

INSTITUTE FOR ADVANCED COMPUTING

AND

SOFTWARE DEVELOPMENT

AKURDI, PUNE

# Documentation On

# “BOOK RECOMMENDATION SYSTEM”

PG-DBDA SEPT23

**Submitted By:**

Group No: 16

Roll No: Name:

**239513 Biradar Vikram**

**239522 Kalpesh Patil**

# Dr. Shantanu Pathak Mr. Rohit Puranik

**Project Guide Centre Coordinator**

## ABSTRACT

In our fast-paced and time-constrained world, the relevance of recommendation systems has surged. With individuals perpetually pressed for time, these systems play a crucial role in facilitating informed decision-making without taxing cognitive resources.

The primary goal of recommendation systems is to sift through an array of options and deliver personalized content tailored to individual interests.

Operating on Artificial Intelligence algorithms, these systems analyze user profiles, search histories, and demographic trends to construct customized lists.

Utilizing predictive modeling and heuristics, recommendation systems optimize content suggestions based on user behaviors and the viewing patterns of similar demographics.

This study sheds light on the imperative role recommendation systems play in providing relevant and engaging content amidst the constant hustle of daily life.

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to everyone who has contributed to the completion of our project.

First and foremost, we would like to thank our project guide **Dr. Shantanu Pathak** for their constant guidance and support throughout the project. We extend our sincere thanks to our respected Centre Co- Ordinator, **Mr. Rohit Puranik**, for allowing us to use the facilities available.

We would also like to express our appreciation to the faculty members of our department for their constructive feedback and encouragement. Their insights and suggestions have helped us to refine our ideas and enhance the quality of our work.

Furthermore, we would like to thank our families and friends for their unwavering support and encouragement throughout our academic journey. Their love and support have been a constant source of motivation and inspiration for us.

Thank you all for your valuable contributions to our project.

**Biradar Vikram (239513)**

**Kalpesh Patil (239522)**

**INTRODUCTION**

* **PROBLEM STATEMENT**

Efficiently tailoring book recommendations to individual preferences remains a challenge, necessitating the development of a more adaptive and personalized book recommendation system that seamlessly integrates user behaviors, preferences, and evolving reading tastes.

## 

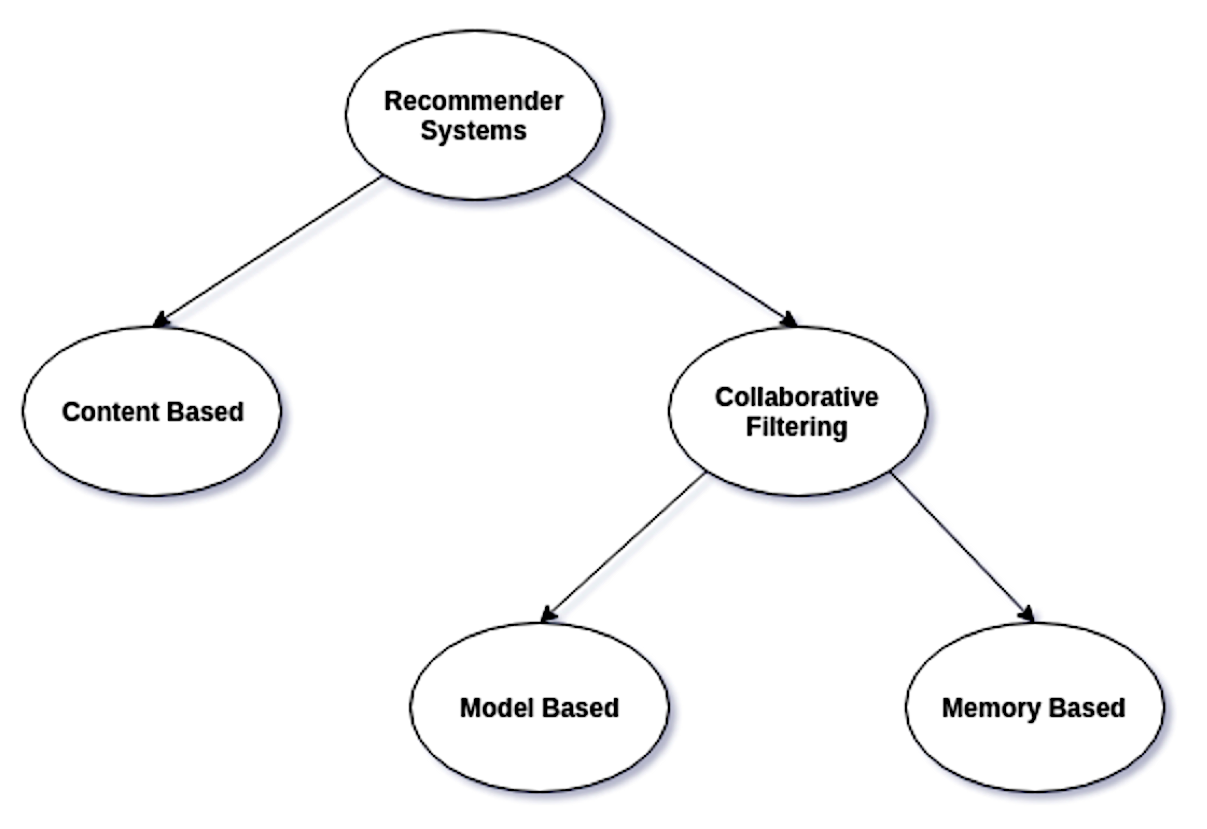
* **FUTURE SCOPE**

The future trajectory of recommendation systems holds immense potential for advancements that can further enhance user experience and adapt to evolving technological landscapes. Key areas for future exploration include:

* Advanced Machine Learning Models:
* Multi-Modal Recommendations:
* Context-Aware Recommendations:
* **AIMS & OBJECTIVES**
* **Aim:**

The aim of the book recommendation system ML project is to develop an intelligent and personalized platform that suggests relevant books to users based on their preferences, past reading history, and similar user behaviors. By leveraging machine learning algorithms and data analysis techniques, the system aims to provide users with tailored recommendations that match their individual tastes and interests, ultimately enhancing their reading experience and satisfaction.

* **Adjectives:**
* Personalized: The recommendation system aims to provide personalized suggestions to users, ensuring that each recommendation aligns with their unique preferences and interests.
* Intelligent: The system utilizes advanced machine learning algorithms to analyze user behavior, identify patterns, and generate intelligent recommendations that adapt and improve over time.
* Efficient: The recommendation system efficiently processes large volumes of data to deliver timely and relevant book suggestions to users, enhancing their overall experience.
* Accurate: By leveraging data-driven approaches and predictive modeling techniques, the system strives to deliver accurate recommendations that resonate with users' reading preferences.



* **RECOMMENDATION SYSTEM:**

One of the largest application areas of ML. Enable tailoring personalized content for users, thereby generating revenue for businesses.

* **CONTENT BASED RECOMMENDATIONS:**

Based on user past likes & dislikes System recommends items similar to items user have liked based on item

Feature space

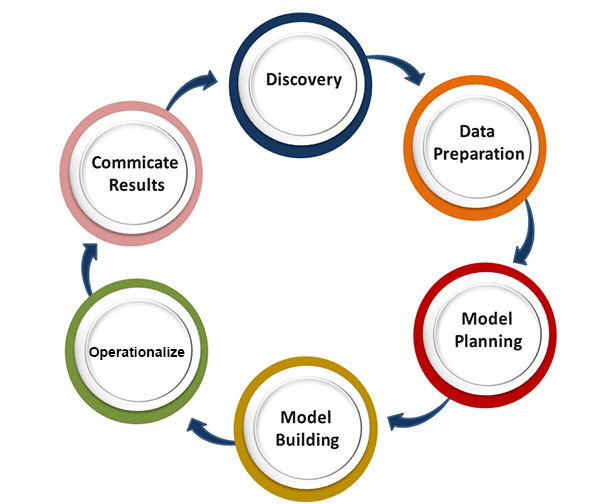
* **COLLABORATIVE BASED RECOMMENDATIONS:**

More robust and widely used. Disregards item & user feature space & solely Based on how different users rate different items.

Overall Description

* **WORKFLOW OF PROJECT:**

The diagram below shows the workflow of this project.

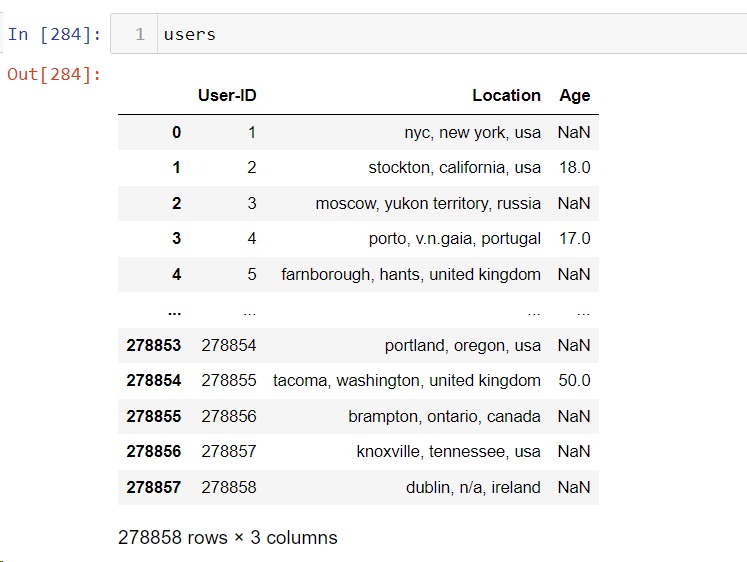


*Figure 1Workflow Diagram.*

## Data Preprocessing and Cleaning:

* DATA CLEANING:

The data can have many irrelevant, missing parts, HTML tags, links. To handle this part, data cleaning is done.

 A screenshot of a computer

Description automatically generated

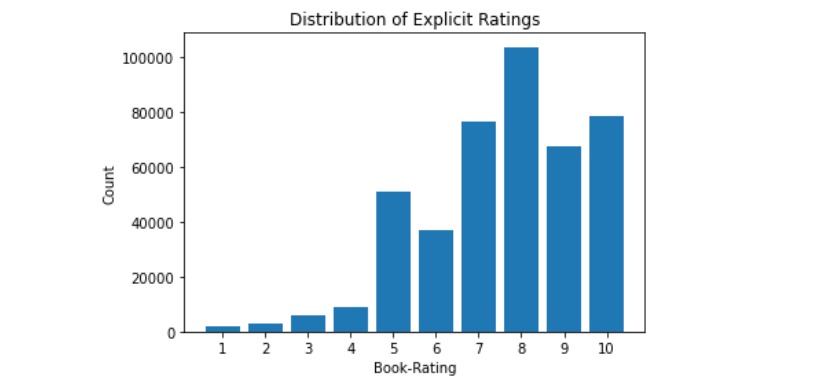
Removing The Null Values Using

numeric\_cols = df.select\_dtypes(include=['float64', 'int64']).columns

df[numeric\_cols] = df[numeric\_cols].fillna(df[numeric\_cols].mean())

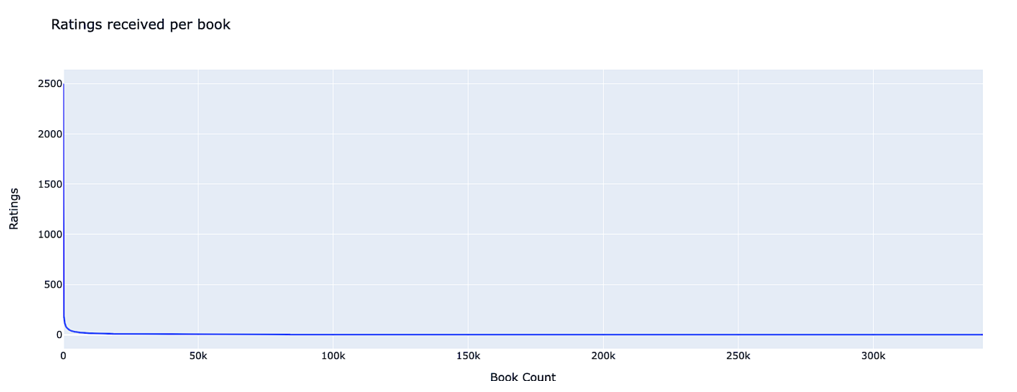
* **EXPLORATORY DATA ANALYSIS:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

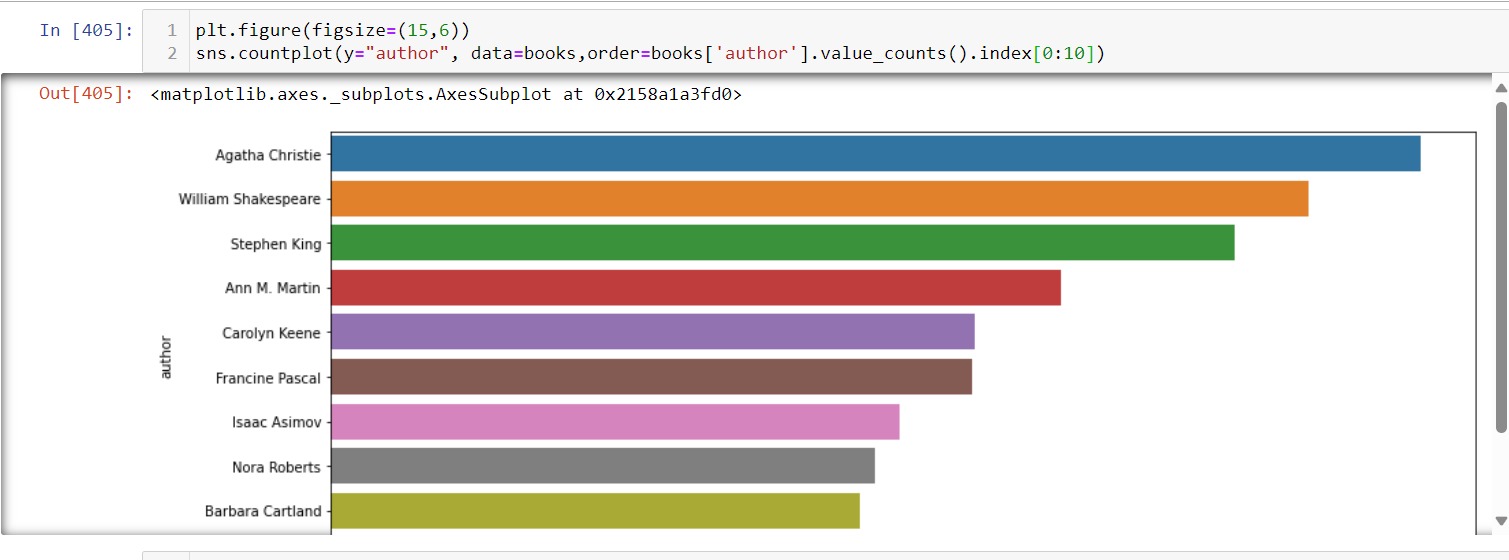


Ratings are of two types, an implicit rating & explicit rating. An implicit rating is based on tracking user interaction with an item such as a user clicking on an item '0'. An explicit rating is when a user explicitly rates an item, i.e., b/w '1-10'

* Majority of ratings are implicit i.e., rating '0'
* Rating of '8' has the highest rating count among explicit ratings '1-10'

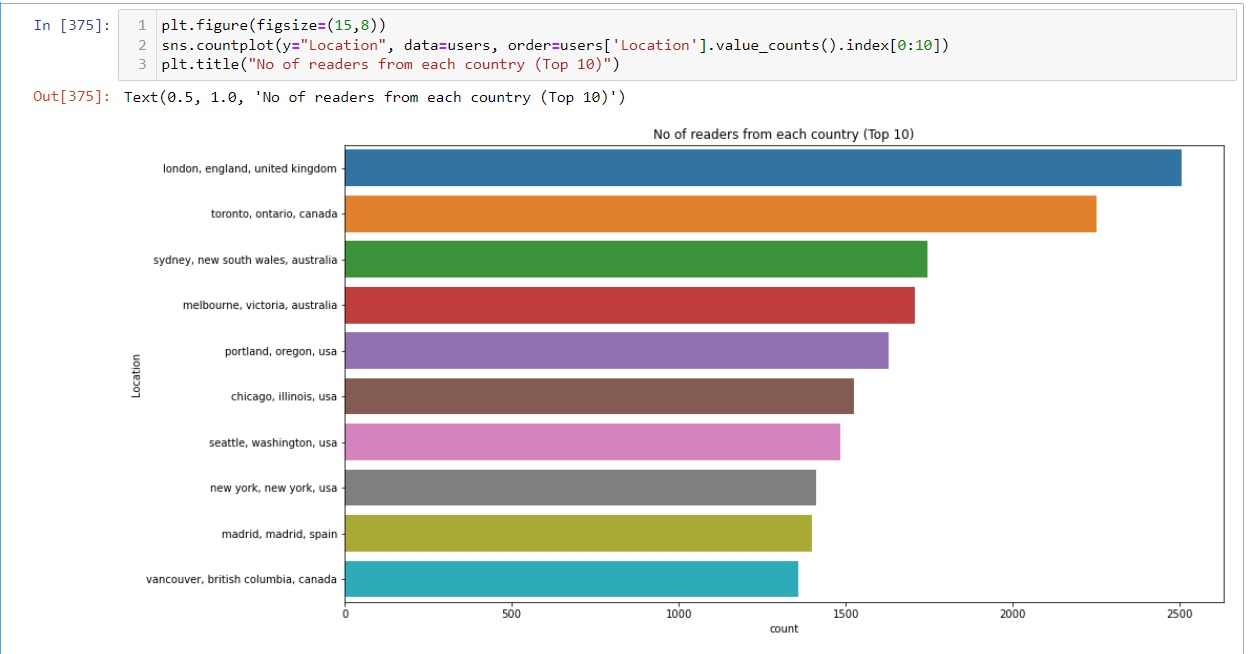


A similar bias is observed. There are a few books that have received many ratings and several books that have received very few ratings. One book has received over 2500 ratings, only ~2100 books (out of 300K+ books) have received more than 50 ratings.



*Figure 3 Bar plot of author who have garnered the most reader attention.*

The authors who have garnered the most reader attention, offering insights into their relative popularity and reader engagement. Understanding these trends can inform strategies for promoting authors' works and tailoring content recommendations to optimize reader satisfaction and engagement within the literary community.



*Figure 3 Bar plot of Readers from each Country (Top 10).*

The visualization provides valuable insights into the geographical distribution of readership, highlighting the countries with the highest reader engagement. Understanding these patterns can inform targeted marketing strategies, content localization efforts, and community-building initiatives to foster a global reading community and enhance reader engagement worldwide.

## Model Building:

* **Train/Test split:**

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into

two parts: first one for training our machine learning model, and second one for testing our model:

* Split the dataset into two pieces: a training set and a testing set.
* Train the model on the training set.
* Test the model on the testing set and evaluate how well our model did.
* **Advantages of train/test split:**
* Model can be trained and tested on different data than the one used for training.
* Response values are known for the test dataset, hence predictions can be evaluated
* Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

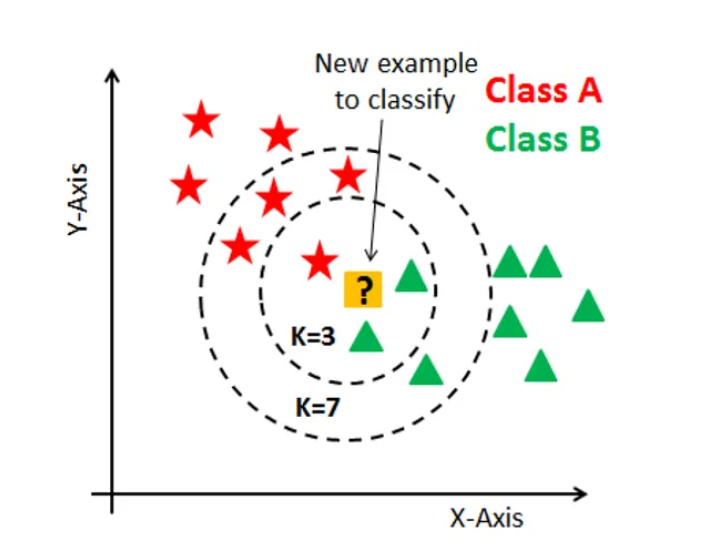
Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.

**We have used the following algorithm to build predictive model.**

* **K nearest neighbor:**

K Nearest Neighbor Algorithm. K nearest neighbor algorithm is very simple. It works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbors.

This algorithm takes into consideration up-to 'K' nearest users (in user based collaborative filtering) or 'K' nearest items (in item based collaborative filtering) for making recommendations. By default, the algorithm is 'user-based', and k is 40 (kmin is 1). This means ratings of 40 nearest users are considered while recommending an an item to a user. Some variants of this algorithm include WithMeans, WithZscore & Baseline wherein the average rating of users, or the normalized Z Scores of ratings or the baseline rating are also considered as the system generates recommendations



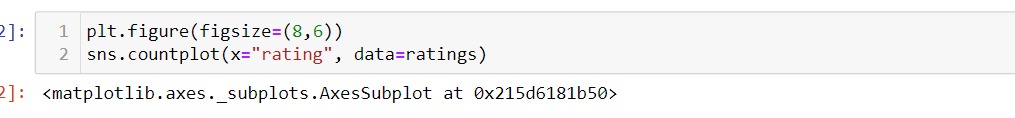
A screenshot of a computer

Description automatically generated

*Figure The accuracy score of KNN model*

## DATA PRE-PROCESSING:

* Handle missing data by imputation, removal, or interpolation techniques.
* Address outliers through trimming, winsorization, or transformation methods.
* Encode categorical variables using techniques such as one-hot encoding or label encoding.
* Scale numerical features to a comparable range using standardization or normalization methods.

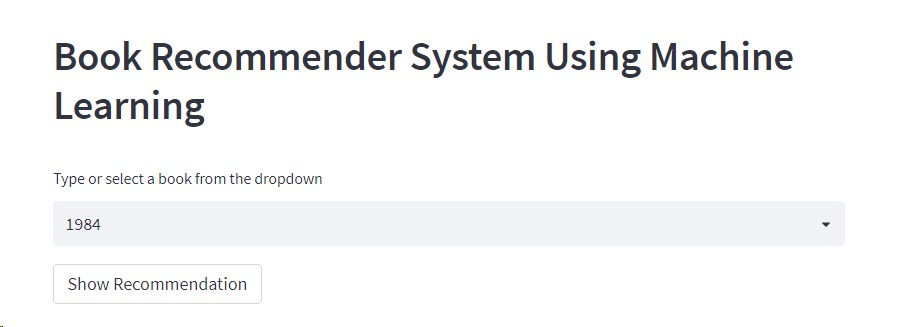


**USER INTERFACE**

After Training the models and finding out the best model for Book Recommended, we can go ahead with building the user interface for our models. This will give us a clean and simple way of accessing our models and make the predictions on our questions.

OUR MACHINE LEARNING PROJECT WAS DEPLOYED USING **STREAMLIT**

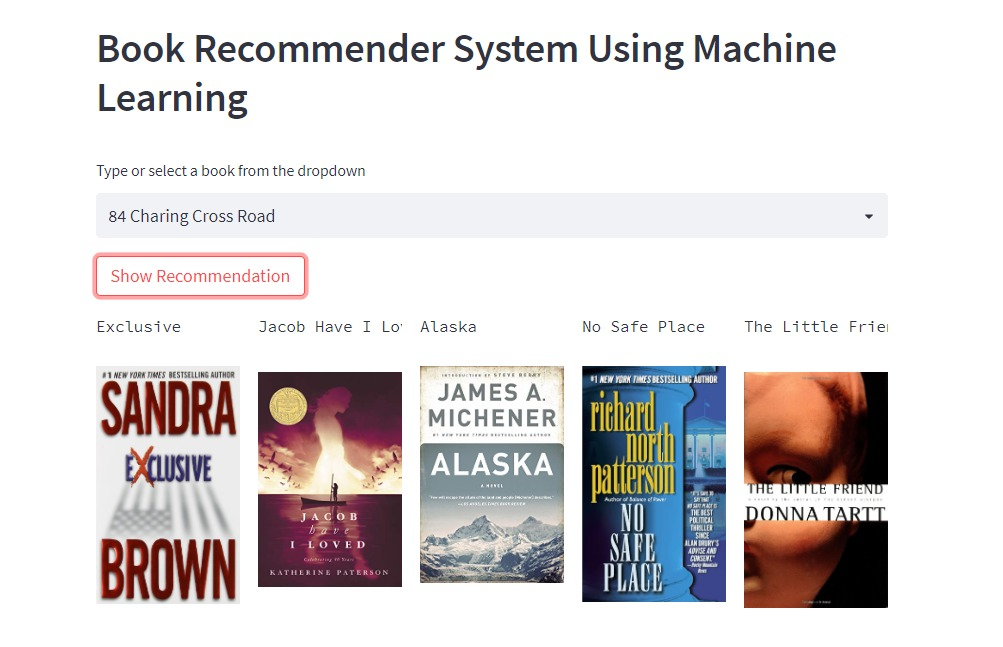
An open-source Python library known for its simplicity and rapid prototyping capabilities. Streamlit allowed us to create an interactive web application seamlessly, showcasing our machine learning models and insights. The user-friendly interface provided end-users with a responsive and intuitive experience, making our project accessible to a wider audience. With Streamlit's integration with popular machine learning libraries and its adaptability, we successfully presented our project outcomes in a centralized and dynamic platform, laying the foundation for future scalability and enhancements.



A screenshot of a computer

Description automatically generated

*Figure 35 This gives us the question’s quality and a comparison of the different models*



* REQUIREMENTS SPECIFICATION
  1. Hardware Requirement:
     + 500 GB hard drive (Minimum requirement)
     + 8 GB RAM (Minimum requirement)
     + PC x64-bit CPU
  2. Software Requirement:
     + Windows/Mac/Linux
     + Python-3.9.1
     + VS Code/Anaconda
     + Python Extension for VS Code
     + Libraries:
       - Numpy 1.18.2
       - Pandas 1.2.1
       - Matplotlib 3.3.3
       - Scikit-learn 0.24.1
       - Streamlit
     + Any Modern Web Browser like Google Chrome
       - To access the web application written in Flask

**CONCLUSION**

* In EDA, the Top-10 most rated books were essentially novels. Books like The little prince and The Secret Life of Bees were very well perceived.
* If we look at the ratings distribution, most of the books have high ratings with maximum books being rated 8. Ratings below 5 are few in number.
* Author with the most books was Agatha Christie, William Shakespeare and Stephen King.
* For modelling, it was observed that for model user based collaborative filtering technique worked .
* A recommendation system helps an organization to create loyal customers. The recommendation system today are very powerful that they can handle the new customer too who has visited the site for the first time. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profitto the company.

**FUTURE SCOPE**

* Given more information regarding the books dataset, namely features like Genre, Description etc, we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.
* We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

References

* https://www.analyticsvidhya.com/blog/2021/06/build-book-recommendation-system-unsupervised-learning-project/
* https://www.w3schools.com
* https://gemini.google.com/app/7c96a52edc4efd92
* https://github.com/entbappy/Books-Recommender-System-Using-Machine-Learning/blob/master/data/BX-Books.csv
* https://www.kaggle.com/search?q=book+recommendation+systemhttps://www.kaggle.com/search?q=book+recommendation+system

**Table of Contents**

[ABSTRACT](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410545) 2

[ACKNOWLEDGEMENT](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410546) 3

[1. INTRODUCTION 4](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410547)

[1.1 Problem Statement 4](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410548)

[1.2 Product Scope 4](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410549)

[1.3 Aim & Objectives 5](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410550)

[2. OVERALL DESCRIPTION 7](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410551)

[2.1 Workflow of Project](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410552) 7

[2.2 Data Description](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410553) 8

[2.3 Importing Dataset.](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410554) 9

[2.4 Project Initiation:](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410555) 10

[2.5 Data Acquisition:](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410556) 11

[2.6 Data Exploration and Understanding:](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410557) 13

[2.7 Data Pre-processing: 15](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410558)

[2.8 Feature Engineering: 16](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410559)

[2.9 Model Development: 17](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410560)

[3. REQUIREMENTS SPECIFICATION 19](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410566)

[4.1 Hardware Requirement: 19](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410567)

[4.2 Software Requirement: 19](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410568)

[4. CONCLUSION 20](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410565)

[5. FUTURE SCOPE 21](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410565)

[6. REFERENCES 22](file:///C:\Users\vikra\AppData\Local\Microsoft\Windows\INetCache\IE\GARLOEZ8\Report%5b2%5d.docx#_Toc159410565)