ПРАКТИКУМ ПО ПРОГРАММИРОВАНИЮ

ЗАДАНИЕ 6

(ДОМАШНЕЕ ЗАДАНИЕ)

Результатом выполнения задания является **архив (или репозиторий) с файлами**.

Допускается выполнение задания в интерактивных средах (Jupiter Notebook, Google Colaboratory).

Допускается выполнение задания на компьютере в одной из средств разработки (JetBrains PyCharm, Visual Studio Code).

Обращаю внимание: задание состоит из последовательности действий, пронумерованных числами. Предполагается после выполнения каждого пункта выводить результат (любым удобным для вас методом, но, чтобы результат каждого пункта был виден).

Желаю успехов!

ХОД РАБОТЫ НАД ЗАДАНИЕМ 6

1. Найдите для себя датасет. Наборов данных очень много, поэтому повторения в группе запрещены. Искать наборы данных в классическом формате csv вы можете на таких ресурсах, как UCI , Kaggle и других.
2. Создайте .md файл (Markdown) с описанием датасета. Кратко опишите, какой смысл имеет каждый столбец данных. Старайтесь использовать возможности формата Markdown (заголовки, подзаголовки, списки и т. д.).
3. Импортируйте датасет в ноутбук средствами pandas. Посмотрите на первые 5 и последние 5 строк датафрейма. Выведите информацию о типах данных в наборе, а также основную статистическую информацию о данных. Удалите дубликаты строк.
4. Удалите из датафрейма (если есть соответствующие поля) такие поля, как ID объектов. Переименуйте произвольное поле датафрейма.
5. С помощью библиотек matplotlib, seaborn и pandas постройте следующие диаграммы:
    a. гистограмму распределения любого числового признака (используется один числовой столбец);

b. диаграмму «ящик с усами» для любого числового признака (используется один числовой столбец);

c. круговую диаграмму (выделите номинативный признак, например пол, и используйте количество соответствующих строк);

d. тепловую карту со значениями взаимной корреляции между всеми парами признаков набора данных (значения корреляции должны быть выведены, размер рисунка подобрать так, чтобы было читаемо);

e. диаграмму countplot с группировкой по двум номинативным признакам.

Сохраните каждую визуализацию как изображение в формате .png.

6. Проверьте наличие пропусков в данных. Если в каких либо столбцах присутствуют пропуски - заполните их в соответствии со следующим правилом:

   a. если значением признака является целое число, заполните значением медианы по данному столбцу;

   b. если значением признака является действительное число, заполните средним значением по данному столбцу;

   c. иначе заполните значением моды по данному столбцу.

7. Выберите произвольный числовой признак, с помощью среза данных выберите не более 200 значений соответствующего признака. Проверьте полученную выборку с помощью normaltest на нормальность распределения.

8. Выполните one-hot кодирование всех категориальных признаков.

9. Сохраните предобработанный набор данных в файл csv.

В рамках данного задания проверяется **архив, состоящий из следующих файлов**:

- файл с датасетом (данные в формате csv);
- файл с описанием датасета (документ в формате .md);
- notebook в формате .ipynb или файл с исходным кодом .py;
- файл с предобработанным датасетом (данные в формате csv);
- 5 изображений в формате .png (все визуализации, описанные в задании).

ОСОБЕННОСТИ ВЫПОЛНЕНИЯ ДОМАШНЕГО ЗАДАНИЯ

С практической точки зрения, домашнее задание представляет собой выполнение задания 6.

Необходимо **составить отчет**, в котором описать ход работы над заданием 6 (полный цикл разведывательного анализа данных).

Отчет включает в себя титульный лист, содержание, введение (1 страница), основная часть, заключение (0.5 страницы), список использованных источников (не менее 4, можно web-сайты), приложения (по желанию).

В **введении** описываете, почему анализ данных это круто и актуально. Рассказываете об инструментах (библиотеках), которые изучили. Для чего они и как помогают в анализе данных.

Основная часть включает в себя следующие разделы:

1) Поиск и загрузка данных (здесь описываете какой датасет выбрали, где его нашли, показывайте скрин .md файла, рассказываете и показываете, как загружали датасет в ноутбук и как он выглядит).
2) Разведывательный анализ данных (здесь описываете и показываете визуализации; **по каждой визуализации должен быть вывод**).
3) Предварительная обработка данных (здесь описываете работу над проверкой наличия пропусков в данных, процесс заполнения пропусков; приводите скриншоты того, что получилось; показываете, как сохраняете предобработанные данные).

В **заключении** описываете, что было сделано в рамках задания.

Напоминаю **о необходимости правильного оформления отчета в соответствии со стандартами**, принятыми на кафедре (оформление аналогично курсовому проекту).

Правила оформления источников ищите здесь.

Не забудьте, что на каждый источник в тексте отчета должна быть ссылка (вида [1] в конце приложения, где 1 - номер источника). Источники в списке располагаются в порядке их упоминания в тексте отчета.

Для проверки выполнения домашнего задания отчет показывается преподавателю в электронном виде, титульный лист печатается. Его **необходимо подписать** у меня (или Ивана Владимировича), после чего загрузить в электронный кабинет до момента аттестации группы.

ПОСЛЕДНЕЕ СЛОВО ГУНЕНКОВА

Дорогой студент, дорогой друг (подруга)!

Поздравляю тебя с завершением прохождения практического курса по программированию. Ты должен понимать, что этот курс создан в первую очередь не преподавателем, а таким же энтузиастом. Возможно, мы изучали что-то не то, но вроде как двигались по программе.

Благодарю тебя за честную работу над заданиями и стремление к получению знаний. Можешь быть уверен - теперь ты готов к погружению в машинное обучение. И мы обязательно совершим это погружение в следующем семестре!

С группой МО-211 мы пройдем этот путь вместе.

Если тебе есть, что сказать касательно предлагаемых заданий, то прошу: https://forms.gle/BwB4mPiizECheDd26. Обязательно учту твое мнение в своей дальнейшей работе. Конечно же тебя попросят пройти традиционный опрос по каждой дисциплине и каждому преподавателю. Но вот этот опрос только между нами)

Хорошего тебе отдыха на праздниках. До скорой встречи!