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 № 3

ТЕМА РАБОТЫ
«Визуализации в RapidMiner»

Москва, 2025

Практическая работа 3: Визуализации в RapidMiner

Цель работы

Познакомиться с расширенными возможностями визуализации данных в RapidMiner. Студенты научатся создавать различные типы диаграмм (гистограммы, круговые, диаграммы рассеяния, пузырьковые, box-plot), настраивать их внешний вид, а также использовать визуализацию для анализа результатов моделей. Главная задача – понять, как визуальный

анализ помогает глубже понять структуру данных и качество построенных моделей машинного обучения.

Целевая аудитория

Студенты курса по информатике и анализу данных, имеющие базовые знания о работе с RapidMiner и предварительной обработке данных. Эта работа подходит для тех, кто уже знаком с основами загрузки и очистки данных (из предыдущих работ) и стремится освоить методы визуального анализа результатов.

Идея и концепция

Для демонстрации возможностей визуализации используется набор данных «Titanic», содержащий информацию о пассажирах «Титаника». Этот набор данных широко используется благодаря сочетанию числовых и категориальных признаков. В нем представлены такие столбцы, как возраст пассажира, пол, класс билета, стоимость билета, порт посадки, а также информация о том, выжил ли пассажир.

Используя данный набор данных, студенты смогут:

- Визуально исследовать распределения признаков: возраст, стоимость билета (Fare), классы обслуживания.
- Изучать соотношение категориальных признаков (пол, класс, порт посадки) с целевой переменной (выжил/не выжил).
- Оценить корреляции между численными признаками и понять, какие факторы, возможно, повлияли на выживаемость.
- После предварительного анализа данных будет продемонстрирован процесс использования AutoModel для построения модели классификации (предсказания выживаемости) с последующей визуальной оценкой её результатов.

Содержание практической работы

Введение в RapidMiner

Краткое напоминание, что RapidMiner – это инструмент для анализа данных, включающий средства предобработки, визуализации и автоматизированного построения моделей.

- При необходимости установить RapidMiner Studio.
- Ознакомиться с панелью инструментов и вкладками «Design», «Results», Turbo Prep», «Auto Model».
- Убедиться в наличии файла с данными Titanic (csv-файл, доступный, например, на Kaggle или в локальном хранилище).

О наборе данных и задаче работы

Датасет «Titanic» содержит следующие столбцы (атрибуты):

- Survived: бинарный признак (0 – не выжил, 1 – выжил).
- Pclass: класс пассажира (1 – первый класс, 2 – второй класс, 3 – третий класс).
- Name: имя пассажира (текстовый признак, можно не использовать для визуализаций).
- Sex: пол пассажира (male/female).
- Age: возраст (числовой признак).
- SibSp: количество братьев/сестер/супругов на борту.
- Parch: количество родителей/детей на борту.
- Ticket: номер билета (текстовый, можно не использовать для визуализаций).
- Fare: стоимость билета (числовой признак).
- Cabin: номер каюты (может содержать много пропусков).
- Embarked: порт посадки (C – Cherbourg, Q – Queenstown, S – Southampton).

Задача: с помощью визуализации понять структуру данных Titanic, увидеть различия между группами пассажиров и их влияние на выживаемость. После этого построить модель машинного обучения (классификацию), оценить ее качество и интерпретировать результаты визуально.

Работа с данными

Загрузка набора данных

- Откройте RapidMiner Studio.
- Нажмите "Create New Process".
- Перетащите оператор **"Read CSV"** на рабочее пространство.
- Загрузите файл titanic.csv (предварительно скачанный).

- Подключите выход **Read CSV** к **Result** и нажмите **Run**, чтобы просмотреть данные.

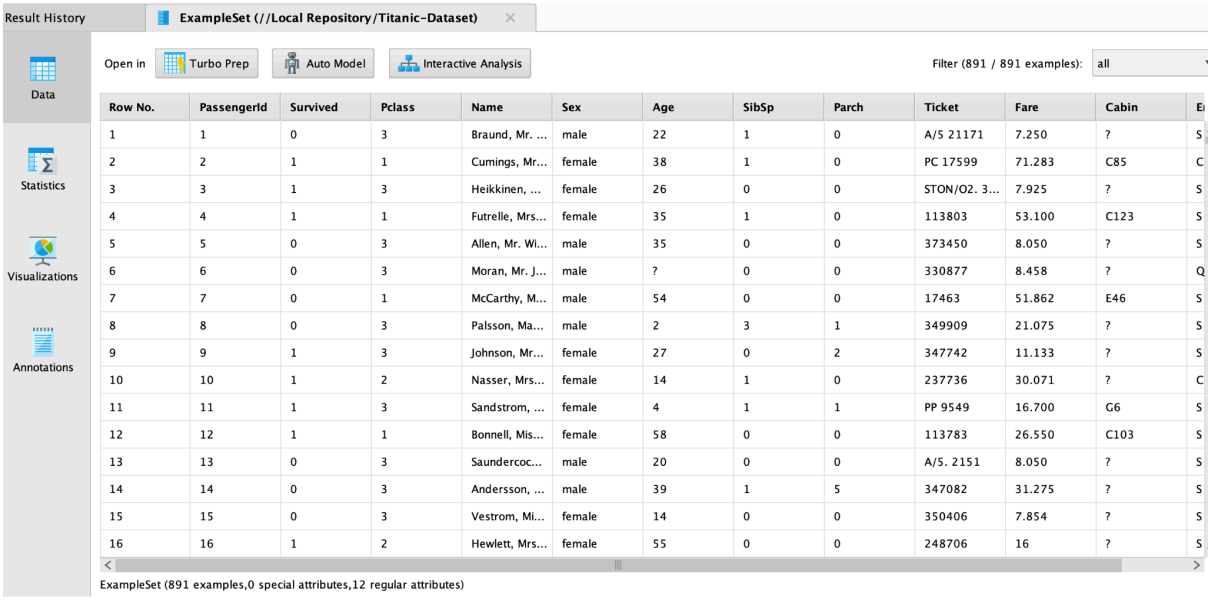

рис. 1: Просмотр датасета после загрузки

Использование TurboPrep для подготовки данных

- Нажмите кнопку **"Turbo Prep"** в верхней панели инструментов.
- Загрузите набор данных Titanic.
- Проверьте данные на пропуски (особенно в столбцах Age, Cabin).
- При необходимости используйте функции **Cleanse** для замены пропусков в Age на среднее значение.

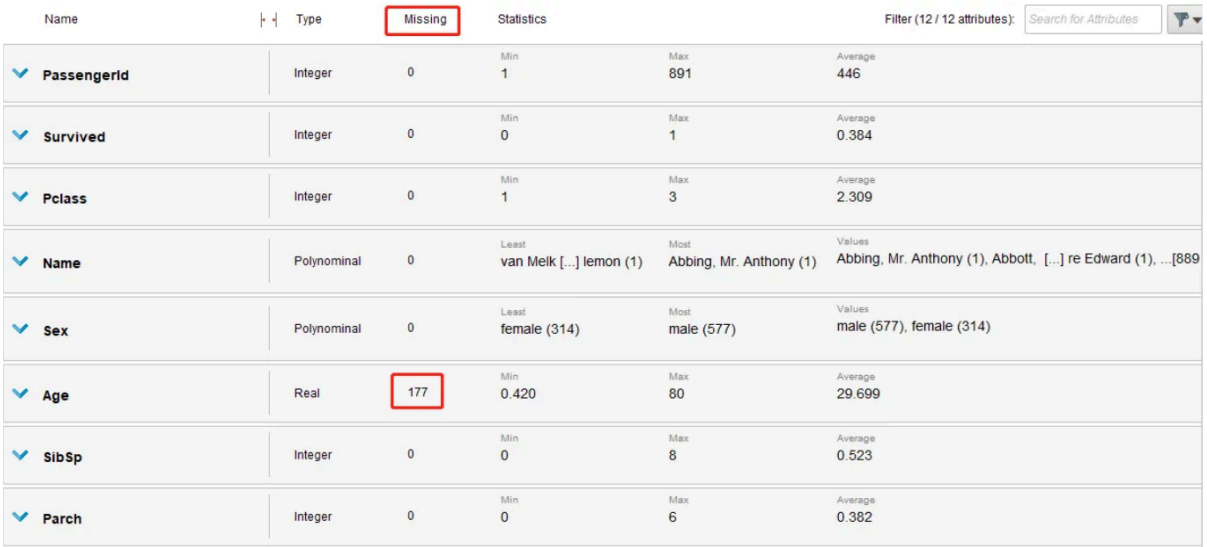

рис. 2: Просмотр информации о датасете, включающей ключевые статистические показатели

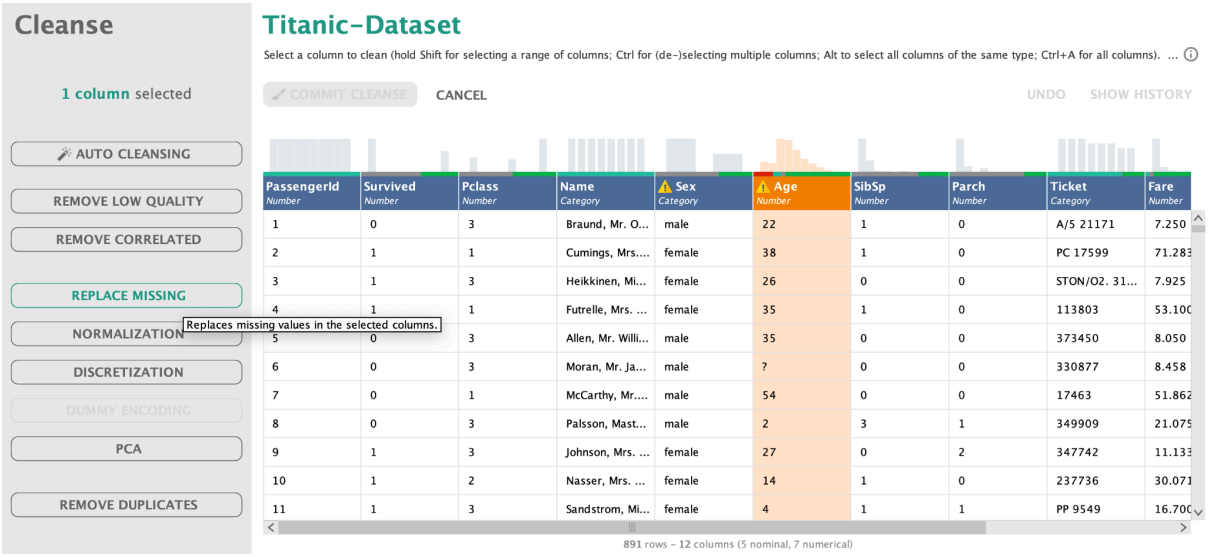

рис. 3: Заполнение пропусков в столбцах с помощью "Replace Missing"

- Вы увидите, что признак «Пол» (Sex) не является числовым типом данных, поэтому мы не сможем использовать этот признак в модели. Для решения этой проблемы стоит использовать Replace. Применим бинарное кодирование, заменив female (женский пол) на 1, а male (мужской пол) на 0.

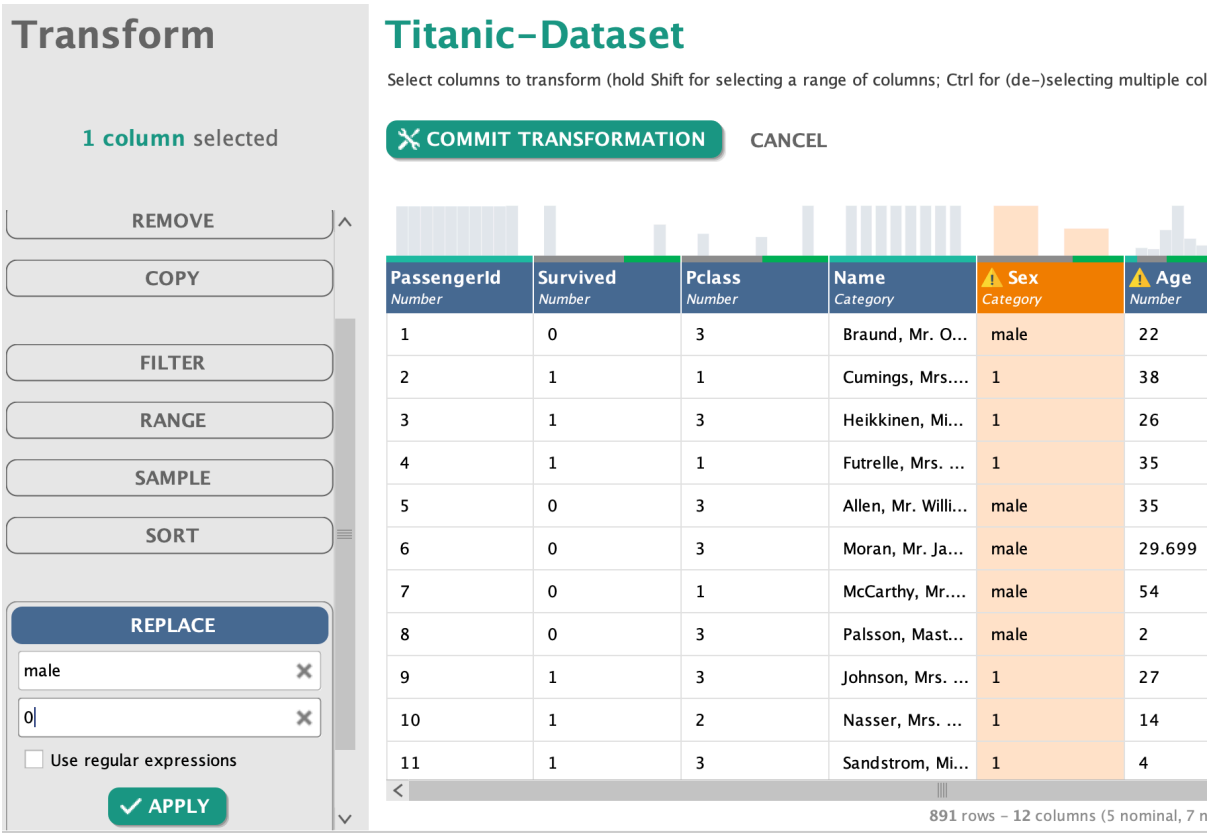

рис. 4: Использование оператора Replace для бинарного кодирования категориальных столбцов

- Обозначьте целевую переменную. В данном случае, мы предсказываем выживаемость, поэтому таргетом будет являться столбец «Survived». Делаем это с помощью оператора «Set role».

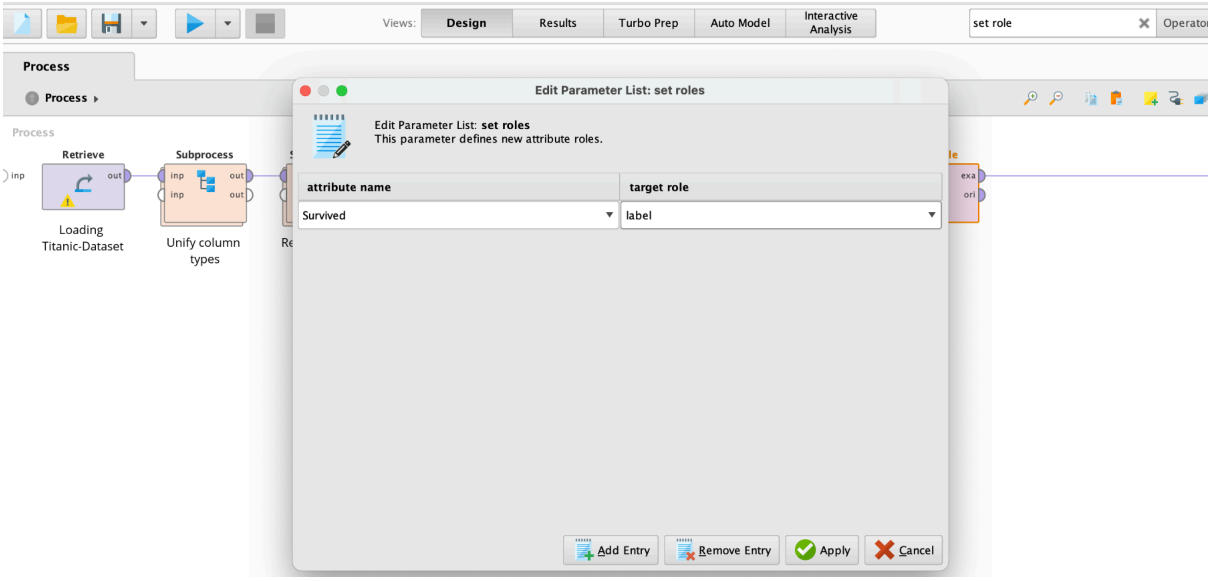

рис. 5: Выбор целевой переменной с помощью оператора "Set role«

- Нормализуйте числовые признаки для облегчения сравнения на графиках.

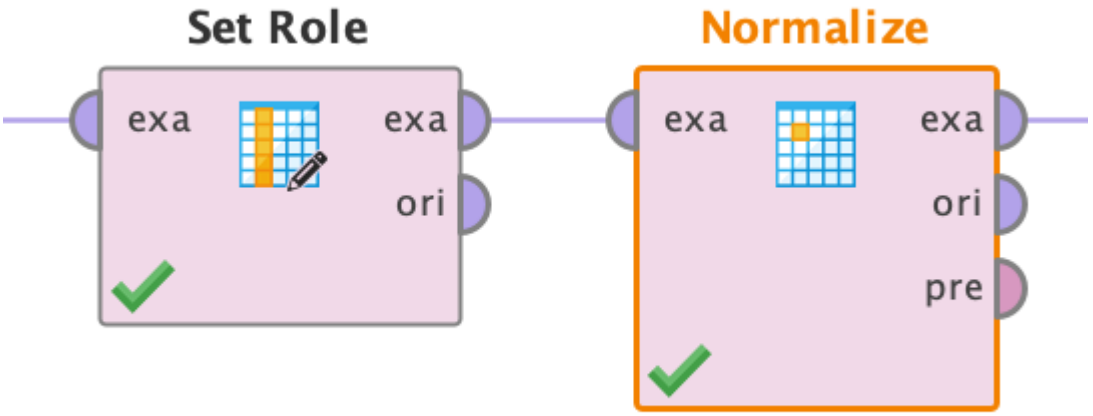

рис. 6: Добавление оператора "Normalize" в цепочку операторов для масштабирования значений и приведения к единой шкале

- Нажмите **"Commit Cleanse"**, чтобы применить изменения.

Использование Transform, Cleanse, Generate, Pivot, Merge

- **Transform**: при желании можно создать новый признак, например, категоризировать возраст по группам (молодые, взрослые, пожилые).
- **Cleanse**: уже применено для пропусков и нормализации.
- **Generate**: мсоздать новый признак, например, является ли возраст детским (< 16 лет).
- **Pivot, Merge**: для данного датасета не обязательны.

Визуализация данных

Откройте вкладку с подготовленными данными (TurboPrep → Charts).

Примеры визуализаций, которые необходимо создать

1. *Гистограмма (Histogram) для Age*

- Plot type: Histogram
- Value column: Age
- Настройка: измените количество корзин (bins) для более детального анализа распределения возрастов.
- Это позволит оценить распределение пассажиров по возрастным группам.

2. *Столбчатая диаграмма (Bar Chart) для Pclass или Sex*

- Plot type: Bar
- Value column: Pclass
- X-axis column: Survived
- Аналогично построить столбчатый график для Sex.
- Это покажет, какова зависимость выживаемости от типа каюты или пола.

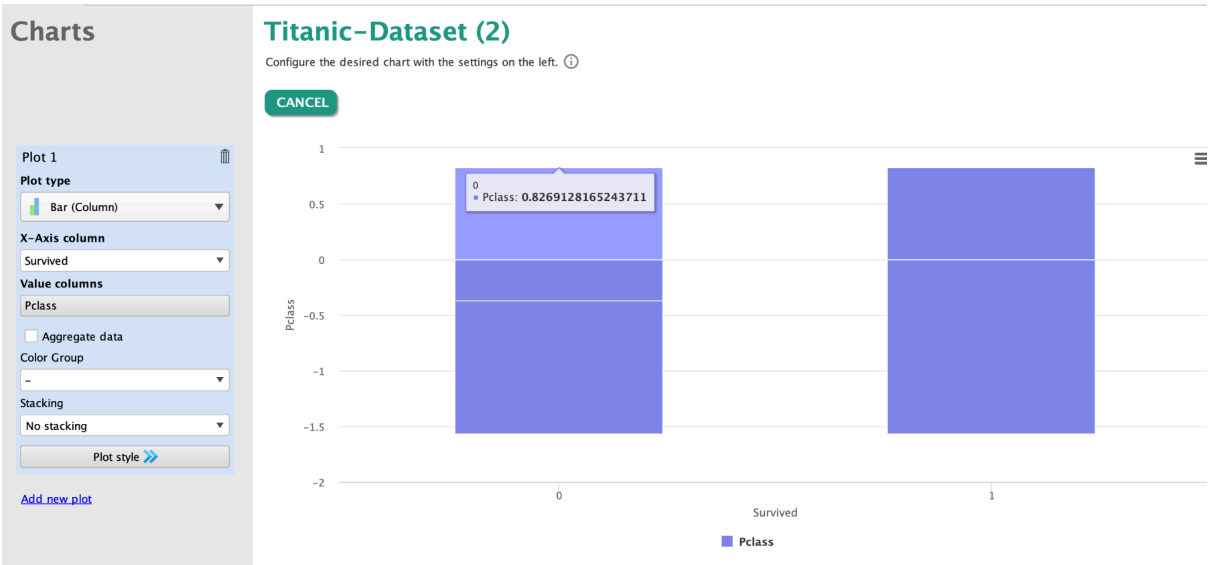

рис. 7: Построение столбчатой диаграммы "Зависимость выживаемости от типы каюта"

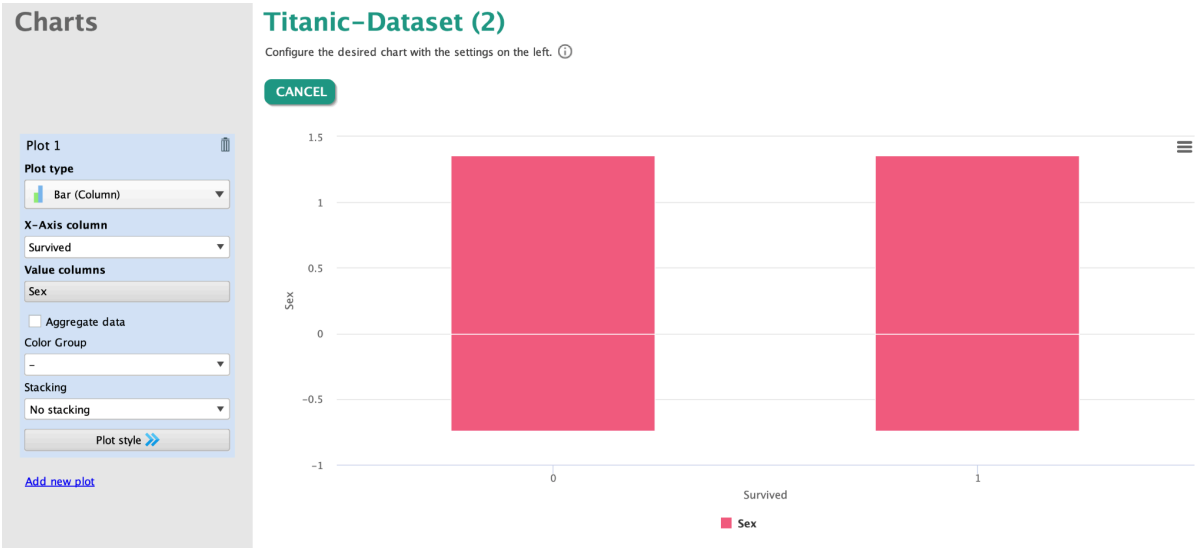

рис. 8: Построение столбчатой диаграммы "Зависимость выживаемости от типы от пола"

3. *Круговая диаграмма (Pie Chart) для Pclass*

- Plot type: Pie/Donut
- Value column: Survived
- Aggregation: Count
- Group by: Pclass
- Позволяет оценить долю выживших пассажиров в разных классах обслуживания (1-й, 2-й, 3-й классы). Цветовую схему можно изменить для лучшей наглядности. На графике будет видно, что выживаемость премиального класса выше.

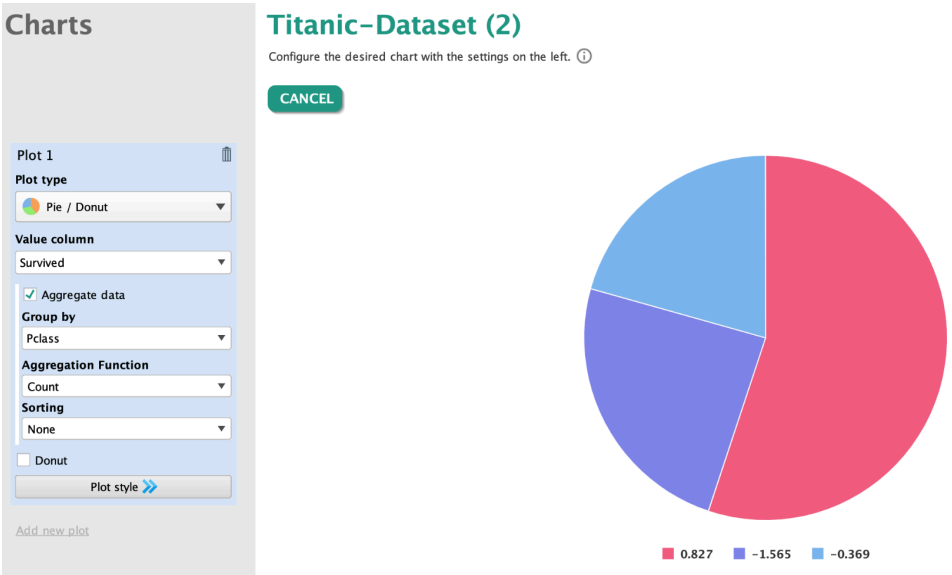

рис. 9: Построение круговой диаграммы "Доля выживших пассажиров в разных

классах обслуживания"

4. *Диаграмма рассеяния (Scatter Plot) для Age и Fare, цвет по Survived*

- Plot type: Scatter/Bubble
- X-Axis column: Age
- Value columns: Fare
- Color by: Survived
- Определите, помогают ли эти признаки различать выживших и погибших. Измените форму и размер точек для лучшего восприятия.

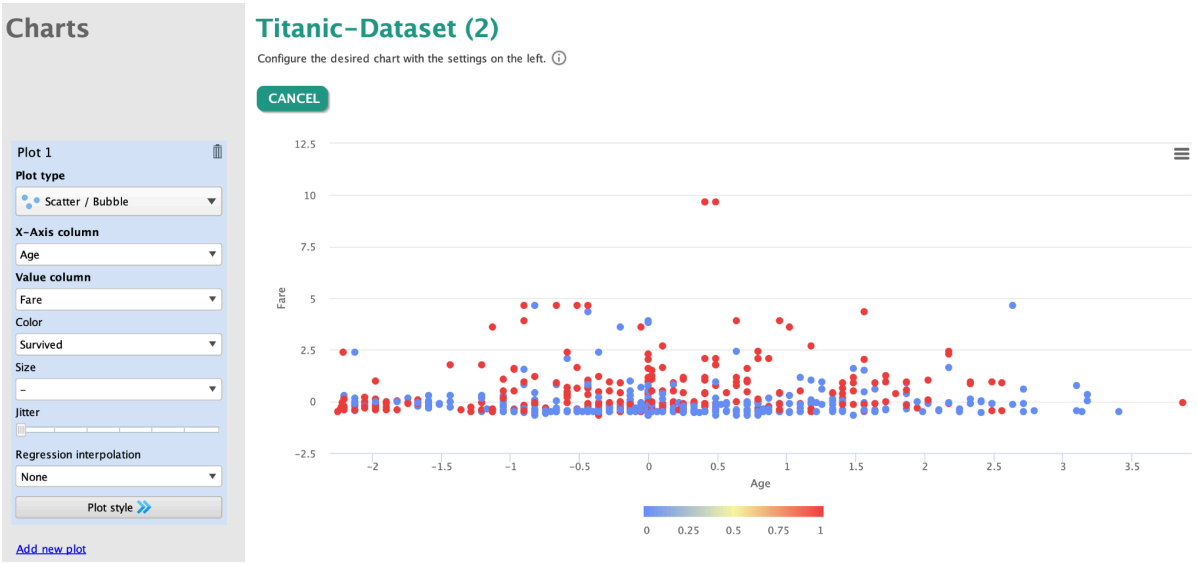

рис. 10: Построение диаграммы рассеяния

5. *Box Plot (ящик с усами) для Fare по Embarked*

- Plot type: Box
- Value column: Fare
- Group by: Embarked (порт отправления)
- Показывает, как стоимость билета распределена по портам отправления, есть ли сильные выбросы.

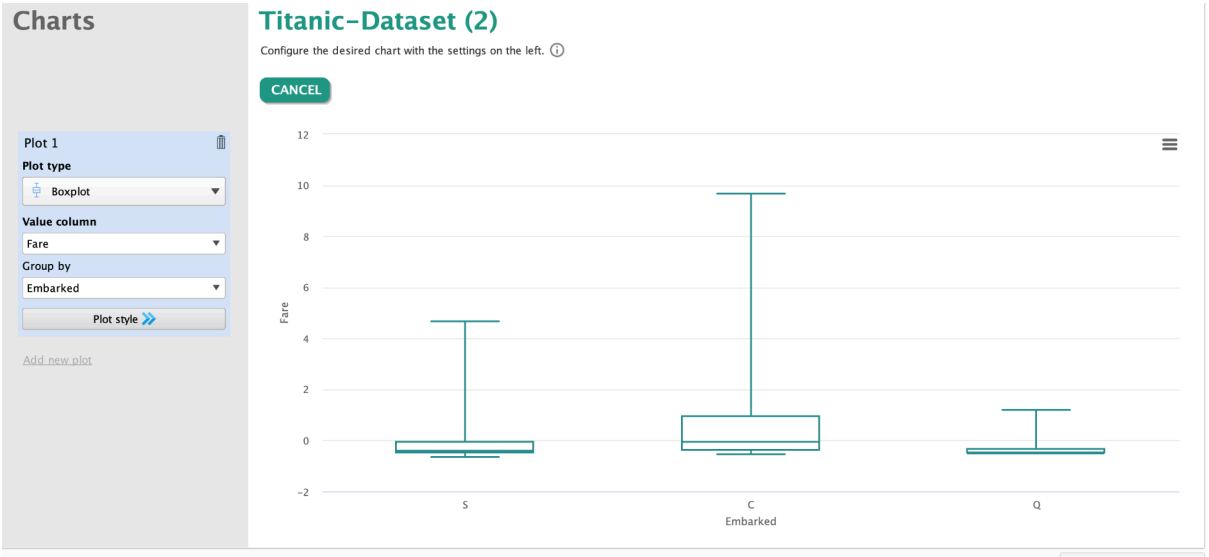

рис. 11: Построение ящика с усами (Box Plot)

6. *Тепловая карта (Heatmap) корреляций для числовых признаков*

- Plot type: Heatmap
- Выберите матрицу корреляций между Age, Fare, SibSp, Parch.
- Показывает степень взаимосвязи признаков друг с другом, что важно при выборе признаков для модели.

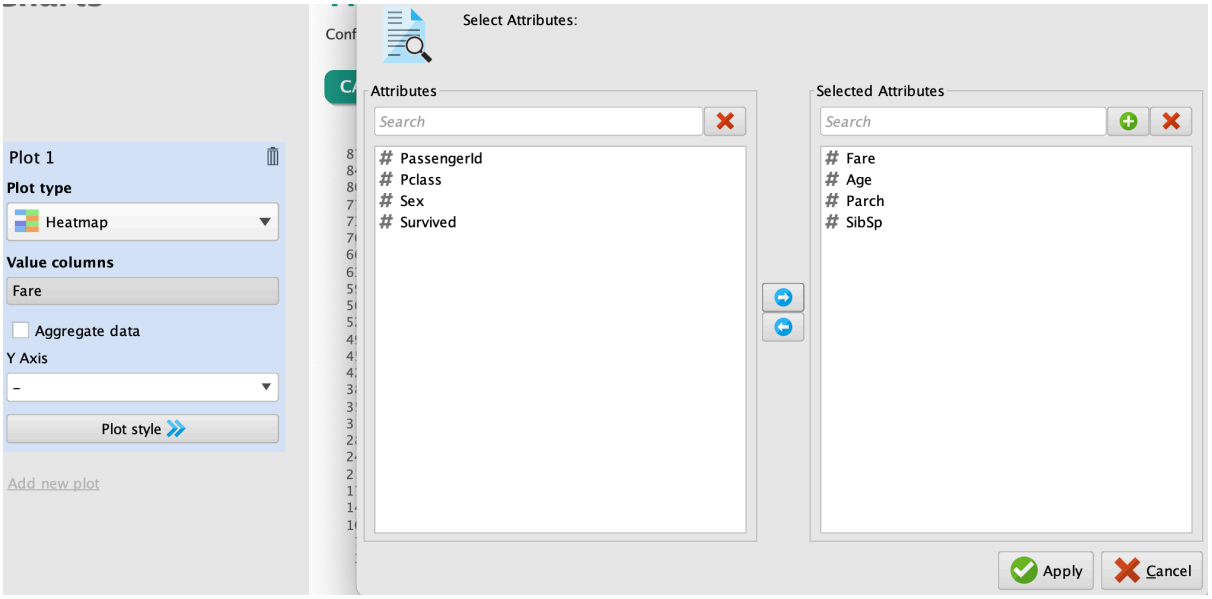

рис. 12: Построение тепловой карты: выбор признаков

Настройки графиков

- Изменяйте заголовки осей (Axis Title), подписи к легенде (Legend), цветовые схемы (Color Settings).
- Экспериментируйте с фильтрами: например, отображать только пассажиров определенного класса или пола.

- Сортируйте столбчатые диаграммы по убыванию или возрастанию значений.

Автоматизированное создание модели с помощью AutoModel

- Нажмите **"AutoModel"** в верхней панели.
- Загрузите подготовленный набор данных Titanic.
- Выберите задачу классификации с целевым признаком Survived.

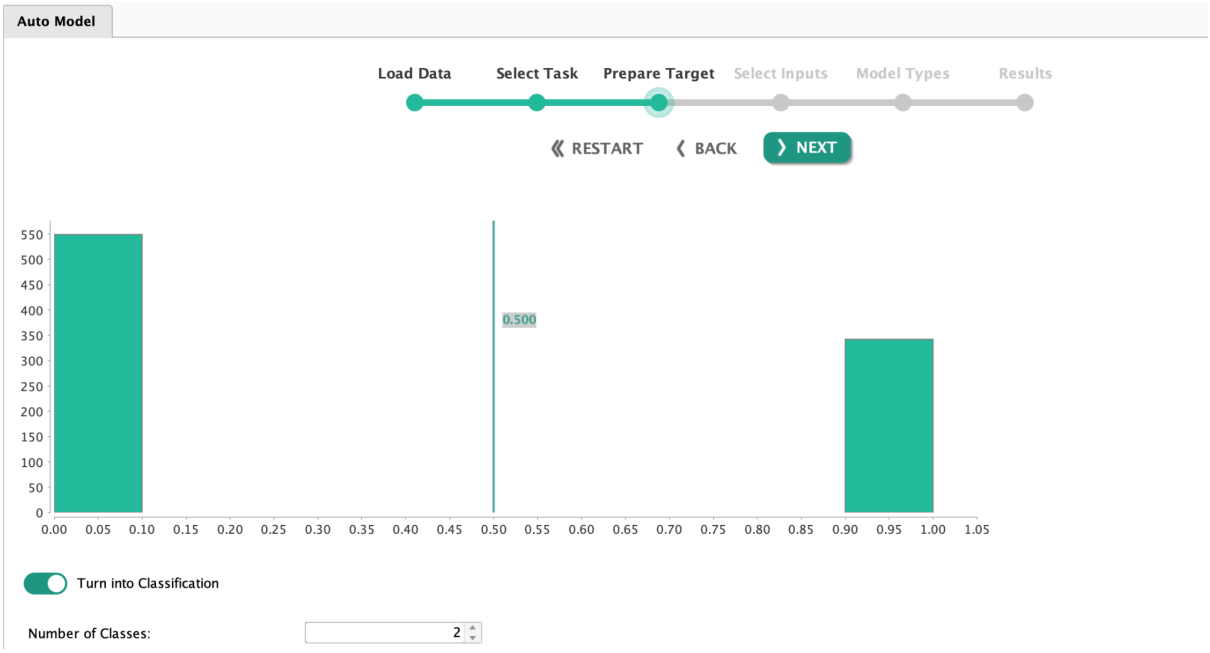

рис. 13: Использование AutoModel

- Выберите модели: Decision Tree, Logistic Regression, Random Forest, Naive Bayes.
- Нажмите **"Run"**, чтобы построить модели.

Анализ результатов

- Перейдите во вкладку Results после работы AutoModel.
- Сравните точность (accuracy).

Model *Category*	Classification Error *Number*	Standard Deviation *Number*	Gains *Number*	Total Time *Number*	Training Time (1,0... *Number*	Scoring Time (1,00... *Number*
Naive Bayes	0.216	0.020	96	19192	117.845	271.709
Logistic Regression	0.176	0.028	102	6429	141.414	224.090
Decision Tree	0.173	0.032	98	7056	74.074	207.283
Random Forest	0.169	0.022	104	102091	404.040	7050.420

рис. 14: Результат использования AutoModel и сравнение различных моделей по

ключевым метрикам качества

- В разделе Model для дерева решений (Decision Tree) можно увидеть визуализацию, демонстрирующую важность признаков и логику классификации.

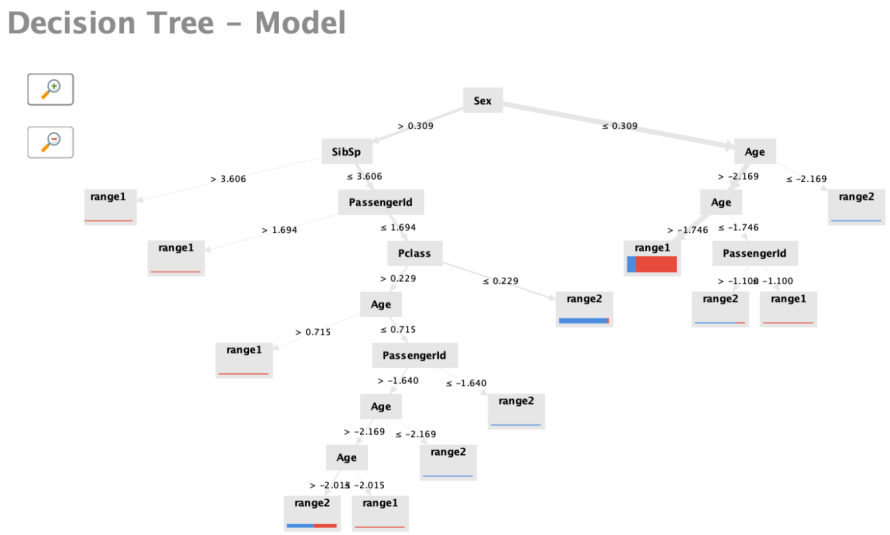

рис. 15: Просмотр архитектуры модели "Дерево решений"

- Также можно рассмотреть другие метрики качества, помимо accuracy. Например, можно выбрать Recall, F Measure или AUC, которые также будут релевантными показателями.

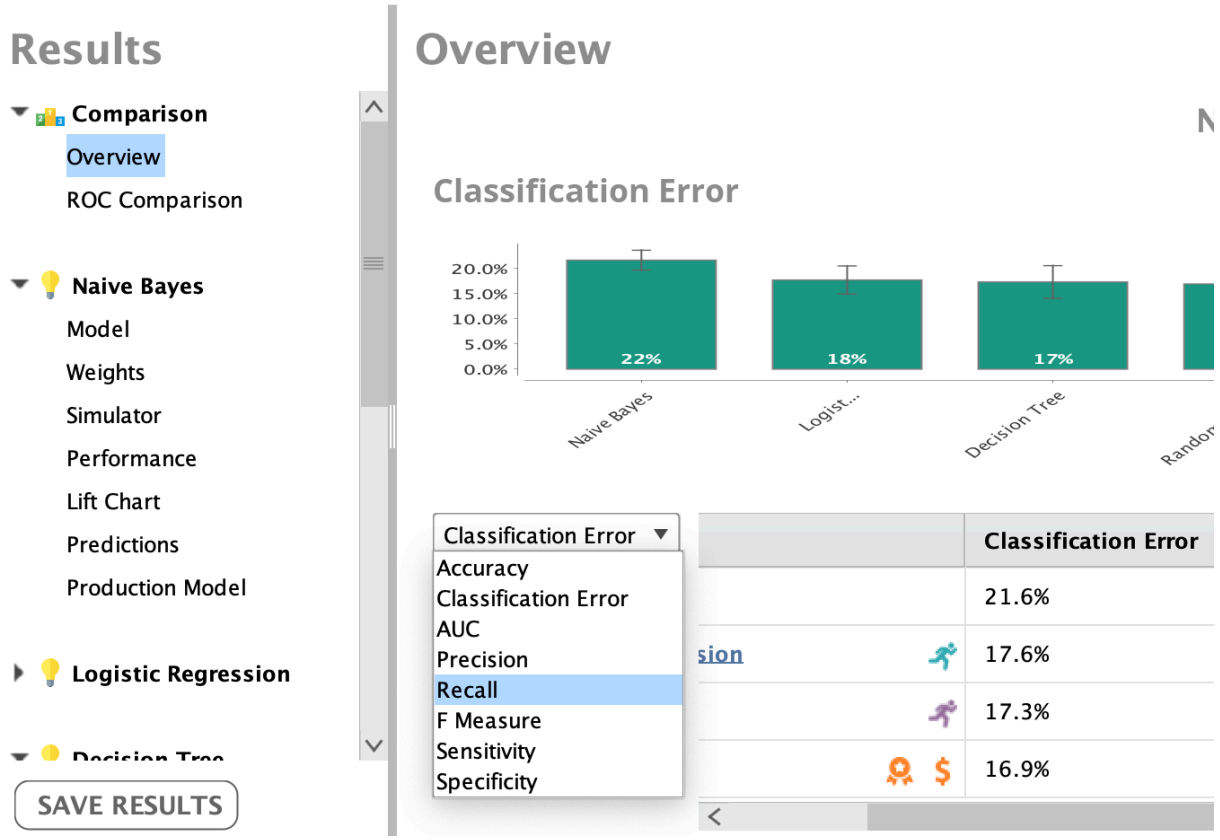

рис. 16: Сравнение моделей по выбранной метрике качества

Приобретенные навыки

- Умение работать с разными типами графиков в RapidMiner (Altair AI Studio) для исследования структуры данных.
- Опыт настройки параметров визуализации: количество корзин в гистограмме, цветовые схемы, легенды, фильтры.
- Навыки интерпретации результатов моделей с помощью визуальных инструментов (дерево решений, матрица ошибок, метрики точности).
- Способность применять визуальный анализ для принятия обоснованных решений в процессе машинного обучения.

Обобщенная задача для выполнения индивидуального варианта

Предложенный вам в индивидуальном варианте набор данных будет содержать как числовые, так и категориальные признаки. В ходе выполнения практической работы вам требуется выполнить в RapidMiner (Altair AI Studio) следующие шаги:

Загрузка и предобработка данных:

- Импортируйте данные через оператор Read CSV.
- Проверьте корректность типов столбцов, устраните пропуски и, при необходимости, закодируйте категориальные признаки (Replace / Nominal to Numerical).
- Выполните нормализацию числовых признаков (оператор Normalize).

Изучение структуры данных через визуализацию:

- Постройте гистограмму для одного из ключевых числовых признаков, чтобы оценить распределение.
- Создайте столбчатую диаграмму для одной категориальной переменной и исследуйте ее связь с целевой меткой (если задача кластеризации — без привязки к целевой).
- Постройте круговую диаграмму для распределения классов (или категорий).
- Постройте диаграмму рассеяния по двум числовым признакам, раскрашенную по одной из категорий или по предсказанию простейшей модели.
- Постройте ящик с усами (Box Plot) для числового признака, сгруппированного по категории.
- Постройте тепловую карту корреляций для всех числовых признаков, чтобы оценить их взаимосвязь.

Сравнение моделей:

- Запустите AutoModel для задачи классификации или регрессии, оцените несколько моделей.
- Сравните их точность и другие метрики, визуализируйте структуру лучшей модели (например, дерево решений) и важность признаков.

Интерпретация результатов и выводы:

- На основе визуализаций сформулируйте ключевые наблюдения о распределении данных и влиянии признаков.
- Опишите, какие виды графиков оказались наиболее информативными и почему.

Распределение вариантов