



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.



Real Data Science, Fast & Simple



Что такое RapidMiner

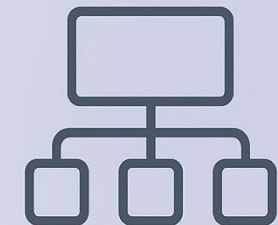
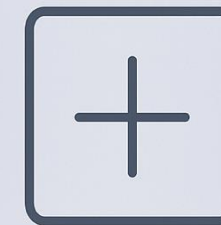
RapidMiner – это платформа для анализа данных, машинного обучения и предиктивного моделирования без необходимости программирования.



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

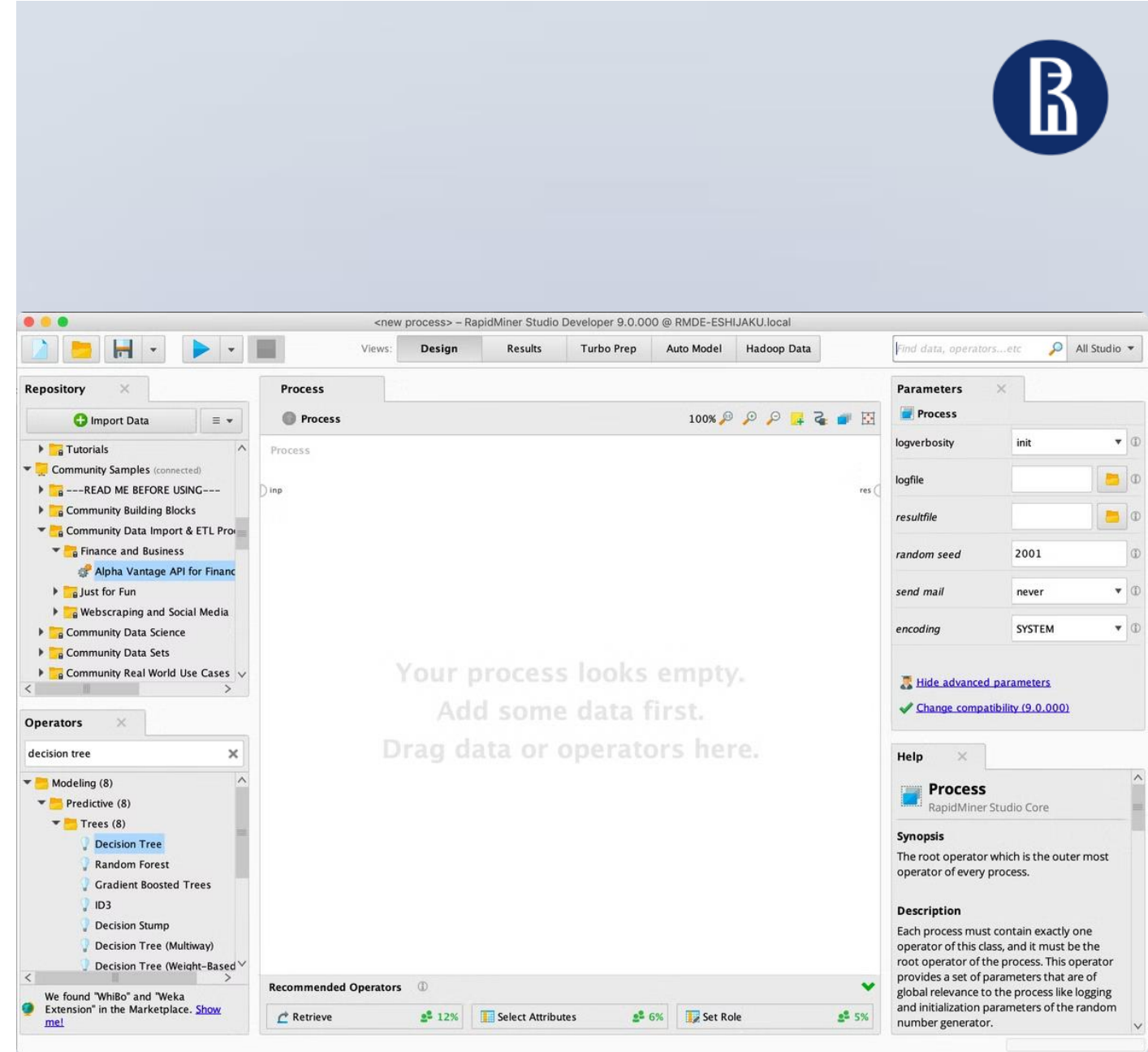
В анализе используются числовые, категориальные и текстовые данные.

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



Интерфейс RapidMiner Studio

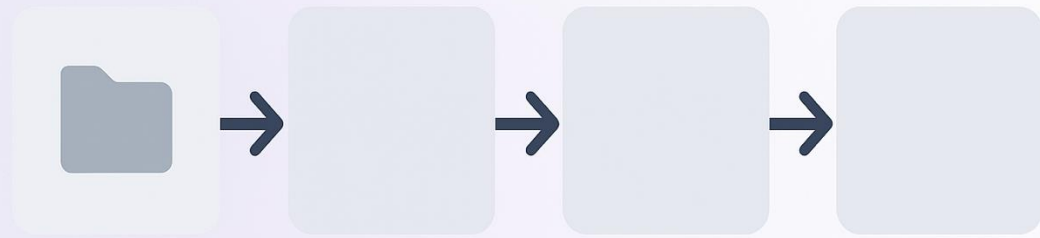
Рабочая среда RapidMiner состоит из рабочего пространства для построения процессов, библиотеки операторов, панели результатов и панели управления для навигации между этапами анализа.





Понятие операторов в RapidMiner

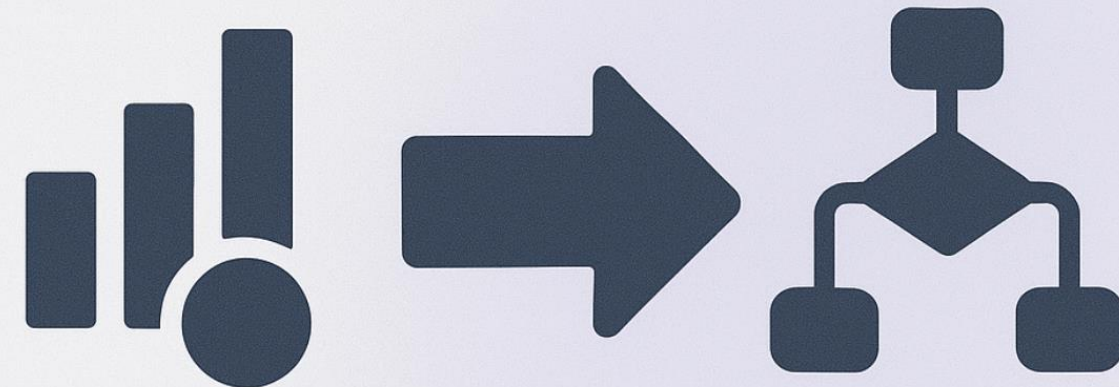
Операторы – это компоненты, реализующие отдельные действия в анализе данных: загрузку файлов, очистку, нормализацию, построение моделей и визуализацию.



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.





<new process*> – Altair AI Studio Educational 2024.1.0 @ MacBook-Air-Nasta.local

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Result History ExampleSet (//Local Repository/iris) x

Turbo Prep Auto Model Interactive Analysis Filter (150 / 150 examples): all

Row No.	sepal.length	sepal.width	petal.length	petal.width	variety
1	5.100	3.500	1.400	0.200	Setosa
2	4.900	3	1.400	0.200	Setosa
3	4.700	3.200	1.300	0.200	Setosa
4	4.600	3.100	1.500	0.200	Setosa
5	5	3.600	1.400	0.200	Setosa
6	5.400	3.900	1.700	0.400	Setosa
7	4.600	3.400	1.400	0.300	Setosa
8	5	3.400	1.500	0.200	Setosa
9	4.400	2.900	1.400	0.200	Setosa
10	4.900	3.100	1.500	0.100	Setosa
11	5.400	3.700	1.500	0.200	Setosa
12	4.800	3.400	1.600	0.200	Setosa
13	4.800	3	1.400	0.100	Setosa
14	4.300	3	1.100	0.100	Setosa
15	5.800	4	1.200	0.200	Setosa
16	5.700	4.400	1.500	0.400	Setosa

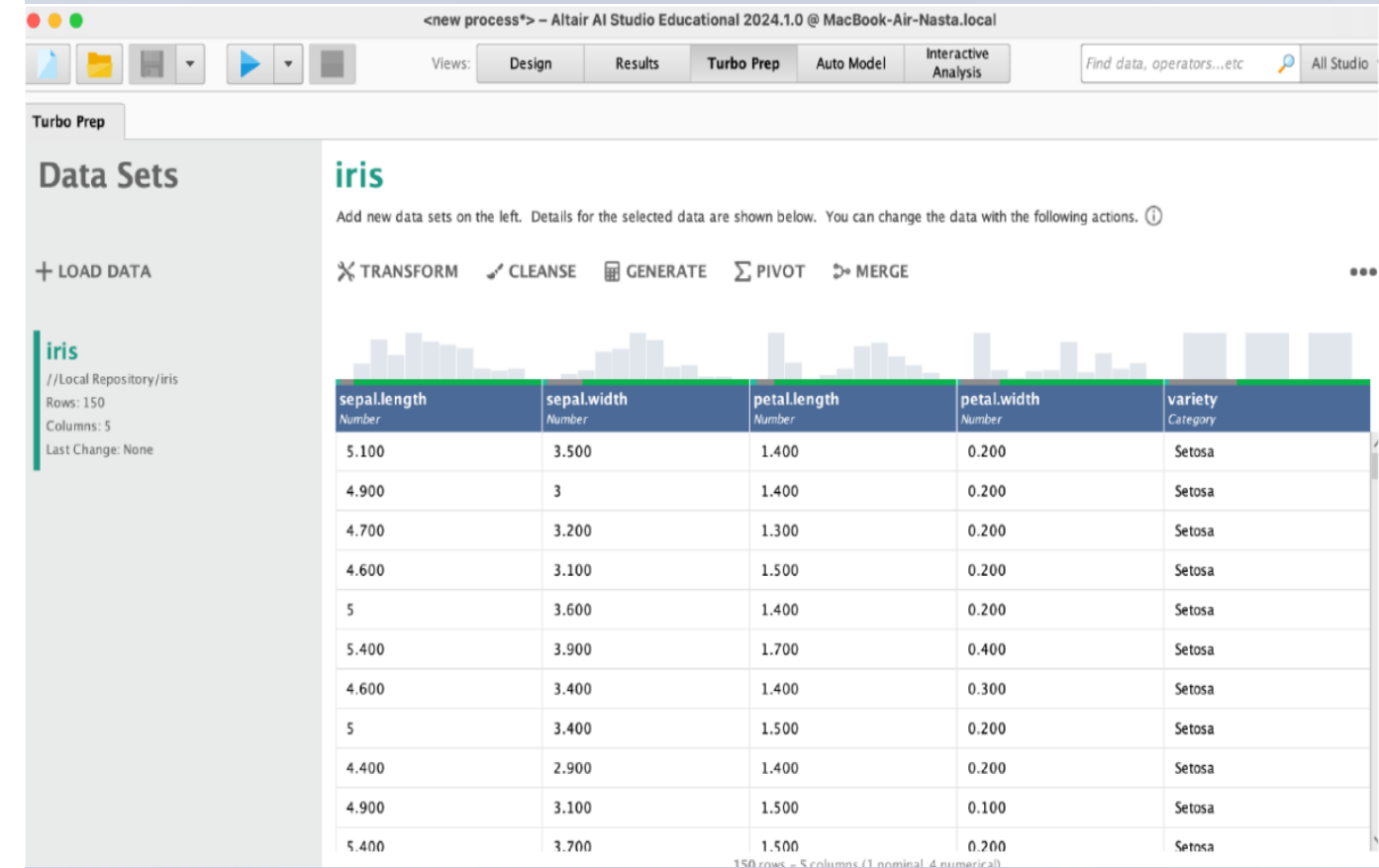
ExampleSet (150 examples, 0 special attributes, 5 regular attributes)

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.



Auto Cleansing

Define Target Improve Quality Change Types Handle Numbers Summary

AI Studio can automatically perform common data cleansing techniques on your data to better prepare it for machine learning. In case you want to predict a column later on, please select it below.

No target column, thanks!

sepal.length Number	sepal.width Number	petal.length Number	petal.width Number	variety Category
5.100	3.500	1.400	0.200	Setosa
4.900	3	1.400	0.200	Setosa
4.700	3.200	1.300	0.200	Setosa
4.600	3.100	1.500	0.200	Setosa
5	3.600	1.400	0.200	Setosa
5.400	3.900	1.700	0.400	Setosa
4.600	3.400	1.400	0.300	Setosa
5	3.400	1.500	0.200	Setosa
4.400	2.900	1.400	0.200	Setosa
4.900	3.100	1.500	0.100	Setosa

< BACK > NEXT



Auto Cleansing

Define Target Improve Quality Change Types Handle Numbers Summary

Finally, AI Studio offers two choices to potentially improve the quality of numerical columns. Principal Component Analysis is a way to reduce the number of columns by mapping the data points into a new space. Normalization is often useful to bring all columns to roughly the same scale. Again, if you are not sure about this, just leave the settings as they are.

☐ Perform PCA
☒ Perform normalization

Information: Normalization is a common technique which ensures that all numeric columns of your data set are roughly on the same scale. Each column is rescaled so that the average of the resulting column is 0 and the standard deviation for all columns becomes 1. By doing this, different scales won't impact machine learning models which is in particular important for distance-based methods. However, the resulting models are somewhat harder to interpret since the scales have changes to something which does not occur in reality. If you use Auto Model, it is usually better to let Auto Model do the normalizations only when they are necessary.

[< BACK](#) [NEXT >](#)

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

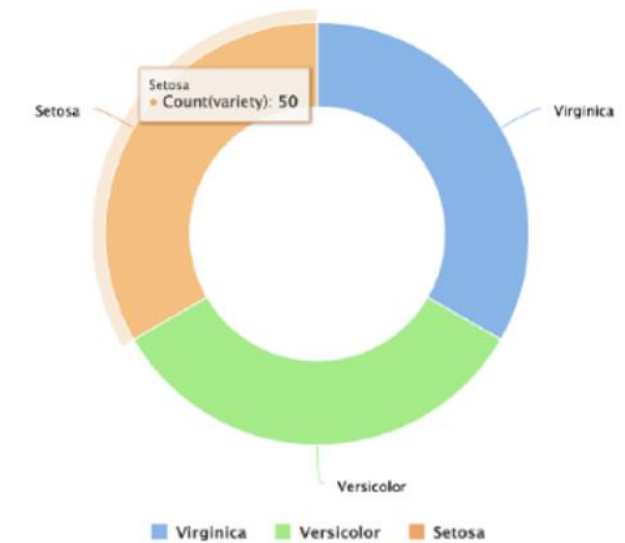
В RapidMiner доступны различные визуализации: круговые диаграммы (доли категорий), диаграммы рассеяния (зависимости признаков) и гистограммы (распределение значений признаков).



iris

Configure the desired chart with the settings on the left. ⓘ

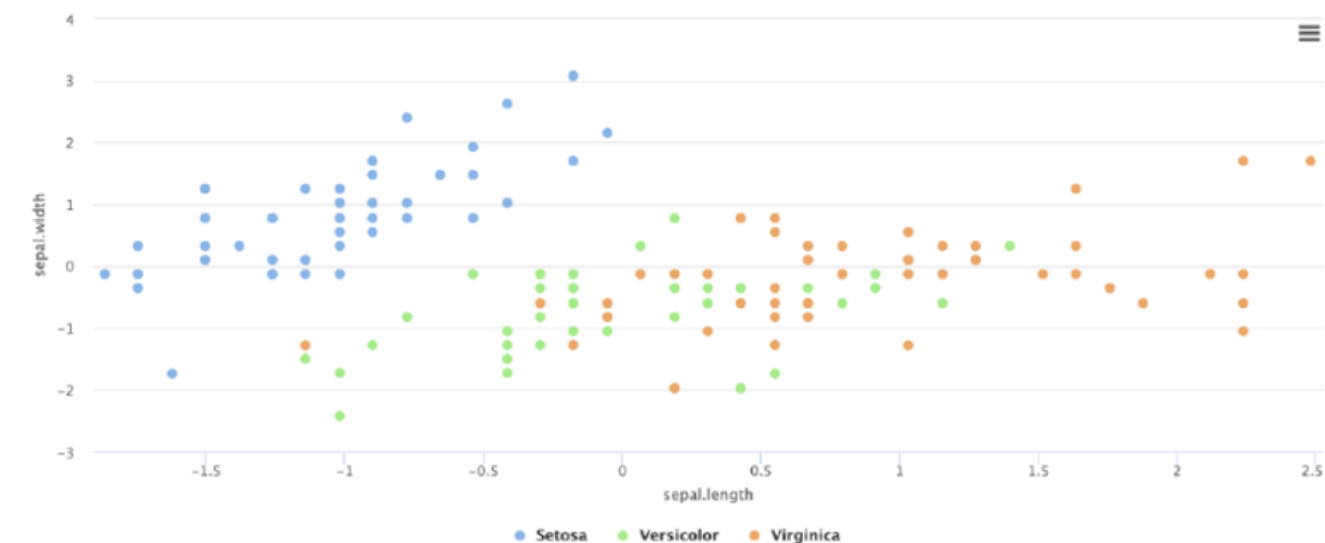
CANCEL

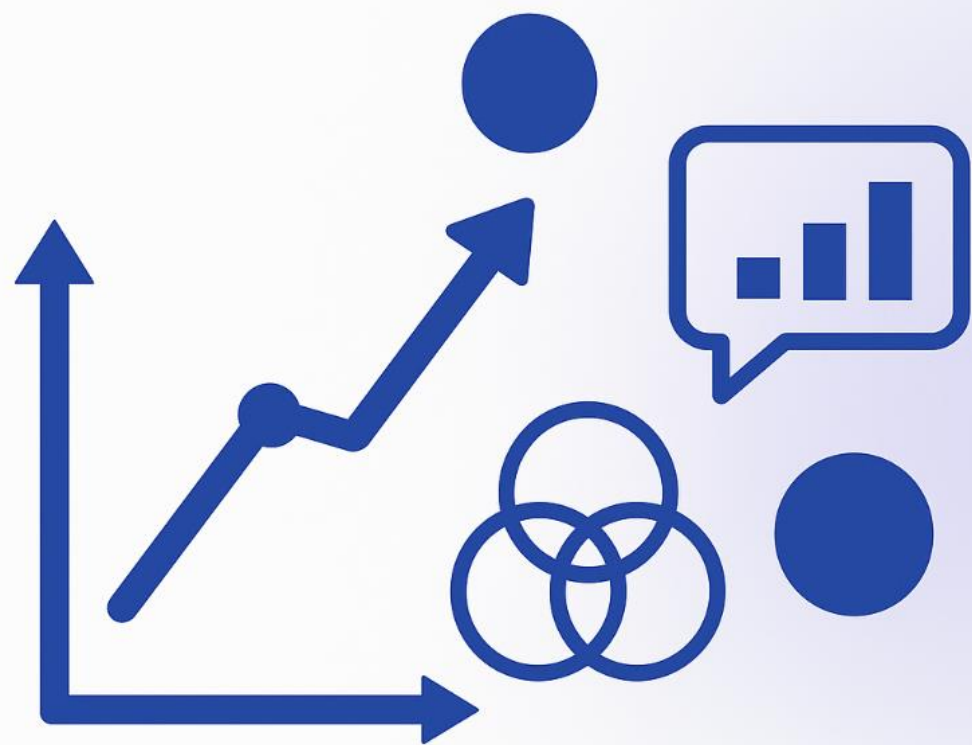


iris

Configure the desired chart with the settings on the left. ⓘ

CANCEL





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.



Auto Model



Models

- ☒ Naive Bayes
- ☐ Generalized Linear Model
 - ☒ Use Regularization ☐ Calculate p-Values
- ☒ Logistic Regression
- ☐ Fast Large Margin
 - ☒ Automatically Optimize
- ☒ Deep Learning
- ☒ Decision Tree
 - ☒ Automatically Optimize Maximal Depth: 20
- ☒ Random Forest

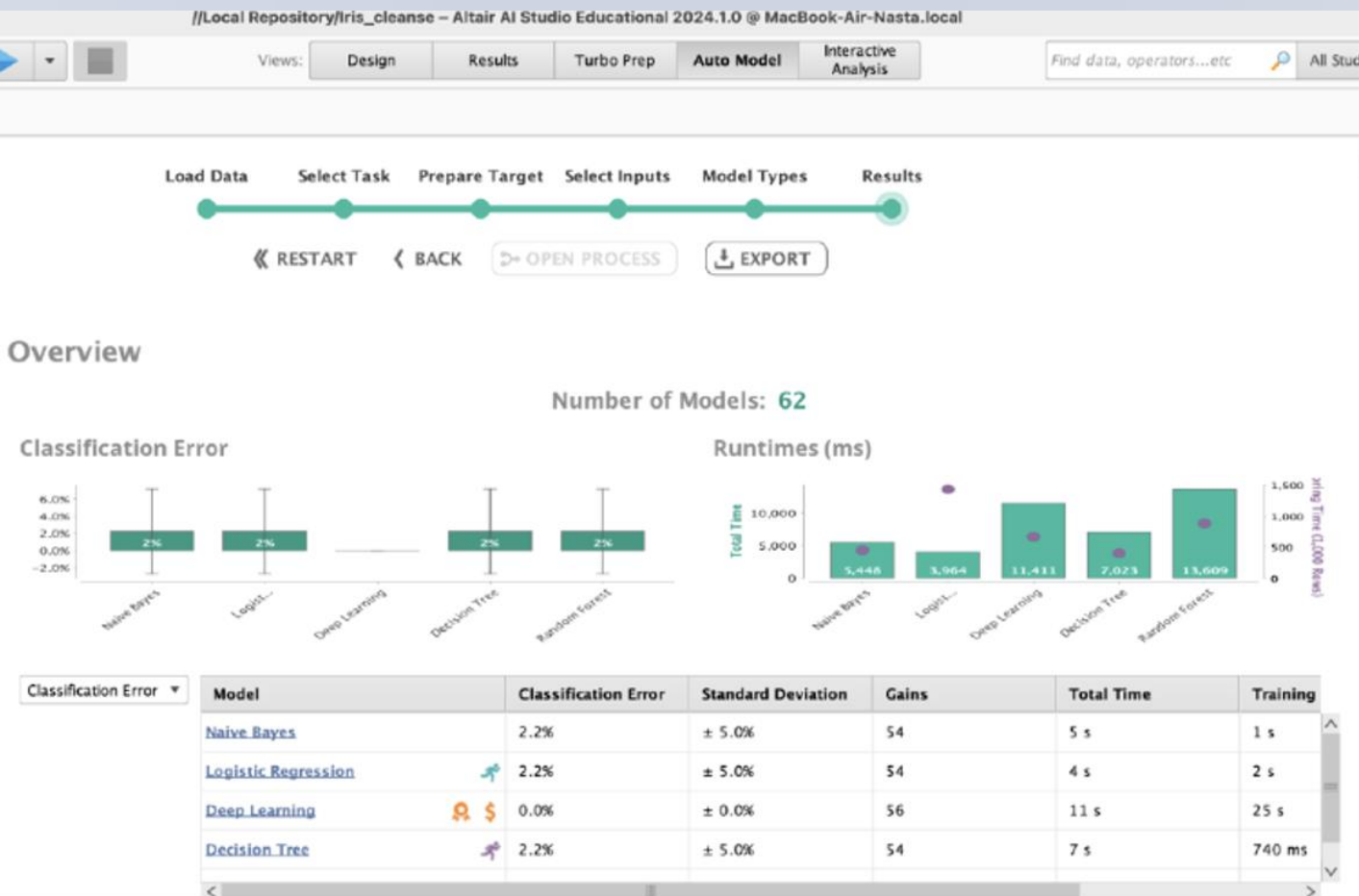
Data Preparation

- ☐ Remove Columns with Too Many Values
 - Maximum Number of Values: 50
- ☐ Extract Date Information
- ☐ Extract Text Information
 - Select Text Columns (0)...
 - Number of Extracted Features: 1,000
- ☐ Automatic Feature Selection
 - Additional Minutes (Maximum): 60
 - Final Feature Set should be: Accurate
- ☐ Automatic Feature Generation



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.

Attribute	Weight
petal.width	0.112
petal.length	0.096
sepal.width	0.060
sepal.length	0.010



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.