Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest

Devika R, Sai Vaishnavi Avilala, V. Subramaniyaswamy

School of Computing,SASTRA Deemed University, India

srdevika@cse.sastra.edu

Abstract— **Chronic kidney disease (CKD), is also known as chronic nephritic sickness. It defines constrains which affects your kidneys and reduces your potential to stay healthy. There will be various complication concerns like increased levels in your blood, anemia (low blood count), weak bones, and nerve injury. Detection and treatment should be done prior so it will typically keep chronic uropathy from obtaining a worse condition. Data processing is the term used for information discovery from big databases. The task of knowledge mining is to generate regular patterns from historical data and emphasize future conclusions , follows from the convergence of many recent trends: the decreased value of huge knowledge storage devices and therefore the tremendous ease of aggregation knowledge over networks; the development of robust and economical machine learning algorithms to method this data; and therefore the decrease value of machine power, enabling use of computationally intensive strategies for knowledge analysis. Machine learning is an important task as it benefits many applications such as analyzing life science outcomes, sleuthing fraud, sleuthing faux users etc. varied knowledge mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. Therefore, this paper examines the performance of Naive Bayes, K-Nearest Neighbour (KNN) and Random Forest classifier on the basis of its accuracy, preciseness and execution time for CKD prediction. Finally, the outcome after conducted research is that the performance of Random Forest classifier is finest than Naive Bayes and KNN**

Keywords—Data Mining, Machine learning, Chronic kidney disease, Classification, K-Nearest Neighbour, Naive Bayes, Random Forest.

1. INTRODUCTION

Data mining deals with extraction of helpful data from vast amounts of knowledge. several alternative terms are getting used to knowing data mining, like mining of data from databases, knowledge extraction, information analysis, and information anthropology. Basically, data {processing} could be a crucial step within the process of data discovery in databases or KDD. the info mining techniques of classification, clump, and association facilitate in extracting knowledge from an outsized quantity data|of information}. Machine Learning could be a rising field involved with the study of big and multiple variable information. The process of learning in identifying patterns and similarities in the information in artificial intelligence lead to the evolution of machine learning. This involves process ways, algorithms, and techniques for analysis. In perspective, Machine Learning guarantees to help physicians to create near-perfect diagnoses,.There are varied applications for Machine Learning, the foremost importance of that is data processing. Machine learning at the side of data processing will usually be effectively applied to such issues, as they improve the potency of the systems and their styles. The same group of options is benefited for the illustration of each case, in any dataset employed by Machine learning algorithms. These options are often continuous, categorical or binary. If the instances are given with category labels or celebrated labels i.e. with the corresponding correct outputs, then the training is named supervised learning, on the opposite hand comes unsupervised learning, wherever instances or category are unlabeled. Researchers hope to get heaps of data victimization these supervised and unsupervised learning.

Classification could be a data processing performance that assigns things during an assortment to focus on classes or categories. The goal is to reliably estimate the target class for every instance within the information.

Various data mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. Here we are concerned about Chronic kidney disease (CKD), also known as chronic renal disease, is an abnormal function of kidney or progressive failure of short-term loss of kidney over a period of months or years. Often, people will be undergoing screening tests when they

are at risk of kidney problems, such as those with hypertension cardiovascular disease or high blood sugar and those with a blood relative with CKD. It is differentiated from acute kidney disease in that the reduction in kidney function must be present for over 3 months. Hence, the prediction of chronic kidney disease is one of the most important tasks.

Chronic Kidney disease is predicted using classification methods of data mining. The classifiers used here are, Naive Bayes, K-Nearest Neighbor (KNN), Random Forest classifier. Their performance is assessed in terms of accuracy, precision, and F-measure.

2. Literature Survey

In 2015, Konstantina Kourou et.al [1] planned a study of Machine learning applications in cancer prognosis and prediction. in this paper, they need to bestow a review of assorted recent mil approaches that square measure applied for the prediction of cancer detection. Here they need bestowed review of new printed content for the work done to this point in cancer detection.

In 2015 P.Swathi Baby et. al [2] planned a project to identification and prediction system supported prognostic mining. Here nephropathy information set is employed and analyzed exploitation rail and Orange computer code. Here the Machine learning algorithms like AD Trees, J48, K star, Naive Bayes, Random forest were studied. As a result, K-Star and Random Forest are the simplest among all the algorithms for the used dataset.

In 2014 K.R.Lakshmi ET.AL [3] projected performance analysis of 3 data processing techniques for predicting urinary organ chemical analysis survivability. during this analysis, numerous data processing techniques (Artificial Neural Networks, call tree, and Logical Regression) area unit accustomed to extracting information concerning the interaction between these variables and patient survival. Information is extracted by comparing the performance of 3 data processing techniques. The ideas introduced during this analysis are engaged and experimented employing a knowledge gathered at completely different chemical analysis sites. The outcomes area unit reported. Finally, ANN is recommended for urinary organ chemical analysis to induce higher results with accuracy and performance

In paper [30] they projected an intellectual distribution framework to observe typical and abnormal magnetic resonance imaging brain pictures.

In the paper [32] by Leena Vig had conferred Associate in Nursing analysis victimization Random Forest classifiers, Artificial Neural Networks, Naïve Bayes a and Support Vector Machines. Results show that ANN's, Random Forests and SVMs square measure ready to provide models with high accuracy, sensitivity, and specificity. However, Naïve Bayes accomplish weak performance.

In 2015, Mister S Dayanand [5]proposed the analysis work to predict internal organ diseases by exploitation Support Vector Machine (SVM) and Artificial Neural Network (ANN). The goal of this task is to match the performance of the pair of algorithms on the idea of its accuracy and execution time. Finally, it's concluded that the performance of the ANN is healthier than the alternative formula.

The existing prediction system for chronic kidney disease is okay with some limitations. Below is that the table is shown, describing the worked in serious trouble prediction and detection of assorted urinary organ diseases. a replacement CKD prediction system continues to be the requirement. a call network for chronic renal disorder continues to be the requirement for early prediction, as not abundant work is finished for constant.

3. Proposed Methodology

The work planned here uses three classification techniques to predict the presence of chronic renal disorder in humans. The classifiers used area unit Naive Bayes, KNN and Random Forest classifier. the information set for chronic sickness| was gathered and applied on every classifier to predict the disease and therefore the performance of the classifier is examined based on their accuracy, precision and F measure.

The operating of the design is as follows: The dataset for CKD patients are collected and fed into the classifier named Naive Bayes, KNN and Random Forest. The prediction of CKD is done with the algorithms performed in C Sharp Language. In this paper, the data set is collected from the UCI machine learning repository, because of the input for prediction. The dataset consists of attributes and values. We can predict the result from the accuracy that what number patients square measure having the chronic nephropathy at intervals a specific time. The experimental results retrieved, that shows the most effective classifier among the three.

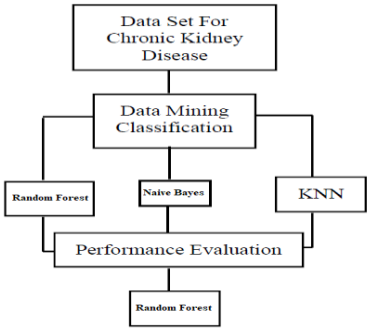

Fig 1. Workflow diagram of the classifier for chronic kidney disease prediction using Navie Bayes, KNN and Random

3.1 Data Mining Techniques

3.1.1 K-nearest neighbour Classification

K-Nearest Neighbor algorithm (K-NN) is widely used for classification and regression in recognition of patterns and consistency in data.In a family of algorithms, K-NN is a supervised learning algorithm. K-NN is a non-parametric technique and is employs for evaluation of statistics. In both cases, the input is taken from a training block that is the training input then there will be a related target and output model will be generated as a model. K-NN is a type of memory-based learning because hypotheses are built directly from training instances The neighbours are derived from the set of objects of a known class. If K = 1, then the single nearest neighbour is assigned to the class. In a common weighting scheme. A straight line will always be produced when there is the shortest distance among 2 neighbours and the distance is called Euclidean distance [7].In K-NN Regression the output will be the mean of the values of its k nearest neighbours The drawbacks of K-NN algorithm are it is not a fastest algorithm, works with less number of inputs, require homogeneous features, delicate for the local alignment of the data.

3.1.3 Random Forest

Random forests [32] are an aggregate of tree predictors so that each one timber rely on the values of a random vector sampled autonomously and with the similar distribution of all bushes inside the forest.
Estimation of random forests algorithm is done or class project can be described as follows:

1. tree bootstrap is figured out by taking out distinctive samples of data
2. for each of the bootstrap samples gives an unpruned category tree, by the following way

amendment: at each node, among all predictors selection of first-class split is not preferable, arbitrarily pattern m attempt of the predictors and from the one's variables, select the acceptable breakup and new data is expected by means of summing up the predictions of the n-tree trees the use of majority votes for types.

4. RESULTS and EVALUATION

Underneath are the figures displaying the performance of numerous assessment parameters. the x-axis denotes the share of performance completed whereas the y-axis denotes the ratio of information set taken for evaluation. for evaluation the subsequent algorithms are run:

4.1 Evaluation Parameters:

A. Precision and recall

Precision:

It is additionally called positive prognostic worth. it's outlined because of the mean chance of appropriate retrieval.

Precision = variety of true positives/Number + False positives

Recall

It is outlined because of the mean chance of absolute retrieval.

Recall= True positives/True positives + False negative

Accuracy

Accuracy is outlined in means of properly classified instances divided by the whole variety of instances gift within the dataset. Accuracy=Number of properly classified samples/Total variety of samples.

Confusion Matrix

It displays the quantity of accurate and inaccurate predictions created by the model compared with the particular classifications within the take a look at information. The matrix is depicted within the type of n-by-n, wherever n is that the variety of categories. The accuracy of every classification algorithms is calculated from that.

For The Formulas: tp is truly positive, FP is false positive, fn Is false negative, TN is a true negative.

$$\text{Average Accuracy} = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

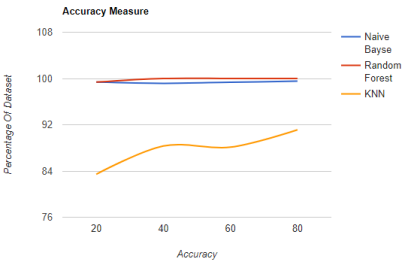

Fig. 5.1 Accuracy for Percentage of Dataset

$$\text{F-Measure} = \frac{(\beta^2 + 1)\text{precision}_M \text{recall}_M}{\beta \text{precision}_M \text{recall}_M}$$

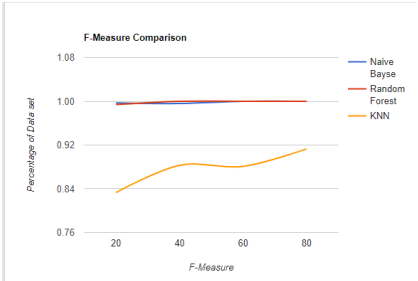

Fig. 5.2 F-measure for Percentage of Dataset

Fig 5.1 and Fig 5.2 illustrates Accuracy and F-Measure. Accuracy refers to the ability of a classifier to predict the class label correctly. F-measure defines as the weighted harmonic mean of precision and recall. Hence, among three classifiers Random Forest has achieved highest in terms of accuracy and F-measure.

$$\text{Precision} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$$

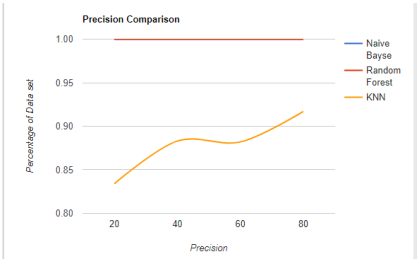

Fig. 5.3 Precision graph

$$\text{Recall} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$

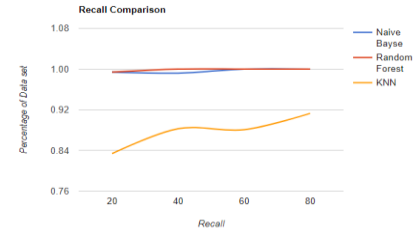

Fig. 5.4 Recall for percentage of Dataset

Fig 5.3 and fig 5.4 depicts Precision and Recall. Precision is a metric of relevant results while recall refers to how many truly relevant results are correctly classified. As a whole, Naive Bayes has elevated in terms of precision and whereas random forest outrageous in recall.

To sum up, The below figures show that Random Forest finished better in phrases of accuracy, and f degree over distinctive datasets, whereas Naive Bayes, shows better Precision. hence we are able to say that Random Forest achieved higher than KNN and Naive Bayes in the prediction of CKD in our analysis.

TABLE I: Performance Evaluation Metrics

Name of *Classifier*	**Evaluation Parameter**			
	Accuracy	***Precision***	***Recall***	***F-measure***
Naive Bayes	99.635	1	0.996	0.998
KNN	87.78	0.879	0.877	0.8775
Random Forest	99.844	0.9985	0.99	0.99

This work is performed in C-Sharp tool. Matrix manipulations, plotting of functions and data, implementation of algorithms, the creation of user interfaces, and interacting with programs written in other languages and tools. The experimental comparison of KNN, Naive Bayes and Random Forest are implemented through the performance measures of classification accuracy and precision.

Datasets

The dataset is gathered from many medical labs, pharmacies, community health centers and hospitals. From this, the mock urinary organ operates to take a look at (KFT) dataset are fashioned for the study of nephrosis. This dataset contains four hundred instances and twenty 5 entities are employed in this comparative study. The features during this KFT dataset age, pressure, relative density, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood sugar random, blood urea, blood serum creatinine, sodium, potassium, haemo-protein, packed cell volume, white somatic cell count, red somatic cell count, high blood pressure, DM, arterial blood vessel sickness, appetite, pedal lump, anaemia, class. This dataset consists of nephritic affected sickness info. This is binary classification, as we've used 2 categories for predicting CKD and NOT CKD.

The datasets are extracted from UCI device mastering repository benchmarks. The UCI system gaining knowledge of repository is a collection of databases, are theories, and data mills which can be utilized by the system learning network for the empirical evaluation of gadget studying algorithms.

5. Conclusion

As we've already seen the programs of statistics mining and system learning inside the clinical region. in this paper, a new selection aid device is carried out for the prediction of CKD. Even though the classifiers labored efficiently within the

prediction of another sickness additionally. In this paper, chronic kidney ailment is anticipated using specific classifiers and a comparative study of their overall performance is finished. from the evaluation, we found that out of classifiers Naive Bayes, Random Forest, and KNN, Random Forest classifier performed higher than the alternative. the price of prediction of CKD is advanced. There are different viable evolutionary strategies that can be used to enhance the results of the proposed classifiers. In this paper, Naive Bayes, Random Forest, and KNN are implemented to locate CKD. We also can compare and compare the overall performance of the used classifiers with other current classifiers. CKD early detection helps in the well-timed treatment of the sufferers affected by the sickness and also to avoid the ailment from getting worse. early prediction of the ailment and timely treatment are the need for the medical zone. New classifiers may be used and their performance may be evaluated to locate higher solutions of the objective feature in destiny paintings.

REFERENCES.

1. P.Swathi Baby, T.Panduranga Vital, "Statistical Analysis and Predicting Kidney Diseases using Machine learning algorithms", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.

2. K.R.Lakshmi, Y.Nagesh and M.Veerakrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering & Technology, Mar. 2014, Vol. 7, Issue 1, pp. 242-254.

3. AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page no 311 327

4. Dr.S.Vijaiyanti, Mr.s.Dayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms" ISSN 22296166, Volume 6 Issue 2 March 2015.

5. Mahfuzah Mustafa, Mohd Nasir Tab, et.al "Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image"Journal of Computer Science& Computational Mathematics, Volume 2, Issue 4, April 2012, pp 17-22.

6. Ross K K Leung, Ying Wang et.al."Using a multi-staged strategy based on machine learning and mathematical modelling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case-control cohort analysis"-BMC-Nephrology 2013 pp 1-9.

7. Bendi Venkata Ramana, "Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis-"International Journal of Database Management Systems(IJDMS), Vol.3, No.2, May 2011 (pp101 114).

8. M. Bramer, Principles of Data Mining: Springer-Verlag, 2007.

9. H.C.Koh and G.Tan."Data MiningAspects on Application on Healthcare" Journal of Healthcare Information Management, vol. 19, no. 2, 2005

10. DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao, Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis", Open Access Scientific Reports, Volume I, 2012.

11. Jinn-Yi Yeha, Tai-Hsi Wu et.al, "Using data mining techniques to predict hospitalization oh hemodialysis patients", Elsevier, Vol. 50 Issue 2, January 2011, Pages 439 448.

12. J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van Den Berghe K U Lauren" Data mining techniques for predicting acute kidney injury after elective cardiac surgery "Springer 2012

13. K.R.Lakshmi, Y.Nagesh and M.Veerakrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering & Technology, Mar. 2014

14. Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad M" Data Mining Performance in Identifying Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis 1,2013, pp 49-54.

15. Xudong Song, Zhanzhi Qui, Jianwei Mu, "Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Technology(IJACT), Volume4, Number3, February 2012.

16. N.Shriram, V. Natasha and H.Kaur", Approaches for Biology, Volume 06, Issue 02, June 2006.

17. Salha M.Alzahani, Afnan Authority et.al, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Taif University, Taif, Saudi Arabia 310-315.

18. Anu Chaudhary, Puneet Garg,(2014) Detecting and Diagnosing a Disease by Patient Monitoring System, International Journal of Mechanical Engineering And Information Technology, Vol. 2 Issue 6 //June //Page No: 493-499.

19. Fadzilah Siraj, Mansour Ali Abdallah, (2011). Mining Enrollment Data Using Descriptive and Predictive Mining.

20. George Dimitoglou, Comparison of the C4.5 and a Naive Bayes The classifier for the Prediction of Lung Cancer Survivability

21. Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orrù, Carlo Carcassi and Giorgio La Nasa, Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long-Term Kidney Transplantation Outcome, Chapter 5, an open-access article distributed under the terms of the Creative Commons Attribution License.

22. Gualtieri. J. A, Chettri. S. R, Cromp. R. F and Johnson.L. F, (1999) Support vector machine classifiers as applied to AVIRIS data, in Summaries 8th JPL Airborne Earth Science Workshop, JPL Pub. 99- 17, pp. 217 227.

23. Ian H. Witten and Eibe Frank. (2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition.

24. Suman Bala, Krishna Kumar," A Literature Review on Kidney Disease Prediction using Data Mining ClassificationTechnique" IJCSMC, Vol3 Issue. 7, July 2014, pg.960 967.

25. Neha Sharma, Hari Om, "Data Mining For predicting Oral cancer survivability" Springer-Verlag Wien 2013, Netw Model Anal Health Inform Bioinforma (2013) 2:285 295.

26. K. Machhale, H. B. Nandpuru, V. Kapur, and L. Kosta, "MRI brain cancer classification using hybrid classifier (SVM-KNN)," in 2015 International Conference on Industrial Instrumentation and Control (ICIC), Pune, 2015, pp. 60–65

27. LeenaVig, "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal, Volume 1 | e660,2014.

28. Breiman, L. (2001) Random forests. Mach. Learning, 45, 5–32.