Лабораторная работа

# Деревья решений в Deductor

**Исходные данные:**

Файлы train.csv, test.csv из соревнования Kaggle "Titanic: Machine Learning from Disaster" <https://www.kaggle.com/c/titanic>

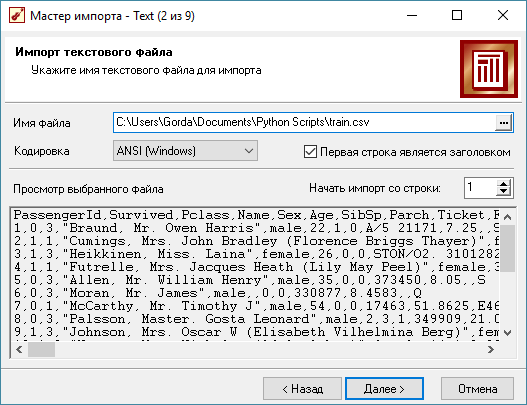
**Описание полей:**

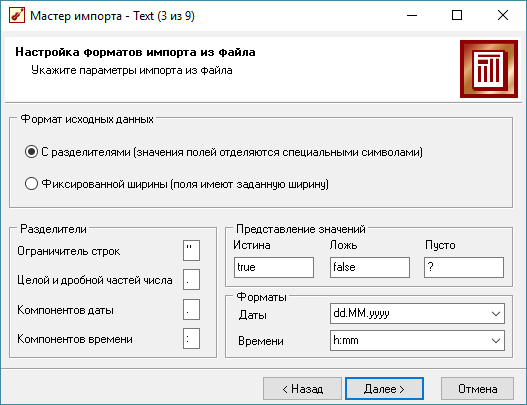
|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Признак выживания | 0 = No, 1 = Yes |
| pclass | Класс билета | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Пол |  |
| Age | Возраст |  |
| sibsp | Кол-во братьев и сестер / супругов на борту Титаника |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

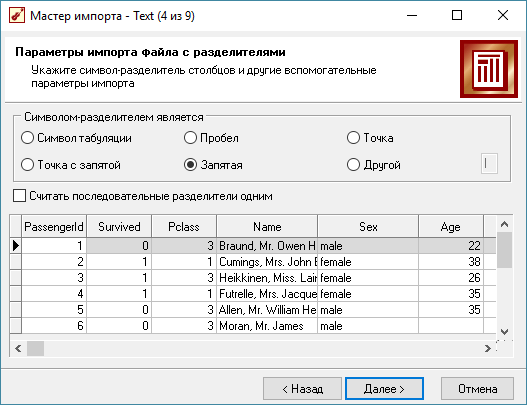
Порядок действий:

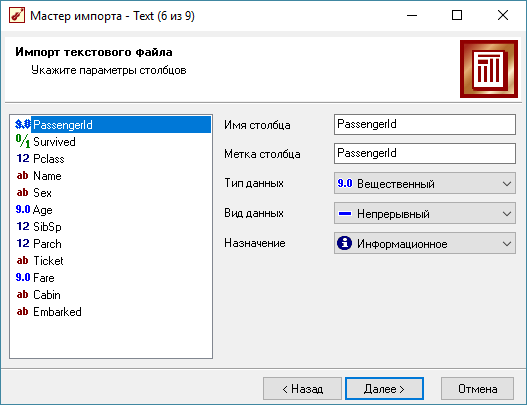
Загрузка обучающих данных, определение типов полей, заполнение пропусков, построение дерева решения (целевое поле survived), загрузка тестовых данных, применение полученной модели на тестовых данных.

1. Загрузка обучающих данных

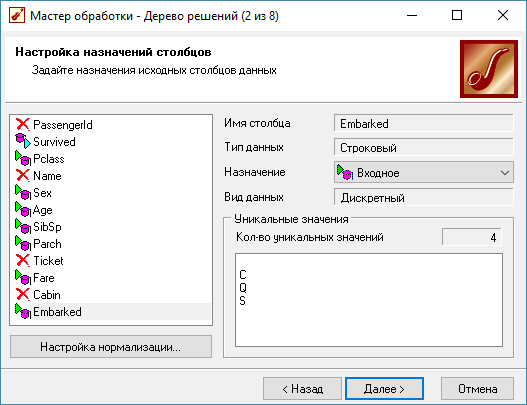


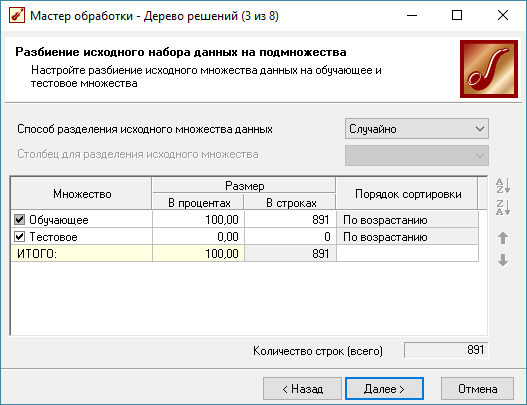




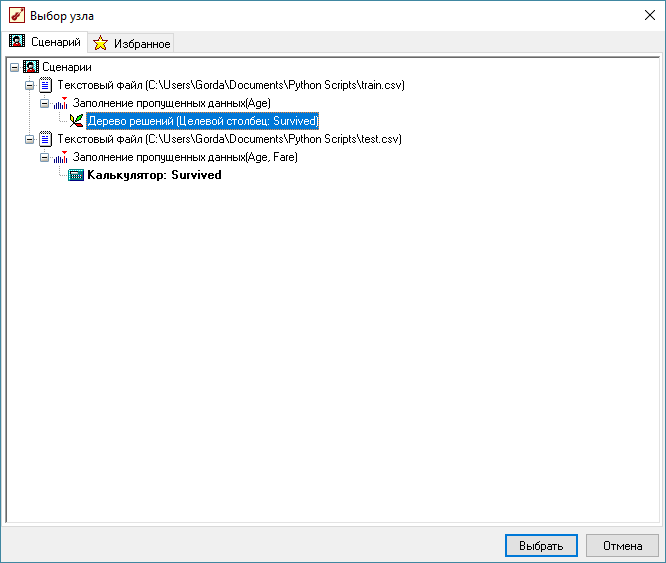


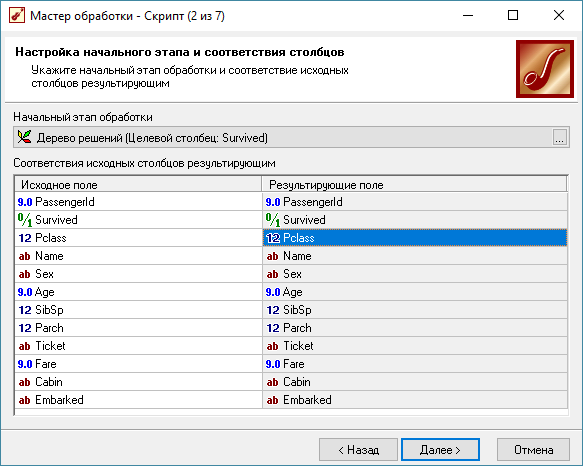
1. Далее необходимо заполнить пропуски с помощью мастера обработки. Метод обработки пропущенных данных – замена медианой.
2. Обработчик «Дерево решений» с настроенными назначениями:





1. Тестовые данные загружаются аналогично обучающим, включая заполнение пропущенных данных.
2. С помощью мастера обработки «Калькулятор» добавить новое поле Survived логического типа.
3. Для применения разработанной модели к новым данным следует использовать обработчик «Скрипт»





**Вопросы:**

1. Для проверки вывести таблицу сопряженности, объяснить ее.
2. Сколько значимых атрибутов среди общего числа?
3. Какой атрибут является наиболее значимым?

**Дополнительное задание:** преобразовать предсказанное поле в формат csv, принимаемый Kaggle и проверить качество классификации.