International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT20652

Chronic Kidney Disease Prediction Using Different Algorithms

Harsh Vardhan Singh

Department of Computer Science & Engineering IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Article Info
Volume 6, Issue 5
Page Number: 06-13
Publication Issue :
September-October-2020

Article History
Accepted : 01 Sep 2020
Published : 04 Sep 2020

ABSTRACT

Chronic Kidney Disease (CKD) is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be lead to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. The objective of paper is to build the model for predicting the Chronic Kidney Disease using various machine learning classification algorithm. Classification is a powerful machine learning technique that is commonly used for prediction. Some of the classification algorithm are Logistic Regression, Support Vector Machine, Naïve Bayes, Random Forest Classifier, KNN. This paper investigate which algorithm is used for the improving the accuracy in the prediction of Chronic Kidney Disease. And, a comparative analysis on the accuracy and mean squared error is to done for predicting the best model.

Keywords: CKD, SVM, Machine Learning, Random Forest Classifier, KNN, Naïve Bayes.

I. INTRODUCTION

Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation. In simple words we can say that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure by then. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Santosh A. Shinde and Dr P. Raja Rajeswari presented a machine learning concept map and review on applications of machine learning

in healthcare domain in order to predict different disease, intellectually [5].

Machine Learning is one such tool which is widely utilized in different domains because it doesn't require different algorithm for different dataset, algorithms such as Naive Bayes, Decision Tree, KNN, Neural Network, are used to predicate risk of CKD each algorithm has its specialty such as Naive Bayes, KNN are used to probability for predicting CKD. All these techniques are using old patient record for getting prediction about new patient. This prediction system for CKD helps doctors to predict heart disease in the early stage of disease resulting in saving millions of lives.

II. LITERATURE SURVEY

There are many researchers who work on prediction of CKD with the help of many different classification algorithm. And those researchers get expected output of their model.

Siddheshwar Tekale , Pranjal Shingavi , Sukanya Wandhekar, Ankit Chatorikar[6] has worked on CKD prediction using SVM and Decision tree. the used the CKD data set to predict the results.

Gunarathne W.H.S.D et.al. [7] Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes

S.Dilli Arasu and Dr. R. Thirumalaiselvi [1] has worked on missing values in a dataset of chronic Kidney Diease. Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing so they got up with unknown values. They replaced missing values with recalculated values.

Asif salekin and john stankovic [2] they use novel approach to detect CKD using machine learning algorithm. They get result on dataset which having 400 records and 25 attributes which gives result of patient having CKD or not CKD. They use k-nearest neighbours, random forest and neural network to get results. For feature reduction they use wrapper method which detect CKD with high accuracy.

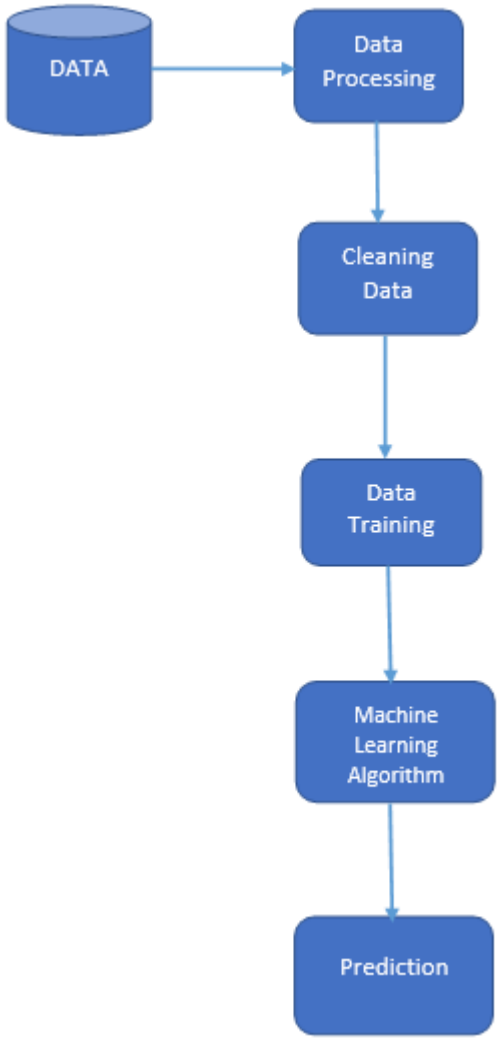
Sahil Sharma, Vinod Sharma, and Atul Sharma [4], has assessed 12 different classification algorithms on dataset which having 400 records and 24 attributes. They had compared their calculated results with actual results for calculating the accuracy of prediction results.

Pinar Yildirim [3] searches the effect of class imbalance when we train the data by using development of neural network algorithm for making medical decision on chronic kidney disease. In this proposed work, a comparative study was performed using sampling algorithm. This study reveals that the performance of classification algorithms can be improved by using the sampling algorithms. It also reveals that the learning rate is a crucial parameter which significantly effect on multilayer perceptron.

(CVD) like hypertension, diabetes mellitus, dyslipidaemia, and metabolic syndrome. CKD also leads to End Stage Renal Disease (ESRD) which has no cure. U. N. Dulhare.et al [4] extracted action rules based on stages but also predicted CKD by using naïve bayes with One R attribute selector which helps to prevent the advancing of chronic renal disease to further stages.

III. Methodology

Flow Chart:



Description of Dataset:

The Dataset has been taken from the Kaggle named 'Chronic Kidney Disease UCI'. This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period.

Description of Dataset:

The Dataset has been taken from the Kaggle named 'Chronic Kidney Disease UCI'. This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period.

Source: Dr.P.Soundarapandian.M.D.,D.M (Senior Consultant Nephrologist), Apollo Hospitals, Managiri, Madurai Main Road, Karaikudi, Tamilnadu, India. This dataset includes 400 patients' records with 25 attributes. All this 25 attributes are main attributes which are related to CKD disease.

Data Set Information:

age – age
bp - blood pressure 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
sc - serum creatinine sod – sodium
pot – potassium hemo – hemoglobin
pcv - packed cell volume wc - white blood cell count
rc - red blood cell count htn – hypertension
dm - diabetes mellitus
cad - coronary artery disease appet – appetite
pe - pedal edema ane – anemia class – class

Dataset needed to be cleaned as there were null "NaN" values present and

Classifier:

Support Vector Machine (SVM)

A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

An SVM algorithm should not only place objects into categories, but have the margins between them on a graph as wide as possible.

Some applications of SVM include:

- Text and hypertext classification
- Image classification
- Recognizing handwritten characters
- Biological sciences, including protein classification

Naive Bayes algorithm (NB)

This is a classification algorithm which is used when the dimensionality of the input is very high. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is based on Bayes theorem.

The Bayes theorem is as follows:

$$P(Y/X) = P(X/Y)\ P(X)$$

This calculates the probability of Y given X where X is the prior event and Y is the dependence event.
It needs less training data. It can be used for binary classification problems and is very simple.

Decision trees

Decision trees is one of the ways to display an algorithm. It is a classic machine learning algorithm. In heart disease, there are several factors such as cigarette, BP, Hypertension, age etc. The challenge of the decision tree lies in the selection of the root node. This factor used in root node must clearly classify the data. We make use of age as the root node The decision tree is easy to interpret. They are non-parametric and they implicitly do feature selection.

Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. Solution for classification is logistic regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.

The logistic function is defined as:

$$\text{logistic}(\eta)=11+\exp(-\eta)$$

Logistic regression can suffer from complete separation. If there is a feature that would perfectly separate the two classes, the logistic regression model can no longer be trained. This is because the weight for that feature would not converge, because the optimal weight would be infinite. This is really a bit unfortunate, because such a feature is really useful. But you do not need machine learning if you have a simple rule that separates both classes. The problem of complete separation can be solved by introducing penalization of the weights or defining a prior probability distribution of weights.

K-nearest neighbors (KNN)

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well –

- Lazy learning algorithm – KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm – KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

Random Forest

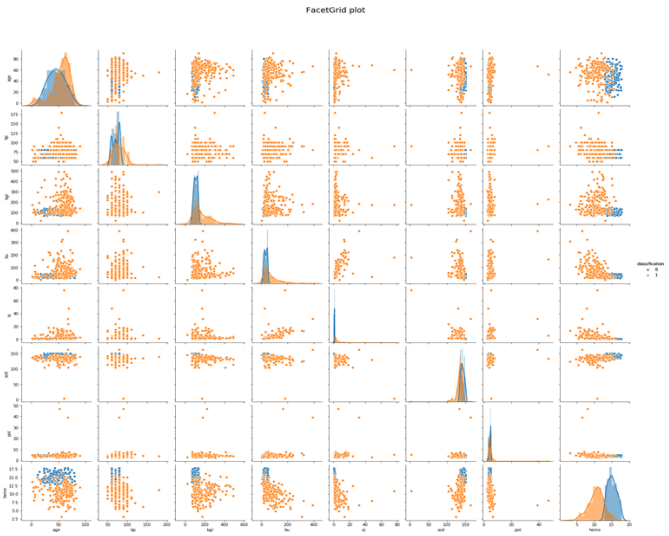
Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

IV. Implementation

Before applying any algorithm data needed to be cleaned as there were lots of "NULL" values and attributes were in 'object' datatype . To solve this I first fill t he null values using fillna() function from panda library and then datatypes wer e changed. After data cleaning algorithm such as KNN, Navie-Bayes, Logistic Re gression, Decision Tree, Random Forest, Support Vector Machine were applied.

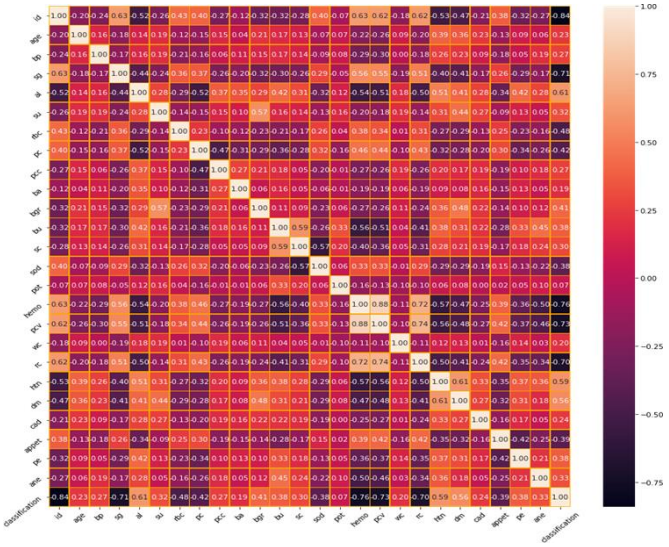
V. Result

FacetGrid Plot:



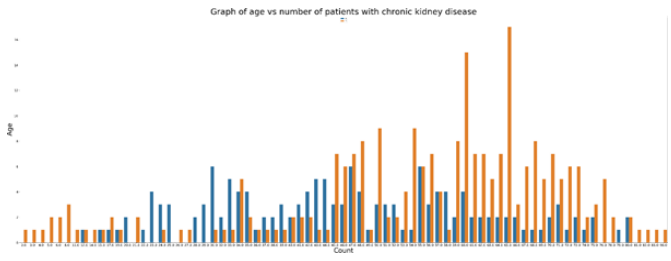
- We do observe distinct classification of positive and negative result in each targeted health attribute.
- We have fairly similar mean age among these 2 test subjects but we do observe significantly more test subjects tested negative for chronic kidney disease.
- The trade-off for choosing to retain data is the creation of less well- behaved results.

Heat Map:

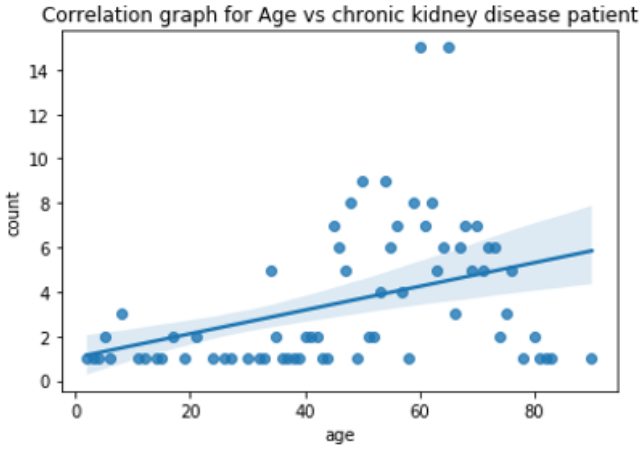


- I will say that an absolute value of more than 0.4 is considered to be significant.
- It seems like there are a significant negative correlation between rbc, pc, etc. and whether the patient has chronic kidney disease.
- Even so, I will look into age, red blood cell, pus cell, blood glucose random, serum creatinine, diabetes mellitus, coronary artery disease, blood urea, sodium, pedal edema and anemia.

Correlation between age and whether a patient has chronic kidney disease



	age	count
age	1.000000	0.387084
count	0.387084	1.000000

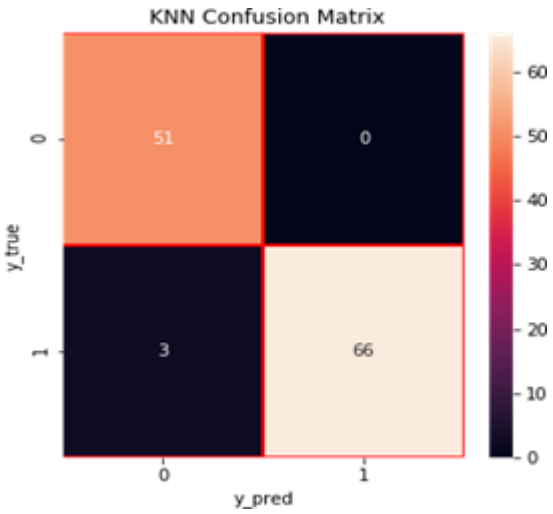


	age	count
age	1.00000	-0.13443
count	-0.13443	1.00000

- From what we know, we have approximately 150 healthy subjects and 250 chronic kidney disease patients.
- There is a weak positive correlation between age and chronic kidney disease patients. We obtained an R value of approximately 0.387 and an R-square value of approximately 0.150. This means that only 15% of variation can be explained by the relationship between the 2 variables.
- The National Kidney Foundation has associated aging with kidney disease, stating that "more than 50 percent of seniors over the age of 75 are believed to have kidney disease. Kidney disease has also been found to be more prevalent in those over the age of 60 when compared to the rest of the general population."

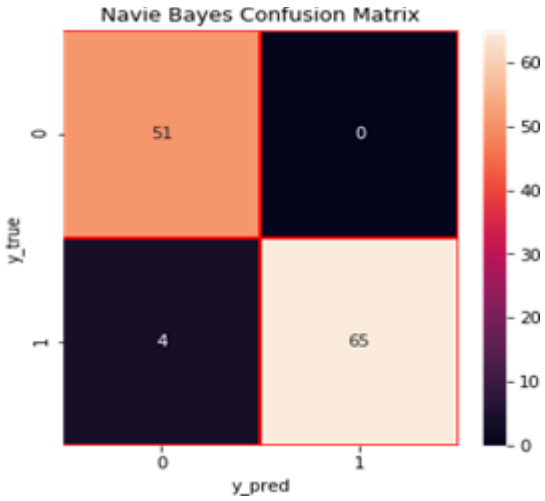
Algorithms:

1. KNN:



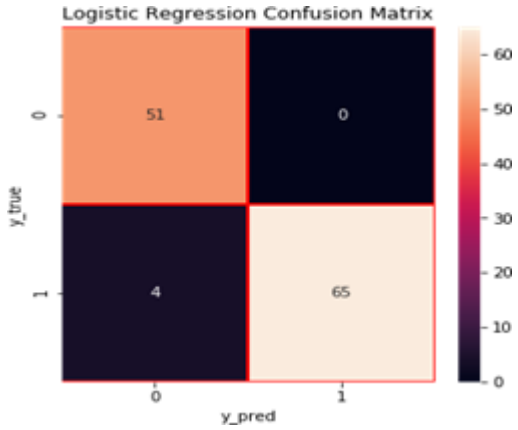
KNN accuracy = 97.5

2. Navie-Bayes:



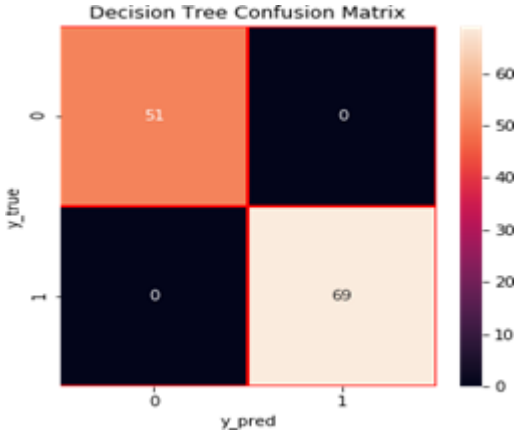
Navie Bayes accuracy = 96.66 666666666667

3. Logistic Regression:



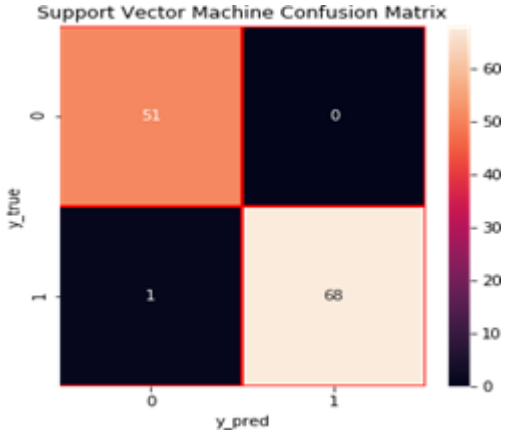
test accuracy 96.66666666666667

4. Decision Tree:

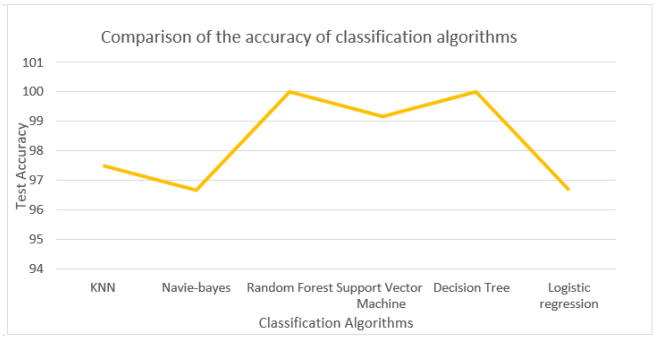


Decision Tree accuracy: 100. 0

5. Support Vector Machine:



SVM test accuracy = 99.16666 666666667



VI. CONCLUSION

In this paper I have studied different machine learning algorithms. I have analy sed 14 different attributes related to CKD patients and predicted accuracy for d ifferent machine learning algorithms like Decision tree, Support Vector Machin e, KNN, Navie-bayes, Support Vector Machine and Logistic Regression. From th e results analysis, it is observed that the decision tree and Random Forest algor ithms gives the accuracy of 100%, SVM gives 99.16667%, KNN gives 97.5% and Navie-byes and logistic Regression gives accuracy of 96.66666667. Limitations of this study are the strength of the data is not higher because of the size of th e data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with more accurate prediction, will need milli ons of records with zero missing values.

VII. REFERENCES

[1]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, “Review of Chronic Kidney Disease b ased on Data Mining Techniques”,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505

[2]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Sele cting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[3]. Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE Inte rnational Conference on Computer Software and Applications (COMPSAC), IEE E, Jul. 2017, doi: 10.1109/COMPSAC.2017.84

[4]. Sahil Sharma, Vinod Sharma, Atul Sharma, “Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Dise ase Diagnosis,” July18, 2016.

[5]. S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning : a review,” IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[6]. Siddheshwar Tekale1 , Pranjal Shingavi2 , Sukanya Wandhekar3 , Ankit Chat orikar4, “Prediction of Chronic Kidney Disease Using Machine Learning Algorith m” IJARCCE Vol. 7, Issue 10, October 2018

[7]. Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, “Perform ance Evaluation on Machine Learning Classification Techniques for Disease Clas sification and Forecasting through Data Analytics for Chronic Kidney Disease (C KD)”,2017 IEEE 17th International Conference on Bioinformatics and Bioengine ering.

[8]. U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

Author Profile:

Harsh Vardhan Singh is a B. tech 3rd year Student in department of Computer Science and Engineering College, Ghaziabad, UP, India. He has also worked as Trainee data scientist in Edulyt India also completed his 4 projects which were website blocker, Interactive Dictionary, Credit Card Approval and Loan Approva l Prediction with ML.
[9].

Cite this article as :

Harsh Vardhan Singh, "Chronic Kidney Disease Prediction Using Different Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 06-13, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT20652
Journal URL : http://ijsrcseit.com/CSEIT20652