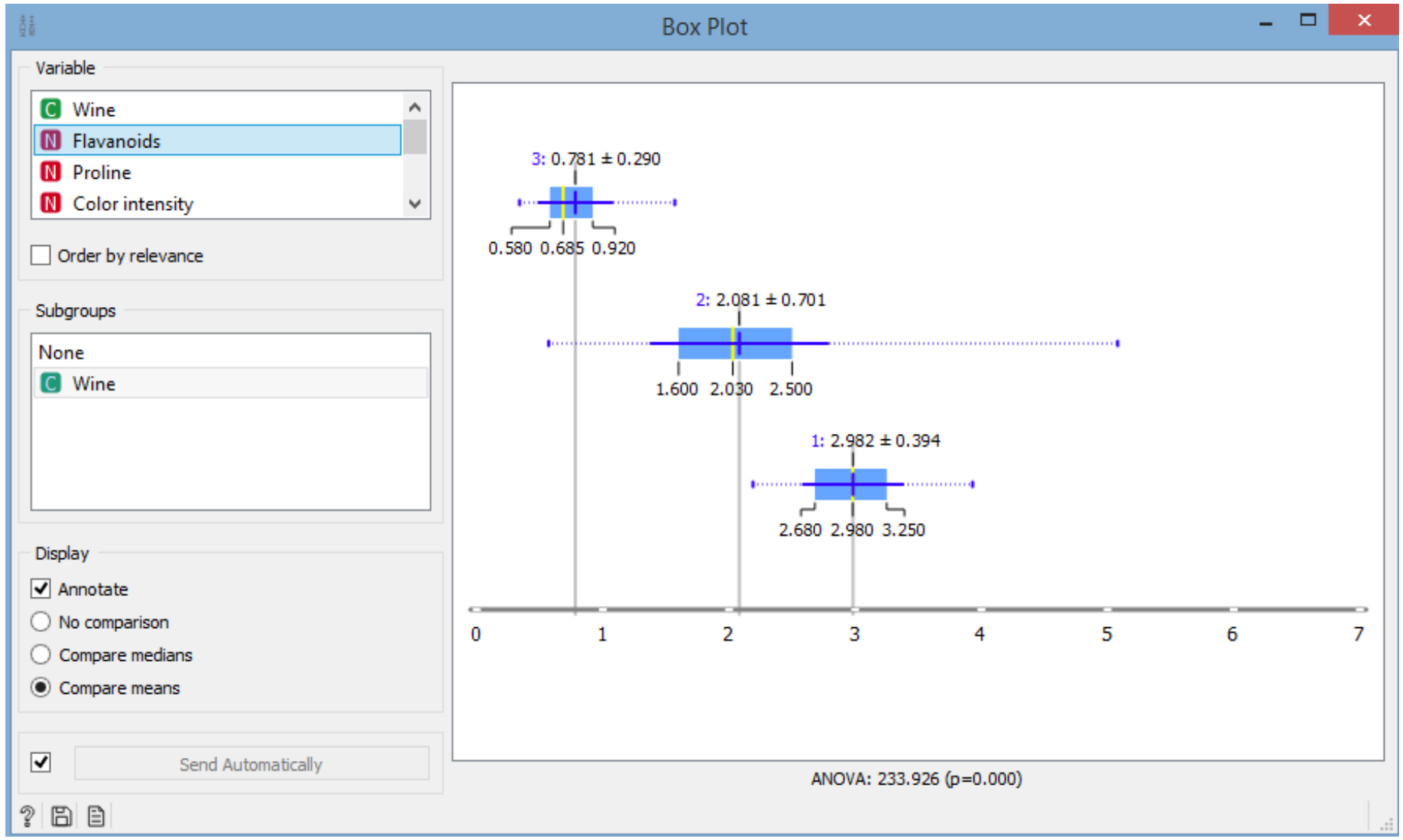
- [x] Information Gain
- [x] Information Gain Ratio
- [ ] Gini Decrease
- [x] ANOVA
- [x] χ^2
- [ ] ReliefF
- [ ] FCBF

	#	Info. gain	Gain ratio	ANOVA	χ^2
Flavanoids		0.858	0.429	233.926	104.972
Proline		0.781	0.391	207.920	99.272
Color intensity		0.664	0.332	120.664	93.382
OD280/OD315 of diluted wines		0.642	0.321	189.972	87.992
Alcohol		0.587	0.294	135.078	86.687
Hue		0.586	0.293	101.317	79.255
Total phenols		0.539	0.270	93.733	75.922
Proanthocyanins		0.338	0.169	30.271	52.177
Alcalinity of ash		0.329	0.165	35.772	53.725
Malic Acid		0.307	0.154	36.943	38.299
Magnesium		0.286	0.143	12.430	37.378
Nonflavanoid phenols		0.238	0.119	27.575	32.767
Ash		0.108	0.054	13.313	17.746

Получилось 5 наиболее значимых фич.

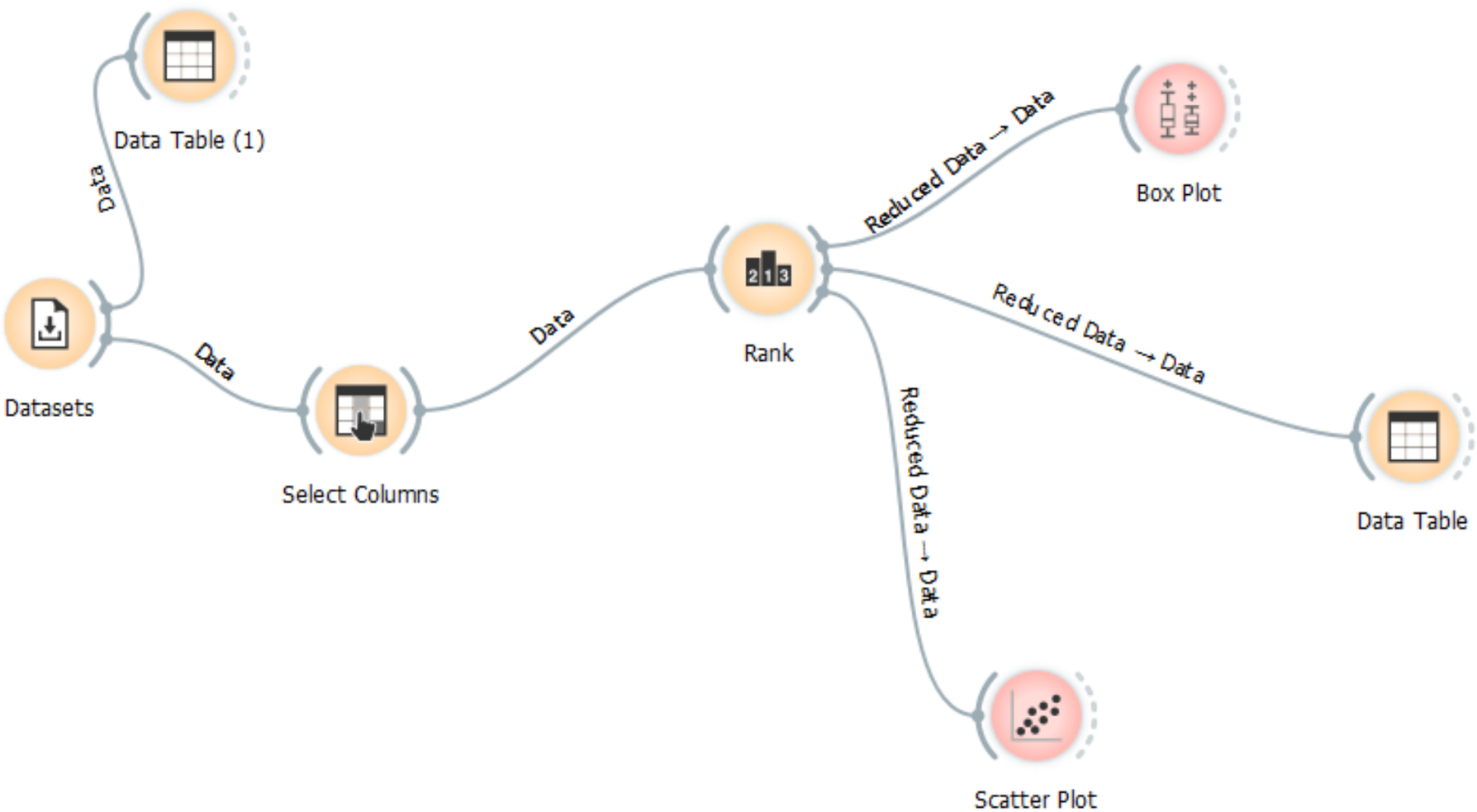
RANK – EXAMPLE WINE - ВИЗУАЛИЗАЦИЯ

Box Plot widget shows the distributions of attribute values

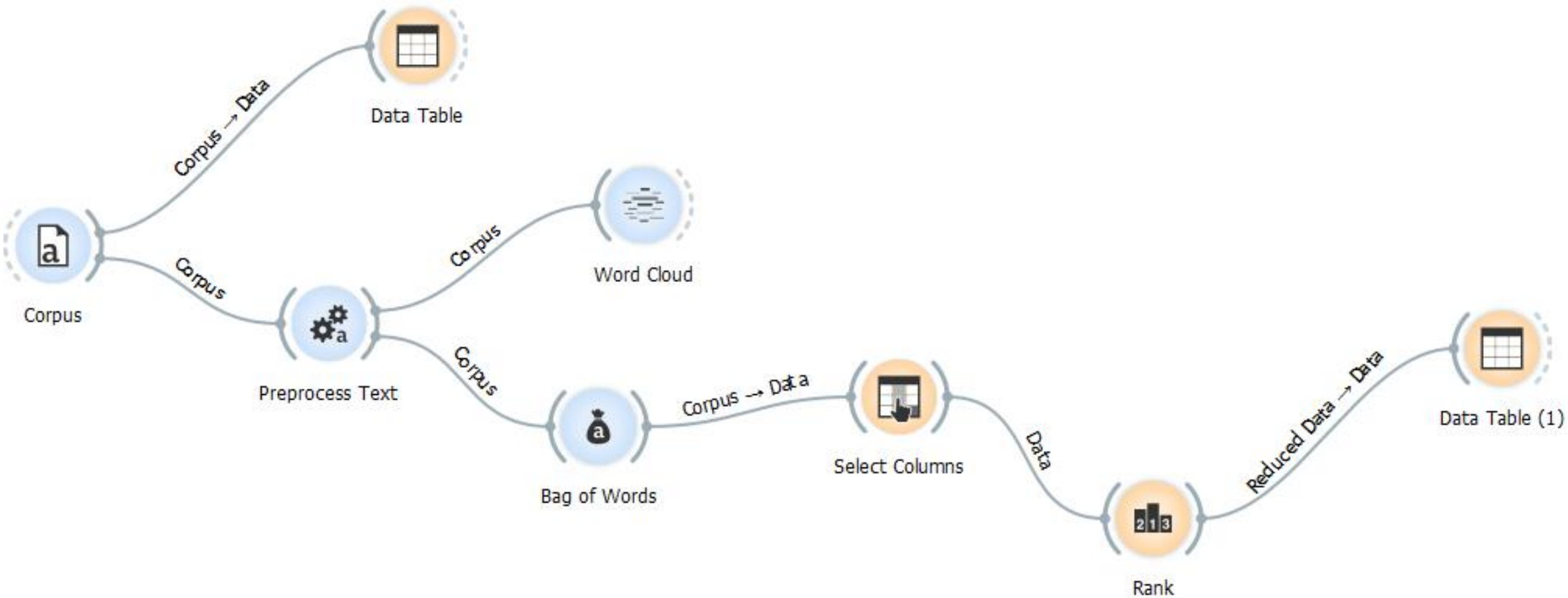


1. The mean (the dark blue vertical line). Толстая голубая линия - стандартное отклонение величины.
2. The median (yellow vertical line)

RANK – EXAMPLE WINE – ОБЩАЯ СХЕМА

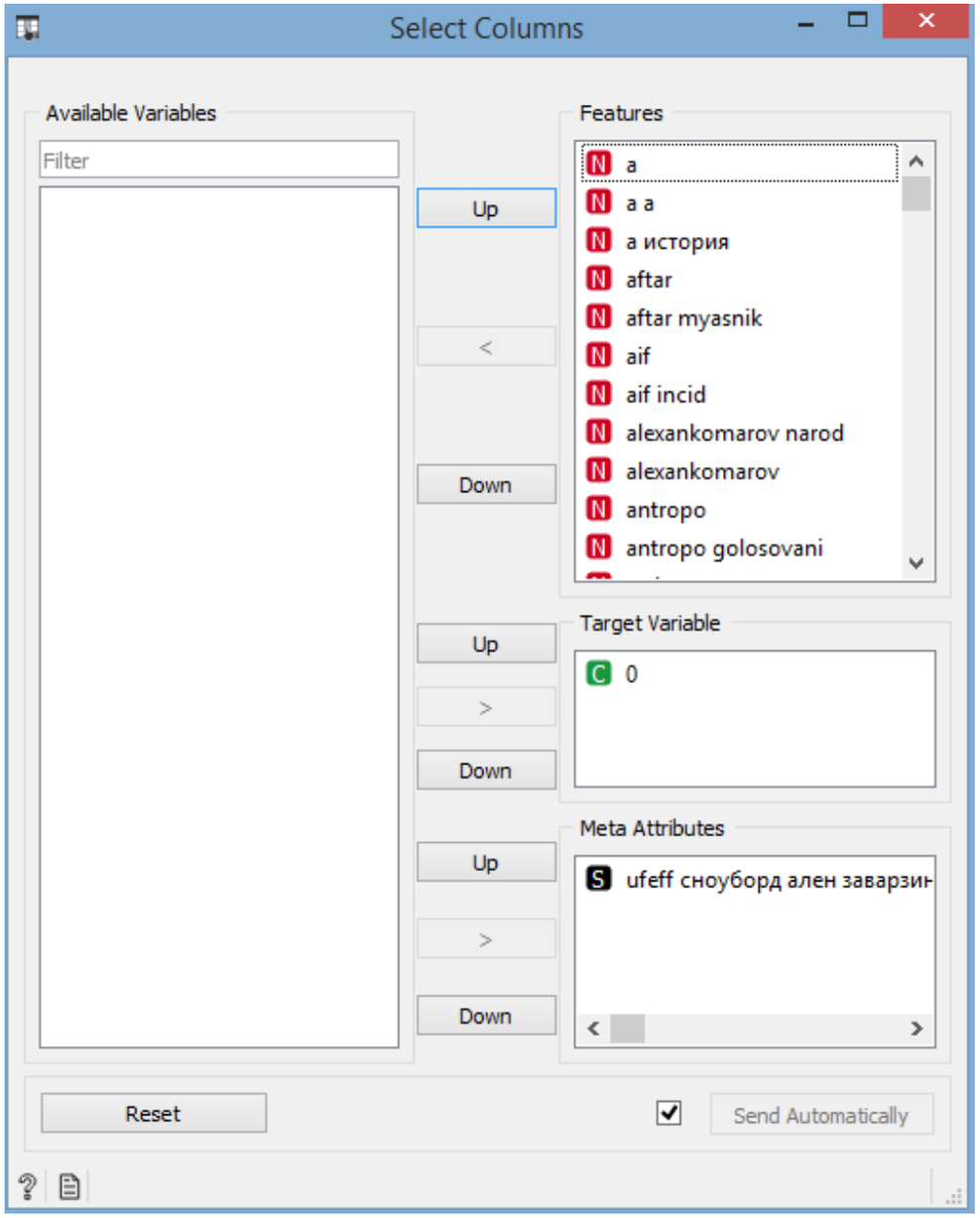


RANK – РАНЖИРОВАНИЕ ФИЧ-СЛОВ
ОБЩАЯ СХЕМА



RANK – РАНЖИРОВАНИЕ ФИЧ-СЛОВ

Выбор колонок в датасете

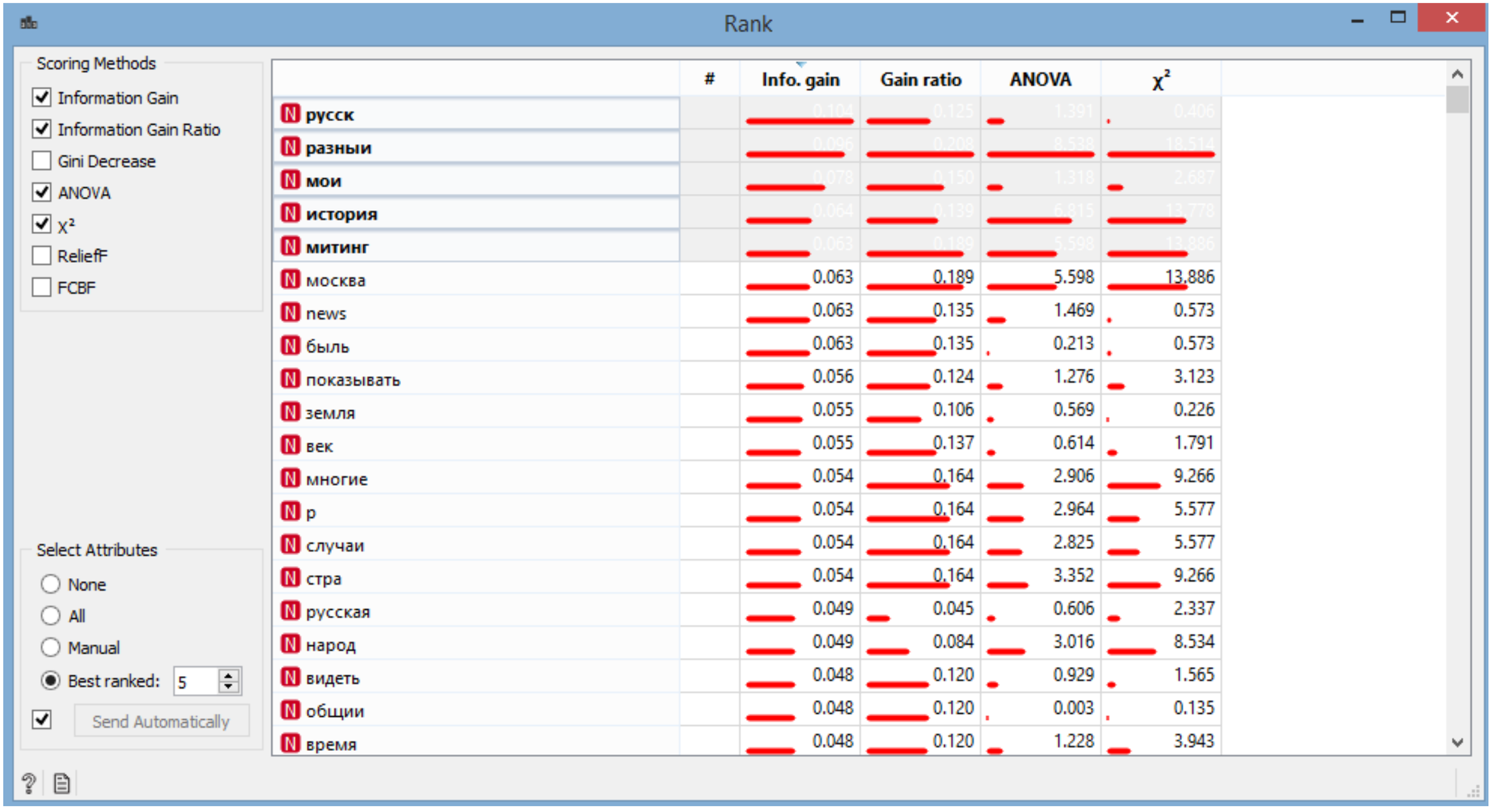


Облако слов

русская
хохол
россия
русские
русск
е
ю
год
хотеть
жить
такои
сторона
общии
blog
news
история
народ
наш
украина
финно
разныи
тех
место
казак
славянин
великии
маленькии
utm
htm
быль
мочь
против
женщина
показывать
человек
земля
info news
группа
язык
идти
крым
запад
ковы
секс
видеть
очень
забывать
татарин
стоить
самыи
мои
семья
душ
сердце
время
оставаться
писать
крым россия
эта
тысяча
мень
скоро
info
styagom htm
aftar myasnik
toptuha
pod sskim
midgard info
toptuha recepti
slavyan
sskim
slavyan pod
белорусия
летнии
митинг кавказ
секс казале
via midgard
styagom
аннексия
таджикскии
казахстан
edineni slavyan
финно угорскии
советскии
украинец
территория
цыган
самыи красивыи
интервью
karaimov aftar
путин
edineni
киев
многие
казале
recepti kibinay
midgard
news nasha
мономах
западныи
nasha cel
кровь
trakayskih
nasha
kibinay trakayskih
sskim styagom
аннексия крым
karaimov
cel edineni
интервью боец
trakayskih karaimov
немцаи
пусть
kibinay
угорскии
москва
жид
боец
дякст
хохол оставаться
нагрелать
кипящыи
северо
myasnik
полуж
случаи
суд
еврона
весьма
настоящии
часть
aftar
век
pod
плохо
знать
германа
сундук
будь
посмотреть
завещать
съедать
блядь
немецкии
баже
хлеб
руся
сегодня
жена
состав
сколькиим
город
происходить
представлять
красивым
стоить
сторона
жена
жена
жена

RANK – РАНЖИРОВАНИЕ ФИЧ-СЛОВ

Результат ранжирования



	#	Info. gain	Gain ratio	ANOVA	χ^2
русск		0.104	0.125	1.391	0.406
разныи		0.096	0.208	8.538	18.514
мои		0.078	0.150	1.318	2.687
история		0.064	0.139	6.815	13.778
митинг		0.063	0.189	5.598	13.886
москва		0.063	0.189	5.598	13.886
news		0.063	0.135	1.469	0.573
быль		0.063	0.135	0.213	0.573
показывать		0.056	0.124	1.276	3.123
земля		0.055	0.106	0.569	0.226
век		0.055	0.137	0.614	1.791
многие		0.054	0.164	2.906	9.266
р		0.054	0.164	2.964	5.577
случаи		0.054	0.164	2.825	5.577
стра		0.054	0.164	3.352	9.266
русская		0.049	0.045	0.606	2.337
народ		0.049	0.084	3.016	8.534
видеть		0.048	0.120	0.929	1.565
общии		0.048	0.120	0.003	0.135
время		0.048	0.120	1.228	3.943

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru