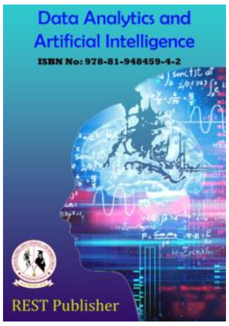
Data Analytics and Artificial Intelligence

Vol: 3(2), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/3/2/28


Data Analytics and Artificial Intelligence
ISBN No: 978-81-948459-4-2
REST Publisher

Predicting Chronic Kidney Disease using RF Algorithm for Big Data

K. Jeya gowri

Shri Shankarlal Sundarbai Shasun Jain College for Women Chennai, Tamil Nadu, India.

*Corresponding Author Email: jeyagowri@shasuncollege.edu.in

Abstract. *Chronic Kidney Disease (CKD) involves a slow loss of kidney function. Kidneys remove wastes and fluids from your blood, which are later eliminated in urine, covering over a period of months to years, signs and symptoms of kidney disease are usually indistinct, and are a serious disease. It is enlightened in six stages congenial to the severity level. It is categorized into various stages based on the Glomerular Filtration Rate (GFR), which in turn utilizes several attributes, like age, sex, race and Serum Creatinine. Among multiple available models for estimating GFR value, Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI), which is a linear model, has been found to be quite efficient because it allows detecting all CKD stages. Random Forest had 99.24% truthfulness. The model's best result is created by considering the 10 most reelected features. When compared to previous studies, our results are amid the good for assessment metrics and the ranking accuracy.*

Keywords: *Random forest, Chronic kidney disease, Disease prediction, Data analysis, Feature selection, Machine learning.*

1. INTRODUCTION

Initial detection and cure of CKD is highly desirable as it can start the prevention of unwanted consequences. Machine learning methods are being broadly approved for early detection of symptoms and diagnosis of a lot of diseases recently. As an end result, to predict the different stages of CKD using machine learning classification algorithms on the dataset collected from the medical records of troubled people. In addition, for looking over's in definite variables, facts received imputation may have a massive deviation from the real values to detect various stages of CKD with all-inclusive medical accuracy. This research used the Random Forest algorithms to access a feasible and applicable model. The concern of reliable diagnostic equipment that can discriminate between CKD affected and NOT CKD one cannot be highlighted. Because the best part of patient evidence practiced in illness identification is varied in character and complicated to evaluate manually beneficial to identify diseases and make original conclusions. Machine Learning has been presented to be an effective tool in medical applications. The scope of detecting patterns in medical records that humans are unable to detect. It may provide guidance with analysis and predicting in the future. People may take advantage of these predictions to assist them avoid infective certain illnesses.

2. RANDOM FOREST ALGORITHM

Random Forest belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Machine Learning Repository, dataset consists of only two classes, i.e., CKD affected and NOTCKD indicating people with no chronic kidney disease. CKD is divided into five stages depends on Glomerular Filtration Rate. (GFR)

Stage 1 - Normal or high and GFR > 90 mL/min
Stage 2 - Mild CKD and GFR = 60-89 mL/min
Stage 3A - Moderate CKD and GFR = 45-59 mL/min
Stage 3B -Moderate CKD and GFR = 30-44 mL/min
Stage 4 Severe CKD and GFR = 15-29 mL/min
Stage 5 End Stage CKD and GFR <15 mL/min

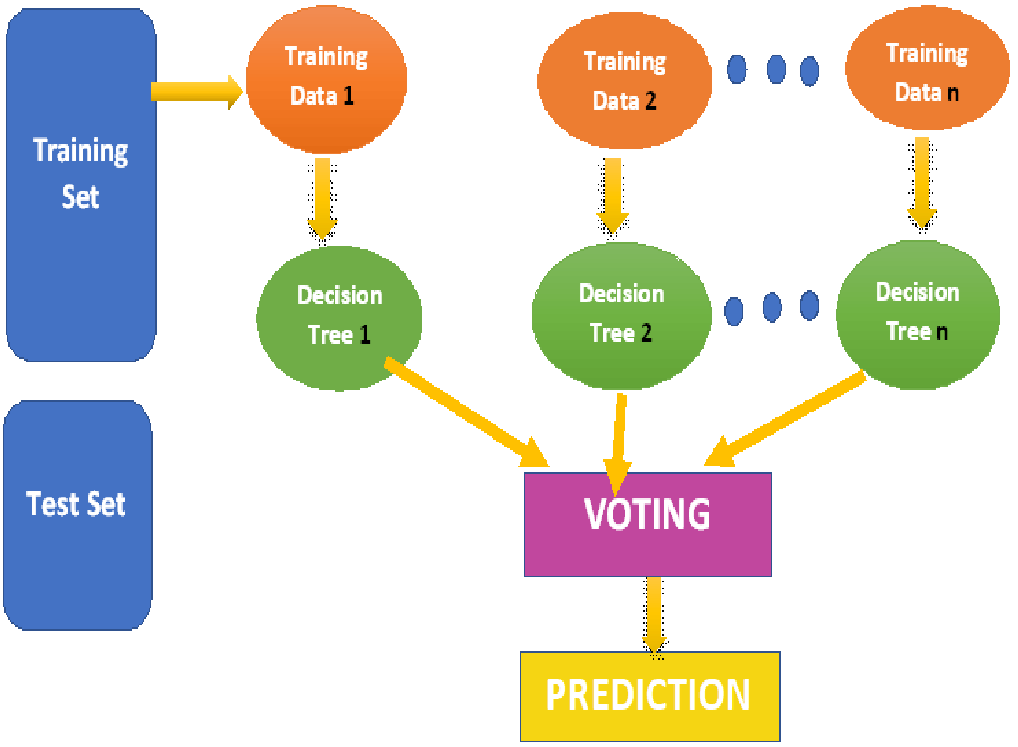


Figure 1

Step 1: Select random data from a given data or training set.
Step 2: Construct a decision tree for all training set data
Step 3: Choose the particular data for decision trees that you want to build.
Step 4 : Repeat the step 2 and 3
Step 5: Find the prediction of decision trees and assign to new data that match with all vote
Step 6 : Finally, select the most voted prediction result as the final prediction result.

In Medicine, With the help of this algorithm, identify the disease trends and risks. Hyperparameters are used in random forests to either enlarge the performance and predictive capacity of models or to make changes quickly.

n_estimators: Using algorithm built Number of trees before averaging the products.

max_features: Maximum number of features random forest uses before considering splitting a node.

mini_sample_leaf: Determines the minimum number of leaves required to split an internal node.

n_jobs: Conveys to the engine how many processors are allowed to use. If the value is 1, it can use only one processor, but if the value is -1. there is no limit.

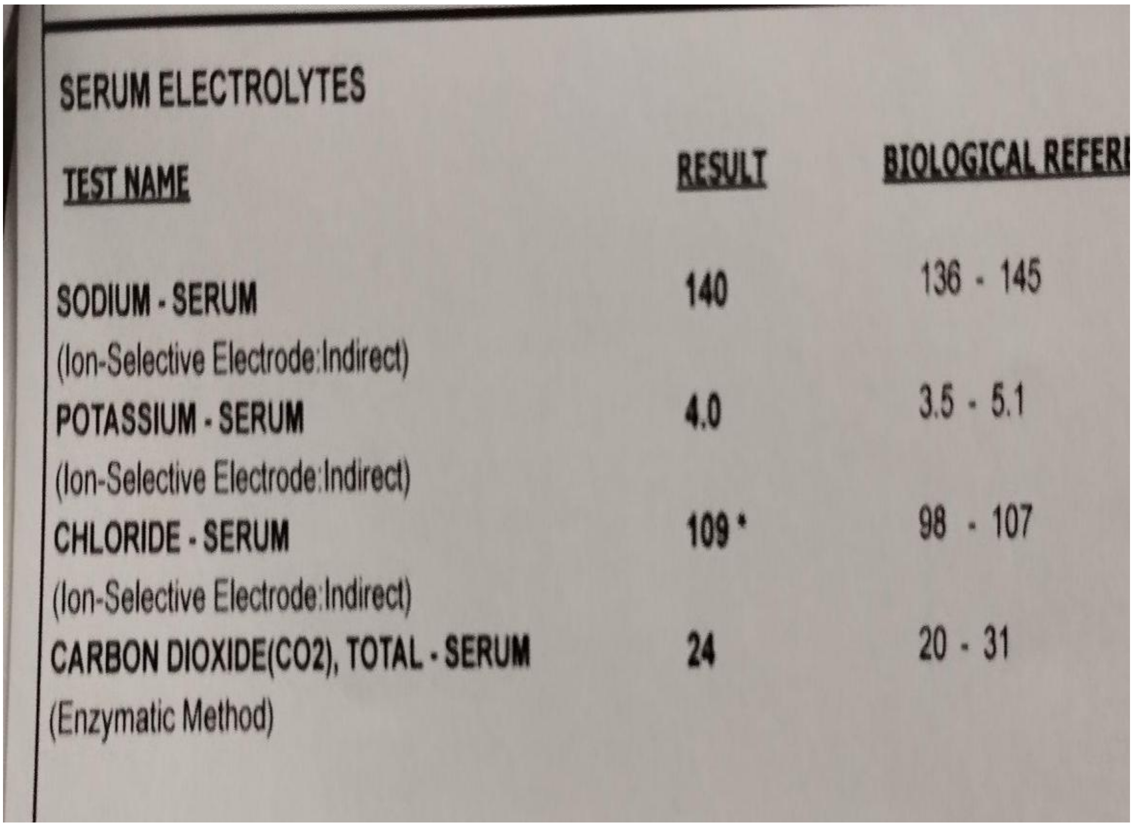
Random state: Controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and if it has been given the same hyperparameters and the same training data.

Obscure: OOB (Out Of the Bag) is a random forest cross-validation method. In this, one-third of the sample is not used to train the data but to evaluate its performance.

TABLE 1. Data Set Information

age	age
bp	blood pressure

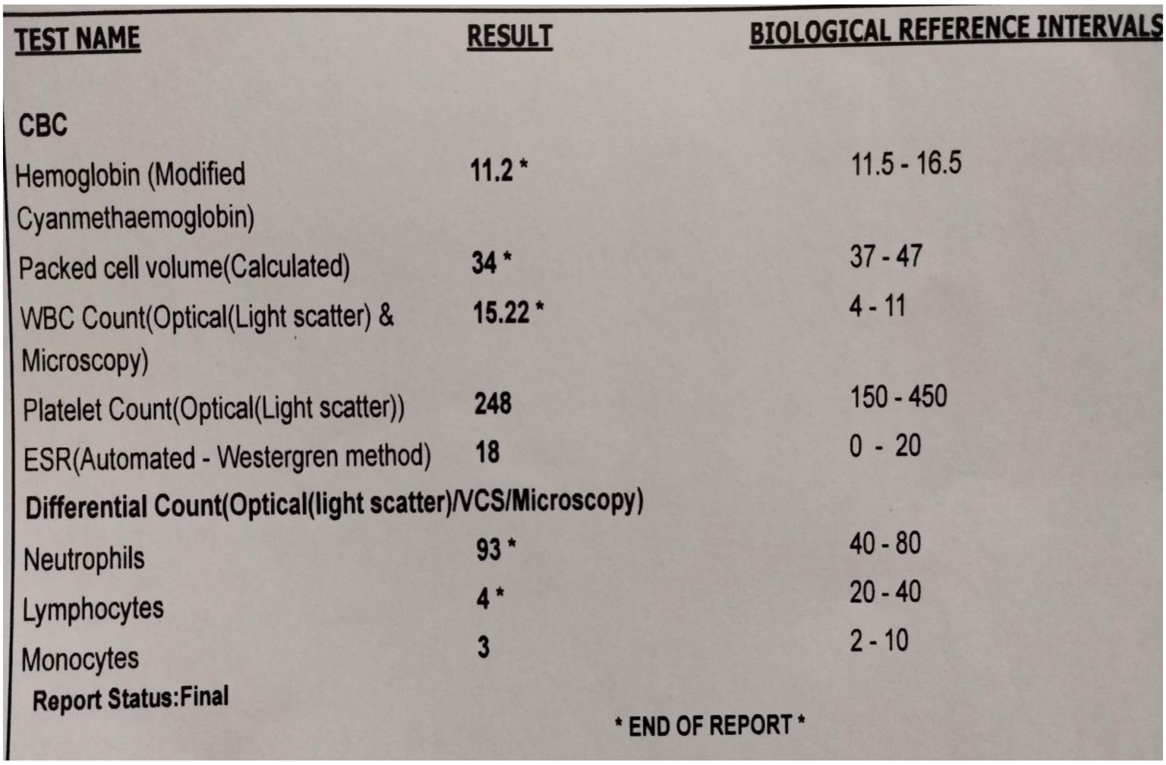
sg	specific gravity
al	albumin
su	sugar
pc	pus cell
pcc	pus cell clumps
bgr	blood glucose random
bu	blood urea
sc	serum creatinine
sod	sodium
pot	potassium
pcv	packed cell volume
hemo	hemoglobin
wc	white blood cell count
cs	chloride serum
ts	Total Serum



SERUM ELECTROLYTES

TEST NAME	RESULT	BIOLOGICAL REFER
SODIUM - SERUM (Ion-Selective Electrode:Indirect)	140	136 - 145
POTASSIUM - SERUM (Ion-Selective Electrode:Indirect)	4.0	3.5 - 5.1
CHLORIDE - SERUM (Ion-Selective Electrode:Indirect)	109 *	98 - 107
CARBON DIOXIDE(CO2), TOTAL - SERUM (Enzymatic Method)	24	20 - 31

FIGURE 2. Sample Data Set



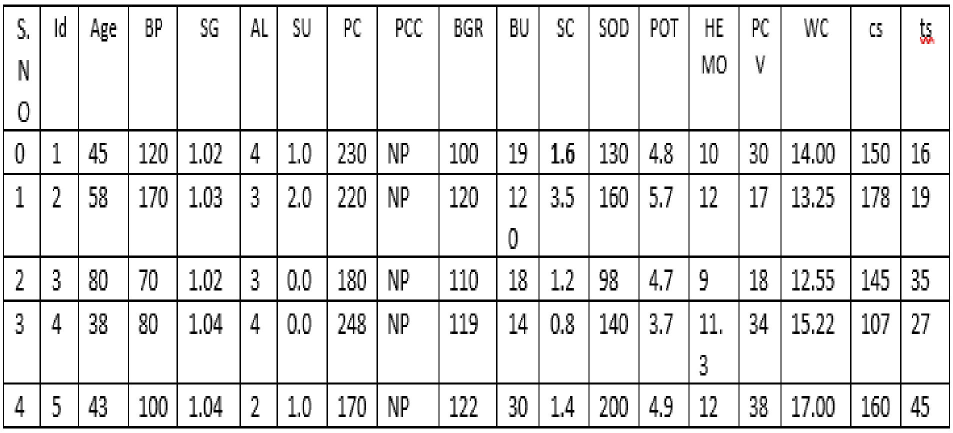
TEST NAME	RESULT	BIOLOGICAL REFERENCE INTERVALS
CBC		
Hemoglobin (Modified Cyanmethaemoglobin)	11.2 *	11.5 - 16.5
Packed cell volume(Calculated)	34 *	37 - 47
WBC Count(Optical(Light scatter) & Microscopy)	15.22 *	4 - 11
Platelet Count(Optical(Light scatter))	248	150 - 450
ESR(Automated - Westergren method)	18	0 - 20
Differential Count(Optical(light scatter)/VCS/Microscopy)		
Neutrophils	93 *	40 - 80
Lymphocytes	4 *	20 - 40
Monocytes	3	2 - 10

Report Status:Final

*** END OF REPORT ***

FIGURE 2. CBC

Table 2

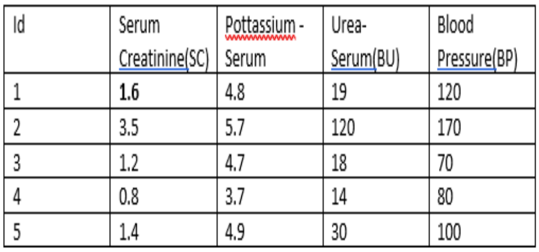


S.NO	Id	Age	BP	SG	AL	SU	PC	PCC	BGR	BU	SC	SOD	POT	HEMO	PCV	WC	cs	ts
0	1	45	120	1.02	4	1.0	230	NP	100	19	**1.6**	130	4.8	10	30	14.00	150	16
1	2	58	170	1.03	3	2.0	220	NP	120	120	3.5	160	5.7	12	17	13.25	178	19
2	3	80	70	1.02	3	0.0	180	NP	110	18	1.2	98	4.7	9	18	12.55	145	35
3	4	38	80	1.04	4	0.0	248	NP	119	14	0.8	140	3.7	11.3	34	15.22	107	27
4	5	43	100	1.04	2	1.0	170	NP	122	30	1.4	200	4.9	12	38	17.00	160	45

3. DATA MANIPULATION: CLEAN THE DATA

Now we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain. Next we drop or remove all columns except for the columns that we want to retain. Finally we drop or remove the rows that have missing values from the data set.

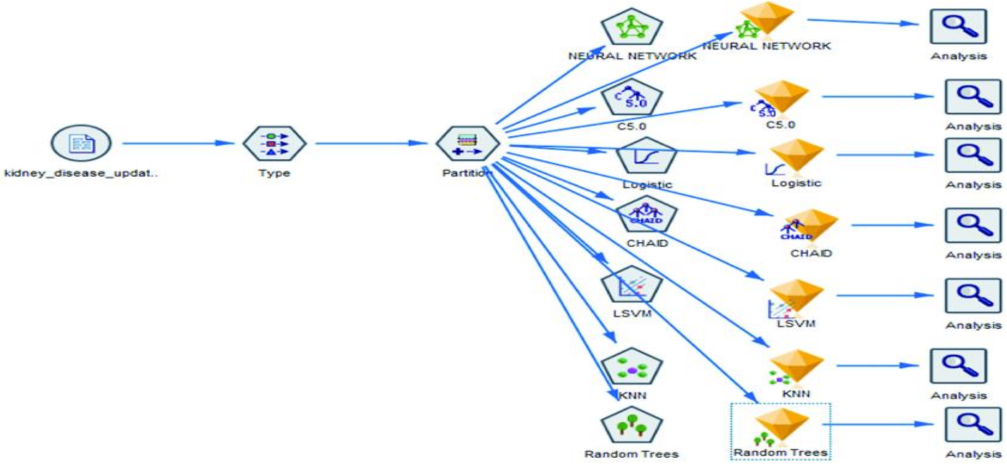
Table 3



Id	Serum Creatinine(SC)	Pottassium - Serum	Urea-Serum(BU)	Blood Pressure(BP)
1	**1.6**	4.8	19	120
2	3.5	5.7	120	170
3	1.2	4.7	18	70
4	0.8	3.7	14	80
5	1.4	4.9	30	100

4. BUILD THE MODEL USING RANDOM FOREST ALGORITHM

Random Forest model accuracy<-c(83.16, 92, 78.95,95,86)



5. CONCLUSION AND RECOMMENDATIONS

In this study, we established Random Forest to predict the various stages of CKD. It is observed that the ratio of correctly classified instances by Random Forest is 88.45%. On the other hand, the time taken by Random forest is 0.38/s. Random forest gets biased in favor of the attributes with categorical values. Random forest builds multiple decision trees, merges them together to get a stable prediction model. Random Forest Algorithm, it reduces the risk of over fitting and the required training time, it offers a high level of accuracy predictions by estimating missing data. After unsupervised imputation of lacking values within side the statistics set via way of means of the use of KNN imputation, the incorporated version may want to acquire a high satisfactory accuracy. In the future, a huge variety of greater complicated and consultant statistics can be amassed to enhance the generalization overall performance whilst permitting it to come across the severity of the ailment

REFERENCES

[1]. SiddheshwarTekale ,PranjalShingavi , Sukanya Wandhekar , Ankit Chatorikar,“Prediction of Chronic Kidney Disease Using Machine Learning Algorithm",IJARCC,October 2018

[2]. Manish Kumar, “Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm”, IJARCC,February 2016

[3]. Zhiyong Z. Zhang, Kun Hua, Arun Kumar Sangaiah,”Advanced Soft Computing Methodologies and Applications in Social Media Big Data Analytics” 25 January 2018

[4]. Teo BW, Xu H, Wang D, Li J, Sinha AK, Shuter B, et al. GFR estimating equations in a multiethnic asian population. Am J Kidney Dis. 2011;58(1):56– 63. https://doi.org/10.1053/j.ajkd.2011.02.393.

[5]. Ponum M, Hasan O, Khan S. EasyDetectDisease: an android app for early symptom detection and prevention of childhood infectious diseases. Interact J Med Res. 2019;8(2):e12664. https://doi.org/10.2196/12664.

[6]. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

[7]. V. Jha, G. Garcia-Garcia, K. Iseki et al., “Chronic kidney disease: global dimension and perspectives,” The Lancet, vol. 382, no. 9888, pp. 260–272, 2013.

[8]. C. A. Johnson, A. S. Levey, J. Coresh, A. Levin, and J. G. L. Eknoyan, “Clinical practice guidelines for chronic kidney disease in adults: part I. Definition, disease stages, evaluation, treatment, and risk factors,” American Family Physician, vol. 70, no. 5, pp. 869–876, 2004

[9]. S. Krishnamurthy, K. KS, E. Dovgan et al., “Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan,” Healthcare, vol. 9, no. 5, p. 546, 2021.