Journal of Physics: Conference Series

Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques

Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques

I.A. Pasadana[1], D. Hartama[12], M. Zarlis[1], A.S. Sianipar[1], A. Munandar[1], S. Baeha[1], A.R.M. Alam[1]

[1]Department of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[2]Department of Computer Science, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

* ignazioahmadpasadana@gmail.com

Abstract. Early detection and proper management of Chronic Kidney Disease (CKD) are solicited for augmenting survivability due to fact that CKD is one of the life-threatening diseases. The UCI's CKD dataset which is selected for this study is consisting of attributes like age, blood pressure, specific grativity, albumin, sugar, red blood cells, plus cell, pus cell clumps, bacteria, blood glucose random, and blood urea. The main purpose of this work is to calculate the performance of various decision tree algorithm and compare their performance. The decision tree techniques used in this study are DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPTree, and SimpleCart. Hence, the results show that RandomForest serves the highest accuracy in identifying CKD.

1. Introduction

Chronic Kidney Disease (CKD) is a termed generally to describe disorders affecting kidney structure and function[1]. CKD becomes a major health problem for the underdeveloped countries of southeast Asia, home to more than 2 billion people [2]. CKD may be triggered by diabetes, high blood pressure and other disorders. Early detection and treatment can often keep CKD from getting worse to save human life.

This paper aims to compare the decision tree algorithms such as DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPTree, and SimpleCart in predicting CKD. The CKD dataset are analyzed using above decision tree algorithms and compare their performance with respect to seven performance metrics (FACC, MAE, PRE, REC, FME, Kappa Statistics and Runtime).

2. Related Works

Subashini, et al. [3] measured the performance of classification techniques using classifier such as Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and Fuzzy. Using Matlab R2016a as analyzing tool, they validated and tested the classification parameters of six methodologies such as Accuracy, Precision, Recall, Specifity, and F-Measure on 400-instances CKD dataset. From the evaluation parameters, they concluded that the fuzzy technique serve the best performance for CKD classification.

IOP Conf. Series: Journal of Physics: Conf. Series **1255** (2019) 012024 doi:10.1088/1742-6596/1255/1/012024

P. Sinha, et al. [4] compared two data mining techniques in predicting CKD by using Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifier. The experimental analysis performed by Matlab tool showed that SVM classifier outperforms three of four evaluation parameters than KNN classifier. E. Celik, et al. [5] diagnosed and estimated of CKD using J48 and Sequential Minimal Optimization (SMO) algorithm in Weka tool. They proposed to compare Decision Tree to Support Vector Machine (SVM) technique in predicting CKD. The classification stage revealed that Decision Tree has more successful than SVM recognition in correct classification.

K.R. Lakshmi, et al. [6] performed a comparative study in predicting kidney dialysis survivability by using thee data mining techniques such as Artificial Neural Network (ANN), Decision Tree, and Logistic Regression. Data analysis used AVF data of 193 under-hemodialysis patients in Hashimenejad Kidney Center (HKC) of Tehran. Using 10-fold cross validation fold, they concluded that ANN show the high level compare than other two techniques in accuracy and sensitivity parameter. S. Ramya, et al. [7] compared Back-Propagation Neural Network (BPN), Radian Basis Function (RBF), and RandomForest (RF) in diagnosing CKD. The dataset including 1000 instances and fifteen attributes were obained from medical reports of patients from different laboratories in Combatore. Using R tool, the experimental results revealed that the RBF gains the highest accuracy than other two algorithm.

Sunil D., et al. [8] addressed the use of two data mining techniques in Rapidminer tool for CKD prediction such as Naïve Bayes and Artificial Neural Network (ANN). The results show that Naïve Bayes outperform ANN in classification accuracy. A. Subas, et al. [9] compared five data mining classifiers such as Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), C.45 Decision Tree, and RandomForest (RF) Decision Tree in Weka tool. The evaluation parameters show that RF serves the highest accuracy than other three techniques. S. Vijayarani, et al. [10] used Support Vector Machine (SVM) and Artificial Neural Network (ANN) to classify kidney disease such as Acute Nephritic Syndrome, Chronic Kidney Disease, Acute Renal Failure, and Chronic Glomerulonephritis. From the results of classification accuracy and execution time, SVM can be considered as better classifier compared to SVM.

3. Method and Framework

Main objective of this study is to identify that whether the patient has CKD or not. Some of the parameter are used for predicting the CKD and compare the performance of the various decision tree techniques using Weka. Weka is a data mining tool which is written in Java and developed at Waikato University [11]. The result found from the CKD dataset by using Weka tool are in section 4. 10-fold cross validation performed on the dataset.

The objective of this study is CKD prediction using data mining tool. The main task in this study is (1) various decision tree techniques are used for the prediction of the CKD; (2) comparing different decision tree techniques; and (3) finding best decision tree for the CKD prediction.

3.1. Decision Tree

Eleven decision tree techniques have been used in this study are DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPTree, and SimpleCart. Their performance was analyzed using Accuracy (ACC), Precision (PRE), Recall (REC), Mean Absolute Error (MAE), F-Measure (FME), Kappa Statistic and Runtime.

3.2. Dataset

The data are collected from UCI Machine Learning Repository and it predicts CKD based on the given attributes [12]. The dataset has twenty four attributes which predict the CKD. The dataset is built on both numerical and nominal data types. As per the UCI's CKD dataset, this contains the attribute such as age, blood pressure, specific grativity, albumin, sugar, red blood cells, plus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, appetite, pedal edema, and anemia.

Table 1. Attribute Description

Attribute Name	Attribute Type	Attribute Values	Attribute Code
Age	numeric	years	age
Blood Pressure	numeric	mm/Hg	bp
Spesific Gravity	numeric	1.005, 1.010, 1.015, 1.020, 1.025	sg
Albumin	numeric	0, 1, 2, 3, 4, 5	al
Sugar	numeric	0, 1, 2, 3, 4, 5	su
Red Blood Cells	nominal	normal, abnormal	rbc
Plus Cell	nominal	normal, abnormal	pc
Plus Cell Clumps	nominal	present, notpresent	pcc
Bacteria	nominal	present, notpresent	ba
Blood Glucosa Random	numeric	mgs/dl	bgr
Blood Urea	numeric	mgs/dl	bu
Serum Creatine	numeric	mgs/dl	sc
Sodium	numeric	mEq/l	sod
Potassium	numeric	mEq/l	pot
Hemoglobin	numeric	gms	hemo
Packed Cell Volume	numeric	-	pcv
White Blood Cell Count	numeric	cells/cumm	wbcc
Red Blood Cell Count	numeric	millions/cumm	rbcc
Hypertension	numeric	yes, no	htn
Diabetes Mellitus	numeric	yes, no	dm
Coronary Artery Disease	nominal	yes, no	cad
Appetite	nominal	good, poor	appet
Pedal Edema	nominal	yes, no	pe
Anemia	nominal	yes, no	ane
Class	nominal	ckd, notckd	class

age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
48	80	1.02	1	0	?	normal	notprese	notprese	121	36	1.2	?	?	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
7	50	1.02	4	0	?	normal	notprese	notprese	?	18	0.8	?	?	11.3	38	6000	?	no	no	no	good	no	no	ckd
62	80	1.01	2	3	normal	normal	notprese	notprese	423	53	1.8	?	?	9.6	31	7500	?	no	yes	no	poor	no	yes	ckd
48	70	1.005	4	0	normal	abnorma	present	notprese	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
51	80	1.01	2	0	normal	normal	notprese	notprese	106	26	1.4	?	?	11.6	35	7300	4.6	no	no	no	good	no	no	ckd
60	90	1.015	3	0	?	?	notprese	notprese	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
68	70	1.01	0	0	?	normal	notprese	notprese	100	54	24	104	4	12.4	36	?	?	no	no	no	good	no	no	ckd
24	?	1.015	2	4	normal	abnorma	notprese	notprese	410	31	1.1	?	?	12.4	44	6900	5	no	yes	no	good	yes	no	ckd
52	100	1.015	3	0	normal	abnorma	present	notprese	138	60	1.9	?	?	10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
53	90	1.02	2	0	abnorma	abnorma	present	notprese	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
50	60	1.01	2	4	?	abnorma	present	notprese	490	55	4	?	?	9.4	28	?	?	yes	yes	no	good	no	yes	ckd
63	70	1.01	3	0	abnorma	abnorma	present	notprese	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
68	70	1.015	3	1	?	normal	present	notprese	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
68	70	?	?	?	?	?	notprese	notprese	98	86	4.6	135	3.4	9.8	?	?	?	yes	yes	yes	poor	yes	no	ckd
68	80	1.01	3	2	normal	abnorma	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
40	80	1.015	3	0	?	normal	notprese	notprese	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
47	70	1.015	2	0	?	normal	notprese	notprese	99	46	2.2	138	4.1	12.6	?	?	?	no	no	no	good	no	no	ckd
47	80	?	?	?	?	?	notprese	notprese	114	87	5.2	139	3.7	12.1	?	?	?	yes	no	no	poor	no	no	ckd
60	100	1.025	0	3	?	normal	notprese	notprese	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
62	60	1.015	1	0	?	abnorma	present	notprese	100	31	1.6	?	?	10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
61	80	1.015	2	0	abnorma	abnorma	notprese	notprese	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
60	90	?	?	?	?	?	notprese	notprese	?	180	76	4.5	?	10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
48	80	1.025	4	0	normal	abnorma	notprese	notprese	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
21	70	1.01	0	0	?	normal	notprese	notprese	?	?	?	?	?	?	?	?	?	no	no	no	poor	no	yes	ckd
42	100	1.015	4	0	normal	abnorma	notprese	present	?	50	1.4	129	4	11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd
61	60	1.025	0	0	?	normal	notprese	notprese	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd
75	80	1.015	0	0	?	normal	notprese	notprese	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes	no	poor	no	no	ckd
69	70	1.01	3	4	normal	abnorma	notprese	notprese	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	ckd

Figure 1. Dataset for CKD Prediction

4. Results and Discussion

The comparison of various decision tree algorithms performed on CKD data is shown in Table 2. RamdomForest has the highest accuracy rate. The accuracy rate of this algorithm is 100%. It is seen that RandomForest is the most powerful classifier for this dataset. This result shows that the CKD of a new patient is predicted successfully with an acceptable ratio 100%. J48, J48graft and NBTree has the accuracy rate as 99%, 98.75% and 98.5%. It is also seen that DecisionStump has the worst accuracy rate with 92%.The comparison of decision tree algorithm with respect to accuracy shown in Figure 2.

The comparison of decision tree algorithm with respect to kappa statistics and runtime is shown in Figure 3 and Figure 4 respectively. RandomTree, DecisionStump and J48 are faster than other algorithms. NBTree algorithm takes a long time even though a small dataset is used.

Table 2. Comparing The Various Decision Tree Algorithm Carried Out CKD

Techniques	Tree Size	ACC	MAE	PRE	REC	FME	Kappa Statistics	Run-time
DecisionStump	Single Level	92%	0.1362	0.926	0.920	0.921	0.8338	0.01
HoeffdingTree	1	95.75%	0.043	0.962	0.958	0.958	0.9113	0.07
J48	14	99%	0.0225	0.990	0.990	0.990	0.9786	0.01
CTC	9	97%	0.0387	0.972	0.970	0.970	0.937	0.03
J48graft	14	98.75%	0.0244	0.987	0.988	0.987	0.9733	0.03
LMT	1	98%	0.0222	0.981	0.980	0.980	0.9577	0.72
NBTree	15	98.5%	0.0289	0.985	0.985	0.985	0.9682	2.99
RandomForest	-	100%	0.0414	1	1	1	1	0.14
RandomTree	67	95.5%	0.045	0.956	0.955	0.955	0.905	0
REPTree	8	96.75%	0.0634	0.968	0.968	0.967	0.9302	0.02
SimpleCart	-	97.5%	0.037	0.975	0.975	0.975	0.9462	0.22

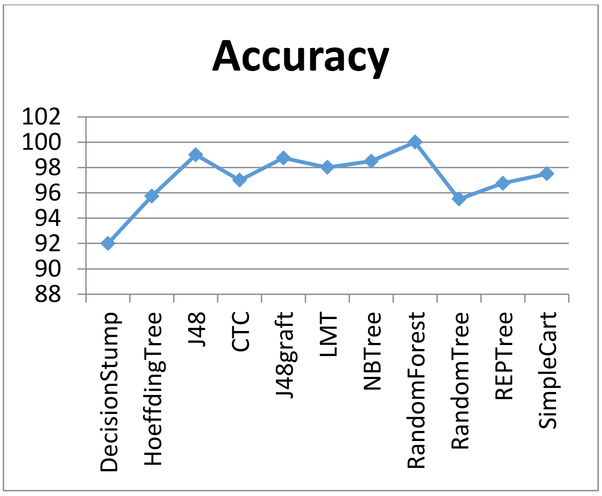

Figure 2. Comparison of The Decision Tree Algorithm According to Accuracy

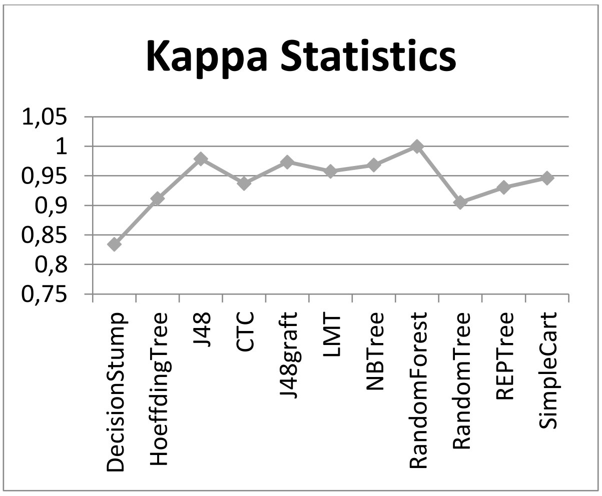

Figure 3. Comparison of The Decision Tree Algorithms According to Kappa Statistics

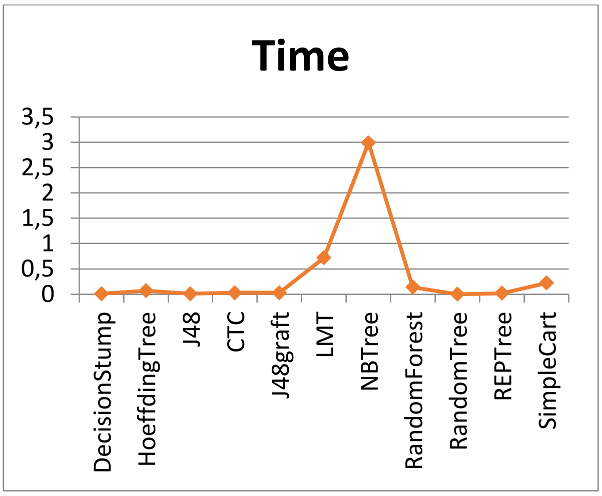

Figure 4. Comparison of The Decision Tree Algorithms According to Time

5. Conclusion

The research employed some decision tree algorithm such as DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPTree, and Simple Cart to predict the CKD at an earlier stage. These algorithm provide various experimental result based on Accuracy, Mean Absolute Error, Precision, Recall, Kappa Statistics and Runtime. These techniques were evaluated and their performance was compared. From the analysis, RandomForest outperforms well than other algorithms and its achieved accuracy is 100%. The application of decision tree in predicting CKD will benefit in maintain the health. However, in future, we will collect the very recent data from various regions across the world for CKD diagnosis. The results of this study will encourage us to continue developing other advanced decision trees algorithm.

References

[1] Levey AS, Coresh J. *Chronic Kidney Disease*. Lancet 2012;379:165-80.
[2] V. Jha. 2009. *Current Status of Chronic Kidney Disease Care in Southeast Asia*. Seminars in Nephrology Journal, Volume 29, Issue No. 5, pp. 487-496.
[3] R. Subashini and M.K. Jeyakumar. 2017. *Peformance Analysis of Different Classification Techniques for The Prediction of Chronic Kidney Disease*. International Journal of Pharmacy and Technology, Volume 9, Issue No. 4, pp. 6563-6582.
[4] P. Sinha and P. Sinha. 2015. *Comaparative Study of Chronic Kidney Disease Prediction by Using KNN and SVM*. International Journal of Engineering Research and Technology (IJERT), Volume 4, Issue No. 12.
[5] E. Celik, M. Atalay, and A. Kondiloglu. 2016. *The Diagnosis and Estimate of Chronic Kidney Disease Using Machine Learning Methods*. International Journal of Intelligent System and Applications in Engineering, Special Issue No. 4, pp. 27-31.
[6] K.R. Lakshmi, Y. Nagesh, and M. VeeraKhrisna. 2014. *Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability*. International Journal of Advances in Engineering and Technology, Volumen 7, Issue No.1, pp. 242-254.
[7] S. Ramya and N. Radha. 2016. *Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms*. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Volume 4, Issue No. 1.
[8] Sunil D. and B.P. Sowmya. 2017. *Chronic Kidney Disease Analysis Using Data Mining*. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology (IJSRCSEIT), Volume 2, Isssue No. 4.

[9] A. Subas, E. Alickovic, and J. Kevric. 2017. *Diagnosis of Chronic Kidney Disease by Using Rabdo Foret*. International Federation of Medical and Biology Engineering (IFMBE), Volume 62.
[10] S. Vijayarani and S. Dhayanand. 2015. *Kidney Disease Prediction Using SVM and ANN Algorithms*. International Journal of Computing and Business Research (IJCBR), Volume 6, Issue No. 2.
[11] S. S. Aksenova, 2004. *Machine Learning with WEKA - WEKA Explorer Tutorial for WEKA Version 3.4*.
[12] Machine Learning Repository, Center for Machine Learning and Intelligent Systems https://archive.ics.uci.edu/ml/index.php.