Privrank for Social Media

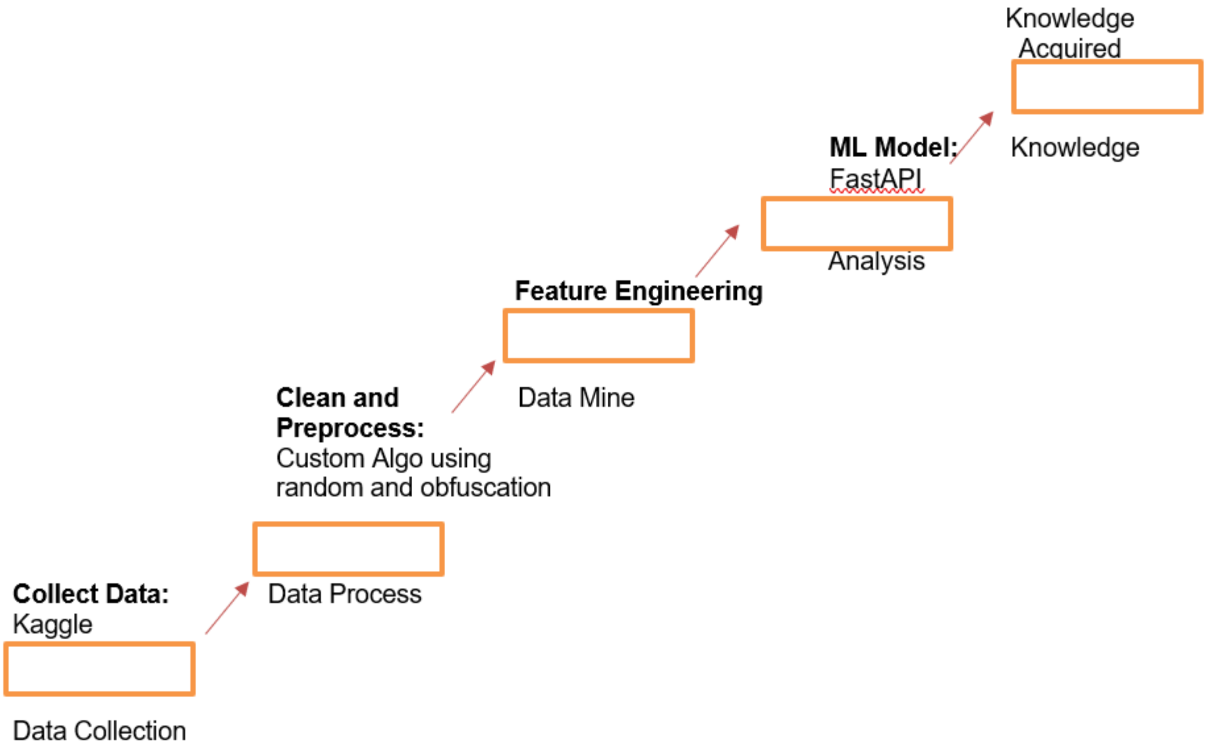
- **Team Name**: Data Freaks
- **Team Members**:

Harika Satti (016260504)

Tanuja Reddy Maligireddy (016715348)

Ravi Teja Reddy Dodda (016693196)

KNOWLEDGE DATA DISCOVERY



Selection:

Selection of raw data from Kaggle Sources:

1) https://www.kaggle.com/code/ashishlepcha/semantic-text-similarity/notebook
2) https://www.kaggle.com/code/gauden/top-2-30dml-dabbler-to-kaggler

Preprocessing:

Preprocessing the data by removing irrelevant data and dropping the null, duplicate values and clearing off the data which is not in proper format or converting the format is taken care of by generating random values.

Cleaning and Preprocess:

Cleaning a preprocessing using custom algorithm and statistics eliminating null values to make the dataset more efficient

Feature Engineering:

Using “Obfuscation” as a tool to extract features that are required to train the model

Interpretation/Evaluation/Analysis:

Using FastAPI and ML Models for evaluation and putting Subjectivity forward would like to push this forward.

Knowledge Representation:

Visualizing under tabular format grouping the product profiles ranking as per the percentage of their risk.

FEATURE ENGINEERING

Features:

Feature Engineering is needed to prepare the input data properly so that our algorithm to determine a favorable candidate can run properly. In the data pre-processing stage of our project, we conducted the following:

1. **Tagging:**
    a. Dividing the dataset for implementation containing 4 things: UserID, gender, age and posts they liked.
2. **Normalizing:**
    a. Firstly, we normalize the data using Standard Scalar. Then used the PCA (Principle Component Analysis) for dimensionality reduction from 3 to 1 and then we created the similarity matrix or we can say correlation matrix.
3. **Recommend person to user:**
    a. Now that we have users that are most similar to us, we need to find the likeability of each post not seen by our current user. For this we use the following likeliness equation using the custom ‘X’ factor algorithm i.e.

$$\sum_{i=1}^{n} x_i \cdot r_i$$

4. **Obfuscation:**
   a. We obfuscated the data 10 - 20% like converting the likes to dislikes and vice versa. So, we could not easily backtrack to the user from their choices. This process was done on the user’s end.