**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра ИС**

**ОТЧЕТ**

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

**Тема: Исследование набора данных**

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

Санкт-Петербург

2024

**Цель работы.**

Выбор набора данных, который будет использоваться в дальнейшем при исследовании алгоритмов кластеризации и классификации.

**Задание на работу.**

1. Выбрать исследуемый датасет: числовые данные, с разметкой.

2. Представление краткого описания датасета:

• предметная область, источник данных, характер данных;

• атрибуты датасета, их тип, значение;

• представить описание задачи анализа.

3. Для каждого атрибута нужно определить:

• среднее значение, ско;

• построить гистограмму распределения значений, определить есть ли выбросы;

• есть ли пропущенные значение, сколько;

• предложить вариант обработки пропущенных значений.

4. Определить корреляцию между параметрами

• какие атрибуты высококоррелированы, характер корреляции;

• определить какие атрибуты не имеют корреляцию;

• построить графики рассеивания (предпочтительнее матрицу графиков рассеивания);

• проанализировать полученные результаты.

**Выполнение работы.**

Для исследования был выбран набор данных о студенческом образе жизни.

Этот набор данных дает подробное представление о моделях образа жизни учащихся и их корреляции с академической успеваемостью, представленной средним баллом успеваемости.

Предметная область: образование.

Источник данных: https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset

Характер данных: реальные.

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.

Тип представленных данных – числовой.

Задача анализа: выявить зависимость успеваемости – среднего балла студентов от количества часов учёба в день.

Исследуемые атрибуты:

Атрибут 'Study\_Hours\_Per\_Day' - количество часов учебы в сутки

Атрибут 'GPA' - средний балл

Получаем доступ к атрибутам df['GPA'], df['Study\_Hours\_Per\_ Day'], вычисляем стандартное отклонение .std() и среднее значение .mean().

Среднее значение атрибута 'Study\_Hours\_Per\_Day': 7.4758000000000004

Стандартное отклонение атрибута 'Study\_Hours\_Per\_Day': 1.423888369586789

Среднее значение атрибута 'GPA': 3.11596

Стандартное отклонение атрибута 'GPA': 0.2986735363916102

Строим гистограмму hist() по тем же атрибутам.

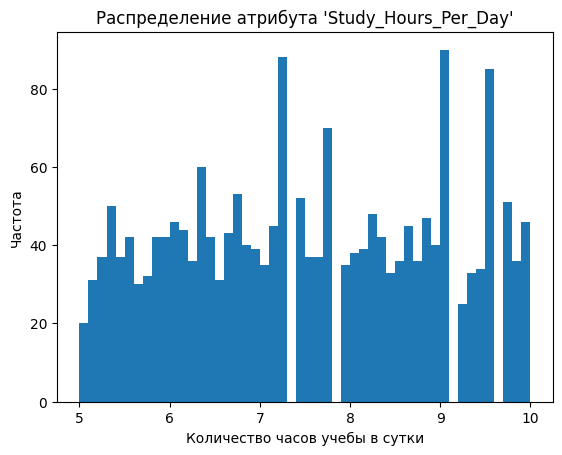
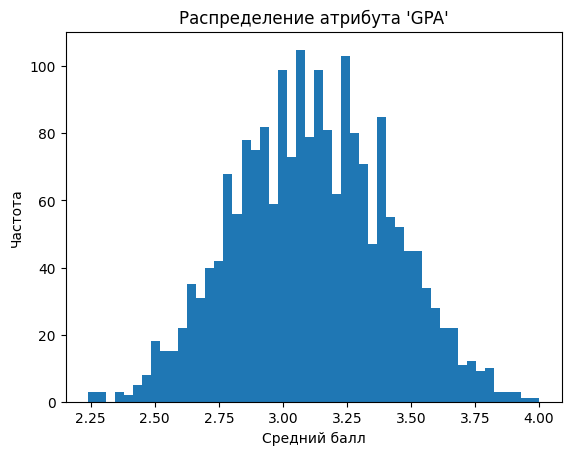
 

Рис1. Гистограммы

Из гистограмм видим, что выбросов данных у исследуемых атрибутов нет.

Функция isnull() заменяет пропущенные значения на True, если значение есть, то False. Применяется по умолчанию к столбцам .sum(), складывает полученые isnull() значения, получает количество пропущенных значений в каждом столбце.

Количество пропущенных значений:

Student\_ID 0

Study\_Hours\_Per\_Day 0

Extracurricular\_Hours\_Per\_Day 0

Sleep\_Hours\_Per\_Day 0

Social\_Hours\_Per\_Day 0

Physical\_Activity\_Hours\_Per\_Day 0

GPA 0

Stress\_Level 0

Пропущенных значений у нет.

Вычисляем корреляцию между параметрами corr().

Значение корреляции 0.734468 положительное, по модулю блишзкое к 1 означает, что зависимость прямая и достаточно сильная: чем больше учебных часов в сутки, тем выше средний балл.

Строим график рассеивания scatterplot() по параметрам 'Study\_ Hours\_Per\_Day' и 'GPA'.

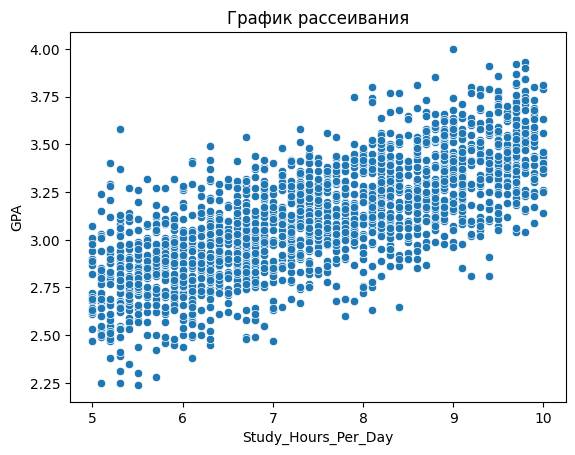


Рис2. График рассеивания

График рассеивания позволяет визуально оценить взаимосвязь между значениями этих атрибутов. Из графика видно, что взаимосвязь ярко выражена, точки на графике имеют тенденцию к расположению на прямой вида некоторой линейной функции.

**Выводы.**

В ходе работы был выбран набор данных для исследования. Получены практические навыки работы с Jupyter Notebook, анализа и визуализации данных с помощью инструментов pandas, seaborn, matplotlib.pyplot.