



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 16

Студент ИУ5И-23М
(Группа)

(Подпись, дата) Лю Цзычжан
(И.О.Фамилия)

Преподаватель

(Подпись, дата) Ю.Е.Гапанюк
(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.

```
Основная информация о наборе данных:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1025 entries, 0 to 1024  
Data columns (total 14 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   age         1025 non-null   int64  
1   sex         1025 non-null   int64  
2   cp          1025 non-null   int64  
3   trestbps    1025 non-null   int64  
4   chol        1025 non-null   int64  
5   fbs         1025 non-null   int64  
6   restecg     1025 non-null   int64  
7   thalach     1025 non-null   int64  
8   exang       1025 non-null   int64  
9   oldpeak     1025 non-null   float64  
10  slope       1025 non-null   int64  
11  ca          1025 non-null   int64  
12  thal        1025 non-null   int64  
13  target      1025 non-null   int64  
dtypes: float64(1), int64(13)  
memory usage: 112.2 KB  
None
```

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

Обработанный набор данных:							
	age	sex	cp	trestbps	chol	fbs	restecg \
0	-0.268437	0.661504	-0.915755	-0.377636	-0.659332	-0.418878	0.891255
1	-0.158157	0.661504	-0.915755	0.479107	-0.833861	2.387330	-1.004049
2	1.716595	0.661504	-0.915755	0.764688	-1.396233	-0.418878	0.891255
3	0.724079	0.661504	-0.915755	0.936037	-0.833861	-0.418878	0.891255
4	0.834359	-1.511706	-0.915755	0.364875	0.930822	2.387330	0.891255
	thalach	exang	oldpeak	slope	ca	thal	target
0	0.821321	-0.712287	-0.060888	0.995433	1.209221	1.089852	-1.026698
1	0.255968	1.403928	1.727137	-2.243675	-0.731971	1.089852	-1.026698
2	-1.048692	1.403928	1.301417	-2.243675	-0.731971	1.089852	-1.026698
3	0.516900	-0.712287	-0.912329	0.995433	0.238625	1.089852	-1.026698
4	-1.874977	-0.712287	0.705408	-0.624121	2.179817	-0.522122	-1.026698

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.

```

Основная информация о наборе данных:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         1025 non-null   int64
1   sex         1025 non-null   int64
2   cp          1025 non-null   int64
3   trestbps    1025 non-null   int64
4   chol        1025 non-null   int64
5   fbs         1025 non-null   int64
6   restecg     1025 non-null   int64
7   thalach     1025 non-null   int64
8   exang       1025 non-null   int64
9   oldpeak     1025 non-null   float64
10  slope       1025 non-null   int64
11  ca          1025 non-null   int64
12  thal        1025 non-null   int64
13  target      1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
None

```

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

	Make	Model	Year	Engine	Fuel Type	Engine HP	\
0	BMW	1 Series M	2011	premium	unleaded (required)	335.0	
1	BMW	1 Series	2011	premium	unleaded (required)	300.0	
2	BMW	1 Series	2011	premium	unleaded (required)	300.0	
3	BMW	1 Series	2011	premium	unleaded (required)	230.0	
4	BMW	1 Series	2011	premium	unleaded (required)	230.0	
	Engine	Cylinders	Transmission	Type	Driven_Wheels	Number of Doors	\
0		6.0	MANUAL	rear wheel drive		2.0	
1		6.0	MANUAL	rear wheel drive		2.0	
2		6.0	MANUAL	rear wheel drive		2.0	
3		6.0	MANUAL	rear wheel drive		2.0	
4		6.0	MANUAL	rear wheel drive		2.0	
	Market Category	Vehicle Size	Vehicle Style	\			
0	Factory Tuner, Luxury, High-Performance	Compact	Coupe				
1	Luxury, Performance	Compact	Convertible				
2	Luxury, High-Performance	Compact	Coupe				
3	Luxury, Performance	Compact	Coupe				
4	Luxury	Compact	Convertible				
	highway MPG	city mpg	Popularity	MSRP			
0	26	19	3916	46135			
1	28	19	3916	40650			
2	28	20	3916	36350			
3	28	18	3916	29450			
4	28	18	3916	34500			

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

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# 加载数据集
data = pd.read_csv('E:\BMSTU\一下\jqxx\heart.csv')

# 定义一个函数来绘制诊断图（直方图和Q-Q图）
def diagnostic_plots(df, variable):
    plt.figure(figsize=(15, 6))

    # 直方图
    plt.subplot(1, 2, 1)
    df[variable].hist(bins=30, edgecolor='black', alpha=0.7)
    plt.title(f'Histogram of {variable}')

    # Q-Q图
    plt.subplot(1, 2, 2)
    stats.probplot(df[variable], dist="norm", plot=plt)
    plt.title(f'Q-Q Plot of {variable}')

    plt.show()

# 对原始的trestbps列进行诊断
diagnostic_plots(data, 'trestbps')

# 应用Box-Cox变换
data['trestbps_boxcox'], param = stats.boxcox(data['trestbps'])

print(f'Optimal  $\lambda$  value for Box-Cox transformation: {param}')

# 对变换后的price_boxcox列进行诊断
diagnostic_plots(data, 'trestbps_boxcox')
```

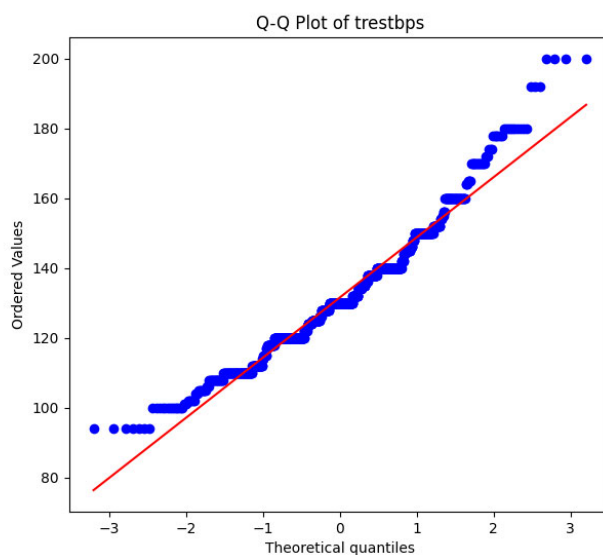
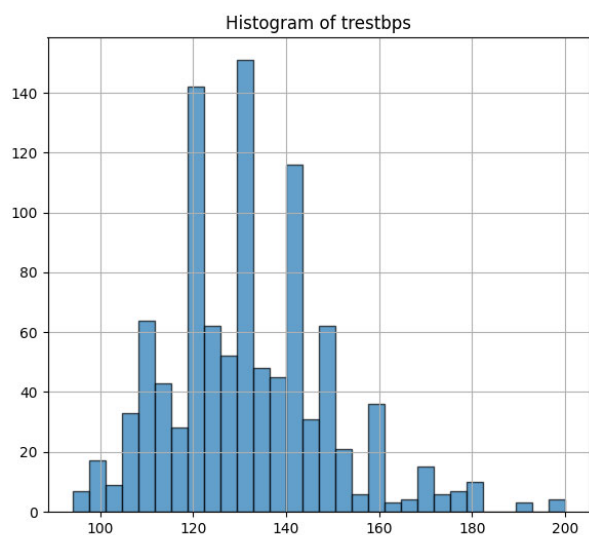



Рисунок 5: Гистограмма и график Q-Q перед преобразованием данных

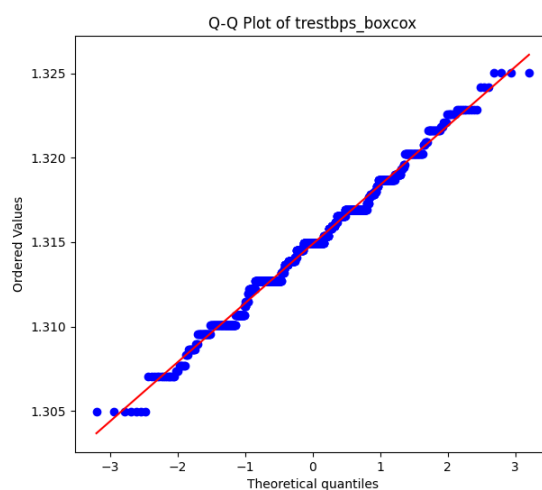
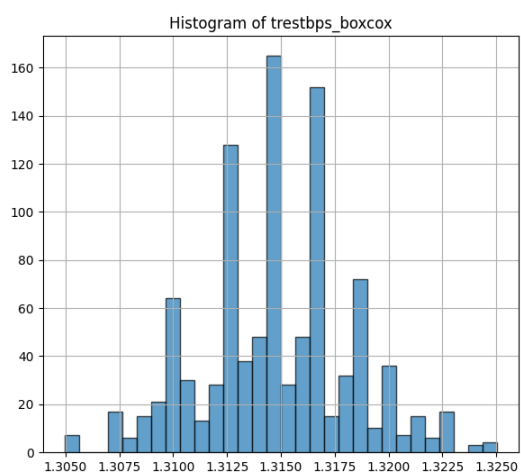


Рисунок 6: Гистограмма и график Q-Q после преобразования данных

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

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_selection import SelectKBest, mutual_info_regression, f_regression
from sklearn.impute import SimpleImputer

# 加载数据集
data = pd.read_csv('E:\BMSTU\一下\jqxx\waves_month_1.csv')
df = pd.read_csv('E:\BMSTU\一下\jqxx\waves_month_1.csv')

# 查看数据集的结构
print(df.info())

# 检查缺失值
print(df.isnull().sum())

# 填充缺失值
imputer = SimpleImputer(strategy='median')
numeric_columns = df.select_dtypes(include=[np.number]).columns.tolist()
df[numeric_columns] = imputer.fit_transform(df[numeric_columns])

# 假设我们要预测的目标变量是 'Hs', 其他数值型列为特征
X = df[numeric_columns].drop('Hs (m)', axis=1)
y = df['Hs (m)']

# 使用互信息方法选择5个最佳特征
selector_mutual_info = SelectKBest(score_func=mutual_info_regression, k=5)
X_new_mutual_info = selector_mutual_info.fit_transform(X, y)

# 获取选中的特征名称
selected_features_mutual_info = X.columns[selector_mutual_info.get_support()]

print("\nSelected features using mutual information:")
print(selected_features_mutual_info.tolist())

# 可视化特征分数
def plot_feature_scores(selector, title):
    scores = selector.scores_
    features = X.columns
    plt.figure(figsize=(10, 6))
    plt.barh(features, scores)
    plt.xlabel('Feature Scores')
    plt.title(title)
    plt.gca().invert_yaxis()

plot_feature_scores(selector_mutual_info, 'Feature Scores using Mutual Information')
plt.show()

```

OUTPUT:

Selected features using mutual information:

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

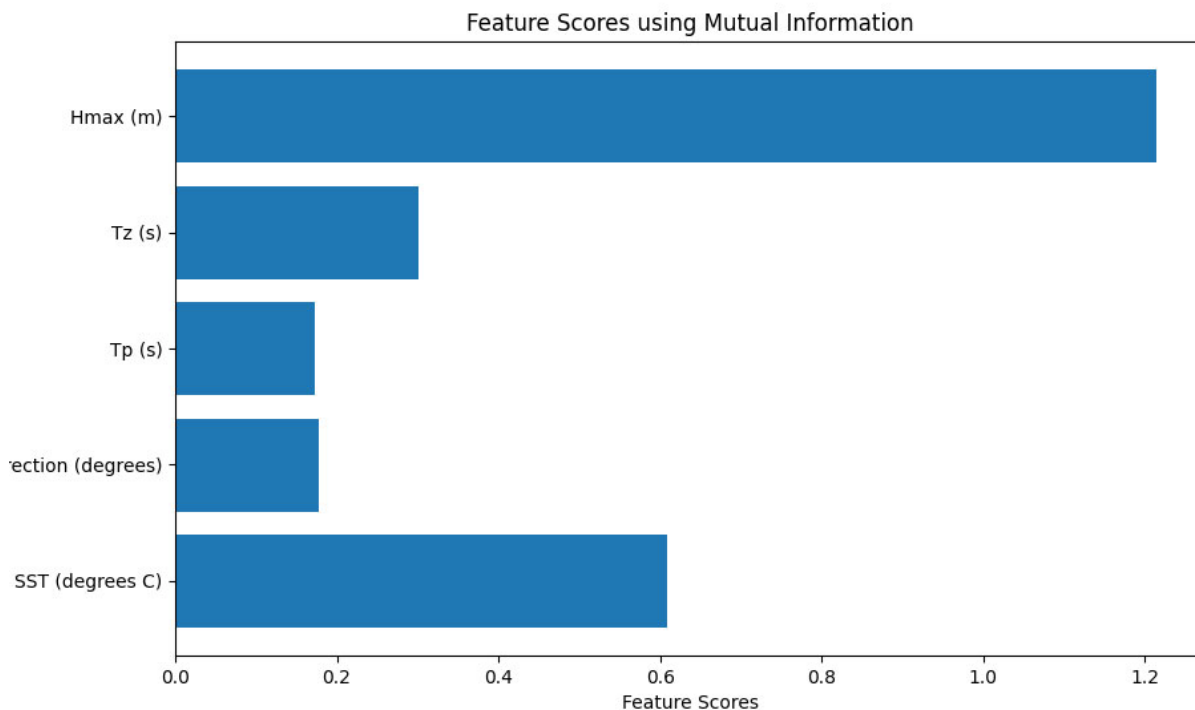


Рисунок 7: Результаты с использованием методов А и В

Часть 4. Дополнительные требования

Для произвольной колонки данных построить гистограмму.

Используя набор данных № 1: heart.csv

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv('E:\BMSTU\—\jqxx\heart.csv')

data['chol'].hist(bins=30, edgecolor='black', alpha=0.7)

plt.title('Histogram of chol')
plt.xlabel('chol')
plt.ylabel('Frequency')

plt.show()
```

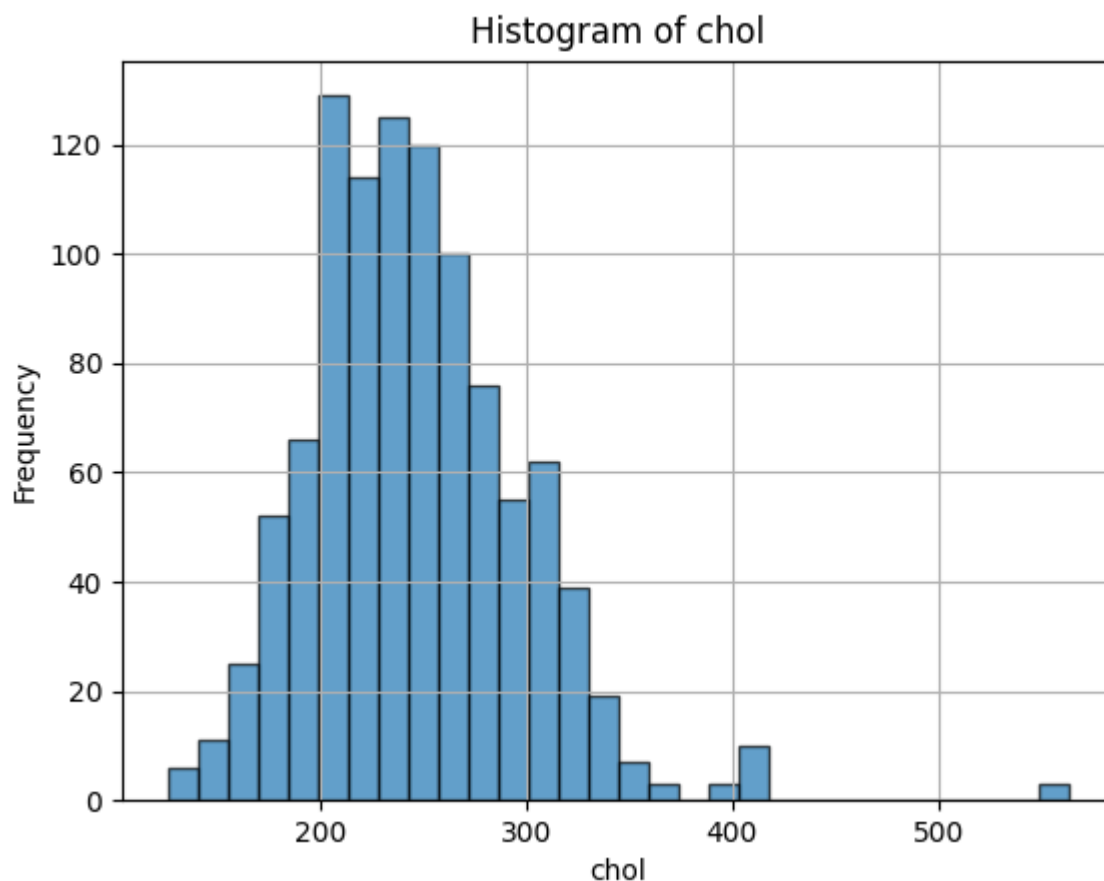


Рисунок 8: Гистограмма столбца Price

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

```
# resting blood pressure 的分布箱线图
plt.figure(figsize=(10, 6))
plt.boxplot(df['trestbps'].dropna())
plt.title('resting blood pressure Distribution')
plt.xticks([1], ['trestbps'])
plt.ylabel('target')
plt.grid(True)
plt.show()
```

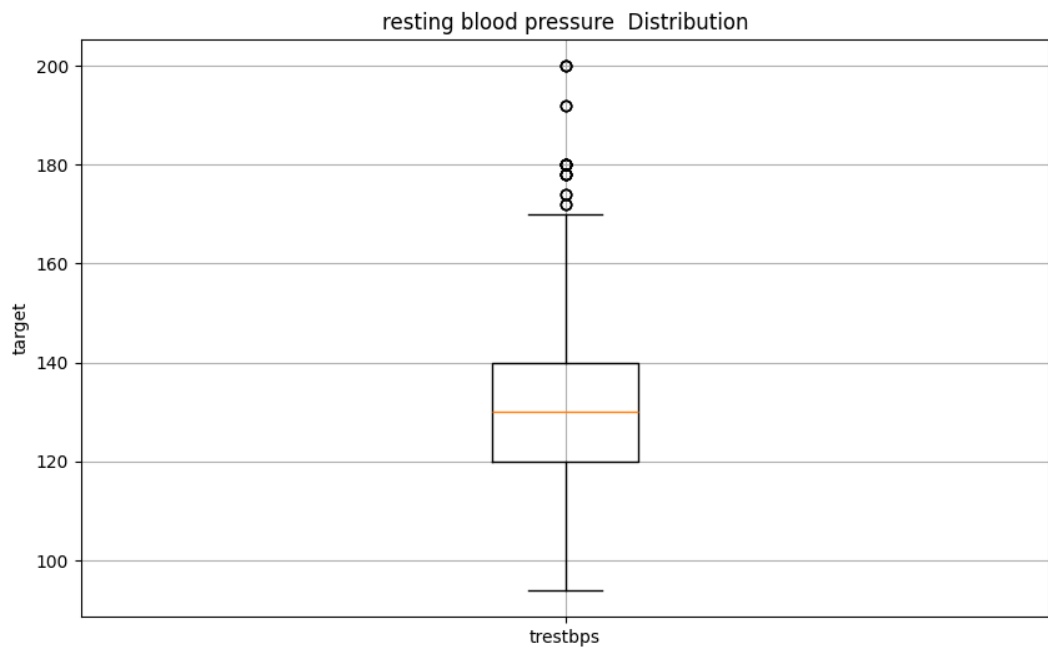


Рисунок 9: Прямоугольный график распределения артериального давления

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.