Домашнее задание: Практическая работа 2

Дисциплина	Применение ML в кибербезопасности
Тема	Применение статистики в ML и визуализация данных
Форма проверки	**Домашнее задание с автопроверкой** *Совет: выполняйте домашнее задание сразу после изучения темы*
Имя преподавателя	Юрий Иванов
Время выполнения	2 часа
Цель задания	1. Научиться исследовать данные из открытых источников 2. Научиться методам визуализации полученных данных
Инструменты для выполнения ДЗ	Для выполнения задания используйте jupyter notebook или google colab. А также библиотеки request и bs.
Правила приема работы	Для выполнения задания прикрепите ссылку на ваше исследование (ссылка на jupyter notebook в github или google colab) Важно: убедитесь в том, что по ссылке есть доступ.
Чек-лист самопроверки	**Задание считается выполненным, если:** - выполнены все этапы задания - получены все требуемые данные - ссылка на исследование размещена в ответ на задание - к заданию открыт доступ **Задание не выполнено, если:** **-**- выполнены не все этапы задания - получены не все требуемые данные - ссылка на исследование не размещена в ответ на задание - к заданию закрыт доступ
Дедлайн	*24 сентября*

Описание задания:

Этап 1

Установите jupyter notebook либо используйте google colab

Этап 2

Скачайте набор данных Wine:

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Набор состоит из 2 датасетов: для красного и белого вина.

Каждый датасет содержит:

- результаты химического анализа вин - Признаки (их 12)
- оценку качества вина - Метки (в диапазоне от 0 до 10)

Этап 3

Задача: **сбор, предварительная обработка и анализ данных**

1. Загрузите данные о красном вине
2. Проверьте на наличие недостающих данных (достаточно проверить на наличие пропусков)
3. Преобразуйте метки к бинарным классам считая, что «хорошее» вино имеет качество 6 и выше. Для этого необходимо добавить в датафрейм новый столбец с бинарными классами 0,1.
4. Рассчитайте и постройте графики:

   - найдите количество выбросов по столбцу Качество. Для этого найдите точки данных с экстремально высокими или низкими значениями; рассчитайте 25-й и 75-й процентили; вычислите диапазон выбросов, используя межквартильный диапазон 1,5 * (Q3 - Q1)

   - удалите найденные выбросы

   - постройте график распределения по Качеству (distplot)

   - постройте график и определите по нему баланс бинарных классов, используя бинарные метки

   - найдите медиану по каждому признаку

   - постройте график “ящик с усами” по показателю качества

Этап 4

Задача: **визуализация**

- постройте графики распределений значений каждого из 12 признаков
- постройте матрицу корреляции между признаками

Этап 5

Внести полученные в результате расчетов ответы в автоматический тест по теме, размещённый на платформе.

Внимание! Все ответы для теста вычисляются после удаления выбросов

Этап 6

Прикрепите ссылку на google колаб или на github в ответ на задание
Проверьте доступ по ссылке