

Автоматизация обработки данных

Цель

Разработать Python-скрипт для автоматизации обработки данных, который будет загружать исходные данные из различных источников, очищать их, проводить анализ, строить простые ML-модели, генерировать отчеты и визуализации для бизнес-задач. Проект поможет бизнес-аналитикам и специалистам по данным автоматизировать рутинные задачи, быстро получать ключевые метрики и прогнозы.

Требования к проекту

1 Загрузка данных

- Поддержка загрузки данных из файлов форматов CSV, Excel, Подключение к базе данных через SQL-запросы (PostgreSQL), включая сложные запросы с JOIN и агрегатами, Возможность загрузки данных из внешних API (REST).

2 Валидация данных при загрузке

- Проверка на дубликаты, Проверка пропусков и корректности типов данных, Выявление и обработка выбросов (методы IQR, Z-score), Логирование результатов валидации.

3 Очистка данных

- Обработка пропущенных значений: замена средними, медианными значениями либо удаление строк, Удаление дубликатов, Кодирование категориальных признаков (One-Hot Encoding, Label Encoding), Преобразование и нормализация данных (масштабирование числовых признаков), Работа с форматами дат (преобразование строковых дат в datetime).

4 Анализ данных

- Подсчет базовых статистик: среднее, медиана, мода, стандартное отклонение,
Анализ временных рядов: выявление трендов, сезонных колебаний (например, с помощью декомпозиции),
Выявление аномалий и выбросов,
Построение и обучение базовых моделей машинного обучения (регрессия, классификация),
Вывод метрик качества моделей (accuracy, precision, recall, F1, ROC-AUC, RMSE и т.п.)

5 Отчетность

- Автоматическая генерация отчетов с ключевыми метриками,
Визуализация данных с помощью Matplotlib, Seaborn и интерактивных графиков Plotly,
Форматирование отчетов в PDF и Excel (использование ReportLab, openpyxl и др.),
Автоматическая отправка отчетов по email через SMTP.

6 Интеграция с бизнес-процессами и автоматизация

- Регулярный запуск скрипта с использованием планировщика задач (cron, Task Scheduler),
Логирование процесса выполнения скрипта,
Возможность интеграции с системами бизнес-анализа через API,
Сохранение результатов анализа и моделей в базу данных.

7 Документация и тестирование

- Подробная документация по использованию всех функций модуля,
Описание параметров и примеров использования,
Создание README-файла для быстрого старта,
Разработка юнит-тестов для основных функций загрузки, очистки и анализа.

Алгоритм выполнения задачи

1 Загрузка данных

- Реализовать функции загрузки данных из CSV, Excel, баз данных через SQL-запросы,
- Добавить возможность загрузки из внешних API,
- Валидация данных: проверка дубликатов, пропусков, типов данных, выявление выбросов,
- Логирование результатов проверки.

2 Очистка данных

- Обработка пропущенных значений: замена средним, медианой или удаление,
- Удаление дубликатов,
- Кодирование категорий, масштабирование числовых признаков,
- Преобразование строк дат в формат datetime.

3 Анализ данных

- Подсчет статистик: среднее, медиана, мода, стандартное отклонение,
- Анализ трендов и сезонности во временных рядах,
- Выявление аномалий (выбросов),
- Построение и обучение базовой ML-модели (регрессия или классификация),
- Вывод метрик качества модели.

4 Отчетность

- Автоматическая генерация текстовых и графических отчетов,
- Визуализация данных с помощью Matplotlib, Seaborn, Plotly,
- Форматирование отчетов в PDF и Excel,
- Автоматическая отправка отчетов по email.

5 Интеграция и автоматизация

- Настройка планировщика задач для регулярного запуска,
- Логирование всех этапов обработки,
- Интеграция с API и сохранение результатов в БД.

6

Документация и тестирование

- Описание всех функций, параметров и примеров,
- Создание README,
- Юнит-тестирование ключевых компонентов.

Результат

ссылка на git

Критерии оценивания

К1	Загрузка данных: корректная работа функций загрузки, валидация, поддержка SQL и API	10 баллов
К2	Очистка данных: обработка пропусков, дубликатов, выбросов, кодирование и нормализация	10 баллов
К3	Анализ данных: расчет статистик, анализ временных рядов, построение ML-модели и оценка её качества	5 баллов
К4	Отчетность: генерация полноценных отчетов с визуализацией, автоматическая рассылка	5 баллов
К5	Автоматизация: регулярный запуск, логирование, интеграция с API и БД	10 баллов
К6	Документация и тестирование: полная документация и наличие юнит-тестов	10 баллов

Максимальное количество баллов

50 баллов

Минимальное количество баллов, чтобы преподаватель смог зачесть вашу работу

25 баллов