**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ 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. 2373 Никитина М.Ю.

Преподаватель Татчина Я.А.

Санкт-Петербург

2024

Задание:

1. Создать Jupyter Notebook, переименовать его «Lab 1, № Группы, ФИО»

1. Выбор исследуемого датасета.

2. Для каждого датасета представить краткое его описание в вашем Jupyter Notebook:

• предметная область, источник данных, характер данных (реальные или имитационные)

• какие атрибуты представлены в датасете, их тип (числовой, строковый (категории)), что они обозначают

• есть ли описание задачи анализа, если есть - представить

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, определить характер корреляции

• какие атрибуты не имеют корреляцию

• постройте графики рассеивания (предпочтительнее матрицу графиков рассеивания)

• проанализируйте полученные результаты.

Для выполнения данной лабораторной работы был взята база данных Abba Music Characteristics (Spotify), скаченная с сайта <https://www.kaggle.com/datasets/thedevastator/abba-music-characteristics-spotify?resource=download>.

В данном датасете представлена информация насчет самых популярных песен (хитов) шведской группы ABBA на известной платформе Spotify. Все данные реальны.

Датасет включает в се6я следующие атрибуты:

album - название альбома (String)

track\_number - трек-номер песни в альбоме (Integer)

uri - уникальный идентификатор песни (String)

acousticness - мера акустики песни (Float)

danceability - насколько хорошо песня хороша для танцев (Float)

energy - энергичность песни (Float)

instrumentalness - инструментальность песни (Float)

liveness - оригинальность песни (Float)

loudness - громкость песни (Float)

speechiness - мера кол-ва текста в песни (Float)

tempo - темп песни (Float)

valence - позитивность песни (Float)

popularity - популярность песни (Integer)

artist - исполнитель песни (String)

Находить среднее значение и СКО атрибутов будем с помощью библиотеки **numpy**, которая предоставляет поддержку для работы с многомерными массивами и математическими функциями.

Изображение выглядит как текст, снимок экрана, документ, Шрифт

Автоматически созданное описаниеНаходим данные значения через встроенные функции .mean() и .std().

Изображение выглядит как текст, документ, снимок экрана, Шрифт

Автоматически созданное описание

Полученные гистограммы:

* Изображение выглядит как диаграмма, График, линия, снимок экрана

  Автоматически созданное описаниеАкустика песен:
* Изображение выглядит как диаграмма, График, линия, текст

  Автоматически созданное описаниеТанцевальность песен:
* Изображение выглядит как диаграмма, График, снимок экрана, линия

  Автоматически созданное описаниеЭнергичность песен:
* Изображение выглядит как текст, линия, снимок экрана, График

  Автоматически созданное описаниеИнструментальность песен:
* Изображение выглядит как текст, График, диаграмма, линия

  Автоматически созданное описаниеОригинальность песен:
* Изображение выглядит как диаграмма, График, линия, снимок экрана

  Автоматически созданное описаниеГромкость песен:
* Изображение выглядит как текст, График, линия, снимок экрана

  Автоматически созданное описаниеКол-во текста в песнях:
* Изображение выглядит как диаграмма, График, текст, линия

  Автоматически созданное описаниеТемп песен:
* Изображение выглядит как диаграмма, График, линия, снимок экрана

  Автоматически созданное описаниеПозитивность песен:
* Изображение выглядит как диаграмма, График, линия, текст

  Автоматически созданное описаниеПопулярность песен:

Для определения корреляции были использованы следующие зависимости:

1. танцевальность песен от их энергичности
2. акустика песен от инструментальности
3. популярность песен от оригинальности
4. позитивность песен от темпа
5. громкость песен от кол-ва текста

Судя по коэффициентам, можно сказать, что:

1. коэф. = 0.13. При большей энергичности танцевальность песен возрастает
2. коэф. = -0.12. Акустика убывает при увеличении инструментальности
3. коэф. = 0.027. Популярность песен не сильно зависит от оригинальности
4. коэф. = 0.05. Позитивность также не сильно зависит от темпа
5. коэф. = 0.03. Громкость также не сильно зависит от кол-ва текста

Полученные графики рассеивания:

* Изображение выглядит как снимок экрана, диаграмма

  Автоматически созданное описаниетанцевальность песен от их энергичности:
* Изображение выглядит как текст, снимок экрана, линия, График

  Автоматически созданное описаниеакустика песен от инструментальности:
* Изображение выглядит как снимок экрана, текст, диаграмма, линия

  Автоматически созданное описаниепопулярность песен от оригинальности:
* позитивность песен от темпа:

Изображение выглядит как снимок экрана, диаграмма, текст, линия

Автоматически созданное описание

* Изображение выглядит как текст, снимок экрана, линия, График

  Автоматически созданное описаниегромкость песен от кол-ва текста

Проанализировав данные графики, можно сказать, что каждый атрибут не сильно зависит от других.