**Model Development Phase Template**

|  |  |
| --- | --- |
| Date | 20 July 2024 |
| Team ID | SWTID1720438678 |
| Project Title | Early prediction of chronic kidney disease |
| Maximum Marks | 5 Marks |

**Feature Selection Report**

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | **Selected (Yes/No)** | **Reasoning** |
| Feature 1 | initial features in the dataset | Yes | Age, Blood Pressure, Specific Gravity, Albumin, Sugar, etc.  Derived or calculated features such as Packed Cell Volume, White Blood Cell Count, and Red Blood Cell Count. |
| Feature 2 | preprocessing steps | Yes | **Removing Columns**: You dropped the 'id' column as it was presumably irrelevant to the analysis.  **Renaming and Corrections**: You renamed columns for clarity and corrected mislabeled or inconsistent values in classification, diabetes\_mellitus, and coronary\_artery\_disease.  **Handling Missing Values**: You filled missing values in numerical columns with the mean and categorical columns with the mode, which is crucial for maintaining data integrity and ensuring the model functions correctly. |
| Feature 3 | Feature Transformation | Yes | **Encoding Categorical Features**: You converted categorical variables into a format suitable for modeling using label encoding. This includes transforming non-numeric features into numeric codes, which is essential for logistic regression.  **Numerical Corrections**: Converted strings in numerical columns like 'packed\_cell\_volume', 'white\_blood\_cell\_count', and 'red\_blood\_cell\_count' to numeric types, handling errors with coercion. |