**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра ИС**

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

**Тема: Метод кластеризации K-Mean**

|  |  |  |
| --- | --- | --- |
| Студентка гр. 2372 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Мельникова М. А. |
| Преподаватель | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

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

**Задание на работу.**

1. К текущему датасету необходимо добавить новый аттрибут, например, если есть дата рождения, то посчитать возраст или определить популярность фильма по количеству просмотров и т.п.
2. "Причесать" датасет: удалить выбросы и дубли, обработать пропущенные значения, найти кривые данные и т.п.
3. Построить графики зависимости одной переменной от другой (plt.scatter), описать какие кластеры (группы) вы видите. Построить несколько разных графиков, найти ярковыраженные группы.
4. Изучить перечисленные ноутбуки, по примеру, попробовать применить метод KMeans к вашему датасету.

**Выполнение работы.**

Исследование продолжим на наборе данных о студенческом образе жизни.

Атрибуты датасета: идентификатор студента, количество учебных часов в день, время дополнительной активности в день, время сна в день, время социальной активности в день, часы физической активности в день, учебный балл, уровень стресса.

Добавим новый атрибут: процент времени, затрачиваемого на учёбу от суточного времени 'Study\_Percentage'.

df['Study\_Percentage'] = (df['Study\_Hours\_Per\_Day'] / 24) \* 100

Из прошлой лабораторной знаем, что в датасете нет выбросов и пропущенных значений, поэтому этап приведения дадасета мы пропускаем –он уже имеет пригодный к исследованию вид.

Только значение стресса в датасете текстовое. Поменяем значения атрибута 'Stress\_Level' в числовые значения соответственно "Low" в 0,2, "Moderate" в 0,6, "High" в 0,9.

stress\_mapping = {

    "Low": 0.2,

    "Moderate": 0.6,

    "High": 0.9

}

df['Stress\_Level'] = df['Stress\_Level'].map(stress\_mapping)

Построим графики зависимости одной переменной от другой plt.scatter(). Построим несколько разных графиков, чтобы найти ярковыраженные группы.

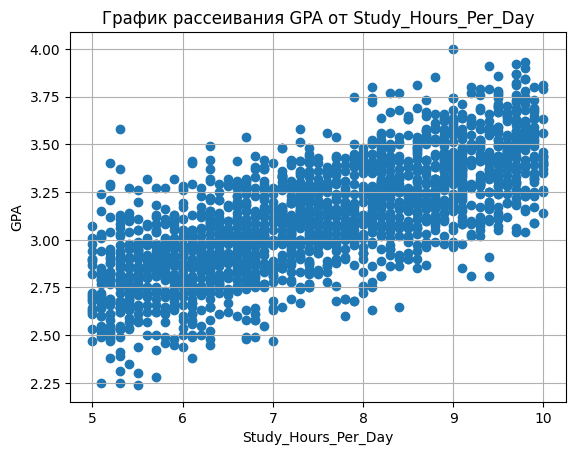
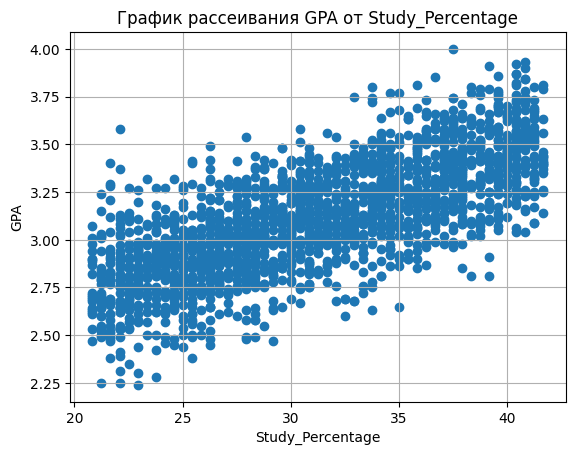
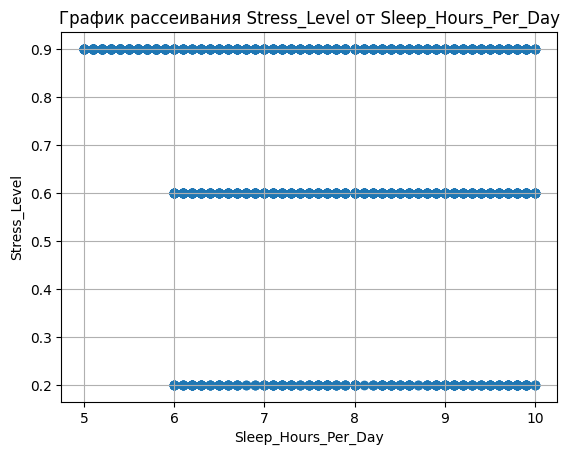
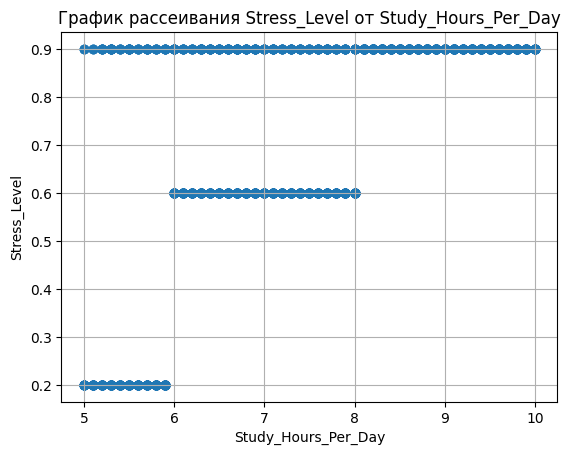
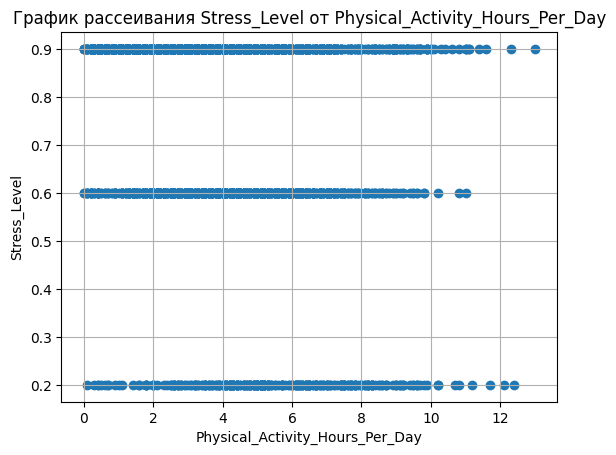
     

Рис1–6. Графики рассеивания

Никаких ярко выраженных кластеров нет.

График рассеивания зависимости Study\_Hours\_Per\_Day от Stress\_Level демонстрирует, что уровень стресса не напрямую зависит от времени учёбы, однако можно выявить такие закономерности: самый низкий уровень стресса испытывают только студенты, тратящие минимальное количество времени, из данной выборки, на учёбу 5-6 часов. Средний уровень стресса испытывают только те студенты, которые тратят среднее по данной выборке время на учёбу 6-8. Самый сильный уровень стресса могут испытывать все студенты. Но те, кто учится наибольшее время 8-10 часов испытывают исключительно сильный стресс.

К зависимости Study\_Hours\_Per\_Day от Stress\_Level и применим метод кластеризации. Визуально выделяется 3 основных группы, описанные выше, поэтому применяя метод кластеризации KMeans зададим количество кластеров параметром n\_clusters = 3.

from sklearn.cluster import KMeans

X = df[['Study\_Hours\_Per\_Day', 'Stress\_Level']]

kmeans = KMeans(n\_clusters=3, random\_state=0)

kmeans.fit(X)

labels = kmeans.labels\_

df['Cluster'] = labels

plt.scatter(df['Study\_Hours\_Per\_Day'], df['Stress\_Level'], c=labels, cmap='viridis')

plt.scatter(kmeans.cluster\_centers\_[:, 0], kmeans.cluster\_centers\_[:, 1], s=150, c='green', marker='X', label='Центроиды')

plt.title('KMeans Кластеризация Study Hours, Stress Level')

plt.xlabel('Study Hours Per Day')

plt.ylabel('Stress Level')

plt.legend()

plt.show()

Получим график кластеризации:

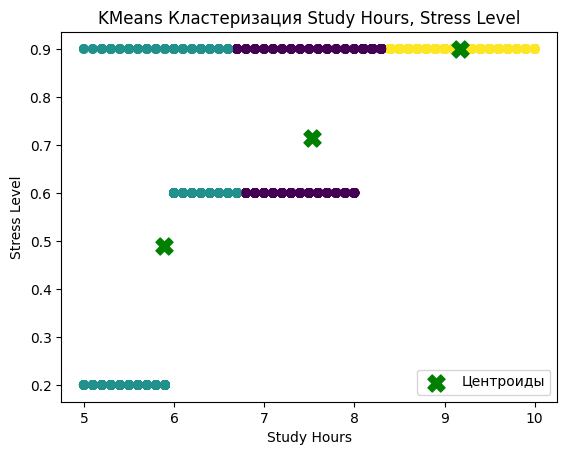


Рис7. График кластеризации для зависимости Stress Level от Study Hours

Построен график кластеризации: данные сгруппированы кластеры и обозначены цветом, найдены центроиды кластеров. Кластеры почти совпадают с описанной выше зависимостью.

Построим еще графики кластеризации, для зависимостей, на графиках которых нет выраженных групп. Значения n\_clusters возьмём 5 и 4, в данном случае выбор параметр количества кластеров не обусловлен естественной выраженностью групп, служит для иллюстрации того, как будет проходит разбиение на кластеры при разных вводных данных, может применяться для равномерного распределения выборки на заданное количество кластеров.

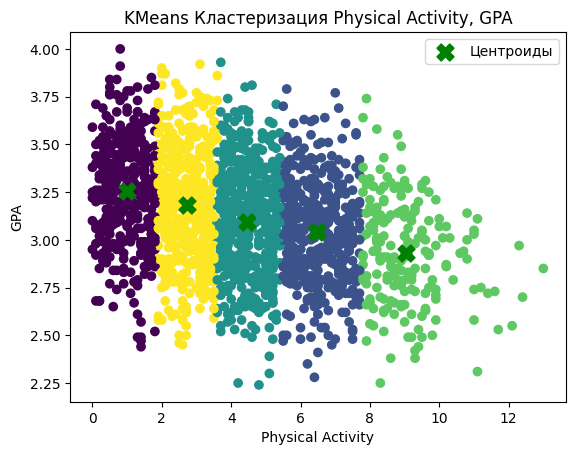


Рис8. График кластеризации для зависимости GPA от Physical Activity

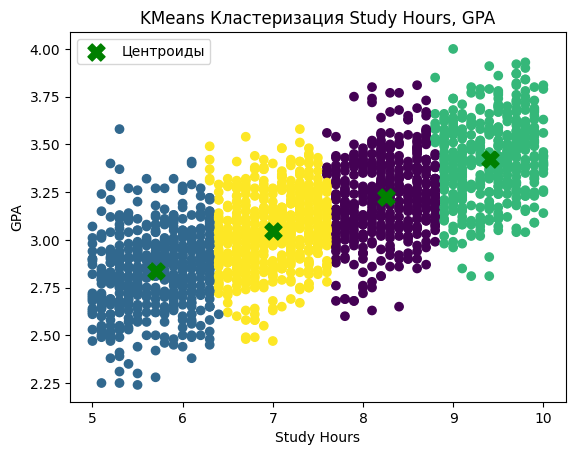


Рис8. График кластеризации для зависимости GPA от Study Hours

Из двух последних построенных графиков кластеризации видим, что если кластеров в явном виде нет, то точки на графике группируются с ближайшими просто на равные части.

**Выводы.**

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