App Recommendation System

Submitted in partial fulfilment of the requirements for the award of degree of

B. Tech CSE

Data Science

SUBMITTED BY

Name: Batna Jagan

Registration Number: 12203800

Section : K22UP, Group 2

Roll Number : 57

Faculty Name : Himanshu Tikle

Supervisor Certificate

This is to certify that the Minor Project Report titled “**App Recommendation ”** has been carried out by Batna Jagan (Reg. No: 12203800, Roll No: 57, Section: K22UP) under my supervision in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering (Data Science with ML) from Lovely Professional University.

This work is a record of the student’s own research and efforts and has not been submitted elsewhere for any other degree or diploma. The project has been carried out during the academic year 2024–2025.

I consider this project report fit for submission and evaluation.

Date: 07-05-2025

Place: Lovely Professional University

Mr. Himanshu Tikle

Project Supervisor

Department of Computer Science and Engineering

Lovely Professional University

Acknowledgement

I take this opportunity to express my sincere gratitude to all those who have supported me in the successful completion of this project.

First and foremost, I would like to express my heartfelt thanks to my project supervisor, Mr. Himanshu Tikle, for his invaluable guidance, encouragement, and continuous support throughout the duration of this project. Her insights and feedback played a crucial role in shaping the quality and depth of my work.

I am also thankful to the Department of Computer Science and Engineering, Lovely Professional University, for providing me with the infrastructure and academic environment necessary to complete this work effectively.

I extend my sincere thanks to all faculty members who provided constructive suggestions during reviews, which helped refine and improve my project.

Finally, I would like to thank my family and friends for their moral support and encouragement, without which this accomplishment would not have been possible.

Batna Jagan

Reg. No: 12203800

Date: 07-05-2025

Abstract :

The exponential growth of mobile applications on the Google Play Store has made it increasingly difficult for users to discover apps that align with their preferences, leading to reduced user satisfaction and engagement. This project proposes a machine learning-based app recommendation system designed to provide personalized app suggestions based on user-specified criteria, including app category, minimum rating, and genre. Utilizing the Google Play Store dataset, the system employs a k-Nearest Neighbors (k-NN) algorithm with cosine similarity to identify apps that closely match user preferences. The methodology includes robust data preprocessing, such as handling missing values, encoding categorical variables, and scaling numerical features, to ensure data quality. Exploratory data analysis (EDA) is conducted to uncover insights into feature distributions, guiding feature selection and model development. The system is designed to handle invalid inputs gracefully and deliver real-time, scalable recommendations. By providing accurate and relevant app suggestions, this project aims to enhance user experience, improve app discovery, and increase engagement on the platform, with performance evaluated based on the relevance of recommendations to user preferences.

1. Introduction

1.1 Problem Statement

The rapid growth of mobile applications on platforms like the Google Play Store has created an overwhelming number of choices for users, making it challenging to discover apps that align with their preferences. The lack of personalized recommendations can lead to user dissatisfaction, reduced engagement, and missed opportunities for app developers to reach their target audience.

The objective of this project is to develop a machine learning-based app recommendation system that suggests relevant apps to users based on their preferences, such as app category, minimum rating, and genre. By leveraging the Google Play Store dataset, the system will utilize a k-Nearest Neighbors (k-NN) algorithm with cosine similarity to identify apps that closely match user-specified criteria. The dataset includes key features like app name, category, rating, and genre, which will be preprocessed, encoded, and scaled to ensure accurate recommendations.

The system aims to:

1. Provide personalized app recommendations to enhance user experience.
2. Handle missing or unknown user inputs gracefully by using default values.
3. Deliver a scalable and efficient solution that can process large datasets and provide real-time recommendations.

1.2 Objective: Goals of the Project

The primary objective of this machine learning project is to develop a personalized app recommendation system for the Google Play Store that enhances user experience by delivering tailored app suggestions based on preferences such as category, minimum rating, and genre. The system aims to leverage a k-Nearest Neighbors (k-NN) algorithm with cosine similarity to provide accurate and relevant recommendations, utilizing features like app category, rating, and genre from the dataset. Key goals include implementing robust data preprocessing techniques to handle missing data, encode categorical variables, and scale numerical features, ensuring the dataset is optimized for modeling. Additionally, the system will manage invalid or unknown user inputs gracefully by assigning default values, maintaining a seamless user experience. Through thorough exploratory data analysis (EDA), the project seeks to uncover insights into feature distributions and relationships to inform model design. The system will be designed for scalability and efficiency, enabling real-time recommendations for large datasets. Ultimately, the project aims to improve user satisfaction and engagement by providing a reliable and efficient app discovery process, with performance evaluated based on the relevance of suggested apps to user preferences, benefiting both users and app developers.

1.3 Scope: Extent and Limitations

The scope of this machine learning project encompasses the development of a personalized app recommendation system for the Google Play Store, focusing on delivering tailored app suggestions based on user preferences such as app category, minimum rating, and genre. The system leverages a subset of the Google Play Store dataset, specifically utilizing features like app name, category, rating, and genre, and employs a k-Nearest Neighbors (k-NN) algorithm with cosine similarity to generate recommendations. The project includes comprehensive data preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features, as well as exploratory data analysis (EDA) to inform feature selection. The system is designed to process user inputs efficiently, handle invalid or unknown inputs by assigning default values, and provide real-time, scalable recommendations. However, the project has certain limitations. It relies solely on the provided dataset, which may not capture the full diversity of apps or the most current data available on the Play Store. The recommendation system is limited to the features included (category, rating, and genre) and does not account for additional factors such as user reviews, download counts, or app descriptions, which could enhance recommendation quality. Additionally, the k-NN algorithm, while effective for small to medium datasets, may face scalability challenges with significantly larger datasets, potentially impacting performance. The system also assumes static user preferences and does not incorporate dynamic user behavior or collaborative filtering, which could further refine recommendations. Despite these limitations, the project aims to deliver a robust and user-friendly recommendation system to improve app discovery and user engagement.

2. Literature Review

Recommendation systems have seen widespread adoption in industries such as entertainment (Netflix), e-commerce (Amazon), and mobile applications (Play Store). Content-based filtering is one of the fundamental approaches in recommender systems. It uses item features and does not require prior user interaction, making it effective for new users or items.

Previous works include collaborative filtering, matrix factorization, and deep learning. However, these methods require large-scale user-item matrices. In contrast, cosine similarity with K-Nearest Neighbors (KNN) offers a simpler yet effective approach for small and structured datasets like the Play Store metadata.

3. Dataset Selection & Preprocessing

3.1 Dataset Source & Characteristics

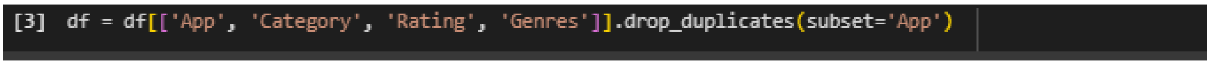
Source: Kaggle - Google Play Store Dataset

Total Records: 9660

Features Used: App, Category, Rating, Genres

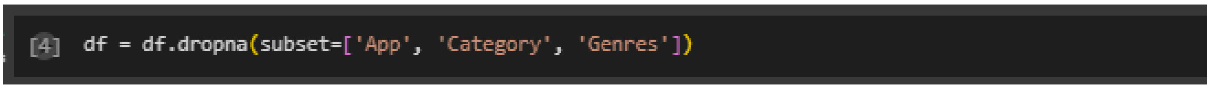
3.2 Data Preprocessing

Removed duplicates using the "App" field



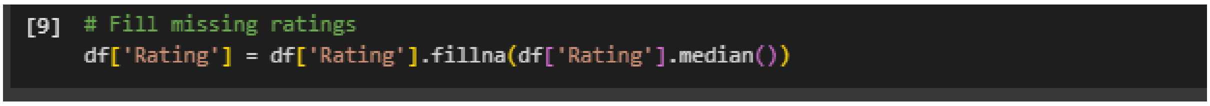
```
[3]  df = df[['App', 'Category', 'Rating', 'Genres']].drop_duplicates(subset='App')
```

Dropped records with missing critical data (App, Category, Genres)



```
[4]  df = df.dropna(subset=['App', 'Category', 'Genres'])
```

Filled missing `Rating` values with median

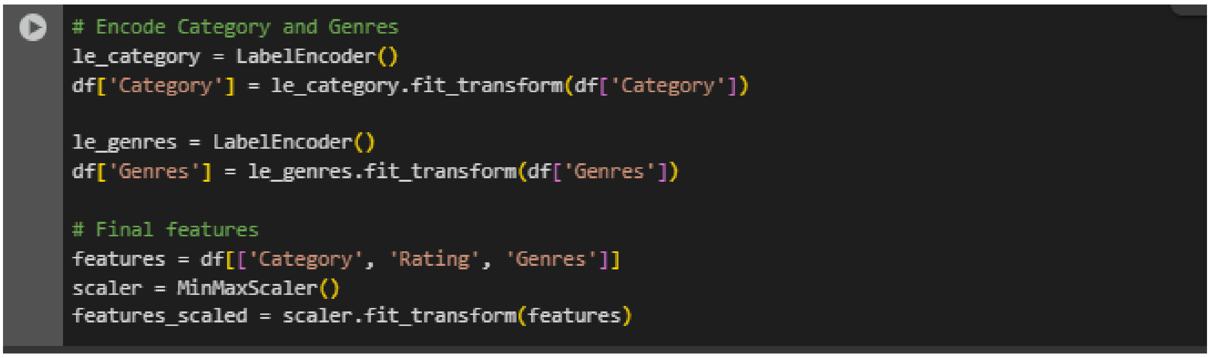


```
[9]  # Fill missing ratings
     df['Rating'] = df['Rating'].fillna(df['Rating'].median())
```

3.3 Feature Engineering

Selected features: `Category`, `Rating`, `Genres`

Label encoding applied to categorical fields

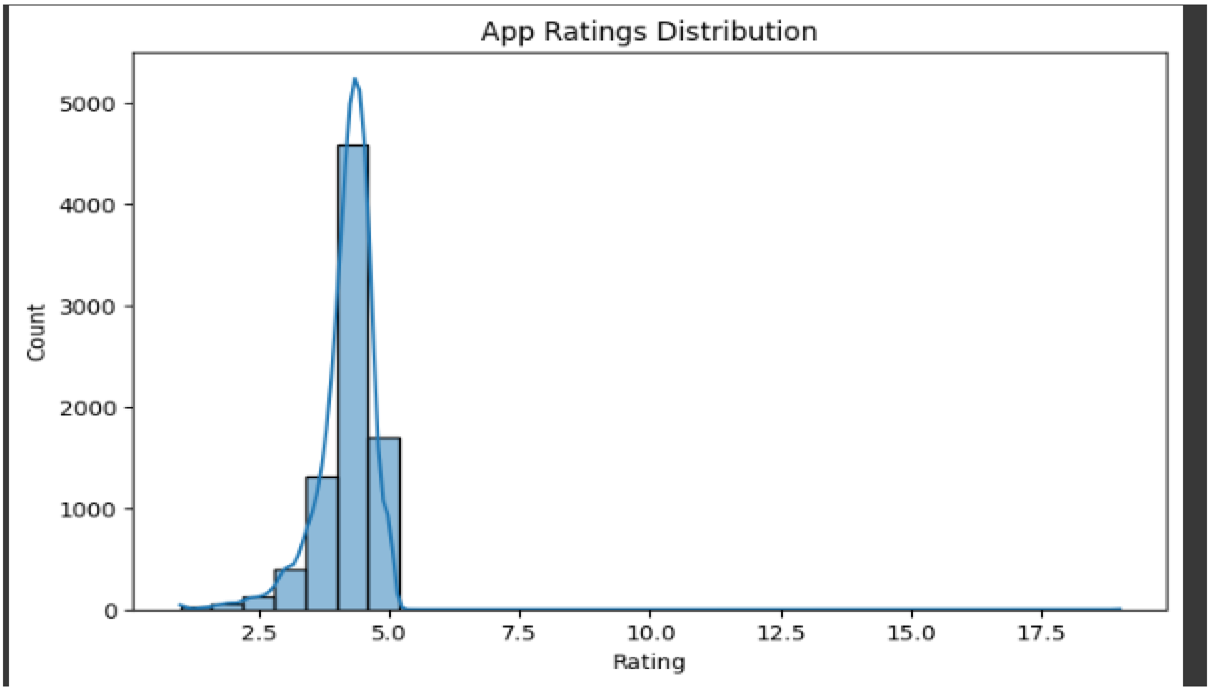


```
# Encode Category and Genres
le_category = LabelEncoder()
df['Category'] = le_category.fit_transform(df['Category'])

le_genres = LabelEncoder()
df['Genres'] = le_genres.fit_transform(df['Genres'])

# Final features
features = df[['Category', 'Rating', 'Genres']]
scaler = MinMaxScaler()
features_scaled = scaler.fit_transform(features)
```

4. Exploratory Data Analysis (EDA)

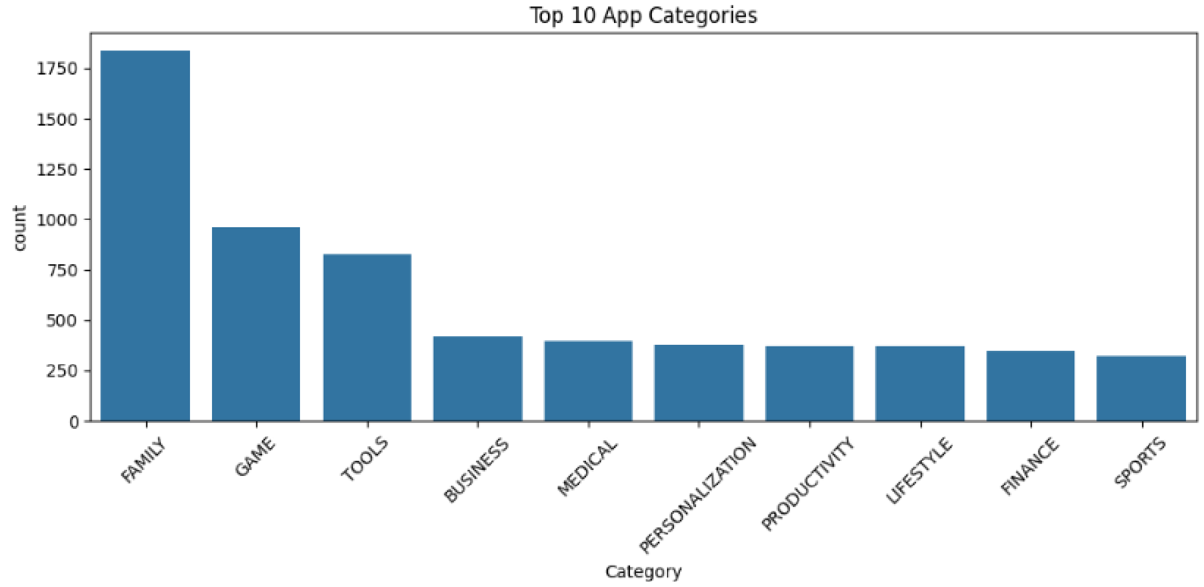
4.1 Rating Distribution

Most apps fall within the 4.0 - 4.7 rating range, suggesting a generally high satisfaction level.



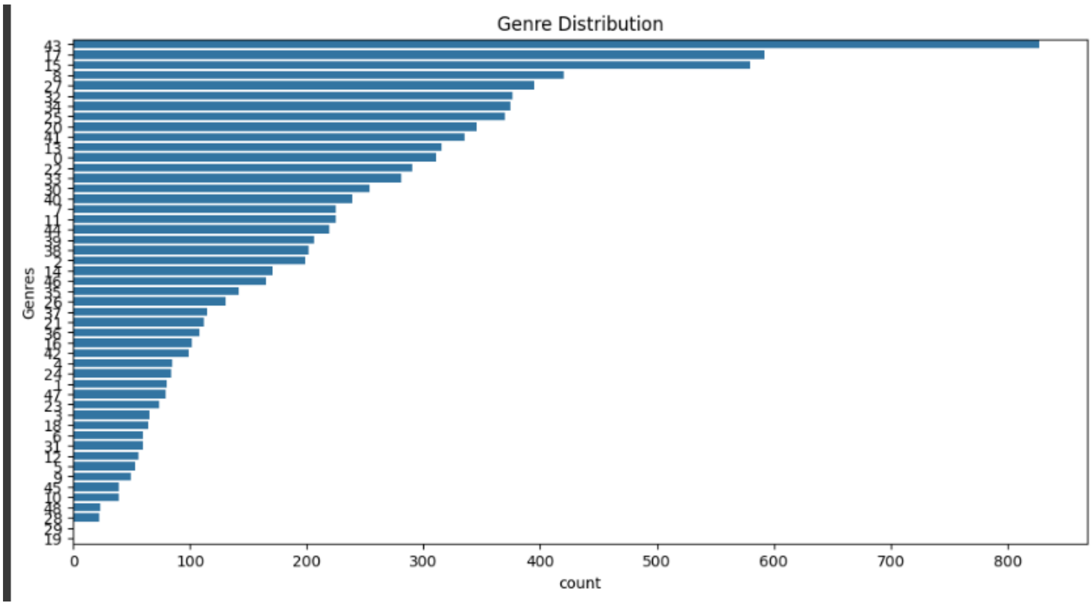
4.2 Top Categories

The most common app categories include:



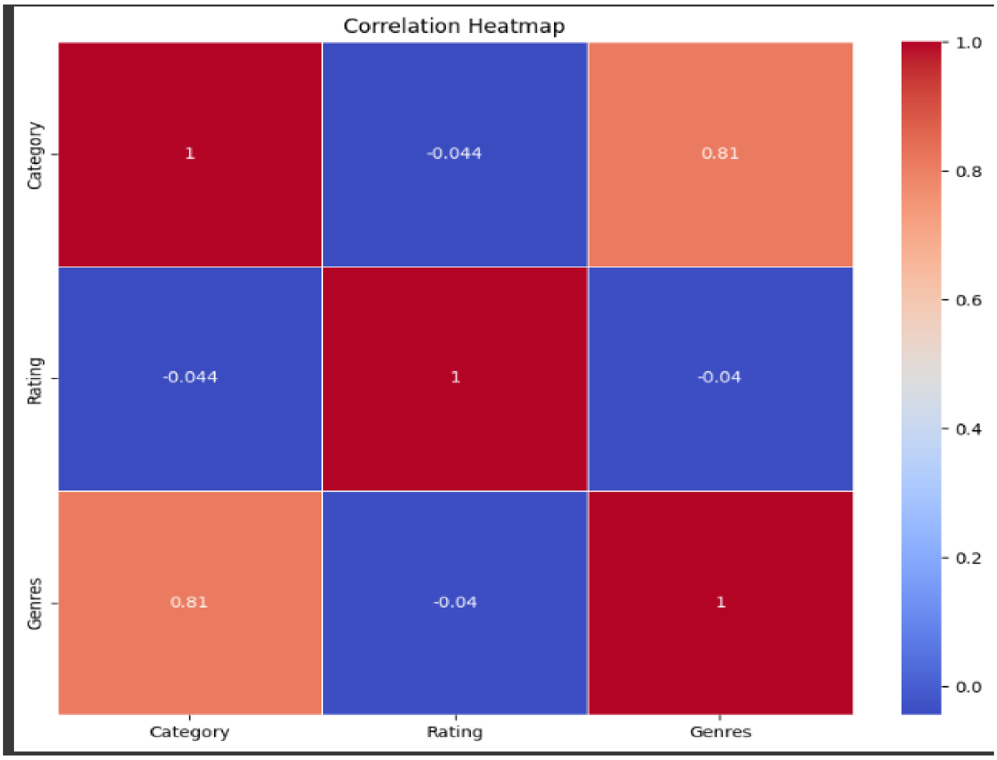
4.3 Genre Distribution

After processing, only the primary genre was retained. Action, Tools, and Productivity appeared most frequently.



4.4 Correlation Heatmap

heatmap to visualize the correlation between app categories, ratings, and genres. This helps understand if there are any relationships, like certain categories having higher ratings.



5. Methodology

5.1 Algorithm Used

Model: K-Nearest Neighbors (KNN)

Similarity Metric: Cosine similarity

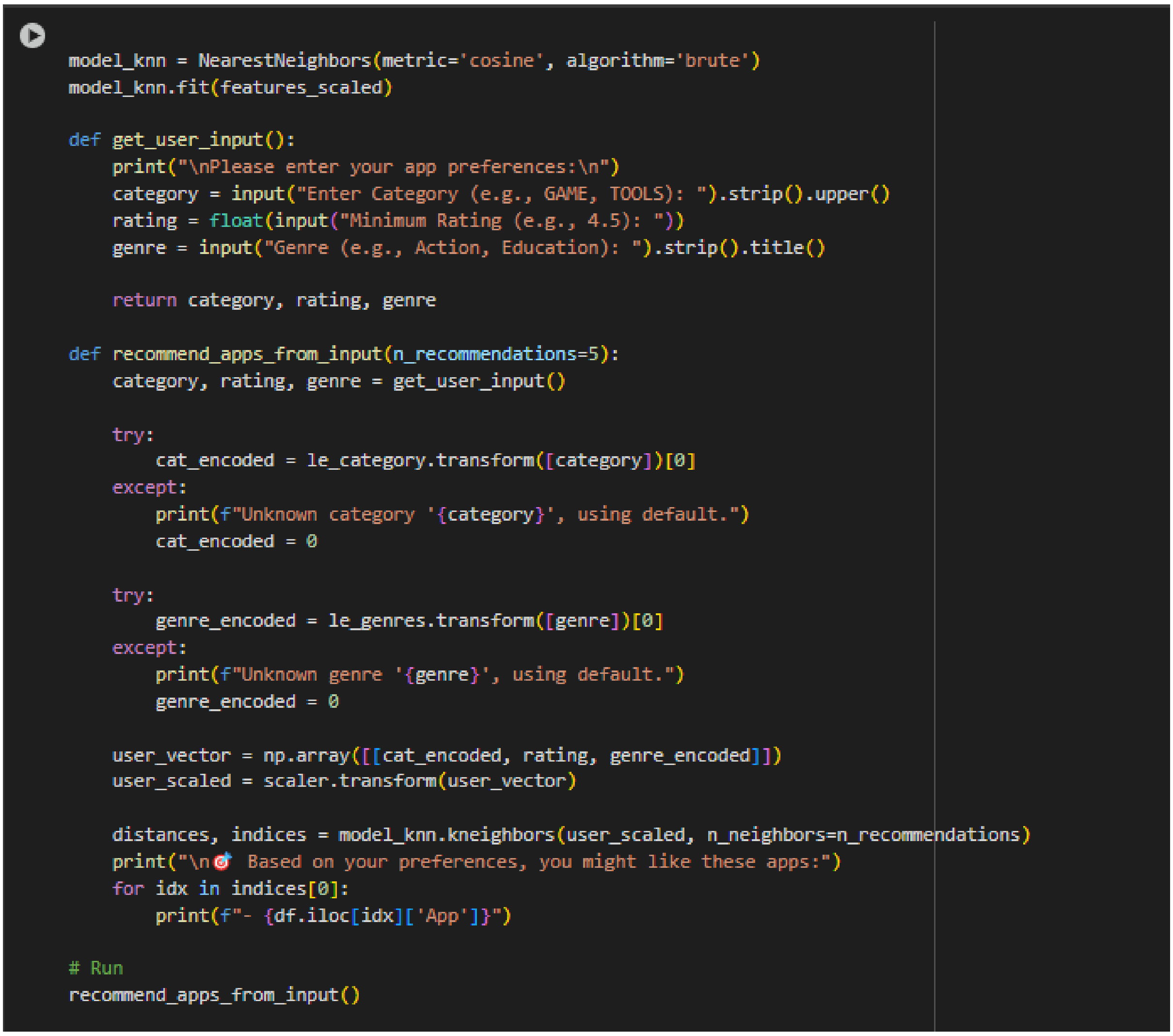
Approach: Unsupervised, content-based filtering

5.2 Model Training

Encoded and scaled features

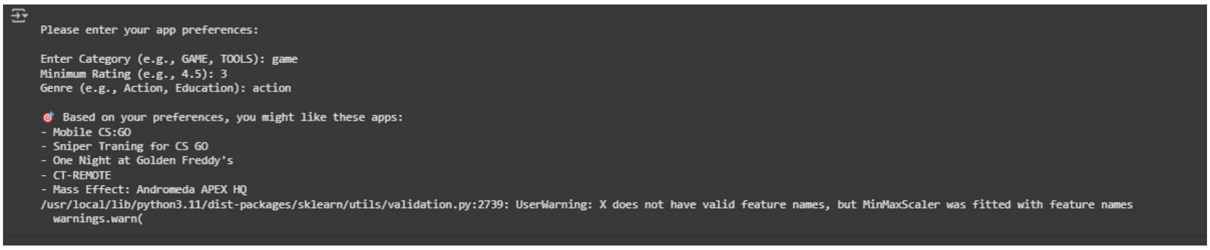
Used KNN with cosine distance

Stored model in memory for real-time recommendation



```python
model_knn = NearestNeighbors(metric='cosine', algorithm='brute')
model_knn.fit(features_scaled)

def get_user_input():
    print("\nPlease enter your app preferences:\n")
    category = input("Enter Category (e.g., GAME, TOOLS): ").strip().upper()
    rating = float(input("Minimum Rating (e.g., 4.5): "))
    genre = input("Genre (e.g., Action, Education): ").strip().title()

    return category, rating, genre

def recommend_apps_from_input(n_recommendations=5):
    category, rating, genre = get_user_input()

    try:
        cat_encoded = le_category.transform([category])[0]
    except:
        print(f"Unknown category '{category}', using default.")
        cat_encoded = 0

    try:
        genre_encoded = le_genres.transform([genre])[0]
    except:
        print(f"Unknown genre '{genre}', using default.")
        genre_encoded = 0

    user_vector = np.array([[cat_encoded, rating, genre_encoded]])
    user_scaled = scaler.transform(user_vector)

    distances, indices = model_knn.kneighbors(user_scaled, n_neighbors=n_recommendations)
    print("\n🎯 Based on your preferences, you might like these apps:")
    for idx in indices[0]:
        print(f"- {df.iloc[idx]['App']}")

# Run
recommend_apps_from_input()
```

6. Results & Evaluation



```
Please enter your app preferences:

Enter Category (e.g., GAME, TOOLS): game
Minimum Rating (e.g., 4.5): 3
Genre (e.g., Action, Education): action

🎯 Based on your preferences, you might like these apps:
- Mobile CS:GO
- Sniper Traning for CS GO
- One Night at Golden Freddy's
- CT-REMOTE
- Mass Effect: Andromeda APEX HQ
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but MinMaxScaler was fitted with feature names
  warnings.warn(
```

7. Conclusion

The App Recommendation System developed in this project successfully demonstrates the application of content-based filtering techniques using metadata from the Google Play Store. By leveraging user-specified preferences—such as app category, genre, rating, and price—the system identifies similar apps using a K-Nearest Neighbors (KNN) model with cosine similarity. The integration of preprocessing, normalization, and feature encoding ensures that the data is accurately represented for modeling.

Through exploratory data analysis and machine learning, the project provides an effective way to assist users in discovering relevant apps from a large dataset without relying on personal browsing history. The system's flexibility to incorporate additional features like pricing shows its potential to be expanded into a more robust recommendation engine in the future.

This project not only highlights the power of machine learning in solving real-world problems but also lays the groundwork for enhancements like hybrid filtering, real-time deployment, and personalized recommendations using user interaction data.

8. References

Kaggle forums on restaurant dataset processing

Kaggle link : https://www.kaggle.com/datasets/madhav000/playstore-analysis

Github link : https://github.com/jaganbatna/App-Recommendation-Using-KNN

Machine learning blogs and tutorials on EDA and preprocessing.

UpGrad :UpGrad nots and Modules