Data Collection and Preprocessing Phase

Date	18 June 2025
Team ID	xxxxxx
Project Title	sloan digital sky survey (sdss) galaxy classification using machine learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification

Enhance your data strategy through the Data Collection Plan and Raw Data Sources Report, promoting careful data curation and maintaining integrity to support accurate analysis and well-informed decisions.

Data Collection Plan Template

Section	Description
Project Overview	To develop a machine learning solution that classifies galaxies by their star formation type and predicts redshift values using photometric features from the Sloan Digital Sky Survey (SDSS) dataset. The solution will be accessible via a simple web interface.
Data Collection Plan	• Explore datasets containing information on galaxy morphology, AGN characteristics, and redshift values. • Give preference to datasets that include a wide range of demographic attributes.
Raw Data Sources Identified	The raw data sources for this project consist of datasets acquired from well-known platforms like Kaggle and UCI, which are widely used for data science competitions and repositories. The sample data provided

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset used from Kaggle does not contain a specific column labeled "morphology" to directly indicate the galaxy's shape. However, it includes a "subclass" feature, which provides insights into the galaxy's formation and indirectly reflects its morphological characteristics.	https://www.kaggle.com/datasets/bryancimo/sdss-galaxy-classification-dr18?resource=download	CSV	42000 KB	Public