**Отчет по проекту:**

Проект Классификации вин

1. Введение
   1. Описание проекта

В данном датасете представлены физико-химические свойства красных и белых сортов вин "Vinho Verde" [ref.Cortez et al., 2009]. В целях скрытия конфиденциальной информации все названия вин, мест производства, отпускных цен, а также какие-либо идентификационные особенности были удалены.

В качестве целевой переменной было предложено использовать показатель «качество», полученный по индивидуальным вкусовым ощущениям.

* 1. **Цели проекта**

Цель проекта была декомпозирована на две, по стоимостному приоритету:

1. За наилучшие вина можно получить сверх высокую прибыль.

Необходимо создать алгоритм определения наилучших вин. В качестве лучших использовать quality=9.

1. Необходимо определять прогнозируемую прибыль.

Необходимо создать алгоритм определения средних по качеству вин

1. Объединить два вышеуказанных алгоритма в один для унификации в применении.
2. Порядок решения задачи
3. Исследовать имеющиеся признаки,
4. Определить скрытые признаки (из данных в свободном доступе)
5. Определить коррелирующие признаки
6. Ввиду несбалансированного целевого класса сгенерировать новых представителей целевого класса.
7. Подобрать модель ML, NN. (Критерием подходящего результата были выбраны метрики точности precision>50%, recall=100%.)
8. Сравнить результаты, выбрать максимальный.
9. Чем проект отличается от остальных

Генерацией новых экземпляров несбалансированного класса.

На **рисунке 1** представлено соотношение количества элементов каждого класса вин. Класс «высокого качества» Состоит из 5 экземпляров.



**Рисунок 1**. Соотношение количества элементов каждого класса

Задача поиска миноритарного класса в несбалансированных классах (задача поиска иголки в стоге сена) решалась в следующем порядке:

* Определить по каким признакам «иголка» отличается от «стога сена».
* Для качественного разделения нужно больше элементов. Определить, как порождать новые иголки.

В результате:

* Выделено два статистически значимых признака «иголок»
* Новые «иголки» создавались из «сена» (medium quality) + замена ключевых признаков

1. Определение важных признаков

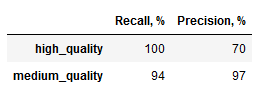
Определялись критерии «качественного» вина согласно данным из открытых источников.

* Каждый из признаков в датасете был конкретизирован, статистически проверен
* Все признаки объединены в две категории «positive» и «negative» и проведен детальный статистический анализ внутри каждой категории, а также на влияние на целевую переменную

В результате проведенной работы были определены три важных признака, которые использовались при выборе алгоритма классификации.

1. Результат

Ввиду малого количества элементов класса «high\_quality», в обучающую выборку было добавлено 500 сгенерированных элементов (10% обучающей выборки) и 100 элементов (10%) в вариационную выборку. В отложенную выборку были включены только оригинальные элементов класса «high\_quality».

Созданный алгоритм имеет следующие метрики точности на отложенной выборке (**Таблица 1**.)

**Таблица 1**. Метрики точности финального алгоритма