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

РК №1

По курсу
«Технологии машинного обучения»

Подготовил:

Студент группы

ИУ5-63Б Борисов А.М.

08.04.2022

Проверил:

2022 г.

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.

Дополнительные требования по группам:

Для студентов групп ИУ5-63Б, ИУ5Ц-83Б - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Решение:

Рубежный контроль №1

Тема: Технологии разведочного анализа и обработки данных.

Вариант №5

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.

Дополнительные требования по группам:

Для студентов групп ИУ5-63Б, ИУ5Ц-83Б - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

1) Текстовое описание набора данных:

В качестве набора данных мы будем использовать набор данных: "Heart Disease Dataset." (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

Файл содержит следующие колонки:

1. Возраст
2. Пол
3. Тип боли в груди
4. Кровяное давление в состоянии покоя
5. Сывороточный холестерин
6. Уровень сахара в крови натощак
7. Результаты электрокардиографии в состоянии покоя
8. Достигнутая максимальная частота сердечных сокращений
9. Стенокардия, вызванная физической нагрузкой
10. Депрессия
11. Наклон пикового сегмента ST упражнения
12. Количество крупных сосудов (0-3), окрашенных флуороскопией
13. Имена и номера социального страхования пациентов

Импорт библиотек:

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

Загрузка данных:

Загрузим файлы датасета в помощью библиотеки Pandas.

```
[ ] data = pd.read_csv('/content/heart.csv', sep=",")
```

Первые 5 строк датасета:

```
[ ] data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Размер датасета:

```
[ ] data.shape
```

```
(1025, 14)
```

Список колонок с типами данных:

```
[ ] data.dtypes
```

```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

Подготовка данных к анализу:

Проверим наличие пустых значений. В выбранном датасете отсутствуют пустые значения.

```
for col in data.columns:

    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
age - 0
sex - 0
cp - 0
trestbps - 0
chol - 0
fbs - 0
restecg - 0
thalach - 0
exang - 0
oldpeak - 0
slope - 0
ca - 0
thal - 0
target - 0
```

Целевым признаком текущего датасета является столбец "target".

Определим уникальные значения для целевого признака:

```
[ ] data['target'].unique()
```

```
array([0, 1, 2, 3])
```

Целевой признак является бинарным и содержит только значения 0 и 1.

При помощи pandas построим корреляционную матрицу:

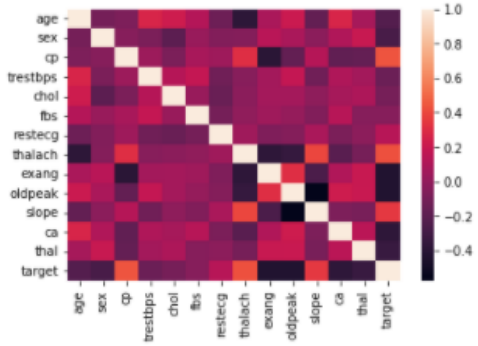
```
corr = data.corr()
print(corr)
```

```
                age       sex        cp  trestbps      chol       fbs  \
age        1.000000 -0.103240 -0.071966  0.271121  0.219823  0.121243
sex       -0.103240  1.000000 -0.041119 -0.078974 -0.198258  0.027200
cp        -0.071966 -0.041119  1.000000  0.038177 -0.081641  0.079294
trestbps   0.271121 -0.078974  0.038177  1.000000  0.127977  0.181767
chol       0.219823 -0.198258 -0.081641  0.127977  1.000000  0.026917
fbs        0.121243  0.027200  0.079294  0.181767  0.026917  1.000000
restecg   -0.132696 -0.055117  0.043581 -0.123794 -0.147410 -0.104051
thalach   -0.390227 -0.049365  0.306839 -0.039264 -0.021772 -0.008866
exang      0.088163  0.139157 -0.401513  0.061197  0.067382  0.049261
oldpeak    0.208137  0.084687 -0.174733  0.187434  0.064880  0.010859
slope     -0.169105 -0.026666  0.131633 -0.120445 -0.014248 -0.061902
ca         0.271551  0.111729 -0.176206  0.104554  0.074259  0.137156
thal       0.072297  0.198424 -0.163341  0.059276  0.100244 -0.042177
target    -0.229324 -0.279501  0.434854 -0.138772 -0.099966 -0.041164

             restecg   thalach     exang   oldpeak     slope        ca  \
age       -0.132696 -0.390227  0.088163  0.208137 -0.169105  0.271551
sex       -0.055117 -0.049365  0.139157  0.084687 -0.026666  0.111729
cp         0.043581  0.306839 -0.401513 -0.174733  0.131633 -0.176206
trestbps  -0.123794 -0.039264  0.061197  0.187434 -0.120445  0.104554
chol      -0.147410 -0.021772  0.067382  0.064880 -0.014248  0.074259
fbs       -0.104051 -0.008866  0.049261  0.010859 -0.061902  0.137156
restecg    1.000000  0.048411 -0.065606 -0.050114  0.086086 -0.078072
thalach    0.048411  1.000000 -0.380281 -0.349796  0.395308 -0.207888
exang     -0.065606 -0.380281  1.000000  0.310844 -0.267335  0.107849
oldpeak   -0.050114 -0.349796  0.310844  1.000000 -0.575189  0.221816
slope      0.086086  0.395308 -0.267335 -0.575189  1.000000 -0.073440
ca        -0.078072 -0.207888  0.107849  0.221816 -0.073440  1.000000
thal      -0.020504 -0.098068  0.197201  0.202672 -0.094090  0.149014
target     0.134468  0.422895 -0.438029 -0.438441  0.345512 -0.382085

              thal    target
age       0.072297 -0.229324
sex       0.198424 -0.279501
cp       -0.163341  0.434854
trestbps  0.059276 -0.138772
chol      0.100244 -0.099966
fbs      -0.042177 -0.041164
restecg  -0.020504  0.134468
thalach  -0.098068  0.422895
exang     0.197201 -0.438029
oldpeak   0.202672 -0.438441
slope    -0.094090  0.345512
ca        0.149014 -0.382085
thal      1.000000 -0.337838
target   -0.337838  1.000000
```

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap которая показывает степень корреляции различными цветами:

```
sns.heatmap(data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2cf8966ad0>
```



На основе корреляционной матрицы можно сделать следующие выводы:

- Целевой признак наиболее сильно коррелирует с типом боли в груди(0.43) и с максимальной частотой сердечных сокращений(0.42).Эти признаки стоит обязательно оставить в модели.
- Целевой признак совсем не коррелирует с возрастом (-0.22), полом(-0.27), кровяным давлением(-0.14) и с результатом электрокардиографии в состоянии покоя(-0.43). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат качество модели.

Дополнительное задание по группам.

Для колонки "age" построен график "Ящик с усами (boxplot)":

```
data.boxplot(column='age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc01f2e1d90>
```