IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86138

Prediction of Chronic Kidney Disease using Logistic Regression

Afsha Firdose[1], Priyanka M[2], Sowmya N[3], Tejashwini S[4], Tulasi K N[5]

Assistant Professor, Information Science, BGS Institute of Technology, Mandya, India[1]

Student, Information Science, BGS Institute of Technology, Mandya, India[2]

Student, Information Science, BGS Institute of Technology, Mandya, India [3]

Student, Information Science, BGS Institute of Technology, Mandya, India [4]

Student, Information Science, BGS Institute of Technology, Mandya, India [5]

Abstract: Chronic Kidney Disease (CKD) harms the kidneys. Kidneys have the ability to dispense with squander from the body. On the off chance that the present circumstance happens, the waste gets amassed in the body. Chronic Kidney Disease (CKD) is one affliction which could pulverize the human body. It tends to be forestalled by means of inspecting not many pointers like RBC tally, explicit gravity esteem, Blood Pressure (BP), egg whites levels in pee, sugar content, sickliness and WBC check. Different conditions like coronary course illness, Diabetes Mellitus (DM) and bacterial diseases could straightforwardly influence the kidneys. [1] In this paper we have gathered some examples from a public medical clinic and chose fields have been examined for planning an expectation model for CKD. Information investigation and representation are completed to work on the measurable examination of given information. Calculated relapse is completed on the information since it contains parcel of sections with downright qualities. Exactness, accuracy, and f1 score of the model have been estimated. Different ends can be drawn from this reliant informational index and can be put away as authentic information for future examination.

Keywords: Chronic Kidney Disease (CKD), RBC count, Blood Pressure (BP), anaemia, WBC count, coronary artery disease, Diabetes Mellitus (DM) & bacterial infections, categorical values, data analysis & visualization.

I. INTRODUCTION

The Chronic Kidney Disease (CKD) is the most common disease in the world. The normal kidney has one million filtering units. Each filtering unit is called Glomerulus. The High Blood Pressure and Diabetes play a vital role in damaging the kidney. These diseases can damage the following functions of kidney. And they are: (1) Damaging their Filtering Units, (2) Collecting Tubules and (3) Causing Scarring. The Normal Kidney or a Healthy kidney can remove the waste products from one's blood and maintain equal chemical level in a human body [1]. In this work the CKD original dataset can be used to identify the disease.

We have identified several factors contributing to the failure of kidneys. Few of the listed fields are

i. Age
ii. Blood Pressure (BP)
iii. Specific gravity
iv. Albumin levels in urine
v. Diabetes Mellitus
vi. RBC count, WBC count, pus cell and packed cell volume vii. Presence or absence of hypertension, coronary artery disease and pedal edema

II. PROBLEM STATEMENT

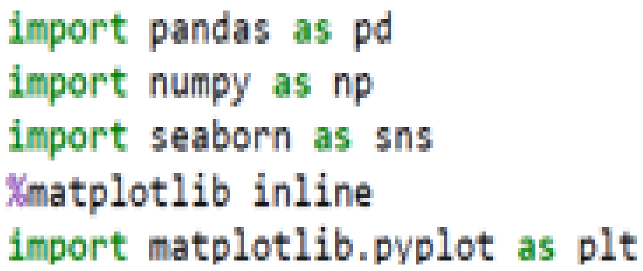
Chronic Kidney Disease prediction is one of the most important issues in healthcare analytics. The most interesting and challenging tasks in day to day life is prediction in medical field. we employ some machine learning techniques for predicting the chronic kidney disease using clinical data. The performance of the above models are compared with each other in order to select the best classifier in predicting the chronic kidney disease for given data.

III.METHODOLOGY

1 Importing Libraries

DOI: 10.17148/IARJSET.2021.86138

- **Numpy:** Numpy stands for Numerical Python. This library is used for fast mathematical computation on arrays and it is referred as 'np' in our project.
- **Pandas:** pandas is imported as pd and it is used for data analyzing.
- **Matplotlib.pyplot:** a collection of command style functions that make matplotlib work like MATLAB. It is imported as plt.
- **Seaborn:** a Python data visualization library based on matplotlib for attractive and informative statistical graphics.



```
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline
import matplotlib.pyplot as plt
```

Fig 1: shows the Python code to import libraries.

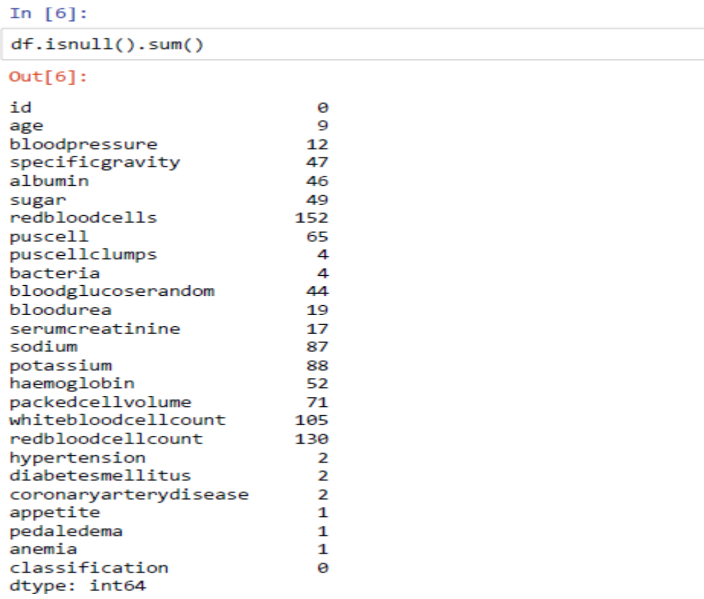
2 Importing data: the Python code to import data from respective directory/ file and assigning it to DataFrame df. The data stored in CSV format is being imported.

3 Checking for NaN: Inorder to find out if there is any missing values present in the dataset we are supposed to perform this process.

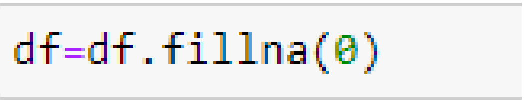
4 Manipulating NaN values

It is essential to remove the NaN values. This can be done by :

- Removing the entire column containing many NaN values.
- Forward fillna method.
- Mean method.



```
In [6]:
df.isnull().sum()
Out[6]:
id                          0
age                         9
bloodpressure              12
specificgravity            47
albumin                    46
sugar                      49
redbloodcells             152
puscell                    65
puscellclumps               4
bacteria                    4
bloodglucoserandom         44
bloodurea                  19
serumcreatinine            17
sodium                     87
potassium                  88
haemoglobin                52
packedcellvolume           71
whitebloodcellcount       105
redbloodcellcount         130
hypertension                2
diabetesmellitus            2
coronaryarterydisease       2
appetite                    1
pedaledema                  1
anemia                      1
classification              0
dtype: int64
```



```
df=df.fillna(0)
```

Fig 2: shows the check for NaN values.

Fig 3: filling NaN values.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86138

5 Plotting a Heatmap: Correlation between the fields of the recorded data is analyzed by plotting a heatmap. The values may be negative or positive and the magnitude plays a key role in designing various predictive models in AI.

6 Splitting: Splitting the data plays a vital role in prediction process. The data is being split as train data and test data. The train data is used to train the model and the test data is used for predicting the output.

7 Algorithm: Logistic Regression which is a statistical model uses a logistic function to model a binary dependent value. It is used to predict the categorical dependent variable using independent variables.

Sklearn.linear_model: sklearn stands for scikit learn. Sklearn is the library/module and logistic regression is the class inside this model.

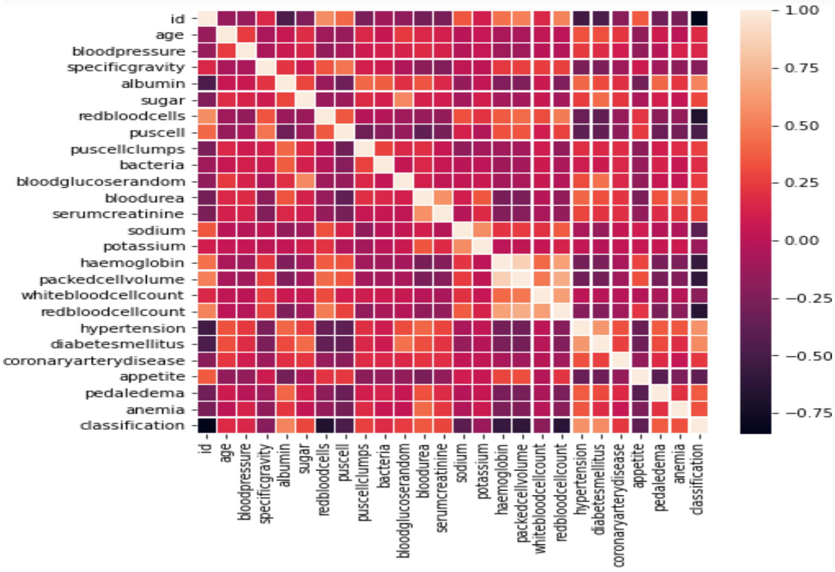


Fig 4: Heatmap and correlation model.

Algorithm used: Logistic Regression

- Logistic regression predicts the output of a categorical dependent variable.
- It can be either Yes or No , 0 or 1, true or false etc. but instead of giving the exact values as 0 or 1, it gives the probabilistic values which lie between 0 and 1.

Steps in LR:

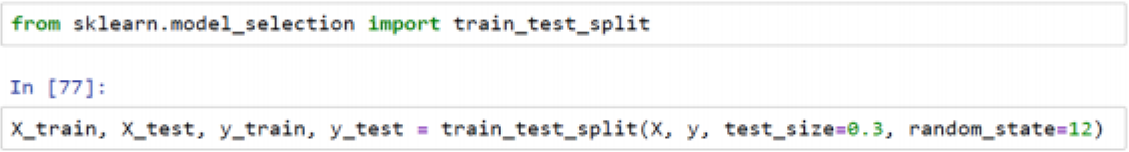
1. Data Pre-processing step.

2. Fitting Logistic Regression to the Training set.

3. Test accuracy of the result(Creation of Confusion matrix).

4. Predicting the test result.

5. Visualizing the test set result.



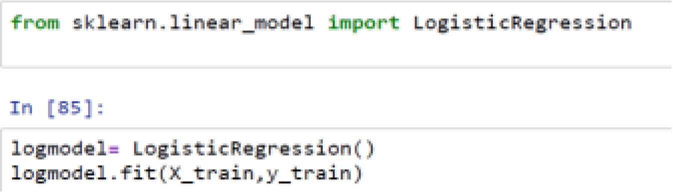
```
from sklearn.model_selection import train_test_split

In [77]:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=12)
```

Figure 5: Splitting data into train data and test data.

DOI: 10.17148/IARJSET.2021.86138



```
from sklearn.linear_model import LogisticRegression

In [85]:

logmodel= LogisticRegression()
logmodel.fit(X_train,y_train)
```

Figure 6: shows logistic regression on given data set.

IV.DATA VISULATION

Visualisation is a piece of information and Machine Learning. When there is a tremendous informational index, manual investigation turns out to be practically unimaginable. visualisation assumes a crucial part in examination in such circumstance. It includes utilization of different plots – reference diagram, pie outlines, box plots, line charts and some more. Figure 7 shows the count of ckd and non ckd patients and figure 8 shows the kidney diseases frequency among ages.

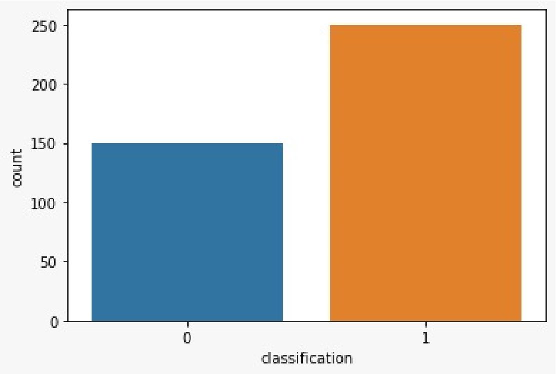


Figure 7: countplot for number of CKD & non-CKD patient.

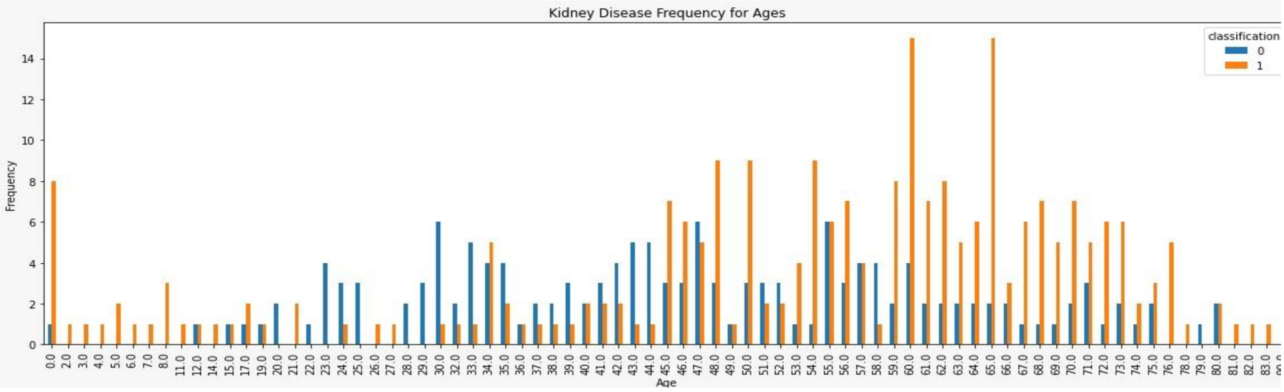
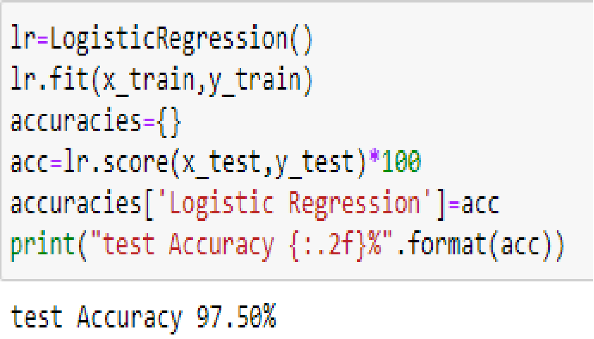


Figure 8: kidney disease frequency for ages.

V. RESULTS

In the wake of investigating the heatmap and sorting out the connection between's various segments/physiological boundaries, Calculated relapse should be done to make a forecast model. Figure 9 shows the after effects of logistic regression model. Figure shows the Accuracy score of the planned model. From this information, exactness, f1 score and unwavering quality can be determined.

DOI: 10.17148/IARJSET.2021.86138



```
lr=LogisticRegression()
lr.fit(x_train,y_train)
accuracies={}
acc=lr.score(x_test,y_test)*100
accuracies['Logistic Regression']=acc
print("test Accuracy {:.2f}%".format(acc))
```

```
test Accuracy 97.50%
```

Figure 9: shows the Accuracy score of the designed model.

VI. CONCLUSION

Chronic Kidney Disease is fatal, but can be treated and cured when identified at an early stage. Few samples were considered to design a predictive model using Logistic Regression. The data set was taken from a trusted source, pre-processed, statistically analysed and graphs plotted. A heatmap was plotted to identify the correlation between different fields of interest. The data being cleansed (removing NaN values) was subjected to division as train and test data. 70% of the data was fed for training and the remaining considered for test. We have calculated the accuracy of the model and were happy to conclude with 97.50% accuracy. Any new samples taken can be predicted with this model with high reliability, accuracy and precision..

REFERENCES

[1] Interactions b/w kidney disease & diabetes- dangerous liaisons- Roberto Pecoits-Filho, Hugo Abensur, Carolina C.R. Betônico, Alisson Diego Machado, Erika B. Parente, Márcia Queiroz, João Eduardo Nunes Salles, Silvia Titan and Sergio Vencio- 2016- article 50.

[2]. The Python Standard Library — Python 3.7.1rc2 documentation https://docs.python.org/3/library/

[3]. Data Warehousing Architecture & Pre-Processing- Vishesh S, Manu Srinath, Akshatha C Kumar, Nandan A.S.- IJARCCE, vol6, iss5, May 2017.

[4]. Data Mining and Analytics: A Proactive Model - http://www.ijarcce.com/upload/2017/february-17/IJARCCE% 20117.pdf topics