The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 138 - 143

Early Detection of Chronic Kidney Disease Using Machine Learning

[1]M. Prathyusha, [2]R. Mahinoor, [3]S. Reshma, [4]Ummy Thahareen
[1,2,3,4] UG Student, Department of Electronics and Communication Engineering,
Dr K V Subba Reddy College Of Engineering For Women, Kurnool, Andhra Pradesh, India

Abstract
The condition known as chronic kidney disease (CKD) is widespread worldwide and accounts for a significant number of deaths. With 1.2 million deaths annually, chronic kidney disease (CKD) ranks 11th on the global death toll. According to the kidney Foundation of Bangladesh, approximately 40,000 CKD patients experience kidney failure each year and several thousand die in the early stages of their lives as a result of CKD. It is challenging to use machine learning in predictive analytics for healthcare to assist physicians in selecting the most effective treatments for saving lives. The majority of the collaborative research conducted by scientists on chronic kidney diseases relied solely on statistical models, resulting in numerous development gaps for machine-learning models. In this project, we evaluated two pre-processing scenarios, combined significant characteristics of the F1 scores, and discussed the current methods and suggested improved technology based on the correlation. In addition, we provided clinical information-based machine learning techniques for anticipating chronic renal disease. The Random Forest Classifier (RFC) and the Logistic Regression (LR) are two master teaching strategies that are investigated. The components are derived from the UCI chronic kidney disease dataset, and the best regression model for the prediction is chosen by comparing the results of these models. From these two preprocessing scenarios, choosing important features and replacing missing values with mean values for each column was the most logical choice because it lets you train with more data without dropping. However, correlation produced the best results in both instances, with an accuracy of 98%. As a result, the system can be used to predict early stage CKD at a low cost, which will be helpful for developing and underdeveloped nations.

1. Introduction

The condition known as chronic kidney disease (CKD) is widespread worldwide and accounts for a significant number of deaths. With 1.2 million deaths annually, chronic kidney disease (CKD) ranks 11th on the global death toll. According to the kidney Foundation of Bangladesh, approximately 40,000 CKD patients experience kidney failure each year and several thousand die in the early stages of their lives as a result of CKD. It is challenging to use machine learning in predictive analytics for healthcare to assist physicians in selecting the most effective treatments for saving lives. The majority of the collaborative research conducted by scientists on chronic kidney diseases relied solely on statistical models, resulting in numerous development gaps for machine-learning models. In this project, we evaluated two pre-processing scenarios, combined significant characteristics of the F1 scores, and discussed the current methods and suggested improved technology based on the

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 138 - 143

correlation. In addition, we provided clinical information-based machine learning techniques for anticipating chronic renal disease.

The Random Forest Classifier (RFC) and the Logistic Regression (LR) are two master teaching strategies that are investigated. The components are derived from the UCI chronic kidney disease dataset, and the best regression model for the prediction is chosen by comparing the results of these models. From these two preprocessing scenarios, choosing important features and replacing missing values with mean values for each column was the most logical choice because it lets you train with more data without dropping. However, correlation produced the best results in both instances, with an accuracy of 98%. As a result, the system can be used to predict early stage CKD at a low cost, which will be beneficial to developing and less developed nations.

A type of machine learning algorithm known as semi-supervised learning sits somewhere in between supervised and unsupervised machine learning. It employs a mix of labeled and unlabeled datasets during the training phase and serves as the middle ground between Supervised (With Labeled Training Data) and Unsupervised (Without Labeled Training Data) algorithms. Although semi-supervised learning is in the middle of supervised and unsupervised learning and works with data that only has a few labels, the majority of the data it works with is unlabeled. Labels are expensive, but they might only be used for business purposes. It is very different from supervised and unsupervised learning, which are both based on whether or not a label is present.

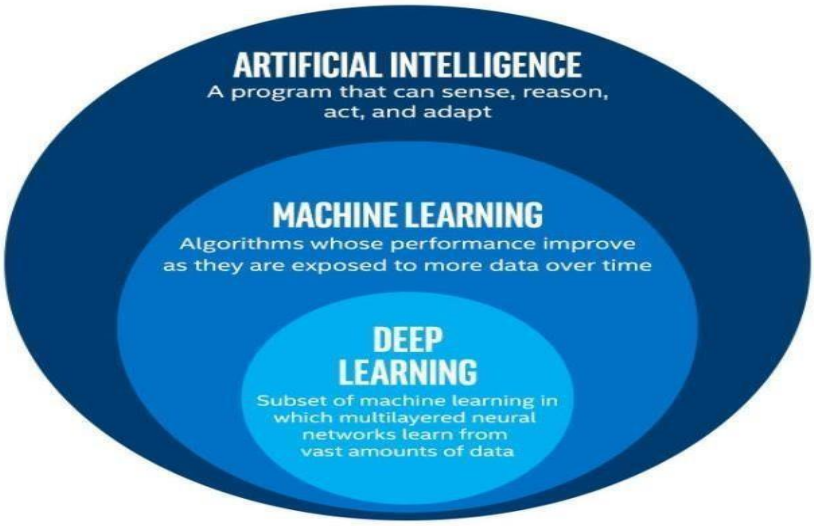

Fig.1 Machine Learning Subset of AI

2. Literature Review

A system based on the Random Forest algorithm and the Back propagation neural network (J. Snegha, 2020) was proposed. Using a supervised learning network known as a feed forward neural network, the Back Propagation algorithm outperforms the other algorithm in this comparison.

[Mohammed Elhoseny, 2019] described a CKD treatment system that combines ACO with density-based feature selection. The system selects features using wrapper methods.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 138 - 143

Using machine learning methods like K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, and Multi-Layer Perceptron Algorithm, [Baisakhi Chakraborty, 2019] was suggested for the creation of a CKD prediction system. These are put into use, and their results are compared to the results for accuracy, precision, and recall. In the end, Random Forest is selected as the system's implementation.

[Arif-Ul-Islam, 2019] proposed a method that makes use of Ant-Miner, J48 Decision Tree, and Boosting Classifiers to predict disease. Deriving rules that illustrate relationships between the characteristics of CKD and analyzing the performance of boosting algorithms for detecting CKD are the two objectives of this paper. The findings of the experiments demonstrate that the performance of AdaBoost was marginally lower than that of LogitBoost.

S.Belina V. (2018)] proposed a method for predicting CKD that makes use of an extreme learning machine and an ACO. The MATLAB tool is used for classification, and ELM has few constraints in the optimization. Under the Sigmoid additive SLFN type, this method is an improvement.

A Decision tree SVM-based machine learning system was described by Siddheshwar Tekale (2018). After comparing the two methods, it was determined that SVM produces the best results. Because its prediction procedure takes less time, doctors can analyze patients more quickly.

A system that uses the Back Propagation Neural Network algorithm for prediction was described in [Nilesh Borisagar, 2017]. Scaled conjugate, Levenberg, Bayesian regularization, and the resilient back propagation algorithm are all discussed in this section. The purpose of the implementation is to use Matlab R2013a. Scaled conjugate gradient and resilient back propagation outperform Levenberg and Bayesian regularization in terms of training time.

[Guneet Kaur, 2017] proposed a system for predicting the CKD using Data Mining Algorithms in Hadoop. They use two data mining classifiers like KNN and SVM. Here the predictive analysis is performed based upon the manually selected data columns. SVM classifier gives the best accuracy than KNN in this system.

[Neha Sharma, 2016] proposed a system in which the kidney disease of a patient is analyzed and the results are to compute automatically using the data set of the patient. Here Rule based prediction method is used. This system uses neuro-fuzzy method and obtained the outcome by mathematical computation.

[Kai-Cheng Hu, 2015] proposed a system which uses a multiple pheromone table based on ACO for clustering. Here they divided the problem into a set of several different patterns based on their features. Two pheromone tables are used here one for keeping the track of the promising information and the other to hold the details of unpromising information which in turn increases the probability of searching directions

3. Proposed System

The prediction of the outcome of a categorical dependent variable is made through the use of logistic regression. Therefore, categorical or discrete output is required. It could be true or false, yes or no, 0 or 1, etc. However, probabilities ranging from 0 to 1 are provided. The applications of logistic regression and linear regression are very similar. Logistic regression is used to solve classification issues, whereas linear regression is used to solve regression

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 138 - 143

issues. We use a logistic function in the shape of an "S," which predicts two maximum values instead of a regression line. The logistic function's curve tells you how likely something is, like whether a cell is cancerous or not or whether an animal is fat or not. A common ML method is logistic regression, which can classify new data using both discrete and continuous datasets..

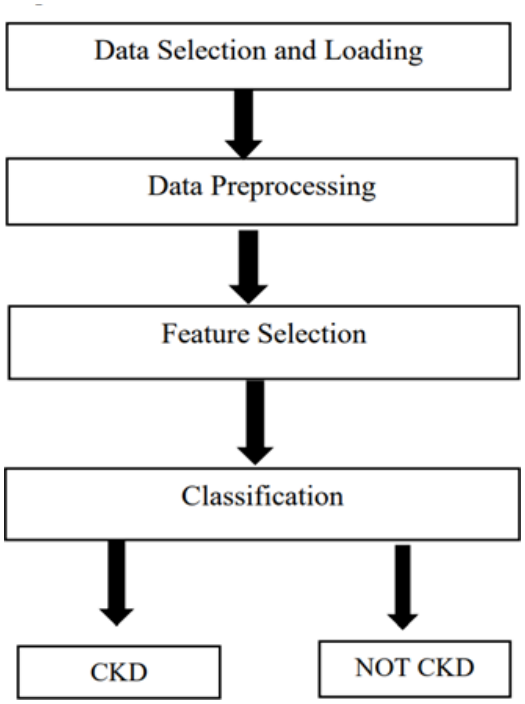

Fig.2 Flowchart

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic which function predicts two maximum values (0 or 1).

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 138 - 143

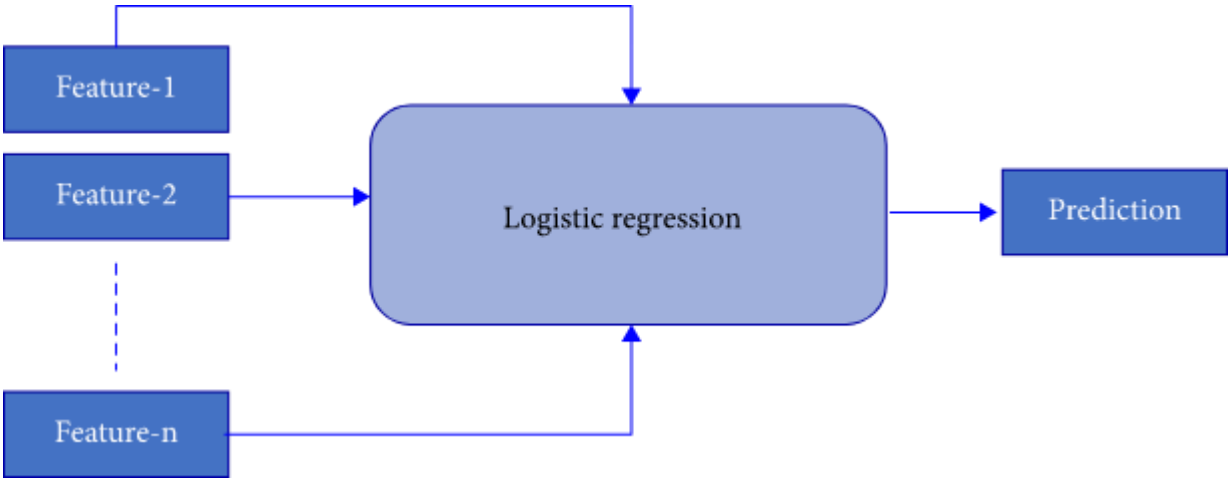

Fig.3 Block diagram of LR Classifier

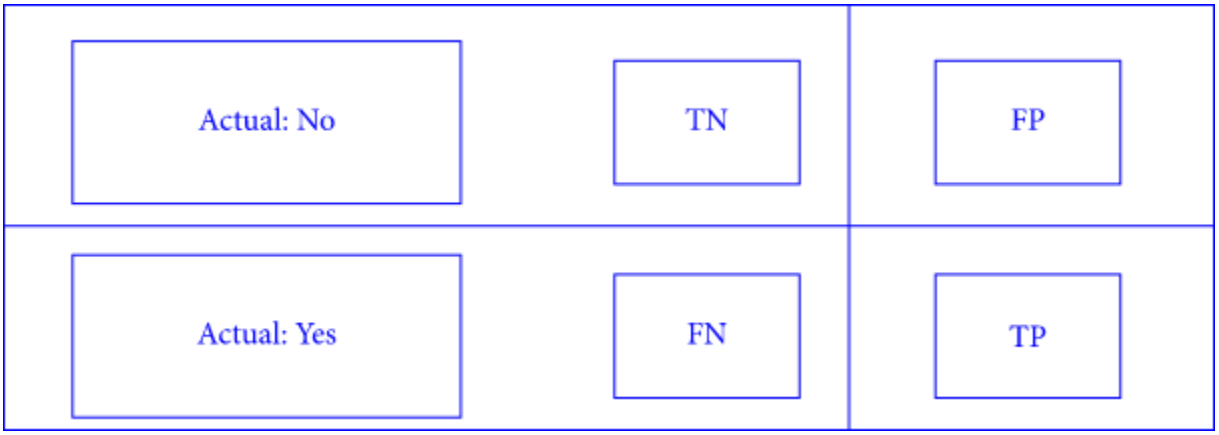

Fig.4 Output Of The Project

The classifier and word-count approaches had sensitivities of 95.4 percent and 99.8 percent, respectively. 99.8% was the specificity of both. 96.9 percent of the time, each patient was correctly categorized using the appropriate documentation. 32 (22 percent) of the 107 patients with manually verified moderate CKD did not have adequate documentation. Patients whose CKD had not been properly documented were significantly less likely to be taking renin-angiotensin system inhibitors, less likely to have their urine protein measured, and had the disease for half as long (15.1 vs. 30.7 months; p0.01) in contrast to patients who had documentation. In order to use the results of each classifier in the FLASK-based web application, they were evaluated using the Confusion Matrix and saved in pickle form.

5.Conclusion

Using machine learning algorithms and the fewest possible tests or features, the ability to detect chronic kidney disease (CKD) is the focus of this study. The Random Forest Classifier model and Logistic Regression are two machine learning classifiers that we use to achieve this goal. The relationship between variables has been investigated in an effort to reduce the number of features and eliminate redundant information. Final results were presented in the

form of accuracy scores, recall scores, precision scores, and F1 scores. We discovered that among all classification algorithms, the Random Forest Classifier Model has the highest F1 Score and the highest accuracy. According to the feature selection by correlation method, albumin and hemoglobin have the greatest impact on predicting CKD. Additionally, we discovered that hemoglobin contributes the most to the detection of CKD.

References

1. V. Jha , G. Garcia-Garcia , K. Iseki , Z. Li , S. Naicker, B. Plattner, R. Saran, A. Y. Wang,
2. C.W. Yang (2013),"Chronic kidneydisease: global dimension and perspectives", The Lancet.
3. Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 300-305)
4. L. Xun, Wu Xiaoming, Li Ningshan and Lou Tanqi, "Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. V15-332-V15-335.
5. A. Salekin and J. Stankovic, (2016) "Detection of chronic kidney disease and selecting important predictive attributes," IEEE International Conference on Healthcare Informatics.
6. D. Gupta, S. Khare, and A. Aggarwal, (2016)"A method to predict diagnostic codes for chronic diseases using machine learning techniques," 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 281–287.
7. A. Y. Al-Hyari, A. M. Al-Taee and M. A. Al-Taee, (2013) "Clinical decision support system for diagnosis and management of Chronic Renal Failure," 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, pp. 1-6.
8. Q. Zheng, G. Tasian, and Y. Fan, "Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on Ultrasound imaging data," International Symposium on Biomedical Imaging, vol. abs/1801.0, no. Isbi, pp. 1487–1490, 2018.
9. S. P. Deng, S. Cao, D.-S. Huang, and Y.-P. Wang, (2017) "Identifying Stages of Kidney Renal Cell Carcinoma by Combining Gene Expression and DNA Methylation Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 5, pp. 1147–1153.
10. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2).