НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

CLASSIFICATION IN ORANGE

1. KNN
2. SVM
3. LOGISTIC REGRESSION

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 1.10.2018

Задача классификации

Задача классификации в машинном обучении — это задача отнесения объекта к одному из заранее определенных классов на основании его формализованных признаков. Каждый из объектов в этой задаче представляется в виде вектора в N-мерном пространстве, каждое измерение в котором представляет собой описание одного из признаков объекта.

Для **обучения классификатора** необходимо иметь набор объектов, для которых заранее определены классы. Это множество называется обучающей выборкой, её разметка производится вручную, с привлечением специалистов в исследуемой области.

Например, у нас есть набор текстов, и у каждого текста есть оценка тональности. Алгоритм классификации может обучится на этих текстах, и в дальнейшем, обученный алгоритм можно использовать для другого набора текстов. В этом случае, многомерное пространство признаков представляет собой матрица частот слов в текстах.

Другой пример, предположим есть таблица пациентов, с медицинскими показателями (виды болей, различные анализы) и диагноз, который был подтвержден. В этом случае можно обучить алгоритм распознавать диагноз у вновь поступивших пациентов.

ТИПЫ КЛАССИФИКАТОРОВ

Типичная задача статистического обучения – есть набор объектов с наблюдаемыми свойствами, и не наблюдаемыми свойствами. Нужно построить алгоритм, который бы позволял вычислить ненаблюдаемые свойства при помощи наблюдаемых, при этом хотелось бы что бы алгоритм ошибался не очень часто и не очень сильно.

Классификаторы основанные на таблице частот.

1. ZeroR (алгоритм строит таблицу частот и выбирает максимальную частоту).
2. OneR (Алгоритм строит таблицу частот и строит одно правило для каждой класса. Выбирает правило, которое дает минимальную ошибку. Это правило применяется для всего датасета)
3. **Naive Bayesian**
4. Decision Tree (Алгоритм разбивает датасет на все меньшие куски данных, формируя тем самым дерево).

Классификаторы основанные на ковариационной матрице

1. Линейный дискриминационный анализ (Linear Discriminant Analysis)
2. **Логистическая регрессия (Logistic Regression)**

Классификатор основанный на функции сходства.

1. **Метод ближайших соседей (K Nearest Neighbors)**

Другие

1. Нейронные сети.
2. **Метод опорных векторов (Support Vector Machine)**

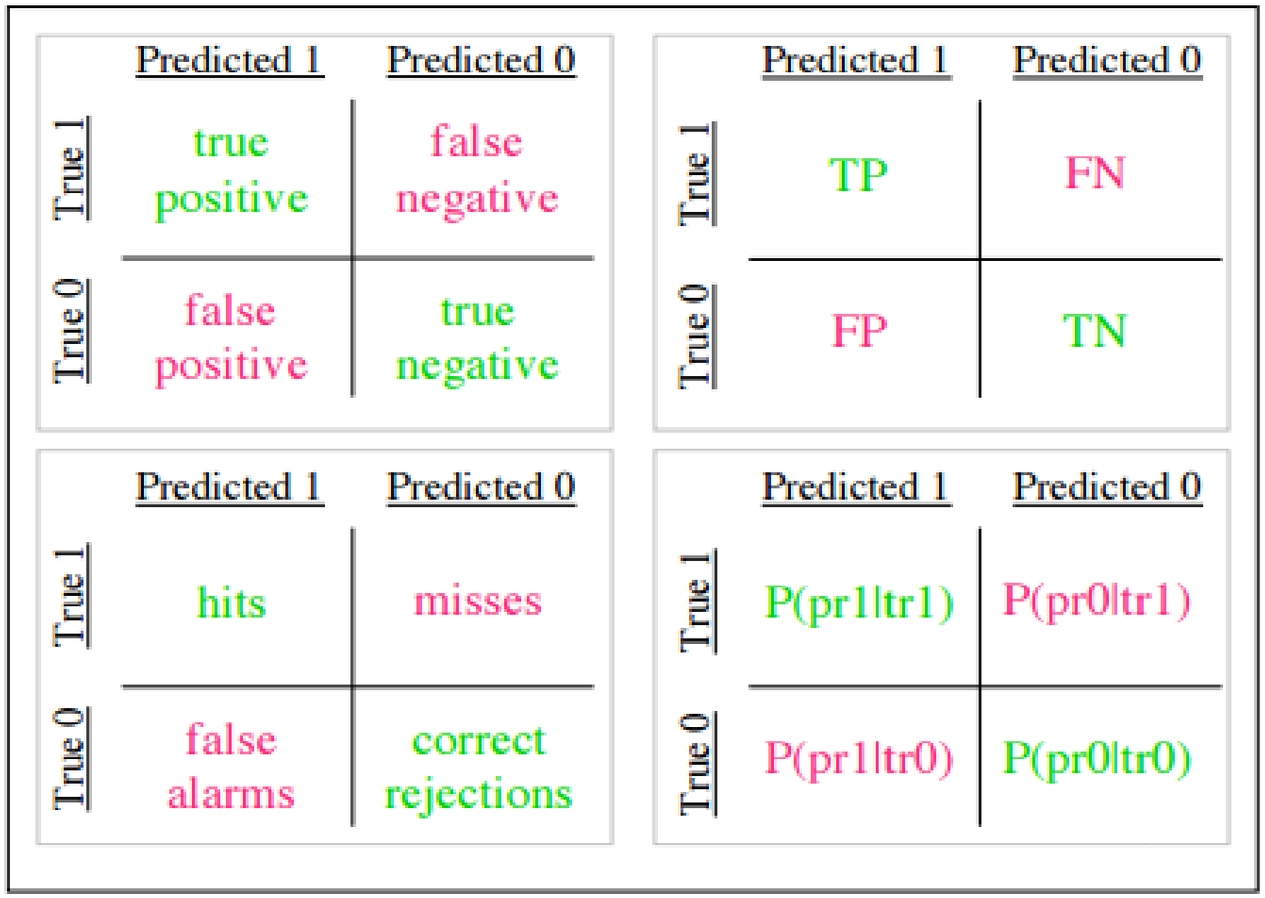
Меры качества классификаторов для бинарных классов

$$\text{Precision} = \frac{\text{TP}}{TP + FP}$$

Precision: число правильно предсказанных положительных значений поделенных на число предсказанных классификатором положительных значений.

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

Recall : число правильно предсказанных положительных значений поделенных на число положительных ответов в данных.



Confusion Matrix

TP – число правильно предсказанных положительных значений

FN – число неправильно предсказанных положительных значений

FN – число правильно предсказанных негативных значений

FP – число неправильно предсказанных негативных значений

Меры качества классификаторов для многих классов

Precision = $\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fp_i)}$

Recall = $\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)}$

F measure = $\frac{(\beta^2 + 1) Precision_\mu Recall_\mu}{\beta^2 Precision_\mu + Recall_\mu}$

	OBTAINED CLASSES				
TRUE CLASSES	-2	-1	0	1	2
-2	5	5	5	5	5
-1	5	5	5	5	5
0	5	5	5	5	5
1	5	5	5	5	5
2	5	5	5	5	5

Class	-2	-1	0	1	2	mean
Tp	5	5	5	5	5	
FP	=B9+B10+B11+B12			20	20	
Fn	20	20	20	20	20	
Precision	0,20	0,20	0,20	0,20	0,20	0,40
Recall	0,20	0,20	0,20	0,20	0,20	0,20
F – measu	0,20	0,20	0,20	0,20	0,20	0,27

	OBTAINED CLASSES				
TRUE CLASSES	-2	-1	0	1	2
-2	5	5	5	5	5
-1	5	5	5	5	5
0	5	5	5	5	5
1	5	5	5	5	5
2	5	5	5	5	5

Class	-2	-1	0	1	2	mean
Tp	5	5	5	5	5	
FP	20	20	=D8+D9+D11+D12			
Fn	20	20	20	20	20	
Precision	0,20	0,20	0,20	0,20	0,20	0,40
Recall	0,20	0,20	0,20	0,20	0,20	0,20
F – measu	0,20	0,20	0,20	0,20	0,20	0,27

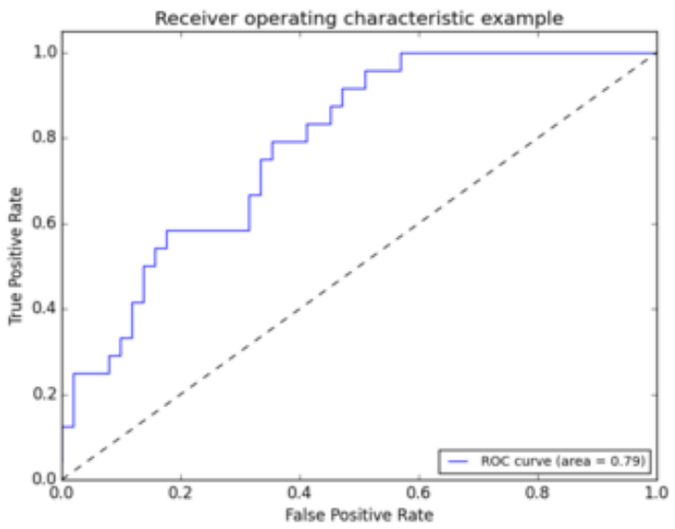
МЕРЫ КАЧЕСТВА КЛАССИФИКАТОРОВ ДЛЯ БИНАРНЫХ КЛАССОВ

F − measure

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \quad \Longleftrightarrow \quad \frac{(\beta^2+1)tp}{(\beta^2+1)tp+\beta^2 fn+fp}$$

β– обычно берут равной 1. F measure = 2 * (precision * recall) / (precision + recall)

The F measure (F1, Fscore) можно интерпретировать как взвешенное среднее precision и recall. Если F1=1, то классификатор отработал на 100% и F1=0 тогда классификатор не справился с задачей.



$$ROC = \frac{P(x|positive)}{P(x|negative)}$$

Рассчитывает отношение числа правильно распознанных случаев к числу не правильных. Процесс расчета таков: берутся данные, последовательно, и в них вычисляется это отношение. В какой то момент отношение становится константой.

AUC – интеграл под кривой.

ОБЗОР ВИДЖЕТОВ

Виджет для работы с методом ближайших соседей (K Nearest Neighbors)

Виджет для работы с методом опорных векторов (SVN)

Виджет для работы с моделью Naïve Bayes

Виджет для работы с моделью логистическая регрессия

ОБЗОР ВИДЖЕТОВ

Виджет для расчета метрик качества

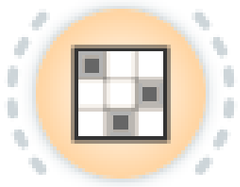
Виджет для расчета для расчета новых значений (предсказание)

Виджет для расчета для расчета confusion matrix

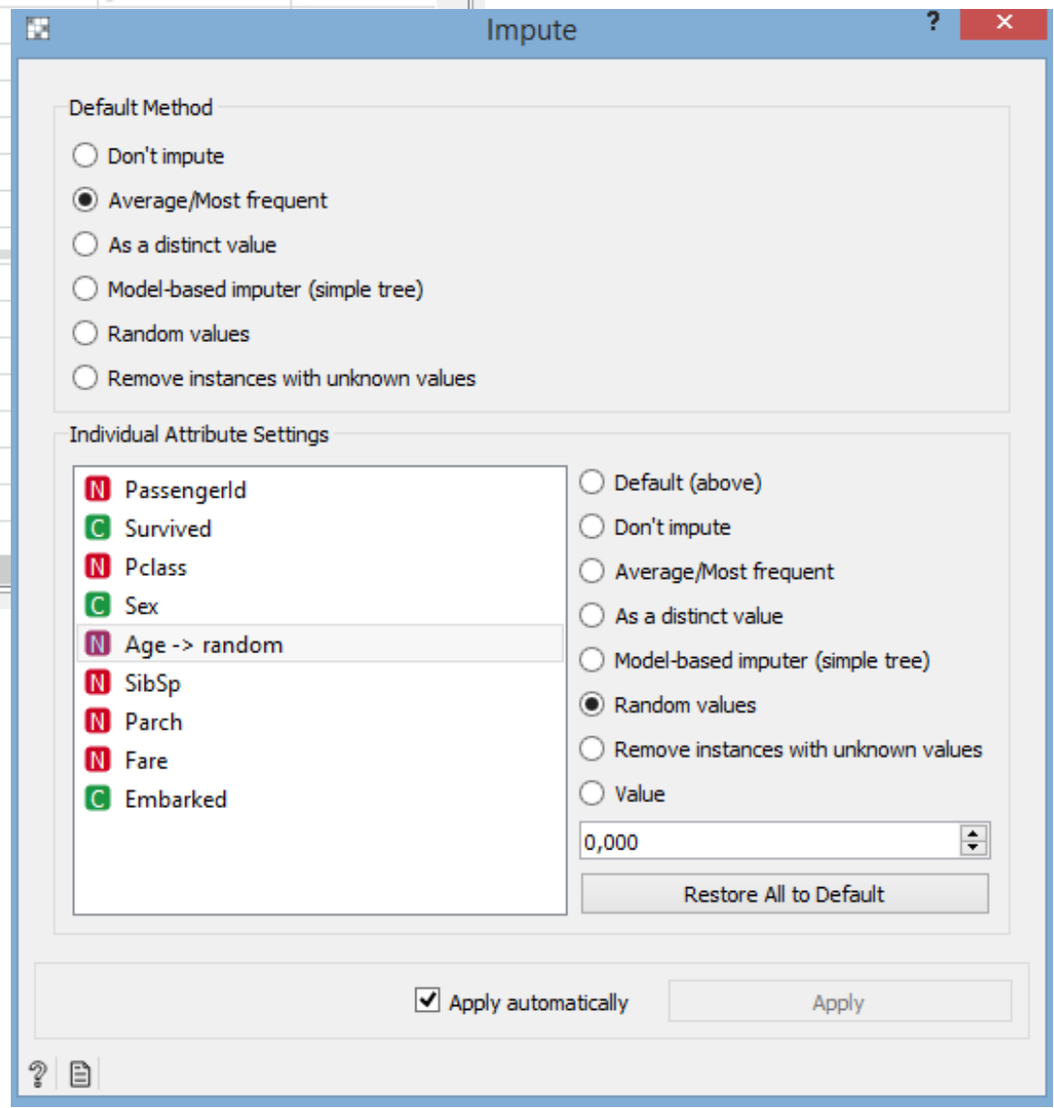
Виджет для расчета для расчета ROC

ОБРАБОТКА ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

	in	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarke
1		1.0	0.0	3.0	1.0	22.00	1.0	0.0	45839.0000	S
2		2.0	1.0	1.0	0.0	38.00	1.0	0.0	71.2833	C
3		3.0	1.0	3.0	0.0	26.00	0.0	0.0	7.9250	S
4		4.0	1.0	1.0	0.0	35.00	1.0	0.0	53.1000	S
5		5.0	0.0	3.0	1.0	35.00	0.0	0.0	43228.0000	S
6		6.0	0.0	3.0	1.0	?	0.0	0.0	30529.0000	Q
7		7.0	0.0	1.0	1.0	54.00	0.0	0.0	51.8625	S
8		8.0	0.0	3.0	1.0	2.00	3.0	1.0		
9		9.0	1.0	3.0	0.0	27.00	0.0	2.0		
10		10.0	1.0	2.0	0.0	14.00	1.0	0.0		
11		11.0	1.0	3.0	0.0	4.00	1.0	1.0		
12		12.0	1.0	1.0	0.0	58.00	0.0	0.0		
13		13.0	0.0	3.0	1.0	20.00	0.0	0.0		
14		14.0	0.0	3.0	1.0	39.00	1.0	5.0		
15		15.0	0.0	3.0	0.0	14.00	0.0	0.0		
16		16.0	1.0	2.0	0.0	55.00	0.0	0.0		
17		17.0	0.0	3.0	1.0	2.00	4.0	1.0		
18		18.0	1.0	2.0	1.0	?	0.0	0.0		
19		19.0	0.0	3.0	0.0	31.00	1.0	0.0		
20		20.0	1.0	3.0	0.0	?	0.0	0.0		
21		21.0	0.0	2.0	1.0	35.00	0.0	0.0		
22		22.0	1.0	2.0	1.0	34.00	0.0	0.0		

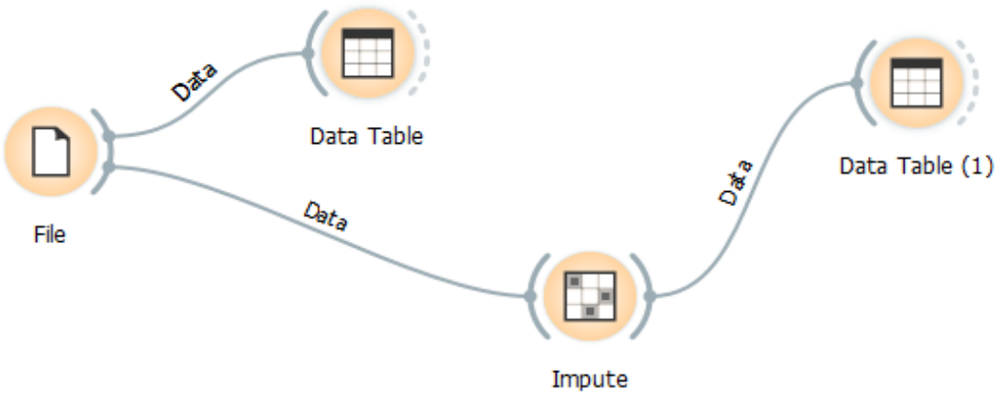


Виджет для заполнения пропущенных значений



ОБРАБОТКА ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

	in	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarke
1		1.0	0.0	3.0	1.0	22.00	1.0	0.0	45839.0000	S
2		2.0	1.0	1.0	0.0	38.00	1.0	0.0	71.2833	C
3		3.0	1.0	3.0	0.0	26.00	0.0	0.0	7.9250	S
4		4.0	1.0	1.0	0.0	35.00	1.0	0.0	53.1000	S
5		5.0	0.0	3.0	1.0	35.00	0.0	0.0	43228.0000	S
6		6.0	0.0	3.0	1.0	?	0.0	0.0	30529.0000	Q
7		7.0	0.0	1.0	1.0	54.00	0.0	0.0	51.8625	S
8		8.0	0.0	3.0	1.0	2.00	3.0	1.0	21.0750	S
9		9.0	1.0	3.0	0.0	27.00	0.0	2.0	11.1333	S
10		10.0	1.0	2.0	0.0	14.00	1.0	0.0	30.0708	C
11		11.0	1.0	3.0	0.0	4.00	1.0	1.0	43297.0000	S
12		12.0	1.0	1.0	0.0	58.00	0.0	0.0	26.5500	S
13		13.0	0.0	3.0	1.0	20.00	0.0	0.0	43228.0000	S
14		14.0	0.0	3.0	1.0	39.00	1.0	5.0	31.2750	S
15		15.0	0.0	3.0	0.0	14.00	0.0	0.0	15523.0000	S
16		16.0	1.0	2.0	0.0	55.00	0.0	0.0	16.0000	S
17		17.0	0.0	3.0	1.0	2.00	4.0	1.0	29.1250	Q
18		18.0	1.0	2.0	1.0	?	0.0	0.0	13.0000	S
19		19.0	0.0	3.0	0.0	31.00	1.0	0.0	18.0000	
20		20.0	1.0	3.0	0.0	?	0.0	0.0	7.2250	
21		21.0	0.0	2.0	1.0	35.00	0.0	0.0	26.0000	
22		22.0	1.0	2.0	1.0	34.00	0.0	0.0	13.0000	



Sex	Age	SibSp
1.0	22.00	1.0
0.0	38.00	1.0
0.0	26.00	0.0
0.0	35.00	1.0
1.0	35.00	0.0
1.0	36.00	
1.0	54.00	
1.0	2.00	3.0
0.0	27.00	0.0
0.0	14.00	1.0
0.0	4.00	1.0
0.0	58.00	0.0
1.0	20.00	0.0
1.0	39.00	1.0
0.0	14.00	0.0
0.0	55.00	0.0

МЕТОД БЛИЖАЙШИХ СОСЕДЕЙ (K NEAREST NEIGHBORS)

Все объекты расположены в многомерном пространстве

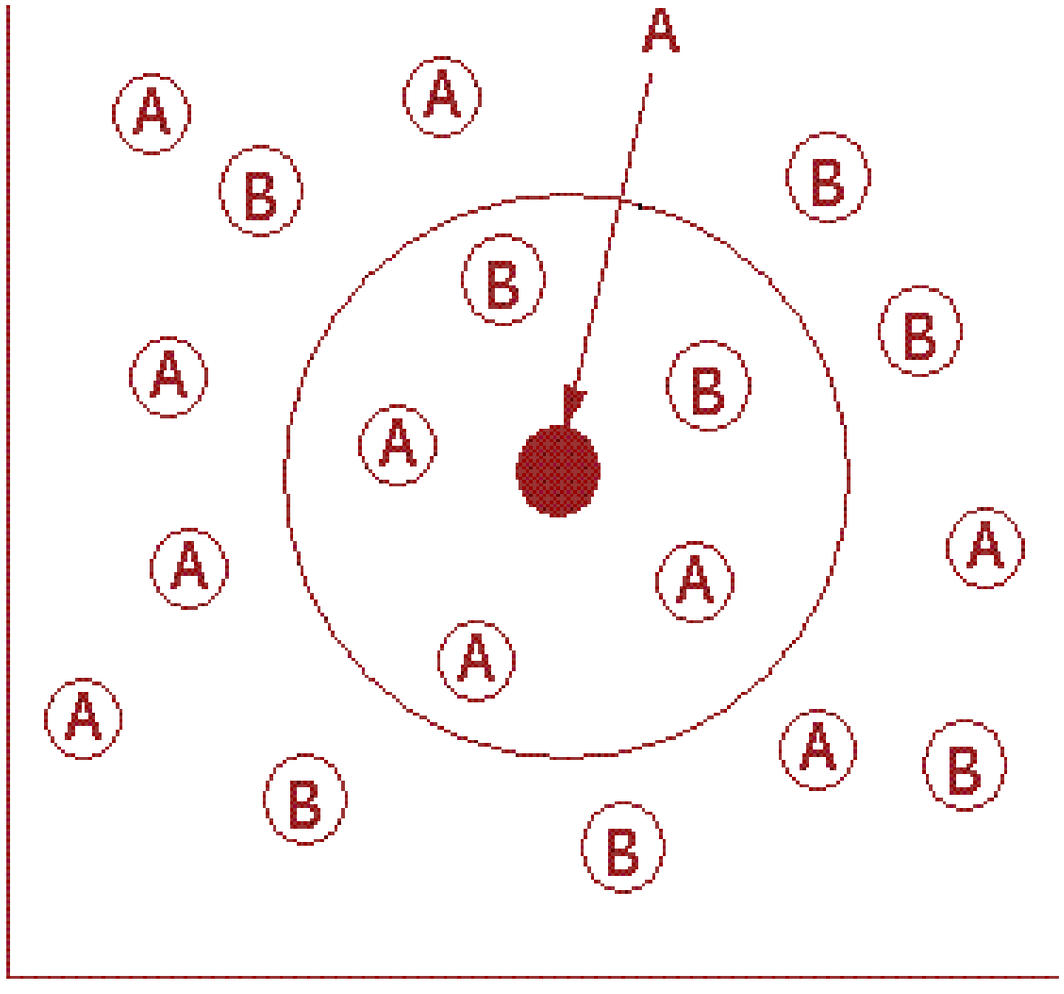
1. Задаем число *k* – количество ближайших соседей.

2. Ищем *k* объектов с минимальным расстоянием до нашего нового объекта. Используем меру для расчета расстояний.

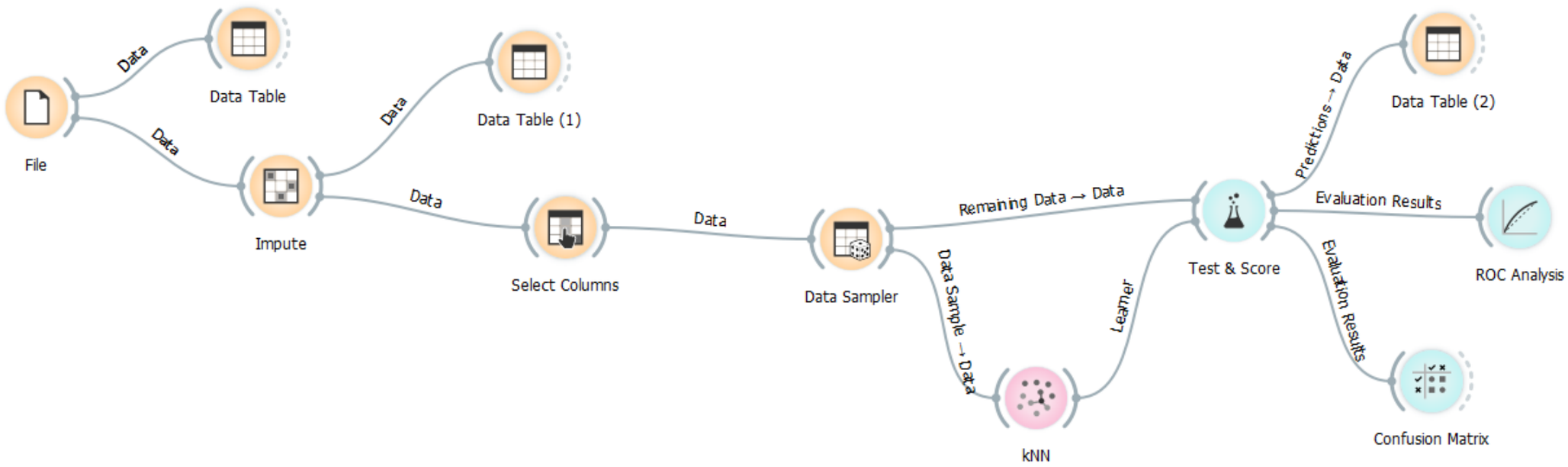
3.1 Простое невзвешенное голосование. Считаем сколько объектов с классами присутствует внутри заданного расстояния. Например, если число объектов с классом А большинство, то новый объект относится к классу А.

3.2. Взвешенное голосование

В такой ситуации учитывается также и расстояние до новой записи. Чем меньше расстояние, тем более значимый вклад вносит голос.



КЛАССИФИКАЦИЯ ТАБЛИЧНЫХ ДАННЫХ



	kNN	kNN (0.0)	kNN (1.0)	Fold	PassengerId	Age	Pclass	Sex	SibSp
870		0.600	0.400	10	362.0	29.00	2.0	1.0	1.0
871		0.800	0.200	10	839.0	32.00	3.0	1.0	0.0
873		0.800	0.200	10	526.0	40.50	3.0	1.0	0.0
874		0.800	0.200	10	574.0	2.00	3.0	0.0	0.0
875		0.600	0.400	10	493.0	55.00	1.0	1.0	0.0
877		0.800	0.200	10	268.0	25.00	3.0	1.0	1.0
878		0.800	0.200	10	198.0	42.00	3.0	1.0	0.0
879		1.000	0.000	10	822.0	27.00	3.0	1.0	0.0
880		1.000	0.000	10	818.0	31.00	2.0	1.0	1.0
881		0.800	0.200	10	24.0	28.00	1.0	1.0	0.0
882		0.600	0.400	10	708.0	42.00	1.0	1.0	0.0
883		0.600	0.400	10	234.0	5.00	3.0	0.0	4.0
886		0.800	0.200	10	859.0	24.00	3.0	0.0	0.0

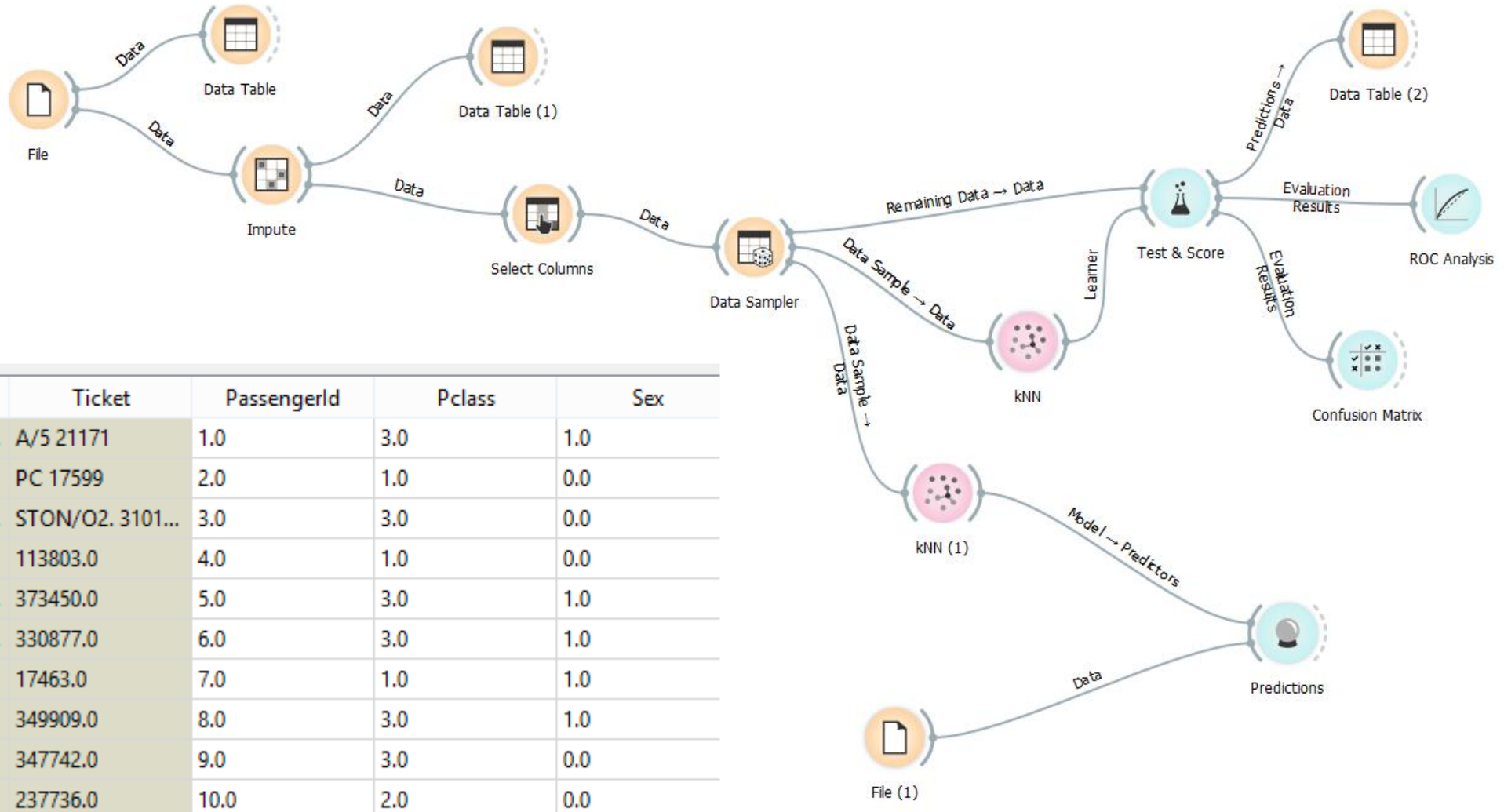
Predicted

Actual	0.0	1.0	Σ
0.0	422	128	550
1.0	235	125	360
Σ	657	253	910

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.591	0.601	0.584	0.584	0.601

ПРЕДСКАЗАНИЯ С ИСПОЛЬЗОВАНИЕМ КЛАССИФИКАТОРА KNN



	kNN	Name	Ticket	PassengerId	Pclass	Sex
1	1.00 : 0.00 → 0.0	Braund, Mr. Ow...	A/5 21171	1.0	3.0	1.0
2	0.60 : 0.40 → 0.0	Cumings, Mrs. ...	PC 17599	2.0	1.0	0.0
3	0.00 : 1.00 → 1.0	Heikkinen, Miss...	STON/O2. 3101...	3.0	3.0	0.0
4	0.40 : 0.60 → 1.0	Futrelle, Mrs. Ja...	113803.0	4.0	1.0	0.0
5	1.00 : 0.00 → 0.0	Allen, Mr. Willia...	373450.0	5.0	3.0	1.0
6	0.60 : 0.40 → 0.0	Moran, Mr. Jam...	330877.0	6.0	3.0	1.0
7	0.60 : 0.40 → 0.0	McCarthy, Mr. ...	17463.0	7.0	1.0	1.0
8	0.60 : 0.40 → 0.0	Palsson, Master...	349909.0	8.0	3.0	1.0
9	0.20 : 0.80 → 1.0	Johnson, Mrs. ...	347742.0	9.0	3.0	0.0
10	0.40 : 0.60 → 1.0	Nasser, Mrs. Ni...	237736.0	10.0	2.0	0.0
11	1.00 : 0.00 → 0.0	Sandstrom, Mis...	PP 9549	11.0	3.0	0.0
12	0.60 : 0.40 → 0.0	Bonnell, Miss. E...	113783.0	12.0	1.0	0.0

ПРЕДСКАЗАНИЯ АКЦИЙ ГАЗПРОМА

<DATE>	<TIME>	<OPEN>	<HIGH>	<LOW>	<CLOSE>	difference
20160104	110000	135.89	136.65	135.62	135.96	0.07
20160104	120000	135.92	135.92	134.42	134.67	-1.25
20160104	130000	134.71	135.24	134.42	135.15	0.44
20160104	140000	135.12	135.22	134.7	135.22	0.1
20160104	150000	135.23	135.49	134.88	135.11	-0.12
20160104	160000	135.11	135.35	134.94	134.97	-0.14
20160104	170000	134.97	135.3	134.8	135.13	0.16
20160104	180000	135.13	135.4	134.96	135.15	0.02
20160104	190000	135.14	135.45	134.9	134.91	-0.23
20160105	110000	134.85	135.35	134.85	135.06	0.21
20160105	120000	135.06	136.16	135	135.93	0.87
20160105	130000	135.92	135.94	135.1	135.31	-0.61
20160105	140000	135.3	135.3	134.61	135.07	-0.23
20160105	150000	135.08	135.23	134.76	135.22	0.14
20160105	160000	135.21	135.75	135.01	135.64	0.43
20160105	170000	135.64	137.28	135.53	136.74	1.1
20160105	180000	136.71	137.15	136.18	136.19	-0.52
20160105	190000	136.23	136.58	136.08	136.45	0.22
20160106	110000	136.35	136.93	136.17	136.51	0.16
20160106	120000	136.51	136.72	136	136.46	-0.05

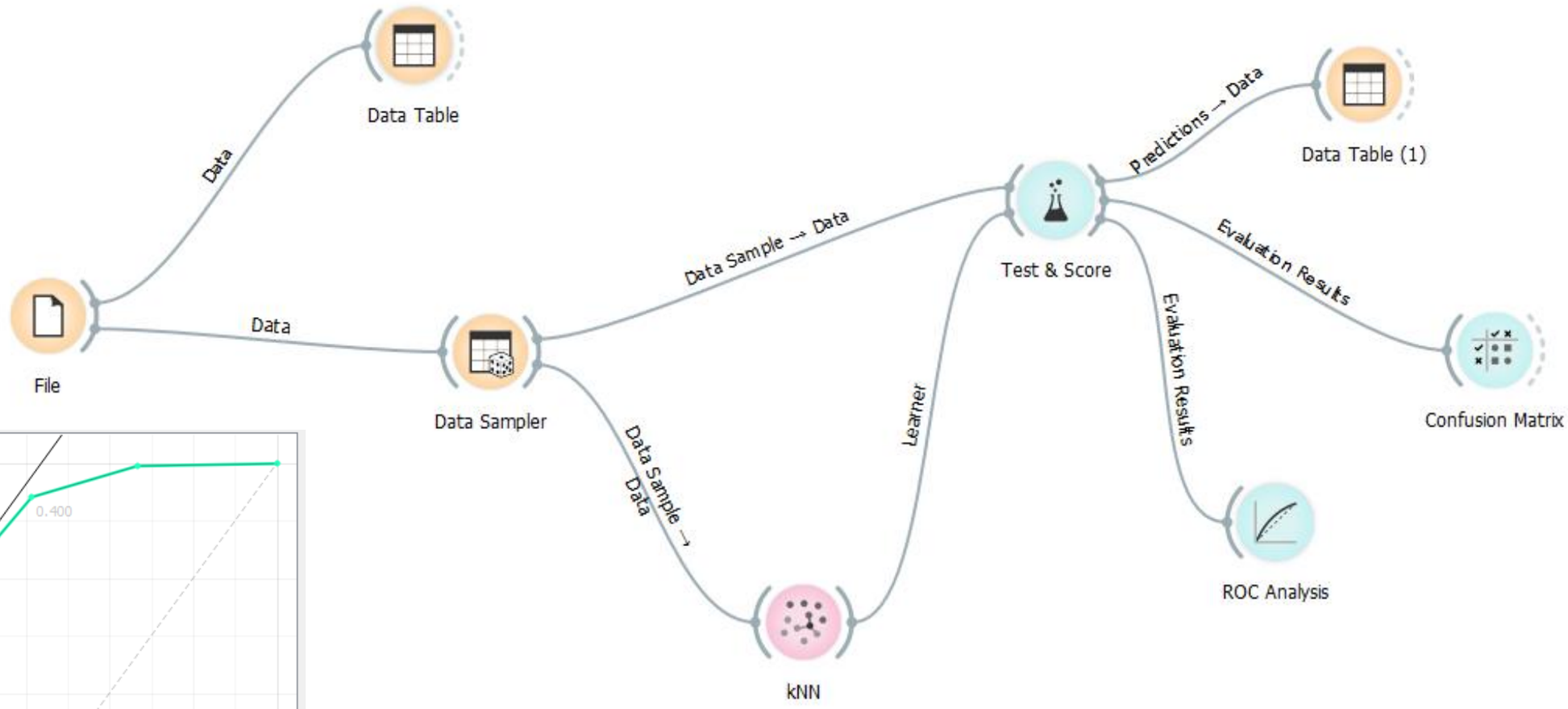
http://www.finam.ru/profile/moex-akcii/gazprom/export/

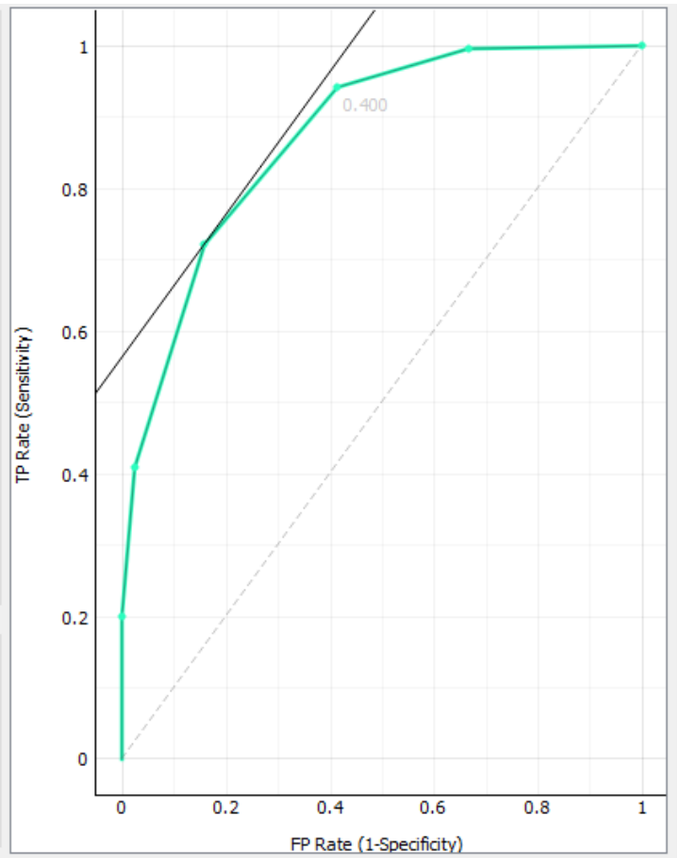
Задача заключается в том ,что бы построить на основе KNN алгоритм предсказания котировок ценных бумаг.

Решение: будем использовать пакет Orange.

1. Возьмем исходный датасет.
2. Обучим классификатора на наших данных.
3. Посмотрим как классификатор предсказывает на наших данных котировки акций Газпрома.

ПРЕДСКАЗАНИЯ АКЦИЙ ГАЗПРОМА





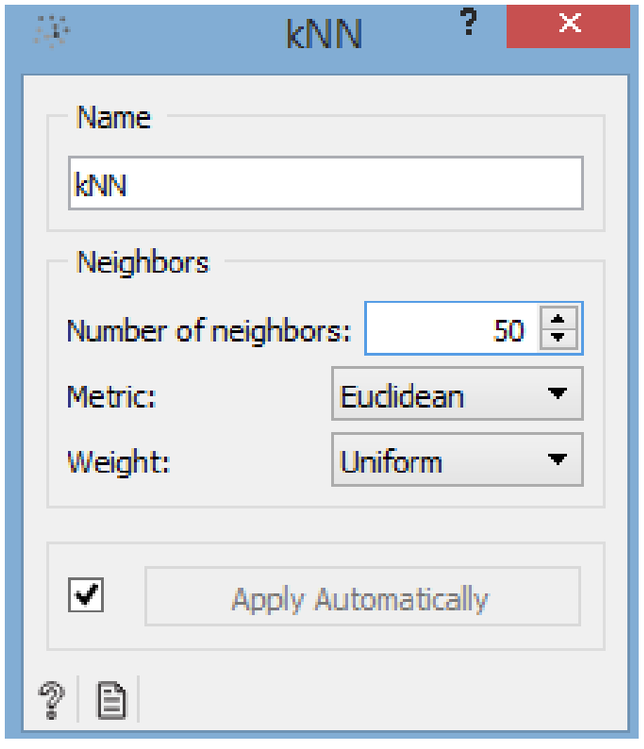
Predicted

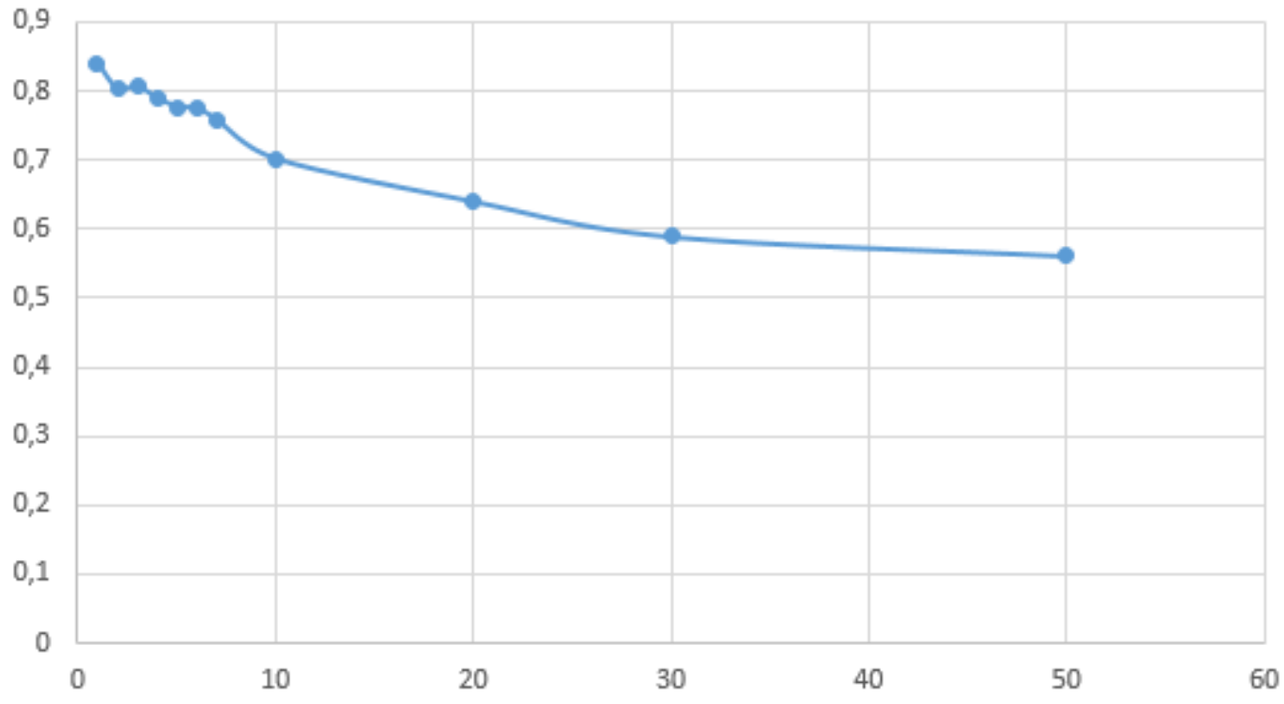
Actual	down	flat	up	Σ
down	363	0	137	500
flat	4	0	6	10
up	95	0	505	600
Σ	462	0	648	1110

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.864	0.782	0.777	0.775	0.782

ВЛИЯНИЕ ЧИСЛА СОСЕДЕЙ НА КЛАССИФИКАЦИЮ

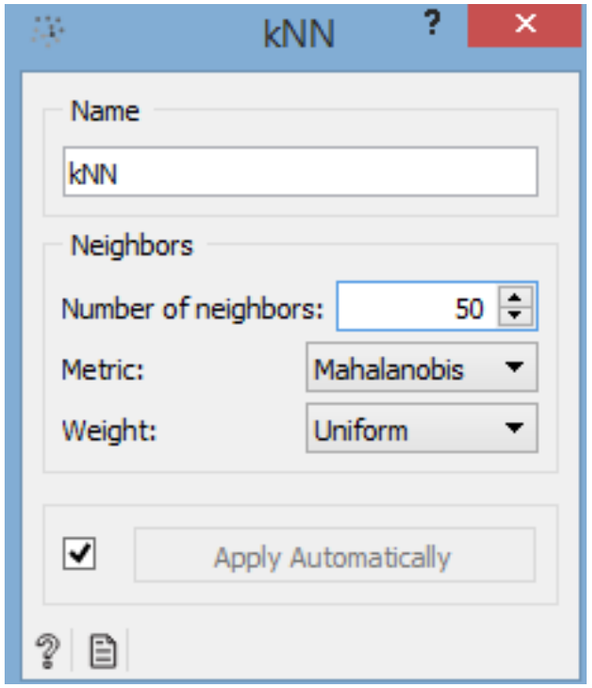


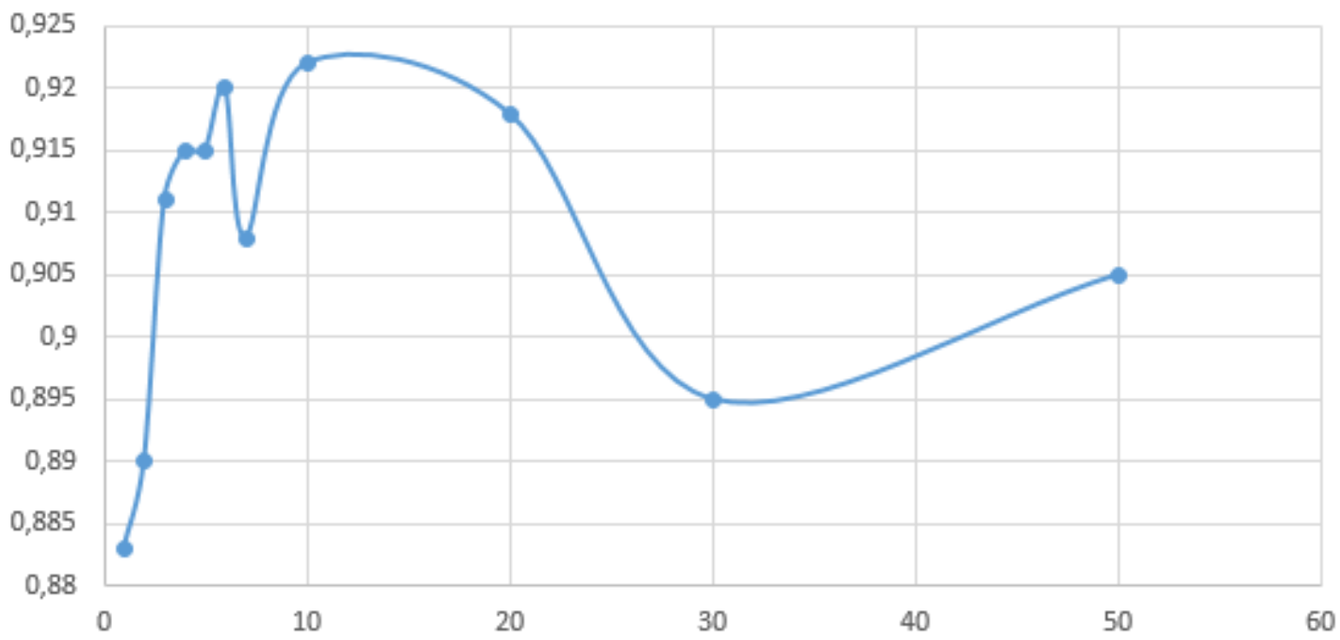


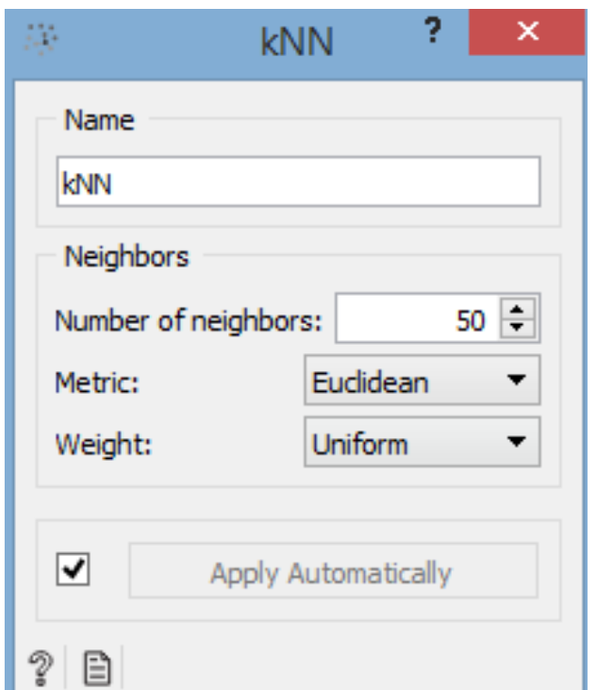
Evaluation Results

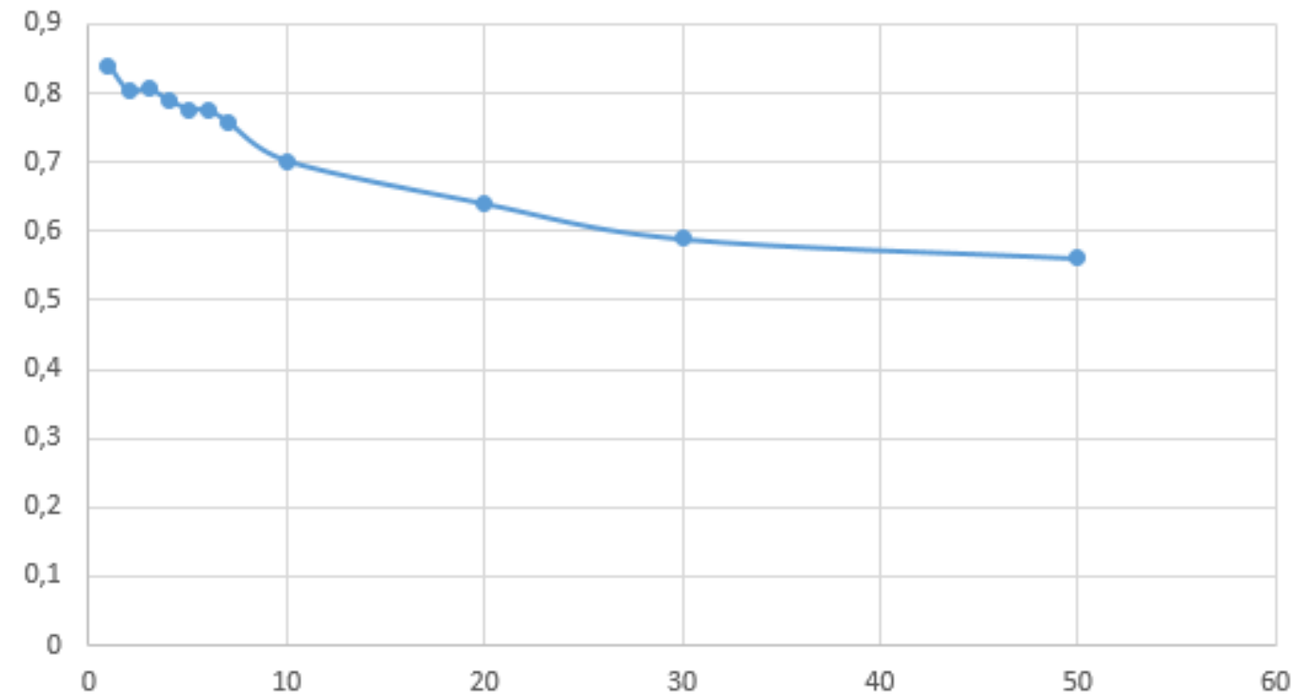
Method	AUC	CA	F1	Precision	Recall
kNN	0.625	0.561	0.559	0.556	0.561

ВЛИЯНИЕ ТИПА РАССТОЯНИЯ НА КЛАССИФИКАЦИЮ

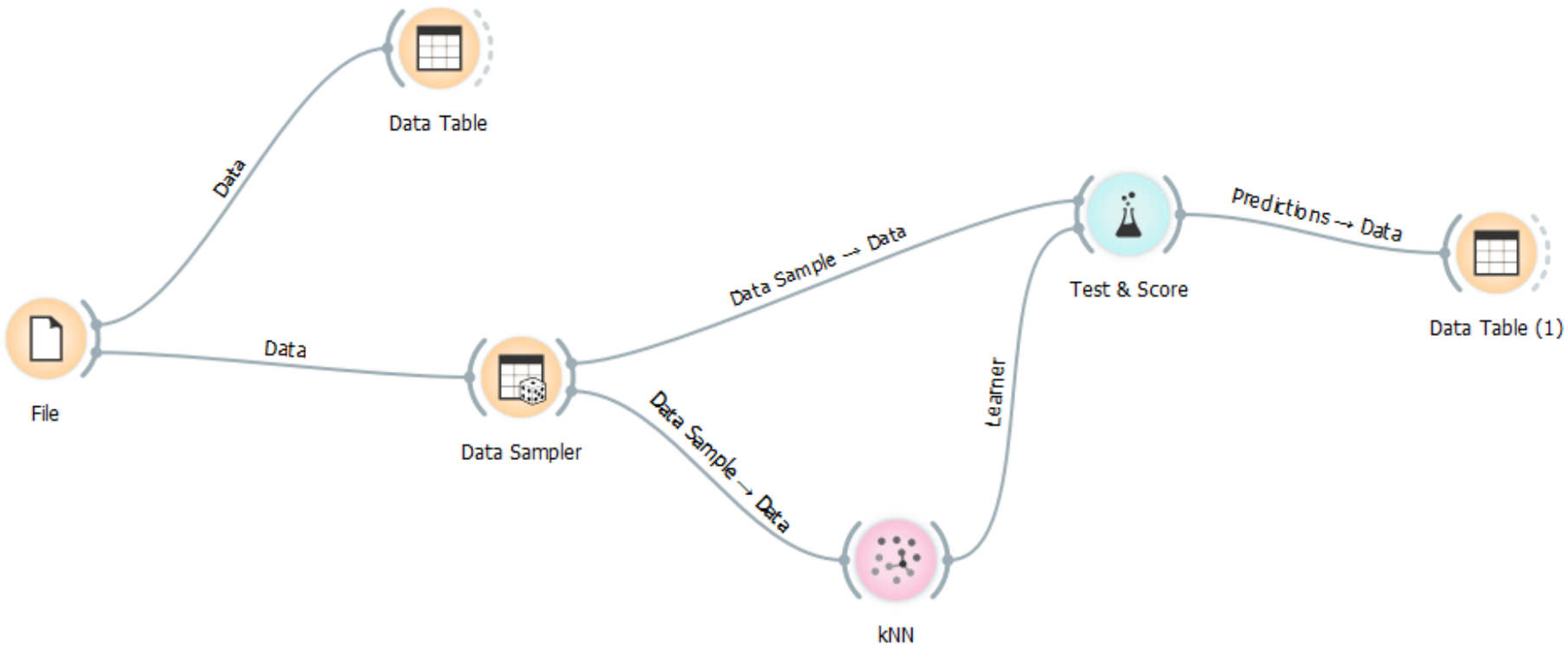








НЕФТЬ - РУБЛЬ



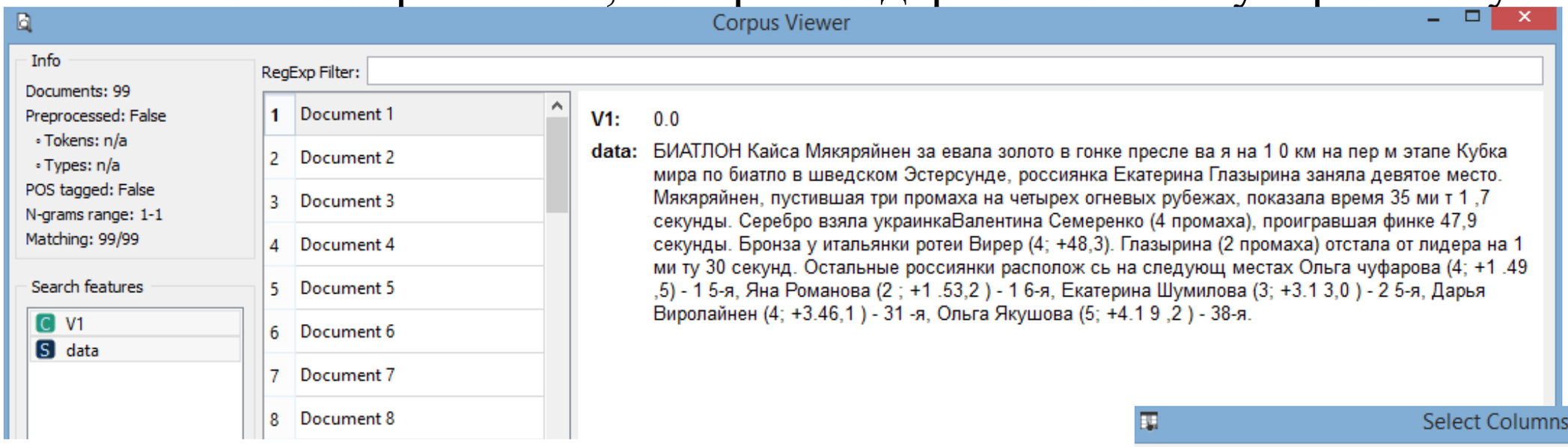
	RubUsd	Date	Brent	UsdRub
1	0.030202	41641.0	107.65	33.1100
2	0.030120	41642.0	106.65	33.2000
3	0.030130	41645.0	106.67	33.1900
4	0.030175	41646.0	107.01	33.1400
5	0.030166	41647.0	106.83	33.1500
6	0.030111	41648.0	106.00	33.2100
7	0.030312	41649.0	106.83	32.9900
8	0.030312	41651.0	106.92	32.9900
9	0.030048	41652.0	105.59	33.2800
10	0.030030	41653.0	105.25	33.3000
11	0.029940	41654.0	105.97	33.4000
12	0.029958	41655.0	105.48	33.3800
13	0.029771	41656.0	106.32	33.5900
14	0.029621	41659.0	106.20	33.7600

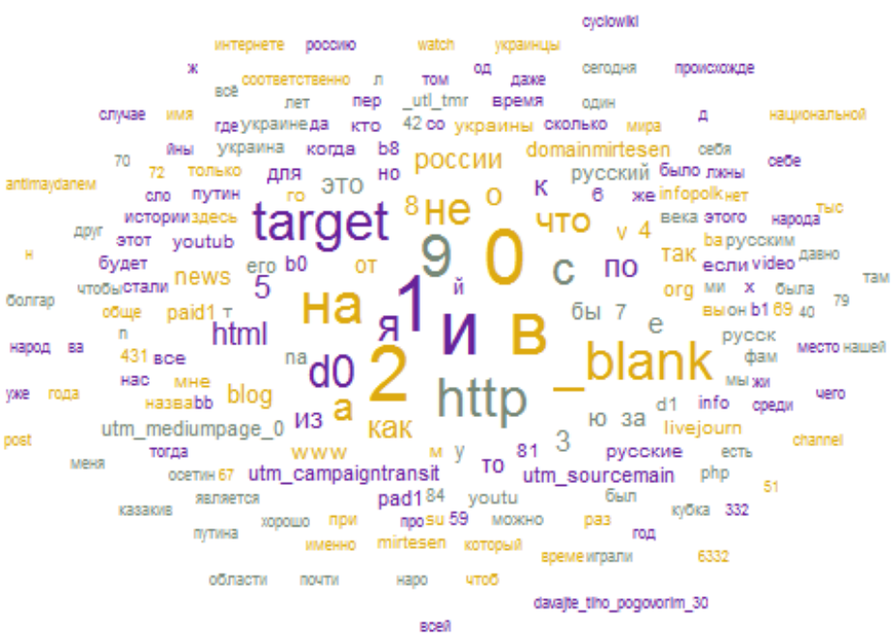
Evaluation Results

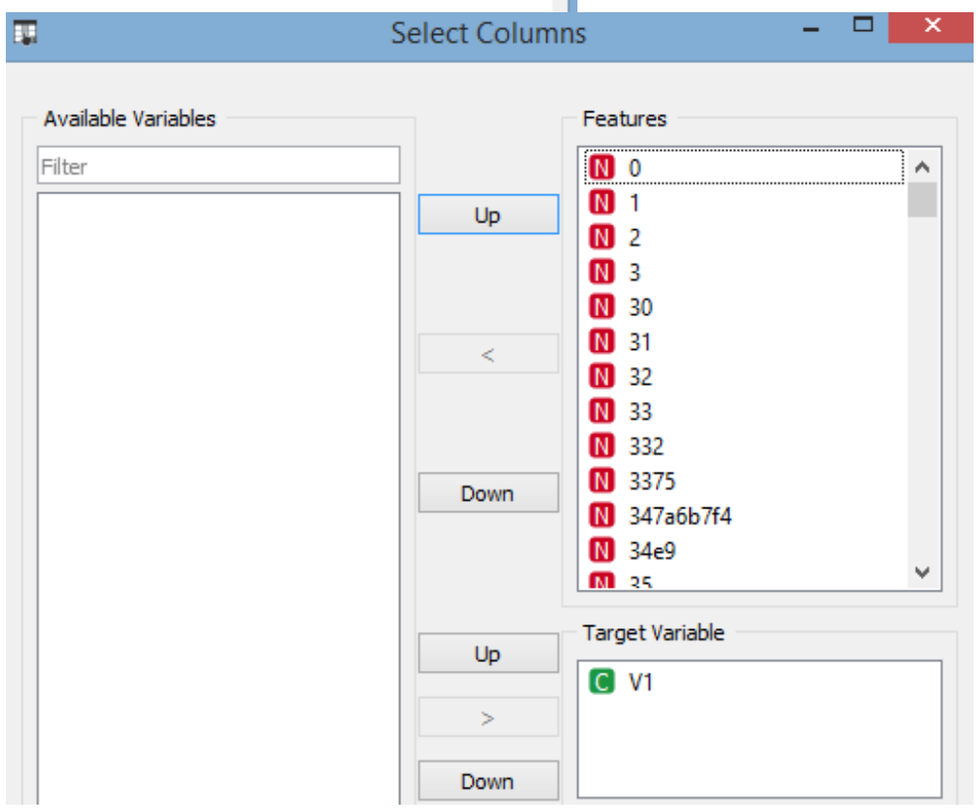
Method	MSE	RMSE	MAE	R2
kNN	0.000	0.001	0.001	0.964

КЛАССИФИКАЦИЯ РУСКОЯЗЫЧНЫХ ТЕКСТОВ

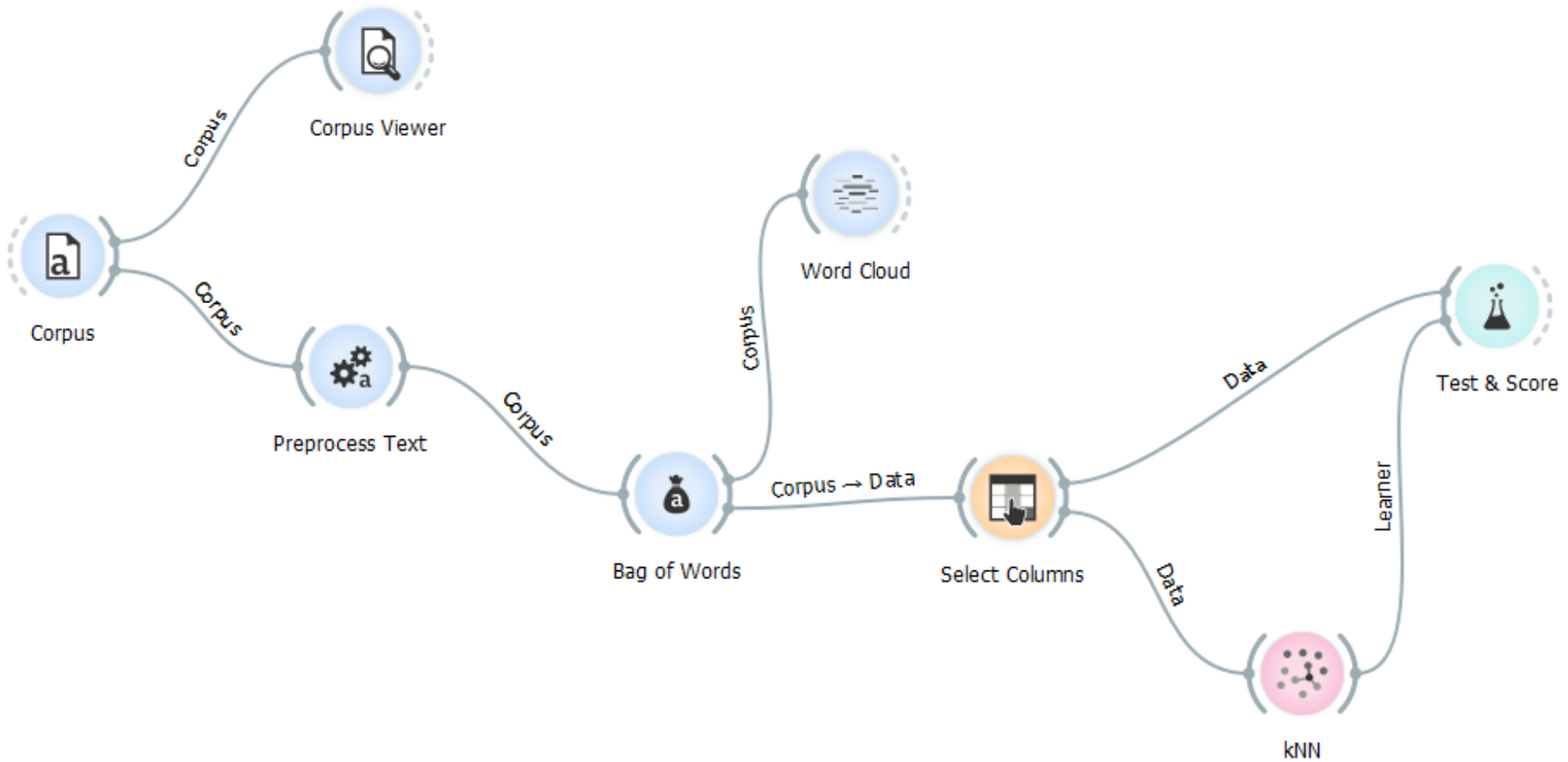
Возьмем набор текстов, которые содержат тональную разметку.







ЭФФЕКТ УДАЛЕНИЯ СТОП СЛОВ НА КЛАССИФИКАЦИЮ



Результат классификации без очистки

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.514	0.506	0.351	0.626	0.506

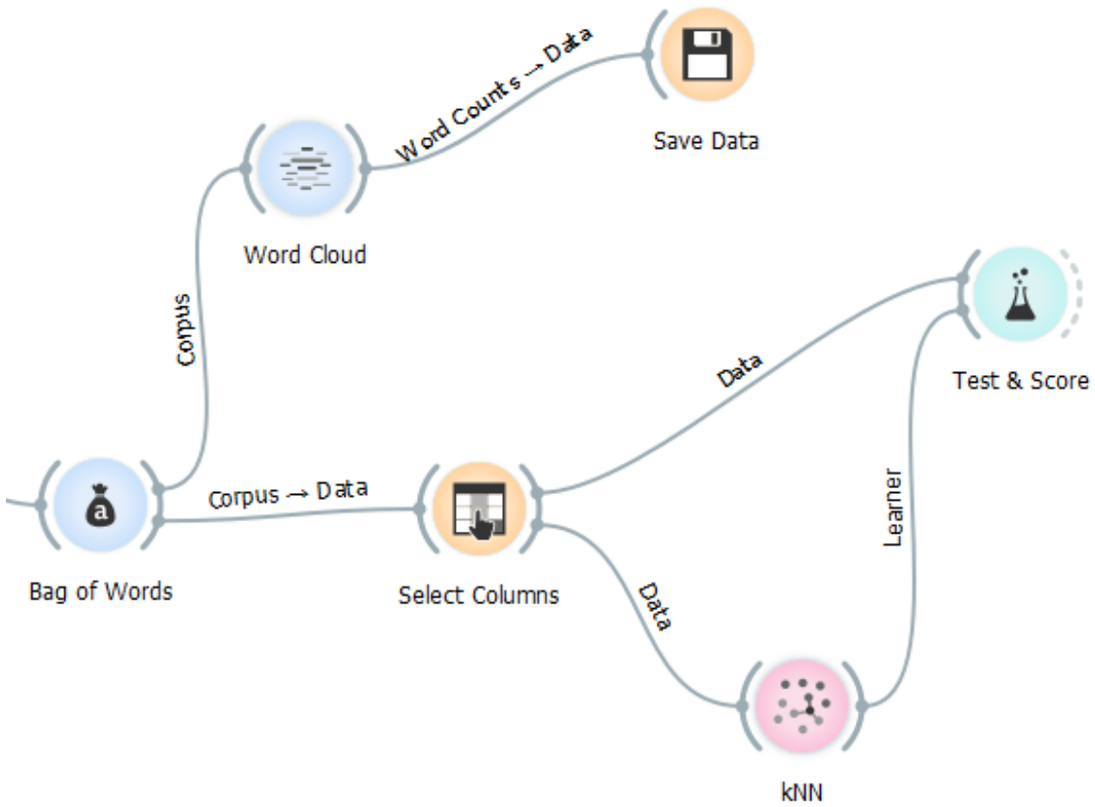
Результат классификации после удаления стоп слов

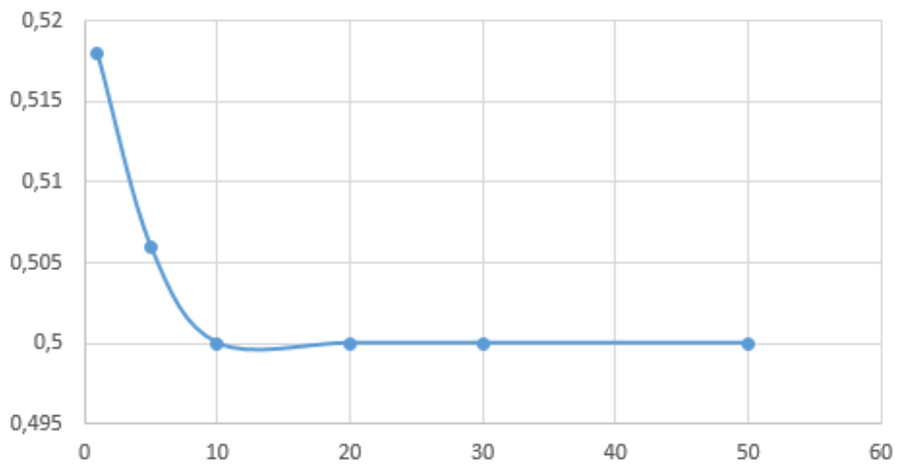
Evaluation Results

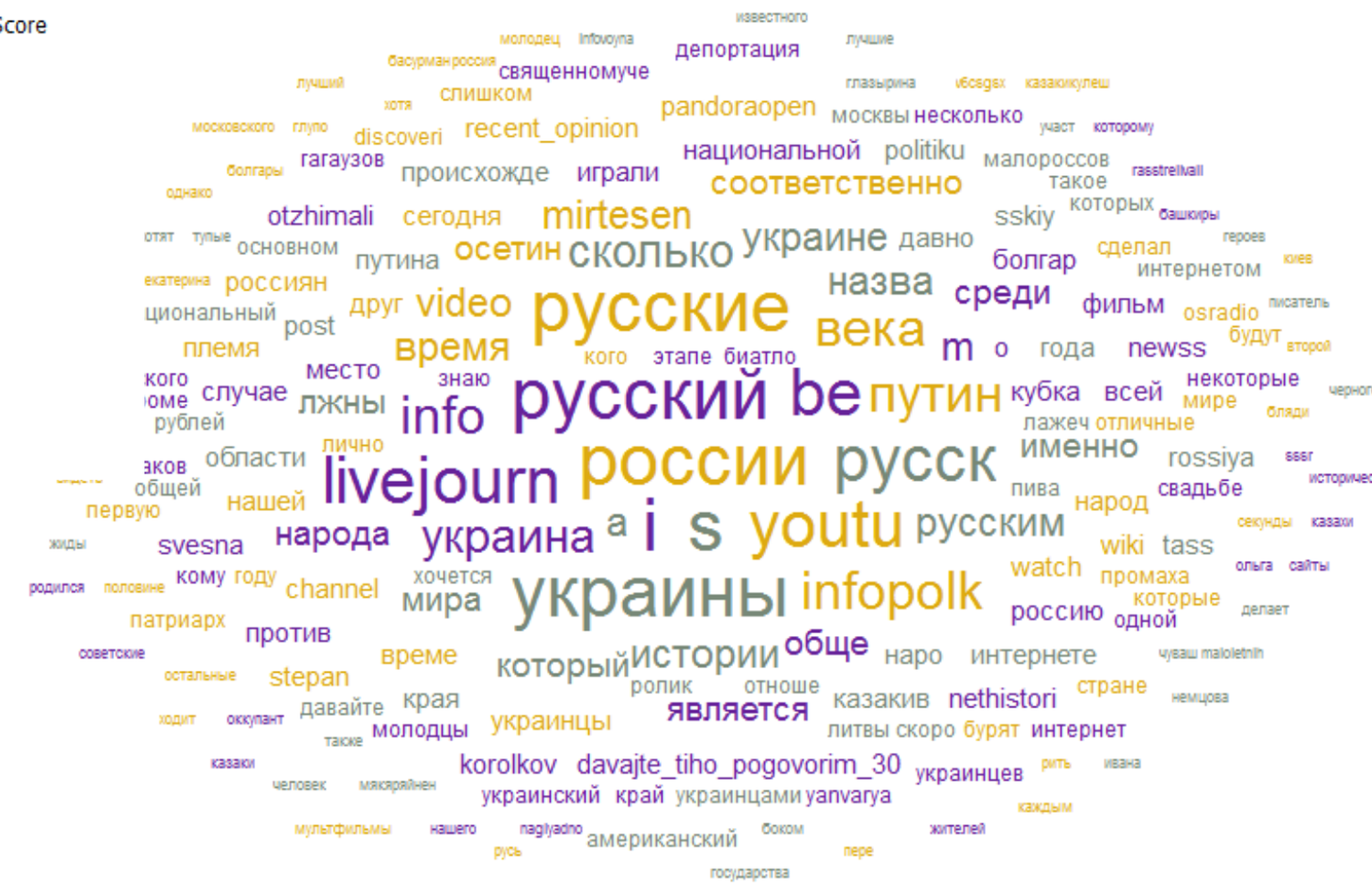
Method	AUC	CA	F1	Precision	Recall
kNN	0.539	0.500	0.401	0.500	0.500

ЭФФЕКТ ВЫБОРА ЧИСЛА СОСЕДЕЙ

Используется очищенный датасет и эвклидово расстояние







КЛАССИФИКАЦИЯ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

1	Document 1
2	Document 2
3	Document 3
4	Document 4
5	Document 5
6	Document 6
7	Document 7
8	Document 8
9	Document 9
10	Document 10

children:

the house Jim says he rum ; and as he spoke he reeled a little and caught himself with one hand against the wall Are you hurt? cried I Rum he repeated I must get away from here Rum! Rum! I ran to fetch it but I was quite unsteadied by all that had fallen out and I broke one glass and fouled the tap and while I was still getting in my own way I heard a loud fall in the parlour and running in beheld the captain lying full length upon the floor At the same instant my mother alarmed by the cries and fighting came running downstairs to help me Between us we raised his head He was breathing very loud and hard but his eyes were closed and his face a horrible colour Dear deary me cried my mother what a

children

has lived rough and I'll raise Cain Your doctor hisself said one glass wouldn't hurt me I'll give you a golden guinea for a noggin Jim He was growing more and more excited and this alarmed me for my father who was very low that day and needed quiet; besides I was reassured by the doctor's words now quoted to me and rather offended by the offer of a bribe I want none of your money said I but what you owe my father I'll get you one glass and no more When I brought it to him he seized it greedily and drank it out Aye aye said he that's some better sure enough And now matey did that doctor say how long I was to lie here in this old berth? A week at least said I Thunder! he cried A week! I can't do that; they'd have the black spot on me by then The lubbers is going about to get the wind of me this blessed moment; lubbers as couldn't keep what they got

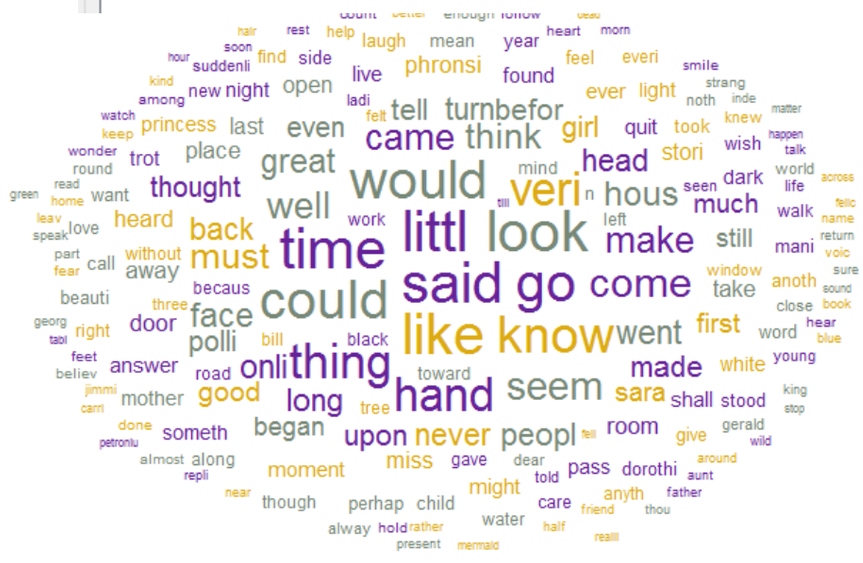
Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.878	0.762	0.760	0.774	0.762

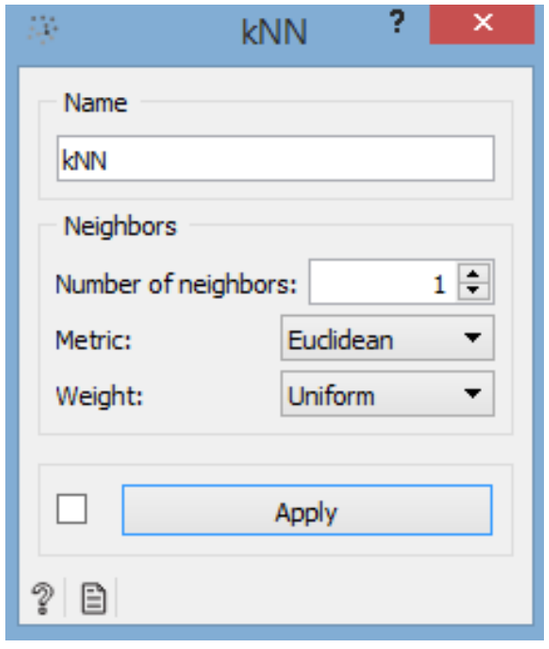
Удаление стоп слов

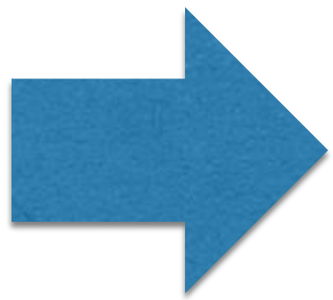
Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.896	0.787	0.786	0.797	0.787



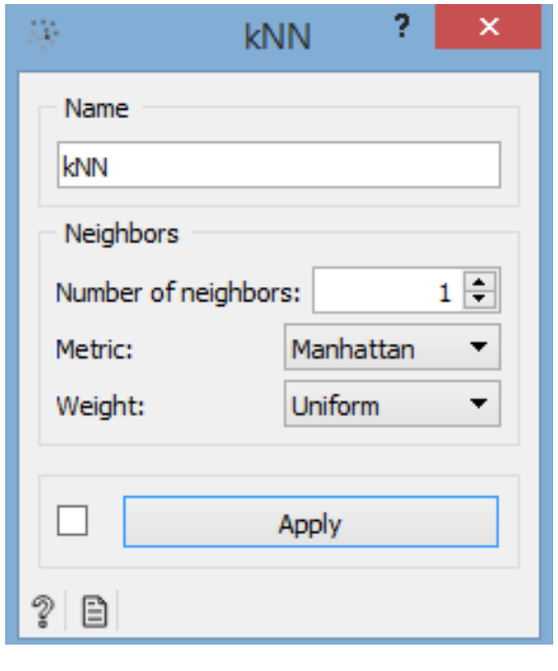
ЭФФЕКТ ВЫБОРА РАССТОЯНИЯ НА ТЕКСТОВУЮ КЛАССИФИКАЦИЮ





Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.925	0.925	0.925	0.926	0.925



Evaluation Results

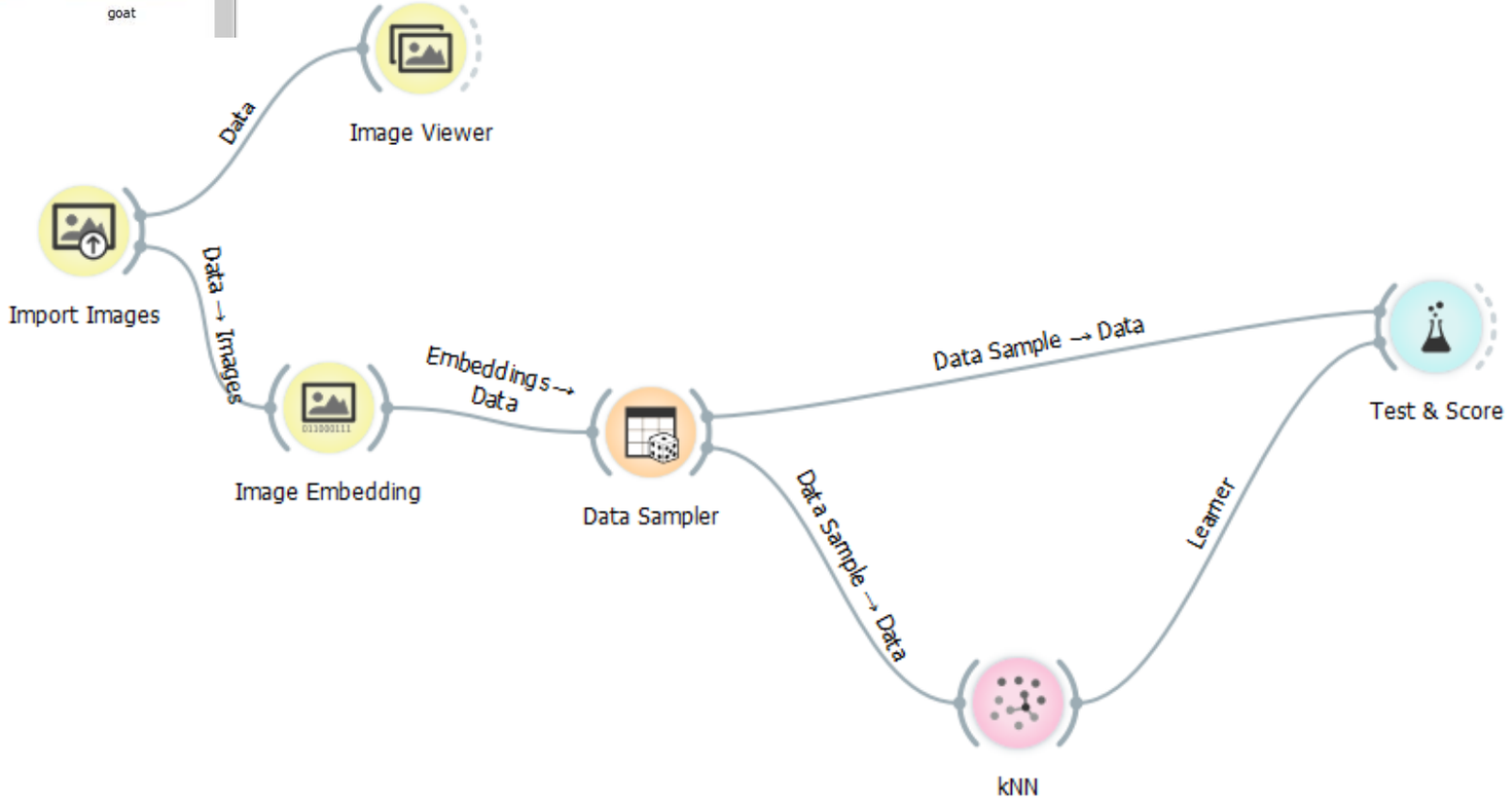
Method	AUC	CA	F1	Precision	Recall
kNN	0.673	0.673	0.637	0.786	0.673

КЛАССИФИКАЦИЯ ИЗОБРАЖЕНИЙ

Evaluation Results

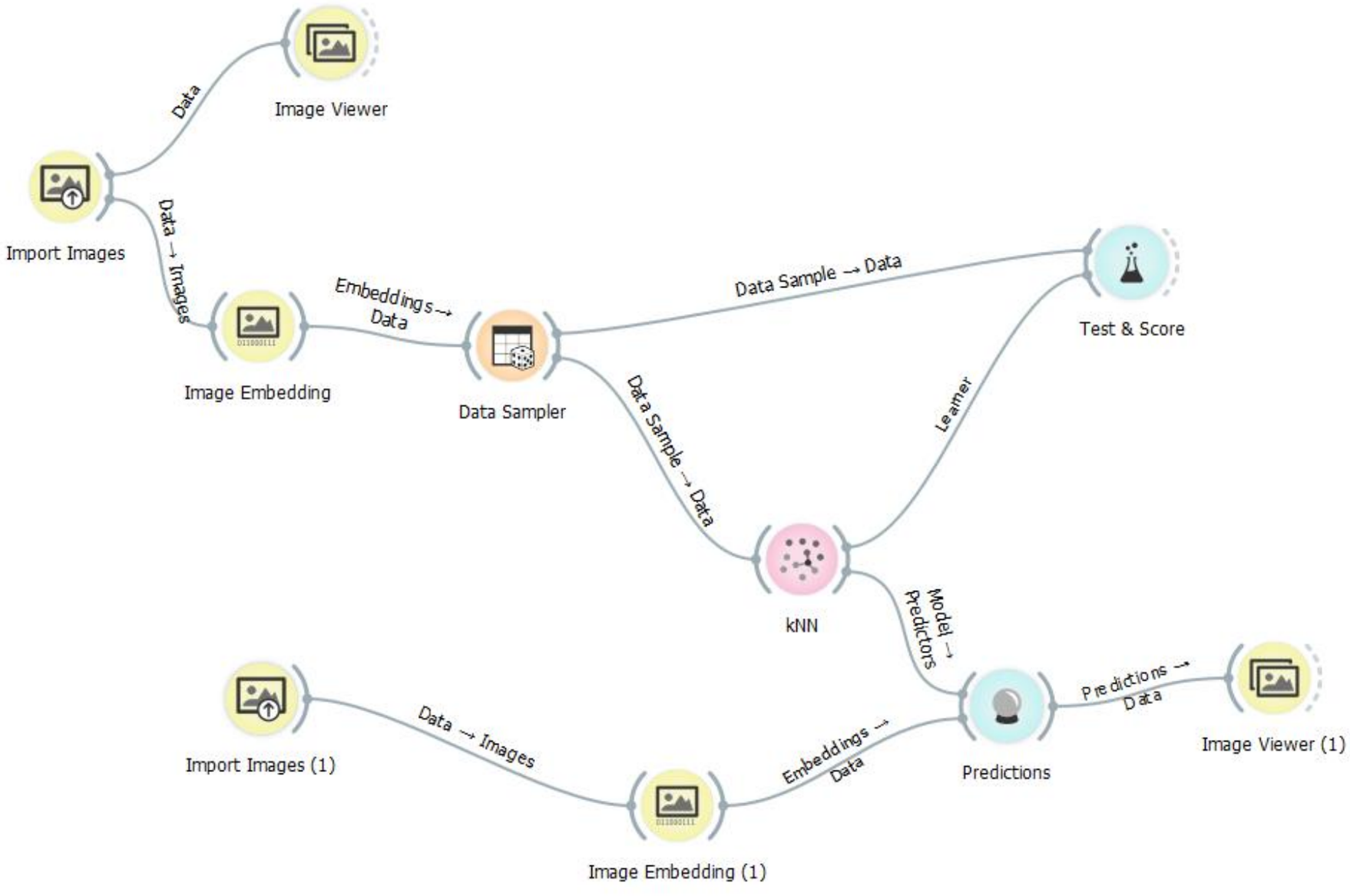
Method	AUC	CA	F1	Precision	Recall
kNN	0.777	0.625	0.592	0.597	0.625

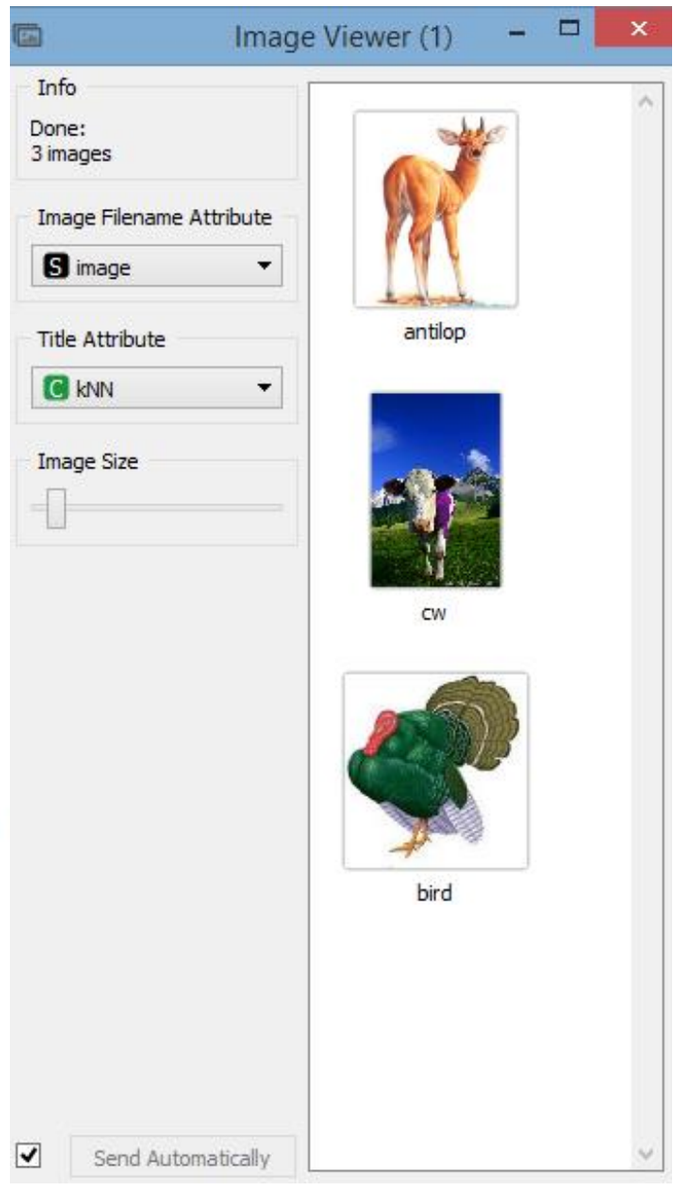
Изображения разложены по категориям. Каждая категория это отдельный каталог.



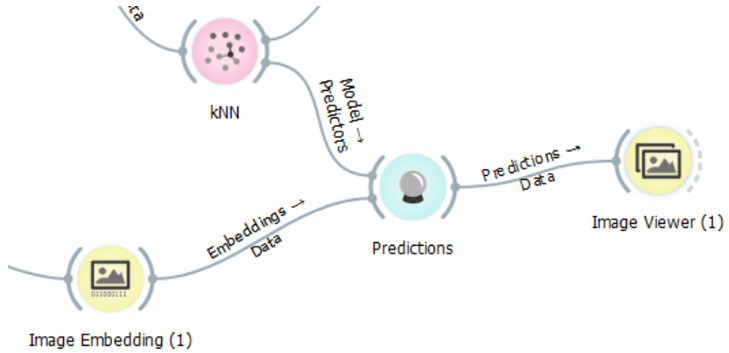
https://www.youtube.com/watch?v=lvgx62a8XQk&index=15&list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy

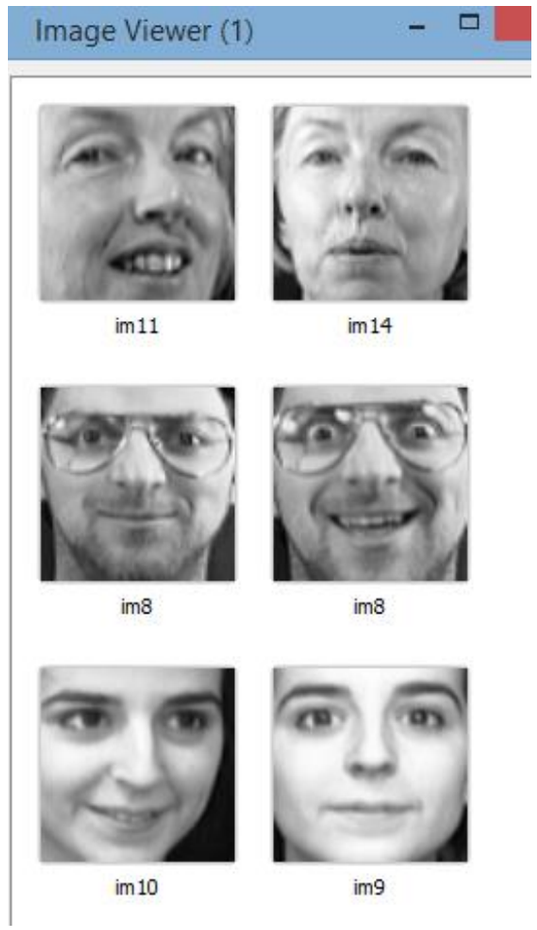
КЛАССИФИКАЦИЯ ИЗОБРАЖЕНИЙ





КЛАССИФИКАЦИЯ ИЗОБРАЖЕНИЙ

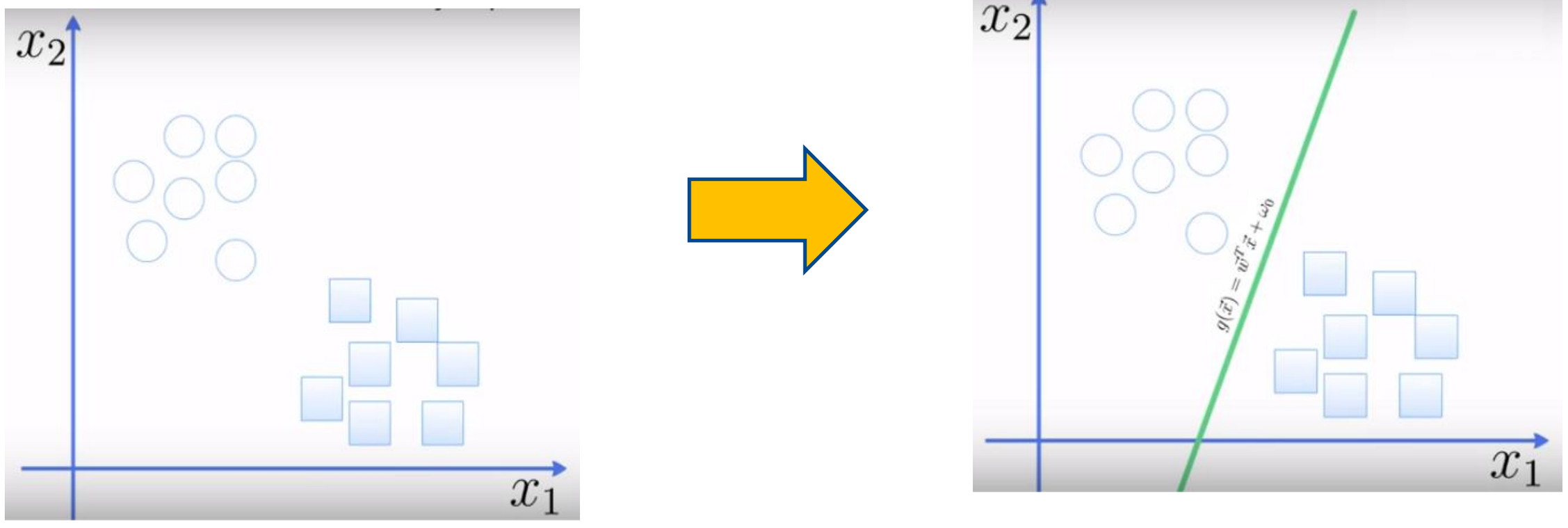




МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM)

Каждый объект данных (например, документ, котировки ценных бумаг или компании) представлен как вектор в **P** мерном пространстве (последовательность чисел). Пусть у нас есть тестовая коллекция, в которой есть набор объектов (features) и есть набор классов. Математическая задача обучения заключается в том что бы найти функцию, которая адекватно сопоставляла объекты и классы, то есть найти такую функцию, которая эффективно разделяла бы объекты в пространстве features.

Рассмотрим пример на плоскости: У нас есть два класса с двумя features (x1, x2). Нужной найти прямую линию, которая оптимально разделяла два класса.

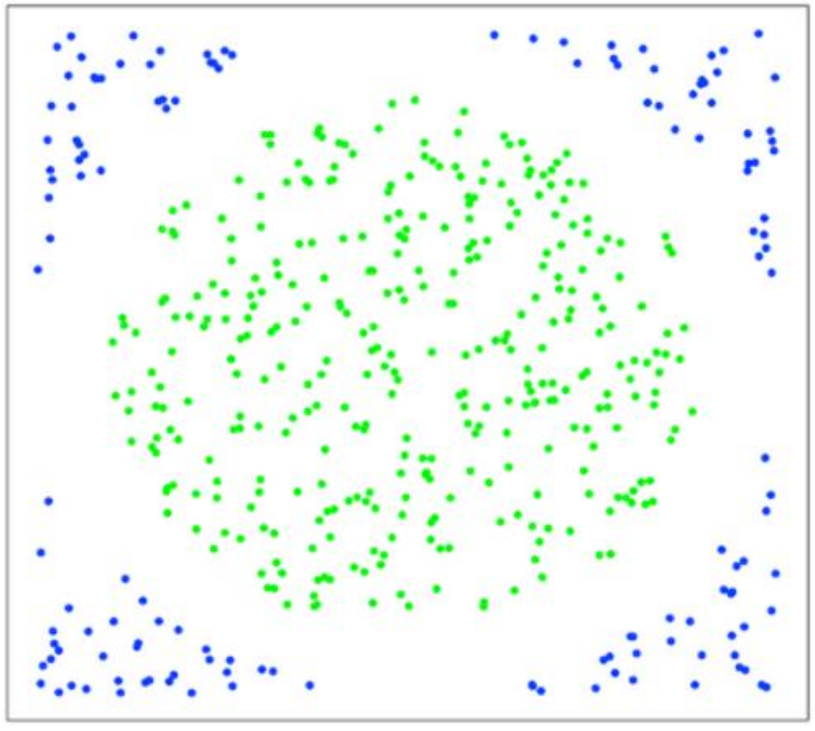


МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM)

Нахождение уравнения плоскости является стандартной задачей квадратичного программирования и решается с помощью множителей Лагранжа. Собственно в этом заключается процесс обучения.

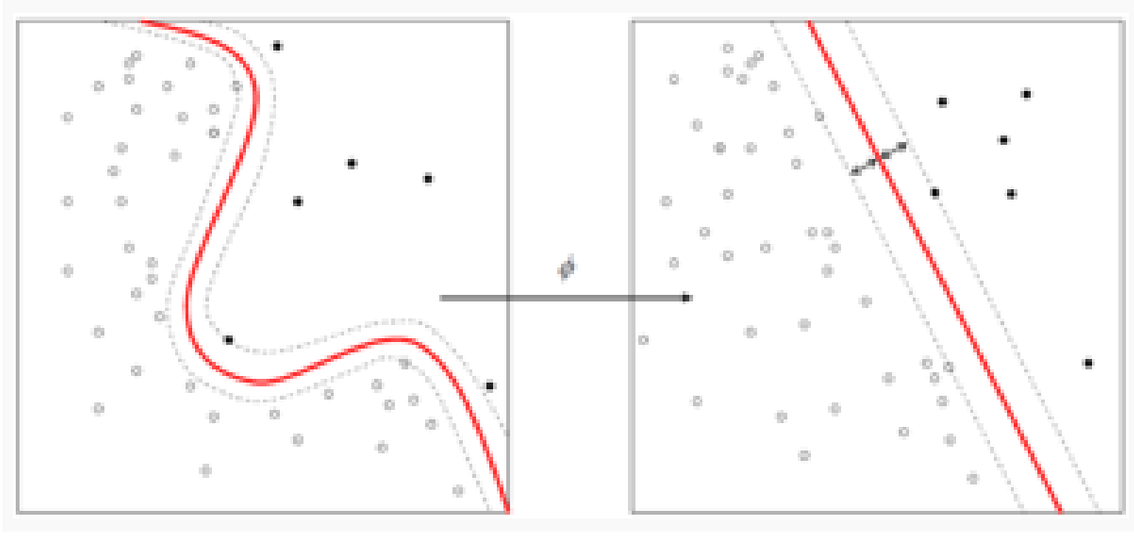
Как только плоскость найдена, берем новый объект и смотрим где он расположен относительно плоскости. Если справа, то принадлежит одному классу, если слева, то наш объект принадлежит другому классу.

Однако, ка правило на практике встречаются случаи когда объекты расположены, так что на плоскости невозможно провести разделяющую прямую. В этом случае плоскость вкладывается в пространство большей размерности. При вложении плоскость трансформируется таким образом, что бы появилась возможность провести разделяющую плоскость.

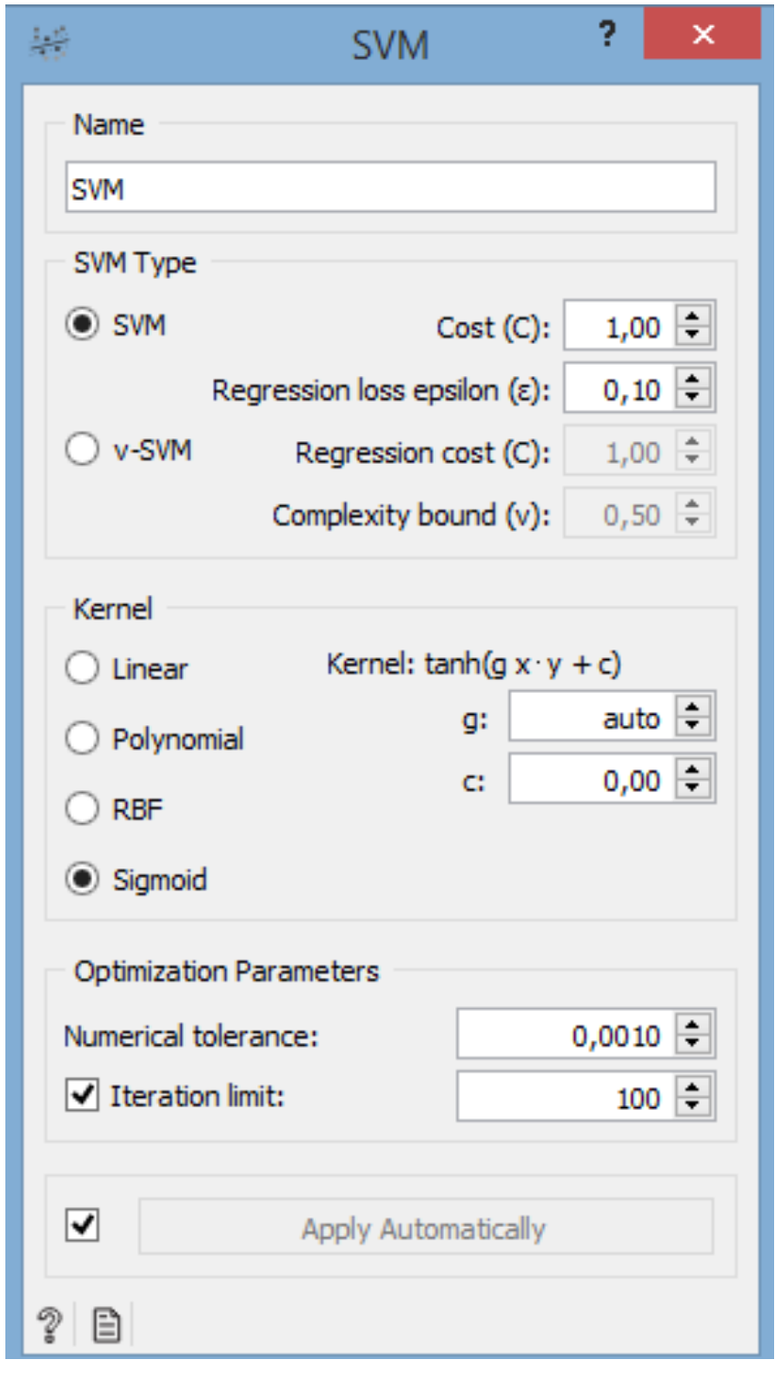


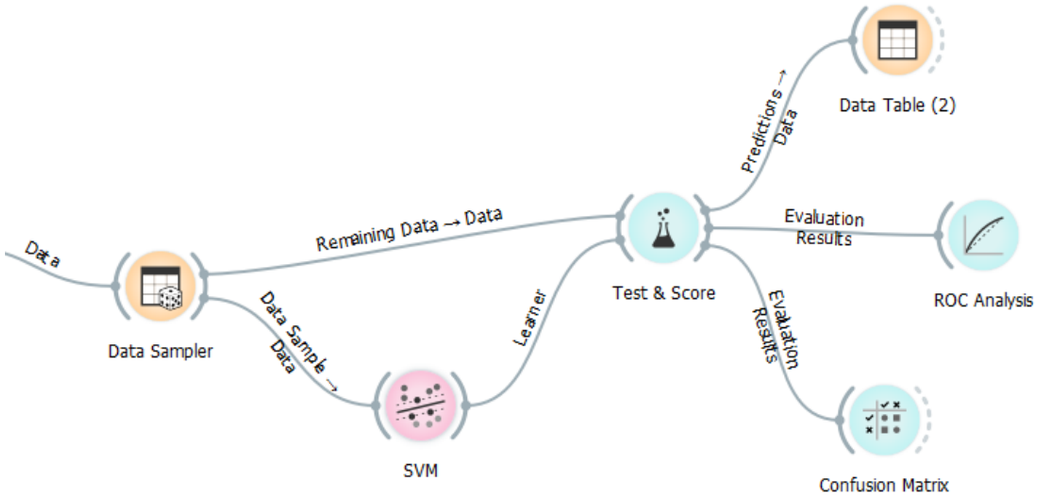
Демонстрация подобного преобразования

https://youtu.be/3liCbRZPrZA



МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM)





Типы kernels (ядер):

1. Линейное ядро.

2. Полиномиальное ядро. $k(x, y) = (\alpha x^T y + c)^d$

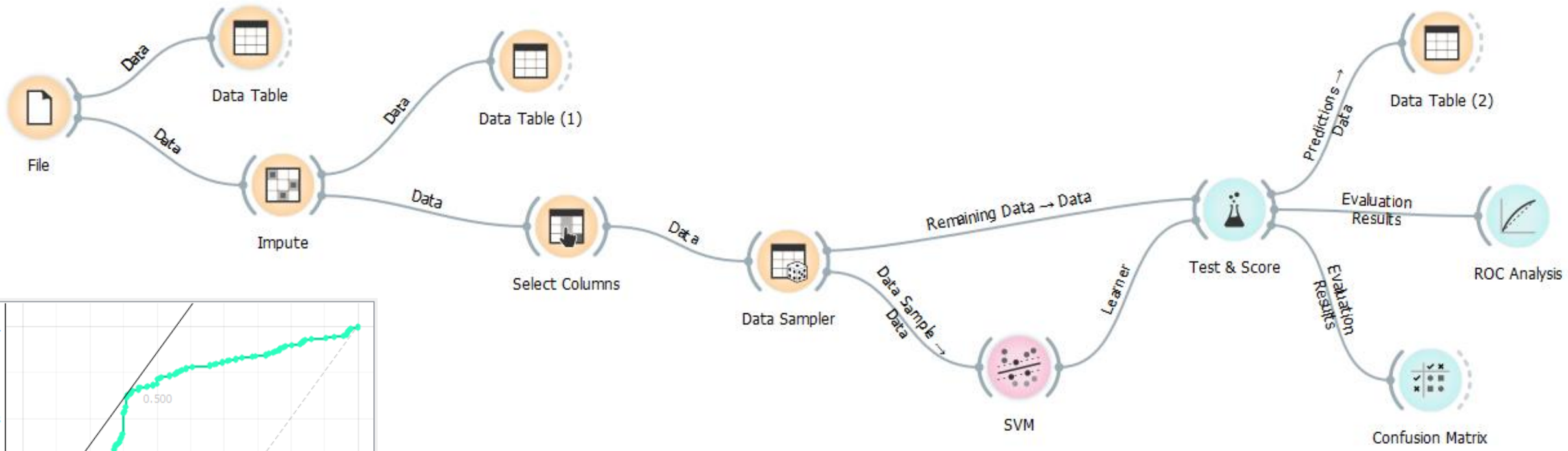
3. Radial basis function kernel $k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$

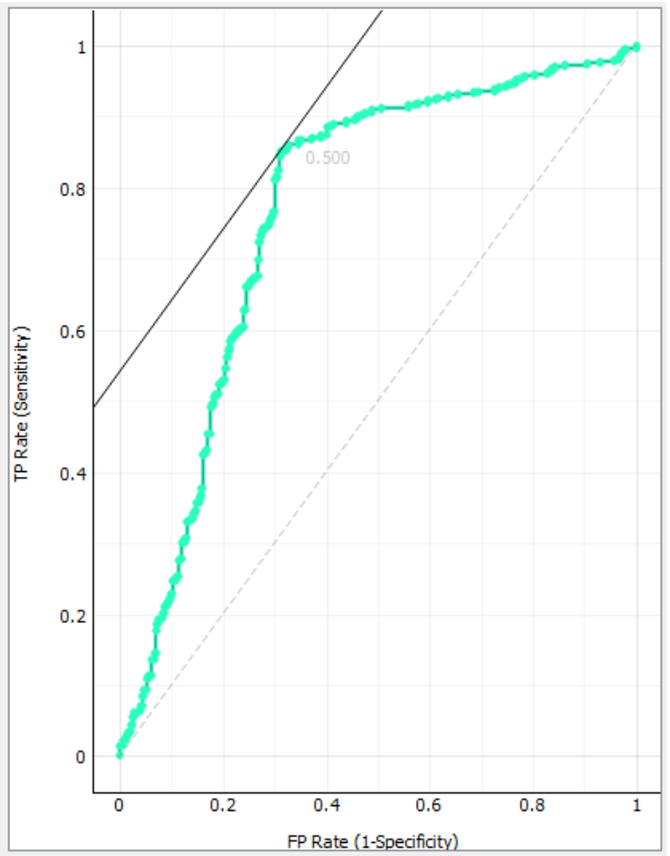
4. Hyperbolic Tangent (Sigmoid) Kernel $k(x, y) = \tanh(\alpha x^T y + c)$

c - const (default =0)

d - степень ядра

SVM - ТАБЛИЧНЫЕ ДАННЫЕ





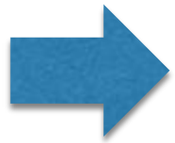
Predicted

Actual	0.0	1.0	Σ
0.0	473	77	550
1.0	120	240	360
Σ	593	317	910

Evaluation Results

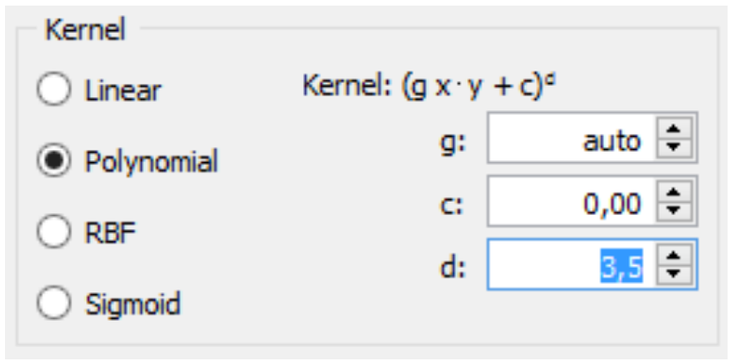
Method	AUC	CA	F1	Precision	Recall
SVM	0.770	0.784	0.781	0.782	0.784

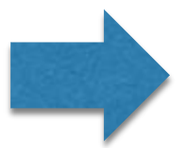
SVM – ТАБЛИЧНЫЕ ДАННЫЕ – ЭФФЕКТ ВЫБОРА ЯДРА



Evaluation Results

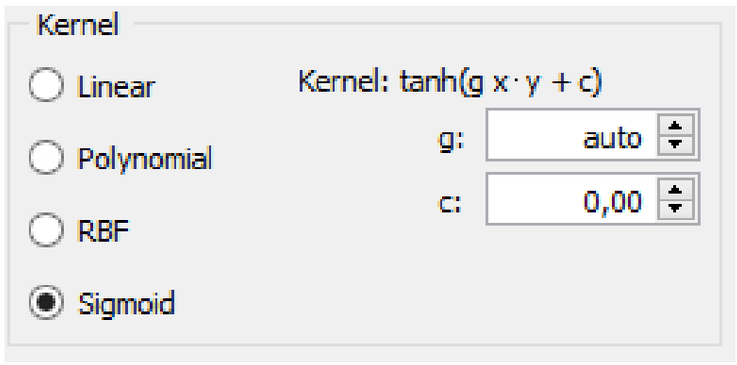
Method	AUC	CA	F1	Precision	Recall
SVM	0.776	0.780	0.776	0.779	0.780

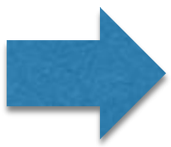




Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.803	0.795	0.789	0.796	0.795

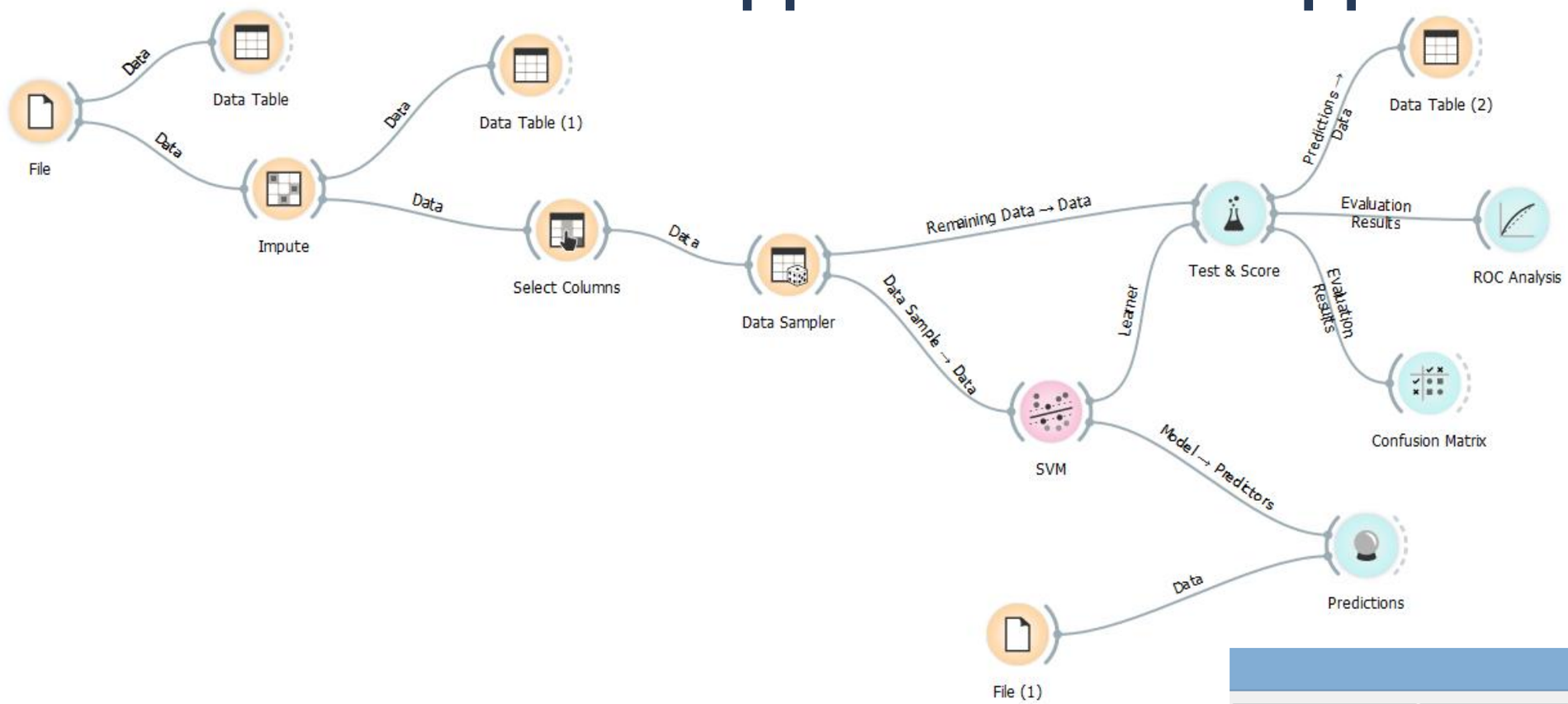




Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.770	0.784	0.781	0.782	0.784

SVM – ТАБЛИЧНЫЕ ДАННЫЕ - ПРЕДСКАЗАНИЕ

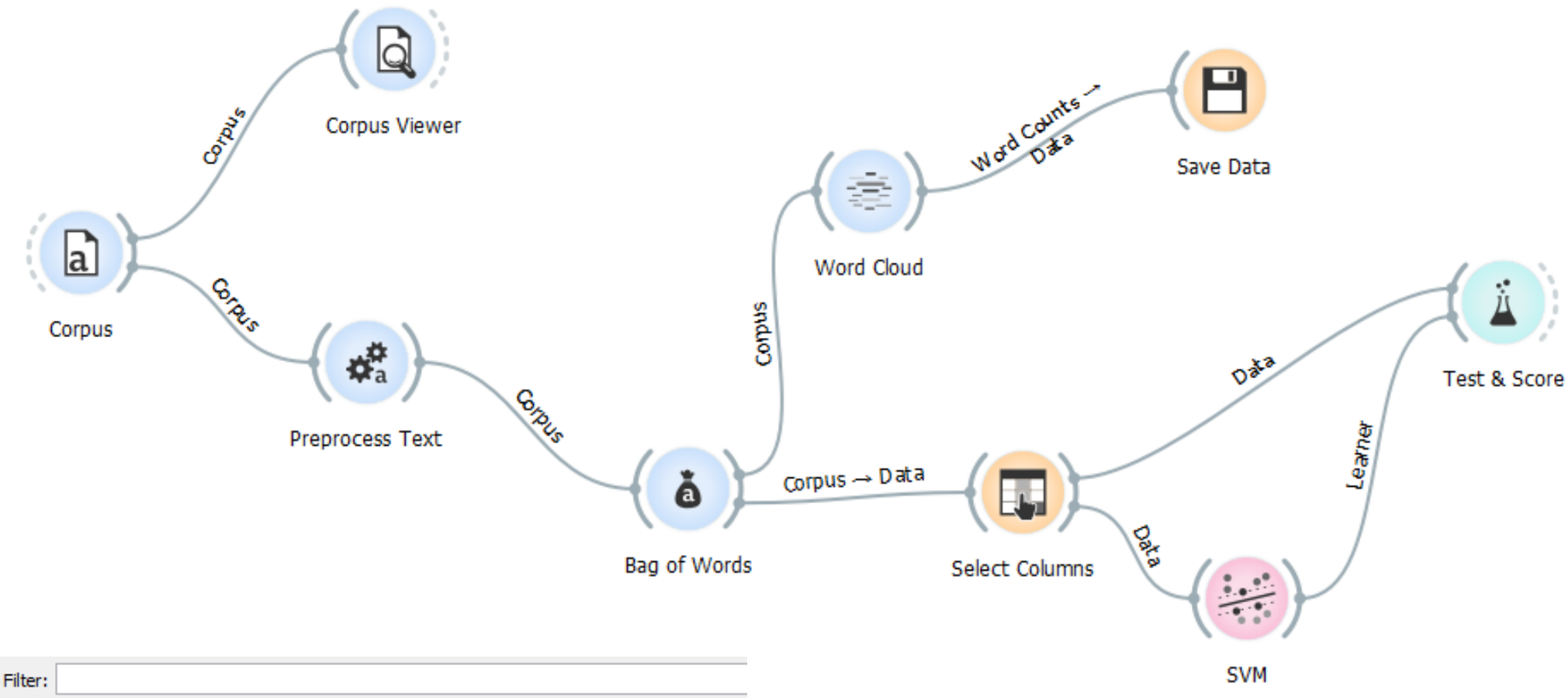


	Name	Ticket	PassengerId	Pclass	Sex	Age
1	Braund, Mr. Ow...	A/5 21171	1.0	3.0	1.0	22.0
2	Cumings, Mrs. ...	PC 17599	2.0	1.0	0.0	38.0
3	Heikkinen, Miss...	STON/O2. 3101...	3.0	3.0	0.0	26.0
4	Futrelle, Mrs. Ja...	113803.0	4.0	1.0	0.0	35.0
5	Allen, Mr. Willia...	373450.0	5.0	3.0	1.0	35.0
6	Moran, Mr. Jam...	330877.0	6.0	3.0	1.0	?
7	McCarthy, Mr. ...	17463.0	7.0	1.0	1.0	54.0
8	Palsson, Master...	349909.0	8.0	3.0	1.0	2.0
9	Johnson, Mrs. ...	347742.0	9.0	3.0	0.0	27.0
10	Nasser, Mrs. Ni...	237736.0	10.0	2.0	0.0	14.0
11	Sandstrom, Mis...	PP 9549	11.0	3.0	0.0	4.0
12	Bonnell, Miss. E...	113783.0	12.0	1.0	0.0	58.0
13	Saundercock, ...	A/5. 2151	13.0	3.0	1.0	20.0
14	Andersson, Mr. ...	347082.0	14.0	3.0	1.0	39.0

Predictions

	SVM	Name	Ticket	PassengerId
1	0.78 : 0.22 → 0.0	Braund, Mr. Ow...	A/5 21171	1.0
2	0.34 : 0.66 → 1.0	Cumings, Mrs. ...	PC 17599	2.0
3	0.33 : 0.67 → 1.0	Heikkinen, Miss...	STON/O2. 3101...	3.0
4	0.42 : 0.58 → 1.0	Futrelle, Mrs. Ja...	113803.0	4.0
5	0.63 : 0.37 → 0.0	Allen, Mr. Willia...	373450.0	5.0
6	0.67 : 0.33 → 0.0	Moran, Mr. Jam...	330877.0	6.0
7	0.38 : 0.62 → 1.0	McCarthy, Mr. ...	17463.0	7.0
8	0.59 : 0.41 → 0.0	Palsson, Master...	349909.0	8.0
9	0.44 : 0.56 → 1.0	Johnson, Mrs. ...	347742.0	9.0
10	0.41 : 0.59 → 1.0	Nasser, Mrs. Ni...	237736.0	10.0
11	0.69 : 0.31 → 0.0	Sandstrom, Mis...	PP 9549	11.0
12	0.26 : 0.74 → 1.0	Bonnell, Miss. E...	113783.0	12.0
13	0.62 : 0.38 → 0.0	Saundercock, ...	A/5. 2151	13.0
14	0.12 : 0.88 → 1.0	Andersson, Mr. ...	347082.0	14.0

SVM – ТЕКСТОВЫЕ ДАННЫЕ



RegExp Filter:

1	Document 1
2	Document 2
3	Document 3
4	Document 4
5	Document 5
6	Document 6
7	Document 7
8	Document 8

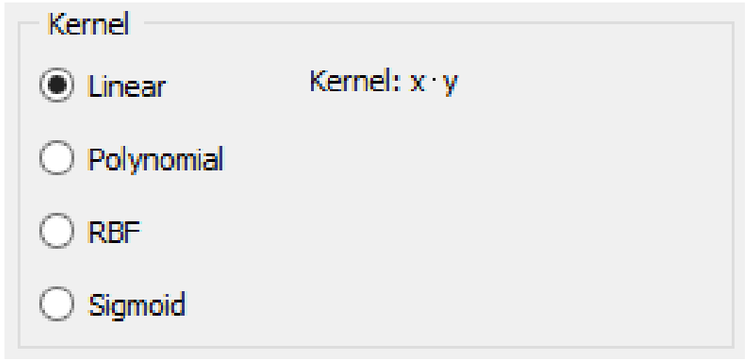
children:

the house Jim says he rum ; and as he spoke he reeled a little and caught himself with one hand against the wall Are you hurt? cried I Rum he repeated I must get away from here Rum! Rum! I ran to fetch it but I was quite unsteadied by all that had fallen out and I broke one glass and fouled the tap and while I was still getting in my own way I heard a loud fall in the parlour and running in beheld the captain lying full length upon the floor At the same instant my mother alarmed by the cries and fighting came running downstairs to help me

Evaluation Results

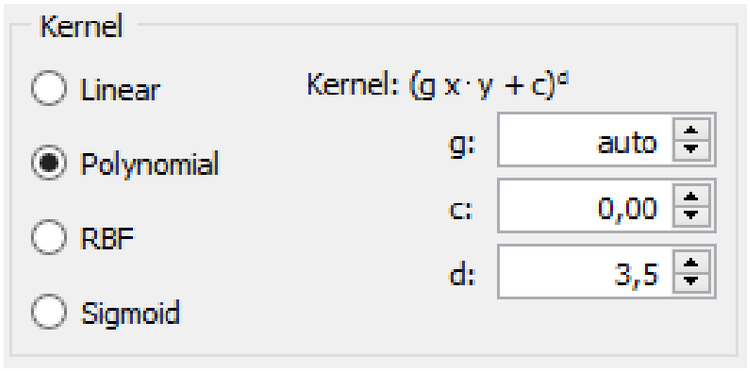
Method	AUC	CA	F1	Precision	Recall
SVM	0.998	0.965	0.965	0.965	0.965

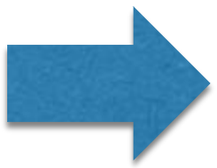
SVM – ТЕКСТОВЫЕ ДАННЫЕ – ЭФФЕКТ ВЫБОРА ЯДРА



Evaluation Results

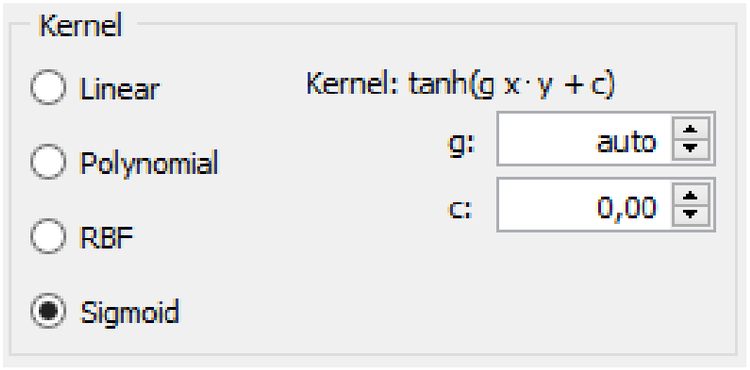
Method	AUC	CA	F1	Precision	Recall
SVM	0.998	0.965	0.965	0.965	0.965

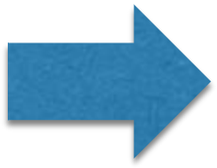




Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.033	0.500	0.333	0.250	0.500

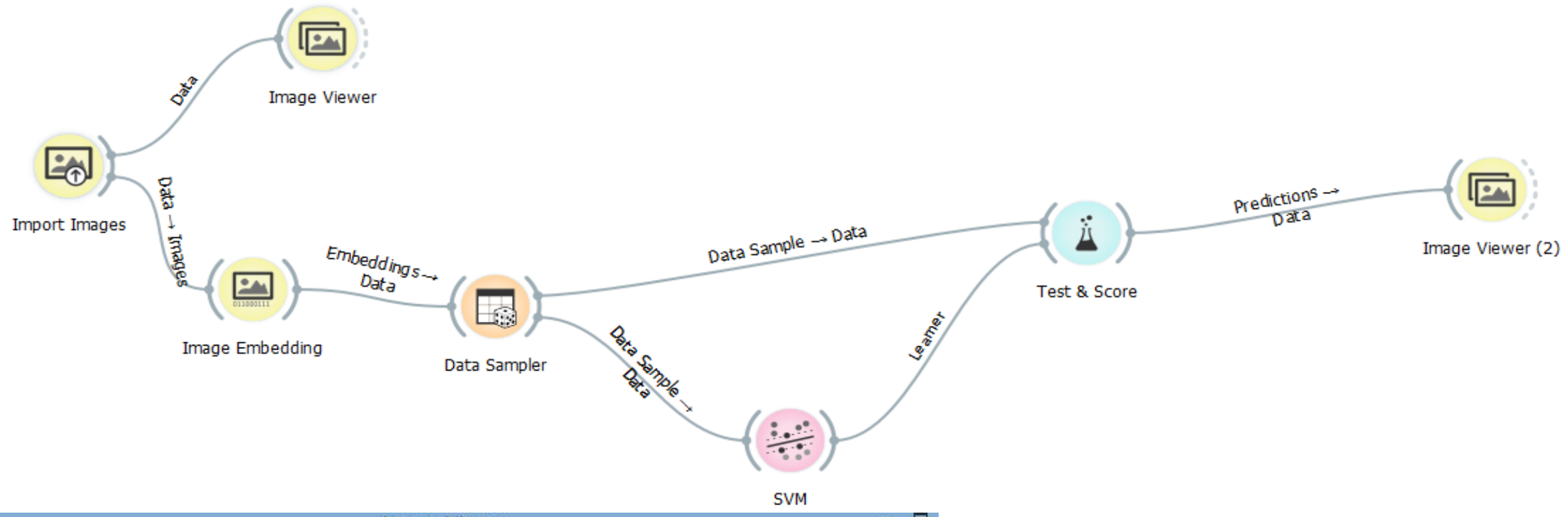




Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.953	0.625	0.565	0.778	0.625

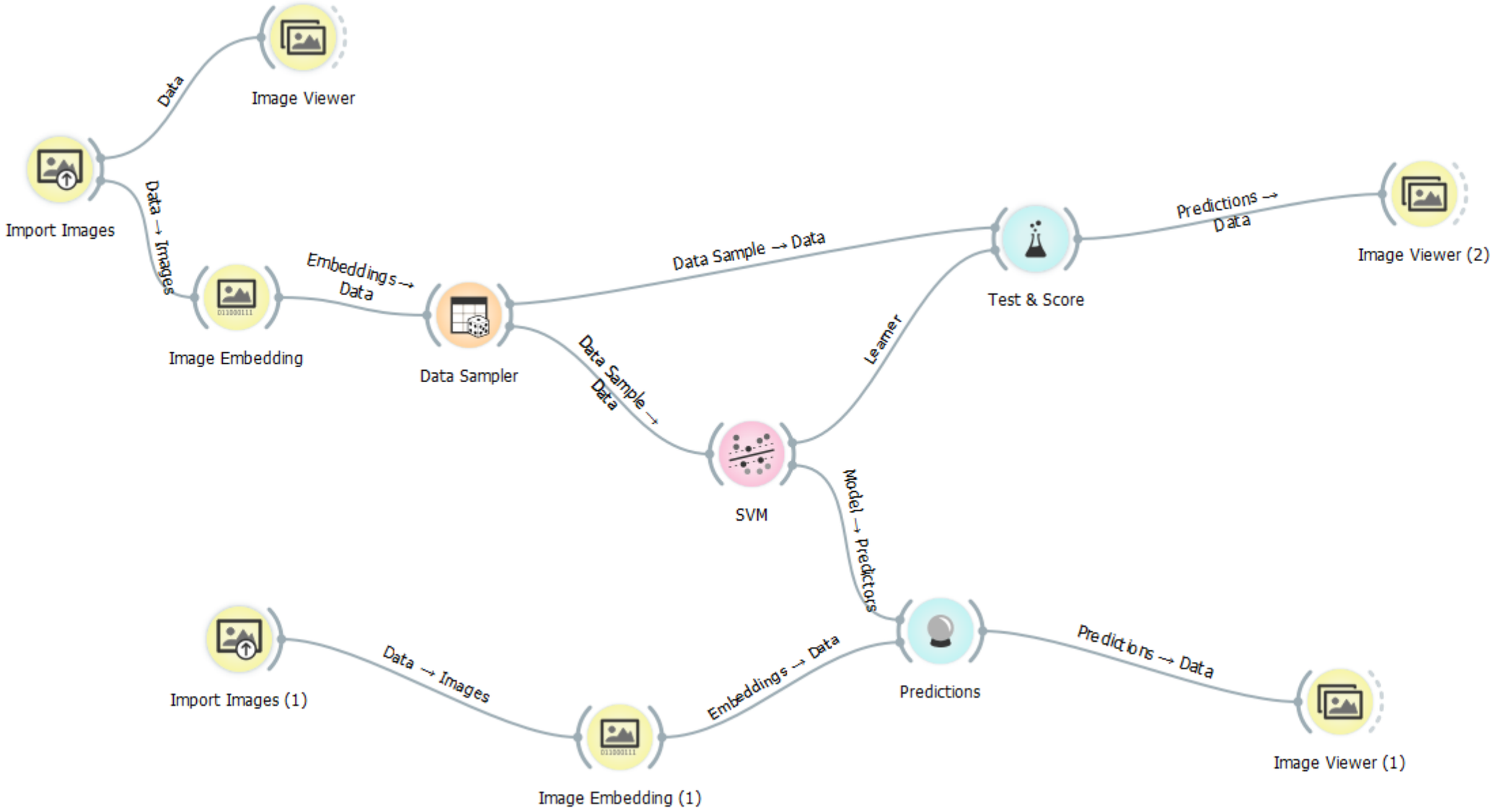
МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM) ИЗОБРАЖЕНИЯ



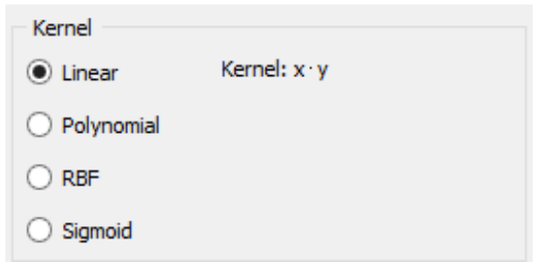
Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.852	0.803	0.770	0.833	0.803

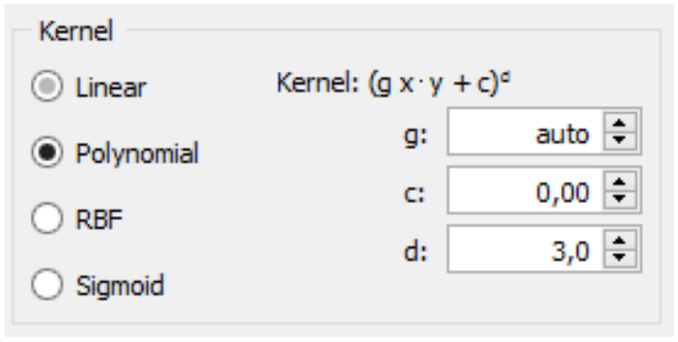
МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM) ИЗОБРАЖЕНИЯ - ПРЕДСКАЗАНИЯ



МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM)
ЭФФЕКТ IMAGE EMBEDDING

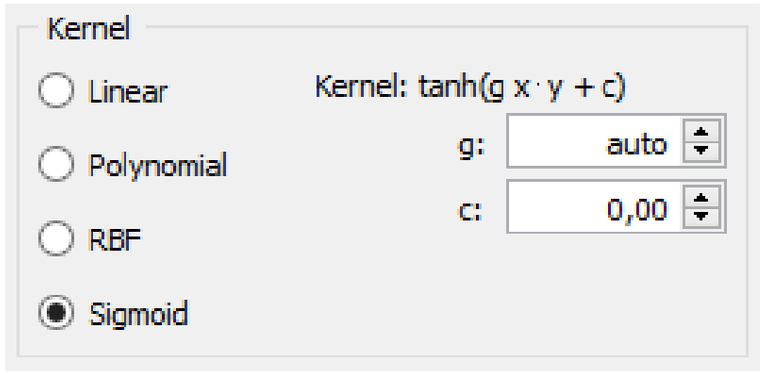


Method	AUC	CA	F1	Precision	Recall
SVM	0.999	0.917	0.914	0.926	0.917



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.272	0.240	0.236	0.612	0.240



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.956	0.883	0.880	0.897	0.883

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Бинарная логистическая регрессия, как следует из названия, применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения). Иными словами, с помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т.д.).

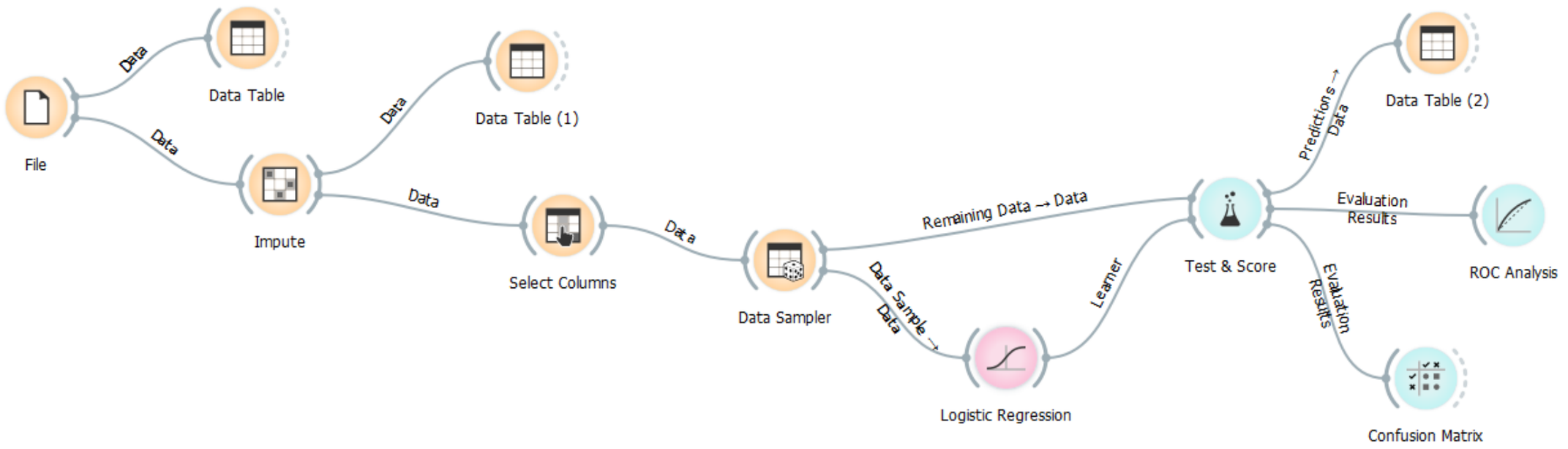
$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

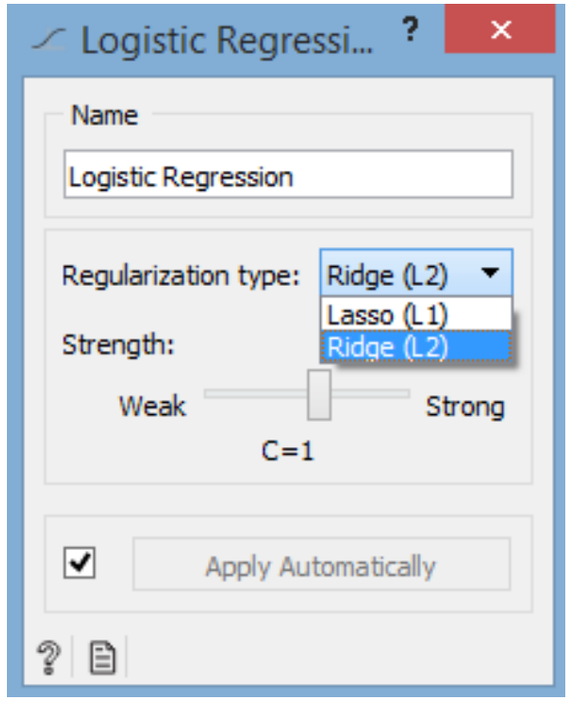
Для решения проблемы задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной, мы предсказываем непрерывную переменную со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением следующего регрессионного уравнения (логит-преобразование):

$$p = \frac{1}{1+e^{-y}},$$

где P – вероятность того, что произойдет интересующее событие; e – основание натуральных логарифмов 2,71…; y – стандартное уравнение регрессии.

LOGISTIC REGRESSION



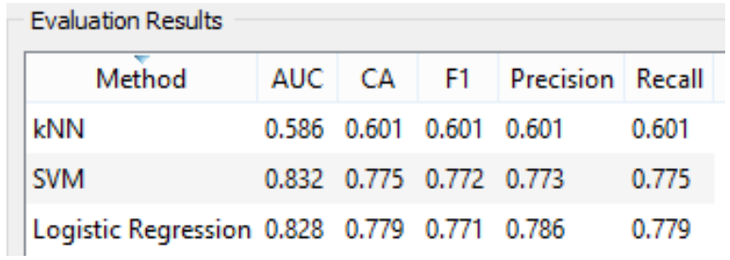


	Name	Ticket	Cabin	PassengerId	Survived
1	Braund, Mr. Ow...	A/5 21171	?	1.0	0.0
2	Cumings, Mrs. ...	PC 17599	C85	2.0	1.0
3	Heikkinen, Miss...	STON/O2. 3101...	?	3.0	1.0
4	Futrelle, Mrs. Ja...	113803.0	C123	4.0	1.0
5	Allen, Mr. Willia...	373450.0	?	5.0	0.0
6	Moran, Mr. Jam...	330877.0	?	6.0	0.0
7	McCarthy, Mr. ...	17463.0	E46	7.0	0.0
8	Palsson, Master...	349909.0	?	8.0	0.0
9	Johnson, Mrs. ...	347742.0	?	9.0	1.0
10	Nasser, Mrs. Ni...	237736.0	?	10.0	1.0
11	Sandstrom, Mis...	PP 9549	G6	11.0	1.0
12	Bonnell, Miss. E...	113783.0	C103	12.0	1.0
13	Saundercock, ...	A/5. 2151	?	13.0	0.0

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.828	0.779	0.771	0.786	0.779

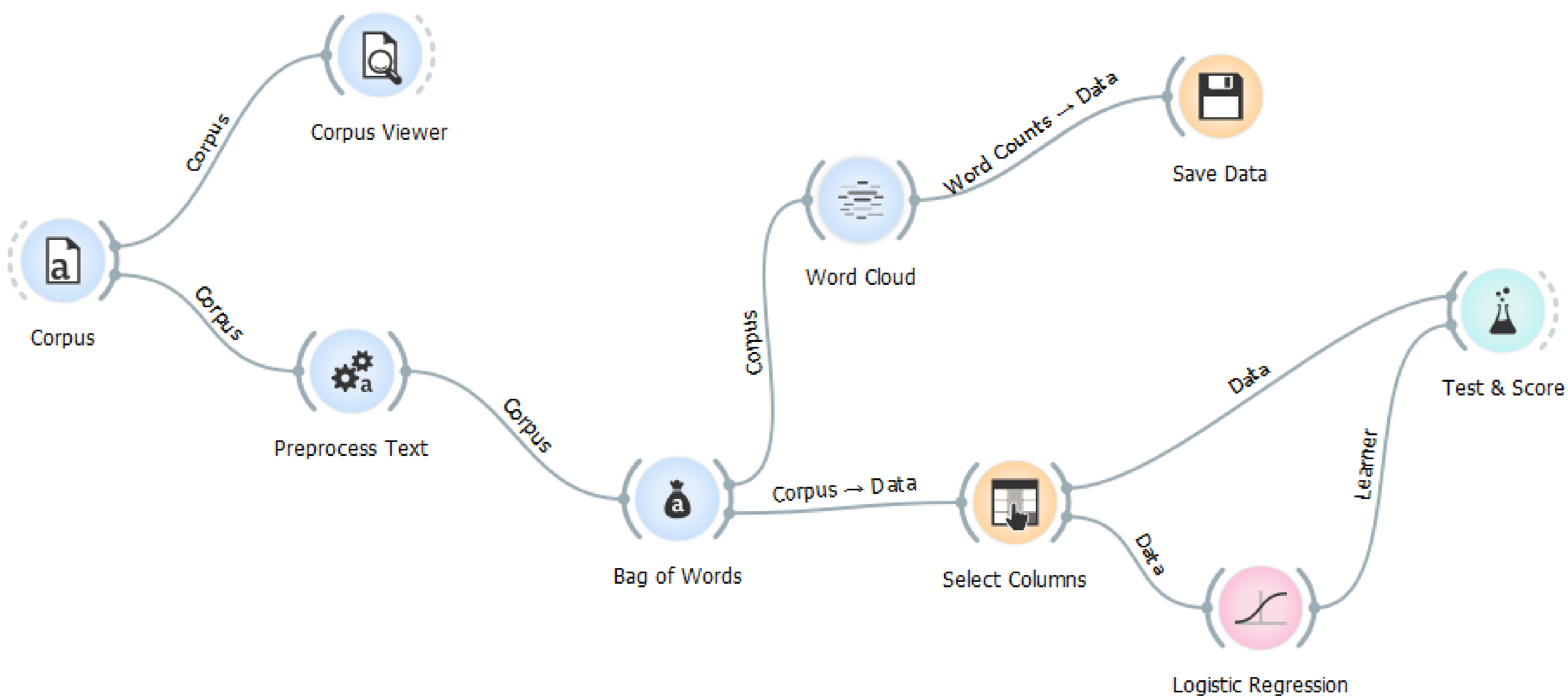
СРАВНЕНИЕ РАБОТЫ ТРЕХ КЛАССИФИКАТОРОВ



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.586	0.601	0.601	0.601	0.601
SVM	0.832	0.775	0.772	0.773	0.775
Logistic Regression	0.828	0.779	0.771	0.786	0.779

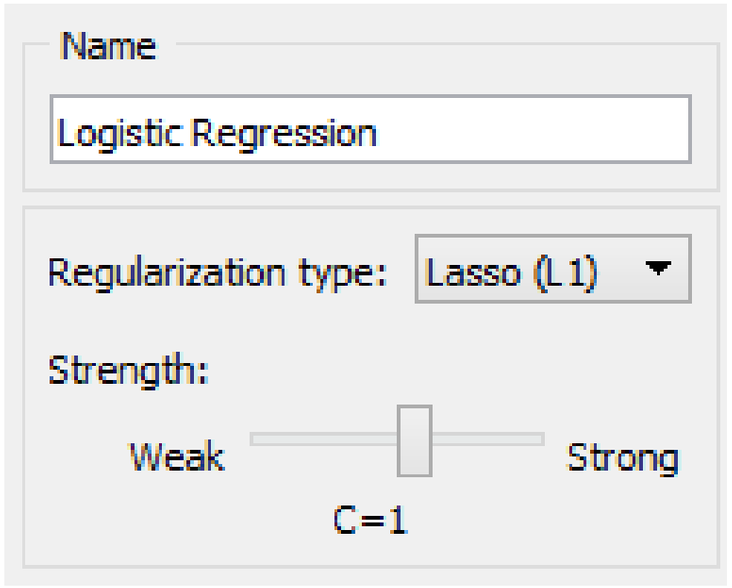
LOGISTIC REGRESSION – ТЕКСТОВЫЕ ДАННЫЕ

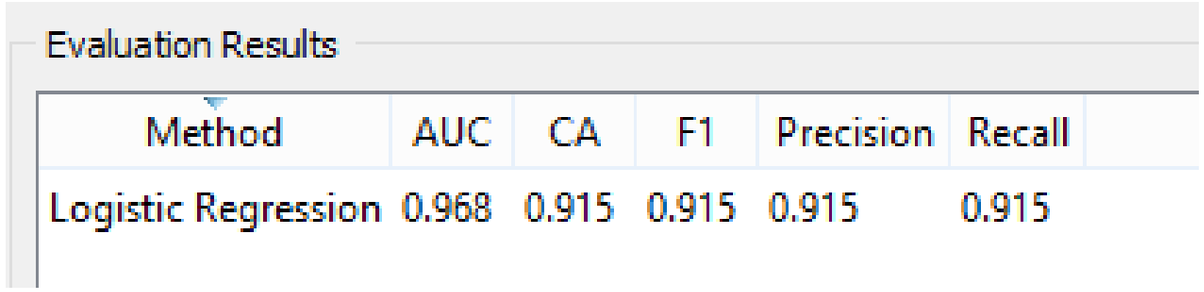


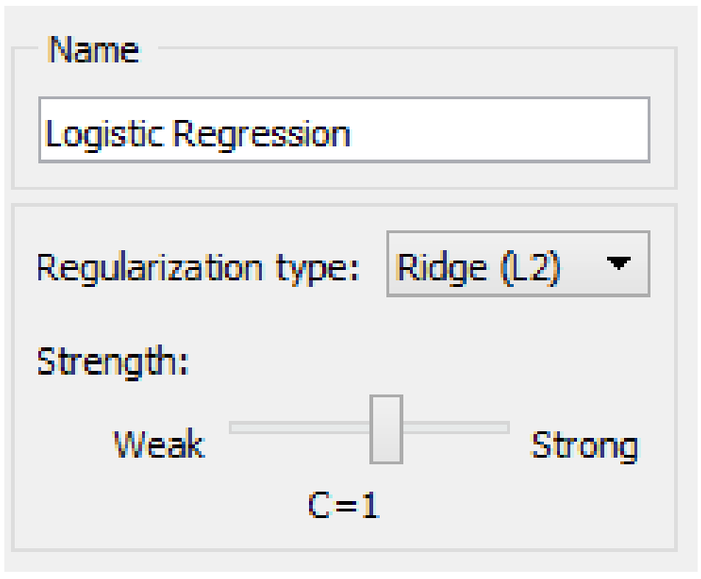
Evaluation Results

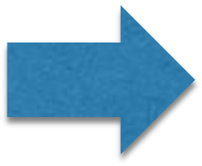
Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.997	0.956	0.956	0.956	0.956

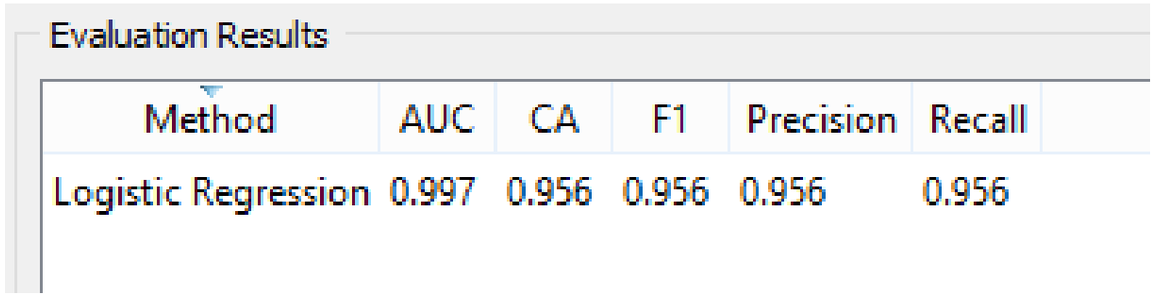
LOGISTIC REGRESSION – ЭФФЕКТ РЕГУЛЯРИЗАЦИИ



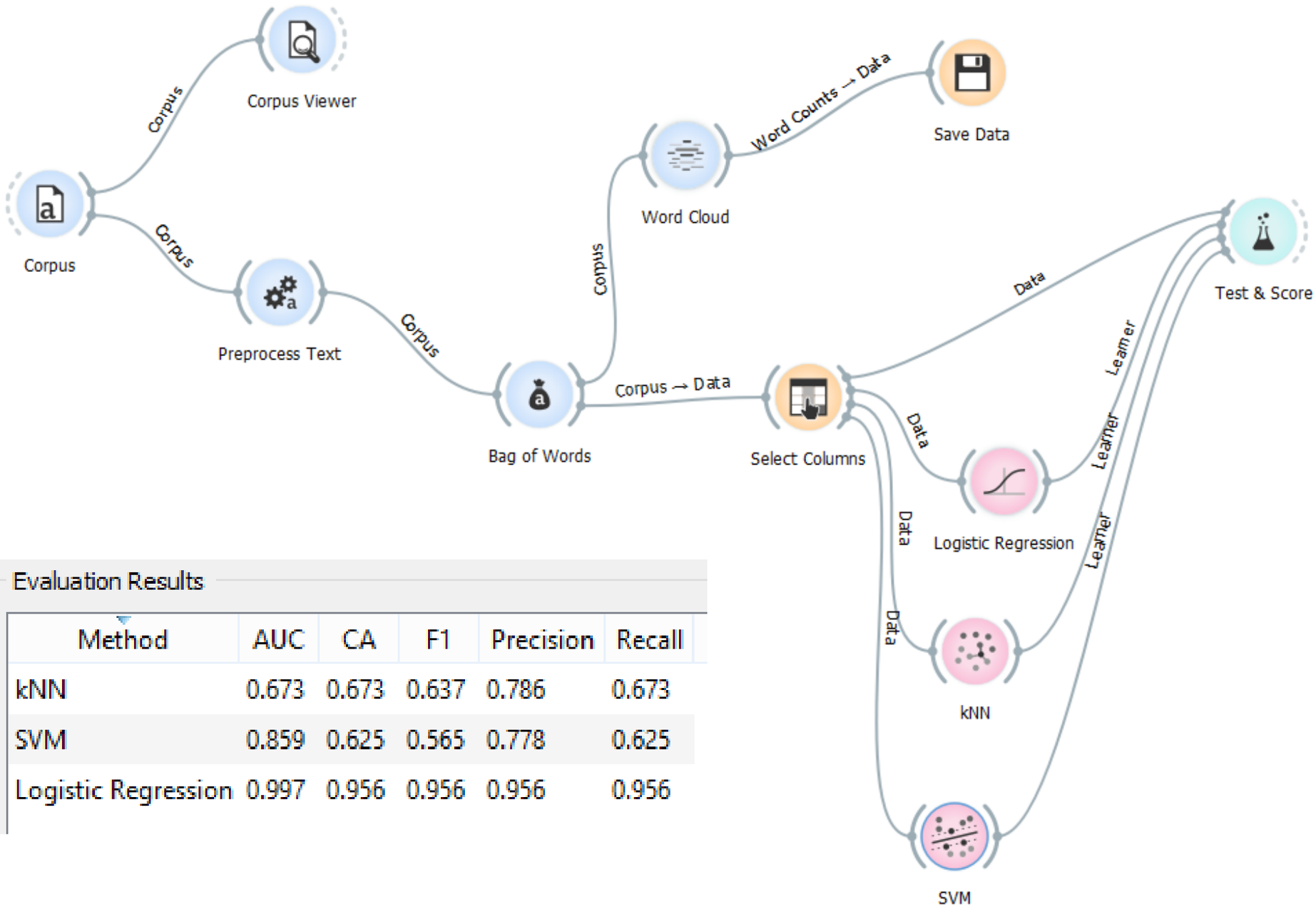








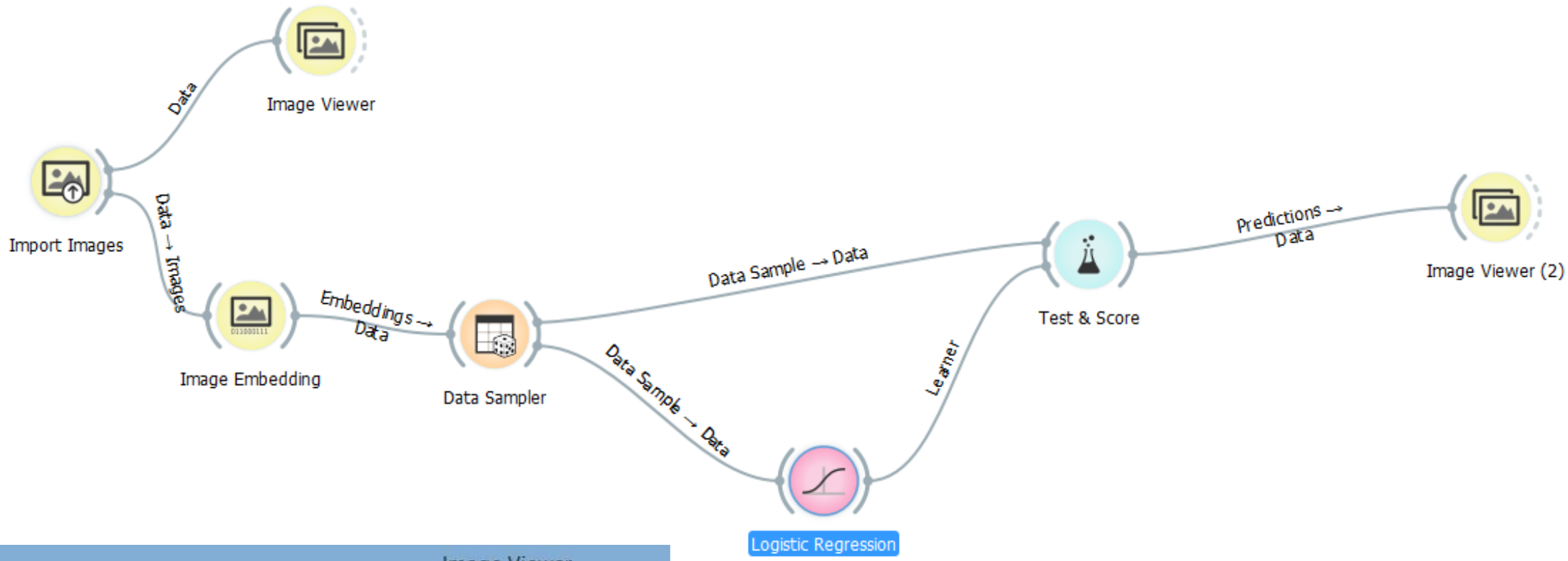
LOGISTIC REGRESSION – KNN – SVM



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.673	0.673	0.637	0.786	0.673
SVM	0.859	0.625	0.565	0.778	0.625
Logistic Regression	0.997	0.956	0.956	0.956	0.956

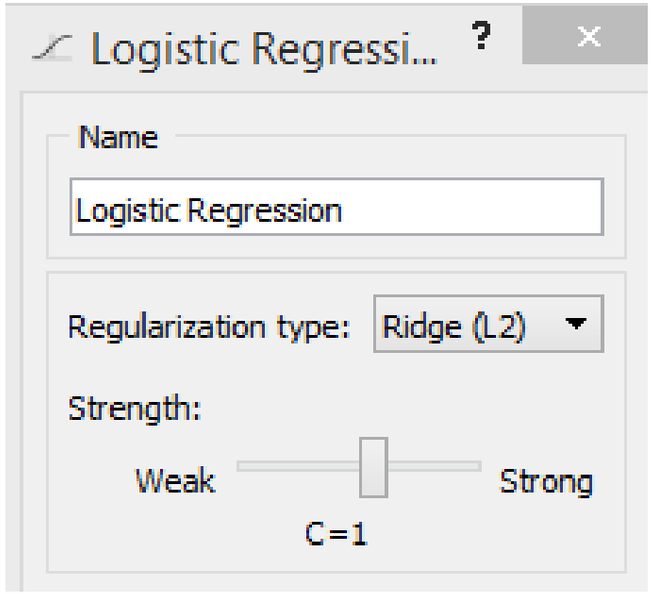
LOGISTIC REGRESSION – IMAGE CLASSIFICATION

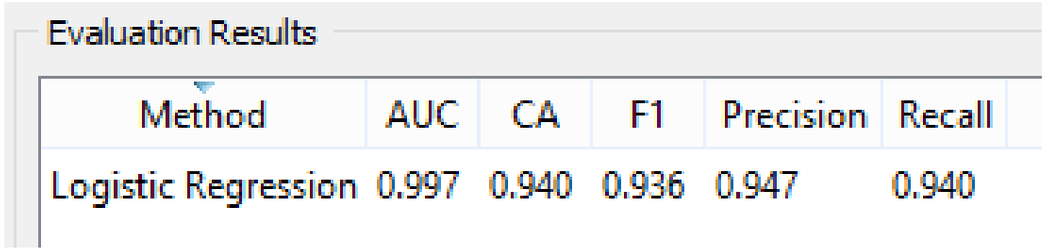


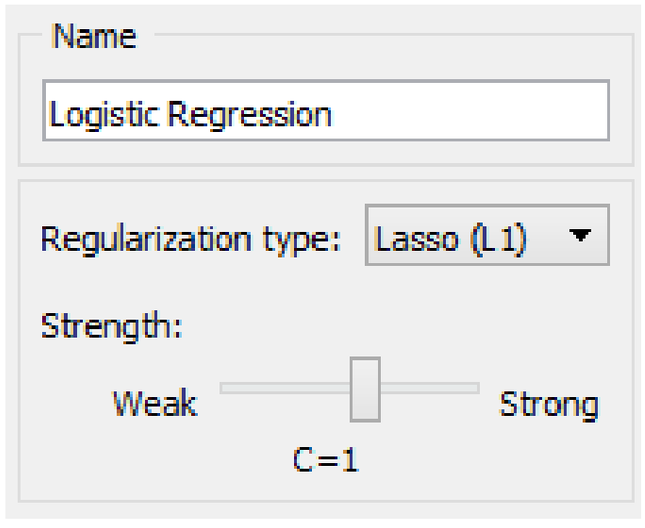
Evaluation Results

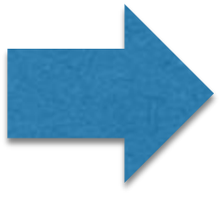
Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.997	0.940	0.936	0.947	0.940

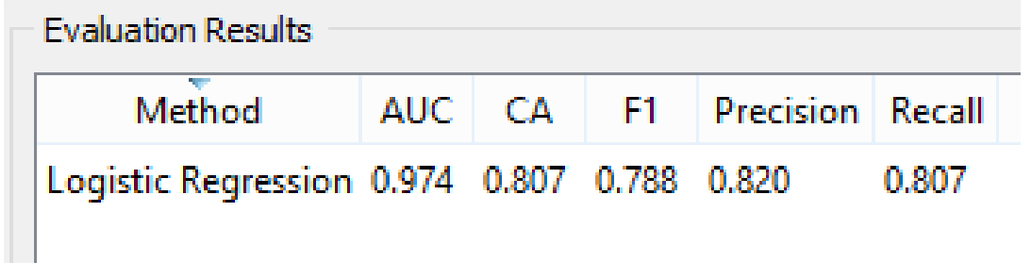
LOGISTIC REGRESSION – ЭФФЕКТ РЕГУЛЯРИЗАЦИИ



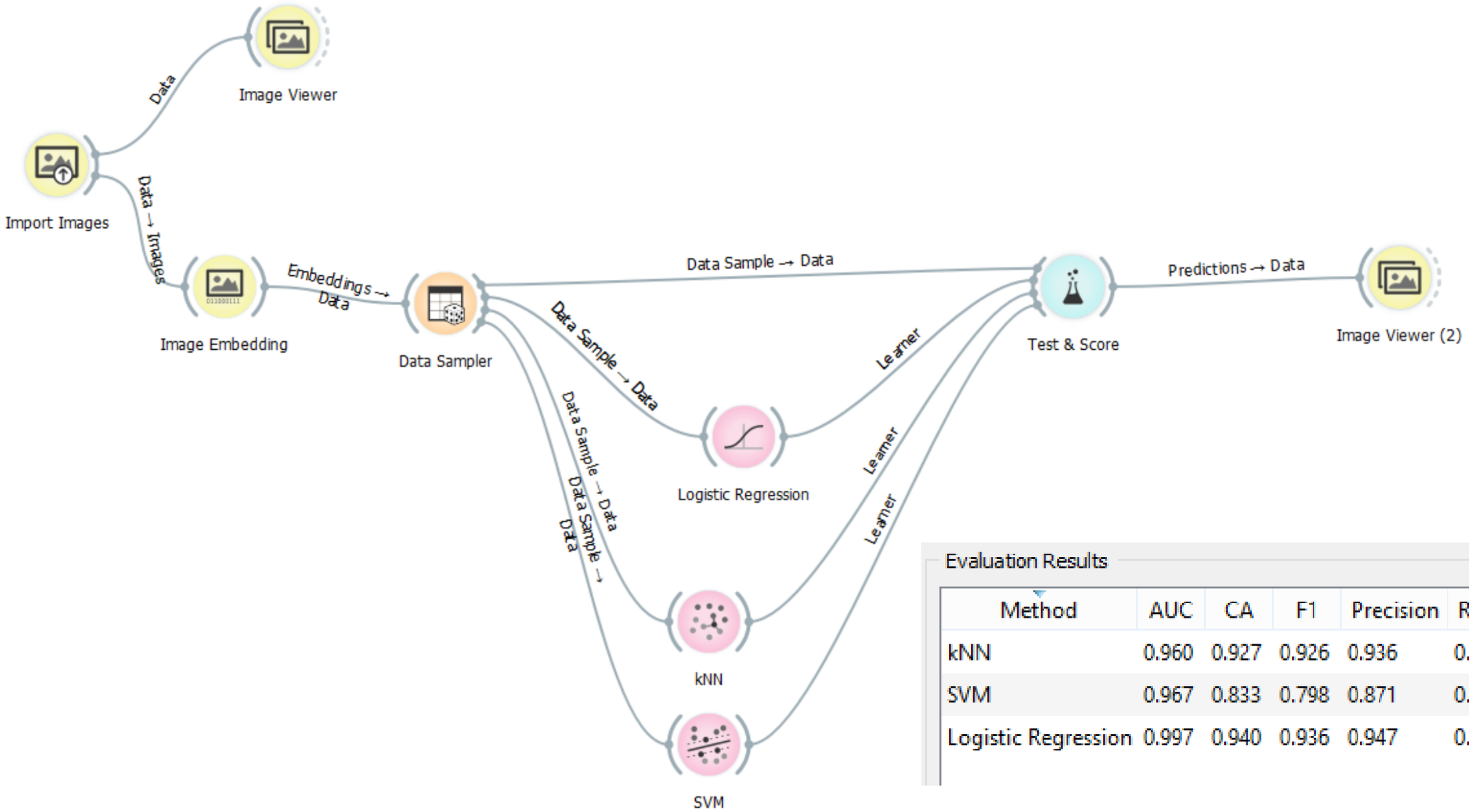








LOGISTIC REGRESSION – KNN - SVM



Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.960	0.927	0.926	0.936	0.927
SVM	0.967	0.833	0.798	0.871	0.833
Logistic Regression	0.997	0.940	0.936	0.947	0.940

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru