Tuijin Jishu/Journal of Propulsion Technology
ISSN:1001-4055
Vol. 44 No.3 (2023)

Ensemble Methods Toward Prediction Kidney Disease

[1]S. Naga Raju, , [2]Erra.Nikhil, , [3]Dr.V.Chandra Sheakr Rao , [4]Dr.C.Srinivas,

[1]Associate Professor, [2]M. Tech Scholar, [3]Associate Professor, [4]Associate Professor

1,2,3,4 Department of Computer Science and Engineering

1,2,3,4 Kakatiya Institute of Technology and Science,

1,2,3,4 Warangal, India.

Abstract— Healthcare transactions produces massive amounts of data that are so overwhelming and voluminous that it cannot be managed or evaluated using standard means. The technique and creativity needed for transforming these mountains of data into data that can be used to make decisions are provided by data mining. Most of the healthcare sector is "data rich," making physical management impractical. In order to separate out useful details and make links between the attributes, data mining needs these vast volumes of data. Understanding renal disease is a challenging task that requires much insight and expertise. As one of the top causes of illness in developing nations is kidney disease, which is also a silent killer in advanced nations. The medical industry uses data mining mostly to find diseases in databases. On the Kidney illness data set, the different data mining processes Decision trees, random forest, logistic-regression, and Naive Bayes are evaluated.

Keywords- *Renal illness, machine learning, prediction, decision trees, random forests, and naive bayes.*

Introduction

Loss to or functioning of the kidneys is referred to as kidney disease, often known as renal disease. On either side of the spine, right below the ribcage, are two bean-shaped organs called the kidneys. Their primary job is to remove waste and extra fluid from the blood so that urine can be expelled. Moreover, the body's natural electrolyte balance, blood pressure, and red blood cell formation are all significantly regulated by the kidneys. Kidney disease can be caused by a wide range of things, including genetics, infections, autoimmune illnesses, drugs, and other medical issues. The most typical kidney conditions are as follows: 1. CKD is a chronic disorder that causes the kidneys to progressively lose function over time. Diabetes, high blood pressure, and other medical disorders are frequently the root of CKD. 2.Acute kidney injury (AKI) is a quick, transient decrease of kidney function that can be brought on by infections, dehydration, drug toxicity, or other conditions. 3.Glomerulonephritis is an inflammation of the kidney's glomeruli, which are tiny blood arteries that filter trash from the blood. Cysts (fluid-filled sacs) develop in the kidneys as a result of the genetic syndrome polycystic kidney disease, which eventually results in kidney

failure. Depending on the kind and severity of the problem, kidney disease symptoms might vary, but they may include weariness, swelling in the legs or ankles, high blood pressure, changes in urine output, and difficulties focusing. If kidney illness is not addressed, it can lead to renal failure, a potentially fatal condition that calls for dialysis or a kidney transplant. Depending on the underlying cause, kidney illness may be treated with medicine, dietary modifications (such as a

healthy diet and exercise), and, in some circumstances, dialysis or kidney transplantation. Early detection and management of kidney disease are important to prevent complications and preserve kidney function.

II.Literature Survey

Prediction of kidney illness is a current field of study in machine learning and healthcare. To create and assess predictive models for the early detection and management of kidney disease, several studies have been carried out. A quick review of several recent studies in this area is provided here: 1 "ML techniques for the prediction of chronic kidney disease: A systematic review" (Ahmed et al., 2020): This systematic review evaluated 24 research that predicted chronic kidney disease using techniques for machine learning. (CKD). The authors found that several algorithms, including random forest, support vector machine, and logistic regression, showed promising results in CKD prediction. 2."Prediction of acute kidney injury with machine learning algorithms: A systematic review" (Majumdar et al., 2020): This systematic review evaluated 28 studies that used machine learning algorithms to predict acute kidney injury (AKI). The authors found that several algorithms, including neural networks and decision trees, showed good performance in AKI prediction.3."Early detection of chronic kidney disease using machine learning techniques" (Nabi Zadeh et al., 2021): This study developed and evaluated a predictive model for early detection of KD using machine learning techniques. The authors found that a SVM alg with radial basis function kernel had the highest performance in KD prediction. 4."A hybrid machine learning model for the prediction of chronic kidney disease" (Aldemir et al., 2021): In this work, a blended ML model was proposed. that combines decision tree and logistic regression algorithms for KD prediction. The authors found that their model achieved higher accuracy and specificity than individual algorithms. 4."Deep learning for prediction of chronic kidney disease progression" (Nagata et al., 2021): This study developed and evaluated a deep learning model for predicting KD progression. The authors found that their model had better performance than traditional machine learning algorithms in predicting the decline in kidney function. Overall, these studies highlight the potential of machine learning algorithms and deep learning models in predicting kidney disease and identifying patients at risk of developing kidney failure. However, further studies are needed to validate these models in clinical practice and integrate them into healthcare systems for early detection and management of kidney disease.

Iii.Methodology

A. Dataset

Dataset and Attributes: The KD dataset used in this work is taken from the UCI repository. 400 patient records with 25 attributes are part of this collection. All of these 25 characteristics are the primary characteristics associated with KD illness.

1.age in numbers 2. Numbers indicating blood pressure 3 Random numerical blood sugar Number 4. Blood Urea 5.serum creatinine 6.potassium 7.red blood cell count 8.White blood cell count 9.Hemoglobin 10.Sodium 11.Sugar 12.Hypertension 13. Anemia 14. pedal edema 15. Appetite 16.Albumin 17.Specific Gravity 18.Bacteria value 19. coronary artery disease 20.Diabetes mellitus 21.Pus cell clumps 22.Pus cell count.

S.NO	ATTRIBUTE	TYPE
1	Age	Numerical
2	Blood pressure	Numerical
3	Blood glucose	Numerical
4	Blood urea	Numerical
5	Serum creatinine	Numerical
6	Potassium	Numerical

7	Red blood cell count	Numerical
9	White blood cell count	nominal
10	Hemoglobin	nominal
11	sodium	nominal
12	Sugar	nominal
13	Hypertension	nominal
14	Anemia	nominal
15	Pedal edema	nominal
16	Appetite	nominal
17	Albumin	nominal
18	Specific Gravity	nominal
19	Bacteria	nominal
20	Coronary artery disease	nominal
21	Diabetes mellitus	nominal
22	Pus cell clumps	nominal
23	Pus cell count	nominal

B. Data cleaning

Compile online raw data from open sources of CKD patients. Since the names of the attributes are missing from data downloaded from the internet, we first gave the attributes names. Missing values in the dataset, such as NAs or blank values, are removed by using the Classification function "Replace Missing Values," which replaces NAs with the mean values of that property. Data compression We have chosen 14 key attributes from the dataset's 25 total attributes that are necessary to develop a predictive model.

C. Training and Testing Dataset

The dataset is broken up into two sub datasets, each of which contains 14 properties.

D. Training data

300 out of the 400 records in the Kidney disease dataset remain important are contained in the training dataset, which is produced from the occupy a central place.

E. Testing data

100 out of the core KD dataset's 400 records make up this dataset.

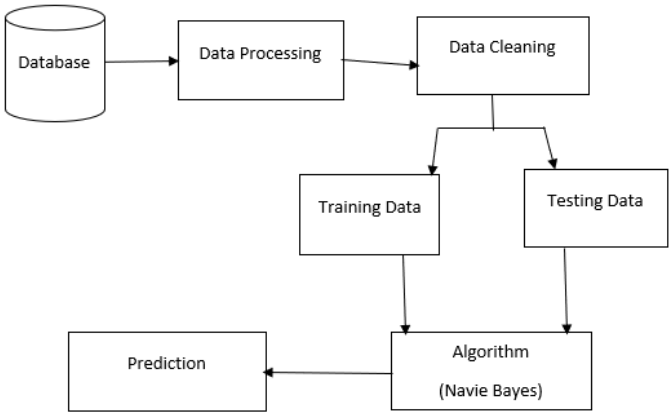

Figure: Architecture Diagram

IV. Method

After collecting the relevant data, we convert the raw it to remove any irregularities, such as missing value. To identify the features that best describe the dataset, feature selection techniques including unilateral selection and association matrices are used. We use the data to shrink the size of our attributes for testing three different categorization techniques. These algorithms include Logistic Regression, Random Tree, and Decision Trees. Then, we compare the techniques using various performance metrics. Counts of cases in this data set are 600 out of which 76.54% have KD and 24.46% do not have CKD. There are 25 qualities altogether. KD should be used for classification, NOT CKD. There are some missing values in the dataset that need to be filled up. Preprocessing and Data Mining We need to clean the dataset we downloaded from the internet because it contains empty or Blank values for a few of the attributes for a specific instance. Using the Pandas libraries "dropna" and "fillna," which drop the column or row with incomplete information and replace NAs with the mean values of such a attribute, accordingly, missing values in the dataset, such as NAs or null values, are eliminated. We significantly alter the features that will be utilized to teach the model by removing those that have a low or negative impact on the data. It is a fundamental principle of machine learning. . The initial step in model design is expected to be image enhancement and data cleaning. To acquire significant features in this paper, we shall use two suggested selection methods. The traits that were determined to be important and pertinent by both selection algorithms would then be subjected to our categorization algorithms.

A. Decision Tree

A decision tree is a tree-based model that uses binary choices to represent options and their potential outcomes, such as utility, resource costs, and chance event outcomes. It is used in data mining and machine learning. A decision tree is a tree-like structure that represents decisions and potential outcomes. Regression and classification issues can both be solved using decision trees. In a classification problem, the goal is to predict the class of an input sample. In a regression issue, the algorithm divides the input data recursively into subsets according to the values of the input variables, and chooses the optimum variable to split the data at each internal node. Decision trees provide a number of advantages over other machine learning methods, such as being easy to understand and interpret, managing numerical and categorical data, and enhancing performance of ensemble methods. However, they have drawbacks, such as overfitting and a predilection for characteristics with high cardinality or many categories. They are a popular tool for predictive modelling and decision-making.

B. Random Forest

ecision trees are combined in the Random Forest ensemble machine learning technique to increase forecast accuracy and decrease overfitting. A huge number of decision trees are created using Random Forest, and each tree is trained using a random subset of the data and features. Afterwards, the forecasts of all the trees in the forest are combined to produce the final prediction, which is commonly determined by taking a majority vote for classification or an average for regression. By generating a more varied set of trees, the random subsets of

data and features utilized in Random Forest assist in lowering overfitting. As a result, each tree has a unique perspective on the data and is less likely to commit the same errors. By lowering the variance and bias of the model, the usage of many trees also aids in improving predictive accuracy. Comparing Random Forest to other machine learning algorithms, there are various benefits. Random Forest can handle category and numerical data and is resistant to noise and missing values, but it has drawbacks such as difficulty understanding decision-making and non-linear correlations. Random Forest is a powerful machine learning method that combines different decision trees to achieve high predicted accuracy and lessen over-fitting.

C. Logistic Regression

it is a analysis used to forecast a binary result (e.g. diseased or not diseased). It functions by fitting a curve to the data that distinguishes between positive and negative cases. The binary dependent variable has just two possible values, often 0 or 1, and both continuous and categorical independent variables are acceptable. To convert the result of a linear equation into a number between 0 and 1, logistic regression employs a function known as the sigmoid function. This converted number shows the likelihood that the dependent variable will have a particular value. It is used to forecast whether a buyer will purchase a product or whether a patient will contract a specific disease. It is frequently employed in a variety of industries, including business, marketing, medicine, and social sciences.

D. Naive-bayes

Credulous Bayes is a probabilistic method utilized in order positions. It depends on the Bayes hypothesis, which expresses that the probability of the proof given the speculation, duplicated by the earlier likelihood of the speculation, decides the likelihood of a theory, (for example, a class mark) given specific proof. The calculation computes the earlier probabilities of each class mark to prepare a Gullible Bayes classifier, and works out the restrictive probabilities of each info highlight given each class name. It then picks the class name with the most elevated back likelihood to make an expectation for another info test. A fast and viable strategy can deal with inadequate and high-layered information, and can be utilized for an assortment of grouping errands, including record classification, opinion investigation, and spam separating. In any case, the reason of restrictive freedom may not be doable all the time.

Iv. Results Analysis And Discussion

Machine learning algorithms have become increasingly popular in healthcare applications as they can aid in the diagnosis and treatment of diseases. In this context, the accuracy of the algorithm becomes a critical factor, as any incorrect predictions could have significant consequences for the patient's health. Therefore, the evaluation of these algorithms is an essential step before their deployment in real-world applications. In this particular experiment, four different classification algorithms were compared, namely decision tree (DT), random forest (RF), logistic regression (LR), and naive Bayes (NB). The results showed that the naive Bayes classifier achieved the highest accuracy rate of 96.25%, making it the most suitable algorithm for this specific task. However, it is worth noting that the decision tree classifier had a significantly lower accuracy rate of 78.48%, indicating that it may not be the best choice for this application. The random forest classifier and logistic regression classifier had relatively high accuracy rates of 90.82% and 91.25%, respectively. While these rates may not be as high as that of the naive Bayes classifier, they are still quite promising and suggest that these algorithms may also be viable options in this context.

The accuracy of machine learning algorithms is a critical factor to consider when deploying them in healthcare applications. The naive Bayes classifier achieved the highest accuracy rate in this experiment, making it the most suitable algorithm for this specific task. However, it is important to consider other metrics and factors, such as precision, recall, and computational efficiency, when selecting the most appropriate algorithm for a particular application.

Algorithm	Accuracy
Decision Tree	**78.48%**
Random Forest	**90.82%**
Logistic-Regression	**91.25%**
Naive-bayes	**96.25%**

A. OUTPUT

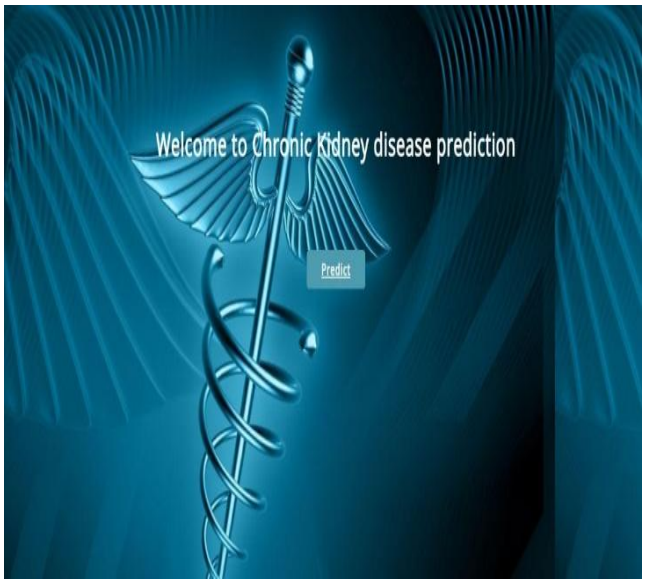

Fig 1: User interface

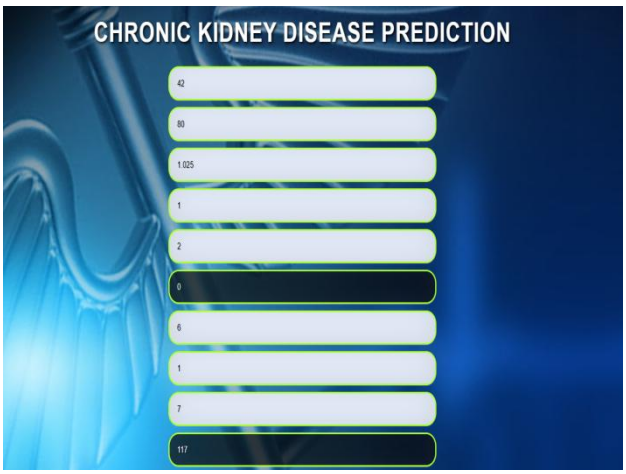

Fig 2: Entering the values

Fig 3: predicted that patient has kidney disease

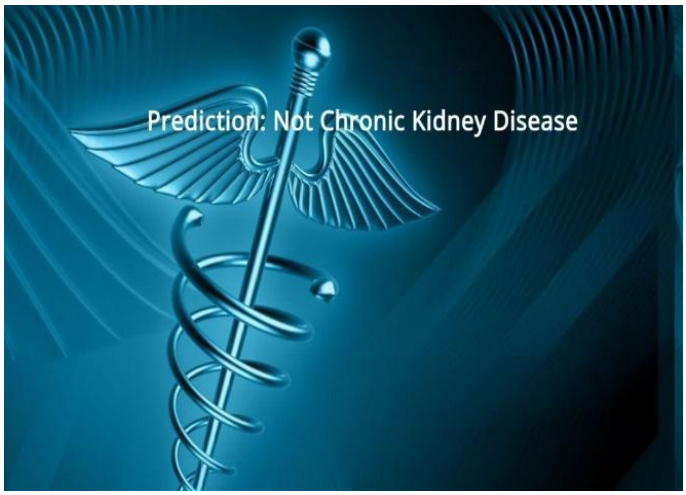

Fig 4: predicted that patient has no kidney disease

V.CONCLUSION

Using the dataset for kidney failure that we obtained from the machine learning repository at UCI, I can assess how well various ML algorithms performed. We used the filter technique of feature selection, which uses stepwise selection, correlation matrices, and further importance, to find the best structures from the data after preprocessing it. The suggested technique, which combines logistic regression, decision tree, random forest, and naive bayes, has produced accurate results of 78%, 90%,91% and 96% respectively have the highest accuracy.

References

[1] Jumani, S.Z., Ali, F., Guriro, S., Kandhro, I.A., Khan, A. and Zaidi, A., 2019. Facial Expression Recognition with Histogram of Oriented Gradients using CNN. Indian Journal of Science and Technology, 12, p.24.

[2] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn. Active clustering with ensembles for social structure extraction. In Winter Conference on Applications of Computer Vision, pages 969–976, 2014

[3] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[4] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In International Conference on Analysis and Modelling of Faces and Gestures, pages 255–263. 2005.

[5] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In Computer Vision and Pattern Recognition, pages 529–534, 2011

[6] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patchbased probabilistic image quality assessment for face selection and improved video-based face recognition. In Computer Vision and Pattern Recognition Workshops, pages 74–81, 2011.

[7] B. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In Biometrics: Theory Applications and Systems, pages 1–8, 2013

[8] N. D. Kalka, B. Maze, J. A. Duncan, K. A. O Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus Surveillance Video Benchmark. In IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2018.

[9] M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa. Cross-spectral cross-resolution video database for face recognition. In IEEE International Conference on Biometrics Theory, Applications and Systems, 2016

[10] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2879–2886.

[11] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, ‘‘On deep generative models with applications to recognition,’’ in Proc. CVPR, Jun. 2011, pp. 2857–2864.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ‘‘ImageNet: A large-scale hierarchical image database,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[13] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, ‘‘Densely connected convolutional networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2017, Jul. 2017, pp. 2261–2269.

[14] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, ‘‘Automatic recognition of facial actions in spontaneous expressions,’’ J. Multimedia, vol. 1, no. 6, pp. 22–35, 2006.

[15] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, ‘‘Toward emotion recognition in car-racing drivers: A biosignal processing approach,’’ IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, vol. 38, no. 3, pp. 502–512, May 2008.

[16] C.-C. Hsieh, M.-H. Hsih, M.-K. Jiang, Y.-M. Cheng, and E.-H. Liang, ‘‘Effective semantic features for facial expressions recognition using SVM,’’ Multimedia Tools Appl., vol. 75, no. 11, pp. 6663–6682, Jun. 2016.

[17] L. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, ‘‘Facial expression recognition from video sequences: Temporal and static modeling,’’ Comput. Vis. Image Understand., vol. 91, nos. 1–2, pp. 160–187, 2003.