Predictive Analytics of Chronic Kidney Disease

Solution Approach

The objective of this project is to develop a robust predictive analytics framework for early detection of Chronic Kidney Disease (CKD) using clinical and laboratory data. The solution follows a systematic pipeline involving data preprocessing, exploratory data analysis (EDA), statistical validation, feature engineering, and model preparation.

Initially, the dataset was explored to understand its structure, data types, and missing value patterns. Missing data analysis revealed varying degrees of incompleteness across clinical parameters, particularly laboratory-based features. Missingness pattern evaluation using correlation analysis indicated a predominantly Missing At Random (MAR) mechanism. Accordingly, a hybrid imputation strategy was adopted: median imputation for low-missing numerical features, mode imputation for categorical attributes, KNN-based imputation for moderately missing variables, and iterative multivariate imputation (MICE) for features exhibiting complex dependencies. This ensured minimal information loss while preserving important clinical relationships.

Outlier detection was performed using clinically informed statistical techniques. Modified Z-score based detection was applied to skewed laboratory parameters such as serum creatinine, blood urea, electrolytes, and estimated glomerular filtration rate (eGFR) to robustly identify extreme pathological values. For demographic and physiological parameters such as age, blood pressure, haemoglobin, and packed cell volume, clinical threshold-based rules were employed. Instead of discarding clinically meaningful extremes, a winsorization strategy was adopted to cap implausible values, thereby maintaining data integrity while preventing model instability.

Comprehensive exploratory data analysis was conducted using interactive visualization techniques. Bivariate analysis using box plots, violin plots, and scatter plots revealed clear distributional separation between CKD and non-CKD populations. Statistical hypothesis testing using independent t-tests and Mann–Whitney U tests identified key biomarkers such as serum creatinine, blood urea, eGFR, haemoglobin, sodium, and packed cell volume as highly significant ($p < 0.001$), confirming their strong discriminatory ability and clinical relevance.

Furthermore, comorbidity analysis was performed to investigate the impact of diabetes mellitus and hypertension on CKD. Prevalence analysis, chi-square tests, and odds ratio estimation demonstrated an extremely strong association between these comorbidities and CKD. Perfect class separation observed in the dataset highlighted the dominant etiological role of diabetes and hypertension in disease progression.

Finally, clinical reference range visualization was utilized to compare observed biomarker distributions against standard physiological thresholds. This enabled direct medical validation of abnormal disease patterns observed in CKD patients. The integrated preprocessing, statistical validation, and visualization framework ensures high data quality, clinical interpretability, and a robust foundation for building accurate and explainable predictive models.