Data Collection and Preprocessing Phase

Date	23 September 2024
Team ID	LTVIP2024TMID24997
Project Title	Cereal Analysis Based on Ratings by using Machine Learning Techniques
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset	Mention the issues faced in the selected dataset.	Low/ Moderate/ High	Give the solution for that issue technically.
Cereal Dataset	Negative values in `carbo`, `sugars`, and `potass`	High	Remove or impute negative values with appropriate statistical measures (e.g., median or mean).
Cereal Dataset	Presence of categorical variables (`mfr`, `type`)	Moderate	Encode categorical variables using techniques like one-hot encoding.

Cereal Dataset	Outliers in `potass` and `fiber`	Moderate	Use techniques like z-score or IQR to detect and handle outliers appropriately.
Cereal Dataset	Inconsistent scaling of features	Low	Normalize or standardize features to ensure they are on a similar scale.
Cereal Dataset	Missing unique identifier for cereals	Low	Ensure that the `name` column is used as a unique identifier or create a new unique identifier if needed.