# Project and Data Management (PDM) Plan

## 1. Project Overview

### Project Title:

### Startup Success/Failure Prediction using the Crunchbase Dataset

### Summary:

This project aims to predict the success or failure of startups based on Crunchbase data using supervised machine learning models. The goal is to develop an explainable classification model that can be used by stakeholders (investors, accelerators, founders) to estimate the likelihood of startup success using early-stage indicators like funding rounds, investor count, sector, and location.

### Research Question:

Can we predict the success or failure of a startup using structured historical data such as funding details, sector, and location? What are the most influential factors contributing to a startup's success, and how can we visualize and explain these predictions effectively?

### Objectives:

* Clean and preprocess Crunchbase startup data for analysis.
* Perform EDA to understand success distribution by category, region, funding, etc.
* Engineer features like funding velocity, startup age, and sector popularity.
* Build and compare three ML models: Logistic Regression, Random Forest, and XGBoost.
* Use SHAP values for model explainability.
* Deploy the model using Streamlit as a decision-support tool for investors (optional).
* Publish the final project on GitHub and submit a written report for academic review.

### Relevant Literature References:

* Ravuri & Olsen (2017) – Startup Success Prediction Using Decision Trees and Random Forests
* Singh et al. (2020) – Startup Success Prediction: A Machine Learning Approach
* Guzman & Stern (2016) – The State of American Entrepreneurship
* Rizwan et al. (2019) – Survival and Success Prediction of Startups
* Lundberg & Lee (2017) – A Unified Approach to Interpreting Model Predictions (SHAP)

## 

## 2. Project Plan (Weekly Timeline)

| **Week** | **Tasks** |
| --- | --- |
| Week 1 | Define research question, collect and explore dataset, perform literature review |
| Week 2 | Data cleaning, handle missing values, format dates, derive initial features |
| Week 3 | Exploratory Data Analysis (EDA) — funding trends, sector/country success patterns |
| Week 4 | Feature engineering (e.g. funding velocity, time to first funding, sector encoding) |
| Week 5 | Build baseline models (Logistic Regression, Random Forest) |
| Week 6 | Train XGBoost, evaluate all models (Confusion Matrix, ROC-AUC, etc.) |
| Week 7 | Model explainability using SHAP values |
| Week 8 | Optional: Develop and test Streamlit deployment (web app interface) |
| Week 9 | Finalize GitHub repository with code, visuals, and README |
| Week 10 | Prepare and submit PDF report for supervisor |
| Week 11–12 | Revisions, backup all work |

## 3. Data Management Plan

### Dataset Description:

**Name**: Crunchbase Startup Dataset  
**Source**: Kaggle – <https://www.kaggle.com/datasets/PromptCloudHQ/startup-investments-crunchbase>  
**Format**: CSV  
**Size**: ~66,000 records  
**Fields Include**: Company name, country, city, category (industry), status, funding amount, funding rounds, first/last funding dates

### Metadata / Preprocessing Notes:

* **Categorical features**: status, country\_code, category\_list, etc.
* **Numerical features:** funding\_total\_usd, funding\_rounds, etc.
* Dates parsed into founded\_at, first\_funding\_at, last\_funding\_at
* **Engineered features:** age, time to first funding, funding per year

### GitHub Repository:

**Link** : [Github-Link](https://github.com/AshishKhandekar99/Startup-Success-Fail-Dataset-from-Crunchbase)

### Data & Code Security:

Project files are stored in OneDrive and GitHub. Optional backup on Google Drive or external USB.