# Performance Comparison of Classification Algorithm in Data Mining Techniques using Chronic Kidney Dataset

**A.Ajeeth[1] D.Ramya Chitra[2]**
[1] M. Phil Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Bharathiar University, Coimbatore

*Abstract*— The problem of chronic kidney disease is getting worsened day by day. It is also known as chronic renal disease and is a life threatening disease; it has various symptoms such as high blood pressure, anemia, rashes, muscle pain, conjuctivitus, etc. So, in order to tackle this problem it has to be detected at earliest stages possible and given suitable treatment before it get worsened. We have used 24 symptoms of chronic kidney disease in this paper which help us to accurately detect this disease with the help of eight classification algorithms i.e. SGD, Random subspace, SMO, JRIP rules, Hoeffding tree, NaiveBayes, Locally weighted learning, oneR in data mining tool WEKA. We conclude the results by introducing the medical datasets to all three algorithms separately with the help of knowledge flow interface of WEKA data mining tool, the parameters which are used to compare the results of these three different algorithms are mean absolute error, kappa statistics and total number of instances studied either correctly or incorrectly. The main aim of this paper is present a clear view of the chronic kidney disease, its symptoms and the criteria to detect it at earliest stages possible which will help the mankind to get safe from this life threatening disease.

***Key words:*** Chronic kidney disease, classification, data mining, WEKA

## I. INTRODUCTION

Data mining is an approach which dispense an intermixture of technique to identify a block of data or decision making knowledge in the database and eradicating these data in such a way that they can be put to use in decision support, forecasting and estimation [1]. The data is often voluminous, but it has data that is useful. Two major preferred models that can be created in data mining are predictive and descriptive. Under these two models there are various tasks that are used in the data mining process. On basis of various historical data a predictive model makes estimation about values of data using recognized results found from various data. On the other side, descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model obliges as way to explore the properties of the data observed, not to predict new properties [2]. The algorithms are many in every single task under both the data mining models which are used for various purposes according to the convenient of the use requirements. The various tasks of the predictive and descriptive models are classification, clustering, summarization, prediction, time series analysis, association rules and regression [3].

In order to anticipate solution set for various problems data mining technique endeavors distinctive data mining tasks such as classification and clustering.

In this paper data mining technique is used in order to detect the chronic kidney disease at earliest stages possible and give proper attention to it. Eight classification algorithms are used in this paper i.e. SGD, Random subspace, SMO, JRIP rules, Hoeffding tree, NaiveBayes, Locally weighted learning, oneR to detect it at earliest stages possible. The tool or software that is used to conduct experiments on this datasets with these algorithms is WEKA data mining tool that is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [4].

The 24 symptoms of chronic kidney disease which are considered while detecting this disease are as follows

| S.No | SYMPTOM | SHORT FORM |
|---|---|---|
| 1 | Age | age |
| 2 | Blood pressure | bp |
| 3 | Specific gravity | sg |
| 4 | Albumin | al |
| 5 | Sugar | su |
| 6 | Red blood cells | rbc |
| 7 | Pus cell | pc |
| 8 | Pus cell clumps | pcc |
| 9 | Bacteria | ba |
| 10 | Blood glucose random | bgr |
| 11 | Blood urea | bu |
| 12 | Serum creatinine | sc |
| 13 | Sodium | sod |
| 14 | Potassium | pot |
| 15 | Hemoglobin | hemo |
| 16 | Packed cell volume | pcv |
| 17 | White blood cell count | wc |
| 18 | Red blood cell count | rc |
| 19 | Hypertension | htn |
| 21 | Diabetes mellitus | dm |
| 21 | Coronary artery disease | cad |
| 22 | Appetite | appet |
| 23 | Pedal edema | pe |
| 24 | Anemia | ane |
| 25 | Class | class |

Table 1: List of symptoms used to detect chronic kidney disease

The symptoms used in this paper for detecting chronic kidney disease are given above with their names and short forms which are used in this paper [5]. The further detailed information about this paper is as follows like section 2 provides the literature survey which explains the previous works done in this field, section 3 describes the methodology used in this paper in detecting the disease

 

accurately and the datasets used for results in this paper, section 4 gives the results of this paper and section 5 gives the conclusion and future scope.

## II. LITERATURE SURVEY

Lambodar Jena et al [6] has suggested the use of various algorithms on chronic kidney disease datasets and compared the results based on different parameters but only on single interface of WEKA and it has been calculated that multilayer perceptron algorithm performs the best among used algorithms.

Morteza Khavanin Zadeh et al [7] suggested the prediction of early chronic kidney disease and data of 193 patients who underwent homodialysis in Hasheminejad Kidney Center were explored. Eight common attributes of the patients including age, sex, hypertension level, Diabetes Mellitus state, hemoglobin level, smoking behavior, location of Arteriovenous fistula, and thrombosis state were used in the machine learning process and only two algorithms are used for prediction process.

Shital Shah et al [8] suggested the data mining approach which helps in detecting the chronic kidney effectively and relating it to the survival of the patient but again only with the help of limited algorithms and single interface.

L.Jerlin Rubini et al [9] used only three different algorithms such as radial basis function network, multilayer perceptron, and logistic regression. Also the interface used in this study only one.

Asim Roy et al [10] in his paper describes about the high-dimensional stored big data and streaming data.

Pushpa M. Patil et al [11] in her review paper on chronic kidney disease researched that CKD can very well be predicted using many classification algorithms using many data mining tools.

Arturo Gil et al[12] has suggested the SGD algorithm for the classification of the chronic kidney disease dataset.

Mario et al[13] gives a survey about the pattern classification in medical dataset. It clearly explains the algorithm for the disease datasets like diabetes, heart, hepatitis etc.

Parul sinha et al [14] introduces a new decision support system to predict chronic kidney disease. They compare the performance of SVM and KNN classifier on the basis of its accuracy precision and execution time for CKD prediction. It has been concluded that KNN is better than SVM classifier.

## III. METHODOLOGY

In this research the algorithms such as SGD, Random subspace, SMO, JRIP rules, Hoeffding tree, NaiveBayes, Locally weighted learning, oneR are compared to predict kidney disease. From the result it is inferred that SGD algorithm provides better results than the other algorithms.

### *A. Dataset*

Dataset is a collection of data or a single statistical data where every attribute of data represents variable and each instance has its own description. For prediction of Chronic kidney disease we used datasets [15] for prediction and classification of algorithms in order to compare their accuracy using wekas knowledge flow interface. The datasets used by us contains 25 attributes and 400 instances out of which 250 are suffering from the disease and 150 are not suffering from the disease. We have applied different algorithms using WEKA data mining tool for our analysis purpose.

| S.No | Major Attributes | Data Type |
|---|---|---|
| 1 | Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packaged Cell Volume, WBC count, RBC count. | Numerical |
| 2 | Specific Gravity, Albumin, Sugar, RBC,Pus cell, Pus cell clumps, Bacteria,Hypertension, Diabetes Mellitus,Coronary Artery Disease, Appetite, Padal Edema, Anemia and class. | Nominal |

Table 2: Data Type of Attribute

### *B. Performance Analysis Factors*

The purpose of this evaluation is to find a suitable method for dataset. In addition to prediction accuracy error rate, Sensitivity, Specificity, F-score and Kappa are also examined as the performance analysis factors. These factors are used to evaluate each classifier for ckd prediction analysis. The Factors are as follows:

*1) Sensitivity*

It is a statistical measure of the performance of a binary classification test, also known in statistics as classification function. Sensitivity (also called the true positive rate, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

*2) Specificity*

It (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition). These two measures are closely related to the concepts of type I and type II errors.

*3) Accuracy*

Accuracy is the percentage of correctly classified instances. It is one of the most widely used classification performance metrics.

*4) F-score*

It is important to calculate the F-score, defined as the weighted harmonic mean of precision and recall. the best F-score, which was chosen because this is an average measure. Special care was taken against overtraining, because some algorithms can be affected by this effect. Overtraining can be detected if the training has higher accuracy than the prediction.

*5) Correctly Classified Accuracy*

It shows the accuracy percentage of test that is correctly classified.

*6) Incorrectly Classified Accuracy*

It shows the accuracy percentage of test that is incorrectly classified.

 

*7) Mean Absolute Error*

It shows the number of errors to analyze algorithm classification accuracy.

*8) Kappa statistics*

It measures inter-rater agreement for qualitative items.

*C. Algorithms*

*1) Naive Bayes*

This is a classification technique based algorithm which is based on Bayes' Theorem and its main feature is that it doesn't have inter dependence among the various predictors to be used in the experiments.

*2) SMO*

This algorithm divides the large quadratic problems into smallest possible sets and solve data analytically which helps in saving the excess time, reduced memory space is required for data and allow to use huge datasets very efficiently.

*3) SGD*

Stochastic gradient descent (often shortened in SGD), also known as incremental gradient descent, is a stochastic approximation of the gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In other words, SGD tries to find minimums or maximums by iteration.

*4) Random subspace*

In machine learning the random subspace method, also called attribute bagging or feature bagging, is an ensemble learning method that attempts to reduce the correlation between estimators in an ensemble by training them on random samples of features instead of the entire feature set.

*5) JRIP rules*

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

*6) Hoeffding tree*

A Hoeffding tree (VFDT) is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time. Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quantifies the number of observations (in our case, examples) needed to estimate some statistics within a prescribed precision (in our case, the goodness of an attribute).

*7) Locally weighted learning*

Locally Weighted Learning is a class of function approximation techniques, where a prediction is done by using an approximated local model around the current point of interest.

*8) oneR*

Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes.

## IV. RESULTS AND DISCUSSIONS

The classification data mining technique is used in this paper by various algorithms such as SGD, Random subspace, SMO, JRIP rules, Hoeffding tree, NaiveBayes, Locally weighted learning, oneR with only one interface that is Knowledge flow. The parameters that have been used to validate the results are correctly classified instances, kappa statistics and mean absolute error.

The comparison of performance factors for the classification algorithms are shown in Fig. 2 and the comparison of accuracy measure are shown in Fig. 3.

| ALGORITHMS | KAPPA STATISTIC | TP RATE | FP RATE | PRECISION | F-MEASURE | ROC AREA |
|---|---|---|---|---|---|---|
| SGD | 0.9841 | 0.993 | 0.005 | 0.993 | 0.993 | 0.994 |
| Random subspace | 0.9788 | 0.99 | 0.006 | 0.99 | 0.99 | 1 |
| SMO | 0.9735 | 0.988 | 0.008 | 0.988 | 0.988 | 0.99 |
| JRIP rules | 0.968 | 0.985 | 0.017 | 0.985 | 0.985 | 0.989 |
| Hoeffding tree | 0.9113 | 0.958 | 0.026 | 0.962 | 0.958 | 1 |
| Naive Bayes | 0.8911 | 0.948 | 0.032 | 0.954 | 0.948 | 1 |
| Locally weighted learning | 0.8488 | 0.928 | 0.059 | 0.932 | 0.928 | 0.996 |
| oneR | 0.8413 | 0.925 | 0.077 | 0.077 | 0.925 | 0.924 |

Table 3: Performance Factors For The Classification Algorithms

| ALGORITHMS | CORRECTLY CLASSIFIED | INCORRECTLY CLASSIFIED |
|---|---|---|
| SGD | 99.25% | 0.75% |
| Random subspace | 99% | 1% |
| SMO | 98.75% | 1.25% |
| JRIP rules | 98.50% | 1.50% |
| Hoeffding tree | 95.75% | 4.25% |
| NaiveBayes | 94.75% | 5.25% |
| Locally weighted learning | 92.75% | 7.25% |
| oneR | 92.50% | 7.50% |

Table 4: Accuracy Measures for Classification Algorithms

 

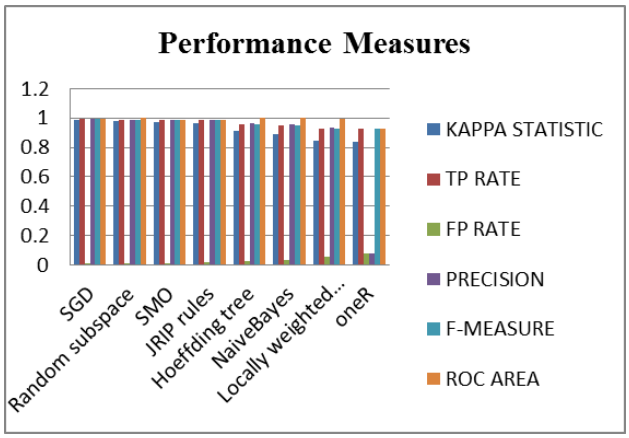


Fig. 1: Performance Measures for the Classifier algorithms

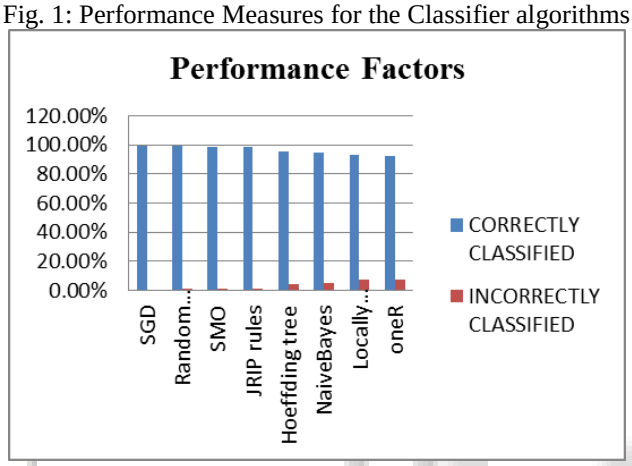


Fig. 2: Accuracy Measure for the Classifier algorithms

For Locally weighted learning algorithm it is inferred that for the cross validation parameter, the Precision, ROC, F-Measure, TP Rate values gives poor results than other algorithms. The Error rate measure for the classification is depicted in Table 3. And also Accuracy error rate measure for the classifier is shown in the Fig. 4.

| ALGORITHMS | MEAN ABSOLUTE ERROR | ROOT MEAN SQUARED ERROR |
|---|---|---|
| SGD | 0.0075 | 0.0866 |
| Randomsubspace | 0.0595 | 0.1041 |
| SMO | 0.0125 | 0.1118 |
| JRIP rules | 0.027 | 0.1162 |
| Hoeffdingtree | 0.042 | 0.1912 |
| NaiveBayes | 0.0498 | 0.2108 |
| Locally weighted learning | 0.1063 | 0.2383 |
| oneR | 0.075 | 0.2739 |

Table 5: Error Rate Measure for Classification Algorithm

| ALGORITHMS | RELATIVE ABSOLUTE ERROR | ROOT RELATIVE SQUARED ERROR |
|---|---|---|
| SGD | 1.60% | 17.89% |
| Randomsubspace | 12.68% | 21.50% |
| SMO | 2.67% | 23.09% |
| JRIP rules | 5.76% | 24.00% |
| Hoeffdingtree | 8.95% | 39.49% |
| NaiveBayes | 10.61% | 43.55% |
| Locally weighted learning | 22.67% | 49.23% |
| oneR | 15.99% | 15.99% |

Table 6: Error Rate Measure For Classification Algorithms

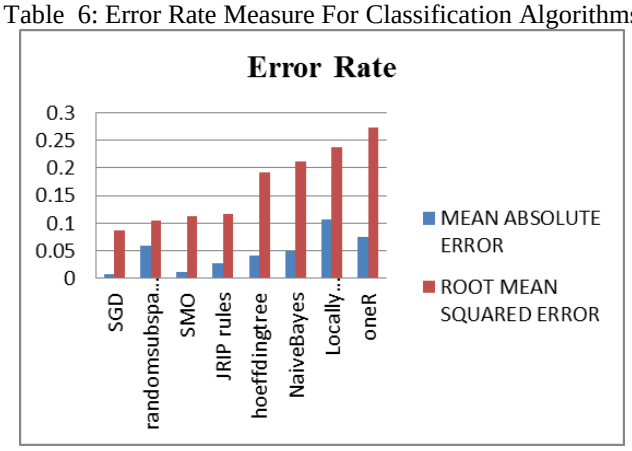


Fig. 3: Accuracy error rate measure for classification algorithms

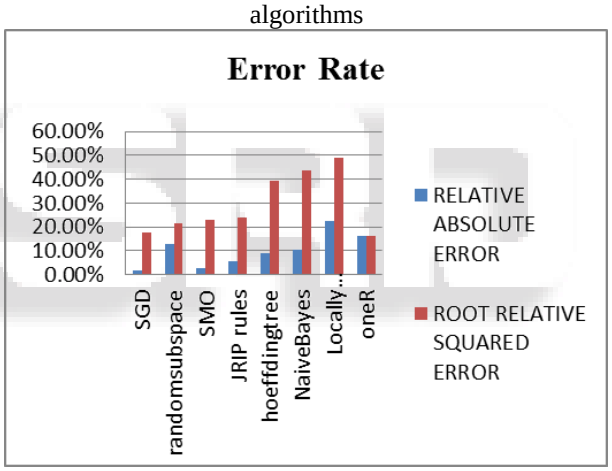


Fig. 4: Accuracy error rate measure for classification algorithms

From the Table 3 and 4 it is clearly visible that SGD is the best performing algorithm because it has classified maximum number of correct instances i.e. 396, has the least mean absolute error i.e 0.0075 and has maximum kappa statistics i.e 0.9841. So, from knowledge flow interface it is clear that SGD is the best performing algorithm in case of detecting chronic kidney disease.

## V. CONCLUSION

The algorithm with highest accuracy value and the lowest error rate is declared as the best algorithm. This paper presents the performance of 8 different classification algorithms like SGD, Random subspace, SMO, JRIP rules, Hoeffding tree, NaiveBayes, Locally weighted learning, oneR on the Chronic kidney disease dataset which is downloaded from the UCI repository. Hence it is proved that SGD algorithm performs better for the Chronic kidney disease dataset. In the future course of this study one can try to further improve the two-class classification accuracy by evaluating some hybrid or ensemble techniques.

REFERENCES

[1] Mahesh Mudhol Purushothama Gowda,( 2004) Data Mining in the Process of Knowledge Discovery in Digital Libraries, 2nd Convention PLANNER, Manipur Uni., Imphal, 4-5 November, 2004, page no 164-167
[2] Fadzilah Siraj, Mansour Ali Abdoulha, (2011). Mining Enrollment Data Using Descriptive and Predictive
[3] Approaches, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-1541,InTech,http://www.intechopen.com/books/knowledge-oriented-applications-in-datamining/mining-enrollment data-using-descriptive-and-predictive-approaches
[4] www.cs.waikato.ac.nz/ml/weka/downloading.html archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
[5] International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11) Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease Lambodar Jena, Narendra Ku. Kamila
[6] International Journal of Hospital Research 2012, 2(1):49-54 Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients Morteza Khavanin Zadeh , Mohammad Rezapour , Mohammad Mehdi Sepehri
[7] S. Shah, A. Kusiak, and B. Dixon, Data Mining in Predicting Survival of Kidney Dialysis Patients, in Proceedings of Photonics West - Bios 2003,Bass, L.S. et al. (Eds), Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems XIII, Vol. 4949, SPIE, Belingham, WA, January 2003, pp. 1-8.
[8] International Journal Of Modern Engineering Research (IJMER) | IJMER | ISSN: 2249–6645 | www.ijmer.com | Vol. 5 | Iss. 7 | July 2015 | 49 | Generating comparative analysis of early stage prediction of Chronic Kidney Disease L.Jerlin Rubini, Dr.P.Eswaran
[9] Asim Roy, Elsevier, Procedia CS, 'A Classification algorithm for high- dimensional data'2015
[10] International journal of computer science and mobile computing volume 5 issue 5, may 2016 (ijcsmc) Review on prediction of chronic kidney disease using data mining techniques Pushpa M. Patil
[11] Arturo et al, "Occupancy grid based graph using the distance transformation, SGD features.
[12] Mario Aldape – Perez et.al "Collaborative learning based on associative models: Application to pattern classification in medical dataset", Elsevier, Computers in human Behavior 51(2015) 771-779
[13] International journal of engineering research and technology volue 4 issue 12, December 2015 (ijert) Comparative study of chronic kidney disease prediction using SVM and KNN Parul sinha Poonam sinha.
[14] International journal of computer science and mobile computing volume 5 issue 5, may 2016 (ijcsmc) Review on prediction of chronic kidney disease using data mining techniques Pushpa M. Patil.

 