# Book Recommendation System (MLBD project)

## 1. Name and Team Members

- Project Name: Book Recommendation System
- Teacher Name: Dip Sankar Banerjee
- Team Members: **Gaurav Sharma(G23AI2051), Sneha Sawla (G23AI2113) , Jojo Joseph(G23AI2100)**

---

## 2. Broad Overview

We propose building a Book Recommendation System that leverages machine learning and big data technologies to suggest personalized book selections to users based on their interests and browsing history. The system will ingest and process large-scale book metadata and user interaction data from sources like Amazon and Google Books. Using collaborative filtering (user-item interactions) and content-based filtering (book descriptions, genres), the system will generate recommendations. The project will emphasize scalability using Google Cloud Platform (GCP) for distributed data processing and model training, with a user-friendly UI for seamless interaction.

---

## 3. Objectives

1. Data Ingestion & Processing: Build a scalable pipeline to collect, clean, and preprocess book metadata and user interaction data from Amazon and Google Books.
2. ML Model Development: Train and deploy a hybrid recommendation model (collaborative + content-based filtering) using GCP's ML tools.
3. UI Integration: Develop a simple web interface to display recommendations and collect user feedback for iterative improvements.

---

## 4. Tech Stack and Justification

- GCP Services (BigQuery, Dataflow, Compute Engine): For scalable data storage, ETL pipelines, and model training. GCP's serverless architecture ensures cost-effective big data processing.
- Machine Learning: Scikit-learn/TensorFlow for model development; Vertex AI for deployment.
- UI: Streamlit (Python-based) for lightweight, interactive UI.
- Justification: GCP simplifies big data handling with managed services, while Streamlit ensures rapid UI prototyping. Hybrid ML models balance accuracy and scalability.

---

## 5. Deliverables

1. Functional System: A working book recommendation system with a UI for input and output.
2. ML Model: Trained model with ensemble methods.
3. Documentation: Codebase, API endpoints, and user guide.
4. Presentation: Demo showcasing data flow, model performance, and UI.

---

## 6. Data Sources, Format, and Size

- Primary Data:
    - Amazon Book Reviews (CSV/JSON, ~5GB): Includes book titles, user ratings, and reviews.
    - Google Books API (JSON, ~1GB): Metadata like descriptions, genres, and authors.
- Processing:
    - Data will be stored in GCP BigQuery for SQL-based analysis.
    - Apache Beam/Dataflow will preprocess data (cleaning, feature engineering).
- Scalability: GCP's autoscaling ensures efficient handling of large datasets.