Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

Лабораторная работа №1 по курсу «Методы машинного обучения в автоматизированных системах обработки информации и управления»

Подготовили:

Чжан Чжиси

ИУ5И-25М

21.03.2024

Проверил:

Гапанюк Ю. Е.

2024 г.

Цель лабораторной работы:

изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание:

Набор данных: набор данных о вине из библиотеки машинного обучения UCI Набор данных содержит 13 признаковых переменных и одну категориальную переменную, представляющую тип вина.Функциональные переменные включают алкоголь, яблочную кислоту, золу, щелочность золы, магний, общие фенолы, флаваноиды, нефлаваноидные фенолы, проантоцианы, интенсивность цвета. Оттенок, OD280/OD315 разбавленных вин

Задание:

1.Выбрать набор данных (датасет). Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

2.Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

- История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
- На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
- Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
- Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
- История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

3. Сформировать отчет и разместить его в своем репозитории на github.

Текст программы

Шаг 1: Распределение классов вин

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_wine

# Load wine dataset
data = load_wine()
X = data.data
y = data.target

# Count the number of samples in each class
class_counts = np.bincount(y)

# Plot the distribution of wine classes
plt.figure(figsize=(8, 6))
plt.bar(np.unique(y), class_counts, color='skyblue')
plt.title('Distribution of Wine Classes')
plt.xlabel('Class')
plt.ylabel('Count')
plt.xticks(np.unique(y))
plt.grid(True)
plt.show()
```

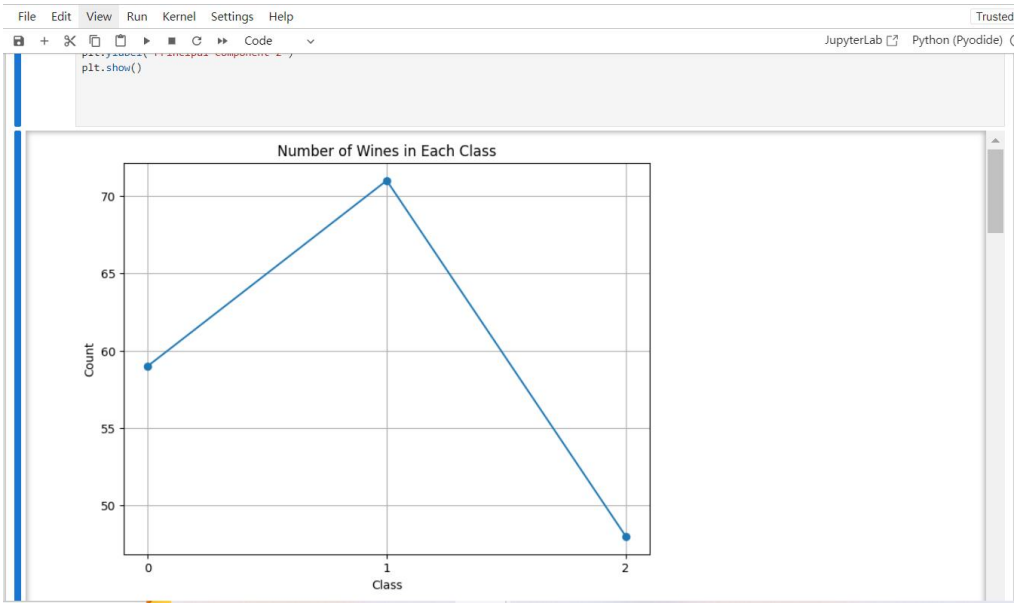

Рис.1-Распределение классов вин

На этом этапе мы покажем распределение объемов выборки для каждой категории вин. Этот график необходим нам для понимания размера выборки для каждой

категории в наборе данных. Мы видим, что категории 0, 1 и 2 имеют 59, 71 и 48 образцов соответственно. Такое относительно сбалансированное распределение является хорошим знаком, поскольку если количество образцов в одной категории значительно меньше, чем в других, это может привести к снижению эффективности модели для этой категории.

Шаг 2: Корреляция между характеристиками вина

```
import pandas as pd

# Convert dataset to pandas DataFrame for correlation analysis
df = pd.DataFrame(X, columns=data.feature_names)

# Calculate correlation matrix
corr = df.corr()

# Plot correlation matrix
plt.figure(figsize=(10, 8))
plt.imshow(corr, cmap='coolwarm', interpolation='nearest')
plt.colorbar()
plt.xticks(np.arange(len(data.feature_names)), data.feature_names, rotation=45)
plt.yticks(np.arange(len(data.feature_names)), data.feature_names)
plt.title('Correlation Between Wine Features')
plt.show()
```

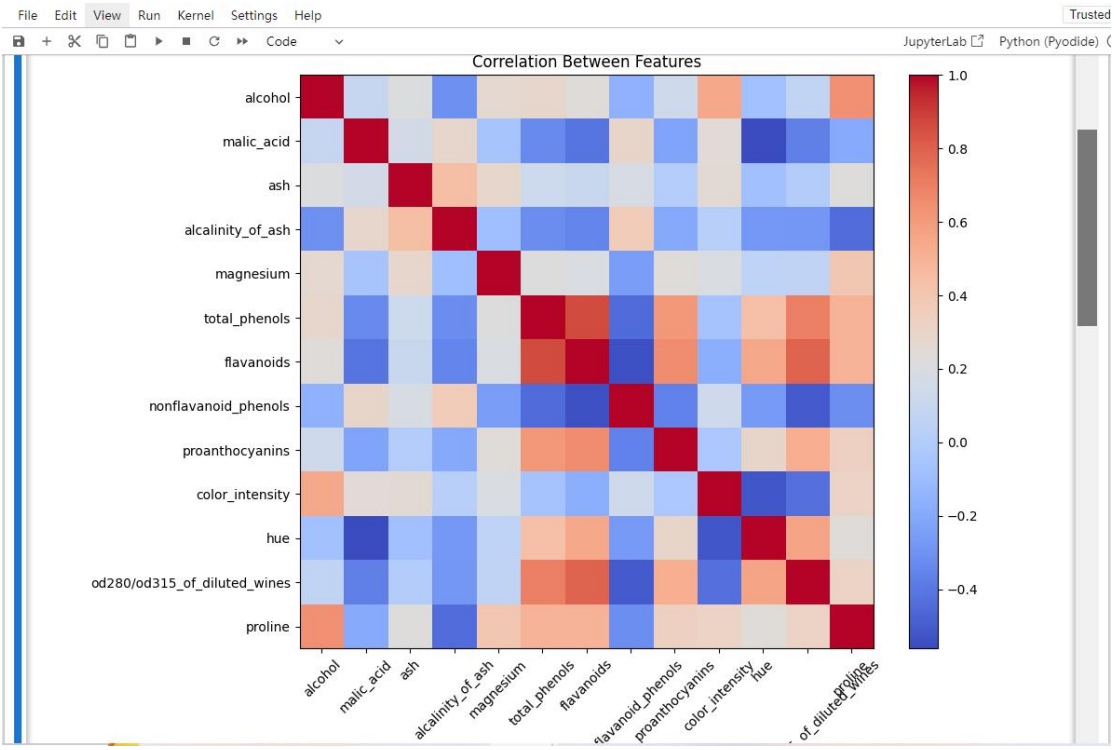

Рис. 2 – Корреляция между характеристиками вина

На этом этапе мы показываем матрицу корреляции между отдельными признаками в наборе данных о вине. Значения в корреляционной матрице находятся в диапазоне от -1 до 1. Значения, близкие к 1, указывают на сильную положительную корреляцию, значения, близкие к -1, - на сильную отрицательную корреляцию, а значения, близкие к 0, - на отсутствие линейной корреляции. Мы наблюдали сильную корреляцию между некоторыми признаками, например, коэффициент корреляции между флаваноидами и общими фенолами составил 0,86, что указывает на сильную положительную корреляцию между ними. Эта корреляция может помочь нам принимать более обоснованные решения при выборе признаков и уменьшении размерности.

Шаг 3: Распределение характеристик вина

```
plt.figure(figsize=(12, 10))
for i in range(X.shape[1]):
    plt.subplot(3, 5, i + 1)
    plt.hist(X[:, i], bins=20, color='skyblue', edgecolor='black')
    plt.title(data.feature_names[i])
plt.tight_layout()
plt.show()
```

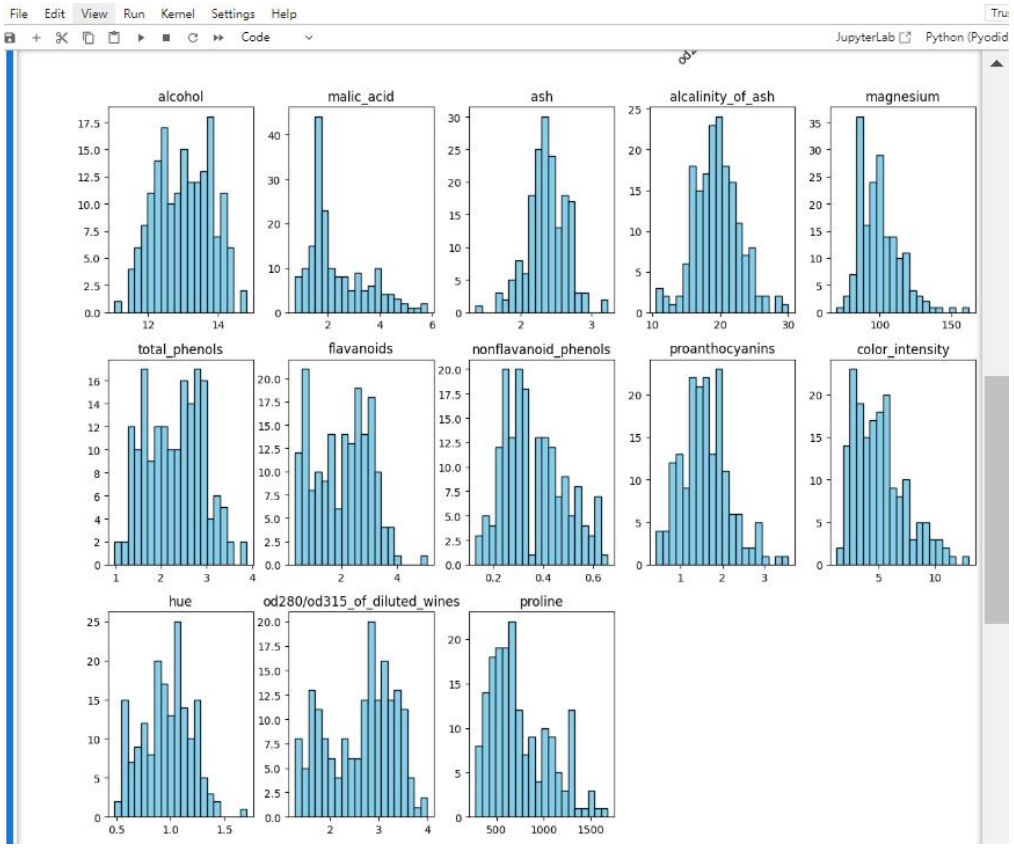

Рис.3– Распределение характеристик вина

На этом этапе мы показываем распределение каждой характеристики вина.

Например, гистограмма характеристики "Алкоголь" показывает распределение этой характеристики в наборе данных. Мы видим, что большинство образцов сосредоточено между 12 и 14 % алкоголя. Такое визуальное представление о распределении характеристик важно для понимания данных и проведения дальнейшего анализа.

Шаг 4: Боксплот по характеристикам вина в зависимости от класса

```
plt.figure(figsize=(10, 6))
for i in range(X.shape[1]):
    plt.subplot(3, 5, i + 1)
    plt.boxplot([X[y == k, i] for k in range(3)], positions=[1, 2, 3], widths=0.5)
    plt.title(data.feature_names[i])
plt.tight_layout()
plt.show()
```

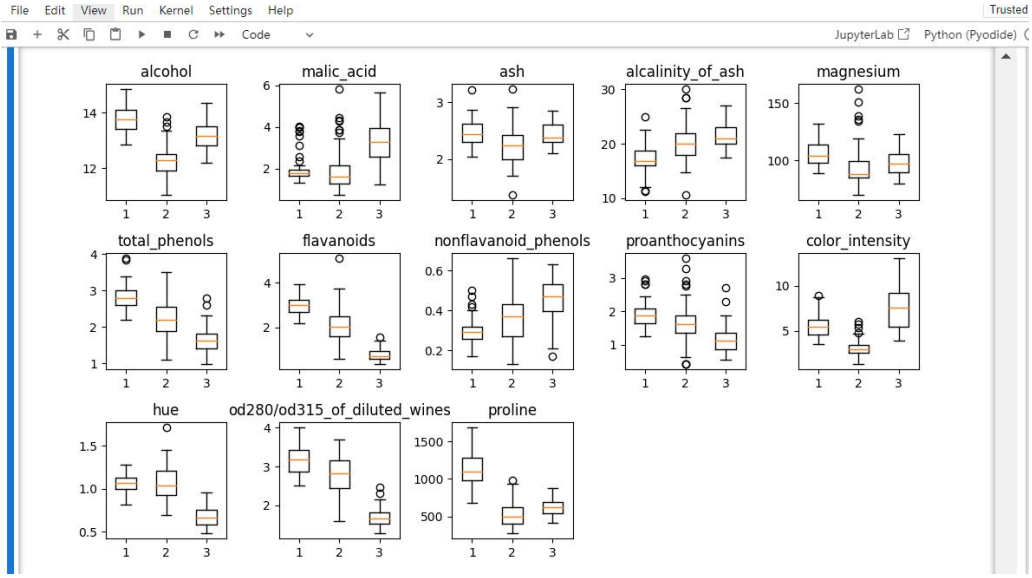

Рис.4– Боксплот по характеристикам вина в зависимости от класса

На этом этапе мы сравнили распределение каждого признака в разных категориях вин с помощью боксплотов. Например, для характеристики "Алкоголь" можно заметить, что медиана для категории 0 немного ниже, чем для категорий 1 и 2, что может означать, что образцы вина в категории 0 могут иметь относительно более низкое содержание алкоголя. Кроме того, мы также можем наблюдать относительно более широкое распределение для категорий 1 и 2, что может означать, что образцы в этих двух категориях имеют больший разброс содержания алкоголя.

Шаг 5: Кластерный анализ данных о вине

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans

# Standardize the data
```

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Reduce dimensionality using PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Perform KMeans clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X_scaled)
cluster_labels = kmeans.labels_

# Visualize clustering results
plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=cluster_labels, cmap='Set1', s=100)
plt.title('Clustering Analysis of Wine Data')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.show()
```

Рис.5–Кластерный анализ данных о вине

На этом этапе мы провели кластерный анализ данных о вине и визуализировали результаты кластеризации с помощью диаграммы рассеяния. Алгоритм кластеризации разделяет данные на 3 кластера и визуализирует их в двумерном пространстве. Каждая точка представляет собой образец, а ее цвет указывает на кластер, к которому она принадлежит. Рассматривая результаты кластеризации, мы можем определить различные кластеры в данных и далее исследовать различия и сходства между этими кластерами.

ВЫВОДЫ

- Набор данных о вине состоит из трех категорий с относительно сбалансированным распределением выборки.
- Некоторые признаки демонстрируют сильные корреляции, что может свидетельствовать о наличии избыточной информации.
- Распределение признаков вина варьируется, и понимание этих распределений необходимо для дальнейшего анализа.
- Распределение некоторых признаков значительно различается между категориями вин, что указывает на их важность для категоризации.
- Кластерный анализ показывает, что в данных есть четкие кластеры, что позволяет выявить потенциальные группы или закономерности.

Выполнив эти шаги и проанализировав результаты визуализации, мы сможем получить полное представление о наборе данных о вине и извлечь значимые выводы для дальнейшего анализа или принятия решений.