INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Impact Factor: 7.542

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Chronic Kidney Disease Prediction using Supervised Machine Learning Approach

M.Narmadha[1] , P.Rajalakshmi[2], D.Yuvashree[3] , Mrs.G.Bharathi M.E[4]

U.G. Student, Department of Electronic and Communications Engineering, Meenakshi Sundararajan Engineering College, Arcot Road, Chennai, TamilNadu, India[1,2,3]

Associate Professor, Department of Electronic and Communications Engineering, Meenakshi Sundararajan Engineering College, Arcot Road, Chennai, TamilNadu, India[4]

ABSTRACT: The term "chronic kidney disease" means lasting damage to the kidneys that can get worse over time. If the damage is very bad, your kidneys may stop working. This is called kidney failure, or end-stage renal disease (ESRD). Kidney disease patients have the potential to get into the chronic phase and chronic kidney disease (CKD) is a decrease in kidney function gradually. So, doctors can diagnose kidney disease patients. So, we are predicting whether patients with renal disease have entered a phase of chronic kidney disease or not by showing the best accuracy result of comparing supervised classification machine learning algorithms in real time applications. The aim is to investigate machine learning based techniques for CKD forecasting by prediction results in best accuracy. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multivariate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Additionally, to compare and discuss the performance of various machine learning algorithms from the given hospital dataset with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the GUI based effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, recall, F1 score, sensitivity and specificity.

KEYWORDS: Dataset ,Machine learning-Classification method, python, Prediction of Accuracy result.

I. INTRODUCTION

Domain Overview: Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves the use of specialized algorithms. It feeds the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated into three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning programs are both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning has no labels. It provided the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they "learn" about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observations.

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the

mapping function from the input to the output is y = f(X). The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression,Random Forest tree, k-Nearest Neighbour, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems.

II. METHODOLOGY

EXPLORATORY DATA ANALYSIS

This analysis is not meant to be providing a final conclusion on the reasons leading to the hospital sector as it does not involve using any inferential statistics techniques/machine learning algorithms. Machine learning supervised classification algorithms will be used to give the CKD/ NOT CKD dataset and extract patterns, which would help in predicting the likely patient affected or not, thereby helping the hospitals for making better decisions in the future. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.
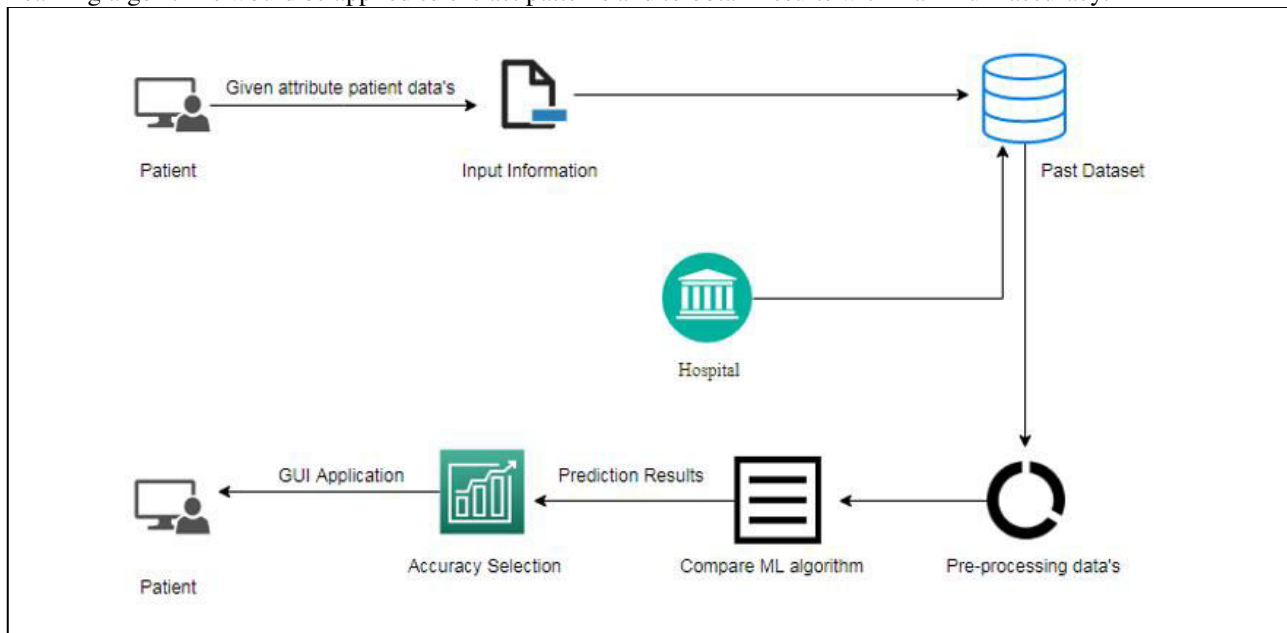

Figure 1. Block diagram

DATA WRANGLING:

In this section of the report will load in the data, check for cleanliness, and then trim and clean the given dataset for analysis. Make sure that the document steps carefully and is justified for cleaning decisions.

DATA COLLECTION:

The data set collected for predicting loan customers is split into Training set and Testset. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms, K-NearestNeighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data needs to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion needs to be done. Based on the correlation among attributes it was observed that attributes that are significant individually include property area, education, loan amount, and lastly credit history, which is the strongest among all.

Variable	Description
age	Age of patient (in years)
bp	Blood pressure(mm/Hg)
sg	Specific gravity
al	Albumin
su	Sugar
rbc	Red blood cells
pc	Pus cell
pcc	Pus cell clumps
ba	Bacteria
bgr	Blood glucose random
bu	Blood urea (mgs/dl)
sc	Serum creatinine (mgs/dl)
sod	Sodium (mEq/L)
pot	Potassium (mEq/L)
hemo	Hemoglobin (gms)
pcv	Packed cell volume
wc	White blood cell count (cells/cumm)
rc	Red blood cell count (millions/cmm)
htn	Hypertension
dm	Diabetes mellitus
cad	Coronary artery disease
appet	Appetite
pe	Pedal edema
ane	anemia
class	Class of ckd/notckd

Table 1. Variable description

Some variables such as applicant income and co-applicant income are not significant alone, which is strange since by intuition it is considered as important.The correlation among attributes can be identified using a plot diagram in data visualization process. Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of loan data removed several attributes that have no significance about the behavior of a customer. Data integration, data reduction and data transformation are also applicable for loan data. For easy analysis, the data is reduced to some minimum amount of records. Initially the Attributes which are critical to make a loan credibility prediction are identified with information gain as the attribute-evaluator and Ranked as the search-method.
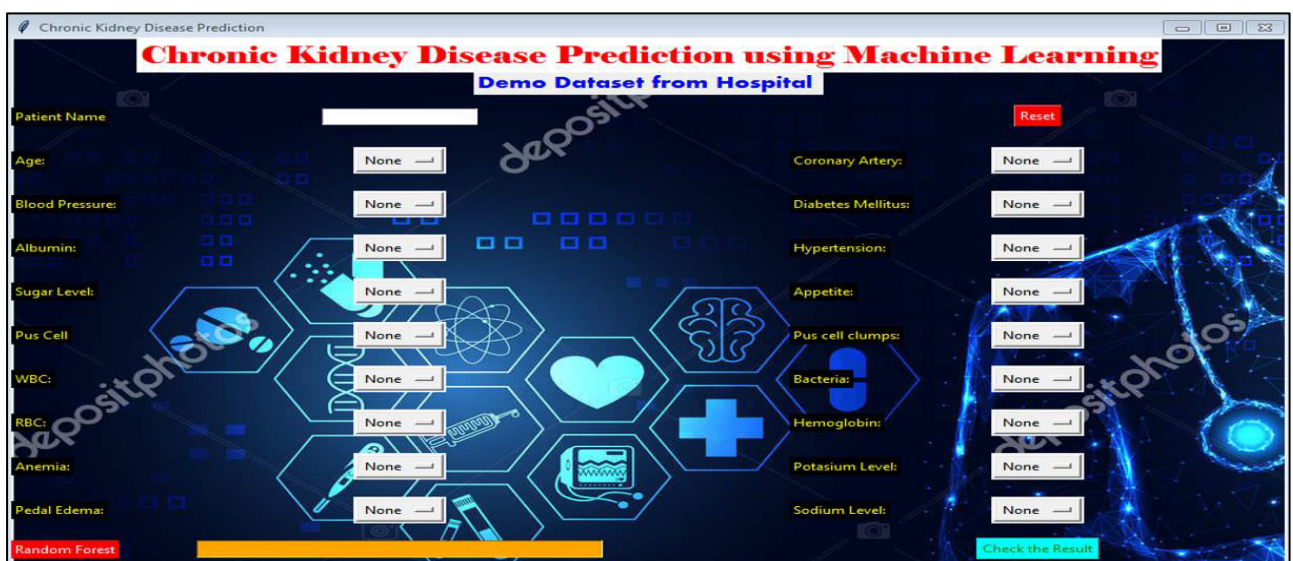
IV. EXPERIMENTAL RESULTS

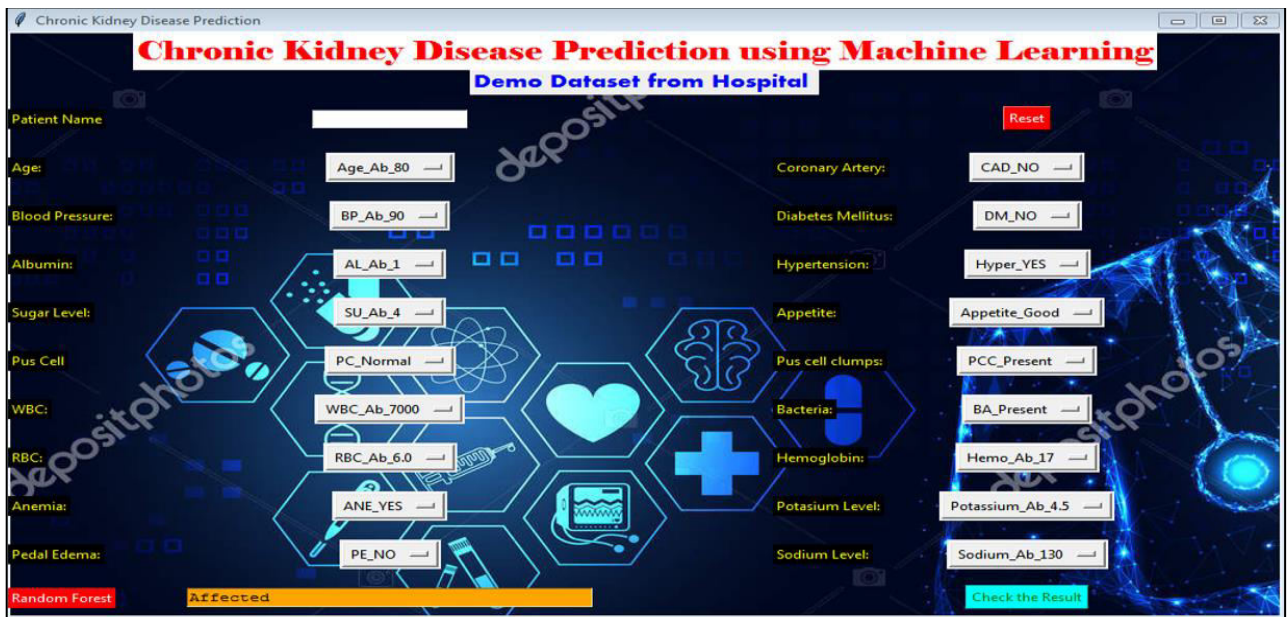**Figure 2**. Output displayed in GUI where patient details should be entered

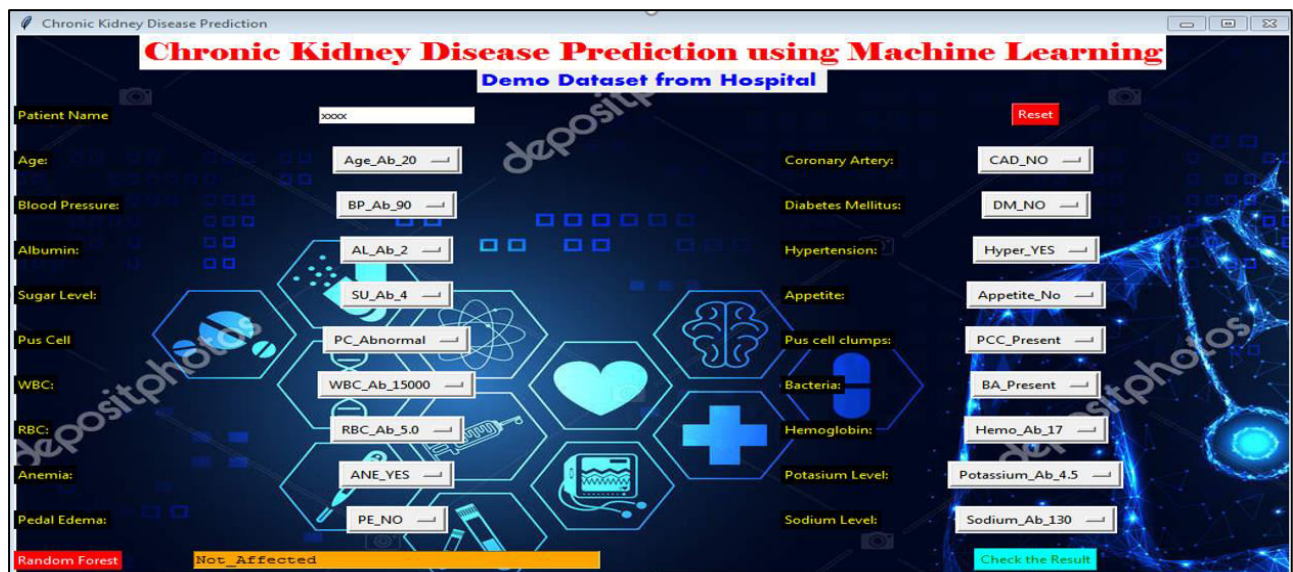**Figure 3**. Output for CKD affected patient

Figure 3. Output for CKD not affected patient

V. CONCLUSION

The implementation of Chronic Kidney disease prediction model based on supervised method of machine learning to predict whether the person is affected by ckd or not. By learning from the past dataset's input and results, the model can predict the result for a newly entered dataset. The dataset is supplied to the machine learning model on the basis of this data set the model is trained. In the process of accumulating information, data from previously CKD affected patients datasets from online sources are gathered together. These datasets are merged to form a common dataset, on which analysis will be done. Various attributes like age,bp,albumin,sugar levels etc are used to predict the result. We found that Random Forest has the greatest accuracy of 96.67% .We then uses this algorithm to predict the result from newly entered data.We show that this approach can further improve the prediction performance.

REFERENCES

[1] Zeinab Sedighi, Ebrahimpour-Komleh, Seyed Jalaleddin Mousavirad,''Feature selection effects on kidney disease analysis'',Second International Congress on Technology, Communication and Knowledge (ICTCK 2015) November, 11-12, 2015 - Mashhad Branch, Islamic Azad University, Mashhad, Iran.

[2] Fadil Iqbal, Aruna S. Pallewatte and Janaka P. Wansapura,"Texture Analysis of Ultrasound Images of Chronic Kidney Disease",2017 International Conference on Advances in ICT for Emerging Regions (ICTer): 299 - 303.

[3] Natalia Myakina, Alla Vigel and Igor Lots,"Time in Range Analysis and Glucose Variability in Type 1 Diabetic Patients with Different Stages of Chronic Kidney Disease",2018 11th International Multiconference Bioinformatics of Genome Regulation andStructure\Systems Biology (BGRS\SB).

[4] Sankhadeep Chatterjee, Simona Dzitac and Soumya Sen,"Hybrid Modified Cuckoo Search-Neural Network in Chronic Kidney Disease Classification",2017 14th International Conference on Engineering of Modern Electric Systems (EMES).

[5] Pinar Yildirim,"Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron",2017 IEEE 41st Annual Computer Software and Applications Conference.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462   6381 907 438   ijircce@gmail.com

Scan to save the contact details