**Отчет по проведенному исследованию**

**Введение**

*Целью исследования* является разработка регрессионной модели для предсказания уровня качества вина на основе его физико-химических параметров.

Набор данных загружен с сайта Kaggle (<https://www.kaggle.com/rajyellow46/wine-quality>). В наборе данных представлена информация о красном (25 %) и белом (75 %) вариантах португальского вина Vinho Verde. Набор данных был сформирован в 2009 году.

**Разведочный анализ и подготовка данных**

В результате проведенного разведочного анализа данных было установлено количество пропущенных значений по каждой переменной и порог потенциальных выбросных значений (см. табл. 1).

Таблица 1 – Результаты разведочного анализа данных

|  |  |  |  |
| --- | --- | --- | --- |
| Название переменной | Описание переменной | Количество пропущенных значений | Порог выбросных значений |
| type | тип вина | 0 | --- |
| fixed acidity | постоянная кислотность | 10 | --- |
| volatile acidity | переменная кислотность | 8 | --- |
| citric acid | содержание лимонной кислоты | 3 | >1,2 |
| residual sugar | остаточный сахар | 2 | >40 |
| chlorides | хлориды | 2 | --- |
| free sulfur dioxide | свободный диоксид серы | 0 | >160 |
| total sulfur dioxide | общий диоксид серы | 0 | --- |
| density | плотность | 0 | >1,02 |
| pH | уровень кислотности | 9 | --- |
| sulphates | сульфаты | 4 | --- |
| alcohol | содержание спирта | 0 | --- |
| quality | уровень качества | 0 | --- |

Исходный размер набора данных включал 6 497 значений. В связи с небольшим количеством пропущенных и выбросных значений было принято решение об их удалении, в результате которого объем данных сократился менее, чем на 1 %. Размер преобразованного набора составил 6 459 значений.

В наборе присутствует бинарная переменная type, которая была преобразована в числовую с помощью OneHotEncoder.

Кроме того, была реализована нормировка числовых переменных с помощью метода min-max. Необходимость нормировки была обусловлена существенным различием значений переменных. В результате проведенных операций все предикторы имеют значения в диапазоне от 0 до 1.

На заключительном этапе подготовки данных были созданы обучающая и контрольная выборки в объеме 80 % и 20 % от исходной выборки.

**Построение и оценка модели**

В качестве модели была использована множественная линейная регрессия. Зависимая переменная – качество вина (quality), независимые переменные (предикторы) – все остальные переменные набора данных. Обучение модели производилось с помощью обучающей выборки.

В качестве оценки модели использовалась средняя абсолютная ошибка (MAE). В таблице 2 представлены результаты оценки модели.

Таблица 2 – Результаты оценки модели

|  |  |
| --- | --- |
| Обучающая выборка (Train MAE) | 0,5677 |
| Контрольная выборка (Test MAE) | 0,5675 |

**Обсуждение результатов**

Таким образом, в среднем ошибка составляет 0,5 балла при определении качества вина. Так как качество вина принимает целочисленные значения в диапазоне от 0 до 10, то ошибка в 0,5 может интерпретироваться как допустимая.