|  |  |
| --- | --- |
|  | **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 ДИСЦ******ИПЛИНЕ***

***«МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»***

***ВАРИАНТ 16***

Студент \_\_ИУ5И-23М\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ Лю Цзычжан\_**

(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 + 1 =16
* Номер задачи №1: 16

Задача №16 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

* Номер задачи №2: 36

Задача №36 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

* Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

**ХОД ВЫПОЛНЕНИЯ РАБОТЫ**

# Часть 1. Текстовое описание набора данных

Набор данных № 1: heart.csv

Этот набор данных датируется 1988 годом и состоит из четырех баз данных: Кливленд, Венгрия, Швейцария и Лонг-Бич V. Он содержит 76 атрибутов, включая прогнозируемый атрибут, но все опубликованные эксперименты ссылаются на использование подмножества из 14 из них. Поле «target» относится к наличию у пациента заболевания сердца. Оно имеет целочисленное значение 0 = нет заболевания и 1 = заболевание.

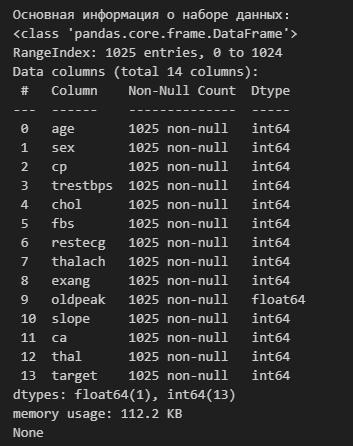


Рисунок 1: Информация о наборе данных (heart.csv)

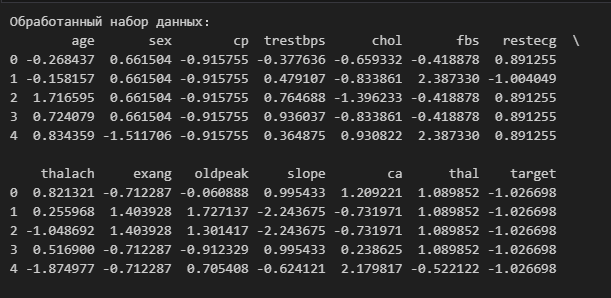


Рисунок 2: Первые 5 строк набора данных (heart.csv)

Набор данных № 2: heart.csv

**The dataset includes the following columns:**

* age
* sex
* chest pain type (4 values)
* resting blood pressure
* serum cholestoral in mg/dl
* fasting blood sugar > 120 mg/dl
* resting electrocardiographic results (values 0,1,2)
* maximum heart rate achieved
* exercise induced angina
* oldpeak = ST depression induced by exercise relative to rest
* the slope of the peak exercise ST segment
* number of major vessels (0-3) colored by flourosopy
* thal: 0 = normal; 1 = fixed defect; 2 = reversable defect  
  The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

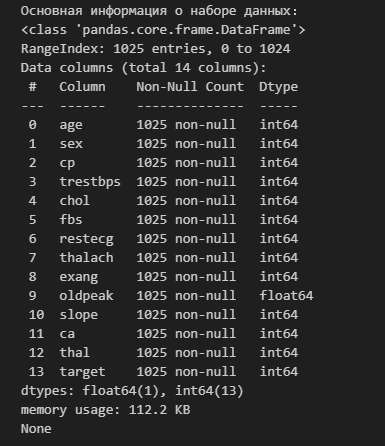


Рисунок 3: Информация о наборе данных (heart.csv)

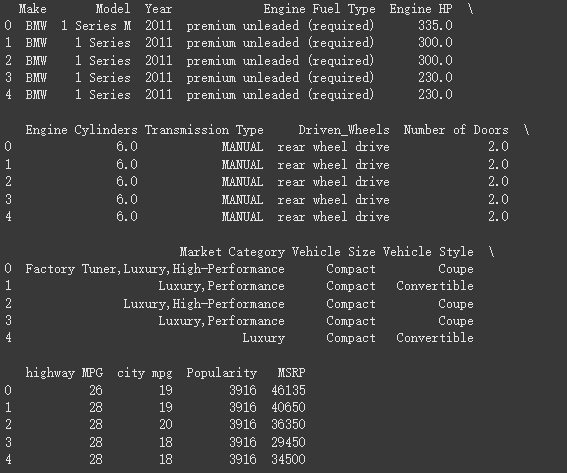


Рисунок 4: Первые 5 строк набора данных (heart.csv)

# Часть 2. Задача №16

Задача №16 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

Используя набор данных № 1: heart.csv



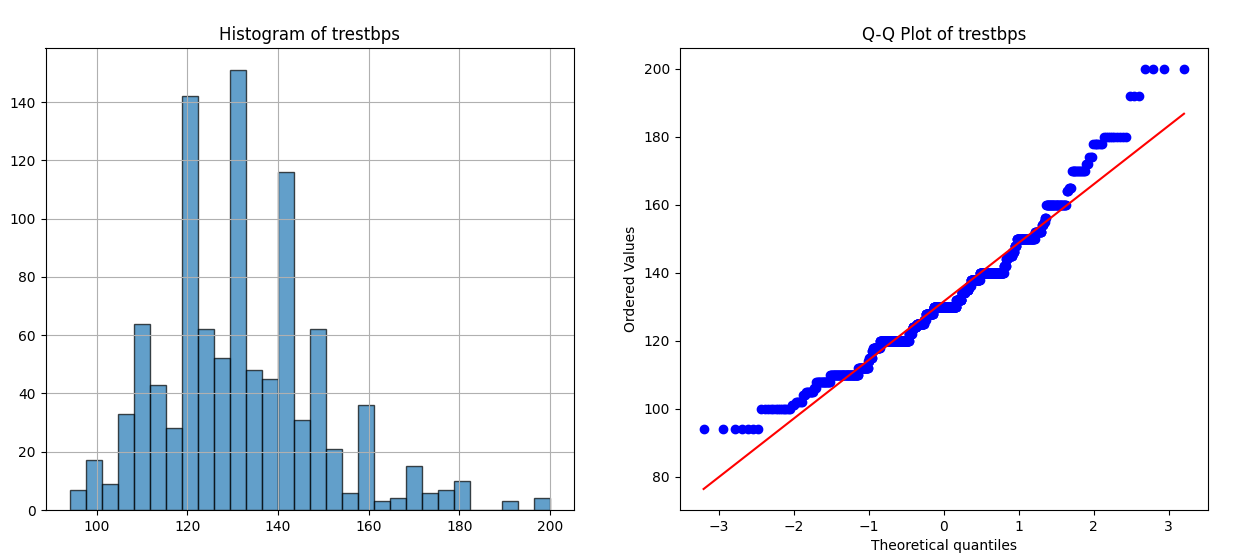


Рисунок 5: Гистограмма и график Q-Q перед преобразованием данных

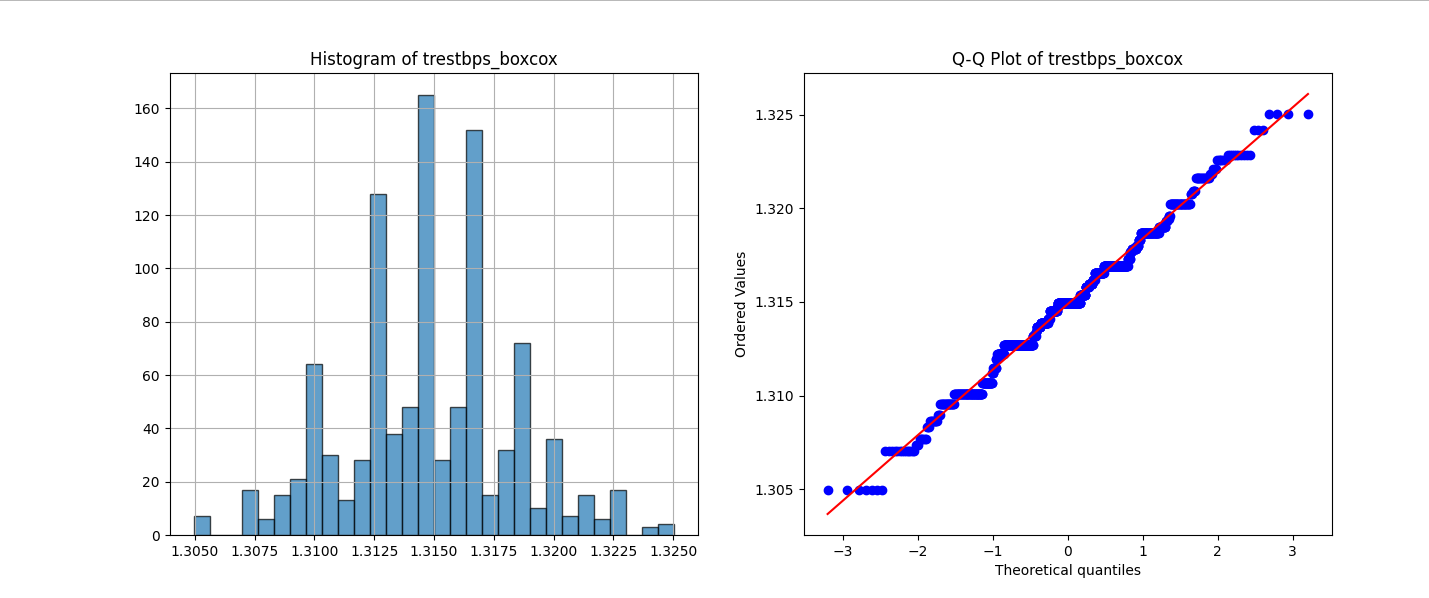
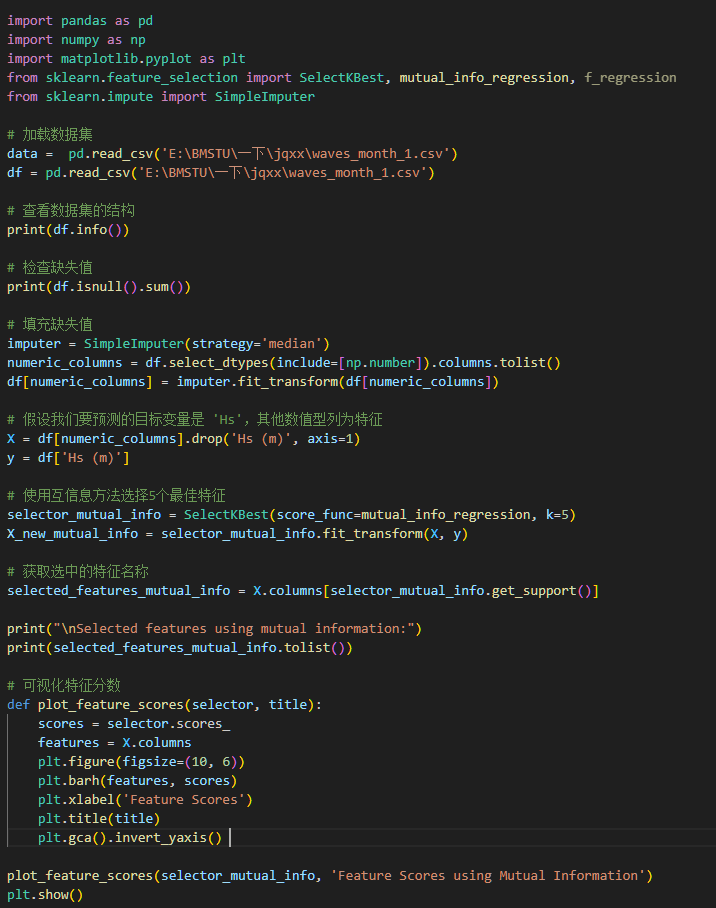


Рисунок 6: Гистограмма и график Q-Q после преобразования данных

# Часть 3. Задача №36

Задача №36 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Используя набор данных № 2: waves\_month\_1.csv



OUTPUT:

Selected features using mutual information:

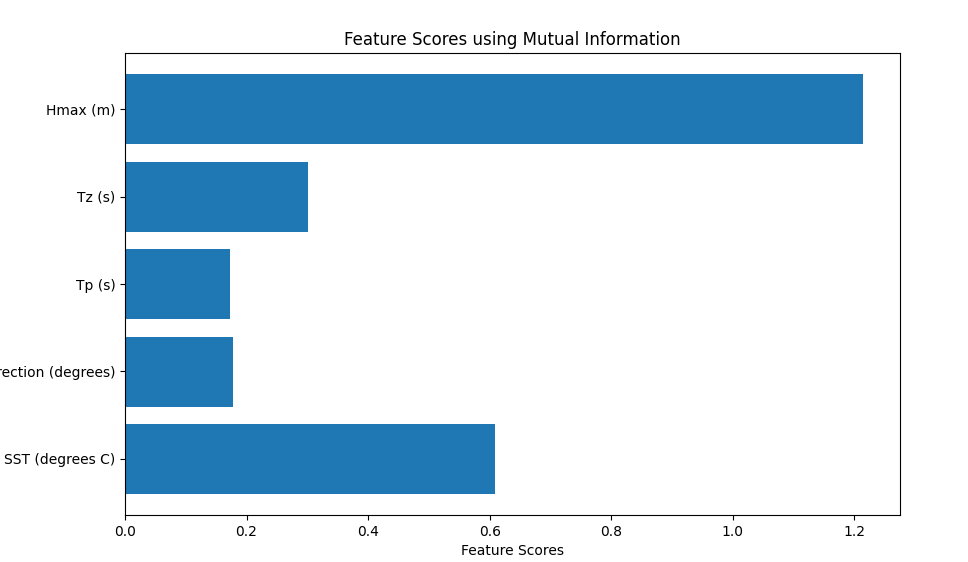
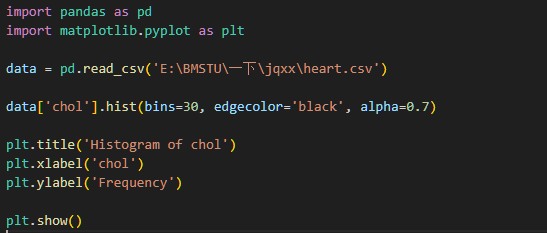
['Hmax (m)', 'Tz (s)', 'Tp (s)', 'Peak Direction (degrees)', 'SST (degrees C)'] 

Рисунок 7: Результаты с использованием методов А и В

# Часть 4. Дополнительные требования

Для произвольной колонки данных построить гистограмму.

Используя набор данных № 1: heart.csv



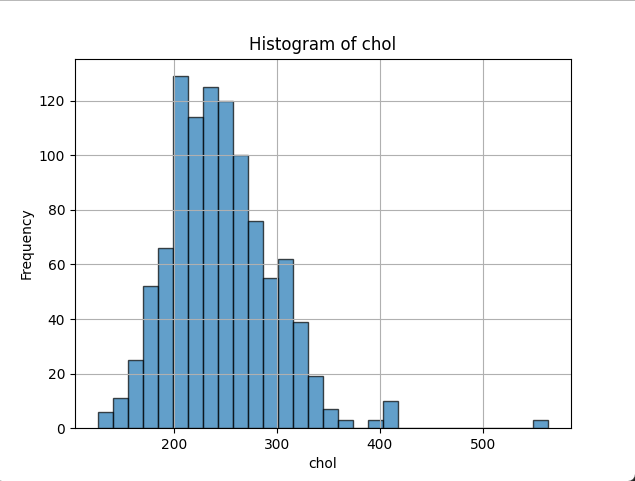
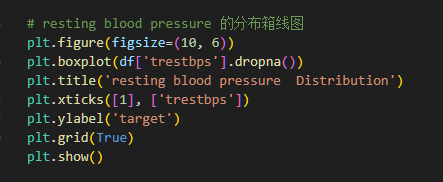


Рисунок 8: Гистограмма столбца Price

Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".



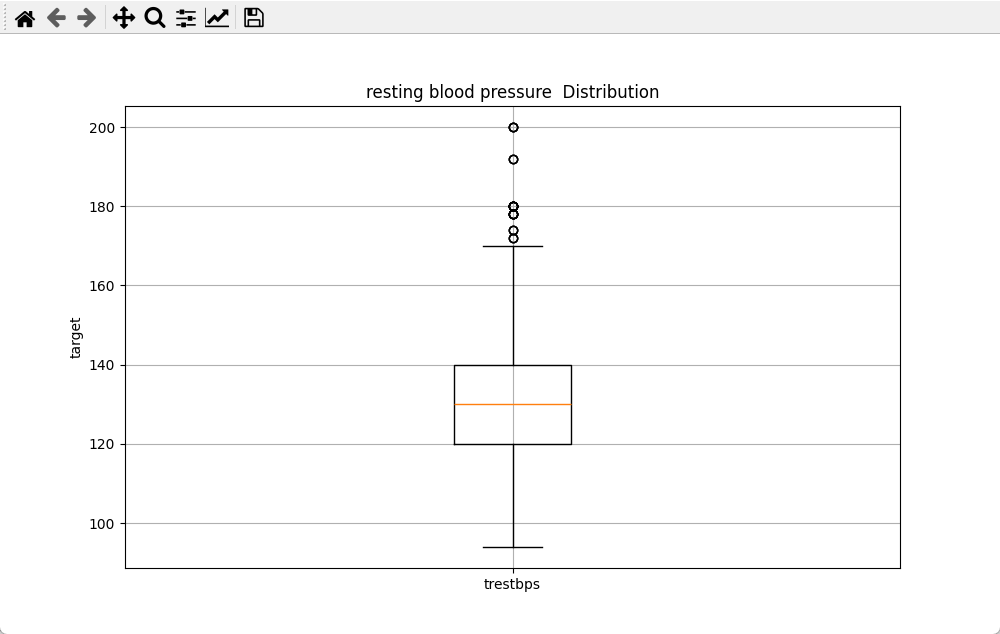


Рисунок 9:Прямоугольный график распределения артериального давления

**ЗАКЛЮЧЕНИЕ**

В ходе выполнения итогового контроля №1 по дисциплине «Методы машинного обучения» была проведена сложная работа по обработке и анализу набора данных по заболеваниям сердца heart.csv и набора данных по волнам waves\_month\_1.csv. В задании № 16 нормализация данных с использованием преобразования Бокса-Кокса была успешно применена к числовому признаку «trestbps» в наборе данных heart.csv. Это позволило нам преобразовать распределение артериального давления в состоянии покоя в нормальное распределение, что является важным шагом в подготовке данных для многих алгоритмов машинного обучения, чувствительных к масштабу и распределению признаков.

Для задачи № 36 была выполнена процедура выбора признаков на наборе данных waves\_month\_1.csv с использованием класса SelectKBest и метода, основанного на взаимной информации. В результате были выявлены пять наиболее важных признаков для прогнозирования волн, что демонстрирует эффективность данных методов в задаче прогнозирования и позволяет упростить модель за счет исключения менее важных признаков, что может улучшить ее производительность и интерпретируемость.

Еще одним запросом от группы было построение гистограммы для любого столбца данных, что было сделано на примере столбца «chol» в наборе данных heart.csv. Гистограмма дает четкую картину распределения цен и помогает визуально оценить форму распределения и возможные выбросы или аномалии в данных.