**ОТЧЕТ  
о результатах исследования химического анализа красных и белых вин и экспертной оценки их качества**

1. Постановка задачи и данные  
     
    К исследованию предлагаются данные, содержащие результаты химического анализа ряда красных и белых вин Португалии, а также экспертную оценку качества каждого из них.   
    Данные представлены единым массивом, признаковое пространство содержит тип вина «белое/красное», оценку качества «quality» и измерения ряда химических параметров.  
    Вопросы от заказчика:  
    а. разработать алгоритм определения признаков, важных для решения задачи предсказания качества вина по его химическим параметрам  
    б. разработать алгоритм выявления выбросов в данных для определения качества вина «excellent/poor»  
    в. разработать алгоритм предсказания качества вина по его химическим параметрам
2. Результаты исследования  
     
    2.1 Определение важности признаков проводилось методами корреляционного и многофакторного дисперсионного анализа, которые позволили утверждать, что для определения качества вина значимыми являются следующие характеристики:  
    а. для белого - fixed acidity, volatile acidity, chlorides, residual sugar, sulphates, alcohol;  
    б. для красного - fixed acidity, volatile acidity, chlorides, total sulfur dioxide, sulphates, alcohol.  
     
    2.2 Определение выбросов проводилось с применением метода машинного обучения IsolationForest в два этапа – сначала исследовалось влияние выбросов в отдельных, наиболее значимых признаках volatile acidity  и chlorides, затем использовалось все значимое признаковое пространство. В результате, удалось выяснить, что:  
    а. использовать для определения вин с выдающимися характеристиками выбросы по отдельным признакам не верно, ввиду комплексного характера оценки качества вин;  
    б. анализ выбросов с использование значимого признакового пространства позволяет дать примерную оценку качества вина, при этом, учитывая субъективный характер экспертной оценки, результат анализа характеристик вина математическими методами не будет отражать картину качества вин с точки зрения потребителей.  
     
    2.3 Предсказание качества вина по его химическому анализу производилось с применением метода классификации RandomForestClassifier, не чувствительного к дисбалансу целевой переменной «quality» (вин среднего качества значительно больше, чем вин отличных и плохих). Было достигнуто качество предсказания 0.69 для красных вин и 0.72 для белых. Основные трудности наблюдались в предсказании качества вин на границах классов качества, там, где характеристики вин формально разного, но близкого качества, могут быть сходными.
3. Выводы  
   а) актуальные для наборов данных признаки:   
    - красное вино - fixed acidity, volatile acidity, chlorides, total sulfur dioxide, sulphates, alcohol  
    - белое вино - fixed acidity, volatile acidity, chlorides, residual sugar, sulphates, alcohol   
   б) анализ выбросов по нескольким значимым признака одновременно, в целом, позволяет сделать несколько выводов:   
    - экспертная оценка качества отражает либо преобладание значения одного параметра, либо некоторое сочетание значений нескольких параметров   
    - экспертная оценка качества сильно снижается при увеличении значений «volatile acidity» и «chlorides», либо одного их этих параметров  
    - качественная оценка вина «строгими» методами плохо отражает картину качества вин с точки зрения потребителей  
   в) в результате работы алгоритма классификации лучшее качество предсказания целевой переменной «quality» достигается для "средних" значений 5, 6 и 7, что объясняется наличием большего количества наблюдений с этими значениями «quality» и меньшей дисперсией значений важных признаков в этих наблюдениях  
   г) из матрицы ошибок видно, что наиболее проблемные зоны - границы классов качества вина  
   д) классы с малым количеством наблюдений не определились, что говорит о необходимости собирать как можно большее количество наблюдений

Ссылки на дополнительные материалы:  
а) ноутбук с описанием деталей решения, алгоритмами и графиками  
 <https://colab.research.google.com/drive/14X91SybLqh7kqRA-eIpi3bqKj3T6f5ji?usp=sharing>  
б) описание задачи на KAGGLE  
 <https://www.kaggle.com/rajyellow46/wine-quality>