Example Project:

Sample Project Proposal:

[Teams need to prepare the Project Proposal in alignment with their chosen project.]

Project Title: *A Classification Model for Predicting Employee Attrition*

1. **Problem Statement:**

   A **significant challenge** for an organization/company/business is employee attrition, which leads to high recruitment and training costs, loss of institutional knowledge, and decreased team morale. The HR department currently relies on exit interviews and general surveys, which are reactive rather than proactive. We need a data-driven approach to identify the key factors driving attrition and predict which employees are at the highest risk of leaving.

2. **Business Goal:**

   The primary goal is to develop a machine learning model that accurately predicts employee attrition with a significant accuracy. This model will empower the HR department to implement targeted retention strategies for at-risk employees, aiming to reduce the overall attrition rate by 15% within the next fiscal year.

3. **Data Source**

   We will use the "IBM HR Analytics Employee Attrition & Performance" dataset. It's a well-structured, fictional dataset created by IBM data scientists, making it ideal for this type of predictive modeling project.

   - **Source Platform: Kaggle**
   - **Full Citation:**
     - IBM. (n.d.). *IBM HR Analytics Employee Attrition & Performance* [Data set]. Kaggle. Retrieved September 27, 2025, from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

4. **Tools & Technologies**

   - Programming Language: Python
   - Core Libraries:
     - Data Manipulation & Analysis: Pandas, NumPy
     - Machine Learning: Scikit-learn
     - Data Visualization & Storytelling: Matplotlib, Seaborn

- Development Environment: [Our team will primarily use: _____]
  (Recommended options: Google Colab, VS Code, Jupyter Notebook, etc.)
- BI Tools: Tableau or Power BI to create a final dashboard.

5. Project Workflow

The project will follow a structured data science lifecycle, visualized as follows:

Data Acquisition → Data Cleaning & Preprocessing → Exploratory Data Analysis (EDA) → Feature Engineering → Model Building & Training → Model Evaluation → Reporting & Visualization

1. Data Acquisition: Fetch the dataset from Kaggle using its API.
2. Preprocessing: Handle missing values (if any), encode categorical variables, and check for data inconsistencies.
3. EDA: Analyze features to understand their relationship with attrition using statistical summaries and visualizations.
4. Feature Engineering: Create new features from existing ones if necessary to improve model performance.
5. Modeling: Train several classification models (e.g., Logistic Regression, Random Forest, Gradient Boosting).
6. Evaluation: Assess model performance using metrics like Accuracy, Precision, Recall, and F1-Score. Select the best-performing model.
7. Visualization: Create an interactive dashboard in Tableau/PBI to present the key findings and predictions to stakeholders.

6. Data Extraction

- "IBM HR Analytics Employee Attrition & Performance" dataset is acquired directly from the Kaggle repository. To ensure a professional and reproducible workflow, manually downloading the files is not done.
- Instead, we will perform the following steps:
  - **Automate the Process:** We will write a Python script that utilizes the official Kaggle API to connect to the source and download the dataset.
  - **Ensure Reproducibility:** This scripted approach guarantees that the data extraction process is consistent and can be easily re-run by any team member or reviewer.
  - **Prepare for Analysis:** The script will handle the unzipping of the downloaded files and load the data directly into a Pandas DataFrame, making it immediately available for the next phase of our project.
  - Notebook: ***data_extraction.ipynb***

7. Schema/Data Dictionary:
   This data dictionary is created after inspecting the dataset.
   Excel sheet: ***Data_Dictionary_Emp_Attrition.xlsx***