УДК 004

# Разработка в Knime регрессионной модели, предсказывающей вес по полу и росту человека

*Звайгзне Алексей Юрьевич*
*Приамурский государственный университет имени Шолом-Алейхема*
*Студент*

**Аннотация**
Целью исследования является разработка регрессионной модели, позволяющей предсказать вес, опираясь на пол и рост человека. Для реализации модели применялась программная система Knime. Данные были взяты из открытых источников на Github. В результате построена регрессионная модель с коэффициентом детерминации R^2=0.9.
**Ключевые слова:** База данных, Knime, линейная регрессия, гистограмма, узел

# Development of a regression model in Knime that predicts weight by gender and height of a person

*Zvaigzne Alexey Yurievich*
*Sholom-Aleichem Priamursky State University*
*Student*

**Abstract**
The aim of the study is to develop a regression model that allows predicting weight based on a person's gender and height. The Knime software system was used to implement the model. The data was taken from open sources on Github. As a result, a regression model with a coefficient of determination R^2=0.9 is constructed.
**Keywords:** Database, Knime, Linear Regression, Histogram, Node

## 1 Введение

### 1.1 Актуальность

Исходя из современного оборота данных человек может потратить кучу времени на расчеты статистических данных именно поэтому появилась актуальность автоматизации данного процесса с возможностью оптимизации указания формул подборки и определения вида и типа данных.

### 1.2 Обзор исследований

М. Р. Бертхолд и др. в публикации рассказывают об нововведениях и возможностях платформы Knime [1]. А. Филлбрунн и др. исследуют возможности воспроизведения полученных данных на любых системах позволяющая быстро передавать и повторять полученные данные из

исследований [2]. С.Бейскин и др. с помощью дополнительного плагина для Knime автоматизируют рутинные расчеты в рабочей среде [3]. М. Мазанец и др. сравнивают с другими продуктами хеминформатики выделяя Knime как лучший инструмент для ученых [4].

### 1.3 Цель исследования

Цель исследования - разработать регрессионную модель в Knime, позволяющую предсказать вес, на основе данных о поле и росте человека.

## 2 Материалы и методы

Данные в формате CSV были взяты с сайта Github [5]. Для изучения узлов программы основным источником информации служил официальный форум платформы Knime [6].

## 3 Результаты и обсуждения

В Knime все элементы называются Узлами (Nodes). Особенностью работы в программе является, то, что после каждого изменения в любом из узлов, его необходимо Execute (выполнить) вручную, так же в программе есть несколько видов связывания узлов, порядок связывания крайне важен для получения конечного результата. Названия всех узлов, описанных в статье, являются точными, так как в программе по категориям есть множество аналогичных необходимо искать именно по названию, а не по изображению. Для начала необходимо загрузить исходные данные в Knime, так как база данных храниться в CSV формате в программу можно её загрузить посредству CSV Reader, там необходимо нажать «Browse…» и через файловый менеджер выбрать нужную базу данных (рис. 1).

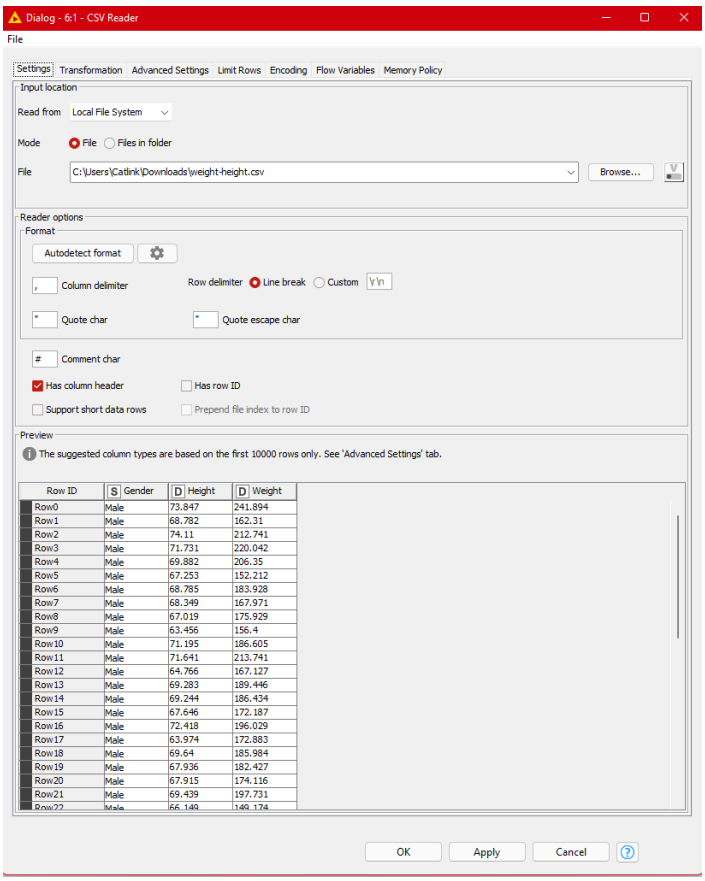


Рисунок 1. Вид окна Configuration в Узле CSV Reader

После этого необходимо перевести данные в столбцах в сантиметры и килограммы. Для этого на рабочую область нужно добавить 2 узла «Math Formula» и соединить их последовательно, чтобы соединить узлы достаточно зажать левую кнопку мыши на значке выходящих данных узла и довести указателем мыши до значка входящих второго узла (рис. 2).

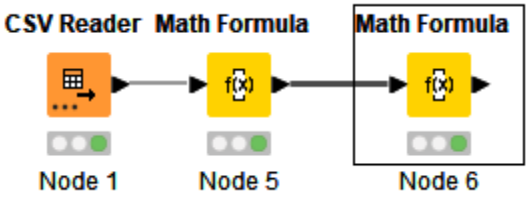


Рисунок 2. Вид рабочей области с связанными узлами

В узлах необходимо указать (рис. 4 и 5).

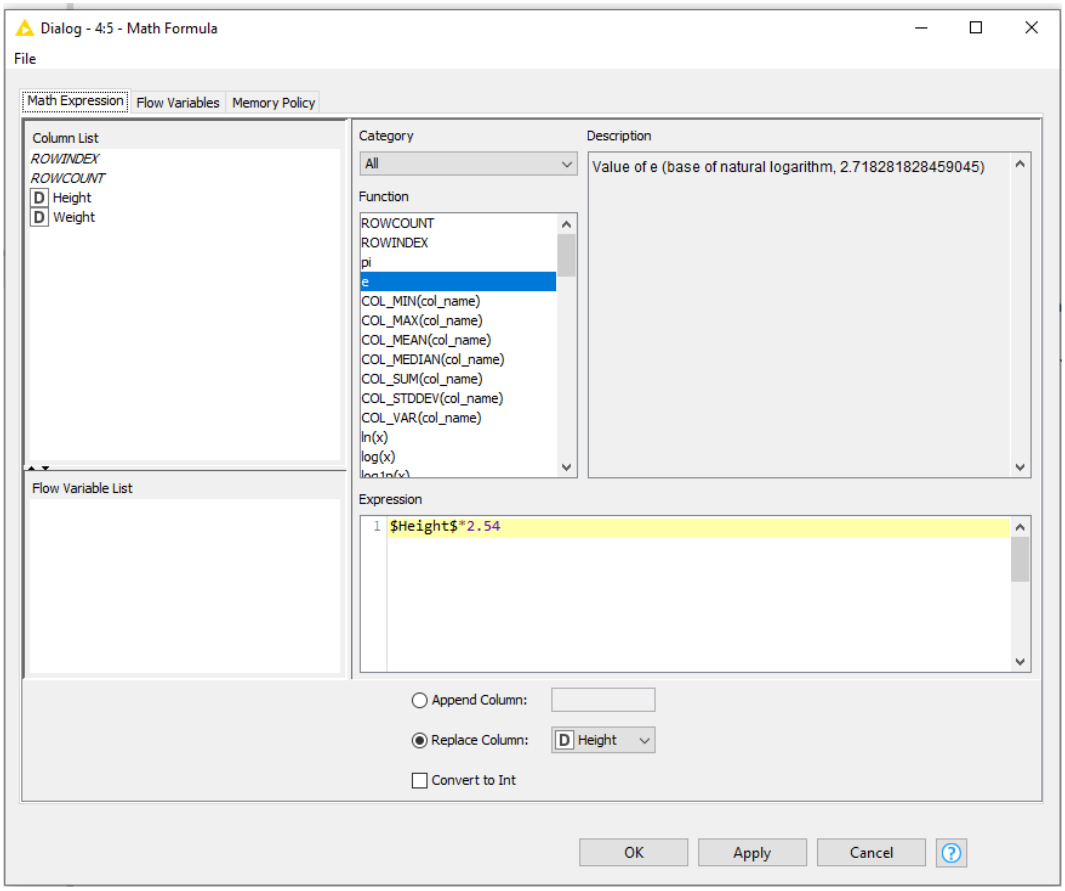


Рисунок 3. Вид первого узла математической формулы

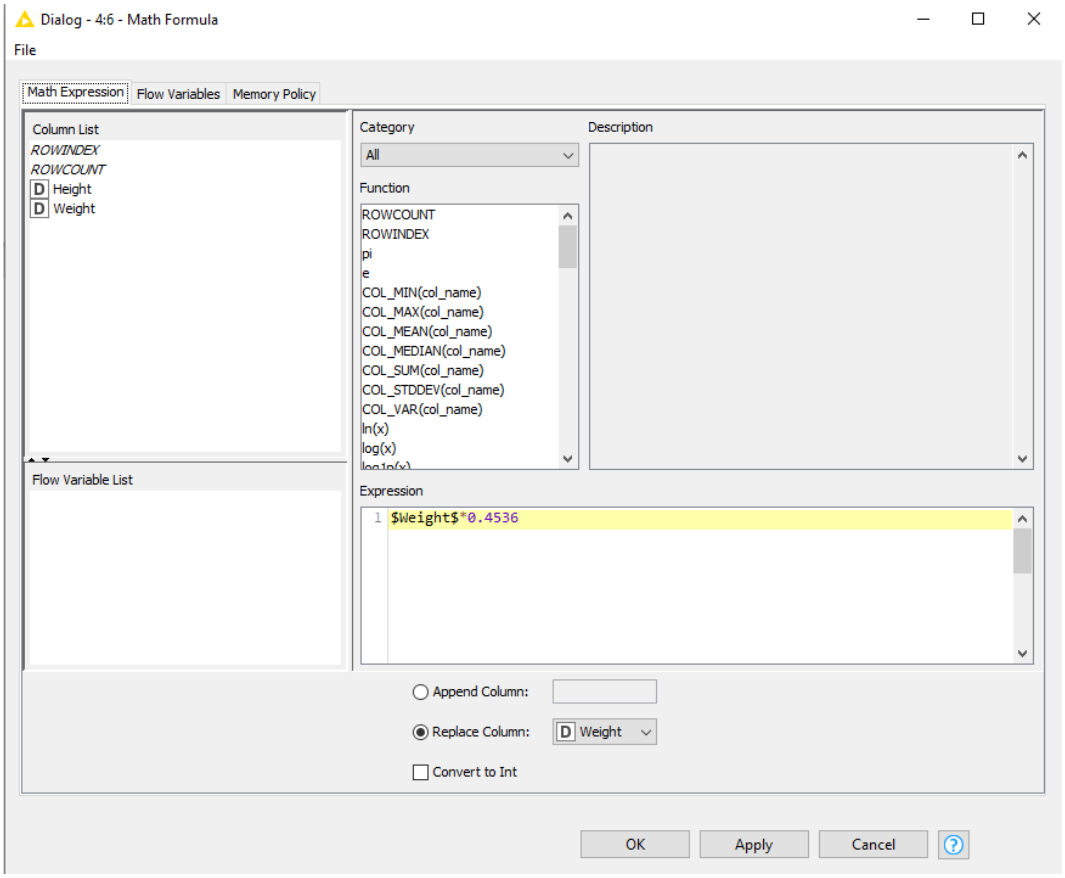


Рисунок 4. Вид второго узла математической формулы

В каждом из узлов по очереди необходимо вручную набрать команду, выбрать столбец для быстрого добавления в поле ввода можно двойным кликом по нему (рис. 5).

| Row ID | S Gender | D Height | D Weight |
|---|---|---|---|
| Row0 | Male | 187.571 | 109.723 |
| Row1 | Male | 174.706 | 73.624 |
| Row2 | Male | 188.24 | 96.499 |
| Row3 | Male | 182.197 | 99.811 |
| Row4 | Male | 177.5 | 93.6 |
| Row5 | Male | 170.823 | 69.043 |
| Row6 | Male | 174.714 | 83.43 |
| Row7 | Male | 173.605 | 76.192 |

Рисунок 5. Новый вид исходной таблицы с уже обновленными данными роста и веса

Пол человека делит данные на 2 категории для дальнейшей работы с данными необходимо закодировать пол с помощью узла Category to Number (из категории в число), где 0 — это мужской пол, а 1 — это женский пол, так как в таблице лишь один столбик с символьными классифицированными данными, то Knime автоматически подберет необходимые параметры и его можно выполнить (рис. 6).

| Row ID | I Gender | D Height | D Weight |
|---|---|---|---|
| Row0 | 0 | 187.571 | 109.723 |
| Row1 | 0 | 174.706 | 73.624 |
| Row2 | 0 | 188.24 | 96.499 |
| Row3 | 0 | 182.197 | 99.811 |
| Row4 | 0 | 177.5 | 93.6 |
| Row5 | 0 | 170.823 | 69.043 |
| Row6 | 0 | 174.714 | 83.43 |

Рисунок 6. Вид таблицы с закодированными значениями пола

Чтобы посмотреть гистограммы полученных данных необходимо добавить соответствующие узлы и соединить с последним узлом. В каждом узле гистограммы необходимо выбрать в поле Histogram column (Столбец гистограммы) данные, которые нужно отобразить (рис 7 и 8), а во вкладке Bining (Деления), в поле Number of bins нужно указать количество делений, (рис. 9, 10, 11).

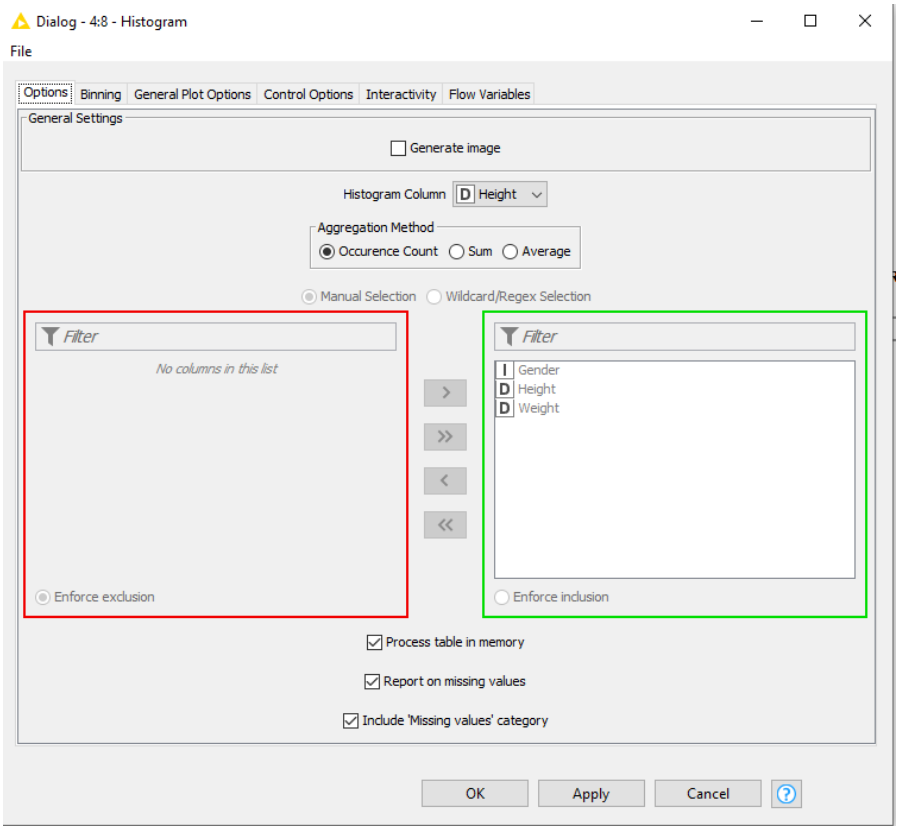


Рисунок 7. Окно настроек гистограммы с указание роста

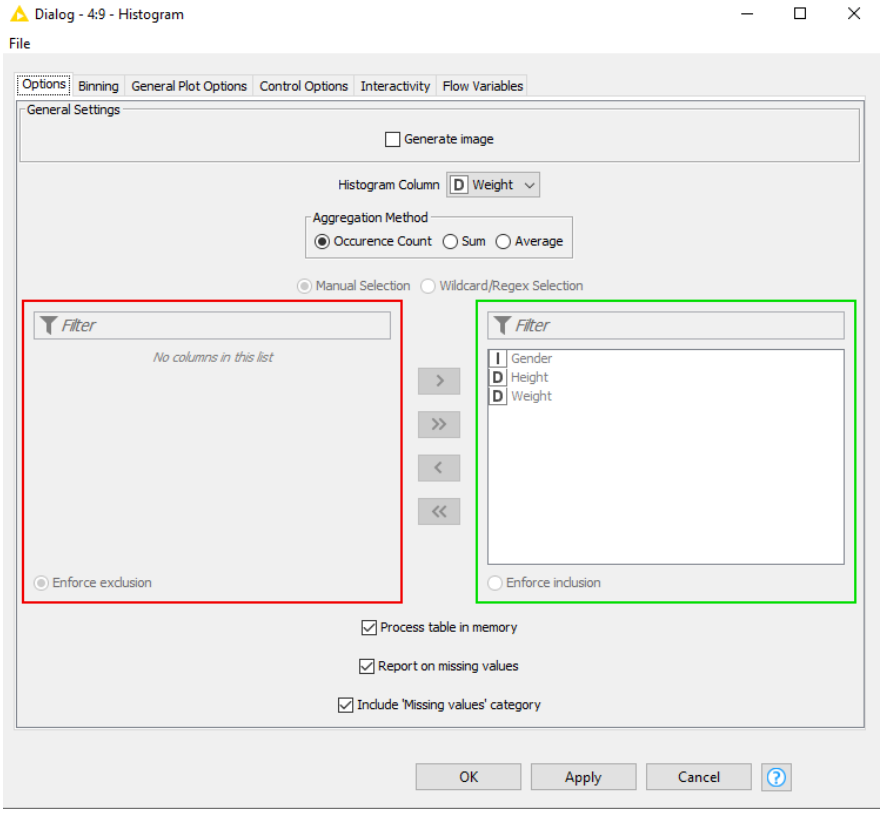


Рисунок 8. Окно настроек гистограммы с указание веса

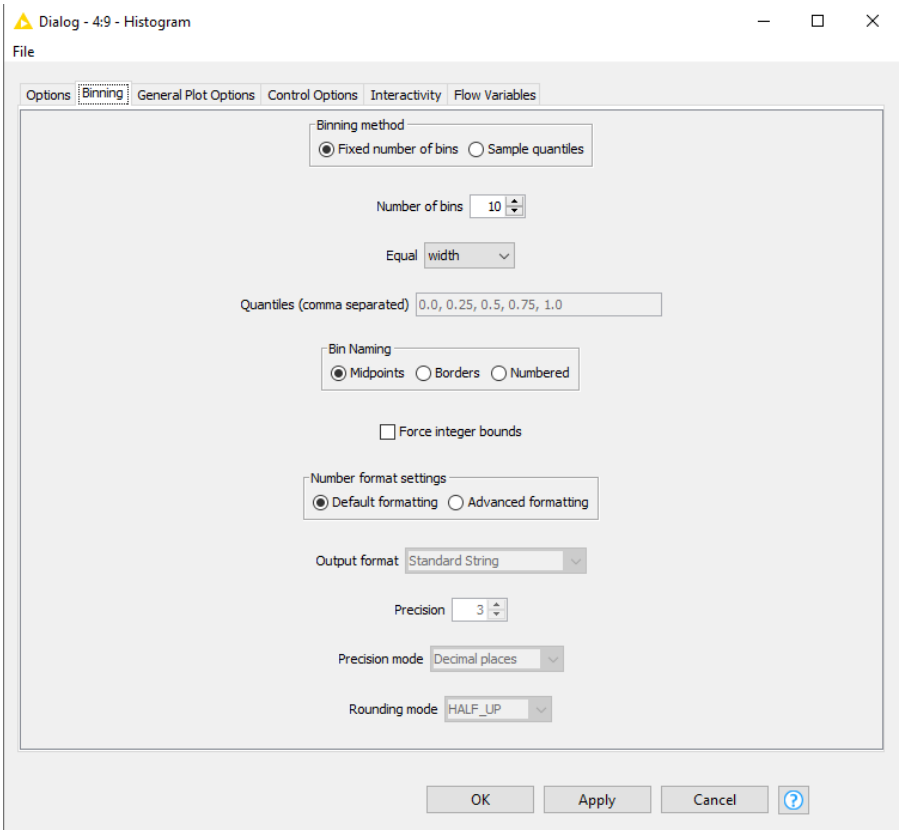


Рисунок 9. Окно настроек гистограммы, параметры вкладки Binning

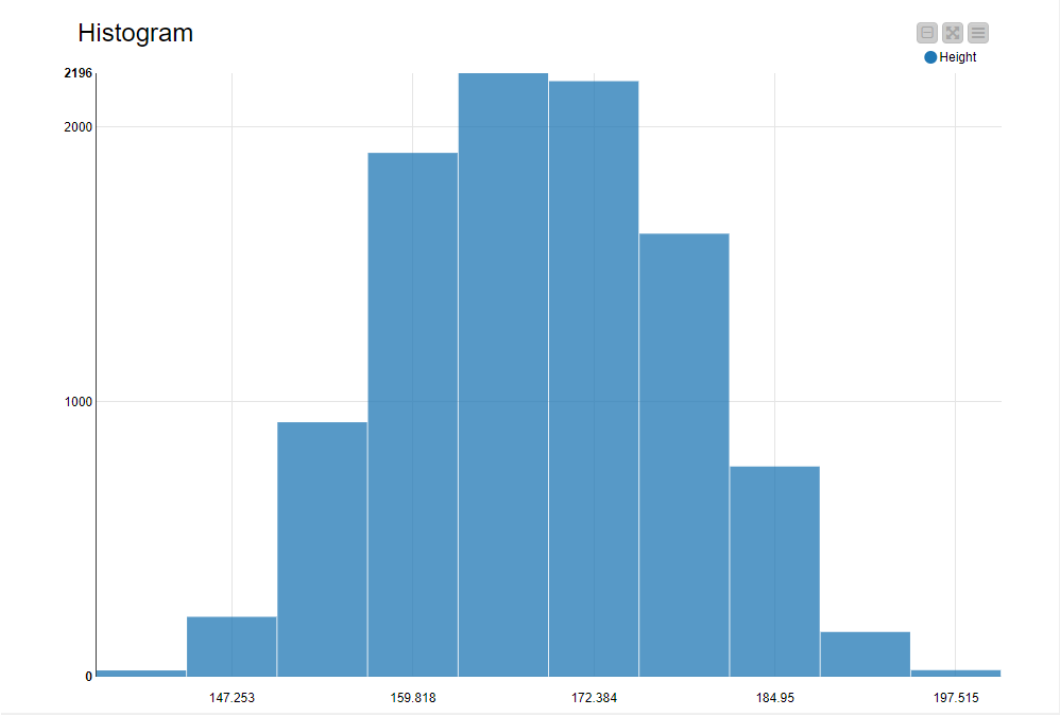


Рисунок 10. Гистограмма роста

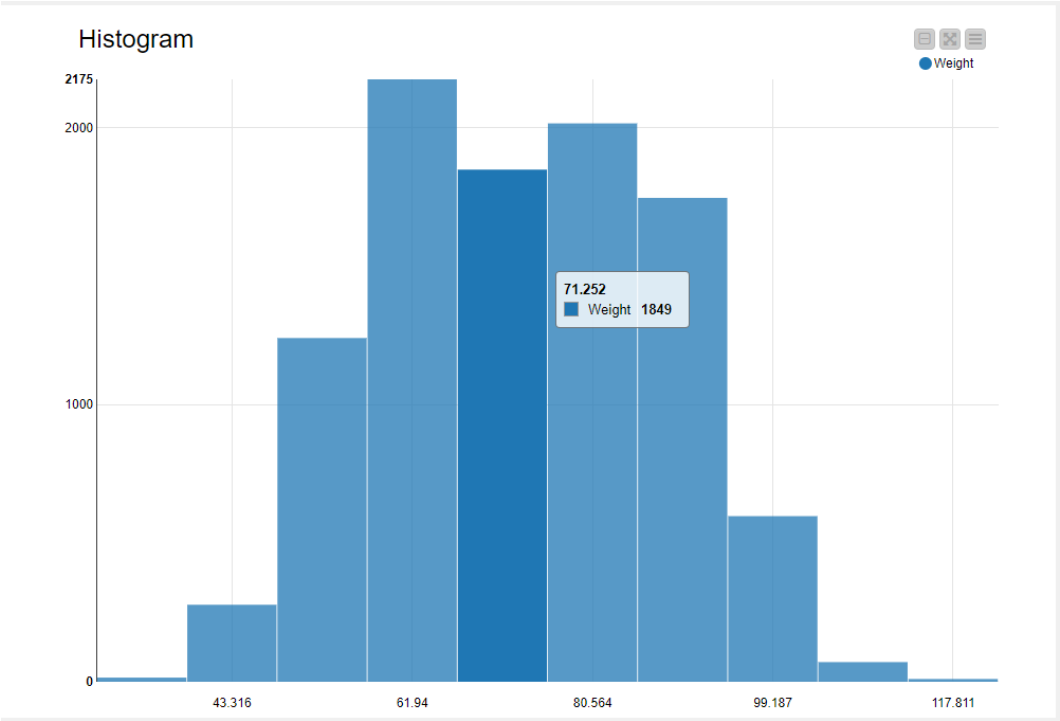


Рисунок 11. Гистограмма веса

Теперь можно попробовать вычислить влияние роста на вес без использования данных о поле человека для этого необходимо на рабочую область добавить узел Linear Regression Learner, соединить с последним интерпретирующим узлом (рис. 12) и настройках узла необходимо указать Target (Цель) – Weight, а также убедиться, что в поле Include (включая), остались поля веса и роста, в противном случае проверить правильность настройки узла Category to Number (рис. 13).

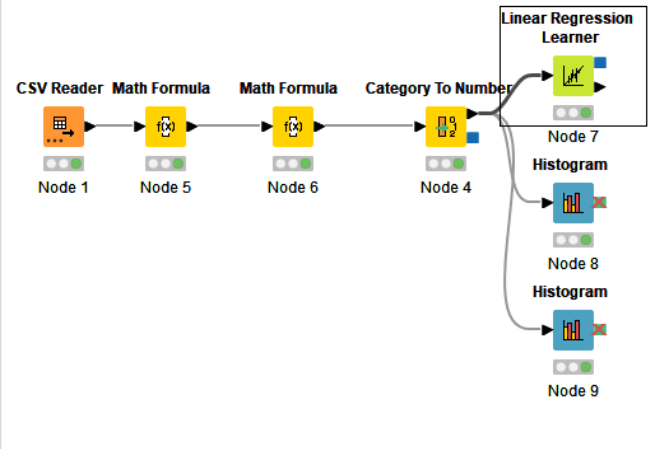


Рисунок 12. Вид рабочей области с новыми узлами

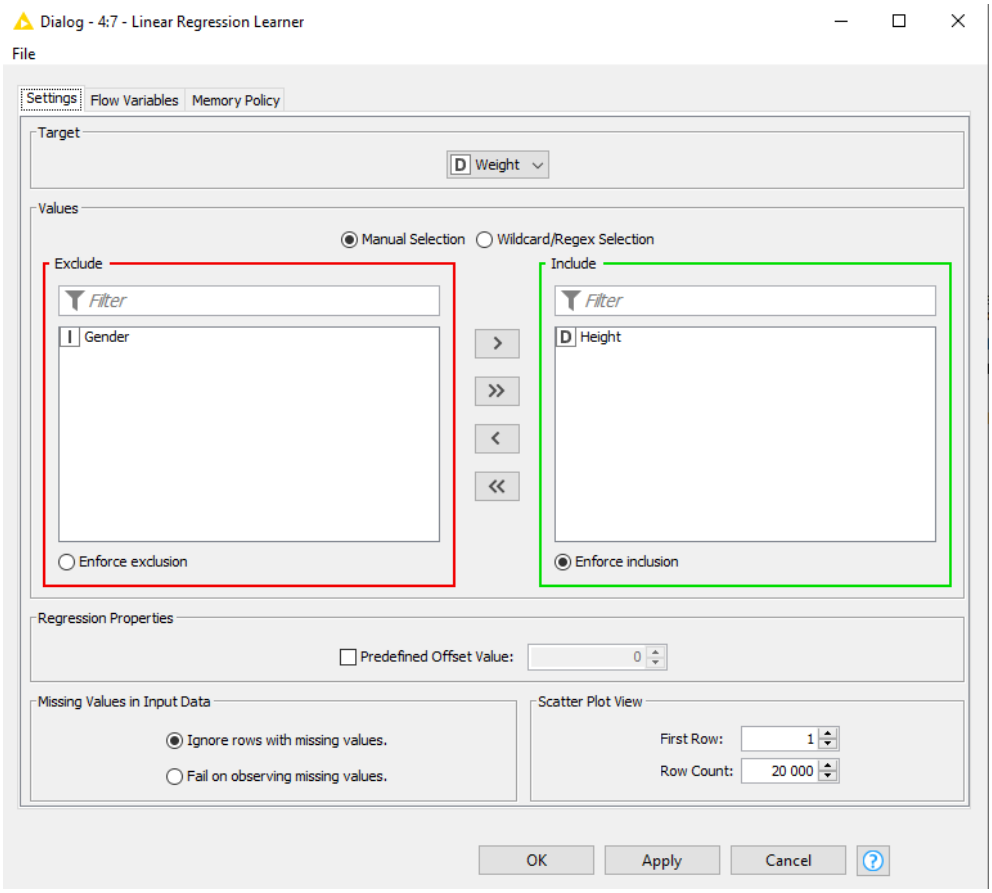


Рисунок 13. Окно настроек узла линейной регрессии без учета пола

После выполнения узла можно посмотреть график линейной регрессии и результаты вычислений (рис. 14 и 15).

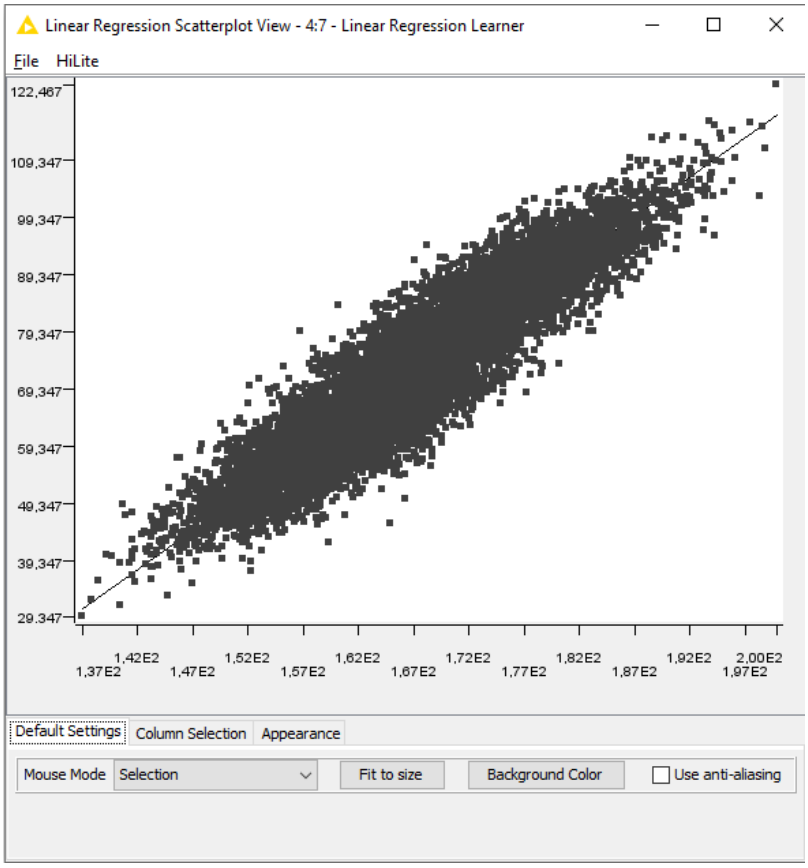


Рисунок 14. График линейной регрессии

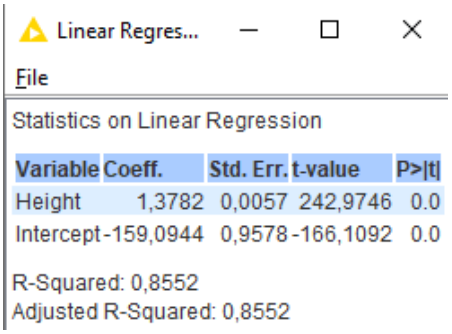


Рисунок 15. Статистика по линейной регрессии

Из результатов следует что уравнение линейной регрессии:

y = 1.38*x - 159.1

где x-рост в см, y-вес в кг

R^2 - Точность модели – 0,8552

Коэффициенты статистически значимы, так как параметр P>|t| меньше 0,05.

Далее с помощью линейной регрессии необходимо высчитать влияние значений роста и пол на вес человека с помощью узла Linear Regression Learner, для этого необходимо вернуть параметр Gender в рабочую область, остальные параметры не меняются (рис. 15).

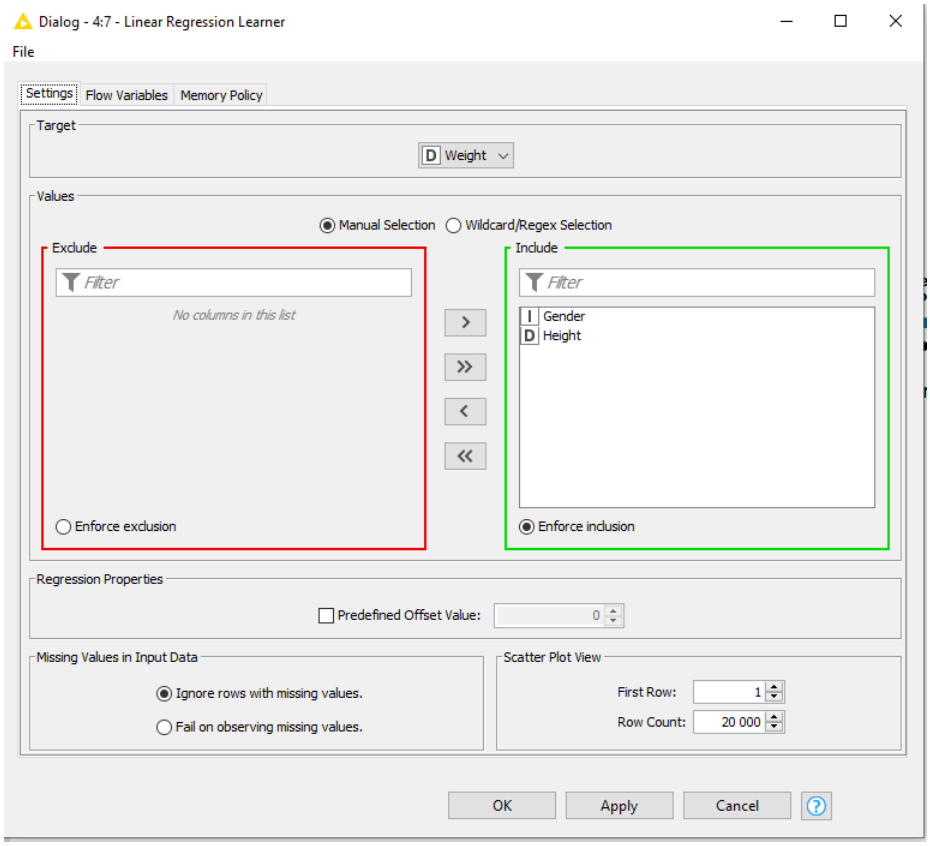


Рисунок 15. Вид окна настроек линейной регрессии включая пол и рост

После выполнения узла можно снова посмотреть график расчета линейной регрессии и результаты (рис. 16 и 17).

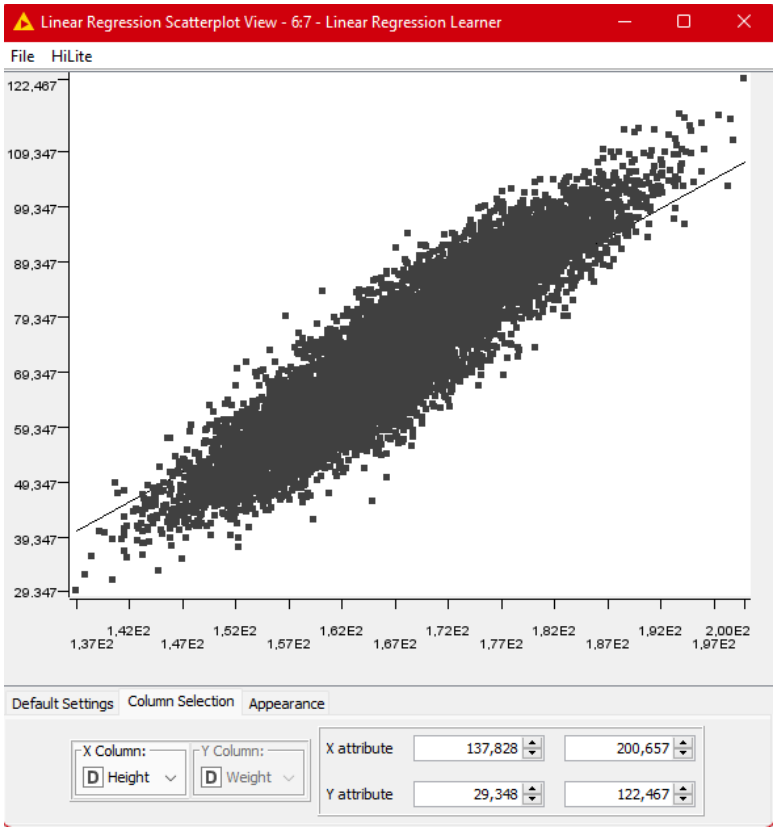


Рисунок 16. График расчета линейной регрессии

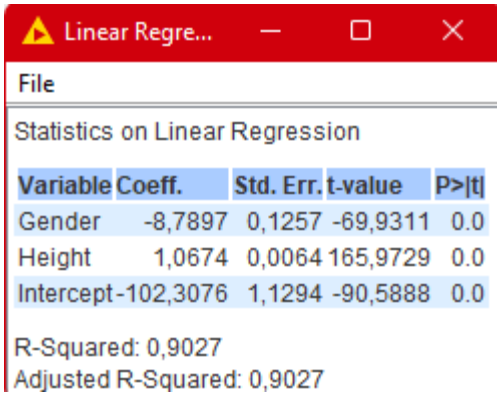


Рисунок 17. Результаты по линейной регрессии

Из результатов следует что уравнение линейной регрессии:
y= - 8.79 *x1 + 1.1 *x2 - 102.3
где x1-пол, x2-рост в см, y-вес в кг
R^2 - Точность модели – 0,9027
Коэффициенты статистически значимы, так как параметр P>|t| меньше 0,05.

Видим по значению R^2, что добавление параметра «Gender/Пол» улучшило модель.

Далее необходимо предсказать вес с помощью узла Regression Predict, после добавления нужного узла, его необходимо соединить с исходными данными через связь вход/выход узла из последнего преобразования таблицы и входа/выхода модели линейной регрессии данный узел не имеет настроек, важно лишь правильное соединение узлов (рис. 18).

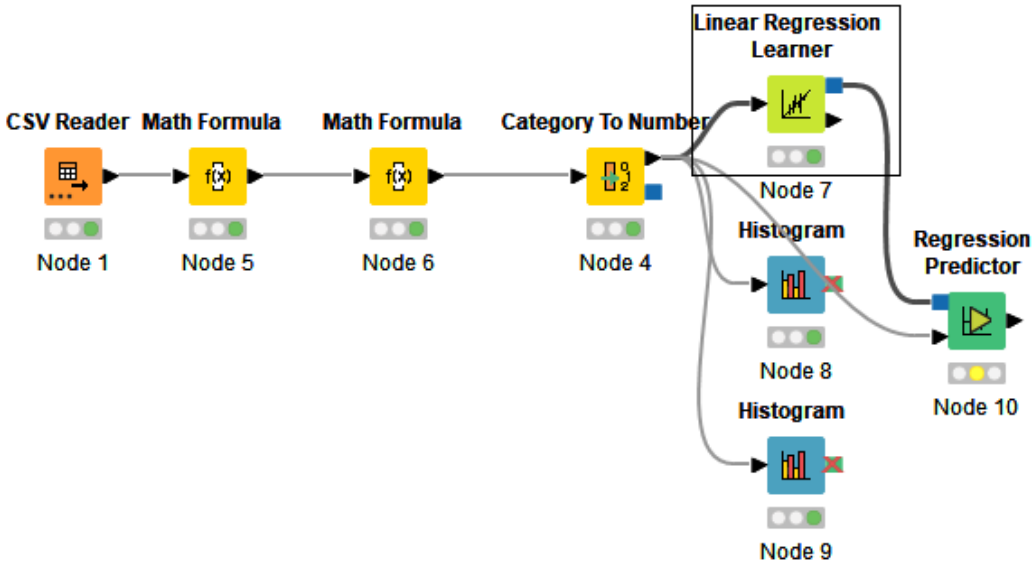


Рисунок 18. Вид рабочей области при соединении нужных узлов

После выполнения последнего узла можно посмотреть полученные значения веса выбрав опцию Predicted data, в таблице появится новый столбец с обозначенными данными (рис. 19).

| Row ID | Gender | Height | Weight | Predicti... |
|---|---|---|---|---|
| Row0 | 0 | 187.571 | 109.723 | 97.902 |
| Row1 | 0 | 174.706 | 73.624 | 84.17 |
| Row2 | 0 | 188.24 | 96.499 | 98.615 |
| Row3 | 0 | 182.197 | 99.811 | 92.165 |
| Row4 | 0 | 177.5 | 93.6 | 87.152 |
| Row5 | 0 | 170.823 | 69.043 | 80.025 |
| Row6 | 0 | 174.714 | 83.43 | 84.178 |
| Row7 | 0 | 173.605 | 76.192 | 82.995 |
| Row8 | 0 | 170.228 | 79.802 | 79.39 |
| Row9 | 0 | 161.179 | 70.943 | 69.732 |
| Row10 | 0 | 180.836 | 84.644 | 90.713 |
| Row11 | 0 | 181.968 | 96.953 | 91.921 |
| Row12 | 0 | 164.506 | 75.809 | 73.283 |
| Row13 | 0 | 175.979 | 85.933 | 85.529 |
| Row14 | 0 | 175.879 | 84.567 | 85.422 |
| Row15 | 0 | 171.82 | 78.104 | 81.089 |
| Row16 | 0 | 183.943 | 88.919 | 94.029 |

Рисунок 19. Конечный вид таблицы с предсказанным весом

**Выводы**

Knime – удобная платформа для анализа данных, получения отчетности и решения многих задач связанных с большим объемом различных данных.

Решение линейной регрессии один из множества задач, которых может решить Knime.

**Библиографический список**


1. Berthold M. R. et al. KNIME-the Konstanz information miner: version 2.0 and beyond //AcM SIGKDD explorations Newsletter. 2009. Т. 11. №. 1. С. 26-31.
2. Fillbrunn A. et al. KNIME for reproducible cross-domain analysis of life science data //Journal of biotechnology. 2017. Т. 261. С. 149-156.
3. Beisken S. et al. KNIME-CDK: Workflow-driven cheminformatics //BMC bioinformatics. 2013. Т. 14. №. 1. С. 1-4.
4. Mazanetz M. P. et al. Drug discovery applications for KNIME: an open-source data mining platform //Current topics in medicinal chemistry. 2012. Т. 12. №. 18. С. 1965-1979.
5. nstokoe/weight-height.csv URL: https://gist.github.com/nstokoe/7d4717e96c21b8ad04ec91f361b000cb (дата обращения 1.09.2022)
6. KNIME Community Forum URL: https://forum.knime.com/ (дата обращения 15.08.2022)