**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: Исследование набора данных

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Лемешко А. Д. |
| Преподаватель |  | Татчина Я. А. |

Санкт-Петербург

2024

**Цель работы.**

Выбор набора данных для дальнейшего исследования алгоритмов кластеризации и классификации. Получение практических навыков работы с инструментами анализа и визуализации данных.

**Постановка задачи.**

1. Создать Jupyter Notebook
2. Выбрать исследуемый набор данных из предложенных источников
3. Описать выбранный набор данных, а именно:
   1. Предметную область, источник данных, характер данных;
   2. Атрибуты, их тип, что они обозначают;
   3. Задачу анализу.
4. Для каждого атрибута:
   1. Определить среднее значение, СКО;
   2. Построить гистограмму распределения значений, определить наличие выбросов;
   3. Определить наличие пропущенных значений и их количество;
   4. Предложить вариант обработки пропущенных значений.
5. Определить:
   1. Высоко коррелированные атрибуты, характер корреляции;
   2. Не коррелированные атрибуты;
   3. Графики рассеивания.
6. Проанализировать полученные результаты.

**Выполнение работы.**

Для исследования был выбран реальный набор данных, описывающих затраты времени студентами, их оценки и уровень стресса. Проанализированы признаки, которые содержит набор (см. приложение А).

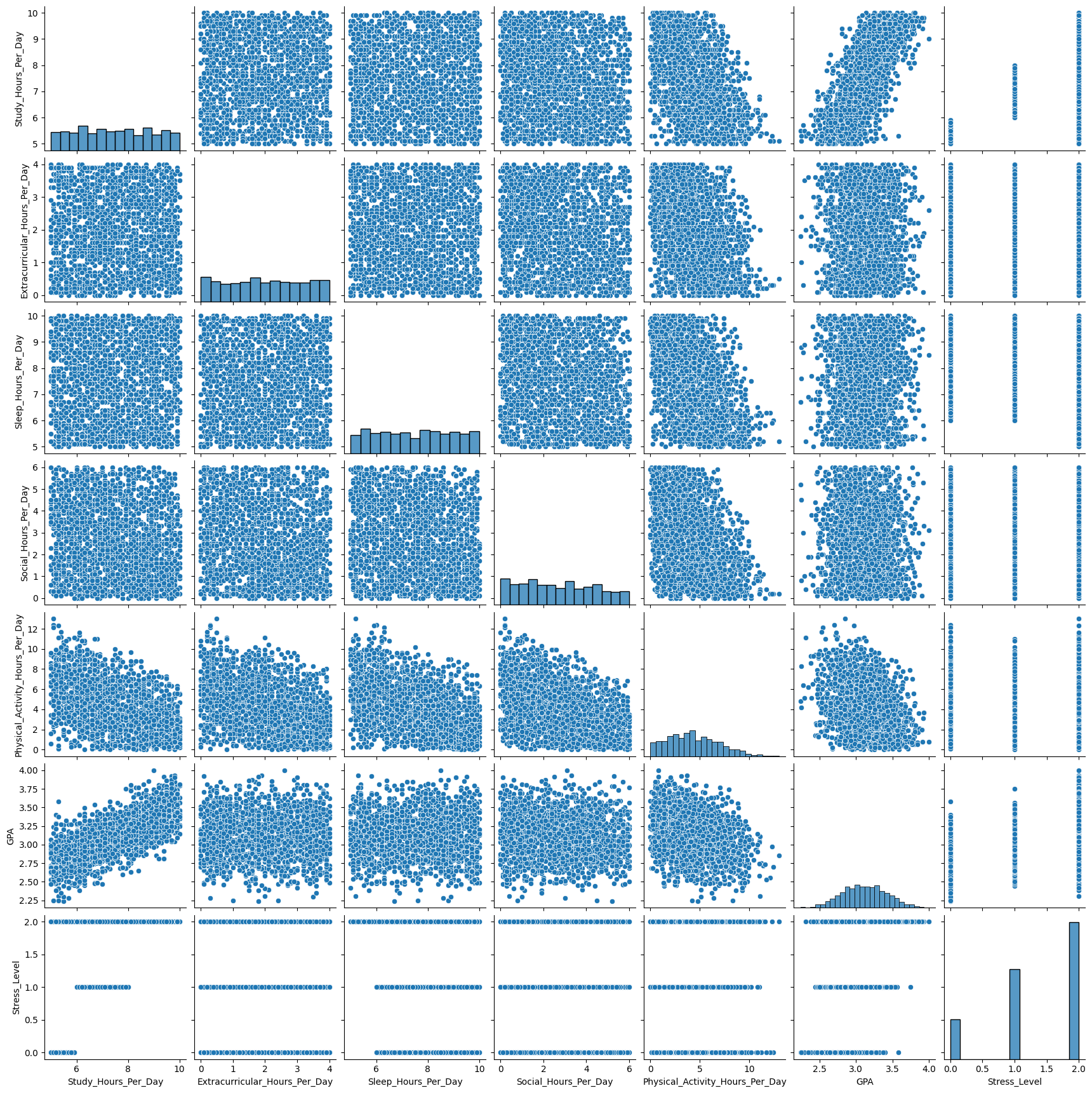
Далее прочитаны данные, значениям атрибута Stress\_Level были присвоены числовые значения: Low – 0, Moderate – 1, High - 2.

С помощью средств pandas.DataFrame.describe и seaborn.histplot вычислены основные параметры признаков и построены гистограммы распределения соответственно. Из гистограмм видно, что набор данных содержит выбросы в атрибуте с временем физической активности и оценках, но не содержит пропусков.

Применив seaborn.heatmap() были выделены коррелированные атрибуты и выдвинуты основные предположения для предобработки, а именно:

1. Сильная положительная корреляция между GPA и Study\_Hours\_Per\_Day (0.734)
2. Умеренная отрицательная корреляция между Physical\_Activity\_Hours\_Per\_Day и Study\_Hours\_Per\_Day (-0.488)
3. Сильная положительная корреляция между Stress\_Level и Study\_Hours\_Per\_Day (0.739)
4. Умеренная отрицательная корреляция между Physical\_Activity\_Hours\_Per\_Day и Extracurricular\_Hours\_Per\_Day (-0.37)
5. Умеренная отрицательная корреляция между Physical\_Activity\_Hours\_Per\_Day и Sleep\_Hours\_Per\_Day (-0.47)
6. Умеренная отрицательная корреляция между Physical\_Activity\_Hours\_Per\_Day и Social\_Hours\_Per\_Day (-0.417)
7. Умеренная отрицательная корреляция между Physical\_Activity\_Hours\_Per\_Day и GPA (-0.341)
8. Умеренная положительная корреляция между GPA и Stress\_Level (-0.341)

Далее была построена матрица графиков рассеивания (рис. 1).



*Рис. 1: Матрица графиков рассеивания*

На графиках подтвердилась умеренная корреляция между временем физической активности и временем учебы, внеучебной деятельности, сна и социального взаимодействия, а также оценкой. Также частично подтвердилась корреляция между уровнем стресса и оценкой, временем учебы и сна (при сильном уровне стресса корреляции не обнаружено)

**Выводы.**

В ходе работы были выбран набор данных для исследования. Получены практические навыки работы с анализом и визуализацией данных с помощью инструментов Jupyter Notebook, NumPy, Pandas, matplotlib и seaborn.

Приложение А

**Атрибуты набора данных**

|  |  |  |  |
| --- | --- | --- | --- |
| Название атрибута | Тип данных | Описание | Пример |
| Student\_ID | integer | Идентификатор студента | 37 |
| Study\_Hours\_Per\_Day | float | Часы учебы | 4.5 |
| Extracurricular\_Hours\_Per\_Day | float | Часы внеучебной деятельности | 4.5 |
| Sleep\_Hours\_Per\_Day | float | Часы сна | 4.5 |
| Social\_Hours\_Per\_Day | float | Часы социального взаимодействия | 4.5 |
| Physical\_Activity\_Hours\_Per\_Day | float | Часы физической активности | 4.5 |
| Stress\_Level | string | Уровень стресса | 1: Moderate |
| GPA | float | Средний балл | 2.5 |