Спринт 7. Машинное обучение. Задание

Основные задания

Если выполнены все задания, вы получаете **20 баллов**.

1. Выбрать два любых датасета (один для регрессии, другой - для классификации). Искать датасеты рекомендуется на Kaggle, но можно искать в других местах. Данные должны быть структурированы, представимы в виде таблиц. **Важно:** данные должны быть посвящены таким темам, как производство (не обязательно продуктов нефтепереработки), нефтехимия, статистика выхода из строя оборудования и т. д. Данные будут использоваться в финальныом проекте.
2. Загрузить датасет средствами pandas.
3. С помощью matplotlib и seaborn сделать 5 любых визулизаций на ваш выбор (5 на оба датасета). Используйте данные, которые нашли. В рамках одного датасета виды визуализаций не должны повторяться (например, вы можете сделать 2 круговые диаграммы - это будет считаться за 2 визуализации; но эти круговые диаграммы должны быть выполнены для разных датасетов).
4. Выполнить предварительную обработку данных. Если в данных присутствуют пропуски - заполнить (или удалить) их. Привести все категориальные признаки к числовым.
5. Решить задачи регрессии и классификации на выбранных датасетах (вы сами выбираете целевой признак). В рамках каждой задачи обучить не менее 3 моделей (без нейронных сетей). Для каждой из них выполнить подбор гиперпараметров (хотя бы 1 параметр для каждой модели подобрать). Качество моделей не повлияет на баллы. Каждую модель оценить с использованием соответствующих метрик.
6. Решить задачи с помощью глубоких нейронных сетей (для каждой задачи построить архитектуру сети, скомпилировать, обучить сеть и проверить качество с помощью метрик).

Дополнительные задания

1. Решить задачу классификации на своем датасете с использованием собственной реализации KNN. Реализация должна быть выполнена в виде класса. Должны быть методы fit() и predict(). **3 дополнительных балла**.
2. Попробовать использовать модели градиентного бустинга из библиотек XGBoost и CatBoost. Решить с использованием каждой модели каждую задачу (то есть всего 4 модели обучить). Оценить качество полученных моделей. Подобрать параметры. **2 дополнительных балла**.