IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Kidney Disease Prediction using Machine Learning

Ms. K. Sri Vijaya[1], P. N. S. Sowmya[2], S. Dimpu Aman[3], V. Pavan Kumar[4], M. Keerthi[5]
Assistant Professor, Department of Information Technology[1]
B. Tech Students, Department of Information Technology[2,3,4,5]
Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *Chronic Kidney Disease is a serious, life-long disorder characterized by either kidney pathology or impaired kidney function. Early detection and treatment can potentially halt or slow the progression of chronic disease to the stage where dialysis or kidney transplantation are the only options for saving patients' lives. In this study, we look into the ability of various machine-learning techniques to detect chronic kidney disease early. Predictive analytics is used to evaluate the relationship between data parameters and target class attributes. It enables us to add the optimal subset of parameters to machine learning, which aids in the creation of a collection of predictive models. The experiment's findings indicate that advances in machine learning and analytic, represent a promising model to recognize the intelligent solutions, which in turn prove the ability of prediction in the kidney disease.*

Keywords: Chronic Kidney Disease, Confusion Matrix, KNN Classifier, Random Forest, Decision tree, LGBM, Classification

I. INTRODUCTION

Chronic Kidney Disease(CKD) is a significant public health problem for worldwide, especially for a low and medium-income countries. Chronic Kidney Disease(CKD) means that the kidney does work and cannot correctly filter the blood. About 10 percent of the population for a worldwide suffering from CKD, and millions of people die each year because they cannot get the affordable treatment, with the number increasing the elderly. On Global Burden Disease in 2010 a study was conducted by that International Society of Numerology; they reported that chronic kidney disease has been raised an important cause of the morality worldwide with the number of deaths increasing by 82.3 percent in the last two decades. Also, the number of patients reaching end-stage renal disease(ESRD) is increasing, which requires kidney transplantation or dialysis to save the patients' lives.

II. PROPOSED SYSTEM

It is better to search for an alternative approach with promising results and we are going to do that. We are going to compare four commonly known machine learning algorithms and are going to state the best among them using a case study on Chronic Kidney Disease prediction. As stated, the four machine learning algorithms are: Random Forest, Decision Tree, LGBM, and K-Near Neighbor(KNN).

III. TECHNOLOGIES USED

3.1 Google Colaboratory

Colaboratory, or "Colab" for short allows the users to write and execute the python code through the browser that runs entirely in the cloud. Unlike other python notebooks, colab does not require any setups to run the python code. It is the free cloud resource that can be viewed by anybody which is hosted by Google. Through this Google encourages the Machine Learning and artificial intelligence research.

3.2 Machine Learning

Machine Learning (ML) is a type of artificial intelligence (AI) that allows software application to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine Learning algorithms use historical data as input to predict new output values. Machine Learning is important because it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

3.3 Python Libraries

There are several libraries used in the data visualization in this application. The numpy is used for array operations and matrices. Pandas is used widely used for data analysis and ML tasks. The train_test_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. A classification_report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.

3.4 Data

The data is taken in the form of CSV file which contains the data in the form of rows and columns. The data has 26 columns and 400 rows. Each row indicates the details of individual person. Here we focus on classifying if each person is having a CKD or not using their data.

3.5 Supervised Learning

Supervised learning is the types of machine learning in which machines are trained using "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input is already tagged with the correct output. In our dataset the data is tagged with classification whether a person is having disease or not.
There are two types of supervised machine learning algorithms. They are:

1. Regression
2. Classification

In the project we are using classification algorithm. Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-False etc….

3.6 Methodologies

A. Random Forest

A Random Forest is a meta estimator that fits the number of the decision tree classifiers of various sub-samples of the dataset and uses averaging to improve the predictive accuracy & its control over fitting. The sub-sample size was always the same as the original input sample size, but the samples are drawn with replacement if bootstrap= True(default).

B. KNN Classifier

K-Nearest Neighbors is simplest algorithms used in machine learning for regression and classification problem. KNN algorithm is used and classify the new data points based on the similarity measures.

C. LGBM

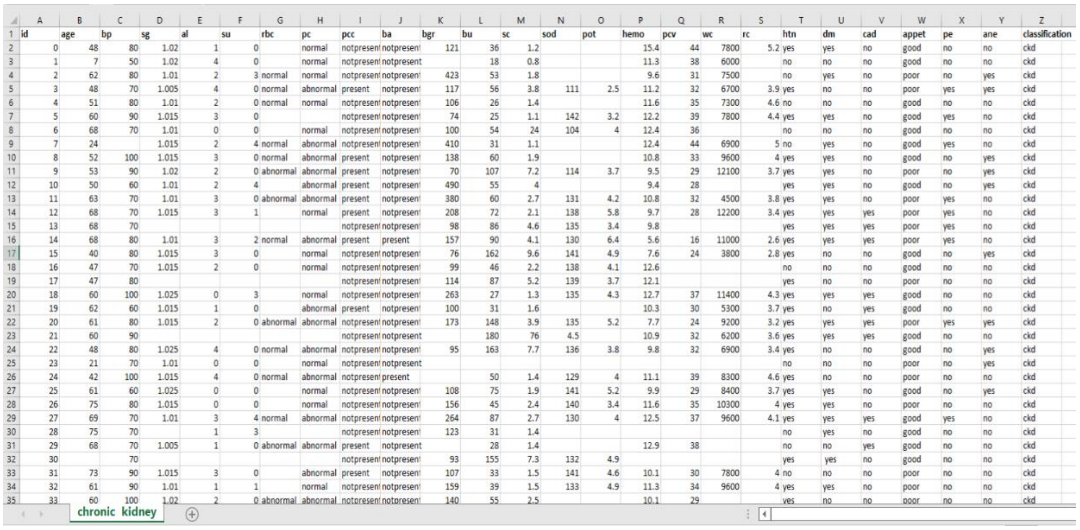
Light GBM can handle the large size of the data and takes least memory to run. Another reason of why Light GBM is popular is because it focuses on the accuracy of results. LGBM also supports GPU learning and thus data scientists are widely using LGBM for data science application development.

D. Decision Tree

A decision tree is a classifier understood as an instance space recursive partitioning. It is composed of nodes that to build a rooted tree, more precisely, it is a directed tree with a node named "root" that does not have any incoming edge.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

IV. RESULTS



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48	80	1.02	1	0		normal	notpresent	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notpresent	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
5	60	90	1.015	3	0			notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68	70	1.01	0	0		normal	notpresent	notpresent	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
7	24		1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
9	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50	60	1.01	2	4		abnormal	present	notpresent	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
11	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68	70	1.015	3	1		normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	no	poor	yes	no	ckd
13	68	70						notpresent	notpresent	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
15	40	80	1.015	3	0		normal	notpresent	notpresent	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
16	47	70	1.015	2	0		normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
17	47	80						notpresent	notpresent	114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
18	60	100	1.025	0	3		normal	notpresent	notpresent	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
19	62	60	1.015	1	0		abnormal	present	notpresent	100	31	1.6			10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
21	60	90						notpresent	notpresent		180	76	4.5		10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
23	21	70	1.01	0	0		normal	notpresent	notpresent										no	no	no	poor	no	yes	ckd
24	42	100	1.015	4	0	normal	abnormal	notpresent	present		50	1.4	129	4	11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd
25	61	60	1.025	0	0		normal	notpresent	notpresent	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd
26	75	80	1.015	0	0		normal	notpresent	notpresent	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes	no	poor	no	no	ckd
27	69	70	1.01	3	4	normal	abnormal	notpresent	notpresent	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	ckd
28	75	70		1	3			notpresent	notpresent	123	31	1.4							no	yes	no	good	no	no	ckd
29	68	70	1.005	1	0	abnormal	abnormal	present	notpresent		28	1.4			12.9	38			no	no	yes	good	no	no	ckd
30		70						notpresent	notpresent	93	155	7.3	132	4.9					yes	yes	no	good	no	no	ckd
31	73	90	1.015	3	0		abnormal	present	notpresent	107	33	1.5	141	4.6	10.1	30	7800	4	no	no	no	poor	no	no	ckd
32	61	90	1.01	1	1		normal	notpresent	notpresent	159	39	1.5	133	4.9	11.3	34	9600	4	yes	yes	no	poor	no	no	ckd
33	60	100	1.02	2	0	abnormal	abnormal	notpresent	notpresent	140	55	2.5			10.1	29			yes	no	no	poor	no	no	ckd

chronic kidney

Fig: 4.1 First few rows of the dataset used

The above figure gives the overview of the details of the patients' blood and urine test data along with the classification if the patient is having the disease or not.

	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo
count	400.000000	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	199.500000	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	115.614301	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	0.000000	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	99.750000	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	199.500000	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	299.250000	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	399.000000	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

Fig: 4.2 Description of the dataset

The above table gives the description of the dataset that is taken by us. The description includes the total count, mean, standard deviation(std), minimum value, percentiles(25%,50%,75%), maximum value of each and every column present in the dataset taken.

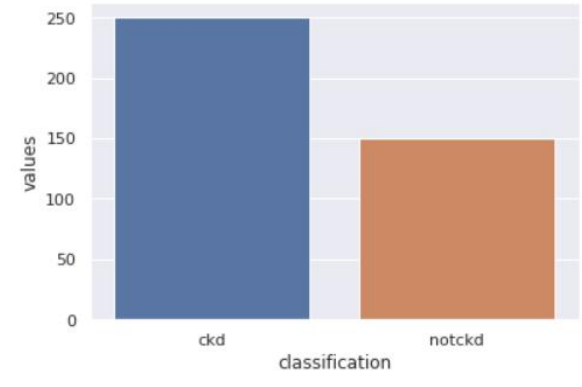


Fig: 4.3 Classification plot analysis

The above bar graph shows us that there are large number of persons having the Chronic Kidney Disease(ckd) i.e., almost 250 persons than the persons not having the disease(notckd) i.e., up to 150 persons.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

```
print(f"Confusion Matrix :- \n{confusion_matrix(y_test, knn.predict(x_test))}\n")
print(f"Classification Report :- \n {classification_report(y_test, knn.predict(x_test))}")
```

```
Training Accuracy of KNN is 0.98
Test Accuracy of KNN is 0.98

Confusion Matrix :-
[[46  2]
 [ 0 52]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        48
           1       0.96      1.00      0.98        52

    accuracy                           0.98       100
   macro avg       0.98      0.98      0.98       100
weighted avg       0.98      0.98      0.98       100
```

Fig: 4.4 KNN Testing Data

Here we have precision and Recall, upon using the confusion matrix we have received the accuracy of 98 percentage

```
[ ]
print(f"Confusion Matrix :- \n{confusion_matrix(y_test, dtc.predict(x_test))}\n")
print(f"Classification Report :- \n {classification_report(y_test, dtc.predict(x_test))}")
```

```
Training Accuracy of Decision Tree Classifier is 1.0
Test Accuracy of Decision Tree Classifier is 1.0

Confusion Matrix :-
[[48  0]
 [ 0 52]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        52

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Fig: 4.5 Decision Tree Testing data

Here we have precision and Recall, upon using the confusion matrix we have received the accuracy of 100 percentage

```
print(f"{confusion_matrix(y_test, lgbm.predict(x_test))}\n")
print(classification_report(y_test, lgbm.predict(x_test)))
```

```
Training Accuracy of LGBM Classifier is 1.0
Test Accuracy of LGBM Classifier is 1.0

[[48  0]
 [ 0 52]]

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        52

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Fig: 4.6 LGBM Testing Data

Here we have precision and recall, upon using the confusion matrix we have received the accuracy of 100 percentage

```
print(f"Confusion Matrix :- \n{confusion_matrix(y_test, rd_clf.predict(x_test))}\n")
print(f"Classification Report :- \n {classification_report(y_test, rd_clf.predict(x_test))}")
```

```
Training Accuracy of Random Forest Classifier is 1.0
Test Accuracy of Random Forest Classifier is 1.0

Confusion Matrix :-
[[48  0]
 [ 0 52]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        52

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Fig: 4.7 Random Forest Testing data

Here we have precision and recall, upon using the confusion matrix we have received the accuracy of 100 percentage.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

V. SCOPE OF FUTURE USE

As a complementary solution, we could build a mobile application, to make it available to everyone to detect whether they are CKD or not and as of now if we have a correct dataset, we could use it to predict many more diseases and if the application collaborates with a highly reputed company, then we can allow making more advancements to the existing model. There would be disadvantages because of that we need to have higher graphic card compatibility to run the model and the solution for that would be, the involvement of a framework such as Kara's or Tensor-Flow at the backend.

VI. CONCLUSION

1. The disease prediction performs depending on the machine learning algorithms (which we have used) and anticipates whether the patients had any disease or not.
2. Comparing to Decision Tree, LGBM, Random Forest we receive less accuracy in k-Nearest Neighbor algorithm.
3. Data collection and prediction is carried out thoroughly and tested.

REFERENCES

[1]. J. Alijaaf et al, "Early prediction of chronic kidney disease using machine learning supported by predictive analytic", in 2020 IEEE Congress on Evolutionary Computation (CEC), 2020.
[2]. S. Vijayarani et al, "International Journal of Computing and Business Research (IJCBR), vol 6, no 2,(2019).
[3]. T Shaikhina, Torgyn, et al, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation". Biomedical Signal Processing and Control (2019).
[4]. Jaymin Patel, prof.Tejal Upadhyay, Dr. Samir Patel et al, "Heart Disease Prediction Using Machine Learning and Data Mining Technique", International Journal of Computer Science Communication, Vol 7, No 1, pp.129-137(2019).
[5]. C.T. Tranet al., Multiple Imputation and Ensemble Learning for Classification with Incomplete Data, Springer International Publishing, pp. 401-415(2020)
[6]. Sathiya Priya S, Suresh Kumar M, Chronic Kidney Disease Prediction Using Machine Learning, Sri Ramakrishna Engineering College, Coimbatore., International Journal of Computer Science and Information Security (IJC-SIS), vol 16, no 4, April 2019.
[7]. J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression", Journal of Transnational Medicine, vol 17, (1), pp. 119,2019.