Diagnosing Chronic Kidney Disease using Machine Learning

Palla Tejaswi[1],Cheedi Dharani[2],Rayavarapu Siva[3],Kakara Shivani[4]

[1234]Final Year B.Tech, CSE, Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P, India
Guided by: Mrs. Dr. K.N.S. Lakshmi, Professor, SVPEC, Visakhapatnam, A.P, India

---***---

Abstract - *Occurance of Chronic Kidney Disease doesn't have any obvious symptoms in the earlier stages to identify, so patients most often fail to notice the disease in kidney. So that the patients are fail to take the timely treatment which enables kidney failure. CKD have an high ill and death rate in the world. Data Mining methods and Machine Learning Algorithms play a major role in this aspect of life of living beings. Machine Learning plays an key role in diagnosing of any diseases. Chronic Kidney Disease(CKD) is a condition in which the kidneys are failed and cannot filter blood properly, also cannot have an electrolyte balance steady. The causes for CKD may also includes a family history of kidney diseases or failure, high blood pressure, hypertension, obesity, age, diabetes. This is a lasting failure to the kidney when it reaches to high time or months. The complications that involves when kidney failure are heart diseases, anemia, bone diseases, high potassium and calcium, leg swelling. The last stage of situation leads to complete kidney failure and needs kidney transplant to live. An early detection of CKD can improve the great quality of life to a greater extension. So we choose a great prediction algorithms(Integrated Model-Naïve Bayes Classifier and Random Forest Algorithms) to predict CKD. This paper uses data preprocessing, data selection and various classifiers (Random Forest and Naive Bayes) to predict CKD and finally it results best prediction accuracy framework for CKD. The results of the prediction accuracy framework show promising results of better prediction at an early stage of CKD.*

Key Words: *Chronic Kidney Disease, K-Nearest Neighbour Imputation, Feature Selection, Integrated Model-Naïve Bayes Classifier and Random Forest Algorithms.*

1. INTRODUCTION

Our investigations have reached great outcomes in the identification of CKD. Dataset was taken from the University of California Irvine Machine Learning Repository which consists of many missing values in it. Missing values are caused due to the patients emergency or by their forgetiveness. The K-Nearest Neighbour Imputation was used to fill missing values in the dataset,that is mean statistical algorithm was utilized to fill missing numerical values whereas mode statistical algorithm was utilized to fill missing nominal values.

CHRONIC KIDNEY DISEASE:

Kidney illness (CKD) is a kind of kidney sickness where there is continuous loss of kidney filteration over a time of months to years. At first there are no obvious symptoms to identify, later manifestations might incorporate leg enlarging, feeling tired, heaving, loss of hunger, and frequent urination, nausea., hypertension, bone sickness, and anemia. Causes of kidney infection incorporate diabetes, hypertension, glomerulonephritis, and polycystic kidney illness. Hazard factors incorporate a family background of persistent kidney illness. Investigations be done by taking blood tests to quantify the assessed Glomerular Filtration Rate (GFR), and a pee test to gauge egg whites, Ultrasound or kidney biopsy might be performed to decide the hidden reason which may lead to CKD.

MACHINE LEARNING:

The investigations on computer calculations that work on naturally through experience that says it learns from the past data whithout explicitly programmed for everytime when it needed. Machine Learning is highly responsible for Diagnosing and Predictions.

1.1 EXISTING SYSTEM:

Existing System of diagnosing of chronic kidney disease includes an integrated model of two algorithms namely Logistic Regression and Random Forest Algorithms. In this System,it have an problem of linear boundaries and unable to predict for higher dataset.

1.2 PROPOSED SYSTEM:

Proposed System, includes an integrated model of Naïve Bayes Classifier and Random Forest Algorithms. The dataset was taken from the University of California Irvine Machine Learning Repository. The dataset consists of many missing values due to the patients forgetiveness or by their any personal issues. Those missing values can be filled using K-Nearest Neighbour imputation algorithm, Mean Statistical Algorithm was used to fill numerical values whereas Mode Statistical Algorithm was used to fill nominal values in the dataset.

2.METHODOLOGY:

The methodology is used for building multiple classification models at a time. The model methods are divided into two sub-modules. The first module has been used to predict the error rates namely AAE and ARE for dataset. Also, the Kruskal-wallis test has been conducted to find the significant difference in the performance. The second sub module used for the selection of important features.

Using Python Tool on Standalone machine Environment: The Python computer programs are essential tool for progression in the numeric examination and machine learning spaces. Python is a perfect way to deal with reproducible, extraordinary examination programme. Python is extensible and offers rich value for architects or programmers to manufacture their own specific gadgets and procedures for examining data. It was easy for programming to build models also efficient in implementing. The vastness of package organic framework is indisputable one of the Python's most grounded qualities are if a true technique exists, odds are there was presently an Python package out there for it. Python's positive conditions false its packages natural framework. Here, the accuracy of different machine learning algorithms has been look using Python Tool on the Standalone machine. Initial analysis has been done using Microsoft excel. A csv dataset file has been provided as an input for Python. The data is gathered from web sources after that Pre-processing of Dataset can be done which involves Data cleaning, Data Integration and Data Transformation.

Confusion Matrix:

The confusion matrix is also known as Error matrix. It's in the format of table that's often accustomed describe the performance of a classification method on a collection of test data that actual value are known. Each tuple of the confusion matrix denotes the occurrences within the actual class. Each column of the matrix denotes the occurances in an extremely predicted class. The confusion matrix is represented as shown below:

		Predicted	
		Yes	No
Actual	No	FP	TN
	Yes	TP	FN

Accuracy and Precision:

In classification, it includes two important evaluation parameters namely accuracy and precision Accuracy is defined by the aggregation of true positive and true negative instances divided by 100 whereas Precision was defined by the fraction of true positive and predicted yes instances. The Accuracy and Precision can be calculated by the given below formulae:

Accuracy: (TP+TN)/100

Precision: TP/Predicted yes

Recall and F-Square: Recall is defined by the fraction of True Positive instances and Actual yes instances whereas F-Square defined by the fraction between product of the recall and precision to the summation of recall and precision parameter of classification. The recall and precision can be calculated by the given below formulae:

Recall: TP/Actual Yes

F-Square: (2*Recall+Precision)/(Recall+Precision)

Sensitivity, Specificity and ROC:

Sensitivity is defined by the fraction of true positive and actual yes instances whereas specificity is defined by the difference between one and false positive rate value by the actual no instances.. ROC is defined by the fraction between truth positive rate and the false positive rate.

The Sensitivity,Specificity and ROC can be calculated by the given below formulae:

Sensitivity: TP/Actual Yes

Specificity: (1-FP)/Actual No

ROC: TPR/FPR

MCC:MCC is a measure that capable of studying both true and false positives and negatives. The MCC can be calculated as

MCC:((TP)(TN)(FP)(FN))

$$\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}$$

Dataset:

The CKD dataset was collected from the University of California Irvine Machine Learning Repository which consists of 400 patients records.The dataset consists of 24 features which were divided into 11 numerical features and 13 categorical features, in addition to the class features,such as “ckd” and “notckd” for classification. Features involves age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. The diagnostic class contains two values namely ckd and notckd. All features contained missing values except for the diagnostic feature. The dataset is unbalanced because it consists of 250 cases of “ckd” class by 62.5% and 150 cases of “notckd” by 37.5%.

Preprocessing:

The dataset contained outliers,duplications and noise, so it must be cleaned up in a preprocessing stage. The preprocessing stage included by estimating missing values and eliminating noise, such as outliers, normalization, and checking of unbalanced data. Some measurements or values in dataset may be missed when patients are undergoing tests, thereby causing missing values. The dataset had 158 completed instances, and the remaining instances had missing values. The simplest method to handle missing values is to ignore or remove the records, but it is inappropriate with small dataset. So we can use algorithms to compute missing values instead of removing records. The missing values for numerical features can be filled through one of the statistical measures, such as mean, median, and standard deviation. But we used mean statistical measures to fill numerical values in the dataset.The missing values of nominal features can be computed using the mode statistical measures, in which the missing values is replaced by the most common value of the features.While numerical features are the values that can be measured which have two types, either separate or continuous.

Features Selection:

After computing the missing values in the preprocessing stage move on to identifying the important features having a strong and positive correlation with features of importance for disease diagnosis is required. Taking out the vector features eliminates useless features and unrelevant features for predicting.we used the Recursive Feature Elimination method to take out the most important features of prediction. The Recursive Feature Elimination (RFE) algorithm was very popular with its ease of use and configurations and its effectiveness in selecting features in training datasets relevant to predicting target variables and eliminating weak features. The RFE method is used to select the most significant features by finding high correlation between specific features and target.

Evaluation Metrics:

Evaluation metrics were used to estimate the performance of the four classifiers. One of these measures is from the confusion matrix, from which the accuracy, precision, recall, and F1-score are extracted by computing the correctly classified samples (TP and TN) and the incorrectly classified samples (FP and FN), as shown in the following equations:

Accuracy: (TN+TP)*100/(TN+TP+FN+FP)

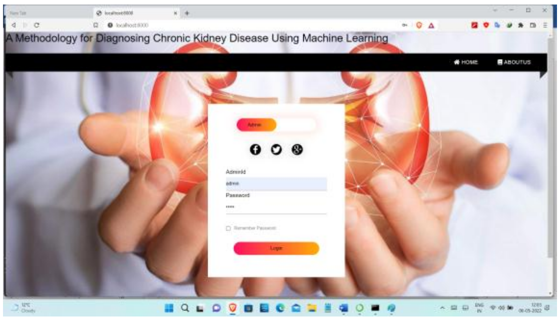
Precision: TP*100/(TP+FP)

Recall: TP*100/(TP+FN)

F1-Score: 2*Precision*recall*100/precision*recall

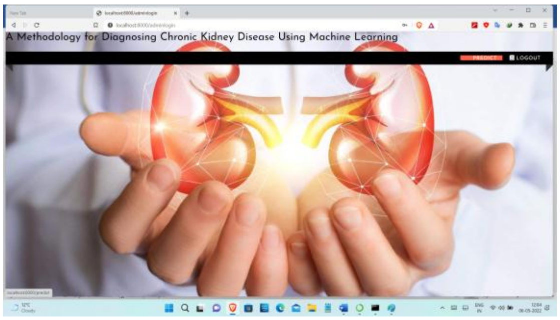
Where, TN is True Negative, TP is True Positive, FN is False Negative, and FP is False Positive.

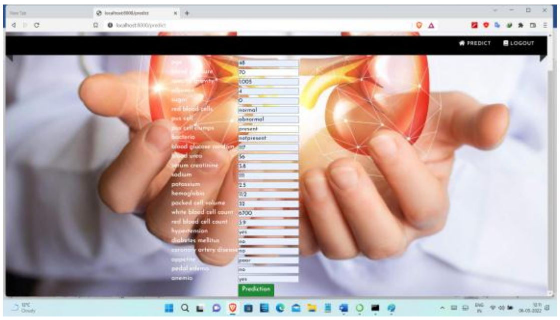
Results and analysis:

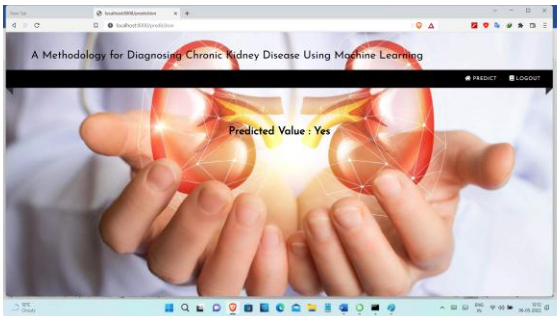

A Methodology for Diagnosing Chronic Kidney Disease Using Machine Learning

AdminId admin

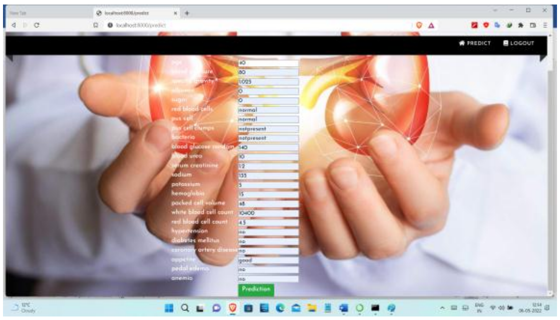
Password ••••

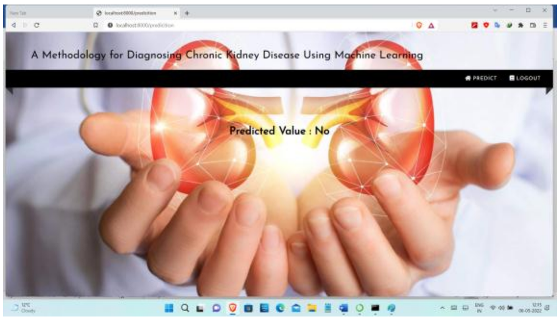

A Methodology for Diagnosing Chronic Kidney Disease Using Machine Learning


age 48
blood pressure 70
specific gravity 1005
albumin 4
sugar 0
red blood cells normal
pus cell abnormal
pus cell clumps present
bacteria notpresent
blood glucose random 117
blood urea 56
serum creatinine 3.8
sodium 111
potassium 2.5
hemoglobin 11.2
packed cell volume 32
white blood cell count 6700
red blood cell count 3.9
hypertension yes
diabetes mellitus no
coronary artery disease no
appetite poor
pedal edema no
anemia yes


A Methodology for Diagnosing Chronic Kidney Disease Using Machine Learning

Predicted Value : Yes


age 40
blood pressure 80
specific gravity 1025
albumin 0
sugar 0
red blood cells normal
pus cell normal
pus cell clumps notpresent
bacteria notpresent
blood glucose random 140
blood urea 10
serum creatinine 1.2
sodium 135
potassium 5
hemoglobin 15
packed cell volume 48
white blood cell count 10400
red blood cell count 4.5
hypertension no
diabetes mellitus no
coronary artery disease no
appetite good
pedal edema no
anemia no



localhost:8000/prediction

A Methodology for Diagnosing Chronic Kidney Disease Using Machine Learning

Predicted Value : No

Conclusion:

This study provides awareness into the diagnosis of CKD patients to handle their condition and receive timely treatment of the disease. The dataset was collected from the University of California Irvine Machine Learning Repository, consists of 400 patients records with 24 features. The dataset was divided into 75% training and 25% testing and validation data. The dataset was pre proocessed to remove outliers and replace missing numerical and nominal values using mean and mode statistical algorithms, respectively. The Recursive Feature Elimination algorithm was applied to select the most strongly identified features of CKD. Selected features were fed into classification algorithms namely Naive Bayes and Random Forest. The parameters of induced classifiers were tuned to perform the best classification, so algorithms reached promising results for diagnosing CKD.

Future Scope :

In future we will be focusing on other algorithms like CNN and RNN.

REFERENCES

[1] AKINYELU, A. A., & ADEWUMI, A. O. (2014). on “Classification of phishing email using random forest machine learning technique”. Journal of Applied Mathematics.

[2] A. Subasi, E. Alickovic, J. Kevric, on “Diagnosis of chronic kidney disease by using random forest,” in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.

[3] L. Zhang et al., on “Prevalence of chronic kidney disease in china: a crosssectional survey,” Lancet, vol. 379, pp. 815-822, Aug. 2012.

[4] A. Singh et al., on “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015.

[5] A. M. Cueto-Manzano et al., on “Prevalence of chronic kidney disease in an adult population,” Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.

[6] Koushal Kumar and Abhishek, on ”Artificial Neural Networks for Diagnosis of Kidney Stones Disease”, I.J. Information Technology and Computer Science, 2012, 7, pp 20- 25.

[7] Tom Fawcett, (2003). on ROC graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories

[8] J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven, on ” used Data mining techniques for predicting acute kidney injury after elective cardiac surgery”, Springer, 2012.

[9] DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi, on ” used Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis”, Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.

[10] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, on ” done performance on comparison of three data mining techniques for predicting kidney disease survivability”, International Journal of Advances in Engineering & Technology, Mar. 2014.

[11] Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri, on ” used Data Mining for funding performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients”, International journal of hospital research, Volume 2, Issue 1,2013, pp 49-54.

[12] Abeer Y. Al-Hyari, on ” chronic kidney disease prediction system using classifying data mining techniques”, library of university of Jordan, 2012. Manish Kumar, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, February- 2016, pg. 24-33 © 2016, IJCSMC All Rights Reserved 31.

[13] Xudong Song, Zhanzhi Qiu, Jianwei Mu, on ” Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field”, International Journal of Advancements in Computing Technology(IJACT) ,Volume4, Number3, February 2012.

[14] N. SRIRAAM, V. NATASHA and H. KAUR, on ” data mining approaches for kidney dialysis treatment”, journal of Mechanics in Medicine and Biology, Volume 06, Issue 02, June 2006.

[15] Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh, on ” An Efficient Diagnosis of Kidney Images using Association Rules”, International Journal of Computer Technology and Electronics Engineering (IJCTEE),Volume 2, Issue 2,april 2012.

[16] Divya Jain, Sumanlata Gautam, on ” Predicting the Effect of Diabetes on Kidney using Classification in Tanagra”,

International Journal of Computer Science and Mobile Computing, Volume 3, Issue 4, April 2014.

[17] M. M. Hossn et al., on "Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts," IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.

[18] M. Alloghani et al., on "Applications of machine learning techniques for software engineering learning and early prediction of students' performance," in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246- 258.

[19] D. Gupta, S. Khare, A. Aggarwal, on "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in Proc. Int. Conf. Computing, Communication and Automation, Apr. 2016, pp. 281-287.

[20] L. Du et al., on "A machine learning based approach to identify protected health information in Chinese clinical text," Int. J. Med. Inform., vol. 116, pp. 24-32, Aug. 2018.

BIOGRAPHIES

Dr. K.N.S. Lakshmi

Currently working as professor in Department of Computer Science and Engineering at Sanketika Vidya Parisad Engineering College.

P.Tejaswi

Pursuing B.Tech final year in Department of Computer Science and Engineering at Sanketika Vidya Parisad Engineering College.

Ch.Dharani

Pursuing B.Tech final year in Department of Computer Science and Engineering at Sanketika Vidya Parisad Engineering College.

R.Siva

Pursuing B.Tech final year in Department of Computer Science and Engineering at Sanketika Vidya Parisad Engineering College.

K.Shivani

Pursuing B.Tech final year in Department of Computer Science and Engineering at Sanketika Vidya Parisad Engineering College.