**Project**

**Movielens**

**DESCRIPTION**

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Members of the GroupLens Research Project are involved in many research projects related to the fields of information filtering, collaborative filtering, and recommender systems. The project is led by professors John Riedl and Joseph Konstan. The project began to explore automated collaborative filtering in 1992 but is most well known for its worldwide trial of an automated collaborative filtering system for Usenet news in 1996. Since then the project has expanded its scope to research overall information by filtering solutions, integrating into content-based methods, as well as, improving current collaborative filtering technology.

**Problem Objective :**

Here, we ask you to perform the analysis using the Exploratory Data Analysis technique. You need to find features affecting the ratings of any particular movie and build a model to predict the movie ratings.

**Domain: Entertainment**

**Analysis Tasks to be performed:**

* Import the three datasets
* Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating. (Hint: (i) Merge two tables at a time. (ii) Merge the tables using two primary keys MovieID & UserId)
* Explore the datasets using visual representations (graphs or tables), also include your comments on the following:

1. User Age Distribution

2. User rating of the movie “Toy Story”

3. Top 25 movies by viewership rating

4. Find the ratings for all the movies reviewed by for a particular user of user id = 2696

* **Feature Engineering:**

Use column genres:

1. Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)
2. Create a separate column for each genre category with a one-hot encoding ( 1 and 0) whether or not the movie belongs to that genre.
3. Determine the features affecting the ratings of any particular movie.
4. Develop an appropriate model to predict the movie ratings

ANALYSIS:

After look at the case study, initially import the data sets to a jupyter notebook.After that chosen the required libraries.

In next step by using movies.dat file we built new data set in data frame format by naming movies\_df. Similarly done ratings\_df and users\_df.

By using merge function here combine both **ratings\_df , movies\_df**  data frame fileson using common ‘MovieID’ column data to form new data frame file **Movie\_rating**

Later again using merge function combine **Movie\_rating** and **users\_df** on using common ‘UserID’ to one master data frame name it as **Master\_Data**

After frame the Master\_Data initially find

1.User Age Distribution:here separate the Age column by using values() function,later plot the graph for User Age Distribution function

2.User rating of the movie “Toy Story”: By using re function and create new lower case toystory function plotted a bar graph for “Toy Story”

3.Top 25 movies by viewership rating:with the help of sort() function movie ratings can be sorted here,that sorted can be assigned to new data frame named as **Master\_Data1**. By using **head(25)** to find the top 25 movie rating names.

4. Find the ratings for all the movies reviewed by for a particular user of user id = 2696

Master\_Data[Master\_Data[‘UserID’]==2696] by using ‘**UserID**’ list out the all uniques UserID 2696 data.

* **Feature Engineering:**

1. Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)

By using unique() function separate out the ‘Genres’ column.After that with the help of Feature Engineering perform label encoding.For this data assign to new data frame name it as Master\_Data2

1. Create a separate column for each genre category with a one-hot encoding ( 1 and 0) whether or not the movie belongs to that genre.

Here for Master\_Data2 perform the **One-hot-encoding**