**Когнитивная бизнес-аналитика**

**План самостоятельных проектов**

1. Описание бизнес-области: для анализа выбран рынок производства яблок.
2. Краткое описание данных. Ссылка на источник.

Датасет содержит информацию о различных атрибутах набора яблок, отражающих их характеристики. Набор данных включает в себя такие измерения, как: ID, размер, вес, сладкость, хрустящесть, сочность, спелость, кислотность, качество.

Ссылка: https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality?resource=download

1. Используемые методы и технологии. Класс/классы задач Data Mining. Выбранные метрики, обоснование.

Анализ данных выполняется на low-code платформе анализа данных Loginom

Заполнение пропусков

Для анализа использованы методы: логистическая регрессия, нейросеть (классификация).

Факторы:

размер, вес, сладкость, хрустящесть, сочность, спелость, кислотность.

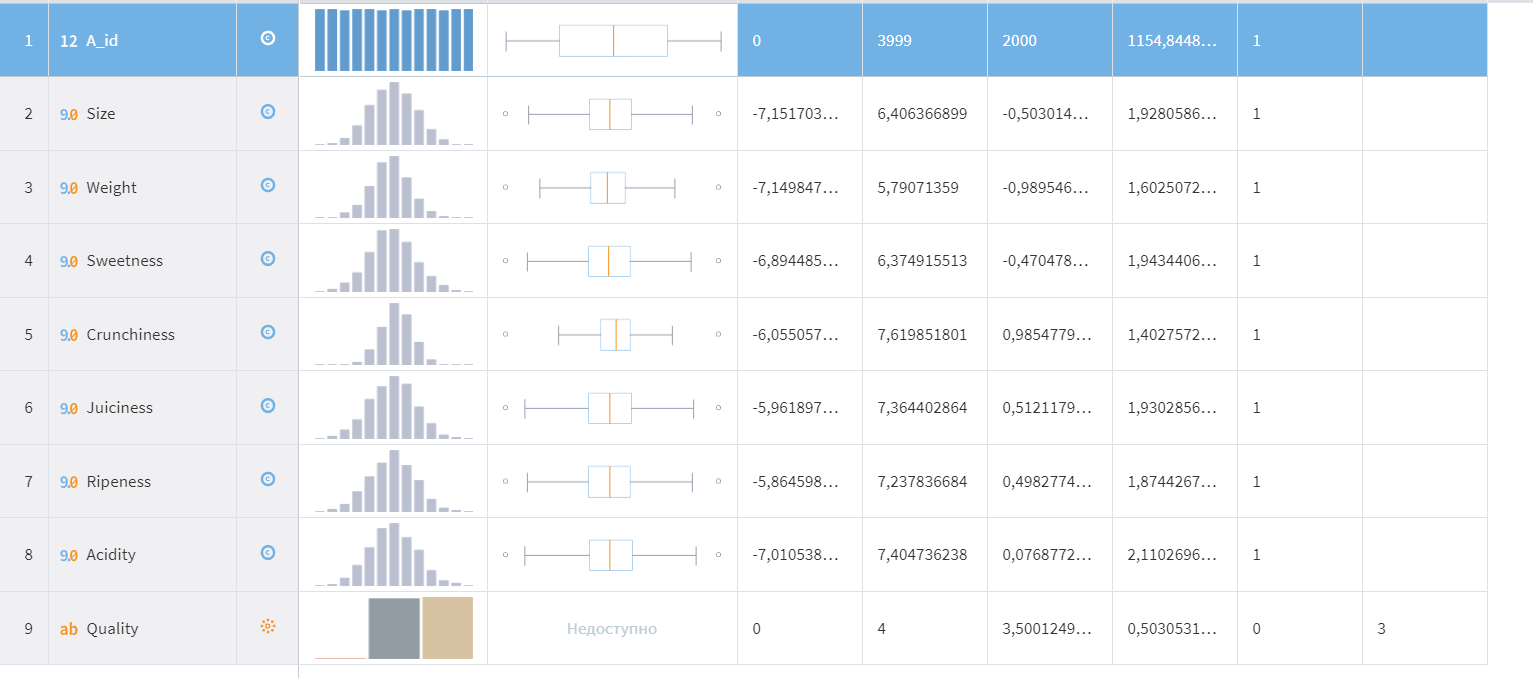
Для определения не влияющих факторов проведён факторный анализ.

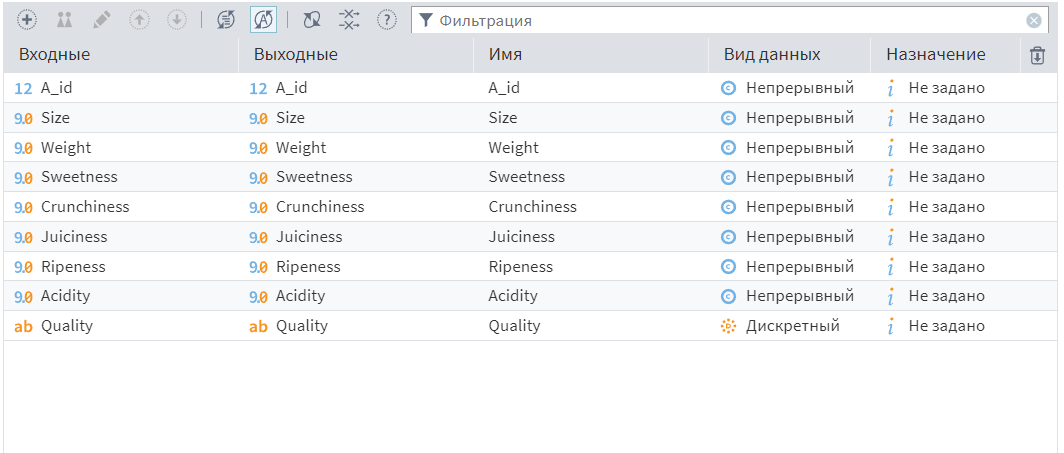
Проведена линейная регрессия и определены факторы, влияющие на спелость.

1. Реализация:

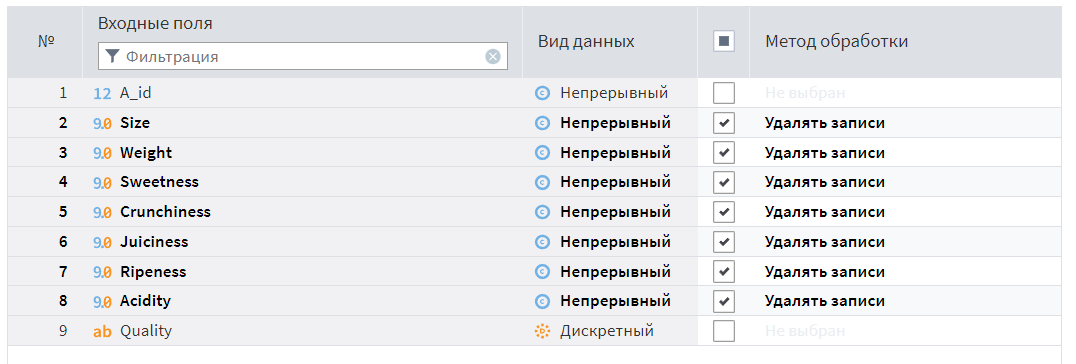
разработка сценариев анализа, разработка системы прогнозирования и т.п.:

* Исследование данных – данные включают в себя 9 столбцов по 1 пропуску во всех столбцах кроме качества.





* Предобработка: заполнение пропусков.



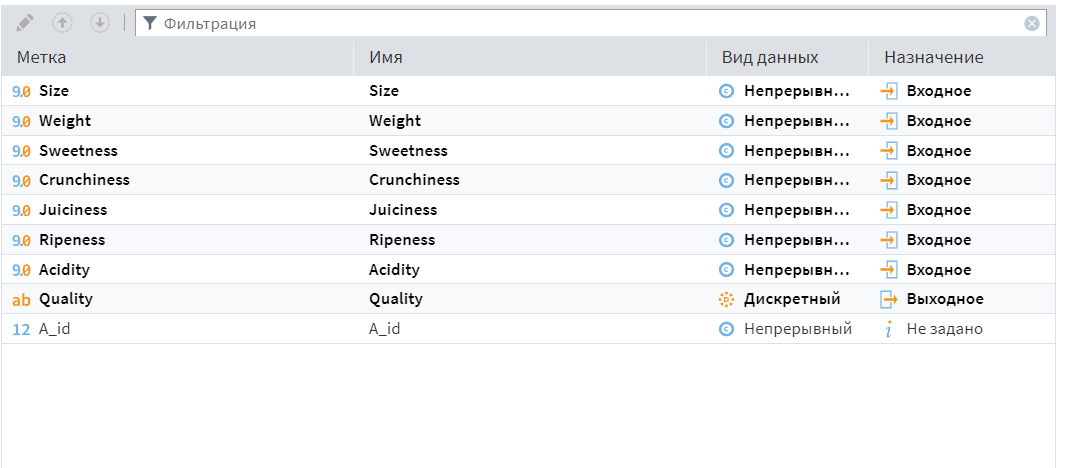
* Применение моделей: Логистическая регрессия - определение влияния факторов на определение состояния яблок хорошим либо плохим.

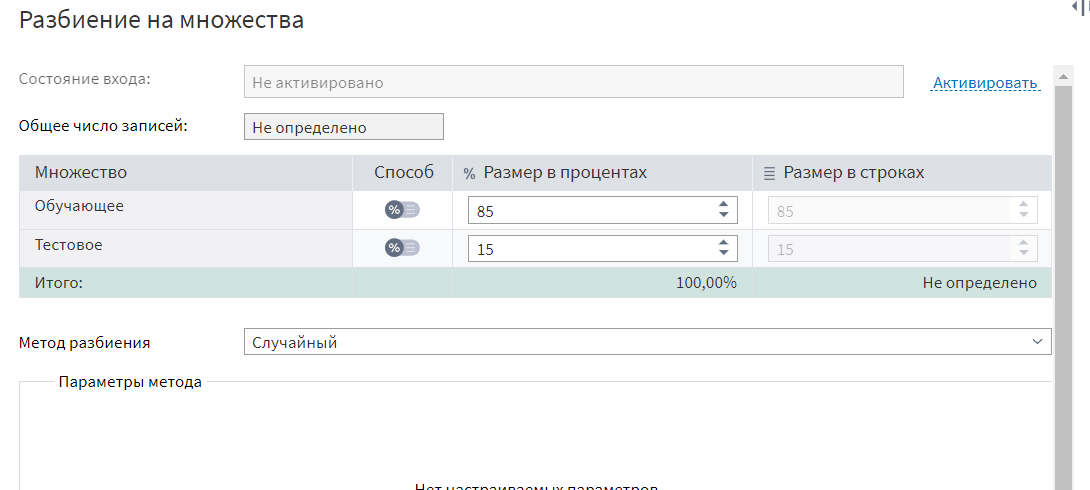
Уравнение логистической регрессии для определения влияния факторов на принятие решения о качестве яблок:

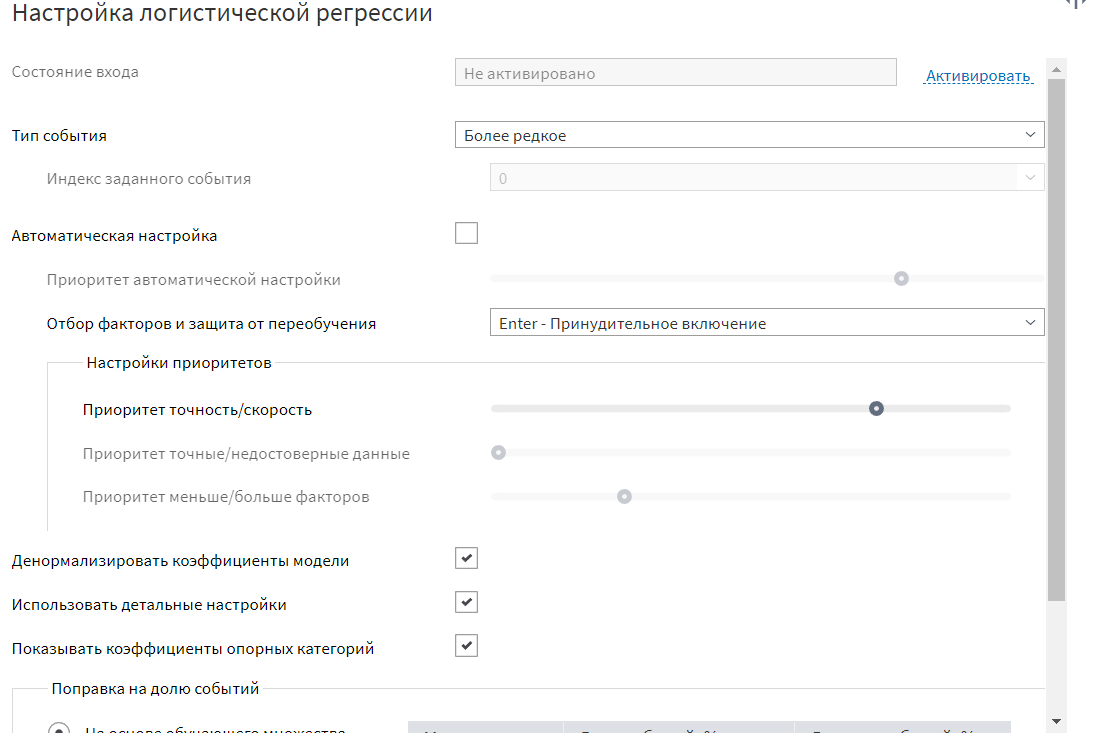
X = -0,6426-0,6365\*x1-0,2751\*x2-0,5819\*x3-0,03666\*x4-0,4385\*x5+0,1096\*x6+0,2905\*x7

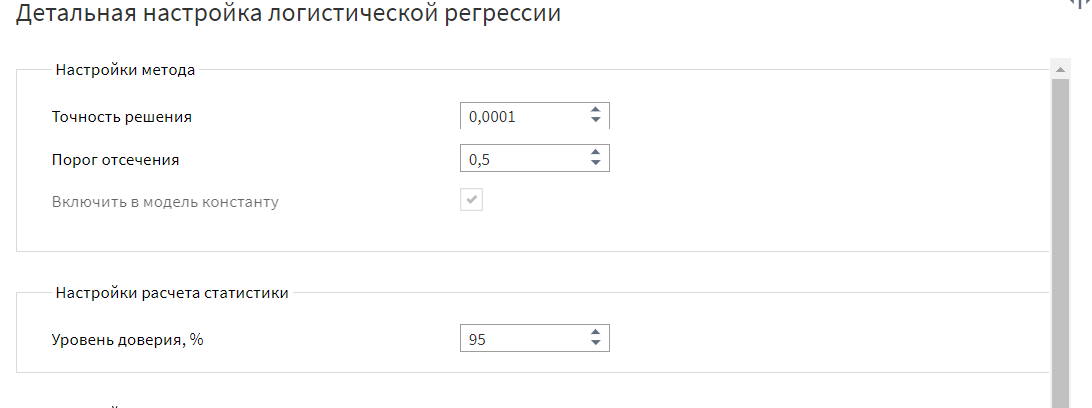
Факторы:

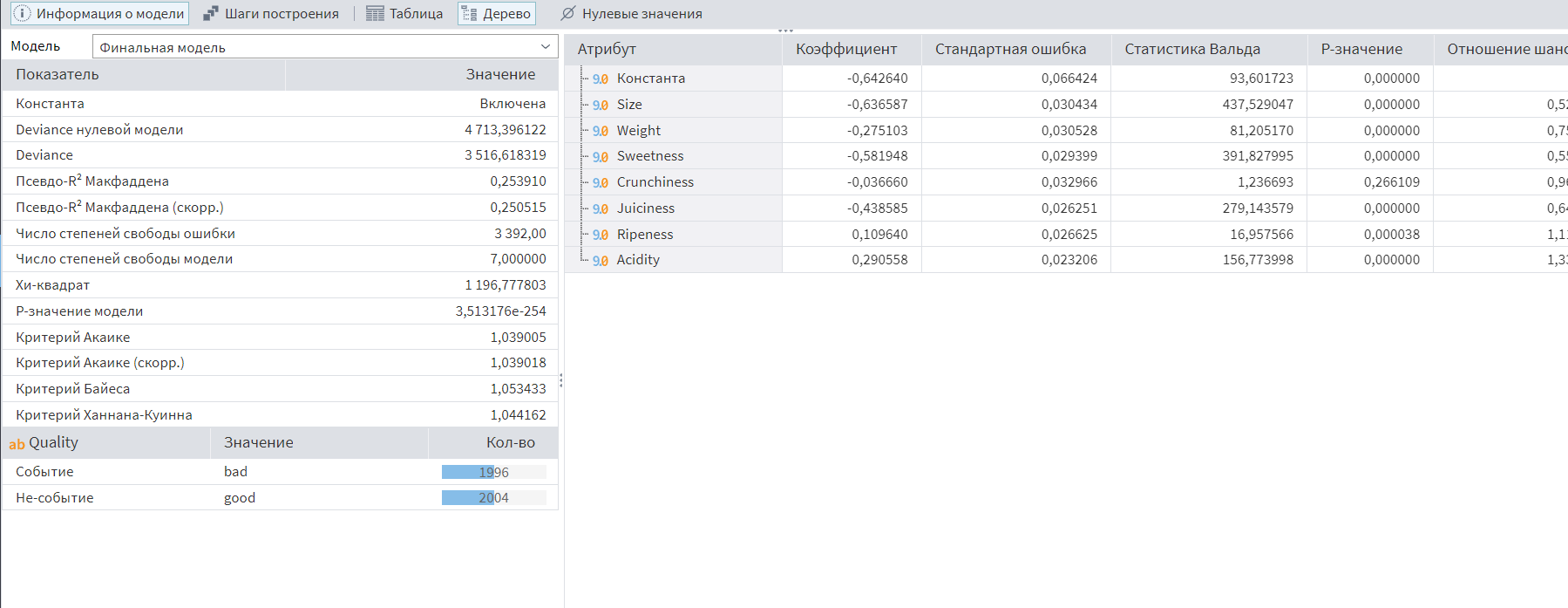
х1 – размер, х2 – вес, х3 – сладкость, х4 – хрустящесть, х5 – сочность, х6 – зрелость, х7 – кислотность.

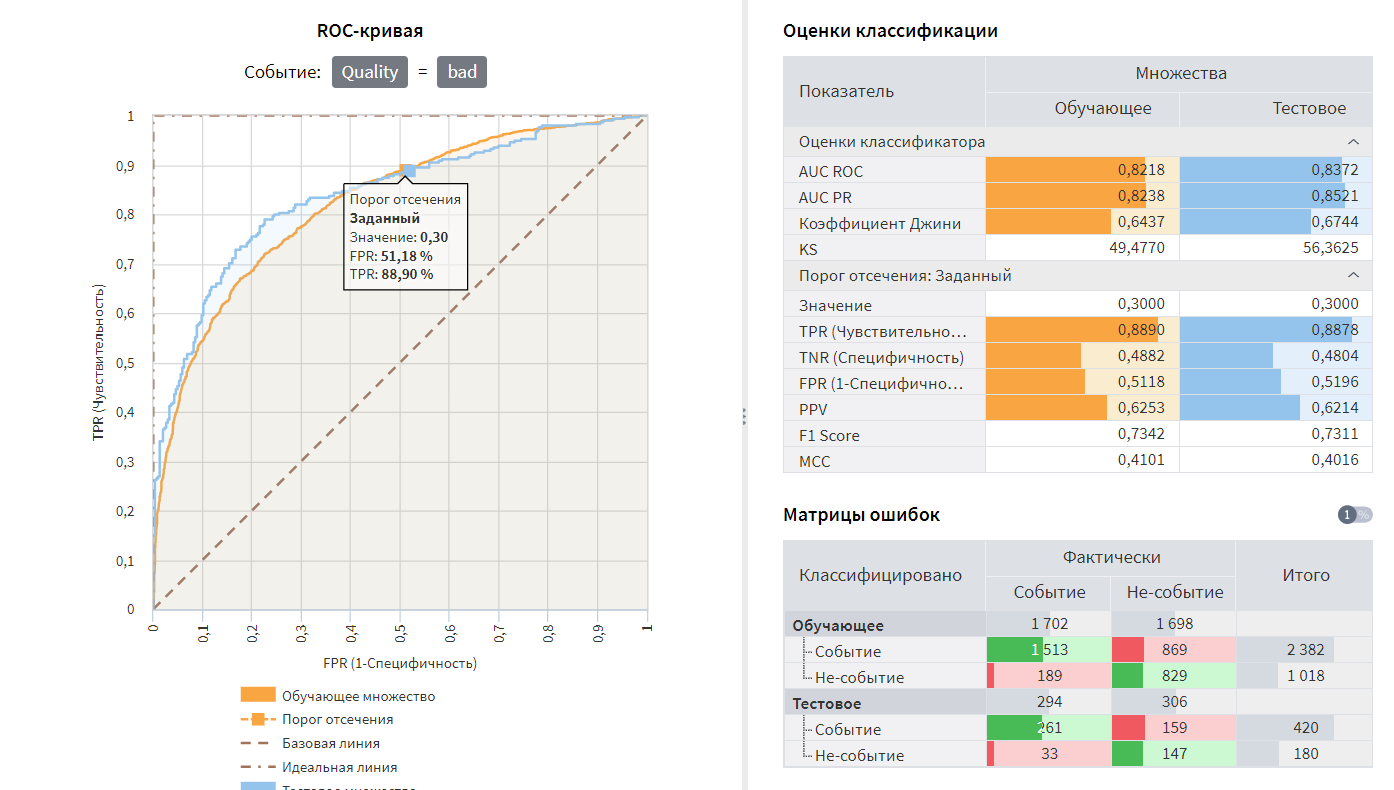








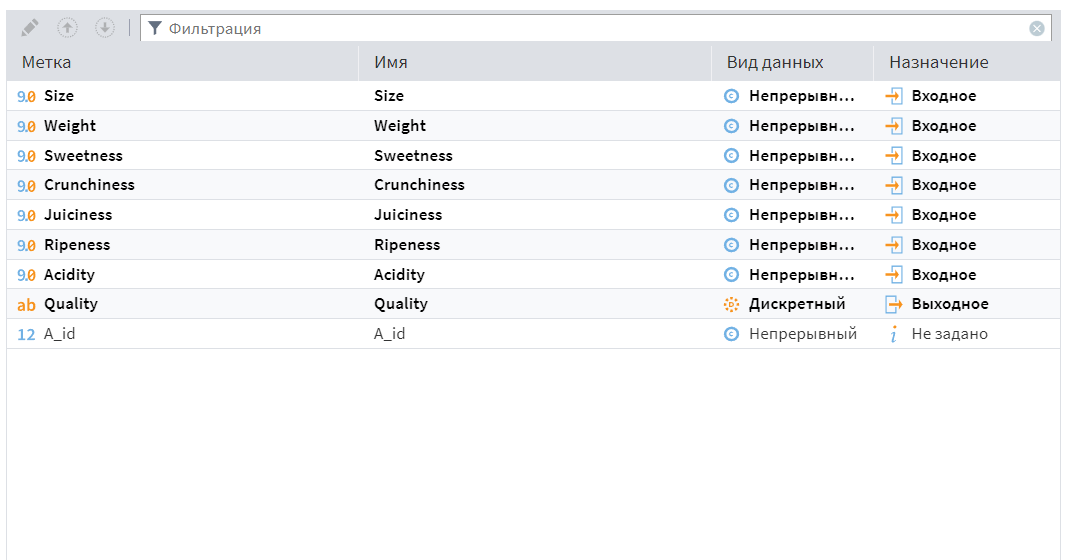


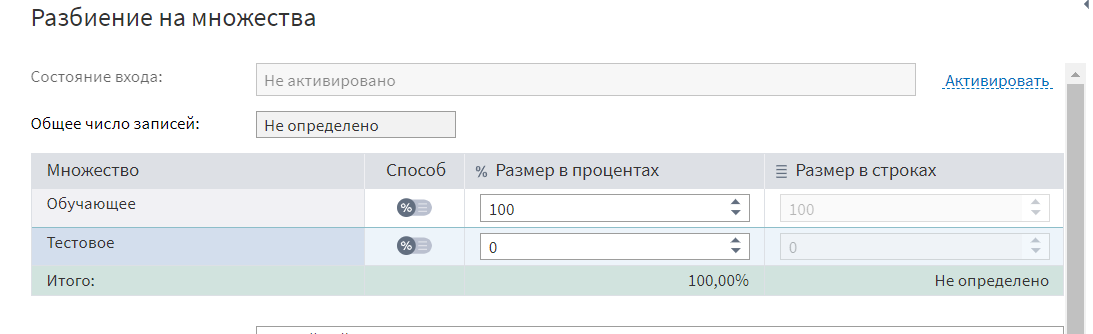


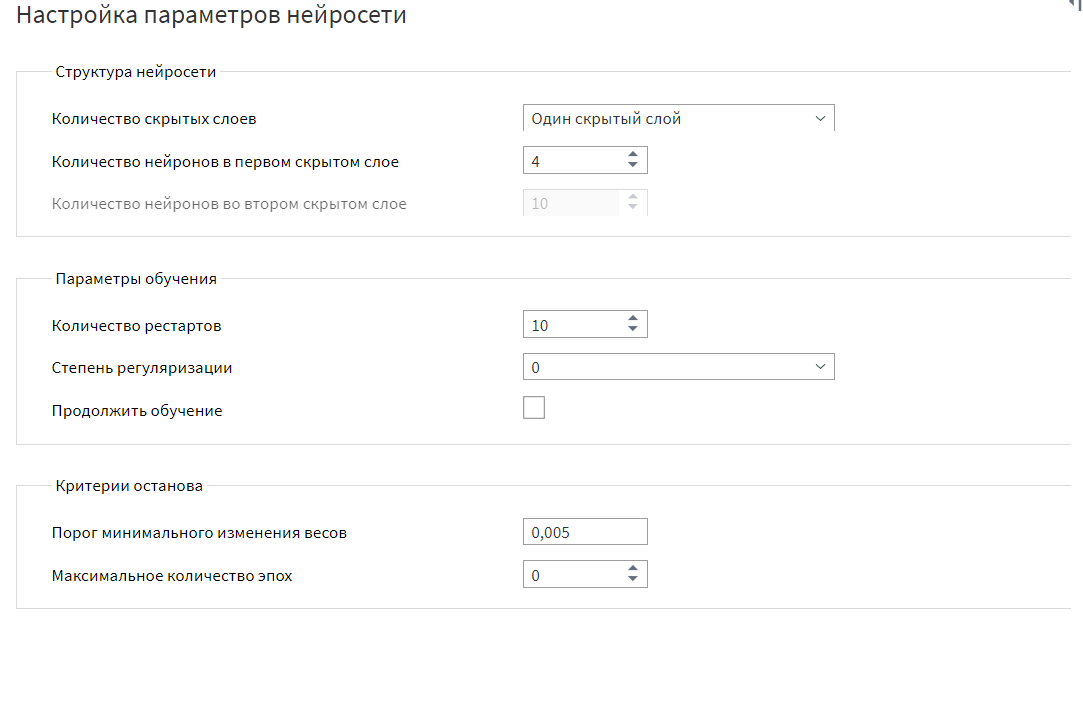
Recall: 1513 / (1513 + 189) = 0,888

Precision: 1513 / (1513 + 869) = 0,635

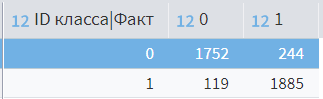
Нейросеть (классификация), позволяет классифицировать хорошие и плохие яблоки.







Классифицировав данные в нейросети выведем выходные данные в кросс-таблицу.



Recall: 1885 / (1885 + 119) = 0,940

Precision: 1885 / (1885 + 244) = 0,885

* Сравнение моделей по выбранным метрикам

Логистическая регрессия:

Recall = 0,888

Precision = 0,635

Нейросеть:

Recall = 0,940

Precision = 0,885

Нейросеть показала лучший результат

1. Выводы

Для набора данных по качеству яблок, влияющие на спелость яблок с разным уровнем влияния были применены методы классификации данных.

Выявлено, что Нейросеть лучше справляется с классификацией качества яблок по факторам.