Statistical ML Project

Discussion:

Chronic kidney disease (CKD) is a critical health condition that requires early detection and management for optimal patient outcomes. In this project, I delve into the realm of medical prediction using machine learning techniques applied to a dataset containing various clinical features associated with CKD.

The project began with meticulous data preprocessing to ensure data quality and integrity. Missing values were addressed using appropriate imputation techniques, and class imbalance was mitigated using the Synthetic Minority Over-sampling Technique (SMOTE). Exploratory Data Analysis (EDA) provided valuable insights into feature relationships and distributions across CKD positive and negative cases, aiding in feature selection and model interpretation. The Support Vector Machine (SVM) algorithm was initially chosen for CKD prediction due to its effectiveness in handling complex decision boundaries. However, some challenges were encountered, particularly related to data compatibility issues between the training and testing sets. These challenges were successfully resolved, leading to a trained SVM model with an accuracy of 80%. The model exhibited promising performance metrics, including 71% precision, 85% recall, 78% F1-score and 81% ROC-AUC score indicating its ability to correctly classify CKD cases.

After applying the SVM model, I further tried to evaluate the performance using Linear Regression model to assess any potential improvement in accuracy.

Following the SVM model, a linear regression model was implemented, yielding remarkable results with an accuracy of 97.5%. The precision score was 100%, recall score: 94%, F1-score 96.5% and ROC-AUC score was 97%. The linear regression model achieved high precision and recall scores for both positive and negative CKD cases, highlighting its robust performance across different classes.