CODRELATE-2025

PROBLEM STATEMENT: Ai-powered content analysis and recommendation.

DATASET:

- **Link:** https://drive.google.com/drive/folders/1uF6Bzle6YF6cth_LaYVKcEb-bPSeX4lt?usp=drive_link
- **Data description**

  Each row in the data is a different article published on Medium. For each article, you have the following features:

  - **title** *[string]*: The title of the article.
  - **text** *[string]*: The text content of the article.
  - **url** *[string]*: The URL associated to the article.
  - **authors** *[list of strings]*: The article authors.
  - **timestamp** *[string]*: The publication datetime of the article.
  - **tags** *[list of strings]*: List of tags associated to the article.

TASK:

- **Tag Modeling -** Identify the tags based on the title and text.
- **Engagement Prediction** – Predict article popularity based on features like title, tags, and reading time.
- **Keyword Extraction** – Identify the most relevant keywords to summarize articles effectively.

INNOVATION:

- **Personalized Article Recommendations** – Suggest articles based on user reading history and preferences.
- **Author Influence Analysis** – Evaluate the impact of authors by analyzing engagement metrics across their articles.
- **Content Optimization Assistant** – Provide recommendations to authors for improving article reach and engagement.

Round 1: Data Analysis & Exploration

Problem Statement

Participants will be provided with a raw dataset containing real-world data. The objective is to **clean, explore, and extract meaningful insights** from the data to support decision-making. Teams must apply **data preprocessing techniques, exploratory data analysis (EDA), and feature selection** to uncover patterns, trends, and relationships within the dataset. The goal is to prepare a well-structured dataset that can be used effectively in the next round for model building.

Workflow / Methodology

1) **Understanding the Dataset**

- Identify the type of data (structured, unstructured, categorical, numerical).
- Analyze column names, data types, and the significance of each feature.
- Check for inconsistencies, missing values, and data imbalances.

2) **Data Cleaning & Preprocessing**

- Handle missing values using imputation techniques (mean, median, mode, interpolation).
- Detect and manage outliers using statistical methods (IQR, Z-score).
- Standardize or normalize numerical data if necessary.
- Encode categorical variables using techniques like one-hot encoding or label encoding.

3) **Exploratory Data Analysis (EDA)**

- Generate summary statistics (mean, median, standard deviation, skewness, kurtosis).
- Identify relationships between features using correlation matrices.
- Analyze class distribution and detect any data imbalances.
- Apply dimensionality reduction techniques (if needed) to improve interpretability.

4) **Data Visualization & Insight Extraction**

- Use **histograms, box plots, scatter plots, and bar charts** to visualize trends.
- Utilize **heatmaps and pair plots** to analyze relationships between variables.
- Identify key factors affecting the problem statement using interactive dashboards.

5) **Feature Engineering**

- Identify the most relevant features that contribute to predictions.
- Perform **feature scaling (Min-Max Scaling, Standardization)** to improve consistency.
- Create new meaningful features through **domain knowledge or data transformations**.

6) **Efficient Use of Data**

- Ensure that no unnecessary features or redundant data points are included.
- Optimize memory usage to handle large datasets efficiently.
- Justify each preprocessing step with logical reasoning.

7) **Final Report Submission**

- Document all findings, insights, and decisions in a structured format.
- Present key visualizations to support the analysis.
- Provide a well-commented Jupyter Notebook or script explaining each step.

Evaluation Criteria

📌 **Problem Understanding & Approach (20%)**

- Clarity in defining the problem and relevance of chosen techniques.

📌 **Data Cleaning & Preprocessing (20%)**

- Efficient handling of missing values, outliers, and inconsistencies.

📌 **Exploratory Data Analysis (EDA) (20%)**

- Quality and depth of statistical analysis and insights.

📌 **Visualization & Interpretation (15%)**

- Use of effective charts and visual storytelling to communicate findings.

📌 **Feature Engineering & Selection (15%)**

- Identification of key features contributing to model performance.

📌 **Report & Code Quality (10%)**

- Well-documented code and structured report with justifications.