Department of Data Science

DTSC 3010: Project Report on Machine Learning

Predictive Modeling for Chronic Kidney Disease Using Machine Learning Techniques

Author:
Modesola Giwa
Student ID: 900333377

Supervisor:
Dr. Nawa Raj Pokhrel

May 2024

Contents

1 Background

Chronic Kidney Disease (CKD) represents a significant global health concern due to its increasing prevalence and substantial implications on morbidity and mortality. CKD is characterized by a gradual loss of kidney function over time, leading to complications such as hypertension, anemia, bone disease, cardiovascular disease, and eventually end-stage renal disease (ESRD) requiring dialysis or transplantation. The early stages of CKD are typically asymptomatic, making early detection and intervention challenging but crucial for altering disease progression and improving patient outcomes. Traditional diagnostic methods rely on biomarkers such as serum creatinine and urine albumin, which may not fully capture the early subtle changes in kidney function. The application of machine learning (ML) techniques in healthcare presents a promising approach to enhance the predictive accuracy of CKD diagnosis. By analyzing complex and large datasets to identify patterns and predict outcomes, ML can potentially enable earlier detection and more personalized interventions for CKD patients. This study aims to evaluate the effectiveness of various ML models in predicting the presence of CKD, thereby contributing to better clinical decision-making and ultimately improving patient care in nephrology.

2 Related Work

The integration of machine learning techniques in predicting and managing Chronic Kidney Disease (CKD) has garnered significant attention in recent medical research. Studies such as those by Coca et al. (2012)[1] and Cruz et al. (2011)[2] provide essential context and evidence supporting the potential of machine learning in enhancing CKD diagnostic and prognostic processes. Coca et al. (2012) conducted a systematic review and meta-analysis to understand the long-term impacts of Acute Kidney Injury (AKI) on the progression to CKD and End-Stage Renal Disease (ESRD). Their findings underscored a significant association between AKI and increased risks of CKD and ESRD, with pooled adjusted hazard ratios indicating a manifold increase in the risk of CKD and ESRD post-AKI. This study highlights the critical need for early detection and continuous monitoring of kidney function to prevent chronic complications resulting from acute injuries. The systematic review emphasizes the utility of predictive modeling in identifying individuals at higher risk of progression to CKD following AKI, which could be crucial for early intervention strategies. In a similar vein, Cruz et al. (2011) explored the quality of life in patients across different stages of CKD and identified various factors influencing the progression and outcomes of the disease. Their research utilized the SF-36 survey and other tools to measure the quality of life and functional status, linking these metrics with clinical, laboratory, and sociodemographic data. The study found a decrease in quality of life with advancing stages of CKD and demonstrated significant associations between lower quality of life scores and factors such as lower hemoglobin levels, higher numbers of comorbidities, and poorer sociodemographic status. These insights are vital for developing machine learning models that can predict CKD progression based on a comprehensive range of indicators, beyond traditional clinical markers.

Both studies reflect a broader trend towards employing advanced analytical

methods in nephrology. They illustrate how machine learning can leverage large datasets from diverse patient populations to improve predictive accuracy and patient-specific outcomes. This trend is supported by other literature which suggests that machine learning models are particularly adept at handling complex interactions between various risk factors and biomarkers in chronic diseases such as CKD. Moreover, these studies lay a groundwork for future research by delineating the critical factors that should be considered in the development of predictive models. This includes the severity of AKI as a significant predictor for CKD, as well as the socio-economic and demographic variables impacting patient outcomes, which could be crucial for personalized medicine approaches in CKD management. In conclusion, the integration of machine learning in CKD research, as exemplified by these studies, provides a promising pathway to enhance early diagnosis, optimize treatment strategies, and ultimately improve the prognostic outcomes for patients with or at risk of CKD. Such advancements underscore the importance of interdisciplinary approaches that combine nephrology with predictive analytics to address the complexities of kidney diseases.

3 Data-set and Features

For the purposes of this study, a dataset comprising clinical data from 400 individuals suspected of having Chronic Kidney Disease (CKD) was utilized. This dataset contains a total of 24 features per patient, encompassing demographic information, laboratory test results, and clinical parameters, all of which are essential in diagnosing and assessing the progression of CKD. The demographic details include age and gender, which are basic yet crucial factors that influence the risk and progression of kidney disease. Blood pressure data is also included, given its dual role as both a potential cause and a consequence of CKD. The dataset is rich in laboratory test results such as serum creatinine, which is pivotal for estimating the glomerular filtration rate (GFR)—a key indicator of kidney function. Blood Urea Nitrogen (BUN) levels, which can be elevated due to kidney dysfunction or other factors like diet, are also included alongside hemoglobin levels to check for anemia associated with chronic kidney conditions. Glucose levels are monitored to identify risks of diabetic kidney disease. In addition to blood tests, the dataset includes urine test results, which are critical for identifying early signs of kidney damage. This includes the Albumin to Creatinine Ratio (ACR), an important marker where higher levels suggest more severe kidney injury. Urine specific gravity, which reflects the kidney's ability to concentrate urine, and the presence of blood in urine—a potential indicator of urinary tract infections or kidney damage—are also documented. The integrity and utility of the dataset are ensured through meticulous preprocessing steps. This includes handling missing data points through appropriate imputation strategies, where continuous variables were imputed with mean values and categorical variables with the mode. To ensure that all input features contribute equally to the analyses, normalization and scaling techniques were applied. Moreover, non-numerical features were transformed into a machine-readable format using one-hot encoding. These preprocessing efforts are crucial for enhancing the performance of machine learning models, ensuring that they can effectively learn from the data

and identify complex patterns indicative of CKD. The comprehensive nature of this dataset, coupled with rigorous data preparation, provides a strong foundation for developing predictive models that could significantly aid in the early detection and management of chronic kidney disease.

4 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) performed on the dataset commenced with loading the data using Pandas in a Python environment and displaying the initial few records to understand the available features. This dataset, comprising 400 records with 25 attributes each, included both physiological measurements and laboratory test results. Essential preprocessing steps were executed to ensure data quality, including handling missing values where modes and means were imputed for categorical and continuous variables, respectively. Histograms were generated for all numerical variables to visualize their distribution and detect any outliers. This step was crucial for identifying patterns and potential data issues that could influence the outcomes of the predictive models. Additionally, bar charts were utilized to examine the distribution of categorical variables, providing insights into the data's composition and any imbalances that might affect model performance. Correlation analysis was conducted to identify relationships between variables and their impact on CKD. This analysis helped in pinpointing significant predictors for CKD This thorough analysis ensured that the predictive models developed later in the project were based on clean, well-understood, and appropriately prepared data, maximizing their effectiveness in diagnosing Chronic Kidney Disease.

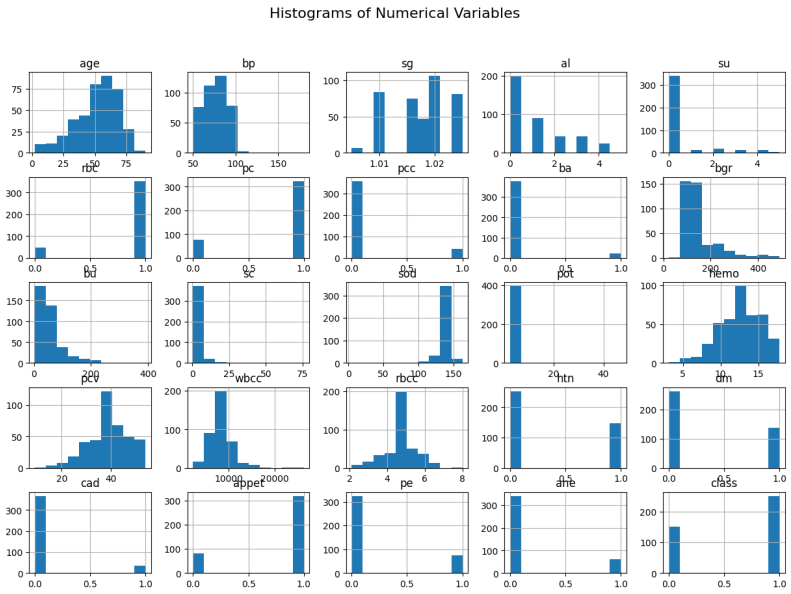

Figure 1: Histograms of Numerical Features

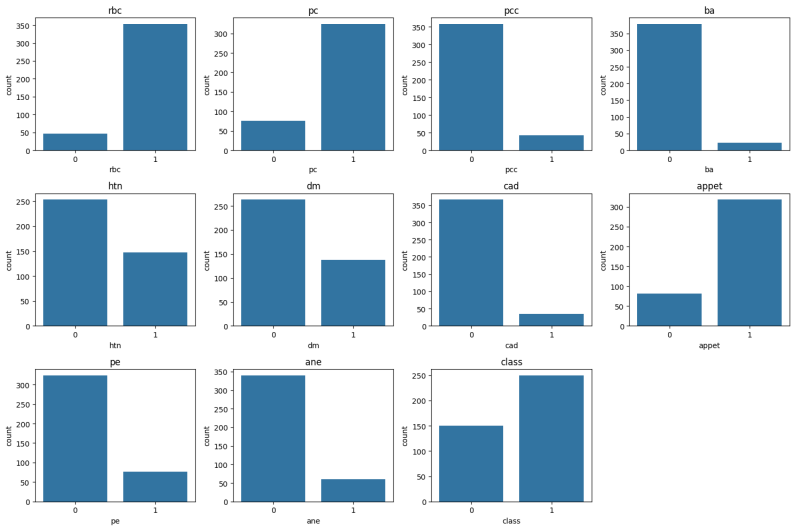

Figure 2: Bar Charts of Categorical Features

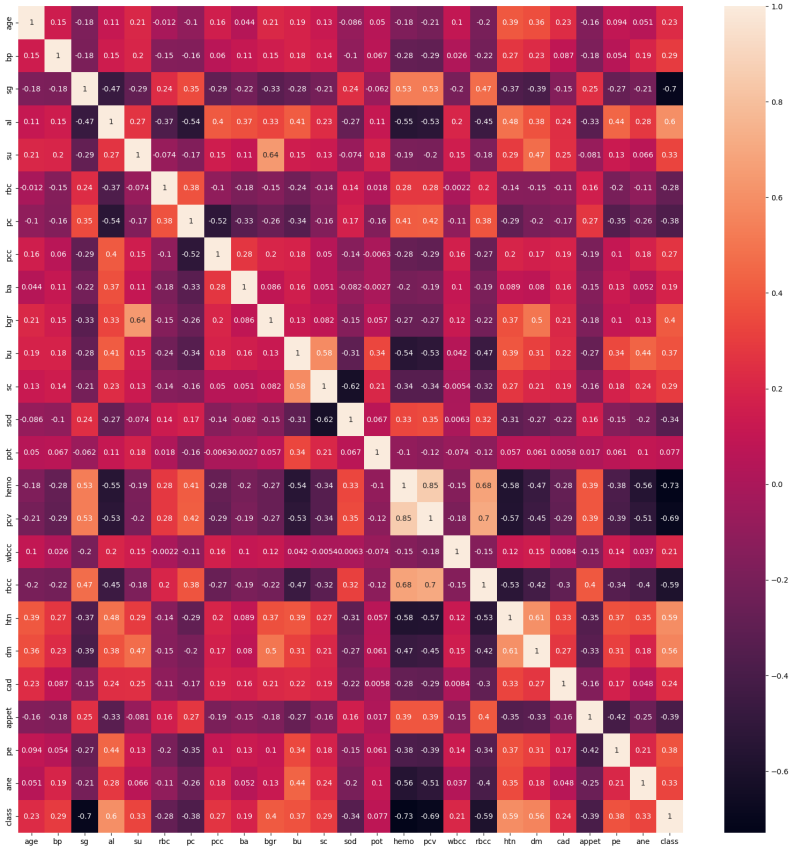

Figure 3: Correlation Heatmap

5 Methods

The methodology for this project was structured to encompass several critical phases, including data preprocessing, model development, training, validation, and evaluation, aimed at developing and assessing the efficacy of machine learning models in predicting Chronic Kidney Disease (CKD). Four different models—K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and Logistic Regression—were selected based on their diverse mechanisms and suitability for binary classification tasks. Initially, data preprocessing was undertaken after loading the dataset using Pandas. This stage involved a preliminary examination to understand the data structure and identify cleaning requirements. Missing values were addressed by imputing modes for categorical variables and means for continuous variables to preserve data integrity. Categorical variables were transformed into a numerical format through binary encoding, facilitating their use in machine learning algorithms and enhancing model interpretability. Additionally, all numerical features underwent standardization using the StandardScaler from scikit-learn, which normalized the data to have a mean of zero and a standard deviation of one. This step was crucial, particularly for distance-based models like KNN, which are sensitive to the scale of input features. For model development, the project utilized four distinct machine learning algorithms. KNN was chosen for its simplicity and effectiveness in handling nonlinear relationships between features. Decision Trees were used for their ability to create transparent models that replicate human decision-making logic. The Random Forest model, an ensemble of multiple decision trees, was implemented to improve prediction robustness and mitigate overfitting. Logistic Regression was included as a baseline for performance comparison due to its widespread application and interpretability in binary classification tasks. The dataset was split into a training set and a testing set, with 75% of the data allocated for training to ensure that the models were exposed to a comprehensive range of data scenarios. The remaining 25% served as a testing set to evaluate model performance on unseen data. This approach helped in tuning the models' hyperparameters through cross-validation, optimizing their performance based on the training data. Model evaluation was conducted using several metrics to assess each model's performance comprehensively. Accuracy was measured to determine the proportion of correctly predicted instances. Precision and recall were calculated to understand the models' effectiveness in managing false positives and false negatives. Furthermore, the AUC-ROC curve was employed as a critical evaluation tool for measuring the models' ability to discriminate between the classes across various threshold settings. Through this structured methodological approach, from data preprocessing to the final evaluation, the project not only compared the predictive capabilities of different machine learning techniques but also ensured that the findings were robust and applicable for clinical use in diagnosing CKD.

6 Result Discussion

The evaluation of the machine learning models on the Chronic Kidney Disease (CKD) prediction task provided insightful outcomes that highlight the strengths

and limitations of each model used in this study. The analysis was based on metrics such as accuracy, precision, recall, and the AUC-ROC curve, which facilitated a comprehensive assessment of each model's predictive capabilities. The Logistic Regression model, often considered a baseline in binary classification tasks, demonstrated high accuracy and provided interpretable results through its coefficients, indicating a strong linear relationship between some features and the likelihood of CKD. However, it faced challenges in convergence, suggesting that either an increase in the number of iterations or further feature scaling might be necessary for optimal performance. The Decision Tree model achieved perfect accuracy indicating that it could perfectly classify the test data without any misclassification. While impressive, this result raises concerns about potential overfitting, where the model might overly adapt to the training data's nuances at the expense of its generalizability to new, unseen data. Similarly, the Random Forest model, an ensemble of decision trees, also scored perfect marks across all metrics. This model's strength lies in its ability to reduce overfitting through the ensemble approach, making it more robust than a single decision tree. Its high performance across both precision and recall suggests that it effectively handles both classes without bias. The K-Nearest Neighbors (KNN) model showed slightly lower performance compared to the ensemble methods, with an accuracy of 99%. This slight drop in metrics might be due to the model's sensitivity to the dataset's feature scale despite prior standardization, or it could stem from the inherent noise within the data. Across all models, the AUC-ROC values were exceptionally high, suggesting excellent model performance in distinguishing between patients with and without CKD across various threshold settings. This is particularly important in a clinical setting where the threshold for classifying a condition can significantly impact patient diagnosis and subsequent treatment plans.

7 Conclusion/Future Work

This study has demonstrated the efficacy of various machine learning models in predicting Chronic Kidney Disease (CKD), with each model offering strengths that make them viable for clinical applications. Logistic Regression provided a strong baseline with interpretable outputs, while Decision Trees and Random Forest showcased exceptional performance, ideal for practical deployment due to their accuracy and robustness against overfitting. The Random Forest model, in particular, stood out as the most effective, combining reliability with high predictive power. K-Nearest Neighbors, although slightly less accurate, still performed well, emphasizing the importance of feature scaling in distance-based models. Looking ahead, future work should focus on external validation of these models across larger, more diverse datasets to ensure their generalizability and effectiveness in different demographic settings. Further optimization of model parameters and exploration of advanced machine learning techniques could enhance their predictive accuracy. Additionally, integrating these models into clinical practice through user-friendly interfaces could facilitate real-time diagnostic support, improving early detection and personalized management of CKD. Developing systems for real-time monitoring using wearable technology could also be explored, providing continuous data to predict CKD onset

and progression more effectively. These advancements could significantly improve patient outcomes and transform CKD management in clinical settings.

References

[1] Steven G Coca, Swathi Singanamala, and Chirag R Parikh. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. *Kidney international*, 81(5):442–448, 2012.

[2] Maria Carolina Cruz, Carolina Andrade, Milton Urrutia, Sergio Draibe, Luiz Antônio Nogueira-Martins, and Ricardo de Castro Cintra Sesso. Quality of life in patients with chronic kidney disease. *Clinics*, 66(6):991–995, 2011.