Risk Factor Prediction of Chronic Kidney Disease Based on Machine Learning Algorithms.

Bijivemula Chenna Krishna Reddy
School of Computer Science and Engineering
Lovely professional University
Phagwara, Punjab, India, bijivemula.11914197@lpu.in

***Abstract*— Chronic kidney disease (CKD) is an increasing medical issue that declines the productivity of renal capacities and subsequently damages the kidneys. CKD is very common nowadays; cardiovascular infection and end-stage renal illness are two life threatening diseases that can be caused as aftereffects of CKD. These are conceivably preventable through early recognizable conditions and treatment of people who are in danger. The expectation of medical problems is a very troublesome assignment. CKD is particularly one of the most lethal diseases in the clinical field. Before it becomes too late to recognize CKD forecast, to get rid of risks, the prediction of risk factor is a major necessary step in the immediate stage. In this paper, we have applied six algorithms. Naïve Bayes, Random forest, Simple logistic regression, Decision Stump, Linear regression model, simple linear regression is using for predicting the risk factors of CKD. Considering the orderly execution and investigations of these strategies, six algorithms give a superior and quicker characterization execution. Six individual algorithms are applied to the dataset and the best outcomes have been acquired through the classification of predicting risk factors.**

I. INTRODUCTION

Now a days, the ratio of chronic kidney disease is rapidly progressive. The current state of CKD is hampering human's day to day life, and it cause for heart failure. Many people are facing this problem in Bangladesh. In most cases rural areas people are not aware about it for deficiency of unbounded sense, few sensations are the main reason for CKD. Technology are increasing rapidly but people are not alert about this. So, they have face huge risk to their kidney. When the utility of a kidney did not work properly, people needs transplantation of the kidney that is not much suitable. Several kidney diseases occur with various symptoms as well kidneys will be damaged, it cannot filter blood the way it should. Sometimes it goes incurable, chronic. Many of several symptoms can be used to predict risk factor for kidney diseases. In this paper the proposal is to analyze the risk factors of CKD and warn patients to stay healthy. Mostly it can help the doctor to identify the symptoms easily and take proper steps to reduce it in before long stage. For this prediction analysis, using several algorithms named Naïve Bayes, Random Forest, Simple Logistic regression, Decision Stump, Linear regression model, Simple linear regression to predict the risk factor.

II. RELATED WORK

Different sort of work has been accomplished for gathering helpful fact from Chronic Kidney Disease dataset utilizing information mining methods. This was done to decrease the hour of the examination and what is more, it would expand the exactness of the expectation with the assistance of the information mining classification technique. Data Mining is likewise utilized for the goal and prognostic of a few infections . K. Ero˘glu and T. Palaba ş proposed guidance that connected six classifiers-KNN, NB, SVM, choice tables, RF, J48, and three outfits measure. Creators of try different things with ongoing kidney sickness utilizing the k-means algorithm and Apriori. An examination was introduced to recognize CKD utilizing SVM, DT, NB, and KNN calculations. Ani R et al altered different characterization of calculations, for example, DT, NB, LDA classifier, Back Propagation Network (BPN), Random Subspace, and KNN. For counteraction of death rate brought about by CKD were applied DT and NB characterization methods to anticipate CKD planned which can estimate Chronic Kidney Disease at a beginning phase? They utilized a few neural networks algorithm. A trial led by M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra test that truncation of KNN, CFS, and AdaBoost. Its prosperity was 98.1%. M.P.N.M. Wickramasinghe et al Presents an exploration concentrate by bringing information from a patient's clinical records and afterward applying an arrangement calculation to these records, which has given CKD patients a reasonable eating regimen plan. Arora, M., and Sharma, E. A. proposed a technique for information mining that has Identification capacities of release window to execution in weka's apparatus. Ms. Astha Ameta et al essentially retained information mining strategies and the techniques by which it can foresee persistent kidney infection. So unmistakably information mining was a more practical instrument for foreseeing long term kidney illnesses. Our proposed technique will investigate the portrayal of Naïve Bayes, Random woods, simply calculated relapse classifiers discover the better exactness of CKD and quest the best answer for identifying CKD. A. J. Aljaaf and Deepika B et al, analysis the early stage of CKD based on machine learning algorithm and find out the most significant factor. Siddheswar Tekale and S. Pitchumani Angayarkanni et al, they are predicted the early stage of CKD and find out the better accuracy to prevent it. Marwa Almasoud et al , detected the CKD using least numbering prediction using machine learning. From the exploratory consequences of Decision Stump, Linear regression model, simple linear regression calculations discover the preferred factor positioning over different Algorithm.

III. PROPOSED SCHEME

Data mining for diagnostic has become an existent tendency in our technological advanced world. In human body there are many survival organs, if they are not working properly human's life are in danger. Kidney is one of the major organs of them. It helps us to reduce the waste product that flow in our body. It is not only filtering the excess fluids but also filter the toxic from our blood. Kidney can control the body's red blood cells, blood pressure and realizes erythropoietin, enzymes such as kallikreins. Chronic kidney illness has ended up a worldwide wellbeing issue concern with rising predominance. Chorionic kidney disease, also called chronic kidney miscarriage, describes the continual decrease of kidney function. Must need

to take a few steps to anticipate and control it. By utilizing different information data mining strategies. The proposed model is to predict kidney diseases with a large dataset to increase the model accuracy and find out the significant and non-significant risk factor of CKD. Naive Bayes algorithm, Simple logistic regression, Random forest are using to predict the accuracy of the model and linear regression model, Decision stump, Simple Linear regression model re using to find out most significant and non-significant risk factor of CKD.

Naive Bayes classification and dissect the foremost viable method. Naive Bayes classifiers are straightforward classifiers with likelihood based on Bayes theorem. The Random Forest accomplished higher than Naive Bayes within the prediction of CKD in our analysis. The quantity of accuracy predictions in Naive Bayes 93.9056%, Random Forest 98.8858%, Simple logistic 94.7679%. So, anticipating the result from the exactness that what number of patients' unit of measurement having the persistent nephropathy at interims a particular time.

The Random Forest wrapped up way better in expressions of precision, and f degree over datasets, though Naive Bayes, appears way better Accuracy. Subsequently, it can be said that Random Forest accomplished higher than Simple logistic and Naive Bayes within the expectation of CKD in the analysis.

In the methodology, six algorithms have been implemented to predict the risk factors of CKD. Associated the algorithms to make a hugely effective method of predicting the CKD risk factors, ensuring very less defects while predicting.

In the primary step, the data was prepared by pre-processing for doing the actual operation. At first, some information has been elected to integrate it into very small parts. Then the data cleaning was done and separated. Finally, a Synthetic Dataset for CKD have been obtained.

After getting the synthetic data set, allocation of the dataset occurs of two individual actions which are Normalized data, Formatting data. After these two actions are completed then combined those and examined for finding the "Z" score. Applied a condition of if the Z>-2 or not. If not, then disease is not found, and the process ends. But if it matches the condition then disease is found.

Z-score normalization may be a strategy of normalizing data that avoids this outlier issue. The formula for Z-score normalization is below: v a l u e − μ σ \frac{value - \mu}{\sigma} σ value−μ Here, μ is that the mean of the feature and σ is that the variance of the feature. -2 is the formal value of Z-score, it is the minimum threshold condition that can fulfill the standard normalization process. Z- Score is very helpful to understanding the probability of data to normalize it easily.

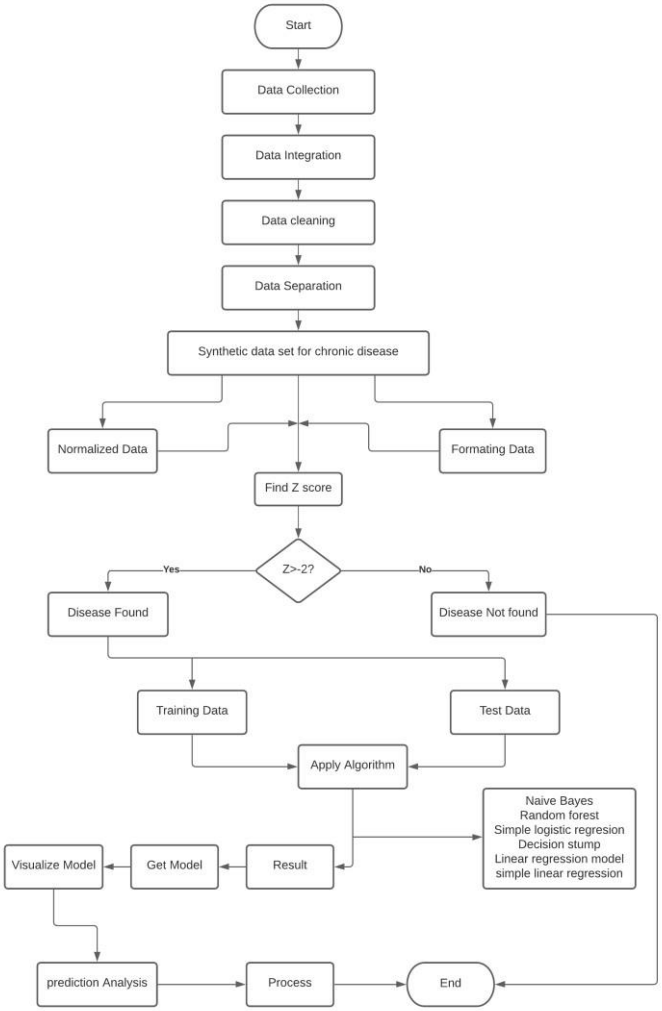

Fig. 1. Flowchart of proposed model.

After finding out the disease, splitted the data into two sections, Test data and Train data. After this, six algorithms were applied to find out the risk factors. Naïve Bayes, Random forest, Simple logistic regression is using to find out the predicting accuracy of the CKD and Decision Stump, Linear regression model, Simple linear regression is using to calculate the risk factor of CKD. After that, the result was found, and a model was got, and the visualization of the model have been done. After discerning this prediction, analyzing is done. Finally, processing and the closure of the operation is performed.

IV. DATASET DESCRIPTION

From the Previous study information was gathered from a survey and created a questionnaire for the data collection. Then, from a reputed medical college in Bangladesh, patient medical data was collected. In the questionnaire both case and control type question were added. The age, blood pressure, hypertension etc. were also included in the questionnaire. After creating the questionnaire, the data was collected through this questionnaire and formatted the data set into CSV format have collected 1032 patient data. We are applying 68% data for the training process and 32% data for the testing process. The overall data preprocessing process is easily maintain and we can get our valuable output to analyze it.

TABLE 1: Stages of CKD risk level

Attribute	Description
Blood Pressure	Given in mm/Hg
Specificity Gravity	Ranges from 1005 to 10025 (the higher the risk)
Albumin	Range is 0 to 5 (the higher the better)
Sugar level	5 levels indicating severity
Red Blood Cells	Is abnormal or normal
Blood urea	It is in mgs/dl
Serum creatinine	High level is not good
Sodium	It is measured in me/L
Potassium	It is measured in me/L
Hemoglobin	Less than 15 is kidney failure
White Blood Cell Count	This is numerical cell count
Red Blood Cell Count	Should not be higher or less than normal
Hypertension	It is categorical (yes or no)
Class	Given as CKD or not CKD

A. Dataset Preprocessing

After collecting the data set, we preprocessed the data because in real-time data set, data often missing or contain garbage value. So, we fill out the missing value by the mean of its attribute column, smooth the noisy values. Data encoding was also applied to convert the data from string to numeric. Some attributes like age, specific gravity, blood glucose regulator, hemoglobin, etc. were organized in a continuous format.

B. Algorithms

Decision stump: A decision stump is a Decision Tree. It utilizes just a solitary trait for parting. This commonly implies that the tree comprises just a solitary inside hub for discrete credits, point to be noted that the root has just leaves as replacement hubs. On the off chance that the property is mathematical, the tree might be more mind boggling. A decision stump is an AI model. It consists of a one-level decision tree. That is, it is a decision tree with one inside hub which is quickly associated with the terminal hubs. Decision stumps perform shockingly well on some regularly utilized benchmark datasets from the UCI vault, which outlines that student with a high Bias and low Variance may perform well since they are less inclined to Over fitting.

Simple linear regression: Simple linear regression is a measurable technique. It permits to sum up and study connections between two nonstop (quantitative) factors: One variable, signified x, is viewed as the indicator, logical, or free factor. Regression permits to appraise the way a
reliant variable change as the free variable(s) change. Simple linear regression is utilized to appraise the connection among two quantitative factors. However, due to its specific nature, this strategy is one of the quickest with regards to simple linear regression. Aside from the fitted coefficient and capture term, it likewise returns fundamental measurements, for example, R^2 coefficient and standard blunder.
Equation: $Y = a + bX + e$ *Xor ……..(1)*

Here, Y is a Dependent variable of (Y), and alpha is a constant; X is the Independent variable of (X) which is the coefficient of X; e is the error term.

Naive Bayes: Naive Bayes could be a machine learning probability algorithm that will be used for a spread of classification tasks. Typical applications include classifying documents, sentiment prediction, etc. it is going to be a probabilistic demonstration, the calculation is often coded up effectively, and therefore, the forecasts made genuine fast. It has been successfully used for several purposes; naive Bayes may be a probabilistic machine learning algorithm supported the Bayes Theorem. Bayes' Theorem is a law of conditional probabilities. It is used to classify the parameter estimation of small training data. It performs well in multiple class prediction.

Simple logistic regression: It is an easy Algorithm that you simply can use as a performance baseline, it is easy to implement, and it will have the best enough in many tasks. Therefore, every Machine Learning engineer should be conversant in its concepts. Like many other machine learning techniques, it is borrowed from the sector of statistics and despite its name, it is not an algorithm for regression problems, where you would like to predict the endless outcomes. It gives you a discrete binary outcome between 0 to 1. To mention it in simpler words, it is the result is either one thing or another.

Random Forest: The "random forest" may be a classification algorithm comprising of various choice trees. It utilizes stowing and highlight arbitrariness when constructing every individual tree to aim to form an uncorrelated wood of trees whose expectation by the panel is more exact than that of a person tree. In the training set hundred to thousand trees are counting on the dimensions, the number of sample trees are B. The error of prediction in each training sample X_i, X_i only using in bootstrap sample to fit the training and test error tend in some number of trees.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}. \quad \text{……(2)}$$

Linear Regression: Linear regression may be a fundamental and frequently utilized quite prescient investigation. The overall thought of regression is to seem at two things: (1) the indicator factors work superbly in anticipating subordinate volatile? (2) Which factors specifically are huge indicators of the result variable, and the way would they–showed by the extent and indication of the beta appraisals sway the result volatile? The smallest amount complex sort of the regression condition are characterized by the recipe y = c + b*x, where y = assessed subordinate variable score, c = consistent, b = parametric statistic, and x = score on the free factor.

V. RESULT AND ANALYSIS

In this Analysis, we have obtained results from three distinct calculations Naïve Bayes, Random Forest, Simple Logistic which are regulated calculations. We examined the result and got an aftereffect of up to 90%, so we can say that these models are highly proficient for this Dataset. Our trained information collection was 63%. At this point, when we prepared up, we got the results by testing 37% information. Furthermore, we identified reasons for Kidney ailment by administering calculations by Decision Stump, Linear Regression Model, and Simple Linear Regression. The fundamental driver is hemoglobin, and it is the principal factor of kidney ailments.

TABLE 2: Accuracy Table of CKD

Algorithm	Accuracy
Naïve Bayes	93.9056 %
Random Forest	98.8858 %
Simple Logistic	94.7679 %

From the result analysis, we can see that Random forest algorithm get the high accuracy 98.8858%. So, our approach model better than any other models. By using this model, we can easily predict the chronic kidney disease. It is an invention that can help the medical community to progress our biomedical science.

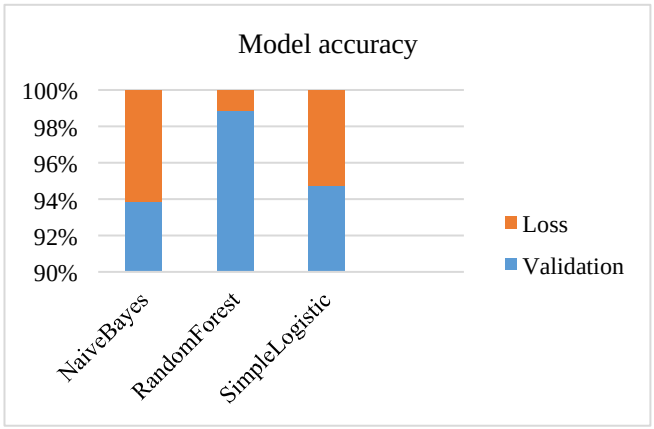

Fig. 2. Accuracy chart diagram of CKD.

TABLE 3: Most significant risk factor prediction table

Algorithm	Factor
Decision Stump	Hemo <= 13.05 : 0.9578606158833063 Hemo > 13.05 : 0.12310286677908938
Linear Regression Model	-0.0002 * Bu + -0.091 * Hemo + 0 * Wbcc + -0.0076 * Rbcc + 0.288 * Htn + 1.5892
Simple Linear Regression	-0.12 * Hemo + 2.13

From the overview of the result, we can easily identify the most significant and non-significant risk factor. In decision stump and simple linear regression model can analyze that hemoglobin is the most significant risk factor.

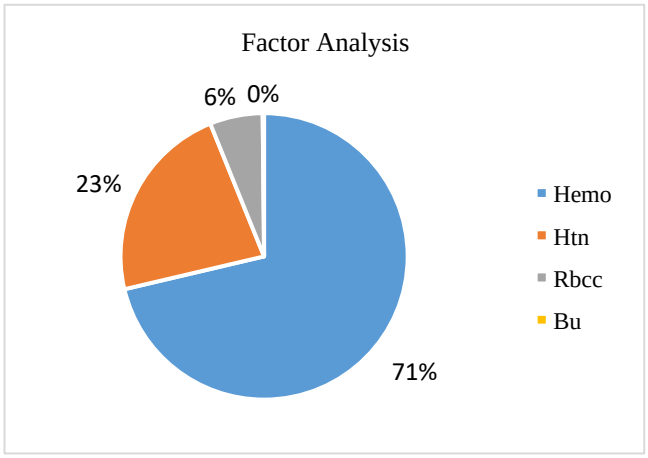

Fig. 3. Pie chart of risk factor prediction.

From the factor analysis we can find out the most significant and non-significant risk factor of CKD. From analyzing report, we can find that hemoglobin is the most significant risk factor for CKD and Hypertension is the less significant risk factor in CKD.

A. Comparison Table

TABLE 4: Comparative analysis of approach model

Algorithms	Precision	Recall	F-Measure	ROC Area
Naïve Bayes	0.940	0.939	0.939	0.972
Random Forest	0.989	0.989	0.989	0.999
Simple Logistic	0.948	0.948	0.948	0.976

Through the comparative analysis between three algorithms, we have obtained the best accuracy from Random forest algorithm, and it is the best fit for our dataset. Our approach model is better than other model to find out the most significant and non-significant risk factor of CKD.

VI. FUTURE WORK

In this paper, we have completed our work with a large dataset. In the future, a new and unprecedented aspect of CKD prevention and control will be revealed through CKD risk prediction, which will play a vital role in diagnosing CKD in our medical science. Doctors will be able to predict CKD by observing the results. Medical scientists will be able to use this dataset and observe the results to play a special role in controlling and preventing CKD. Through this study, a new horizon in CKD control will be opened in the future.

VII. Conclusion

In this paper, we predicted chronic kidney disease (CKD) risk factors and predicted the progression of CKD. Risk factor predictions perform an essential induction in recognizing the risk of getting rid of chronic kidney disease (CKD). Using algorithms to predict risk factors to get the best results are achieved by categorizing every single strategy. We are getting high accuracy with Random forest algorithm (98.8858 %). In this context, chronic kidney disease (CKD) will be particularly effective in predicting outcomes by identifying or listing people at risk. It will be especially effective to treat people by listing people at risk for scores to predict outcomes. However, a significant portion of the population at lofty hazard of chronic kidney disease (CKD) can still be recognized or identified within the community using CKD risk factor predicting without admittance for taking care in hospital.

References

[1] Ramya, S., & Radha, N. (2016). Diagnosis of chronic kidney disease using machine learning algorithms. International Journal of Innovative Research in Computer and Communication Engineering, 4(1), 812-820..J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Purushottam Sharma, Kanak Saxena and Richa Sharma, "Heart Disease Prediction System Evaluation Using C4. 5 Rules and Partial Tree," in Computational Intelligence in Data Mining—Volume 2, ed: Springer, pp. 285-294, 2016.

[3] Ritika Chadha, Shubhankar Mayank, Anurag Vardhan and Tribikram Pradhan, "Application of Data Mining Techniques on Heart Disease Prediction: A Survey," in Emerging Research in Computing, Information, Communication and Applications, ed: Springer, pp. 413-426, 2016.

[4] Moloud Abdar, Mariam Zomorodi-Moghadam, Resul Das and I-Hsien Ting, "Performance analysis of classification algorithms on early detection
of liver disease," Expert Systems with Applications, vol. 67, pp. 239-251, 2017.

[5] K. Ero ̆glu and T. Palaba ̧s, "The impact on the classification performance of the combined use of different classification methods and different ensemble algorithms in chronic kidney disease detection," 2016 National Conference on Electrical, Electronics and Biomedical Engineering(ELECO), Bursa, 2016, pp. 512-516.

[6] N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection tech-nique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6. doi:10.1109/MEDITEC.2016.7835365.

[7] R. Ani, G. Sasi, U. R. Sankar and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," 2016 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), Jaipur, 2016,pp. 1287-1292. Doi: 10.1109/ICACCI.2016.7732224 .