МИНОБРНАУКИ РОССИИ
Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
Кафедра Вычислительной техники

отчет

по лабораторной работе № 1

по дисциплине «Основы машинного обучения»

Тема: **«Предобработка данных для машинного обучения»**

Студенты гр. 3311 Аршин А. Д
Баймухамедов Р. Р.
Пасечный Л. В.

Преподаватель Петруша П. Г.

Санкт-Петербург

2025

Цель работы

Получение и закрепление навыков предобработки данных для дальнейшего применения методов машинного обучения для решения задач.

Задание работы

Самостоятельно выбрать набор сырых (не разобранных) данных на сайте kaggle.com с следующими критериями:

- Число столбцов признаков – не менее 10;
- Число записей – не менее 10000;
- Набор данных имеет пропуски.

Очистить данные (удалить пропуски, нормализовать данные, удалить дубликаты).

Визуализировать значимые признаки с помощью таких методов как:

- Диаграммы рассеяния
- Ящики с усами
- Гистограммы

Проанализировать корреляцию данных (матрица корреляций)

Описание решения задачи

В качестве датасета был выбран https://www.kaggle.com/datasets/mahdimashayekhi/disease-risk-from-daily-habits/data в виду не разобранности, большого количества признаков и записей, хорошего целевого признака (здоров / болен).

Детальное и подробное описание выполнение лабораторной работы №1 представлено в среде Google Colab: https://colab.research.google.com/drive/1gVbkpqhRaBVxy9vG7_4ODncTRRoH5k_Z?usp=sharing

Альтернативный вариант на GitHub: https://github.com/brick1ng5654/course-3/blob/main/boml/lab_01/boml_lab01.ipynb

В датасете более 100+ тысяч строк и 41 признак. Выберем необходимое и отсеем лишнее.

Целевым признаком для предсказания будет target (Бинарная классификация): healthy (здоров) / diseased (болен). Важное уточнение, что данное понятие достаточно расплывчатое.

Из числовых признаков возьмём:
- age (Возраст)
- bmi (Индекс массы тела)
- blood pressure (Артериальное давление)
- calorie intake (Потребление калорий)
- cholesterol (Холестерин)
- daily steps (Ежедневное количество шагов)
- glucose (Глюкоза)
- heart rate (Пульс)
- insulin (Инсулин)
- sleep hours (Количество часов сна)
- stress level (Заявленный уровень стресса)
- sugar intake (Потребление сахара)
- water intake (Потребление воды)
- work hours (Количество часов работы)
- meals per day (Количество приемов пищи в день)

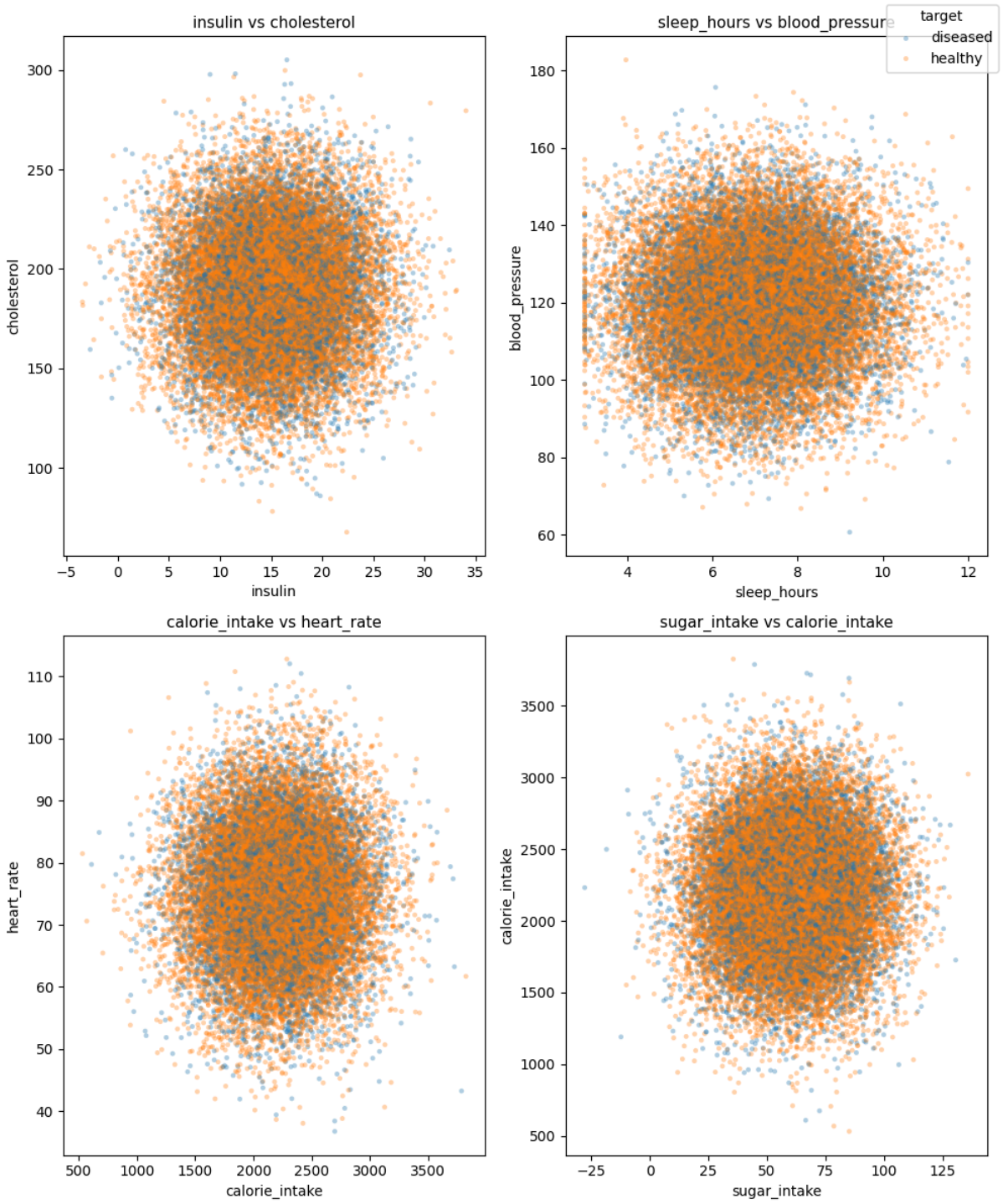
Из категориальных признаков возьмём:
- alcohol consumption (Потребление алкоголя, [Occasionally, Regularly, None])
- caffeine intake (Потребление кофеина, [Moderate, High, None])
- diet type (Тип диеты, [Vegan, Omnivore, Vegetarian, Keto])
- exercise type (Тип тренировок, [Strength, Cardio, None, Mixed])
- gender (Пол, [Male, Female])
- sleep quality (Качество сна, [Good, Excellent, Fair, Poor])
- smoking level (Потребление сигарет, [Light, Non-smoker, Heavy])
- sunlight exposure (Ежедневное воздействие солнечного света, [Low, Moderate, High])

Таким образом были убраны следующие признаки:
- bmi ?* (В виду взятия признака bmi, вследствие чего остальные показатели bmi не нужны)
- daily supplement dosage (В виду не до конца понятной классификации)

- device usage (В виду наименьшего влияния на результат заболевания)
- education level (В виду наименьшего влияния на результат заболевания)
- electrolyte level (В виду бесполезности признака, т.к. все строки одинаковые)
- enviromnent risk score (В виду бесполезности признака, т.к. все строки одинаковые)
- family history (В виду не до конца понятной бинарной классификации)
- gene marker flag (В виду бесполнезности признака, т.к. все строки одинаковые)
- healthcare access (В виду наименьшего влияния на результат заболевания)
- height (В виду ненужности признака из-за наличия признака bmi)
- income ((В виду наименьшего влияния на результат заболевания)
- insurance (В виду наименьшего влияния на результат заболевания)
- job type ((В виду наименьшего влияния на результат заболевания)
- mental health score (В виду схожести с заявленным уровнем стресса)
- mental health support (В виду наименьшего влияния на результат заболевания)
- occupation (В виду наименьшего влияния на результат заболевания)
- pet owner (В виду наименьшего влияния на результат заболевания)
- physical activity (В виду не до конца понятного получения значения признака)
- screen time (В виду наименьшего влияния на результат заболевания)
- waist size (В виду наименьшего влияния на результат заболевания)
- weight (В виду ненужности признака из-за наличия признака bmi)

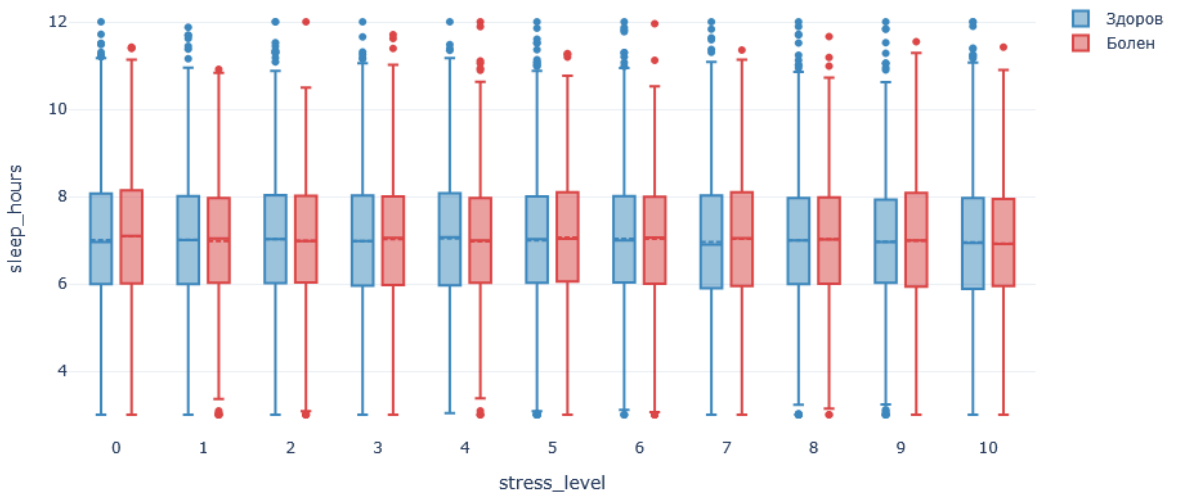
Основная работа и детальное описание действие представлено в Google Colab. Дальше в отчете будет представлен уже результат работы

В результате работы получили следующие
Диаграммы рассеяния:



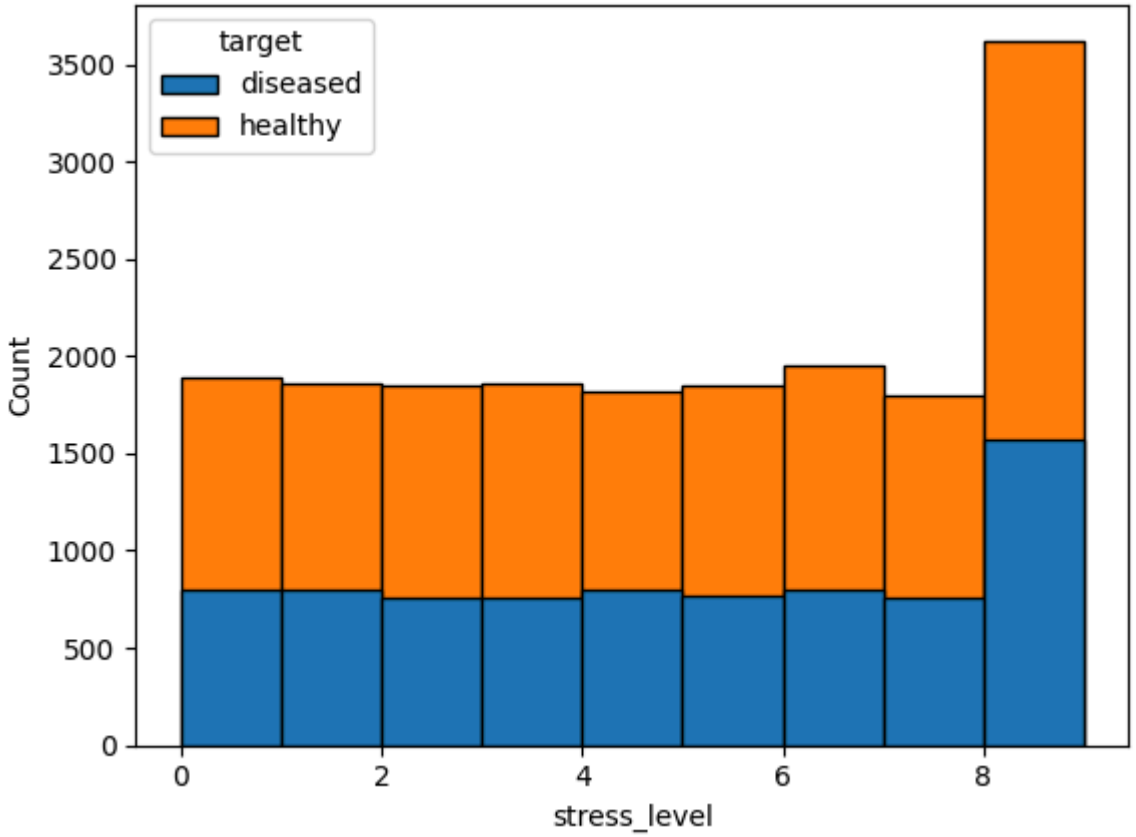
На диаграммах рассеяния не просматривается выраженных трендов - ни линейных, ни нелинейных. Зависимости выглядят слабыми и статистически неубедительными. Как итог сильной взаимосвязи между выбранными парами признаков не обнаружено

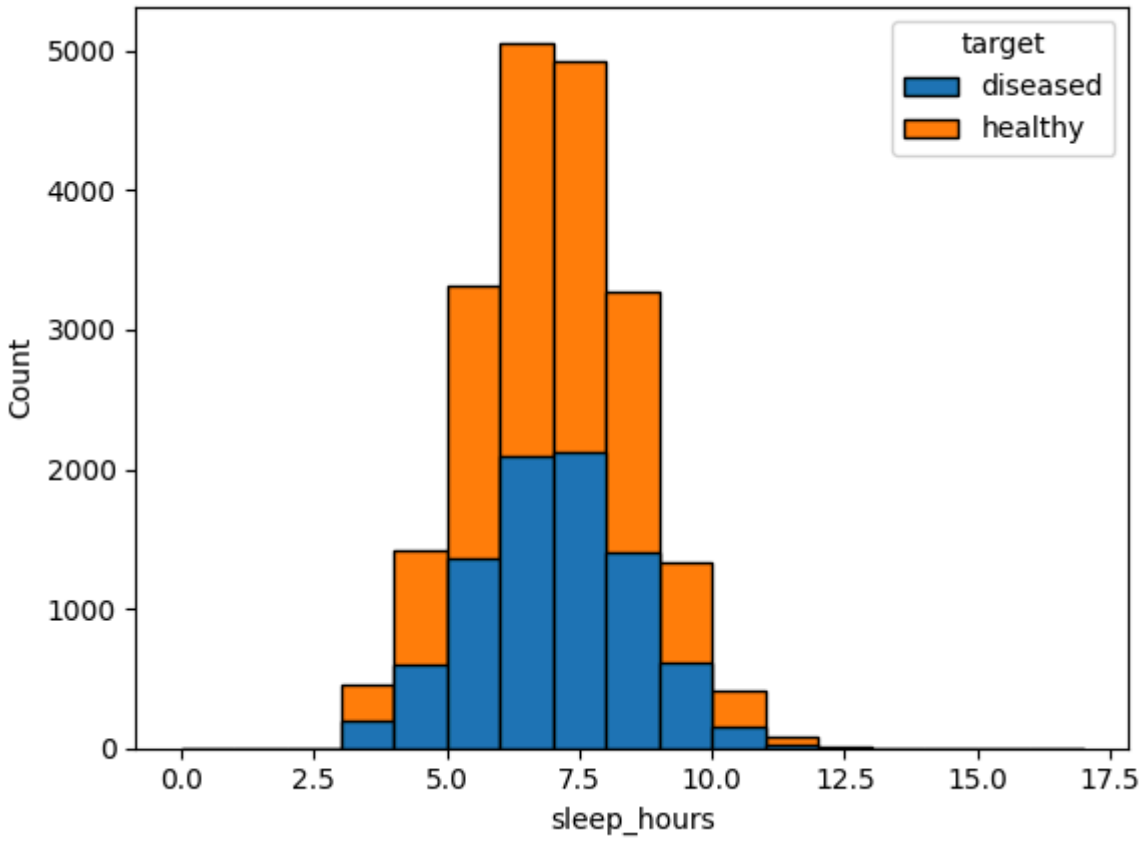
Ящик с усами:

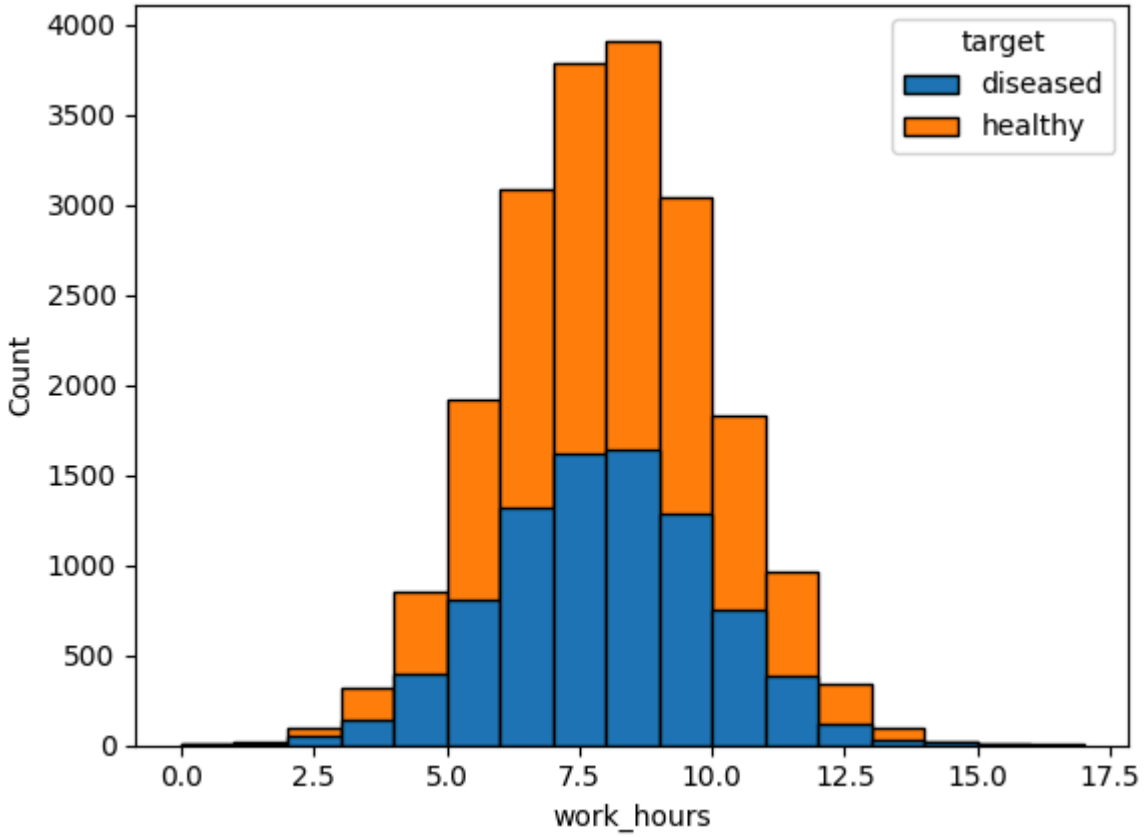


Больные в среднем спят немного меньше, стресс слегка укорачивает сон, однако сильной взаимосвязи на основе полученных данных не было обнаружено

Гистограммы:


target
diseased
healthy
Count
stress_level
0
500
1000
1500
2000
2500
3000
3500
0
2
4
6
8


target
diseased
healthy
Count
sleep_hours
0
1000
2000
3000
4000
5000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5



На этих признаках никакой сильной связи с целевым признаком нет. Некоторые наблюдения возможны для sleep_hours, но это слишком незначительно

Вывод

В ходе лабораторной работы мы самостоятельно выбрали набор данных, удалили в нем пропуски и дубликаты, нормализовали численные и категориальные признаки, визуализировали значимые признаки при помощи диаграмм рассеяния, ящика с усами и гистограмм, сделали и проанализировали матрицу корреляций. Получили и закрепили навыки предобработки данных для дальнейшего применения методов машинного обучения для решения задач