**МИНОБРНАУКИ РОССИИ САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №1**

**по дисциплине «Машинное обучение»**

**Тема: «Исследование набора данных»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чесноков М. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Татчина Я. А.

Санкт-Петербург

2024

**Цель работы:** Провести статистический анализ данных выбранного датасета.

**Краткое описание датасета**

Датасет об использовании смартфонов. Датасет был взят с сайта [www.kaggle.com](http://www.kaggle.com/). Этот набор данных дает представление о ежедневном использовании мобильных устройств 1000 пользователей, охватывая такие аспекты, как время использования экрана, использование приложений и вовлеченность пользователей в различных категориях приложений. Он включает в себя разнообразный спектр пользователей по возрасту, полу и местоположению. Данные фокусируются на общем использовании приложений, времени, проведенном в социальных сетях, приложениях для повышения производительности и игровых приложениях, а также на общем времени использования экрана.

Этот набор данных полезен для анализа мобильного взаимодействия, привычек использования приложений и влияния демографических факторов на мобильное поведение. Он может помочь определить тенденции для маркетинга, разработки приложений и оптимизации пользовательского опыта.

В датасете представлены следующие атрибуты:

- User\_ID - уникальный идентификатор для каждого пользователя (ключевой атрибут) (числовой тип)

- Age - возраст пользователя от 18 до 60 лет (числовой тип)

- Gender - Пол пользователя (строковый тип)

- Total\_App\_Usage\_Hours - общее количество часов, проведенных в мобильных приложениях за день (числовой тип)

- Daily\_Screen\_Time\_Hours - общее время, проведенное перед экраном, включая все действия, в часах (числовой тип)

- Number\_of\_Apps\_Used - количество различных приложений, используемых пользователем за день (числовой тип)

- Social\_Media\_Usage\_Hours - количество часов, посвященное приложениям социальных сетей (числовой тип)

- Productivity\_App\_Usage\_Hours - количество часов, посвященныое приложениям для повышения производительности (числовой тип)

- Gaming\_App\_Usage\_Hours - количество часов, посвященное игровым приложениям (числовой тип)

- Location - город или регион, где проживает пользователь (строковый тип)

Этот датасет был считан с помощью библиотеки pandas, и с ним была произведена дальнейшая работа.

**Определение параметров**

Для каждого числового атрибута были определены среднее значение и СКО. Для этого использовались функции библиотеки numpy: np.mean() и np.std() (данные представлены в файле с программой).

Также для каждого числового атрибута была построена гистограмма распределения значений при помощи функции библиотеки matplotlib plt.hist() (гистограммы представлены в файле с программой).

Проверка на наличие выбросов проводилась в функции FindOutliers с использованием квартилей. У всех атрибутов данного датасета отсутствуют выбросы.

Проверка на наличие пропущенных значений проводилась при помощи функции библиотеки pandas pd.isna().sum(), которая выводит количество пропущенных значений для данного атрибута. У всех атрибутов данного датасета отсутствуют пропущенные значения.

**Корреляция между параметрами**

Корреляция — это взаимосвязь между разными показателями в статистике. Её используют, чтобы оценить зависимость переменных друг от друга. Если два показателя коррелируют друг с другом, выше вероятность, что они как-то связаны: например, один зависит от другого или они оба зависят от третьей переменной.

Корреляция может быть:

* положительной — когда один показатель растет, другой тоже растет;
* отрицательной — когда одна переменная растет, другая уменьшается;
* нейтральной — изменения не связаны друг с другом.

Была определена корреляция для всех пар числовых атрибутов. Эти данные представлены в корреляционной матрице (данные представлены в файле с программой).

По полученным данным видно, что корреляция между атрибутами нейтральная, то есть в данном датасете не удалось выявить взаимосвязь между ними.

Для демонстрации отсутствия корреляции между параметрами в программе представлены графики рассеивания всех величин в зависимости от возраста пользователей. На графиках прекрасно видно отсутствие взаимосвязи между атрибутами (графики представлены в файле с программой).

**Вывод**

В данной работе я провёл статистический анализ датасета о ежедневном использовании мобильных устройств 1000 пользователей, охватывая такие аспекты, как время использования экрана, использование приложений и вовлеченность пользователей в различных категориях приложений.

Определив корреляцию атрибутов, стало понятно, что в данном датасете между ними отсутствует взаимосвязь. Я считаю, что на большей выборке могут появиться взаимосвязи между атрибутами, например, положительная корреляция между общим количеством часов, проведенных в мобильных приложениях за день и количеством различных приложений, используемых пользователем за день.