Feature selection criteria and machine learning algorithms for analysing chronic kidney databases

Josué Nguinabé[1,2☯], Khadiime Jhumka[3], Naushad Mamode Khan[4☯], Muhammad Muzzammil Auzine[3*], Maleika Heenaye-Mamode Khan[3‡], Zahra Mungloo-Dilmohamud[5‡], Aboo Swalay Fedally[6], Yuvraj Sunecher[7]

1 Department of Computer Science and Mathematics, Faculty of Science, University of Ngaoundere, Cameroon **2** African Institute for Mathematical Sciences (AIMS), Limbe, South-West, Cameroon
3 Department of Software and Information Systems, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius
4 Department of Economics and Statistics, Faculty of Social Sciences and Humanities, University of Mauritius, Reduit, Mauritius
5 Department of Digital Technologies, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius
6 Consultant Nephrologist, Sir Seewoosagur Ramgoolam National (SSRN) Hospital Pamplemousses, Mauritius
7 Department of Finance and Accounting, University of Technology, Mauritius

☯These authors contributed equally to this work.
‡These authors also contributed equally to this work.
¤Current Address
†Deceased
¶Membership list can be found in the Acknowledgments section.
*muhammad.auzine2@umail.uom.ac.mu

Abstract

Chronic kidney disease poses a significant global health threat and is increasingly recognized as a silent killer worldwide. Unfortunately, lack of awareness about the disease and the challenges in early detection contribute to high mortality rates and severe suffering among affected individuals. Machine learning techniques have emerged as valuable tools for early disease identification and prediction, providing medical experts with fast and accurate diagnostic support. In this study, we employed machine learning techniques to predict chronic kidney disease. The dataset used in our analysis contained 25 features, encompassing both numerical and nominal values, alongside the corresponding class labels for each instance. To ensure data quality, we performed preprocessing steps to handle missing values. Subsequently, we employed logistic regression as a feature selection method to identify the most relevant predictors. Our analysis revealed that among the initially considered 19 features, the following 8 features demonstrated strong associations with chronic kidney disease: coronary artery disease, appetite, bilateral pedal edema, anemia, specific gravity, albumin, blood urea, and hemoglobin. We have used 10 different Machine Learning models to classify CKD and found that using the logistic feature selection increases the accuracy of some models as well as reducing model overfitting in the case of Random Forest and Extra tree.

Introduction

Kidney plays an important role in the human body, from maintaining fluid balance to maintaining a role in regularizing blood pressure [1]. So, a kidney failure not only affects the kidney but it also gives rise to a lot of diseases such as cardiovascular and diabetes [2]. There are different kidney diseases which can cause kidney failure for example kidney cancer which has seen 403,262 new worldwide cases in the year 2018 [3] or even kidney stones which have been associated with an increased risk of lots of diseases such as cardiovascular diseases, diabetes, and hypertension [4]. The kidney disease which leads to more kidney failure is the Chronic Kidney Disease (CKD). CKD affects over 10% of the population worldwide and it was ranked 16th among the leading causes of death in 2016, and is expected to rise to 5th ranked by 2040. CKD has been an important contributor to chronic non-communicable diseases (NCDs) and Goal 3.4 of SDGs shows that the world sees CKD as a significant burden for individuals, health care systems and societies [5]. According to the Global Burden of Disease CKD Collaboration, Mauritius is ranked second in CKD prevalence and death due to CKD with 218,092 CKD patients and 1070 deaths in 2017(GBD Chronic Kidney Disease Collaboration, 2020). CKD is normally defined as an abnormality of kidney function or structure for ≥ 3 months. CKD is a silent disease, as most sufferers have no symptoms until kidney function drops to 15–20% of normal. A blood or urine test is used to verify whether the estimated glomerular filtration rate (eGFR) is less than $15ml/min/1.73m^2$ or the presence of albuminuria (ie, urine albumin $\geq 30mg$ per 24 hours or urine albumin-to-creatinine ratio [ACR] $\geq 30mg/g$). Several studies have concluded a lot of causes for CKD such as cardiovascular disease, diabetes, hypertension, Inherited diseases or systemic infections and others [6, 7]. However, there are no clear risk factors which can conclude that a person will have CKD or not. Early diagnosis of CKD is a big challenge for nephrologists. With the growing number of chronic kidney patients, and the high costs of diagnosis and treatment, computer-assisted diagnostics to assist medical experts in making diagnostic decisions are being looked into.

With the emergence of the big data era, new ways for constructing a predictive model that previously relied on classical statistics become accessible. Machine learning (ML) is a subset of artificial intelligence (AI) that enables a machine to execute a task without explicit instructions. ML algorithms may be taught to capture the underlying patterns of sample data and generate predictions about actual data based on the learned knowledge when employed in predictive modelling. ML reflects more complicated maths functions than standard statistics and typically leads in superior performance in predicting a result that is determined by a broad number of factors with non-linear, complex interactions. Recently, ML has been used in a number of research and exhibited a high degree of performance that outperformed conventional statistics and even humans. Further research implemented ML to predict early detection of diseases such as colorectal cancer [8], skin cancer [9], alzheimer disease [10], early-stage cancer (Machine-Learning Models Can Help Detect Early-Stage Cancer) and so on. In this study, we will be focusing on some ML algorithms to predict the binary classification of CKD disease while also exploring a regression approach to identify the potential risk factors contributing to CKD.

The organization of the paper is thus as follows: In Section 2, an overview of related works, followed by the research methodology in which the the CKD data preparation and the logistic regression as feature selection are provided. Section 3 focuses on the fitting of the various ML models and providing the possible classification of CKD by making uses the best features resulting from the logistic regression. Section 4 comprises the discussions on the several significant factors. The concluding remarks and some limitations are provided in Section 5.

Materials and methods

Related works

A variety of studies have been conducted to predict CKD phases using various classifications systems. Current research approaches for detecting CKD using Machine Learning algorithms and performance indicators are summarised in this section.

Using the UCI Machine learning Repository CKD dataset, [11] conducted a research to compare different machine learning algorithms in detecting CKD. Before the data is being trained, they applied several preprocessing steps; removing all rows having null data, encode all the nominal features, finally a feature selection is used. The principal component analysis (PCA) algorithm was used as the feature selection tool and out of 23 features, only 10 features were selected. Those 10 features include serum creatinine, albumin, specific gravity, sugar, blood glucose random, potassium, packed cell volume, white blood cells count, Red blood cell count and diabetes. They obtained the best performance with XgBoost with an accuracy of 0.9916. The adaBoost, random forest, gradient boosting, LGBM, and Extra Tree show the same accuracy (0.9833). [12] have explored using different parameters to develop a system to classify CKD. In this study, they have analysed classifying CKD in 160 different scenarios with 5 different algorithms namely Neural Networks (ANN), Naive Bayes, k-Nearest Neighbors, Support Vector Machine (SVM) and J48. The best result of 97.66% of accuracy, 96.13% of sensitivity, 98.78% of specificity and 98.31% of precision was achieved with J48 with an oversampled data and using cross-validation of 10 folds.

Meanwhile, [13]designed an improved version of the Teacher Learner Based Optimization (TLBO) algorithm in the classification of the UCI CKD dataset. The algorithm has identified 16 features including blood pressure, specific gravity, albumin, red blood cells, pus cell, pus cell clumps, blood glucose random, blood urea, serum creatinine, potassium, hemoglobin, red blood cell count, diabetes mellitus, appetite, pedal edema and anemia. They used 3 classifiers namely the SVM, Gradient boosting and a CNN model and the latter scored the highest accuracy of 95.25% with features selected from the improved TLBO algorithms. [14] designed a deep learning model and compared it to traditional algorithms such as Support Vector Machine , K-Nearest Neighbor, Logistic regression, Random Forest, and Naive Bayes classifier. The proposed model achieved the highest accuracy and they also performed a feature importance using Recursive Feature Elimination (RFE) on the UCI CKD dataset and found that hemoglobin, Specific Gravity, Serum Creatinine, Red Blood Cell Count, Albumin, Packed Cell Volume, and Hypertension were the key features in determining CKD with the said dataset. Using the same dataset, [15] have used another feature selection tool known as the Correlation-based feature subset selection (CFS). Before the feature selection, the rows having null values were removed. Then, using CFS, the dataset was reduced from 23 features to 8 features which includes specific gravity, albumin, serum creatinine, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension. The reduced dataset was classified using the Levenberg–Marquardt classifier and obtained an average accuracy of 99.78%.

A deep neural network-based Multi-Layer Perceptron Classifier is proposed in [16] to identify CKD in patients. The algorithm was trained using data from 400 people and took into account a variety of symptoms and indicators such as age, blood sugar, red blood cell count, and so on. Experiments show that the suggested model performs flawlessly in classification tasks. The authors' goal is to facilitate to introducing Deep Learning algorithms to learning from dataset attribute reports and effectively detecting CKD. The suggested Deep Neural Network model for chronic kidney disease diagnosis outperforms common machine learning models such as support vector machines and naive Bayes classifiers by 100% of accuracy. In [17] the authors use functional magnetic resonance imaging (fMRI) as a new noninvasive method to identify early stages of CKD. A total of 21 articles with 1472 patients were included for analysis. The results indicated fMRI techniques had great efficacy in assessing early stages and different stages of CKD, among which DTI, IVIM, and BOLD exerted great superiority in differentiating early CKD patients from the general population, while DWI showed the advantage in distinguishing different CKD stages.

In [18] PC-AKI was defined based on the serum creatinine criteria of the Kidney Disease. Six feature selection methods were used to identify the most influential predictors from 79 candidate variables. Deep neural networks (DNNs) were used to establish the model and compared with logistic regression analyses. Model discrimination was evaluated by area under the receiver operating characteristic curve (AUC). Low-risk and high-risk cutoff points were set to stratify patients. 14 variables-based DNN model had significantly better performance than the logistic regression model with AUC being 0.939 (95% confidence interval: 0.916–0.958) and 0.940 (95% confidence interval: 0.909–0.954) in the internal and external validation cohorts, respectively, and showed promising discrimination in subgroup analyses (AUC $\geq$ 0.800).

A large number of works have been carried out as listed above including works on the comparison of the performance of algorithms in classification including classic algorithms and neural networks on the one hand and on the other hand, some authors seek to optimise certain parameters of the algorithms in order to obtain better performance in CKD classification. However, the use of logistic regression as a feature selection to select the appropriate variables before using the classic model machines remains undeveloped. This is the reason which lead to this study. The theory behind this study is that classic methods can produce better performance if implemented with the right variables (hence the idea of logistic regression as feature selection).

Data specifications

In this study, the Chronic Kidney dataset from UCI machine repository is being used. The dataset includes 400 patient records with 25 different attributes such as diabetes, serum creatinine and others as shown in the x-axis of Fig 1. This is a binary classification, as we have used two classes for predicting CKD and NOT CKD. The data needs to be processed before using Logistic regression as the dataset has some missing values and some typing errors. First of all, the number of missing values is calculated for each attribute as shown in Figure X. Dealing with null values becomes a critical step in data preprocessing. As per [19], if an attribute contains more than 20 percent of missing values and imputation is used, it can have substantial differences on the prediction model. Thus, attributes such as red blood cells, blood sodium, blood potassium, white blood cell count and red blood cell count, having more than 80 missing values, are removed from the data. The data is then divided in two categories: Nominal and numerical columns. The method of imputation used is essential as it can have a large impact on the performance of the models. KNN imputer is a popular technique for imputing missing values and it is frequently used in place of traditional imputation methods such as mean and median imputation [20, 21]. For numerical columns, the missing data are replaced using the kNN imputer. As in the case of nominal columns, the missing data are imputed with the mode value of this attribute. Fig 1 shows the missing data in the original dataset

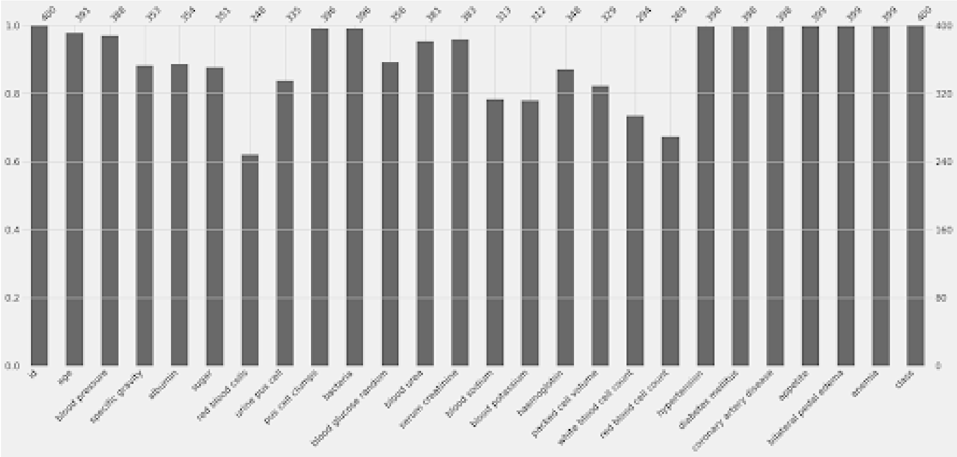

Fig 1. Missing Values in dataset

Logistic regression

Logistic regression is a classification algorithm that is often used in machine learning for feature selection. It is a straightforward and quick method for identifying the most significant features in a dataset and developing a predictive model capable of reliably classifying new observations. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one, more nominal, ordinal, interval, or ratio-level independent variables. To build a logistic regression model, we select the features that will be used to predict the target variable. The selection of features is an important step in the modeling process, as it can significantly impact the performance of the model [22].

Logistic regression equation

In 1972, Nelder and Wedderburn proposed Generalized Linear Model (glm) which include Logistic Regression Model. GLM model was developed in an effort to provide a method for applying linear regression to issues that were not directly suited for linear regression application. In fact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression etc) which included logistic regression as a special case [23]. The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x_1 + Y x_2 \tag{1}$$

Where, $g()$ is the link function, $E(y)$ is the expectation of target variable and $\alpha + \beta x_1 + x_2$ is the linear predictor ($\alpha, \beta,$ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Logistic function

Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1. Furthermore, if the sigmoid function output (estimated probability) exceeds a predetermined threshold, the model predicts that the instance belongs to that class. The model predicts that the instance does not belong to the class if the calculated probability is less than the predefined threshold. For example, if the output of the sigmoid function is above 0.5, the output is considered as 1. On the other hand, if the output is less than 0.5, the output is classified as 0. Furthermore, if the graph is drawn further to the left, the anticipated value of y will be 0 and vice versa. In other words, if the sigmoid function returns 0.65, it means that the event, has a 65% chance of occuring [24]. The sigmoid function, often known as an activation function in logistic regression, is defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

where, e = base of natural logarithms value = numerical value one wishes to transform.

The following equation represents logistic regression:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{3}$$

As, the equation of the best fit line in linear regression is:

$$y = \beta_0 + \beta_1 x \tag{4}$$

Assume we're using probabilities (P) instead of y. But there is a problem here: the value of (P) will either exceed 1 or fall below 0, and we know that the Probability range is (0-1). To overcome this issue we take "odds" of P (Odds can always be positive which means the range will always be $(0, +\infty)$. It's defined as the ratio of the probability of success and probability of failure):

$$P = \beta_0 + \beta_1 x \tag{5}$$

$$\frac{P}{1-P} = \beta_0 + \beta_1 x \tag{6}$$

The issue here is that the range is confined, and we don't want a restricted range because it would reduce our correlation. By limiting the range, we are reducing the quantity of data points, and as we reduce our data points, our correlation will fall. A variable with a narrow range is difficult to model. To control this we take the log of odds which has a range from $(-\infty, +\infty)$.

$$log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x \tag{7}$$

We want to predict probability rather than the log of odds so we just need a function of P. To do so, multiply both sides by the exponent and then solve for P.

$$e^{ln[\frac{P}{1-P}]} = e^{(\beta_0 + \beta_1 x)} \tag{8}$$

$$\frac{P}{1-P} = e^{(\beta_0 + \beta_1 x)} \tag{9}$$

$$P = e^{(\beta_0 + \beta_1 x)} - Pe^{(\beta_0 + \beta_1 x)} \tag{10}$$

$$P = P\Big(\frac{e^{(\beta_0 + \beta_1 x)}}{P} - e^{(\beta_0 + \beta_1 x)}\Big) \tag{11}$$

$$1 + e^{(\beta_0 + \beta_1 x)} = \frac{e^{(\beta_0 + \beta_1 x)}}{P} \tag{12}$$

$$P = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} \tag{13}$$

Now, divide each term by $e^{(\beta_0+\beta_1 x)}$ and we will get:

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{14}$$

This is our logistic function, often known as a sigmoid function.

Machine learning algorithms

In this section, we will give a brief description of the ten machine learning algorithms that have been used in this work.

(1)-K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is a non-parametric algorithm originally proposed by Fix and Hodges Jr. in 1951. This algorithm operates based on the principle of classifying new data points by considering the majority class of their k-nearest neighbors. To determine the class label of a new data point, KNN calculates the distances between the new point and the existing data points. The class assignment is then determined through a majority vote among the k-nearest neighbors. Over the years, KNN has been widely applied across various domains, including educational settings and intrusion detection in computer networks, showcasing its versatility and effectiveness [25, 26].

(2)-Decision Tree Classifier The Decision Tree Classifier, introduced by Quinlan in 1986, is a machine learning model that constructs a tree-like structure to represent decisions and their potential outcomes. This algorithm partitions the dataset by evaluating different features, resulting in branches that correspond to decision rules. At each node of the tree, the algorithm selects the most informative feature to split the data, with the goal of maximizing information gain or reducing impurity. Decision trees have demonstrated their utility in diverse domains, including data stream mining and comparative analyses of various decision tree algorithms [27, 28].

(3)-Random Forest Classifier. The Random Forest Classifier, proposed by Breiman in 2001, is an ensemble learning technique that leverages multiple decision trees for classification tasks. This model generates a collection of decision trees by utilizing random subsets of the training data and random subsets of the features. Each decision tree in the forest independently predicts the class, and the final prediction is determined through a majority voting scheme. Random Forest Classifier has found applications in diverse domains, including the prediction of student performance and the identification of financial distress [29, 30].

(4)-AdaBoost Classifier. The AdaBoost (Adaptive Boosting) Classifier, introduced by Freund and Schapire in 1996, is a boosting algorithm designed to construct a robust classifier by combining multiple weak classifiers. In this approach, each training instance is assigned a weight, and weak classifiers are iteratively trained on misclassified instances, with their importance determined by the weights. The final prediction is then generated by aggregating the weighted predictions of the weak classifiers. AdaBoost has been successfully applied in various domains, including face detection and financial distress prediction [31–33].

(5)-Gradient Boosting Classifier. The Gradient Boosting Classifier, originally proposed by Friedman in 2001, is an ensemble learning technique that sequentially combines weak classifiers. Each subsequent classifier in the sequence focuses on rectifying the errors made by the preceding classifiers. This approach employs gradient descent optimization to minimize the loss function by iteratively introducing weak classifiers trained on the negative gradients of the loss function. Gradient Boosting has demonstrated its effectiveness in a wide range of tasks, including efficient training of convolutional neural networks and prediction of high-dimensional sparse outputs [34, 35].

(6)-Stochastic Gradient Boosting (SGB). Stochastic Gradient Boosting (SGB), an extension of the Gradient Boosting algorithm proposed by Friedman in 2002, introduces a stochastic element into the training process. In contrast to the conventional Gradient Boosting approach, SGB randomly selects a subset of samples from the training set at each iteration. This random sampling not only reduces computation time but also helps prevent overfitting. By incorporating stochasticity, SGB promotes increased diversity among the weak learners and enhances the overall generalization performance of the model. SGB has been successfully applied in various domains, including recommender systems and anomaly detection tasks, showcasing its versatility and effectiveness [36–38].

(7)- XGBoost. XGBoost (Extreme Gradient Boosting), introduced by Chen and Guestrin in 2016, is an optimized gradient boosting framework that incorporates several advancements over traditional gradient boosting methods. These enhancements include a regularized learning objective, parallel tree construction, and hardware optimization. The XGBoost algorithm is renowned for its efficiency, scalability, and exceptional performance across a wide range of machine learning tasks. It has found successful applications in diverse domains, including fraud detection and recommendation systems, where its capabilities have proven valuable [39, 40].

(8)-CatBoost Classifier. The CatBoost Classifier, proposed by Prokhorenkova et al. in 2018, is a gradient boosting algorithm that stands out for its ability to effectively handle categorical features. This classifier incorporates innovative techniques such as categorical feature encoding, ordered boosting, and gradient-based ranking. By leveraging these techniques, CatBoost aims to deliver accurate predictions while minimizing the need for extensive preprocessing of categorical data. The algorithm has demonstrated successful applications in diverse domains, including customer behavior prediction and click-through rate prediction, highlighting its versatility and effectiveness [41, 42].

(9)-Extra Trees Classifier. The Extra Trees (Extremely Randomized Trees) Classifier, introduced by Geurts et al. in 2006 [43], is an ensemble learning technique that extends the Random Forest algorithm. The Extra Trees Classifier introduces an extra level of randomness by selecting random feature subsets and random thresholds for splitting at each node of the decision tree. This added randomness enhances the diversity among the individual trees, resulting in reduced variance and improved robustness to noise in the data. The Extra Trees Classifier has been successfully applied in various domains, such as drug resistance prediction (Gupta et al., 2019) [44], highlighting its versatility and effectiveness in different contexts.

(10)-LGBM Classifier. The LGBM (Light Gradient Boosting Machine) Classifier, proposed by Ke et al. in 2017 [45], is a gradient boosting framework renowned for its efficiency and scalability. This classifier leverages a technique called Gradient-based One-Side Sampling (GOSS) to intelligently select and prioritize data instances during the training phase, leading to expedited and more precise model construction. The LGBM Classifier has been successfully employed in a diverse range of machine learning tasks, including disease prediction and high-dimensional sparse output prediction, demonstrating its versatility and efficacy [45, 46].

Results and discussion

Logistic regression

Logistic regression is a type of regression model that can be use to understand the relationship between one or more predictor variables and a response variable when the response variable is binary. In this study, we deals multiple predictor variables and one response variable, then we have used multiple logistic regression, which uses the following formula to estimate the relationship between the variables:

$$Log\Big[\frac{p(X)}{(1-p(X))}\Big] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{15}$$

Multiple logistic regression uses the following null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \tag{16}$$

$$H_A : \beta_1 = \beta_2 = \ldots = \beta_k \neq 0 \tag{17}$$

The null hypothesis states that all coefficients in the model are equal to zero. In other words, none of the predictor variables have a statistically significant relationship with the response variable, y. The alternative hypothesis states that not every coefficient is simultaneously equal to zero. After implementation of the logistic regression, we find the results bellow, from this results, we focus on P-Value to selection our best feature.

```
    Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.715
Dependent Variable: class            AIC:              139.3832
Date:               2023-05-02 13:07 BIC:              207.6120
No. Observations:   268              Log-Likelihood:   -50.692
Df Model:           18               LL-Null:          -178.05
Df Residuals:       249              LLR p-value:      8.9678e-44
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
-------------------------------------------------------------------
           Coef.    Std.Err.      z      P>|z|     [0.025    0.975]
-------------------------------------------------------------------
x1        -0.2432     0.3131   -0.7765   0.4375   -0.8569    0.3706
x2        -0.1962     0.3268   -0.6004   0.5483   -0.8367    0.4443
x3        -0.0518     0.2826   -0.1833   0.8546   -0.6058    0.5021
x4         0.0308     0.3591    0.0857   0.9317   -0.6731    0.7347
x5         0.2247     0.3361    0.6686   0.5038   -0.4340    0.8834
x6        -0.4681     0.3113   -1.5038   0.1326   -1.0782    0.1420
x7         0.4757     0.3209    1.4825   0.1382   -0.1532    1.1047
x8         0.6983     0.3429    2.0362   0.0417    0.0261    1.3705
x9        -0.5067     0.3246   -1.5611   0.1185   -1.1428    0.1295
x10        0.1260     0.2926    0.4307   0.6667   -0.4475    0.6996
x11        0.3650     0.3218    1.1343   0.2567   -0.2657    0.9957
x12       -2.1280     0.3725   -5.7132   0.0000   -2.8581   -1.3980
x13        1.1430     0.3559    3.2118   0.0013    0.4455    1.8405
x14        0.3750     0.3354    1.1180   0.2636   -0.2824    1.0323
x15        0.3719     0.4125    0.9016   0.3673   -0.4366    1.1804
x16       -0.5996     0.3453   -1.7363   0.0825   -1.2764    0.0772
```

```
x17        0.3100    0.4010    0.7732   0.4394   -0.4759    1.0959
x18       -1.6174    0.7778   -2.0795   0.0376   -3.1418   -0.0929
x19       -0.1744    0.6785   -0.2571   0.7971   -1.5042    1.1554
=================================================================
Where x1 = urine pus cell, x2 = pus cell clumps,
      x3 = bacteria, x4 = hypertension, x5= diabetes mellitus,
      x6 = coronary artery disease,
      x7 = appetite, x8 = bilateral pedal edema,
      x9 = anemia, x10 = age, x11 = blood pressure,
      x12 = specific gravity, x13 = albumin,
      x14 = sugar, x15 = blood glucose random,
      x16 = blood urea, x17 = serum creatinine,
      x18 = haemoglobin, x19 = packed cell volume.
```

If p-value ($P > |z|$) is less than 0.05, we reject the null hypothesis. In other words, there is a statistically significant relationship between the combination the the predictors and the output variable. According our analysis, the most important features show the p values less or equal to $0.1 (\leq 0.1)$, in others words the most important features are those who are significant at 10%, 5% and 1%. This shows that these independent variables have a significant impact on dependent variable prediction.

Finally, the following 8 features out of 19 features are the best one for our models according to logistic regression result: **coronary artery disease, appetite, bilateral pedal edema, anemia, specific gravity, albumin, blood urea, haemoglobin**. Studies have indicated that these features either contribute to the development of CKD or are consequences of patients already having CKD:

1. Coronary artery disease (CAD) is a well-known risk factor for the development and progression of CKD. The presence of CAD in individuals with CKD can exacerbate renal dysfunction due to reduced blood flow to the kidneys.

2. Appetite is another important clinical indicator in CKD patients and can serve as a proxy for nutritional status. Decreased appetite, also known as anorexia, is commonly observed in individuals with CKD.

3. Bilateral pedal edema, characterized by swelling in both feet and ankles, is a clinical manifestation commonly associated with CKD. Bilateral pedal edema can serve as a visible sign of fluid overload and can be indicative of advanced CKD stages.

4. Anemia is a prevalent complication in CKD patients and is primarily attributed to impaired production of erythropoietin, a hormone produced by the kidneys that stimulates red blood cell production. Effective management of anemia in CKD involves monitoring haemoglobin levels which is also another one of the features.

5. Specific gravity is a measure of urine concentration and is often used as a marker to assess kidney function. In CKD, impaired renal tubular function can result in the inability to concentrate urine effectively, leading to decreased specific gravity. Monitoring specific gravity can help evaluate the kidneys' ability to concentrate and dilute urine, providing insights into the severity and progression of CKD.

6. Albumin is a protein normally present in the blood and plays a crucial role in maintaining fluid balance. In CKD, damage to the glomerular filtration barrier can result in increased urinary excretion of albumin, leading to albuminuria. Persistent albuminuria is considered a hallmark of kidney damage and is associated with the progression of CKD and increased cardiovascular risk.

7. Blood urea, specifically blood urea nitrogen (BUN), is a commonly measured parameter in assessing kidney function. In CKD, impaired kidney function leads to a decreased ability to filter and excrete urea, resulting in elevated blood urea levels.

By considering these eight features—coronary artery disease, appetite, bilateral pedal edema, anemia, specific gravity, albumin, blood urea, and haemoglobin—you have obtained valuable insights into various aspects related to Chronic kidney disease. These features provide a comprehensive view of the disease process, ranging from cardiovascular implications to nutritional status, fluid balance, renal function, and complications such as anemia. So will will uses those selected variable in the next section for model building.

Models implementation

After having found our best features, we proceed to the Model building. To do so, we choose to build these models in two ways: the first using the features from logistic regression result (only 8 features) and the second using all the features from original dataset (19 features at all). In this study, 10 machine learning algorithms namely KNN, Decision Tree Classifier, Random Forest Classifier,Ada Boost Classifier, Gradient Boosting Classifier, Stochastic Gradient Boosting, XgBoost, Cat Boost, Extra Trees Classifier, LGBM Classifier has been used to classify chronic kidney disease data.

As performance metrics evaluations, we used confusion matrix which is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. Confusion matrix incorporate 4 elements namely, True Positive or TP (The total number of observations that are normal and the model classifies them as normal), False Negative or FN (The total number of observations that are good but the model classifies them as bad), False Positive or FP (The total number of observations that are bad but the model classifies them as good), True Negative or TN (The total number of observations that are bad and the model classifies them as bad).

From the confusion matrix, the accuracy of each model is evaluated as follow :

$$A_i = \frac{TP_i + TN_i}{TP_i + TP_i + FP_i + FN_i} \tag{18}$$

where $i = 1, 2, ...10$ our different models.

From that same confusion matrix, others performance metrics like Specificity knowing as True Negative Rate (TNR), Sensitivity knowing as True Positive Rate(TPR), False Positive Rate(FPR) and False Negative Rate (FNR) can be derived as illustrated below:

$$TNR = \frac{TN}{TN + FP} \tag{19}$$

$$FPR = 1 - TNR \tag{20}$$

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FNR = \frac{FN}{FN + TP} \tag{22}$$

It should be noted that TNR + FPR should be equal to 1 and the same for TPR + FNR.

Specificity and Sensitivity plays a crucial role in deriving ROC curve. Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, we assume $p < 0.5$ since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of $p > 0.5$. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

Results analysis

In this section, we will show some findings of this study. As mentioned earlier, we have implemented a total of ten (10) different algorithms. First, we trained these models with the eight variables resulting from the logistic regression and second, we trained these same models with all nineteen (19) variables of the original data-set. We used the confusion matrix as a performance evaluation metric because of the use of the classification models. From this confusion matrix we extracted the accuracy of the models as shown in table 2 with the corresponding plot in Fig 2 and the different ROC curves as well as the area under the ROC curve named AUC as show in table 3 as well as the corresponding plot line in Fig 3.

Table 1. Accuracy of Models.

Accuracy of Models with Logistic Regression		**Accuracy of Models without Logistic Regression**	
Model's Names	**Accuracy**	**Model's Names**	**Accuracy**
KNN	0.931818	KNN	0.840909
Decision Tree	0.986970	Decision Tree	0.986012
Random Forest	0.994848	Random Forest	1.000000
Ada Boost	0.992121	Ada Boost	0.992424
Gradient Boosting	0.987273	Gradient Boosting	0.984848
Stochastic Gradiant Boost	0.997273	Stochastic Gradiant Boost	0.984848
XgBoost	0.989697	XgBoost	0.992424
Cat Boost	0.962121	Cat Boost	0.977273
Extra Trees	0.997273	Extra Trees	1.000000
LGBM	0.997273	LGBM	0.992424

Table showing the accuracy of the models, after logistic regression (left) and before logistic regression (right).

Table 2. AUC of Models.

AUC of Models with Logistic Regression		**AUC of Models without Logistic Regression**	
Model's Names	**AUC**	**Model's Names**	**AUC**
KNN	0.970362	KNN	0.924355
Decision Tree	0.986012	Decision Tree	0.989583
Random Forest	0.999752	Random Forest	1.000000
Ada Boost	0.976845	Ada Boost	0.994048
Gradient Boosting	0.998760	Gradient Boosting	1.000000
Stochastic Gradiant Boost	0.998264	Stochastic Gradiant Boost	1.000000
XgBoost	0.997148	XgBoost	0.999752
Cat Boost	0.999008	Cat Boost	0.999504
Extra Trees	0.999504	Extra Trees	1.000000
LGBM	0.999256	LGBM	1.000000

Table showing the AUC of the models, after logistic regression (left) and before logistic regression (right).

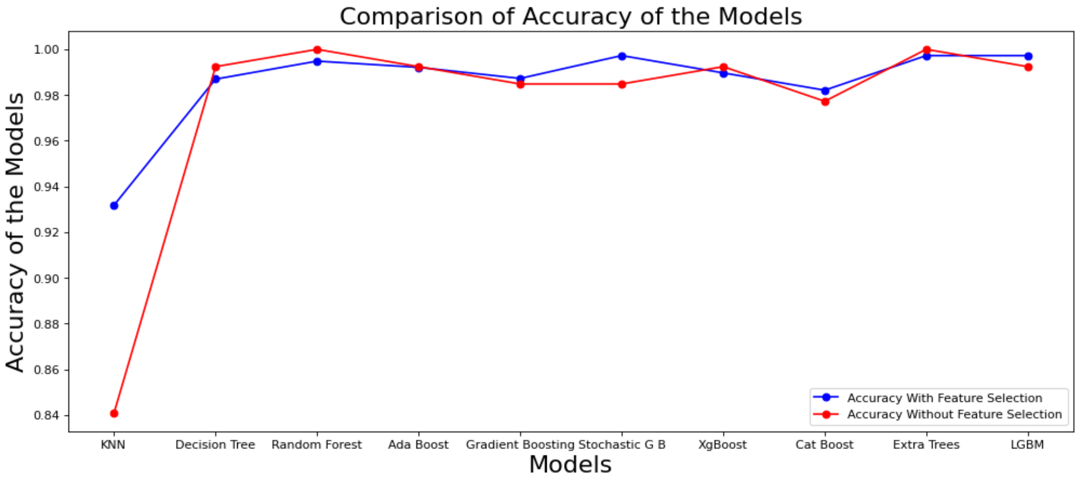

Fig 2. Comparison of model performance in terms of accuracy criteria

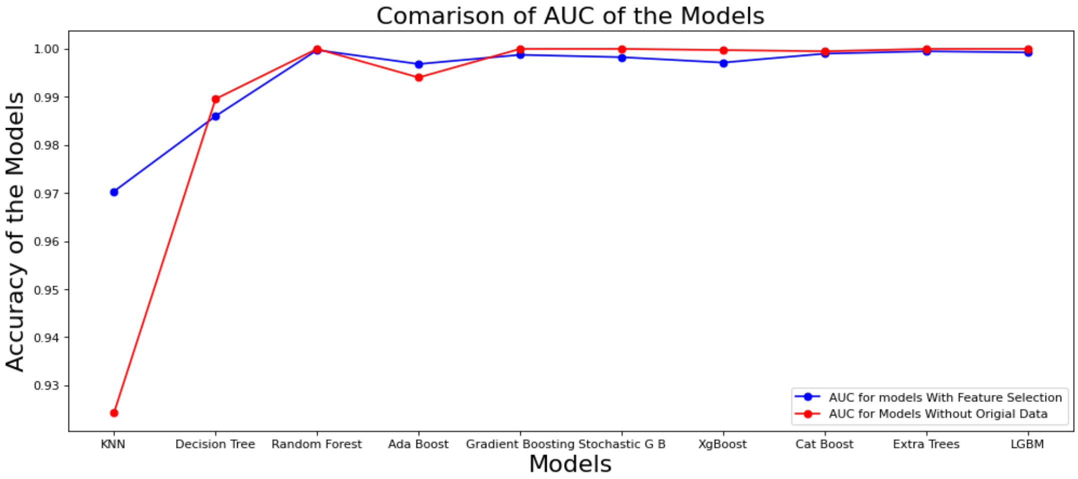

Fig 3. Comparison of model performance in terms of AUC criteria

Fig 2 shows that the proposed method (after logistic regression) is quite efficient as out of the ten models built, five of them, namely KNN, Gradient Boosting, Stochastic Gradient Boost, Cat Boost and LGBM, give higher accuracy than the models before logistic regression. It should be noted that the models after logistic regression are trained with only 8 variables unlike the models before logistic regression which are trained with all 19 variables but the results show that 5 constructed models offer superior accuracy which shows the power of logistic regression as a feature selection. We found a slight case of over-fitting for the models before logistic regression especially for Random Forest and Extra tree which gave an accuracy of 100% but logistic regression corrected this and gave accuracy of 0.999752 and 0.999504 respectively after logistic regression. Fig 3 also shows that different AUCs of the models before and after logistic regression are almost the same overall.

Conclusion

The aims of this study was to analyse Chronic Kidney Disease data using logistic regression as feature selection. Firstly, prepossessing steps to handle missing values was carried out. Secondly, we employed Logistic Regression method to identify the most relevant predictors. Our analysis revealed that among the initially considered 19 features, 8 of them demonstrated strong associations with chronic kidney disease. Ten (10) different classification Machine Learning algorithms was used to classify CKD and we found that using the logistic regression as feature selection, we can increases the accuracy of models as well as reducing model overfitting. Among the models build, five of them suit best namely: Random Forest, Ada Boost,Stochastic Gradiant Boost, Extra Trees and LGBM classifiers. Those models are well suited to classify CKD. The future work include:

- Extending the model to others type of disease.
- Extending the length of the forecast horizon taking into account the noisiness.
- Applying the Prediction Interval techniques, bootstrap strategies and other types of machine learning models including Gated Recurrent Units (GRUs), Long Short Term Memory (LSTM) and Transformers to CKD prediction.

Author Contributions

Methodology : Josué Nguinabé, Naushad Mamode Khan
Literature review : Khadiime Jhumka
Data Collection : Muhammad Muzzammil Auzine
Discussion : Maleika Heenaye-Mamode Khan, Zahra Mungloo-Dilmohamud.
Medical Interpretation : Swalay Aboo Fedally
Conclusion and review : Yuvraj Sunecher

References

1. Ivy JR, Bailey MA. Pressure natriuresis and the renal control of arterial blood pressure. The Journal of physiology. 2014;592(18):3955–3967.

2. Robson L. The kidney–an organ of critical importance in physiology. The Journal of physiology. 2014;592(Pt 18):3953.

3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394–424.

4. Alelign T, Petros B. Kidney stone disease: an update on current concepts. Advances in urology. 2018;2018.

5. Elshahat S, Cockwell P, Maxwell AP, Griffin M, O'Brien T, O'Neill C. The impact of chronic kidney disease on developed countries from a health economics perspective: a systematic scoping review. PloS one. 2020;15(3):e0230512.

6. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. Jama. 2019;322(13):1294–1304.

7. Jha V, Wang AYM, Wang H. The impact of CKD identification in large countries: the burden of illness. Nephrology Dialysis Transplantation. 2012;27(suppl_3):iii32–iii38.

8. Waljee AK, Weinheimer-Haus EM, Abubakar A, Ngugi AK, Siwo GH, Kwakye G, et al. Artificial intelligence and machine learning for early detection and diagnosis of colorectal cancer in sub-Saharan Africa. Gut. 2022;71(7):1259–1265.

9. Jones O, Matin R, van der Schaar M, Bhayankaram KP, Ranmuthu C, Islam M, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. The Lancet Digital Health. 2022;4(6):e466–e476.

10. Diogo VS, Ferreira HA, Prata D, Initiative ADN. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. Alzheimer's Research & Therapy. 2022;14(1):107.

11. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics. 2023;14:100189.

12. Pinto A, Ferreira D, Neto C, Abelha A, Machado J. Data mining to predict early stage chronic kidney disease. Procedia Computer Science. 2020;177:562–567.

13. Balakrishnan S, et al. Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset. Procedia Computer Science. 2020;171:1660–1669.

14. Singh V, Asari VK, Rajasekaran R. A deep neural network for early detection and prediction of chronic kidney disease. Diagnostics. 2022;12(1):116.

15. Misir R, Mitra M, Samanta RK. A reduced set of features for chronic kidney disease prediction. Journal of pathology informatics. 2017;8(1):24.

16. Sawhney R, Malik A, Sharma S, Narayan V. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decision Analytics Journal. 2023;6:100169. doi:https://doi.org/10.1016/j.dajour.2023.100169.

17. Zhou H, Si Y, Sun J, Deng J, Yang L, Tang Y, et al. Effectiveness of functional magnetic resonance imaging for early identification of chronic kidney disease: A systematic review and network meta-analysis. European Journal of Radiology. 2023;160:110694. doi:https://doi.org/10.1016/j.ejrad.2023.110694.

18. Yan P, Duan SB, Luo XQ, Zhang NY, Deng YH. Development and validation of a deep neural network–based model to predict acute kidney injury following intravenous administration of iodinated contrast media in hospitalized patients with chronic kidney disease: a multicohort analysis. Nephrology Dialysis Transplantation. 2022;38(2):352–361. doi:10.1093/ndt/gfac049.

19. Feng S, Hategeka C, Grépin KA. Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. Population Health Metrics. 2021;19(1):1–14.

20. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence. 2019;33(10):913–933.

21. Sania A, Pini N, Nelson M, Myers MM, Shuffrey LC, Lucchini M, et al. The K nearest neighbor algorithm for imputation of missing longitudinal prenatal alcohol data. Available at SSRN 4065215. 2021;.

22. Medium. Logistic Regression for Feature Selection: Selecting the Right Features for Your Model; January 01, 2023. https://medium.com/@rithpansanga/logistic-regression-for-feature-selection-selecting-the-right-features-fo

23. Vidhya A. Simple Guide to Logistic Regression in R and Python; March 30, 2023. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/.

24. Spiceworks. What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices; April 18, 2022. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/amp/.

25. Dhanabal S, Chandramathi S. A review of various k-nearest neighbor query processing techniques. International Journal of Computer Applications. 2011;31(7):14–22.

26. Fix E, Hodges J. Discriminatory analysis, nonparametric discrimination. 1951;.

27. Gama J, Fernandes R, Rocha R. Decision trees for mining data streams. Intelligent Data Analysis. 2006;10(1):23–45.

28. Quinlan JR. Induction of decision trees. Machine learning. 1986;1:81–106.

29. Nachouki M, Abou Naaj M. Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm. International Journal of Distance Education Technologies (IJDET). 2022;20(1):1–17.

30. Breiman L. Random forests. Machine learning. 2001;45:5–32.

31. Huang S. Applying the adaboost face detection algorithm to detect inattentive states. In: 2021 4th International Conference on Information Systems and Computer Aided Education; 2021. p. 1624–1628.

32. Kim SY, Upneja A. Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling. 2014;36:354–362.

33. Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: icml. vol. 96. Citeseer; 1996. p. 148–156.

34. Si S, Zhang H, Keerthi SS, Mahajan D, Dhillon IS, Hsieh CJ. Gradient boosted decision trees for high dimensional sparse output. In: International conference on machine learning. PMLR; 2017. p. 3182–3190.

35. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001; p. 1189–1232.

36. He X, Zhang H, Kan MY, Chua TS. Fast matrix factorization for online recommendation with implicit feedback. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval; 2016. p. 549–558.

37. Khan ZA, Chaudhary NI, Zubair S. Fractional stochastic gradient descent for recommender systems. Electronic Markets. 2019;29:275–285.

38. FRĠEDMAN J. Stochastic gradient boosting. Computational statistics and data analysis. 2002;.

39. Zhang R, Li B, Jiao B. Application of XGboost algorithm in bearing fault diagnosis. In: IOP Conference Series: Materials Science and Engineering. vol. 490. IOP Publishing; 2019. p. 072062.

40. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–794.

41. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. 2018;31.

42. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. Journal of big data. 2020;7(1):1–45.

43. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63:3–42.

44. Gupta S, Arango-Argoty G, Zhang L, Pruden A, Vikesland P. Identification of discriminatory antibiotic resistance genes among environmental resistomes using extremely randomized tree algorithm. Microbiome. 2019;7:1–15.

45. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.

46. Ahamed BS, Arya S, et al. LGBM classifier based technique for predicting type-2 diabetes. European Journal of Molecular & Clinical Medicine. 2021;8(3):454–467.