Project:

Books Recommendation System

An example of unsupervised learning model

Book Recommendation System

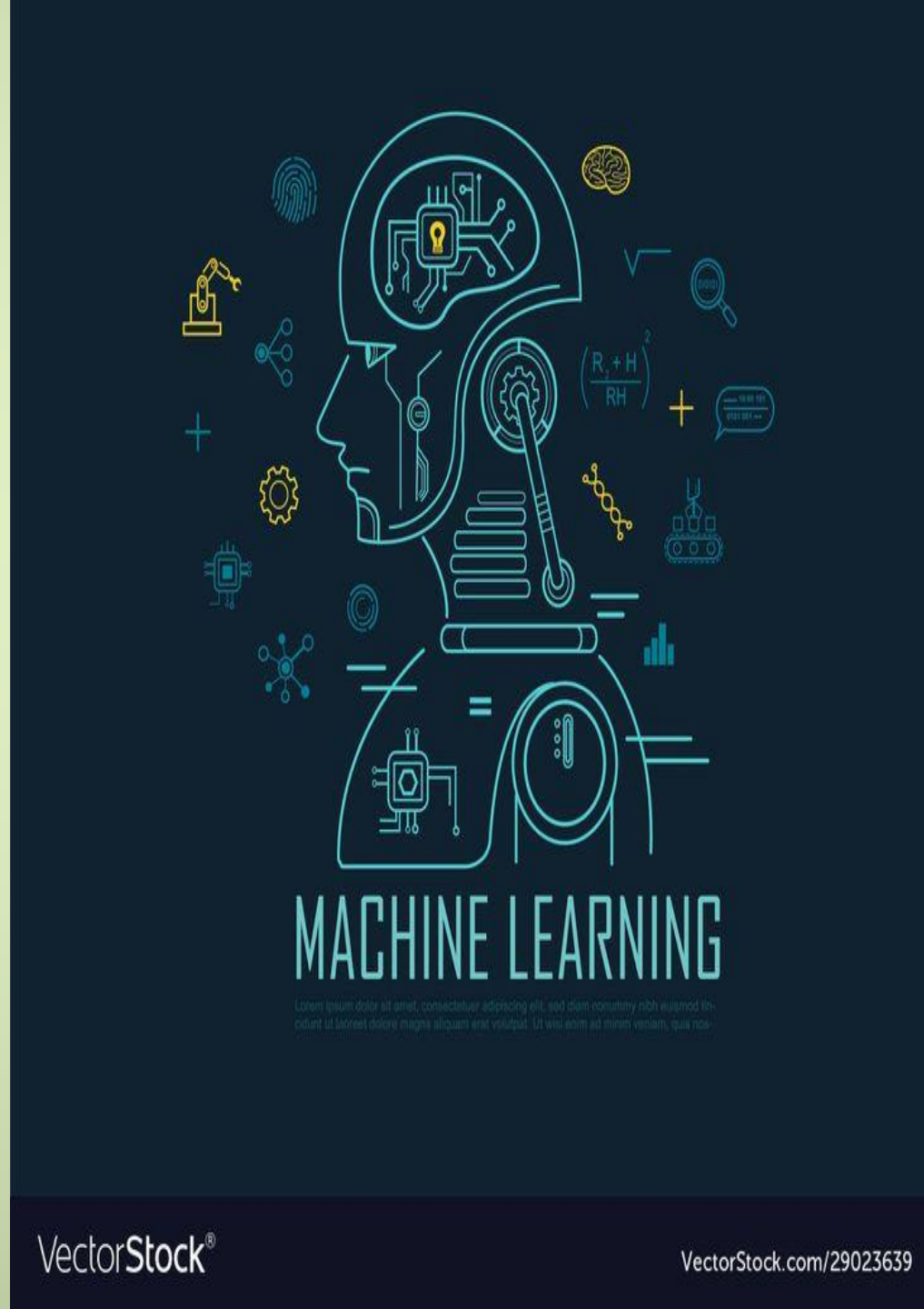
- A Book Recommendation System is a data-driven application designed to suggest books to users based on their preferences, reading history, and behavior.
- It employs various data science and machine learning techniques to provide personalized book recommendations, enhancing the reading experience for users.
- Our Book Recommendation System provides a list of similar books to the book title provided by the user.

Type of ML Problem

- The project focuses on building a book recommendation system.
- The problem addressed is the challenge of helping users discover relevant books based on their preferences.
- The goal is to enhance the user experience in finding books by providing personalized recommendations.
- Book Recommendation System in this case is

treated as

  - **Un-Supervised** Machine Learning Problem



Dataset Used for Evaluation

The dataset used for evaluation has following features:

- Total no. of instances/books = 11128
- There are 4 columns in the dataset comprising of
- Book ID, Title, Authors, and Average ratings.

The dataset serves as the foundation for training and evaluating the book recommendation system.

Preprocessing Steps:

- **Data Collection**
- **Data Cleaning :** Data is already clean in our case.
- **Exploratory Data Analysis (EDA):** The project starts by using the Pandas library to read and manipulate the dataset from the CSV file.
- **Handling Missing Values:** The **pd.to_numeric** function is used to convert the 'average_rating' column to numeric data type, handling any potential errors with errors='coerce’.
- **Text Feature Creation:** A new column 'book_content' is created by combining 'title' and 'authors' for text-based analysis.
- **Text Vectorization:** The **TF-IDF Vectorizer** from scikit-learn is used to convert the 'book_content' text into a numerical format **(TF-IDF matrix).**

Preprocessing Steps:

- **Feature Extraction:**
  - The Sample Data contains Four Attributes
  - Book ID
    - Title
    - Author
    - Average Rating
  - Input

    Input comprises of Three Attributes
    - Title
    - Author
    - Average Rating

Pre-Processing Steps:

Splitting Data:

- Train-Test Split Ratio : 80% -20%

- Training Data:

  Total Instances = 8902

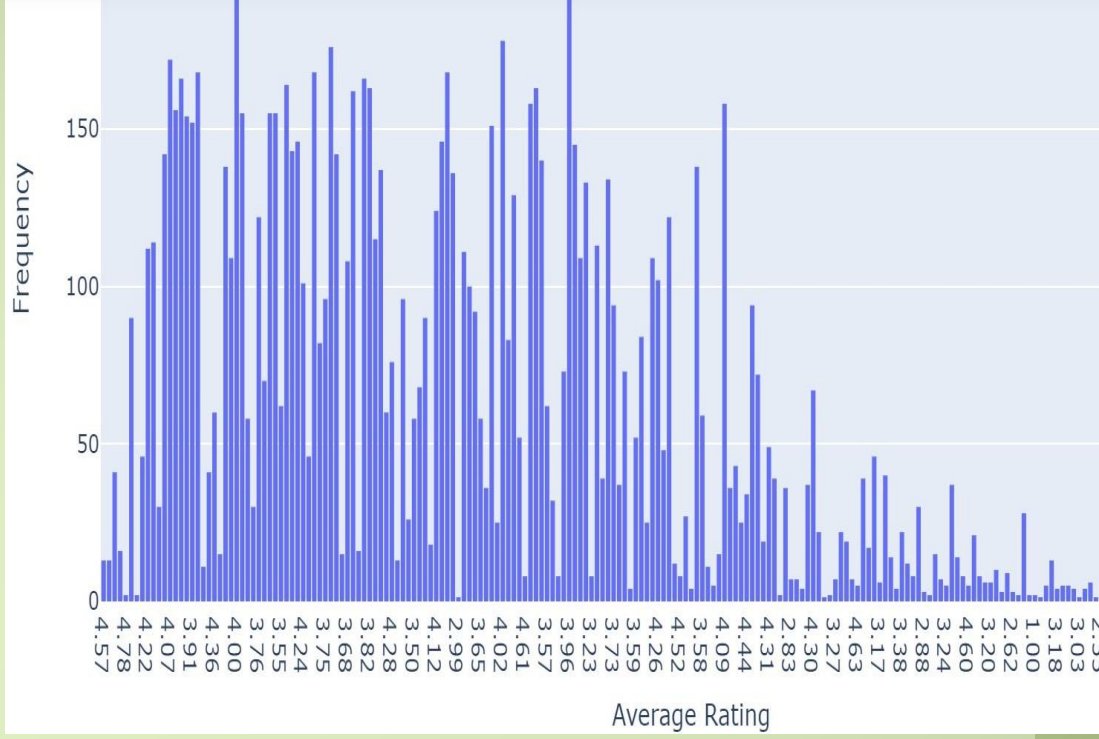
- Test Data:

  Total Instances = 2226

Data Visualization:

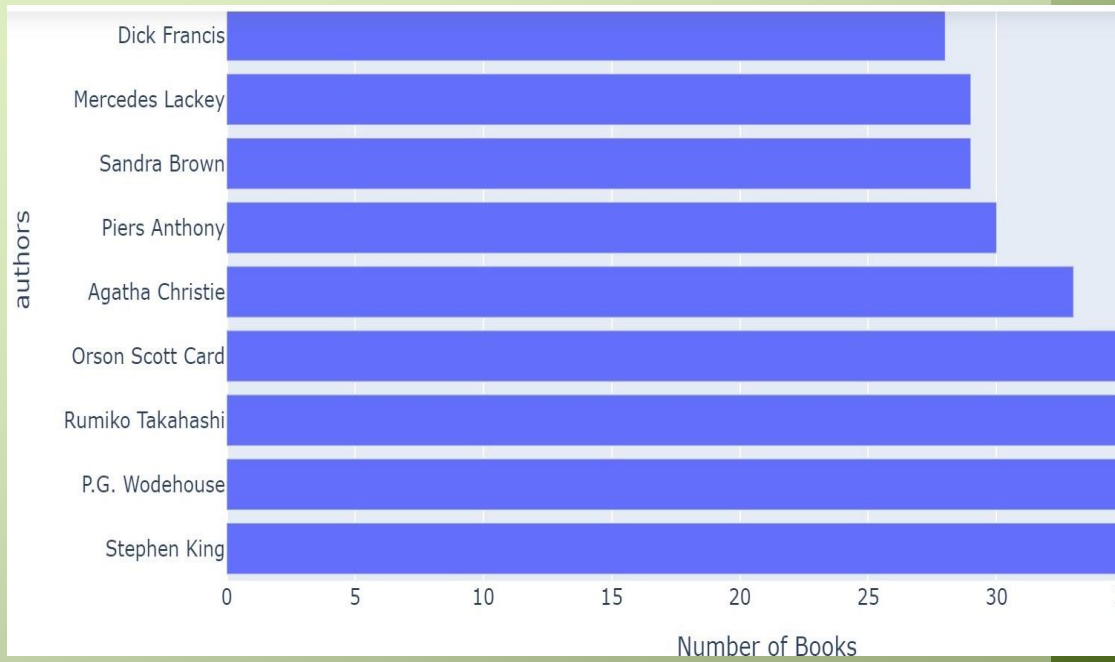
Distribution of Average Ratings:

- The histogram of average ratings provides an overview of how ratings are distributed across the dataset.
- The majority of books seem to have ratings concentrated within a certain range, as indicated by the peaks in the histogram.



Number of Books per Author:

- The bar chart depicting the number of books per author gives an idea of the distribution of authors and their contributions to the dataset.
- It identifies the top 10 authors with the highest number of books, helping us understand which authors are prolific in terms of book production.



Algorithm Used:

- **In this model, we used**

TF-IDF with Cosine Similarity

TF-IDF and Cosine Similarity:

- **Cosine similarity is a metric that measures the cosine of the angle between two vectors.**
- **In the TF-IDF representation, each book is represented as a vector, and the cosine similarity between two books is calculated based on the angles between their vectors.**
- **A higher cosine similarity indicates that two books are more similar in terms of their content.**

Evaluation Measure:

- **Mean Reciprocal Rank is used as an evaluation measure in our model.**
- **Mean Reciprocal Rank (MRR) is a metric commonly used in information retrieval and recommendation systems to evaluate the effectiveness of a ranked list of items.**
- **Formula:**

$$MRR = \frac{1}{N} \sum_{i=1}^{N} RR_i$$

- ***N* is the total number of recommendation scenarios or queries.**
- ***RRi* is the Reciprocal Rank for the *i*-th scenario, indicating how well the system ranked the first relevant book in the list of recommended books for that scenario.**
- ***MRR* is the average of the Reciprocal Ranks across all scenarios, providing a summary measure of the system's overall performance.**

Summary

- The dataset includes book information, and the system suggests books based on textual content similarity.
- The project involves creating a content-based book recommendation system using TF-IDF with cosine similarity.
- Evaluation is performed using Mean Reciprocal Rank, showcasing an approach for personalized book recommendations.
- The output consists of a list of similar books according to the user's preference.

Thank you for your attention!

-Group Members-

Maleeha Asghar
Meerab Jalil
Momina Farrukh
Zainab Jahanzeb
Ayesha Tariq