

Data Collection and Preprocessing Phase

Date	18 June 2025
Team ID	SWTID1749880888
Project Title	Prosperity Prognosticator: Machine Learning for Startup Success Prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This machine learning project aims to predict the success of startups by analyzing factors such as funding history, business category, location, and market trends. The model will support investors, entrepreneurs, and policymakers in making informed, data-driven decisions.
Data Collection Plan	The data will be sourced from publicly available datasets containing startup profiles, funding rounds, and business attributes. The primary dataset will be obtained from Kaggle.
Raw Data Sources Identified	Kaggle – Startup Success Prediction Dataset Link This dataset contains detailed information on startups, including

	company names, founding years, categories, funding rounds, investment types, and final status (acquired, closed, operating, etc.), suitable for training classification models.
--	---

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	The dataset includes details of 9,000+ startups, covering founding year, category, location, funding rounds, and investor types. It labels each startup by status as operating, acquired, or closed to support success prediction modeling.	https://www.kaggle.com/datasets/manishkc06/startup-success-prediction	CSV	210 B	Public