Лабораторная работа №1 (дедлайн 29 апреля 2024 Anywhere on Earth, или 15.00 30 апреля по Москве)

Данные: датасет leetcode (https://www.kaggle.com/datasets/jaydeepagravat94583/leetcode/data).

Задание 0. Предобработайте данные: из нужных колонок удалите ряды с пустыми ячейками, разделите выборку на тренировочное и тестовое множество, векторизуйте тексты, закодируйте (**если это требуется**) зависимую переменную. (1 балл: 0.5 за предобработку для задачи 1, 0.5 за предобработку для задачи 2)

Задание 1: Предскажите сложность (difficulty) по тексту задания (problem_description). Оцените качество предсказаний. (3 балла: 2 за моделирование, 1 за оценку)

Задание 2: Предскажите процент принятых решений (acceptance) по тексту задания (problem_description). Оцените качество предсказаний. (3 балла: 2 за моделирование, 1 за оценку)

Задание 3: Хотя бы для одной из задач 1 или 2 выполните кросс-валидацию и подбор гиперпараметров. Не обязательно делать и то и другое для одной и той же задачи: допустим, вы можете выполнить cross_validate для задачи 1 и GridSearchCV для задачи 2. (2 балла: 1 за кросс-валидацию, 1 за поиск гиперпараметров)

Дополнительные баллы можно получить:
1) За применение нескольких векторайзеров для одной задачи и сравнение их результатов (0.5);
2) За применение нескольких моделей для одной задачи и сравнение их результатов (0.5);
3) За детальный анализ выдачи модели: например, анализ самых “важных” для каждого класса слов по коэффициентам (1).

Формат сдачи: один или несколько .ipynb или .py-файлов. Возможно, вам будет удобнее делать задачи 1 и 2 в разных файлах. Если вы сдаете файлы в формате .py, пожалуйста, выгрузите картинки, если вы их рисовали, и сдайте их вместе с решениями в любом формате.

Куда сдавать: в форму https://forms.yandex.ru/u/661bf2b643f74f8dcd3b8aa3/ либо в тг Даниилу @letit66. **При сдаче через форму, пожалуйста, пишите имя и фамилию в названии файла.**