



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

Кафедра информационной
безопасности киберфизических
систем

Москва 2025

Визуализации в RapidMiner

Введение в визуализацию данных

Визуализация данных – графическое представление информации, позволяющее быстро выявлять закономерности и тенденции, которые трудно обнаружить, анализируя таблицы или числовые данные напрямую.





Значение визуализации при анализе данных

Грамотная визуализация значительно упрощает интерпретацию результатов. Визуальное представление данных помогает выявить зависимости между признаками, определить структуру набора данных и принять верные решения в дальнейшем анализе.



Возможности RapidMiner в области визуализации

RapidMiner – это среда анализа данных, предлагающая инструменты для визуализации. Пользователю доступны различные типы диаграмм: гистограммы, диаграммы рассеяния, столбчатые диаграммы и тепловые карты.





Turbo Prep

Cleanse

1 column selected

AUTO CLEANSING

REMOVE LOW QUALITY

REMOVE CORRELATED

REPLACE MISSING

NORMALIZATION

DISCRETIZATION

DUMMY ENCODING

PCA

REMOVE DUPLICATES

Titanic-Dataset

Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). ...

COMMIT CLEANSE

CANCEL

UNDO

SHOW HISTORY

PassengerId Number	Survived Number	Pclass Number	Name Category	Sex Category	Age Number	SibSp Number	Parch Number	Ticket Category	Fare Number
1	0	3	Braund, Mr. O...	male	22	1	0	A/5 21171	7.250
2	1	1	Cumings, Mrs....	female	38	1	0	PC 17599	71.283
3	1	3	Heikkinen, Mi...	female	26	0	0	STON/O2. 31...	7.925
4	1	1	Futrelle, Mrs. ...	female	35	1	0	113803	53.100
5	0	3	Allen, Mr. Willi...	male	35	0	0	373450	8.050
6	0	3	Moran, Mr. Ja...	male	?	0	0	330877	8.458
7	0	1	McCarthy, Mr....	male	54	0	0	17463	51.862
8	0	3	Palsson, Mast...	male	2	3	1	349909	21.075
9	1	3	Johnson, Mrs. ...	female	27	0	2	347742	11.133
10	1	2	Nasser, Mrs. ...	female	14	1	0	237736	30.071
11	1	3	Sandstrom, Mi...	female	4	1	1	PP 9549	16.700

891 rows - 12 columns (5 nominal, 7 numerical)

Turbo Prep

Transform

1 column selected

REMOVE

COPY

FILTER

RANGE

SAMPLE

SORT

REPLACE

male

o|

Use regular expressions

APPLY

Titanic-Dataset

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns).

COMMIT TRANSFORMATION

CANCEL

PassengerId Number	Survived Number	Pclass Number	Name Category	Sex Category	Age Number
1	0	3	Braund, Mr. O...	male	22
2	1	1	Cumings, Mrs....	1	38
3	1	3	Heikkinen, Mi...	1	26
4	1	1	Futrelle, Mrs. ...	1	35
5	0	3	Allen, Mr. Willi...	male	35
6	0	3	Moran, Mr. Ja...	male	29.699
7	0	1	McCarthy, Mr....	male	54
8	0	3	Palsson, Mast...	male	2
9	1	3	Johnson, Mrs. ...	1	27
10	1	2	Nasser, Mrs. ...	1	14
11	1	3	Sandstrom, Mi...	1	4

891 rows - 12 columns (5 nominal, 7 numerical)

Этапы работы с данными в RM

Перед визуализацией данные проходят предварительную обработку: очистку от пропусков, нормализацию признаков и кодирование категориальных переменных. После этого происходит построение моделей и анализ результатов.

Особенности набора данных Titanic

Набор данных Titanic популярен благодаря сочетанию категориальных и числовых признаков: возраст, пол, класс обслуживания, стоимость билета. Это позволяет изучать разнообразные подходы визуального анализа и моделирования.





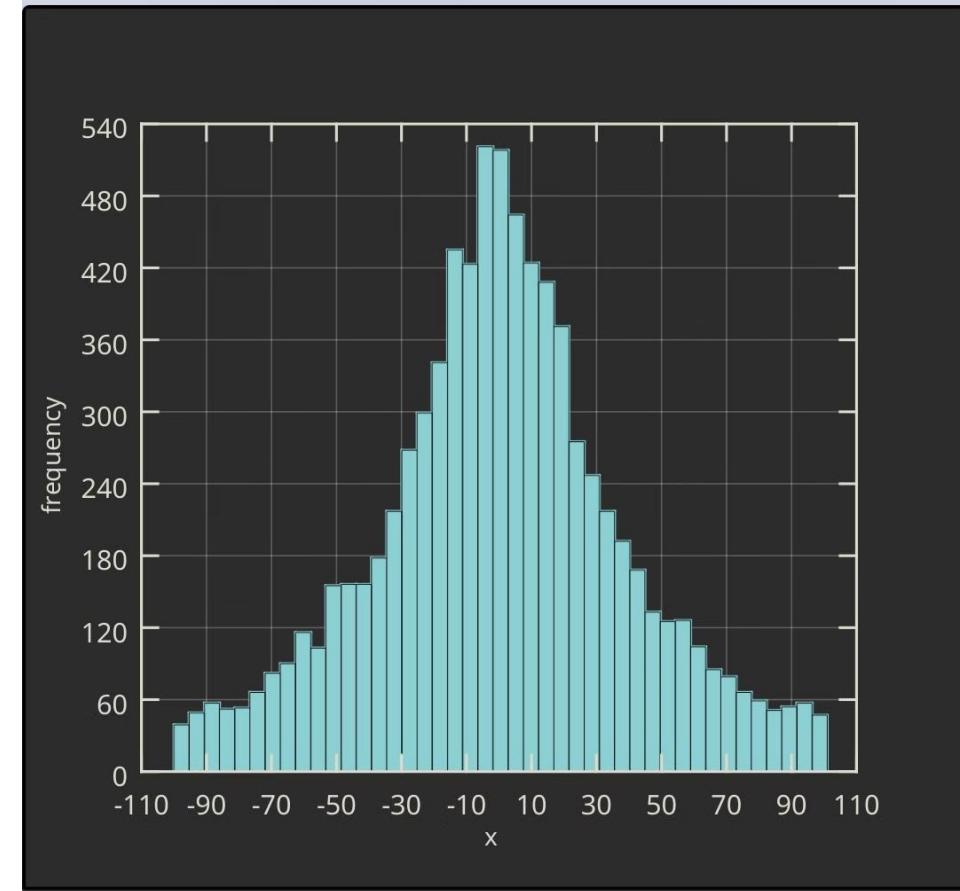
Основные ошибки при визуализации данных

Распространённые ошибки включают игнорирование пропусков и выбросов, использование неподходящих диаграмм, отсутствие подписей осей и неправильный выбор масштабов, что существенно затрудняет интерпретацию графиков.



Гистограмма для анализа числовых данных

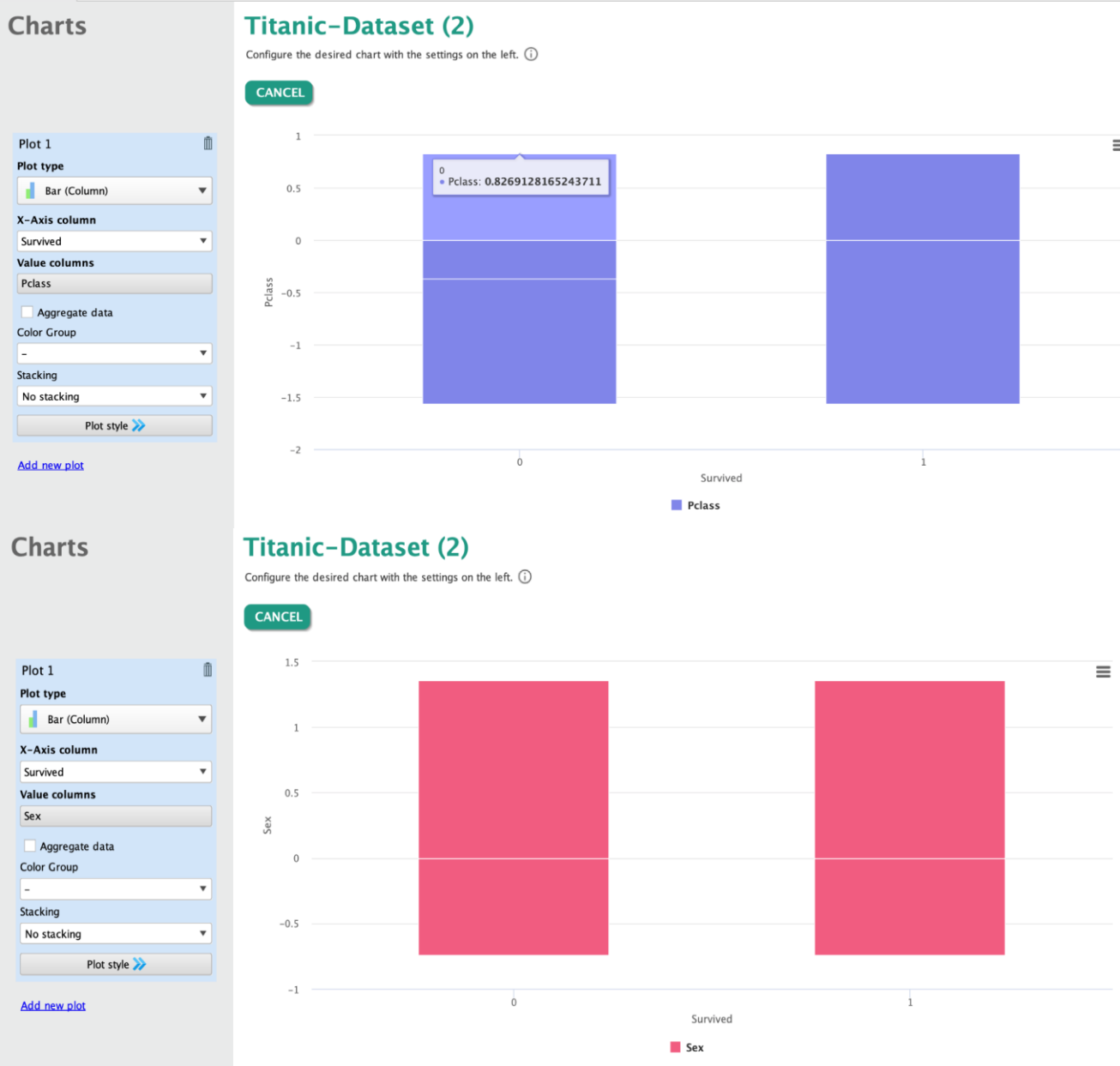
Гистограмма позволяет визуализировать распределение числовых признаков, таких как возраст пассажиров. С её помощью легко определить частоту попадания данных в определенные интервалы и выявить выбросы.





Столбчатая диаграмма (Bar Chart)

Столбчатая диаграмма эффективно отображает различия между группами категориальных признаков. Например, легко сравнить выживаемость пассажиров различных классов или полов по относительному количеству выживших.



Особенности и применение круговых диаграмм (Pie Chart)

Круговая диаграмма предназначена для представления долей категорий в общем объёме. Подходит для небольшого количества категорий, демонстрируя соотношение, например, пассажиров разных классов.



Charts

Titanic-Dataset (2)

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

CANCEL

Plot 1

Plot type
Pie / Donut

Value column
Survived

☒ Aggregate data

Group by
Pclass

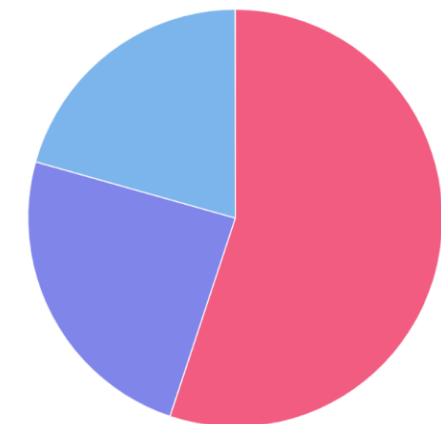
Aggregation Function
Count

Sorting
None

☐ Donut

Plot style >>

[Add new plot](#)

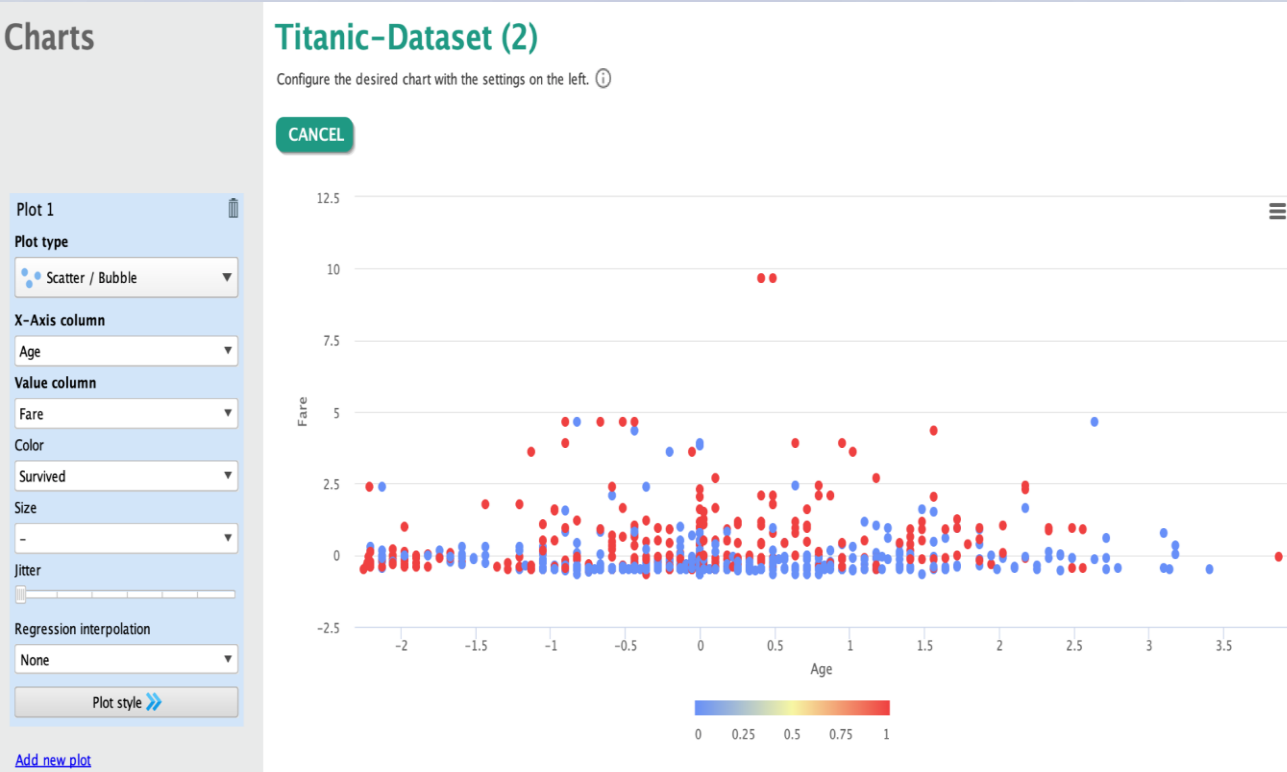


0.827 -1.565 -0.369



Диаграмма рассеяния для выявления корреляций

Диаграмма рассеяния позволяет определить взаимосвязь между двумя числовыми признаками, такими как возраст и стоимость билета, и понять, влияют ли они на выживаемость пассажиров.



Box Plot и его ВОЗМОЖНОСТИ

Box Plot («ящик с усами») визуализирует медиану, квартильный разброс и выбросы числовых данных. Позволяет сравнить, например, распределение стоимости билетов по портам отправления пассажиров.





Charts

Titanic-Dataset (2)

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

CANCEL

Plot 1 

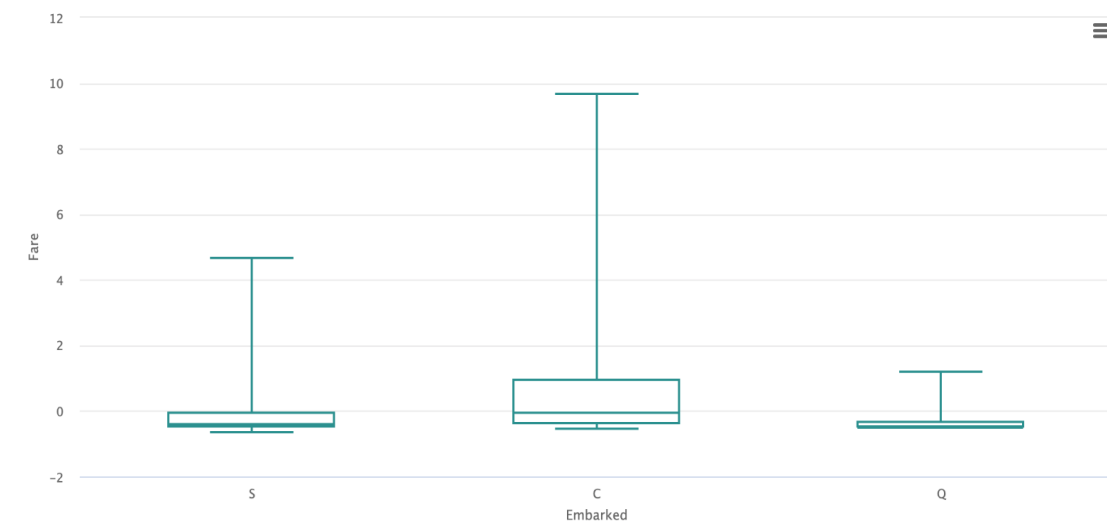
Plot type
 Boxplot ▼

Value column
Fare ▼

Group by
Embarked ▼

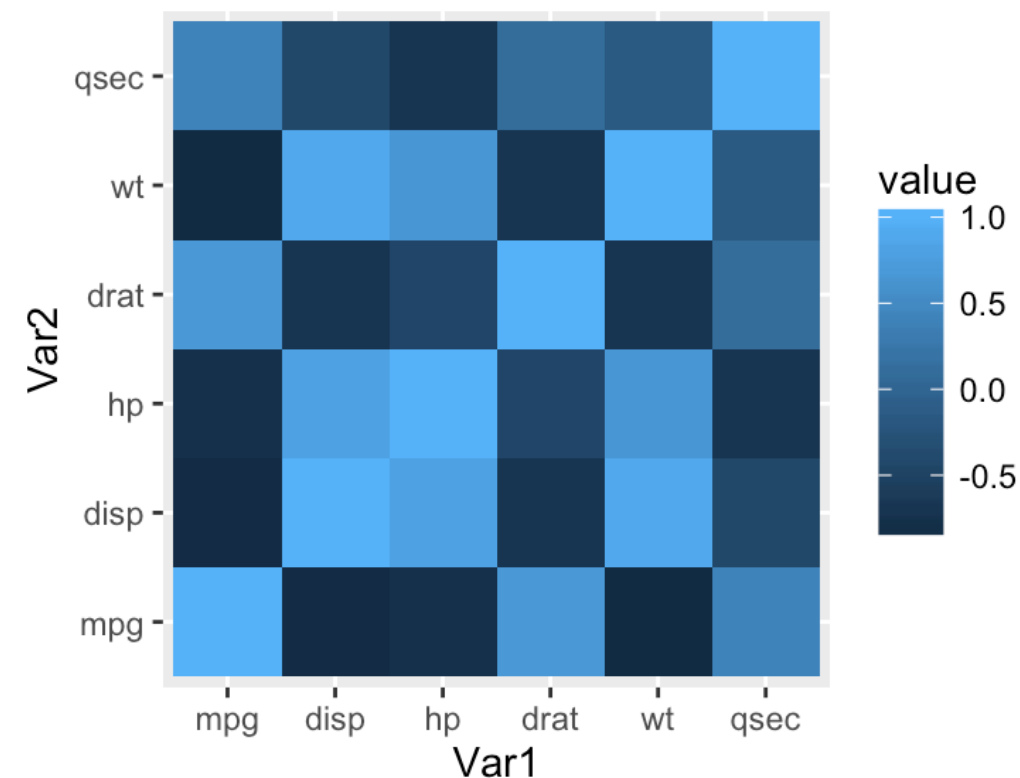
Plot style >>

[Add new plot](#)





Тепловая карта корреляций (Heatmap)

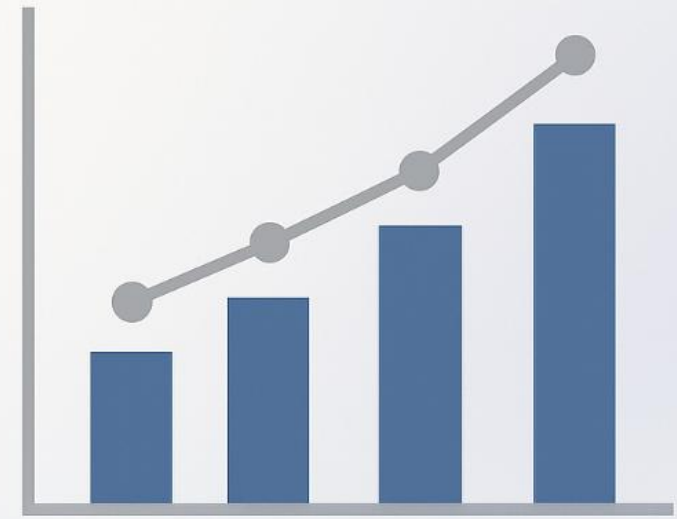


Тепловая карта наглядно показывает корреляцию между несколькими признаками одновременно. Цветовая шкала отображает силу и направление связи признаков, что помогает при отборе переменных для модели.

Как корректно настроить визуализацию



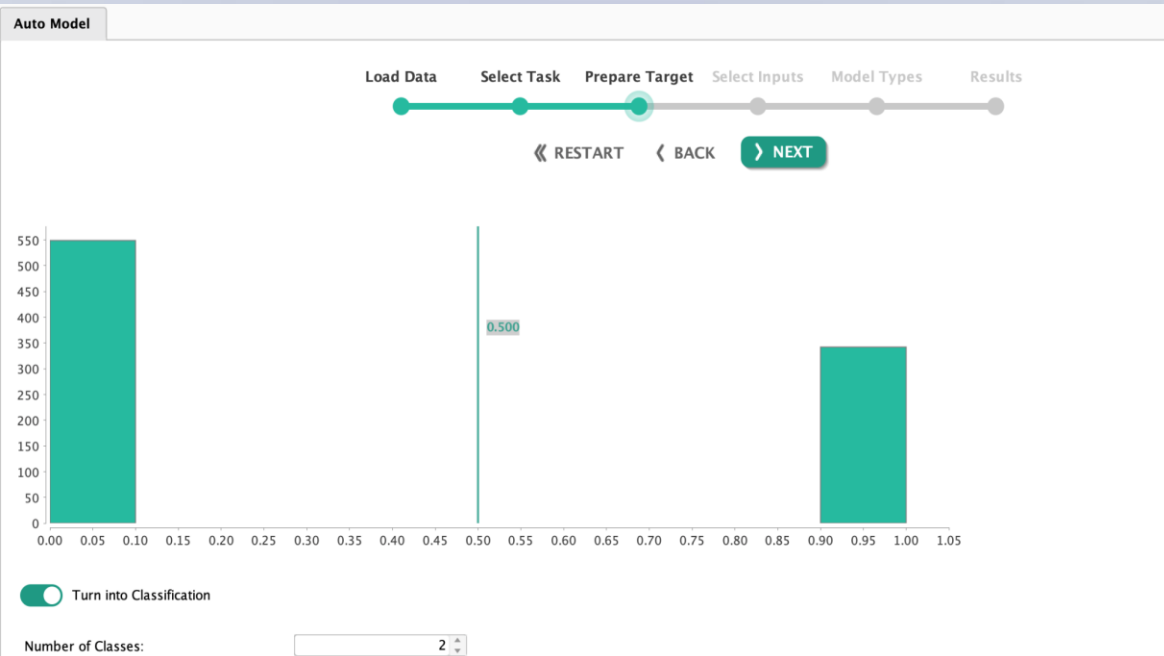
Для повышения информативности графиков необходимо использовать подписанные оси, понятные легенды и подходящую цветовую гамму. Это значительно улучшает читаемость и понимание данных.





Инструмент AutoModel в RapidMiner

AutoModel автоматизирует создание нескольких моделей одновременно. Он также визуализирует их результаты, облегчая анализ и выбор наилучшей модели по показателям качества классификации или регрессии.

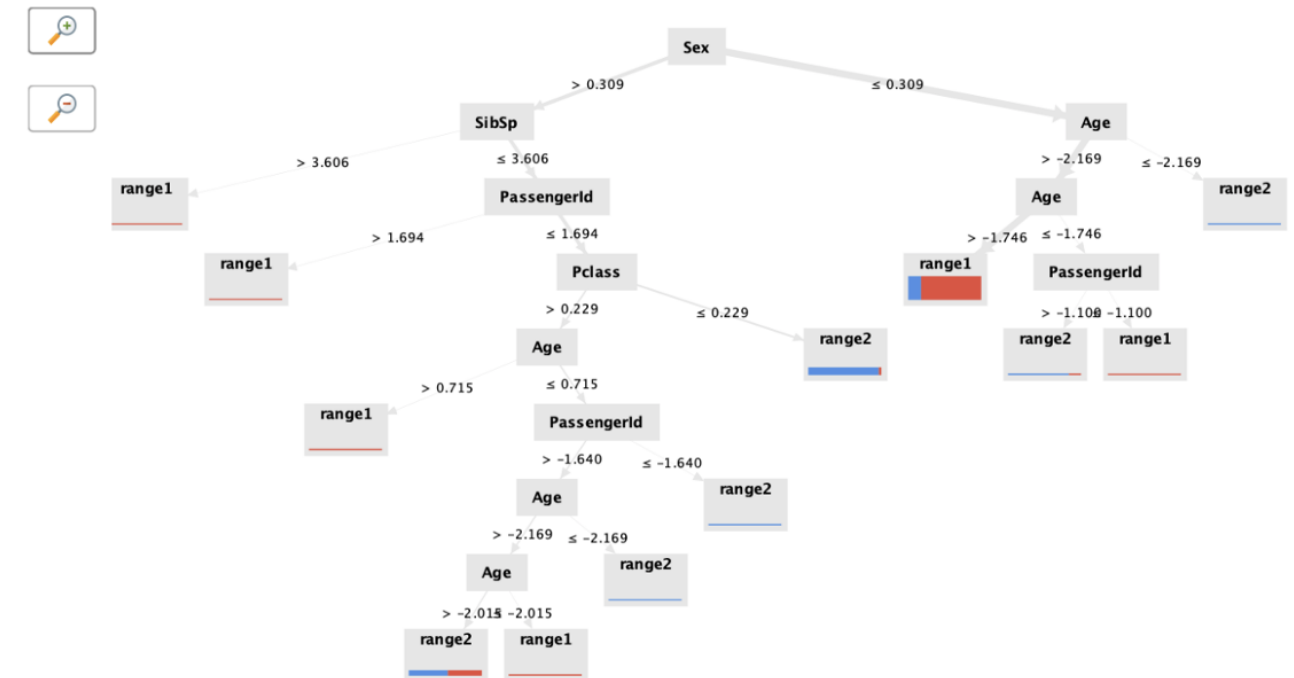


Визуализация модели дерева решений



Дерево решений – понятная модель благодаря графическому представлению, отображающему логику принятия решений, последовательность проверок признаков и их относительную значимость для итогового прогноза.

Decision Tree – Model





Примеры удачных и неудачных визуализаций



Удачные визуализации чётко демонстрируют структуру и закономерности данных. Неудачные же перегружены лишними деталями, не имеют ясных подписей и затрудняют восприятие информации.

Как избежать ошибок при визуализации

Избежать типичных ошибок можно, предварительно определив цель анализа, правильно выбрав тип графика, корректно настроив оси и цвета, и удалив шумы и выбросы из данных.





Значение визуализации в предварительном анализе данных



На предварительном этапе визуализация помогает сформировать гипотезы о зависимости признаков, выявить потенциальные проблемы в данных и спланировать дальнейшие действия по анализу.

Роль визуализации после построения модели

После построения модели визуализация используется для оценки её качества и интерпретации результатов. Например, графики метрик помогают выявить сильные и слабые стороны модели.





Дополнительные возможности визуализации в RapidMiner



Помимо основных графиков, в RapidMiner есть интерактивные функции, позволяющие быстро изменять настройки визуализации: фильтрация данных, выбор подмножеств и динамическое отображение результатов.

Визуальный анализ как основа принятия решений

Визуальный анализ данных позволяет принимать более обоснованные и быстрые решения. Наглядные графики и диаграммы упрощают процесс интерпретации сложной информации.

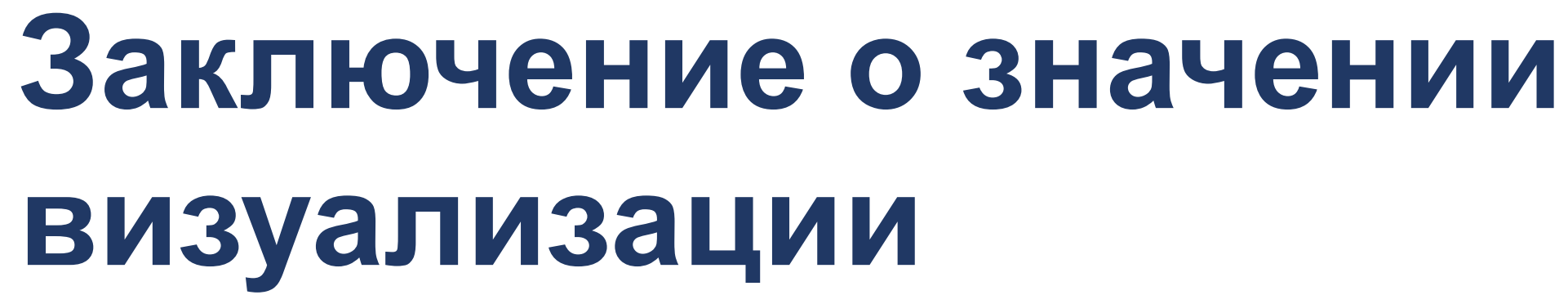




Практическая польза визуализации (пример Titanic)

Анализ данных Titanic визуально подтверждает исторические факты: пассажиры первого класса и женщины имели гораздо более высокие шансы на выживание, чем мужчины и пассажиры третьего класса.



[illegible]