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 № 8

ТЕМА РАБОТЫ
«Анализ текста (Text Mining) в RapidMiner»

Москва, 2025

Цель работы

Освоить инструменты Text Mining в RapidMiner для обработки, анализа и классификации текстов. В ходе лабораторной работы студенты:

- Познакомятся с основами анализа текста;
- Научатся загружать текстовые данные и проводить их предобработку;
- Освоят ключевые методы работы с текстом, такие как токенизация, удаление стоп-слов, лемматизация;
- Построят модель классификации текстов на основе машинного обучения;
- Проанализируют результаты и оценят качество модели.

Целевая аудитория

Работа предназначена для студентов, заинтересованных в освоении анализа текстовых данных, закреплении навыков использования алгоритмов машинного обучения и изучении обработки естественного языка (NLP) в среде RapidMiner.

Идея и концепция

В этой практической работе студенты изучат основы текстового анализа, применяя его к задаче анализа тональности SMS. В качестве примера используется набор данных с текстовыми сообщениями (SMS Spam Collection).

В процессе выполнения работы студенты:

- Загрузят и предобработают текстовые данные;
- Применят основные методы Text Mining (разбиение текста на слова, удаление стоп-слов, векторизация текста);
- Создадут модель машинного обучения для классификации текстовых сообщений на «полезные» и «спам-сообщения»;
- Проведут анализ и оценку результатов.

Содержание практической работы

О наборе данных и задаче работы

Набор данных:
SMS Spam Collection – содержит текстовые сообщения с метками о назначении.

Формат данных – CSV-файл с двумя столбцами:
- message – текст сообщения.
- spam or not – поля («spam» / «ham»), ham — «нормальное» сообщение, spam — рекламное или мошенническое. Датасет сбалансирован: 4 827 ham (86 %) и 747 spam (14 %).

Задача:
Построить модель, классифицирующую текстовые сообщения на положительные и отрицательные (спам).

Работа с данными

Загрузка набора данных

1) Загрузка CSV-файла с отзывами:

- В панели «Operators» найдите оператор «Read CSV».
- Перетащите оператор в центральную область процесса.
- В панели Parameters справа нажмите на кнопку … около поля file и выберите требуемый CSV-файл.
- Убедитесь, что разделитель (separator) указан правильно и флаг First row as column names установлен.
- Правой кнопкой по оператору Read CSV → Show Data — убедитесь, что загружены два столбца: spam or not (Nominal) и message (Nominal).

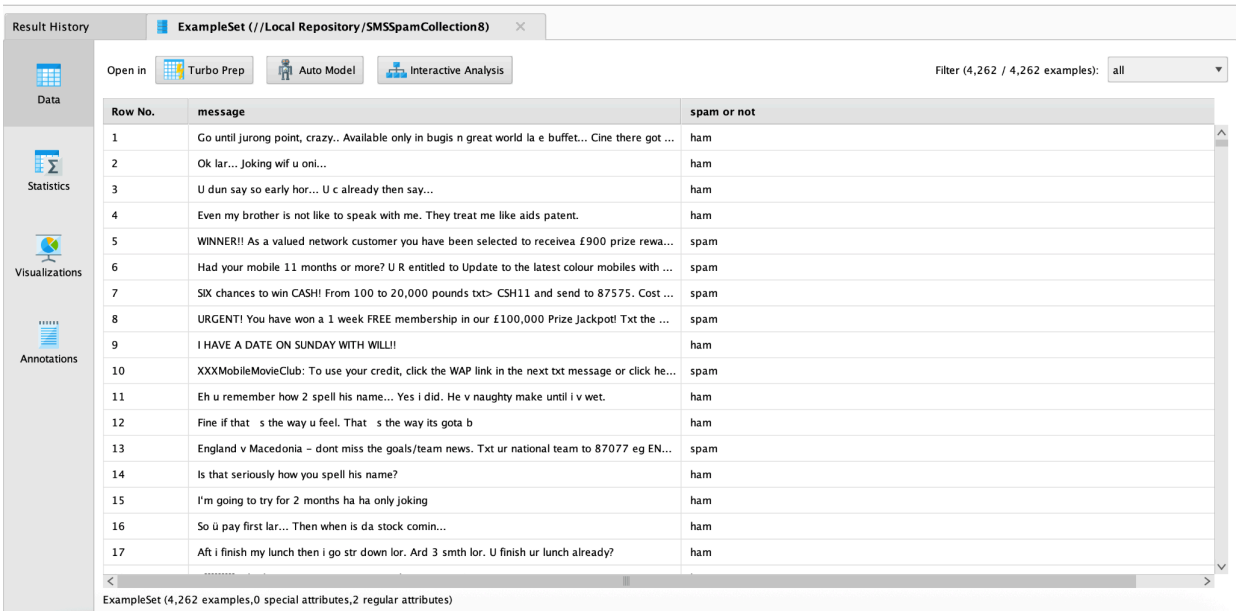

рис.1: Загрузка датасета

рис.2: Просмотр вкладки Statistics

Предварительная обработка текстов

1) Назначение ролей (выбор целевой переменной):
   - Перетащите Set Role; соедините Read CSV (out) → Set Role (exa).
   - Параметр set roles → Edit List → attribute name =spam or not, target role = label.

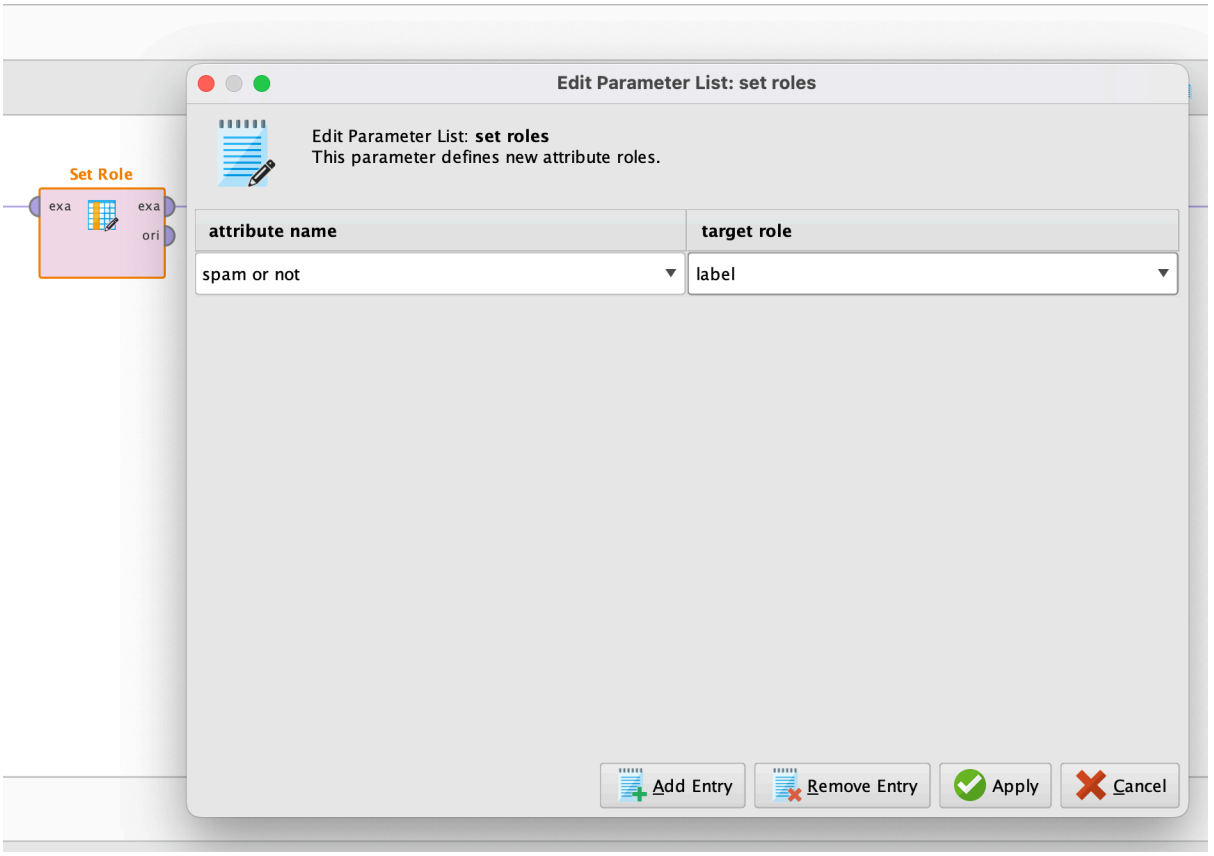

рис.3: Настройки оператора Set Role

2) Преобразование типов столбцов: так столбец message имеет тип Nominal, сначала требуется сконвертировать его в Text-формат:

- В панели Operators найдите Nominal to Text.
- Перетащите оператор Nominal to Text в главный процесс.
- В параметрах оператора Nominal to Text установите attribute filter type = single и attribute = message.

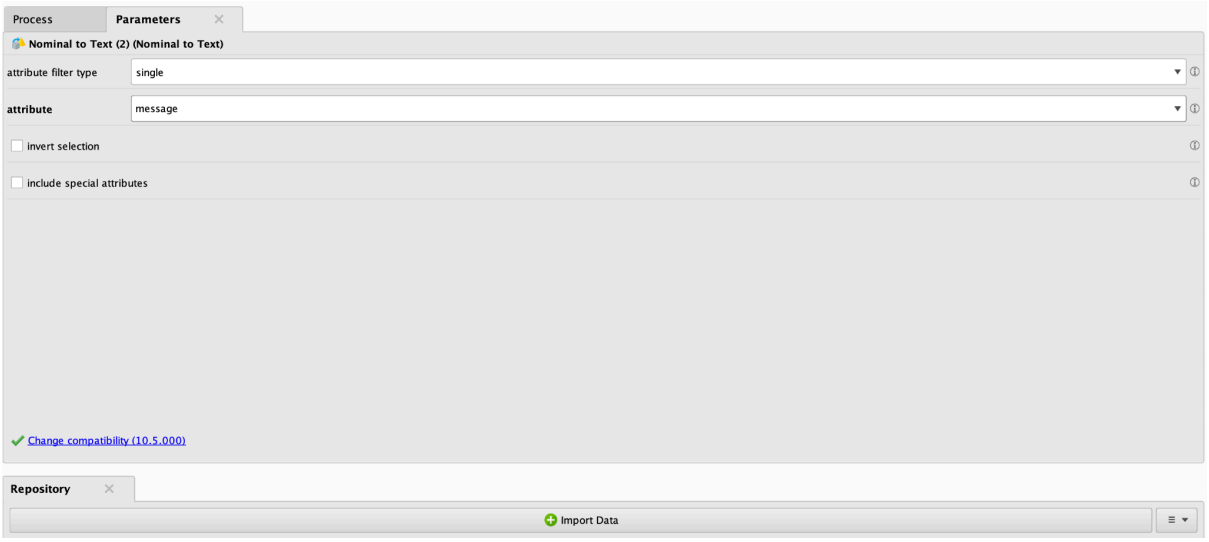

рис.4: Настройка параметров оператора Nominal to Text

- Теперь выход exa у Nominal to Text содержит атрибут message в формате Text.

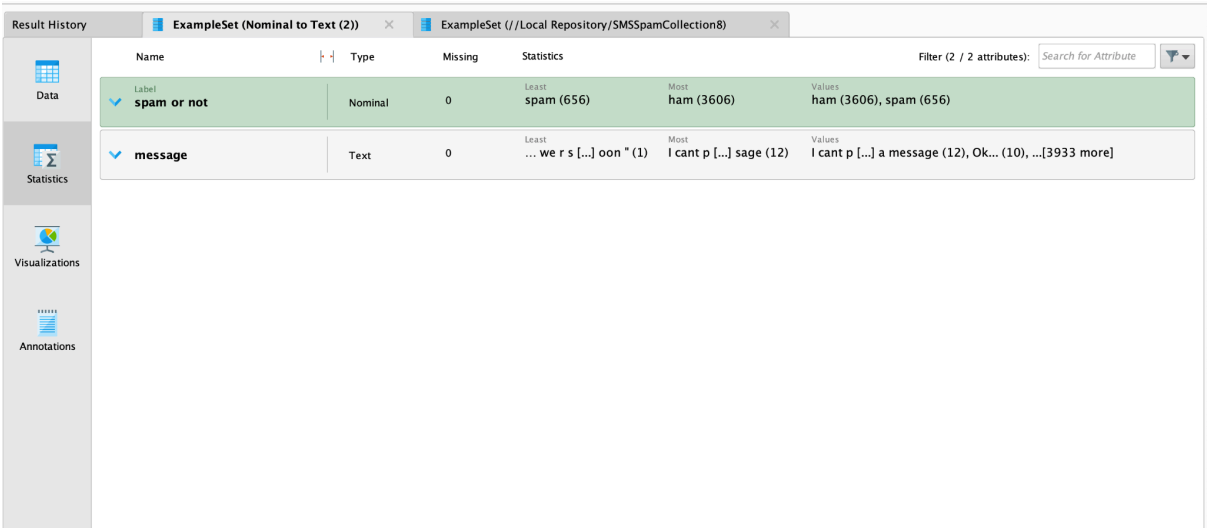

рис.5: Результат преобразований с помощью оператора Nominal to Text

3) Использование оператора Process Documents from Data:

- В панели Operators найдите Process Documents from Data и перетащите его после Nominal to Text.
- Подключите выход exa у Nominal to Text к входу exa у Process Documents from Data.

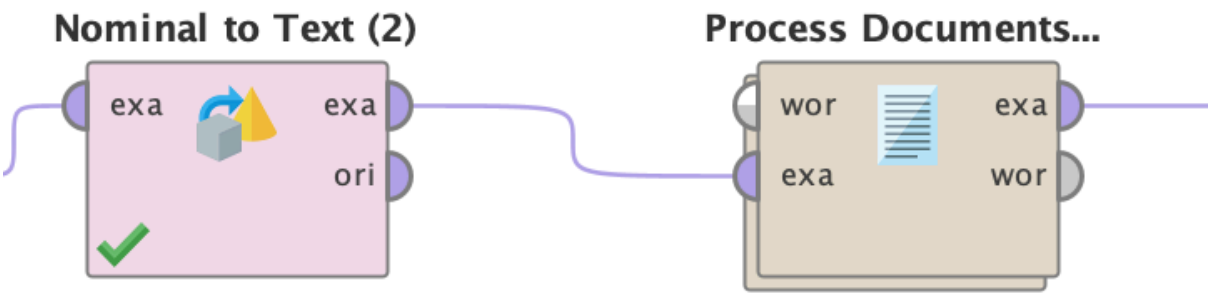

рис.6: Подключение оператора Process Documents from Data

- Дважды щелкните по оператору Process Documents from Data, чтобы открыть подпроцесс.
- Внутри подпроцесса добавьте следующие операторы в порядке:
  - Tokenize (режим Non-letters) — разбивает текст по небуквенным символам.
  - Transform Cases (lowercase) — приводит все токены к нижнему регистру.
  - Filter Stopwords (English) — удаление английских стоп-слов.
  - Stem (Porter) — стемминг.

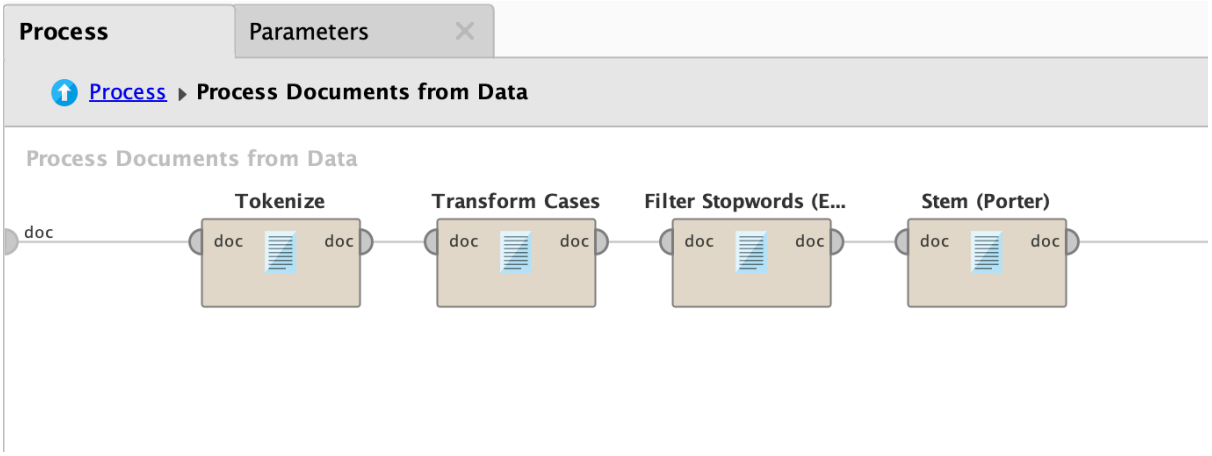

рис.7: Настройка субпроцессов внутри оператора Process Documents from Data

- Нажмите Return to Parent Process для возврата к основному процессу.
- Откройте параметры оператора Process Documents from Data и выберите Vector Creation – (TF-IDF), для формирование векторного представления. TF(term, d) – относительная частота токена в документе; IDF(term) = log (N / df); итоговый признак = TF × IDF.
- 

рис.8: Настройка параметров оператора Process Documents from Data

- Проверьте последовательность операторов и убедитесь, что выход word list у TF-IDF соединен с выходом.

Разделение данных на обучающую и тестовую выборки

1) В основном процессе найдите оператор Split Data и перетащите его после Process Documents from Data.

2) Подключите выход exa у Process Documents к входу exa у Split Data.

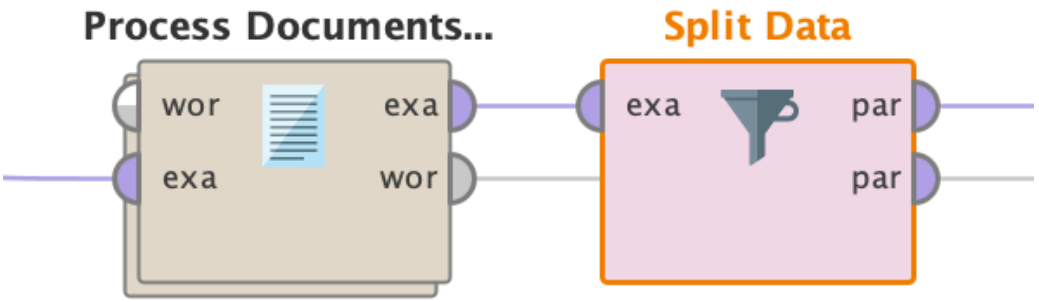

рис.9: Подключение оператора Split Data

3) В правой панели Parameters щелкните по полю partitions → откроется окно «Edit Parameter List: partitions».
4) Нажмите Add Entry два раза, чтобы появилось две строки.
5) В первой строке ratio введите 0.8 – это будет первая выборка (обучающая).
6) Во второй строке ratio введите 0.2 – это будет вторая выборка (для тестирования).

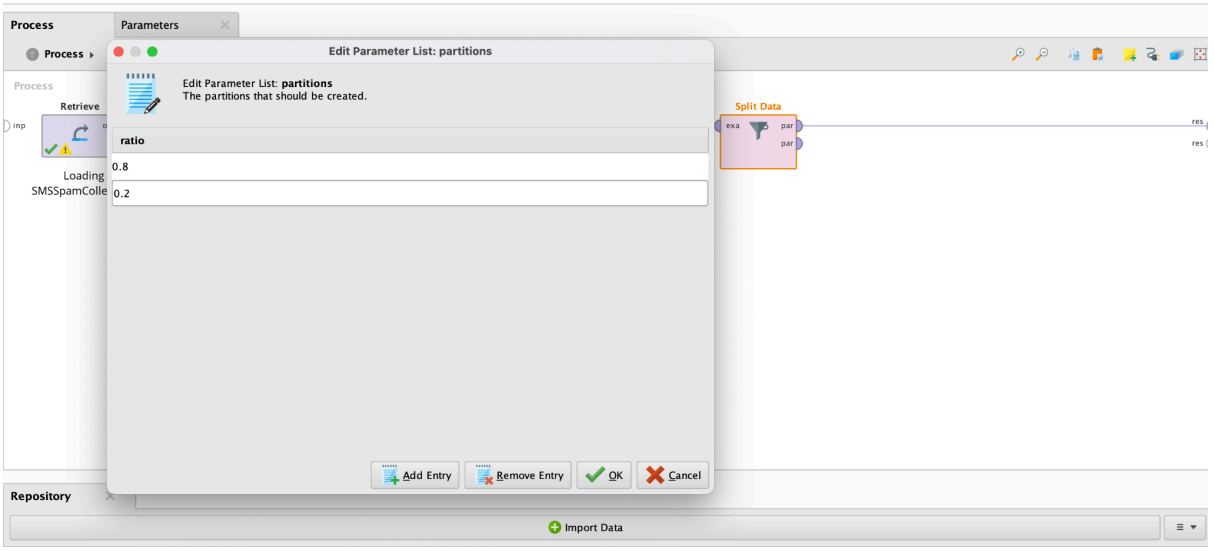

рис.10: Настройки оператора Split Data для разделения на обучающую и тестовую выборки

7) После этого у блока Split Data появятся два выходных порта:

- exa (0) — содержит 80 % данных;
- exa (1) — содержит 20 % данных.

8) После этого для подачи в модель требуется использовать оператор Nominal to Binominal → оба потока:
- Split Data (exa0) → Nominal to Binominal (exa)
- Split Data (exa1) → Apply Model (unlabeled).
- В настройках параметров укажите: attribute filter type = single, attribute = spam or not (целевая переменная).

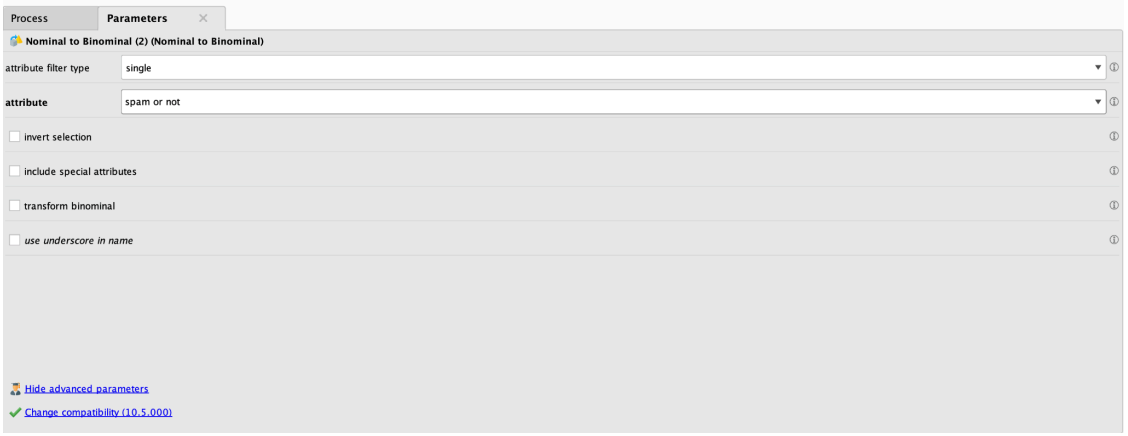

рис.11: Настройки параметров оператора Nominal to Binominal

Применение модели и прогнозирование

1) Обучение с помощью модели логистической регрессии:
- Найдите параметр Naive Bayes;

- Подключите вход tra ← Nominal to Binominal (exa0);

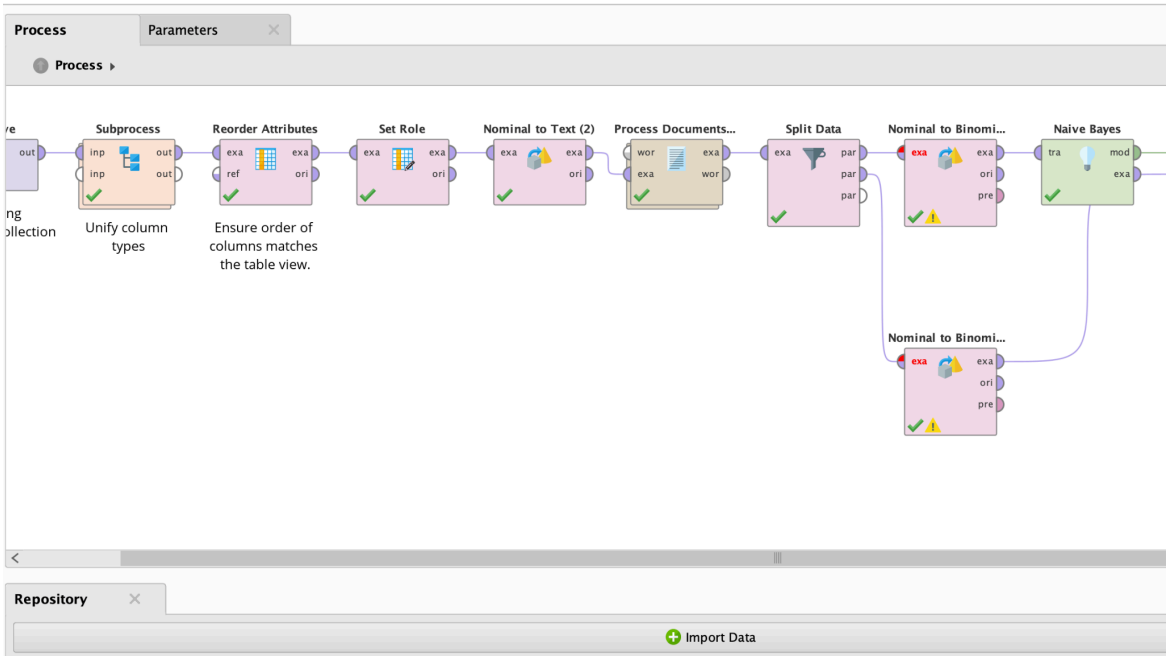

рис.12: Подключение оператора Naive Bayes

2) Прогнозирование:
- Добавьте оператор Apply Model;
- Подключите model ↔ Naive Bayes, unlabelled data ↔ Nominal to Binominal (exa1).

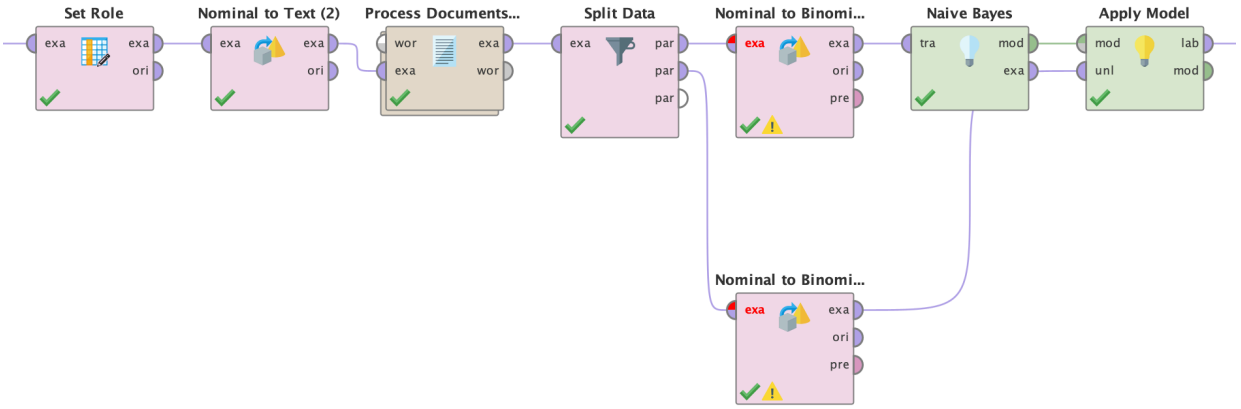

рис.13: Подключение оператора Apply Model

3) Оценка качества прогнозирования:
- Подключите оператор Performance (универсальный) после Apply Model;

- Соедините Apply Model (labelled data) → Performance (labelled data) (второй вход Performance оставить пустым).
- Запустите процесс.

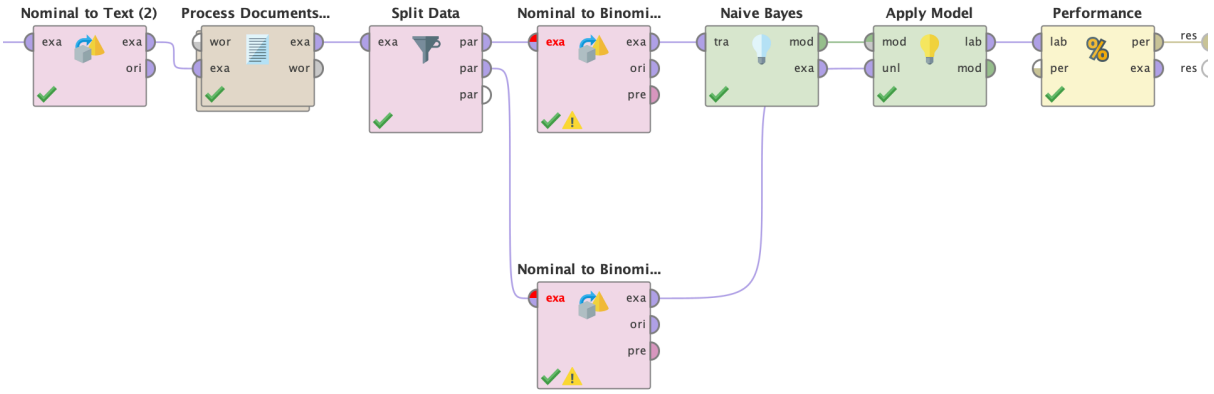

рис.14: Подключение оператора Performance

4) Просмотр результатов во вкладке Performance (Result):
- Accuracy
- Precision (spam)
- Recall (spam)
- AUC

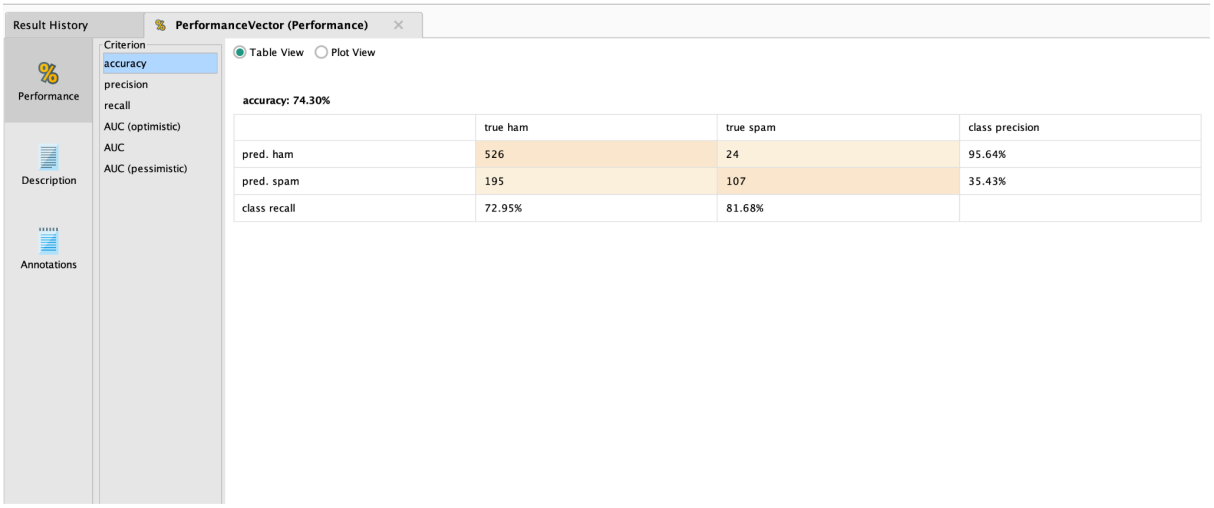

рис.15: Результаты прогнозирования

5) Построение столбчатой диаграммы:
- Для просмотра распределения классифицированных бинарных классов (полезное сообщение или спам) необходимо построить столбчатую диаграмму во вкладке Visualizations;

- Выберите тип диаграммы – Bar(column);
- Value columns – spam or not (целевая прогнозируемая переменная);
- Aggregation Function – Count;

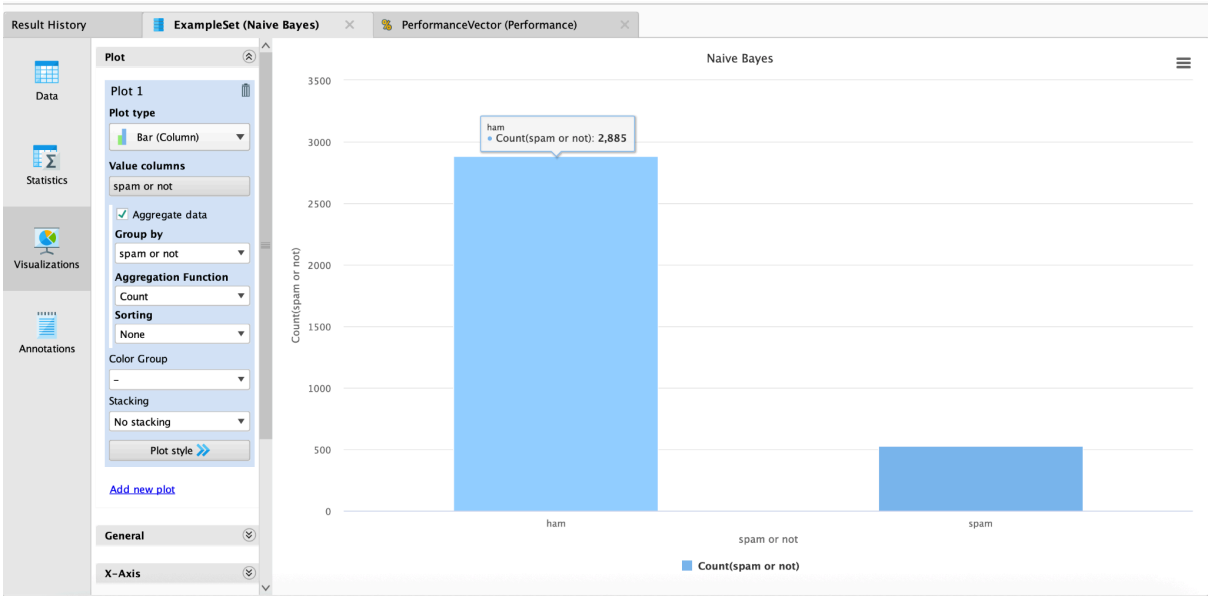

рис.16: Столбчатая диаграмма для распределения бинарных классов целевой переменной

6) Построение круговой диаграммы:

- Для просмотра распределения по детекции определенного слова / словосочетания (является ли сообщение, содержащее данный токе, полезным или спамом) необходимо построить круговую диаграмму во вкладке Visualizations;
- Выберите тип диаграммы – Pie / Donut;
- Value columns – goodnight (приведено в качестве примера, можно выбать любое слово);
- Aggregation Function – Count; Group By – spam or not.

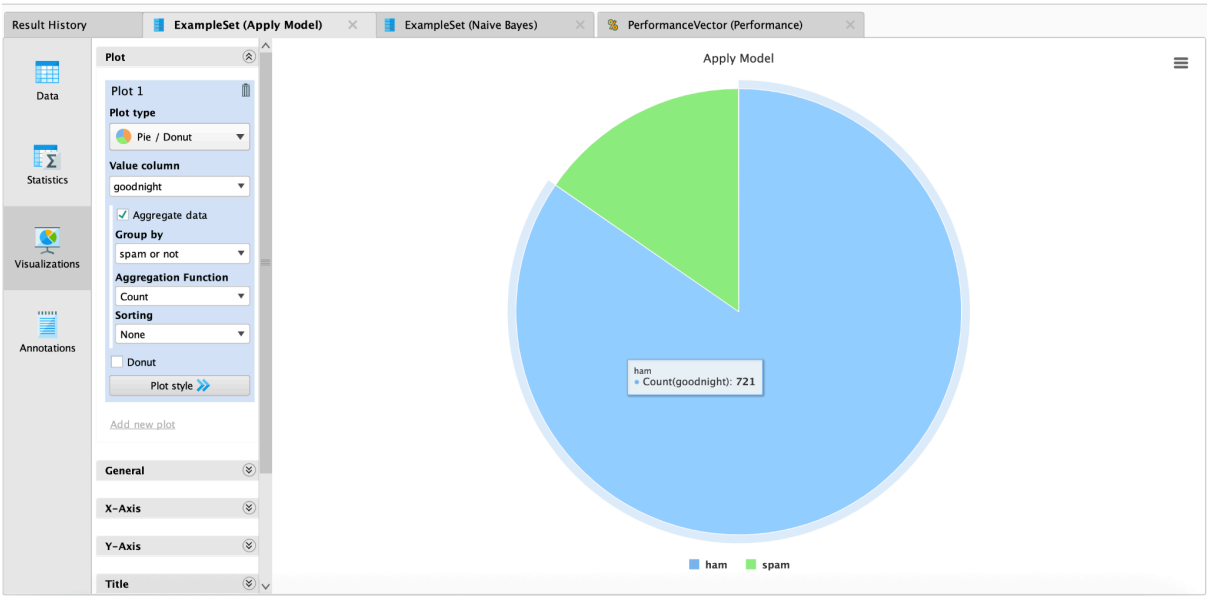

рис.17: Круговая диаграмма

Приобретенные и закрепленные навыки

- Импорт и первичная проверка текстовых данных в RapidMiner (Read CSV, Show Data).
- Назначение целевой переменной и преобразование типов (Set Role, Nominal to Text).
- Настройка субпроцесса текстовой обработки (Tokenize, Transform Cases, Filter Stopwords, Stem/Lemmatize).
- Преобразование текста в числовые признаки с помощью TF-IDF.
- Разделение выборки на обучающую и тестовую (Split Data).
- Обучение и применение классификаторов (Naive Bayes, Logistic Regression, SVM).
- Оценка качества модели по ключевым метрикам (Accuracy, Precision, Recall, AUC).
- Анализ ошибок классификации и практические рекомендации по улучшению текстового пайплайна.

Обобщенная задача для выполнения индивидуального варианта

Практическая работа направлена на анализ и классификацию текстовых данных на наборе сообщений или документов (например, отзывы о товарах, комментарии в социальных сетях, письма службы поддержки, новости). В вашей работе должны быть реализованы следующие этапы:

1) Загрузка и подготовка данных
- Импорт CSV-файла.
- Проверка корректности распознавания типов: преобразование текстового столбца в формат Text (оператор Nominal to Text), а целевую метку в Label (оператор Set Role).

2) Предобработка текстовых данных
- Используйте оператор Process Documents from Data и внутри субпроцесса реализуйте:
  • Tokenize (разбиение по небуквенным символам),
  • Transform Cases (приведение к нижнему регистру),
  • Filter Stopwords (удаление стоп-слов на соответствующем языке),
  • Stem или Lemmatize (сведение слов к основе).
  • В главных параметрах Process Documents выберите Vector Creation (TF-IDF) для получения числового представления документов.

3) Построение и оценка модели
- Разделите данные на обучающий и тестовый наборы (оператор Split Data, например, 80/20).
- Обучите классификатор (Naïve Bayes, Logistic Regression или SVM) на обучающих данных.
- С помощью Apply Model и Performance (Classification) оцените качество по метрикам: Accuracy, Precision, Recall, AUC.

4) Анализ результатов
- Сравните полученные метрики и проанализируйте, какие ошибки совершает модель.
- При необходимости поэкспериментируйте с альтернативными методами векторизации или с другими классификаторами.

Распределение вариантов


RAPIDMINER