Employing Machine Learning Techniques and Feature Selection to Predict Chronic Kidney Disease and Highlight Critical Diagnostic Indicators

Suejit Pechprasarn[1]*, Peeraya Wetchasit[2], and Suphornthip Pongsuwan[2]

[1]College of Biomedical Engineering, Rangsit University, Pathumthani, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok, Thailand
The last two authors have contributed equally.
*Corresponding author; E-mail: suejit.p@rsu.ac.th

Abstract

Chronic kidney disease (CKD), also known as chronic kidney failure, is a medical illness with substantial morbidity and no treatment that involves the kidneys losing their ability to operate optimally. Furthermore, patients with CKD are frequently asymptomatic until the terminal stage, meaning early detection is critical. This paper utilized a publicly available clinical dataset comprising 27 clinical attributes from Enam Medical College, sourced from the UCI repository. This study aimed to (1) preprocess the data for unbiased machine learning model training, (2) train and test the CKD datasets and compare the performance of different classification models, and (3) identify crucial factors in CKD to reduce the model complexity using statistical approaches. The dataset was meticulously preprocessed and refined to ensure balance, and then divided into training and test sets at an 80:20 ratio. Subsequently, 22 machine-learning models were employed to classify CKD. Performance evaluation encompassed various metrics such as confusion matrix, accuracy, precision, sensitivity, recall, F1-score, and receiver operating curve (ROC), with the area under the curve (AUC) computed via 5-fold cross-validation on both training and test sets. The Kernel Naive Bayes model emerged as the optimal classifier, achieving a training dataset accuracy of 96.55%, precision of 95.00%, recall of 98.28%, and F1-score of 96.61%. Performance on the test dataset exhibited comparable results with metrics of 97.62% accuracy, 97.22% precision, 100.00% recall, and 98.59% F1-score, maintaining a discrepancy within 1% relative to the training dataset. Additionally, feature selection methodologies were employed to discern predictor importance, revealing that only four predictors were necessary to attain comparable model performance. Through the utilization of contemporary computational tools, healthcare providers can enhance their capacity to identify and address this persistent ailment, leading to improved patient outcomes and overall healthcare provision. Herein, a systematic approach to preprocessing data is discussed and demonstrated, different machine learning models are compared, and statistical approaches are employed to reduce the complexity of the trained model without significantly compromising the classification performance.

Keywords**:** *Chronic Kidney Disease Classification, Chronic Kidney Disease, Machine Learning, Feature Selection Methods, Artificial Intelligence*