IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 4, May 2022

Prediction of Chronic Kidney Disease Using Machine Learning

Himani Hatwar[1], Divya Chhaprwal[2], Dvarkesh Rokade[3], Santosh Kale[4]
Department of Computer Science
NBN Sinhgad School of Engineering, Pune, Maharashtra, India

Abstract: *Chronic Kidney Disease is one of the most serious illnesses nowadays, and it is vital to have a good diagnosis as soon as possible. Machine learning has proven to be effective in medical therapy. The doctor can diagnose the ailment early with the use of machine learning classifier algorithms. This article has examined Chronic Kidney Disease prediction from this standpoint. The Chronic Kidney Disease dataset was obtained from the University of California at Irvine's repository. The artificial neural network, C5.0, Chi-square. After effectively filling out the incomplete data set, 4 machine learning algorithms (Random forest, Support vector machine, K-nearest neighbor and Decision tree) were used to establish models. The dataset was also subjected to the significant feature selection technique. The results were computed for each classifier using I full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, and (vi) synthetic minority over-sampling technique with full features. The results show that in synthetic minority over-sampling technique with full features, LSVM with penalty L2 has the maximum accuracy of 98.86 percent. Along with precision, recall, F-measure, and area, accuracy, precision, recall. Among these machine learning models, random forest achieved the best performance with 99.75% diagnosis accuracy. Hence, we speculated that this methodology could be applicable to more complicated clinical data for disease diagnosis..*

Keywords: Chronic Kidney Disease, Machine Learning, Prediction.

I. Introduction

Chronic kidney disease (CKD) is a condition in which your kidneys are impaired and are unable to change your blood as they should. Kidneys' major function is to filter excess water and waste from the body. Haruna Chiroma was the associate editor in charge of coordinating the review of this article and authorising it for publication. Your blood produces urine, and if you have CKD, it signifies that wastes are accumulating in your body.

Because of the cumulative damage over time, this condition is chronic. It's a common ailment all across the world. Some health problems may arise as a result of CKD. Diabetes, high blood pressure, and heart disease are just a few of the causes of CKD. CKD is affected by age and gender, in addition to these serious disorders. If your kidneys aren't performing properly, you may experience abdominal pain, back pain, diarrhoea, fever, nosebleeds, rash, and vomiting.CKD is caused by two primary diseases: diabetes and high blood pressure. As a result, addressing these two disorders is the key to preventing CKD. CKD usually does not manifest itself until the kidneys have been severely damaged. According to studies, hospitalisation cases are increasing at a rate of 6.23 percent per year, but the global fatality rate stays constant. There are a few tests that can be used to determine the severity of CKD:
(i) EGFR (estimated glomerular filtration rate)
(ii) urine test
(iii) Blood pressure.

1.1 EGFR

EGFR value shows that how your kidney cleaning the blood. If your eGFR value is greater than 90, that means the kidney is normal. If eGFR value is less than 60, that means you have CKD [5].

DOI: 10.48175/IJARSCT-3878

1.2 Urine Test

The doctor also asks for urine test for kidney functionality because kidneys make urine. If the urine contains blood and protein [6], that means your kidney is not working properly.

1.3 Blood Pressure

Doctor measures blood pressure as Blood pressure range shows how your heart is pumping blood. If eGFR value reaches less than 15, that means the patient has end-stage kidney disease. At this point, there are only available treatments: (i) dialysis and (ii) kidney transplant. Patient's life after dialysis depends on such factors as age, gender, frequency and duration of dialysis, physical movement of the body and mental health [7]. If dialysis is not possible, the doctor has only one solution, i.e., kidney transplantation. However, it is extremely expensive [8].

Medical care is a major application area for cognitive intelligence systems. Following that, data mining can be used to extract hidden information from large patient medical and treatment datasets that doctors regularly collect from patients in order to obtain pieces of information about symptomatic data and to carry out exact treatment plans. The method of obtaining hidden information from a large dataset is known as data mining. Data mining strategies are interconnected and widely used in a variety of contexts and fields. We can anticipate, classify, change, and cluster data using data mining approaches. The goal specifies how an algorithm will handle a training set with a set of attributes and targets. Data mining is appropriate for data mining when the dataset is large, but we can also do it with machine learning when the dataset is small. Data analysis and pattern recognition are also possible with machine learning. Because there are so many different health datasets, machine learning algorithms are the greatest way to increase diagnosis prediction accuracy. Machine learning techniques are becoming more widespread in healthcare as the electronic dataset expands rapidly.

Padmanaban and Parthiban proposed that machine learning classifier methods might be used to diagnose CKD in diabetic individuals early. They gathered data from a diabetic research centre in Chennai and used Naive Bayes and Decision Tree to analyse the data. They used the Weka tool to determine the accuracy and found that the Nave Bayes classifier had the maximum accuracy of 91 percent. De Almeida et al. employed Decision Trees, Random Forests, and Support Vector Machines (SVM) using linear, polynomial, sigmoid, and RBF functions in their research. They used the MIMIC-II database for their research. They came to the conclusion that random forest and decision tree produced the greatest results, with prediction accuracy of 80% and 87 percent, respectively.

The primary goal of this study is to determine whether or not a person has Chronic Kidney Disease. The dataset was subjected to seven different machine learning classifiers for this perception. All of the algorithms were performed with both full and selective features enabled. Oversampling was done with SMOTE, and all of the findings were recorded. The outcomes of each machine learning model were compared to a single deep neural network algorithm. Two hidden layers were employed in a deep learning neural network. For the computation, IBM SPSS Modeler was used. When using a deep neural network on the dataset, the contribution shows a 99.6% accuracy estimate.

To summarise the prior CKD diagnostic models, we find that the majority of them suffer from either a narrow applicability range or low accuracy due to the strategy employed to impute missing information. As a result, we suggest an approach for expanding the applicability range of CKD diagnostic models in this paper. At the same time, the model's accuracy is being increased. The following are the contributions of the planned work.

The remainder of the paper is laid out as follows. The preliminaries are described in Section II. Section III explains how the individual model and the combined model are created. The integrated model's performance is evaluated.

1.4 The Importance of an Early CKD Diagnosis

In the instance of Brazil, around 10% of persons were aware that they had some kidney disease in 2009, while approximately 70% were undiagnosed. Around 500-650 persons per million need dialysis or kidney transplantation in 2010. According to the Brazilian chronic dialysis study, the number of dialysis patients in 2016 was 122,825, an increase of 31,000 people over the previous five years. A more recent dialysis survey found that the frequency and incidence rates of dialysis patients are increasing.

In 2017, they were found to be 610 and 194 per million persons, respectively. In addition. As a result, in Brazil, the incidence and prevalence of CKD is clearly still a public health issue. Unfortunately, most developing countries have high CKD incidence and prevalence, which is increasing morbidity and mortality rates as well as the cost of public health care.CKD has afflicted a considerable portion of the young and middle-aged population in other emerging nations, such as India, Bangladesh, Pakistan, Nepal, Bhutan, Sri Lanka, and Afghanistan. Early detection of CKD is even more difficult for persons who reside in distant and difficult-to-reach areas, due to the precarious nature of primary care. Diagnosing CKD in its early stages can assist developing countries reduce morbidity and mortality rates, as well as public health expenses.

II. Research Gap

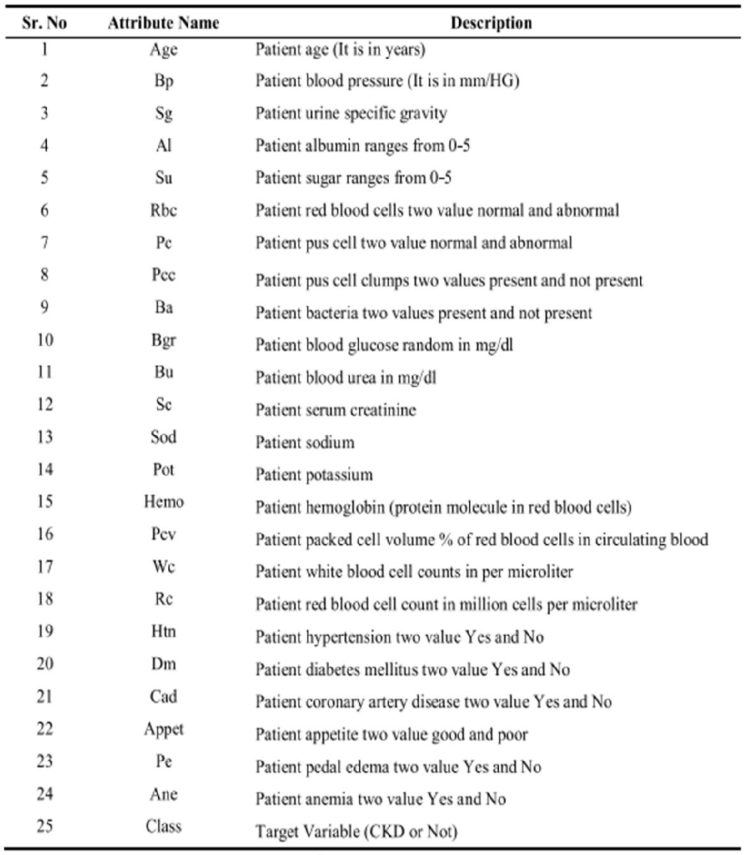
Until now, full features have been considered in the majority of cases. Feature optimization was carried out in this study, with three distinct feature selection algorithms used to find the most useful approach for extracting the relevant feature for the prediction of Chronic Kidney Disease. Due to the fact that many datasets include unbalanced classes, class balancing is required to improve the performance of the classifier model. SMOTE was employed as a class balancein this study. On the same dataset, the highest accuracy of 99.6 percent was reached, while the article claims an accuracy of 99.1 percent. According to, the model's maximum accuracy was 99.7%, however they were working on patient risk calculations, whereas the main goal of the paper is to forecast Chronic Kidney Disease.

III. Proposed Model

3.1 Dataset

This study makes use of the Chronic Kidney Disease dataset. This dataset had already been used by a number of researchers. The UC Irvine Machine Learning Repository has donated this dataset, which is available on the UCI website. There are 400 instances in this dataset, with 24 attributes and one target attribute. The target attribute is divided into two classes: CKD and non-CKD. In 2015, the data was gathered from a number of hospitals. It also has a value that is missing.

TABLE 1. Description of Attributes in the Dataset.



Sr. No	Attribute Name	Description
1	Age	Patient age (It is in years)
2	Bp	Patient blood pressure (It is in mm/HG)
3	Sg	Patient urine specific gravity
4	Al	Patient albumin ranges from 0-5
5	Su	Patient sugar ranges from 0-5
6	Rbc	Patient red blood cells two value normal and abnormal
7	Pc	Patient pus cell two value normal and abnormal
8	Pcc	Patient pus cell clumps two values present and not present
9	Ba	Patient bacteria two values present and not present
10	Bgr	Patient blood glucose random in mg/dl
11	Bu	Patient blood urea in mg/dl
12	Sc	Patient serum creatinine
13	Sod	Patient sodium
14	Pot	Patient potassium
15	Hemo	Patient hemoglobin (protein molecule in red blood cells)
16	Pcv	Patient packed cell volume % of red blood cells in circulating blood
17	Wc	Patient white blood cell counts in per microliter
18	Rc	Patient red blood cell count in million cells per microliter
19	Htn	Patient hypertension two value Yes and No
20	Dm	Patient diabetes mellitus two value Yes and No
21	Cad	Patient coronary artery disease two value Yes and No
22	Appet	Patient appetite two value good and poor
23	Pe	Patient pedal edema two value Yes and No
24	Ane	Patient anemia two value Yes and No
25	Class	Target Variable (CKD or Not)

DOI: 10.48175/IJARSCT-3878

3.2 Data Processing

To make computer processing easier, each categorical (nominal) variable was coded. Normal and abnormal values for rbc and pc were coded as 1 and 0, respectively. Present and notpresent were recorded as 1 and 0, respectively, for the values of pcc and ba. Yes and no were coded as 1 and 0 for the values of htn, dm, cad, pe, and ane,respectively. Good and poor were coded as 1 and 0, respectively, for the value of appet. Although the three variables sg, al, and su are defined as categorical types in the original data description, the values of these three variables are still numeric, hence these variables were regarded as numeric variables. Every category variable was converted into a factor. Each sample was assigned an independent number between 1 and 400. The data collection contains a substantial amount of missing values, with only 158 complete instances. In general, patients may miss certain measurements before receiving a diagnosis for a variety of reasons.When the diagnostic categories of samples are unknown, missing values will show in the data, necessitating the use of an imputation procedure.

The missing values in the original CKD data set were handled and lled at rst after encoding the categorical variables. KNN imputation was used to determine the K full samples with the least Euclidean distance for each sample with missing values. The missing values for numerical variables are calculated using the median of the corresponding variable in K full samples, whereas missing values for category variables are calculated using the category with the highest frequency in the corresponding variable in K complete samples. People with similar physical conditions should have similar physiological data, which is why the method based on a KNN is used to fill in the missing values for physiological measurements. For healthy people, physiological measurements, for example, should be steady within a particular range. The physiological measurements of a person with a similar degree of the same disease should be similar in diseased persons. Differences in physiological measurement data, in particular, should not be significant for people in similar settings. This strategy, which has been used in the field of hyperuricemia, should be modified to the diagnostic data of other disorders.

Data preprocessing is a technique for converting raw data into a clean dataset. Every machine learning classifier algorithm begins with this stage. This method completes tasks such as handling missing values, rescaling the dataset, converting it to binary data, and standardising the dataset. Rescaling is used to scale the dataset when it contains attributes with different scales. To convert the value into 0 and 1, the binary transformation was used. Every attribute's value is interpreted as 1 if it is above the threshold and 0 if it is below the threshold. Each attribute in the standardised method has a mean of 0 and a standard deviation of 1.

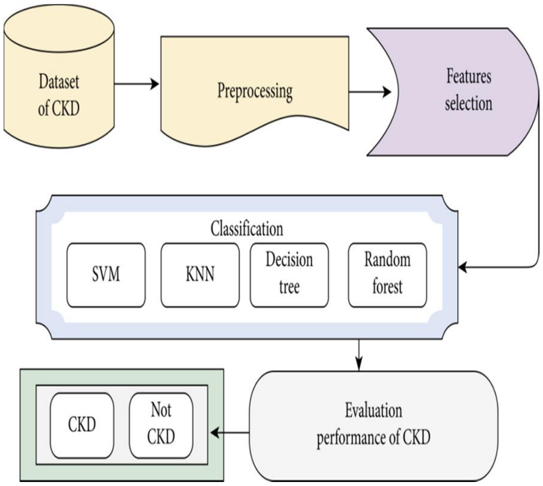


Figure: Data Preprocessing

3.3 Classification Algorithms

A key component of supervised learning is the classification technique. Classifiers learn from the training dataset and use it to locate the target attribute on the testing dataset. The classification techniques used in study are listed below.

A. Linear Support Vector Machine (LSVM)

The linear support vector machine (LSVM) is a recent, particularly fast machine learning method that uses a simple iterative strategy to solve multiclass classification problems for huge datasets. The SVM model is built in linear CPU time for the dataset. The sparse and dense formats of LSVM can be used for high-dimensional datasets. It's used to solve machine learning problems with massive datasets using less expensive computational resources. A supervised classifying algorithm is the Support Vector Machine. The kernel trick is used to solve the categorization problem.The optimum edge between the possible outputs is found using these adjustments.For nonlinear kernels, such as RBF, SVM is utilised. LSVM is an excellent candidate for the linear kernel. For all linear problems, the LSVM classifier is sufficient.

B. K- Nearest Neighbors (KNN)

The KNN method is a simple supervised algorithm. It can be used to solve problems involving classification and regression. It is, however, mostly employed to solve categorization difficulties. KNN is a lazy learning algorithm that does not employ a specific training stage and instead uses all of the data for training. It also does not consider anything about the underlying data, making it anonparametric learning algorithm. KNN saves the entire dataset. Because it lacks a model, there is no need to learn it. When fresh data is entered for predicting outcomes, it compares K neighbours, therefore choosing K's value is crucial. The distance between two already labelled data is determined. The distance aids in locating the new data's nearest neighbour. The distance is calculated using the Euclidian approach.

C. Random Forest

A type of supervised classifier is the random tree. It creates a diverse group of students. The tree is created using a stochastic method. It's a classification technique that uses ensemble learning. It functions in the same way as a decision tree, except each split is based on a random subset of qualities. This approach can be applied to both classification and regression issues. 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 bulk of votes are chosen by the random tree's output. Every leaf node in the tree has a linear model. The model is trained using the bagging training technique.

D. Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

3.4 Performance Evaluation Measure

The performance of the classifier was evaluated using a variety of assessment matrices. The confusion matrix was utilised for this purpose. Due to two classes in the dataset, it is a 22 matrix. The confusion matrix shows two types of correct classifier predictions and two types of wrong classifier predictions.

A. Confusion Matrix Description

- TP: True Positive means output as positive such that predictedresult is correctly classified.
- TN: True Negative means output as negative such thatpredicted result is correctly classified.
- FP: False Positive means output as positive such that predictedresult is incorrectly classified.
- FN: False Negative means output as negative such thatpredicted result is incorrectly classified.

B. Classification Accuracy

The right rate of prediction findings is shown by classification accuracy. It's based on the confusion matrix. The classification accuracy is found by equation 2:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

C. Classification Error

The rate of inaccurate prediction outcomes is shown by the classification error. The confusion matrix is used to calculate it. The classification error is found by equation 3:

$$Error = \frac{FP + FN}{TP + TN + FP + FN} * 100$$

D. Precision

Precision is a key criterion for evaluating model performance. It's the proportion of related instances in the overall number of retrieved items. It is a forecasted value that is positive. The precisionis calculated as follows in equation 4:

$$Precision = \frac{TP}{TP + FP} * 100$$

E. Recall

Recall is an important criterion for evaluating model performance. It's the proportion of related instances in the overall number of instances retrieved. The recall is calculated as follows in equation 5:

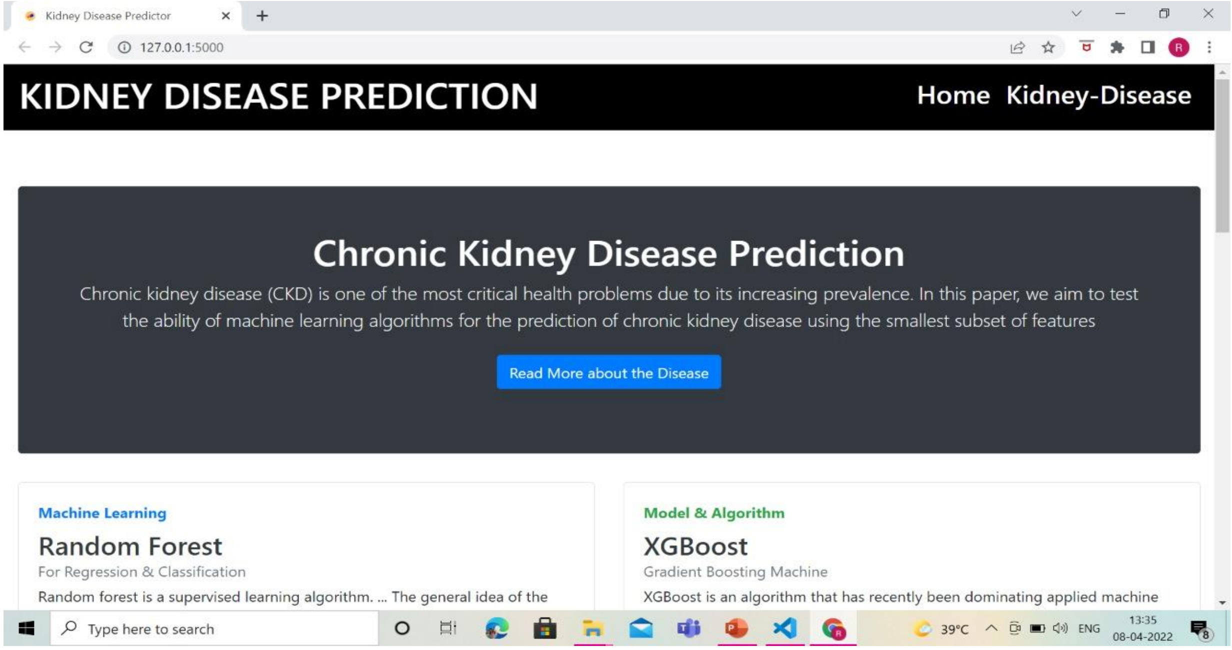
$$Recall = \frac{TP}{TP + FN} * 100$$

F. F-Measure

F Score is another name for it. The F-measure is used to determine the correctness of a test. Precision and recall are used to compute itby equation 6:

$$F - Measure = 2 \frac{* Precision * Recall}{Precision + Recall}$$

IV. Project Frontend



Chronic Kidney Disease Prediction

Chronic kidney disease (CKD) is one of the most critical health problems due to its increasing prevalence. In this paper, we aim to test the ability of machine learning algorithms for the prediction of chronic kidney disease using the smallest subset of features

Machine Learning

Random Forest

For Regression & Classification

Random forest is a supervised learning algorithm. ... The general idea of the

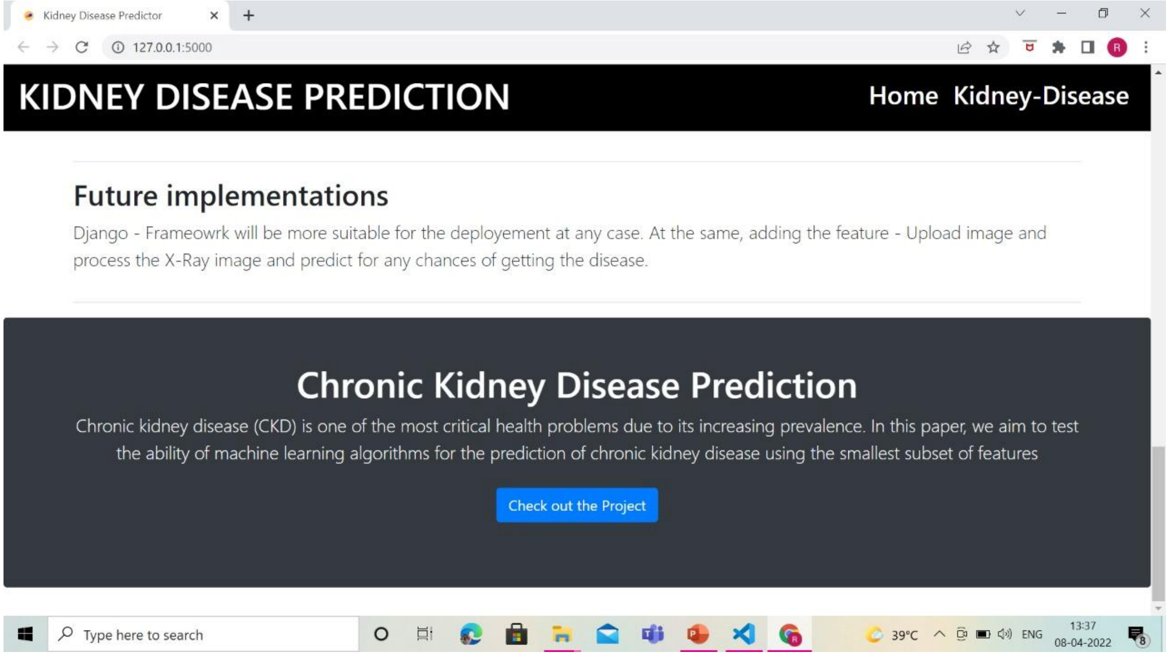
Model & Algorithm

XGBoost

Gradient Boosting Machine

XGBoost is an algorithm that has recently been dominating applied machine

DOI: 10.48175/IJARSCT-3878

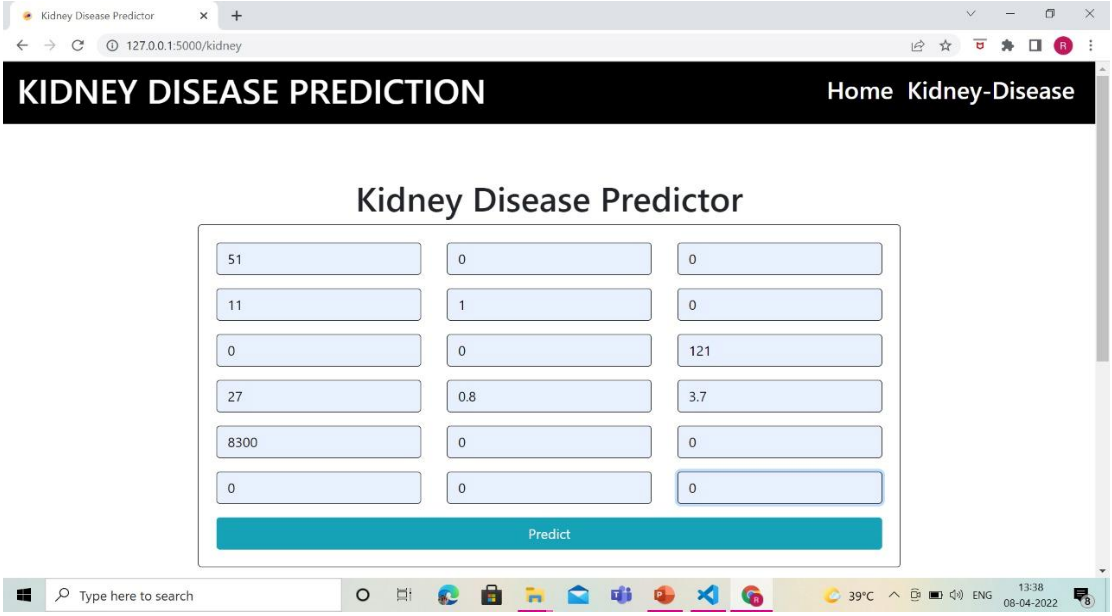


Future implementations

Django - Frameowrk will be more suitable for the deployement at any case. At the same, adding the feature - Upload image and process the X-Ray image and predict for any chances of getting the disease.

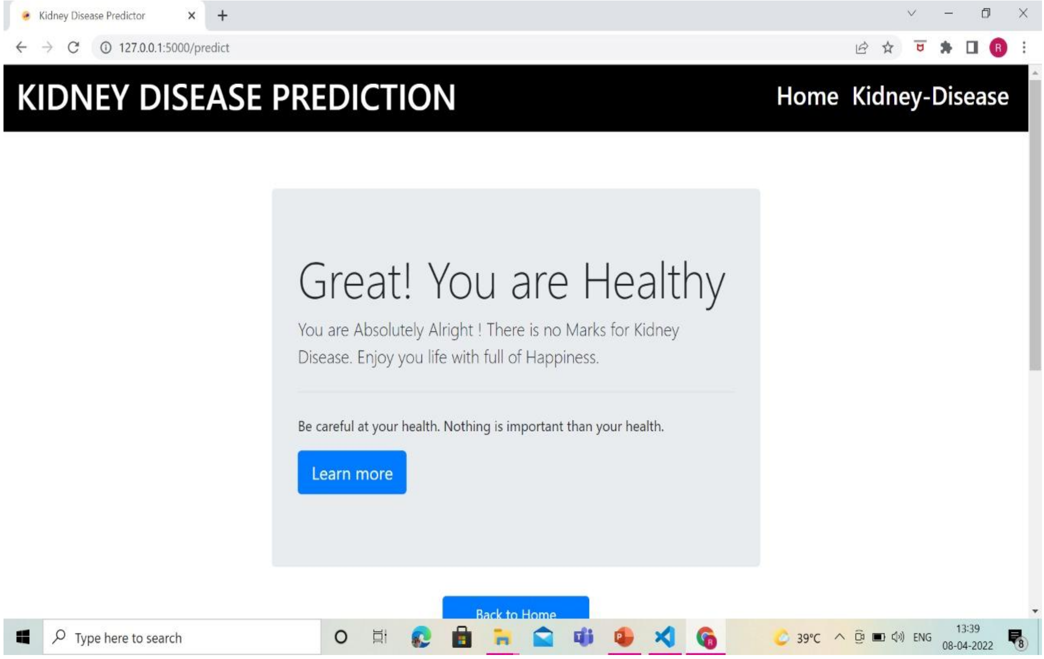
Chronic Kidney Disease Prediction

Chronic kidney disease (CKD) is one of the most critical health problems due to its increasing prevalence. In this paper, we aim to test the ability of machine learning algorithms for the prediction of chronic kidney disease using the smallest subset of features



Kidney Disease Predictor

51	0	0
11	1	0
0	0	121
27	0.8	3.7
8300	0	0
0	0	0



Great! You are Healthy

You are Absolutely Alright ! There is no Marks for Kidney Disease. Enjoy you life with full of Happiness.

Be careful at your health. Nothing is important than your health.

V. Limitations To Validity

The study's findings have certain limitations in terms of their validity. The dataset size, for example, may be limited; nevertheless, there are no publicly available datasets that give CKD risk assessments of participants using well-accepted medical guidelines and nephrologists with experience in underdeveloped countries. Even though there were only 60 participants, the nephrologists were obliged to assess each one, which is a time-consuming procedure. Furthermore, the dataset's size necessitated data enhancement by changing CKD biomarker values to better confidence in the classifications. Despite the possibility of bias in the supplemented data, this work was meticulously carried out to ensure that the simulated participants remained in the same CKD risk class. The experiment was conducted using the k-fold cross-validation approach to improve confidence, and the validity of the supplemented data was assessed by an expert nephrologist (used only in the training set). As a result, the enhanced data had no negative impact on the machine learning classifiers' testing phase.

VI. Conclusion

This article aims to predict Chronic Kidney Disease using the CKD dataset's entire characteristics and significant features. Three alternative strategies for feature selection were used: correlation-based feature selection, Wrapper method, and LASSO regression. Artificial neural network, C5.0, logistic regression, CHAID, linear support vector machine (LSVM), K-Nearest neighbours, and random tree were used to classify this perception. Complete features, selected features by CFS, selected features by Wrapper, selected features by LASSO regression, SMOTE with selected features by LASSO, and SMOTE with full features were used to compute the results for each classifier. In SMOTE with complete features, it was discovered that LSVM had the maximum accuracy of 98.86 percent. On characteristics picked by LASSO regression with SMOTE and without SMOTE, all classifier algorithms performed well. For all five classifiers, SMOTE with all characteristics produced the best results. A total of seven classifiers were utilised in this study. Logistic and KNN, on the other hand, did not produce satisfactory results, which is why they were not used in SMOTE. According to the findings, SMOTE is the best strategy for balancing a dataset. It should be observed that when using LASSO regression with specified features, SMOTE produced better results than when using LASSO regression without SMOTE. When compared to other classifier algorithms, LSVM scored the greatest accuracy in all testing.

References

[1]. Q.-L. Zhang and D. Rothenbacher,``Prevalence of chronic kidney diseasein population-based studies: Systematic review," BMC PublicHealth,vol. 8, no. 1, p. 117, Dec. 2008.

[2]. W. M. McClellan, D. G. Warnock, S. Judd, P.Muntner, R. Kewalramani,M. Cushman, L. A. McClure, B. B. Newsome, and G. Howard, ``Albuminuriaand racial disparities in the risk for ESRD," J. Amer. Soc. Nephrol.,vol. 22, no. 9, pp. 1721_1728, Aug. 2011.

[3]. M. K. Haroun, ``Risk factors for chronickidney disease: A prospectivestudy of 23,534 men and women in Washington County, Maryland,"J. Amer. Soc. Nephrol., vol. 14, no. 11, pp. 2934_2941, Nov. 2003.

[4]. W. D. Souza, L. C. D. Abreu, L. G. D. SilvaI,and I. M. P. Bezerra,``Incidence of chronic kidney disease hospitalisations and mortality inEspírito Santo between 1996 to 2017," Wisit Cheungpasitporn, Univ.Mississippi Medical Center, Rochester, MN, USA, Tech. Rep., 2019, doi:10.1371/journal.pone.0224889.

[5]. B. Zupan, A. J. Halter, and M. Bohanec,``Qualitative model approach tocomputer assisted reasoning inphysiology," in Proc. Intell. Data Anal.

[6]. Med. Pharmacol. (IDAMAP), Brighton, U.K.,2018, pp. 1_7.

[7]. T. Xiuyi and G. Yuxia, ``Research on application of machine learningin data mining," in Proc. IOP Conf., Mater. Sci. Eng., 2018, doi:10.1088/1757-899X/392/6/06220.

[8]. A. Dhillon and A. Singh, ``Machine learning in healthcare data analysis:A survey," J. Biol. Today's World, vol. 8, no. 2, pp. 1_10, Jan. 2018.

[9]. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, andR. Ranganath, ``A reviewof challenges and opportunities in machine learningfor health," in Proc. AMIA Joint Summits Transl. Sci., 2020, p. 191.

[10]. K. R. A. Padmanaban and G. Parthiban, ``Applying machine learningtechniques for predicting the risk of chronic kidney disease," Indian J. Sci.Technol., vol. 9, no. 29, Aug. 2016.

[11]. L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino,``Kidney failure detection using machine learning techniques," in Proc. 8thInt. Workshop ADVANCEs ICT Infrastructures Services, 2020, pp. 1_8.

DOI: 10.48175/IJARSCT-3878