

Лабораторная работа №11. Анализ данных в среде Orange.

1. В данной лабораторной работе студенту предлагается ознакомиться с возможностями, предоставляемыми средой анализа данных Orange (<https://orangedatamining.com/>).
2. Необходимо установить предлагаемую среду, провести манипуляции над данными согласно заданию (начинается со следующей страницы) и оформить письменный отчет.
3. В качестве набора данных для анализа можно взять предлагаемый в самом задании набор данных либо любой другой по выбору студента, **позволяющий полностью выполнить описанные в задании действия.**
4. Отчет оформляется в свободной форме и должен содержать:
 - a) ФИО и группу студента;
 - b) Название дисциплины и лабораторной работы;
 - c) Краткую формулировку задания;
 - d) Краткую характеристику выбранного набора данных (где взят, какие объекты описывает и сколько их, какими признаками эти объекты обладают);
 - e) Пошаговое описание действий, **сопровожаемое скриншотами.**
5. Отчеты, в которых отсутствуют один или несколько элементов из списка выше, не засчитываются.
6. Данная лабораторная не предполагает очной защиты.

1 Знакомство с системой

Запуск системы сопровождается открытием стартового окна. Здесь пользователю предлагается создать новый проект, открыть существующий или выбрать из недавних; просмотреть видеоуроки, заранее приготовленные примеры или перейти на страницу системы на портале (Рисунок 1).

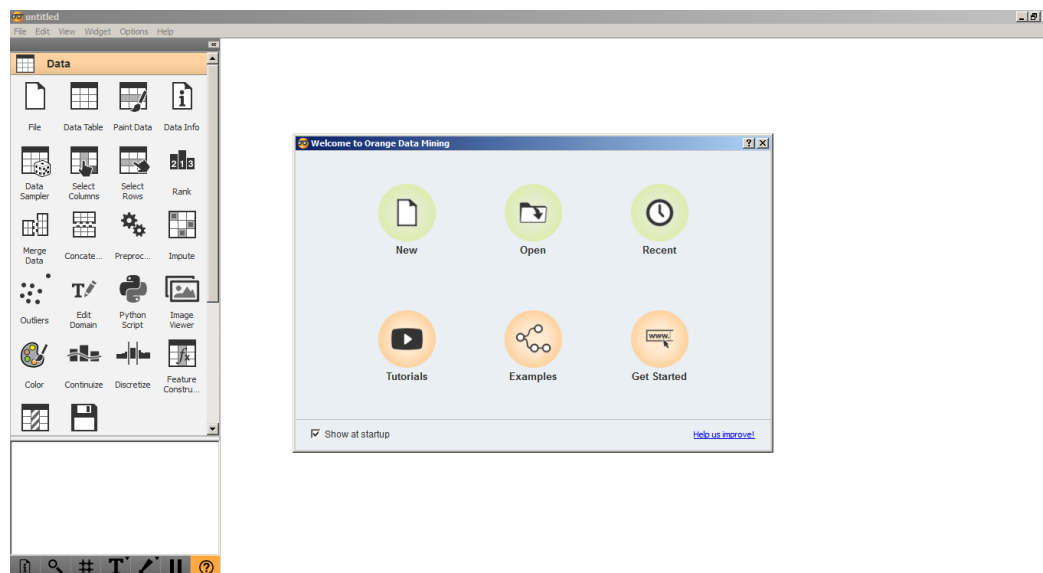


Рисунок 1. Стартовое окно для запуска системы

1.1 Миссия «Выжить на Титанике»

В рамках знакомства с системой рассмотрим небольшой проект. На основе данных пассажиров Титаника необходимо построить модель для предсказания вероятности выживания в зависимости от класса пассажира, пола и возраста.

Для начала необходимо создать новый проект (File ☰ New). Открывшееся окно программы состоит из двух областей: область виджетов, являющихся, по своей сути, рабочими инструментами, и рабочей области (канваса), на которой осуществляется взаимодействие виджетов с данными.

Для построения модели требуется:

- 1 Загрузить данные. Для этого нужно перетащить виджет File в рабочую область.

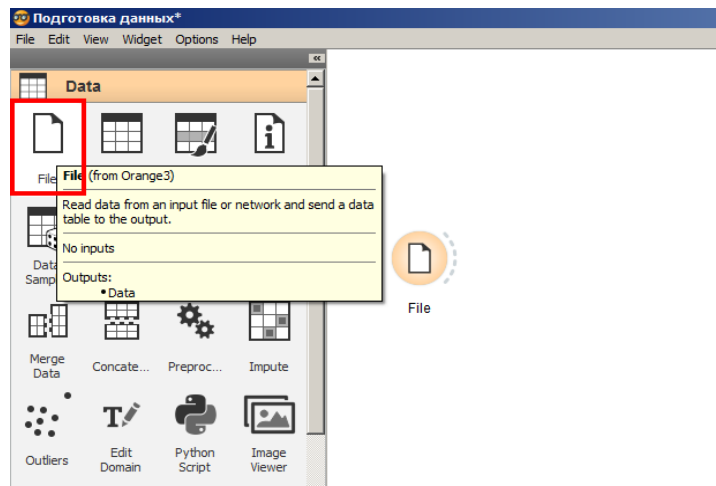


Рисунок 2. Виджет для загрузки данных

2 Дважды нажать на виджет и выбрать требуемый набор данных (датасет) (Рисунок 3).

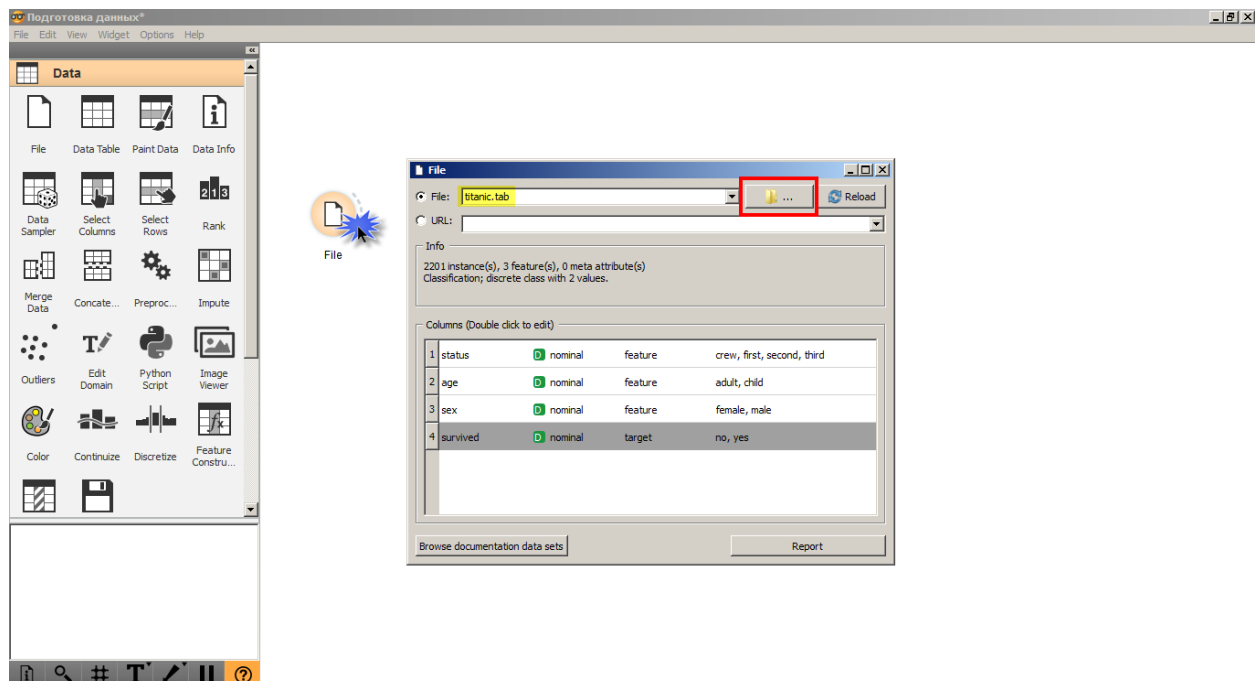


Рисунок 3. Выбор набора данных

Обратите внимание, что Orange автоматически определил типы данных и вид элементов.

Поскольку поля класс пассажира (status), возраст и пол являются элементами описания, им были присвоены значение feature (свойство). Поле выживший (survived) является расчетным значением модели, поэтому имеет вид target. Для изменения типа данных и вида элементов требуется дважды нажать на поле и выбрать значение в выпадающем списке.

3 Для просмотра значений в таблице воспользуемся виджетом Data table. Для передачи загруженных данных из виджета File возможны два варианта:

3.1 Перетащить виджет Data table из панели виджетов и соединить с виджетом File.

3.2 Протянуть линию из виджета File на пустое место в рабочей области и выбрать в появившемся окне требуемый виджет (Рисунок 4).

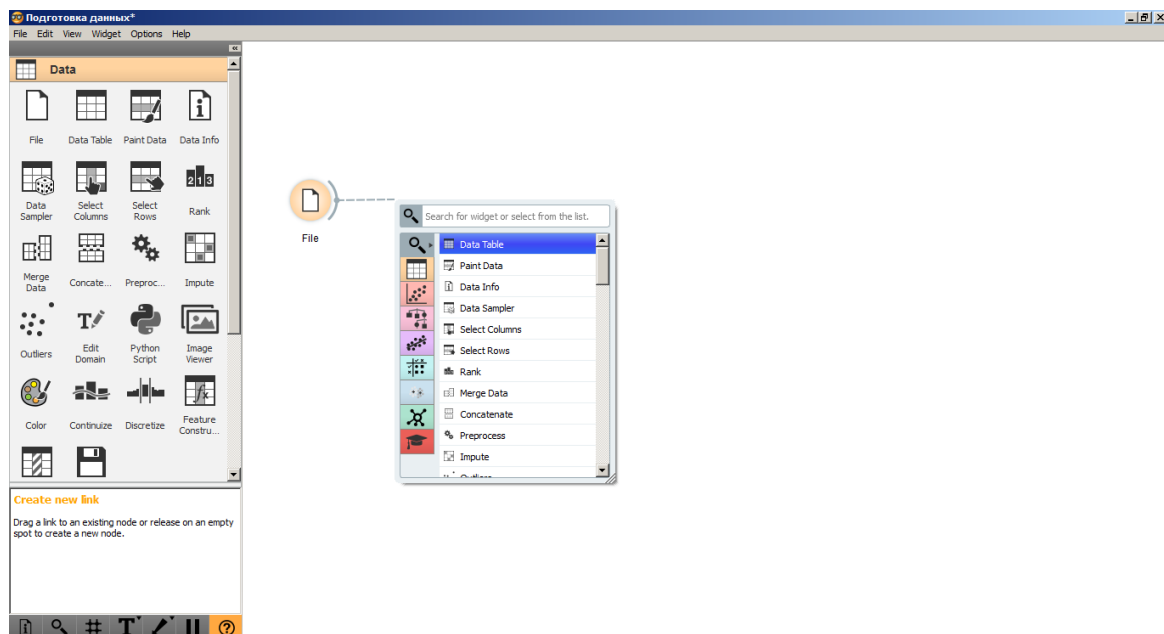


Рисунок 4. Просмотр набора данных в табличном варианте

- 4 По двойному нажатию на виджет Data table будет открыта таблица, содержащая загруженные данные (Рисунок 5).

	survived	status	age	sex
1	yes	first	adult	male
2	yes	first	adult	male
3	yes	first	adult	male
4	yes	first	adult	male
5	yes	first	adult	male
6	yes	first	adult	male
7	yes	first	adult	male
8	yes	first	adult	male
9	yes	first	adult	male
10	yes	first	adult	male
11	yes	first	adult	male
12	yes	first	adult	male
13	yes	first	adult	male
14	yes	first	adult	male
15	yes	first	adult	male
16	yes	first	adult	male
17	yes	first	adult	male
18	yes	first	adult	male
19	yes	first	adult	male
20	yes	first	adult	male
21	yes	first	adult	male
22	yes	first	adult	male
23	yes	first	adult	male
24	yes	first	adult	male
25	yes	first	adult	male
26	yes	first	adult	male
27	yes	first	adult	male
28	yes	first	adult	male
29	yes	first	adult	male
30	yes	first	adult	male
31	yes	first	adult	male

Рисунок 5. Табличное представление набора данных

- 5 Для построения наглядной модели зависимости вероятности выживания от класса, пола и возраста воспользуемся виджетом Sieve Diagram (решето). Для этого необходимо расположить виджет на канвасе и соединить с виджетом File, содержащим загруженные данные. Дважды нажмите на виджет Sieve Diagram для просмотра результатов.
- 6 Попробуйте изменить комбинацию атрибутов. Для этого воспользуйтесь командой «Score Combinations», затем выберите один из предложенных вариантов (Рисунок 6).

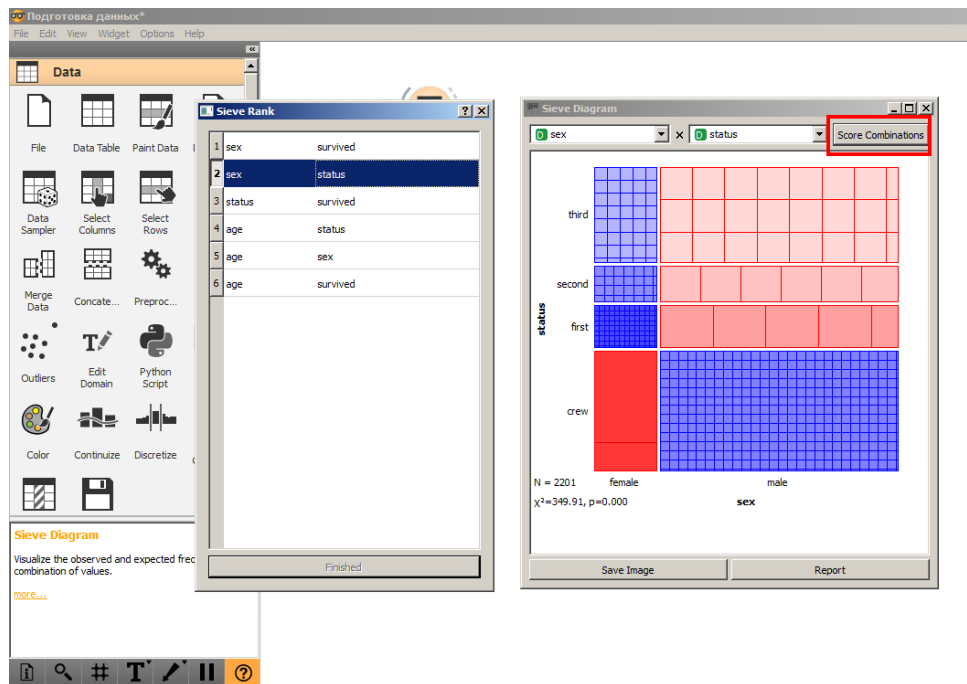


Рисунок 6. Выбор комбинации атрибутов

Задание: определите наиболее информативную комбинацию атрибутов. Сделайте выводы о полученных результатах моделирования.

2 Примеры виджетов

2.1 Очистка от выбросов

- 1 Загрузите набор данных «Iris» при помощи виджета File. Просмотрите данные в табличном варианте.
- 2 Для обнаружения выбросов необходимо воспользоваться виджетом «Outliers». Расположите виджет в рабочей области и соедините с загруженными данными. Настройки для виджета приведены на Рисунок 7.

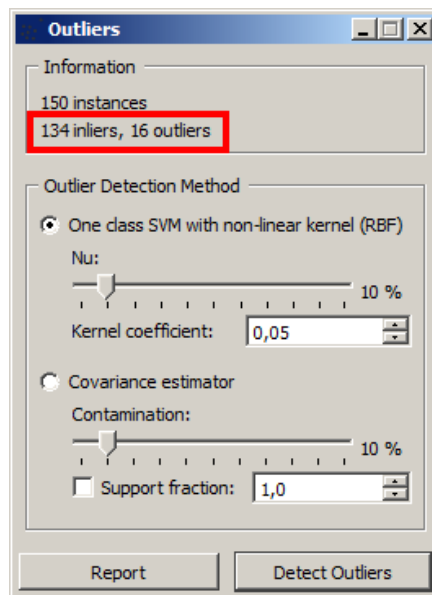


Рисунок 7. Настройки для виджета outliers

- 3 Для просмотра результатов работы виджета воспользуемся сразу несколькими инструментами. Во-первых, посмотрим строки с выбросами в таблице. Для этого требуется соединить виджет «Выбросы» с виджетом «Таблица». После двойного щелчка на созданную связь появится окно для редактирования (Рисунок 8).

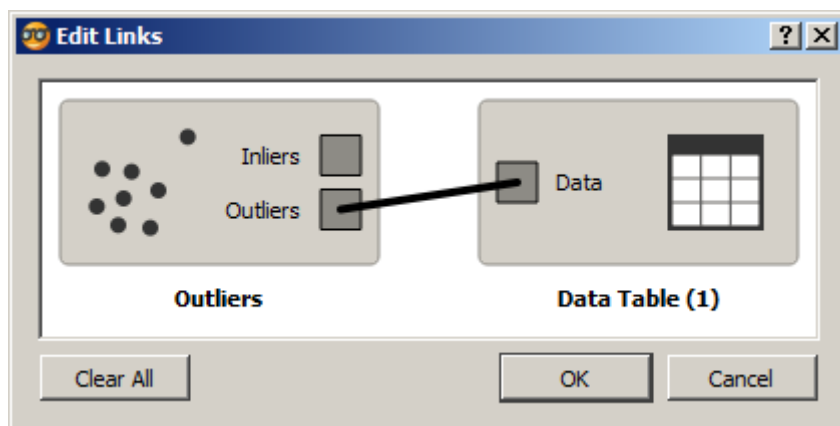


Рисунок 8. Окно редактирования связи

- 4 Во-вторых, сравним исходные и очищенные данные при помощи инструмента Scatter plot. Соедините виджет File и Scatter plot, затем протяните связь от Outliers к тому же Scatter plot. Настройки связи: Data – Data и Inliers – Data subset. Таким образом, будут цветом будут выделены только те элементы датасета, которые не являются выбросами (Рисунок 9).

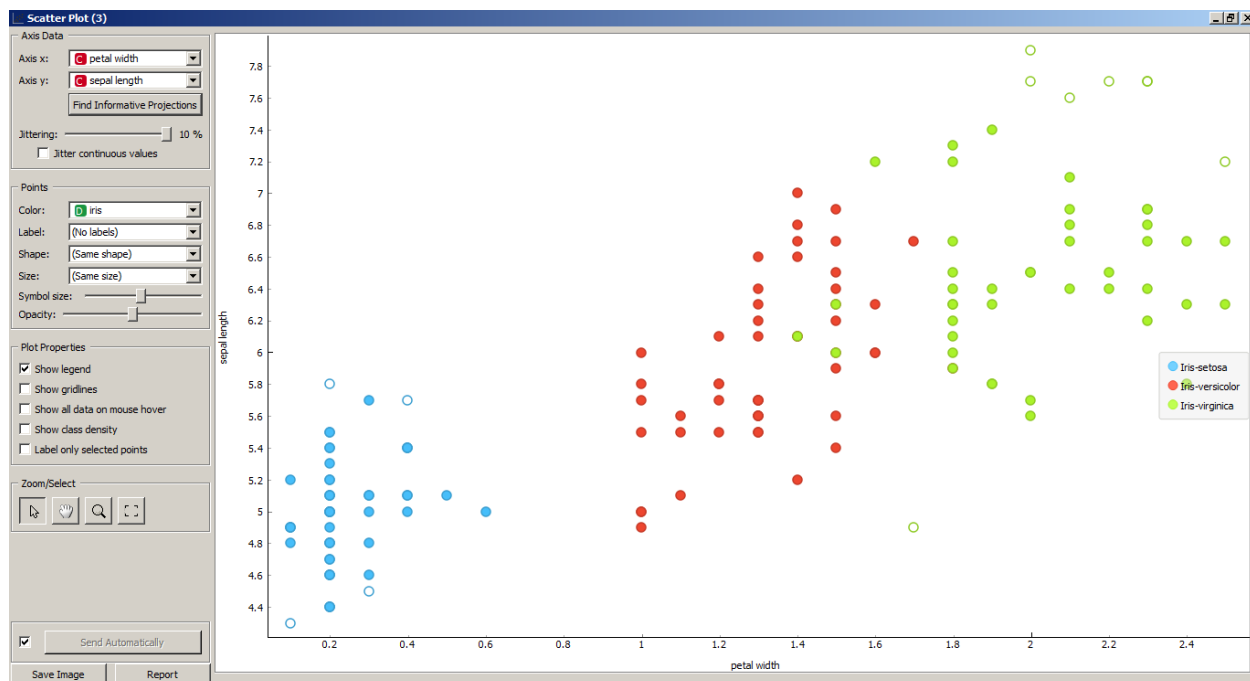


Рисунок 9. Виджет Scatter plot

- 5 Помимо виджета Scatter plot для наглядности воспользуемся виджетом Linear Projection. Для просмотра выбросов соедините виджет File и Linear Projection, а затем протяните связь от Outliers к тому же Linear Projection. Настройки связи: Data – Data и Outliers – Data subset. Таким образом, будут цветом будут выделены только те элементы набора, которые являются выбросами (Рисунок 10).

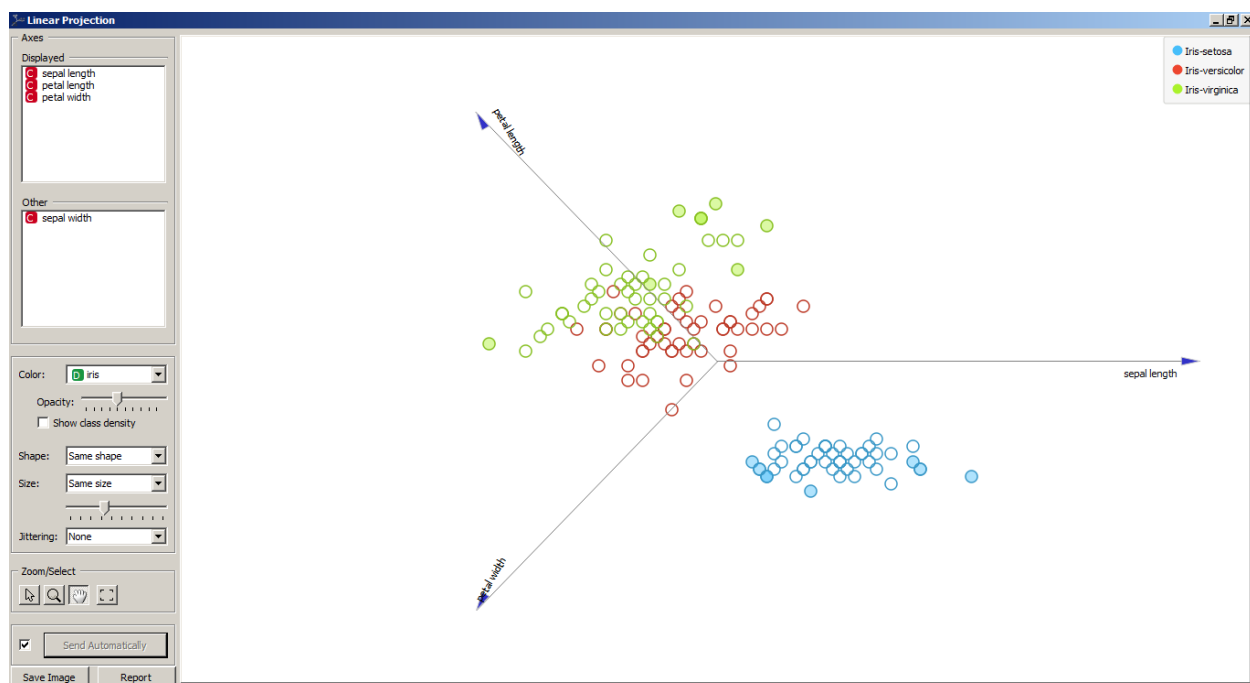


Рисунок 10. Виджет Linear Projection

Задание: Выделите цветом в виджете Scatter plot элементы, являющиеся выбросами. Выделите цветом в виджете Linear Projection элементы, не являющиеся выбросами. Измените в виджете Linear Projection выбранные оси координат.

2.2 Виджет Classification

Данный раздел содержит пример решения задачи классификации с учителем. Для этого требуется:

- 1 Загрузить набор данных «fruits-and-vegetables-train» и соединить с виджетом Classification Tree. В системе предусмотрена возможность распознавания не только специально подготовленные форматы файлов, но и таблицы в Excel.
- 2 Выявленная ошибка означает, что виджет не может обнаружить целевого (target) признака для классификации. Вернитесь к виджету File и проверьте настройки полей, обратив внимание на их виды. Поле classification указано как свойство (Feature), однако в данном случае оно является целевой величиной. Измените вид на соответствующий, а затем примените изменения (Apply) (Рисунок 11).

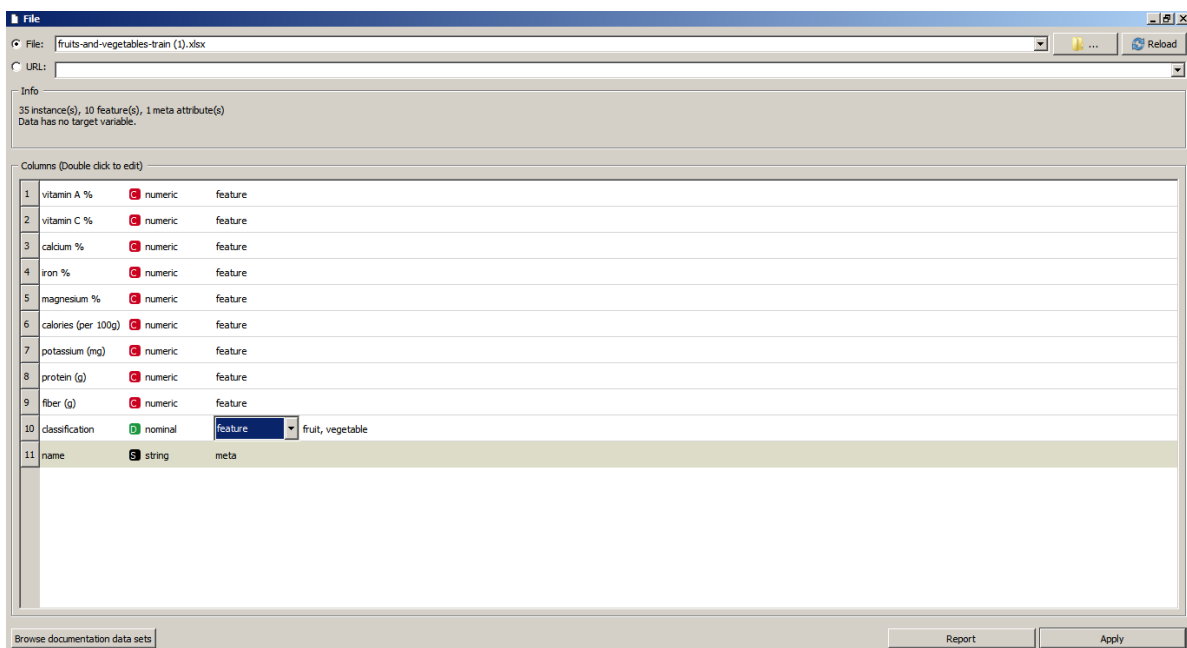


Рисунок 11. Изменение видов полей

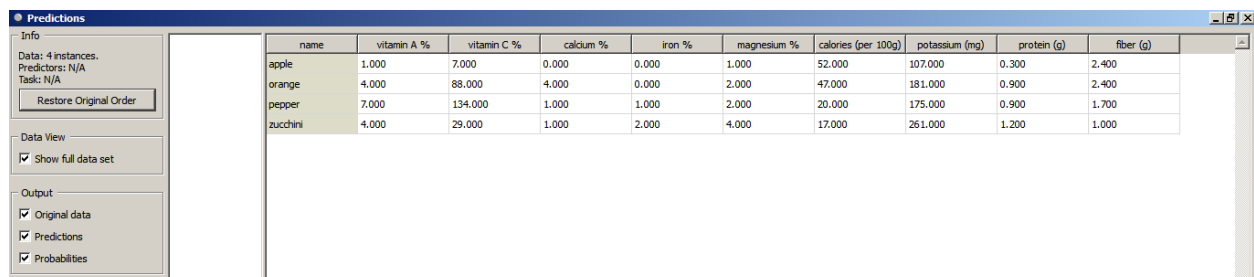
- 3 Для отображения результатов классификации следует воспользоваться виджетом Classification Tree Viewer. Для этого требуется соединить его с виджетом Classification Tree.

Задание: Просмотрите результаты и сделайте выводы, какие свойства являются основными для классификации.

2.3 Виджет Predictions

Помимо непосредственной классификации программа предусматривает возможность прогнозирования на основе созданной модели. Исследуем виджет Predictions на основе Набора данных «test». Данный набор содержит те же самые поля, что и «fruits-and-vegetables-train», однако в нем отсутствует целевая величина.

- 1 Соедините виджет file (с загруженным тестовым набором данных) и виджет Predictions. Данные прогноза отсутствуют, поскольку не была создана модель (учитель), согласно которой виджет прогнозирования смог бы осуществить классификацию (Рисунок 12).



name	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (per 100g)	potassium (mg)	protein (g)	fiber (g)
apple	1.000	7.000	0.000	0.000	1.000	52.000	107.000	0.300	2.400
orange	4.000	88.000	4.000	0.000	2.000	47.000	181.000	0.900	2.400
pepper	7.000	134.000	1.000	1.000	2.000	20.000	175.000	0.900	1.700
zucchini	4.000	29.000	1.000	2.000	4.000	17.000	261.000	1.200	1.000

Рисунок 12. Данные прогноза отсутствуют (слева от таблицы)

- 2 Для обучения виджета Predictions необходимо предоставить ему обучающую модель: соединить его с виджетом Classifications Tree (Рисунок 13).

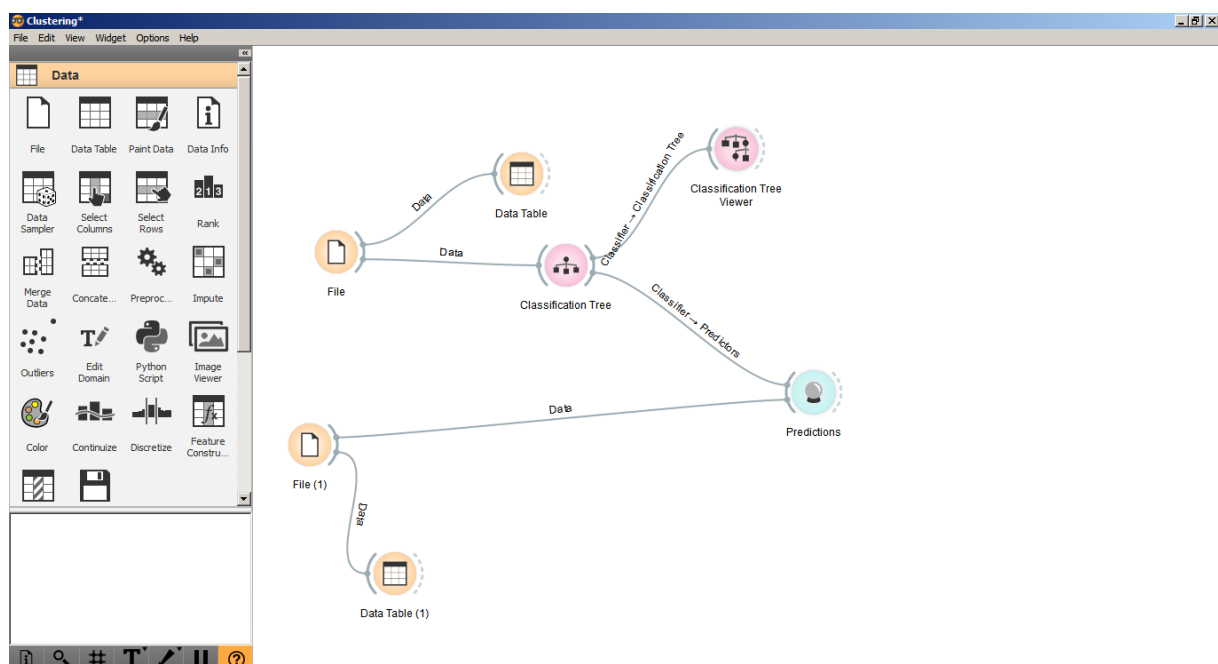


Рисунок 13. Обучение виджета

- 3 После обучения в виджете Predictions появятся прогнозы (Рисунок 14).

Predictions

Info
Data: 4 instances.
Predictors: 1
Task: Classification
[Restore Original Order](#)

Options (classification)
☒ Show predicted class
☒ Show predicted probabilities
fruit
vegetable
☒ Draw distribution bars

Data View
☒ Show full data set

Output
☒ Original data
☒ Predictions
☒ Probabilities

	Classification Tree	name	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (per 100g)	potassium (mg)	protein (g)	fiber (g)
1	1.00 : 0.00 → fruit	apple	1.000	7.000	0.000	0.000	1.000	52.000	107.000	0.300	2.400
2	1.00 : 0.00 → fruit	orange	4.000	88.000	4.000	0.000	2.000	47.000	181.000	0.900	2.400
3	0.67 : 0.33 → fruit	pepper	7.000	134.000	1.000	1.000	2.000	20.000	175.000	0.900	1.700
4	0.00 : 1.00 → vegetable	zucchini	4.000	29.000	1.000	2.000	4.000	17.000	261.000	1.200	1.000

Рисунок 14. Результаты прогнозирования

Задание: проведите обучение виджета Predictions на основе виджета Logistic Regression. Сравните полученные результаты.

2.4 Виджет Test & Score

В предыдущей задаче мы выяснили, что разные методы классификации могут давать разные результаты. Поэтому одной из важных задач является сравнение результатов классификации, полученных при использовании различных методов.

Для сравнения воспользуемся набором данных «Iris». Сравним результаты работы виджетов Logistic Regression, Random Forrest Classification, SVM при помощи виджета Test & Score (Рисунок 15).

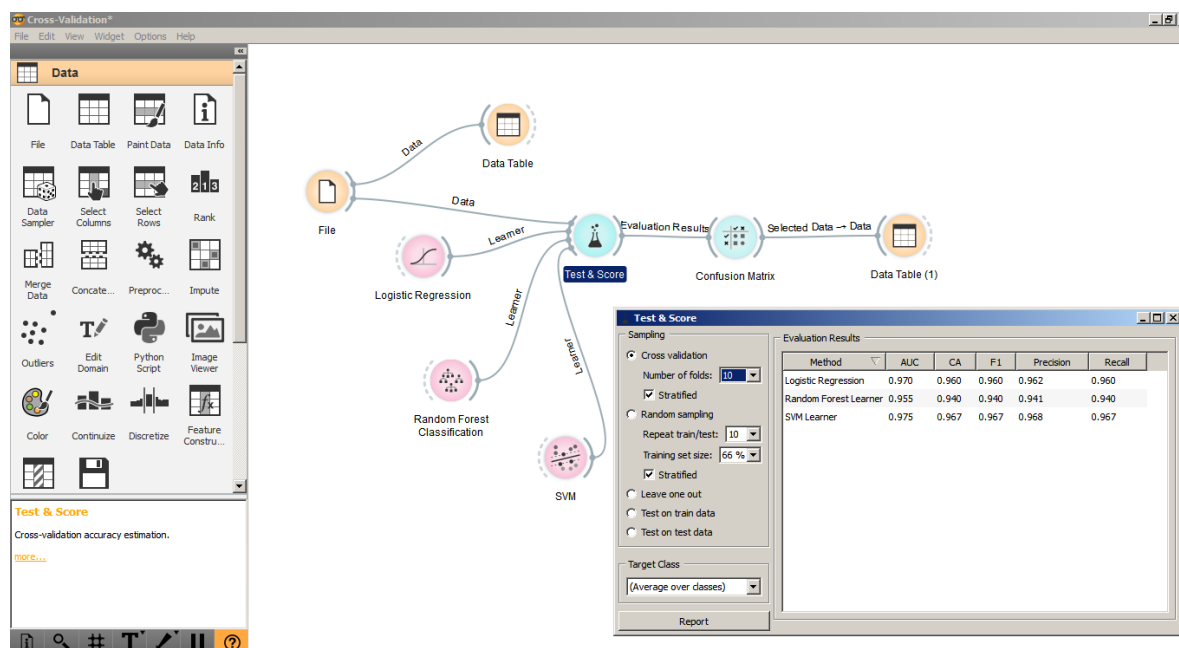


Рисунок 15. Сравнение результатов

Ни один из предложенных методов классификации не имеет точность 100% (CA – classification accuracy), что значит, что часть элементов были классифицированы неправильно. Для просмотра неверно классифицированных элементов следует воспользоваться виджетом Confusion Matrix (Рисунок 16).

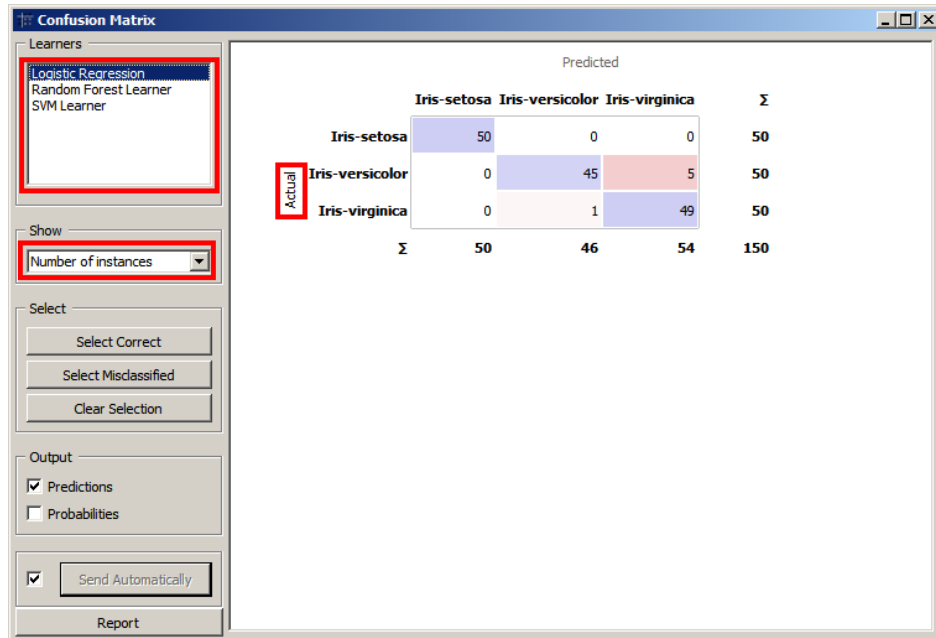


Рисунок 16. Неверно классифицированные данные

Задание: при помощи виджета установите какой процент каждого из видов ирисов был классифицирован неверно (подсказка — Show). Затем выведите в таблицу все неверно классифицированные элементы.

2.5 Виджет PCA

В рассмотренных выше примерах были использованы наборы данных, имеющие низкую размерность. Однако довольно часто встречаются наборы данных, имеющие большое число атрибутов. В таком случае перед аналитиками встает задача уменьшить размерность данных, потеряв наименьшее количество информации. Для решения данной задачи в среде Orange воспользуемся инструментом PCA – principal component analysis (Метод главных компонент).

В качестве набора данных в рамках данной задачи будет использован набор «Wine», имеющий 13 измерений. Загрузите датасет и соедините его с виджетом PCA.

В окне настроек виджета требуется обратить внимание на число компонент и процент данных, верно классифицируемых при использовании данного числа компонент (Рисунок 17).

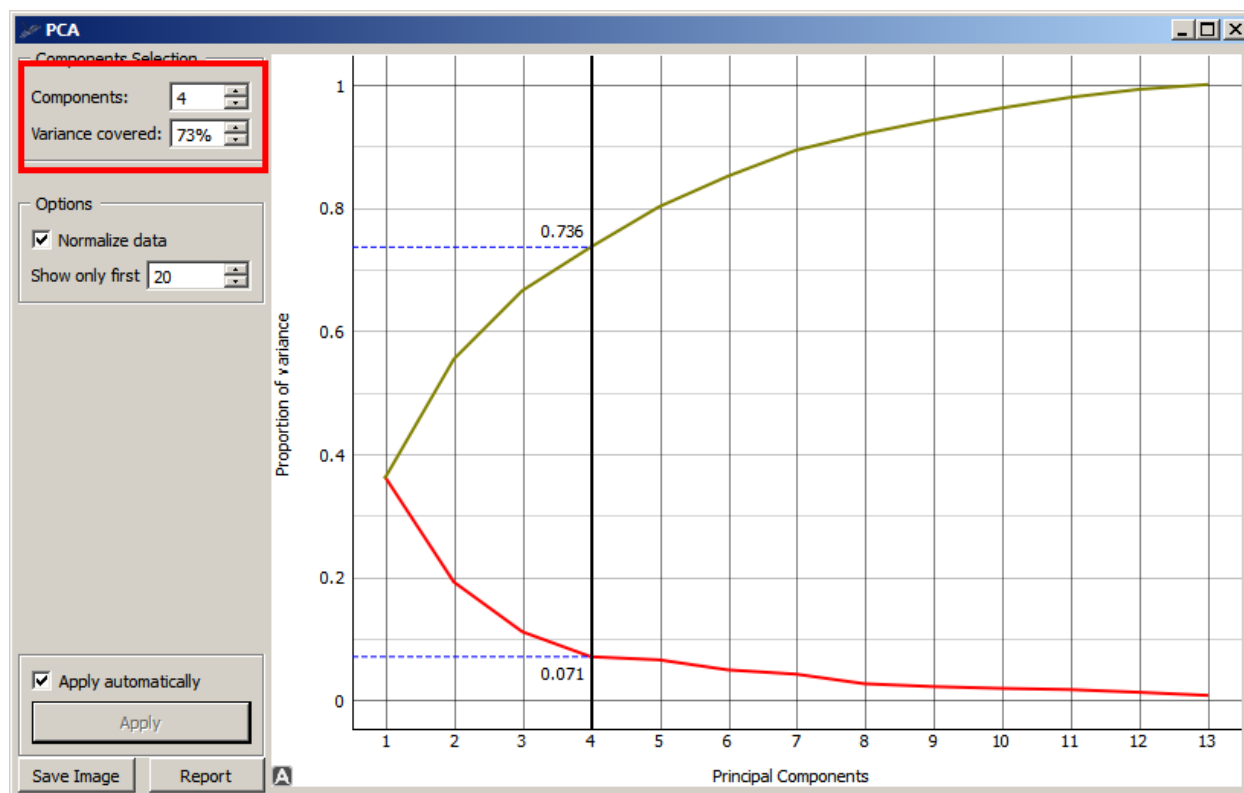


Рисунок 17. Окно настроек виджета Principal Component Analysis

Задание: при помощи виджета установите добейтесь значения **Variance Covered** равного **80%**.