Predictive Analysis for Chronic kidney Disease Using Machine Learning Techniques

Bhumika M, Divya Shivakumar, Meghana B, Meghana M,
Student, Department of Computer Science & Engineering,
GSSSIETW, Mysuru

Rajashekar MB
Assistant Professor, Department of Computer Science &
Engineering, GSSSIETW, Mysuru

Abstract—In this study, dataset named "Chronic Kidney Disease" obtained from UCI database is used. The dataset consists of 400- individuals information and contains 25 features. With WEKA software, this dataset is classified according to whether it is chronic kidney disease using Naïve Bayes (NB), and C4.5 classifiers used in data mining. Accuracy, precision, sensitivity, and F- measure values are used for performance comparisons of the performed classifications. According to the obtained results, more successful results were obtained in Naïve bayes algorithm with 99%accuracy.

Keywords— ***chronic kidney disease; weka; performance comparison; naive bayes, c4.5;***

I. INTRODUCTION

The amount of data emerging along with the development of information technologies also shows a rapid increase. It is estimated that the data stored in the world database is being doubled every 20 months [1]. With the amount of data that increases each passing day, the processing of this data has become a challenge. Various data mining algorithms have been developed to solve this problem. In the literature, there are many algorithms used in data mining and different field of studies for comparing these algorithms. The health sector is also one of these areas. It is used to find more accurate results in diagnostics and treatments in this sector and to prevent human errors[2].

In our study, the dataset named "Chronic Kidney Disease" obtained from the UCI database [3] of chronic kidney disease, which affects human life negatively and is one of the most common public health problems in the world, has been used[4]. According to Tanrıverdi and his colleagues, chronic kidney disease [5] has been described as "the end of the reduction of glomerular filtration value, adjusting the fluid-solute balance of the kidney, and chronic and progressive impairment of metabolic endocrine functions". Chronic kidney disease occurs when there is a major reduction in kidney function[6]. According to the data of the Turkish Nephrology Association, it is 390 per million population prevalence of chronic kidney disease in Turkey. Compared to other countries, this rate is quite low. The most important cause for this case is the experienced difficulties

in data collection [5]. The purpose of the study is to compare the performances of some classifiers on the extracted dataset in order to develop more effective software for this area.

B. Used Classifiers

The dataset used in this study is deployed for performance comparison using Naive Bayes (NB) and C4.5 classifiers. The The rest of this work is organized as follows. The material and method are presented in section II, the dataset used and the performance criteria used in the classification algorithms are discussed in section III, the experimental results in are provided in section IV, and finally the conclusions and the recommendations are offered in section V.

II. MATERIAL AND METHOD

A. Data Extraction

In this study, the dataset named "Chronic Kidney Disease" extracted from UCI database was used. The dataset consists of a total of 25 attributes that 24 attributes + class (11 numeric, 14 nominal). In addition, there are a total of 400 records for individuals, whose ages range from2-90.

Table I contains the attributes descriptions and the value ranges of the dataset used.

TABLE I. USED DATA SET

ttribute Name	***Value Range***	***Description***
age	2, .., 90	age
bp	50, …, 180	blood pressure
sg	1.005,1.010,1.015,1.020,1.025	specific gravity
al	0,1,2,3,4,5	albumin
su	0,1,2,3,4,5	sugar
rbc	2.1, …, 8	red blood cells
pc	normal,abnormal	pus cell
pcc	present,notpresent	pus cell clumps
ba	present,notpresent	bacteria
bgr	22, …, 490	blood glucose random
bu	1.5, …, 391	blood urea
sc	0.4, …, 76	serum creatinine
sod	4.5, …, 163	sodium
pot	2.5, …, 47	potassium
hemo	3.1, …, 17.8	hemoglobin
pcv	9, …, 54	packed cell volume
wc	2200,…, 26400	white blood cell count

rc	2.1,…, 8	red blood cell count
htn	yes, no	hypertension
dm	yes, no	diabetes mellitus
cad	yes, no	coronary artery disease

following information is given about these classifiers.

Naive Bayes Classifier: This classifier is based on the Bayes theorem, which is one of the simple and widely used methods, and it can handle any number of properties or classes. Although the model is simple, it performs well with some problems [7]. This classifier assumes that the data is already classified. When a new data arrives, it calculates the probability that this data belongs to one of the classes. When calculating these probability values, it is assumed that each feature is independent of the other, and each feature has the same degree of importance. When it is desired to find out which class the externally entered data belongs to, the probability of belonging to that class for each class of the data is calculated by using the formula in equation (1). The class having the highest probability among these calculated values is regarded as the class to which that data belongs[2]. **III.**

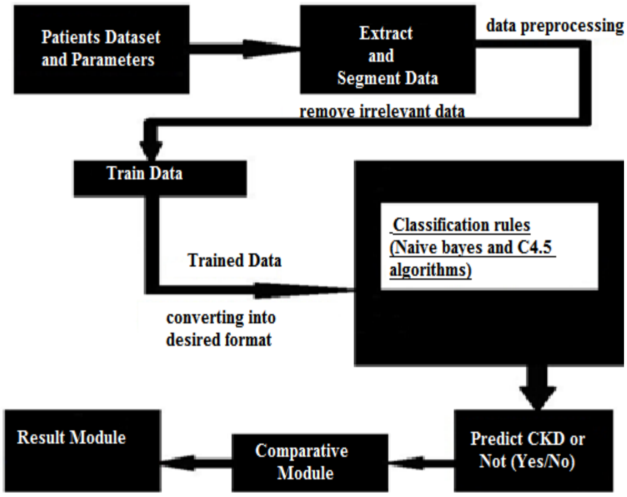
Patients Dataset and Parameters
Extract and Segment Data
data preprocessing
remove irrelevant data
Train Data
Trained Data
converting into desired format
Classification rules (Naive bayes and C4.5 algorithms)
Predict CKD or Not (Yes/No)
Comparative Module
Result Module

Fig 1. System Architecture

$$(S|X) = \frac{P(X|S_i)*P(S_i)}{P(X)} \qquad (1)$$

Where,

P(Si|X): The probability of occurrence of the *Si* event when the *X* event occurs,

P(X|Si): The probability of occurrence of the *X* event when the Si event occurs,and

P(Si), P(X): The prior probability of *Si* and *X* events.

The optimal hyperplane, given in Figure 1, can be defined as a linear decision function with the maximum distance between the vectors of the two classes [8].

C4.5 Classifier: Step 1: Scan the dataset (storage servers) Step 2: for each attribute a, calculate the gain [number of occurrences]

Step 3: Let a_best be the attribute of highest gain [highest count]

Step 4: Create a decision node based on a_best – retrieval of nodes [patient] where the attribute values matches with a_best. Step 5: recur on the sub-lists [list of patient] and calculate the count of outcomes [Stages] – termed as sub nodes. Based on the highest count we classify the new node.

SampleExample

Attributes(Features) – F1,F2,F3 [m=3]

Subject (outcome) – CKD, NOT CKD [p=1/2=0.5]

TABLE II. TRAINING DATASET

Patient Name	S1(X,Y,Z)	S2(A,B,C)	S3(P,Q,R)	Disease (subject)
Anil	X	A	P	CKD
Ajay	X	B	Q	CKD
Arun	Y	B	P	NOT CKD
Kumar	Z	A	R	CKD
Naveen	Z	C	R	NOT CKD

III. PERFORMANCE MEASURES USED FOR THE CLASSIFICATION ALGORITHMS

There are many classifiers used in data mining. Comparing and analyzing these classifiers are a rather complex process. Because there are various evaluation dimensions and these dimensions must be taken into consideration [14].

Two different classification algorithms (NB and C4.5) are used in this study. In data mining applications, confusion matrix is frequently used to measure the performance of algorithms classification.

A. Accuracy

The correct classification is the total classification ratio [14]

B. Precision

One of the classifiers role is the ability to determine the positive features of the whole classification. That is, as seen in Equation (9), it is obtained by dividing the positive values classified correctly in all positive values.

Precision = TP/TP+FP

C. *Sensitivity*

To describe the class labels of the classifiers, it refers to the average per class activity. In other words, positive value of correctly classified to the sum of correctly classified positive and false classified negative values [14].

$$\text{Sensitivity}=\frac{TP}{TP+FN} \qquad (10)$$

D. *F-Measure*

Precision and sensitivity alone may not be sufficient in performance comparisons. In this case, it is necessary to look at the F-measure. The F-measure is obtained by taking the weighted average of the precision and sensitivity values [14].

IV. EXPERIMENTAL RESULTS

The following tables show the confusion matrixes of the classifiers used. The expression "*ckd*" in the tables means "chronic kidney disease" and "*notckd*" means "non-chronic kidney disease".

TABLE III. NAIVE BAYES CONFUSION MATRIX

		Prediction	
		ckd	***notckd***
Actual	***ckd***	230	20
	notckd	0	150

TABLE IV.C4.5 CONFUSION MATRIX

		Prediction	
		ckd	***notckd***
Actual	***ckd***	241	9
	notckd	0	150

TABLE V.COMPARISON OF PERFORMANCE MEASURES OF THE USED ALGORITHMS

Used Algorithms	**Performance Measures**			
	Accuracy (%)	***Precision***	***Sensitivity***	***F-Measure***
C4.5	97.75	1.0	0.96	0.97
NB	99.00	0.98	0.99	0.98

As it can be seen from Table V, the accuracy rate of the performance measure according to the classifiers was obtained from the NB classifier with the highest 99%. C4.5 with the 0.97. Sensitivity was obtained from NB with a 0.99, and C4.5 with a 0.96. Finally, the F-measure was obtained from NB and C4.5 with the 0.98, and the h 0.97. Judging from the generally obtained results, it can be said that Naïve bayes classifier is better with 99% accuracy.

V. CONCLUSIONS AND RECOMMENDATIONS

Currently, kidney disease is a major problem. Because there are so many people with this disease. Kidney disease is very dangerous if not immediately treated on time, and may be fatal. If the doctors have a good tool that can identify patients who are likely to have kidney disease in advance, they can heal the patients in time. Chronic Kidney Disease has been predicted and diagnosed using data mining classifiers: Naive Bayes and C4.5. In future studies, datasets with more data can be used while the performances of the algorithms are being compared.

Moreover, other classifiers can be also used for the same dataset.

REFERENCES

[1] I. H. Witten and E. Frank, “Data Mining Practical Machine Learning Tools and Techniques,” 2nd ed., San Francisco/ABD,2005.

[2] M. Karakoyun and M. Hacıbeyoglu, “Biyomedikal veri kümeleri kullanarak makine ögrenmesi sınıflandırma algoritmalarının karşılaştırılması,” 2014 October 9-10 [Akıllı Sistemlerde Yenilikler ve Uygulamaları (ASYU) Sempozyumu.Izmir/Turkey].

[3] G. Süleymanlar, “Kronik Böbrek Hastalıgı ve Yetmezlig: Tanımı, Evreleri ve Epidemiyolojisi,” Turkiye Klinikleri J Int Med Sci., vol. III, number: 38, 2007, pp. 1-7.

[4] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_ Disease (Access Date: 2018 February7).

[5] M. H. Tanrıverdi, A. Karadag and E. ?. Hatipoglu, “Kronik Böbrek Yetmezligi,” Konuralp Tıp Dergisi, vol. 2, number:2, 2010, pp.27-32.

[6] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli and I. O. Ellis, “A ‘non-parametric’ version of the naïve Bayes classifier,” Knowledge- Based Systems, vol. 24, 2011, pp. 775-784.

[7] C. Cortes and V. Vapnik, “Support- Vector Networks,” Machine Learning, vol. 20, 1995, pp.273-297.

[8] R. Panja and N. R. Pal, “MS-SVM: Minimally Spanned Support Vector Machine,” Appied Soft Computing, vol. 64, 2018, pp.356-365.

[9] A. Bahri, V. Sugumaran and S. B. Devasenapati, “Misfire Detection in IC Engine using Kstar Algorithm,” CoRR, abs/1310.3717,2013.

[10] S. Piramuthu and R. T. Sikora, “Iterative feature construction for improving inductive learning algorithms,” Expert Systems, vol. 36, 2009, pp. 3401- 3406.

[11] C. K. Madhusudana, H. Kumar and S. Narendranath, “Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal,” Engineering Science and Technology, an International Journal, vol. 19, 2016, pp.1543-1551.

[12] T. Kavzoglu and I. Çölkesen, “Karar Agaçları Ile Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örnegi,” Harita Teknolojileri Elektronik Dergisi, vol.2, number:1, 2010, pp. 36-45.

[13] N. Mehdiyev, D. Enke, P. Fettke and P. Loos, “Evaluating Forecasting Methods by Considering Different Accuracy Measures,” Procedia Computer Science, vol. 95, 2016, pp.264-271.

[14] http://www.datascience.istanbul/2017/07/02/hata- matrisini-confusion- matrix-yorumlama/ (Access Date: 2018 February7).

[15] C. Coşkun and A. Baykal, “Veri Madenciliginde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması,” 2011 February 2- 4 [XIII. Akademik Bilişim Konferansı.Malatya/Turkey].