Project Abstract

Dataset Overview :
This project explores a publicly available dataset of research papers, offering a comprehensive view of academic literature across diverse fields such as computer science, physics, and mathematics. Each paper in the dataset is meticulously documented with the following attributes: (**1 Million** research papers)

- Paper ID: A unique identifier for tracking each paper.
- Title: A concise summary of the research focus.
- Authors: Names of contributors to the study.
- Venue: Publication source, such as journals or conferences.
- Year: The publication year, useful for tracking trends over time.
- Number of Citations: An indicator of the paper's academic impact.
- References: A list of papers cited, enabling analysis of knowledge networks.
- Abstract: A brief overview of the paper's content, essential for text-based analysis.

This dataset provides rich opportunities for understanding relationships between research papers, clustering them into meaningful groups, and uncovering interdisciplinary trends. Its structured nature ensures feasibility for analysis, while its diverse content opens doors for high-impact discoveries.

Exploratory Data Analysis (EDA) Goals Summary:

1. **Clustering Research Papers**: Group research papers based on abstract similarity to identify thematic trends and gaps.
    - **Procedures**: Preprocess text with tokenization and TF-IDF, apply K-Means and Hierarchical Clustering, and evaluate with silhouette scores.
2. **Visualizing Research Themes**: Create visualizations to highlight relationships within clusters and uncover hidden patterns.
    - **Procedures**: Use PCA and t-SNE for dimensionality reduction, visualize clusters in scatter plots, and analyze trends and outliers.
3. **Extracting Dominant Topics**: Identify key themes and keywords from clusters to highlight research priorities and interdisciplinary opportunities.
    - **Procedures**: Apply LDA and NMF for topic modeling, perform keyword analysis, and track temporal trends to identify emerging fields.
4. **Exploring High-Risk, High-Reward Questions**:
    - **Question 1**: Predict citation count based on abstracts and references using regression models and graph-based analysis.
    - **Question 2**: Identify interdisciplinary overlaps through cluster analysis and metadata visualization.
    - **Question 3**: Investigate factors correlating with high-impact papers using correlation analysis and trend exploration.

Risk and Feasibility:

While certain procedures (like topic modeling) may not always yield clear results, low-risk objectives such as clustering and dimensionality reduction are expected to succeed. High-risk questions offer opportunities for groundbreaking insights, making this project both feasible and impactful.