BioGecko Vol 12 Issue 01 2023

ISSN NO: 2230-5807

DETECTION OF CHRONIC KIDNEY DISEASE APPLYING THE RANDOM FOREST ALGORITHM

Melam Naga Raju
Department of Information Technology
Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.
melam.nagaraju5810@gmail.com

D.M.K Venkata Subbarao
Department of Information Technology
Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.
dmkvsubbarao@gmail.com

K.V Kavya Sri
Department of Information Technology
Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.
kalavalakavyateja@gmail.com

E Naga Sudha Rani
Department of Information Technology
Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.
sudharanidcme2@gmail.com

ABSTRACT

The prevalence of morbidity and mortality from chronic kidney disease (CKD), as well as the emergence of additional ailments, makes it a global health concern. People commonly overlook CKD in its early stages since there are no obvious symptoms. Early detection of CKD enables patients to receive immediate treatment to halt the progression of the condition. Doctors can successfully accomplish this goal with the help of machine learning models because of their quick and accurate identifying capabilities. In this research, we provide a machine learning approach for CKD diagnosis. The University of California, Irvine's machine learning repository served as the source of the CKD data set (UCI). Hence, it will establish whether a patient has CKD and, if so, how severe it isfurther drugs should be taken. In this we will use random forest algorithm to determine whether a person is suffering from chronic kidney disease or not.

1. INTRODUCTION

A global public health issue, CHRONIC KIDNEY DISEASE (CKD) affects roughly 10% of the world's population. In China, the prevalence of CKD is 10.8%, whereas in the United States, it ranges from 10% to 15%. Another survey indicates that the overall adult population of Mexico has this percentage at 14.7%. This illness is characterised by a gradual decline in renal function that ultimately results in a total loss of renal function. Early on, CKD does not have noticeable symptoms. Because of this, the illness might not be discovered until the kidney has lost around 25% of its functionality. Moreover, CKD affects the human body globally and has a high rate of morbidity and mortality.It can cause cardiovascular disease to develop. CKD is a pathologic illness that progresses and cannot be reversed. Thus, it is crucial to detect and diagnose CKD in its early stages so that patients can start therapy right away and slow the disease's progression. For instance, models created by machine learning algorithms were used to identify diseases like acute renal injury, cancer, heart disease, diabetes, and retinopathy.

Furthermore, the majority of earlier investigations made use of the CKD data set that was accessible from the UCI machine learning repository.

Early CKD frequently does not exhibit any symptoms. This is because a significant decline in kidney function can usually be tolerated by the human body. Unless a normal examination for another condition, such as a blood or urine test, uncovers a potential concern, kidney disease is frequently not recognised until this stage. Early detection can help prevent it from advancing to a more advanced form, as can drug treatment and continuing monitoring with regular testing. Therefore, it is crucial to identify kidney disease in its earliest stages in individuals..

2. PROBLEM STATEMENT

Training and testing the model are necessary before using the training dataset. For the purpose of detecting CKD, various techniques including decision trees and XGBoost were combined with trained models. The system receives the patient input parameters as input. Employ machine learning techniques to examine the numbers by using the generated graphs, figures and perform a diagnostic. Less exact. greater likelihood of errors occurring.

3.PROPOSED SYSTEM

The component models that performed better while diagnosing the data samples were chosen for inclusion. This section proposes numerous machine learning models. The roles of the component model as prospective biomarkers were identified by analysis of the component model's errors. Very high accuracy. Less Susceptible to Error. Quick and practical.

4. ALGORITHMS

After data preprocessing, modelling, usually referred to as model selection, is a crucial stage. The process of choosing a model for a prediction problem from a large pool of models is known as model selection. Predicting whether a person has chronic kidney disease or not is our dilemma. The techniques KNN, SVM, XGboost, Adaboost, Decision Tree Classifier, Logistic Regression, and Random Forest Classifier can all be used to achieve this. Because it is a decision-making algorithm, the Random Forest classifier algorithm is being used in this study.The entire set of data was divided into two categories: train data (75%), and test data (25%). Next, many models are considered.

4.1 Random Forest

One kind of supervised classifier is the random tree. It generates a large number of unique learners. The tree is created using the stochastic method. It is a particular kind of categorization technique for ensemble learning. Similar to a decision tree, except with each split using a random subset of attributes. Both classification and regression problems are addressed by this approach.A forest is a collection of unrelated trees. The random trees classifier classifies input for each tree in the forest using the input feature set. The random tree's output chooses from the majority of votes. Each leaf node of the tree contains a linear model. The model is trained using the bagging training technique.

Tenfold cross-validation and other performance evaluation measures were used in this work to assess the proposed models' performance. Before implementing feature selection, seven machine-learning models—Logistic Regression, SVM, Decision Tree Classifier and Extreme Gradient Boosting (XGBoost), ADABoost, Random Forest Classifier, and KNN—have been developed. SVM achieved a 99.9% accuracy rate. RF produced 100% accuracy.XGBoost produced 100% accurate results. With 99.0% accuracy, logistic regression produced results. The accuracy of the Decision Tree Classifier was 100%. KNN produced 100% accuracy. Using ADABoost, accuracy was 100%. However, as we can observe in fig. 4.1.4(a), the test accuracy of the ADABoost and KNN is not as accurate as the adaboost, random forest classifier, and xgboost. The outcome is encouraging, and we think it may be used to help medical professionals quickly and accurately detect the disease. In light of its accuracy and performance evaluation in comparison to seven-class classification algorithms, Random Forest is thus advised in our study.

	Model Name	Train Accuracy(%)	Test Accuracy(%)	AUC Score
0	Logistic Regression	99.056604	98.333333	0.999701
1	Decision Tree Classifier	100.000000	100.000000	1.000000
2	AdaBoost	100.000000	98.333333	1.000000
3	Random Forest Classifier	100.000000	100.000000	1.000000
4	kNN	100.000000	94.166667	0.953947
5	SVM	99.528302	98.333333	0.997907
6	XGBoost	100.000000	100.000000	1.000000

Accuracy of seven algorithms

5. IMPLEMENTATION

5.1 DATA COLLECTION

We were seeking for a dataset online that contained details on a person's test results, including their age, blood pressure, sugar levels, and other information. Data was gathered from the Kaggle website [1], which contains about 20 attributes, including the number of red blood cells per millilitre and their normal and abnormal values, patient albumin range, patient sodium range, patient potassium range, patient haemoglobin, patient white blood cell counts per microliter, and patient serum creatinine range. Also, it had about 5 million rows of data.The number of rows we used from that data was around 80 thousand. Some information relates to the ckd, and the rest to the nockd. The following data view is provided.

```
df = pd.read_csv('kidney.csv')
    data = df
data.head()
```

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44.0	7800.0	5.2	yes	yes	no	good	no	no	ckd
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38.0	6000.0	NaN	no	no	no	good	no	no	ckd
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	31.0	7500.0	NaN	no	yes	no	poor	no	yes	ckd
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	32.0	6700.0	3.9	yes	no	no	poor	yes	yes	ckd
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	35.0	7300.0	4.6	no	no	no	good	no	no	ckd

Data collection

5.2 DATA VISUALIZATION:

Using various graphs, charts, plots, and other visual aids, data visualization makes it easier to understand the data. In order to appropriately examine the data, it is depicted here using a scatter plot, bar graph, and histogram.

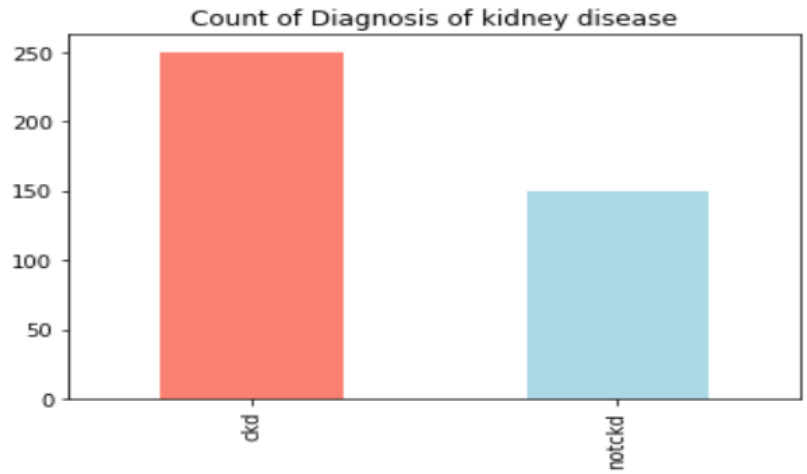
	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc
age	1.000000	0.159480	-0.191096	0.122091	0.220866	0.244992	0.196985	0.132531	-0.100046	0.058377	-0.192928	-0.242119	0.118339	-0.268896
bp	0.159480	1.000000	-0.218836	0.160689	0.222576	0.160193	0.188517	0.146222	-0.116422	0.075151	-0.306540	-0.326319	0.029753	-0.261936
sg	-0.191096	-0.218836	1.000000	-0.469760	-0.296234	-0.374710	-0.314295	-0.361473	0.412190	-0.072787	0.602582	0.603560	-0.236215	0.579476
al	0.122091	0.160689	-0.469760	1.000000	0.269305	0.379464	0.453528	0.399198	-0.459896	0.129038	-0.634632	-0.611891	0.231989	-0.566437
su	0.220866	0.222576	-0.296234	0.269305	1.000000	0.717827	0.168583	0.223244	-0.131776	0.219450	-0.224775	-0.239189	0.184893	-0.237448
bgr	0.244992	0.160193	-0.374710	0.379464	0.717827	1.000000	0.143322	0.114875	-0.267848	0.066966	-0.306189	-0.301385	0.150015	-0.281541
bu	0.196985	0.188517	-0.314295	0.453528	0.168583	0.143322	1.000000	0.586368	-0.323054	0.357049	-0.610360	-0.607621	0.050462	-0.579087
sc	0.132531	0.146222	-0.361473	0.399198	0.223244	0.114875	0.586368	1.000000	-0.690158	0.326107	-0.401670	-0.404193	-0.006390	-0.400852
sod	-0.100046	-0.116422	0.412190	-0.459896	-0.131776	-0.267848	-0.323054	-0.690158	1.000000	0.097887	0.365183	0.376914	0.007277	0.344873
pot	0.058377	0.075151	-0.072787	0.129038	0.219450	0.066966	0.357049	0.326107	0.097887	1.000000	-0.133746	-0.163182	-0.105576	-0.158309
hemo	-0.192928	-0.306540	0.602582	-0.634632	-0.224775	-0.306189	-0.610360	-0.401670	0.365183	-0.133746	1.000000	0.895382	-0.169413	0.798880
pcv	-0.242119	-0.326319	0.603560	-0.611891	-0.239189	-0.301385	-0.607621	-0.404193	0.376914	-0.163182	0.895382	1.000000	-0.197022	0.791625
wbcc	0.118339	0.029753	-0.236215	0.231989	0.184893	0.150015	0.050462	-0.006390	0.007277	-0.105576	-0.169413	-0.197022	1.000000	-0.158163
rbcc	-0.268896	-0.261936	0.579476	-0.566437	-0.237448	-0.281541	-0.579087	-0.400852	0.344873	-0.158309	0.798880	0.791625	-0.158163	1.000000

Correlation matrix and matrix visualization

	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc
age	1.000000	0.159480	-0.191096	0.122091	0.220866	0.244992	0.196985	0.132531	-0.100046	0.058377	-0.192928	-0.242119	0.118339	-0.268896
bp	0.159480	1.000000	-0.218836	0.160689	0.222576	0.160193	0.188517	0.146222	-0.116422	0.075151	-0.306540	-0.326319	0.029753	-0.261936
sg	-0.191096	-0.218836	1.000000	-0.469760	-0.296234	-0.374710	-0.314295	-0.361473	0.412190	-0.072787	0.602582	0.603560	-0.236215	0.579476
al	0.122091	0.160689	-0.469760	1.000000	0.269305	0.379464	0.453528	0.399198	-0.459896	0.129038	-0.634632	-0.611891	0.231989	-0.566437
su	0.220866	0.222576	-0.296234	0.269305	1.000000	0.717827	0.168583	0.223244	-0.131776	0.219450	-0.224775	-0.239189	0.184893	-0.237448
bgr	0.244992	0.160193	-0.374710	0.379464	0.717827	1.000000	0.143322	0.114875	-0.267848	0.066966	-0.306189	-0.301385	0.150015	-0.281541
bu	0.196985	0.188517	-0.314295	0.453528	0.168583	0.143322	1.000000	0.586368	-0.323054	0.357049	-0.610360	-0.607621	0.050462	-0.579087
sc	0.132531	0.146222	-0.361473	0.399198	0.223244	0.114875	0.586368	1.000000	-0.690158	0.326107	-0.401670	-0.404193	-0.006390	-0.400852
sod	-0.100046	-0.116422	0.412190	-0.459896	-0.131776	-0.267848	-0.323054	-0.690158	1.000000	0.097887	0.365183	0.376914	0.007277	0.344873
pot	0.058377	0.075151	-0.072787	0.129038	0.219450	0.066966	0.357049	0.326107	0.097887	1.000000	-0.133746	-0.163182	-0.105576	-0.158309
hemo	-0.192928	-0.306540	0.602582	-0.634632	-0.224775	-0.306189	-0.610360	-0.401670	0.365183	-0.133746	1.000000	0.895382	-0.169413	0.798880
pcv	-0.242119	-0.326319	0.603560	-0.611891	-0.239189	-0.301385	-0.607621	-0.404193	0.376914	-0.163182	0.895382	1.000000	-0.197022	0.791625
wbcc	0.118339	0.029753	-0.236215	0.231989	0.184893	0.150015	0.050462	-0.006390	0.007277	-0.105576	-0.169413	-0.197022	1.000000	-0.158163
rbcc	-0.268896	-0.261936	0.579476	-0.566437	-0.237448	-0.281541	-0.579087	-0.400852	0.344873	-0.158309	0.798880	0.791625	-0.158163	1.000000

Correlation matrix and matrix visualization.

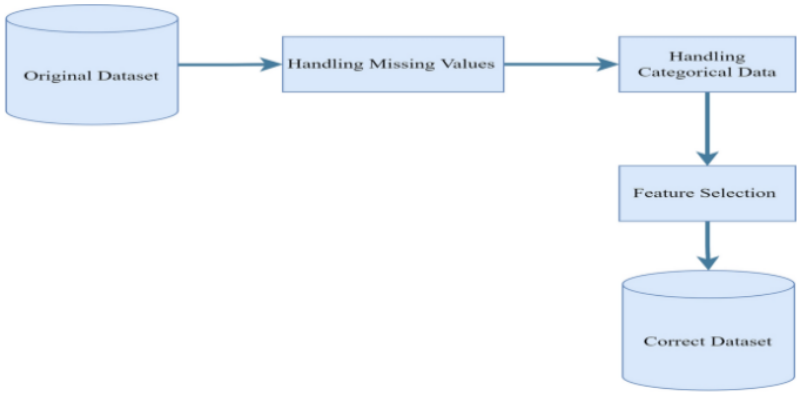
CKD and NON-CKD numbers are represented as a bar graph.



CKD and NON-CDK count

5.3 DATA PREPROCESSING

Data preprocessing is a key step in the cleaning of the data in machine learning. In essence, it is a procedure to transform unclean data into clean data. Here, data cleansing will be carried out in two ways: Without data, Noisy data are those beyond the specified range.



Data preprocessing

5.4 Missing Data

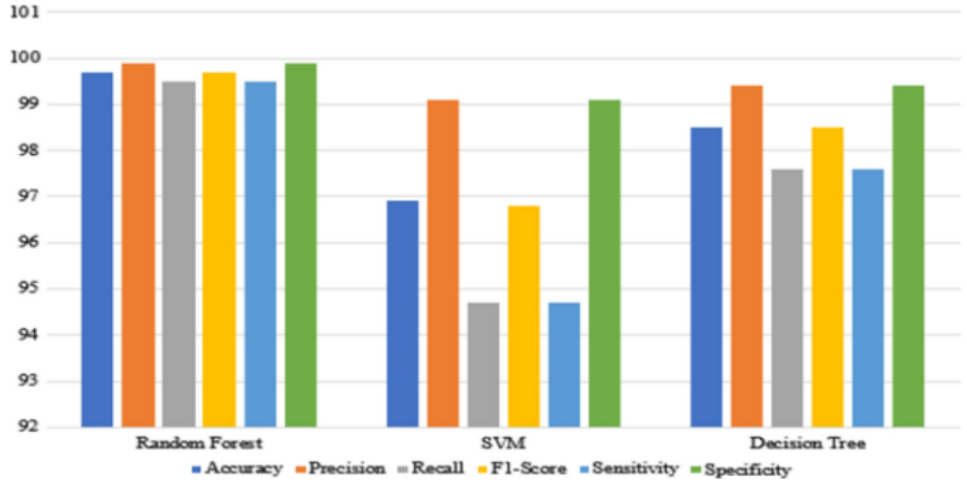
Before cleaning, there were many missing values that were filled using the median approach; total missing values are dropped during cleaning. There are other approaches as well, such as mean and mode, but mode can occasionally produce biassed results. Thus, the mean and median are preferred.

Variables	Explain	Class	Scale	Missing Rate
age	Age	Numerical	age in years	2.25%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020,1.025)	11.75%
al	Albumin	Nominal	(0,1,2,3,4,5)	11.5%
su	Sugar	Nominal	(0,1,2,3,4,5)	12.25%
rbc	Red Blood Cells	Nominal	(normal,abnormal)	38%
pc	Pus Cell	Nominal	(normal,abnormal)	16.25%
pcc	Pus Cell clumps	Nominal	(present,notpresent)	1%
ba	Bacteria	Nominal	(present,notpresent)	1%
bgr	Blood Glucose Random	Numerical	in mgs/dl	11%
bu	Blood Urea	Numerical	in mgs/dl	4.75%
sc	Serum Creatinine	Numerical	in mgs/dl	4.25%
sod	Sodium	Numerical	in mEq/L	21.75%
pot	Potassium	Numerical	in mEq/L	22%
hemo	Hemoglobin	Numerical	in gms	13%
pcv	Packed Cell Volume	Numerical	-	17.75%
wbcc	White Blood Cell Count	Numerical	in cells/cumm	26.5%
rbcc	Red Blood Cell Count	Numerical	in millions/cmm	32.75%
htn	Hypertension	Nominal	(yes,no)	0.5%
dm	Diabetes Mellitus	Nominal	(yes,no)	0.5%
cad	Coronary Artery Disease	Nominal	(yes,no)	0.5%
appet	appet	Nominal	(good,poor)	0.25%
pe	Pedal Edema	Nominal	(yes,no)	0.25%
ane	Anemia	Nominal	(yes,no)	0.25%
class	Class	Nominal	(ckd,notckd)	0%

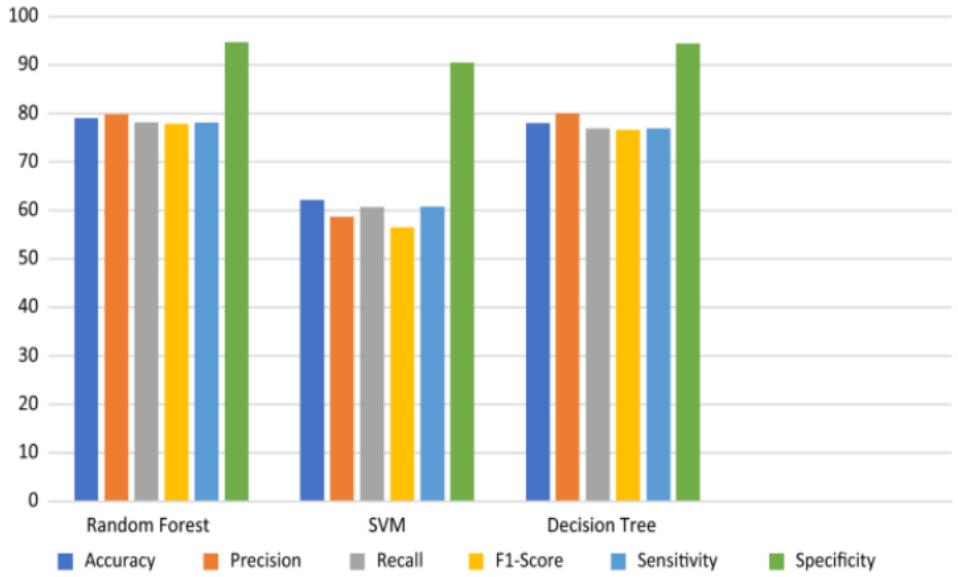
Missing values

5.5 FEATURE SELECTION:

It is often referred to as variable or attribute selection in machine learning. In essence, this is the process of choosing the attributes that have the greatest impact on the goal variable. Thus, whether or not a person has an illness is the target value or prediction value. A correlation matrix shows how highly correlated the variables are. The matrix indicates that each independent variable is significant for the prediction variable because they all influence it. Although there are further methods, we have chosen features using correlation analysis.



Classification without feature selection



Classification with feature selection

6. IMPLEMENTATION SCREENSHOTS

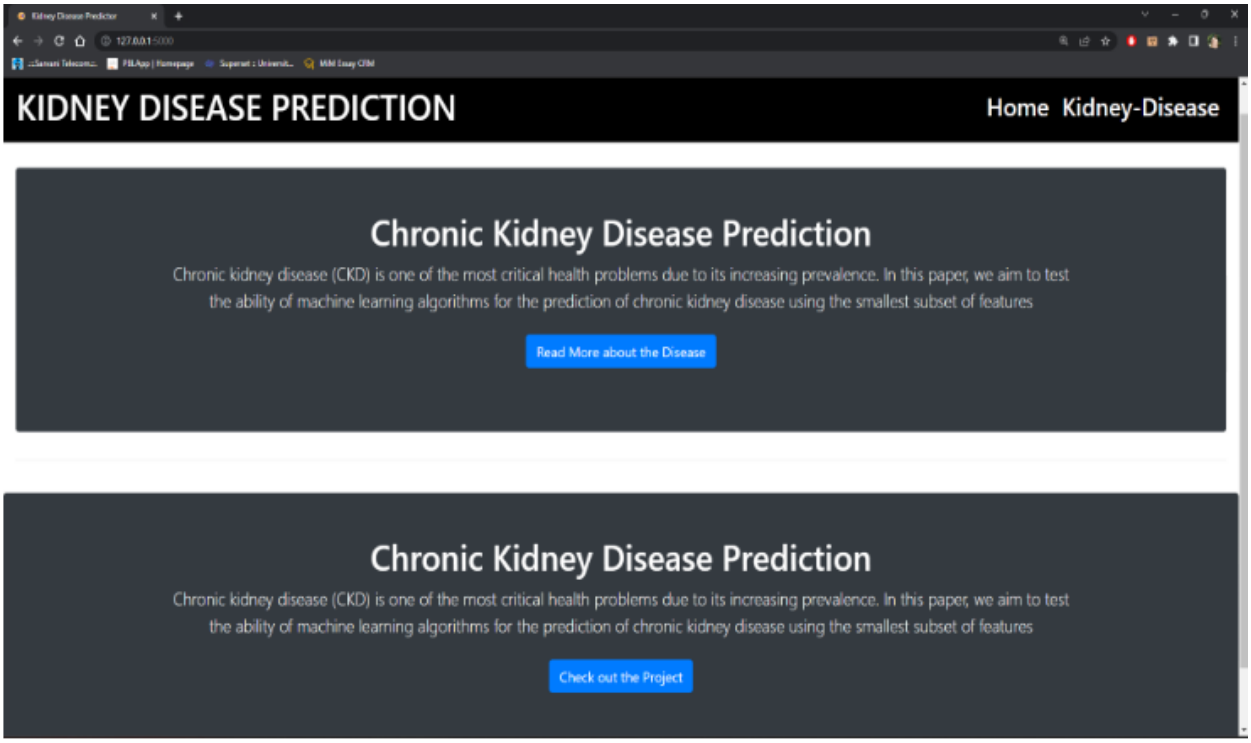


Fig 6.1 Home page is generated

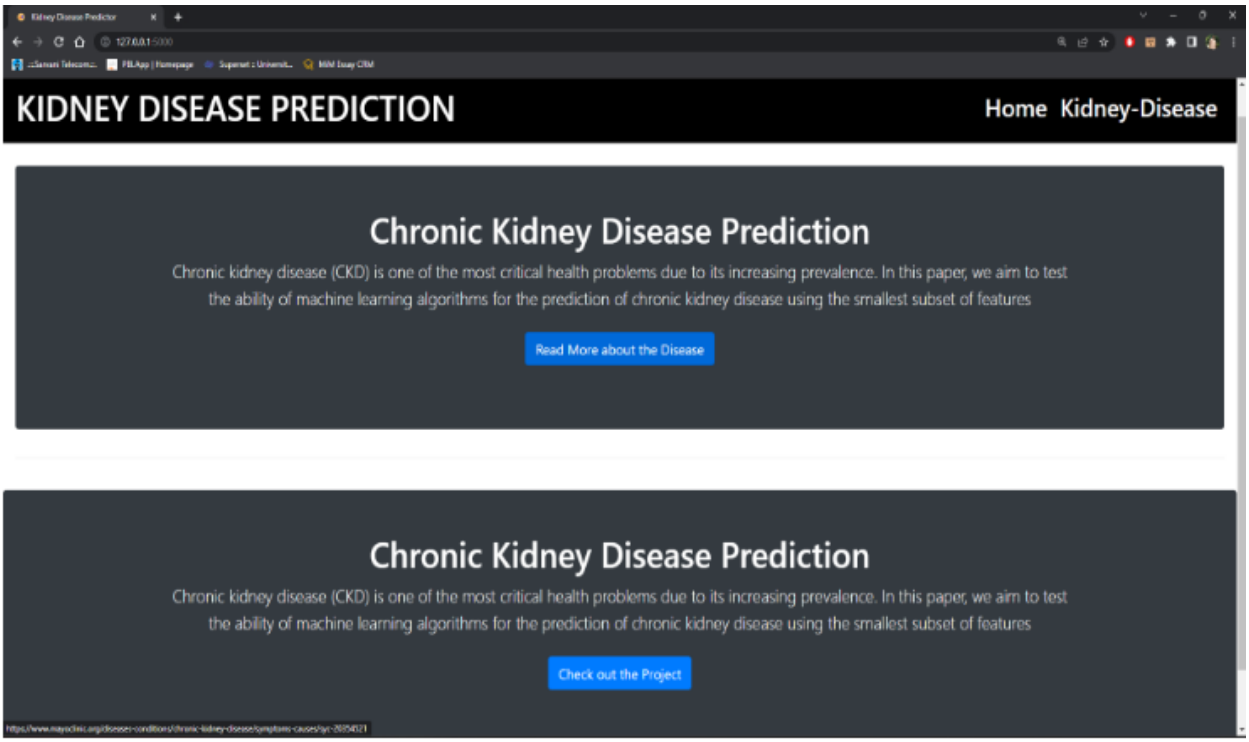


Fig 6.2 Click on read more about the disease to know the details of kidney disease

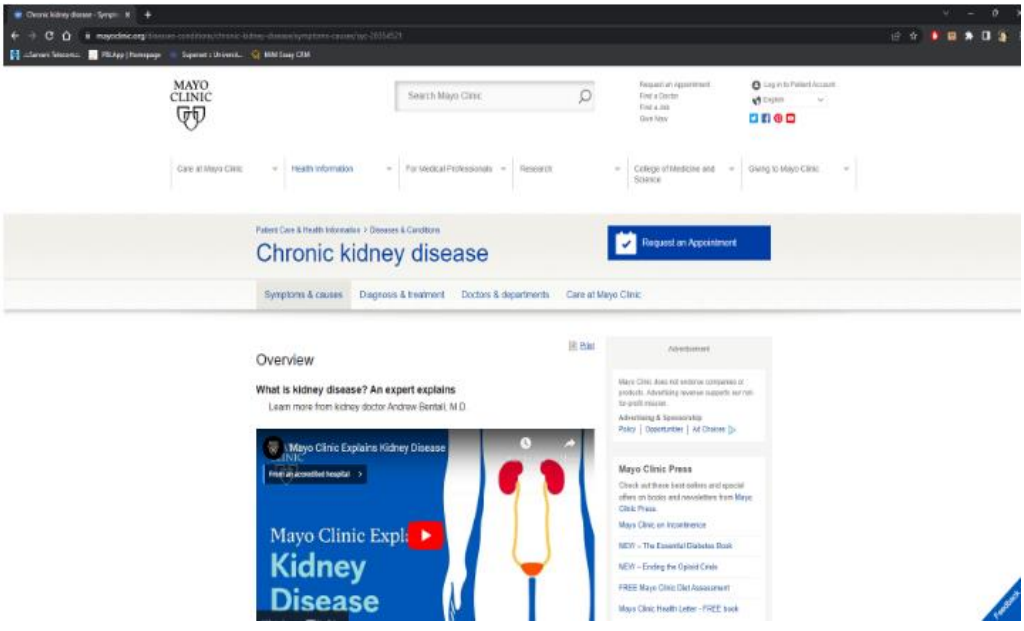


Fig 6.3 Details of the kidney disease will be shown in this page

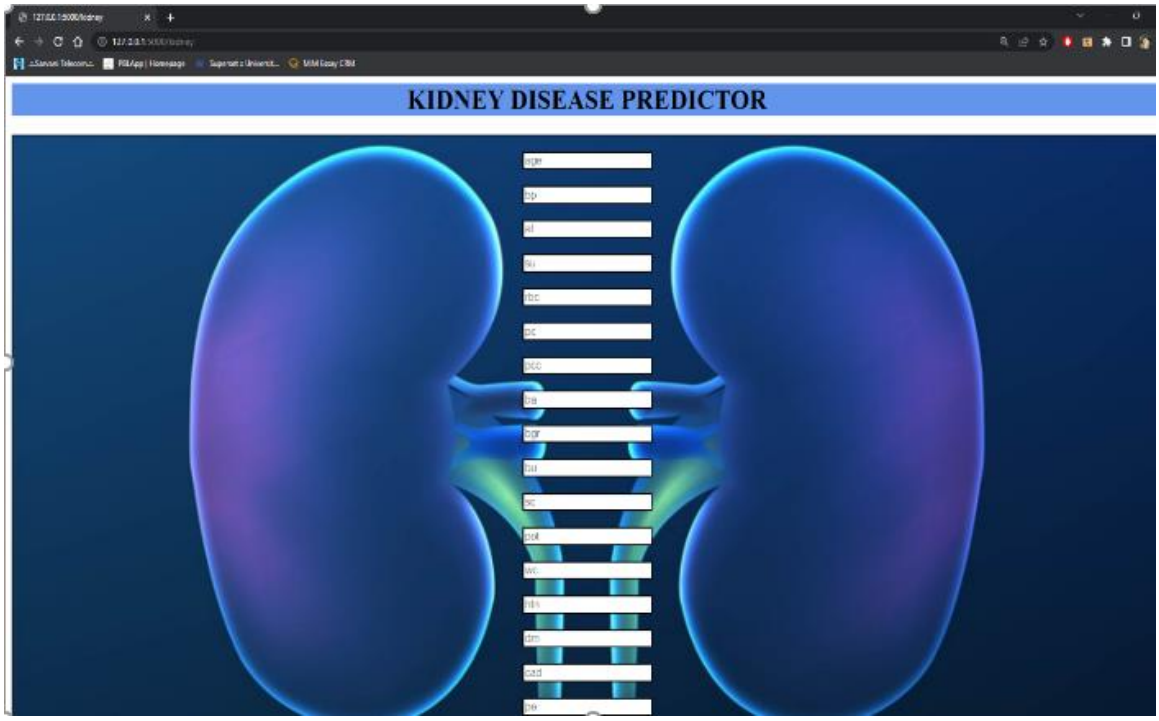


Fig 6.4 Kidney disease predictor web page

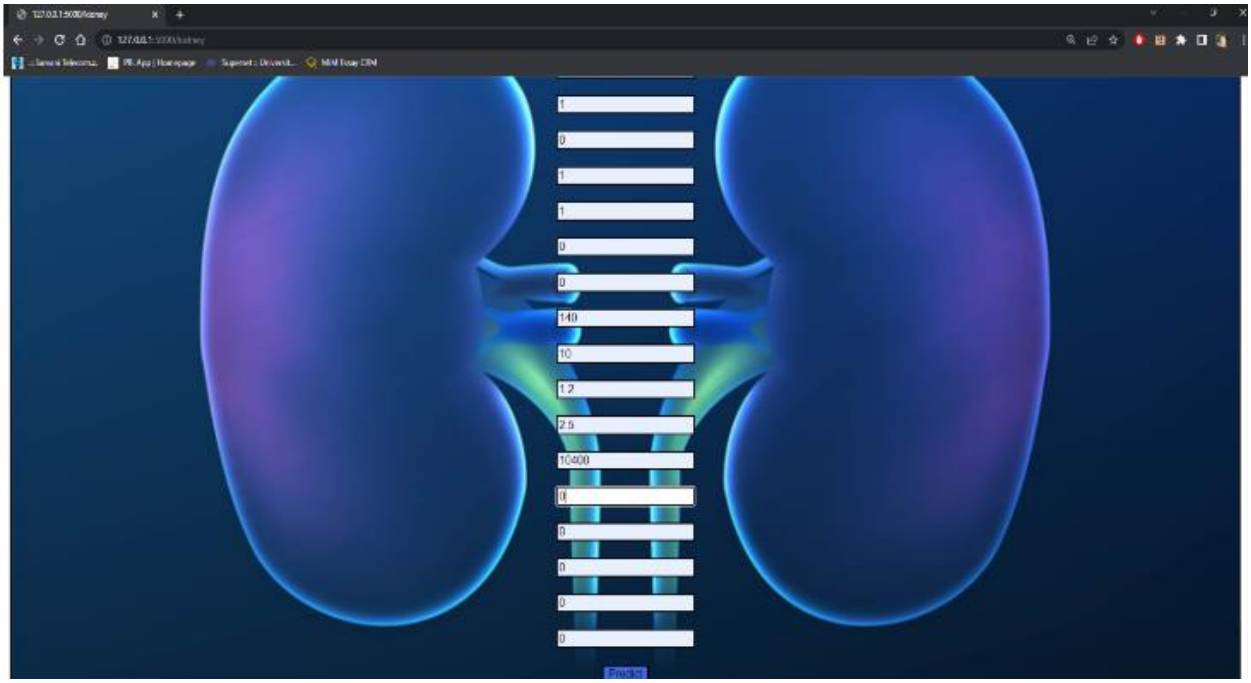


Fig 6.5 After entering value click on predict to predict the results

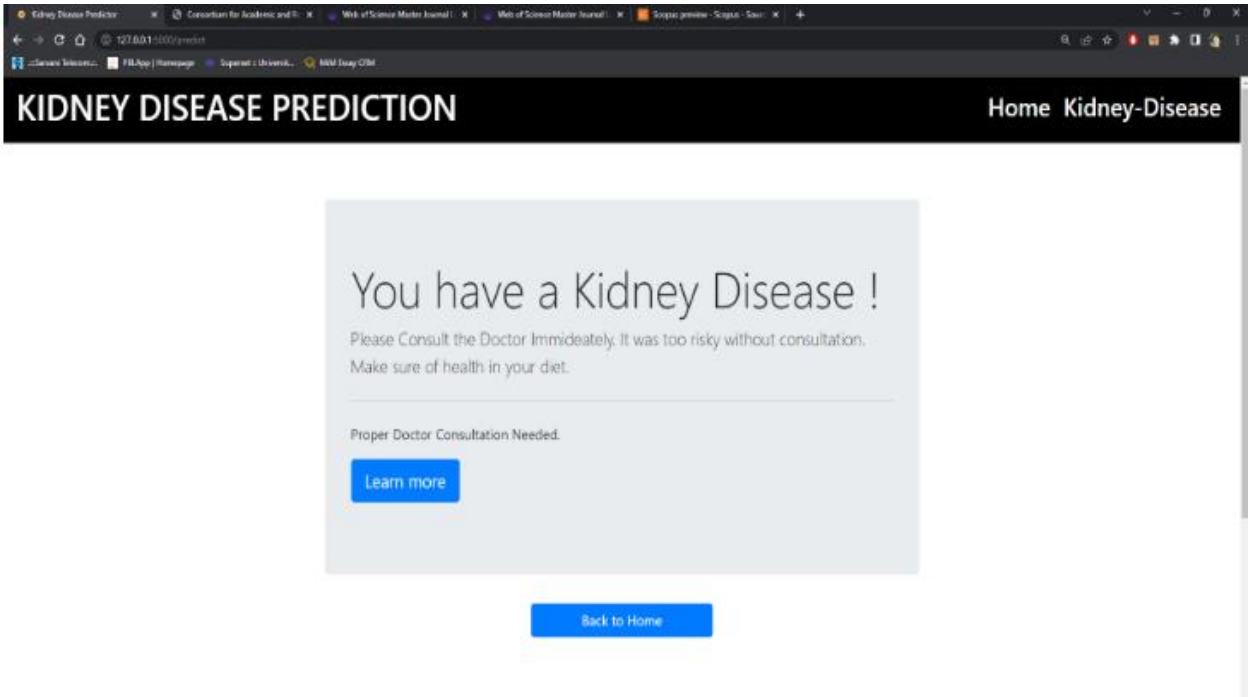


Fig 6.6 Above picture shows that a person is suffering from kidney disease

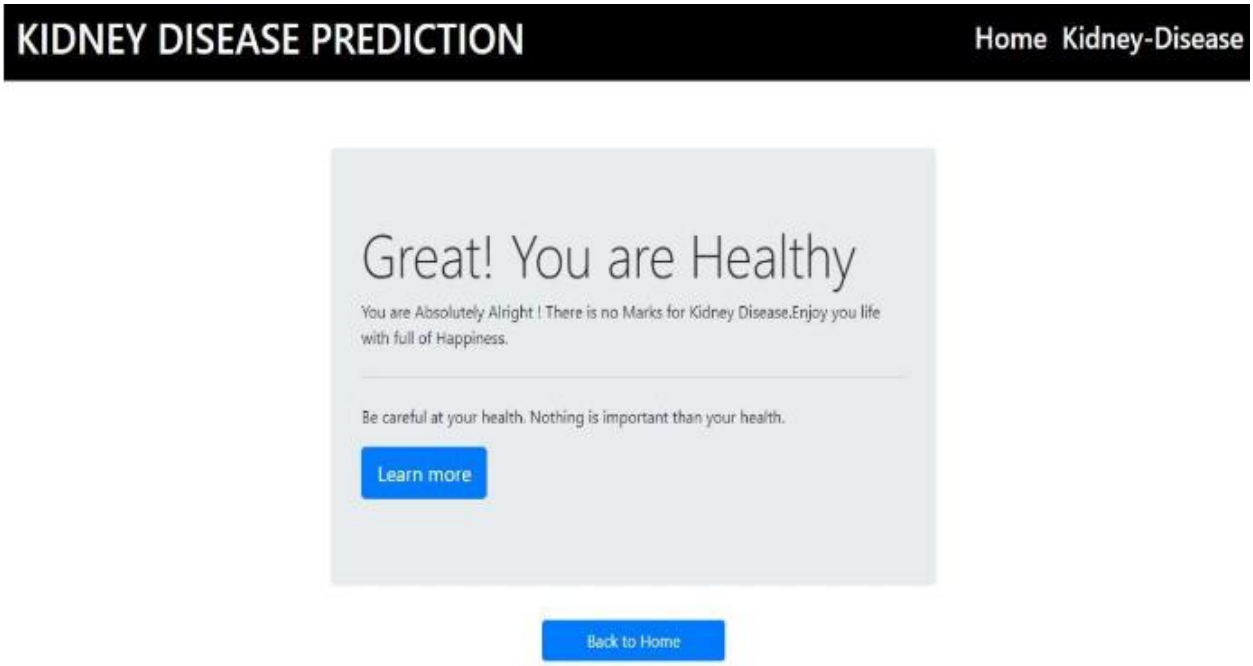


Fig 6.7 Above picture shows that a person is not suffering from kidney disease

7. CONCLUSION

The main goal of this project is to predict Chronic Kidney Disease using the dataset's complete features and significant characteristics. Effective feature engineering aids in reducing the number of features required for the prediction algorithm, in turn reducing the quantity of medical tests that must be performed. So, we looked at various machine learning techniques in this paper. 18 different characteristics of CKD patients were examined, and the projected accuracy for several machine learning methods, such as Decision Tree, SVM, Random Forest, and others, was calculated.The results analysis shows that the accuracy of the decision tree algorithms is 100%, the accuracy of the SVM algorithms is 99.5%, and the accuracy of the random forest algorithms is 100%. When using the decision tree method, the tree is constructed using the complete dataset's features, and when using the random forest algorithm, the determination of whether or not a person has renal disease is made. So, we have thought about using random forest to forecast chronic renal disease. This system's benefit is that the prediction process takes less time. Early therapy initiation for CKD patients will benefit the medical team.

8. FUTURE SCOPE

The ability to forecast chronic renal disease is aided by this method. However, this approach can be enhanced in accordance with future needs, for example, by emphasising the significance of adding domain expertise into feature selection when assessing clinical data linked to CKD. Also, by using information on genetics, water consumption habits, and food kinds in the research, better understanding of CKD can be acquired. In this way, we can enhance the system in accordance with recommendations for the future.

9. REFERENCES

1. C.-S. Lee and M.-H. Wang, “A fuzzy expert system for diabetes decision support application.” IEEE transactions on systems, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, vol. 41, no. 1, pp. 139–153, 2011.
2. C. B. Delahunt, C. Mehanian, L. Hu, S. K. McGuire, C. R. Champlin, M. P. Horning, B. K. Wilson, and C. M. Thompon, “Automated microscopy and machine learning for expert-level malaria field diagnosis,” Proceedings of the 5th IEEE Global Humanitarian Technology Conference, GHTC 2015, pp. 393–399, 2015.
3. B. D. Sekar, C. M. Dong, J. Shi, and X. Y. Hu, “Fused hierarchical neural networks for cardiovascular disease diagnosis,” IEEE Sensors Journal, vol. 12, no. 3, pp. 644–650, 2012.
4. S. Basnet and N. Venkatraman, “A novel fuzzy-logic controller for an artificial heart,” Proceedings of the IEEE International Conference on Control Applications, pp. 1586–1591, 2009.
5. C. Arya and R. Tiwari, “Expert system for breast cancer diagnosis: A survey,” 2016 International Conference on Computer Communication and Informatics, ICCCI 2016, pp. 1–9, 2016.