**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 16th June 2025 |
| Team ID | LTVIP2025TMID44055 |
| Project Title | Revolutionizing Liver Care : Predicting Liver  Cirrhosis Using Advanced Machine Learning  Techniques |
| 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 | **Dimensions** : 950 x 42    **Shape** : (950,42)    **Head:**    **Overview of columns:** |

|  |  |
| --- | --- |
|  | **Duplicate rows:**      **Target value to predict:**          **Object columns:** |
| Univariate Analysis | **Exploration using Distplots:**    **Code:** |

|  |  |
| --- | --- |
|  | **Plots:**            **Inference:** |

|  |  |
| --- | --- |
|  | **Pie charts:** |

|  |  |
| --- | --- |
|  | **Code:**        **Statistical analysis for individual variables:** |

|  |  |
| --- | --- |
|  | **Mean of all numerical columns:**        **Median:** |

|  |  |
| --- | --- |
| Bivariate Analysis | **Violin Plots Between Two Variables:**      **Inference:** |

|  |  |
| --- | --- |
|  | **Inference:** |

|  |  |
| --- | --- |
|  | **Inference:** |

|  |  |
| --- | --- |
|  | **Scatter Plots:** |

|  |  |
| --- | --- |
|  | **Inference:**  No Clear Trend:   * There doesn't appear to be a clear linear relationship between age and Total Bilirubin levels. * Total Bilirubin levels are spread across the age range without a consistent pattern.             **Inference:**    A large cluster of data points is concentrated around Haemoglobin levels of 8 to 12 g/dl and Albumin levels of 1.3 to 3 g/dl.    This suggests that most individuals in the dataset have Haemoglobin levels within this range. |

|  |  |
| --- | --- |
|  | **Inference:** |
| Multivariate Analysis | **Correlation Matrix:** |

|  |  |
| --- | --- |
|  | **Heatmap:**          **Columns having high correlation:** |

|  |  |
| --- | --- |
| Outliers and Anomalies | **Identification using boxplot:**                                            **Removal using IQR:** |

|  |  |
| --- | --- |
|  | **After removing:** |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Handling Missing Data |  |

|  |  |
| --- | --- |
|  | **Missing values in Data:** |

|  |  |
| --- | --- |
|  | **Cleaning Numerical columns:**              **Filling numeric columns with mean:** |

|  |  |
| --- | --- |
|  | **Cleaning Abnormalities found in data:** |

|  |  |
| --- | --- |
|  | **Cleaning A/G Ratio:** |

|  |  |
| --- | --- |
|  | **Cleaning And Transforming Blood Pressure:**      **Cleansing Categorical Columns:** |

|  |  |
| --- | --- |
|  | **Removing all the abnormalities:** |

|  |  |
| --- | --- |
|  | **After cleaning:** |

|  |  |
| --- | --- |
|  | **Cleaning the outcome:** |
| Data Transformation | **Encoding all the categorical columns:**        **Encoded Data:** |

|  |  |
| --- | --- |
| Feature Engineering | **Feature Importance:** |
|  | **Removing Unecessary Features:** |
| Save Processed Data |  |