Milestone 2 | Project Proposal and Plan

The Proposal

1. **Problem Statement:** The project aims to understand **employee turnover** by identifying key factors contributing to turnover in organizations. Employee turnover can be costly for businesses, especially when high-performing employees leave. This project will help predict which employees are likely to leave based on various factors like salary, job satisfaction, and time spent at the company.
2. **Proposed Solution:** The solution will involve a **predictive model** using machine learning algorithms like **logistic regression** or **random forests**. These models will classify employees as either likely to stay or leave, using features such as salary, job satisfaction, time spent at the company, and performance evaluations. The analysis will help companies take proactive measures to reduce turnover.
3. **Potential Stakeholders:**
    - **HR departments** in organizations looking to reduce turnover and retain talent.
    - **Business leaders** aiming to improve employee satisfaction.
    - **Recruiting teams** who will benefit from understanding employee behaviors.
    - **Employees** who can experience better work environments through informed (data-driven) decisions by HR.
4. **Potential Obstacles:**
    - **Data Preparation:** The dataset may contain missing or uneven data, which can affect model performance.
    - **Feature Engineering:** Selecting the right features to include in the model will be crucial. Some factors, like job satisfaction or performance evaluations, may not have straightforward correlations with turnover.
    - **Bias in the Data:** The dataset might reflect biases, which could affect predictions if not handled carefully.
5. **Novelty of the Approach:** What sets this approach apart is the focus on understanding employee turnover through both standard job metrics and a specific emphasis on **job satisfaction** and **performance ratings**. By leveraging these insights, the model can offer a more holistic view of the factors leading to turnover.

The Plan

1. **Data Source:**
    - The primary dataset is the Employee Turnover Analytics Dataset.

    - If additional data is required, I may supplement it with publicly available HR datasets or employee satisfaction surveys. Also, if I come across a better dataset that can help me create a more accurate model, I will most likely switch it with the primary dataset I'm planning to use. I don't plan to create a dataset on my own.
2. **Data Organization:**
    - The dataset will be organized using **Pandas** for data manipulation. Features such as **years at the company**, **last evaluation score**, and **average monthly hours** will be used to understand trends related to turnover.
    - Target variable: **Left** (whether an employee left or stayed)
3. **Data Analysis:**
    - I will conduct **exploratory data analysis** to understand relationships between variables and visualize employee turnover patterns.
    - Potential insights include identifying whether **overwork** or **low performance evaluations** lead to higher turnover rates.
4. **Model:**
    - The initial model will likely be **logistic regression**, a simple yet effective approach for binary classification problems.
    - Depending on performance, I may explore other models such as **random forests** or **decision trees** to improve accuracy.
5. **Model Evaluation:**
    - I will evaluate the model using metrics like **accuracy**, **precision**, **recall**, and **F1-score** to ensure that the classification model performs well.
    - Cross-validation and a **confusion matrix** will help fine-tune the model and ensure validity.
6. **Baseline Models/References:**
    - Yes, there are other employee turnover analysis projects and research to compare results. I will use them as a baseline for my model's performance.