Работа с данными: отбор признаков, преобразование признаков, заполнение пропущенных значений

Домашнее задание 6

Вам предлагается самостоятельно найти датасет на просторах интернетов (kaggle или любой другой источник). Датасет должен содержать минимум 500 записей и иметь минимум три столбца с пропусками в данных. Пропуски должны быть как в числовых, так и в категориальных столбцах. Если нашли крутой датасет, но в нем нет пропусков, на крайний случай можете сымитировать пропуски удалением некоторых значений (только по-честному, случайным образом в коде, чтобы я это видел). Для тех, кто не захочет поискать свой датасет, предлагается взять датасет мобильного оператора из домашнего задания по визуализациям.

Выбранный датасет (название или ссылку) впишите в файл Датасеты, чтобы не было повторений.

Для выбранного датасета нужно организовать пайплайн обработки данных для последующего их использования в обучении модели. Все фичи после обработки должны иметь числовой формат.

Для всех преобразований (скалирование, one-hot-encoder и т.д.), а также для созданных вами новых фич нужно написать обоснование (почему вы решили сделать так, зачем создана фича и т.п.).

Обработку данных организовывать с использованием API пайплайнов в sklearn (любой из рассмотренных способов).

Последним шагом пайплайна вставьте какую-либо модель, например, рассмотренный KNNClassifier. Здесь не будет оцениваться скор, это задание просто на то, чтобы вы потрогали API. Если вдруг успеете по времени пройти какие-то другие модели на лекции, можете поэкспериментировать с ними.

Для созданных классов стоит написать docstring'и, описав их назначение, параметры и прочее. Не забывем про PEP-8, в общем.

Дедлайн: 22 июля, 8:00.

Удачи!