International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206215

Chronic Kidney Disease Prediction System

Ammavajjula Sai Tejaswi, Animilla Swapna Deepika, Yaragani Sowmya

Computer Science and Engineering, GVP College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

Chronic Kidney Disease (CKD) is a very dangerous health problem that has been spreading due to globally due to alterations in lifestyle such as food habits, changes in the atmosphere, etc. So, it is essential to decide any remedies to avoid and predict the disease in an early stage. This paper focuses on predictive analytics architecture to analyze the CKD dataset using feature engineering and classification algorithm. The proposed model incorporates techniques to validate the feasibility of data points used for analysis. The main focus of research work is to analyze the dataset of chronic kidney failure and perform the classification of CKD and Non-CKD cases.

Keywords : Feature Engineering, Classification Algorithm.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a condition characterized by a gradual loss of kidney function over time. CKD includes conditions that damage the kidneys and decrease their ability to keep healthy by performing its responsibilities. If kidney disease gets worse, wastes can be built to high levels in the blood and make the person feel sick. That person may be developed with complications like high blood pressure, anemia, weak bones, poor nutritional health, and nerve problems that may happen slowly over a long period of time. CKD may be caused by diabetes, high blood pressure, and other disorders. Early detection and treatment can often keep CKD from getting worse. The two main causes of CKD are diabetes and high blood pressure, which are responsible for up to two-thirds of cases.

According to the Global Data (2019) a leading data analytics company has specified that CKD had low diagnosis rate till 2026 unless there are effective ways incorporated to diagnose CKD at early stages. Data Mining, a Knowledge Discovery process (KDD) is a very effective approach used to analyse the raw data and extract information from complex data. The application developed uses the CKD dataset of the University of California-Irvine (UCI) repository. This system uses Logistic Regression and SVM algorithms to predict whether a person is having CKD or not.

DATASET SUMMARY

Attribute Used	Type of Attribute	Attribute Description
age	Numerical	Age
bp	Numerical	Blood Pressure
sg	Nominal	Specific Gravity
al	Nominal	Albumin
su	Nominal	Sugar
rbc	Nominal	Red Blood Cell
pc	Nominal	Pus Cell
pcc	Nominal	Pus Cell Clumps
ba	Nominal	Bacteria
bgr	Numerical	Blood Glucose Random
bu	Numerical	Blood Urea
sc	Numerical	Serum Creatinine
sod	Numerical	Sodium
pot	Numerical	Potassium
hemo	Numerical	Hemoglobin
pcv	Numerical	Packed Cell Volume
wc	Numerical	White Blood Cell Count
rc	Numerical	Red Blood Cell Count
htn	Nominal	Hypertension
dm	Nominal	Diabetes Mellitus
cad	Nominal	Coronary Artery Disease
appet	Nominal	Appetite
pe	Nominal	Pedal Edema
ane	Nominal	Anemia
Classification	Flag	Class

The above are the 25 (24 +1 class) attributes. In which 11 are numeric and 14 are attributes.

STATISTICAL ANALYSIS

	age	bp	sg	al	su	bgr
count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000
mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517
std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714
min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000
25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000
50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000
75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000
max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000

Fig 1: Statistical summary of few columns of dataset

II. IMPLEMENTATION

A. System Architecture

The proposed system for CKD analysis and prediction is shown in Fig 2. The raw data is first analyzed and then trained and tested using machine learning models. The models were then evaluated based on performance metrics- precision, accuracy, recall, error-rate, based on the confusion matrix.

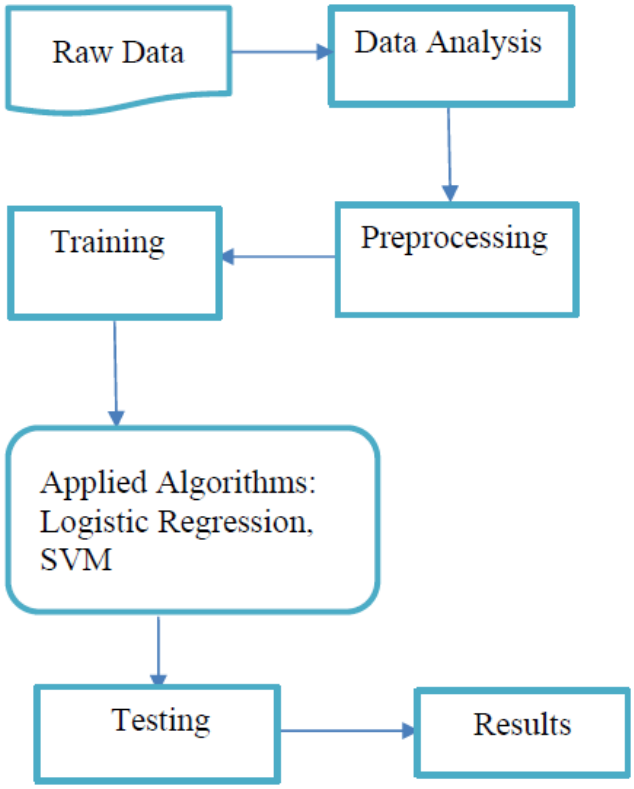

Fig 2 : A proposed system for CKD analysis and prediction.

B. Preprocessing

The original dataset is irregular and has many missing values. The profiling report of the dataset shown in Fig 3 is generated using the module pandas_profiling

```
age has 9 (2.2%) missing values
al has 199 (49.8%) zeros
al has 46 (11.5%) missing values
ba has 4 (1.0%) missing values
bgr has 44 (11.0%) missing values
bp has 12 (3.0%) missing values
bu has 19 (4.7%) missing values
hemo has 52 (13.0%) missing values
pc has 65 (16.2%) missing values
pcc has 4 (1.0%) missing values
pcv has 70 (17.5%) missing values
pot has 88 (22.0%) missing values
rbc has 152 (38.0%) missing values
rc has 130 (32.5%) missing values
sc has 17 (4.2%) missing values
sg has 47 (11.8%) missing values
sod has 87 (21.8%) missing values
su has 290 (72.5%) zeros
su has 49 (12.3%) missing values
wc has a high cardinality: 93 distinct values
wc has 105 (26.2%) missing values
```

Fig 3: Profiling report of the dataset

Based on the report shown in Fig 3, undergo the following steps to preprocess the data

1. Initially, the variables with high missing values and zeros are removed.
2. Then the outliers detected through boxplot and removed.
3. Missing values are imputed using the class-wise mean for numerical data and class-wise mode for categorical data.
4. After imputing the missing data, label encoding is done and data is split into test and train.
5. Then the training data is normalized using Pearson Normalisation and the model is fitted with this data. To predict the test data class label, it is normalized using the train data's mean and standard deviation and its class label is predicted using model.predict() function.

C. Models

Models used for the prediction of CKD are Logistic Regression and Support Vector Machine.

Logistic Regression

Logistic Regression is a simple and powerful algorithm for binary classification. It uses the logit function for classification. The logit or sigmoid function maps the input to either 0|1 or yes|no, unlike linear regression. Before fitting the model, the bias value is to be added to the input data. This model uses the Cross-Entropy loss function, which is a measure of how good the prediction model is and also uses the Gradient Descent optimization technique to minimize the loss function by iteratively moving in the direction of steepest descent as defined by negative of descent. In general, it is used to update the parameters of the model.

Support Vector Machine :

A Support Vector Machine (SVM) is a supervised machine learning algorithm that is employed for classification problems. SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of the dataset. The objective of the SVM algorithm is to find a hyperplane in N-dimensional space (N- the number of features) that distinctly classifies the data points. The kernel function used here is sigmoid to classify.

III. RESULTS AND DISCUSSION

Performance Metrics-Accuracy, Precision, Recall are used to evaluate models. These metrics can be calculated using the confusion matrix shown in Fig 5.

Outcome of the Diagnostic Test		Predicted	
		Positive (1)	Negative (0)
Observed	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 4: Confusion Matrix

In the above Fig 4,

TP- True Positives-Number of correct classifications predicted as positive.

FP- False Positives-Number of correct classifications predicted as negative.

FN- False Negatives-Number of examples that are incorrectly predicted as positive which are actually negative.

TN- True Negatives-Number of examples that are incorrectly predicted as negatives which are actually positive.

Accuracy measures the ability of a model to correctly predict the class label of new or unseen data.

Accuracy measures the ability of a model to correctly predict the class label of new or unseen data.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision is a measure that tells what proportion of people are diagnosed as having CKD, actually having CKD.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is a measure that tells what proportion of patients actually having CKD, was diagnosed by algorithm as having CKD.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The report of models is shown in following Fig 5, Fig 6.

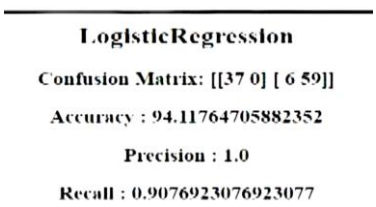

Fig 5: Logistic Regression Performance Report

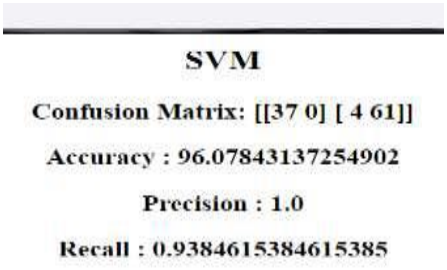

Fig 6: SVM Performance Report

After preprocessing of dataset, the influencing features of the dataset are age, blood pressure, specific gravity, pus cells, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, hemoglobin, hypertension, diabetes mellitus, coronary artery disease, Volume 1 | Issue 1 | July-August 2016 | www.ijsrcseit.com 4 appetite, pedal edema, anemia. The user can give these values to the system as shown in Fig 7 and know in the early stage whether he is having CKD or NOT and take the precautions to avoid or treat it.

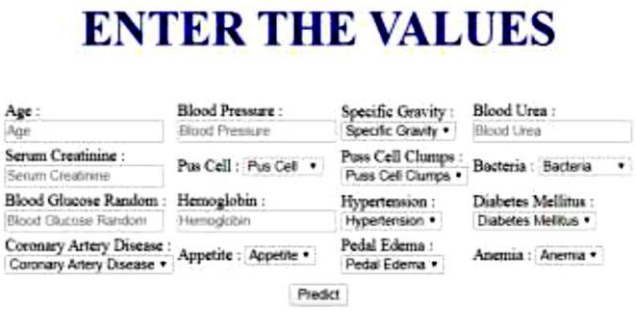

Fig 7 : Predict Webpage developed using flask.

When the user clicks on predict, internally the predicted probability of both logistic regression and SVM are multiplied with weights-0.9and 0.8 based on the highest accuracy and are summed and checked against the given threshold to know whether the patient is healthy or he has chances of getting CKD.

IV.CONCLUSION

Anticipating diseases remains a major challenge in the medical field. In this paper, the application of machine

learning techniques to predict CKD is discussed. Since the data includes missing values data imputation methods are utilized and preprocessed data is fitted to the model and tested by a user by giving the values to know whether he/she is having CKD or not.

V. REFERENCES

[1]. Bala, S., and Kumar, K., 2014, "A Literature Review on Kidney Disease Prediction Using Data Mining Classification Technique," International Journal of Computer Science & Mobile Computing, 3(7), 960-967.

[2]. Tangari, N., Kitsios, G.D., et al., 2013, "Risk Prediction Models for Patients with Chronic Kidney Disease" Annals of Internal Medicine, 158(8), 596-603.

[3]. Andrew Kusiak, Bradley Dixonb, Shital Shaha, (2005) "Predicting survival time for kidney dialysis patients: a data mining approach", Elsevier Publication, Computers in Biology and Medicine, Vol.35 pp 311-327.

[4]. M Hall, E Frank, G Holmes, B Pfahringer, (2009), 'The WEKA data mining software: An update', volume 11, issue 1, pp 10-18.

[5]. Tangri, N., Stevens, L., et al., 2011, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure," Journal of American Medical Association,305 (15), 1553-1559.

[6]. Vijayarani, S., and Dhayanand, S., 2015, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics and Information, 4(4), 13-25.

[7]. Ziyad, A., 2013, "Prediction of Renal End Points in Chronic Kidney Disease," Kidney International, 83(2), 189-191.

[8]. R. Weil, (2014)," Big Data In Health: A New Era For Research And Patient Care Alan R. Weil", Health Affair, Vol. 33, N˚ 7, pp 1110.

Cite this article as :

Ammavajjula Sai Tejaswi, Animilla Swapna Deepika, Yaragani Sowmya, "Chronic Kidney Disease Prediction System", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 2, pp. 43-47, March-April 2020.
Available at doi : https://doi.org/10.32628/CSEIT206215
Journal URL : http://ijsrcseit.com/CSEIT206215