**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. Amir khan   Email: [ak41552010@gmail.com](about:blank)   * Data preprocessing  1. Dropping duplicates 2. Handling null and missing values 3. Handling Outliers with specific range value 4. Handling null values of all three dataframes  * EDA * Univariate Analysis  1. Top 10 book ratings 2. Count Plot of top 10 book titles 3. Book author top 10 4. Top publishers  * Bivariate Analysis  1. Group by columns to get average ratings and plot them  * Feature encoding/ Feature Scaling  1. Log transformation to Smooth Distributions  * Splitting Appropriate dependent and independent features * Train test split on dependent and independent features * Implemented Collaborative Filtering * Collaborative Filtering using KNN * Collaborative Filtering using SVD * Recall used as evaluation matrix  1. Saurabh daund   Email: [sudaund@mitaoe.ac.in](mailto:sudaund@mitaoe.ac.in)   * Data Preprocessing  1. Dropping duplicates 2. Handling Null/nan values of all three datasets 3. Handling Outliers with IQR  * EDA * Univariate Analysis  1. Top 10 publishers according to count plot 2. Top implicit book ratings 3. Top explicit book ratings  * Bivariate Analysis  1. To group by on columns to get average ratings 2. Combined group ratings using grou pby method  * Feature encoding/ Feature Scaling  1. Smoothing distributions  * Splitting Appropriate dependent and independent features * Train test split on dependent and independent features * Implemented Collaborative filtering * Created pivot table to make spars matrix * Used KNN to recommend the books only based on us based dataset * Recall used as evaluation matrix  1. Mouleena Jaiswal   Email: [mouli14112000@gmail.com](about:blank)   * Data preprocessing  1. Dropping duplicates 2. Handling Null/nan values of all the 3 data frames 3. Handling Outliers with mean values  * EDA * Univariate Analysis  1. Age distribution using histogram 2. Book ratings using count plot 3. Top 10 books using count plot  * Bivariate Analysis  1. Used group by to plot average ratings  * Feature encoding/ Feature Scaling  1. Normalizing the distributions  * Splitting dependent and independent features * Train test split on dependent and independent features * Collaborative filtering using KNN * Model based collaborative filtering * Top n accuracy matrix  1. Het Kothari   Email: [het.k123@gmail.com](about:blank)   * Data Preprocessing  1. Dropping Duplicates 2. Handling Null values of all datasets 3. Handling nan values of all datasets 4. handling outliers and dropping the outliers  * EDA * Univariate analysis  1. Top 10 book authors 2. Top 10 publishers using count plot 3. Top book titles 4. Top book ratings  * Bivariate analysis  1. Group by to get final book ratings  * Splitting dependent and independent features * Train test split on dependent and independent features * Final Ratings above 30 for the recommendations * Pivot table * Sparse matrix with the help of pivot table * Implemented KNN to recommend top n books  1. Kamya Malhotra   Email: [malhotra.kamya11@gmail.com](about:blank)   * Data Preprocessing  1. Dropping duplicates 2. Handling null values of all datasets 3. Handling nan values of all datasets 4. Handling outliers with mean values  * EDA * Univariate Analysis  1. Top 10 books using count plot 2. Top 10 book titles 3. Year of publication using pie chart 4. Top 10 location  * Bivariate Analysis  1. Group by plot to get average ratings  * Splitting dependent and independent features * Train test split on dependent and independent features * ML Regression algorithms used * Collaborative filtering * Collaborative filtering using KNN * Sparse matrix with the help of pivot table * Recommending top n books with accuracy matrix recall   During the whole project we continuously discussed all the approach and methods we implemented. Our daily google meets helped us clear our concepts and improved our confidence in group discussions. |
| **Please paste the GitHub Repo link.** |
| [**https://github.com/mouleenajaiswal1/Capstoneproject-4**](https://github.com/mouleenajaiswal1/Capstoneproject-4)  **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **We have been provided with quite a big dataset which had three different files.**  **In this dataset we are provided with datasets of Users, Books and ratings.**  **The Book dataset had 271360 entries with 8 columns, in users we had 278578 entries with 3 columns, and in ratings 1149780 entries with 3 columns.**  **The features given were like book id, user id, user name, user age, user location, book ratings etc.**  **We started with cleaning data that is EDA. In which we first looked into null and nan values, we found that, Quite huge null values in books and users especially age column had many entries with null values so we handled them with the help of EDA functions.**  **After that we dropped all the duplicates in our dataset.**  **Then we plotted some of the graphs like count plot pie plot using Seaborn and matplotlib.**  **We started with book title in which we had almost more then 2 lakh entries which were unique, so out of which we plotted top 10 books using count plot, then we plotted book author in which we used count plot to get count of top 10 book authors out of many unique entries, similarly we plotted pie plot of the year of publications and found that 2002 was the year with maximum publications.** |
| **Then we looked into outliers we had huge outliers in age column and also many invalid entries like age greater than 100 and 200, so as per our requirement we kept only those entries which were greater than 5 age group and less than 90 age.**  **Then we started with implementing models for our book recommendation system.**  **As we know when we talk about recommendations we have two types that is content based and collaborative based, so in our project we implemented collaborative based system.**  **First model we implemented was Collaborative Filtering Using KNN that is K nearest Neighbors, As we know KNN operates on the nearest points based on the distance calculated using metrics like Euclidean distance, so with the KNN it will recommend top n books to the users.**  **Then we implemented Collaborative filtering using SVD that is singular value decomposition, And Finally we evaluated model using Top N accuracy matrix, in which we got recall@15 as 51% which is quite good accuracy as our dataset was huge.** |
| **THANK YOU** |