**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 3 July 2025 |
| Team ID | SWTID1749727925 |
| Project Title | E-Adapt: Predicting Student Adaptability in Online Classes |
| 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 | The dataset, *Students Adaptability Level in Online Education*, is sourced from Kaggle. It contains approximately 1,200 rows and 7 features, including demographic details, education level, mode of study, and the target adaptability level. We will review basic statistics, data types, and overall structure to understand the dataset. |
| Univariate Analysis | We will explore individual variables to calculate measures like mean, median, mode, and frequency distributions. Categorical variables (e.g., Education Level, Institution Type) will be visualized with bar plots and pie charts, while numerical variables (if any) will be summarized using histograms and boxplots. |
| Bivariate Analysis | Relationships between pairs of variables will be analyzed to identify potential predictors for adaptability level. We will use cross-tabulations, grouped bar charts, and correlation matrices (if applicable) to understand how input features relate to the target variable. |
| Multivariate Analysis | Patterns involving multiple features will be explored using pair plots, heatmaps, and advanced visualizations to detect hidden relationships that can enhance model performance. |
| Outliers and Anomalies | We will identify outliers and anomalies using boxplots and IQR methods. If significant outliers are found, we will analyze their impact and decide whether to transform, cap, or remove them to improve data quality. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | We will load the dataset into a Jupyter Notebook using pandas.read\_csv(). |
| Handling Missing Data | We will check for missing values using isnull().sum() and handle them through imputation or removal, depending on their proportion and impact. |
| Data Transformation | Categorical variables will be encoded using Label Encoding or One-Hot Encoding. Numerical variables will be scaled or normalized if required. |
| Feature Engineering | We may create new features or combine existing ones if they improve model accuracy — for example, grouping similar education levels. |
| Save Processed Data | The cleaned dataset will be saved in .csv format for training and model deployment. |