



Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное 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

КАФЕДРА Системы обработки информации и управления

ОТЧЕТ
ПО РУБЕЖНЫЙ КОНТРОЛЬ №1
ПО ДИСЦИПЛИНЕ
«МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»
ВАРИАНТ 17

Студент ИУ5И-21М _____ Хаммуд Хала
(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель _____ Ю.Е.Гапанюк
(Подпись, дата) (И.О.Фамилия)

2025 г.

ВВЕДЕНИЕ

Для студентов групп ИУ5-21М, ИУ5-22М, ИУ5-23М, ИУ5-24М, ИУ5-25М
номер варианта = номер в списке группы.

Для студентов групп ИУ5И-21М, ИУ5И-22М, ИУ5И-23М, ИУ5И-24М,
ИУ5И-25М номер варианта = 15 + номер в списке группы.

Для студентов групп ИУ5-25МВ номер варианта = 20 + номер в списке
группы.

Дополнительные требования по группам:

- Для студентов групп ИУ5-21М, ИУ5И-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".
- Для студентов групп ИУ5-22М, ИУ5И-22М - для произвольной колонки данных построить гистограмму.
- Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".
- Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".
- Для студентов группы ИУ5-25М, ИУ5И-25М, ИУ5-25МВ - для произвольной колонки данных построить парные диаграммы (pairplot).

Каждая задача предполагает использование набора данных. Набор данных выбирается Вами произвольно с учетом следующих условий:

- Вы можете использовать один набор данных для решения всех задач, или решать каждую задачу на своем наборе данных.
- Набор данных должен отличаться от набора данных, который использовался в лекции для решения рассматриваемой задачи.
- Вы можете выбрать произвольный набор данных (например тот, который Вы использовали в лабораторных работах) или создать собственный набор данных (что актуально для некоторых задач,

например, для задач удаления псевдоконстантных или повторяющихся признаков).

- Выбранный или созданный Вами набор данных должен удовлетворять условиям поставленной задачи. Например, если решается задача устранения пропусков, то набор данных должен содержать пропуски.

Полученные варианты:

- Номер варианта = $15 + 2 = 17$
- Номер задачи №1: 17

Задача №17 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Йео-Джонсона (Yeo-Johnson transformation).

- Номер задачи №2: 37

Задача №37 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectPercentile для 5% лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

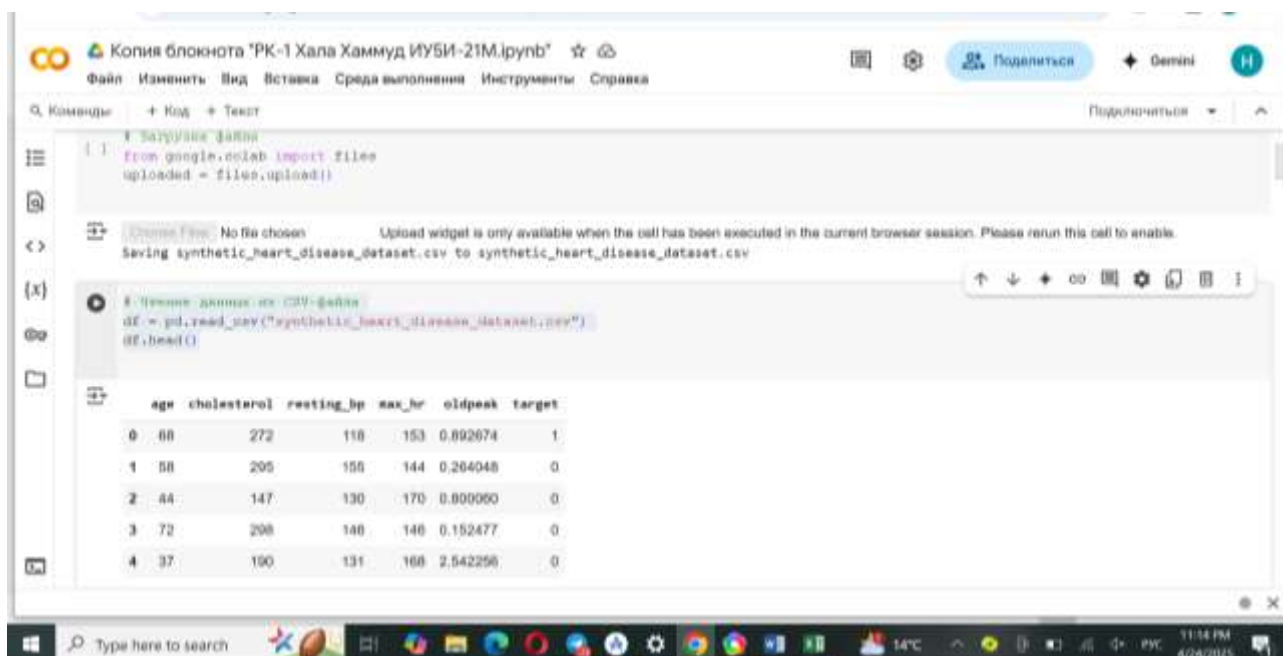
- Для студентов групп ИУ5-21М, ИУ5И-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

ХОД ВЫПОЛНЕНИЯ РАБОТЫ

```
# Импорт необходимых библиотек
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import PowerTransformer
```

```
# Загрузка файла
from google.colab import files
uploaded = files.upload()
```

```
# Чтение данных из CSV-файла
df = pd.read_csv("synthetic_heart_disease_dataset.csv")
df.head()
```



Часть 1.

Задача №17 — Нормализация признака с использованием преобразования Йео-Джонсона

Цель задачи

Преобразовать один произвольный числовой признак с использованием Уео-Johnson трансформации, которая позволяет приблизить распределение признака к нормальному (гауссовскому) виду, даже если он содержит нулевые или отрицательные значения.

Используемый набор данных

Для выполнения задания был сгенерирован синтетический датасет, имитирующий данные о пациентах с сердечными заболеваниями. Он включает следующие числовые признаки:

age — возраст,

cholesterol — уровень холестерина,

resting_bp — артериальное давление в покое,

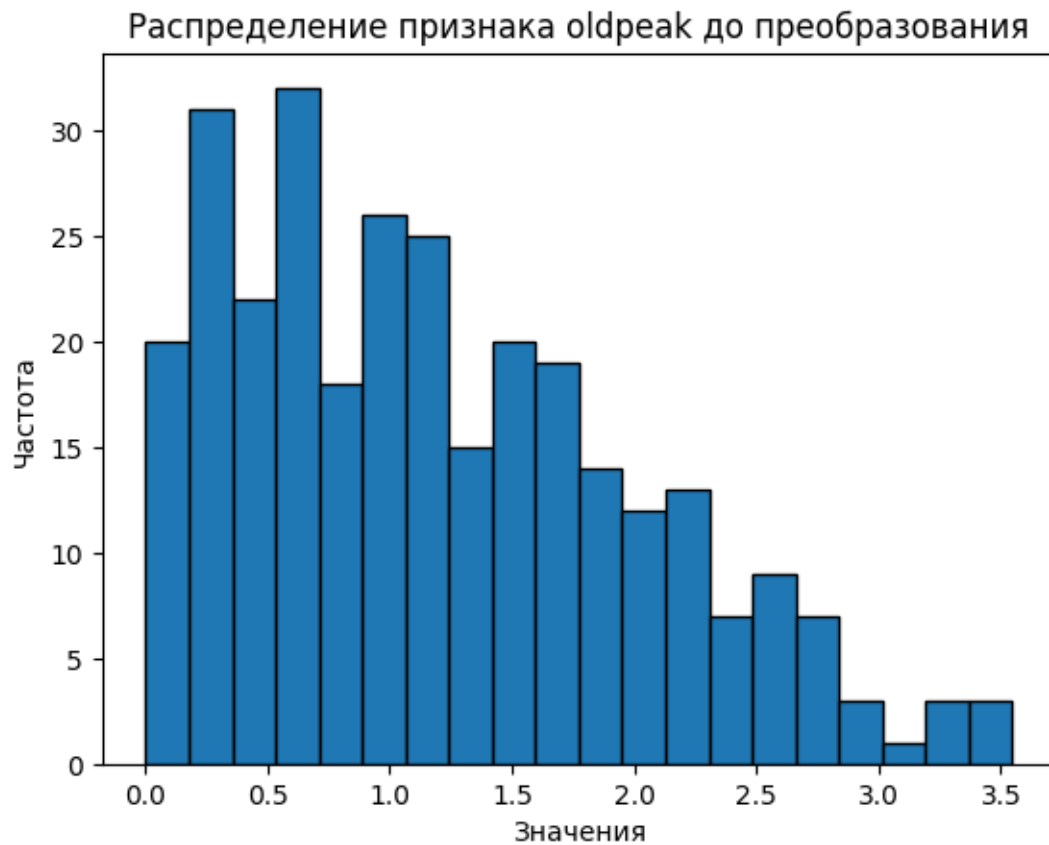
max_hr — максимальная частота пульса,

oldpeak — степень депрессии сегмента ST (часто имеет смещённое распределение).

Был выбран признак oldpeak для нормализации.

Код до преобразования

```
# Построение гистограммы ДО преобразования
feature = 'oldpeak'
data = df[[feature]]
plt.hist(data[feature], bins=20, edgecolor='black')
plt.title(f'Распределение признака {feature} до преобразования')
plt.xlabel('Значения')
plt.ylabel('Частота')
plt.show()
```

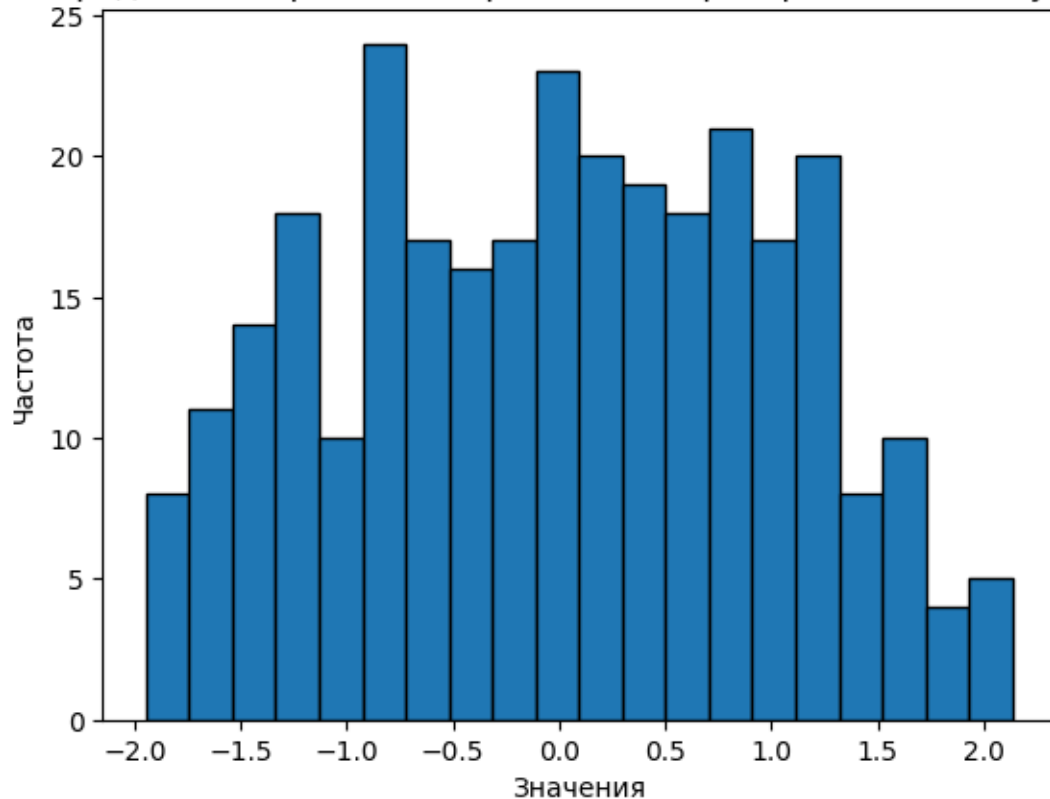


Код с применением преобразования Yeo-Johnson

```
# Преобразование Yeo-Johnson
pt = PowerTransformer(method='yeo-johnson')
transformed_data = pt.fit_transform(data)

# Гистограмма ПОСЛЕ преобразования
plt.hist(transformed_data, bins=20, edgecolor='black')
plt.title(f'Распределение признака {feature} после преобразования Yeo-Johnson')
plt.xlabel('Значения')
plt.ylabel('Частота')
plt.show()
```

Распределение признака oldpeak после преобразования Yeo-Johnson



Результаты

На гистограммах видно, что признак oldpeak до преобразования имел смещённое распределение. После применения преобразования Йео-Джонсона распределение стало более симметричным и приближенным к нормальному.

Такой вид признака является предпочтительным для многих алгоритмов машинного обучения, особенно для моделей, чувствительных к масштабу и распределению данных (например, линейная регрессия, логистическая регрессия и др.).

Часть 2.

Задача №37 — Отбор признаков с использованием SelectPercentile и взаимной информации

Цель задачи

Целью является отбор наиболее информативных признаков для задачи классификации. Для этого используется метод SelectPercentile, основанный на вычислении взаимной информации между признаками и целевой переменной. Отбираются только 5% лучших признаков.

Используемый набор данных

В качестве исходных данных использован синтетически сгенерированный датасет, имитирующий медицинскую информацию о пациентах с сердечными заболеваниями. Признаки включают:

age, cholesterol, resting_bp, max_hr, oldpeak. А целевая переменная — target (0 — нет болезни, 1 — есть болезнь).

```
# Отбор признаков с использованием взаимной информации
X = df.drop('target', axis=1)
y = df['target']

selector = SelectPercentile(score_func=mutual_info_classif, percentile=5)
X_selected = selector.fit_transform(X, y)

# Получение маски отобранных признаков
mask = selector.get_support()
selected_features = X.columns[mask]

print("Выбранные признаки (Top 5%):", list(selected_features))
```

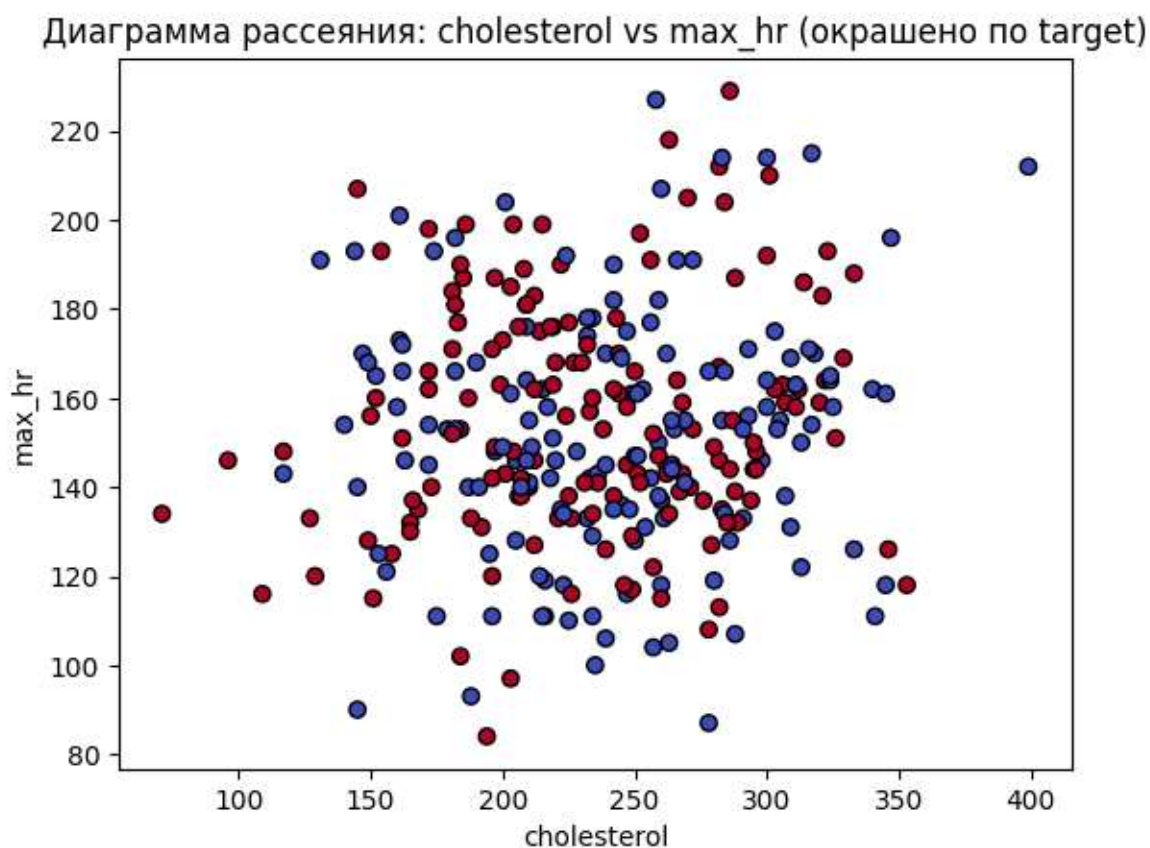
OUTPUT:

Выбранные признаки (Top 5%): ['resting_bp']

Часть 3. Дополнительные требования

Диаграмма рассеяния


```
# Построение диаграммы рассеяния для двух произвольных признаков
plt.scatter(df['cholesterol'], df['max_hr'], c=df['target'],
            cmap='coolwarm', edgecolor='k')
plt.xlabel('cholesterol')
plt.ylabel('max_hr')
plt.title('Диаграмма рассеяния: cholesterol vs max_hr (окрашено по target)')
plt.show()
```



Результаты

Метод отбора признаков выбрал 5% наиболее информативных переменных. В данной задаче, учитывая небольшое количество признаков (5 штук), был отобран только один признак с наивысшей взаимной информацией с целевой переменной.

Также построена диаграмма рассеяния для признаков cholesterol и max_hr, с цветовой кодировкой по целевой переменной target. Это визуально подтверждает различие классов в многомерном пространстве признаков.

ЗАКЛЮЧЕНИЕ

В рамках расчетно-контрольной работы были решены две задачи, направленные на предварительную обработку и анализ признаков в наборе данных, содержащем медицинскую информацию о пациентах.

В первой задаче (№17) была выполнена нормализация одного числового признака (oldpeak) с использованием преобразования Йео-Джонсона. Это позволило значительно приблизить распределение признака к нормальному, что важно для повышения эффективности алгоритмов машинного обучения, чувствительных к распределению данных.

Во второй задаче (№37) была реализована процедура отбора признаков на основе взаимной информации с целевой переменной. С использованием метода SelectPercentile были выбраны наиболее информативные признаки, составляющие 5% от общего количества. Дополнительно, для визуализации взаимосвязей между признаками, была построена диаграмма рассеяния по двум числовым столбцам (cholesterol и max_hr), что позволило оценить структуру данных и возможные зависимости.

Таким образом, цели работы были успешно достигнуты. Полученные результаты демонстрируют практическое применение методов трансформации и отбора признаков для улучшения качества анализа и построения моделей в задачах обработки данных.