**Отчет по проекту обработки и анализа данных: пример машинного обучения для набора данных классификации качества вина**

## 1. Понимание бизнеса

### Постановка задачи

Цель проекта— построение модели машинного обучения для определения качества вина на основе его характеристик.

Набор данных для этого проекта взят с платформы [kaggle](https://www.kaggle.com/datasets/rajyellow46/wine-quality).

### Метрика

Производительность моделей машинного обучения будет оцениваться на тестовом наборе. Точность измеряется и сообщается с помощью accuracy и classification report. Accuracy > 0.8 будет считаться приемлемым и пригодным для развертывания.

## 2. [**Сбор и понимание данных**](https://github.com/Azure/MachineLearningSamples-TDSPUCIAdultIncome/tree/master/code/01_data_acquisition_and_understanding)

### Необработанные данные

Всего существует 6 497экземпляра (до любой фильтрации), сочетание непрерывных и дискретных (train = 4547, test = 1950)

ЦЕЛЬ: оценка качества вина от 3 до 9.

Характеристики: фиксированная кислотность, летучая кислотность, лимонная кислота, остаточный сахар, хлориды, свободный диоксид серы и т.д.

## 3. [**Моделирование**](https://github.com/Azure/MachineLearningSamples-TDSPUCIAdultIncome/tree/master/code/02_modeling)

### Проектирование функций

**Очистка данных и заполнение пропусков. Были проанализированы пропуски в признаках. Для заполнения была использована медиана, которая позволила сохранить распределение данного признака.**

**Категориальные признаки. Признак тип вина представлен в виде категориальной переменной со значениями «red» и «white». С помощью pandas.get\_dummies() признак был закодирован в one-hot представлении.**

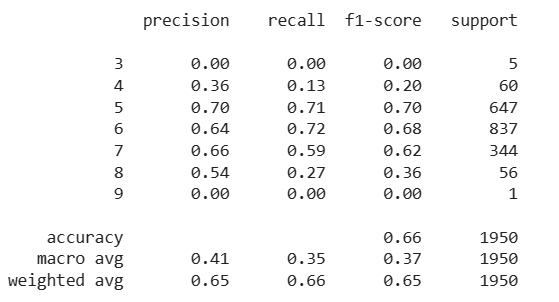
**Анализ выбросов.** Выбросов в данных не было обнаружено, поэтому никаких изменений не было произведено.

### Обучение моделированию

Была создана модель классификации с помощью библиотеки xgboost со 100 оценщиками и максимальной глубиной 7.

### Оценка модели

Точность модели измерялась на тестовом наборе с помощью метрики accuracy, а также был изучен classification report.



## 4. Результаты

В результате исследования не удалось построить достаточно емкую модель, которая позволила бы оценить качества вина. Связано это с переобучением модели, а также сильным дисбалансом классов в выборке.