**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ**

**ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра Вычислительной техники

ОТЧЁТ

по расcчетно-графической работе

по дисциплине: «Интеллектуальный анализ данных и методы машинного обучения»

|  |  |
| --- | --- |
| Выполнили:  Студент гр. АММ-22, АВТФ:  Салиму. М. | Преподаватель:  Яковина И.Н. |

Новосибирск, 2024

ОГЛАВЛЕНИЕ

[**1. Постановка задачи, актуальность темы и результаты анализа изученных источников по теме работы (1-3 стр.) 2**](#_z4eal4dglb30)

[**2. Описание данных и предобработка данных (2-5 стр.) 2**](#_lf6tgp19fb75)

[2.1. Разработанный рабочий процесс Orange3 для предварительной обработки данных 2](#_wm2c7vsw9r2v)

[2.2. Информация о наборе данных 3](#_u8lbnx8lcs2b)

[2.3. Проверка наличия пропущенных значений 3](#_9fyrb5f4om8h)

[2.4. Информация о признаках (статистики и кол-во пропусков) 3](#_l76cq726vqhg)

[2.5. Обработка пропущенных значений 3](#_jxr6j16oopt)

[2.6. Преобразование категориальных признаков в числовые 4](#_9d69vnmof381)

[2.7. Выбор признаков для работы 5](#_78v7crz885s8)

[**3. Наиболее значимые/интересные результаты анализа данных с пояснениями (гистограммы, графики, таблицы и т.п) (2-5 стр.) 5**](#_migs67h5ir8b)

[3.1. Проверка наличия выбросов 5](#_1olfvo283zem)

[3.2. Фильтрация (выбор строк виджетом Select Rows) 5](#_mn4gxai120ot)

[3.3. Графики выбранных признаков после фильтрации 5](#_lo5l0v43obm7)

[**4. Обзор разработанных моделей машинного обучения и примеры их использования (2-5 стр.) 5**](#_3qa6wu2t85qn)

[4.1. Описание выбранных методов МО 5](#_u5t9yipii7k6)

[4.2. Результаты решения задачи кластеризации 5](#_ozw9z6psqpqc)

[4.3. Результаты решения задачи классификации 6](#_8k4w8kn3vczf)

[4.4. Результаты решения задачи регрессии 6](#_ssxvxdk8uo72)

[4.5. Метрики оценки точности полученных моделей регрессии и классификации 6](#_v1djmym1z52h)

[**5. Проект интерфейса пользователя с описанием (2-5 стр.) 7**](#_4g52mqt2eea2)

[**6. Выводы по работе (1-2 стр.) 7**](#_bgcoghcztcul)

———————————————————————————————————

# Постановка задачи, актуальность темы и результаты анализа изученных источников по теме работы (1-3 стр.)

Общие слова об актуальности применения информационных технологий и концепции цифрового двойника для решения выбранной задачи.

Описание данных и предобработка данных

## Разработанный рабочий процесс Orange3 для предварительной обработки данных

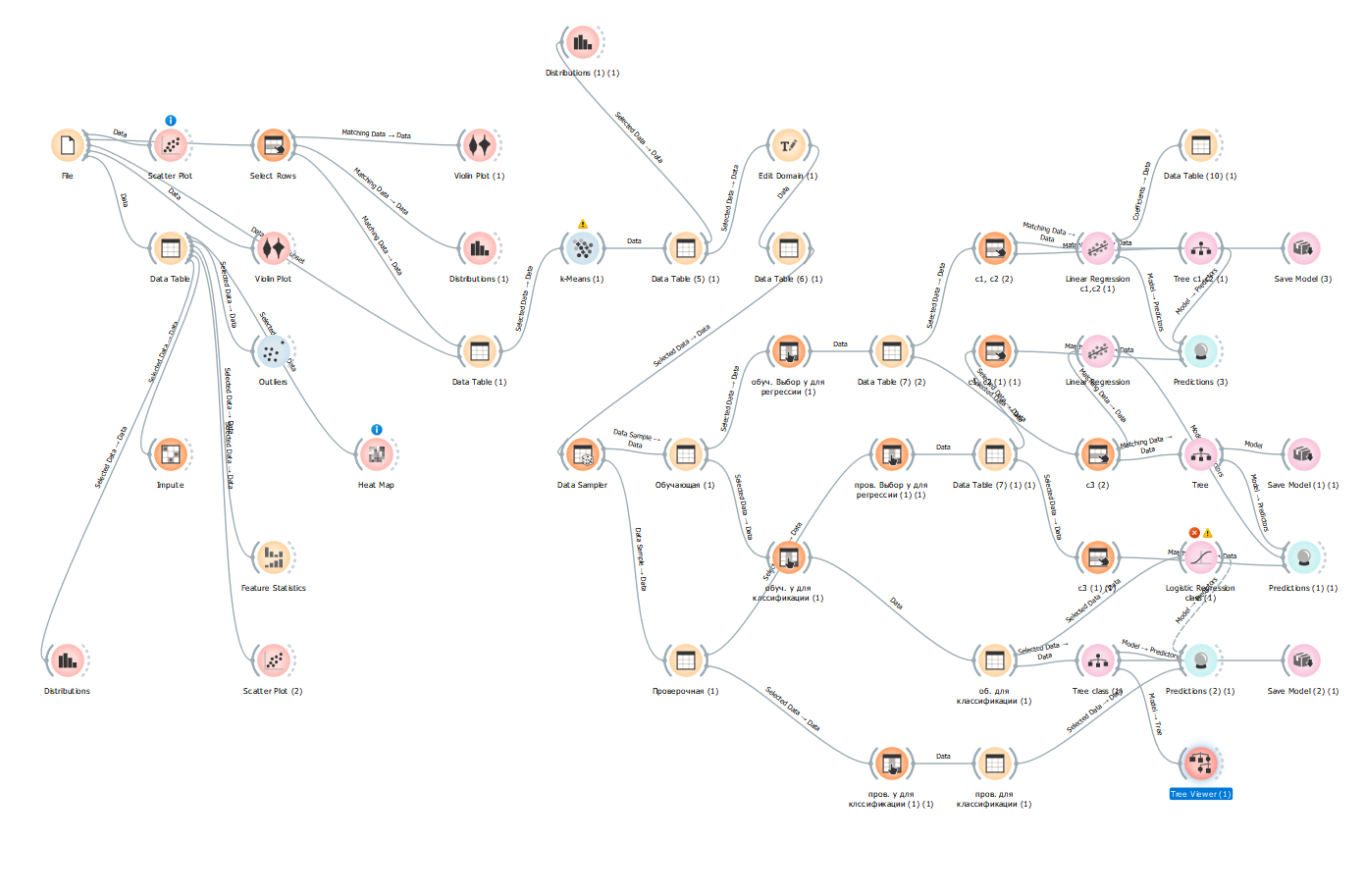


Рисунок 1 – Порядок соединения виджетов для предобработки данных

## Информация о наборе данных

* Кол-во строк (образцов): 20640
* Кол-во столбцов (признаков): 14
* Наименование признаков и тип данных

|  |  |  |  |
| --- | --- | --- | --- |
| № | Наименование | Тип данных  (числовой / категориальный) | Тип признака (целевой Y / независимый X) |
| 1 | Ср. Стоимость | numeric | Feature |
| 2 | Ср. Доход | numeric | Target |
| 3 | Ср. Возраст | categorical | Feature |
| 4 | Всего номеров | numeric | Feature |
| 5 | Всего спален | numeric | Feature |
| 6 | Население | numeric | Feature |
| 7 | Семьи | numeric | Feature |
| 8 | Широта | numeric | Meta |
| 9 | Долгота | numeric | Meta |
| 10 | Расст. до океана | numeric | Feature |
| 11 | Расст. до LA | numeric | Skip |
| 12 | Расст. до SA | numeric | Skip |
| 13 | Расст. до SJ | numeric | Skip |
| 14 | Расст. до SF | numeric | Skip |

## Проверка наличия пропущенных значений

* в текстовом виде

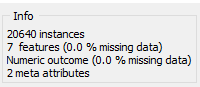


Рисунок 2 – Инф-ия о пропусках из виджета Table

* в графическом виде (Фрагмент)

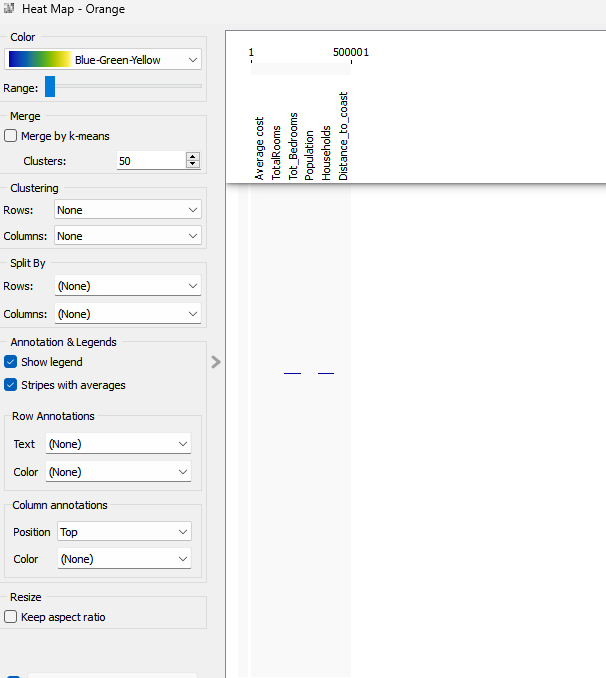


Рисунок 3 – Инф-ия о пропусках из виджета Heatmap

## Информация о признаках (статистики и кол-во пропусков)

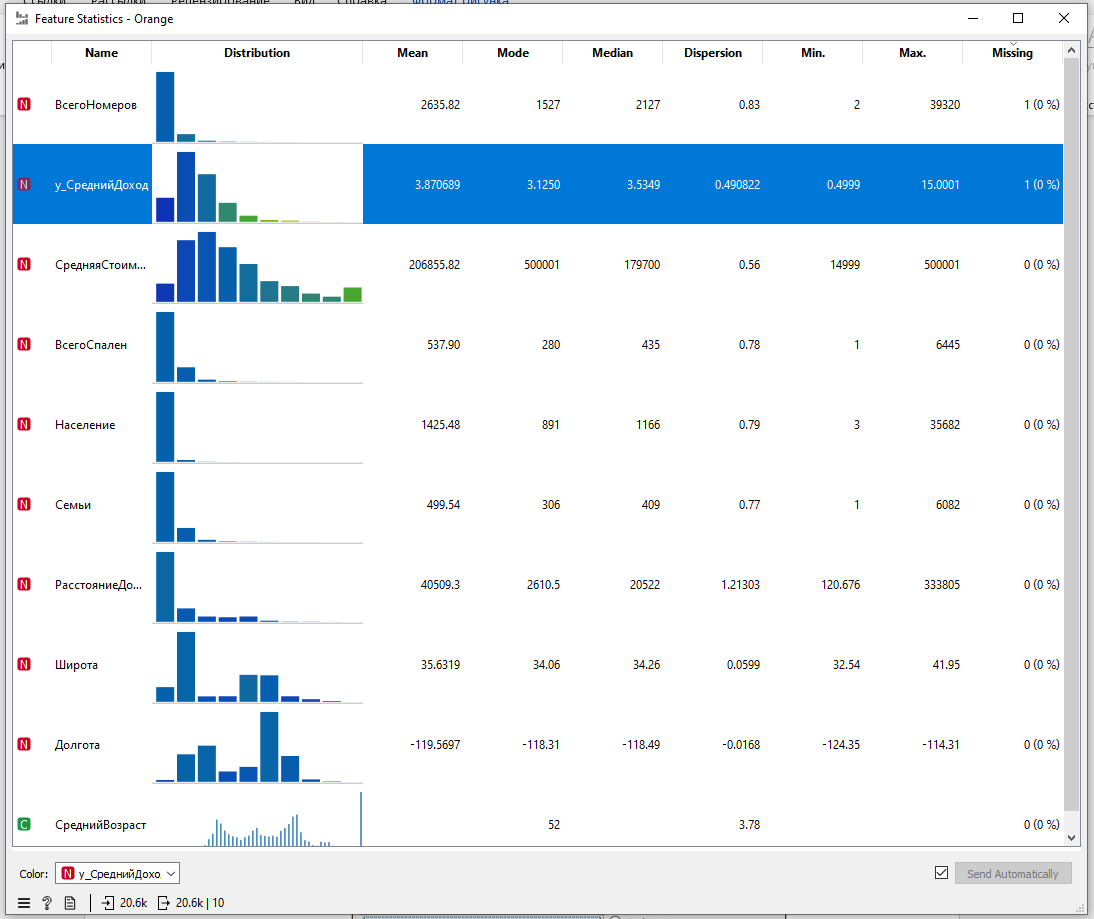


Рисунок 4 – Инф-ия о статистиках виджета Feature Statistics

## Обработка пропущенных значений

* Описание выбранного режима обработки пропусков для признаков:

Так, как во всём наборе данных были пропущены всего две строки, то было принято решение исключить эти две строки из набора.

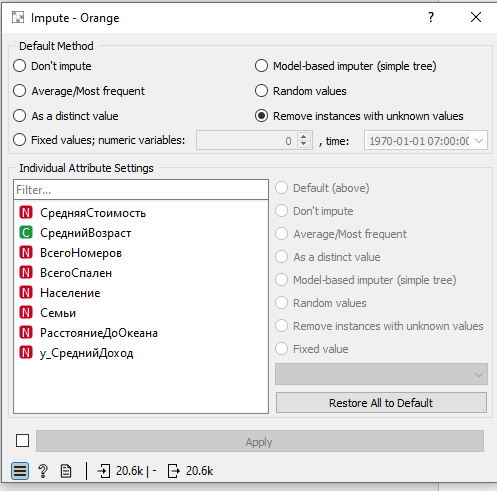


Рисунок 5 – Настройки обработки пропусков виджета Impute

* Кол-во строк (образцов) после обработки пропусков: 20638 шт.
* Осмотр инф-ии о наличии пропусков после обработки

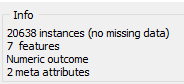


Рисунок 6 – Инф-ия о пропусках из виджета Table

Шаг с преобразованием категориальных параметров в числовые был пропущен, так как все столбцы уже имеют числовой формат.

## Выбор признаков для работы

Для дальнейшей работы выберем назначим состав признаков: целевого Y (target) и независимых X (features).

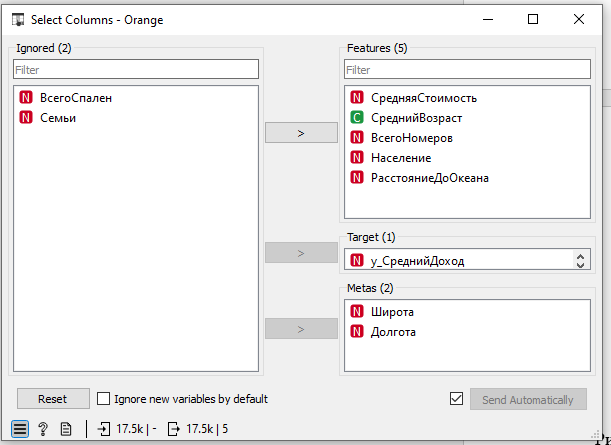


Рисунок 7 – Выбранные зависимые и независимые признаки в виджете Select Columns

# Наиболее значимые/интересные результаты анализа данных с пояснениями (гистограммы, графики, таблицы и т.п)

## Проверка наличия выбросов

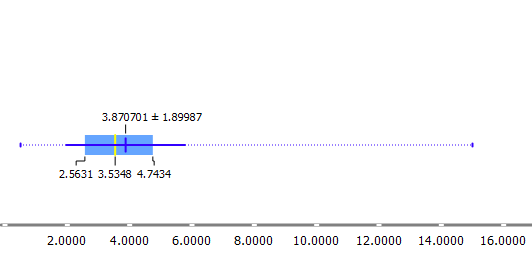


Рисунок 8 – Диаграмма размаха

На данном графике изображено распределение среднего дохода. Верхний и нижний квартили расположены на отметках 4.7434 и 2.5631 тыс. денежных единиц, соответственно. Максимальное значение по данному графику можно считать 6 тыс. денежных единиц, минимальным – 2 тыс. денежных единиц. Всё, что находится за пределами данного диапазона – можно считать выбросами, однако взглянув на гистограмму распределения среднего дохода – можно заметить, что нормальное распределение продолжается до 8 тыс. денежных единиц.

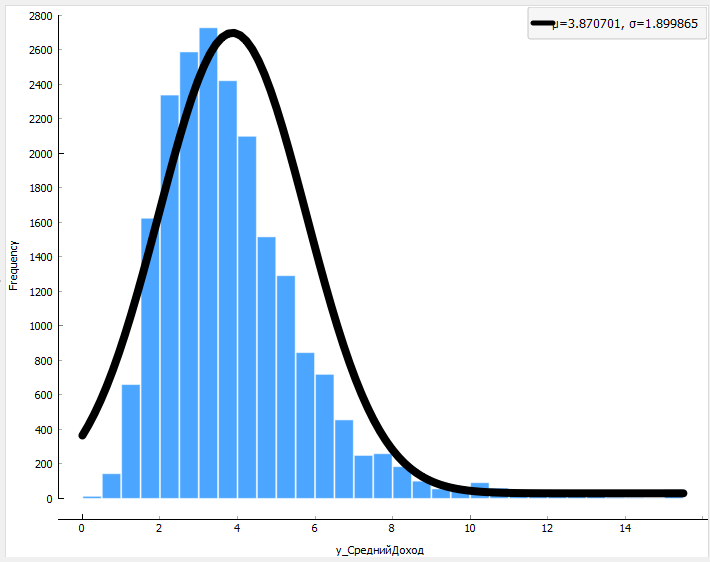


Рисунок 9 – Гистограмма распределения среднего дохода

## Фильтрация

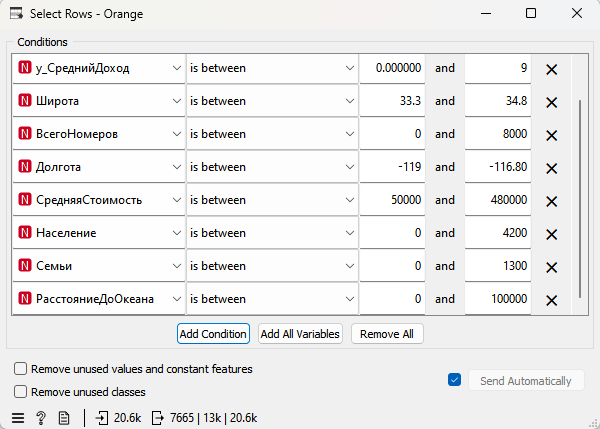


Рисунок 10 – Фильтрация виджетом Select Rows

Для дальнейшей обработки данных была произведена фильтрация при помощи виджета Select Rows, в котором были отброшены строки, значения в которых выходят за заданный диапазон.

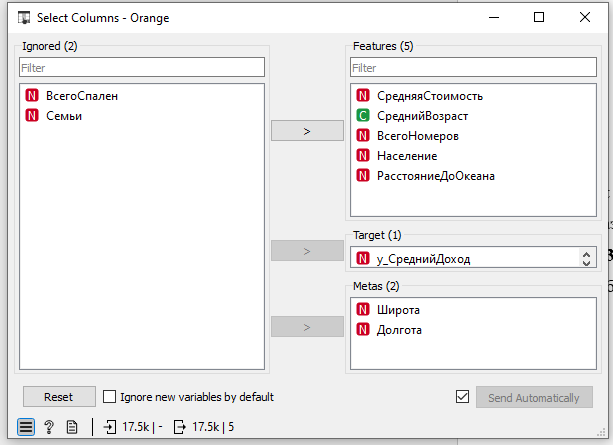


Рисунок 11 – Фильтрация виджетом Select Columns

На следующем шаге были отброшены лишние столбцы, которые имели наименьшую корреляцию с искомой величиной.

## Графики выбранных признаков после фильтрации

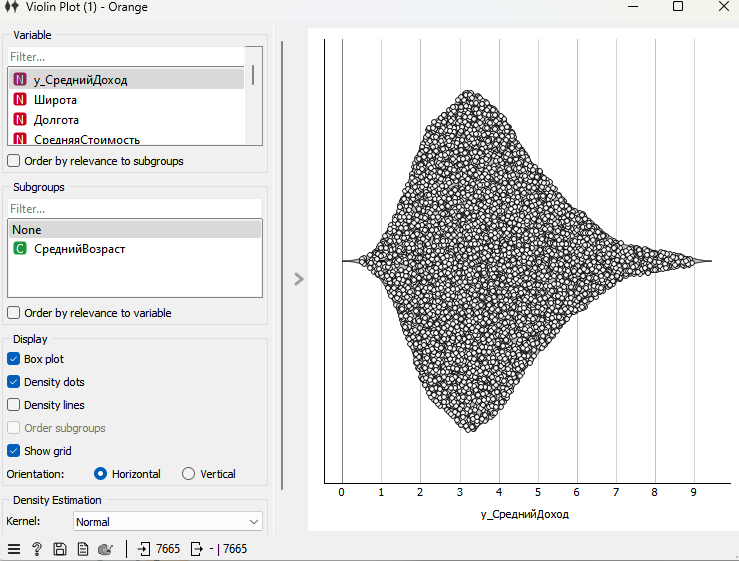


Рисунок 12 – Диаграмма рассеяния для y\_СреднийДоход

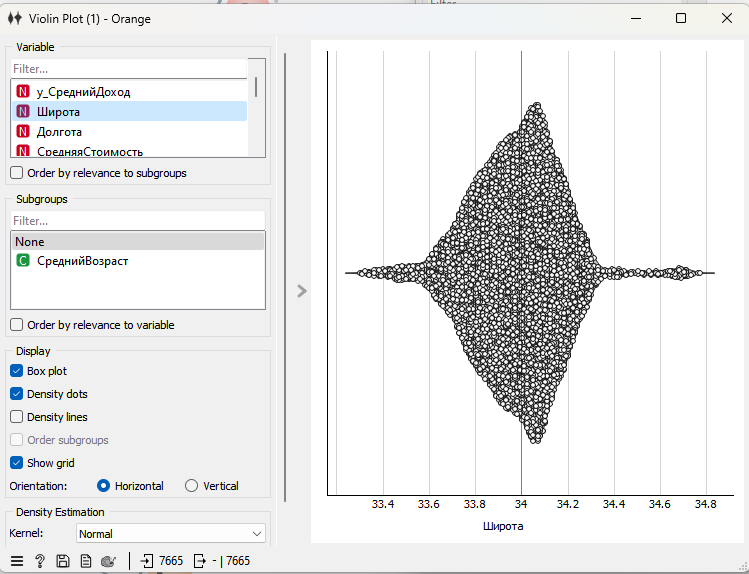


Рисунок 13 – Диаграмма рассеяния Широта

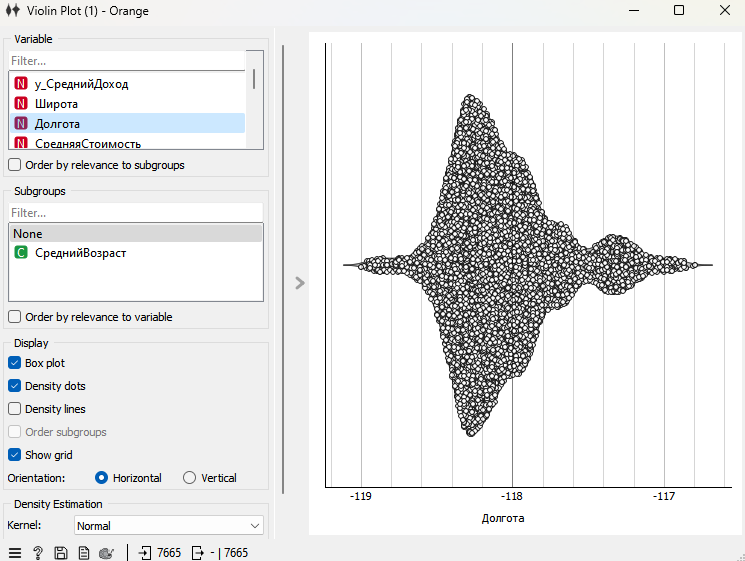


Рисунок 14 – Диаграмма рассеяния Долгота

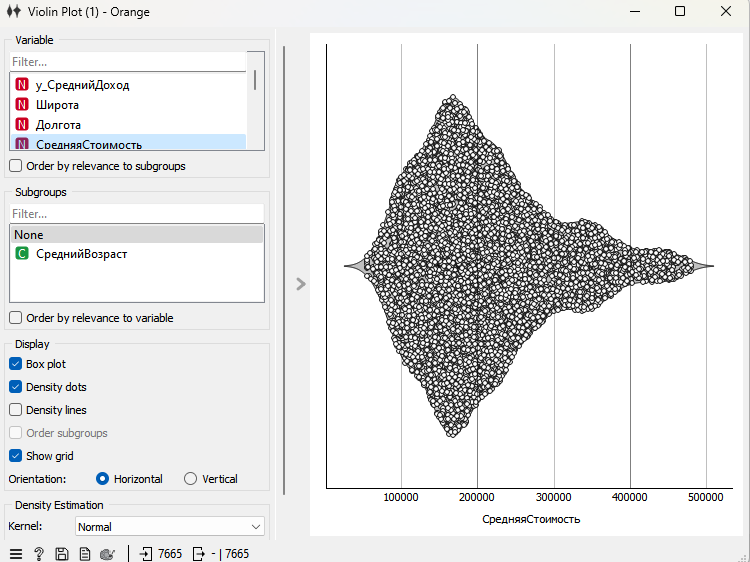


Рисунок 15 – Диаграмма рассеяния СредняяСтоимость

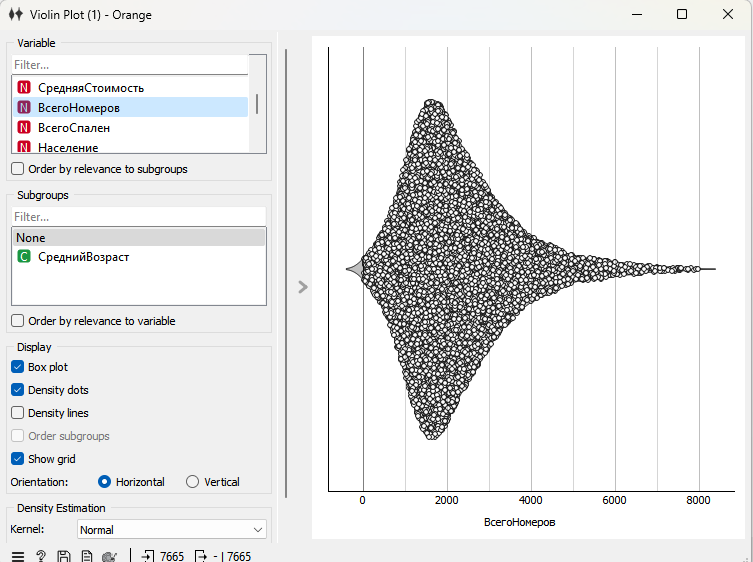


Рисунок 16 – Диаграмма рассеяния ВсегоНомеров

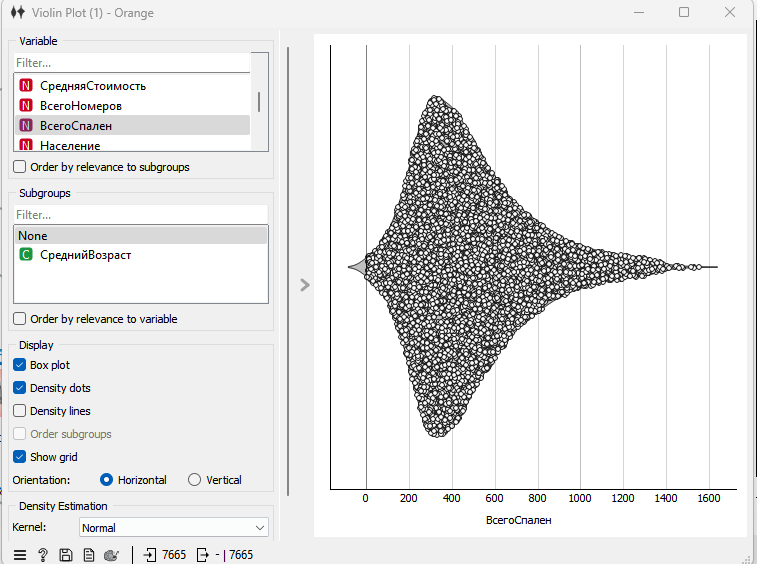


Рисунок 17 Диаграмма рассеяния ВсегоСпален

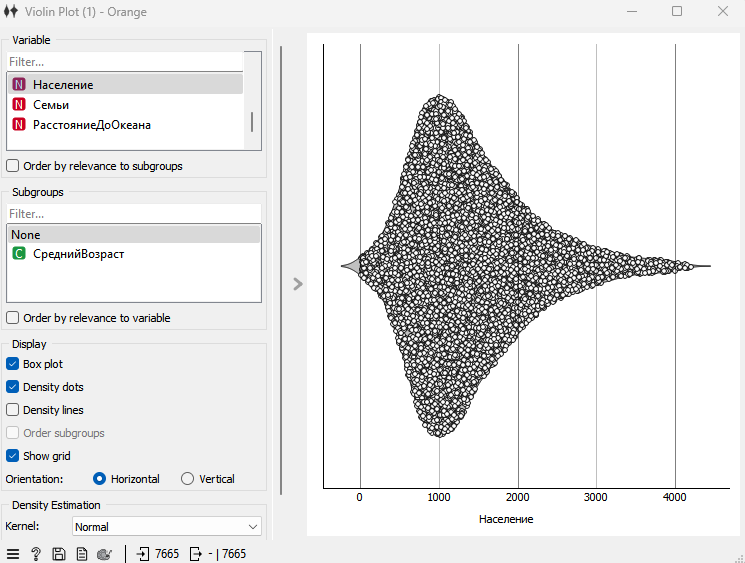


Рисунок 18 – Диаграмма рассеяния Население

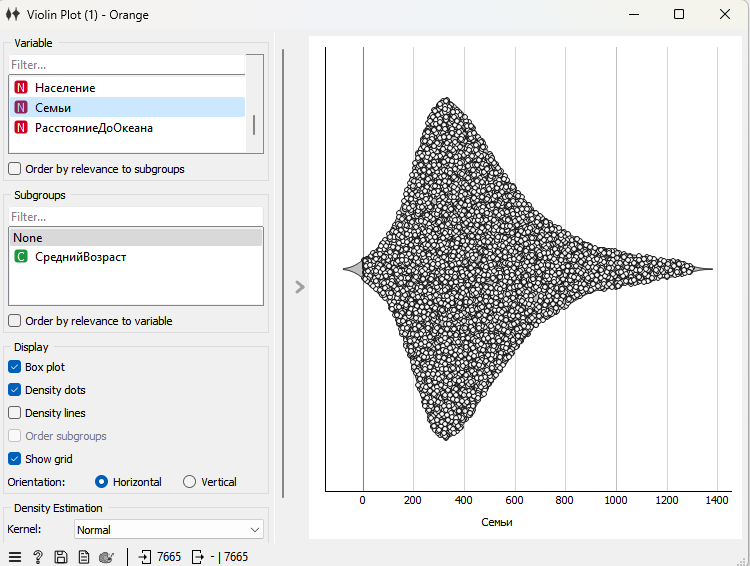


Рисунок 19 – Диаграмма рассеяния Семьи

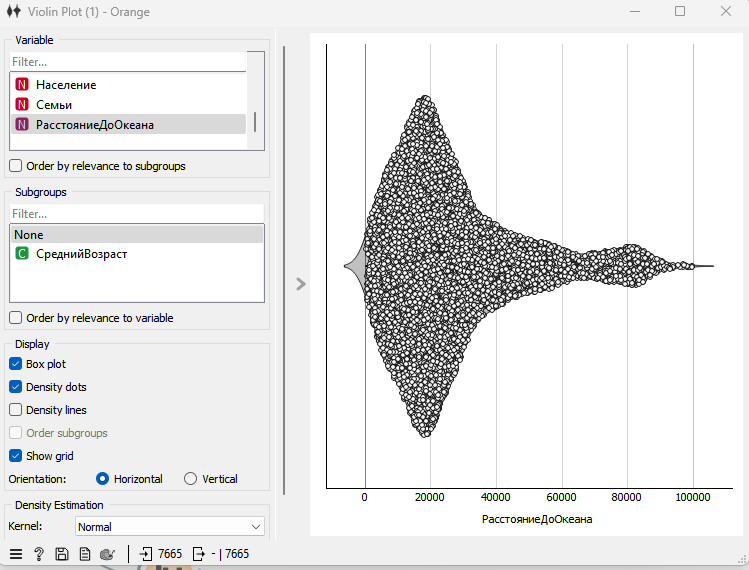


Рисунок 20 – Диаграмма рассеяния РасстояниеДоОкеана

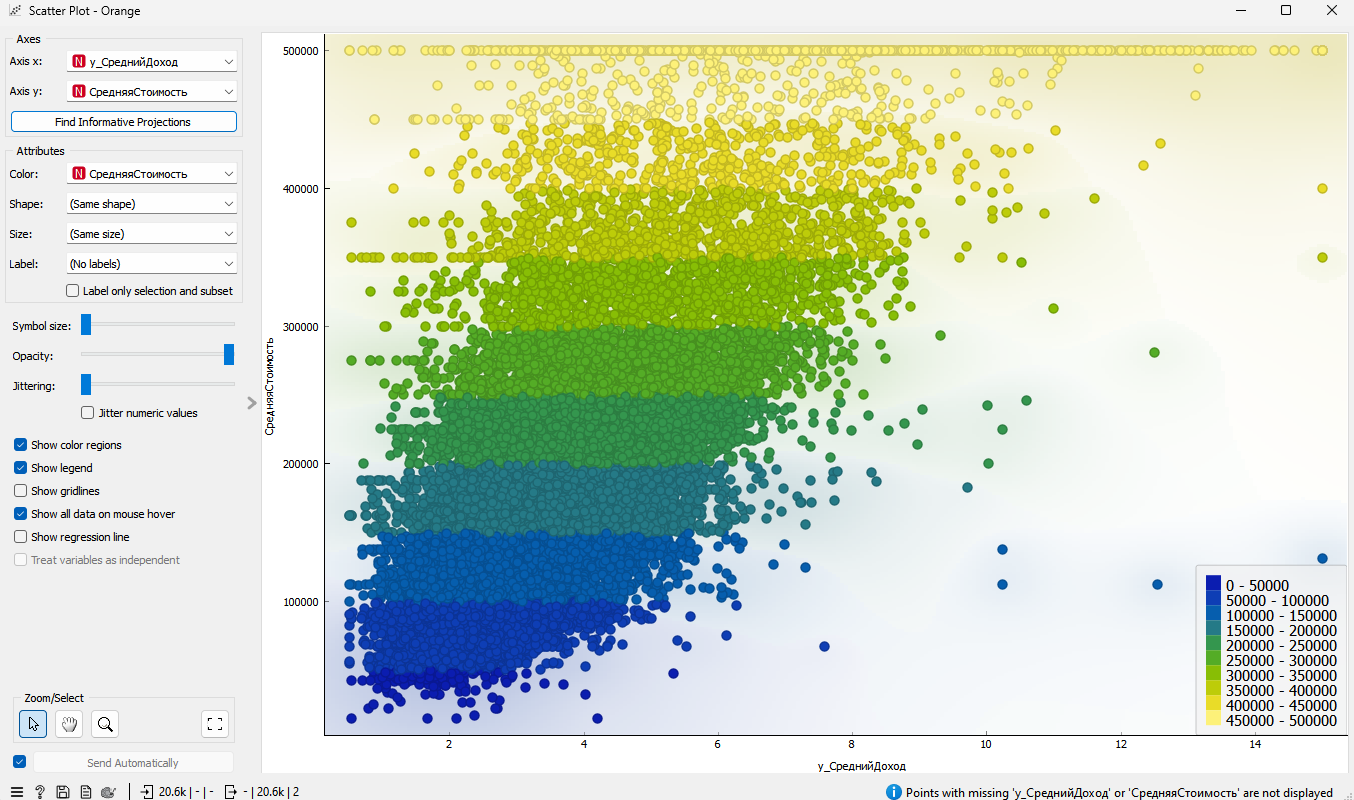


Рисунок 21 – График зависимости дохода от стоимости

# Обзор разработанных моделей машинного обучения и примеры их использования

## Описание выбранных методов МО

Перечень методов МО с краткими пояснениями и виджетов, позволяющих их использовать при решении задачи разработки моделей.

## Результаты решения задачи кластеризации

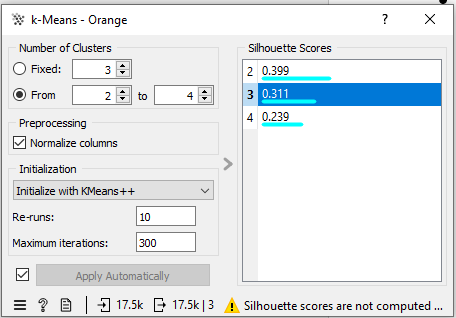


Рисунок 22 – Виджет k-means для выполнения кластеризации.

В рамках решения задачи кластеризации было выбрано 2 кластера.

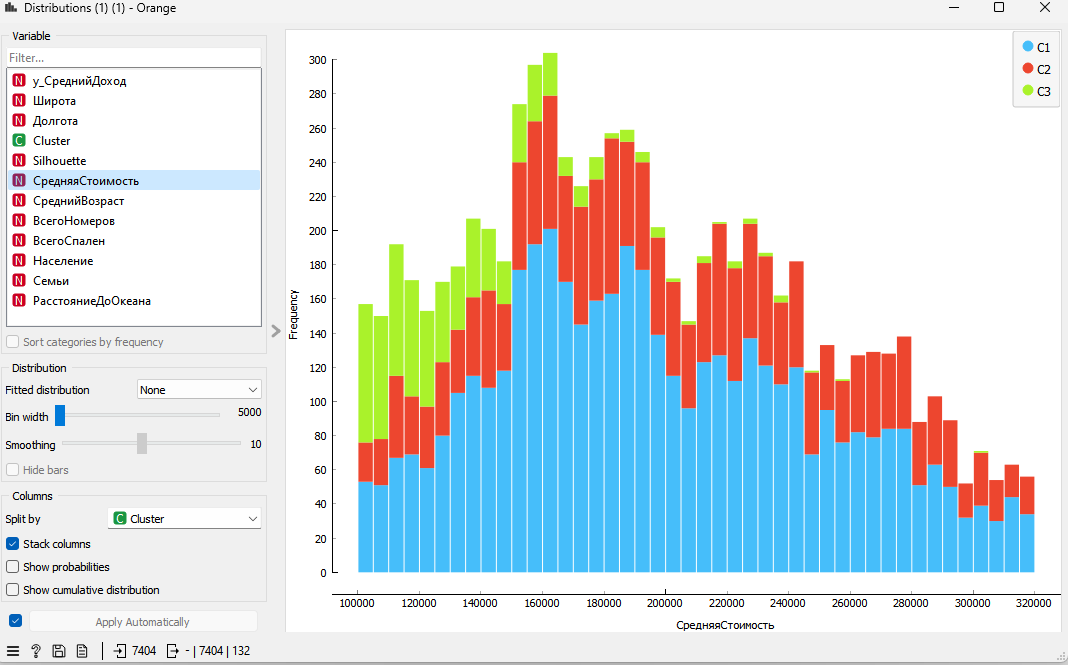


Рисунок 23 – График с цветовой маркировкой по кластерам

В результате получена таблица с новым столбцом “Cluster”, в которой указана принадлежность к определенному кластеру.

## Результаты решения задачи классификации

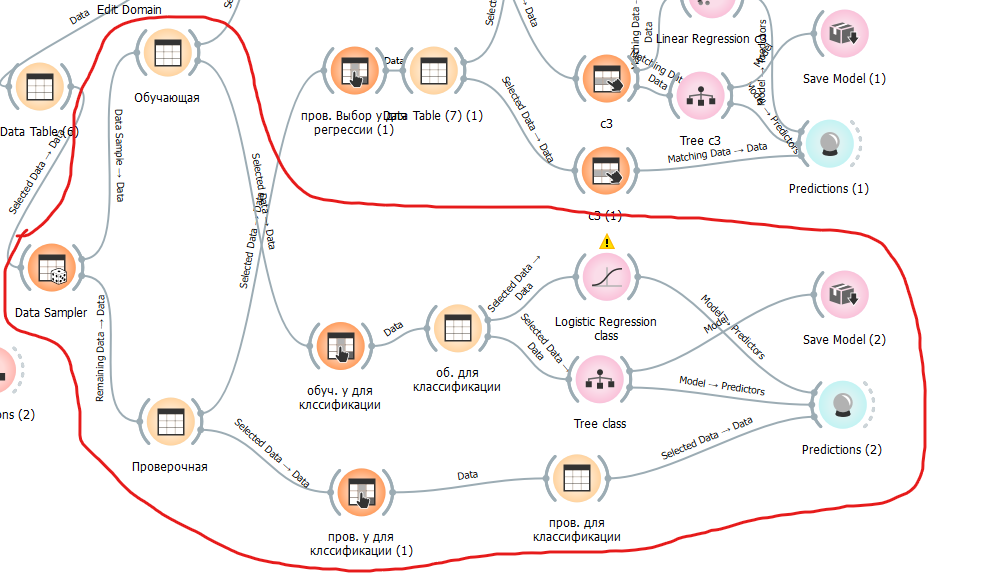


Рисунок 24 – Порядок соединения виджетов для обучения моделей

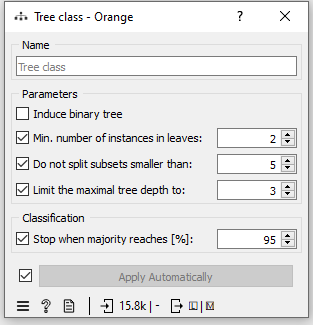


Рисунок 25 – Параметры алгоритма Tree class

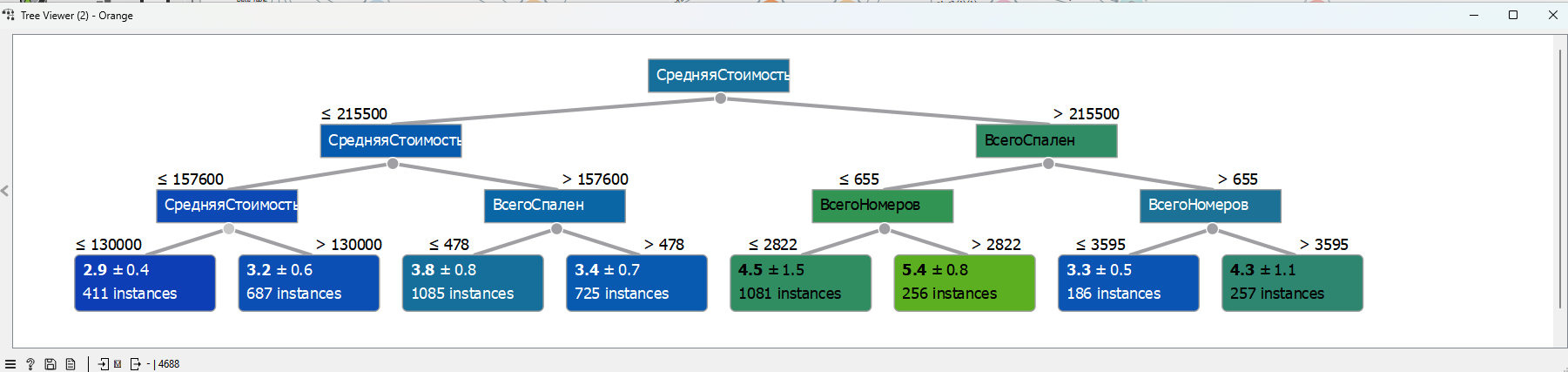


Рисунок 26 – Внутреннее состояние обученной модели

## Результаты решения задачи регрессии

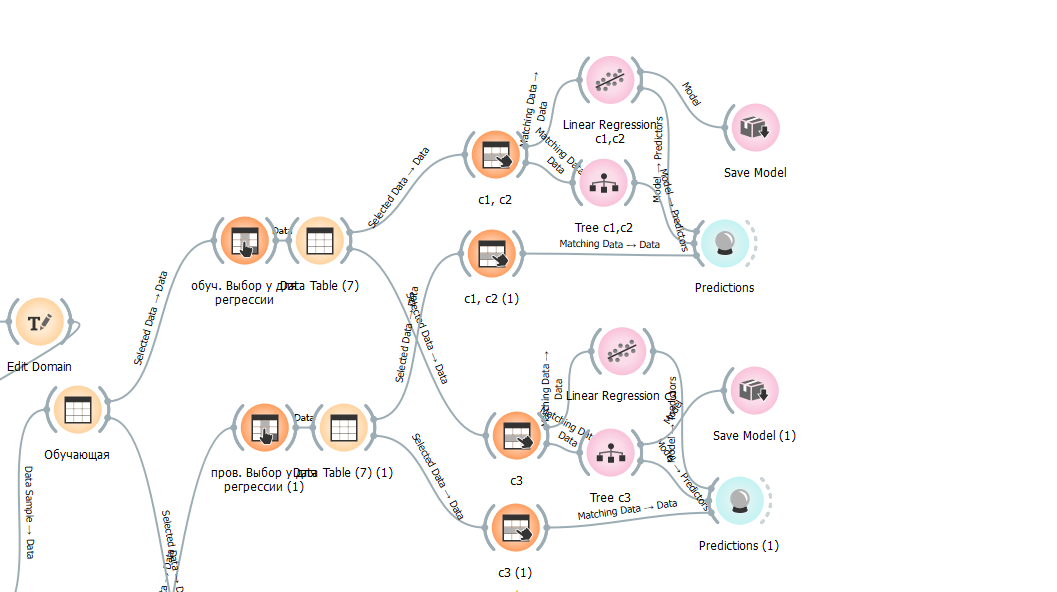


Рисунок 27 – Порядок соединения виджетов для обучения моделей

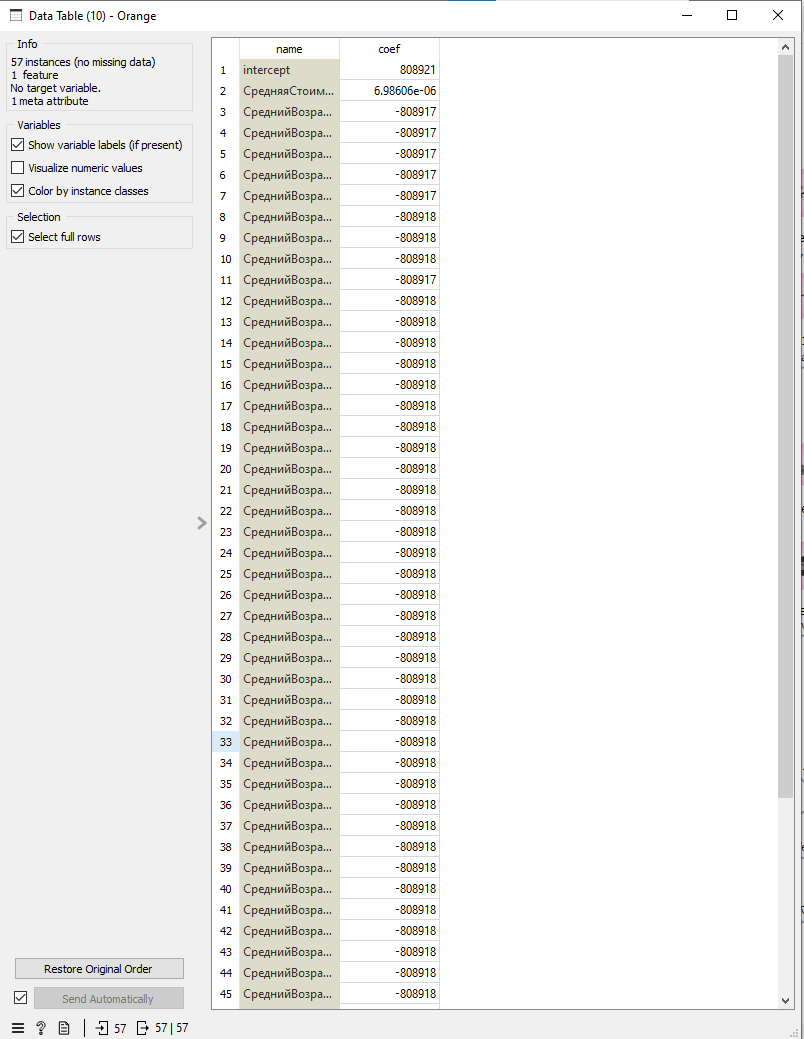
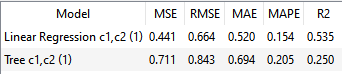


Рисунок 23 – Внутреннее состояние обученной модели

## Метрики оценки точности полученных моделей регрессии и классификации

* Лучше всего со своей задачей справляется модель линейной регрессии с двумя классами. По сравнению с методом дерева решений, она имеет более высокую точность. Для сравнения были взяты параметры MAE и R2.
* Для повышения точности модели можно обучить её на большем количестве данных. Для этого можно разделить данные в виджете Data Sampler в пропорциях, к примеру, 70% и 30%.

 - для модели регрессии

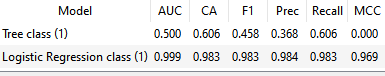
 - для модели классификации

Рисунок 24 – Значения метрик оценки точности обученных моделей в виджете Predictions

# Проект интерфейса пользователя с описанием (2-5 стр.)

* Описание выбранного способа опроса обученных моделей.
* Описание проектируемого интерфейса
* Описание разработки и реализации интерфейса

# Выводы по работе (1-2 стр.)

* Использованное ПО для анализа данных, построения моделей, разработки интерфейса пользователя.
* Описание выбранных методов(алгоритмов) машинного обучения для анализа данных и построения моделей.
* Краткое описание полученных результатов моделирования