Задание по машинному обучению

(на выбор: выполняете 1 или 2)

1. Базовое задание

Для набора данных: **файл CSV**

1. Импортировать data set.
2. Провести предобработку (проверить на наличие выбросов, пропусков). Если есть выбросы, удалить. Если есть пропуски, обработать, как сочтёте нужным.
3. Найти числовые характеристики признаков и зависимой переменной.
4. Проверить на наличие связей зависимой переменной с факторами и факторов между собой, применяя методы корреляционного анализа и сравнения групп.
5. Визуализировать данные и зависимости между ними.

пишите комментарии к коду в Markdown или прямо в коде через #

Комментарий

- В файле для задания должно быть 392 наблюдения и 9 признаков.
  Если импорт выполнен корректно, результат должен давать такую структуру: 392 rows x 9 columns.

Полезные ссылки

1. Блокнот
2. 50 оттенков matplotlib — The Master Plots
3. Предварительная обработка данных

2. Задание для тех, кто хочет попробовать свои силы

(состоит из двух частей)

1. Статистика.

Для набора данных : https://archive.ics.uci.edu/ml/datasets/auto+mpg

1. Импортировать data set.
2. Провести предобработку (проверить на наличие выбросов, пропусков). Если есть выбросы, удалить. Если есть пропуски, обработать, как сочтёте нужным.
3. Найти числовые характеристики признаков и зависимой переменной.
4. Проверить на наличие связей зависимой переменной с факторами и факторов между собой, применяя методы корреляционного анализа и сравнения групп.
5. Визуализировать данные и зависимости между ними.

Если не удалось импортировать, можно взять готовый **файл CSV**.

2. Регрессионный анализ.

На основании результатов ***1. Статистика***, выбрать значимые факторы и построить регрессионную модель зависимой переменной от признаков, предварительно разбив выборку на обучающую и тестовую.
Предложить разные алгоритмы. Выбрать наилучший.

пишите комментарии к коду в Markdown или прямо в коде через #

Комментарий

- В файле для задания должно быть 392 наблюдения и 9 признаков.
  Если импорт выполнен корректно, результат должен давать такую структуру: 392 rows x 9 columns.
- Кто берёт данные сам с сайта, там 398 строк, 6 из которых содержат пропуски.
  Плюс-минус одно наблюдение — некритично. Кто-то будет удалять выбросы, кто-то — нет. Так что решения в любом случае будут немного отличаться.

- В присланном файле csv стоят десятичные точки, как это уже и требуется для питона. Если Вы просто откроете его в Excel, формат покажется странным, так как Excel переведёт десятичные числа в даты и пр. Это нормально! Ничего менять не надо. Если Вы откроете этот файл в блокноте, то увидите, что там всё в порядке. Наш набор про квартиры, кстати, был точно такой же. Там площади были дробные.
- Если сами хотите изменить формат исходного файла, то тогда нужно менять и параметры функции импорта.

Полезные ссылки

1. Блокнот
2. 50 оттенков matplotlib — The Master Plots
3. Предварительная обработка данных
4. Инструкция для установки Anaconda и JupyterLab