IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

Medical data classification for prediction of early chronic kidney disease

Sathyanarayana S[1], Sandeep B[2]

Assistant Professor, Computer Science Department, JNNCE, Shivamogga, India[1]

Assistant Professor, Computer Science Department, JNNCE, Shivamogga, India[2]

Abstract: Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for early prediction of CKD. Machine learning methods are effective in CKD prediction. This document proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. We also consider the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.

Keywords: Chronic Kidney Disease, Data Mining, Classification, Healthcare, Machine Learning.

I. INTRODUCTION

Kidney diseases are increasing day by day among people. Chronic kidney disease is becoming major health issue around the world. Not maintaining proper food habits and drinking less amount of water are one of the major reasons that contribute this condition. Caused due to lack of water consumption, smoking, improper diet, loss of sleep and many other factors. This disease affected 753 million people globally in 2016 in which 417 million are females and 336 million are males. Majority of the time the disease is detected in its final stage, and which sometimes leads to kidney failure. Chronic kidney disease is detected during the screening of people who are known to be in threat by kidney problems, such as those with high blood pressure or diabetes and those with a blood relative chronic kidney disease patient.

II. KDD PROCESS

The KDD process includes selection of relevant data; it's processing, transforming processed data into valid information and then extracting hidden information/pattern from it. The KDD process can be categorized as under:

A. Selection: It includes selecting data relevant for the task of analysis from the database.

B. Pre-Processing: In this phase we remove noise and inconsistency found in data and combine multiple data sources.

C. Transformation: In this phase transformation of data takes place into appropriate forms to perform mining operations.

D. Data Mining: This phase includes applying data mining algorithm appropriate for extracting patterns.

E. Interpretation/Evaluation: Interpretation/Evaluation includes finding the relevant patterns of information hidden in the data.

III.SYSTEM DESIGN

A. The proposed system for CKD analysis and prediction is shown. The raw data is first analysed machine learning works only with numericals so the text data is converted into numerical value.

B. Check for missing values is done. If there is an attribute with more than 20% of missing values, then that attribute is discarded.

C. If it is less than 20%, that missing value is filled using several methods like KNN impute which is inbuilt in python.

D. In feature selection, the attributes which are very corelated to target variable, only those are considered for predicting the CKD.

E. After gaining the data, several models are trained using this data.

F. Model which gives the highest accuracy is considered for predicting the CKD.

G. Then we predict whether patient is having CKD or not CKD.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

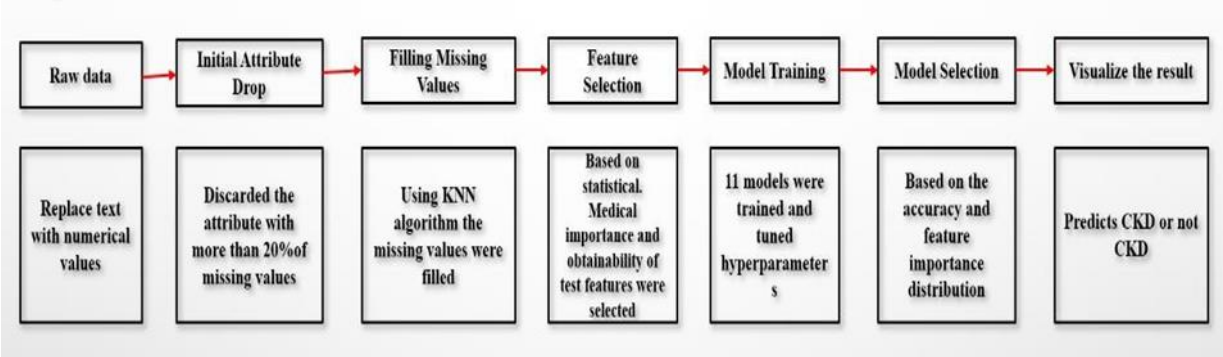

Fig. 1 Architecture of Early Prediction of CKD

IV. IMPLEMENTATION

The dataset is taken from the Kaggle. It contains 410 rows and 26 columns.

```
df.head()
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd

5 rows × 26 columns

Fig. 2 Details of Dataset

A. Outliers Detection

Outlier is the data object that deviates significantly from the rest of the data object and behaves in a different manner. An outlier is a data object that diverges essentially from the rest of the objects as if it were produced by several mechanisms. For the content of the demonstration, it can define data objects that are not outliers as “normal” or expected data. Usually, it can define outliers as “abnormal” data. Outliers are data components that cannot be combined in a given class or cluster. These are the data objects which have several behaviours from the usual behaviour of different data objects. The analysis of this kind of data can be important to mine the knowledge. Outliers are fascinating because they are suspected of not being created by the same structure as the rest of the data. Hence, in outlier detection, it is essential to justify why the outliers identified are produced by several mechanisms.

B. Correlation of all the Attributes

Correlation is used to quantify the degree to which 2 variables are related. Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. Correlation is often used to determine whether there is a cause-and-effect relationship between two variables. Correlation does not necessarily imply causation; other factors may be at play. However, it is important to remember that correlation does not imply causation. For example, there may be a strong correlation between ice cream sales and swimming accidents, but that doesn’t mean that eating ice cream causes people to have accidents.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

Fig. 3 Correlation of Attributes

C. Feature Selection (Using Extra Regressor)

It is the process of reducing the number of input variables when developing the predictive model. Feature selection is also called variable selection or attribute selection. It is the automatic selection of attributes in your data (such as columns in tabular data) that are most relevant to the predictive modelling problem you are working on feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them.

```
# plot graph for feature importance
feat_importance = pd.Series(model.feature_importances_,index=x.columns)
feat_importance.nlargest().plot(kind="barh")
plt.show()
```

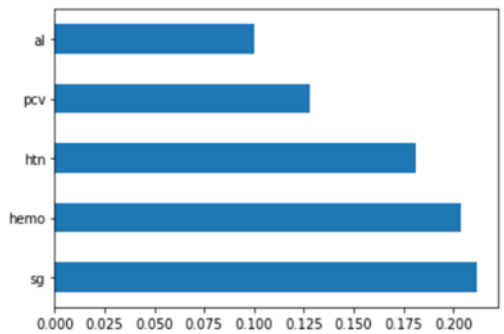

Fig. 4 Feature Selection

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

V. CLASSIFICATION ALGORITHS WITH RESULTS AND SNAPSHOTS

A. KNN: K-nearest neighbor Classification

Step-1: Select the number K of the neighbors.
Step-2: Calculate the Euclidean distance of K number of neighbors.
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.
The Euclidian distance formula is shown below

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

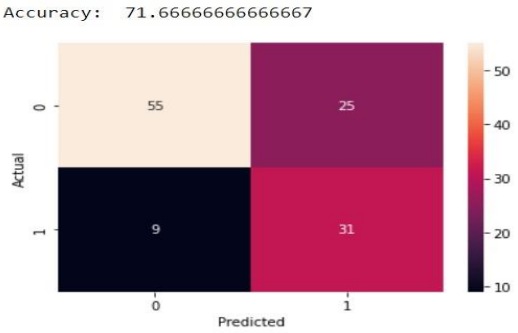

fig 5.1: Confusion matrix of KNN

B. Decision Tree Algorithm(ID3)

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

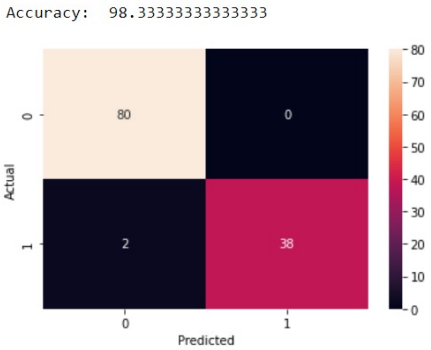

fig 5.2: Confusion matrix of ID3

C. Naïve Bayes

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

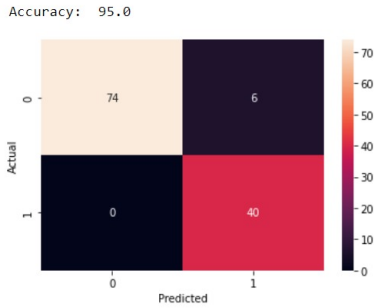

fig 5.3: Confusion matrix of Naïve bayes

D. Support Vector Machine

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

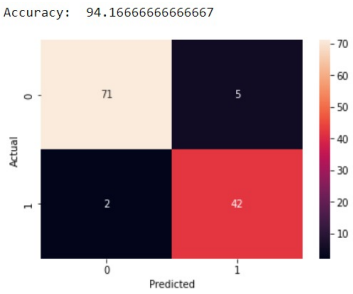

fig 5.4: Confusion matrix of SVM

E. Random Forest

- Random Forests are composed of multiple independent decision trees trained independently on a random subset of data.
- Trees are generated at the time of training, and the outputs are obtained from each decision tree.
- A random forest is a **meta estimator** that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- The accuracy obtained by training the model using this particular algorithm is 100%.
- The sub-sample size is controlled with the max_samples parameter.
- if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

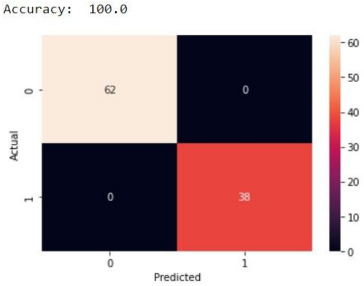

fig 5.5: confusion matrix of random forest

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117105

Accuracy of algorithms

	Algorithms	Accuracy
1	K-nearest neigbour (KNN)	71.66666
2	Decision tree algorithm(ID3)	98.33333
3	Naïve Bayes	95.0
4	Support vector machine	94.16666
5	Random forest	100.0

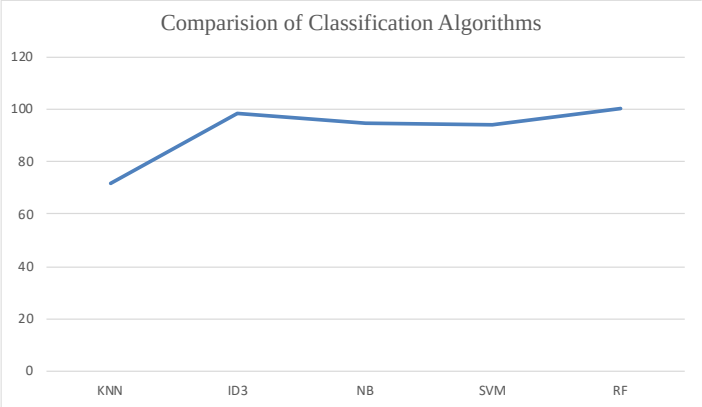

fig 5.6: comparision of different classification algorithms

VI. CONCLUSION

It has been found that the Random forest and Decision tree algorithms are more efficient in the prediction of chronic kidney diseases. Their accuracy was found to be 100% and 98.33 respectively. In the future, this work can be upgraded by building up a web application based on classification algorithms and using large amount of dataset. This will help in giving better outcomes and help healthcare experts in the early prediction of chronic kidney disease.

REFERENCES

[1] Narander Kumar, Department of Computer Science & Engineering "Implementing WEKA for medical data classification and early disease prediction" 3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017)

[2] Kallu Samatha, Muppidi Rohitha Reddy, Pattan Faizal Khan, Rayapati Akhil Chowdary, PVRD Prasada Rao, International Journal of Preventive Medicine and Health (IJPMH) "Chronic Kidney Disease Prediction using Machine Learning Algorithm" ISSN: 2582-7588(Online), Volume- 1 Issue-3, July 2021

[3] Sathiya Priya S (PG Scholar), Suresh Kumar M (Professor) Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, International Journal of Computer Science and Information Security (IJCSIS), "Chronic Kidney Disease Prediction Using Machine Learning" Vol. 16, No. 4, April 2018

[4] Pradeep Nijalingappa and Sandeep B. Machine learning approach for the identification of diabetes retinopathy and its stage, International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE International Conference, 2015.

[5] Imesh Udara Ekanayake, Damayanthi Herath Dept. Computer Engineering University of Peradeniya Peradeniya, 20400, Sri Lanka, "Chronic Kidney Disease Prediction Using Machine Learning Methods" Conference Paper · September 2020

[6] Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A. Dept of Computer Science and Engineering LBS Institute of Technology for Women, Thiruvananthapuram, Kerala, International Journal of Engineering Research & Technology (IJERT) "Chronic Kidney Disease Prediction using Machine Learning" ISSN: 2278-0181 IJERTV9IS070092 Vol. 9 Issue 07, July-2020