IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CHRONIC KIDNEY DISEASE WITH SUPERVISED MACHINE LEARNING ALGORITHMS

Durga Shree B[1],

M.Sc. Computer Science, Nirmala College for Women, Coimbatore-641018.

Lincy Jacquline M[2],

Assistant Professor, Nirmala College for Women, Coimbatore-641018.

Abstract

The Chronic Kidney Disease (CKD) is a major health problem around the world. There are several reasons for this and it results in some serious health issues like renal failure, cardiovascular diseases and ultimately leads to early death. Recent study shows that the CKD has increased around 17% worldwide. It is found that out of 10 people 1 is affected by CKD. The early prediction of this disease helps in control through medicines. The aim of this project is to apply machine learning algorithms in the early prediction and also to find the major factor behind this. The dataset is publicly available at Kaggle, it has 400 rows and 26 features. The dataset is preprocessed (like removing null values, duplicate values, etc.) to make the dataset to be used to build the model. Logistic Regression, Decision Tree classifier and K-Nearest Neighbor algorithms are applied in the dataset to obtain the best fit model. The Logistic Regression outperforms will with the other algorithms and produces the accuracy of 99%.

Keywords: Chronic Kidney Disease, Decision Tree Classifier, K-Nearest Neighbor, Logistic Regression.

I. INTRODUCTION

When the kidneys are suffering from chronic kidney disease, or CKD, they are unable to filter blood as effectively as they should. The primary function of the kidneys is to remove excess water and waste from the circulation. CKD indicates that the body has accumulated waste. Chronic refers to the fact that the damage develops gradually over a long period of time. It is a disease that individuals all around the world are affected by. You could face a range of health issues as a result of CKD. CKD can result from a variety of illnesses, including diabetes, high blood pressure, and heart disease, to name just a few. Who develops a CKD is affected by age and gender in addition to these grave health issues. You may experience a variety of symptoms, including back pain, stomach pain, diarrhoea, fever, nosebleeds, rash, and vomiting, if one or both of your kidneys aren't functioning properly. Diabetes and high blood pressure are the 2 conditions that are most frequently associated with long-term renal impairment [1]. Hence, CKD control can be viewed as the management of these two diseases. Many patients with chronic kidney disease (CKD) do not become aware of their condition until it is too late because it frequently does not exhibit symptoms until it has proceeded to a more severe state.

Levels of CKD

Early stages of CKD

In its early stages, CKD often shows no symptoms. This is because a significant decline in kidney function can usually be tolerated by the human body. Early detection can help prevent it from advancing to a more advanced form, as can drug treatment and continuing monitoring with regular testing.

Advanced stages of CKD

CKD's last stage is kidney failure. It is also known as established renal failure or end-stage renal disease. At some point, it's likely that dialysis or a kidney transplant will be required.

CKD testing

eGFR

The eGFR value tells you how well your kidneys are able to filter the blood. Your kidneys are functioning properly if your eGFR is greater than 90. If your eGFR is less than 60, you have chronic kidney disease [1].

Urine test

The doctor also demands a urine sample to assess kidney function. The kidneys generate urine. It is a sign that one or both of your kidneys are not working properly if your urine contains blood and protein.

Blood pressure

The patient has reached the terminal stage of renal disease if their eGFR result is less than 15. There are currently only two treatments for renal failure: dialysis and kidney transplantation. Age, gender, the frequency and duration of dialysis treatments, the patient's level of physical mobility, and their mental state are all factors that affect the patient's life expectancy following dialysis[3]. If dialysis cannot be completed successfully, the doctor's sole remaining choice is kidney transplantation. But, the cost is outrageously excessive[8].

Experimental data

CKD dataset

This approach makes use of a dataset from the UCI Machine Learning Repository[11] referred to as CKD. A total of 24 features and 1 target variable are included in the CKD Dataset. It can be broken down into 2 categories, yes or no. The dataset has 25 attributes, 11 of which are numerical and 14 of which are nominal. For the purposes of training machine learning algorithms to make predictions, the entire dataset of 400 instances is utilized. Out of a total of 400 cases, 250 are classified as having CKD, and the remaining 150 are classified as having non-CKD.

II.LITERATURE REVIEW

Charleonnan et al. [1] did comparison of the predictive models such as K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree (DT) on Indians Chronic Kidney Disease (CKD) dataset in order to select best classifier for predicting chronic kidney disease. They have identified that SVM has the highest classification accuracy of 98.3% and highest sensitivity of 0.99.

Salekin and Stankovic [2] did evaluation of classifiers such as K-NN, RF and ANN on a dataset of 400. Wrapper feature selection was implemented and five features were selected for model construction in the study. The highest classification accuracy is 98% by RF and a RMSE of 0.11. S. Tekale et al. [3] worked on “Prediction of Chronic Kidney Disease Using Machine Learning Algorithms” with a dataset consisting of 400 instances and 14 features. They have used decision trees and support vector machines. The dataset has been preprocessed and the number of features has been reduced from 25 to 14. SVM is stated as a better model with an accuracy of 96.75%.

Xiao et al. [4] proposed prediction of chronic kidney disease progression using logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network and k-nearest neighbor and compared the models based on their performance. They have used 551 patients’ history data with proteinuria with 18 features and classified the

outcome as mild, moderate, severe. They have concluded that Logistic regression performed better with AUC of 0.873, sensitivity and specificity of 0.83 and 0.82, respectively.

Mohammed and Beshah [6] conducted their research on developing a self-learning knowledge-based system for diagnosis and treatment of the first three stages of chronic kidney disease using machine learning. A small amount of data has been used in this research and they have developed a prototype which enables the patient to query KBS to see the delivery of advice. They used a decision tree in order to generate the rules. The overall performance of the prototype has been stated as 91% accurate.

Priyanka et al. [5] carried out chronic kidney disease prediction through naive bayes. They have tested using other algorithms such as KNN (K-Nearest Neighbor Algorithm), SVM (Support Vector Machines), Decision tree, and ANN (Artificial Neural Network) and they have got Naïve Bayes with better accuracy of 94.6% when compared to other algorithms.

Almasoud and Ward [6] aimed in their work to test the ability of machine learning algorithms for the prediction of chronic kidney disease using a subset of features. They used Pearson correlation, ANOVA, and Cramer's V test to select predictive features. They have done modeling using LR, SVM, RF, and GB machine learning algorithms. Finally, they concluded that Gradient Boosting has the highest accuracy with an F-measure of 97.1.

Yashfi [7] proposed to predict the risk of CKD using machine learning algorithms by analyzing the data of CKD patients. Random Forest and Artificial Neural Network have been used. They have extracted 20 out of 25 features and applied RF and ANN. RF has been identified with the highest accuracy of 97.12%.

Rady and Anwar [8] carried out the comparison of Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Radial Basis Function (RBF) algorithms to predict kidney disease stages. The researchers conducted their research on a small size dataset and few numbers of features. The result of this paper shows that the Probabilistic Neural Networks algorithm gives the highest overall classification accuracy percentage of 96.7%.

Alsuhibany et al. [9] presented an ensemble of deep learning based clinical decision support systems (EDL-CDSS) for CKD diagnosis in the IoT environment. The presented technique involves Adaptive Synthetic (ADASYN) technique for outlier detection process and employed ensemble of three models, namely, deep belief network (DBN), kernel extreme learning machine (KELM), and convolutional neural network with gated recurrent unit (CNN-GRU).

Quasi-oppositional butterfly optimization algorithm (QOBOA) technique is also employed in the study for hyperparameter tuning of DBN and CNN-GRU. The researchers have concluded that the EDL-CDSS method has the capability of proficiently detecting the presence of CKD in the IoT environment.

Poonia et al. [10] employed Various machine learning algorithms, including k-nearest neighbors algorithm (KNN), artificial neural networks (ANN), support vector machines (SVM), naive bayes (NB), and Logistic Regression as well as Re-cursive Feature Elimination (RFE) and Chi-Square test feature-selection techniques. Publicly available dataset of healthy and kidney disease patients were used to build and analyze prediction models. The study found that a logistic regression-based prediction model with optimal features chosen using the Chi-Square technique had the highest accuracy of 98.75%.

Vinod [11] carried out the assessment of seven supervised machine learning algorithms namely K-Nearest Neighbor, Decision Tree, Support vector Machine, Random Forest, Neural Network, Naïve Bayes and Logistic Regression to find the most suitable model for BCD prediction based on different performance evaluation. Finally, the result showed that k-NN is the best performer on the BCD dataset with 97% accuracy.

III.METHODOLOGY

Logistic Regression Algorithm

Logistic regression [**16**] is a well-established supervised learning algorithm in the medical community. Logistic regression predicts the probability of the class output (a target categorical variable with values of Yes, No or 0, 1) using a set of independent features. Assuming that *p* is the probability of a subject being a member of the CKD class, then 1−p is the probability of a subject and is a member of the Non-CKD.

Binary outcomes are modeled using the statistical method of logistic regression, which is well known in the field. Different learning methods are used to execute logistic regression in statistical research. A variant of the neural network method was used to create the LR algorithm. This method resembles neural networks in many ways, but it is simpler to set up and use. Figure 1 shows the block diagram of LR.

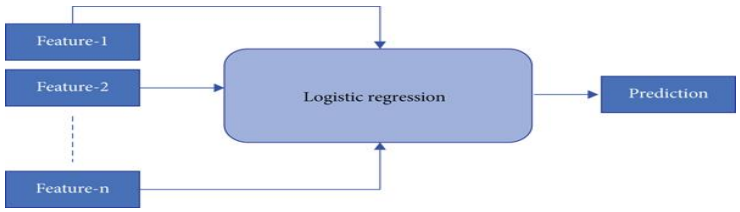

Fig.1. Block diagram of LR algorithms

Utilizing logistic regression, the output of a categorical dependent variable is predicted. So, the output must be discrete or categorical. It may be yes or no, 0 or 1, true or false, etc., but probability values between 0 and 1 are given. Logistic regression and linear regression are used in very similar ways. Classification problems are addressed with logistic regression, and regression problems are addressed using linear regression. Instead of a regression line, we use an "S" shaped logistic function that predicts two maximum values (0 or 1). The logistic function's curve indicates the probability of anything, such as whether cells are malignant or not, or if an animal is fat or not. Since it can classify new data using both discrete and continuous datasets, logistic regression is a common ML technique.

Decision Tree Algorithm

The DT method is a classification and regression technique that can be used to predict both discrete and continuous characteristics. Based on the links between input columns in a dataset, the algorithm predicts discrete characteristics. It predicts the states of a column that you identify as predictable using the values of those columns, known as states. The method specifically finds the input columns that are associated with the predicted column. The DT classifier's block diagram is shown in Figure.

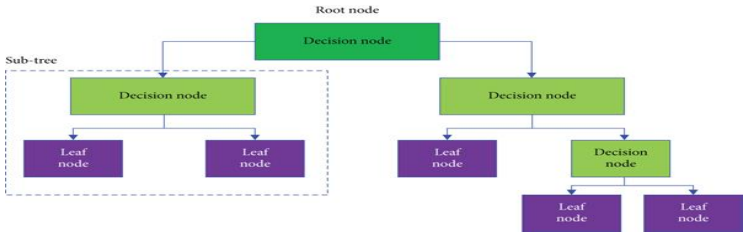

Fig.2. DT Classifier block diagram

Decision Tree (DT): It is one of the most popular supervised machine-learning algorithms that can be used for classification. Decision Tree solves the problem of machine learning by transforming the data into a tree representation through sorted feature values. Each node in a decision tree denotes features in an instance to be classified, and each leaf node represents a class label the instances belong to. This model uses a tree structure to split the dataset based on the condition as a predictive model that maps observations about an item to make a decision on the target value of instances [14]

Decision Tree pseudocode is shown in Fig.3.

Input: Data
1. Set the value of k
2. Loop: 1 to N // To get predicted class
    2.1. Calculate the distance Di (Euclidian/Cosine/Chebyshev) between data instance in training data and test data.
3. Increasingly arrange the computed distances (Di)
4. Populate the upper k results from the arranged list
5. Pick up the most frequent class from the list

Output: resultant class

Fig.3.Pseudocode of DT Algorithm

K Nearest Neighbor Algorithms

The k-Nearest Neighbors classifier measures the distance between an unlabeled instance and every other training instance [15] and designates it into the class where most of its *k* proximal neighbors originate. Figure depicts the whole KNN model's flowchart. One of the simplest ML algorithms is KNN, which uses the supervised learning approach. A new case is assigned to a category based on how closely it resembles prior categories. This is known as the KNN technique. With the KNN method, you can store all the data you have and then classify new data based on how similar it is to the old. This suggests that the KNN technique can rapidly classify new data into well-defined categories. Though it is often utilized for classification problems, the KNN method may be used for regression as well. There are no data assumptions made by the KNN technique, which is nonparametric and also called a "lazy learner algorithm," since it does not instantly learn from the training set but rather keeps and categorizes the data for later. If it receives new data, the KNN classifies it into a category that is quite close to the new data that was stored during training.

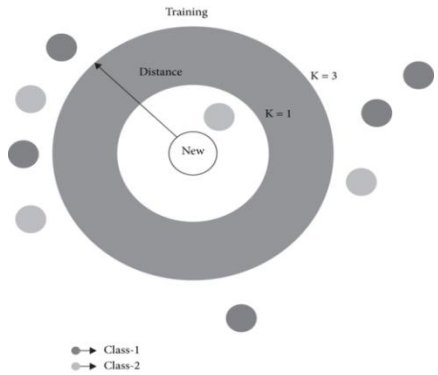

Fig.4. Working of KNN algorithm

For classification, one of the most frequently used ML methods is the -nearest neighbor classifier. The nonparametric slow learning approach, -nearest neighbor, may be used to categorize data. This classifier sorts objects according to how far they are from each other and how close they are. It prioritizes the immediate surroundings of the item above, the dissemination of essential information.

IV.PREDICTION MODEL EVALUATION

Performance evaluation is the critical step of developing an accurate machine-learning model. Prediction model shall be evaluated to ensure that the model fits the dataset and works well on unseen data. The aim of the performance evaluation is to estimate the generalization accuracy of a model on unseen/out-of-sample data. Cross-Validation (CV) is one of the performance evaluation methods for evaluating and comparing models by dividing data into partitions. The original dataset was partitioned into k equal size subsamples called folds: nine used to train a model and one used to test or validate the model. This process is repeated k times and the average performance will be taken. Tenfold cross-validation has been used in this study. Different performance evaluation metrics including accuracy, precision, recall, f1-score, sensitivity, specificity have been computed.

- True positive (TP): are the condition when both actual value and predicted value are positive.
- True negative (TN): are the condition when both the actual value of the data point and the predicted are negative.
- False positive (FP): These are the cases when the actual value of the data point was negative and the predicted is positive.
- False negative (FN): are the cases when the actual value of the data point is positive and the predicted is negative.

Accuracy

Accuracy implies the ability of the classification algorithm to predict the classes of the dataset correctly. It is the measure of how close or near the predicted value is to the actual or theoretical value. Generally, accuracy is the measure of the ratio of correct predictions over the total number of instances. The equation of accuracy is shown in Eq.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

Precision measures the true values correctly predicted from the total predicted values in the actual class. Precision quantifies the ability of the classifiers to not label a negative example as positive. The equation of precision is shown in Eq.

$$Precision = \frac{TP}{TP + FP}$$

Macro average is used for multiclass classification because it gives equal weight for each class. The equation of macro average precision is shown in Eq.

$$Precision\_macro = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TPi}{TPi + FPi}$$

Recall

Recall measures the rate of positive values that are correctly classified. Recall answers the question of what proportion of actual positives are correctly classified. The equation of recall is shown in Eq. $$Recall = \frac{TP}{TP + FN}$$

Since the macro average is used in order to compute the recall value of the models, macro average recall is calculated as follows (Eq.).

$$Recall\_macro = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TPi}{TPi + FNi}$$

F-measure

F-measure is also called F1-score is the harmonic mean between recall and precision. The equation of F1-score is shown in Eq.

$$F1_{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The macro average of F1_score is calculated as follows (Eq.).

$$F1_{score\_macro} = 2 * \frac{\text{Precision\_macro} * \text{Recall\_macro}}{\text{Precision\_macro} + \text{Recall\_macro}}$$

Sensitivity

Sensitivity is also called True Positive Rate. Sensitivity is the mean proportion of actual true positives that are correctly identified [16]. The equation of sensitivity is shown in Eq.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity

Specificity is also called True Negative Rate. It is used to measure the fraction of negative values that are correctly classified. The equation of sensitivity is shown in Eq.

$$Specificity = \frac{TN}{TN + FP}$$

RESULT AND DISCUSSION

The following figure shows some sample rows in the dataset used in this experiment.

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no

Fig.5. First 5 rows

The following figure shows the distribution of two classes' i.e. chronic kidney disease positive and negative in the bar chart representation.

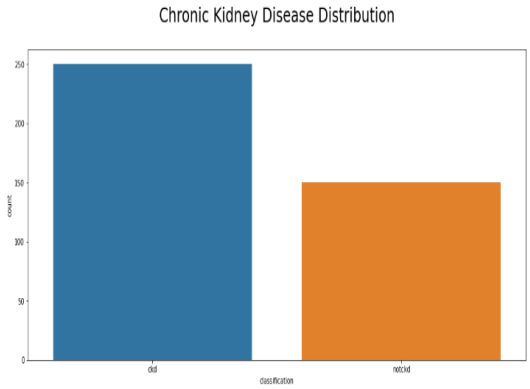

Fig.6. Distribution of two classes

Logistic Regression Algorithm

The following diagram shows the confusion matrix obtained from applying the logistic regression and the accuracy obtained is 99% which is the best one among the three algorithms taken.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	57
1	0.98	1.00	0.99	43
accuracy			0.99	100
macro avg	0.99	0.99	0.99	100
weighted avg	0.99	0.99	0.99	100

Fig.7. Confusion matrix from logistic regression

K-Nearest Neighbor Algorithm

The following is the diagrammatic representation of the confusion matrix as a result form applying the KNN algorithm for the prediction of chronic kidney disease, and the accuracy obtained here is 80%.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	57
1	0.98	1.00	0.99	43
accuracy			0.99	100
macro avg	0.99	0.99	0.99	100
weighted avg	0.99	0.99	0.99	100

Fig.8. Confusion matrix from KNN

Decision Tree Algorithm

Confusion matrix from the decision tree algorithms is shown in the following figure the accuracy is %.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	57
1	0.98	1.00	0.99	43
accuracy			0.99	100
macro avg	0.99	0.99	0.99	100
weighted avg	0.99	0.99	0.99	100

Fig.9. Confusion matrix from Decision Tree

VI. CONCLUSION & FUTURE WORK

This study used a supervised machine-learning algorithm, feature selection methods to select the best subset features to develop the models. Here the results gained were 99%, 97% and 98% for the algorithms logistic regression, K Nearest Neighbor and Decision tree respectively. It is better to see the difference in performance results using unsupervised or deep learning algorithms models. The proposed model supports the experts to make the fast decision, it is better to make it a mobile-based system that enables the experts to follow the status of the patients and help the patients to use the system to know their status.

References

[1] Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. Manag Innov Technol Int Conf MITiCON. 2016;80–83:2017.

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

[3] Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.

[4] Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.

[5] Priyanka K, Science BC. Chronic kidney disease prediction based on naive Bayes technique. 2019. p. 1653–9.

[6] Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Computer. 2019;10(8):89–96.

[7] Yashfi SY. Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms. 2020.

[8] Rady EA, Anwar AS. Informatics in Medicine Unlocked Prediction of kidney disease stages using data mining algorithms. Informatics Med. 2019;15(2018):100178.

[9].Alsuhibany SA, et al. Ensemble of deep learning based clinical decision support system for chronic kidney disease diagnosis in medical internet of things environment. Comput Intell Neurosci. 2021;3:2021.

[10] Poonia RC, et al. Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease. Healthcare. 2022;10:2.

[11] Kumar V. Evaluation of computationally intelligent techniques for breast cancer diagnosis. Neural Comput Appl. 2021;33(8):3195–208.

[12] Amirgaliyev Y. Analysis of chronic kidney disease dataset by applying machine learning methods. In: 2018 IEEE 12th International Conference Application Information Communication Technology, pp. 1–4, 2010.

[13] Kumar V. Evaluation of computationally intelligent techniques for breast cancer diagnosis. Neural Comput Appl. 2021;33(8):3195–208.

[14] Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol. 2017;48(3):128–38.:

[15] Ali, N.; Neagu, D.; Trundle, P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl. Sci. 2019, 1, 1559.

[16] Nusinovici, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.Y. Logistic regression was as good as machine learning for predicting major chronic diseases. J. Clin. Epidemiol. 2020, 122, 56–69.