Определение качества вина

**1. Введение**

Качество вина – это его соответствие нормативным показателям по химическому составу, окраске, прозрачности, аромату и вкусу.

Определение качества вина позволяет не только отличить вино от подделки, но и позволит выбрать хорошее вино по совокупности критериев.

Этот отчет составлен следующим образом. В разделе 2 описывается проблема, а также набор данных, используемый для обучения и тестирования нашего алгоритма. В разделе 3 описываются этапы предварительной обработки данных. В разделе 4 описываются различные модели обучения, которые были применены к определению качества вина, а в разделе 5 представлены результаты работы. Возможные пути улучшения модели указаны в разделе 6.

**2. Описание проблемы и набор данных**

Набор данных относится к красному и белому вариантам португальского вина «Vinho Verde».

Из-за конфиденциальности и логистических проблем доступны только физико-химические (входные данные) и сенсорные (выходные данные) переменные. Окраска, прозрачность, аромат и вкус вина в наборе данных отсутствуют.

В этой работе показана возможность оценки качества вина только по его физико-химическим качествам.

Подробная информация о размере набора данных отображена в таблице 1.

Таблица 1. Размерность набора данных

| Вид данных | Число записей | Количество признаков |
| --- | --- | --- |
| Тестовый | 4547 | 11 |
| Тренировочный | 1950 | 11 |

**3. Предварительная обработка данных**

Перед построением модели категориальная переменная о цвете вина была преобразована в числовую при помощи использования класса LabelEncoder().

Пропущенные значения (их общее количество – 38) заменены средними значениями по переменной для всех наблюдений в наборе данных.

**4. Модели обучения**

При обучении модели были использованы следующие модели обучения: Logistic Regression, Decision Tree Classifier, Stacking Classifier.

В свою очередь Stacking Classifier состоял из Decision Tree Classifier, Linear SVC, KNeighbors Classifier. Для объединения базовых оценок использована модель Logistic Regression.

**5. Результаты работы**

В этом разделе мы сравниваем относительную производительность трёх реализованных моделей обучения в нашем наборе данных. На рисунке 1 показаны качество обучения различных моделей в наборе данных.

**6. Улучшение модели**

Для улучшения работы модели планируется провести следующую предварительную обработку данных: сбалансировать классы целевой переменной, при заполнении пропущенных значений использовать более сложные алгоритмы.

При обучении модели будет оценена необходимость использования каждого признака исходных данных, сформированы новые признаки, позволяющие улучшить показатели модели.

Рисунок 1. Сравнение моделей обучения