# Documentation

## Data Analysis

The dataset consists of two key files:

**1. user\_interaction.csv**

- Contains user engagement data with Pratilipi stories.

**2. metadata.csv**

- Provides details about the stories, including topics and authors.

## Data Exploration

- The dataset was loaded using pandas.

- Converted timestamp columns (updated\_at, published\_at) to datetime format for time-based analysis.

- Checked for missing values and data inconsistencies.

**Outlier Detection & Removal:**

Used Interquartile Range (IQR) method to filter out anomalies in read\_percent.

### Visualizations

- Reading Time Distribution: Histogram to analyze how long users spend reading stories.

- Read Percentage Distribution: Histogram and boxplot (before and after outlier removal) to check how much of a story users typically read.

- Time-Based Analysis: Bar plots showing interactions grouped by hour, month, and year.

- Interactions per User: Histogram to visualize how frequently users engage with stories.

- Multiple Category & Author Check: Identified whether a single story can have multiple topics or authors.

## Training Process

### Feature Extraction

- Engagement Score: read\_percent was scaled between 0 and 1 as implicit feedback.

- User & Item Mappings: Created numerical indices for user\_id and pratilipi\_id for matrix factorization.

- Sparse Interaction Matrix: Constructed a COO matrix using SciPy for efficient computation.

### Model Training

- The model used is **Alternating Least Squares (ALS)**

- Built a confidence-weighted interaction matrix (confidence = 1 + alpha \* engagement, where alpha=40).

- Trained ALS model with 50 latent factors, 0.01 regularization, and 20 iterations.

- Used CSR sparse matrix format for efficient computation.

### Model Evaluation

- Predictions were generated for test user-item pairs.

- Train-Test Split: 75% interactions used for training, 25% for testing (time-based split).

**Evaluation Metrics:**

- Precision@K: Computed for K=5 by checking if recommended stories appeared in test data.

- Used interaction matrices for validation.

## Recommendations & Results

- Generated Top 5 Story Recommendations for Each User:

- Filtered out stories the user had already interacted with.

- Stored recommendations in a dictionary mapping users to their suggested pratilipi\_ids.

**Saved Model & Mappings for Future Use:**

- ALS model (als\_model.pkl)

- User and item mappings (user\_id\_to\_idx.pkl, item\_id\_to\_idx.pkl)

## Conclusion

This project provides insights into user engagement with stories and applies **Alternating Least Squares (ALS)**

for collaborative filtering-based recommendations. The methodology involves thorough data analysis, outlier removal, feature engineering, and machine learning modeling to generate meaningful personalized recommendations.

A blue and white rectangle with black border

AI-generated content may be incorrect.

A blue line with a black border

AI-generated content may be incorrect.

A blue and white line art

AI-generated content may be incorrect.

A blue and white graph

AI-generated content may be incorrect.

A blue and white line

AI-generated content may be incorrect.

A blue bars with black border

AI-generated content may be incorrect.

A blue and white vertical lines

AI-generated content may be incorrect.