IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 10, Issue 4, Nov 2022

Chronic kidney disease prediction based on machine learning

Lavanya Barivi[1], Dr.K. Narayana Rao[2], Dr. R. Sowmya[3]

Article Info

Received: 13-08-2022 Revised: 27 -09-2022 Accepted: 22-10-2022

Abstract:

Chronic kidney disease (CKD) is a serious and potentially lifelong condition, often caused by factors like kidney malfunctions or reduced kidney function. Early detection and appropriate treatment are crucial for slowing down or halting its progression, preventing the need for life-preserving interventions like dialysis or surgeryIn a supervised learning setting, we've evaluated twelve different machine learning classifiers. The XgBoost classifier has emerged as the top performer, boasting an accuracy of 0.983, precision of 0.98, recall of 0.98, and an F1-score of 0.98.Our research underscores the potential of recent advances in machine learning and predictive modeling for discovering innovative solutions, not only for kidney disease but also for broader applications in healthcare and beyond.

Keywords: Chronic kidney disease, Machine learning, XgBoost classifier, Classification model

Introduction

Chronic kidney disease (CKD) is a global health concern, often developing silently with no apparent symptoms in its early stages. Early detection through routine medical tests, including blood and urine assessments, is critical to prevent CKD from progressing to severe stages, which can lead to kidney failure and necessitate treatments like dialysis or transplantation.Recognizing signs of advanced CKD is essential, and individuals should promptlyseek medical advice if they suspect kidney issues. Early detection plays a crucial role in preventing kidney failure.Various diagnostic tests are available to assess kidney function and the progression of CKD, including the estimated glomerular filtration rate (eGFR) test, urine analysis for blood and protein, monitoring blood pressure, and, when necessary, imaging scans and kidney tissue analysis.Studies have shown an alarming increase in hospital admissions related to CKD, highlighting the importance of early detection and intervention to address this growing health concern.

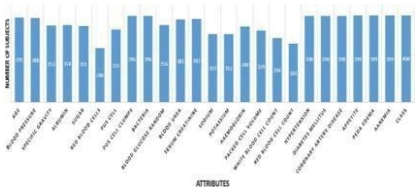
[1] Associate Professor,Department of CSE, RISE Krishna sai Gandhi Group of Institutions, Ongole,[2] Professor, Department of CSE, RISE Krishna sai Prakasam Group of Institutions, Ongole,[3] Professor, Department of CSE, RISE Krishna sai Gandhi Group of Institutions, Ongole

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 10, Issue 4, Nov 2022

Experimental Data:
This study utilizes the CKD dataset from the UCI Machine Learning Repository, specifically designed for predicting Chronic Kidney Disease (CKD). The dataset contains 25 attributes, with one serving as the target variable, categorized as "yes" or "no."
Out of these attributes, 11 are numerical, and 14 are nominal, collectively facilitating
The dataset offers a wide range of attributes essential for CKD prediction and analysis.

CKD Dataset with Principal Component Analysis (PCA):

PCA, or Principal Component Analysis, is a valuable technique for reducing data dimensionality while preserving essential information. In the case of the CKD dataset used for chronic kidney disease detection, it originally contains 24 input features. However, PCA is applied to understand each feature's contribution to the final outcome. You can find an overview of the CKD dataset after applying PCA in Table 1, compared to the original dataset. Further details on the PCA process will be discussed in the following section. In-depth descriptions of eachfeature in the main CKD dataset. In-depth descriptionsof each feature in the main CKD dataset.



Methodology

Machine Learning Classification In our study, we assessed the performance of different machine learning classifiers using various evaluation metrics. We employed a 10-fold cross-validation to avoid overfitting and nested cross-validation to fine-tune the model's parameters. Our analysis was carried out in Jupyter Notebook using Python 3.3 and Scikit-learn libraries, a free platform for machine learning in PythonWe considered several evaluation measures like sensitivity, specificity, AUC, and the F1-score to assess each model's performance. The unique model outputs depended on their specific parameters.

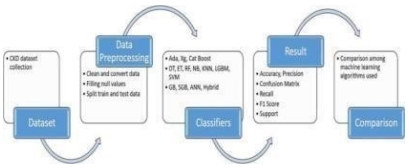
CKD Dataset Analysis

In analyzing the CKD dataset, we experimented with different machine learning algorithms, including SVM, KNN, LGBM, and hybrid models. Figure 2 illustrates the CKD diagnostic process.

Preprocessing Stage

During the preprocessing phase, we filled missing numerical values using the mean method and missing nominal values with the mode method. Relevant features for CKD diagnosis were selected using the Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) algorithms.

These carefully chosen features were used in the classifiers for CKD diagnosis. Our approach aimed to achieve high classification performance with a limited number of features, optimizing performance through PCA.



Data preprocessing is the essential step of modifying raw data to prepare it for machine learning models. In our case, this step involved dealing with the presence of numerous unaccounted-for numbers and addressing missing values, specifically "NaNs." To ensure data quality, any row containing even a single "NaN" value was removed from the dataset.

To make the data more computer-friendly, weassigned numeric codes to categorical variables. For example, "1" and "0" represented "normal" and "abnormal" for "rbc" and "pc," while "1" and "0"indicated "yes" and "no" for variable like "htn," "dm,""cad," "pe," "ane," and "appet."

Although initially described as categorical, we treated variables sg, al, and su as numeric due to their numeric nature. Each categorical variable was converted into a factor, and each sample was assigned a unique number from 1 to 400.

The dataset contained a substantial number of missing values, with only 158 complete instances. Missing values were filled in using imputation techniques, such as filling missing numerical values with the median of the relevant variable and missing categorical values with the most frequently appearing category in the corresponding variable.

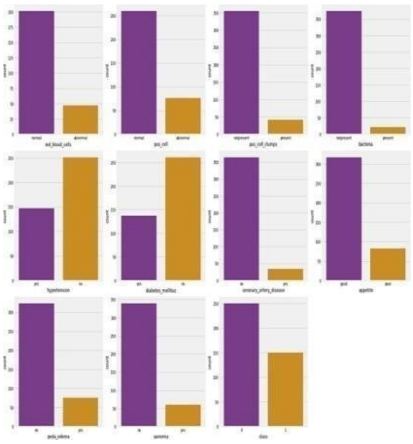
International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 10, Issue 4, Nov 2022

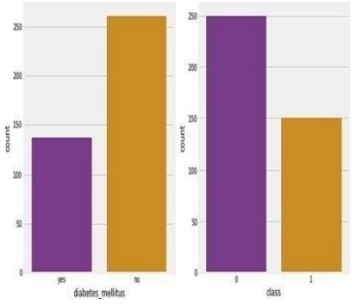
KNN-Based Imputation

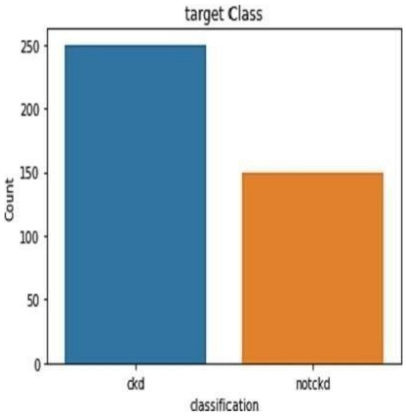
To address missing physiological measurements, we used a k-nearest neighbors (kNN) approach. Since similar physical conditions should result in comparable physiological measurements, kNN imputation was employed to fill in missing numbers for individuals with similar conditions. This approach was adapted from the field of hyperuricemia and applied to diagnostic data for other disorders.

CKD Dataset Overview

The CKD dataset comprises 400 rows and 14 columns, with the "class" column indicating "yes" or "no" for CKD. "Yes" is assigned the value "1," indicating a CKD patient, and "no" is assigned the value "0," signifying a non-CKD patient. The dataset includes categorical columns, as shown in Figure 4, and a view with PCA. Figure 5 demonstrates the distribution of instances with and without diabetes-mellitus and Figure 6 illustrates the distribution of CKD and non-CKD patients in the target class.



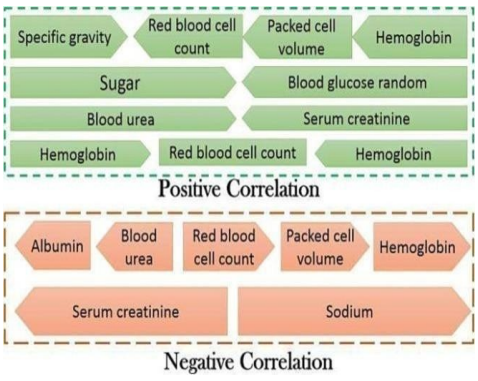




Class Imbalance and Feature Correlations

Some of the features in the dataset have unbalanced categories, necessitating the use of stratified folds in cross-validation. To ensure there's no significant class imbalance, we examined the percentage of patients with chronic renal disease (62.5%) and those without it (37.5%). Fortunately, the classes are reasonably balanced.

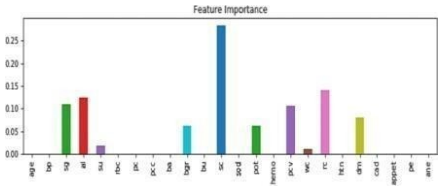
The Figure presents a heatmap illustrating the correlations between the class label and various features. Blood pressure, specific gravity, albumin, sugar, blood urea, serum creatinine, blood glucose random, and sodium display positive correlations, while hemoglobin, potassium, white blood cell count, and red blood cell count exhibit negative correlations. This heatmap offers insights into the relationships between the features in the dataset.



Eliminating variables that are neither helpful for prediction nor connected to response variables can be accomplished by extracting feature vectors or predictors. Because of this, the building of the model would not be affected by variables that are not linked to the problem at hand, which would lead to the

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 10, Issue 4, Nov 2022

models making accurate predictions.displays the results of the procedure for extracting important variables from the data.



Machine Learning Classifiers

In our pursuit of diagnosing chronic kidney disease (CKD), we developed various machine learning models, each employing specific subsets of features from the CKD dataset. Disease diagnosis often involves organizing diagnostic samples in multidimensional space. Predictors used during classification (CKD or non-CKD) are critical in this process, helping data samples cluster into distinct regions. Effective classification techniques are vital in this context.

Classifier Overview

1. AdaBoost: AdaBoost combines and enhances weak and inaccurate rules to create highly accurate prediction rules. It improves performance by allowing weak classifiers to learn and adapt from their own errors. The noise reduction ability is enhanced when AdaBoost reaches a stopping condition.

2. Decision Tree: Decision trees are widely used for supervised machine learning, offering a flowchart-like structure. Each node represents a feature test, branches represent outcomes, and leaves contain class labels. The decision tree uses various techniques to split data into homogeneous sets.

3. XGBoost: XGBoost employs gradient boosted trees to address structured data problems. Each tree acts as a weak learner, iteratively trained to predict residuals of previous trees. The final prediction is a combination of these trees.

4. CatBoost: CatBoost is a gradient boosting technique using decision trees, designed for rapid predictions. It utilizes balanced trees for categorical features and is known for various applications, including self-driving cars and recommendation systems.

5.K-Nearest Neighbor (KNN): KNN is a simple, widely-used supervised learning approach for classification and regression. It calculates Euclidean distances between query and data examples, with the value of "k" determined automatically to increase accuracy.

More Classifiers

1. Random Forest: Random forest combines multiple decision trees on different dataset subsets. It averages the results to improve predictive accuracy.

2. Gradient Boosting: Gradient boosting combines weak learners to build a powerful predictive model. It aims to minimize prediction error by setting desired results for each subsequent model.

3. Stochastic Gradient Boosting: In this method, the training dataset is subsample, learners are trained on random samples, and correlation between results is reduced.

4. Light Gradient Boosting Machine (LGBM): LGBM is a powerful, distributed gradient boosting framework, particularly efficient in handling large datasets. It grows trees vertically, reducing loss more efficiently.

5. Extra Tree*:* The Extra Trees method uses multiple randomized decision trees on different data subsamples to enhance predictive accuracy and reduce over fitting.

6. Support Vector Machine (SVM): SVM is a classical learning method for classification and regression, suitable for various linear and non-linear problems.

7. Artificial Neural Network (ANN): ANN consists of interconnected neurons that process inputs, with weights adjusted through back-propagation for training.

8. Hybrid Machine Learning (HML): HML integrates existing techniques or strategies from various domains to enhance predictive capabilities and overcome the limitations of individual machine learning methods.

These diverse classifiers were employed in our study to achieve the goal of diagnosing CKD efficiently and accurately.

		Predicted CKD	
		ckd	not ckd
Ground Truth	ckd	TP	FP
	not ckd	FN	TN

ISSN 2347–3657
Volume 10, Issue 4, Nov 2022

In evaluating the performance of each classifier for diagnosing chronic kidney disease (CKD), various metrics were employed, including accuracy, sensitivity, specificity, precision, recall, and the F1 score. The formulas for these metrics are provided in equations (1) to (5).

Accuracy: Indicates the overall correct predictions.

Sensitivity: Reflects the ability to correctly identify CKD cases.

Specificity: Measures the ability to accurately identify non-CKD cases.

Precision: Focuses on the accuracy of positive predictions.

F1 Score: Balances precision and recall for model evaluation.

The models' results were assessed using 10-fold cross-validation to prevent overfitting. Layered cross-validation was also used to fine-tune the models' parameter settings. The experiments were conducted in Python, utilizing the Google Colab web application. Scikit-learn, an open-source Python machine learning library, was instrumental in these analyses. Evaluation metrics included accuracy, F1 score, precision, and recall.

Demonstrates that different sets of outputs are produced based on the parameter values assigned to each model. Notably, XgBoost exhibited the best performance with an accuracy of 0.9833 on the original CKD dataset, which improved to 0.9916 after implementing PCA. Some classifiers, such as AdaBoost, Random Forest, Gradient Boosting, LGBM, and Extra Tree, achieved an accuracy of 0.9833 on the original CKD dataset. However, the performance of KNN and MLP was less impressive, with ANN classifiers achieving 60% accuracy for both datasets due to limited data availability.

The results of these experiments, including testing and training accuracy, F1 measure, precision, recall, and confusion matrices, are summarized in the table.

References

1. Akhter T., Islam M.A., Islam S. Artificial neural network based covid-19 suspected area identification. *J Eng Adv.* 2020;1:188–194. [Google Scholar]

2. Aljaaf A.J., Al-Jumeily D., Haglan H.M., et al. *2018 IEEE Congress on Evolutionary Computation(CEC)* IEEE; 2018. Early prediction of chronic kidneydisease using machine learning supported by predictive analytics; pp. 1–9. [Google Scholar]

3. Almasoud M., Ward T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *Int J Soft Comput Appl.* 2019;10 [Google Scholar]

4. Banik S., Ghosh A. Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis. *Int Urol Nephrol.* 2021;53:713–718. [PubMed] [Google Scholar]

5. Charleonnan A., Fufaung T., Niyomwong T., Chokchueypattanakit W., Suwannawach S., Ninchawee N. *2016 Management and Innovation Technology International Conference (MITicon)* IEEE; 2016. Predictive analytics for chronic kidney disease using machine learning techniques. pp. MIT–80. [Google Scholar]

6. Chen Z., Zhang X., Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *Int Urol Nephrol.* 2016;48:2069–2075. [PubMed] [Google Scholar]

7. Chittora P., Chaurasia S., Chakrabarti P., et al. Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access.* 2021;9:17312–17334. [Google Scholar]

8. Cueto-Manzano A.M., Cortés-Sanabria L., Martínez-Ramírez H.R., Rojas-Campos E., Gómez-Navarro B., Castillero-Manzano M. Prevalence of chronic kidney disease in an adult population. *Arch Med Res.* 2014;45:507–513. [PubMed] [Google Scholar]

9. Dua D., Graff C. UCI Machine Learning Repository. 2017. http://archive.ics.uci.edu/ml URL.

10. Gudeti B., Mishra S., Malik S., Fernandez T.F., Tyagi A.K., Kumari S. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* IEEE; 2020. A novel approach to predict chronic kidney disease using machine learning algorithms; pp. 1630–1635. [Google Scholar]

11. Heung M., Chawla L.S. Predicting progression to chronic kidney disease after recovery from acute kidney injury. *Curr Opin Nephrol*

ISSN 2347–3657
Volume 10, Issue 4, Nov 2022

Hypertens. 2012;21:628–634. [PubMed] [Google Scholar]

12. Islam M., Shampa M., Alim T., et al. Convolutional neural network based marine cetaceans detection around the swatch of no ground in the bay of bengal. *Int J Comput Digit Syst.* 2021;12:877– 893. [Google Scholar]

13. Islam, M.A., Akhter, T., Begum, A., Hasan, M.R., Rafi, F.S. Brain tumor detection from MRI images using image processing.

14. Islam M.A., Akter S., Hossen M.S., Keya S.A., Tisha S.A., Hossain S. *2020 3rd International Conference on Intelligent Sustainable Systems(ICISS)* IEEE; 2020. Risk factor prediction of chronic kidney disease based on machine learning algorithms; pp. 952–957. [Google Scholar]

15. Islam M.A., Hasan M.R., Begum A. Improvement of the handover performance and channel allocation scheme using fuzzy logic, artificial neural network and neuro-fuzzy system to reduce call drop in cellular network. *J Eng Adv.* 2020;1:130–138. [Google Scholar]

16. Mahesh B. Machine learning algorithms-areview. *Int J Sci Res (IJSR).[Internet]* 2020;9:381–386. [Google Scholar]

17. Polat H., Danaei Mehr H., Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *J Med Syst.* 2017;41:1–11. [PubMed] [Google Scholar]