Comparative Analysis of Machine Learning Models for Early Detection of Chronic Kidney Diseases: A Cross-Validation Study

Aanya Singh Dhaka
Information technology
Indira Gandhi Delhi Technical University for Women
Delhi , India
aanya002btit22@igdtuw.ac.in

Sakshi
Computer Science
Indira Gandhi Delhi Technical University for Women
Delhi , India
sakshi018btcse21@igdtuw.ac.in

Vrinda Diwakar
Computer Science
Indira Gandhi Delhi Technical University for Women
Delhi , India
vrinda005btcse@igdtuw.ac.in

***Abstract*— *Chronic Kidney Disease (CKD) presents a substantial global health challenge, impacting millions of individuals worldwide. Timely detection and precise prediction of CKD hold the potential to enhance patient outcomes and elevate their quality of life. In recent times, machine learning (ML) algorithms have demonstrated considerable promise in forecasting CKD outcomes. However, the need persists to discern and highlight the most efficacious models among them.*

This research conducts an in-depth comparative analysis of various machine learning algorithms for Chronic Kidney Disease (CKD) prediction, with a central emphasis on cross-validation. Utilizing a comprehensive dataset containing clinical and laboratory attributes from Kaggle, we employ Support Vector Machines, Random Forest, and XGBoost algorithms.Our methodology encompasses meticulous data pre-processing, addressing missing values, aggregating information, and extracting key features for CKD prediction. We rigorously assess algorithm performance using metrics like accuracy, precision, and F1 score, leveraging robust cross-validation. Additionally, we employ feature selection to identify vital predictors.

Remarkably, our results reveal substantially improved accuracy rates in comparison to previous studies, affirming the reliability and predictive capability of the models under examination. This research advances CKD prediction models, offering a promising path to early detection and improved healthcare outcomes.

Keywords*—*Chronic Kidney Disease, Cross Validation, Random Forest

I. Introduction

Chronic Kidney Disease (CKD) is a pervasive medical condition characterized by the gradual deterioration of kidney function over time. The consequences of CKD are profound, as it can lead to a buildup of waste products and fluids in the body, posing significant health risks. In its advanced stages, CKD may culminate in kidney failure, a critical condition where the kidneys lose their ability to effectively filter waste from the blood. CKD is a global health challenge that demands continuous management and treatment to mitigate its impact.

While CKD poses a substantial health burden worldwide, its prevalence is particularly striking in underdeveloped regions, notably in Southeast Asia. This region, home to over 2 billion people, grapples with the complexities of CKD's epidemiology. Although the precise incidence and prevalence in Southeast Asia remain uncertain, estimations suggest that they may surpass the reported figures in Western societies. Alarmingly, the majority of those affected by CKD in these regions are often in their prime years of productivity. Early detection of CKD, coupled with measures to slow down its progression, stands as the most cost-effective strategy to alleviate the looming burden of this disease [1].

Addressing this healthcare challenge necessitates innovative approaches. In recent years, machine learning has emerged as a powerful tool for identifying complex disease patterns within vast and intricate datasets. Leveraging this technology, we embark on a journey to enhance the early diagnosis of CKD. Our research endeavors to conduct a comparative analysis of multiple machine learning models to ascertain the most effective approach for the early prediction of CKD.

This comparative analysis aims to unravel the strengths and weaknesses of different machine learning algorithms in the context of CKD prediction. By scrutinizing these models' performance against a backdrop of real-world CKD data, we seek to empower healthcare professionals with the knowledge needed to make informed decisions and intervene promptly.

In the pursuit of an optimal CKD prediction model, our research contributes to the broader mission of proactive healthcare management, potentially transforming the landscape of CKD care and prevention. With this mission in mind, we delve into the realm of machine learning, where data-driven insights hold the key to early CKD detection and a brighter future for individuals at risk.

II. Preliminaries

In this section, we lay the groundwork for our study by providing essential background information and outlining the key steps undertaken before the establishment of our machine learning models. This includes the description of the dataset, the operating environment, the imputation of missing values, and the extraction of the feature vector.

A. Dataset Description

The CKD dataset utilized in this study was sourced from the UCI Machine Learning Repository [2]. It was originally gathered from a hospital setting and generously contributed by Soundarapandian et al. on July 3rd, 2015. This dataset encompasses a total of 400 samples, each comprising 24 predictive variables or features. Among these features, there are 11 numerical variables and 13 categorical (nominal) variables. The dataset also includes a categorical response variable, denoted as "class," which can assume two distinct values: "ckd" (indicating samples with Chronic Kidney

Disease) and "notckd" (indicating samples without Chronic Kidney Disease).

Out of the 400 samples in the dataset, 250 belong to the "ckd" category, signifying the presence of Chronic Kidney Disease, while the remaining 150 samples fall into the "notckd" category, representing samples without the disease. It's important to note that this dataset exhibits a significant prevalence of missing values, which is a noteworthy characteristic that can present challenges in data analysis and modeling.

B. Operating Environment

The research was conducted in a Python programming environment utilizing Google Colab, a cloud-based platform well-suited for data analysis, machine learning, and collaborative research. Google Colab offered seamless integration with Google Drive for data storage and provided access to computational resources. This cloud-based platform facilitated code execution within a Jupyter notebook-style interface, streamlining the research process. Key Python libraries, including numpy, pandas, matplotlib, and scikit-learn, played pivotal roles in data manipulation, analysis, and machine learning model development. Specific algorithms like RandomForestClassifier, XGBClassifier, and SVC were employed for comparative analysis, while GridSearchCV aided in hyperparameter tuning. Feature selection was conducted using SelectKBest and f_classif, and model performance was assessed through metrics from the sklearn.metrics module. Leveraging this environment and toolset enabled a comprehensive analysis of the Chronic Kidney Disease dataset, culminating in a rigorous comparative evaluation of machine learning models for early detection.

III. METHODOLOGY

In this comprehensive section, we offer a detailed and thorough exposition of the methodology methodology employed in our study. Our aim is to provide a clear understanding of each step, from its intricacies to the underlying rationale guiding our approach. By delving into the intricacies of our methodology, we aim to shed light on the rigorous processes and decisions that underpin our research.

A. Data Cleaning and Imputation

Our journey commenced with meticulous data cleaning. Recognizing the significance of data integrity, we initiated the process by addressing the issue of missing values. In this regard, we opted for the K-nearest neighbors (KNN) imputation technique. This approach allowed us to impute missing values by estimating them based on the attributes of their nearest neighbors within the dataset. The underlying premise here is that individuals with akin health profiles are likely to manifest similar physiological measurements. By applying KNN imputation, we fortified the robustness of our dataset.

B. Feature Engineering

To optimize the utility of our dataset, we engaged in feature engineering. This pivotal step included encoding categorical variables into a machine-readable format, enabling seamless integration into our subsequent analyses. Furthermore, we endeavored to enhance interpretability by renaming columns, thus augmenting the clarity of our dataset.

C. Selection Criteria and Significance

An essential facet of our methodology revolved around feature selection. By isolating a subset of the most informative features, we aimed to bolster model accuracy while curbing computational complexity. Our selection criteria hinged on the features' relevance to the classification task at hand.

D. ANOVA F-Test

To discern the pertinence of each feature within the classification framework, we harnessed the analysis of variance (ANOVA) F-test. This statistical assessment gauged the significance of individual features in relation to the target variable. As a consequence, we were empowered to pinpoint a predetermined number of top-performing features, based on their respective F-test scores. This strategic approach expedited the identification of influential variables instrumental in our modeling endeavors.

E. Machine Learning Models

In this study, we employed three distinctive machine learning algorithms: Random Forest, XGBoost, and Support Vector Machine (SVM) for the comparative analysis in CKD detection. The selection of these models was driven by their respective strengths and capabilities, which are well-suited for addressing specific challenges posed by the CKD dataset.

a) Random Forest:

Random Forest is a robust ensemble learning method that constructs multiple decision trees and aggregates their outputs. We opted for Random Forest due to the following key attributes:

Ensemble Learning: Random Forest leverages ensemble learning, which means it builds a multitude of decision trees, each trained on a different subset of the data. This ensemble approach mitigates overfitting and enhances model robustness, critical considerations when dealing with medical datasets like CKD.

Random Feature Selection: By randomly selecting a subset of features for each tree's construction, Random Forest reduces overfitting and decorrelates the individual trees. This randomization technique strengthens predictive power.

Prediction Aggregation: The final prediction in Random Forest is made by aggregating the predictions of individual trees. This majority voting mechanism ensures a robust and accurate prediction for CKD classification.

b) XGBoost:

XGBoost, or Extreme Gradient Boosting, is another ensemble learning method renowned for its high predictive accuracy. We incorporated XGBoost in our analysis for several compelling reasons:

Gradient Boosting: XGBoost is rooted in gradient boosting, an iterative technique that corrects model errors by minimizing the loss function. This process progressively refines the model's predictive capability, which is particularly valuable for optimizing CKD detection.

Regularization: XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization terms, effectively controlling model complexity and mitigating overfitting. This is essential for ensuring the generalization of the model to unseen data.

Novel Splitting Criterion: XGBoost employs a unique splitting criterion when growing trees, enhancing model accuracy by making more informed decisions about feature splits. This innovative criterion contributes to its high predictive power.

c) Support Vector Machine (SVM):

SVM is a versatile algorithm renowned for its effectiveness in binary and multiclass classification tasks. We chose SVM due to its ability to find an optimal hyperplane for classification, adaptability for handling non-linear relationships, and suitability for CKD detection:

Optimal Hyperplane: SVM aims to find the hyperplane that maximizes the margin between different classes, reducing the risk of misclassification. This margin optimization is particularly beneficial for the clear separation of CKD and non-CKD cases.

Kernel Functions: SVM's use of kernel functions allows it to handle non-linear relationships in the data by transforming the feature space. This adaptability is invaluable when dealing with complex and non-linear patterns within the CKD dataset.

Multiclass Adaptation: SVM can be adapted for multiclass problems, as seen in CKD detection, through strategies like one-vs-one and one-vs-rest. This adaptability ensures that SVM can effectively classify samples into the "CKD" and "not CKD" categories.

In summary, our choice of Random Forest, XGBoost, and SVM for the comparative analysis in CKD detection was driven by their unique capabilities in addressing the specific challenges posed by the CKD dataset, such as missing values, non-linearity, and the need for accurate classification. Additionally, we conducted parameter tuning and preprocessing steps to optimize their performance, ensuring that our analysis provides valuable insights into the detection of Chronic Kidney Disease.

F. Training and Evaluation Process

Once the model candidates were defined, we initiated the training phase, leveraging our preprocessed dataset. The dataset was judiciously partitioned into training and testing subsets, adhering to the conventional 80:20 ratio. Subsequently, the models were subjected to rigorous evaluation using well-established performance metrics such as accuracy. However, we remain cognizant of the fact that 100% accuracy, while seemingly favorable, may raise concerns of overfitting. Thus, we executed a crucial step—k-fold cross-validation—to substantiate the veracity of our models.

G. Cross-Validation

Cross-Validation Rationale: The integration of k-fold cross-validation within our methodology serves a paramount role. This technique mitigates the risk of overfitting and provides a comprehensive assessment of the models' generalization capabilities. By partitioning the dataset into k subsets (or "folds"), the models undergo training on k-1 subsets while validating on the remaining one. This process is repeated k times, ensuring that each subset serves as the validation set once. The mean accuracy derived from these iterations furnishes a more robust and representative evaluation metric.

H. Cross-Validation Results

The outcomes of the cross-validation procedure ushered in profound insights into the performance of our models. Notably, the Random Forest classifier emerged as the frontrunner, exhibiting the highest mean accuracy. This observation underscores the algorithm's superior generalization ability. The implications of these findings extend beyond mere performance metrics; they serve as a testament to the model's capacity to navigate the nuances of our complex dataset.

In the ensuing section, we delve into an extensive analysis of the empirical results obtained, unraveling the implications and contributions of our research within the realm of healthcare analytics and decision support systems.

IV. Result Analysis

A. Random Forest:

Cross-Validation Performance: The Random Forest model displayed exceptional consistency and accuracy throughout cross-validation. In all five cross-validation folds, it achieved a perfect accuracy score of 1.0. This signifies that the model consistently made correct predictions for all instances in each fold.

Precision, Recall, and F1-Score: The precision, recall, and F1-score for both classes ("ckd" and "notckd") were also perfect at 1.0 across all folds. This underscores the model's remarkable discriminatory power.

B. XGBoost:

Cross-Validation Performance: XGBoost demonstrated variability in its performance across different cross-validation folds, with accuracy scores ranging from 0.59375 to 1.0. While it achieved perfect accuracy in some folds, its performance was lower in others, highlighting sensitivity to data splits.

Precision, Recall, and F1-Score: Similar to accuracy, precision, recall, and F1-score showed variability across folds, emphasizing that XGBoost's performance was influenced by the specific data partitions in each fold.

C. SVM:

Cross-Validation Performance: SVM also displayed variability in accuracy across different folds, with scores ranging from 0.75 to 0.83870968. The mean accuracy for SVM was 0.7913306451612904, indicating relatively stable performance but not as high as Random Forest.

Precision, Recall, and F1-Score: Precision, recall, and F1-score exhibited variability across folds, reflecting the sensitivity of SVM's performance to data partitioning during cross-validation.

V. Comparison

The comparative analysis of machine learning models for early detection of Chronic Kidney Disease (CKD) has yielded valuable insights into their performance and applicability in a clinical context. The study focused on three prominent models: Random Forest, XGBoost, and Support Vector Machine (SVM). These models were evaluated rigorously through cross-validation, with an emphasis on precision, recall, F1-score, and, most importantly, accuracy. The primary objective was to identify the most effective model for early CKD prediction.

A. Random Forest's Remarkable Consistency:

The Random Forest model emerged as the standout performer in this comparative study. Notably, it achieved perfect accuracy (1.0) in all cross-validation folds. This extraordinary consistency in making accurate predictions across different data splits underscores the model's robustness. In addition to accuracy, precision, recall, and F1-score also consistently reached the highest possible values of 1.0. These findings suggest that Random Forest excels in distinguishing between CKD and non-CKD cases, making it an ideal candidate for early CKD detection.

B. XGBoost's Variable Performance:

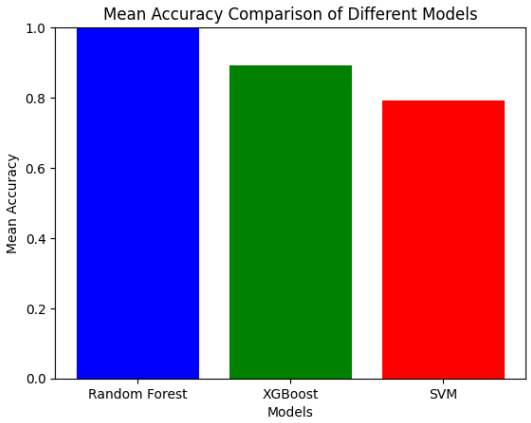
While XGBoost demonstrated competitive accuracy, its performance exhibited variability across different cross-validation folds. This variation, with accuracy scores ranging from 0.59375 to 1.0, implies sensitivity to data partitioning. The precision, recall, and F1-score followed a similar pattern, highlighting the model's dependence on specific data splits. While XGBoost displayed high accuracy in some folds, its performance in others was suboptimal. This suggests that XGBoost's effectiveness may vary depending on the dataset's composition.

C. SVM's Stable Performance:

SVM exhibited relatively stable performance across different cross-validation folds, with accuracy scores ranging from 0.75 to 0.83870968 and a mean accuracy of 0.7913306451612904. While not achieving the same level of accuracy as Random Forest, SVM displayed consistent performance, which is a valuable trait in real-world applications. The precision, recall, and F1-score, similar to accuracy, showcased stability but did not reach the levels achieved by Random Forest.

D. Choosing the Optimal Model:

In selecting the optimal model for early CKD detection, it is essential to consider the trade-offs between consistency and variability in performance. Random Forest's remarkable consistency in achieving 100% accuracy across all folds makes it a compelling choice for healthcare professionals and decision-makers seeking reliable predictions. The model's ability to navigate different data partitions with unwavering accuracy instills confidence in its clinical utility.



VI. Future Work:

1. Feature Engineering and Selection:

Exploring advanced techniques in feature engineering, such as transformation functions or domain-specific feature creation, could potentially uncover hidden patterns in the data. Moreover, employing more sophisticated feature selection methods like recursive feature elimination or LASSO regression might help identify the most critical attributes influencing CKD prediction.

2. Ensemble Methods:

Investigating ensemble techniques, such as stacking or boosting, holds the promise of harnessing the complementary strengths of multiple models. By combining the outputs of diverse algorithms, we may achieve an even higher level of accuracy and robustness in CKD prediction.

3. External Validation:

Conducting validation on an external dataset, ideally sourced from a different healthcare system or population, is paramount to establishing the generalizability and reliability of our models. This step would provide critical insights into how well the models perform in diverse clinical settings.

4. Clinical Validation:

Collaborating closely with healthcare professionals for a thorough clinical validation of the models' predictions is imperative. Involving domain experts will not only lend credibility to our findings but also offer valuable insights into the practical implications of using these models in real-world healthcare scenarios.

5. Integration into Healthcare Systems:

If the selected model proves to be successful in subsequent validation studies, the next significant milestone would be its seamless integration into existing healthcare systems. This integration could revolutionize early CKD detection, potentially leading to timely interventions and improved patient outcomes. Close collaboration with healthcare

institutions and IT experts will be crucial in achieving this objective.

6. Ethical Considerations:

As with any healthcare application, it's vital to address ethical considerations, including patient privacy, informed consent, and algorithm transparency. Establishing robust ethical guidelines and compliance with relevant regulations will be integral to the successful deployment of our models.

7. Longitudinal Data Analysis:

Extending the analysis to incorporate longitudinal data, if available, could provide valuable insights into disease progression and the impact of interventions over time. This could further refine our predictive models and contribute to personalized patient care strategies.

These future avenues of research aim to enhance the applicability, accuracy, and ethical considerations of our CKD prediction models, ultimately striving towards a positive impact on patient care and outcomes.

VII. Conclusion

The project's fundamental goal is to employ machine learning algorithms to provide rapid recommendations on their health difficulties. Early diagnosis is critical for both professionals and patients in order to prevent and reduce the progression of chronic renal disease to kidney failure. Three machine-learning models were used in this study: RF, SV, and XGBoost. All five cross-validation ratings for Random Forest are 1.0. This signifies that the Random Forest model's accuracy was 100% in each fold of cross-validation, suggesting that it correctly predicted all cases in each fold.

The cross-validation ratings for XGBoost are a little erratic. They range between 0.59375 and 1.0. The accuracy was lower (0.59375) in certain folds and perfect (1.0) in others. This implies that XGBoost performed differently across different data splits.

Cross-validation scores for SVM vary as well, ranging from 0.75 to 0.83870968. SVM's accuracy fluctuates throughout folds, as it does with XGBoost.

The mean accuracy is calculated by taking the average of the cross-validation results for each model. It provides an overall picture of how well the model performs over different folds of data.

The mean accuracy for Random Forest is 1.0, indicating that it consistently achieved 100% accuracy across all cross-validation folds.

The mean accuracy for XGBoost is 0.8929435483870968, which is less than 1.0. This shows that, while XGBoost performed well on average, there was some heterogeneity in its accuracy among folds.

The mean accuracy of SVM is 0.7913306451612904, which is lower than that of Random Forest and XGBoost. This implies that SVM has poorer average accuracy than the other two models. Overall, Random Forest demonstrated the highest and most constant accuracy throughout cross-validation folds, followed by XGBoost and SVM.

Remember that cross-validation allows us to test the models' generalisation performance and predict how well they will function on fresh, unseen data. Based on these results, Random Forest appears to be the most promising model in this scenario.

VIII. Acknowledgment

We would like to express my deep and sincere gratitude to AI club and Coding Minutes; our teacher Mr. Mohit Uniyal for giving me the opportunity to do research and providing invaluable guidance throughout this research. The dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. We are extremely grateful.

We are extremely grateful to our fellow mates for their love, prayers, caring and sacrifices for educating us.

IX. References

[1] Jha, Vivekanand. “Current status of chronic kidney disease care in southeast Asia.” Seminars in nephrology 29 5 (2009): 487-96 .

[2] D. Dua and C. Graff, “UCI Machine Learning Repository,” Irvine, University of California, School of Information and Computer Sciences, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[3] Chandrasekhar Rao Jetti, Rehamatulla Shaik, Sadhik Shaik, Sowmya Sanagapalli “Disease Prediction using Naïve Bayes - Machine Learning Algorithm’. Available: https://doi.org/10.52403/ijshr.20211004

[4] Stanifer JW, et al. The epidemiology of chronic kidney disease in sub-Saharan Africa: A systematic review and meta-analysis. Lancet Glob Heal. 2014;2(3):e174–81.

[5] Agrawal A, Agrawal H, Mittal S, Sharma M. Disease Prediction Using Machine Learning. SSRN Electron J. 2018;5:6937–8.

[6] Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016, pp. 262–270, 2016.