Journal of Coastal Life Medicine
www.jclmm.com
ISSN: 2309-5288(Print)/2309-6152(Online)
Volume 8 No.3 (2020), Page No. 01 – 06
Article History: Received: 12 July 2020, Revised: 08 August 2020, Accepted: 18 August 2020, Publication: 30 September 2020

Predicting Diseases via Technology Incorporation

Gavin Ssebuuma
Department of Physiology, Iran

Abstract

Currently, some of the technical innovations that many scholarly investigators have embraced include learning algorithms, machine learning, predictive analytics, and big data analytics. The aim of these innovations has been to aid in useful data extraction for purposes of informed decision-making. Given that the big data section has seen predictive analytics emerge as a promising platform, with machine learning models incorporated, there has been a growing possibility of predicting the future behavior of parameters. In healthcare, this possibility has been felt in terms of disease prediction, as well as cure anticipation and development. The central purpose of this study was to apply a machine learning model on medical data sets in a big data environment. Specifically, the C4.5 algorithm was used towards chronic kidney disease prediction in a healthcare setting marked by big data presence.

Introduction

With the emergence of the big data concept, many conventional information technologies have failed to offer effective management of the large quantity of information. This challenge comes at a time when there is big data dominance in the world's key industries, including the healthcare sector. At each moment, thousands of data attributes are generated [1]. Some of the factors accounting for the big data evolution and dominance include sector digitization, the increasing use of mobile devices and the Internet, and the incorporation of an effective tool, information technology (IT). While there is growth in data speed, variety, and volume or quantity, however, challenges have emerged in terms of how best and effective it (the information) could be managed [2, 3].

The healthcare field has also played a leading role in shaping big data management outcomes. An example is a case in which through cloud computing, big data has paved the way for disease management and monitoring away from physical structures or hospitals [4, 5]. Specific examples of big data applications in healthcare include simple devices for physiological condition monitoring, smart homes for self-care among patients and families, and robotic and laparoscopic surgeries that have come in the place of classical surgeries [6-8]. Software or smart applications have also been incorporated into big data in a quest to analyze the signals that one's body generates, with sensors playing a specifically significant role [9]. For m-Health (mobile health) technologies, they have also been embraced towards collecting environmental, behavioral, and biological data that shapes healthcare operations [10, 11]. With information such as biometric data, lab results, diagnostic images, and other electronic medical records (EMRs) multiplying the data to yield big data in healthcare, this aspect of information explosion has prompted big data analysis procedures aimed at addressing healthcare issues, accomplished through early disease detection, improved treatment processes, hospital quality of service monitoring, and the provision of superior services to patients and their families [10].

Regarding the practices of most of the killer diseases' prediction and classification, many algorithms have been documented. Some of these diseases include diabetes, motor neuron issues, heart disease, and breast cancer [3, 8]. In this study, the main aim lay in the application of the C4.5 algorithm as a data mining framework towards the prediction of chronic kidney disease. The motivation of the study was to evaluate the performance and discern the superiority of such learning models towards disease prediction, eventually reflecting the application of big data in the healthcare sector.

Methodology

As mentioned above, this study employed a learning algorithm, C4.5, to determine and evaluate its capacity towards extracting information, predicting, and classifying data for patients to yield two categories of individuals: those without chronic kidney disease (the notckd group) and those with chronic kidney disease (the ckd group). In the experimental setup, the investigation involved the use of the Weka (Waikato Environment for Knowledge Analysis) approach, which is a Java suite with different models that support data mining and clustering, as well as results analysis, regression, and classification [7]. The choice of this platform was informed, particularly, by its capacity to allow researchers to realize an ideal surrounding in which classification models could be implemented and their performance evaluated. The results of Weka analysis were compared to those reported for ORANGE and TANAGRA, upon which the feasibility of using the selected algorithm towards disease prediction (chronic kidney disease) in such settings would be discerned.

The dataset constituted that which reflected chronic kidney disease. The UCI machine learning repository was the primary site from which data for chronic kidney disease was extracted. In this database or repository, there were 24 integer aspects, as well as 400 instances. Also, classes included notckd and ckd variables.

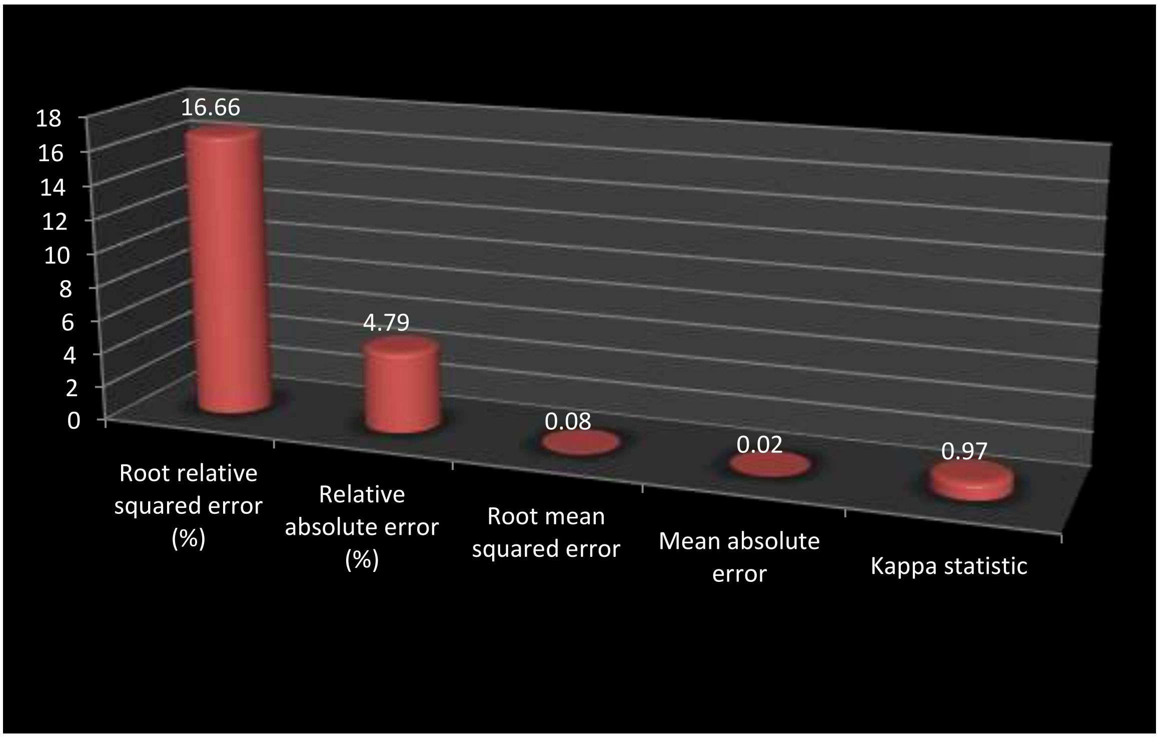
Attributes that were measured in particular, which aided in determining and understanding the performance of the selected C4.5 algorithm, included the false positive (FP), false negative (FN), true negative (TN), and true positive (TP). Indeed, FP was the number of samples that were predicted as positive wrongly, yet they were negative. Regarding the FN attributed, it represented positive samples, but were predicted wrongly. In relation to TN, it represented negative samples, which were also predicted correctly. Lastly, TR reflected positive samples that were also predicted correctly.

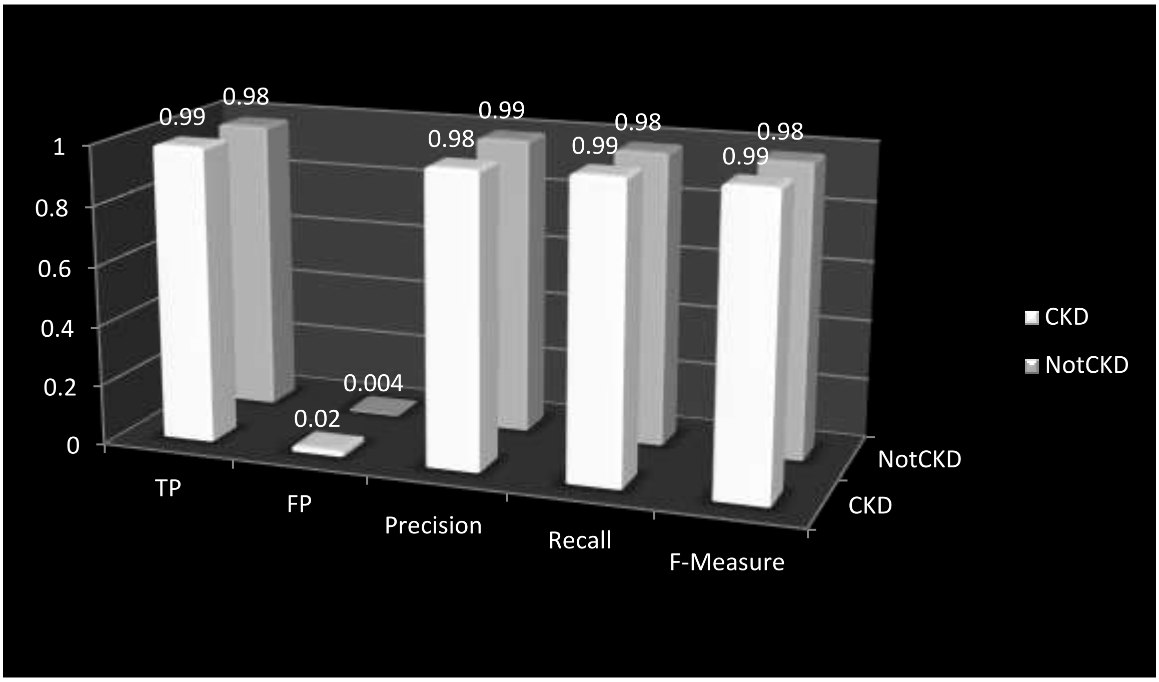
It is also worth indicating that the Mean Absolute Error was obtained via the comparison of eventual outcomes and the predictions or forecasts. On the other hand, the parameter of precision referred to positive predicted aspects. With regard to specificity, it referred to the negative prediction proportions that were identified correctly. For the case of the sensitivity parameter, it constituted positive prediction proportions that were identified correctly. Lastly, the accuracy aspect involved the number of predictions made correctly, relative to the predictions in the entirety. Indeed, these parameters aided in understanding the performance of the algorithm regarding chronic kidney disease prediction and classification.

Results

Given that the objective of the study lay in the evaluation of the performance of the classifier, after its implementation, this study relied on the 10-fold test for classification. The implication is that there were two sets of healthcare data. They included the test set and the training sample, the role of the training sample involved model training. With preparation and pre-processing techniques accomplished, there was a visual analysis of data, which strived to discern value distributions relative to the model's accuracy and performance. From the specific results, the error stood at 0.37, with the accuracy being 63% and the instances classified correctly being 396. Also, the time that was taken for model building was 0.08, the evaluation criterion being the selected C4.5 algorithm.

Indeed, the aforementioned values were obtained after evaluating the performance of the model. Another step involved establishing the simulation error. The figure below represents the findings that were obtained.





Based on the findings demonstrated above, this study confirmed the superiority of C4.5 algorithm as a classifier for the prediction of chronic kidney disease, especially in situations involving the big data phenomenon in healthcare. This inference was informed by the value of 396 as that which represented the algorithm's or model's instances that were classified correctly. As such, instances that were misclassified were only 4. It can also be seen that 0.37, a low value, is the error rate, adding to or confirming the model's superiority. Additional deductions indicate that C4.5 as a classifier exhibits excellent performance, as 0.97 is the KS value, implying that the classifier comes with superior accuracy and overall performance. In relation to precision, the findings demonstrate 0.99 for not ckd and 0.98 for ckd, outcomes that suggest superiority in relation to the precision of the model. Similar promising outcomes are observed when the parameter of recall is considered, which suggests 0.98 not ckd and 0.99 ckd.

Therefore, C4.5 as a prediction model was found to come with superior results or performance, demonstrating its capacity as a powerful classifier in relation to variables such as the minimum execution time and performance or classification accuracy. Therefore, it was inferred that the model, being a good classifier, is worth incorporating into and implementing in the healthcare sector, especially regarding disease prediction and classification. In this case, the classifier as a powerful model was demonstrated when it was applied to a big data scenario in the industry.

From the literature, it can be seen that this study contributes to and extend the work of various scholars, who have strived to implement and evaluate the performance of different machine learning techniques – to discern their applicability in real-life or practical situations. For instance, one of the works that this study extends is that which has focused on data mining techniques, data transformations, and data pre-processing to gain knowledge regarding patient survival [2]. Similar to the current study, such scholarly findings demonstrate that the machine learning techniques pose superior performance and come with beneficial outcomes such as the reduction of the effort and cost of selecting patients for clinical examinations, as the most important parameters discovered by the big data analytics and the prediction results from machine learning technique implementation aid in choosing patients. This study also extended works that have used RF and SVM algorithms as machine learning techniques [3-5]. For the latter investigations, focus has been on the comparison, studying, and classification of data sets for conditions such as heart disease, liver disease, and cancer with different kernel parameters and kernels. From the studies' findings, it has been demonstrated that algorithms such as SVM and RF aid in proper parameter selection when applied to data sets such as the diseases mentioned above, aiding in saving time and costs while emerging as superior classifiers that could aid in informed decision-making about disease diagnoses and severities, as well as information about relevant early interventions that are worth adopting. Thus, this study, which implemented C4.5 algorithm and evaluated its persons, contributed to the current state-of-the-art regarding the implementation of machine learning methods in the healthcare sector, especially in the wake of the dominance of the bug data concept.

Conclusion

In summary, data mining techniques continue to receive growing scholarly attention, as well as industry application in the real-world, the healthcare sector unexceptional. These techniques play an important role in such a way that they aid in achieving predictive analyses. In the healthcare industry for example, the predictive analyses achieved through the implementation of the data mining techniques allow for the provision of early disease prediction, diagnosis, and intervention, having proved to be powerful classifiers and prediction models or algorithms. The implication is that these techniques pave the way for individuals and families to anticipate cures, besides playing supportive roles to the healthcare terms, whereby they support towards informed decision-making. In this study, C4.5 was implemented as the learning algorithm. The model was used towards the prediction of chronic kidney disease in individuals. These individuals were classified under the ckd category. Also, the algorithm was applied to classify individuals not suffering from chronic kidney disease. The latter group was classified under notckd. From the results, the selected classifier exhibited superior performance relative to disease prediction and classification. In particular, the experimental studies witnessed the algorithm yield superior results on when tested on specific parameters such as the minimum execution time and accuracy. Overall, the study is deemed insightful and contributory to the healthcare field, whereby it paves the way for effort increase or increased focus towards the establishment and implementation of machine learning techniques to classify and predict disease occurrence, upon which early interventions could be spearheaded. The resulting inference is that due to its superior performance as a classifier, C4.5 algorithm is seen to aid in informed decision-making among healthcare team members, having exploited intelligent data in an environment marked by the big data attribute.

References

[1] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) "Predicting survival time for kidney dialysis patients: a data mining approach", Elsevier Publication, Computers in Biology and Medicine ,Vol.35, pp 311–327

[2] Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain, (2013) “Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques”, International Journal of Computer Applications Vol. 69, No.11, pp 12-16

[3] Sadik Kara, Aysegul Guvenb, Ayse OztUrk Onerc, (2006) “Utilization of artificial neural networks in the diagnosis of optic nerve diseases”, Elsevier Publication, Computers in Biology and Medicine,Vol. 36, pp 428–437

[4] M Hall, E Frank, G Holmes, B Pfahringer,(2009), ‘The WEKA data mining software: an update’, Volume 11, Issue 1, pp 10-18

[5] R. Weil, (2014),’’ Big Data In Health: A New Era For Research And Patient Care Alan R. Weil’’, Health Affair, Vol. 33, N° 7, pp 1110.

[6] Peter Groves; Basel Kayyali, (2013),’’ The ‘big data’ revolution in healthcare’’, McKinsy and Company. Center for US Health System Reform Business Technology Office. Available at http://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf.

[7] T., Huang, L., Lan, (2015), ‘’Promises and Challenges of Big Data Computing in Health Sciences’’, Big Data Research vol. 2, pp 2-11 available at http://dx.doi.org/10.1016/j.bdr.2015.02.002

[8] Khurshid R., G., Kai, Z., John T., W., and Charles P., F., (2014), ‘’Harnessing Big Data for Health Care and Research Are Urologists Ready? ", Journal of European Urology, vol. N., pp 1-3

[9] Wullianallur Raghupathi,Viju Raghupathi, (2014),’’Big data analytics in healthcare: promise and Potential’’, Health Information Science and Systems. Available at http://www.biomedcentral.com/content/pdf/2047-2501-2-3.pdf

[10] Rashedur M. Rahman, Fazle Rabbi Md. Hasan ‘Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data’’, Elsevier, Vol. 38, pp 11421–11436

[11] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005),’’ Predicting survival time for kidney dialysis patients: a data mining approach’’, Elsevier Publication, Computers in Biology and Medicine, Vol. 35, pp 311–327