Лабораторная работа

Разработка системы идентификации с помощью пакета Orange

Выполнила: Короткова Инга Сергеевна

2020 год

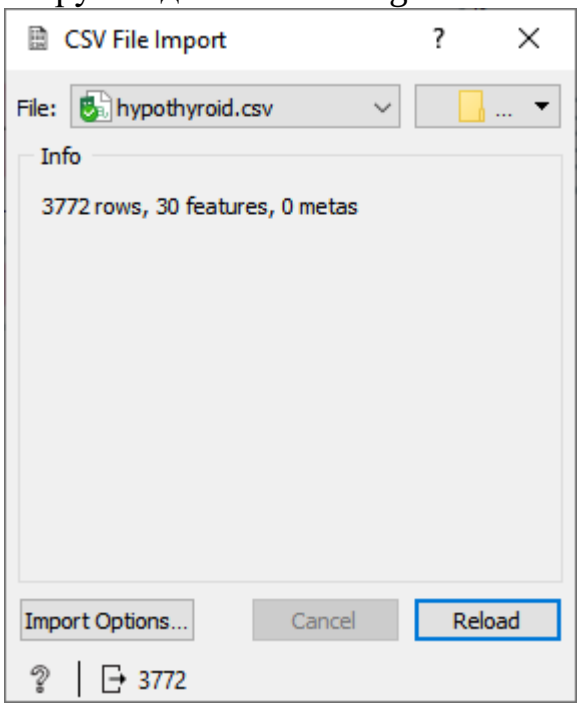
Цель работы – закрепить навыки проектирования интеллектуальных моделей с помощью платформы Orange Data Mining для задач построения идентифицирующих систем и оценки эффективности их работы.

Задачи:

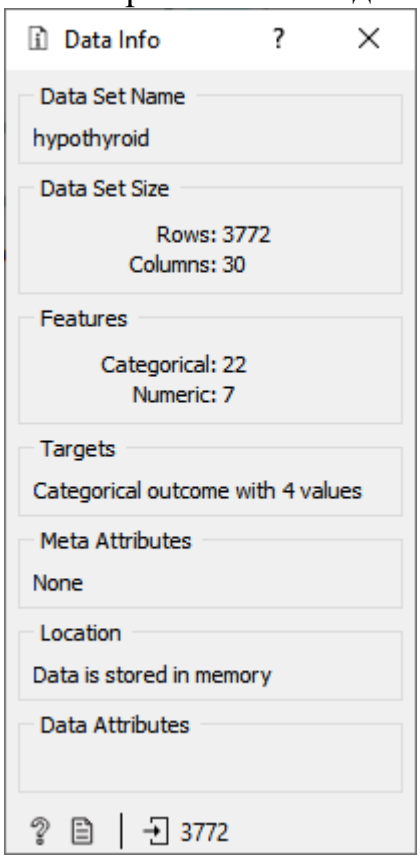
1. Выбрать набор данных, содержащий не менее 500 объектов и не менее 10 атрибутов, отмеченный как набор, для задачи классификации или регрессии с репозитария https://archive.ics.uci.edu/ml/datasets.php
2. Загрузить выбранный набор данных.
3. Произвести исследовательский анализ данных:
   - получить объём исследуемых данных;
   - получить число атрибутов и их типы данных;
   - посмотреть распределение числа примеров классов
   - если необходимо, выполнить преобразование категориальных атрибутов;
   - если необходимо, заполнить пропущенные значения в выборке.
4. Разделить набор данных на обучающую и тестовую выборку и объяснить это разбиение.
5. Обучить несколько моделей с помощью трех любых алгоритмов построения классификаторов или регрессоров (по выбору слушателя курса), отметить почему были выбраны эти алгоритмы
6. Оценить эффективность моделей на тестовой выборке с помощью матрицы неточностей, критериев полноты *Recall* и точности *Precision*, в случае создания классификатора, или критериев средняя квадратичная ошибка *MSE*, средняя абсолютная ошибка *MAE* и коэффициент детерминации *R2*.

В этой работе будем использовать датасет по заболеваниям щитовидной железы.
Датасет взят с сайта: http://archive.ics.uci.edu/ml/datasets/thyroid+disease
Данные собраны Garvan Institue в Австралии

Загрузим данные в Orange.

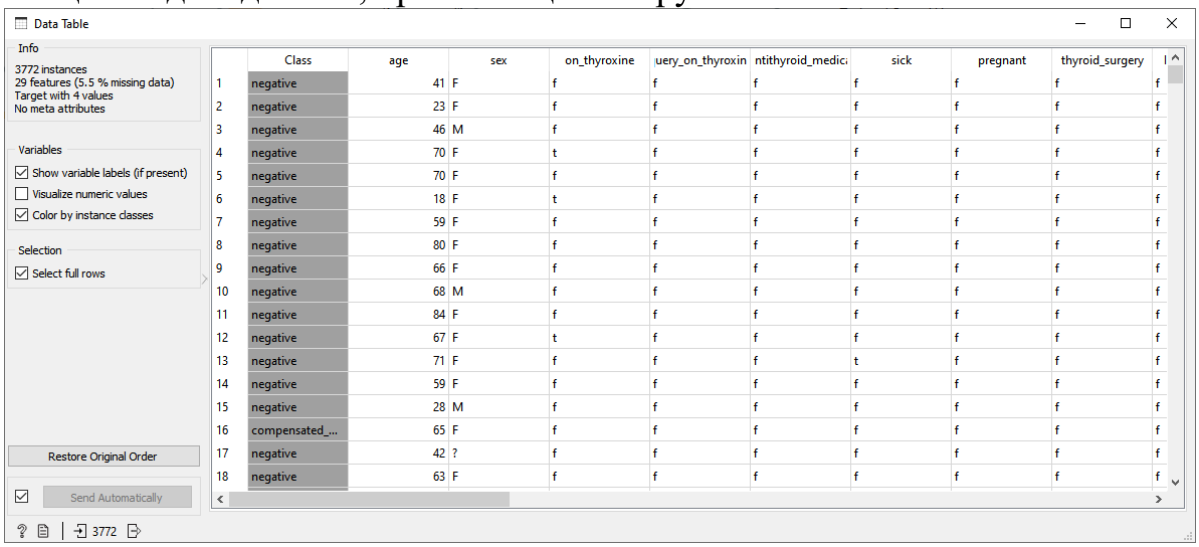


Посмотрим ближе на данные:

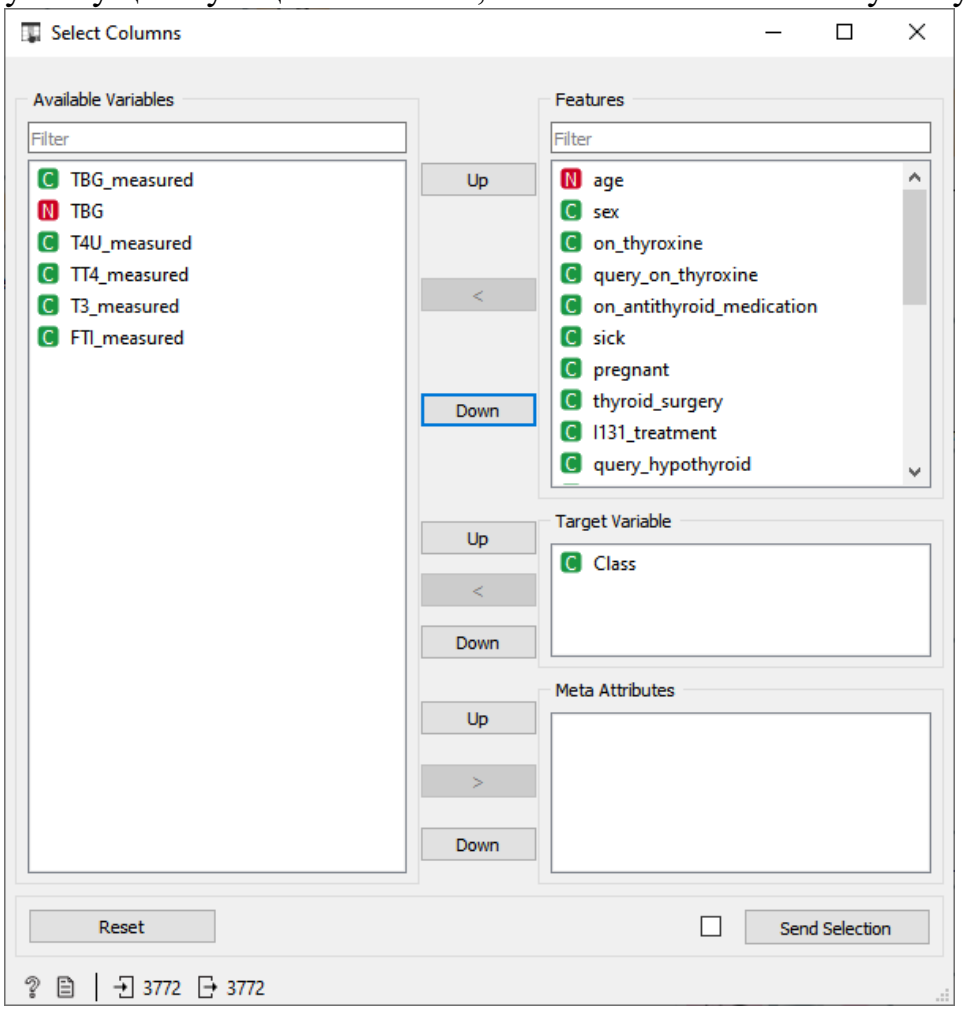


По итогу есть 3772 записи и 30 колонок.
Признаки распределены следующим образом: 7 числовых и 23 категориальных.

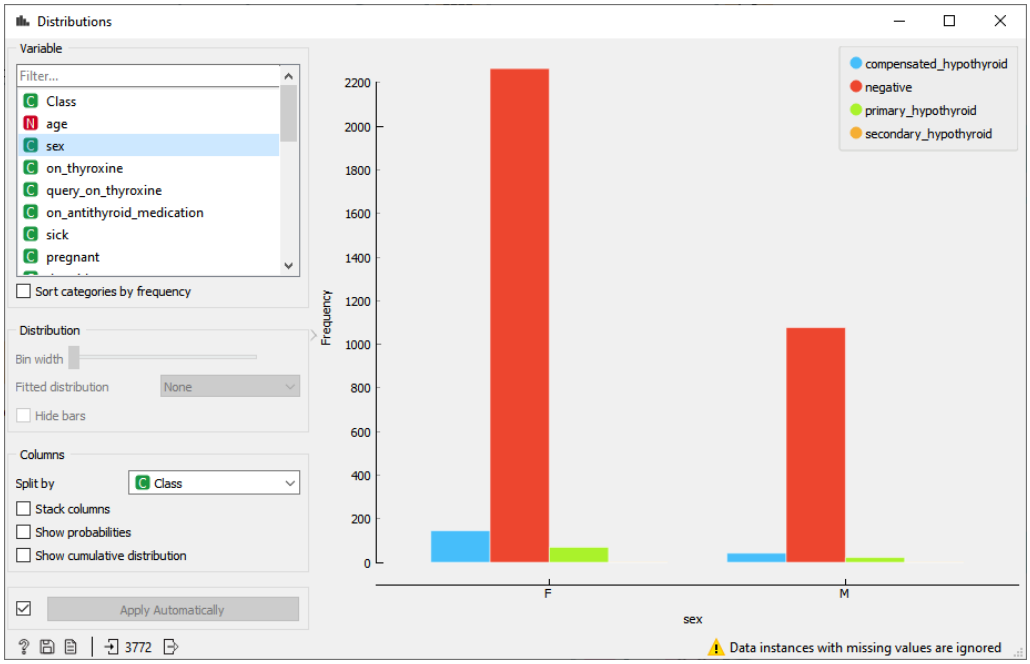
Общий вид на данные, при помощи инструмента Data Table.



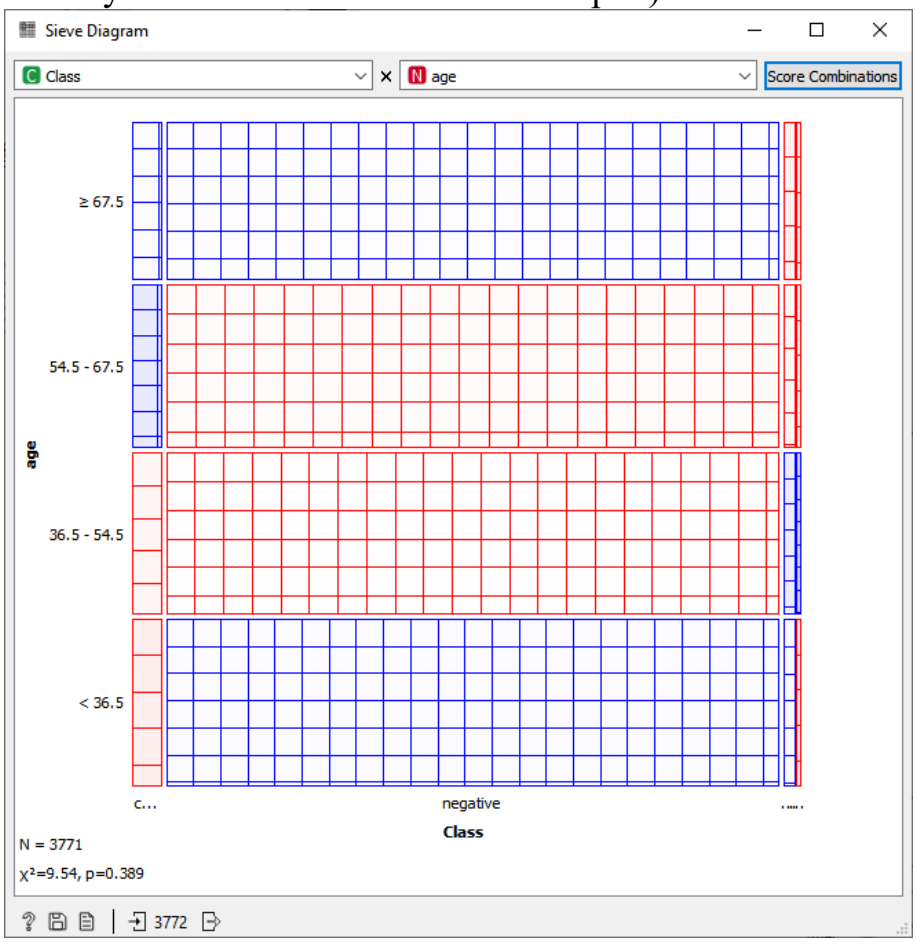
Отбросим часть данных, которые либо же являются по сути дубликатами уже существующих значений, либо же полностью отсутствуют.



Продолжим обзорное исследование данных при помощи инструмента Distributions:

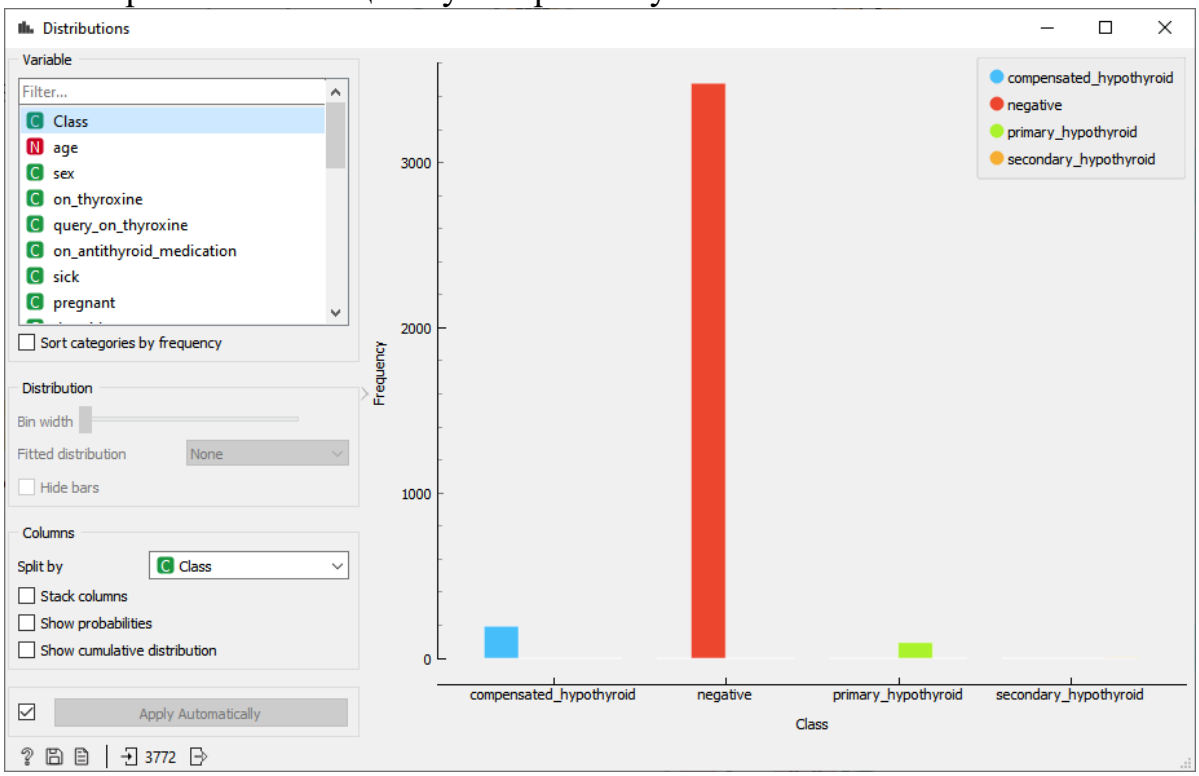


Как видно, в выборке больше женщин и у они чаще болеют (скорее всего это обусловлено особенностью выборки).



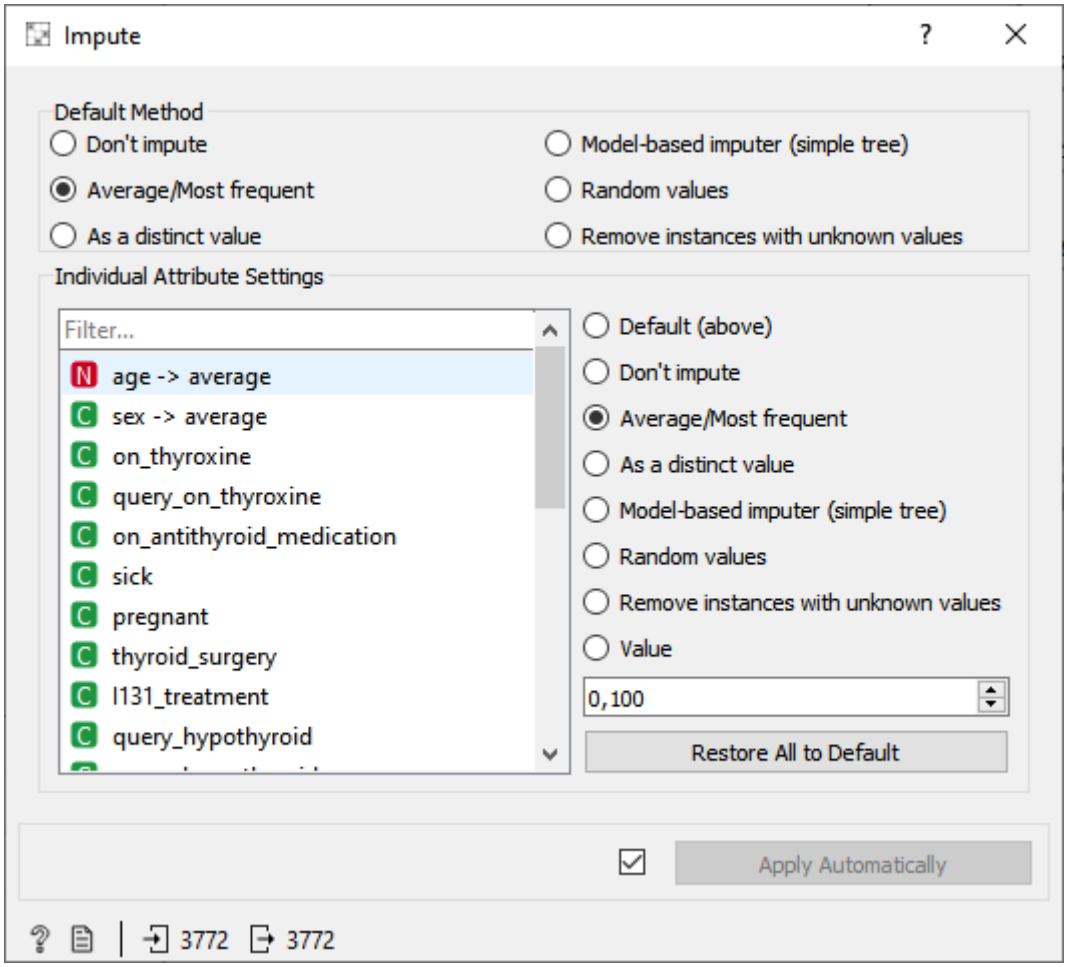
Как видно, большую часть выборки составляют люди старшего возраста и заболевание имеет более “возрастной” характер.

Посмотрим ближе на целевую переменную:

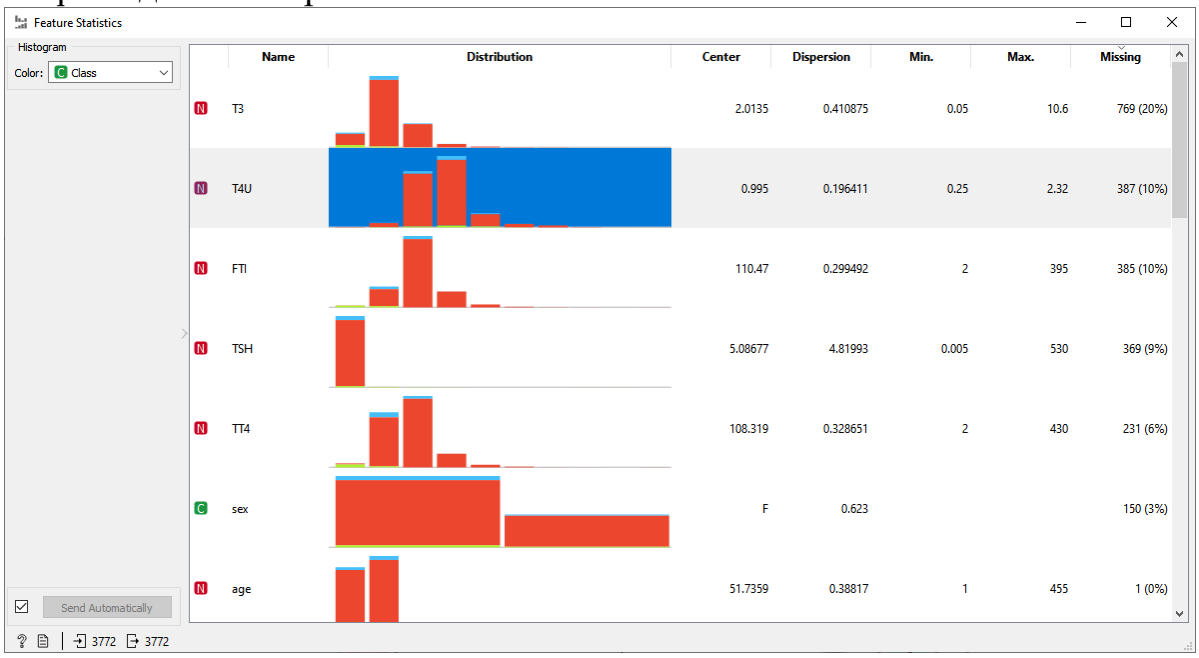


Целевая переменная Class состоит из следующих значений: ('negative', 'compensated_hypothyroid', 'primary_hypothyroid', 'secondary_hypothyroid'). Причем, количество объектов класса 'secondary_hypothyroid' очень мало.

Заполненим пропуски при помощи блока Impute.

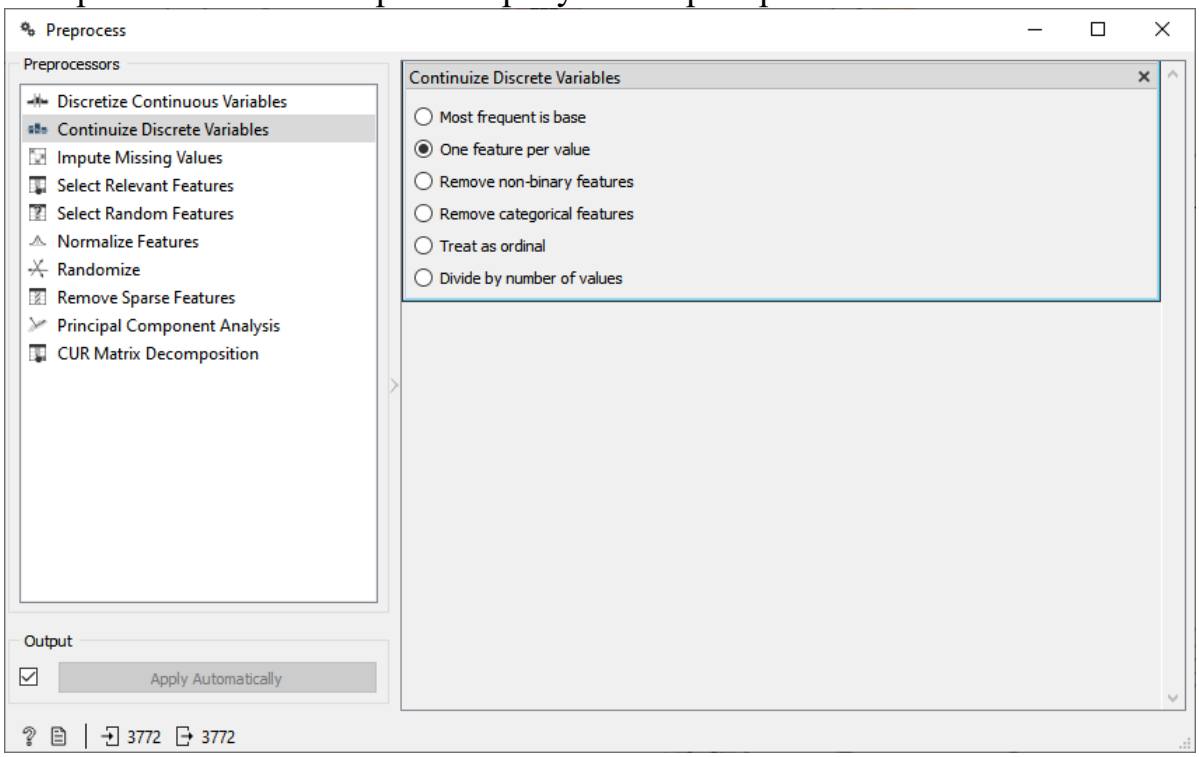


Заполнять будем средним (для численных признаков), либо же most frequent для категориальных.

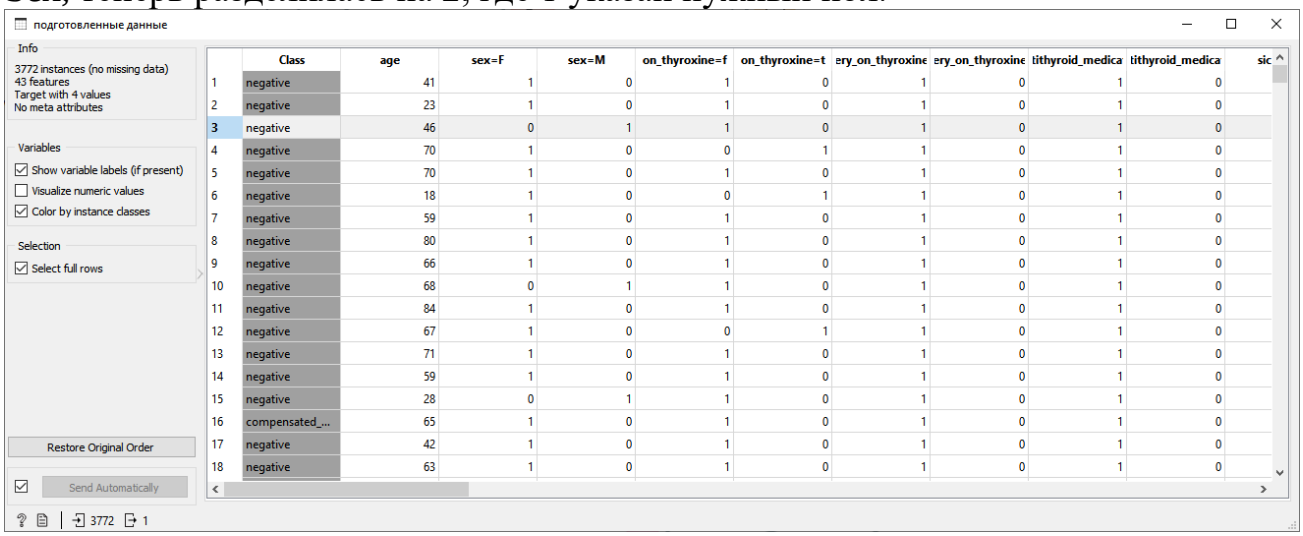


Всего потребовалось заполнять пропуски в 7 признаках:
T3, T4U, FTI, TSH, TT4, Age, Sex.

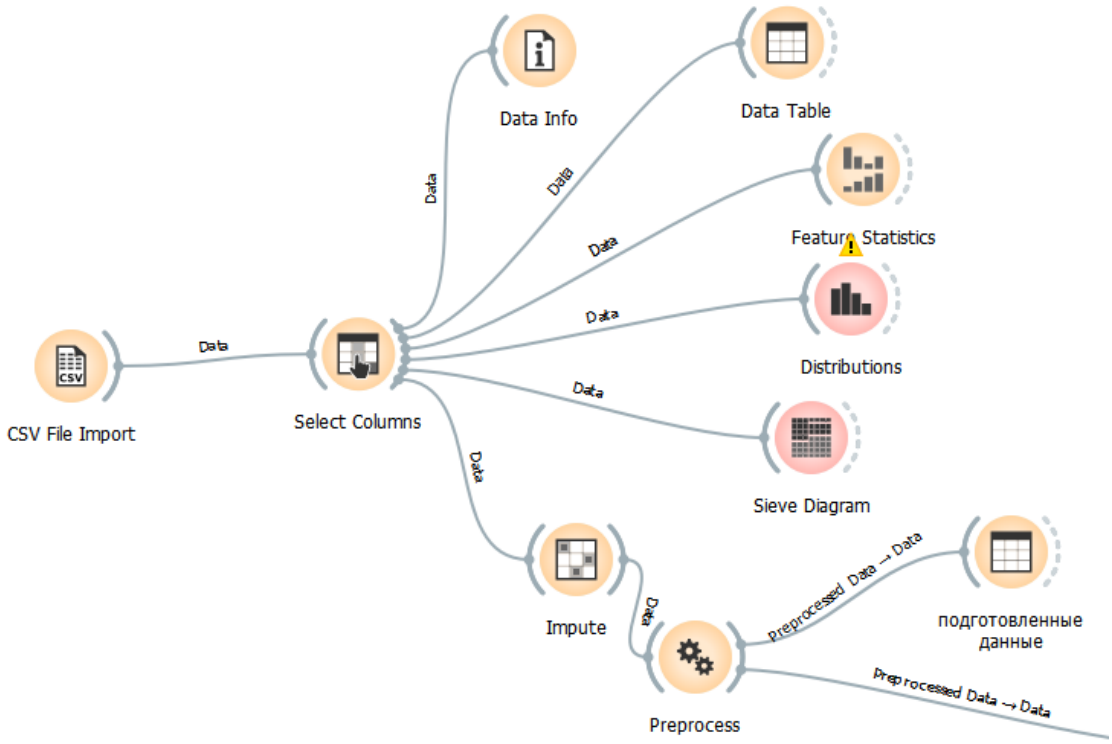
Теперь остается кодировать категориальные переменные.
Для этого воспользуемся блоком Preprocess -> One Feature per value, которая позволяет совершить требуемые преобразования.



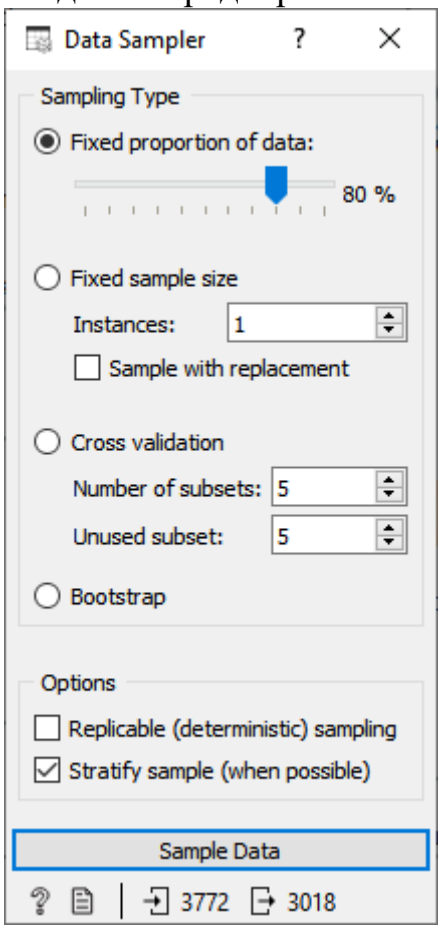
Получившийся результат, после One Hot Encoding, видно, что колонка Sex, теперь разделилась на 2, где 1 указан нужный пол.



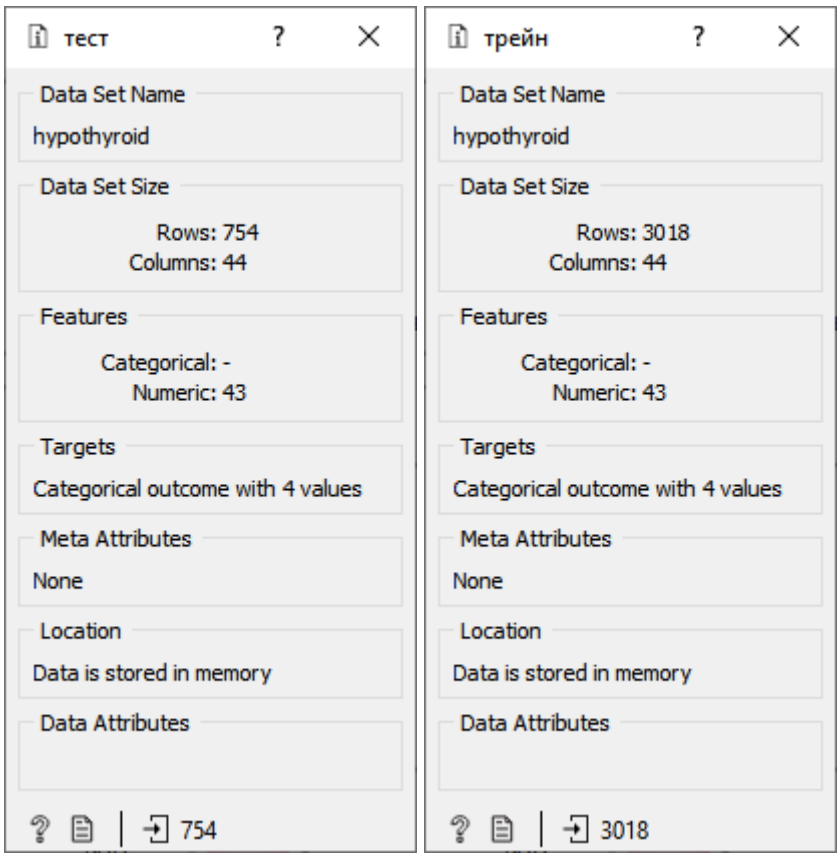
Общий вид dashboard’a Orange перед разделением данных.



Разделим предобработанные данные на обучающую и тестовую выборки.



Данные делим в соотношении 80% / 20 %, где на 80% придется обучающий набор данных, а проверять будем на оставшихся 20%. Такое соотношение является достаточно распространённым и позволит обучить и проверить работу алгоритмов машинного обучения, также не стоит забывать про стратификацию, для того, чтобы пропорция целевых переменных в подвыборках была сохранена.

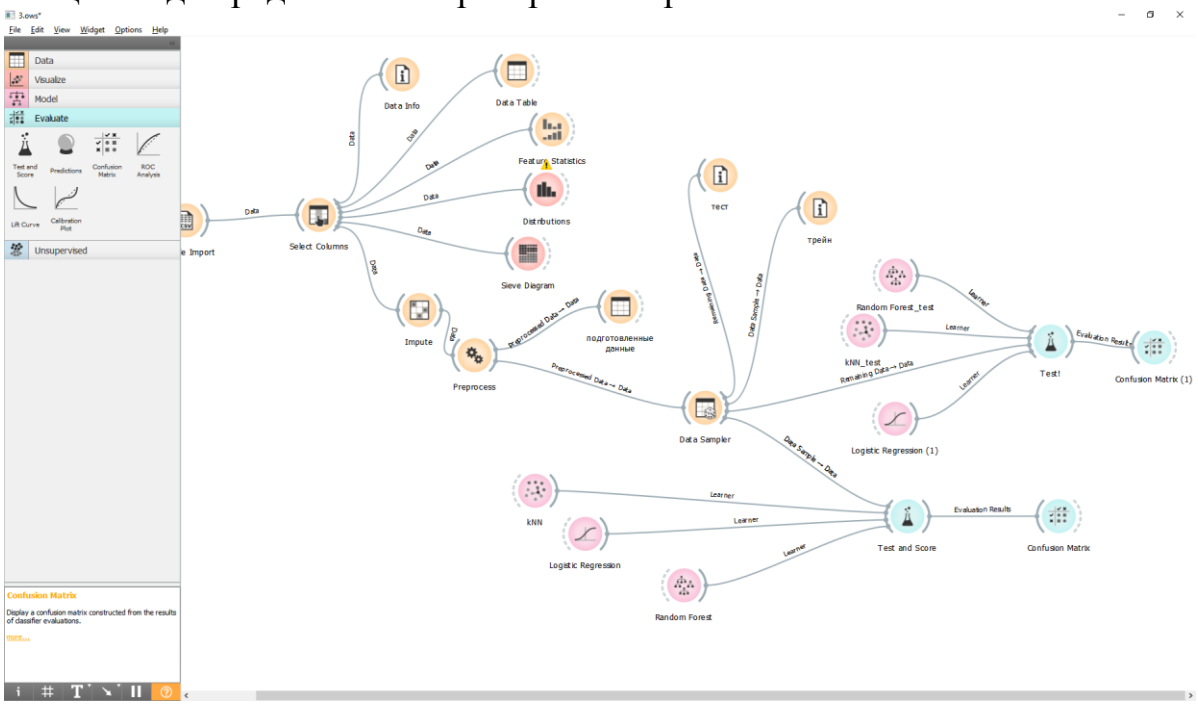


Как видно, разбиение произошло нужным образом.

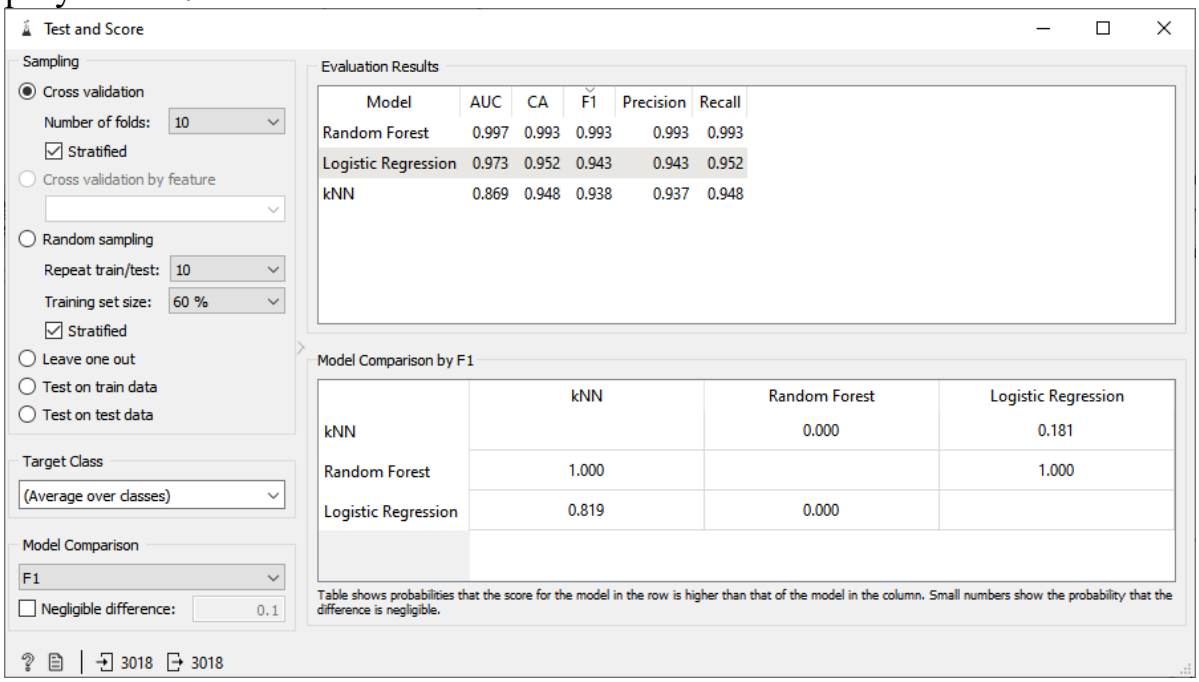
Обучим несколько моделей с помощью трех алгоритмов построения.
В качестве алгоритмов были выбраны:

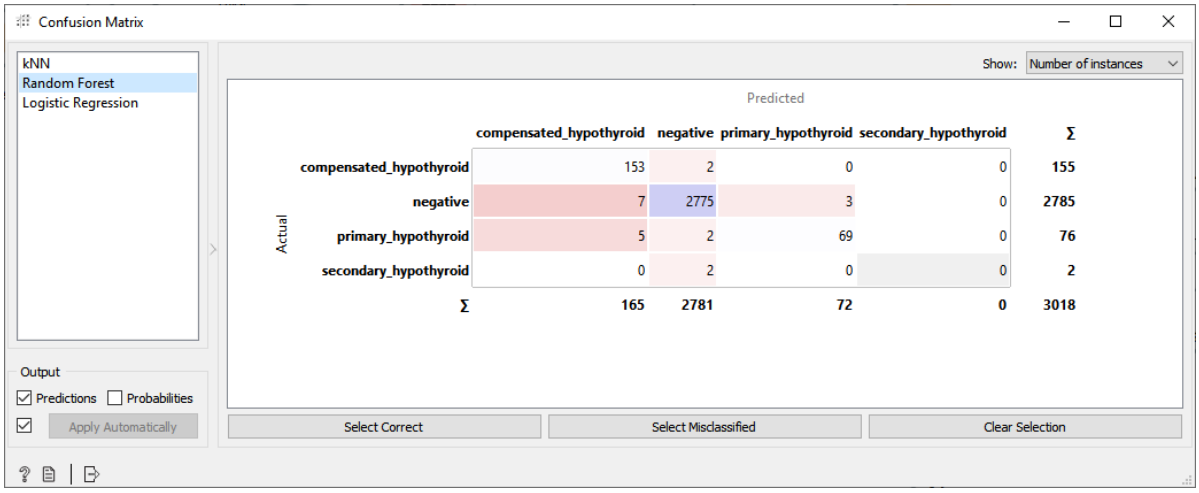
- Алгоритм градиентного бустинга был выбран в силу того, что он сочетает в себе скорость работы и содержит простую, но эффективную идею того, что ансамбль слабых моделей в совокупности дает хороший результат.
- Алгоритм k-ближайших соседей.
  Он был выбран потому, что в основе его лежит простая гипотеза о том, что можно судить о классе объекта (исходя из выбранной метрики) по его соседям.
- Логистическая регрессия, хороший и быстрый алгоритм, который хорошо подходит для быстрого получения результатов.

Общий вид перед началом проверки алгоритмов:



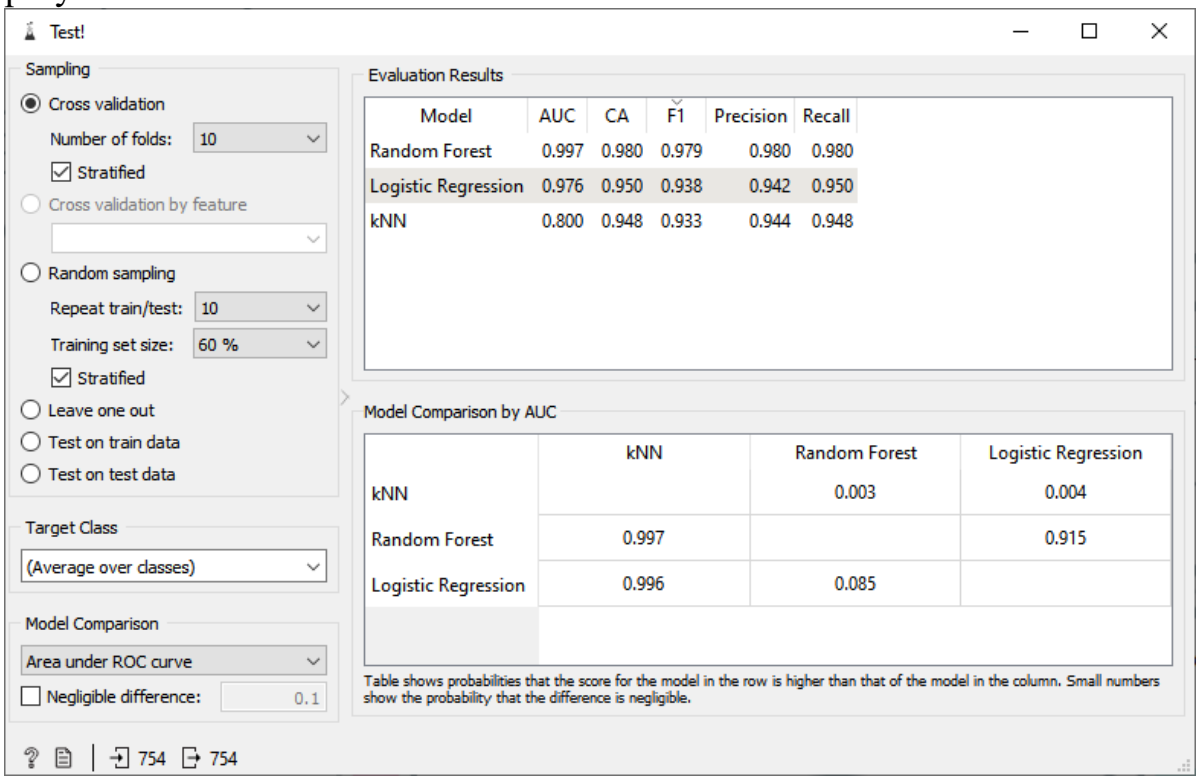
Оценим эффективность моделей на тестовой выборке с помощью матрицы неточностей, критериев полноты Recall и точности Precision. Для обучающей выборки результаты были получение следующие результаты:

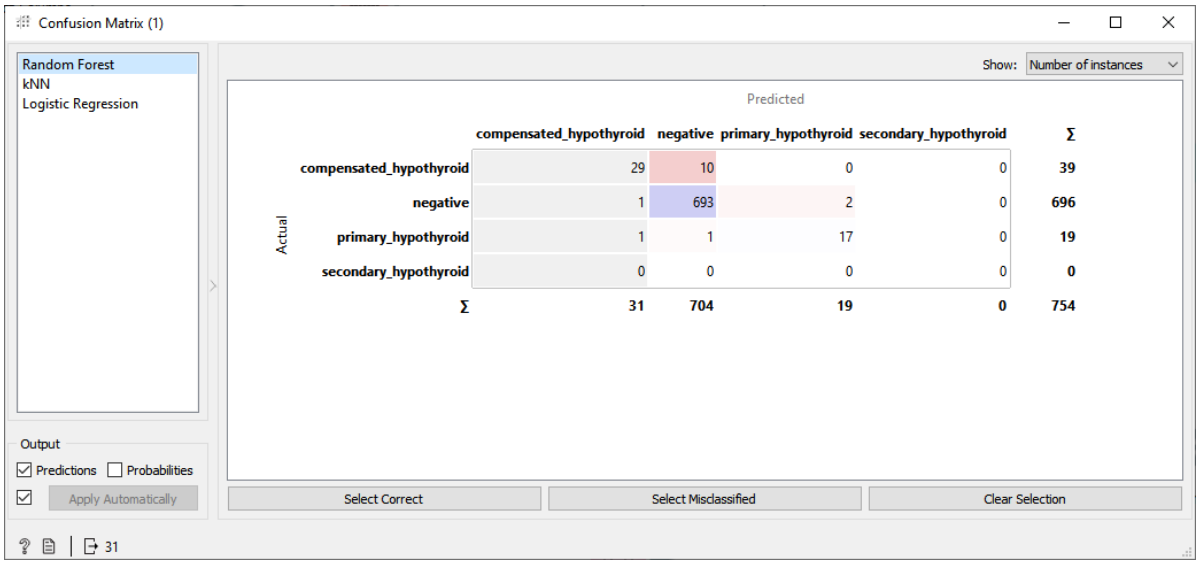




Confusion Matrix для RF.

Для тестовой выборки результаты были получение следующие результаты:





Confusion Matrix для RF.

Заключение.

В данной работы мы провели: работу с датасетом, а именно:

- Загрузили данные в Orange
- Произвели исследовательский анализ данных:
  - получили объём исследуемых данных
  - получили число атрибутов и их типы данных
  - посмотреть распределение числа примеров классов
  - провели исследование других признаков
  - выполнили преобразование категориальных атрибутов
  - заполнили пропущенные значения в выборке
  - отобрали часть данных
- Разделили данные на обучающую и тестовую выборки
- Обучили 3 модели-классификатора машинного обучения
- Оценили эффективность моделей на тестовой выборке с помощью матрицы неточностей, критериев полноты *Recall* и точности *Precision* и гармоничной меры F1 Score.

Среди сравненных классфикаторов и параметров наилучшими оценками обладают (в порядке убывания) по метрике F1 Score (которая по сути является гармоничной мерой, которая совмещает в себе precision и recall):

- Градиентный бустинг на деревьях принятия решений
- Логистическая регрессия
- KNN