**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод кластеризации K-mean

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

Санкт-Петербург

2024

**Цель работы.**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи.**

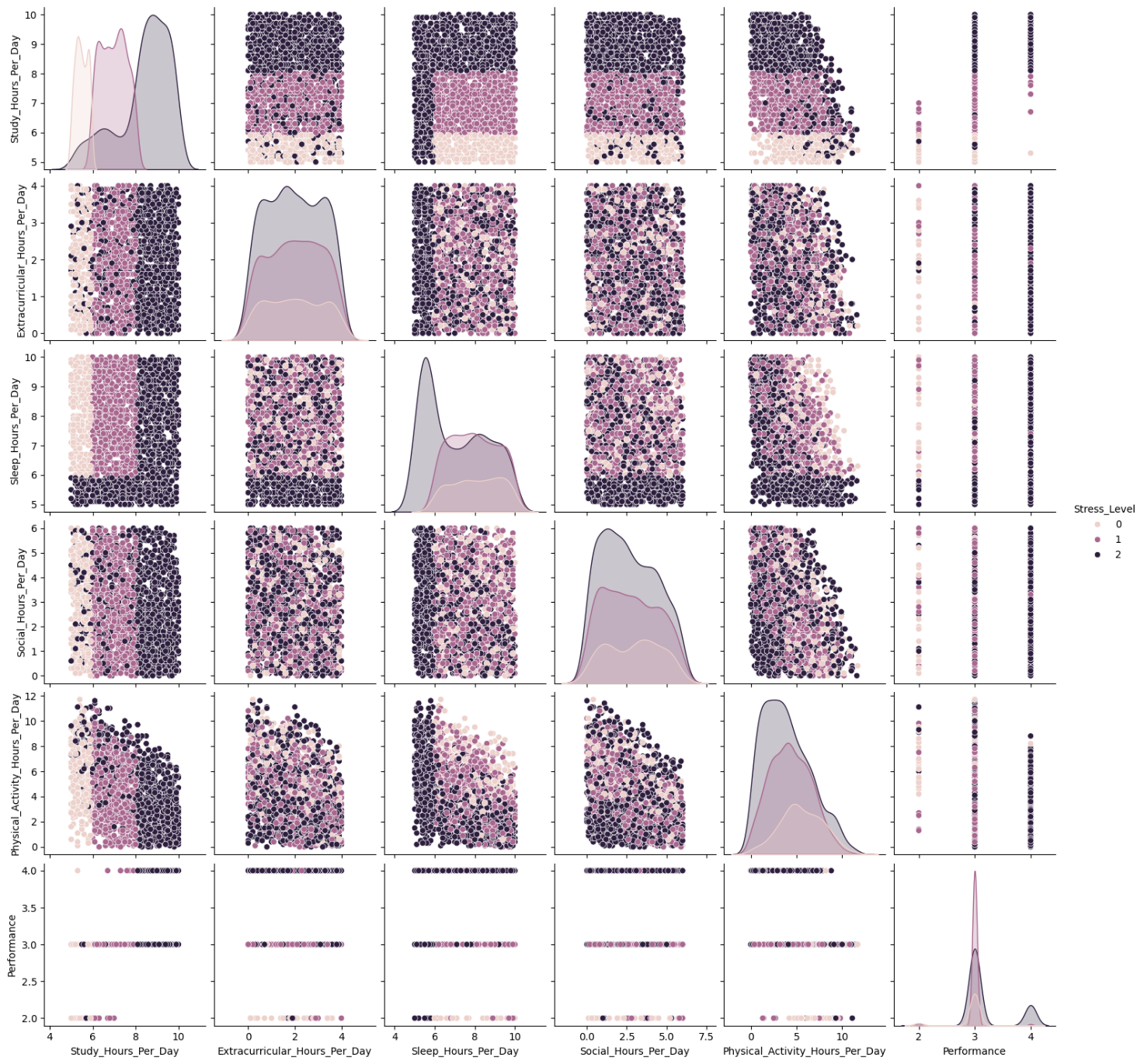
1. Добавить к выбранному набору данных новый атрибут, вычислив его из имеющихся;
2. Очистить данные, удалив выбросы и дубликаты; обработать пропущенные значения; найти искаженные данные;
3. Применить метод кластеризации K-mean к текущему набору данных
4. Проанализировать полученные результаты.

**Выполнение работы.**

Для данной лабораторной работы был выбран старый датасет с информацией о студентах.

Был создан новый атрибут: вещественный атрибут GPA был заменен на целочисленный Performance, где 0 – низкий балл, 1 – средний, 2 - высокий.

После предобработки данных был проведён анализ атрибутов и их визуализация. Так, графики зависимостей между атрибутами можно увидеть на рисунке 1.



*Рисунок 1. Графики зависимостей*

На графиках, как и в первой работе, можно заметить, что уровень стресса сильно зависит от количества сна и количества учебы.

Для кластеризации методом k-means было выбрано число k = 3. После применения метода данные разделились на 3 группы, а именно:

1. Студенты с высоким уровнем стресса;
2. Студенты с умеренным или низким уровнем стресса;
3. Студенты с умеренным или высоким уровнем стресса.

**Выводы.**

В ходе работы были изучены и обработаны данные студентов. Были выявлены зависимости и проведена кластеризация. В результате были получены навыки кластеризации с помощью метода K-means.