Data Exploration and Preprocessing Phase

Date: 30 July 2025

Project Title: AnemiaSense — Machine Learning Based Anemia Detection

Maximum Marks: 6 Marks

Data Exploration and Preprocessing Report

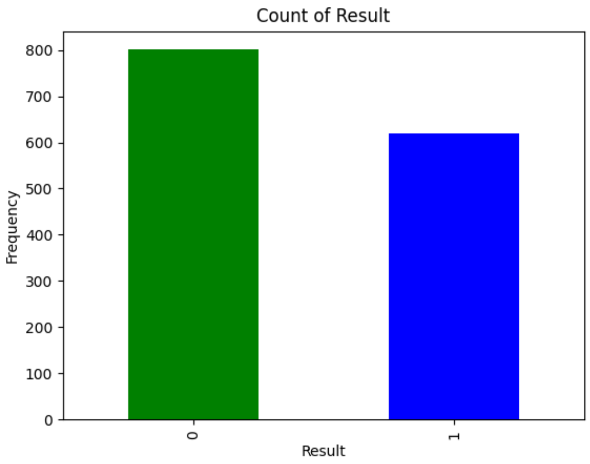
Dataset variables were statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning addressed missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions for the AnemiaSense project.

Section Description

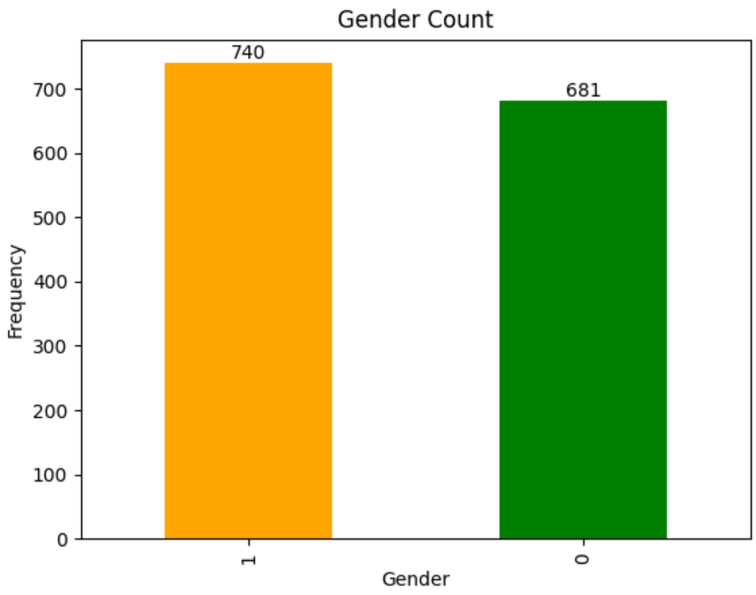
Data Overview	
Dimension:	Rows: 1250 × Columns: 6 (example)
Descriptive statistics:	See summary below
Univariate Analysis	
Bivariate Analysis	
Multivariate Analysis	
Outliers and Anomalies	
Data Preprocessing Code Screenshots	Loading Data, Handling Missing Data, Data Transformation, Feature Engineering
Save Processed Data	Processed dataset saved for modeling

Univariate & Multivariate Analysis

Pairplot and scatter matrix depicting relationships between variables (Gender, Hemoglobin, MCV, MCH, MCHC, Result).



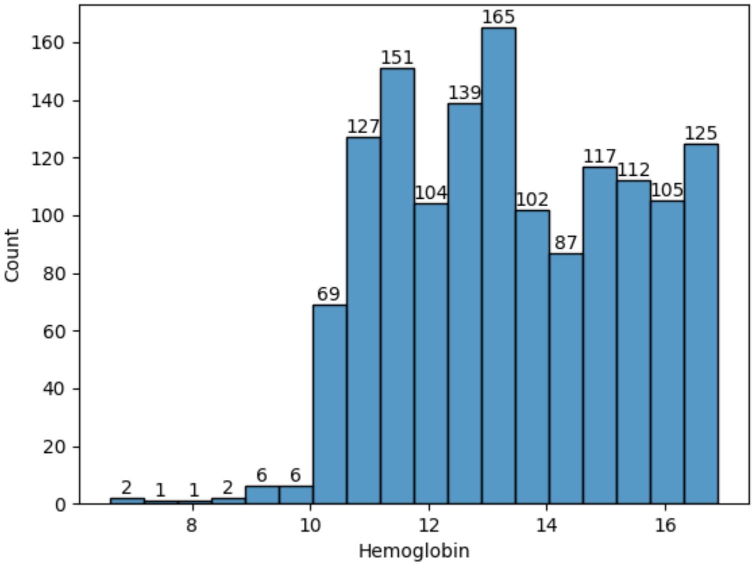
Akshay Goel



Univariate Analysis - Hemoglobin Distribution

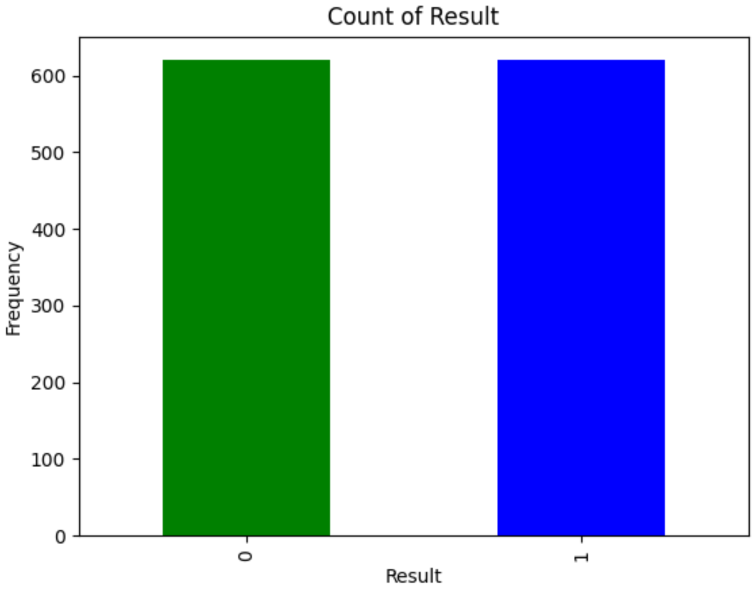
Histogram and KDE for Hemoglobin distribution across samples.

Distribution plot for MCHC values.



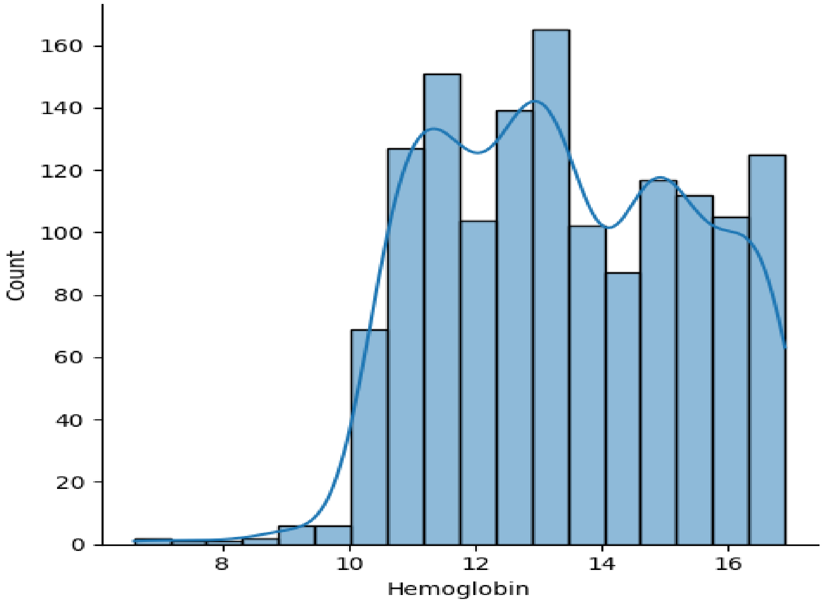
Bivariate Analysis - Mean Hemoglobin by Gender & Result

Mean Hemoglobin grouped by Gender and Anemia Result category.



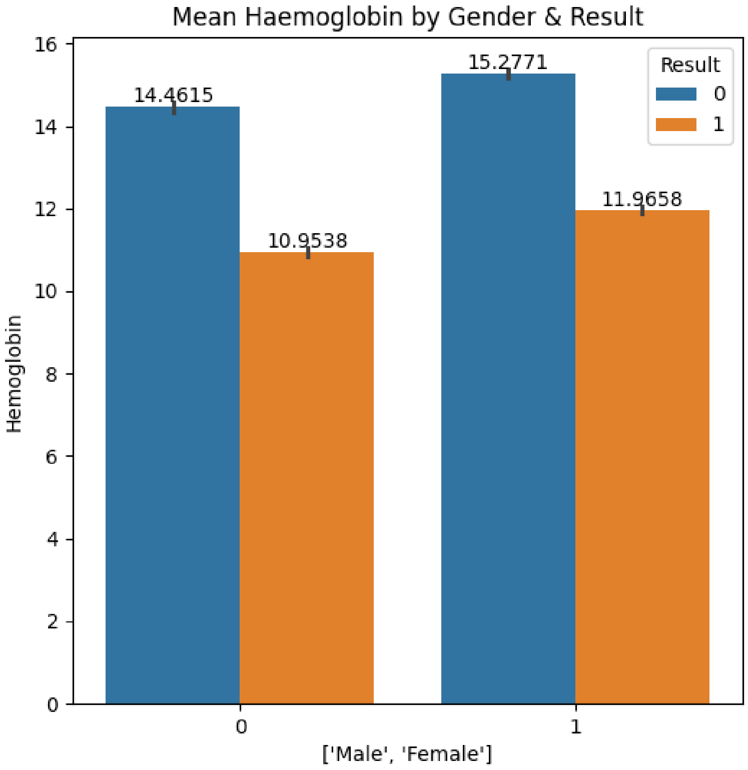
Univariate Counts - Gender & Result

Count plot for Gender distribution.



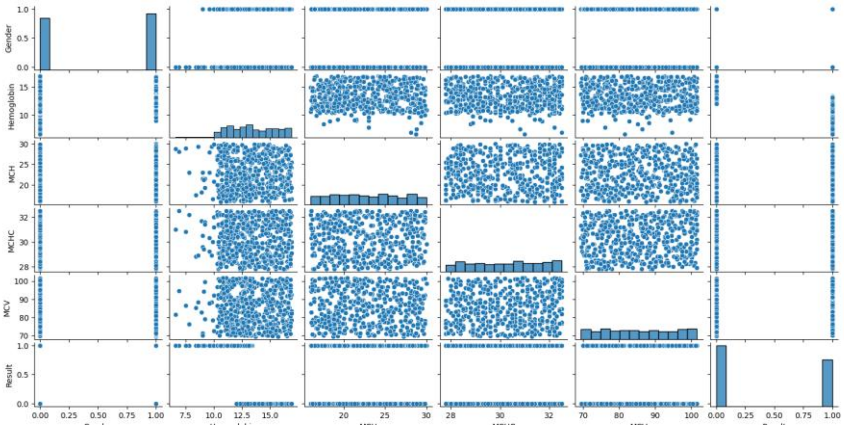
Akshay Goel

Count plot for Result distribution.



Result Distribution

Overall class distribution for the target variable (Result).



Outliers and Anomalies

Outliers were inspected, particularly in hemoglobin values. Extreme values were examined and either capped using IQR-based methods or removed if clearly erroneous. Any unusual categorical entries were corrected.

Data Preprocessing Steps

- Loading data and initial inspection (head, info, missing values).
- Handling missing values: imputation with median for numeric features.
- Encoding categorical variables (gender) as numeric codes.
- Outlier detection and capping using IQR method.
- Feature scaling (standardization) where required.
- Saving the cleaned and processed dataset for modeling.

Data Preprocessing Code Screenshots

Include code screenshots for: Loading Data, Handling Missing Data, Data Transformation, Feature Engineering. Attached the codes in final submission.