Контрольное задание

Компетенция «Машинное обучение и большие данные»

**Общее описание**

В эти три для мы с вами познакомимся с задачами машинного обучения, с которыми могут столкнуться специалисты в реальной жизни: от загрузки и предобработки данных и до вопросов комментирования кода.

**Модуль 3. Машинное обучение.**

В данном модуле участники должны построить модели машинного обучения, а также создать пайплайн предобработки для этих моделей и подготовить отчет о проделанной работе.

В данном модуле участники получают баллы **исключительно за точность полученных алгоритмов на тестовой выборке!** Приведенные ниже в задании шаги являются рекомендательным планом действий. Участники имеют право отклоняться от него.

Входные данные:

1. В архиве Stud\_IT\_input находятся следующие файлы: **train\_data.csv**, **test\_data.csv**, **train\_shapes.csv**, **test\_shapes.csv**.

Подробное описание задания:

1. Загрузите из папки **training\_datasets** файл **train\_data.csv**.
   1. Изучите данные, проведите анализ и обработку признаков.
   2. Разбейте датасет на обучающую и валидационную выборки.
   3. Обучите как минимум 3 алгоритма машинного обучения для предсказания правильного класса (**label**).
   4. Проведите оценку точности обученных алгоритмов, оцените время обучения для каждого из алгоритмов.
   5. Подберите оптимальные гиперпараметры для наилучшего алгоритма.
   6. Выполните предсказание на тестовой выборке. Результат сохраните в формате, полностью идентичном формату файла **sample\_submission.csv. Файл с предсказанием сдается на проверку вместе с отчетом, с названием, описанным ниже. Важно! При несовпадении форматов ваша точность будет приравнена к нулю! После сохранения крайне рекомендуется проверить идентичность форматов!**
   7. **Данная задача имеет систему оценки в зависимости от преодоления барьеров по точности. Список барьеров по метрике accuracy: 0.65, 0.8, 0.88, 0.95, 0.975**
2. Загрузите из папки **training\_datasets** файл **train\_shapes.csv**.
   1. Изучите данные, проведите анализ и обработку признаков.
   2. Разбейте датасет на обучающую и валидационную выборки.
   3. Обучите как минимум 3 алгоритма машинного обучения для предсказания правильного класса (**shape**).
   4. Проведите оценку точности обученных алгоритмов, оцените время обучения для каждого из алгоритмов.
   5. Подберите оптимальные гиперпараметры для наилучшего алгоритма.
   6. Выполните предсказание на тестовой выборке. Результат сохраните в формате, полностью идентичном формату файла **sample\_submission\_shapes.csv. Файл с предсказанием сдается на проверку вместе с отчетом, с названием, описанным ниже. Важно! При несовпадении форматов ваша точность будет приравнена к нулю! После сохранения крайне рекомендуется проверить идентичность форматов!**
   7. **Данная задача имеет систему оценки в виде рейтингового списка. Наибольший балл получает участник с лучшей точностью, наименьший – с худшей. Используемая метрика – f1\_weighted.**
3. Подготовить отчет по модулю (отчет должен быть в формате HTML, отчет должен быть структурирован, описывать ход работы, содержать комментарии).

Выходные данные по модулю:

1. Отчет в формате HTML.
2. Файл .ipynb, **полностью исполняемый и без выпадающих ошибок**.
3. Файл M3\_submission1\_{lastname}\_{first name}.csv с предсказаниями из части 1 задания
4. Файл M3\_submission2\_{lastname}\_{first name}.csv с предсказаниями из части 2 задания

Отчет должен иметь строго определенные название:

M3\_{lastname}\_{first name}

(например, M3\_Ivanov\_Ivan)