IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Predicting Chronic Kidney Disease Classification Using Hybrid Machine Learning Technique"

Kuldeep Patil, Ayush Parmar, Yogiraj Phalke, Raj Patil

(Trinity College Of Engineering & Research, Pune)

Dr. Sujeet More*

Abstract

Chronic kidney disease (CKD) has become a widespread disease nowadays. It is associated with several serious risks including cardiovascular disease, increased risk, and end-stage renal disease and may be preventable through early detection and treatment of those at risk for this disease. Learning algorithms are an important aid to medical professionals in accurately diagnosing disease in its early stages. More recently, big data platforms have been integrated with machine learning algorithms to add value to healthcare. Therefore, in this paper we propose a hybrid machine learning approach that includes a feature selection method and a machine learning classification algorithm based on a big data platform that has been used for chronic kidney disease (CKD) detection.

INTRODUCTION

The present, especially the last 20 years, can be called the era of big data, where digital data is becoming more and more important in various fields such as science, medicine, technology, and society. A huge amount of data is generated and generated by multiple sensor networks and mobile applications in almost every field, especially in healthcare, and this huge amount of data is called big data. The multitude of data sources such as streaming machines, high-end output devices, visualization and knowledge extraction of these vast and diverse data types presents a significant challenge when state-of-the-art technology and tools are underutilized. One of the major technical challenges in big data analytics is finding suitable methods to effectively retrieve useful and relevant information for different categories of users. Various forms and types of data sources in healthcare are currently being collected in both clinical and non-clinical settings. There, the most important data in health analysis is a digital copy of the patient's medical history. As such, the process of designing and building distributed data systems to process big data poses three main challenges. The first challenge is that the data is diverse and massive, making it difficult to collect data from distributed locations. The second challenge is that big data systems require storage with guaranteed performance, so storage becomes a major issue for large heterogeneous datasets. A third challenge, more closely related to big data analytics, especially mining large datasets in real time, involves visualization, prediction, and optimization.

LITERATURE REVIEW

Many authors have used various ML techniques for diagnosis and prediction of chronic kidney disease. They compared the proposed model with

six ML algorithms: LR, RF, SVM, KNN, Naive Bayes, and Feedforward Neural Network (FNN). Their proposed model recorded the highest accuracy of 99.83%. NB, K-Star, SVM, and J48 classifiers were used to predict CKD. Performance comparisons were performed using WEKA software. The J48 algorithm outperformed the other algorithms with 99% accuracy. Some authors used ML algorithms and feature selection methods to predict CKD. [22] used the feature selection method of recursive feature elimination (RFE) to select essential features from the chronic kidney disease (CKD) dataset. Four classification algorithms (SVM, KNN, DT, and RF) were applied to both complete and selected features. Results showed that RF outperformed all other algorithms. In [20], the authors used chi-square, CFS, and lasso feature selection to select significant features from the database. They applied ANN, C5.0, LR, LSVM, KNN, and RF to both full and selected features.

The results showed that the full-featured LSVM recorded the highest accuracy at 98.86%. Used five feature selection methods: random forest feature selection (RF-FS), forward selection (FS), forward depletion selection(FES), backward selection (BS), and backward depletion (BE) the most important features from the database. We used four ML algorithms, RF, SVM, NB, and LR, to predict CKD. Results showed that RF with random forest function selection performed best with 98.8% accuracy. [26] used a genetic search algorithm to select the most significant traits from the CNI dataset. Decision tables, J48, multilayer perceptron (MLP), and NB were applied to both full and selected features. Using the genetic search algorithm improved performance. The MLP classifier performed the best and outperformed the other classifiers using correlation-based feature selection (CFS) to select the number of significant features. Detected CKD using AdaBoost, KNN, NB, and SVM. The proposed CFS with AdaBoost achieved the best performance with 98.1% accuracy. In [25], the author predicted his CKD using two ensemble methods, namely the bagging method and the random subspace method, and he used three basic learners, ANN, NB, and DT.

RESEARCH METHODOLOGY

The proposed chronic kidney disease prediction system consists of two main approaches. The first approach uses feature selection methods to select important features from the chronic kidney disease dataset. The second approach applies ML techniques (LR, RF, SVM) to selected and complete features to predict CKD. The proposed system consists of six steps. In the first step (data collection) his CKD dataset from the UCI machine learning repository is used. The second step, the data pre-processing step, handles null values. The third step is to select important features using feature methods. In the fourth step, grid search with stratified cross-validation is used to optimize the ML parameters and ensemble learning method. Each step is in detail in the following subsections.

3.1. Data collection. The Chronic Kidney Disease (CKD) dataset used in this study was provided by the UCI Machine Learning Repository. The CKD data set contains 400 specimens, 24 traits, and 1 class designation. The dataset contains 400 samples. The class label has two values: ckd (example with CKD) and notckd (example without CKD).

3.2.Data pre-processing. The dataset contained outliers and noise. So it should be cleaned and cleaned in the pre-treatment stage. In the pre-processing phase, we estimated missing values and removed noise such as outliers, normalization, and imbalanced data checks. This is because certain measurements may be lost when examining patients, resulting in missing values. The dataset has 158 closed cases and the remaining occurrences have missing values. Ignoring datasets is the easiest way to handle missing values, but this strategy is ineffective for small datasets. Another approach is to apply an algorithm to extrapolate missing data instead of deleting records. Missing values in nominal properties were filled by mode. Missing values in numerical characteristics were imputed with the mean.

RESULTS

This section describes the results of applying chi-square and Bump F to a dataset to select the most important features. We also describe cross validation performance and test results applying ML algorithms SVM, LR, NB, RF, DT and GBT classifiers to full and selected features. It also shows the optimal parameter valves for each ML algorithm optimised by Grid Search. Two features selection methods were used. The CKD dataset was split into an 80% training set and a 20% testing set. Cross-validation results were registered in the training set and test results were registered in the test set. ML algorithms and function selection methods were implemented using PySpark.

TABLE 2: The CKD dataset description.

Features	Explain
age	Age
bp	Blood pressure
sg	Specific gravity
al	Albumin
su	Sugar
rbc	Red blood cells
pc	Pus cell
pcc	Pus cell clumps
ba	Bacteria
bgr	Blood glucose random
bu	Blood urea
sc	Serum creatinine
sod	Sodium
pot	Potassium
hemo	Hemoglobin
pcv	Packed cell volume
wc	White blood cell count
rc	Red blood cell count
htn	Hypertension
dm	Diabetes mellitus
cad	Coronary artery disease
appet	Appetite
pe	Pedal edema
ane	Anemia
class	Class

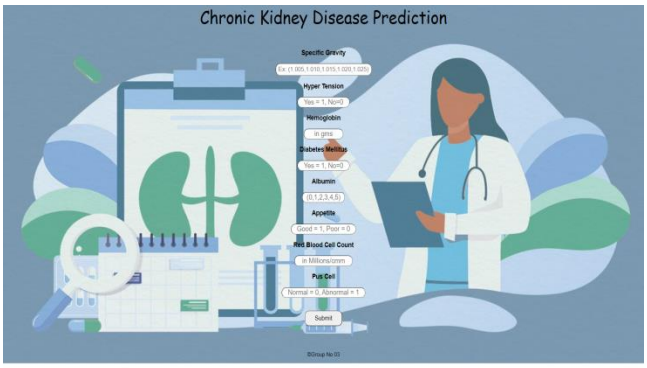

Fig 1: Showing the implementation of the proposed model.

Fig 2: showing the output result of the diagnosis.

CONCLUSION

In this post, we used a hybrid ML technique that integrates a feature selection technique based on a big data platform and a classification ML algorithm to predict CKD. Feature selection techniques were used to select significant features from the dataset. ML algorithms LR, NB, RF, SVM and GBT classifiers ensembles as learning algorithms were applied to the benchmark chronic kidney disease dataset. Moreover, they are applied to all features and selected features. Grid search with cross-validation was used to optimize the parameters of ML. Moreover, we applied four evaluation methods: accuracy, precision, recall, and F1 measures to validate the results and then register the cross-validation results and test data. Results showed that SVM, DT, and GBT classifiers performed best on the selected features. Overall, the performance of selection is better than that achieved by other feature selection.

REFERENCES

[1] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," International Journal of Advanced Intelligence Paradigms, vol. 10, no. 1-2, pp. 118–132, 2018.

[2] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: a technology tutorial," IEEE access, vol. 2, pp. 652–687, 2014.

[3] "Apache Hadoop," 2021, https://hadoop.apache.org/.

[4] A. Kafka, "Apache Kafka," 2021, https://kafka.apache.org/.

[5] A. Storm, "Apache Storm," 2021, https://storm.apache.org/.

[6] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," JAMA, vol. 319, no. 13, pp. 1317-1318, 2018.

[7] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark based machine learning model on streaming big data for health status prediction," Computers & Electrical Engineering, vol. 65, pp. 393–399, 2018.

[8] A. A. Ali, "Stroke prediction using distributed machine learning based on Apache spark," Stroke, vol. 28, no. 15, pp. 89–97, 2019.

[9] World Health Organization, Public Health Agency of Canada, and Canada. Public Health Agency of Canada, Preventing Chronic Diseases: A Vital Investment, World Health Organization, Geneva, Switzerland, 2005.

[10] B. Bikbov, N. Perico, and G. Remuzzi, "Disparities in chronic kidney disease prevalence among males and females in 195 countries: analysis of the global burden of disease 2016 study," Nephron, vol. 139, no. 4, pp. 313–318, 2018.

[11] K. Disease, "Improving global outcomes (kdigo) transplant work group. kdigo clinical practice guideline for the care of kidney transplant recipients," American Journal of Transplantation, vol. 9, no. 3, pp. S1–S155, 2009.

[12] Cdc, "Chronic Kidney Disease in the united states," 2021, https://www.cdc.gov/kidneydisease/publications-resources/c kd-national-facts.html.

[13] L. Deng and X. Li, "Machine learning paradigms for speech recognition: an overview," IEEE Transactions on Audio Speech and Language Processing, vol. 21, no. 5, pp. 1060–1089, 2013.

[14] M. Q. Huang, J. Nini´c, and Q. B. Zhang, "Bim, machine learning and computer vision techniques in underground construction: current status and future perspectives," Tunnelling and Underground Space Technology, vol. 108, Article ID 103677, 2021.

[15] P. Oza, P. Sharma, and S. Patel, "Machine learning applications for computer-aided medical diagnostics," in Proceedings of the Second International Conference on Computing, Communications, and Cyber-SecuritySpringer, New York, NY, USA, 2021.

[16] T. Bikmukhametov and J. J¨aschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," Computers & Chemical Engineering, vol. 138, Article ID 106834, 2020.

[17] R. J. Palma-Mendoza, D. Rodriguez, and L. D. Marcos, "Distributed relieff-based feature selection in spark," Knowledge and Information Systems, vol. 57, no. 1, pp. 1–20, 2018.

[18] M. Nassar, H. Safa, A. A. Mutawa, A. Helal, and I. Gaba, "Chi squared feature selection over Apache spark," in Proceedings of the 23rd International Database Applications & Engineering Symposium, pp. 1–5, Athens Greece, June 2019.

[19] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in Proceedings of the 2016 management and innovation technology international conference (MITicon), Bang-San, Thailand, October 2016.