Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 2

Выполнил:
Беккиев Р.И.
группа ИУ5-64Б

Проверил:
Гапанюк Ю.Е.

Дата: 11.04.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

Задание:

Номер варианта: **2**

Номер задачи: 1

Номер набора данных, указанного в задаче: **2**

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Ход выполнения:

1) Загрузил набор данных, просмотрел начало, проверил пропуски и выяснил, что пропуски отсутствуют.

```
# Проверка на наличие пропущенных значений
print("\nКоличество пропущенных значений в каждой колонке:")
print(df_features.isnull().sum())

# Обработка пропусков (если бы они были)
# В данном наборе данных пропусков нет, но если бы были, можно было бы использовать:
# df_features_cleaned = df_features.dropna() # Удаление строк с пропусками
# df_features_cleaned = df_features.dropna(axis=1) # Удаление колонок с пропусками
# Поскольку пропусков нет, df_features_cleaned = df_features

# Расчет корреляционной матрицы
correlation_matrix = df_features.corr()

print("\nКорреляционная матрица:")
print(correlation_matrix)

# Визуализация корреляционной матрицы с помощью тепловой карты
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Тепловая карта корреляций признаков набора данных Wine')
plt.show()
```

2) Расчет и визуализация корреляционной матрицы

```
# Расчет корреляционной матрицы
correlation_matrix = df_features.corr()

print("\nКорреляционная матрица:")
print(correlation_matrix)

# Визуализация корреляционной матрицы с помощью тепловой карты
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Тепловая карта корреляций признаков набора данных Wine')
plt.show()

# Выводы (как были представлены ранее, можно добавить в текстовую ячейку Markdown в Jupyter)
# 1. Возможность построения моделей машинного обучения:
#    - Наличие мультиколлинеарности (например, total_phenols и flavanoids) может повлиять на линейные модели.
#    - Древовидные модели менее чувствительны.
#    - PCA может быть полезен.
#    - Данные подходят для построения моделей с учетом этих особенностей.

# 2. Возможный вклад признаков в модель:
#    - Группы сильно коррелирующих признаков (фенольные соединения) могут содержать избыточную информацию.
#    - Признаки, такие как proline, alcohol, color_intensity, alcalinity_of_ash,
#      показывают интересные паттерны корреляций и могут быть важны.
#    - Признаки с более слабыми корреляциями (magnesium, ash) могут вносить независимый вклад.
#    - Для точной оценки вклада необходим анализ важности признаков конкретной модели.

# Сохранение корреляционной матрицы в CSV файл (опционально)
# correlation_matrix.to_csv("wine_correlation_matrix.csv")
# print("\nКорреляционная матрица сохранена в 'wine_correlation_matrix.csv'")
```

✓ 2.3s


```
[1]  ✓ 2.3s                                                                    Python
...  Первые 5 строк DataFrame с признаками:
      alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  total_phenols  \
   0    14.23        1.71  2.43               15.6      127.0           2.80
   1    13.20        1.78  2.14               11.2      100.0           2.65
   2    13.16        2.36  2.67               18.6      101.0           2.80
   3    14.37        1.95  2.50               16.8      113.0           3.85
   4    13.24        2.59  2.87               21.0      118.0           2.80

      flavanoids  nonflavanoid_phenols  proanthocyanins  color_intensity   hue  \
   0        3.06                  0.28             2.29             5.64  1.04
   1        2.76                  0.26             1.28             4.38  1.05
   2        3.24                  0.30             2.81             5.68  1.03
   3        3.49                  0.24             2.18             7.80  0.86
   4        2.69                  0.39             1.82             4.32  1.04

      od280/od315_of_diluted_wines  proline
   0                          3.92   1065.0
   1                          3.40   1050.0
   2                          3.17   1185.0
   3                          3.45   1480.0
   4                          2.93    735.0

   Общая информация о DataFrame:
   <class 'pandas.core.frame.DataFrame'>
   RangeIndex: 178 entries, 0 to 177
   ...
   color_intensity                       -0.428815  0.316100
   hue                                    0.565468  0.236183
   od280/od315_of_diluted_wines           1.000000  0.312761
   proline                                0.312761  1.000000
```