Data Collection and Preprocessing Phase

Date	15 August 2024
Team ID	LTVIP2024TMID24776
Project Title	Early Prediction Of Chronic Kidney Disease
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**
1. Data Collection	The process of gathering data from various sources, such as surveys, databases, APIs, or web scraping. This step ensures that relevant and sufficient data is obtained for analysis.
2. Data Inspection	Analyzing the collected data to understand its structure, characteristics, and quality. This includes checking for missing values, data types, and general statistics to identify initial issues.
3.Exploratory Data Analysis (EDA)	A visual and quantitative analysis of the dataset to uncover patterns, trends, and insights. EDA techniques include statistical summaries, data visualization, and correlation analysis, helping to inform further analysis and model selection.
4. Data Cleaning	The process of identifying and correcting errors or inconsistencies in the data. This includes handling missing values, correcting inaccuracies, removing duplicates, and ensuring data integrity.
5. Data Balancing	Addressing class imbalance in the dataset to ensure that models are trained effectively. Techniques may include oversampling minority classes, undersampling majority classes, or using synthetic data generation methods like SMOTE.

Data Preprocessing Code Screenshots	
6. Text Preprocessing	Preparing text data for analysis, which involves tasks such as tokenization, removing stop words, stemming or lemmatization, and converting text to lower case. This step enhances the quality of text inputs for further processing.
7. Label Encoding	Transforming categorical labels into numerical values, allowing models to process these categories effectively. This step is essential for machine learning algorithms that require numerical input.
8. Data Splitting	Dividing the dataset into training and testing subsets. The training set is used to build the model, while the testing set evaluates its performance.
9. Model Building	The phase where machine learning or statistical models are constructed using the training data. Various algorithms can be applied based on the problem type (e.g., classification, regression).
10.Model Evaluation	Assessing the performance of the built model using the testing dataset. Evaluation metrics may include accuracy, precision, recall, F1 score, and AUC-ROC for classification problems.