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

Кафедра информационной безопасности киберфизических систем

Москва 2025

Знакомство с инструментом RapidMiner

Знакомство с инструментом RapidMiner

Цель: изучить возможности платформы RapidMiner для анализа данных и машинного обучения, освоить основные подходы и функции, необходимые для проведения аналитических исследований и построения моделей.

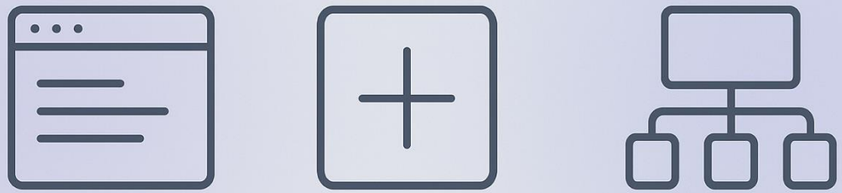
Что такое RapidMiner

RapidMiner – это платформа для анализа данных, машинного обучения и предиктивного моделирования без необходимости программирования.

Платформа используется в академической и коммерческой деятельности с 2001 года.

Преимущества No-code решений

Платформы No-code, такие как RapidMiner, упрощают работу с данными, делая аналитику доступной широкому кругу специалистов без навыков программирования, повышая эффективность решения задач анализа данных.



Основные этапы анализа данных

Этапы включают сбор данных (определение источников информации и её получение) и очистку данных (устранение ошибок, заполнение пропусков и удаление выбросов).

Основные этапы анализа данных

Следующие этапы — это сам анализ данных с применением статистических методов и интерпретация результатов, заключающаяся в формулировании выводов и рекомендаций.

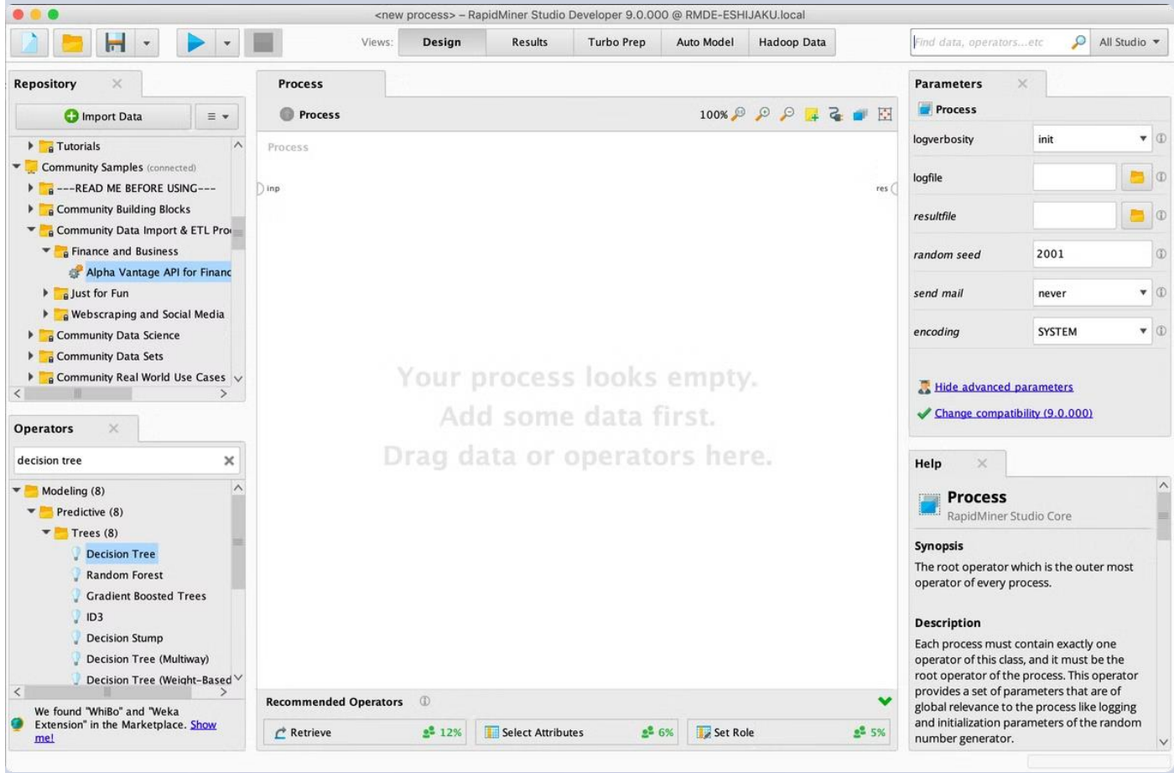
Типы данных и их особенности

В анализе используются числовые, категориальные и текстовые данные.

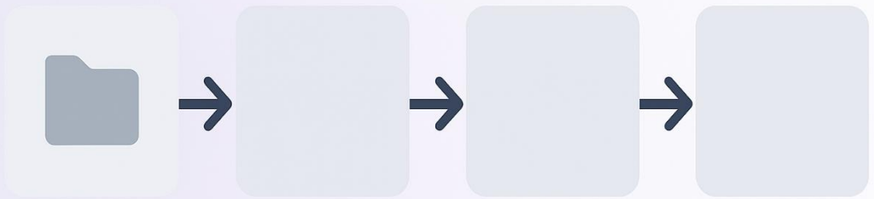
Каждый тип требует особой обработки: нормализации, кодирования или извлечения признаков для эффективного анализа.

Интерфейс RapidMiner Studio

Рабочая среда RapidMiner состоит из рабочего пространства для построения процессов, библиотеки операторов, панели результатов и панели управления для навигации между этапами анализа.



Понятие операторов в RapidMiner

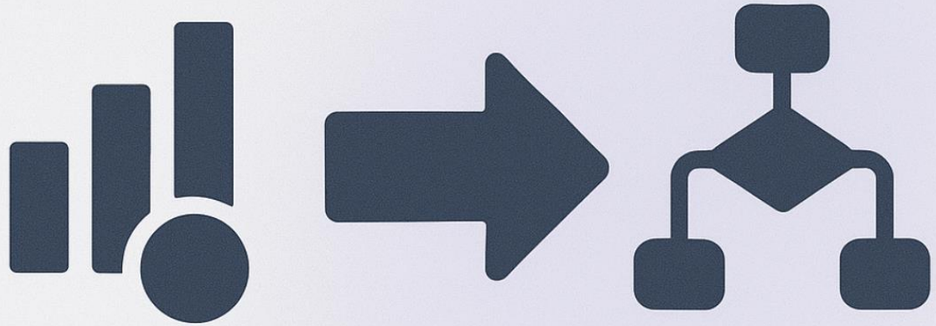


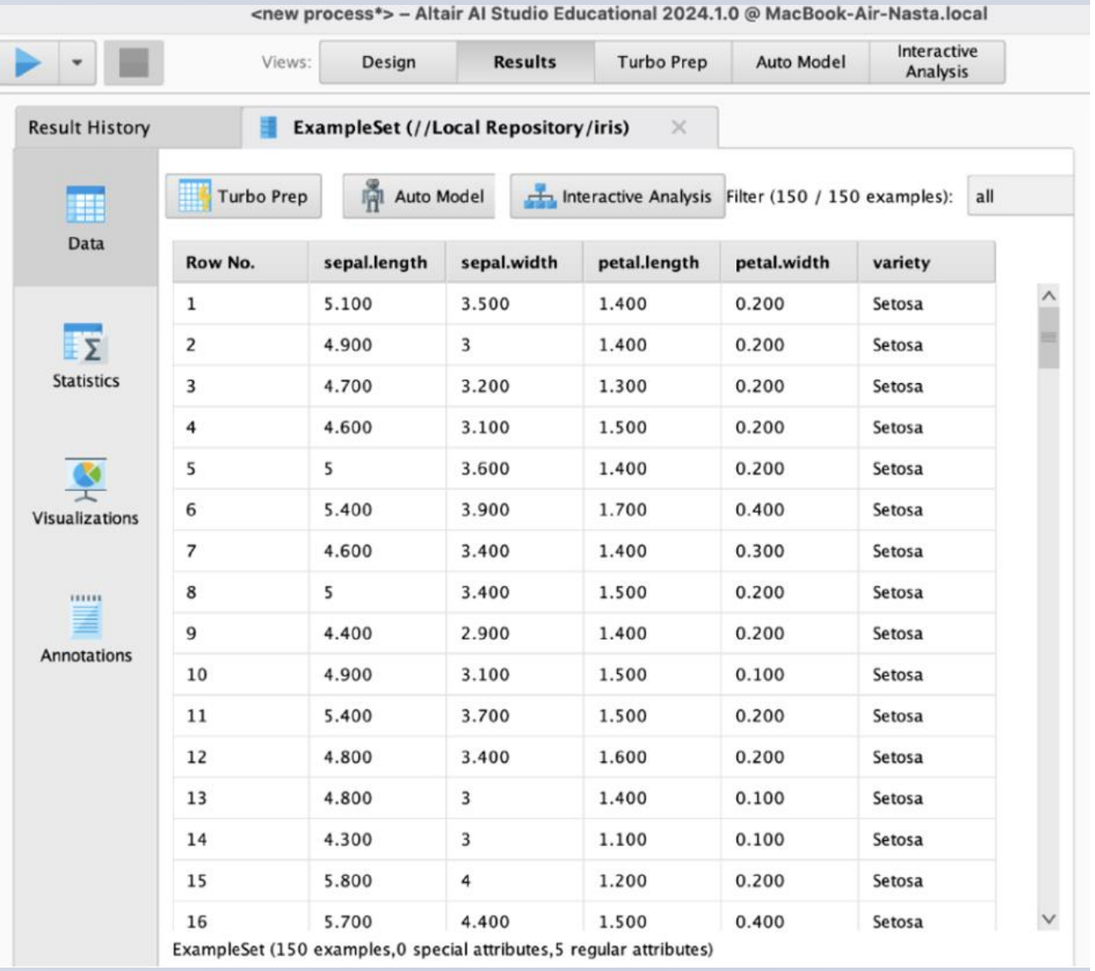
Операторы – это компоненты, реализующие отдельные действия в анализе данных: загрузку файлов, очистку, нормализацию, построение моделей и визуализацию.

Они соединяются в последовательности.

Преимущества графического подхода

Использование графического drag-and-drop интерфейса позволяет визуально контролировать весь процесс анализа, упрощая понимание и интерпретацию даже сложных аналитических сценариев.



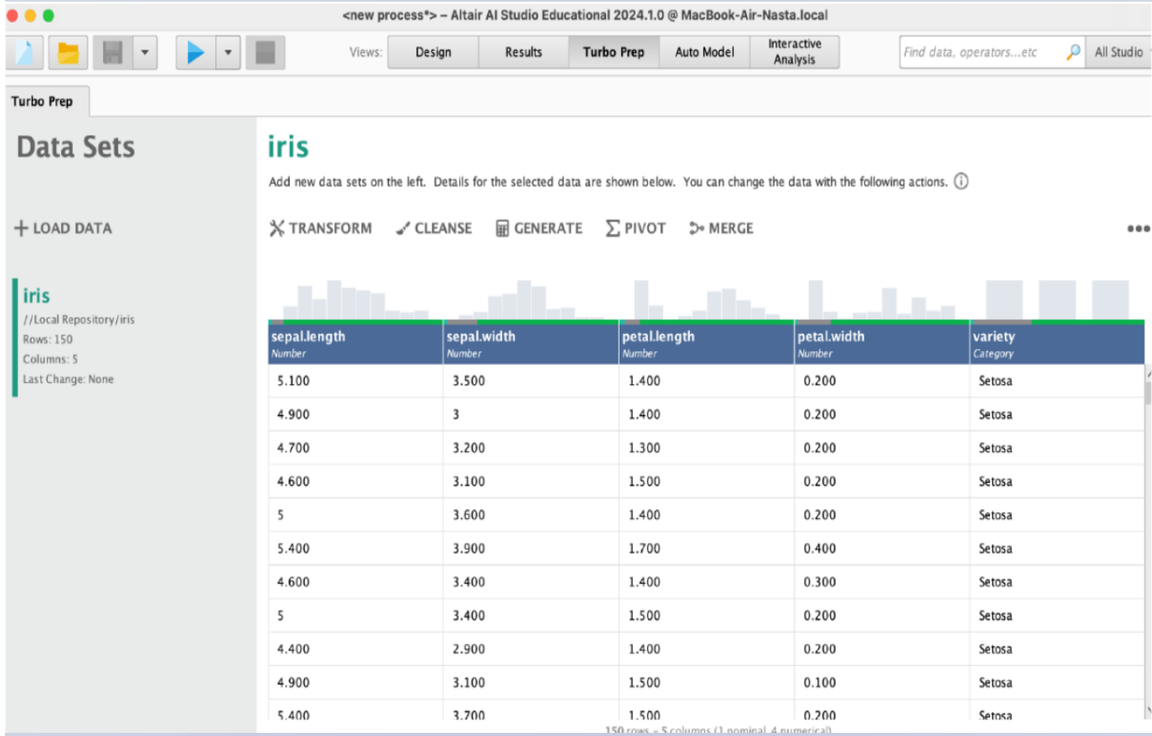


Загрузка и предварительный просмотр данных

В RapidMiner загрузка данных осуществляется через оператор Read CSV. После загрузки данные отображаются в виде таблицы, позволяя предварительно оценить качество и структуру.

Подготовка данных: инструмент TurboPrep

TurboPrep в RapidMiner облегчает процесс предварительной обработки, автоматизируя очистку от пропусков, устранение выбросов и нормализацию данных, что существенно ускоряет аналитику.

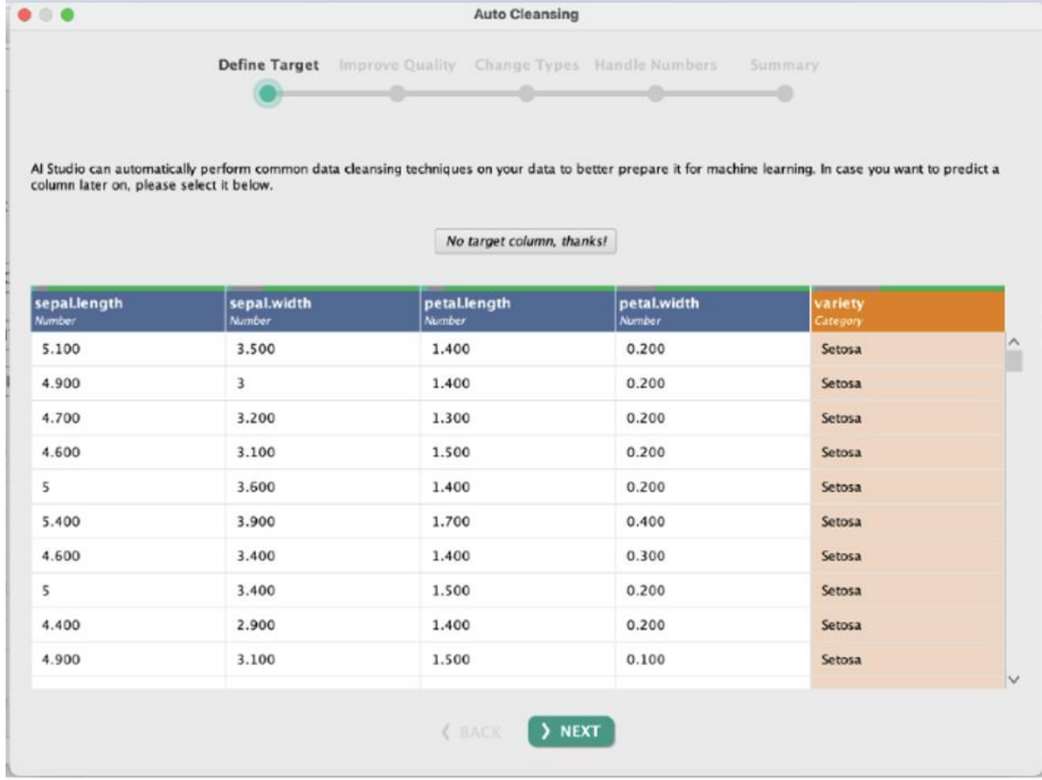


Преобразование и создание признаков

Операторы Transform и Generate позволяют изменять данные, создавать новые признаки и улучшать их качество, что повышает точность моделей машинного обучения.

Очистка данных (Cleanse)

Функция Cleanse используется для автоматического обнаружения и устранения проблем в данных: удаление дубликатов, заполнение пропусков и исправление ошибок в значениях.

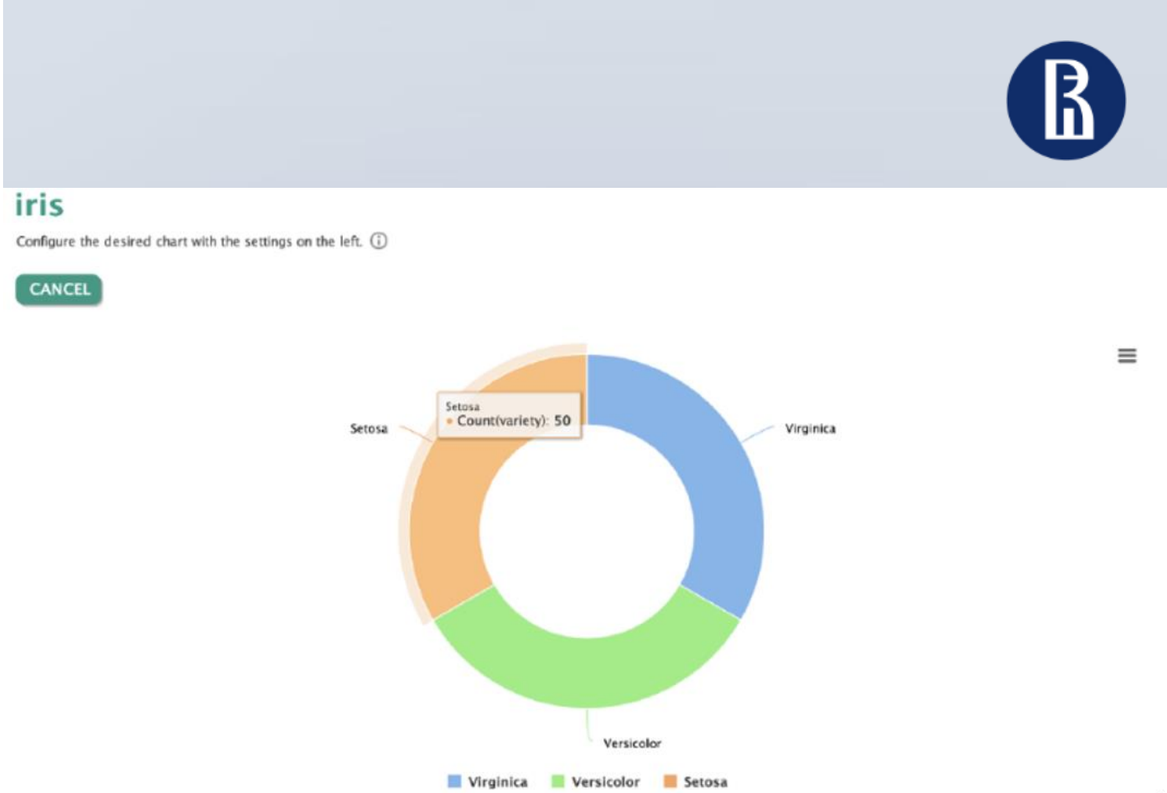


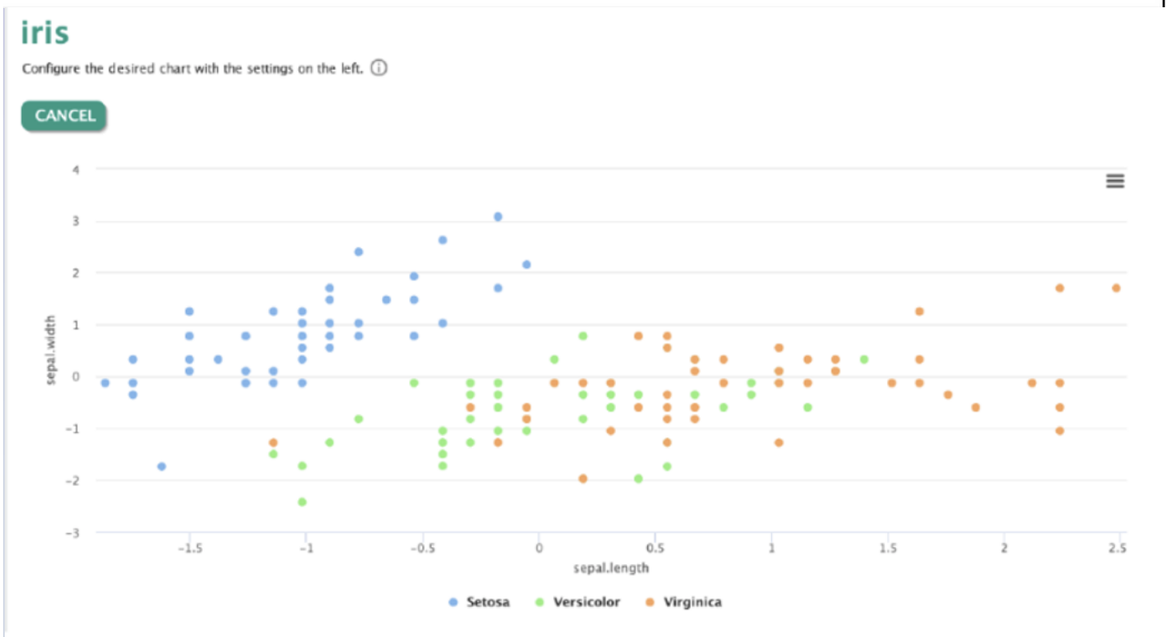
Нормализация данных

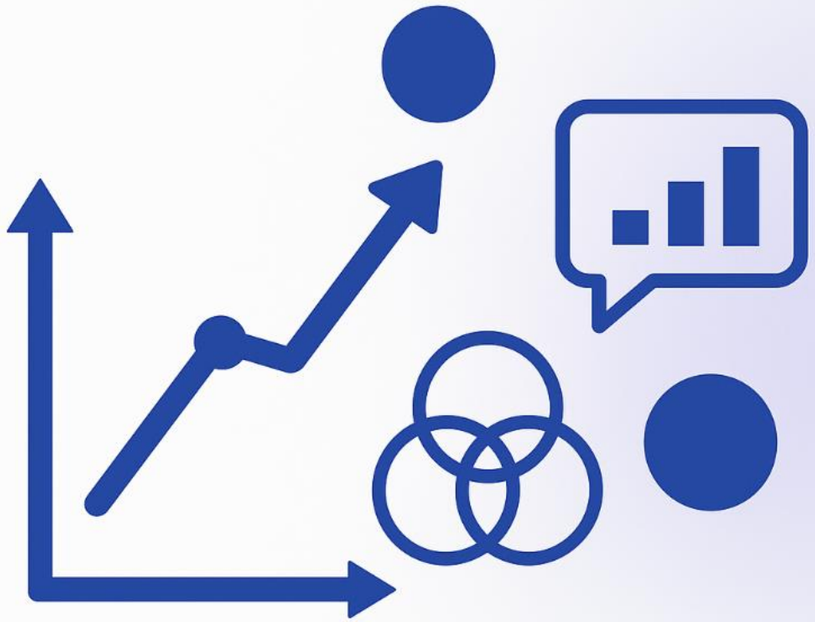
Нормализация данных – это приведение числовых признаков к единой шкале, необходимое для корректного функционирования многих моделей машинного обучения, например, ближайших соседей или регрессии.

Визуализация данных: виды графиков

В RapidMiner доступны различные визуализации: круговые диаграммы (доли категорий), диаграммы рассеяния (зависимости признаков) и гистограммы (распределение значений признаков).





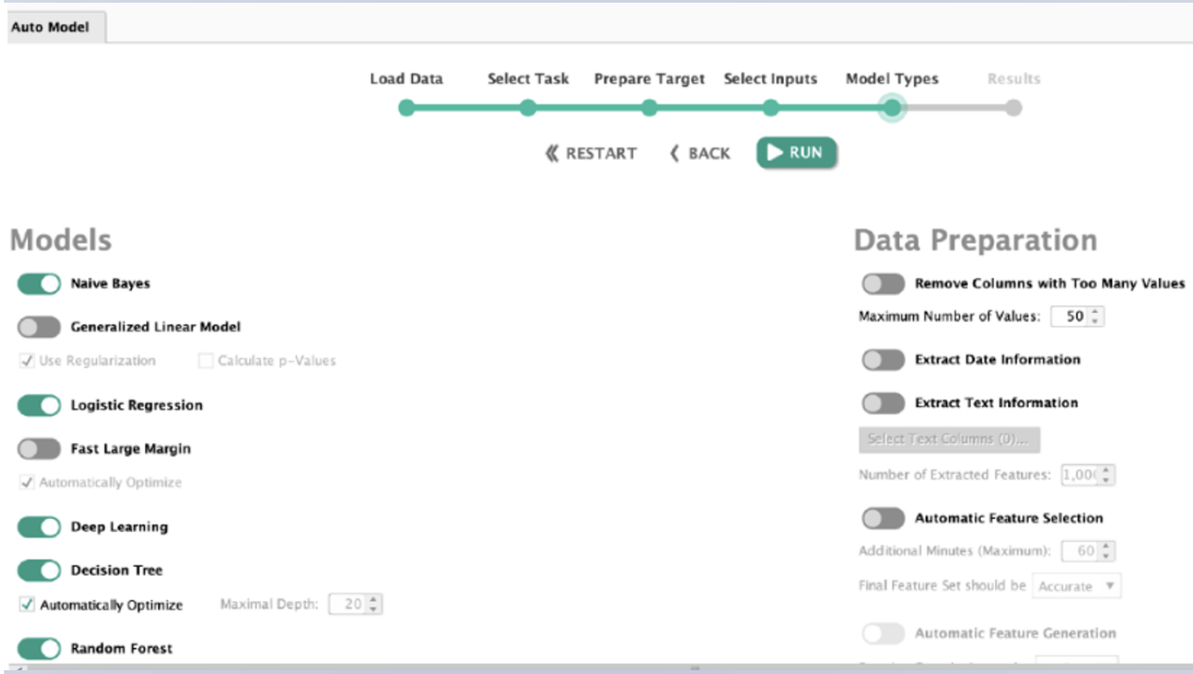


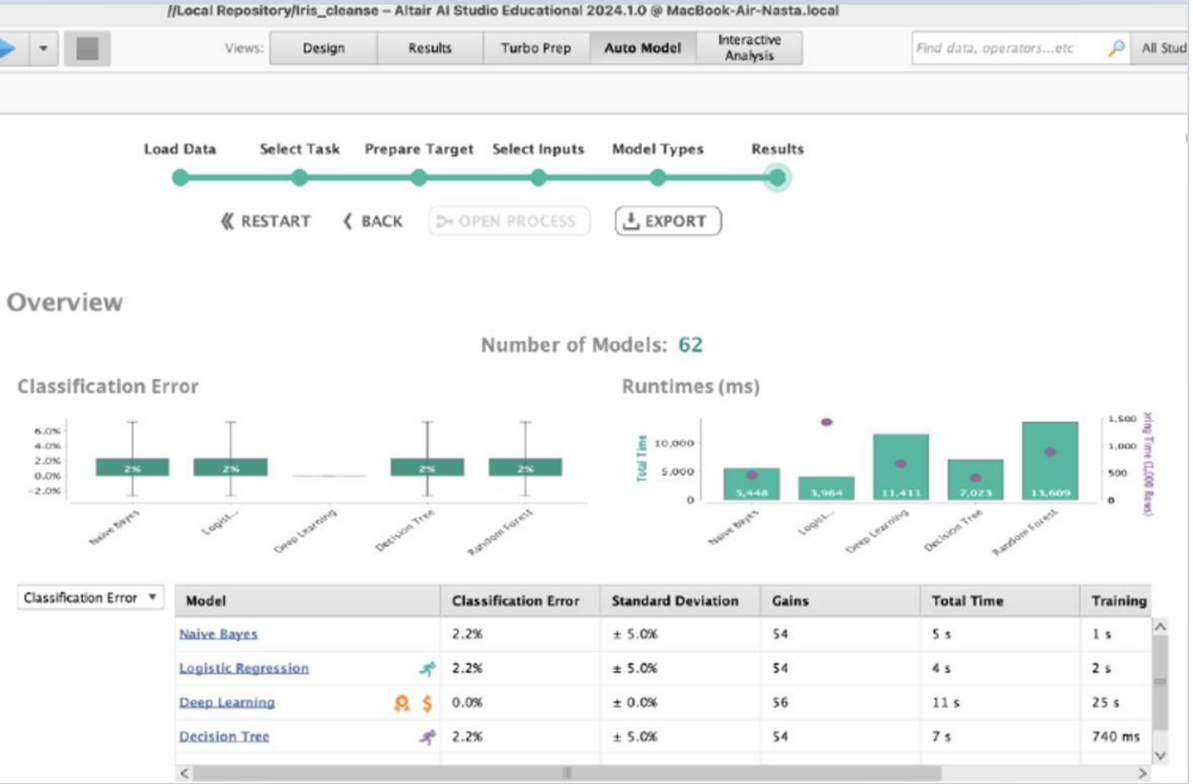
Преимущества визуального анализа данных

Визуализация позволяет быстро выявить тренды, зависимости и потенциальные аномалии в данных, обеспечивая эффективную коммуникацию аналитических результатов.

Инструмент AutoModel: автоматизация анализа

AutoModel позволяет автоматически подбирать оптимальные алгоритмы машинного обучения, проводить их обучение и тестировать результаты без необходимости ручной настройки параметров.





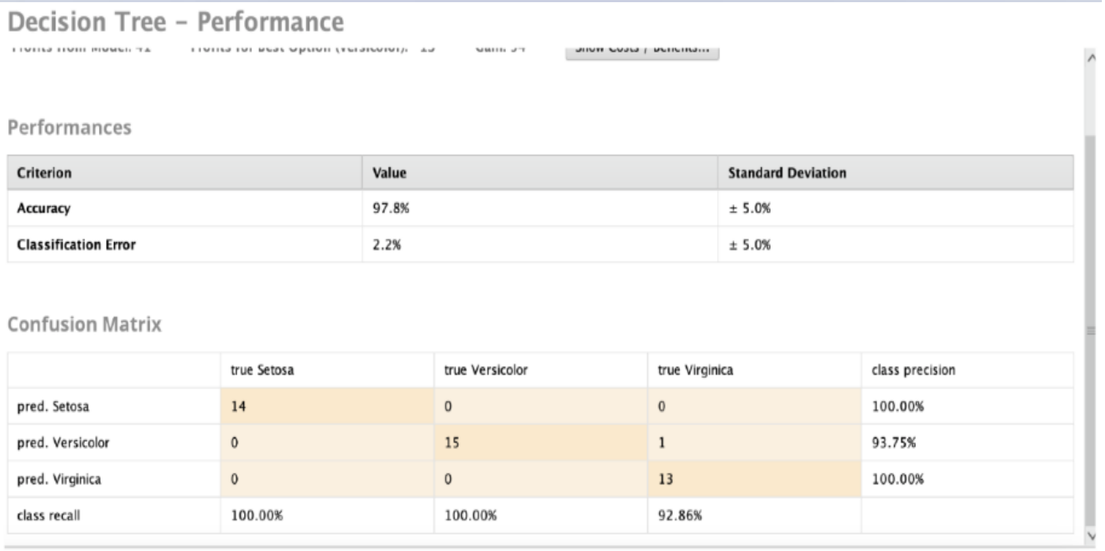
Инструмент AutoModel: автоматизация анализа

Используя AutoModel, пользователь может быстро сравнить эффективность различных моделей, таких как деревья решений, логистическая регрессия и случайные леса, по метрикам качества.

Алгоритмы машинного обучения в RapidMiner

RapidMiner поддерживает различные алгоритмы: деревья решений (интерпретируемость), логистическая регрессия (простота), случайные леса (точность) и глубокое обучение (решение сложных задач).

Оценка качества моделей



Decision Tree - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	97.8%	± 5.0%
Classification Error	2.2%	± 5.0%

Confusion Matrix

	true Setosa	true Versicolor	true Virginica	class precision
pred. Setosa	14	0	0	100.00%
pred. Versicolor	0	15	1	93.75%
pred. Virginica	0	0	13	100.00%
class recall	100.00%	100.00%	92.86%	

Метрики качества, такие как Accuracy (точность), Precision (точность классификации), Recall (полнота) и ROC-кривая, используются для объективной оценки эффективности работы моделей.

Анализ важности признаков в модели

RapidMiner позволяет определить, какие признаки наиболее значимы для модели.

Это помогает исключить ненужные переменные и улучшить качество итогового прогноза.

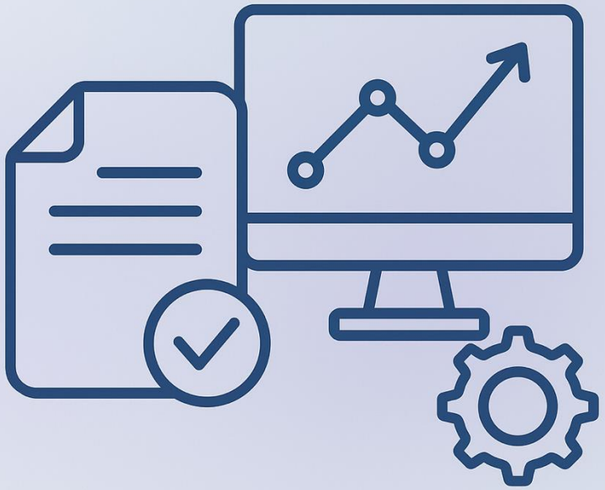
Проблемы в машинном обучении и их решение

При работе с моделями возникают проблемы: пропущенные значения, выбросы и переобучение. RapidMiner предоставляет инструменты для эффективного решения этих задач.

Применение RapidMiner на практике

RapidMiner широко применяется для решения реальных задач в маркетинге (сегментация клиентов), финансах (анализ рисков), медицине (диагностика заболеваний) и промышленности (предсказание поломок оборудования).

Заключение



В результате освоения RapidMiner изучены основные функции анализа данных и автоматизации моделирования. Полученные навыки позволяют эффективно решать задачи анализа и машинного обучения в различных прикладных областях.