Machine Learning Models For Predicting Chronic Kidney Disease

Jobina Kurian
Department Of Computer Application
Amal Jyothi College Of Engineering,
Kanjirappally, Kottayam
jobinakurian@mca.ajce.in

Sruthimol Kurian
Asst.Professor
Amal Jyothi College Of Engineering,
Kanjirappally, Kottayam
sruthimolkurian@amaljyothi.ac.in

***Abstract*— Chronic Kidney Disorder is an extreme lifelong condition that induced via both kidney pathology or reduced kidney features. Chronic Kidney Disorder affects one in every five men and one in every four women globally between the ages of 65 and 74. Early prediction and right treatments can probably prevent the cease-level. Dialysis or kidney transplantation is the only way to save patients life. In my study, I examine how machine-learning can be used to predict chronic kidney disease (CKD) early. Machine learning models are effective in CKD prediction. This work proposes a workflow to predict CKD based on Clinical information's (clinical data's). Data prepossessing, missing value handling method with filtering and attributes selection are the major processes. Naïve Bayes Method/Algorithm which has highest Precision, Recall, Accuracy and F1 score. Also it is less time consuming.**

Keywords—Chronic kidney disease, machine learning, prediction

I. INTRODUCTION

Chronic Kidney Disorder(CKD) is a phenomenon in which your kidneys are damaged and not filtering your blood. The major role of kidneys is to clear out greater water content and waste from our blood and to produce urine. If a person is suffered from CKD, it means that wastes are collected in his/her body. This disease is chronic because the damage gradually over a long period. It is a common disease worldwide. CKD may have some health troubles. There are numerous reasons for CKD like diabetes, excessive blood pressure, heart disease and so forth. Along with these critical diseases, CKD also depend on age and gender. 10% of the population worldwide is affected by CKD, and thousands and thousands die every year because of lack of diagnostic measures .

CKD, in its early stages, has no symptoms. Testing may be the only way to find out if the patient has any kidney disease. Early identification of CKD can assist patients receive appropriate treatment and avoid the development of ESRD(End-Stage Renal Diseases) is suggested that everyone with one of the CKD risk factors, such as a family history of renal failure, hypertension, or diabetes, have their.

The goal of this work was to develop machine-learning models for predicting CKD. This take a look at analyses CKD the use of machine learning strategies with CKD dataset from the ' Kaggle ' data warehouse. By normalizing missing data, the dataset can be pre-processed. The most applicable functions are decided on the dataset to improve accuracy and decrease training time for machine learning strategies.

Test is done through Colaboratory, or "Colab" for short, which is a python library. It allows user to write and execute Python in your browser, with Zero configuration required, Free access to GPUs and Easy sharing. Generate the dataset's accuracy, recall, F1 score, precision, confusion matrix, and classification reports based on the models.

II. LITERATURE REVIEW

A. J. Snegha, 2020 : Proposed a device that makes use of diverse statistics mining strategies like Random forest with set of rules and Back Propagation Neural network. Right here they examine each of the algorithms and found that Back Propagation set of rules offers the first-class end result as it makes use of the Supervised mastering community known as Feedforward Neural community.

B. Siddheshwar Tekale, 2018 : Described a system the usage of device studying which makes use of decision tree and SVM strategies. By evaluating two strategies finally concluded that SVM offers the satisfactory result. Its prediction manner is much less time consuming so that doctors can examine the sufferers inside a much less term.

C. Guneet Kaur, 2017 : Proposed a machine for predicting the CKD. They use statistics mining classifiers like KNN and SVM. Here the predictive analysis is done primarily based upon the manually selected statistics columns. SVM classifier offers the first-class accuracy than KNN set of rules.

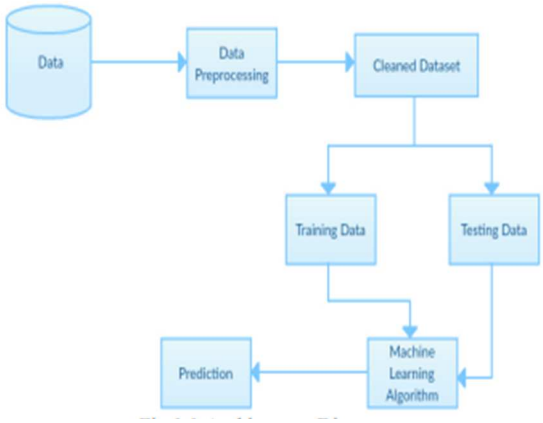
III. DATASET AND METHODS

A. Dataset

We use the CKD Dataset from Kaggle repository, which is publicly available.. There are 25 attributes, 11 of which are numeric, 13 of which are non-numeric, and one is a class attribute. The data set contains missing values. Dataset contains age, blood pressur, specific gravity , albumin , sugar , packed cell volume wc - white blood cell count , red blood cell count , blood glucose level , blood urea , serum creatinine , sodium , potassium , hemoglobin , hypertension ,diabetes mellitus , cad - coronary artery disease , appetite , pedal edema etc.

B. Steps.

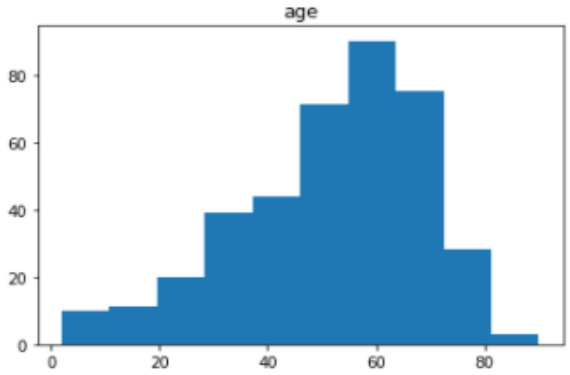
DOI: 10.5281/zenodo.6369500
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam



Exploratory data analysis (EDA)

1) For numeric data

- Made histograms to understand distributions



(Histogram Based on Age)

- Corrplot
- Pivot table comparing survival rate across numeric variables

	age	al	bgr	bp	bu	hemo	pot	sc	sg	sod	su
classification											
ckd	54.425000	1.721154	175.523810	79.705882	72.656170	10.652217	4.883030	4.430720	1.013937	133.882530	0.770732
ckd\t	68.500000	2.000000	164.500000	70.000000	41.000000	9.700000	4.500000	2.550000	1.010000	135.500000	0.000000
notckd	46.516779	0.000000	107.722222	71.351351	32.798611	15.188194	4.337931	0.868966	1.022414	141.731034	0.000000

(Pivot Table)

2) For Categorical Data

- Made bar charts to understand balance of classes
- Made pivot tables to understand relationship with survival

a. Data Pre-processing :

- deal with unwanted text.

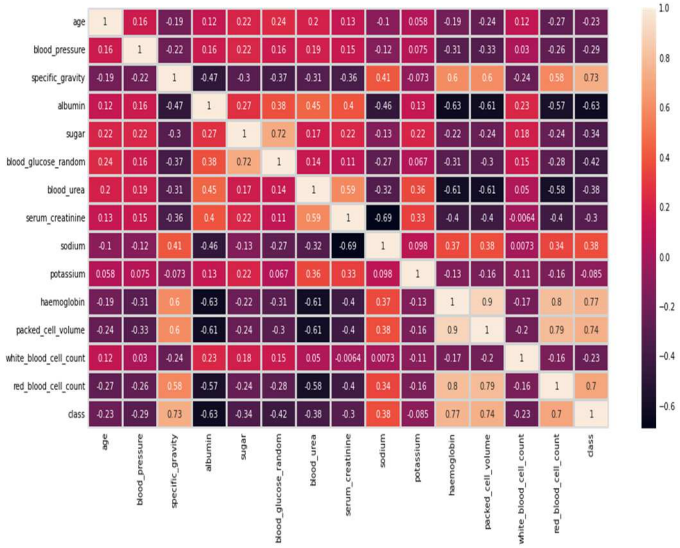
Text data can be converted into numerical data with the help of LabelEncoder. Label encoding is the process of translating labels into numeric format so that they may be read by machines.

- deal with missing values

Missing Value Handling :

In this work, filling null values, we will use two methods, random sampling for higher null values and mean/mode sampling for lower null values. Filling “num_cols” null values using random sampling method and filling "red_blood_cells" and "pus_cell" using random sampling method and rest of the columns are filled using mode imputation.

The absolute values of heat map of correlations of attributes to the class label show that haemoglobin, specific gravity, albumin, hypertension and diabetes mellitus have the highest correlations that is more than 0.5. Then the secondary attributes pus_cell, blood glucose random, appetite, blood urea, pedal edema , sugar, anemia and serum creatinine are the attributes that have correlations of extra than 0.3.



b. Training and Testing Dataset:

The dataset is split into 2 sub datasets each containing 14 attributes.

Training data set: training dataset is derived from main dataset and it carries 300 out of 400 records.

Testing data set: trying out dataset contain 100 out of 400 records.

c. Classifiers:

Logistic Regression: It is used to model a binary categorical outcome. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable has a binary solution. The link between one dependent binary variable and one or more independent variables is assessed using logistic regression. It gives discrete outputs ranging between 0 and 1.

Naive Bayes: It uses probabilities. Probability is defined as a number between 0 and 1 (i.e., a percentage between 0% and 100%). Probability of 0 means event will definitely not occur and probability of 1 means that the event will occur with 100 percent certainty. Bayesian classifiers use training data to calculate the likelihood of each occurrence based on the information provided by feature values.

DOI: 10.5281/zenodo.6369500
ISBN: @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Decision Tree: Decision Data is continuously split according to a parameter in Decision Trees, a type of Supervised Machine Learning. The tree can be explained by two entities, namely decision nodes and leaves.They are also be able to generate the maximum affecting characteristic inside the mass of population. Decision tree is based totally on Entropy and Information Gain. There are two downsides to Decision Trees: Overfitting and Greedy methodology.

Support Vector Classifier: A supervised machine learning algorithm is SVC. It's mainly used in category problems. In this set of guidelines, we plot each records objects as a factor in n-dimensional place (in which n is range of features you have). The hyperplane that clearly distinguishes the two classes is being used to classify the data.

KNN: It is, in fact, one of most fundamental and frequently used machine learning method. In k-NN classification, examples are classified according to their nearest neighbors. k means a variable implying that any number of nearest neighbors could be used.

Random Forests (Decision Tree Forests): It is another ensemble-based method. Focuses only on ensembles of decision trees. Easier to use. Less prone to overfitting, more powerful, more versatile, and one of the most used machine learning approaches.

Confusion Matrix:

The matrix is divided into two axes, one axis represents actual value while the other represents predicted value. It is a 2*2 matrix.

Accuracy:

Based on input data, or training data, the accuracy of machine learning models is determined by the model's ability to identify relationships and patterns among variables.

Accuracy=(TruePositive+TrueNegative)/(TruePpsitive+ TrueNegative+FalsePositive+FalseNegative)

Precision:

As a result, precision p is defined as the total number of positive instances accurately identified divided by total number of positive instances classified.

Precision = TruePositives / (TruePositives + FalsePositives)

Recall:

The recall r indicates how many positive examples were correctly classified, compared to all positive examples in the sample.

Recall = TruePositives / (TruePositives + FalseNegatives)

F-score:

The F-score is also called the F1-score. F1-score is a harmonic mean of recall and precision.. It takes both false positive and false negatives into account. Therefore, it performs well on an imbalanced dataset.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

d. Prediction:

Prediction using Logistic Regression

```
 f1 score:  0.9827586206896551
[[41  1]
 [ 1 57]]
 Accuracy  0.9791733333333335
 precision score:  0.9827586206896551

 recall score:  0.9827586206896551

              precision    recall  f1-score   support

           0       0.98      0.98      0.98        42
           1       0.98      0.98      0.98        58

    accuracy                           0.98       100
   macro avg       0.98      0.98      0.98       100
weighted avg       0.98      0.98      0.98       100
```

Prediction using k-nearest neighbors (K-nn)

```
f1 score:  0.8571428571428571

[[36  4]
 [12 48]]
 Accuracy:  0.8948933333333333
 precision score:  0.9230769230769231

 recall score:  0.8

              precision    recall  f1-score   support

           0       0.75      0.90      0.82        40
           1       0.92      0.80      0.86        60

    accuracy                           0.84       100
   macro avg       0.84      0.85      0.84       100
weighted avg       0.85      0.84      0.84       100
```

Prediction using naive bayes

```
f1 score:  1.0

[[37  0]
 [ 0 63]]
 Accuracy:  1.0
 precision score:  1.0

 recall score:  1.0

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        37
           1       1.00      1.00      1.00        63

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Prediction using support vector classification

DOI: 10.5281/zenodo.6369500
ISBN: @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

```
f1 score:  0.7730061349693252

[[ 0 37]
 [ 0 63]]
Accuracy:  0.7623200000000001
precision score:  0.63

recall score:  1.0

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        37
           1       0.63      1.00      0.77        63

    accuracy                           0.63       100
   macro avg       0.32      0.50      0.39       100
weighted avg       0.40      0.63      0.49       100
```

Prediction using Decision Tree

```
f1 score:  1.0

[[37  0]
 [ 0 63]]
Accuracy:  0.99892
precision score:  1.0

recall score:  1.0

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        37
           1       1.00      1.00      1.00        63

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Prediction using Random Forest

```
[[37  0]
 [ 1 62]]
Accuracy:  0.9998133333333333
precision score:  1.0

recall score:  0.9841269841269841

              precision    recall  f1-score   support

           0       0.97      1.00      0.99        37
           1       1.00      0.98      0.99        63

    accuracy                           0.99       100
   macro avg       0.99      0.99      0.99       100
weighted avg       0.99      0.99      0.99       100
```

IV. CONCLUSION

In this paper we've got understanding of machine learning algorithms. We analysed 14 different attributes related to CKD and predicted F1 score, Accuracy, Precision and Recall with the machine learning algorithms like Logistic Regression, k-nearest neighbors (K-nn), naive bayes, decision tree, Random forest and support vector classification. From the effective evaluation, it is observed that the Logistic Regression offer the accuracy of 0.98 , naive bayes offer the accuracy of 1.0, support vector classification offer the accuracy of 0.76, KNN algorithms offers the accuracy of 0.89, Random forest algorithms gives the accuracy of 0. 99, and Random Forest gives accuracy of 0.99. Also we can see that F1 score is highest for naive bayes algorithm , that is 1.0 . From these observations we can find out that Naïve Bayes Algorithm gives higher accuracy , precision , recall and F1 Score. So we can say that Naïve Bayes Algorithm is a better choice for predicting Chronic Kidney Disease. The benefit of this tool is that, the prediction method take less time consuming. It will help the medical doctors to begin the remedies early for the CKD sufferers and also it'll help to diagnose more patients inside a much less time period.

IV.REFERENCES

[1] "Kidney Disease: The Basics," Aug. 2014. [Online]. Available: https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics

[2] "Kidney Disease: The Basics," Aug. 2014. [Online]. Available: https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics

[3] "Global Facts: About Kidney Disease," [Online]. Available: https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease/

[4] Guneet Kaur, "Predict Chronic Kidney Disease using Data Mining in Hadoop, International Conference on Inventive Computing and Informatics, 2017.

[5] M. K. Haroun, ''Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland,'' J. Amer. Soc. Nephrol., vol. 14, no. 11, pp. 2934–2941, Nov. 2003.

[6] 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.

[7] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, ''A review of challenges and opportunities in machine learning for health,'' in Proc. AMIA Joint Summits Transl. Sci., 2020, p. 191.

[8] L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino, ''Kidney failure detection using machine learning techniques,'' in Proc. 8th Int. Workshop ADVANCEs ICT Infrastructures Services, 2020, pp. 1–8.

DOI: 10.5281/zenodo.6369500
ISBN: @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam