Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
Национальный исследовательский университет ИТМО
Факультет инфокоммуникационных технологий

Отчет
Проектирование интеллектуальных систем в управлении
Практическая работа №2

Студент:
Катаева Вероника Алексеевна
Группа: К3342
Преподаватель:
Бережков Андрей Вячеславович

Санкт-Петербург
2021

Цель работы: Провести анализ датасета «Титаник» в KNIME.

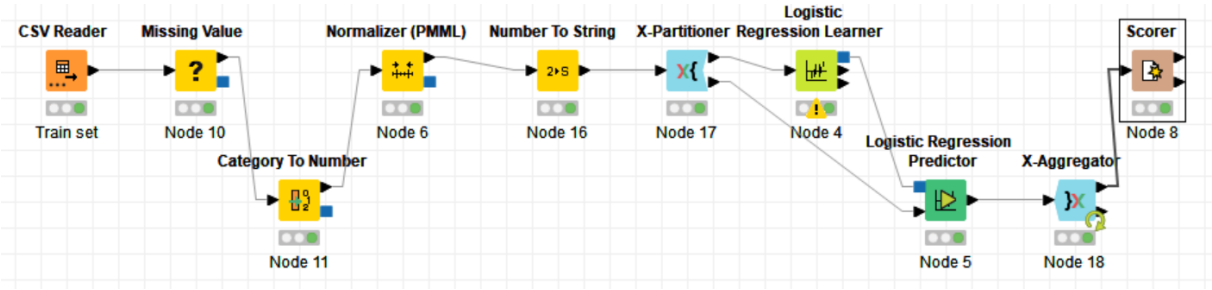
Задачи:

1. Необходимо вытроить процессы аналогично тетрадке (https://www.kaggle.com/pramodsivakumar/titanic-simple-eda-and-predictions).

Выполнение работы:

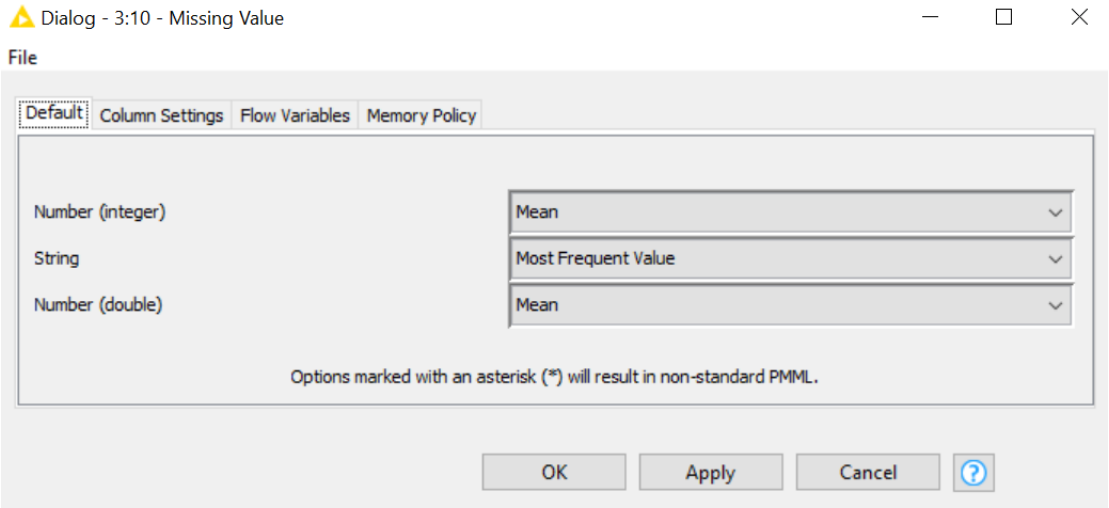
Построение моделей:

Поток работ линейной регрессии:



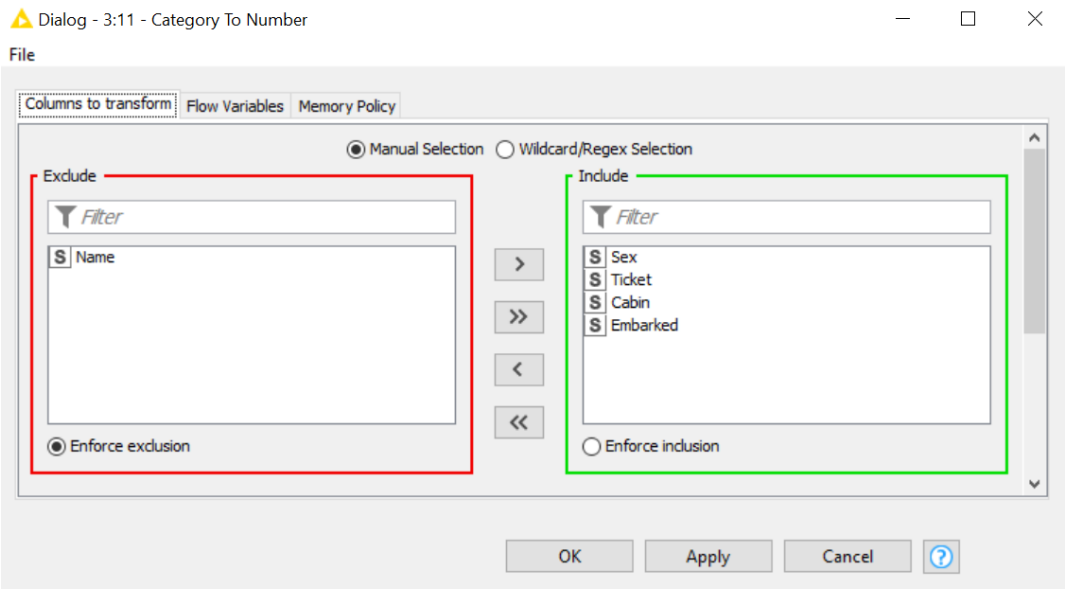
Состоит из следующих узлов:

1) Чтение csv файла – данных о Титанике.
2) Обработка недостающих значений.

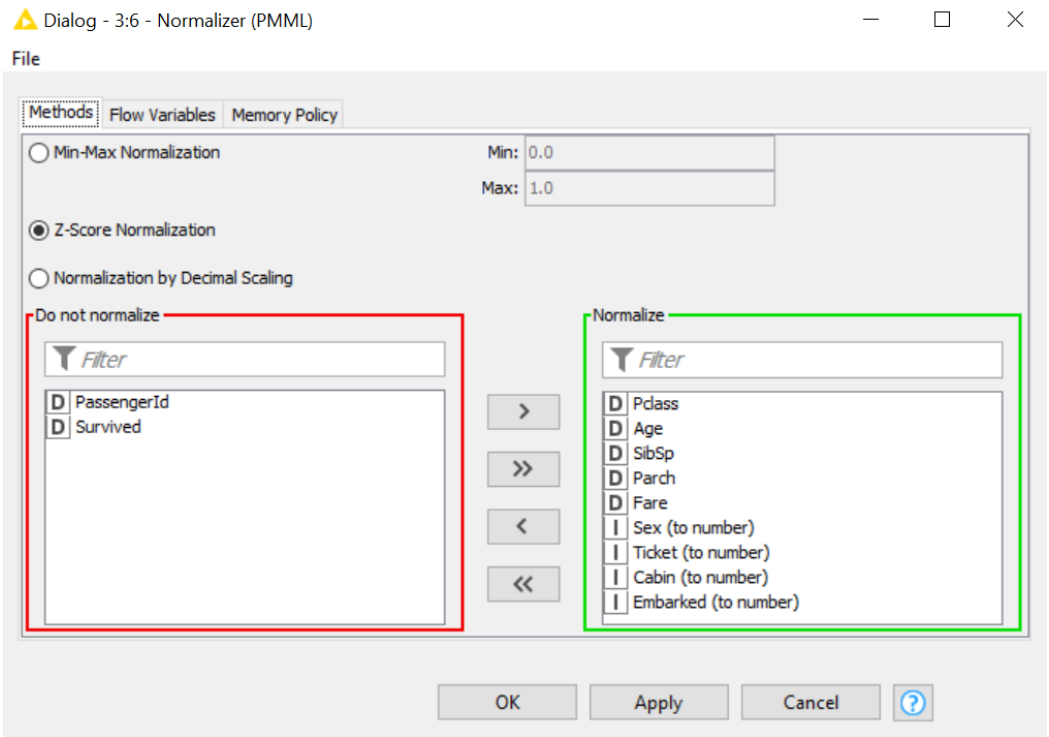


Числа заменяются на среднее значение, строки – на самое популярное.

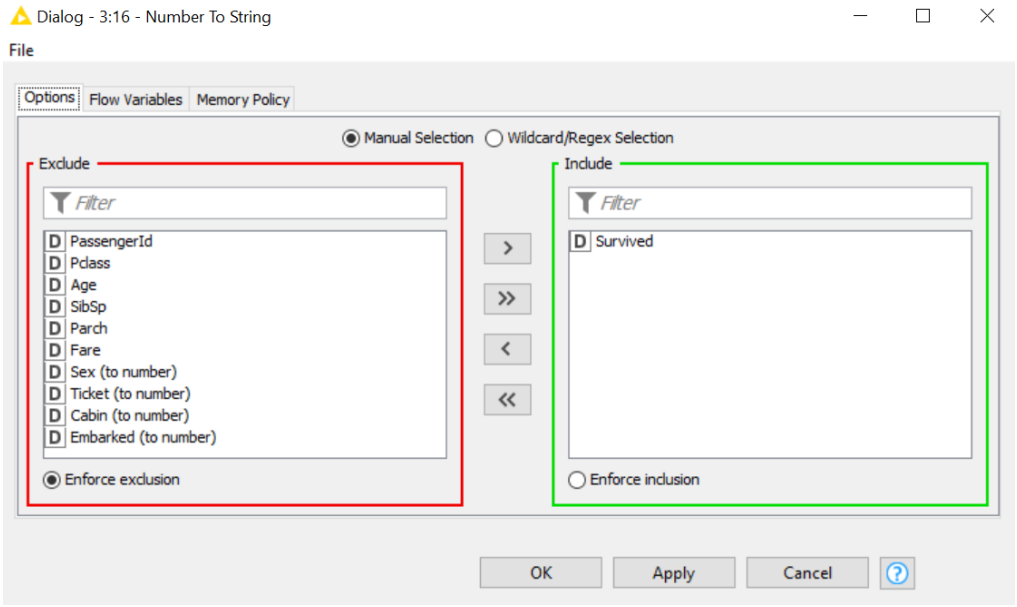
3) Замена категориальных данных на числовые (за исключением имени пассажира).



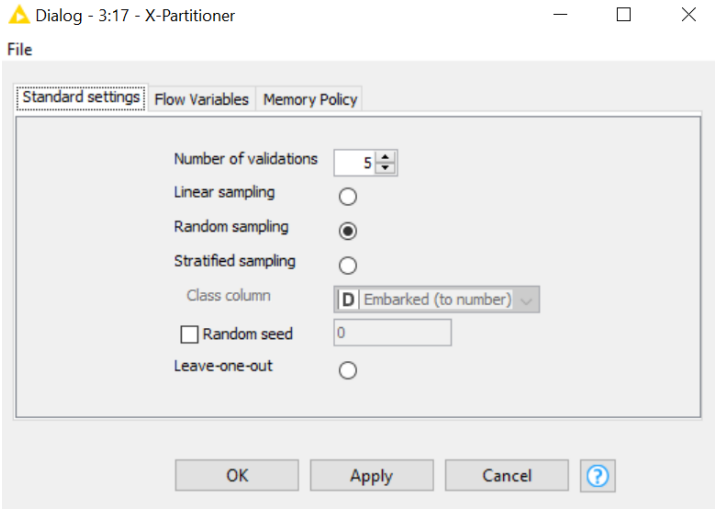
4) Нормализация данных (за исключением ID пассажира и целевую переменную).



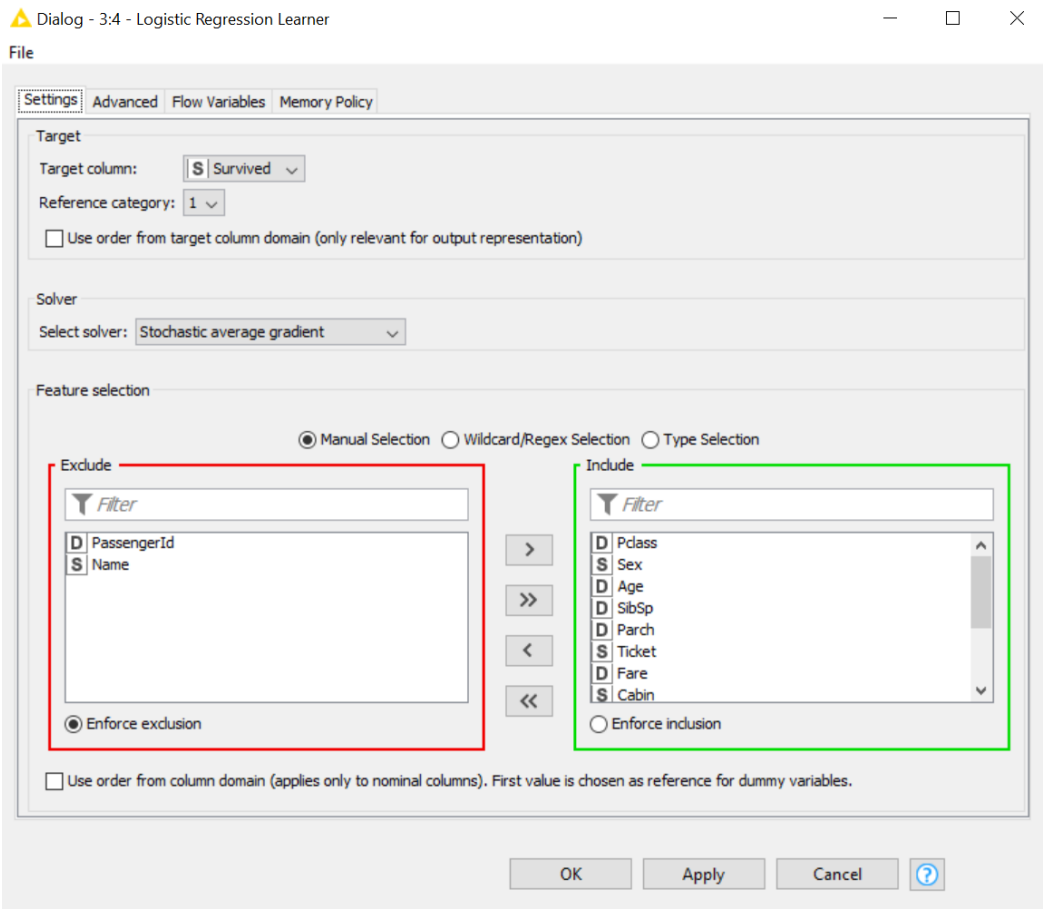
5) Конвертация числа в строку (для создания номинальных данных для предсказания).



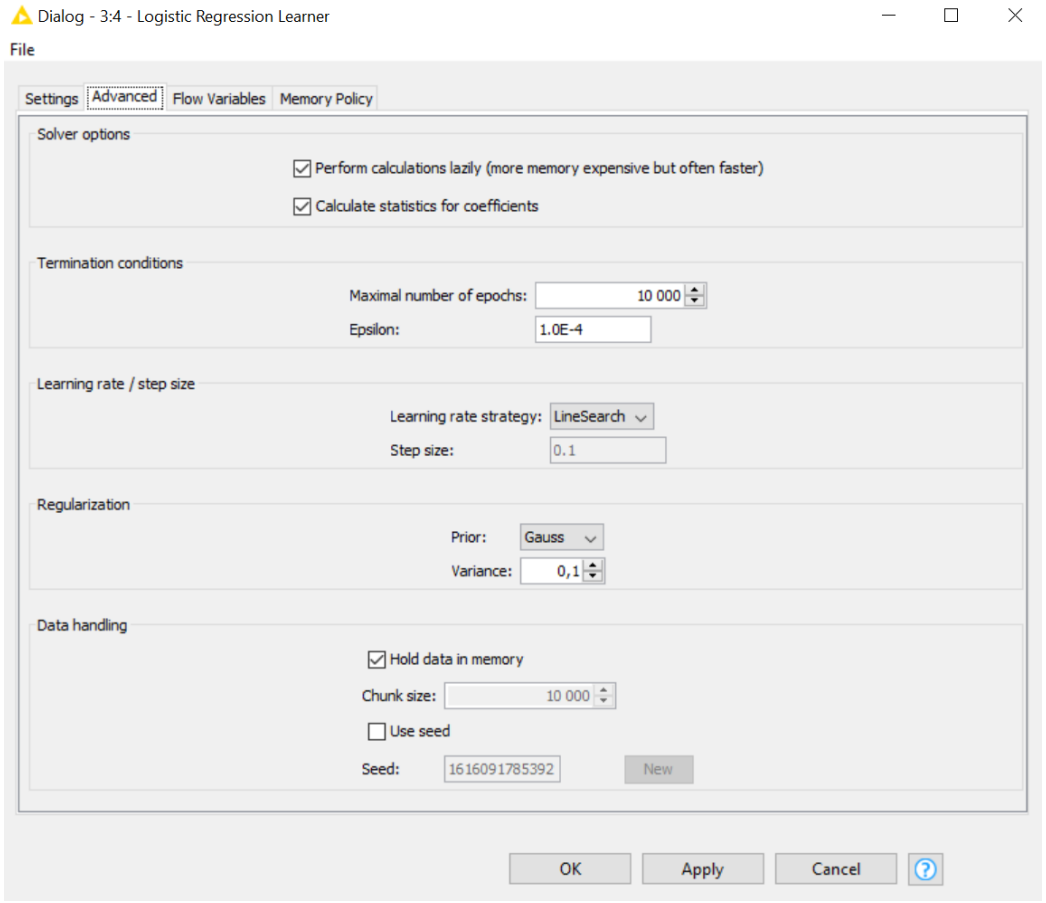
6, 9) 5-кратная кросс-валидация.



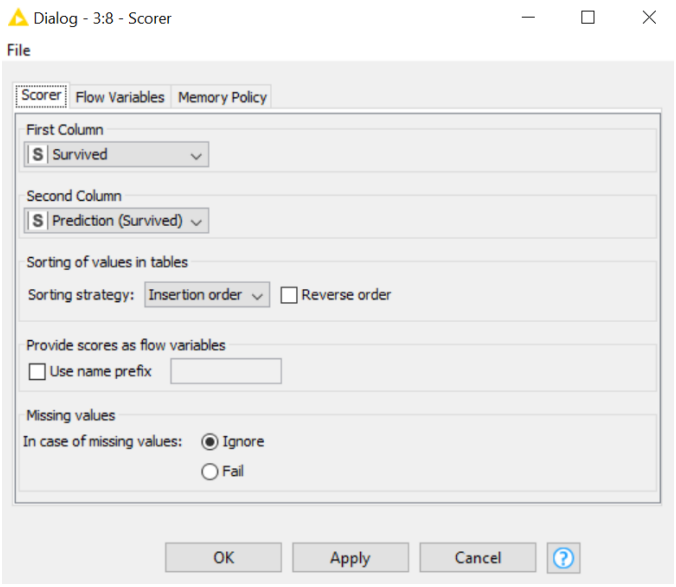
7, 8) Логистическая регрессия.



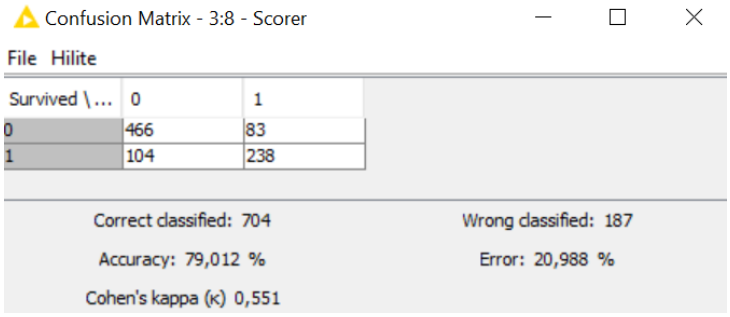
Гиперпараметры логистической регрессии:



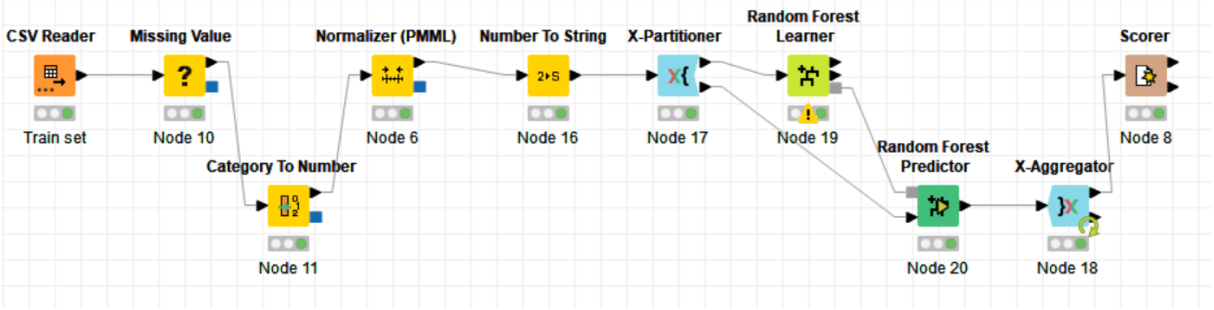
10) Счетчик.



Результаты логистической регрессии:

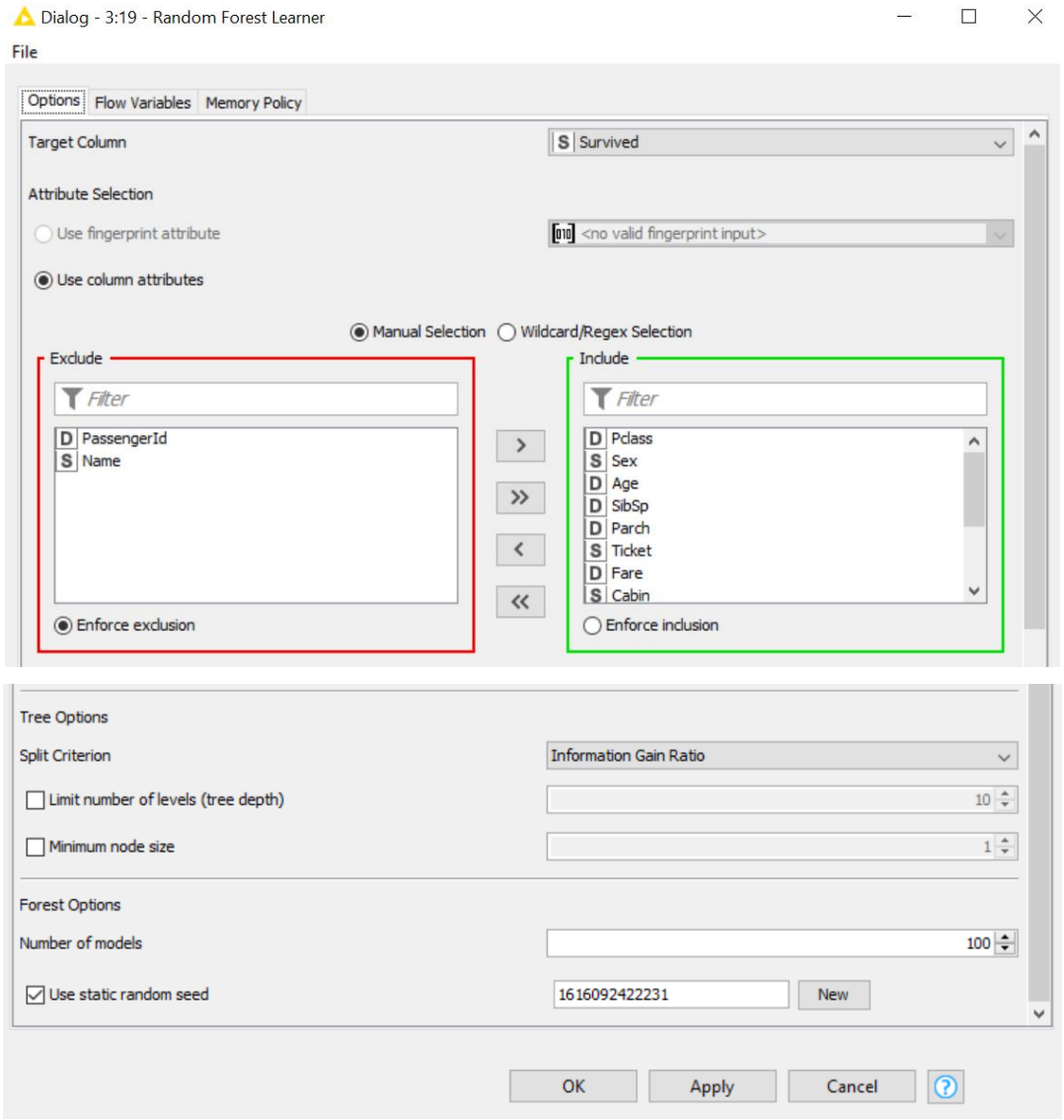


Поток работ классификации с помощью случайного леса:

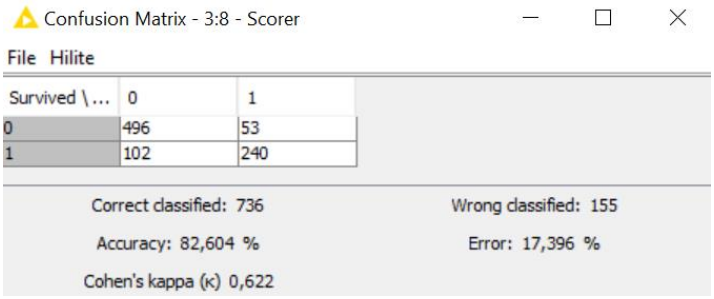


Поток работ содержит схожие узлы – вместо линейной регрессии используется модель случайного леса.

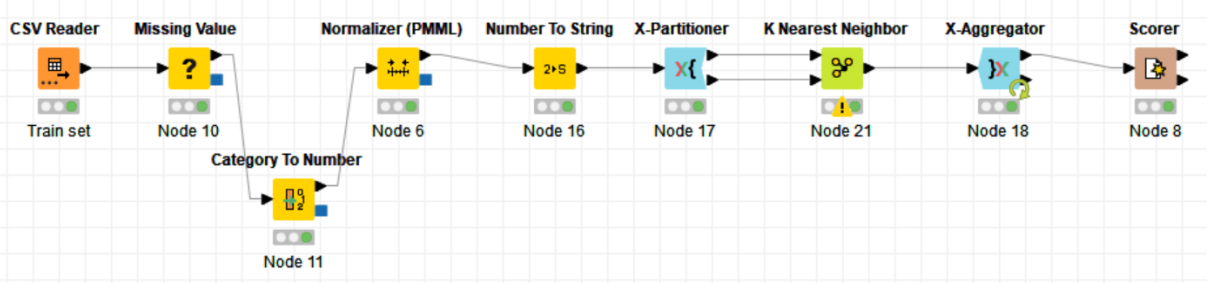
Конфигурация и гиперпараметры случайного леса:



Результаты случайного леса:

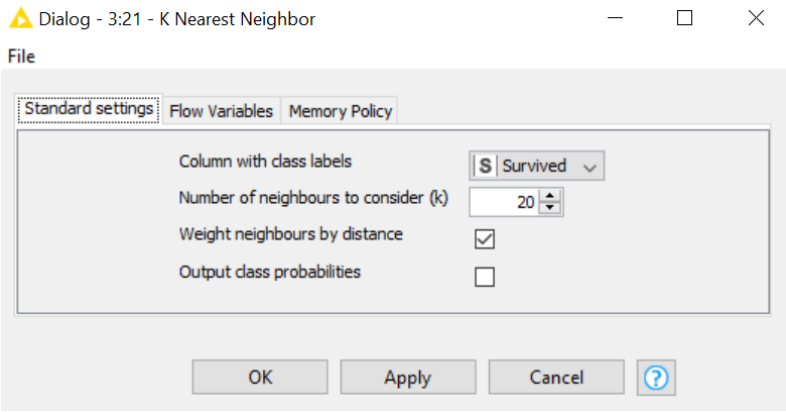


Поток работ метода ближайших соседей:

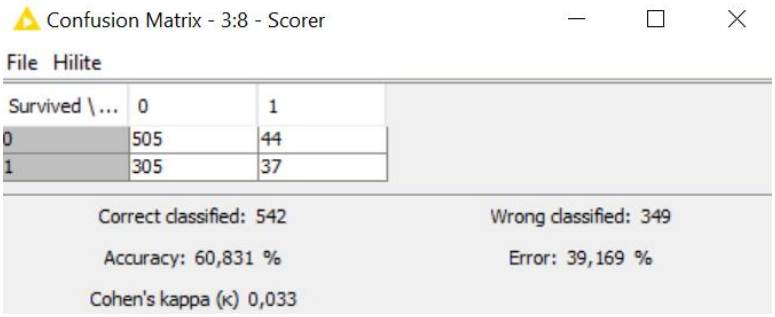


Поток работ содержит схожие узлы – вместо линейно регрессии используется модель *k* ближайших соседей.

Конфигурация и гиперпараметры метода ближайших соседей:



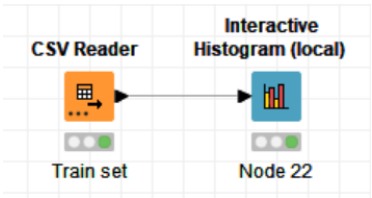
Результаты метода ближайших соседей:



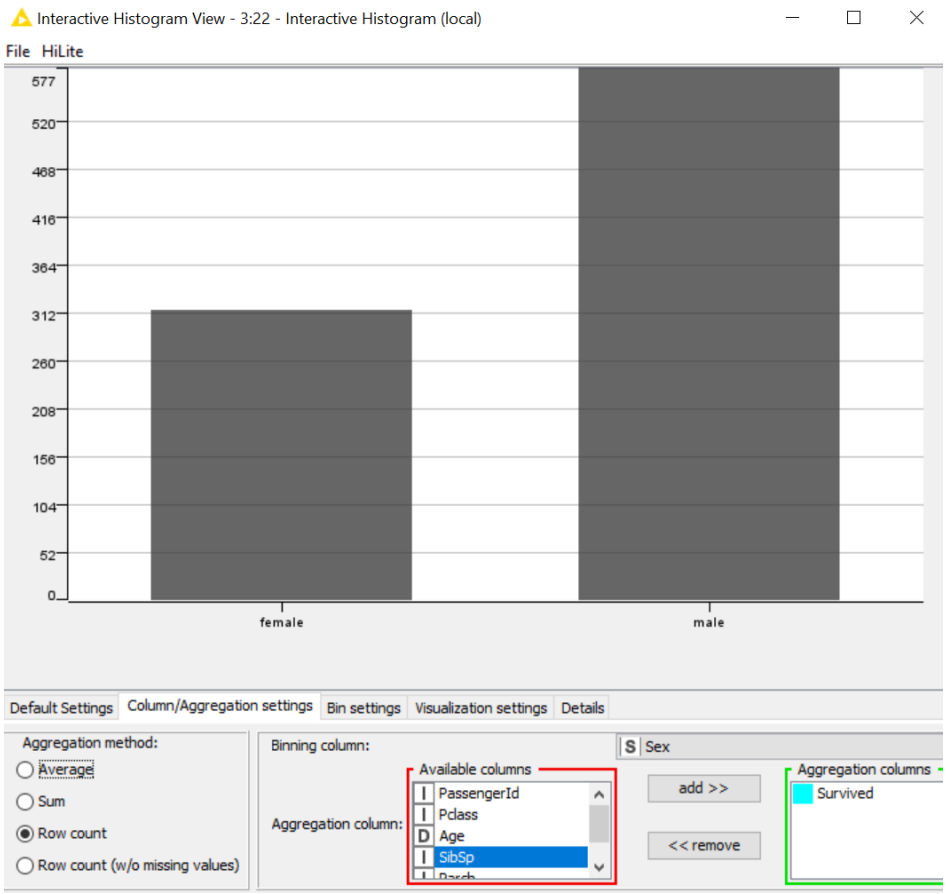
Вывод: таким образом, модель случайного леса в большей степени подходит для представленных данных, так как обладает наивысшей в сравнении с другими моделями долей правильных ответов и лучшими показателями матрицы ошибок.

Визуализация:

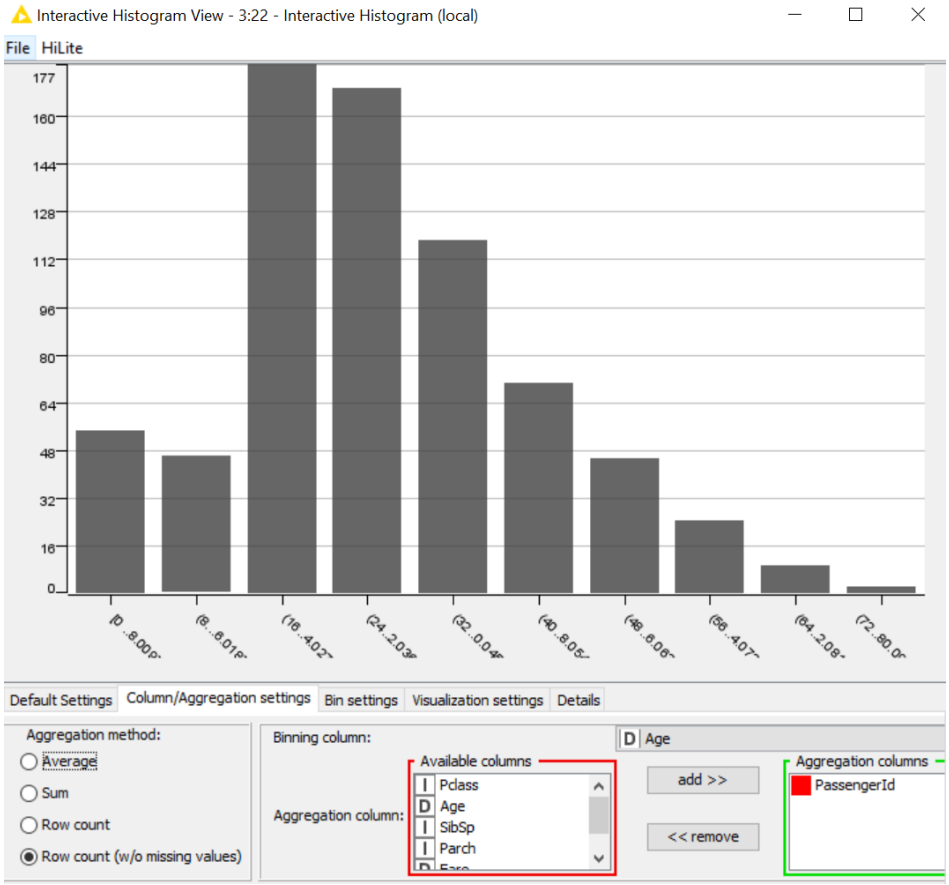
Поток работ гистограммы:



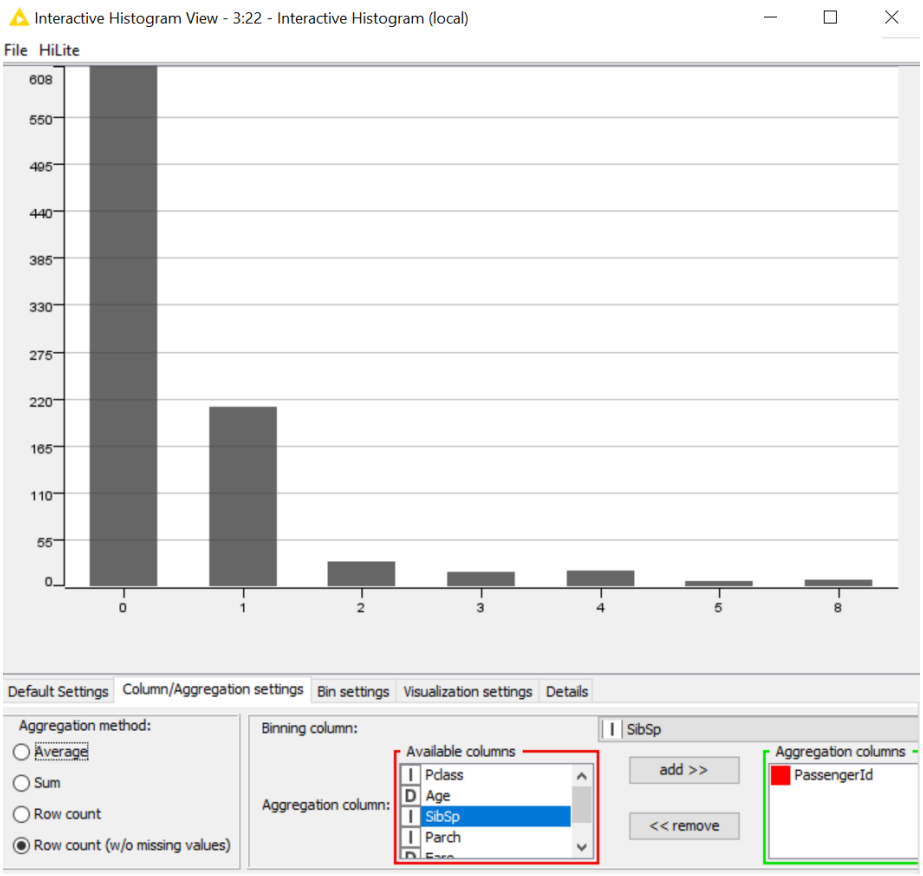
Гистограмма распределения пола в зависимости от того, что человека не спасли:



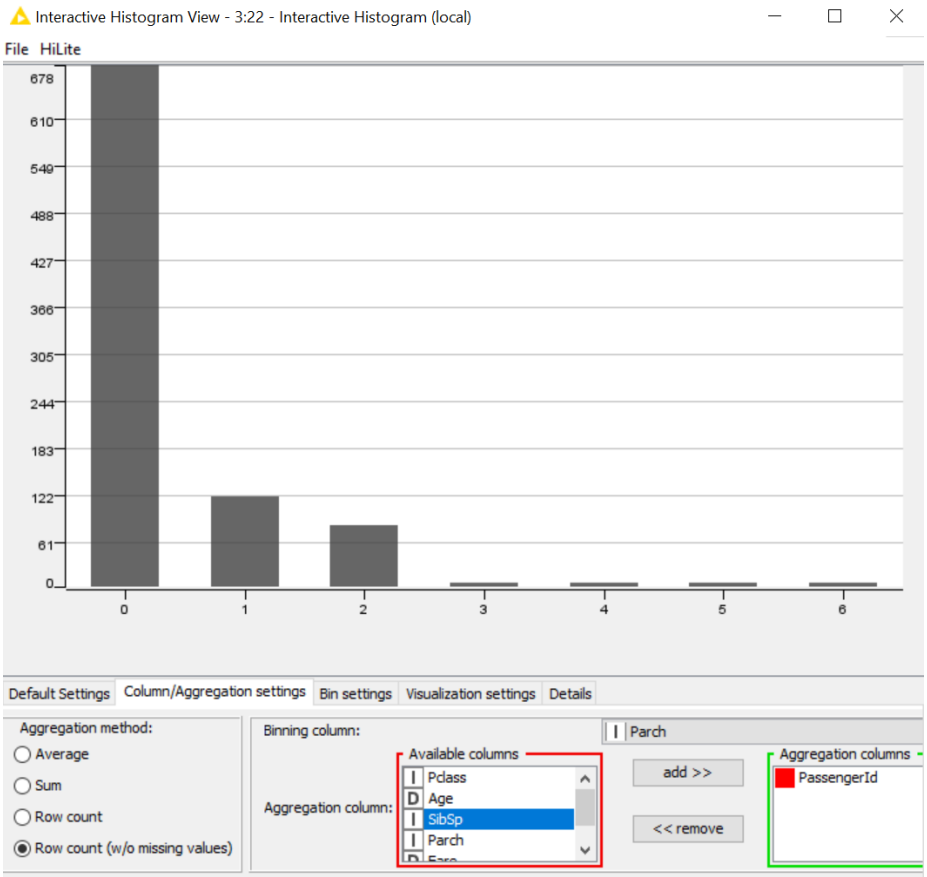
Гистограмма распределения возрастов:



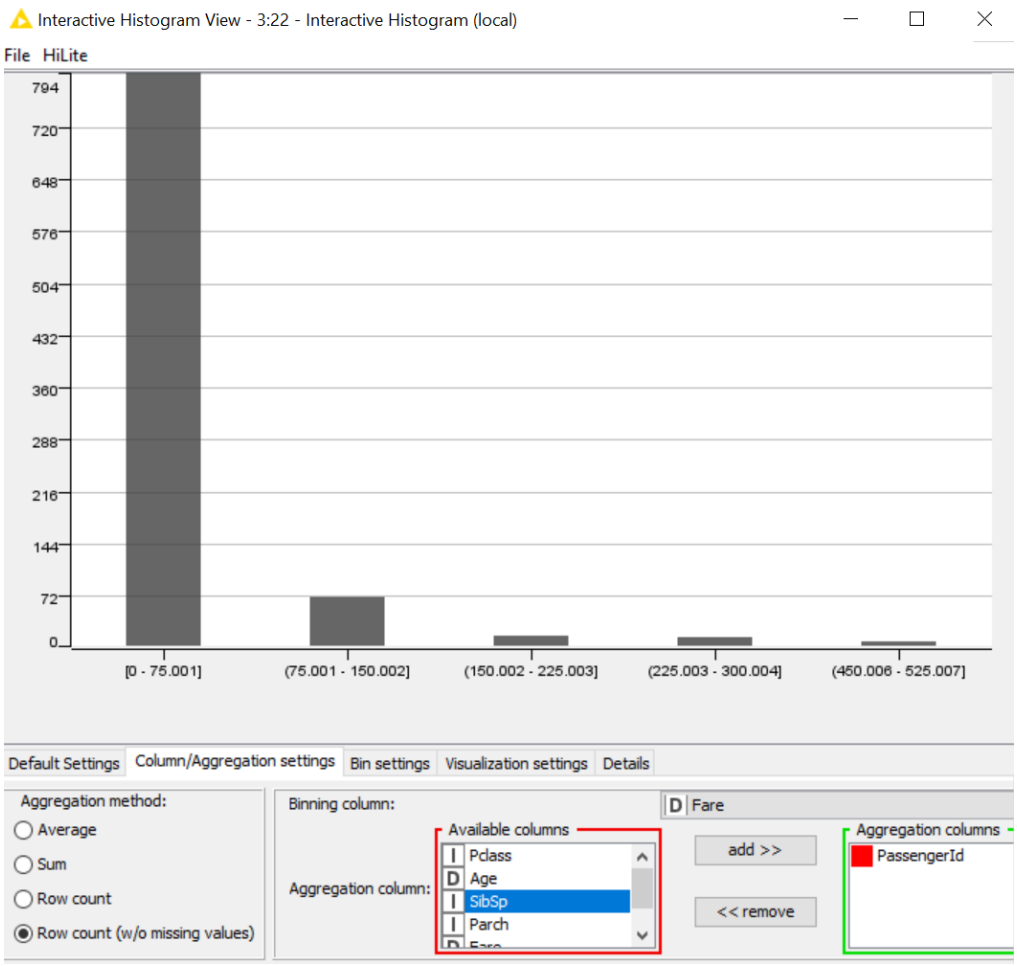
Распределение количества братьев, сестёр или наличия супруга на борту:



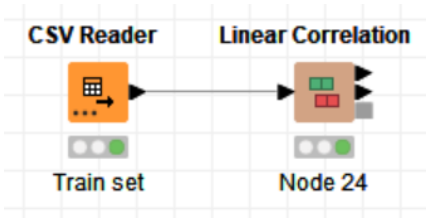
Распределение количества родителей, детей на борту:



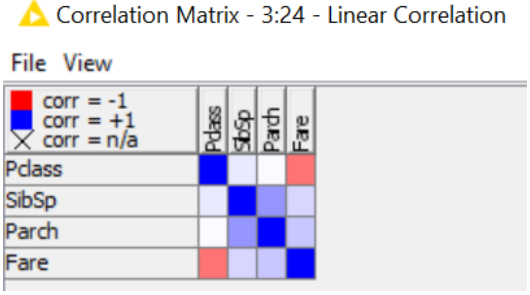
Пассажирский тариф в британских фунтах стерлингов:



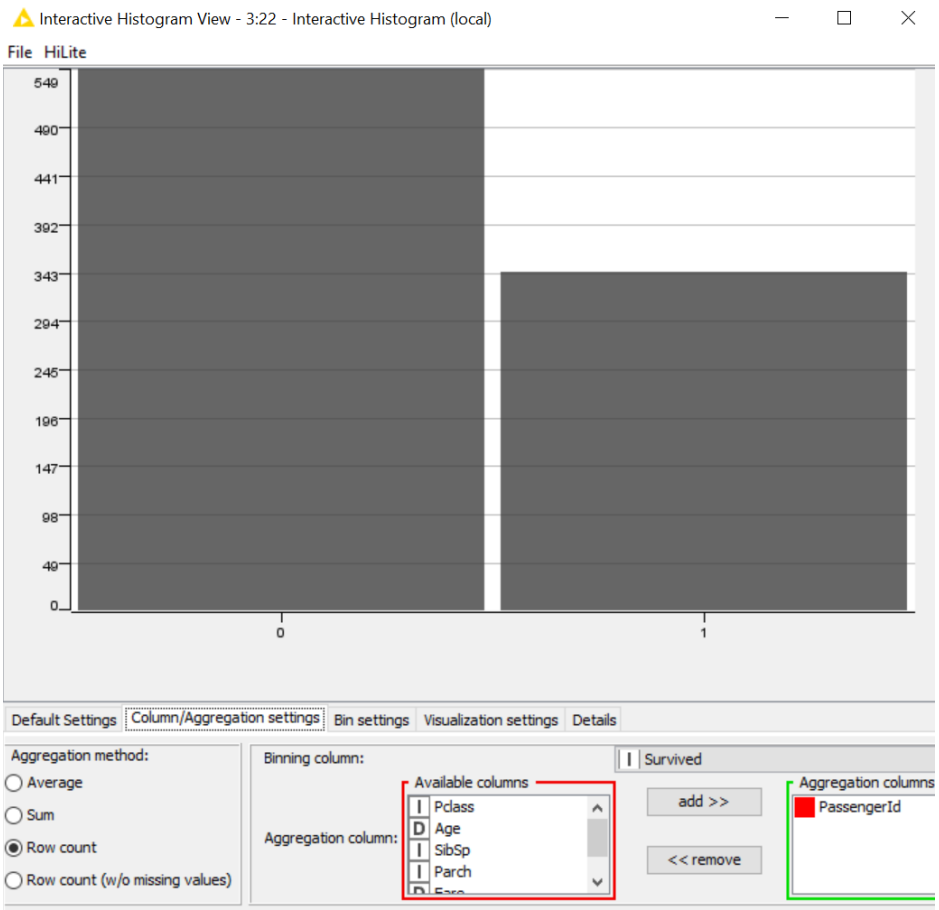
Поток работ линейной корреляции:



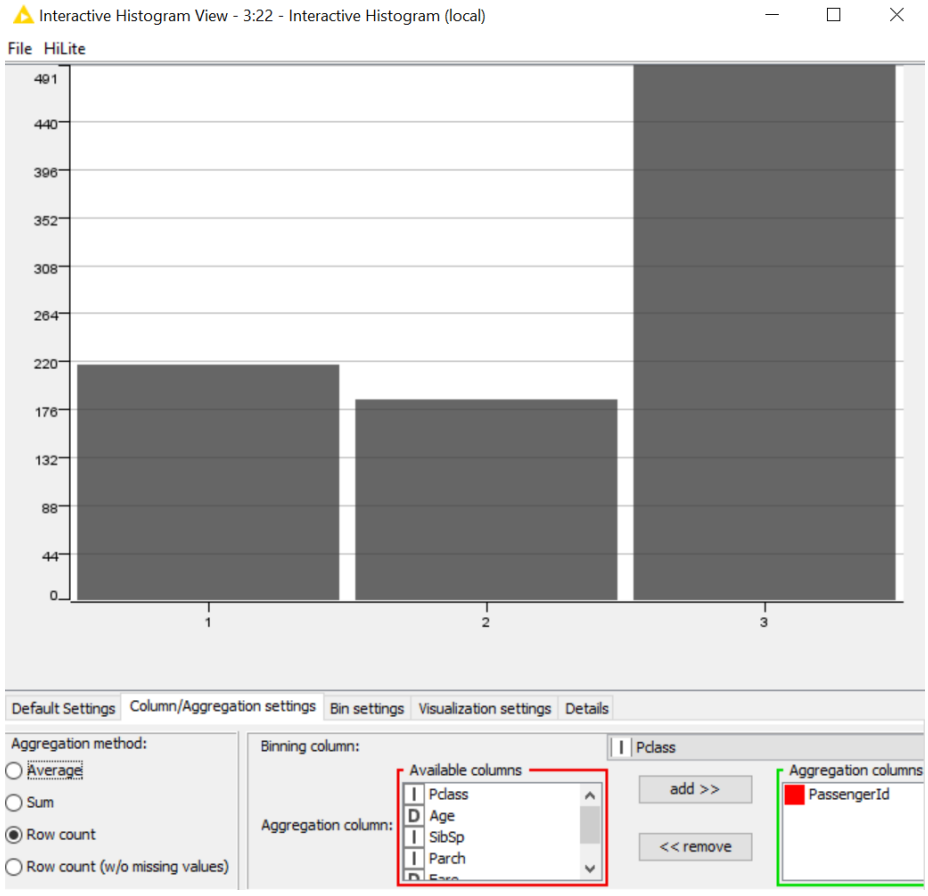
Матрица корреляции между классом пассажира, наличием его братьев/сестер/супруга на борту, наличием его родителей/детей, пассажирским тарифом:



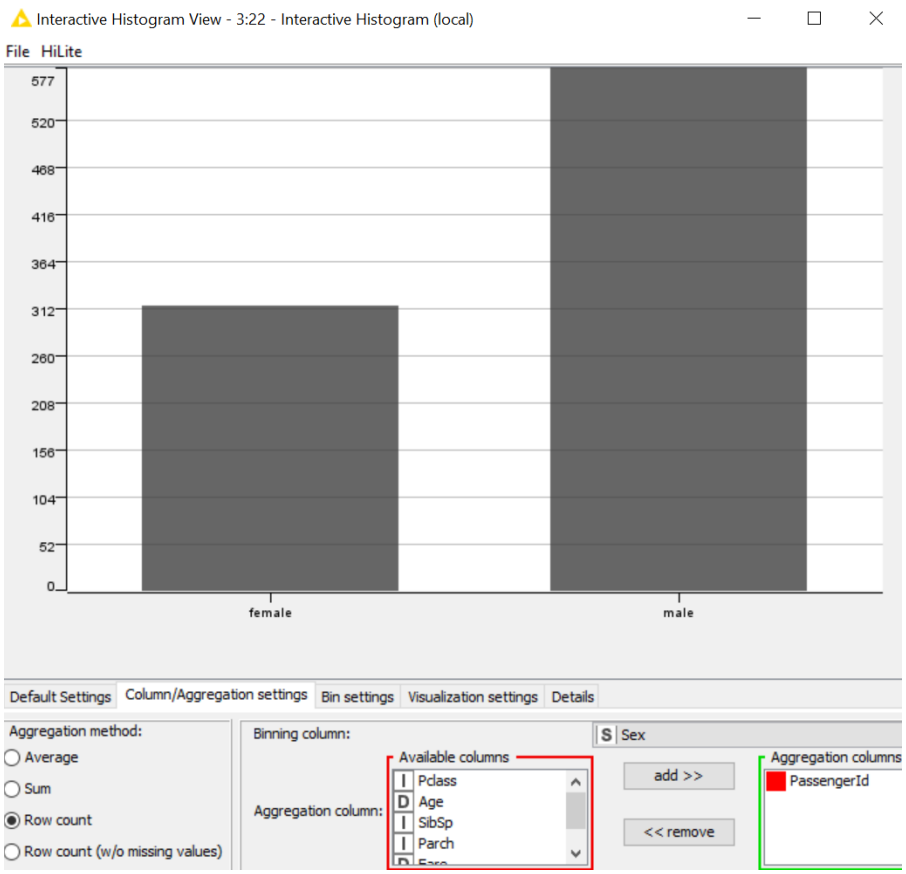
Распределение выживших:



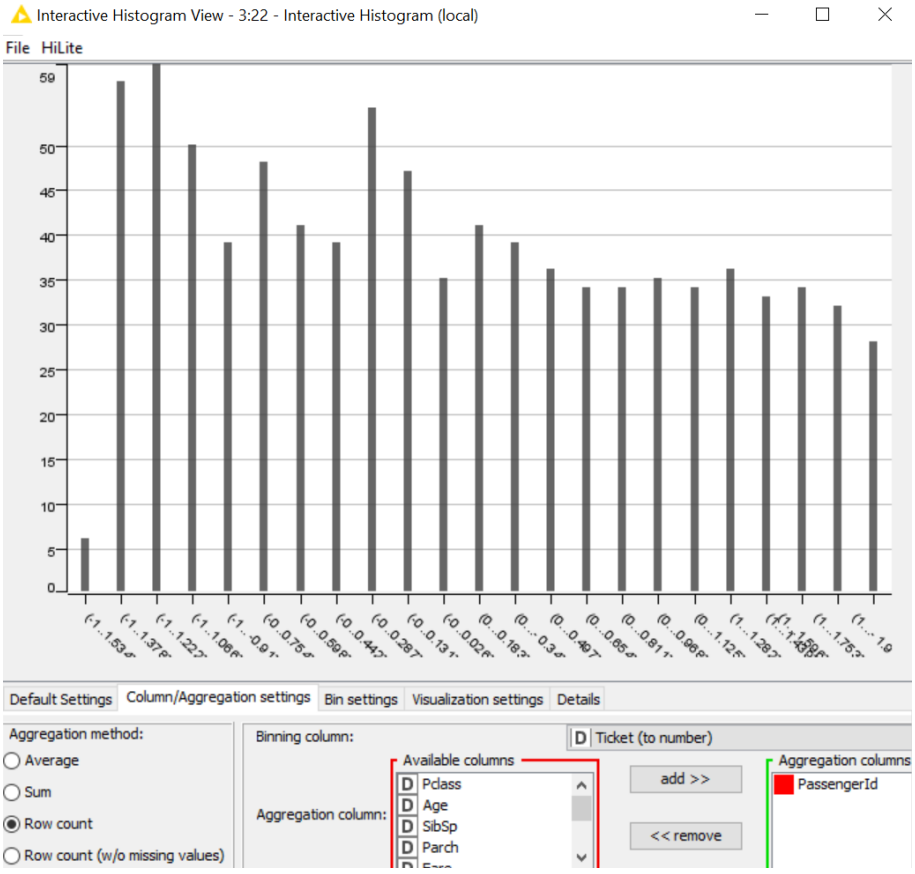
Распределение пассажиров по значению класса:



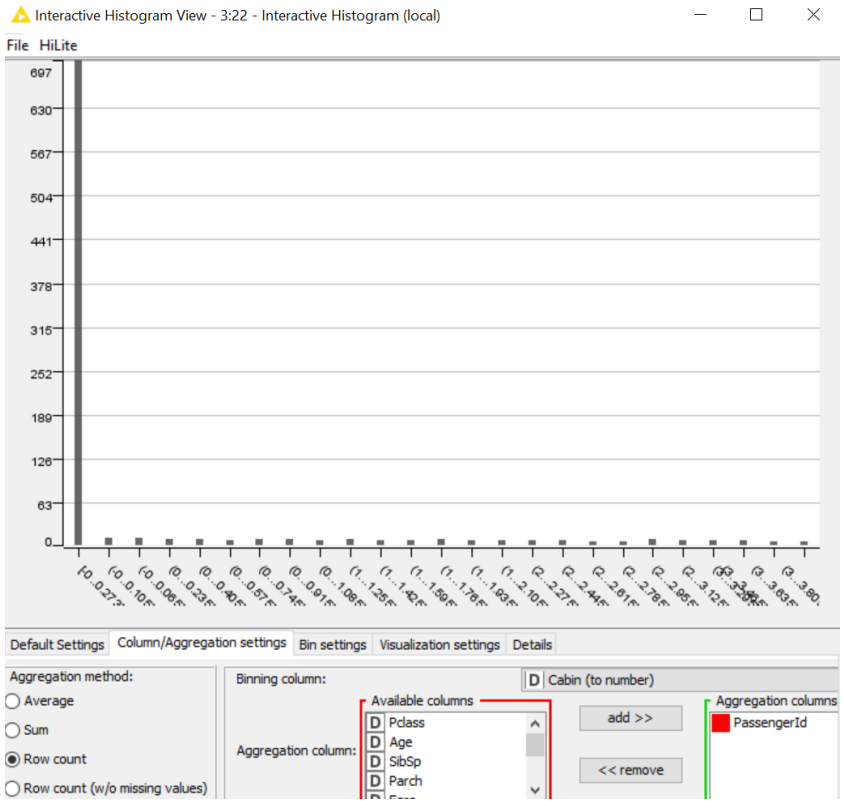
Распределение пассажиров по половому признаку:



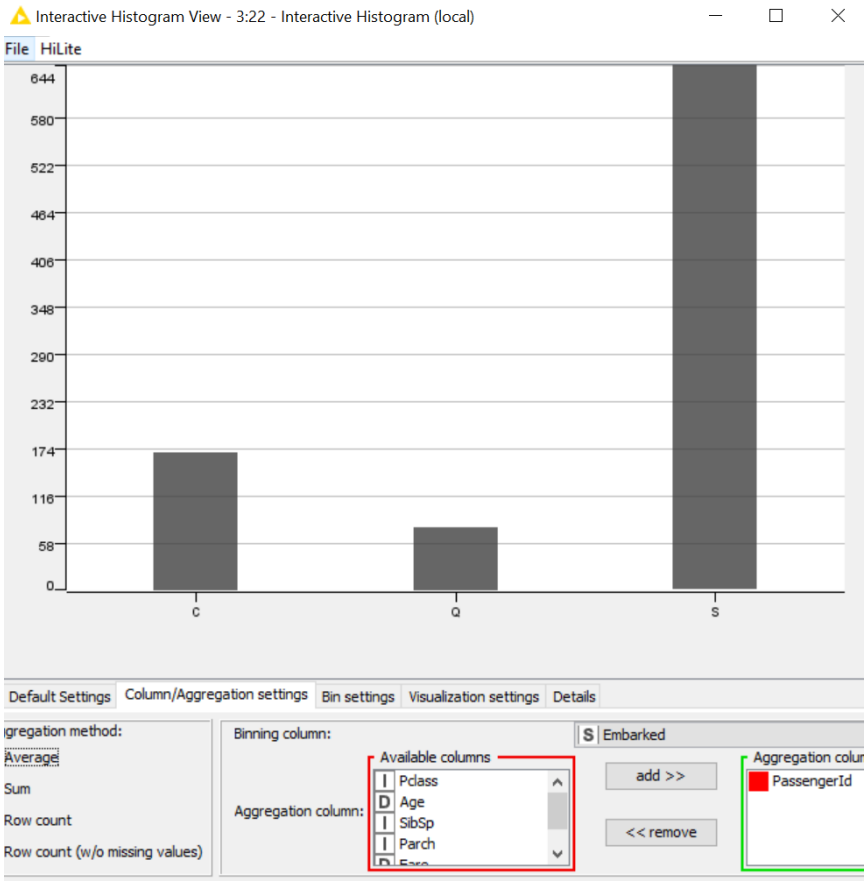
Распределение пассажиров в зависимости от билета (нормализованные значения):



Распределение пассажиров в зависимости от каюты (нормализованные значения):



Распределение пассажиров в зависимости от порта посадки (C = Шербур; Q = Квинстаун; S = Саутгемптон):



Вывод: в ходе лабораторной работы я познакомилась с инструментарием KNIME, а в частности с моделями для анализа и визуализации данных.