PYTHON

АНАЛИЗ ХИМИЧЕСКОГО СОСТАВА ВИН И ПРОГНОЗ ИХ КАЧЕСТВА С ПОМОЩЬЮ МГК И ВИЗУАЛИЗАЦИИ ГРАФИКОВ

Над проектом работали:

Волков Егор
Галстян Ваган
Лим Дмитрий

ЗАДАЧИ

С помощью математических методов и применения наших знаний в языке программирования "Python" получить инструмент, который носит практический характер в сфере производства и продажи винных изделий.

Мы проанализируем имеющийся датасет, в котором указаны химические характеристики различных вин и с помощью метода главных компонент упростим визуализацию и интерпретацию данных, выбрав в качестве "опорных" признаков те, чья корреляция представляется наибольшей.

ПРЕДОБРАБОТКА И ТРАНСФОРМАЦИЯ ДАННЫХ

Эффективная предобработка данных является краеугольным камнем успешного анализа. На этом этапе мы выполняем центрирование и нормирование признаков, что крайне важно для алгоритмов машинного обучения, особенно для тех, которые чувствительны к масштабу данных, таких как PCA и K-Means. Эти операции помогают стабилизировать дисперсию и среднее значение, предотвращая доминирование признаков с большими числовыми диапазонами.

Этот датасет содержит следующие столицы:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxidel
7. total sulfur dioxidel
8. density
9. pH
10. sulphates
11. alcohol
12. quality

КЛЮЧЕВЫЕ ЭТАПЫ:

1. Разделение на признаки и целевую переменную: Отделение независимых переменных (признаков) от зависимой (целевой), что необходимо для обучения моделей.
2. Нормирование данных (StandardScaler): Приведение всех признаков к единому масштабу, что обеспечивает одинаковый вклад каждого признака в анализ и ускоряет сходимость алгоритмов.
3. Кодирование категориальных признаков
4. Преобразование нечисловых данных в числовой формат с использованием One-Hot Encoding, если таковые имеются, что делает их пригодными для анализа

ИССЛЕДОВАНИЕ

Получение статистик по набору данных дает глубокое понимание его структуры, распределения и взаимосвязей между признаками. Это включает описательные статистики, корреляционный анализ, выявление пропусков и проверку типов данных.

ФУНКЦИИ ДЛЯ АНАЛИЗА:

- **get_statistics(data):**
Возвращает описательную статистику, корреляционную матрицу, количество пропусков и типы данных.

- **apply_pca(X_scaled, n_components=None):**
Применяет метод главных компонент для уменьшения размерности данных. Если n_components не указан, PCA вычисляет все компоненты.

- **find_optimal_components):**
Определяет оптимальное число главных компонент, объясняющих заданный порог дисперсии (по умолчанию 95%).

ПРИМЕНЕНИЕ PCA

PCA — это мощный статистический метод для уменьшения размерности данных при сохранении максимально возможной дисперсии. Он преобразует набор, возможно, коррелированных переменных в набор некоррелированных переменных, называемых главными компонентами. Это позволяет выявить наиболее значимые паттерны в данных, снизить вычислительную сложность и улучшить интерпретируемость.

На графике справа показана кумулятивная объясненная дисперсия, которая помогает определить оптимальное количество компонент для сохранения большей части информации в данных.

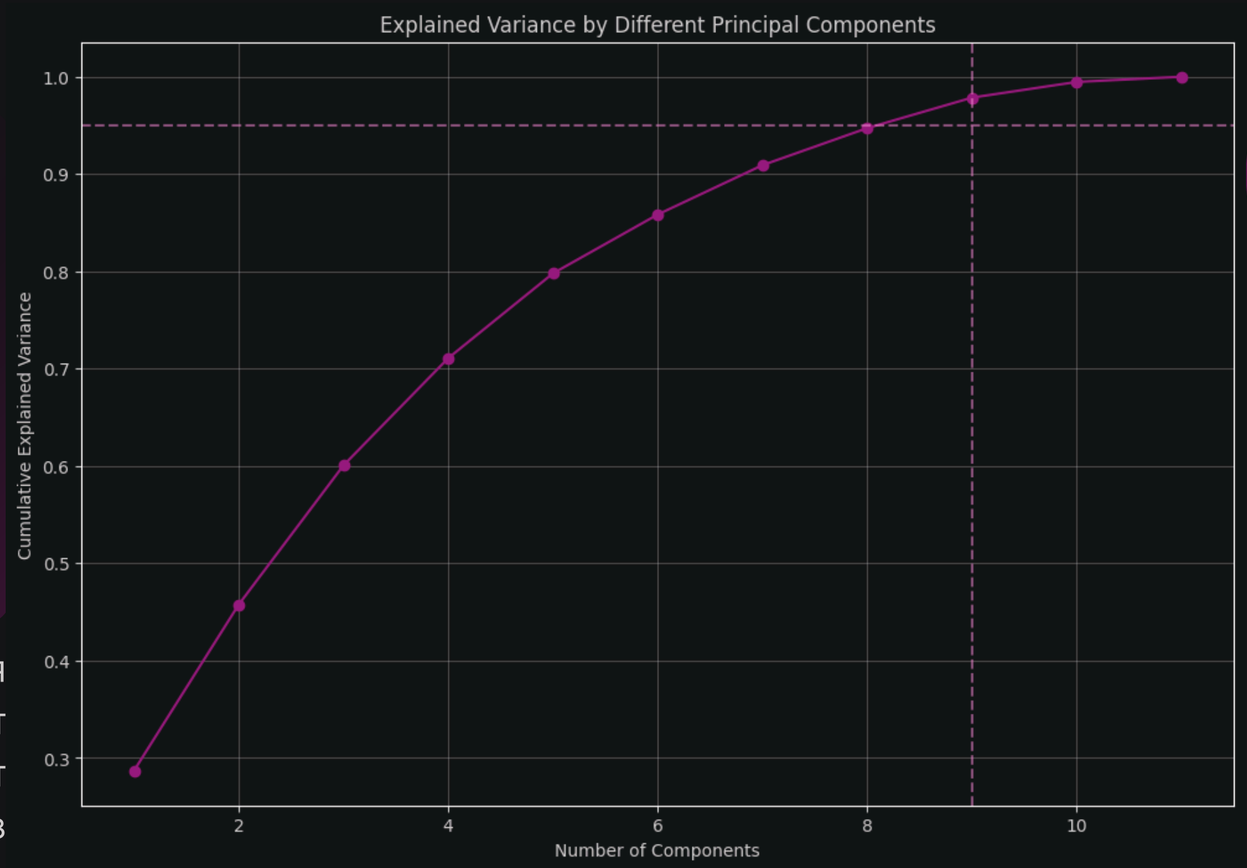
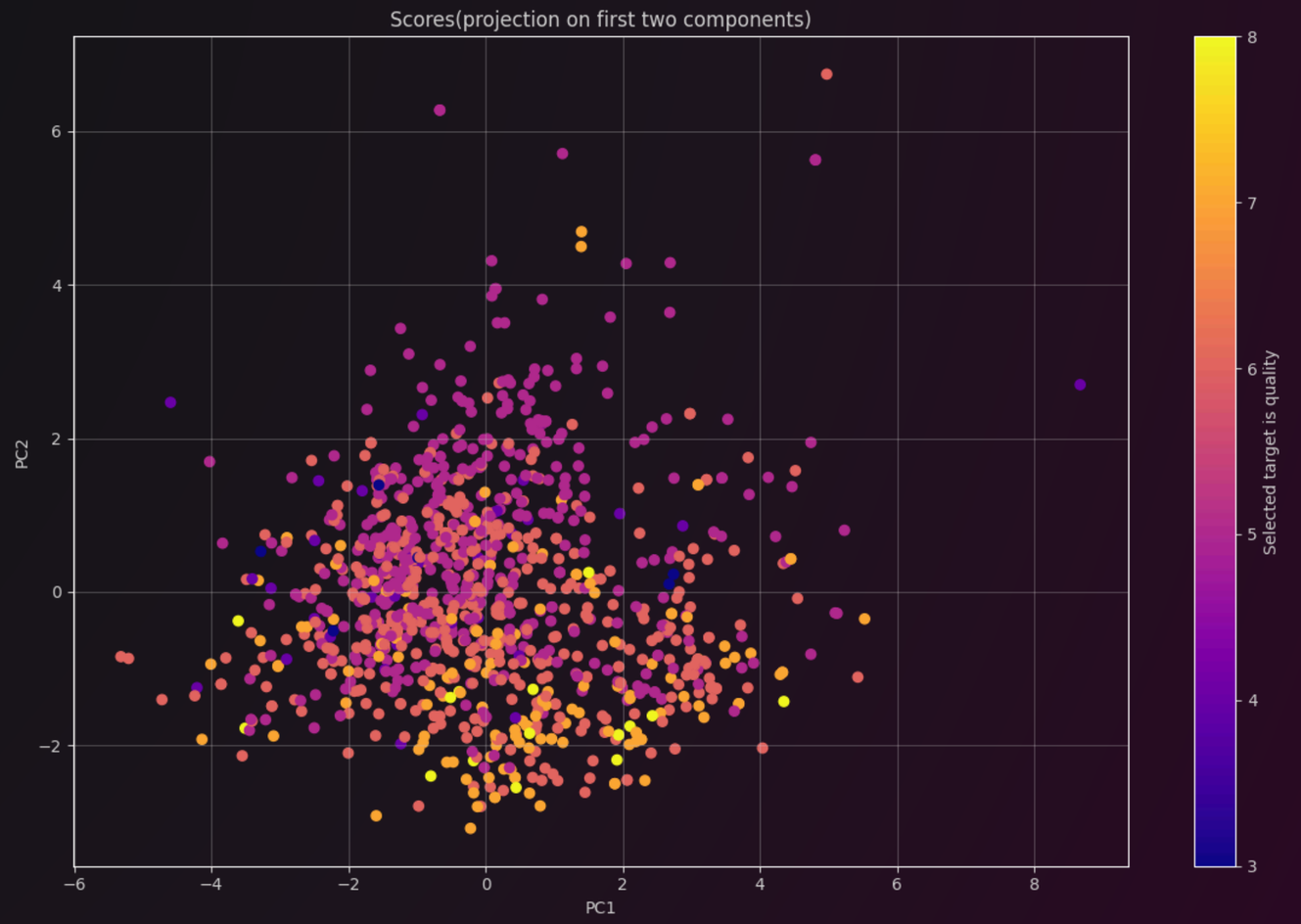


ГРАФИК СЧЕТОВ:

Scores Plot: Отображает объекты в пространстве первых двух главных компонент. Цвет точки отвечает за качество вина. Точки, расположенные близко друг к другу, указывают на схожесть объектов по основным компонентам. Цвет точек отражает качество, что позволяет выявить зависимости.



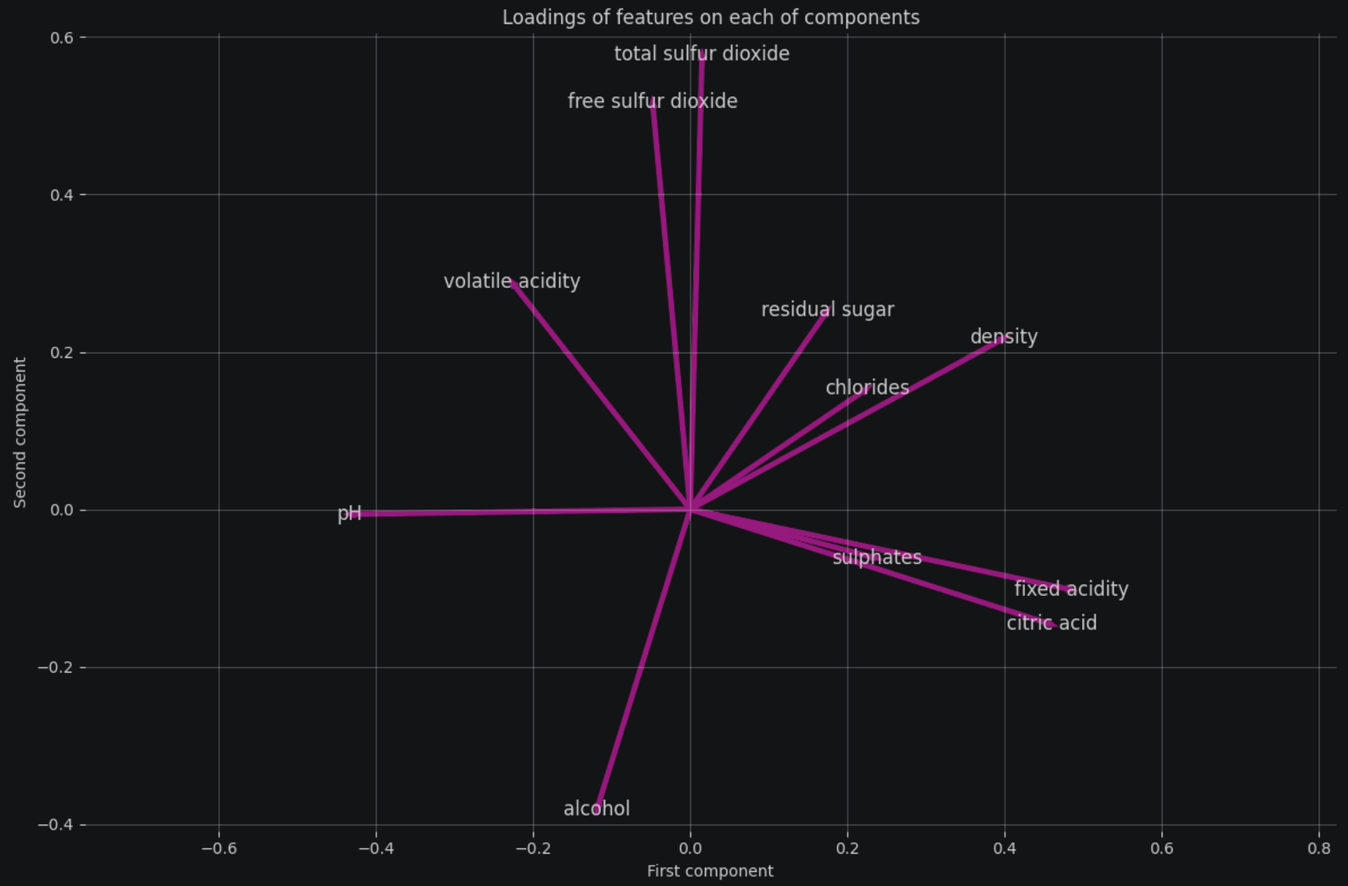


ГРАФИК НАГРУЗОК

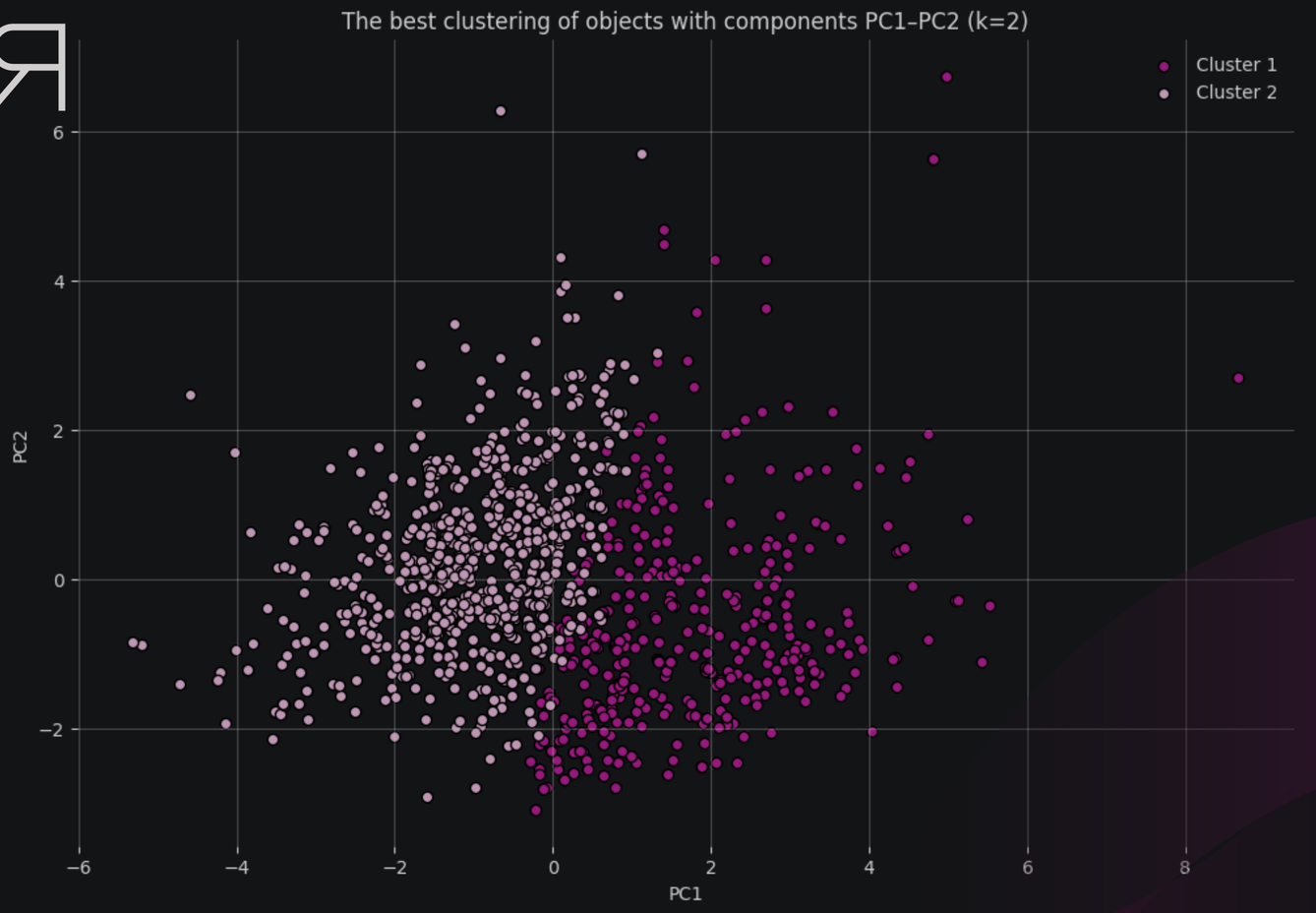
На графике показаны вектора нагрузок признаков на первые две главные компоненты.

- Длина и направление стрелок отражают вклад признаков в формирование компонент(признак может влиять негативно).
- Признаки, направленные в одну сторону, высоко коррелированы.
- Чем длиннее вектор, тем более значимое он оказывает влияние на компоненту.

Например, признаки density, fixed acidity и citric acid сильно влияют на первую компоненту, тогда как volatile acidity и alcohol — в противоположную сторону.

КЛАСТЕРИЗАЦИЯ С K-MEANS:

Алгоритм кластеризации, разбивает данные на k заранее заданных кластеров. Каждая точка данных относится к кластеру с ближайшим средним значением (центроидом), и мы обучаем их на наших двух главных компонента.



ВЫВОД

В результате анализа данных был реализован полный цикл обработки, начиная от нормирования признаков и заканчивая снижением размерности. Использован методом главных компонент (PCA). Проведенная кластеризация датасета WineQT с использованием KMeans на первых двух главных (PC1 и PC2) демонстрирует низкую эффективность из-за ограниченной объясняющей способности этих компонент, при этом дальнейшее увеличение количества главных компонент не увеличивает процент объясненной дисперсии, поэтому расширение графика с кластерами для лучшего разбиения в трехмерный(добавление третьей компоненты) не дает особых изменений. По графику “счетов” можно увидеть, что вина со схожим качеством группируются, но есть некоторые выбросы и нечеткость группировки, так как все точки фактически в одной кучке.

ВЫПОЛНЯЛИ

Волков Егор

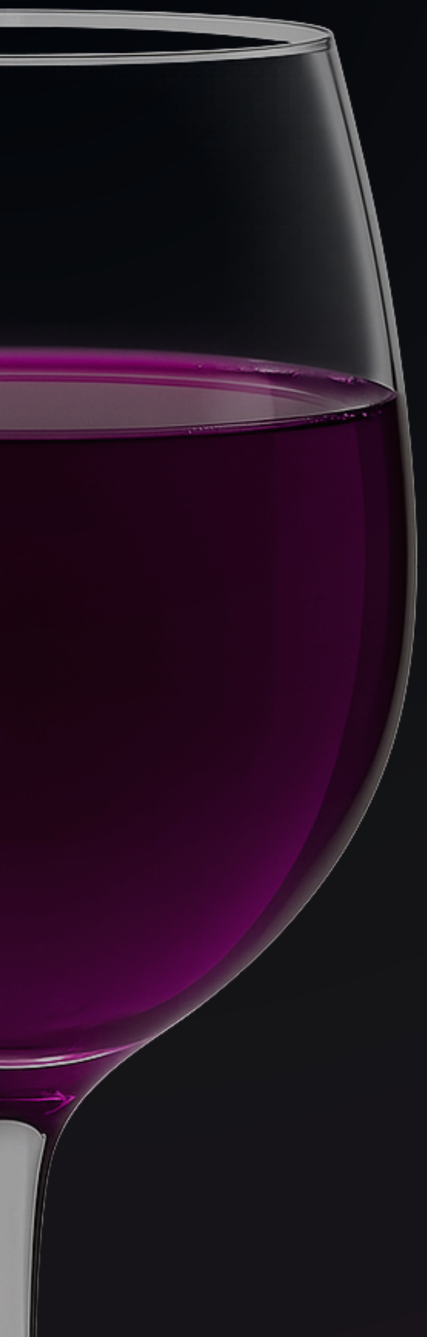
Написал функции для построения графиков счетов и нагрузок, кластеризации, а так же вычисления наиболее эффективного числа для разбиения на кластеры с помощью функции, высчитывающей Silhouette score. Так же написал функцию, которая преоьразует категориальные данные.

Галстян Ваган

Подготовил презентацию, структурировал и разделил материал на логически завершённые смысловые блоки, что обеспечило наглядность и простоту восприятия результатов анализа.

Лим Дмитрий

Подготовил данные (нормировка, удаление лишних столбцов, проверка статистик) и применил PCA для снижения размерности, определив оптимальное число компонент через анализ объяснённой дисперсии, что позволило перейти к кластеризации и визуализации.



СПАСИБО ЗА ВНИМАНИЕ