Detection and Evaluation of Chronic Kidney Disease Using Different Regression and Classification Algorithms in Machine Learning

Chapter · January 2021

DOI: 10.1007/978-981-15-8752-8_10

Avinash Kumar
KIIT University
7 PUBLICATIONS **70** CITATIONS

Sobhangi Sarkar
KIIT University
7 PUBLICATIONS **70** CITATIONS

Detection and Evaluation of Chronic Kidney Disease Using Different Regression and Classification Algorithms in Machine Learning

Anusmita Sarkar, Avinash Kumar, Sobhangi Sarkar, and Chittaranjan Pradhan

Abstract Nowadays, many people are suffering from chronic kidney disease worldwide. Factors responsible for such conditions are food, living standards, and the environment. Detection and identification of chronic kidney disease are costly, time-consuming, and often risky. Therefore, the early detection of such disease is very important. In this research study, we have tried to reduce the clinical effort by automating the process of detection. We have classified whether the person is suffering from chronic kidney disease or not. We have used different classification algorithms and regression algorithms like KNN, SVM, Naive Bayes, and logistic regression. We have got some good results in all the algorithms but KNN performed very well.

Keywords Kidney disease · KNN · Logistic regression · Naive Bayes · Support vector machine

1 Introduction

Kidneys are essential for the normal functioning of the human body. Some of the crucial functions performed by kidneys include filtration of waste materials, excessive water content from the blood. They also help to maintain acid base balance and regulate electrolyte concentrations in the body. The other tasks of the kidneys are to regulate blood pressure by creating hormones, creation of red blood cells, and

A. Sarkar (✉) · A. Kumar · S. Sarkar · C. Pradhan
Kalinga Institute of Industrial Technology, Bhubaneswar, India
e-mail: anusmitasarkar2000@gmail.com

A. Kumar
e-mail: avinash1605@gmail.com

S. Sarkar
e-mail: shobhangisarkar@gmail.com

C. Pradhan
e-mail: chitaprakash@gmail.com

P. K. Mallick et al. (eds.), *Advances in Electronics, Communication and Computing*,
Lecture Notes in Electrical Engineering 709,
https://doi.org/10.1007/978-981-15-8752-8_10

promote bone health. Millions of people belonging to different age groups suffer from kidney disease all over the world. It happens when they become damaged and cannot filter blood the way they should do. Kidneys may get damaged due to high blood pressure, diabetes, etc. It might also cause other medical problems such as malnutrition and weak bones. Chronic kidney disease is the most frequent form of kidney disease. It is commonly induced by high blood pressure. Kidney stones are another typical kidney problem. Kidney stones can make urination extremely painful. There are also other types of kidney diseases like glomerulonephritis and polycystic kidney disease.

Doctors run some investigations to see if the kidneys are working properly. These investigations include glomerular filtration rate (GFR) to determine the stage of the disease, kidney biopsy to determine what type of kidney disease, and to what extent damage has occurred; urine test, blood creatinine to determine whether the kidney is working properly and ultrasound and CT scan to produce images of kidney and urinary tract. These tests are used by doctors to deliver some judgment about the condition of the kidney or kidney disease. To help improve their judgment, techniques namely machine learning, deep learning and artificial intelligence are used [1].

2 Existing Work

Anusorn Charleonnan et al. used machine learning models like K-nearest neighbors (KNN), logistic regression, support vector machine (SVM), and decision tree classifier to predict chronic kidney disease and compared their accuracy. The performance measures used were accuracy, sensitivity, and specificity. SVM had the highest sensitivity was 0.99, and KNN had the highest specificity. SVM was selected as the appropriate model for predicting chronic kidney disease [2].

Engin Avci et al. used WEKA software with Naive Bayes, J48, and some other algorithms to make predictions about chronic kidney disease. The performance measure used were accuracy, precision, sensitivity, and F-measure parameters found that J48 classifier gave 99% accuracy and J48 model was as the most appropriate model for predicting kidney disease or ckd [3].

W. H. S. D. Gunarathne et al. used different types of algorithms such as multi-class decision jungle and multi-class neural network, etc., for the detection of kidney disease [4]. The research approach that had been used is cross industry standard process for data mining (CRISP-DM) [5]. They have used Microsoft Azure machine learning studio to develop this model. The highest accuracy of 99.1% was achieved by multi-class decision forest [4].

Y. Amirgaliyev et al. used SVM classifier with linear kernel for predicting chronic kidney disease. The performance measures used were sensitivity, specificity, and accuracy metrics. SVM classifier with linear kernel got a sensitivity of 93.1% [6].

Mubarik Ahmad used SVM. SVM has been used for detection of kidney disease. The random forest package had been used to calculate the error rate which came out to be 1.66%, and hence, an accuracy of 98.34% was obtained [7].

3 Methodology

We have downloaded the dataset from online source which was publicly available on UCI repository. The data was taken within a period of 2 months in India with 25 features consisting 400 rows. The features that was used in building our model is shown in Table 1.

The different algorithms used here are discussed below.

Table 1 Dataset attributes

Attribute	Short name used
Age	Age
Blood pressure	Bp
Specific gravity	Sg
Albumin	Al
Sugar	Su
Red blood cells	Rbc
Pus cell	Pc
Pus cell clumps	Pcc
Bacteria	Ba
Blood glucose random	Bgr
Blood urea	Bu
Serum creatinine	Sc
Sodium	Sod
Potassium	Pot
Hemoglobin	Hemo
Packed cell volume	Pcv
White blood cell count	Wc
Red blood cell count	Rc
Hypertension	Htn
Diabetes mellitus	Dm
Coronary artery disease	Cad
Appetite	Appet
Pedal edema	Pe
Anemia	Ane
Classification	Class

3.1 K-Nearest Neighbor (KNN)

It is a ML algorithm which can be utilized to work out classification and regression domains. It works by recording existing occurrence and categorizing new occurrences using method which is known as similarity measure also known as distance function. By obtaining the majority of votes of its neighbors, a case is categorized to the most common class. The case belongs to the class of its nearest neighbor if $K = 1$ [8].

3.2 Naive Bayes

For binary-class and multi-class classification problems, Naive Bayes algorithm is used. It is related to Bayes theorem. This classifier computes the probabilities for every factor. It selects the result with the highest probability. This classifier assumes the features are independent. Thus, Naïve Bayes theorem calculates happening event's probability against one already happened event.

3.3 Support Vector Machine (SVM)

For classification and regression problems, SVM is another common supervised machine learning algorithm. In a n-dimensional space, a point is plotted for each data item. Here, "n" represents the number of features. Then classification is accomplished by determining the hyperplane which best distinguishes the two classes. The coordinates of each observation are support vectors.

3.4 Logistic Regression

For classification task, logistic regression is another machine learning algorithm. The cost function of logistic regression can be described as the sigmoid (or logistic) function. The hypothesis of logistic regression usually confines the cost function between 0 and 1. When the inputs are passed through a prediction function, this classifier gives a set of classes and finds the probability.

4 Proposed Work

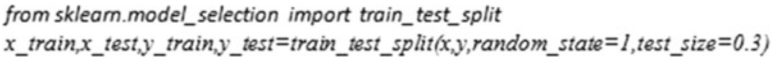
We have used a dataset for the development of the model in our research is available on an open source website. The development of model was initiated with pre-processing of it. Preprocessing is the most important step in development of any model. It helps the model to be more accurate and precise.

First of all, we started with the data cleaning and pre-processing steps. Our dataset had lot of textual values which we converted into numerical values, i.e., 0 or 1. We have also changed the target parameter values to 1 and 0 to be able to use the classification algorithms.

Moreover, we have also transformed some of the attributes into integer with the help of replace method of pandas dataframe. The value abnormal and normal were changed to 1 and 0, respectively, for attributes such as for rbc, pc. The value present and not present were also changed to 1 and 0, respectively, for attributes such as ba, pcc.

The target column was also transformed. If the value is "chronic kidney disease" (ckd) then it is replaced with 1, if not then its equal to 0. Our dataset have some NaN values. We have imputed all those rows which were Null. After the data preprocessing, we have visualized the correlation of parameters with the help of heatmap as shown in Fig. 1.

We have split our dataset into 70:30 ratio, where 30% is test size and 70% is training size and have applied different machine learning algorithms to analyze the accuracy.

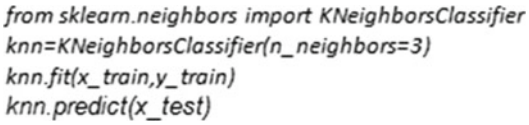


```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=1,test_size=0.3)
```

The following sections give the description of regression and classification algorithms used.

4.1 K-Nearest Neighbor (KNN)

We have applied KNN classification algorithm by using following code and achieved an accuracy of 97.5%.



```
from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier(n_neighbors=3)
knn.fit(x_train,y_train)
knn.predict(x_test)
```

The confusion matrix for the KNN algorithm was visualized on heatmap is shown in Fig. 2.

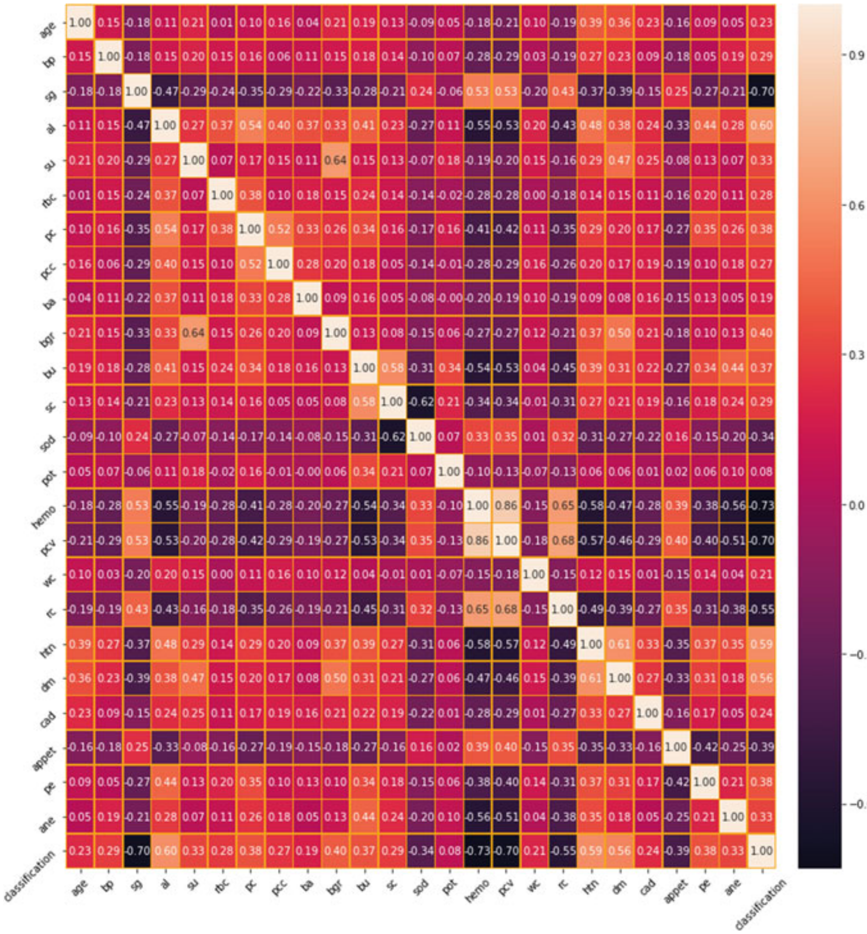


Fig. 1 Dataset without NaN Values

4.2 Naïve Bayes

We have applied Navie-Bayes classification algorithm by using following code and achieved an accuracy of 94.16%.

```
from sklearn.naive_bayes import GaussianNB
nb=GaussianNB()
nb.fit(x_train,y_train)
```

The confusion matrix for the above algorithm was visualized on heatmap is presented in Fig. 3.

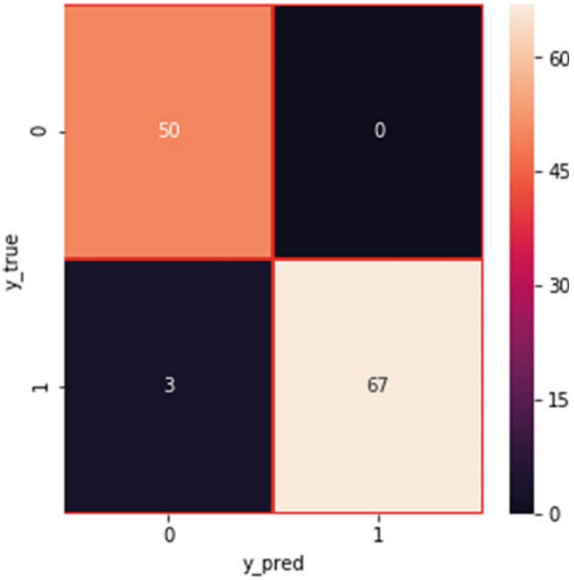


Fig. 2 KNN confusion matrix

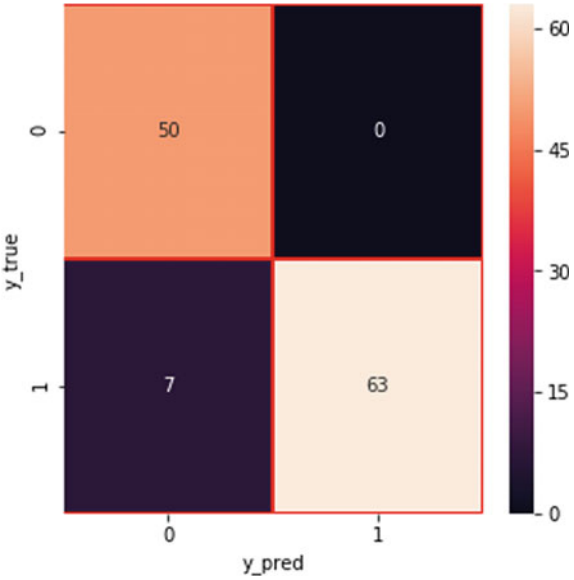


Fig. 3 Naïve Bayes confusion matrix

4.3 Support Vector Machine (SVM)

We have applied support vector machine classification algorithm by using following code and achieved an accuracy of 92.5%.

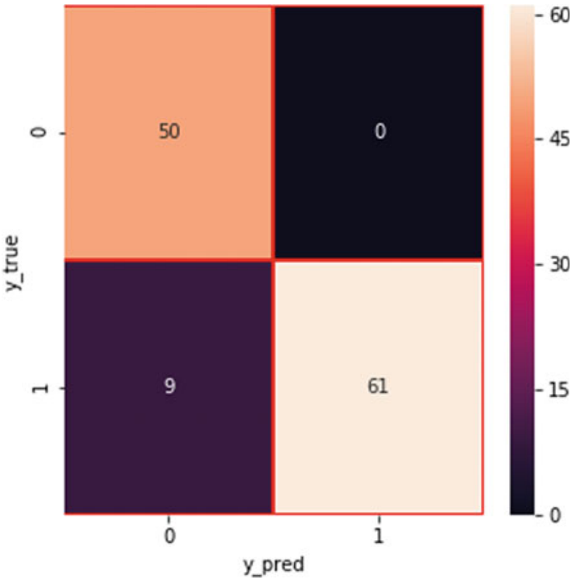
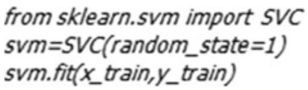


Fig. 4 Support vector machine confusion matrix

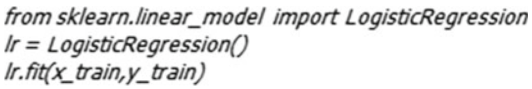


```
from sklearn.svm import SVC
svm=SVC(random_state=1)
svm.fit(x_train,y_train)
```

The confusion matrix for the support vector machine algorithm was visualized on heatmap is shown in Fig. 4.

4.4 Logistic Regression

We have applied logistic regression algorithm by using following code and achieved an accuracy of 95.00%.



```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(x_train,y_train)
```

The confusion matrix for the logistic regression algorithm was visualized on heatmap is shown in Fig. 5.

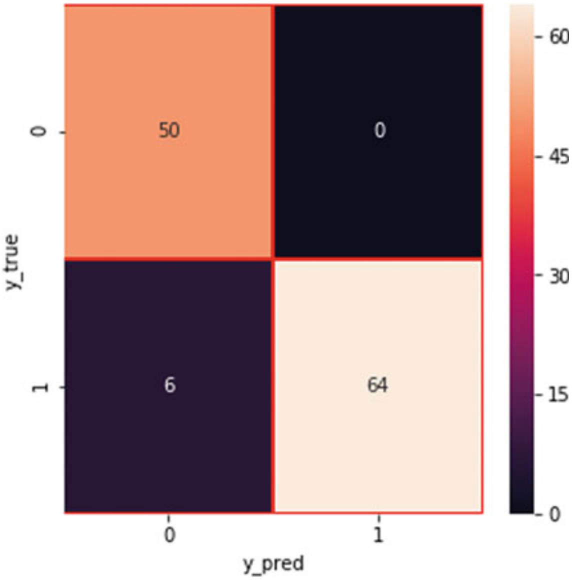


Fig. 5 Logistic regression confusion matrix

5 Result Analysis

The pre-processing of the dataset was the major part which lead to the good accuracy of our model. We have chosen the suitable algorithms for our model and the dataset and applied it.

After applying all the algorithms and analyzing, we got to know that all the algorithms gave good results but KNN performed exceptionally well. Fig. 6 shows the accuracy comparison of all the algorithms.

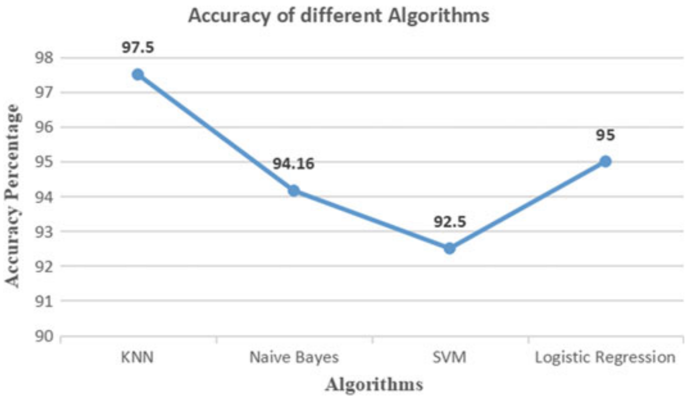


Fig. 6 Accuracy comparison of different algorithms

We have used the dataset from online source which was publicly available on UCI repository. We have used four different algorithms and got good accuracy. KNN algorithm gave us good result which is around 97.5% precision, Naive Bayes showed the precision of 94.16%, SVM also showed the precision of 92.5%, and finally, logistic regression showed the precision of 95%.

6 Conclusion and Future Work

In our research, we tried to obtain a good accuracy by applying different algorithms. From the result analysis, we found KNN algorithm produces the best accuracy of 97.5% followed by the accuracy of 95% by logistic regression algorithm. In future, we can work on huge real life dataset and can develop a healthcare system prototype for chronic kidney disease patients.

References

1. Kidney health and kidney disease basics. https://www.healthline.com/health/kidney-disease.
2. Niyomwong, T., et al. (2016). Predictive analytics for chronic kidney disease using machine learning techniques. In *International Conference on Management and Innovation Technology* (pp. 80–83). Thailand: IEEE.
3. Ozmen, O., et al. (2018). Performance comparison of some classifiers on chronic kidney disease data. In *International Symposium on Digital Forensic and Security* (pp. 1–4). Turkey: IEEE.
4. Gunarathne, W. H. S. D., et al. (2017). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In *International Conference on Bioinformatics and Bioengineering* (pp. 291–296). USA: IEEE.
5. Taylan, D., et al. (2013). Data mining process for modeling hydrological time series. *Hydrology Research, 44*(1), 78–88.
6. Shamiluul, S., et al. (2018). Analysis of chronic kidney disease dataset by applying machine learning methods. In *International Conference on Application of Information and Communication Technologies* (pp. 1–4). Kazakhstan: IEEE.
7. Amalia, P., et al. (2017). Diagnostic decision support system of chronic kidney disease using support vector machine. In *International Conference on Informatics and Computing* (pp. 1–4). Indonesia: IEEE.
8. Penny, W. D., et al. (1997). Neural network modeling of the level of observation decision in an acute psychiatric ward. *Computers and Biomedical Research, 30*(1), 1–17.