International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 594-599 www.ijaem.net ISSN: 2395-5252

A Machine Learning Methodology for Predicting chronic Kidney Disease

Dr.N.Mohanapriya, ASP/CSE, Mrs.A.Thamaraiselvi, AP/CSE
Corresponding Author: Anupriya.M[1], Bhavadharani.B.S[2], Charumathi.G [3], Kousalya.S [4]
Vivekanandha College of Engineering for Women, Tiruchengode, Namakkal,

Date of Submission: 05-04-2023 Date of Acceptance: 15-04-2023

ABSTRACT
Chronic Kidney Disease (CKD) is one of the most critical Illness, with high bleakness and death rate. Intaking of more tablets for various reasons like fever, cold, head ache etc cause kidney diseases like glomerulonephritis, polycystic kidney disease, blockages in the flow of urine.These longstanding diseases of the kidney known as Chronic KidneyDisease (CKD). Since there are no prominent effects during the starting periods of CKD, patients consistently disregard to see the illness. Early detection of CKD enables patients to seek helpful treatment to improve the treatment of this sickness. The aim of this work is prediction, the process of predicting whether the patient has **CKD** or **NO CKD**. A hybrid method is proposed by using **Logistic Regression** and **K-Nearest Neighbour** algorithm, by workingwith these algorithms and get the most accurate resultsdue to its confident prediction level. The CKD data set is taken from the University of California Irvine (UCI) AI store, which has a number of attributes and characteristics.Performance of the proposed algorithm is measured by accuracy,recall,f-measure,precision and the proposed algorithm is achieved 98.6%of accuracy.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a sort of kidney illness wherein there is slow loss of kidney capability over a time of months to years. At first there is no large side effects.Later,side effects might incorporate leg swelling, feeling tired, heaving,loss of craving, and disarray. Complexities incorporate an expanded gamble of coronary illness, hypertension, bone infection, and anemia,causes of Chronic kidney disease incorporate diabetes, hypertension, glomerulonephritis, and polycystic kidney sickness.Risk factors incorporate a family background of persistent kidney illness. Conclusion is by blood tests to gauge the assessed **Glomerular Filtration Rate** (eGFR), and a pee test to quantify egg whites. Ultrasound or kidney biopsy might be performed to decide the fundamental reason. Beginning medicines might incorporate prescriptions to bring down circulatory strain, glucose, and cholesterol.

Logistic regression (LR) and K-Nearest Neighbour (KNN) algorithms are used in this proposed work to predict the disease with 98.6% of accuracy in less time. ML calculations have been a main impetus in recognition of irregularities in various physiological information, and are, with an extraordinary achievement, utilized in various characterization undertakings.

KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm.KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

LOGISTIC REGRESSION

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

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 594-599 www.ijaem.net ISSN: 2395-5252

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 function, which predicts two maximum values (0 or 1).The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

II. LITERATURE REVIEW

Q. Zou et al.,[9] proposed in Predicting diabetes mellitus with machine learning techniques**.** Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used. Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results.

Erlend Hodneland, Eirik Keilegavlen et al., [2] proposed in Vivo detection of chronic kidney disease using tissue deformation fields from dynamic mr imagingPersistent kidney illness is a serious ailment portrayed by continuous misfortune in kidney capability. Moreover, our outcomes show that ongoing picture enlistment strategies are deficient with regards to aversion to recuperate gentle changes in tissue firmness. End Picture enrollment applied to dynamic time series ought to be additionally investigated as a device for obtrusive estimations of arteriosclerosis.

GabrielR.Vásquez-Morales,SergioM. Martínez-Monterrubio et al.,[3]proposed Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoningpresents a neural network-based classifier to predict whether a person is at risk of developing chronic kidney disease (CKD). The model is trained with the demographic data and medical care information of two population groups on the one hand, people diagnosed with CKD in Colombia during 2018, and on the other, a sample of people without a diagnosis of this disease. To combat the effect of overfitting in the network, regularization was used using the dropout technique, in conjunction with early stopping.

Njoud Abdullah Almansour, Hajra Fahim Syed et al.,[4] proposed in Neural network and support vector machine for the prediction of chronic kidney disease. A comparative studyaims to assist in the prevention of Chronic Kidney Disease (CKD) by utilizing machine learning techniques to diagnose CKD at an early stage.The optimized parameters for Support Vector Machine and Artificial Neural Network were identified. Several experiments were performed using different values of the parameters for both techniques.It was found that ANN performs better with an accuracy of 96.75%.

Diego Buenaño-Fernández , David Gil et al.,[5] proposed in Application of machine learning in predicting performance for computer engineering students, a case studypresent work proposes the application of machine learning techniques to predict the final grades (FGs) of students based on their historical performance of grades. To define the project architecture, it is not recommended to use a traditional approach based on a data warehouse; rather, due to the nature of the proposed project, it will be necessary to create a documented, scalable, and flexible database that can support large indexing and data consultation by students, teachers, and educational administrators.

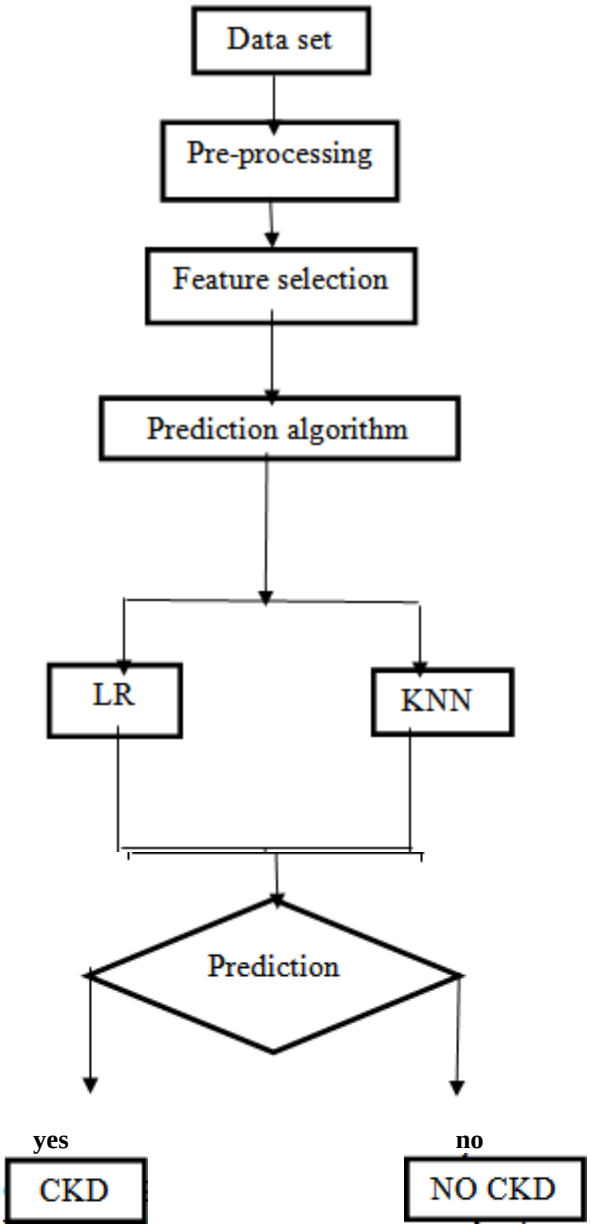
III. PROPOSED SYSTEM

The CKD dataset is given as information which comprise of various attributes. Removal of undesirable information and unknown credits are finished in preprocessing. Feature selection choice is finished. Grouping execution is finished in calculations like logistic regression and KNN algorithm. Accuracy, review, f-measure, exactness will be classified. Those boundaries will be displayed in type of graphical portrayal. They utilized picture enlistment to perceive renal morphologic changes and set up a classifier subject to brain association using huge degree CKD data, and the precision of the model on their test data. Additionally, most of the past inspects utilized the CKD educational list that was gained from the UCI artificial intelligence store. This work examines how CKD can be analyzed by utilizing AI (ML) methods In the current review, various different ML classifiers are tentatively approved to a genuine informational index, taken from the UCI AI Store, and our discoveries are contrasted and the discoveries detailed in the new writing. The outcomes are quantitatively and subjectively talked about and our discoveries uncover that the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 594-599 www.ijaem.net ISSN: 2395-5252

Calculated relapse LR classifier accomplishes the close ideal exhibitions on the distinguishing proof of CKD subjects. Subsequently, we show that ML calculations serve significant capability in analysis of CKD, with palatable strength, and our discoveries propose that LR can likewise be used for the conclusion of comparative diseases. Their assessments have achieved extraordinary results in the finding of CKD. In the above models, the mean attribution is used to fill in the missing characteristics and it depends upon the illustrative groupings of the models. In this way, their strategy couldn't be used right when the decisive outcomes of the models are dark. As a general rule, patients might miss a couple of assessments for various reasons preceding diagnosing.

IV. PREDICTION OF CHRONIC KIDNEY DISEASE:



International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 594-599 www.ijaem.net ISSN: 2395-5252

4.1 DATA PREPROCESSING

Data pre-processing is the first module of our work. In this process, missing valued columns are removed and also converting raw data to meaningful data. Each categorical (nominal) variable was coded to work with the handling. Every one of the straightout factors were changed into factors. Each example was given a free number that is from 1 to 400. There is countless missing qualities in the informational collection, and the quantity of complete occurrences is 158. As a rule, the patients could miss a few estimations in light of multiple factors prior to making a conclusion. Consequently, missing qualities will show up in the information when the demonstrative classes of tests are obscure, and a relating ascription strategy is required.

4.2 FEATURE SELECTION

Extracting feature vectors or predictors could remove variables that are neither useful for prediction nor related to response variables and thus prevent these unrelated variables the models to make an accurate prediction. Here in, we used optimal subset regression and RF to extract the variables that are most meaningful to the prediction. Optimal subset regression detects the model performance of all possible combinations of predictors and selects the best combination of variables. The combinations are ranked from left to right by the degree The vertical axis represents variables. The horizontal axis is the adjusted r-squared which represents the degree to which the combination of variables explains the response variable. To make it easy to distinguish each combination of variables, we used four colors (red, green, blue and black) to mark the selected variables.

4.3 PREDICTION

The utilization of the AI calculation like Logistic regression and KNN algorithm shows the most elevated conceivable exactness alongside the accuracy, review, f-measure.Various evaluation matrices were used for checking the performance of the classifier. For this purpose, the confusion matrix was used. It is a 2∗2 matrix due to two classes in the dataset. The confusion matrix gives two types of correct prediction of the classifier and two types of incorrect prediction of the classifier.

4.3.1 ACCURACY

Accuracy (likewise called positive prescient worth) is the small portion of applicable occurrences among the recovered cases. Classification accuracy shows the correct rate of prediction results. It computes from the confusion matrix. The classification accuracy is :

Accuracy =TP+TN/TP+TN+FP+FN*100
TP – True Positive
TN – True Negative
FP – False Positive
FN – False Negative

4.3.2 PRECISION

Precision is an important model performance evaluation matrix. It is the fraction of related instances among the total retrieved instances. It is a positive predicted value. The precision is calculated as follows:
Precision = TP/ TP + FP ∗ 100

4.3.3 RECALL

Recall is also an important model performance evaluation matrix. It is the fraction of related instances among the total number of retrieved instances. The recall is calculated as follows:
Recall = TP /TP + FN ∗ 100

ALGORITHM	PRECISION(%)	RECALL(%)	FMEASURE(%)	ACCURACY(%)
KNN	84	84.5	84	84.5
Decision Tree	80	81.9	80	81.9
K star	83	83	83	83.8
Logistic	86	86	86	86.45
SVM	80	82	80	82.9

Table 4.1 represents the respected algorithm with the precision, recall, f-measure, and accuracy.

4.3.4F-MEASURE

It is also known as F Score. F-measure is calculated so as to measure the accuracy of test. It is calculated from the precision and recall by:

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 594-599 www.ijaem.net ISSN: 2395-5252

V. EXPERIMENTALSETUP AND PROCEDURE

To assess model execution exhaustively, on account of holding the example dispersion in the first information, a total informational index was separated into four subsets equitably. For the above models, every subset was used once for testing,and different subsets were used for preparing, the general outcome was taken as the last execution.

To check whether the coordinated model can work on the presentation of the part models, our outcomes show the plausibility of the proposed procedure. By the utilization of LR, accomplish preferred execution over the attribution was utilizedthrough the misconceptions investigation, LR were chosen as the part models. The LR accomplished an exactness of around 98.45 which shows most examples in the informational collection are straightly distinct.

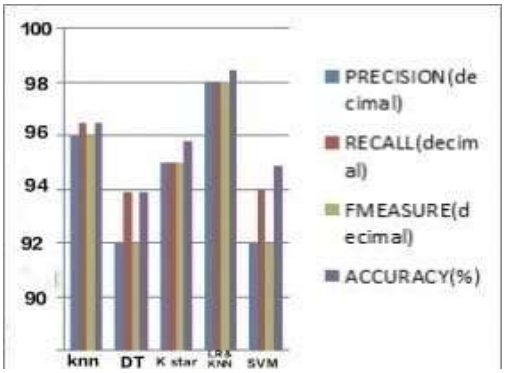


Fig 5.1 Graph showing comparison of algorithms

VI. RESULTS AND DISCUSSION

The conclusion of this technique could be reached out to additional impossible condition. While handling more perplexing information, different various calculations are endeavored to lay out models. After misjudgement test, the better calculations that produce different misjudgements are separated as part models.It is very well seen that the proposed system works on the exhibition of the models and it accomplishes practically identical with the models proposed in past test.

What's more, the CKD informational index is made out of blended factors (numeric and classification), so this techniques in view of blended information could be utilized to compute the comparability between tests, like general similitude coefficient. In this work, we utilized Euclidean distance to assess the comparability among tests, and with calculated could get a proper outcome of 98%.

VII. CONCLUSION

The proposed CKD demonstrative approach is doable regarding information attribution and tests determination. After unaided attribution of missing qualities in the informational index by utilizing strategic a attribution, the coordinated model could accomplish a good exactness.In this evaluation, we propose a calculated relapse, framework for diagnosing CKD consequently, we concluded that applying this technique to the determination of CKD would accomplish a beneficial impact. Furthermore, this approach may be pertinent to the clinical information of different sicknesses in real clinical determination. Be that as it may, during the time spent laying out the model, because of the limits of the circumstances, the accessible information tests are generally little, including just 400 examples.

Accordingly, the speculation execution of the model may be restricted. Likewise, because of there are just two classes (CKD and not CKD) of information tests in the informational collection, the model can't analyze the seriousness of CKD.In the future, an enormous number of impossible and delegate information will be gathered to prepare the model to further develop the speculation execution while empowering it to identify the seriousness of the illness. we accept that this model will be increasingly more ideal by the increment of size and proper information.

REFERENCES

[1]. M. M. Hossain et al., "Mechanical anisotropy evaluation in kidney cortex utilizing ARFI top removal: Preclinical approval and pilot in vivo clinical outcomes in kidney allografts," IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Blemish. 2020.

[2]. E. Hodneland et al., "In vivo discovery of persistent kidney illness utilizing tissue distortion fields from dynamic MR imaging," IEEE Trans. BioMed. Eng., vol. 66, no. 6, pp. 1779-1790, Jun. 2021.

[3]. G. R. Vasquez-Spirits et al., "Logical forecast of persistent renal illness in the colombian populace utilizing brain organizations and case-based thinking," IEEE Access, vol. 7, pp. 152900-152910, Oct. 2021.

[4]. N. Almansour et al., "Brain organization and backing vector machine for the expectation of persistent kidney illness: A near report," Comput. Biol. Prescription., vol. 109, pp. 101-111, Jun. 2020

[5]. M. Alloghani et al., "Uses of AI methods for computer programming learning and early expectation of understudies' exhibition," in Proc. Science, Dec. 2021, pp. 246-258.
[6]. L. Du et al., "A machine learning based approach to identify protected health information in Chinese clinical text," Int. J. Med. Inform., vol. 116, pp. 24-32, Aug. 2018
[7]. R. Abbas et al., "Classification of foetal distress and hypoxia using machine learning approaches," in Proc. Int. Conf. Intelligent Computing, Jul. 2018, pp. 767-776
[8]. M. Mahyoub, M. Randles, T. Baker and P. Yang, "Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance," in Proc. 11th Int. Conf. Developments in eSystems Engineering, Sep. 2018.
[9]. Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, Nov. 2018
[10]. Z. Gao et al., "Diagnosis of diabetic retinopathy using deep neural networks," IEEE Access, vol. 7, pp. 3360-3370, Dec. 2018.