:

Analyzing Google App Store

Author:-Jyoti Kapoor

:

Objective:

Finding key metrics and factors and show the meaningful relationships between attributes present in the dataset.

Benefits:

- Most famous app in the category
- Average app size
- Relation between category and reviews
- Installs in every category.
- Content rating and count.
- Top genre and their number of installs.
- Distribution of rating
- Ratio of paid and free apps in each category
- Sentiment review count in each category

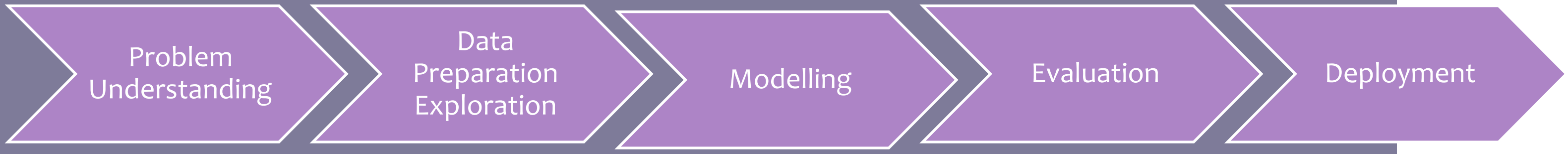
:

➤ Data Source

- playstore_apps.csv - It contains the basic details of the app with the following columns: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content rating, Genres, Last Updated, Current version.
- playstore_reviews.csv - It contains the user reviews for the respective app.
- App: It contains the name of the app.
- Translated Review: It contains the English translation of the review dropped by the user of the app.

:

Architecture

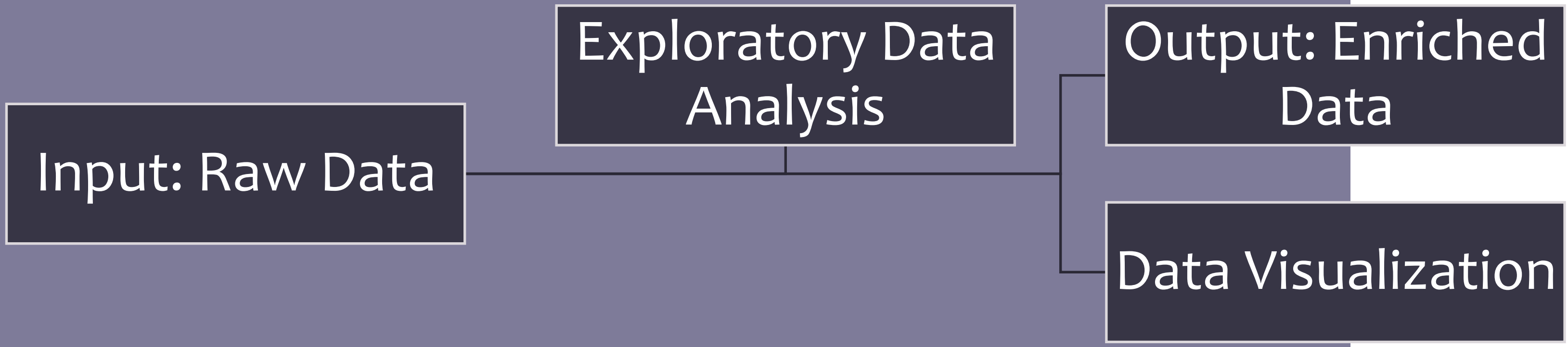

Problem Understanding

Data Preparation Exploration

Modelling

Evaluation

Deployment


Input: Raw Data

Exploratory Data Analysis

Output: Enriched Data

Data Visualization

:

Data Validation and Data Transformation :

- Name Validation - Validation of files name as per the DSA.
- Number of Columns – Validation of number of columns present in the files.
- Name of Columns - The name of the columns is validated and should be the same as given in the schema file.
- Data type of columns - The data type of columns is given in the schema file
- Null values in columns - If any of the columns in a file have all the values as null or missing it is filled or cleaned by python codes mainly with the help of Pandas and Numpy library

:

Data Insertion in Database:-

- Table creation :- Table name “play store apps” is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the “Good Data Folder” are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table.

- Data Preprocessing:-

- Performing EDA to get insight of data like identifying distribution ,
- datatype of each attributes, duplicates handling etc.
- Check for null values in the columns. If present impute the null values.
- Encode the categorical values with numeric values.
- Perform suitable cleansing and transformation operation.

:

- Data Import to Database:-
  - The accumulated data from cleaned dataset is exported in csv encoded in UTF-8 format for running MySQL queries.

- Visualization:-
  - Power BI used for data modelling and visualization of apps and sentiment of users.