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

Кафедра информационной безопасности киберфизических систем

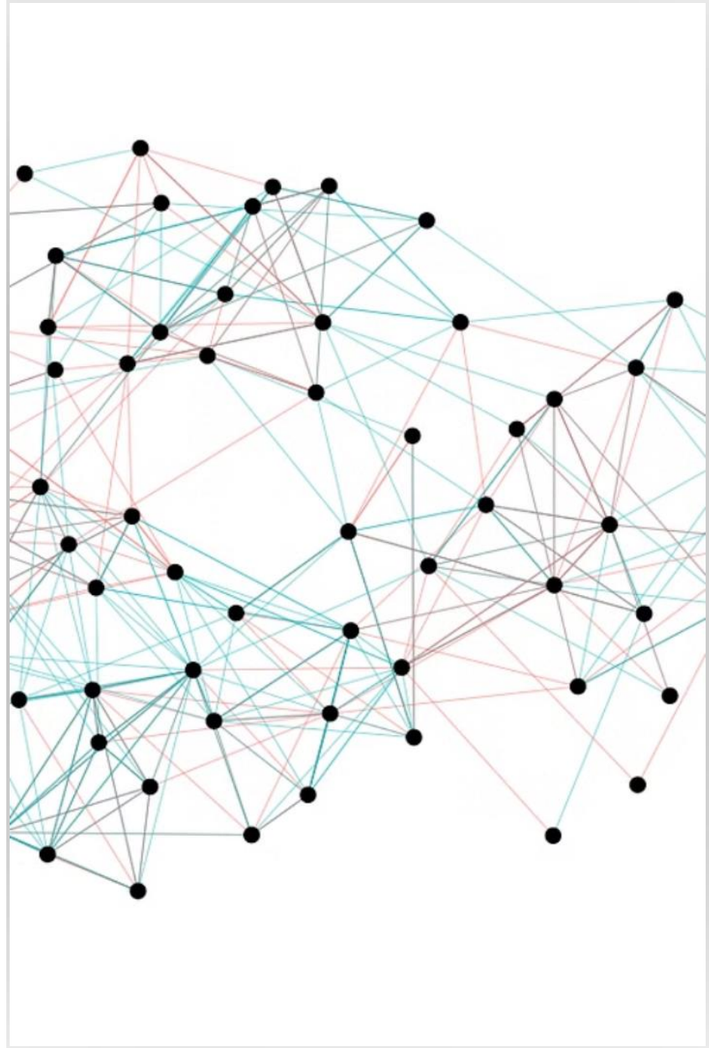
Москва 2025

Анализ данных методом решающих деревьев

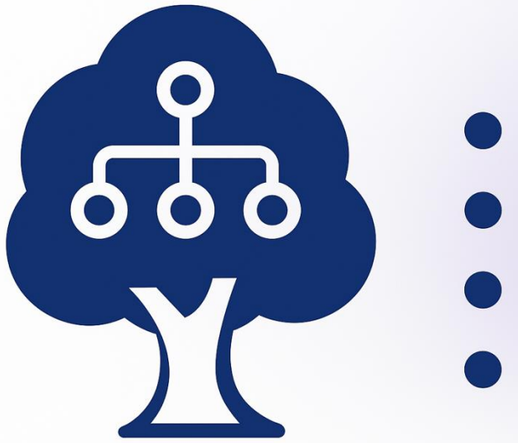
Анализ данных методом решающих деревьев

Решающие деревья – алгоритм машинного обучения, создающий древовидную модель принятия решений на основе данных.

Используется для классификации и регрессии, популярен благодаря высокой интерпретируемости.



Области применения решающих деревьев

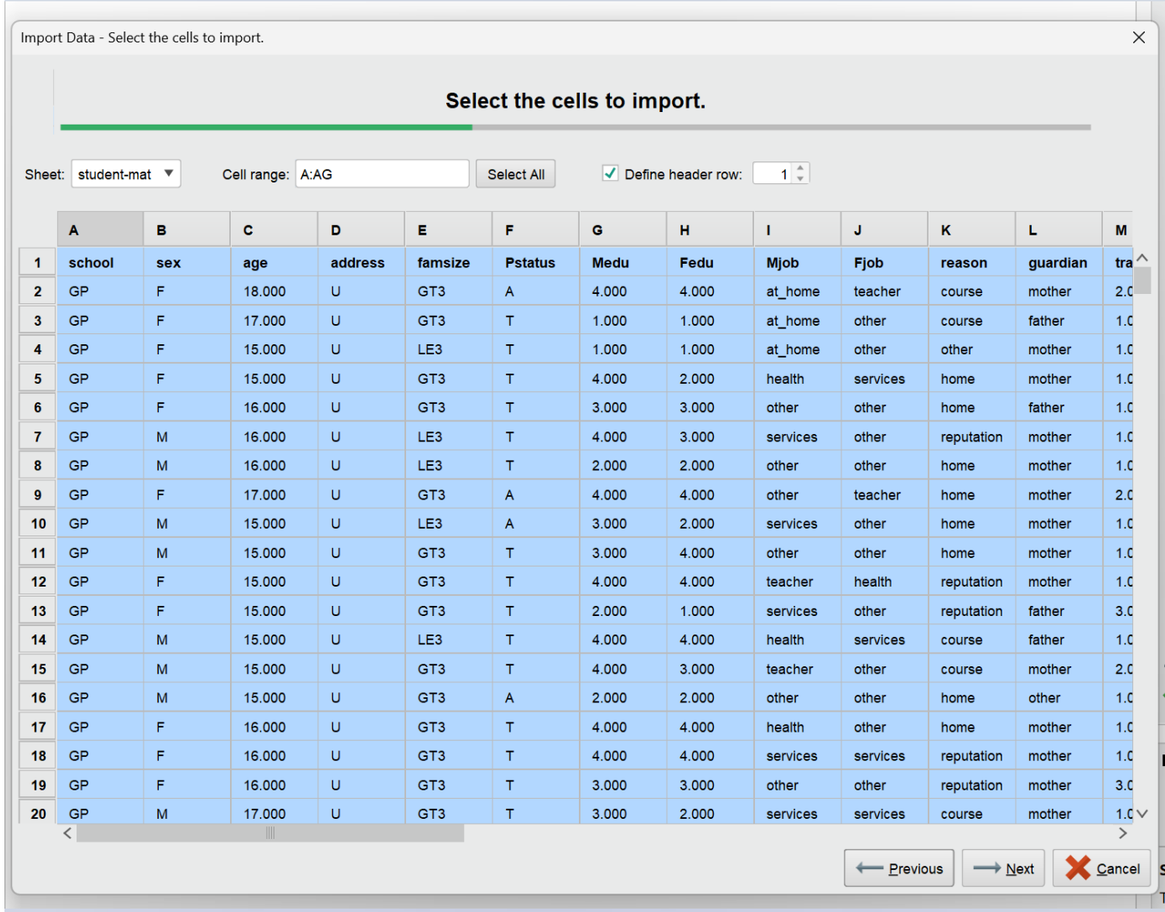


- Финансы: кредитный скоринг, оценка рисков
- Медицина: диагностика заболеваний
- Маркетинг: сегментация клиентов
- Образование: прогноз успеваемости студентов на основе различных факторов

Описание тестовых данных

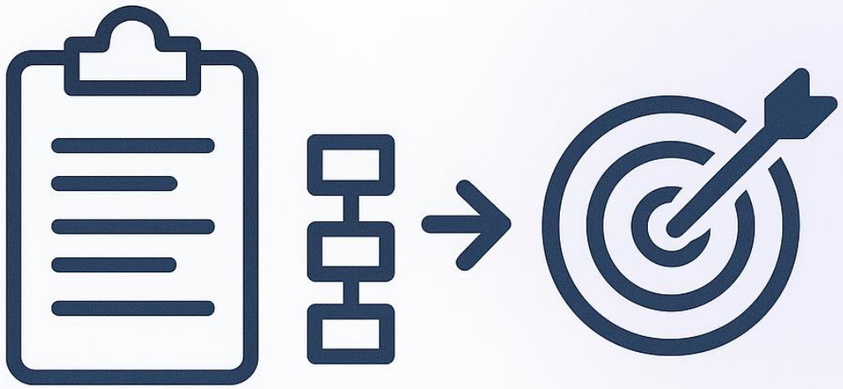
Данные содержат характеристики студентов: возраст, образование родителей, учебное время, потребление алкоголя и пропуски занятий.

Целевая переменная — итоговая оценка (G3).



	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

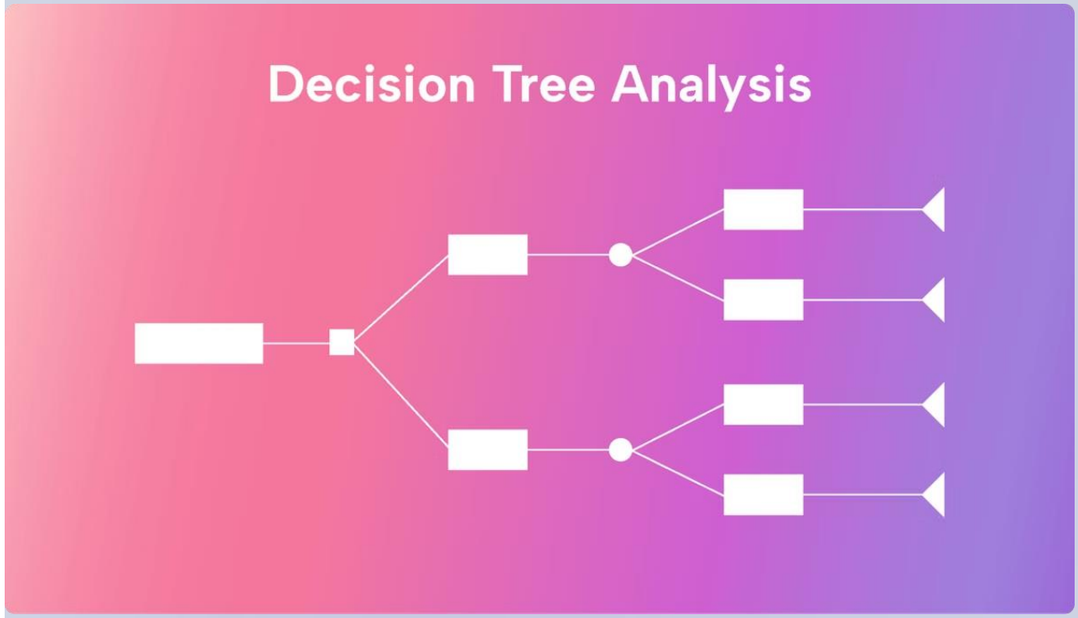
Цель анализа



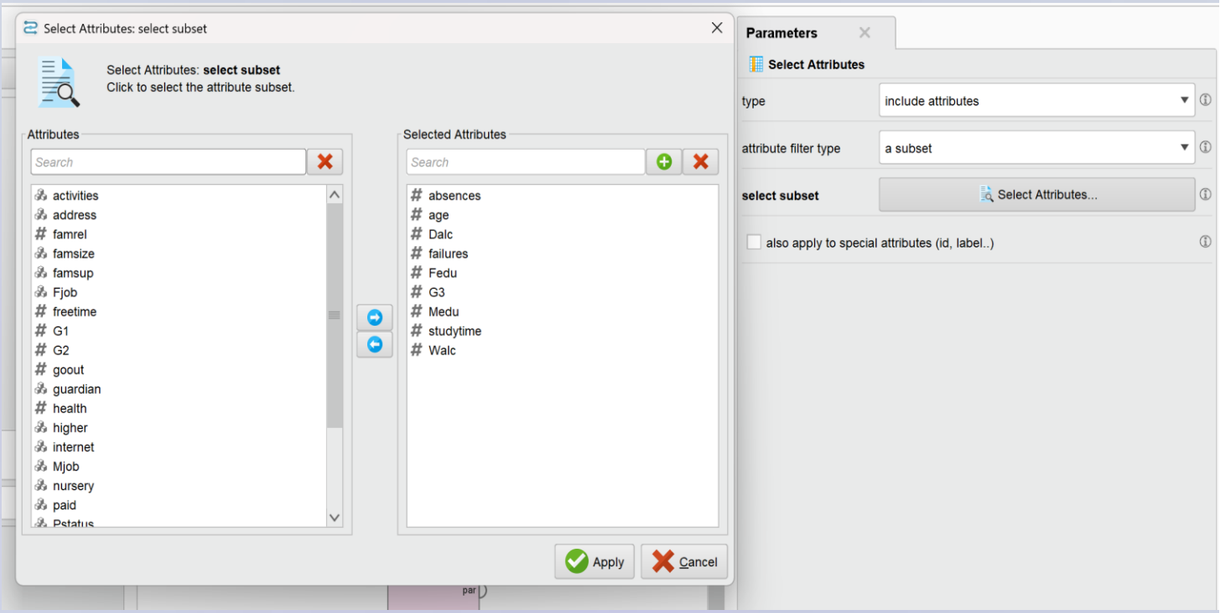
Выявить факторы, наиболее влияющие на успеваемость студентов, и построить модель, способную прогнозировать итоговые оценки на основе представленных признаков.

Преимущества решающих деревьев

- Простая интерпретация
- Работа с категориальными и числовыми признаками
- Устойчивость к выбросам
- Возможность визуализации решений



Подготовка данных: удаление лишних столбцов

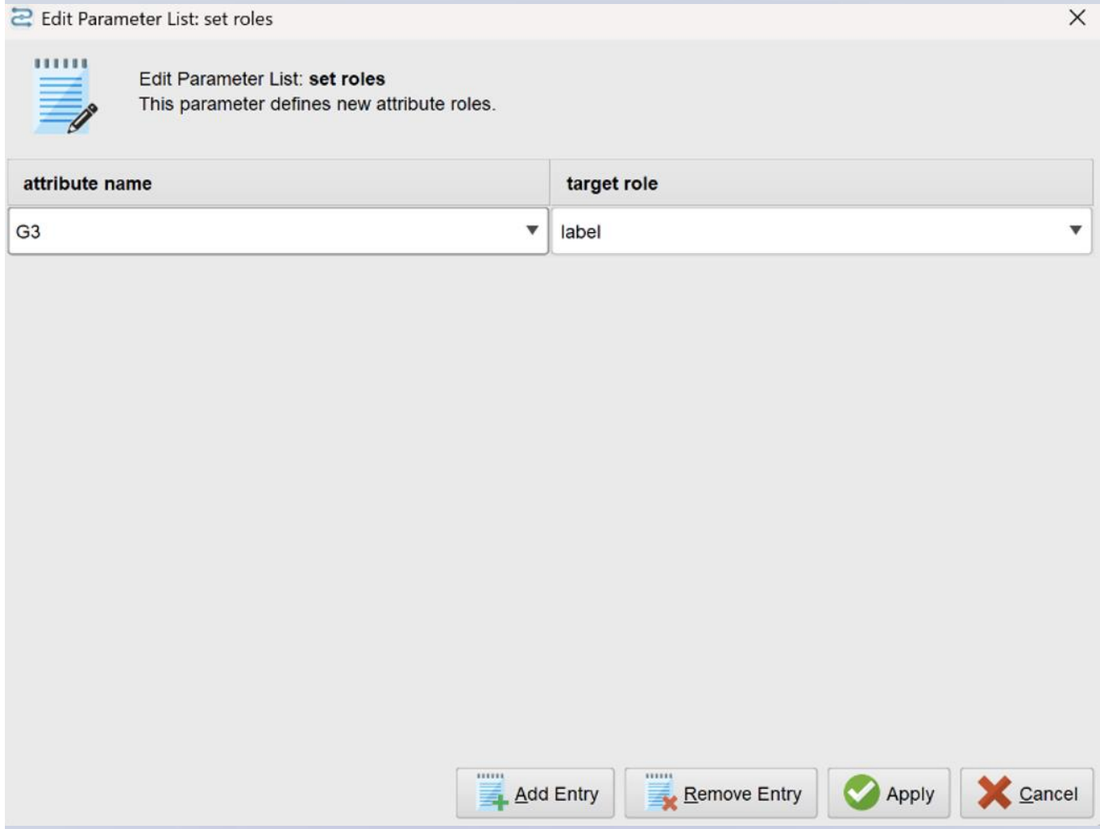


На этапе предварительной обработки удаляются неиспользуемые столбцы.

Это помогает повысить точность анализа, оставив только значимые атрибуты.

Установка целевого признака

Целевая переменная (G3) задаётся явно через оператор «Set Role».
Это необходимо, чтобы алгоритм понимал, какую характеристику он должен прогнозировать на основе других признаков.



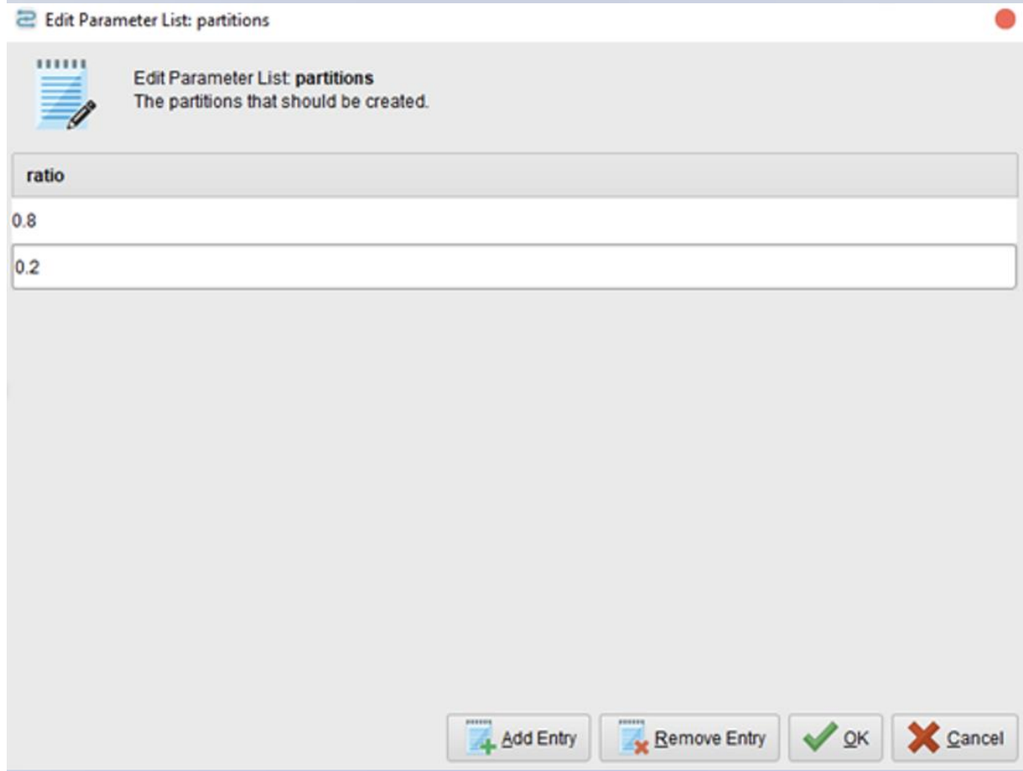
Проверка и очистка данных

На этапе предобработки проверяются данные на наличие пустых или некорректных значений, что улучшает качество модели и предотвращает ошибки во время анализа.

Разделение данных на обучение и тестирование

Данные разделяются на две выборки: обучающую (80%) и тестовую (20%).

Это позволяет объективно оценить точность модели и избежать её переобучения.



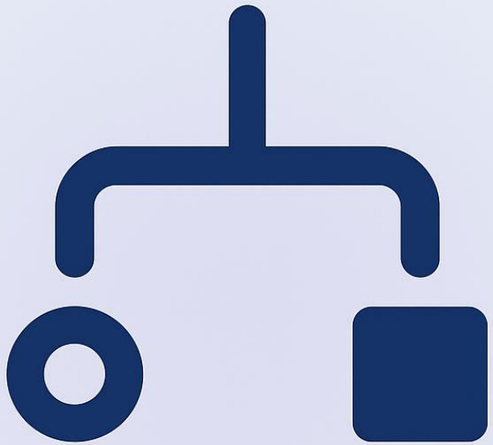
Понятие переобучения модели

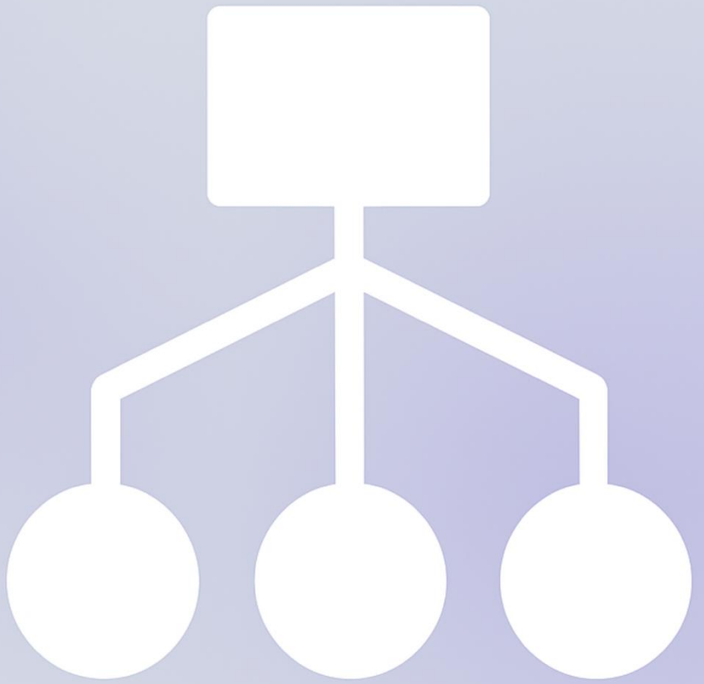
Переобучение происходит, когда модель идеально описывает обучающие данные, но плохо обобщает новые.

Его можно контролировать ограничением глубины дерева или минимального размера листьев.

Типы задач для решающих деревьев

- Классификация (целевой признак категориальный)
- Регрессия (целевой признак числовой)





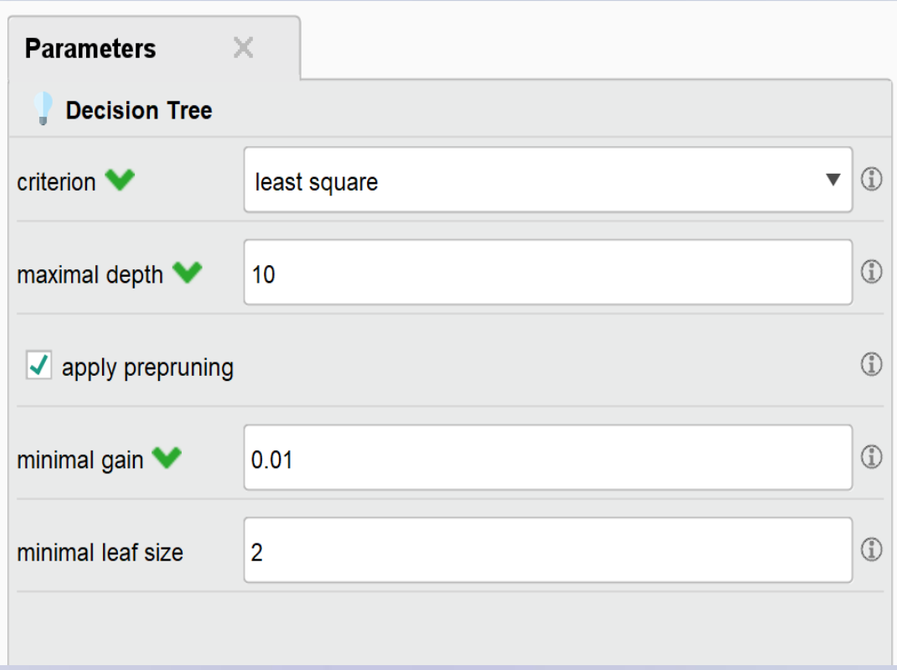
Критерий разбиения: Least Square

Используется критерий Least Square, минимизирующий сумму квадратов ошибок при разделении данных.

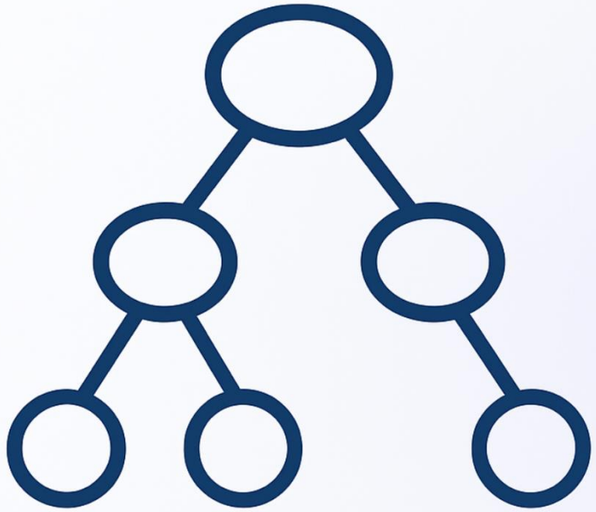
Это позволяет оптимально распределять объекты по ветвям дерева.

Настройки решающего дерева в RapidMiner

- Максимальная глубина (ограничение сложности)
- Минимальный размер листа (контроль переобучения)
- Минимальный прирост информации (минимальное улучшение при разделении)



Построение решающего дерева в RapidMiner



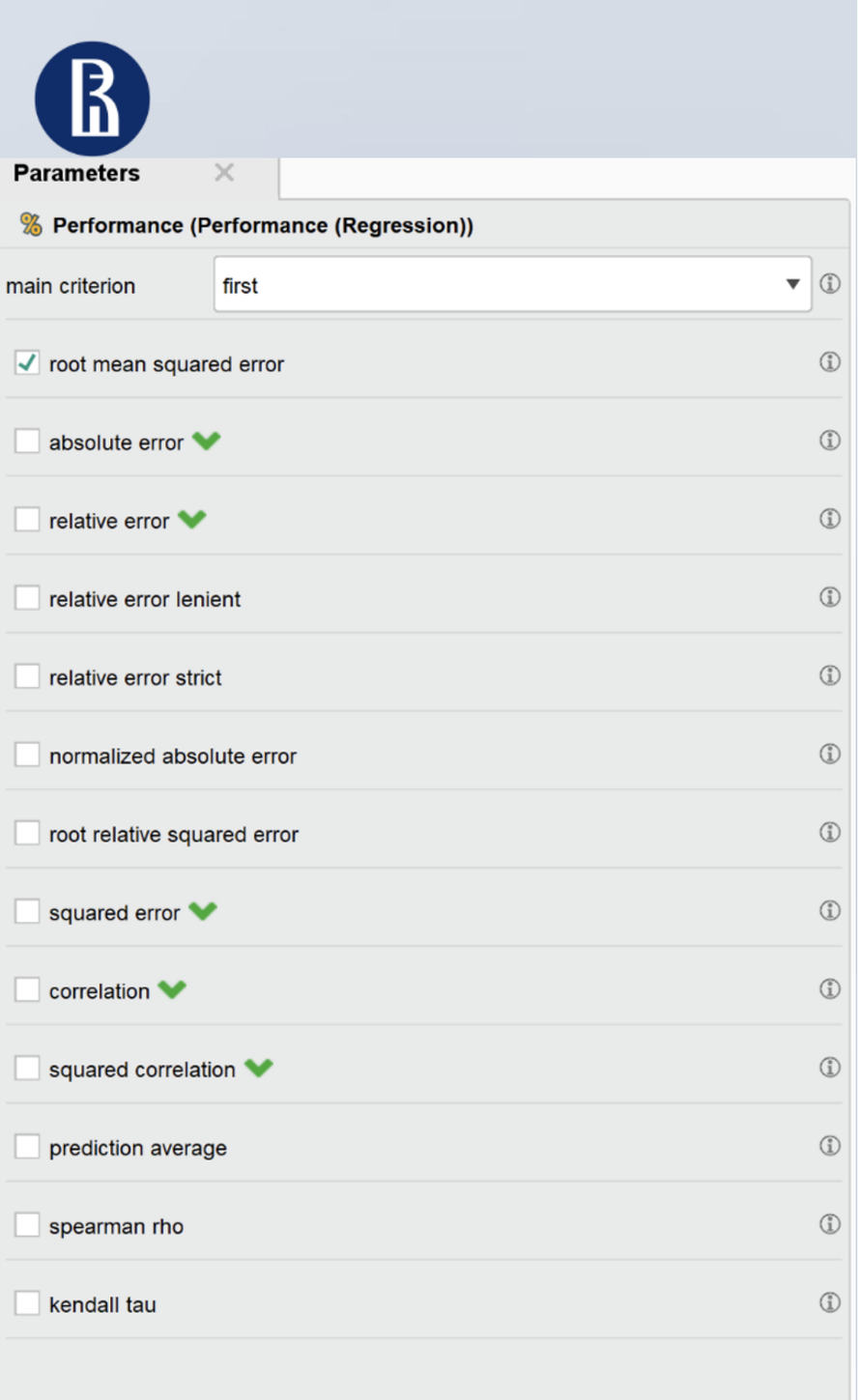
Оператор Decision Tree автоматически генерирует модель, используя заданные параметры.

Глубина дерева и предварительная обрезка помогают найти баланс точности и простоты модели.

Применение модели на тестовой выборке

Тестовая выборка используется для проверки способности модели делать точные прогнозы на новых данных.

Оператор Apply Model генерирует предсказания на основе обученной модели.



Оценка качества модели

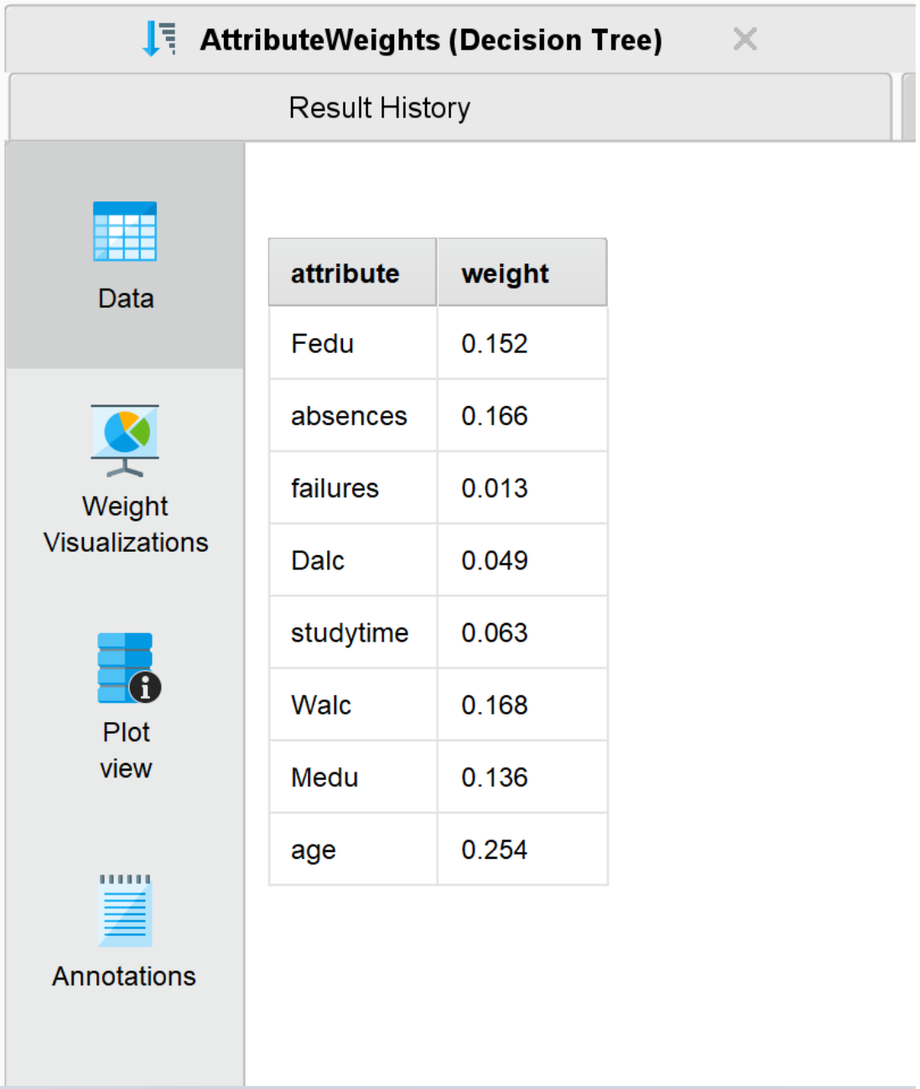
Для оценки точности используются метрики:

- RMSE (среднеквадратичная ошибка)
- Корреляция предсказаний и реальных значений
- Абсолютная и относительная ошибка

Значение метрики RMSE

RMSE показывает среднее отклонение прогнозов от реальных значений. Чем ниже значение RMSE, тем точнее прогнозы модели.
Приемлемым считается значение RMSE менее 5 для диапазона оценок 0–20.

Значимость признаков



attribute	weight
Fedu	0.152
absences	0.166
failures	0.013
Dalc	0.049
studytime	0.063
Walc	0.168
Medu	0.136
age	0.254

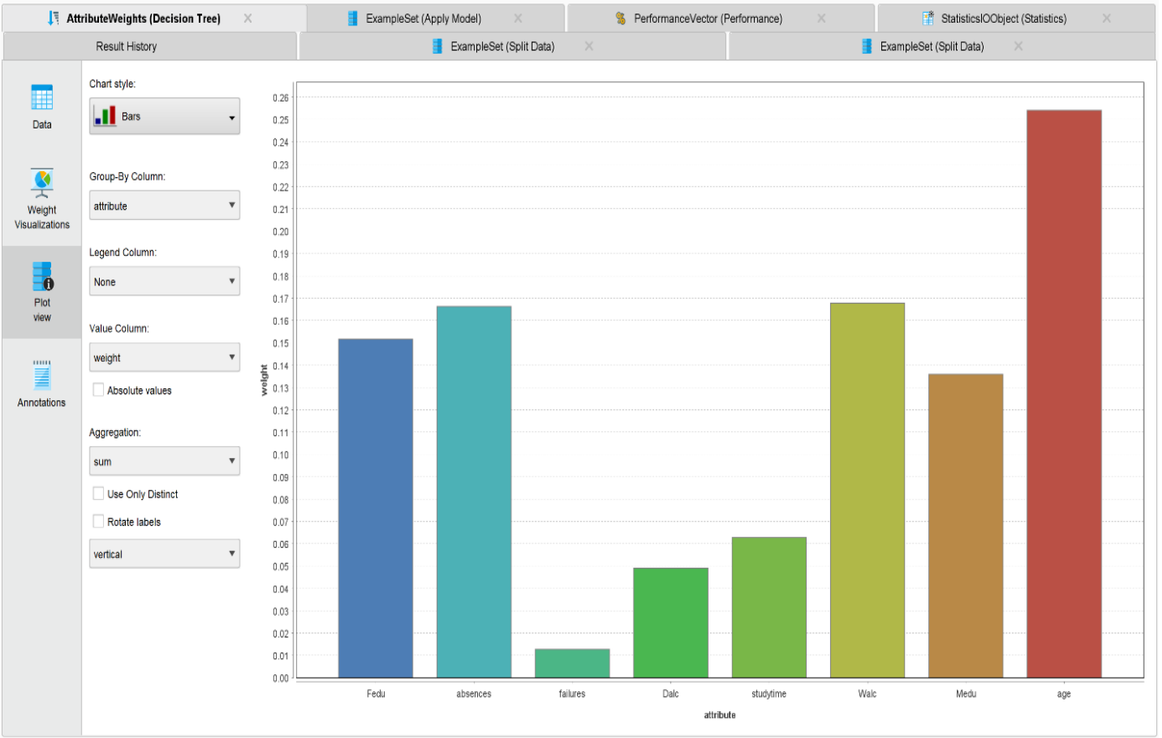
Каждый признак имеет свой вклад в итоговый прогноз.

Наиболее влиятельными оказались возраст (age), пропуски занятий (absences) и употребление алкоголя в выходные (Walc).

Интерпретация факторов

Старший возраст, низкое количество пропусков и низкое потребление алкоголя положительно коррелируют с успеваемостью.
Неудачи в прошлом (failures) имеют низкое влияние на итоговую оценку.

Визуализация важности признаков

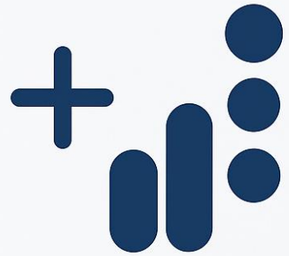


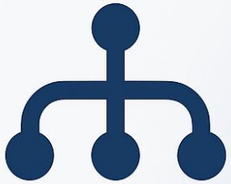
RapidMiner позволяет построить диаграмму значимости признаков.

Это наглядно демонстрирует, какие характеристики студентов сильнее всего влияют на итоговые оценки.

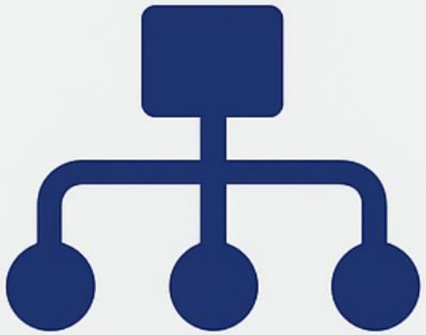
Пути улучшения модели

- Оптимизация параметров (глубина, обрезка)
- Добавление дополнительных признаков
- Использование альтернативных алгоритмов (Random Forest, Gradient Boosted Trees)

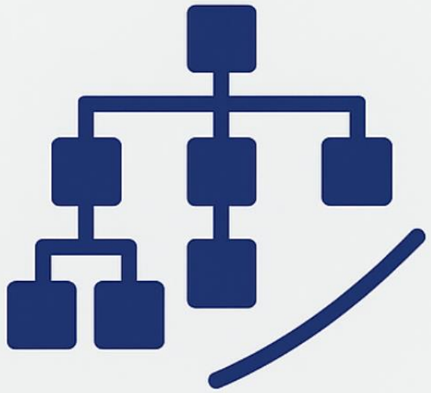




Альтернативные алгоритмы



Random Forest и Gradient Boosted Trees обычно дают более высокую точность, но менее интерпретируемы.



Решающее дерево часто используется именно за прозрачность выводов и простоту объяснения.

Практическое применение тестовой модели

Результаты анализа могут использоваться преподавателями и администрацией для разработки рекомендаций по улучшению успеваемости

Заключение

Решающие деревья в RapidMiner дают понятные и интерпретируемые прогнозы.

Тестовая модель показывает удовлетворительную точность (RMSE ≈ 3.2), выявляет значимые факторы, влияющие на успеваемость студентов.