Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

Московский институт электроники и математики им. А.Н. Тихонова

ПРАКТИЧЕСКАЯ РАБОТА № 5
по дисциплине «Информатика»

ТЕМА РАБОТЫ

«Анализ данных методом решающих деревьев»

Москва, 2024

Оглавление

1. Введение

Целью данной лабораторной работы является освоение методов анализа данных и визуализации на основе алгоритма решающих деревьев с использованием инструмента RapidMiner. Решающие деревья являются одним из наиболее интерпретируемых методов машинного обучения, который позволяет выявлять зависимости между признаками и строить модели для прогнозирования значений целевой переменной.

В рамках работы студенты выполнят предобработку данных, обучат модель на основе Decision Tree, проведут анализ важности признаков и оценят качество модели.

Кроме того, студенты изучат методы оценки качества модели с помощью метрики Root Mean Squared Error (RMSE) и проанализируют полученные предсказанные значения с помощью графика Prediction (G3).

2. Содержание практической работы

Описание работы:

В основе работы лежит анализ данных о студентах с применением метода решающих деревьев. Данные включают множество атрибутов, таких как возраст, уровень образования родителей, количество пропущенных занятий, потребление алкоголя, учебное время, семейные отношения и другие. Основная цель работы – построение модели, способной предсказать итоговую оценку на основе имеющихся факторов.

Этапы выполнения работы:

1. Провести предобработку данных, включая удаление ненужных столбцов и выбор значимых атрибутов для анализа.
2. Разделить данные на обучающую и тестовую выборки с использованием метода "Split Data".
3. Построить модель решающего дерева, настроив критерий разбиения, глубину и параметры обрезки дерева.
4. Применить обученную модель для предсказания итоговых оценок студентов.
5. Оценить качество модели, используя метрики регрессии, включая RMSE.
6. Провести анализ важности признаков, влияющих на итоговую оценку, и визуализировать их значимость.
7. Построить графики и диаграммы, отражающие взаимосвязи между факторами и итоговыми результатами.

О наборе данных:

Анализ проводится на наборе данных, содержащем следующие характеристики:

- **school** — название школы ученика
- **sex** — пол ученика
- **age** — возраст ученика
- **address** — тип места проживания
- **famsize** — размер семьи

- **Pstatus** — статус совместного проживания родителей
- **Medu** — уровень образования матери
- **Fedu** — уровень образования отца
- **Mjob** — профессия матери
- **Fjob** — профессия отца
- **reason** — причина выбора школы
- **guardian** — опекун ученика
- **traveltime** — время в пути до школы
- **studytime** — время на учебу в неделю
- **failures** — количество неудовлетворительных оценок в предыдущих классах
- **schoolsup** — дополнительная образовательная поддержка
- **famsup** — семейная образовательная поддержка
- **paid** — дополнительные платные курсы по предмету
- **activities** — участие в внеклассных мероприятиях
- **nursery** — посещение детского сада
- **higher** — желание получить высшее образование
- **internet** — наличие интернета дома
- **romantic** — наличие романтических отношений
- **famrel** — качество семейных отношений
- **freetime** — свободное время после школы
- **goout** — частота прогулок с друзьями
- **Dalc** — потребление алкоголя в будни
- **Walc** — потребление алкоголя в выходные
- **health** — текущее состояние здоровья
- **absences** — количество пропущенных занятий
- **G1** — оценка за первый учебный период
- **G2** — оценка за второй учебный период
- **G3** — итоговая оценка (0–20, целевая переменная).

Ключевые особенности данных:

Количество записей: содержит данные о 395 студентах.

Формат данных: табличный набор, включающий числовые и категориальные переменные.

Потенциальные проблемы:

Различные форматы признаков (числовые и категориальные данные). Также возможны выбросы в данных (например, высокий уровень прогулов). Ещё одной проблемой могут стать переменные, которые мы не учитываем в построении модели, а именно социальные факторы и их влияние на итоговую оценку.

Специфика работы:

Для выполнения лабораторной работы используются только следующие столбцы:

- **age** (возраст студента),
- **Medu** (образование матери),
- **Fedu** (образование отца),
- **studytime** (учебное время),
- **failures** (количество неудач),
- **Dalc** (потребление алкоголя в будни),
- **Walc** (потребление алкоголя в выходные),
- **absences** (количество пропусков),
- **G3** (итоговая оценка, целевой признак).

Этот набор данных позволяет проанализировать, какие факторы наиболее сильно влияют на итоговую успеваемость студентов и построить предсказательную модель для оценки будущих результатов.

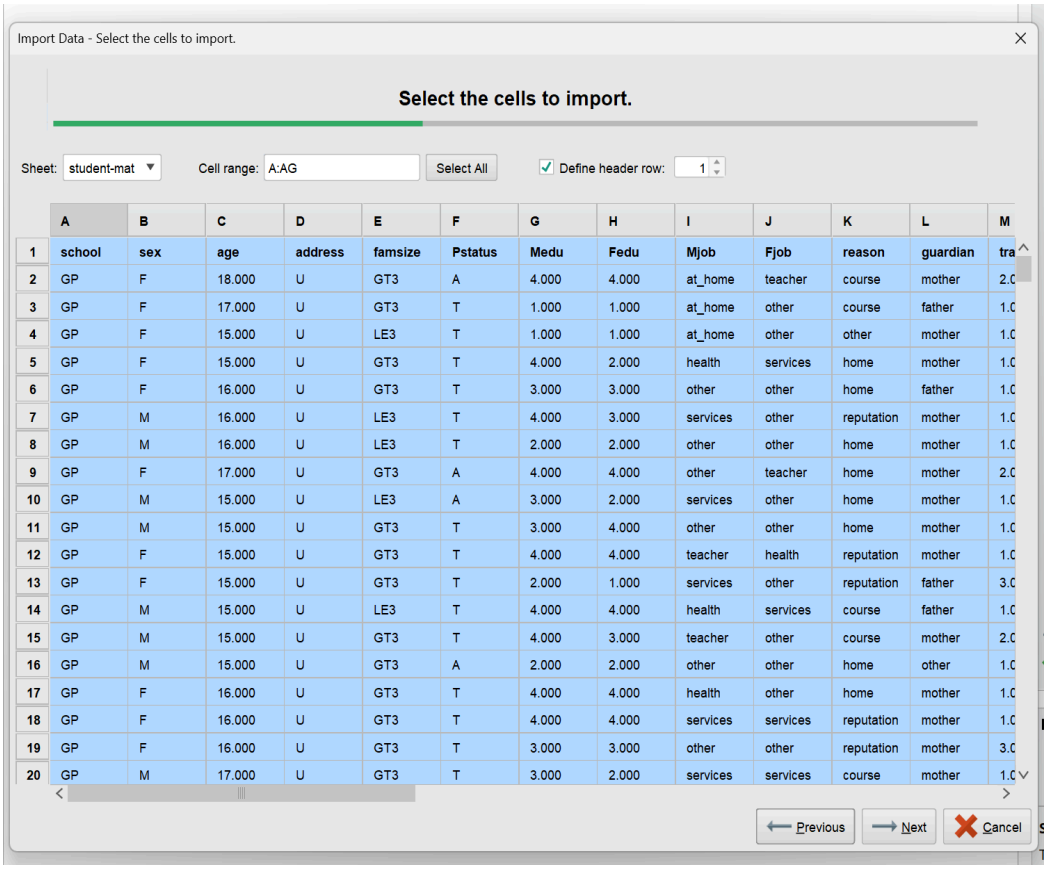
3. Ход работы

Загрузка набора данных

1. Откройте RapidMiner Studio.
2. В главном меню выберите **"Create New Process"**.

3. Воспользуйтесь функцией **"Import data".**
4. Загрузите набор данных о транзакциях, выбрав файл **"student-mat"** в формате xlsx.
5. Сохраните полученную базу данных в папку со своей работой.
6. В результате вы увидите таблицу с данными, содержащими атрибуты: (**school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3**)

Данные успешно загружены, их структура показана на рисунке 3.2.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	tra
2	GP	F	18.000	U	GT3	A	4.000	4.000	at_home	teacher	course	mother	2.0
3	GP	F	17.000	U	GT3	T	1.000	1.000	at_home	other	course	father	1.0
4	GP	F	15.000	U	LE3	T	1.000	1.000	at_home	other	other	mother	1.0
5	GP	F	15.000	U	GT3	T	4.000	2.000	health	services	home	mother	1.0
6	GP	F	16.000	U	GT3	T	3.000	3.000	other	other	home	father	1.0
7	GP	M	16.000	U	LE3	T	4.000	3.000	services	other	reputation	mother	1.0
8	GP	M	16.000	U	LE3	T	2.000	2.000	other	other	home	mother	1.0
9	GP	F	17.000	U	GT3	A	4.000	4.000	other	teacher	home	mother	2.0
10	GP	M	15.000	U	LE3	A	3.000	2.000	services	other	home	mother	1.0
11	GP	M	15.000	U	GT3	T	3.000	4.000	other	other	home	mother	1.0
12	GP	F	15.000	U	GT3	T	4.000	4.000	teacher	health	reputation	mother	1.0
13	GP	F	15.000	U	GT3	T	2.000	1.000	services	other	reputation	father	3.0
14	GP	M	15.000	U	LE3	T	4.000	4.000	health	services	course	father	1.0
15	GP	M	15.000	U	GT3	T	4.000	3.000	teacher	other	course	mother	2.0
16	GP	M	15.000	U	GT3	A	2.000	2.000	other	other	home	other	1.0
17	GP	F	16.000	U	GT3	T	4.000	4.000	health	other	home	mother	1.0
18	GP	F	16.000	U	GT3	T	4.000	4.000	services	services	reputation	mother	1.0
19	GP	F	16.000	U	GT3	T	3.000	3.000	other	other	reputation	mother	3.0
20	GP	M	17.000	U	GT3	T	3.000	2.000	services	services	course	mother	1.0

Рисунок 3.1 – подготовка данных к выгрузке

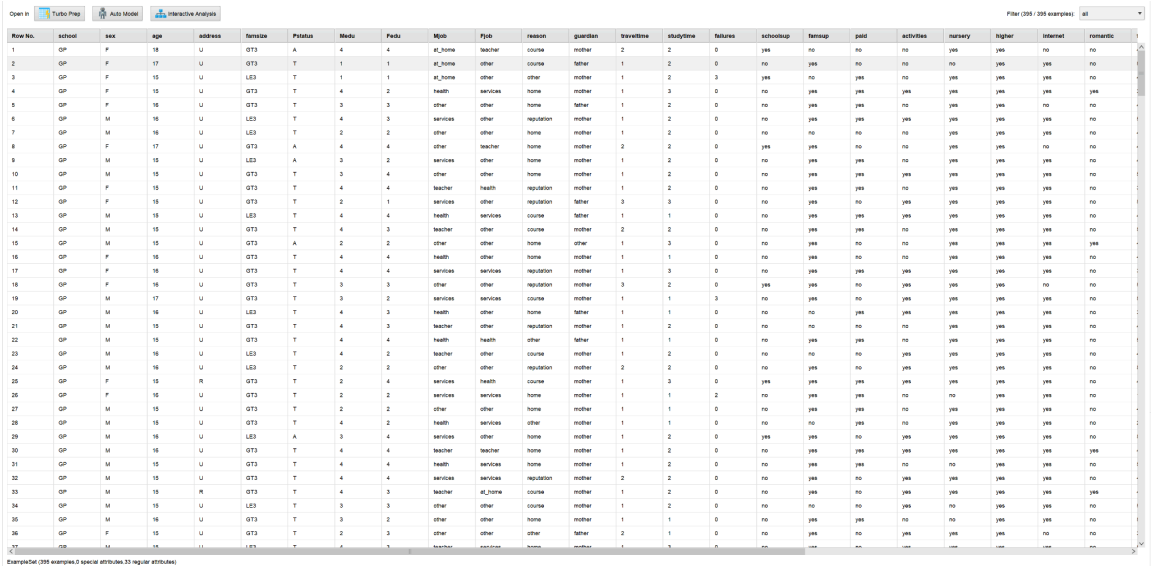


Рисунок 3.2 – выгруженные данные

Фильтрация данных:

Для начала необходимо удалить из исходного набора данных те атрибуты, которые не используются в анализе. Для этого в проект добавляется оператор **"Select Attributes".** Затем **"attribute add filter type"** ставим как “a subset” и из списка всех имеющихся атрибутов выбираем только необходимые как показано на рисунке 3.3

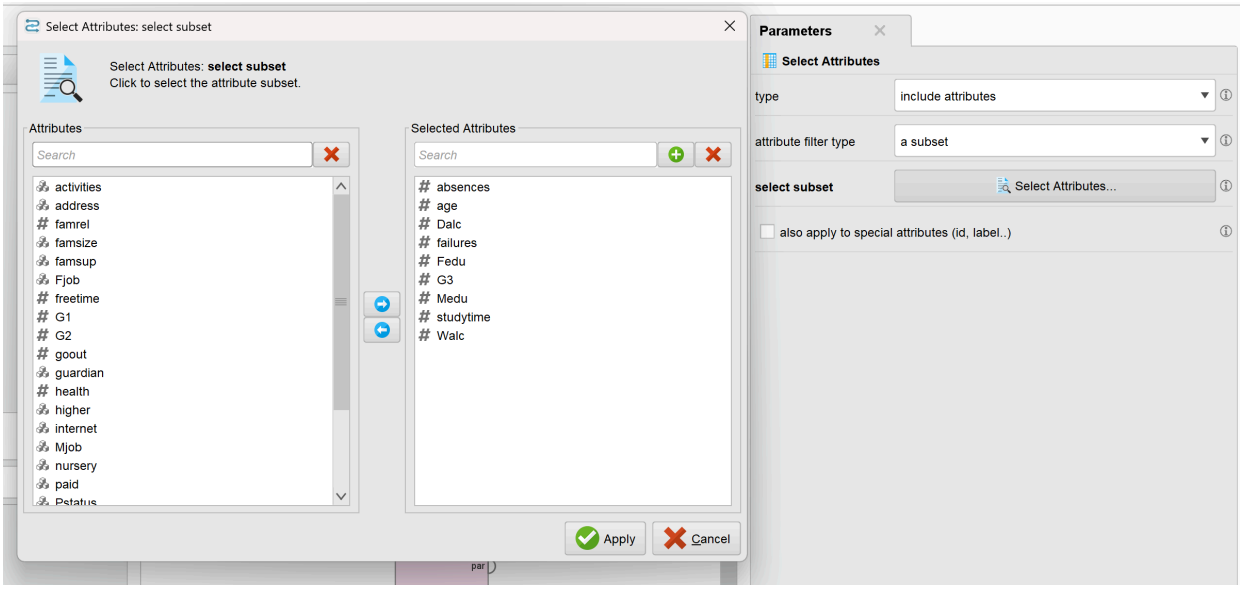


Рисунок 3.3 – настройки для Select Attributes

Разделение данных:

Для корректного обучения и тестирования модели необходимо разделить исходные данные на обучающую и тестовую выборки. Для этого в проект был

добавлен оператор **"Split Data"**, который позволяет случайным образом распределить данные в указанных пропорциях.

В параметрах оператора были заданы следующие значения: 80% данных используется для обучения модели и 20% данных выделяется для тестирования соответственно.

Методом разделения был указан **shuffled sampling**, который обеспечивает случайное перемешивание записей перед разбиением, что важно для создания репрезентативных обучающей и тестовой выборок. Настройки блока представлены на рисунке 3.4.

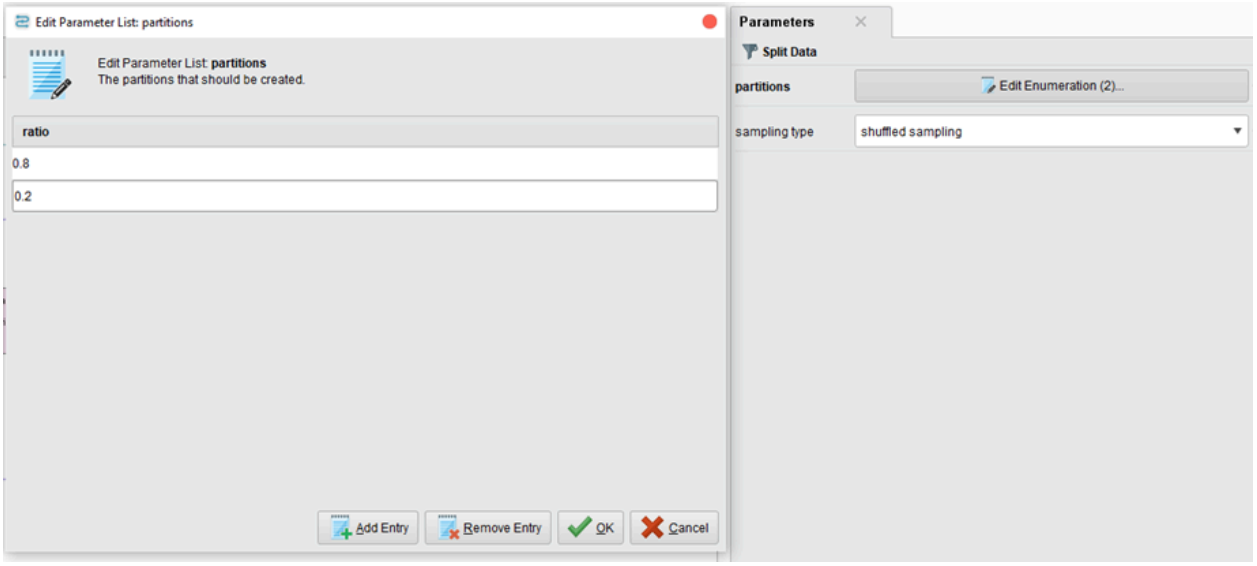


Рисунок 3.4 – настройки для Split Data

Установка целевого атрибута:

Для корректной работы модели машинного обучения необходимо определить целевой атрибут, который будет использоваться в качестве метки (label). В данной лабораторной работе целевым атрибутом является **"G3"** – итоговая оценка студента.

Для этого в проект необходимо добавить новый оператор **"Set Role"**, который позволяет задать специальную роль для атрибутов. Назначение метки позволяет алгоритму машинного обучения понимать, какую переменную необходимо предсказывать на основе остальных данных.

Целевую переменную можно назначить в параметрах оператора, как это показано на рисунке 3.5.

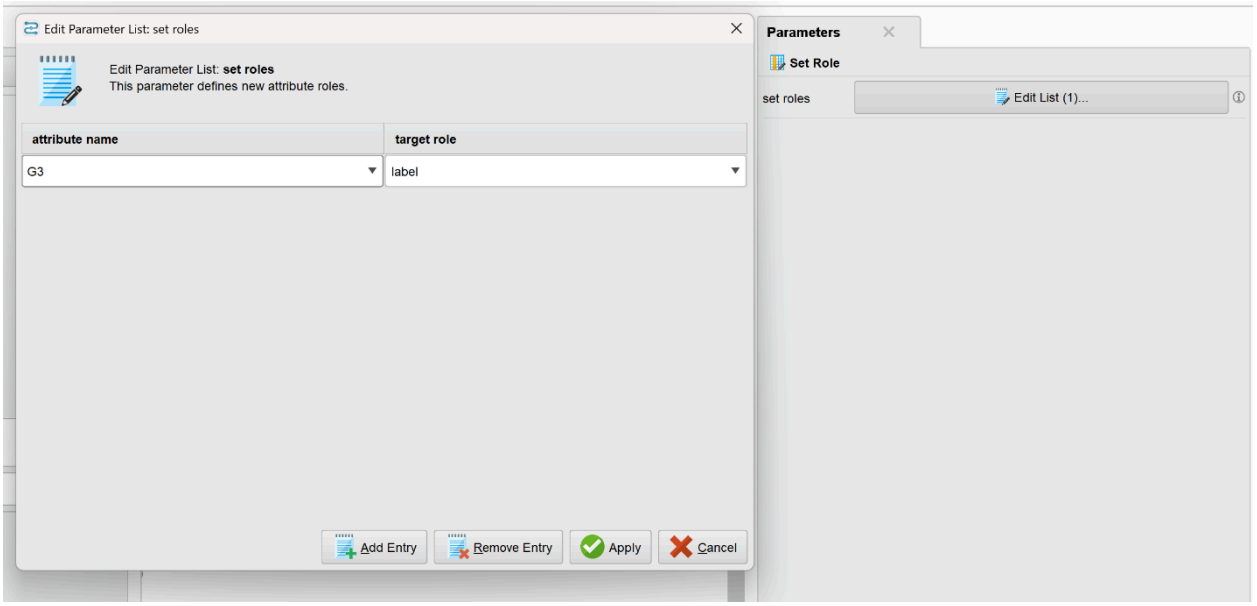


Рисунок 3.5 – настройки для Set roles

Построение модели решающего дерева:

После подготовки данных и задания целевого атрибута необходимо создать модель машинного обучения, которая будет предсказывать итоговую оценку G3. Для этого добавим в проект новый оператор **"Decision Tree"**, который строит дерево решений на основе имеющихся данных.

Решающие деревья позволяют находить зависимости между входными признаками и целевым атрибутом, создавая иерархическую структуру решений. В данной лабораторной работе был выбран алгоритм регрессии с критерием **"least square"**, поскольку наша целевая переменная является числовым значением.

Настройка параметров Decision Tree:

- **Criterion (Критерий разбиения)**: *Least Square* – минимизация суммы квадратов ошибок, используется для построения регрессионного дерева.
- **Maximal Depth (Максимальная глубина дерева)**: *10* – ограничение максимального количества уровней в дереве, предотвращает чрезмерную сложность модели.
- **Apply Prepruning (Применить предварительную обрезку)**: *Включено* – предотвращает избыточное разветвление дерева и снижает вероятность переобучения.

- **Minimal Gain (Минимальный прирост информации)**: *0.01* – минимальное улучшение критерия разбиения, необходимое для создания новой ветви.
- **Minimal Leaf Size (Минимальный размер листа)**: *2* – минимальное количество примеров, необходимых для формирования конечного узла дерева.

Выбранные параметры позволяют построить сбалансированную модель, которая учитывает важные закономерности в данных, но при этом не переусложняет структуру дерева. Финальные настройки для блока представлены на рисунке 3.6.

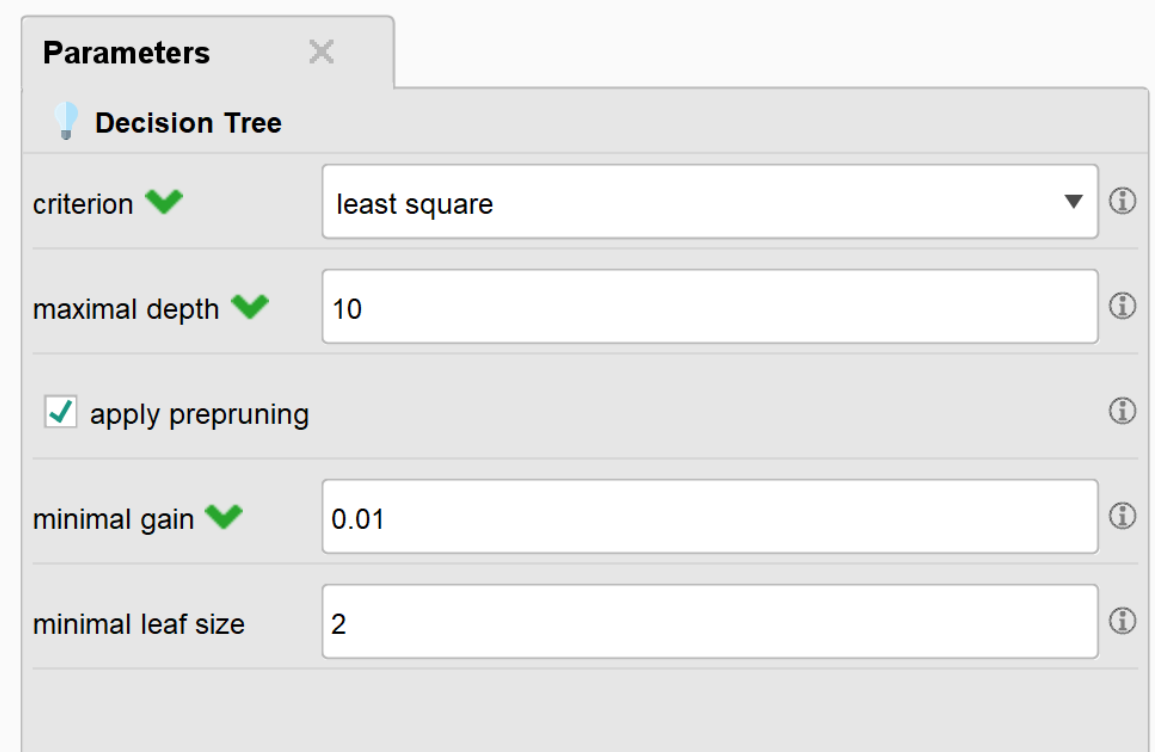


Рисунок 3.6 – параметры для Decision tree

Финальные приготовления к построению дерева:

Для корректного применения обученной модели решающего дерева и последующего анализа её качества необходимо добавить ещё три ключевых оператора: **"Apply Model"**, **"Set Role"** и **"Performance"**.

Оператор **"Apply Model"** необходим для тестирования построенной модели на ранее выделенной тестовой выборке. Этот блок принимает два входных потока данных: обученную модель, созданную с помощью **"Decision Tree"**, и тестовый

набор данных (выход из **"Split Data"**), на который ещё не было наложено предсказание. В результате работы этого оператора к тестовому набору добавляется новый столбец, содержащий предсказанные моделью значения целевой переменной.

После применения модели снова необходимо корректно задать атрибуту G3 роль label для дальнейшего анализа. Это выполняется с помощью оператора **"Set Role (3)"**. Добавление второго оператора необходимо, поскольку аналитические блоки в RapidMiner требуют явного указания целевой переменной, иначе они не смогут правильно интерпретировать данные.

Для оценки точности модели был использован оператор **"Performance"**, предназначенный для анализа регрессионных моделей. В параметрах блока были выбраны ключевые метрики оценки:

Root Mean Squared Error (RMSE) — среднеквадратичная ошибка, измеряющая среднее отклонение предсказанных значений от реальных. Это одна из наиболее распространенных метрик в задачах регрессии.

Squared Error — сумма квадратов отклонений предсказанных значений от реальных. Позволяет оценить общее расхождение модели.

Correlation — коэффициент корреляции, который показывает степень линейной зависимости между предсказанными и фактическими значениями целевого атрибута.

Squared Correlation — квадрат коэффициента корреляции, дающий представление о доле объясненной дисперсии в данных.

Absolute Error — это разница между предсказанным и фактическим значением. Она измеряет точность модели, показывая, на сколько в среднем предсказанные значения отклоняются от реальных.

Relative Error — это отношение абсолютной ошибки к фактическому значению, выраженное в процентах. Эта метрика помогает оценить ошибку относительно масштаба данных.

Все три оператора были последовательно соединены таким образом, чтобы результаты предсказания модели могли быть корректно интерпретированы и

проанализированы. Настройки оператора **"Performance"** и корректное подключение представлены на рисунке 3.7.

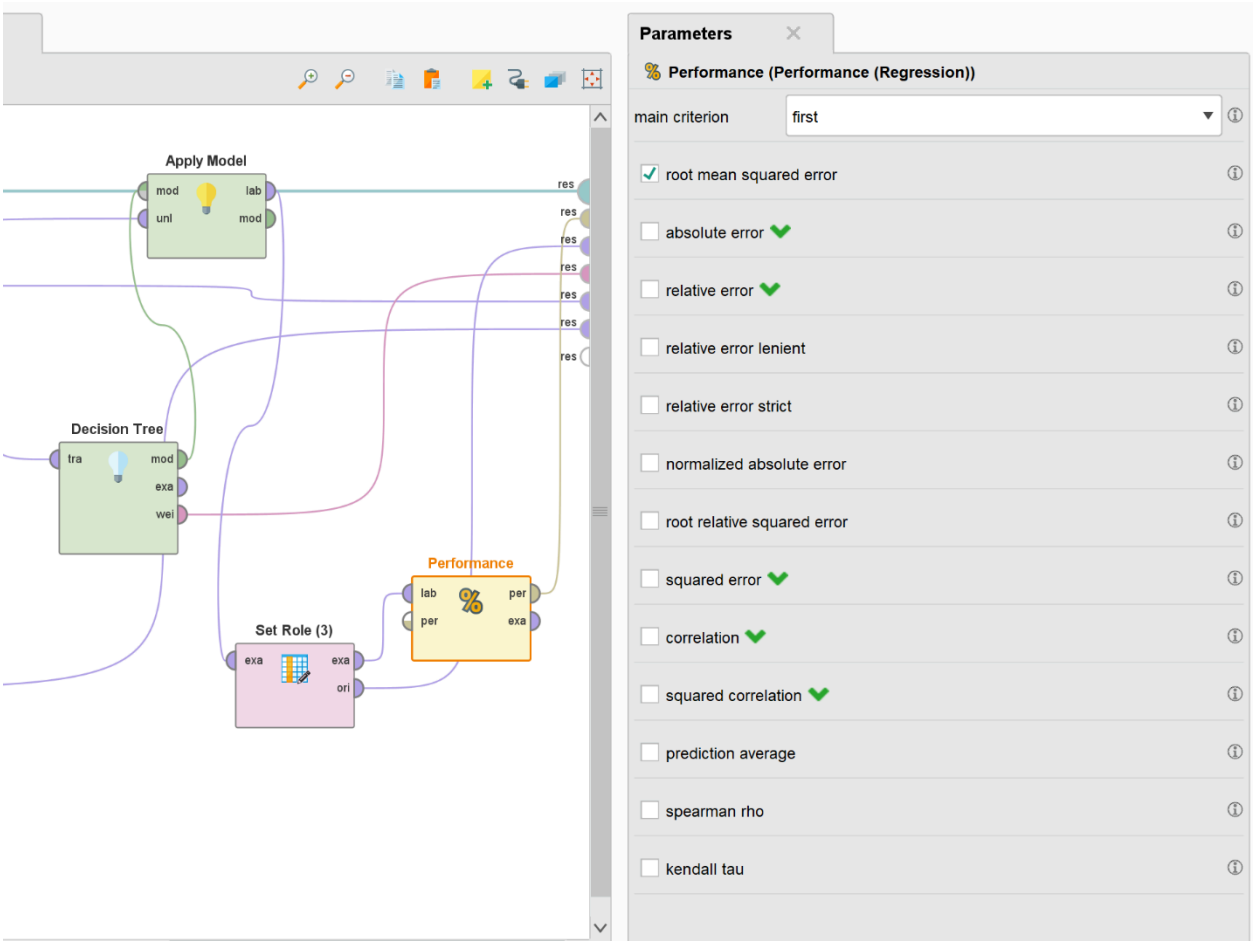


Рисунок 3.7 – Настройка Performance

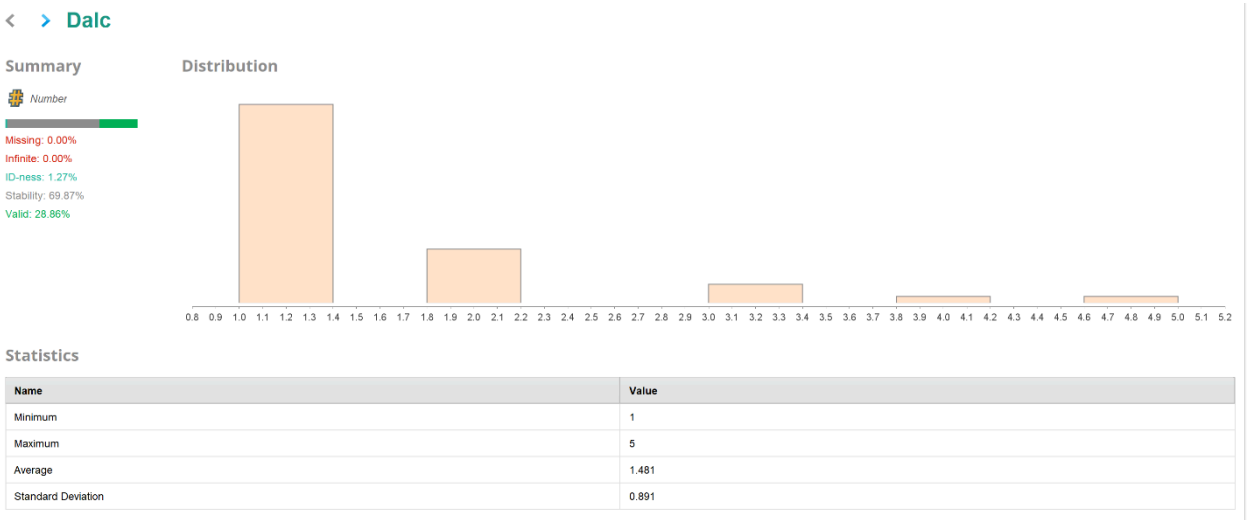
Оператор Statistics:

Перед тем как запустить проект, был добавлен блок **"Statistics"**, который позволяет провести предварительный анализ данных. Этот оператор автоматически вычисляет ключевые статистические показатели для каждого атрибута в наборе данных, включая минимальное, максимальное и среднее значение, а также стандартное отклонение

На рисунке 3.8 показан пример статистического анализа переменной Dalc (потребление алкоголя в будние дни). В данном случае видно, что переменная принимает значения от 1 до 5, а среднее значение составляет 1.481. Это говорит о том, что большая часть студентов редко употребляет алкоголь в будние дни. Блок

предоставляет аналогичный анализ для всех остальных переменных, что может помочь лучше понять исходные данные перед обучением модели.

Финальный вид схемы с новым блоком представлен на рисунке 3.9



Name	Value
Minimum	1
Maximum	5
Average	1.481
Standard Deviation	0.891

Рисунок 3.8 – Статистика для параметра Dalc

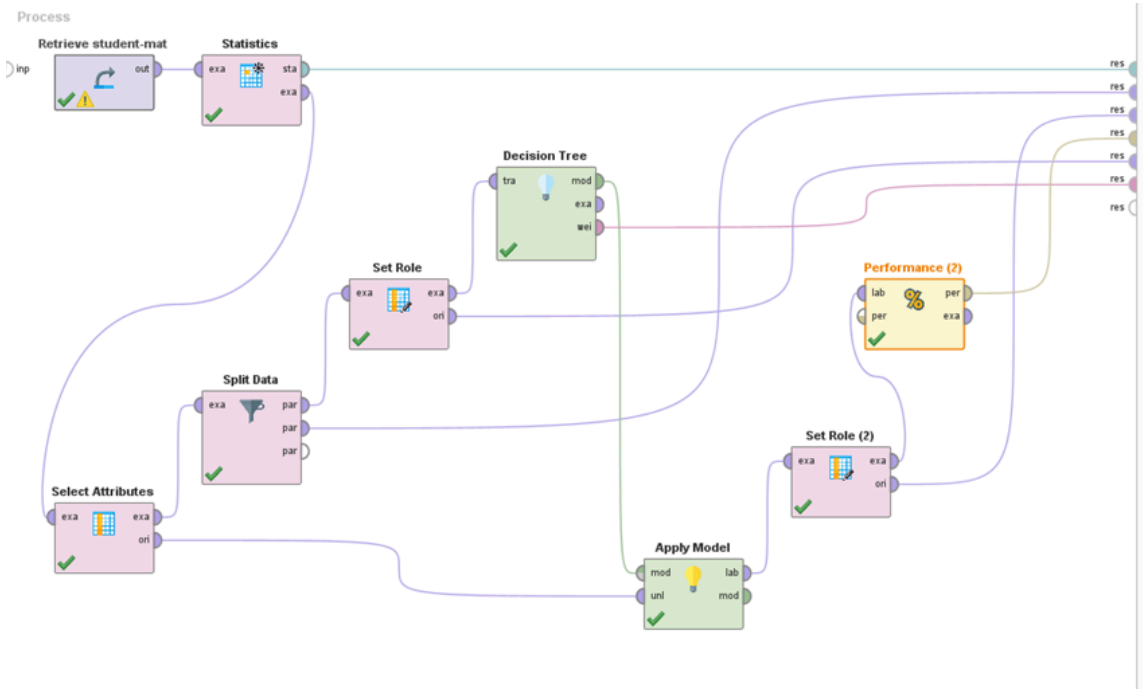


Рисунок 3.9 – Финальный вид схемы

После запуска процесса мы получили несколько результатов, представленных на рисунке 3.10. В частности, был создан **"ExampleSet (Split Data) "**, который разбит на обучающую (80%) и тестовую (20%) выборки, что соответствует заданным параметрам разделения данных.

Далее мы рассмотрим результаты работы **"Decision Tree"** и **"Performance"**, чтобы детально проанализировать качество модели, предсказанные значения и важность атрибутов.

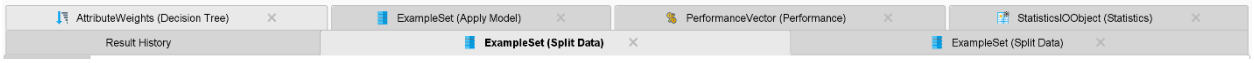


Рисунок 3.10 – Полученные результаты

После запуска модели решающего дерева можно провести анализ значимости атрибутов, влияющих на итоговую оценку студентов. Для этого автоматически была сформирована таблица значимости признаков, которая представлена на рисунке 3.11, где каждому атрибуту соответствует его вклад в предсказание целевого значения.

Для удобства анализа можно воспользоваться вкладкой "Plot view", которая находится в левом меню интерфейса. Здесь можно выбрать тип графика (Chart style), указать оси и сгруппировать данные. В данном случае была выбрана столбчатая диаграмма (Bars), где по оси X представлены атрибуты, а по оси Y – их значимость. Результаты представлены на рисунке 3.12.

Интерпретация результатов:

Наибольшее влияние на итоговую оценку оказывает возраст (age) – 0.254. Вторым по значимости является употребление алгоколя в выходные (Walc) – 0.168, однако влияние алкоголя в будние дни (Dalc) – 0.049 существенно ниже. Также важную роль играет количество пропусков занятий (absences) – 0.166, что подтверждает очевидную зависимость между посещаемостью и успеваемостью. Следующими значительными параметрами являются уровни образования родителей (Fedu – 0.152, Medu – 0.136),

Наименьшее значение имеет количество неуспешных попыток сдачи экзаменов (failures) – 0.013, что в рамках рассмотренного датасета говорит о том, что прошлые неудачи не являются ключевым фактором предсказания финальной оценки.

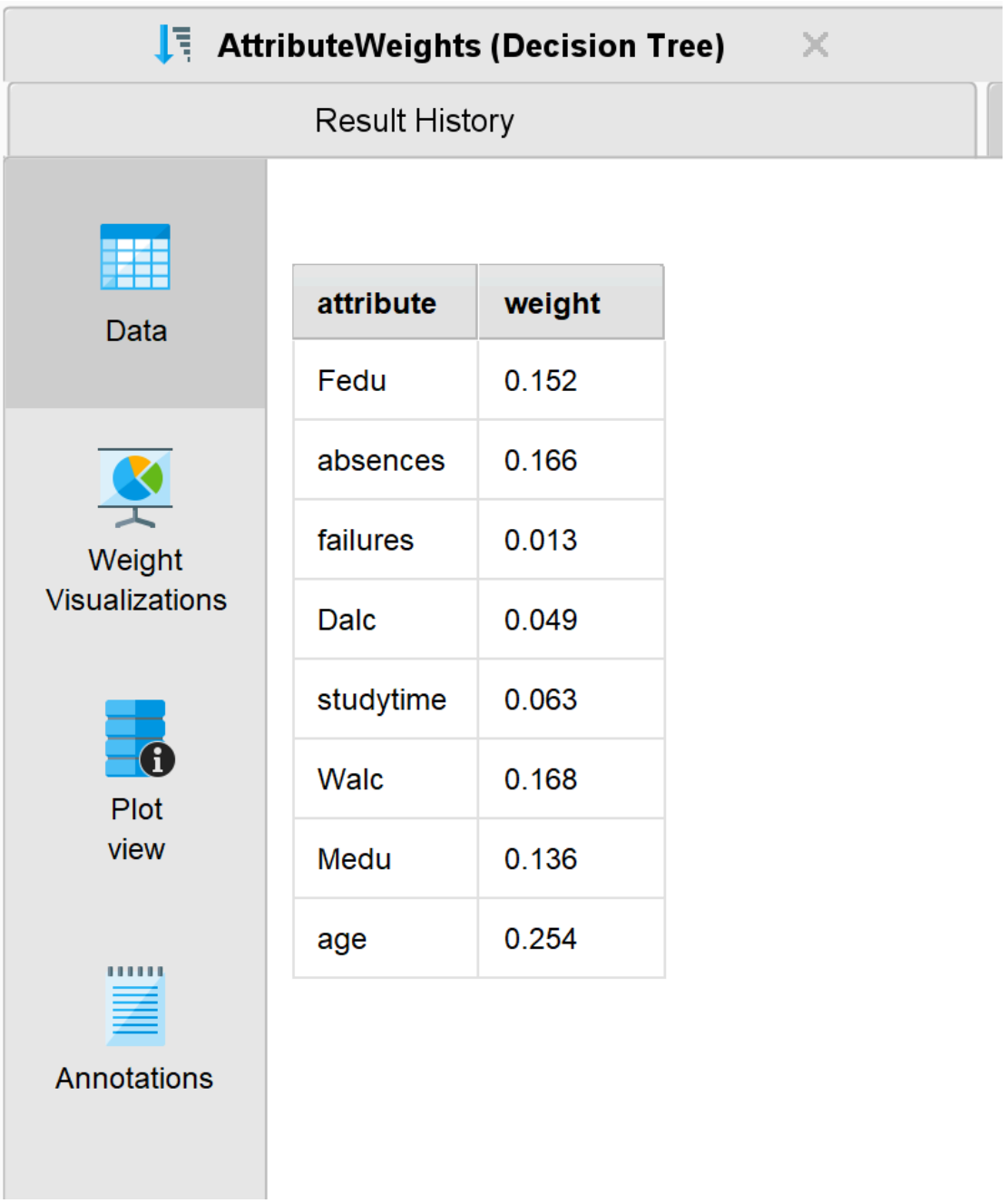


Рисунок 3.11 – таблица значимости признаков

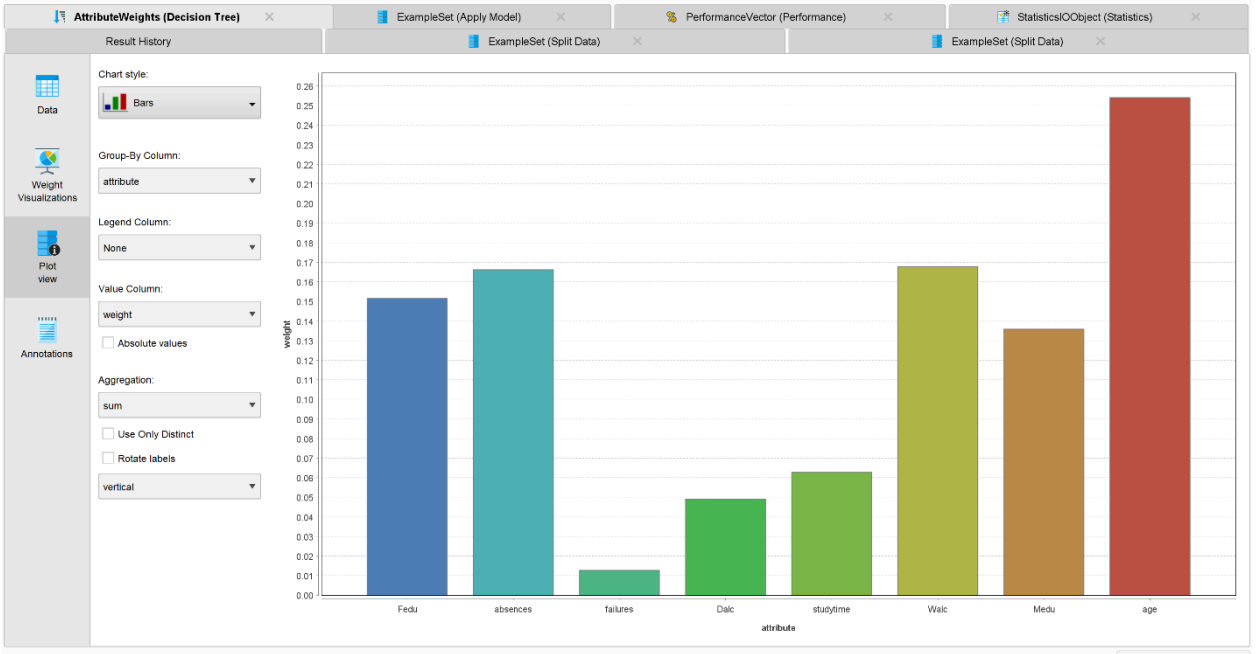


Рисунок 3.12 – визуализация полученных результатов

Оценка качества модели:

Как показано на рисунке 3.13 для данной модели значение root mean squared error (RMSE) — корня средней квадратичной ошибки составляет 3.237, что свидетельствует о допустимом уровне отклонения предсказанных значений от истинных. Низкое значение RMSE указывает на хорошую согласованность модели с данными.

Для оценки полученного результата стоит учитывать диапазон целевой переменной G3 (от 0 до 20). RMSE = 3.237 означает, что средняя ошибка предсказания составляет около трёх баллов, что в данном контексте является удовлетворительным показателем.

Если говорить о RMS в целом:

RMSE < 1 — модель очень точная.

RMSE от 1 до 5 — допустимый уровень ошибки, модель предсказывает значения с некоторым отклонением.

RMSE > 5 — высокая ошибка, предсказания недостаточно точны.

Таким образом, полученная модель имеет удовлетворительный уровень точности, но при необходимости её можно улучшить, например, изменив параметры решающего дерева или используя альтернативные алгоритмы.

4. Приобретаемые навыки

1. Работа с интерфейсом RapidMiner Studio – освоение инструментов для построения процессов, анализа данных и настройки операторов.
2. Построение модели машинного обучения – использование алгоритма Decision Tree для предсказания целевой переменной и настройка его параметров для оптимального разбиения.
3. Применение модели к тестовым данным – освоение оператора Apply Model и корректная интерпретация полученных прогнозов.
4. Оценка качества модели – вычисление метрик точности, таких как root mean squared error (RMSE), и анализ её предсказательной способности.
5. Анализ значимости признаков – использование оператора Attribute Weights для определения наиболее значимых факторов, влияющих на предсказания.
6. Развитие навыков анализа данных – интерпретация полученных результатов, выявление закономерностей и подготовка отчёта по итогам лабораторной работы.

5. Обобщенная задача для индивидуального варианта

Цель работы – предсказать значения целевой переменной на основе алгоритма решающих деревьев. В качестве исходных данных вам предлагается индивидуальный датасет с числовыми и/или категориальными признаками. Этапы выполнения:

1) Загрузка и первичная проверка данных

- Импортируйте CSV-файл через оператор Read CSV.
- Убедитесь в корректности типов столбцов, при необходимости измените их с помощью Type Conversion.
- Проверьте набор на пропуски и выбросы, обработайте их через Replace Missing Values или Filter Examples.

2) Подготовка признаков

- С помощью Select Attributes отберите для анализа только те признаки, которые имеют смысл для предсказания (исключите идентификаторы, текстовые поля и т. п.).
- При необходимости создайте новые признаки через Generate Attributes (например, комбинации или нормированные версии исходных).
- Разбейте данные на обучающую и тестовую выборки (оператор Split Data, соотношение ~70/30).

3) Построение модели решающего дерева

- Добавьте оператор Decision Tree и настройте:
  • критерий разбиения (например, Gini или Information Gain),
  • максимальную глубину,
  • минимальный размер листа или параметр предварительной обрезки.
- Обучите модель на тренировочной выборке.

4) Применение и оценка модели

- С помощью Apply Model сделайте предсказания на тестовой выборке.
- Оцените качество модели через Performance (Regression) или Performance (Classification) в зависимости от типа целевой переменной: RMSE/MAE/R^2 для регрессии или Accuracy/Precision/Recall/F1 для классификации.

5) Интерпретация результатов

- Проанализируйте важность признаков (Attribute Weights) и выясните, какие факторы наиболее влияют на предсказания.
- Постройте визуализацию структуры дерева (Plot View) и кратко опишите ключевые ветвления: на каких признаках модель «разветвляется» в первую очередь.

6) Выводы

- Сформулируйте основные наблюдения о том, какие признаки и в каком порядке влияют на целевую переменную.

6. Распределение вариантов