## Отчет по исследованию

### Введение

Была поставлена задача исследовать, как физико-химические свойства португальского вина марки «Vinho Verde» влияют на его качество. Регрессионная задача решалась в рамках выполнения домашнего задания по курсу «Машинное обучение».

Из-за проблем конфиденциальности и логистики доступны только указанные параметры, и нет данных о типах винограда, марке вина, цене продажи и т.д.

### Описание проблемы и набор данных

Данная проблема открыта для решения на платформе Kaggle в качестве упражнения для начинающих специалистов в области Data Science.

Набор данных представляет собой 12 признаковых переменных (плотность жидкости, содержание алкоголя, количество лимонной кислоты, остаточных сахаров и т.д.) и 1 целевую переменную (качество продукта), которую необходимо предсказать. Сведения о единицах измерения признаковых переменных отсутствуют, значения целевой переменной находятся в диапазоне от 0 до 10.

Всего в наборе 6497 экземпляров. Для построения модели и оценки результатов набор данных был разделен на обучаемую и тестовую выборки. Тестовая выборка составляет 20% от всего набора данных.

## Подготовка данных

Для решения задачи был проведен EDA для выявления маловажных признаков, построена корреляционная матрица для выявления зависимостей между признаками, произведено преобразование категориальных переменных в пригодный для компьютерной обработки числовой вид, а также присутствующие в наборе данных пропуски заменены медианными значениями.

## Построение модели

Для построения модели и решения задачи регрессии был выбран нелинейный алгоритм RandomForestRegressor, т.к. он наиболее подходит для решения подобных задач, в данных отсутствуют выбросы и зашумление, а количество экземпляров достаточно велико.

При построении модели использовались стандартные параметры, за исключением случайного начального состояния.

## Результаты

Для оценки результатов была выбрана метрика RMSE, т.к. она наиболее эффективно позволяет оценивать результаты предсказаний в случае, когда числовые значения данных находятся в маленьких диапазонах.

Полученная оценка составляет 0.63, что говорит о достаточно эффективной работе выбранного алгоритма.

## Возможные улучшения модели

Несмотря на то, что выбранная метрика для оценки работа алгоритма показала хороший результат, думаю, что, изменяя параметры алгоритма (например, максимальную глубину дерева, максимальное количество листьев), результат может быть улучшен.

При необходимости дальнейших исследований для автоматизации подбора параметров можно воспользоваться алгоритмами GridSearchCV или RandomSearchCV.