International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Chronic Kidney Disease Prediction based on Blood Potassium Levels using Machine Learning

S.D.Harish, K.Vinay Kumar, K.Taraka Ram, G.Pradeepini

Abstract: ***Machine learning is an artificial intelligence(AI) technology that provides the systems with the knowledge and capability to learn and evolve automatically from specifically programmed experiences. This focuses on designing computer programs that are able to gain access and use information on their own. Kidney damage or decreased activity for more than three months is known as chronic kidney disease.This illness occurs when the kidneys can no longer expel extra water or waste from human blood. The goal of this research study is to prepare a predictive modeling for chronic kidney disease data to analyze the different open source python module and output the results predicted by machine learning algorithms and determine the accuracy by comparing different algorithms such as KNN and Logistic Regression which are primarily used for classification of data. This algorithm makes predictions on a dataset collected from the patient's medical records. It gives us the clarity that if someone has chronic kidney disease or not primarily based on a person's blood potassium levels present in their body.***

Keywords : Data mining , Machine Learning (ML), levels of blood potassium, Chronic renal disease.

I. INTRODUCTION

Machine Learning's objective is to understand the data structure and fit the data into models that people can understand and use. Machine learning is the artificial intelligence sub-branch, and deep learning is the advance principle of machine learning. Machine learning can predict the future on the basis of past or historical data. Constant kidney illness can likewise be alluded to as renal disappointment. One of every ten individuals worldwide experience the ill effects of kidney infection. 10% of the total populace experience the ill effects of constant kidney ailment; one of every five men and one out of four ladies from the age gathering of 60-75 have CKD as indicated by the National Kidney Foundation. In this examination we have utilized AI to distinguish CKD and non-CKD with 25 attributes contributing to kidney malady but, we mainly sorted the people into CKD and non-CKD based on one's blood potassium level. The information utilized comprised of a record 400 individuals. The informational index has an assortment of missing information. We have utilized this informational collection and order calculations to manufacture prescient models for the characterization of CKD. The model with the best forecast precision is taken. This will accomplish quick and precise outcomes for the expectation of CKD, which will lessen the ideal opportunity for the forecast of infection and give advantages to both specialist and patient in giving early treatment and expedient recuperation. Our model will correctly determine the presence or absence of chronic kidney disease in a patient based on their blood potassium level using machine learning techniques as well as data mining in which searching and processing operations on large data sets will be done. Also, the patterns and the trends will be found and then they will be transformed into data understood by the data pre-processing, visualization. Clinics and hospitals can use this quicker digital methodology for the prediction of chronic kidney disease.

A. Existing method

- In the current existing system so many researchers working with data mining algorithms in different kidney disease survey techniques.
- In these survey techniques they are using different databases available for different sources, but they are unable to analyze data with visualization techniques to identify the correlation between different attributes.
- Using limited techniques and they are unable to optimize by increasing efficiency.

B. Proposed method

- In our proposed system we are collecting data from different data sources and using advanced data visualization modules in python and analyze data with different visualization techniques.
- We use various machine learning algorithms such as KNN and Logistic regression to predict the presence of chronic kidney disease and to determine accuracy & performance.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
S.D.Harish*, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India. Email: harishsd1998@gmail.com
K.Vinay Kumar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India. Email: vinaygowd788kesam@gmail.com
K.Taraka Ram, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India. Email: tarakaramkolli514@gmail.com
G.Pradeepini, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India. Email: pradeepini_cse@kluniversity.in

II. LITERATURE SURVEY

A. M.P.N.M Wickramasinghe , D.M. Perera and K.A.D.C.P. Kahandawaarachchi [1]. The primary objective of their research study was to determine the acceptable diet plan for a patient with CKD by applying the classification algorithms to the test findings obtained from the medical records of patients. Finally, their work indicate that the Multiclass Decision Forest algorithm achieves a better result than the other classification algorithms with an accuracy of 99.17%.

B. In todays world, Individuals are very much conscious and concerned about their health. Individuals only give attention to their health only when have found some symptoms of health issues. These health issues occur majorly due to enormous workload and tensions. Nevertheless, CKD is a type of disease which does not show any kind of symptoms. In some cases, it is very tougher to predict whether the disease is present in the patient or not. In addition, this disease will lead to permanent health deterioration. Tekale, Siddheswar [6] performed a research such that they collected the data of CKD patients, which contains 14 attributes and 400 records and applied various machine learning algorithms like Decision tree and (SVM) through which a model was built to predict whether CKD is present in the patient or not with high accuracy.

C. Wang, Zixian , et al[4] conducted research on chronic kidney disease prediction using the Apriori classification strategy on medical records of patients(400 cases) with ten-fold validation testing. Finally,the results for selected data set features showed 99 percent apriori-based CKD detection accuracy.

III. ALGORITHM

Two machine learning algorithms were implemented on the dataset collected from the medical records of the patient. The goal of these algorithms is predicting whether a patient does have chronic kidney disease.

- K Nearest Neighbours.
- Logistic Regression.

A. K Nearest Neighbours

It is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. If there is a need for an unknown data, it looks for k more similar instances through the whole training dataset and eventually returns the data with a more similar instance as the prediction.To locate its neighbors,it uses the least distance measure.[12]

Working

The following two cases can explain the working in detail.

Case1:

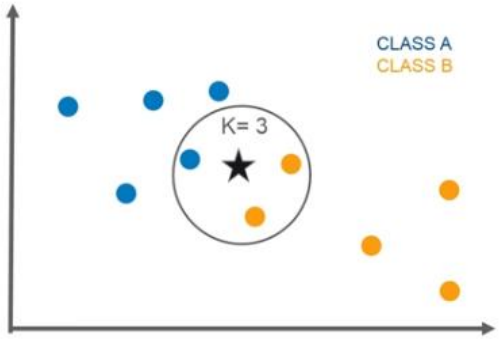

Fig. 1.selecting "k" value=3

In the Fig.1, let us assume that the blue points belong to Class A whereas the orange points belong to Class B. Now, our task is to predict the class to which the new star point belongs to. The initial step is to select the value of "k" which is nothing but the number of nearest neighbours that we want to select from the new point. In this case, for instance, let us consider the value of k is 3. It means that we are selecting three points that are at least distance from the new star point. Once we calculated the distance, we get one blue point and two orange points to be the closer points to the new star point. Since in this case, we are having a majority of orange points, we can confirmly say that for k=3, the new star point belongs to Class B or we can say that the star point is more similar to orange points.

Case2:

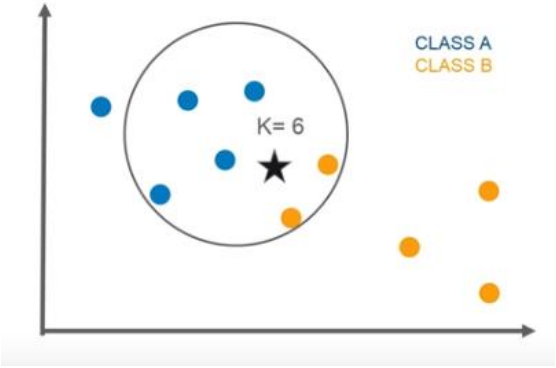

Fig. 2. selecting "k" value=6

In the above graph ie., Fig.2, let us assume that the blue points belong to Class A whereas the orange points belong to Class B. Now, our job is to determine the class to which the new star point belongs. The initial step is to select the value of "k" which is nothing but the number of nearest neighbours that we want to select from the new point. In this case,for instance, let us consider the value of k as 6. It means that we are selecting six points that are at least distance from the new star point. Once we calculated the distance, we get four blue points and two orange points to be the closer points to the new star point. Since in this case, we are having a majority of blue points, we can confirmly say that for k=6, the new star point belongs to Class A or we can say that the star point is more similar to blue points. In order to find the least distance between two points ,the algorithm mainly focuses on two distance measures.

- Euclidean distance

It is defined as the square root of the difference between a new point and the existing one.

Example: Suppose if we want to find the Eucidean Distance between two points P1(1,1) and P2(5,4) then the calculation can be done by using the formula stated in the below diagram.

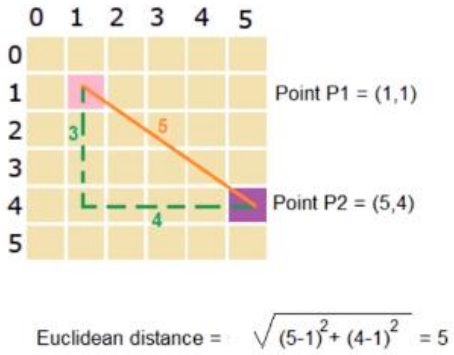

Fig. 3.Euclidean distance

- Manhattan distance

It is defined as the sum of two-point absolute difference.

Example:

The formula stated below in the diagram can be used to calculate the manhattan distance between any two points. In this calculation, we have considered two points as P1(1,1) and P2(5,4).

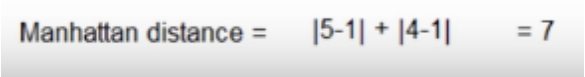

Fig. 4. Manhattan distance

B. Logistic Regression

- This is a technique used for traditional statistics as well as machine learning.
- It is also used for classification of data and it contains only two values as its output.

Eg: {0 & 1} or { True & False}

Working

The following graph can explain the working in detail.

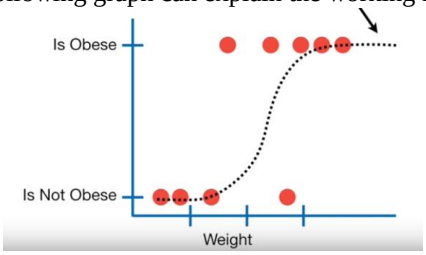

Fig. 5.S-curve

In the above graph , the "S" shaped curve goes from 0 (Is Not Obese) to 1(Is Obese) and this curve tells us the probability that is person is obese or not not based on their weight. Similarly, in our research work this curve tells whether a patient contains ckd or not ,based on the amount of blood potassium.

IV. METHODOLOGY

A. Data Collection

In this research paper we have used Real world data set for predicting CKD status of a patient. The data collected is widely used data and is available at UCI Machine Learning Repository. Real Data is owned by Apollo Hospital in Tamilnadu, India for 2 months. The data set is available exclusively used for research of Chronic Kidney Disease. It consists of a record 400 people with each 25 attributes related CKD them. The data consist of real numbers, decimal values and values Nominal.[3]

B. Data Pre-Processing

It is a way to convert the noisy and huge data into relevant and clean data, as the data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, to expel the conflicting information from the dataset, the proposed framework should tidy up the crude information Building Training and Test this dataset is a significant part to supplement the prescient model. This lessens the measurement and causes the motor to accomplish better outcomes. It is one of the most tedious advances in building an arrangement model. Searching Up for Proper Format: As we make our model utilizing python, so we have to utilise csv (comma isolated esteem) file for our code. Information is downloaded as a RAR document, with the goal that we remove information from a book record that is accessible and store it into a csv record so we can peruse them python code. This is the most significant initial step, if the information isn't accessible in the arrangement necessitates that we can not plan a grouping model. This step includes all actions in this project to develop the final dataset from the original raw dataset and we used blood potassium level as the main contributing attribute to test whether or not a patient has chronic kidney disease. The instances are graded as safe, caution and danger based on the blood potassium level quality.

C. Finding missing values

When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or on the other hand disregarded in the event that they are not huge in number. Missing esteems can likewise be supplanted by zero.

D. Data transformation

In this step, we added an extra attribute "Outcome" to the existing dataset. Also, we have applied a function on the outcome column such that the values of the outcome column will be safe, nodata and low. These values are obtained according to the range of blood potassium level there are certain conditions involved in the function which is applied on the outcome column.

TABLE I : CONDITIONS

Blood Potassium Level	Patients zone
>0 & < 3.5	Low
>=3.5 & <=5.0	safe
>=5.1 & <=6.0	Caution
>=6.1	danger
No Presence(0)	no data

Now,we create another function and apply it on newly created outcome column which has the values like "nodata","safe","low". The conditions of the function are stated in the below table.

TABLE II : ASSIGNMENT OF VALUES

Outcome column	value
safe	0
Low	1
No data	1

TABLE II clearly indicates that all the Nominal type data is converted into numerical binary values 0 & 1.The purpose of this convertion is that machine learning algorithms will be applied on numerical data. Furthermore, we have selected 20 random attributes from the list of 25 attributes and designed a predictive model in such a way that if we give any random values for the selected attributes, our model will detect the presence of ckd in a patient. It can be visually seen ie., if "0" is displayed after the processing of either of the algorithm(KNN & Logistic regression) on the selected attributes,then it is clear that the patient is not having ckd.However,if "1" is displayed ,then the patient is unsafe. Also,the calculation of the error ie., confusion matrix of either of the algorithms has been done. It is the predictor of the performance of our predictive model on classification problem ie., It tells us how much correctly our predictive model is predicting the presence of disease in the patient's body especially based on blood potassium levels. Finally, the classification report has been calculated.

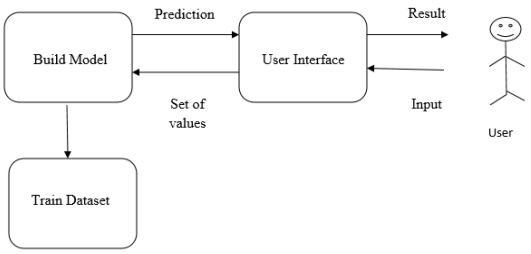

Fig.9. Architectural design

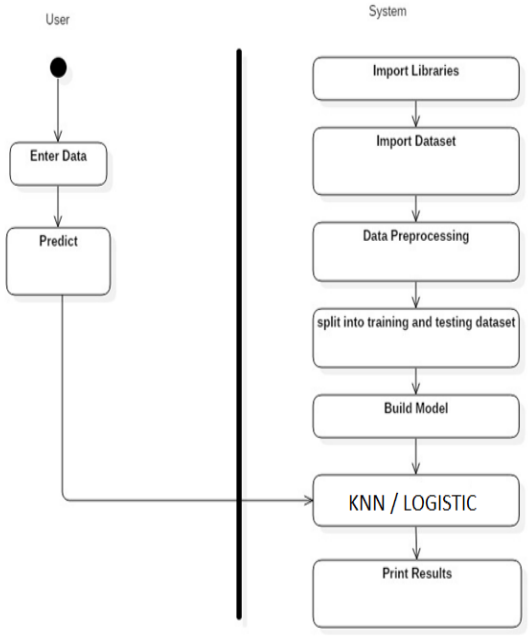

Fig.10.block diagram

E. Data Set

	0	1	2	3	4	5	6	7	8	9	...	390	391
id	0	1	2	3	4	5	6	7	8	9	...	390	391
age	48	7	62	48	51	60	68	24	52	53	...	52	36
bp	80	50	80	70	80	90	70	0	100	90	...	80	80
sg	1.02	1.02	1.01	1.005	1.01	1.015	1.01	1.015	1.015	1.02	...	1.025	1.025
al	1	4	2	4	2	3	0	2	3	2	...	0	0
su	0	0	3	0	0	0	0	4	0	0	...	0	0
rbc	0	0	normal	normal	normal	0	0	normal	normal	abnormal	...	normal	normal
pc	normal	normal	normal	abnormal	normal	0	normal	abnormal	abnormal	abnormal	...	normal	normal
pcc	notpresent	notpresent	notpresent	present	notpresent	notpresent	notpresent	notpresent	present	present	...	notpresent	notpresent
ba	notpresent	notpresent	notpresent	notpresent	notpresent	notpresent	notpresent	notpresent	notpresent	notpresent	...	notpresent	notpresent
bgr	121	0	423	117	106	74	100	410	138	70	...	99	85
bu	36	18	53	56	26	25	54	31	60	107	...	25	16
sc	1.2	0.8	1.8	3.8	1.4	1.1	24	1.1	1.9	7.2	...	0.8	1.1
sod	0	0	0	111	0	142	104	0	0	114	...	135	142
pot	0	0	0	2.5	0	3.2	4	0	0	3.7	...	3.7	4.1
hemo	15.4	11.3	9.6	11.2	11.6	12.2	12.4	12.4	10.8	9.5	...	15	15.6
pcv	44	38	31	32	35	39	36	44	33	29	...	52	44
wc	7800	6000	7500	6700	7300	7800	0	6900	9600	12100	...	6300	5800
rc	5.2	0	0	3.9	4.6	4.4	0	5	4.0	3.7	...	5.3	6.3
htn	yes	no	no	yes	no	yes	no	no	yes	yes	...	no	no
dm	yes	no	yes	no	no	yes	no	yes	yes	yes	...	no	no
cad	no	no	no	no	no	no	no	no	no	no	...	no	no
appet	good	good	poor	poor	good	good	good	good	good	poor	...	good	good
pe	no	no	no	yes	no	yes	no	yes	no	no	...	no	no
ane	no	no	yes	yes	no	no	no	no	yes	yes	...	no	no
classification	ckd	ckd	ckd	ckd	ckd	ckd	ckd	ckd	ckd	ckd	...	notckd	notckd
Outcome	nodata	nodata	nodata	Low	nodata	Low	safe	nodata	nodata	safe	...	safe	safe

Fig.11.Chronic kidney disease dataset

V. RESULT AND DISCUSSION

We used the jupyter notebook environment to make our predictive model because it seemed to be flexibe than R studio and it allowed us to share codes and documents.In addition, the individual cells can be compiled at a faster rate which will allow us to view the immediate result on the screen. It neglects the disadvantage of necessity of executing the code starting from the script.One bigger disadvantage is, we can't merge two jupyer notebooks.

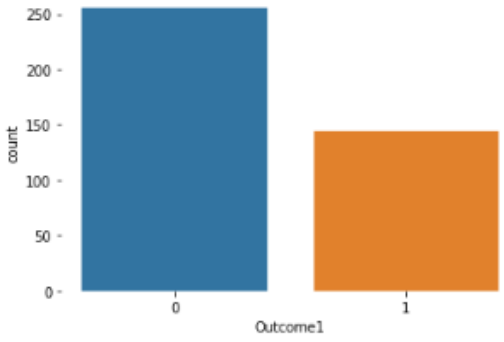

Fig. 12.safe&non-safe patients

Fig.12 indicates the count of safe and non-safe patients. Here, the blue colored barplot represents the proportion of safe patients and orange colored barplot represents the proportion of unsafe patients.

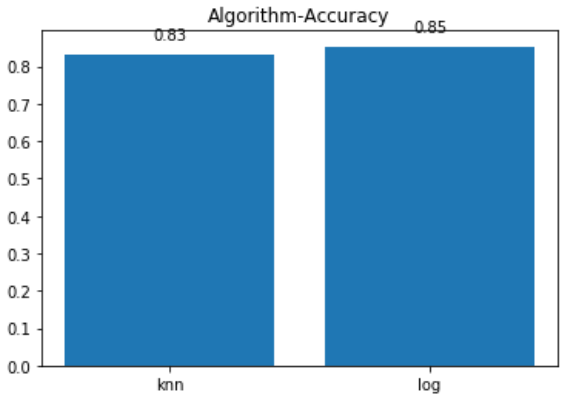

Fig.13. accuracy comparision

Fig.13 represents the bar-plot representation of both the prediction algorithms, KNN with 83% accuracy and Logistic Regression with an accuracy of 85%. In the above figure x-axis represents the calculation utilized and the y-hub is the precision esteem. The above outcomes feature that the precision of Logistic Regression calculation is 2% higher than KNN calculation. The outcomes demonstrated that interminable kidney ailment can be anticipated by utilizing Logistic Regression calculation with 85% exactness. The benefit of this investigation is that it will help specialists to effortlessly anticipate CKD with high exactness and accuracy in less timespan.

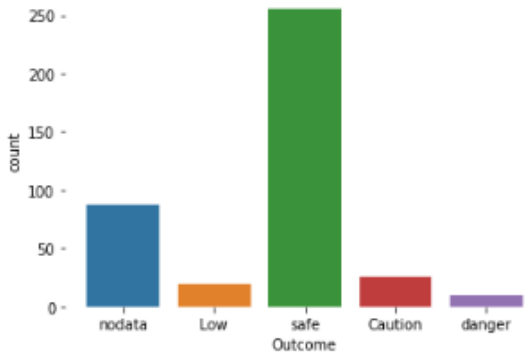

Fig.14.Classification of patients into different zones

Fig.14 indicates the range of patients falling into different zones like nodata ,Low, Safe, Caution and danger. The Classification has been done especially based on the blood potassium level present in the body of patient.

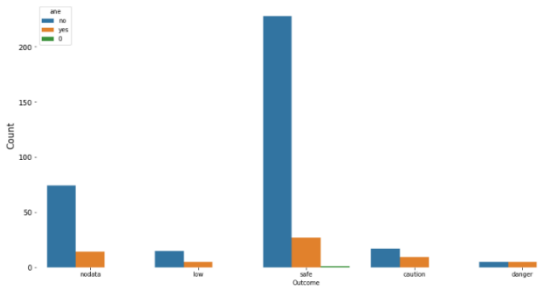

Fig.15. Classification based on anemia

Fig.15 indicates the further classification of outcome attribute based on anemia level. It means the number of patients belonging to different zones like nodata, Caution, Low, safe and danger are classified based on the "anemia" attribute. Likewise, Further classification based on the attributes like bacteria, appet and classification has been done.

	precision	recall	f1-score	support
0	0.79	0.91	0.85	64
1	0.78	0.58	0.67	36
accuracy			0.79	100
macro avg	0.79	0.74	0.76	100
weighted avg	0.79	0.79	0.78	100

Fig.16..Classification report using KNN algorithm

	precision	recall	f1-score	support
0	0.84	0.80	0.82	64
1	0.67	0.72	0.69	36
accuracy			0.77	100
macro avg	0.75	0.76	0.75	100
weighted avg	0.78	0.77	0.77	100

Fig.17.Classification report using Logistic regression

VI. CONCLUSION

We predicted whether not a person chronic kidney disease ,especially based on his blood potassium levels. Also, wehave Classified the patients in to safe and non safe category and also noted their count. We have mainly used two machine learning algorithms namely KNN and Logistic Regression for the purpose of prediction of disease. In our Predictive model, we have selected some main attributes from the dataset and by providing them our own values, we can get a clarity that a patient has chronic kidney disease or not. In addition to that, we have classified the patients on the basis of various attribues like bacteria ,appet, classification, anemia and all other attributes.We have detected the performance of our predictive model through the analysis of confusion matrix. We also compared the training and test data accuracy of the algorithms.Finally,we have made the classification report of both the algorithms which provides the values of key matrics ,which further gives us the information about the prediction of disease.

REFERENCES

1. Dietary Prediction for Patients with Chronic Kidney Disease (CKD) by considering Blood Potassium Level using Machine Learning Algorithms by M.P.N.M. Wickramasinghe, D.M. Perera, and K.A.D.C.P. Kahandawaarachchi
2. "Potassium and Your CKD Diet", *The National KidneyFoundation*.[Online]. Available:https://www.kidney.org/atoz/content/potassium.[Accessed: 24- Aug – 2017].
3. "UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set", Archive.ics.edu, 2015.[Online]. Available:http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.
4. Wang, Zixian, et al. "Machine Learning-Based Prediction System for Chronic Kidney Disease Using Associative Classification Technique." *International Jounal of Engineering & Technology*, www.sciencepubco.com/index.php/ijet/article/view/25377/12942

5. *Predictive Analytics for Chronic Kidney Disease Using MachineLearning Techniques- IEEE Conference Publication,* ieee.org/document/8025242.
6. Tekale, Siddheshwar. "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm." IJARCEE ,2018, ijarcce.com/wp-content/uploads/2018/11/IJARCCE 2018.71021.pdf
7. Pradeepini . G, Pradeepa G, Tejanagasri, B. ,Gorrepati , S.H "data Classification and personal care Management System by Machine Learning Approach " International Journal of Engineering and Technology 2018.
8. Anila M. & Pradeepini G , "Least Square Regression for Prediction Problem in Machine Learning Using R" International Journal of Engineering and Technology 2018.
9. Rajesh , N. Maneesha T. Hafeez & Krishna H. "Prediction of Heart Disease using Machine Learning Algorithms " International Journal of Engineering and Technology(UAE) 2018.
10. Razia S. Swathi Pratyusha P. Vamsi Krishna , N, &Sathya Sumana N. "A Comparative Study of Machine Learning Algorithms on Thyroid Disease Prediction "
11. International Journal of Engineering and Technology(UAE) 2018.
12. Wickramasinghe, M.P.N.M.,
13. M. Perera and K. A.D.C.P.Kahandawaarachchi. "Forecast diet for patients with interminal kidney illness (CKD) by considering the blood potassium levels utilizing AI calculations." In Life Sciences Conference (LSC) 2017 IEEE, pp.300-303. IEEE 2017.
14. Sinha , Parul and Poonam Sinha. "Near Study of constant kidney malady expectation utilizing KNN and SVM." International Journal of Research and Technology 4 Engineering, no.12 (2015):608-12.

AUTHORS PROFILE

S.D.Harish is pursuing his B.Tech in KL university. He is passionate about research and his area of interests are Machine Learning, Artificial Intelligence,Data mining, Network Security and Cloud Computing. In addition, he is also interested to work in health care organisations.

K.Vinay Kumar is pursuing his B.Tech in KL university. He is passionate about research and his area of interests are Artificial Intelligence, Deep Learning, Soft-Computing. He is very good in coding.

K.Taraka Ram is pursuing his B.Tech in KL university
He is passionate about research and his area of interests are Artificial Neural Networks and Deep learning.

Dr.G.Pradeepini is working as Professor in department of CSE in Koneru Lakshmaiah University. She had published papers in National & International Journals. She is having 19 years of teaching experience and 14 years of research experience. Her main research interest includes data mining and data warehouse, Big data and Neural Networks, Machine Learning and Deep Learning.