Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/130101

Comparative Study on Multivariate Methods Using Chronic Kidney Disease

*K. Senthamarari Kannan and S. Anitha**

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli-12, Tamil Nadu, India

Received: 17 Mar. 2023, Revised: 20 May 2023, Accepted: 7 Jun. 2023.
Published online: 1 Jan. 2024.

Abstract: The human being is currently one of the most serious illnesses in the modern world, and accurate diagnosis is necessary as soon as possible. In this modern world, there are numerous diseases that exist. Chronic kidney disease is regarded as the most serious of these disorders in humans. There are several methods in the medical area for disease diagnosis, and the prediction criterion is also significant in the medical field for determining the consequences of the study in the future. Many statistical methods are employed in order to forecast the medical dataset and provide accurate and reliable findings. A lot of models are available in multivariate methods to predict the dataset. In this paper, the computational algorithms for detecting CKD using Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Logistic Regression (LR) are reviewed. The first, based on the association, inference for the study. Decision tree and logistic regression approaches are used to more correctly diagnose chronic renal disease based on the results of the association. Finally, the study came to the conclusion that greatest fit for forecasting chronic renal disease.

Keywords: Chronic renal disease, multivariate approaches, K-Nearest Neighbor (KNN), and Logistic Regression (LR).

1 Introduction

In vertebrates, there are two kidneys, which resemble beans and are reddish-brown in colour. They are about 12 cm (4 + 12 inches) in length in adults and are located on the left and right sides of the retroperitoneal region. Both the paired renal veins and the connected renal arteries supply them with blood. The ureter, a tube that carries urine from the kidneys to the bladder, is attached to each kidney.

CKD, a serious public health issue on a global scale, has devastating effects including kidney dysfunction, heart disease, and early mortality. According to the Global Burden of Disease Study, CKD was the 18th largest cause of death worldwide in 2010, increasing from 27th in 1990. Nearly 500 million people globally suffer from chronic kidney disease, with emerging markets, particularly South Asia and Sub-Saharan Africa, carrying a disproportionately high burden. A 2015 study found that whereas 387.5 million people lived with CKD in low- and middle-income nations, 110 million people with the disease lived in high-income countries (men 48.3 million, women 61.7 million). The kidneys filter waste and extra fluid from the blood, purifying it. As the kidneys fail, waste builds up. Slowly developing symptoms that are not unique to the condition. Some people are diagnosed with a lab test even when they have no symptoms. Taking medication helps with symptom management. Later stages may call for a transplant or mechanical blood filtering (dialysis).

*Corresponding author e-mail:statanita92@gmail.com

2 Review of Literature

This Chapter informs researchers about prior research efforts in the topic, allowing them to determine a proper analytical and methodological issue pertinent to the investigation. It enables researchers to carry out their research in the appropriate direction and to draw conclusions. As part of the inquiry, the researcher should familiarise oneself with previous research on the subject or a related problem by reading professional journals. Articles of research, approved theses, books, official records, documents, and other literature. As a result, in this section, we will analyse various related studies and summarise their key findings.

Osamah (2021) suggested using data from prolonged brain infraction disease to enhance the disease's prediction systems. Researcher discovered that a model's precision decreases when data are missing. They rebuilt the missing data with a latent component model, Their results showed so as to the ID3 algorithm worked better than the evolutionary strategy. However, the researchers found that using the K- nearest neighbours classifier, they could compare how well renal illness might be predicted using sophisticated ML techniques. The accuracy rate for the most majority of studies was about 90%, which was regarded as excellent. The uniqueness of our study is that we used a wide range of techniques and attained a 97% Precision rate. The decision tree technique and logistic regression, according to the study's findings, can be used to better correctly forecast chronic renal disease.

Ravindra (2021) utilised a simple K-means algorithm to learn about the relationship between many of these CKD markers and patient survival. He concluded that the grouping procedure forecasts dialysis patients' longevity.

Misir R, et al. (2021) used feature selection algorithms to identify a set of features that accurately predict renal problems. The smaller feature set lowers costs, saves time, and reduces uncertainty. Diabetes-related kidney impairment is a chronic and sluggish process that has serious consequences for the patient. Hyperglycemia interfere with the kidney's ability to operate properly.

3 Research Methodology

- The study's goal is to establish the relationship between gender and chronic renal disease.
- To investigate the relationship between all conceivable pairings of values.
- To determine the optimum model and forecast the accuracy of Chronic Kidney Disease.

3.1 Data sources:

Secondary data is used in this study. The information was obtained from https://www.kaggle.com/chronic-kidney-disease. The CKD dataset was used in the investigation. This dataset contains 400 rows and 14 columns.

3.2 K-Nearest Neighbor

The supervised learning methodology is used by KNN, among the mainly essential algorithms. Depending on how similar a new case is to earlier categories, it is categorized. The KNN approach is what is used in this. You can save all of your data using the KNN approach, and new data will be categorized depends on how similar it seems to the old. These show that the KNN methods can swiftly categories new data into distinct groups. Although classification problems are where the KNN algorithm is most typically used, it can also be used for regression. Since the KNN approach doesn't immediately learn from the training set but instead stores & organizes the data for soon after use, this is nonparametric and sometimes alluded to as a "lazy learner algorithm." As a result, it do not make any assumptions about the data. A category that the KNN creates for newly acquired data is pretty similar to the category it creates for newly acquired data that was saved during training.

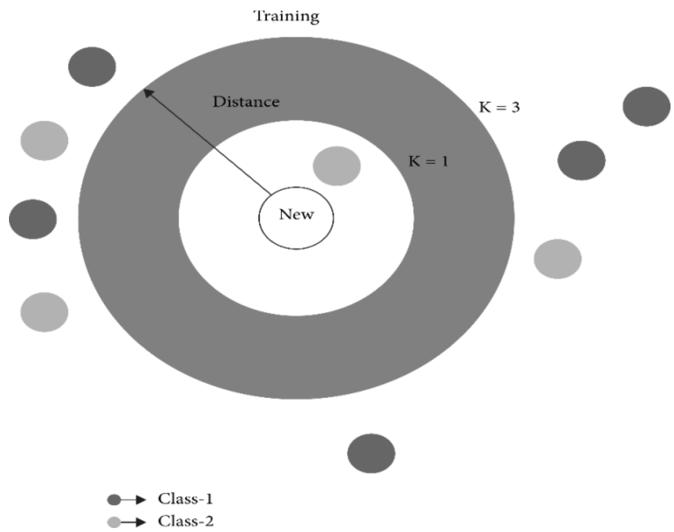

Figure 1: This figure show the flowchart K-Nearest Neighbor

The K-nearest neighbour classifier is one of the most admired ML techniques for categorization. Data can be categorised by the K-nearest Neighbor nonparametric slow learning method. This classifier divides items into divisions according to how far and how close they are to one another. It gives the delivery of essential information and the previous item's immediate surroundings priority.

3.3 Logistic Regression

The statistical technique of logistic regression, a very well technique in the field, is used to simulate binary outcomes. In statistical research, LR is carried out using a variety of learning techniques. An advanced form of the neural network approach was used to create the LR algorithm. Although this approach is simple to set up and utilise, it shares many similarities with neural networks.

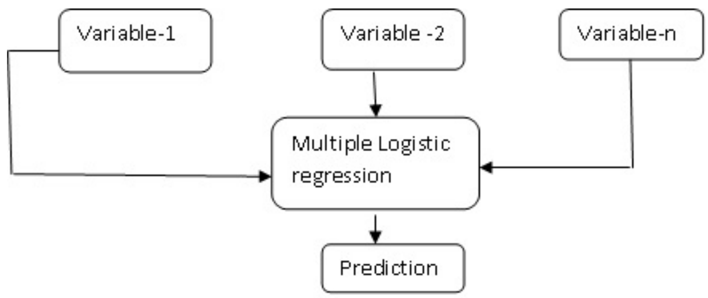

Figure 2: This figure show that block diagram of Logistic Regression.

The output of a dependent qualitative variable is forecasts by LR. As a result, the output should be quantitative or qualitative. There are provided probability values between 0 and 1, however it could also be true or false, 0 or 1, etc. Both logistic regression and linear regression have a wide range of uses. Classification problems are handled by LR & regression problems are handled by linear regression. Rather than a regression line, we use a logistic function in the shape of a "S" that predict 2 highest values 0 otherwise 1.

The logistic function curve displays the possibility of any event, including whether a cell is malignant or an animal is overweight. LR is a popular ML technique; it can describe new information using quantitative datasets.

3.4 Decision Tree:

The classification and regression method known as the decision tree approach may be used to predict quantitative features. The method uses a relationship between input columns in a data to forecast discrete qualities. Based on the values of those columns, which have been referred to as states, it predicts the states of a column you identify as predictable. The approach precisely identifies the input columns related to the anticipated column. The decision tree is simple to understand since it mimics the steps a person would take to produce a real-life conclusion. When dealing with problems impacting decision-making, it might be quite helpful. It is a good idea to think about all options for resolving a problem. Data cleansing may not be as significant as other methods.

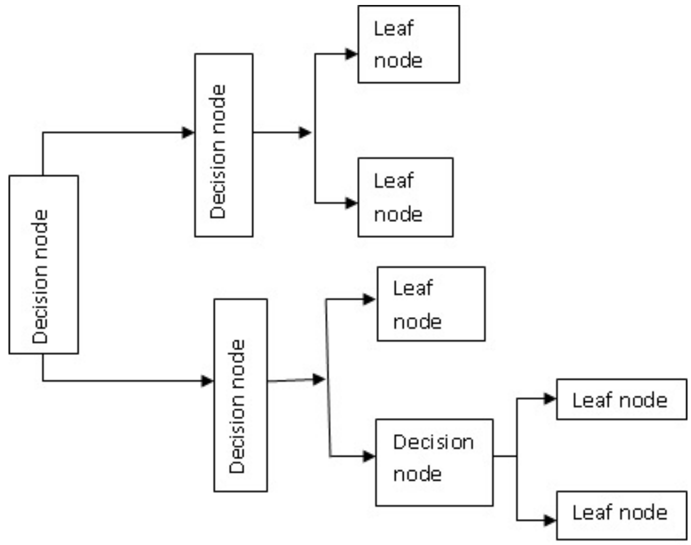

Figure 3: This figure show that Decision Tree classifier's block diagram.

3.5 Random Forest:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. It is a popular machine learning classifier for developing prediction models in many research settings. Random forests are a collection of trees which are constructed using randomly selected training datasets and random subsets of predictor variables for modeling outcomes. Random forest often gives higher accuracy compared to a single decision tree model.

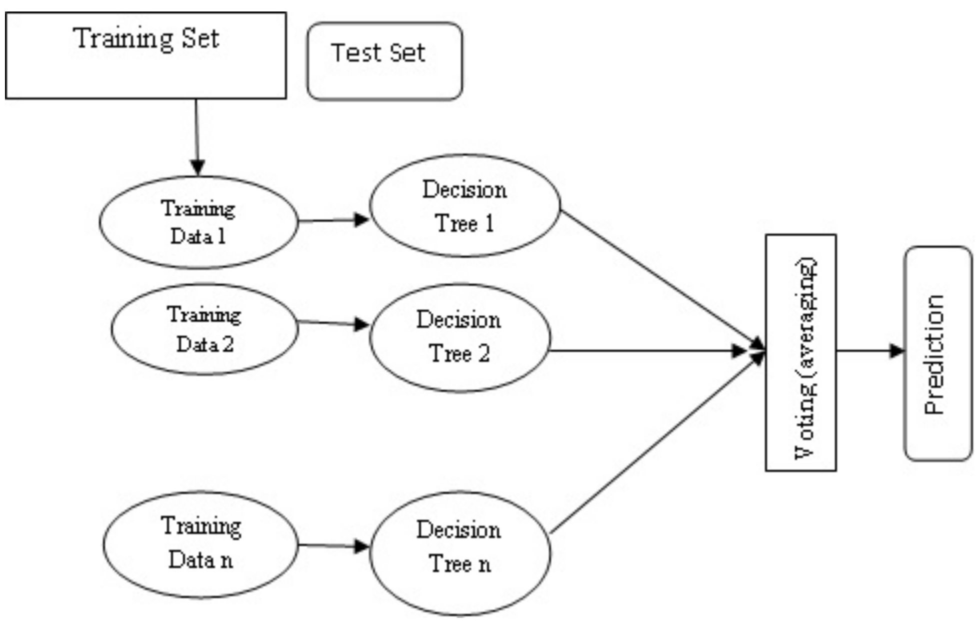

Figure 4: This figure show that Random Forest classifier's block diagram.

4. Result and Discussion

In this study, we will analyse the data, which consists of 16 variables, using multiple methods, including DT, KNN technique, and LR. Based on the accuracy of the target variable prediction, we will select the best model from among these. The variables used to produce the forecast are detailed in the following information.

4.1 Exploratory Data Analysis:

A pair plot is a pair wise relationship in a dataset. The seaborne library's pair plot module offers a high-level interface for instructive statistical visuals.

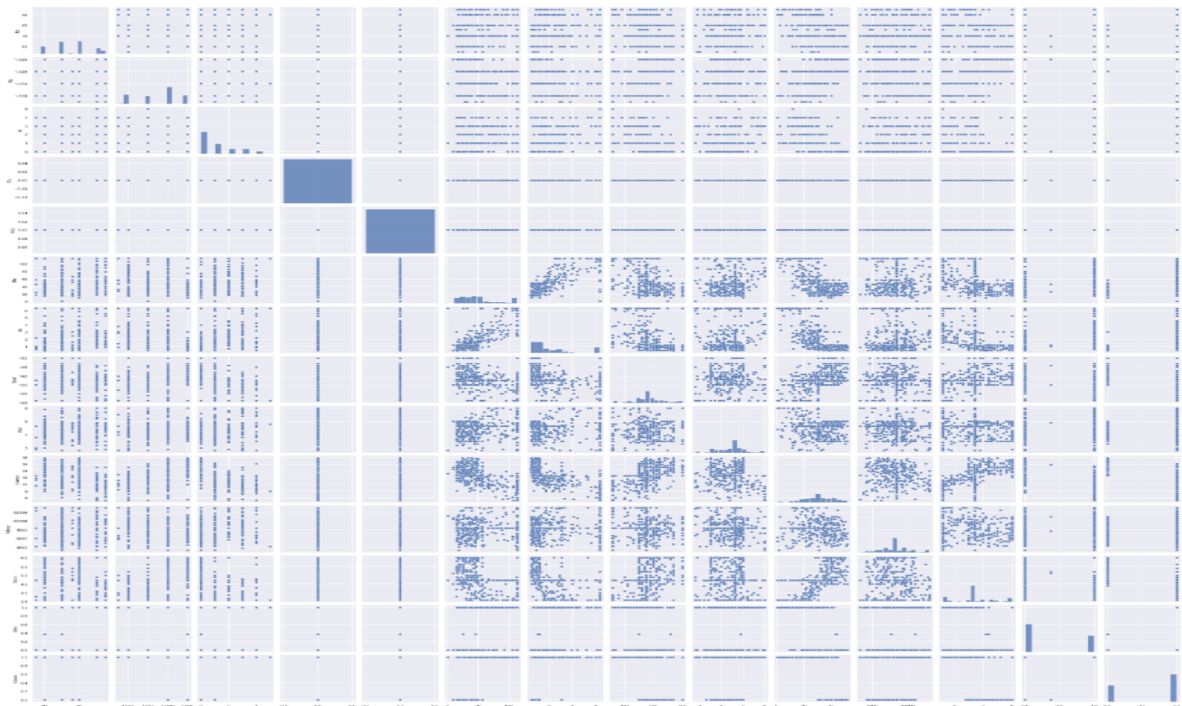

Figure 5: The pair wise associations in a dataset are as follows.

4.2 Correlation Matrix

I used the Correlation matrix analysis to figure out how the variables were related.

Table 1: The above table represents correlation matrix analysis to figure out how variables were related.

	Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
Bp	1.000000	-0.164057	0.146060	0.190277	-0.151478	0.184173	0.144469	-0.103383	0.066791	-0.279441	0.025963	-0.220827	0.268003	0.290145
Sg	-0.164057	1.000000	-0.460835	-0.292053	0.253894	-0.249263	-0.176141	0.217456	-0.063450	0.492103	-0.206880	0.443437	-0.318956	-0.659504
Al	0.146060	-0.460835	1.000000	0.262564	-0.374484	0.405035	0.229396	-0.270709	0.114484	-0.548681	0.200664	-0.454131	0.478309	0.598389
Su	0.190277	-0.292053	0.262564	1.000000	-0.092940	0.126074	0.094568	-0.053448	0.180098	-0.156875	0.159033	-0.163825	0.253179	0.294555
Rbc	-0.151478	0.253894	-0.374484	-0.092940	1.000000	-0.236270	-0.138391	0.140568	0.018164	0.280991	-0.002205	0.202298	-0.139342	-0.282642
Bu	0.184173	-0.249263	0.405035	0.126074	-0.236270	1.000000	0.581176	-0.307357	0.336954	-0.540699	0.041530	-0.465947	0.387503	0.371982
Sc	0.144469	-0.176141	0.229396	0.094568	-0.138391	0.581176	1.000000	-0.624493	0.205361	-0.342053	-0.005420	-0.323056	0.273904	0.294076
Sod	-0.103383	0.217456	-0.270709	-0.053448	0.140568	-0.307357	-0.624493	1.000000	0.067414	0.333604	0.006334	0.316883	-0.306501	-0.342268
Pot	0.066791	-0.063450	0.114484	0.180098	0.018164	0.336954	0.205361	0.067414	1.000000	-0.100612	-0.074057	-0.120418	0.057028	0.077063
Hemo	-0.279441	0.492103	-0.548681	-0.156875	0.280991	-0.540699	-0.342053	0.333604	-0.100612	1.000000	-0.153806	0.681864	-0.576932	-0.729537
Wbcc	0.025963	-0.206880	0.200664	0.159033	-0.002205	0.041530	-0.005420	0.006334	-0.074057	-0.153806	1.000000	-0.151380	0.123790	0.205266
Rbcc	-0.220827	0.443437	-0.454131	-0.163825	0.202298	-0.465947	-0.323056	0.316883	-0.120418	0.681864	-0.151380	1.000000	-0.527051	-0.590248
Htn	0.268003	-0.318956	0.478309	0.253179	-0.139342	0.387503	0.273904	-0.306501	0.057028	-0.576932	0.123790	-0.527051	1.000000	0.586340
Class	0.290145	-0.659504	0.598389	0.294555	-0.282642	0.371982	0.294076	-0.342268	0.077063	-0.729537	0.205266	-0.590248	0.586340	1.000000

According to the heatmap's exact values of the connection among characteristics and the class label, HST, BP, AL, Su, BU, SC, Pot, and WBCC are all positively correlated. For subsequent prediction, all positively linked factors are taken into account.

4.2.1 Heatmap for correlation matrix

Figure 6: This figure represents the heatmap for correlation matrix.

The upper darker coloured section represents the positively correlated, the bottom darker coloured half represents the negatively correlated, and the middle part represents no correlation. The actual numbers of the relationship between characteristics and the class label in the heat map demonstrate that BP, ALB, Sg, Bu, Sc, Pot, WBCC and HST all have positive relationships.

4.3 Decision Tree Classifier

The precision of the Decision Tree classifier. The precision in this instance is 95.5%. Even after fine-tuning, this precision remained unchanged.

Table 2: classification report for decision tree

	Precision	recall	Feature- score	Support
0	0.96	0.93	0.945	34
1	0.97	0.96	0.965	46
accuracy			0.95	80
Macro average	0.95	0.95	0.95	80
Weighted average	0.95	0.95	0.95	80

The classification report from the DT classifier. Decision trees have an accuracy rate of 0.95. Overall, the F1-value is 95%. Personal F1-values for CKD ranges between 96.5% and 94.5%. In the picture above, precision and recall are also displayed.

4.3.1 Decision Tree Confusion Matrix:

Table 3: confusion matrix

		Predicted Label	
		0	1
True Label	0	31	3
	1	1	45

The values 31 and 45 were correct predictions in the aforementioned confusion matrix, while 4 was an incorrect prediction.

4.4 K-Nearest Neighbor

The classification precision of the KNN classifier. Compared to other methods, the accuracy is lower in this case. This precision did not become much better even after fine-tuning. Accuracy of KNN Method: 0.725

Table 4: KNN algorithm classification report

	precision	recall	F1-score	support
0	0.61	0.79	0.688	34
1	0.81	0.66	0.727	46
accuracy			0.725	80
Macro average	0.72	0.725	0.725	80
Weighted average	0.73	0.73	0.725	80

KNN algorithm classification report. KNN's overall performance falls short of expectations. Here, an overall F1 score of 72.5% was attained. Individual F1 outcomes are 68.8% for people without CKD, whereas values for people with CKD are 72.7%.

4.4.1 KNN Confusion Matrix

Table 5: KNN Confusion Matrix

		Predicted Label	
		0	1
True Label	0	26	7
	1	16	31

In the above confusion matrix, the values 26 and 31 were the accurate predictions and then 23 were erroneous predictions.

4.5 Logistic Regression

The LR model report is displayed. The most accurate model for classifying items is this one.

Table 6: Logistic Regression Classification Report

	Precision	Recall	F1-Score	Support
0	0.95	0.66	0.78	34
1	0.82	0.79	0.81	46
Accuracy			0.975	80
Macro average	0.975	0.985	0.975	80
Weighted average	0.975	0.975	0.975	80

In this instance, a 97.5 percent overall F1 score was attained. The F1-score for those without CKD is 78 percent, whereas the score for people with CKD is 81 percent.

4.5.1 Confusion matrix for logistic regression

Table 7: Confusion matrix for logistic regression

		Predicted Label	
		0	1
True Label	0	33	1
	1	1	45

Two correct predictions, represented by the numbers 33 and 45 in the confusion matrix above, were preceded by two incorrect predictions. The logistic regression model's concluding prediction. The expected outcome and the estimated performance of the model are depicted in the confusion matrix.

5. Random forest

The precision of the Random Forest classifier. The precision in this instance is 92.5%. Even after fine-tuning, this precision remained unchanged.

Table 8: Random Forest Classification Report

	Precision	Recall	F1-Score	Support
0	0.83	0.77	0.798	34
1	0.72	0.61	0.661	46
Accuracy			0.925	80
Macro average	0.93	0.925	0.925	80
Weighted average	0.925	0.925	0.925	80

In this instance, a 92.5 percent overall F1 score was attained. The F1-score for those without CKD is 80 percent, whereas the score for people with CKD is 66 percent.

Table 9: Confusion matrix for Random Forest

		Predicted Label	
		0	1
True Label	0	30	4
	1	2	44

In the above confusion matrix, the values 30 and 44 were the accurate predictions and then 6 were erroneous predictions.

Model Comparison

The visual makes it clear that among the models in the framework, LR is the most finest.

Table 10: Model comparison

Name of the Model	Accuracy (%)
Decision Tree	**95**
K-nearest neighbor	**72.5**
Logistic Regression	**97.5**
Random Forest	**92.5**

It got 97.5 percent accuracy using Logistic Regression. The Decision Tree classifier & Random Forest also performed well, with an accuracy of 95 & 92.5 percent.

6. Conclusions

Chronic kidney disease is a huge strain on the medical industry due to its increasing prevalence, propensity of developing into end-phase kidney dysfunction, and poor incidence and death prognosis. This is slowly being a major worldwide physical condition issue. The main contributors to this illness include poor eating patterns and a deficiency in water intake. A person is able to live without kidneys for around 18 days before needing dialysis or a kidney transplant. Initial CKD requires the necessitates of precise prediction techniques. Approaches based on ML are quite good at foretelling CKD. Four different models for accurate prediction were trained in this study using a variety of physiological variables and ML methodologies LR, DT, RF and KNN. With an accuracy of almost 97.5% in this examination, the LR cataloging technique was shown to be the most accurate in this capacity. Their durability has been demonstrated by numerous model comparisons, and the scheme can be extrapolated from the study's results.

According to the study's findings, gender and chronic renal illness are unrelated. BP, Al, Su, Bu, Sc, Pot, WBCC, and HST are all positively connected, as seen by the exact values of the relationship between features and the class label. The study's findings imply that better predictions of chronic kidney disease can be made using the logistic regression. Considering on the study, Decision Tree accuracy is 95 percent, the K-NN Method accuracy is 72.5%, the Logistic Regression accuracy is 97.5% and the Random Forest accuracy is 92.5 percent. As a consequence, for forecasting Chronic Kidney Disease, Logistic Regression is the best fit.

References

[1] Ahmad .I, M. Basheri, M. J. Iqbal, and A. Rahim, “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection,” *IEEE Access, vo*l. 6, pp. 33789–33795, 2018.

[2] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina ,"Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," *2018 IEEE Congress on Evolutionary Computation (CEC),* Rio de Janeiro, 2018, pp. 1-9.

[3] Arif-Ul-Islam and S. H. Ripon, *"Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree,"* 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

[4] B. Bharathi, S.Prince Mary (2019), Neural Computation based general disease prediction model, *International journal of Recent Technology and Engineering*, vol8(2),pp 5646-5649.

[5] R. Devika, S.V. Avilala and V. Subramanian, *"Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest,"* 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 679-684.

[6] U. N. Dulhare and M. Ayesha, *"Extraction of action rules for chronic kidney disease using Naïve bayes classifier,"* 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[7] V. Garcia, E. Debreuve, and M. Barlaud, "*Fast k nearest neighbor search using GPU,"* in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6, Anchorage, AK, USA, 2008.

[8] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, *"Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)*," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 291–296, Washington, DC, USA, 2017.

[9] Surya Krishnamurthy, Kapeleshh KS, Erik Dovgan, Mitja Luštrek, Barbara Gradisek Piletic, Kathiravan Srinivasan, Yu-Chuan (Jack) Li, Anton Gradišek, and Shabbir Syed-Abdul, "Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan," *Healthcare*, vol. 9, no. 5, p. 546, 2021.

[10] E. Sakthivel, K. Senthamarai Kannan, and S. Arumugam, "Optimized Evaluation of Students Performances using Fuzzy logic," *International Journal of Scientific & Engineering Research*,Vol.4,Issue 9, p 1128-1133.

[11] Misir R, Mitra M, Samanta RK. A Reduced Set of Features for Chronic Kidney Disease Prediction. J Pathol Inform, 2017.

[12] Sakthivel. E And Anitha. S, "Survival Analysis For Diagnosing Tuberculosis, *Stochastic Modeling and Applications*Vol. 25 No. 1 (January-June, 2021)

[13] S. Nair, S. V. O'Brien, K. Hayden et al., "Effect of a cooked meat meal on serum creatinine and estimated glomerular filtration rate in diabetes-related kidney disease," *Diabetes Care*, vol. 37, no. 2, pp. 483–487, 2014.

[14] P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," 2016 Second Asian Conference on Defence Technology (ACDT), *Chiang Mai,* pp. 145-150, 2016.

[15] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked,* vol. 15, article 100178, 2019.

[16] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering and Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.

[17] V. Ravindra, N. Sriraam and M. Geetha, *"Discovery of significant parameters in kidney dialysis data sets by K-means algorithm,"* International Conference on Circuits, Communication, Control and Computing, Bangalore,2014,pp.452-454.

[18] G. Sharma Kaur and A., "*Predict chronic kidney disease using data mining algorithms in*

hadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, pp. 973-979, 2017,

[19] N. Tazin, S. A. Sabab and M. T. Chowdhury,"Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, *Health Informatics and Technology (MediTec)*, Dhaka, 2016, pp. 1-6.

[20] S. Vijayarani,S.Dhayanand, "Kidney Disease Prediction Using SVM AND ANN ALGORITHMS", *International Journal of Computing and Business Research* (IJCBR), vol. 6, no. 2, 2015.

[21] M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "*Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms,"* 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 300-303.

[22] H. A.Wibawa, I. Malik and N. Bahtiar,*"Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease,"* 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2018, pp. 1-4

[23] P. Yildirim, *"Chronic kidney disease prediction on imbalanced data by multilayer perceptron: chronic kidney disease prediction,"* in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), pp. 193–198,Turin, Italy, 2017.

[24] H. Zhang,C. Hung, W. C. Chu,P. Chiu and C. Y. Tang, *"Chronic Kidney Disease Survival Prediction with Artificial Neural Networks,"* 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 1351-1356.