Рубежный контроль №1 по предмету "Технологии машинного обучения"

Имя студента: Ергалиев Аслан, РТ5-61Б
Датасет №4: Heart Disease Dataset

Описание датасета

Данный набор данных датируется 1988 годом и состоит из четырех баз данных: Cleveland, Hungary, Switzerland и Long Beach V. Он содержит 76 атрибутов, включая предсказуемый атрибут, но все опубликованные эксперименты используют подмножество из 14 атрибутов. Поле "target" указывает на наличие сердечного заболевания у пациента. Оно принимает целочисленные значения:

- **0** = отсутствие заболевания
- **1** = наличие заболевания

Основные атрибуты:

1. **age** — возраст
2. **sex** — пол
3. **chest pain type** — тип боли в груди (4 значения)
4. **resting blood pressure** — артериальное давление в состоянии покоя
5. **serum cholestoral in mg/dl** — уровень холестерина в сыворотке крови (мг/дл)
6. **fasting blood sugar > 120 mg/dl** — уровень сахара в крови натощак > 120 мг/дл
7. **resting electrocardiographic results** — результаты ЭКГ в состоянии покоя (значения 0, 1, 2)
8. **maximum heart rate achieved** — максимальная достигнутая частота сердечных сокращений
9. **exercise induced angina** — стенокардия, вызванная физической нагрузкой
10. **oldpeak** — депрессия ST сегмента, вызванная физической нагрузкой относительно состояния покоя

11. **the slope of the peak exercise ST segment** — наклон пикового ST сегмента при физической нагрузке

12. **number of major vessels (0-3) colored by flourosopy** — количество крупных сосудов (0-3), окрашенных флюороскопией

13. **thal** — состояние талия:
    - 0 = нормальное
    - 1 = фиксированный дефект
    - 2 = обратимый дефект

Имена и номера социального страхования пациентов были недавно удалены из базы данных и заменены фиктивными значениями.

Вариант задания №1

1. Проведите корреляционный анализ для заданного набора данных.

2. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски.

3. Сделайте выводы о:
    - Возможности построения моделей машинного обучения.
    - Возможном вкладе признаков в модель.

Импорт библиотек

```
import pandas as pd
import numpy as np
```

Загрузка данных

```
data = pd.read_csv('heart.csv')
```

Обработка данных

```
data.head(20)
```

```
    age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang
oldpeak  \
0    52    1   0       125   212    0        1      168      0
1.0
1    53    1   0       140   203    1        0      155      1
```

```
3.1
2    70    1   0       145   174    0        1      125       1
2.6
3    61    1   0       148   203    0        1      161       0
0.0
4    62    0   0       138   294    1        1      106       0
1.9
5    58    0   0       100   248    0        0      122       0
1.0
6    58    1   0       114   318    0        2      140       0
4.4
7    55    1   0       160   289    0        0      145       1
0.8
8    46    1   0       120   249    0        0      144       0
0.8
9    54    1   0       122   286    0        0      116       1
3.2
10   71    0   0       112   149    0        1      125       0
1.6
11   43    0   0       132   341    1        0      136       1
3.0
12   34    0   1       118   210    0        1      192       0
0.7
13   51    1   0       140   298    0        1      122       1
4.2
14   52    1   0       128   204    1        1      156       1
1.0
15   34    0   1       118   210    0        1      192       0
0.7
16   51    0   2       140   308    0        0      142       0
1.5
17   54    1   0       124   266    0        0      109       1
2.2
18   50    0   1       120   244    0        1      162       0
1.1
19   58    1   2       140   211    1        0      165       0
0.0

    slope  ca  thal  target
0       2   2     3       0
1       0   0     3       0
2       0   0     3       0
3       2   1     3       0
4       1   3     2       0
5       1   0     2       1
6       0   3     1       0
7       1   1     3       0
8       2   0     3       0
9       1   2     2       0
```

```
10      1   0     2        1
11      1   0     3        0
12      2   0     2        1
13      1   3     3        0
14      1   0     0        0
15      2   0     2        1
16      2   1     2        1
17      1   1     3        0
18      2   0     2        1
19      2   0     2        1
```

```
data.isnull().sum()
```

```
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
dtype: int64
```

```
print(data.dtypes)
```

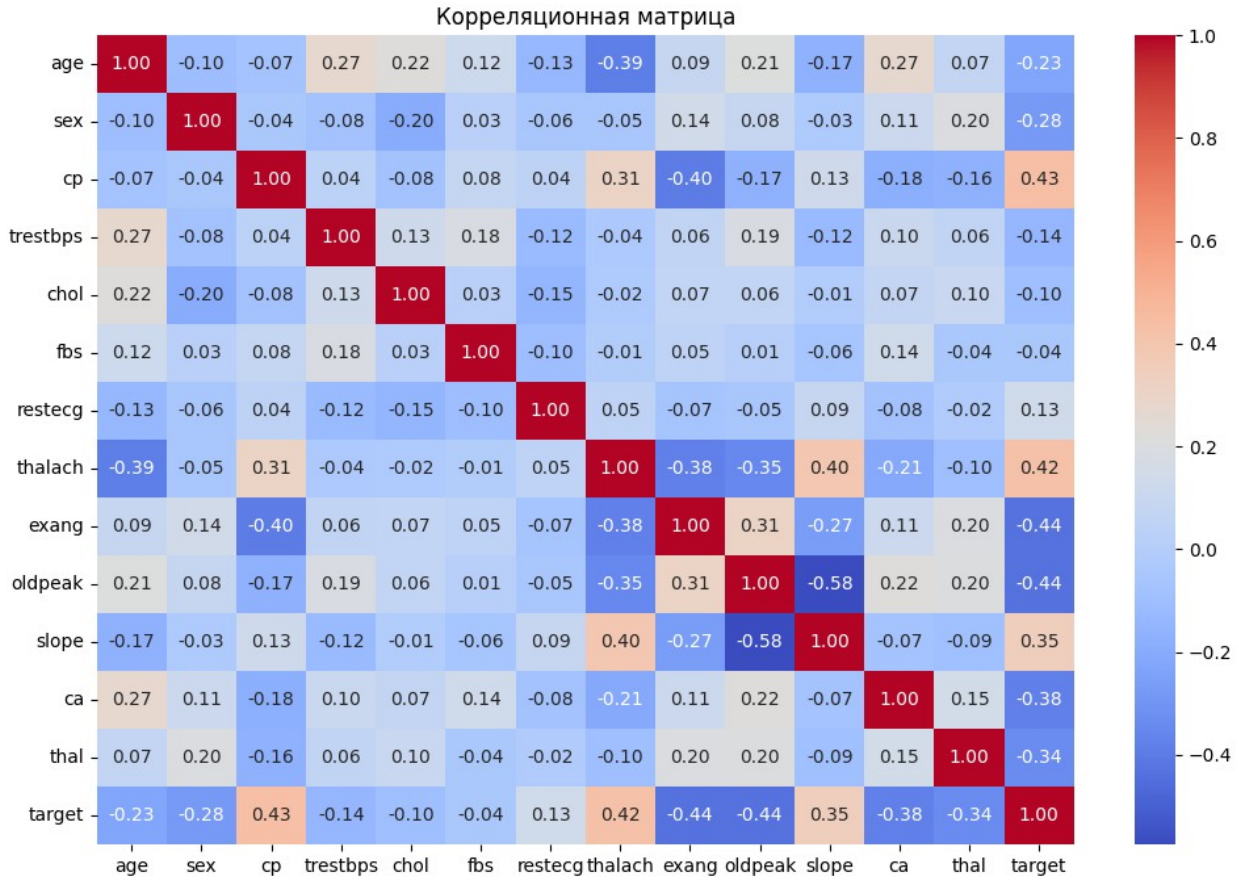
```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
# Построение корреляционной матрицы
correlation_matrix = data.corr()

# Визуализация корреляционной матрицы
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, fmt=".2f",
cmap="coolwarm")
plt.title("Корреляционная матрица")
plt.show()
```



Анализ корреляционной матрицы

На основе корреляционной матрицы можно выделить признаки, наиболее связанные с наличием сердечных заболеваний (столбец `target`). Ниже представлен подробный анализ.

Общий анализ

Признаки с наибольшей положительной корреляцией с `target`:

Признак	Коэффициент	Интерпретация
`cp` (тип боли в груди)	**0.43**	Некоторые типы боли в груди сильнее ассоциированы с болезнями сердца
`thalach` (макс. ЧСС)	**0.42**	Более высокая ЧСС при нагрузке — возможный индикатор патологии
`slope` (наклон ST сегмента)	**0.35**	Некоторые формы наклона чаще встречаются у больных

Признаки с наибольшей отрицательной корреляцией с `target`:

Признак	Коэффициент	Интерпретация
`exang` (нагрузочная стенокардия)	**-0.44**	Пациенты с нагрузочной стенокардией чаще не болеют (возможно, из-за кодировки)
`oldpeak` (депрессия ST сегмента)	**-0.44**	Выраженная депрессия ST сегмента — индикатор патологии
`ca` (кол-во поражённых сосудов)	**-0.38**	Чем больше сосудов поражено, тем выше риск болезни
`thal` (дефекты по таллию)	**-0.34**	Фиксированные/обратимые дефекты — индикатор патологии

Взаимосвязи между признаками

- `exang` и `thalach`: **-0.38**
  → При стенокардии наблюдается более низкая ЧСС.
- `oldpeak` и `slope`: **-0.58**
  → Чем выше депрессия ST, тем менее выражен наклон сегмента (отрицательная зависимость).

Выводы о построении моделей машинного обучения

1. Возможность построения моделей

Корреляционная матрица показывает наличие **заметных линейных зависимостей** между признаками и целевой переменной `target`, что указывает на **высокий потенциал применения методов машинного обучения**. Особенно это касается:

---

2. Возможный вклад признаков в модель

Признак	Возможный вклад в модель	Комментарий
`cp` (тип боли в груди)	Высокий	Сильно коррелирует с `target`, важный диагностический признак
`thalach` (макс. ЧСС)	Высокий	Связан с наличием патологии, особенно при нагрузке
`oldpeak` (депрессия ST)	Высокий	Показатель перегрузки сердца, ярко выражен у больных
`slope` (наклон ST)	Средне	Вносит полезную информацию в модель
`exang` (нагрузочная стенокардия)	Средний	Имеет значимую отрицательную корреляцию
`ca` (число поражённых сосудов)	Средний	Информативный, но может быть дискретным и шумным
`thal`	Средний	Определённые категории явно связаны с патологией
`age`, `sex`	Низкий	Слабая связь, но могут играть вспомогательную роль
`chol`, `fbs`, `restecg`, `trestbps`	Очень низкий	Могут быть исключены при необходимости упрощения модели

Дополнительные требования по группам

Для студентов группы РТ5-61Б - для пары произвольных колонок данных построить график "Jointplot".

```
sns.jointplot(data=data, x='age', y='thalach', kind='scatter',
hue='target', height=8)
plt.show()
```