Data Collection and Preprocessing Phase

Date	16 June 2025
Team Lead Name	Jayanth Srinivas Bommisetty
Project Title	Sloan Digital Sky Survey (SDSS) galaxy classification using machine learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

The dataset was sourced from Kaggle's Galaxy Zoo collection, containing images categorized into five distinct galaxy types. These images were organized into folder structures based on their respective classes, making it suitable for supervised image classification tasks.

Data Collection Plan Template

Section	Description
Project Overview	Manual classification of astronomical galaxies is a labor-intensive and subjective process, often requiring expert interpretation. With the increasing volume of astronomical data from large-scale surveys like SDSS, this task becomes impractical at scale.
Data Collection Plan	Searched for suitable datasets folder form Kaggle for Galaxy Classification.
Raw Data Sources Identified	Many datasets form Kaggle, Galaxy-zoo, Galaxy Morphology datasets.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle	Folder named Galaxy_dataset	Kaggle_dataset	CSV	1.92GB	Private