Detection of Chronic Kidney Disease using Machine Learning and Data Mining techniques Supported by Predictive Analysis

Abstract

Background Chronic kidney disease (CKD) poses a serious threat to global health, with high rates of morbidity and mortality and the emergence of new diseases as a result. CKD develops when a sickness or illness damages the kidney, inhibits kidney function, and interferes with the kidney's normal operation. This occurs over several months or years. Early diagnosis of CKD can help with timely treatment. Doctors can benefit greatly from machine learning models because to their quick and precise recognition capabilities.
Methods In this study the authors propose a machine learning model for the proper diagnosis of CKD. The models are implemented on 400 patients belonging to different classes. The data is hosted at the University of California Irvine (UCI) database for renal disease prediction. There were various models deployed and Random Forest (RF) and Logistic Regression (LR) proved to be most effective among them. Also numerous pre-processing methods like Outlier Analysis, Feature Selection and SMOTE were applied on dataset before implementing models.
Results The models were analysed using 5 different parameters viz. Accuracy, Precision, Recall, F1 Score, and the results demonstrate that both approaches might be successfully used in a clinical situation, with the probabilistic approach producing accuracy of 100% for (RF) and 98.75% for (LR) allowing for more specific grading of a probable prognosis of chronic kidney disease.
Conclusion The current study demonstrated how, when compared to conventional kidney disease screening, the use of prediction machines can help with early diagnosis and prioritised care for individuals with renal disease while also providing cost-saving benefits. The proposed clinical cybernetic loop would now need to be further validated, and the prediction machine would need to be improved by investigating non-linear dimensions embedding and clustering techniques and currently the survey is limited to 400 patients further more research could be done by taking into consideration more diverse patients belonging to different communities and having disparate dietary habits.

Keywords: Machine Learning, Data Mining, Chronic Kidney Disease (CKD), Synthetic Minority Oversampling Technique (SMOTE), Random Forest (RF), Logistic Regression (LR)

1. Introduction

Progressive decrease of renal function is a feature of CKD, often known as chronic kidney failure. If a glomerular filtration rate of less than 60 mL/min per 1.73 m2 for three months or more, with or without kidney damage then it is identified as [1]. Around the world, it affects more than 10% of all people, or more than 800 million people worldwide, 843.6 million people were expected to be affected by CKD in 2017. It has become one of the major factors contributing to death and suffering in the twenty-first century. Human blood is filtered by kidneys to remove wastes and extra fluids, which the body excretes in urine [2]. Advanced chronic renal disease causes your body to amass toxic amounts of fluid, electrolytes, and wastes. Progressive kidney disease signs and symptoms develop over time if kidney damage occurs gradually. Depending on how severe the loss of kidney function is, other symptoms including nausea, vomiting, and loss of appetite may also appear. Weak and exhausted, trouble sleeping, peeing, having trouble thinking clearly, cramping in the muscles, ankle and foot swelling, itchy skin, elevated blood pressure, and possible breathlessness if fluid builds up in the lungs, chest discomfort could develop if fluid builds up around the heart's lining. Kidney disease symptoms and signs are frequently vague. They can therefore also be brought on by different diseases. CKD arises when a disease or condition impairs kidney function. The kidney damage worsens over the course of months or years [3]. CKD can be brought on by the following conditions and ailments: type 1 or type 2 diabetes, elevated blood pressure persistent obstruction of the urinary tract caused by diseases like an enlarged prostate, kidney stones, and certain malignancies Pyelonephritis, a chronic kidney infection, is another name for it.

From 1990 to 2017 the mortality rate related to chronic renal disease increased by 41.5 %. CKD became a leading causes of death, it rose from 36th leading cause in 1990 to the 19th in 2013, it is determined by the number of deaths due to CKD and the life expectancy of people in different age groups at the time of their death from CKD[4]. The majority of CKD patients can manage their illness with medication and routine check-ups. Only around 1 in 50 individuals with CKD advance to renal failure. Even if your CKD is mild, you still run a higher risk of acquiring other significant conditions including cardiovascular disease. Heart

attack and stroke are among the disorders in this group that have an impact on the heart and blood vessels. Although there is no known treatment for CKD, it can be managed to lessen symptoms and prevent further progression. Depending on how bad your disease is, your treatment will vary. The primary therapies are as follows modifying your lifestyle can assist you in maintaining the best level of health, pharmaceuticals to address concurrent problems like high cholesterol and blood hypertension, in situations of advanced chronic renal disease, dialysis—a therapy that duplicates some kidney functions—may be necessary, advanced CKD may possibly require a kidney transplant. In order to keep an eye on your situation, regular check-ups will also be advised.

Machine learning, a field of research, has made it possible for computers to learn without explicit programming. It is one of the most fascinating technologies ever developed. It grants the computer the ability to learn, which, as the name suggests, makes it more like humans. There are probably a lot more locations than one would think where machine learning is currently being actively used. The term "Machine Learning" was created in 1959 by Arthur Samuel, a pioneering American in the fields of artificial intelligence and video games. He described machine learning as "the branch of research that enables computers to learn without being explicitly programmed.". It is possible to predict risk factors for kidney illnesses using a variety of symptoms. The various machine learning methods could be used as technology advances for patient assessment and prediction of renal disease [8]. The prognosis rate of the condition can be investigated, assessed, and appropriate actions can be undertaken based on the findings and conversations using various machine learning algorithms. Due to their precision and quickness in recognising patterns, machine learning models can assist clinicians in achieving this goal [9].

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. [5]. The most popular type of machine learning applied in medical research is supervised learning [6]. With the aid of labels, supervised ML algorithms use what they have learned from the past and present data to predict future events [7]. The system can generate results from input data if given enough training. In order to find flaws and modify the model based on outcomes, the ML algorithm compares the generated results with the actual and expected results. The goal of this study is examining the risk factors of CKD. Most of the time, it can assist the doctor in quickly identifying the symptoms and taking appropriate action to lessen them in the early stages. Four parameters were used for the experiment's needs. Following extensive testing, accuracy for RF and Logistics Regression was determined to be 100% and 98.75%, respectively.

2. Related Works

K. R. Anantha Padmanaban and G. Parthiban created a machine learning method for anticipating the early detection of chronic renal disease [10]. This essay primarily focuses on data from 600 clinical records gathered from a diabetes research facility in Chennai. The accuracy of 90.69% was highest for decision trees using information gain as a split parameter. And accuracy of 85.77% was achieved for Naïve Bayes. The primary output risk class provided the vulnerability of the patients to kidney disease. The results show that decision tree method adopted proves to be the most effective. The Data that is used here is old and based in a single hospital thus taking into consideration the parameters of the people of a small geographical area and thus hindering the generalization of this model.

Anusorn Charlonnan et al. described a machine learning method for anticipating CKD[11]. Here, four machine learning techniques—KNN, SVM, LR, and decision tree classifier—are investigated, and their results are contrasted in order to choose the best classifier. SVM obtained an accuracy of 98.3%, while decision tree and LR models only managed 96.55 and 94.8%, respectively. The accuracy of each class was also evaluated using sensitivity and specificity, and SVM was found to be the most effective at forecasting CKD.

Zewei Chen et al. in this paper has suggested using two internal fuzzy classifiers to diagnose CKD in patients: fuzzy rule-building expert system (FuRES) and fuzzy optimal associative memory (FOAM)[12]. The data used for the research purpose was taken from the UCI Machine learning Repository which was collected for 400 samples. The original data was combined with composite data, which was created by introducing various levels of noise. A final accuracy of 99.2% and 99.0% was achieved respectively for FuRES and FOAM classifiers.

Sai Prasad Potharaju and M. Sreedevi have proposed a framework to improve the accuracy of rule induction and Alternating Decision tree models by using SMOTE for the prediction of kidney disease patients[13]. The data used for this purpose was obtained from Apollo Hospital, Tamil Nadu. In this paper they have proposed a model in which the imbalanced data was first balanced using SMOTE and then the models were trained on the balanced data. An average accuracy of 98.73% was attained by using various rule based algorithms like J48, ADTree etc.
In order to analyse CKD Veenita Kunwar et al. has employed the data mining classification algorithms Nave Bayes and Artificial Neural Network (ANN)[14]. In the suggested method, the accuracy of the Naive Bayes model is 100% and that of the artificial neural network is 72.73% which were compared using Rapidminer toll. Because the authors' ANN model was improperly hyperparameter tuned during training, they were able to achieve this accuracy.

Here, Alvaro Sobrinho et al. has forecasted the probability of CKD using a variety of machine learning techniques[15]. Using the K-fold cross validation approach based on Weka software, comparative comparison of qualitative and quantitative data is carried out. SVM, multilayer perceptron, and k-nearest neighbour techniques all achieved accuracy levels between 76.66% and 75.00%, with J48 Decision Tree achieving the highest accuracy of 95.00%. RF achieved an accuracy level of 93.33%, Naive Bayes of 88.33%. By taking into account deep analysis methods used, the authors could have improved models' performance in terms of disease prediction. With the current methodology, specific features were identified and retrieved, resulting in a considerably higher accuracy for same RF technique and Artificial Neural Network.

Md. Ashiqul Islam et al. has given a proposed methodology for predicting the risk of CKD using six algorithms—Naive Bayes, RF, Simple LR etc.[16]. RF accuracy is 98.8858%, naive bayes accuracy is 93.9056%, and basic LR accuracy is 94.7679%. Using the computations from Decision Stump, Linear Regression Model, and Simple Linear Regression, the causes of kidney disease were observed.

Redhma S et. al, in this paper employs machine learning tools including the SVM classifier and Ant Colony Optimization (ACO) for prediction of CKD[17]. With 12 out of 24 qualities being used for prediction, the main goal of this research was to employ fewer attributes while retaining a greater level of accuracy. The attained accuracy is 96%.

S.Revathy et.al have proposed a model for the prediction of CKD using traditional datamining techniques and three models viz. Decision tree, SVM and RF[18]. The accuracies achieved for each mode are 94.16, 98.33, 99.16 respectively. And final conclusion given is that RF proves to be the best model.

S. Belina , V. J Sara, Dr.K.Kalaiselvi have suggested a method for prediction of CKD using Ant Colony Optimization(ACO) and Extreme Learning Machine(ELM)[19]. The authors aim is to attain the maximum accuracy with minimum number of features. Feature selection was done using Ant colony Optimization and classification was done using ELM, KNN and NN the accuracies achieved were 91%, 76% and 82% respectively.

The method proposed by the authors is different from the above methods because it includes various steps for preprocessing for cleaning the data viz. data visualization, data scaling,data balancing, feature selection and outlier analysis. Thus after preprocessing the data many different models were applied on the processed data. Out of these models RF Classifier and LR proved to give the best accuracies for the early prediction of CKD.

3. Proposed Methodology

The main objective of this suggested methodology is to increase the predictability of CKD. As the data obtained is real life data there are no duplicates found. The proposed methodology is divided into 4 main parts data visualization, Data pre-processing, Train test Split and model building. The architecture of the proposed methodology is shown in Figure 1.

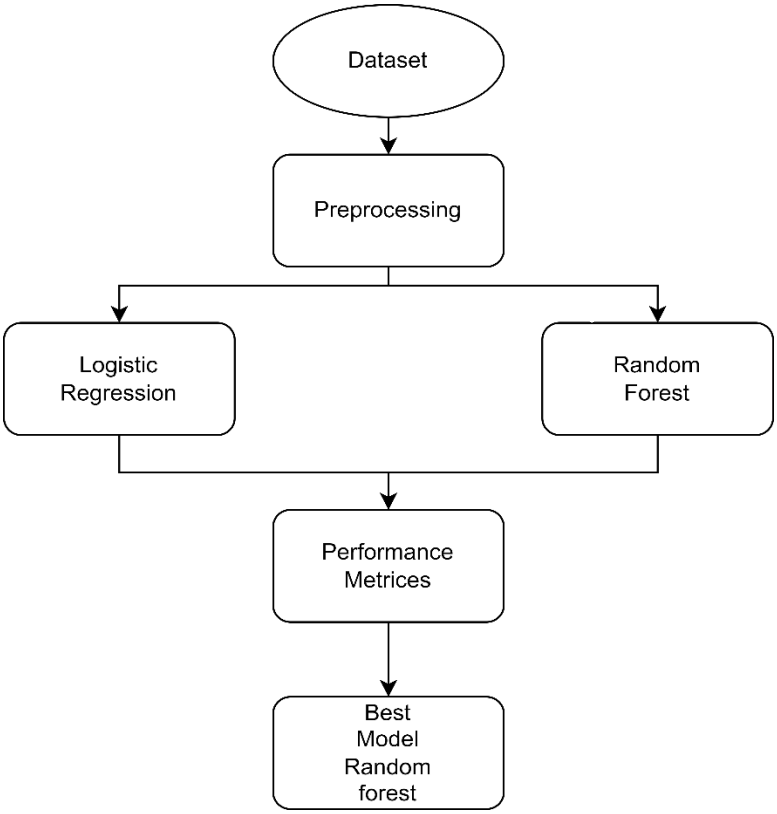

Figure 1: Flow chart of Proposed Methodology

3.1.Dataset

The dataset used for this research purpose is originally from UCI Machine Learning Repository[20]. The main purpose of this dataset is to predict whether a patient has CKD or not. The diagnostic measurements(Features) taken into consideration can be seen in the Figure 2. The class of the target variable is calculated using these features. There are a total of 400 instances taken into consideration.

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Bp      400 non-null    float64
 1   Sg      400 non-null    float64
 2   Al      400 non-null    float64
 3   Su      400 non-null    float64
 4   Rbc     400 non-null    float64
 5   Bu      400 non-null    float64
 6   Sc      400 non-null    float64
 7   Sod     400 non-null    float64
 8   Pot     400 non-null    float64
 9   Hemo    400 non-null    float64
 10  Wbcc    400 non-null    float64
 11  Rbcc    400 non-null    float64
 12  Htn     400 non-null    float64
 13  Class   400 non-null    int64
dtypes: float64(13), int64(1)
```

Figure 2: Features and their data types

3.2.Exploratory Data Analysis

Exploratory data Analysis was done by plotting histograms of all the features and the target variable. Upon observing these plots it was found that the obtained data was noisy and need to be cleaned and so the technique

of Outlier Analysis was used to do the same. And The target class was also found to be imbalanced so SMOTE was used to balance the training data. The histograms are plotted in Figure 3.

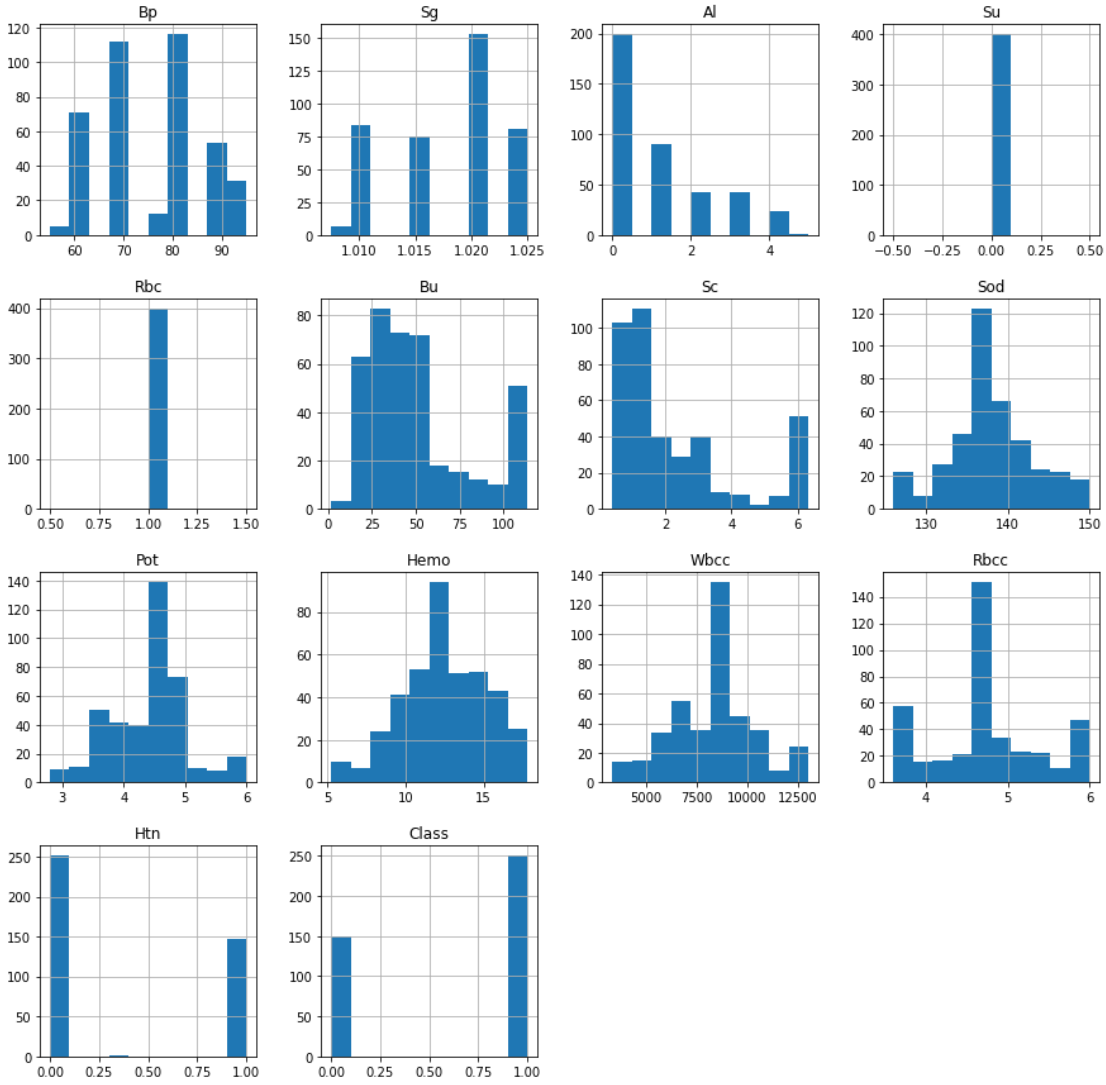

Figure 3: Histograms of all the features

3.3.Data Pre Processing

Pre-processing of data should be done before using data for predictive modelling. Additionally, various pre-processing techniques have been used to clean the data because the data obtained from the repository was noisy in nature. Thus, the part of data pre-processing has been divided into three parts viz. Outlier Analysis, Feature Selection and dealing with imbalanced data using Synthetic Minority Oversampling Technique.

3.3.1. Standard Scaler

Data normalisation is accomplished with a Standard Scaler. The data is normalised using the Standard Scaler so that the mean(μ) is 0 and the data is normally distributed. This is done because not all the features contribute equally in model fitting and such unscaled data would result in bias[21]. And thus to eliminate such a problem standardization is used before fitting the model for any machine learning algorithm. And so, all the features were normalized. The formulas for standardization using various methods are given in equation 1, equation 2 and equation 3.

$$z = \frac{x - \mu}{\sigma}$$

Equation 1: Standardization

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

Equation 2: Standardization with mean

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Equation 3: Standardization with standard deviation

3.3.2. SMOTE

Several real world problems have a requirement of generating rare values of a target variable. Similarly here, the dataset was observed to be imbalanced where the target class had uneven distribution of observations i.e. there were more instances resulting in target value 1. To deal with this imbalanced data Synthetic Minority Oversampling Technique (SMOTE) was used. SMOTE was used as it has proved to be more effective than other simple duplicating techniques for data balancing[22]. In SMOTE new instances are generated using a minority class instance selected at random and the instance belonging to the minority class is selected at random from the k nearest neighbors taken into consideration. And using any one of these neighbors a new instance is created at random[23]. This newly created instance lies on the line segment joining the two instances a and b. Here, a is the instance selected at the start and b is one of the k nearest neighbors selected at random.

Data	No of Outcomes per class	
	CKD	No CKD
Before Oversampling	198	122
After Oversampling	198	198

Table 1: Dataset comparison before and after Oversampling Using SMOTE.

Table 1 shows that before applying SMOTE the data was highly skewed and would have led to improper results. And now that the training data is properly balanced with the same number of instances belonging to both the classes there are fewer chances of miscalculations due to imbalanced data distribution. Figure 4 is a simple representation of the working of SMOTE

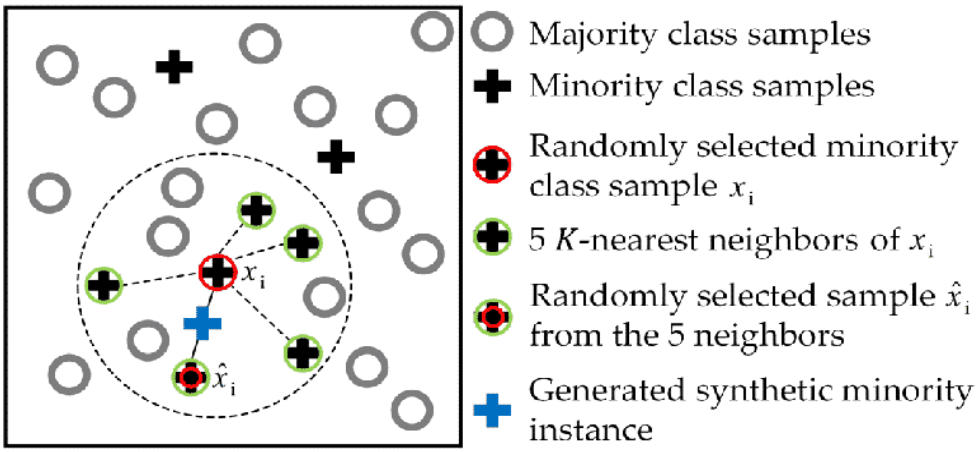

Figure 4: SMOTE Working

3.3.3. Feature Selection Using Correlation filter method

The selection of features was done using the filter method with Pearson correlation matrix[24].The value of the correlation coefficient lies between -1 and 1 and as the value approaches zero it represents a weaker correlation and exactly 0 implies no correlation. As the value approaches 1 it shows a stronger positive correlation and similarly a value closer to -1 implies stronger negative correlation. The correlation of the independent variables with the output variable 'Class' was seen using the correlation heatmap[25]. And the features having 0 correlation coefficient wrt target variable were removed. So, the features Sugar(Su) and Red Blood Cell Count(Rbcc) were dropped. And thus decreasing the total feature count taken into consideration for further steps to 11. The Pearson Correlation Heatmap is shown in Figure 5.

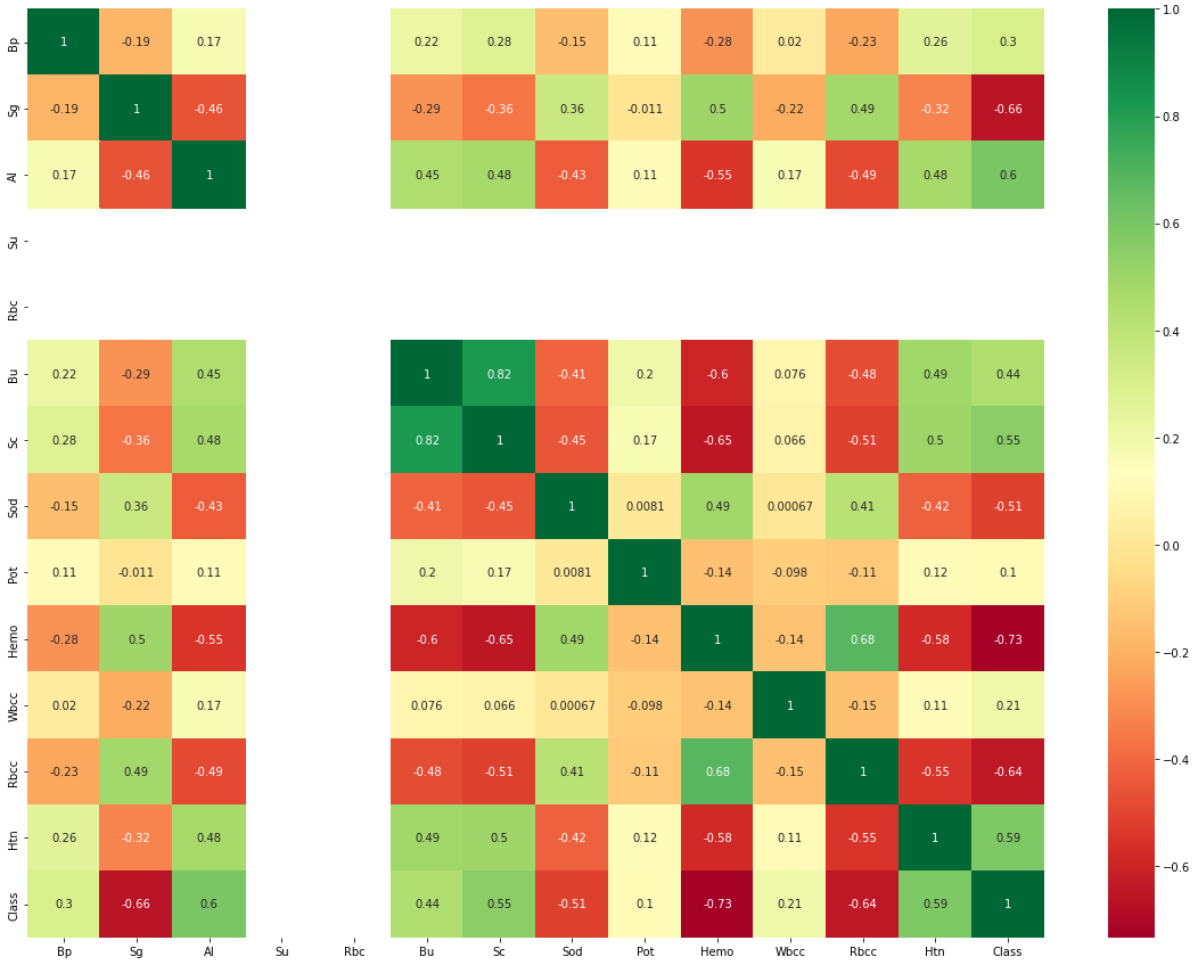

Figure 5: Pearson Correlation Heatmap

3.3.4. Outlier Analysis

An outlier is a datapoint that lies outside the overall data pattern of the features. The process of outlier analysis is used to remove these data points. There are many ways of performing outlier analysis here the method which uses quantile range and iqr is used[26]. Here, the quantile ranges were used where the lower limit was 25^{th} quantile-(1.5*iqr) and the upper limit used was 75^{th} quantile + (1.5*iqr).

3.3.5. Model Building

For classification, many models have been developed and proved useful in various scenarios. Upon experimenting it was observed that RF Classifier and LR were best suited for the prediction of CKD. Along with these other models KNN, SVM was also employed but proved to be less effective due to lower accuracies achieved.

3.3.6. Random Forest(RF)

Random Forest algorithm by Leo Brieman [27] is accurate classification method here as both random feature selection and bagging of features work together. RF has proved to be highly effective for prediction of CKD[28]. RF is an ensemble learning classification algorithm where results of various decision trees are gathered and then the final class of the instance is decided based on most frequently repeated class of multiple decision trees. A top-down splitting approach splitting approach is followed by all the decision trees. The division is done by minimum impurity method starting from the root node, and growth is halted when a node's number of records falls below a certain threshold. The Tree generated for the RF Algorithm is shown in Figure 6. Gini-index is the impurity measure that is used here. Gini Index can be calculate using Equation 4.

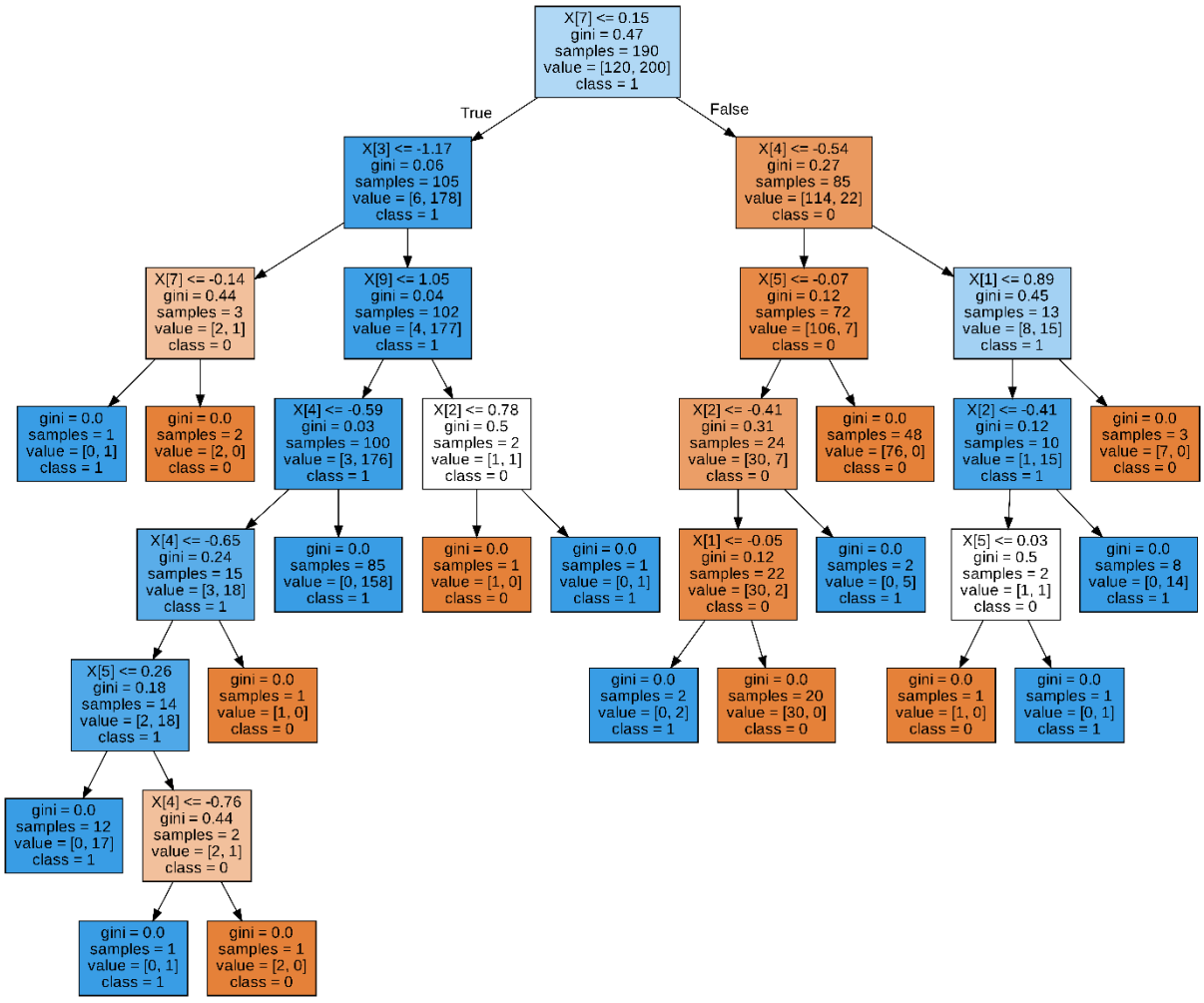

Figure 6: Tree for the Random Forest Algorithm

$$\sum_{k=1}^{K} p_k(1 - p_k)$$

Equation 4: Gini Index

Here, p_k is the number of samples in any class of size k, and K is the total classes. Entropy is another way to understand the branching of nodes in a tree[29].Entropy is calculated using Equation 5.

$$\sum_{i=1}^{c} -p_i * \log_2(p_i)$$

Equation 5: Entropy

3.3.7. Logistic Regression

Logistic Regression was primarily used for many biological science applications in the Early twentieth century. LR is used when the target variable is of categorical form. LR model takes the log of the odds in the form of a regression function of the predictors[30]. With 1 predictor X, it takes the form of the Equation 6.

$$Y = \beta_0 + \beta_1 X$$

Equation 6: Logistic Regression for 1 predictor X

Here, Y is the outcome the value of Y is 1 when the event happens and 0 when it does not. The data is fit into a linear regression function and then is passed through a Sigmoid function giving the output of a categorical target variable. The Equation 7 is the equation of Sigmoid Function.

$$sig(t) = \frac{1}{1 + e^{-t}}$$

Equation 7: Sigmoid Function

4. Performance Metrices

The algorithms were evaluated using various performance metrices viz. Accuracy, Precision Recall, F1 score, Confusion Matrix the equation for calculating these performance metrics can be seen in the Table 2.

Parameter	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F-1 Score	$2 * \frac{(Precision * Recall)}{(Precision + Recall)}$

Table 2: Formulas for calculating values of different Performance Metrices

Here,

TP=No. of observations which are positive and correctly predicted.

TN=No. of observations which are negative and correctly predicted.

FP=No. of observations which are positive and are wrongly predicted

FN=No. of observations which are negative and are wrongly predicted

5. Result and Analysis

The result obtained after following the steps mentioned in the proposed methodology is explained using two metrices which are confusion matrix and classification report. Figures 7 and 8 display the confusion matrix for the algorithms, and Table 3 displays the classification report. Here it can be observed that the number of false positive observations is value using the proposed models. This is very important in the case of medical diagnosis. The same can be seen using the Recall values of the models.

Model	Precision		Recall		F-1 Score		Accuracy
	0	1	0	1	0	1	
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00	100%
Logistic Regression	1.00	0.98	0.98	1.00	0.98	0.99	98.75%

Table 3: Classification Report for Random Forest and Logistic Regression models.

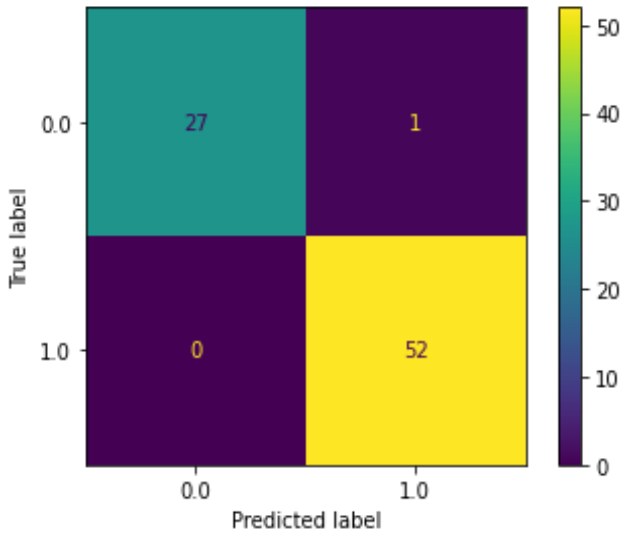

Figure 7: Confusion Matrix for classification using Logistic regression

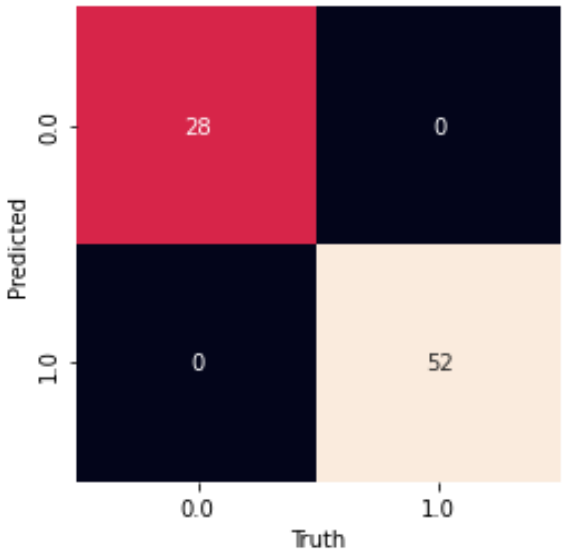

Figure 8: Confusion Matrix for classification using Random Forest.

6. Conclusion

Studies show that CKD is a factor in a significant number of fatalities each year. Therefore, it is crucial to develop methods for early diagnosis of CKD. The classification of CKD has been automated using a variety of machine learning methods. Here, the authors empirically compare Machine Learning models of Logistics Regression and RF. According to the findings of the experiment on the forecasting of CKD, the RF model behaves preferable when various assessment criteria are considered. These results make it quite clear that Logistics Regression might not be able to respond to the CKD prophecy appropriately. Consequently, experimental findings showed that RF, after rigorous testing, is a viable method for predicting CKD.

7. References

[1] Alebiosu, Olutayo & Ayodele, Olugbenga. (2005). The global burden of chronic kidney disease and the way forward. Ethnicity & disease. 15. 418-23.

[2] Csaba P. Kovesdy, Epidemiology of chronic kidney disease: an update 2022, Kidney International Supplements, Volume 12, Issue 1, 2022, Pages 7-11,ISSN 2157-1716, https://doi.org/10.1016/j.kisu.2021.11.003

[3] https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521#:~:text=Chronic%20kidney%20disease%20occurs%20when,High%20blood%20pressure

[4] Epidemiology of chronic kidney disease: an update 2022 Kovesdy, Csaba P. (Kovesdy, 2022) Kidney International Supplements, Volume 12, Issue 1, 7 – 11

[5] Alazzam, M. B., Hamad, A. A., & AlGhamdi, A. S. (2021). Dynamic mathematical models' system and synchronization. *Mathematical Problems in Engineering*, *2021*.

https://www.hindawi.com/journals/mpe/2021/6842071/

[6] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[7] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945-949, doi: 10.1109/ICCONS.2018.8663155.

[8] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, 2018, pp. 1-4, doi: 10.1109/ICAICT.2018.8747140.

[9] Ye, H., Chen, Y., Ye, P. *et al.* Nomogram predicting the risk of three-year chronic kidney disease adverse outcomes among East Asian patients with CKD. *BMC Nephrol* 22, 322 (2021). https://doi.org/10.1186/s12882-021-02496-7

[10] K. R. Anantha Padmanaban, G. Parthiban (2016). Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. *Indian Journal of Science and Technology*, 1-5. DOI: 10.17485/ijst/2016/v9i29/93880.

[11] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," *2016 Management and Innovation Technology International Conference (MITicon)*, 2016, pp. MIT-80-MIT-83, doi: 10.1109/MITICON.2016.8025242.

[12] Zewei Chen, Zhuoyong Zhang, Ruohua Zhu, Yuhong Xiang, Peter B. Harrington, Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers, Chemometrics and Intelligent Laboratory Systems, Volume 153, 2016, Pages 140-145, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2016.03.004.

[13] Sai Prasad Potharaju, M. S. (2016). An Improved Prediction of Kidney Disease using SMOTE. *Indian Journal of Science and Technology*, 1-7. doi:10.17485/ijst/2016/v9i31/95634

[14] V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016, pp. 300-305, doi: 10.1109/CONFLUENCE.2016.7508132.

[15] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro and A. Perkusich, "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," in IEEE Access, vol. 8, pp. 25407-25419, 2020, doi: 10.1109/ACCESS.2020.2971208.

[16] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha and S. Hossain, "Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 952-957, doi: 10.1109/ICISS49785.2020.9315878.

[17] Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, Chronic Kidney Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020),

[18] Ramesh, Revathy. (2020). Chronic Kidney Disease Prediction using Machine Learning Models. International Journal of Engineering and Advanced Technology. 9. 6364. 10.35940/ijeat.A2213.109119.

[19] V.J Sara, S.Belina and Kalaiselvi, K, Ant Colony Optimization (ACO) Based Feature Selection and Extreme Learning Machine (ELM) for Chronic Kidney Disease Detection (2018). International Journal of Advanced Studies of Scientific Research, Vol. 4, No. 1, 2019, Available at SSRN: https://ssrn.com/abstract=3330020

[20] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[21] Loukas, S. (2020, May 26). *Towards Data Science*. Retrieved from Towards Data Science: https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832

[22] Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P. (2013). SMOTE for Regression. In: Correia, L., Reis, L.P., Cascalho, J. (eds) Progress in Artificial Intelligence. EPIA 2013. Lecture Notes in Computer Science(), vol 8154. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40669-0_33

[23] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. *SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research*, *16*, 321-357.

[24] Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[25] Benesty, J., Chen, J., Huang, Y., Cohen, I. (2009). Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

[26] P. S. Femi and S. Ganesh Vaidyanathan, "Comparative Study of Outlier Detection Approaches," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 366-371, doi: 10.1109/ICIRCA.2018.8597395.

[27] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

https://doi.org/10.1023/A:1010933404324

[28]Lin, K., Hu, Y., & Kong, G. (2019). Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *International journal of medical informatics*, *125*, 55-61.

https://doi.org/10.1016/j.ijmedinf.2019.02.002

[29] Shaikh, T.A., Ali, R. (2019). Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk. In: Krishna, C., Dutta, M., Kumar, R. (eds) Proceedings of 2nd International Conference on Communication, Computing and Networking. Lecture Notes in Networks and Systems, vol 46. Springer, Singapore. https://doi.org/10.1007/978-981-13-1217-5_57

[30] LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399.

https://doi.org/10.1161/CIRCULATIONAHA.106.682658