Домашнее задание: EDA и базовое предсказание заболеваний на основе табличных данных

Домашнее задание

Цель

Сделать разведочный анализ данных (EDA), подготовить отчётные таблицы и построить базовую модель машинного обучения для предсказания заболевания или исхода лечения.

Задание

Шаг 1. Подготовка данных

1. Сделать загрузку набора данных с клиническими случаями:
    - использовать данные из предыдущих домашних заданий, расширить их.
2. Сделать проверку структуры таблицы и типов данных.
3. Приложить вывод первых строк таблицы.

Шаг 2. Разведочный анализ данных (EDA)

1. Сделать описание данных:
    - количество строк и столбцов;
    - типы данных;
    - наличие пропущенных значений.
2. Сделать базовые статистики для числовых признаков:
    - среднее значение;
    - медиана;
    - минимум и максимум.
3. Сделать анализ категориальных признаков:

    - частоты диагнозов;
    - распределение исходов лечения.
4. Сделать не менее **двух выводов** по результатам EDA.
5. Приложить результаты анализа.

Шаг 3. Подготовка данных для машинного обучения

1. Сделать выбор целевой переменной:
    - диагноз **или**
    - исход лечения.
2. Сделать выбор признаков для модели.
3. Сделать подготовку данных:
    - обработку пропущенных значений;
    - кодирование категориальных признаков;
    - при необходимости — масштабирование числовых признаков.
4. Приложить итоговый набор признаков.

Шаг 4. Построение модели машинного обучения

1. Сделать разбиение данных на обучающую и тестовую выборки.
2. Сделать обучение **одной базовой модели** (например, логистическая регрессия или дерево решений).
3. Сделать получение предсказаний на тестовой выборке.
4. Приложить код обучения модели.

Шаг 5. Оценка качества модели

1. Сделать расчёт метрик качества:
    - accuracy;
    - precision и recall **или** другую подходящую метрику.
2. Сделать краткий комментарий:
    - что показывает модель;
    - какие ограничения есть у полученного результата.
3. Приложить результаты оценки.

Шаг 6. Подготовка отчётных таблиц

1. Сделать таблицу с:
   - реальными значениями целевой переменной;
   - предсказаниями модели.
2. Сделать сводную таблицу с метриками качества.
3. Приложить полученные таблицы.

Шаг 7. Экспорт и сохранение данных

1. Сделать сохранение:
   - отчётных таблиц в CSV;
   - при необходимости — обученной модели.
2. Убедиться, что сохранённые файлы можно загрузить повторно.
3. Приложить результат (пути к файлам или ссылки).

Шаг 8. Практический сценарий

1. Сделать краткое описание сценария:
   - как подобная модель может использоваться в клинической практике;
   - какие данные нужны для улучшения качества предсказаний.
2. Сделать вывод о применимости модели.
3. Приложить текстовое описание (2–4 абзаца).

Подсказки по ключевым частям

- Для EDA используйте:
  - `.describe()`
  - `.value_counts()`
  - `.isna()`
- Для машинного обучения можно использовать `scikit-learn`.
- Для кодирования категорий:
  - `get_dummies`
  - `LabelEncoder`
- Для сохранения данных:
  - `.to_csv()`

- `joblib` или `pickle`.

Что проверить перед отправкой (чек-лист)

- Сделан разведочный анализ данных.
- Сделаны выводы по EDA.
- Подготовлены данные для модели.
- Обучена базовая модель машинного обучения.
- Рассчитаны метрики качества.
- Подготовлены отчётные таблицы.
- Данные и результаты сохранены.
- Работа оформлена аккуратно и понятно.

Советы по улучшению работы

- Чётко разделяйте EDA, обучение модели и выводы.
- Используйте комментарии и Markdown-ячейки для пояснений.
- Не стремитесь к высокой точности — важна корректность процесса и интерпретация.

Ответ

Ссылка на решение:
https://colab.research.google.com/drive/1PE5sQ0fdSYPMLe057CdxRdDZzLRrq2YC?usp=sharing