**Data Collection and Preprocessing Phase**

| Date | July 2024 |
| --- | --- |
| Team ID | Team-739777 |
| Project Title | Cereal analysis basede on ratings by using mechine learning techinques |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

| **Section** | **Description** |
| --- | --- |
| Data Overview | Data overview enables a thorough analysis of cereal products based on consumer ratings, helping to identify consumer preferences, market trends, and opportunities for product improvement or market positioning. |
| Data Preparation | Gather cereal attributes and ratings from datasets or sources.Handle missing values, outliers, and inconsistencies.Create new features or modify existing ones to improve model performance.Convert categorical variables to numerical using encoding techniques.Train the model using the training data.Optimize model parameters for better performance. |
| Handling missing values | After loading it is important to check the complete information of data as it can indication many of the hidden information such as null values in a column or a row Check whether any null values are there or not. if it is present then following can be done,Imputing data using Imputation method in sklearn Filling NaN values with mean, median and mode using fillna() method.Heatmap:It is way of representing the data in 2-D form.It gives coloured visual summary of the data. |
| Data Visualisation | Data visualization is where a given data set is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data.Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. In fact, even the most sophisticated machine learning models will perform poorly on data that wasn’t visualized and understood properly.To visualize the dataset we need libraries called Matplotlib and Seaborn. |
| Splitting The Dataset Into Dependent And Independent Variable | In machine learning, the concept of dependent variable (y) and independent variables(x) is important to understand. Dependent variable is nothing but output in the dataset and the independent variable is all inputs in the dataset. With this in mind, we need to split our dataset into the matrix independent variables and the vector or dependent variable. Mathematically, Vector is defined as a matrix that has just one column. |
| **Data Preprocessing Code Screenshots** | |
| Import the libraries |  |
| Importing The Dataset |  |
| Analyse The Data |  |
| Handling missing values |  |
| Data visualisation |  |
| Splitting The Dataset Into Dependent And Independent Variable |  |