Prediction of Chronic Kidney Disease Using Machine Learning Methodologies

Keerthana[1], Mageshwari[2], Malathi, M.Tech(Ph.d)[3], Balaji,M.E(Ph.d)[4]
[1, 2] Dept of Computer Science
[3, 4]Assistant Professor, Dept of Computer Science
[1, 2, 3, 4] Anand Institute Of Higher Technology

Abstract- *Chronic Kidney Disease (CKD) is one of the deadliest diseases that slowly damages human kidney. The disease remains undetected in its early stage and the patients can only realize the severity of the disease when it gets advanced. Hence, detecting such disease at earlier stage is a key challenge now. Machine Learning is one of the emerging field used in the health sectors for the diagnosis of different diseases. In this paper, we compute, analyze and compare between Machine Learning classification approaches to determine which classification approach is the optimal for the prediction of CKD. Random Forest Algorithm and Logistic Regression are some renowned machine learning methods which were selected to train the model and based on these results, we can compare and determine which among the following Machine Learning Methods can predict the possibility of CKD at the most accurate level. From this comparative analysis, Random Forest Algorithm is found to be the best approach to predict CKD. Methods can predict the possibility of CKD at the most accurate level. From this comparative analysis, Random Forest Algorithm is found to be the best approach to predict CKD.*

Keywords- Machine Learning, Classification Technique, Prediction System

I. INTRODUCTION

Kidney disease is taken into account a significant drawback for individuals sixty and on top of. the most important cause is that the degeneration of the excretory organ that reduces the speed of capillary filtration. This drawback, once lasting over 3 months, is mostly thought-about as chronic renal disorder (CKD). CKD is hierarchical because the tenth major reason behind death within the world. high blood pressure, diabetes, and aging ar thought-about leading causes of CKD, additionally to different factors like high vital sign, artery malady, and anemia. If the matter are often detected in early stages, then it's thought-about possible to save lots of excretory organ operate for the longer survival of the patient. Early designation of CKD will facilitate its treatment avoid expensive treatment procedures like chemical analysis and transplants. With machine learning techniques, it's attainable to investigate science lab records and different data on patients for the first detection of CKD. Low-level knowledge are often reworked into high-level data through the data discovery in databases (KDD). This transformation will facilitate practitioners higher perceive CKD patterns for its early designation. This study analyzes CKD victimisation machine learning techniques employing a CKD knowledge set from the Kaggle machine learning data warehouse. the foremost relevant options ar hand-picked from the dataset to boost accuracy and cut back coaching time for machine learning techniques. a group of experiments is conducted victimisation numerous WEKA-implemented machine learning techniques to observe CMD supported the CKD dataset from the Kaggle machine. The results ar compared for detection accuracy across completely different machine learning techniques. the remainder of this paper is organized as follows: Section two describes machine learning techniques. Section three presents the state of the art within the field for CKD detection. Section four presents the projected work for investigation machine learning techniques. Section five reports the results, and Section half-dozen concludes with some avenues for future analysis.

II. PROBLEM DEFINITION

The first class is classification; a classification drawback is outlined when the output variable could be a class, like “malignant” or “benign”, “disease”, or “no disease”. The second class is regression, that is outlined when the output variable could be a numerical worth. supervised learning algorithms try to model relationships and dependencies between the target prediction output and therefore the input feature such the output values for brand new knowledge is predicted supported those relationships that it learned from the previous knowledge sets.

Sg	dm	htn	hemo	pcv
1.02	Yes	Yes	15.4	44
1.02	No	No	11.3	38
1.01	Yes	No	9.6	31
1.005	No	Yes	11.2	32
1.01	No	No	11.6	35
1.015	Yes	Yes	12.2	39
1.01	Yes	No	12.4	36
1.015	Yes	No	12.4	44
1.02	Yes	Yes	10.8	33
1.01	Yes	Yes	9.5	29
1.01	Yes	Yes	9.4	28
1.015	Yes	Yes	10.8	32
1.01	Yes	Yes	9.7	28
1.01	Yes	Yes	9.8	
1.015	No	No	5.6	16
1.015	No	Yes	7.6	24
1.02	No	Yes	12.6	
1.025	Yes	Yes	12.1	
1.015	No	Yes	12.7	37

III. PREDICT THE CHRONIC KIDNEY DISEASE METHODOLOGIES

DATA COLLECTION

Data collection is the process of gathering and measuring information from Kaggle countless different sources. In order to use the data we collect to develop practical machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand. Data for which you already know the target answer is called labelled data.

DATA PRE-PROCESSING

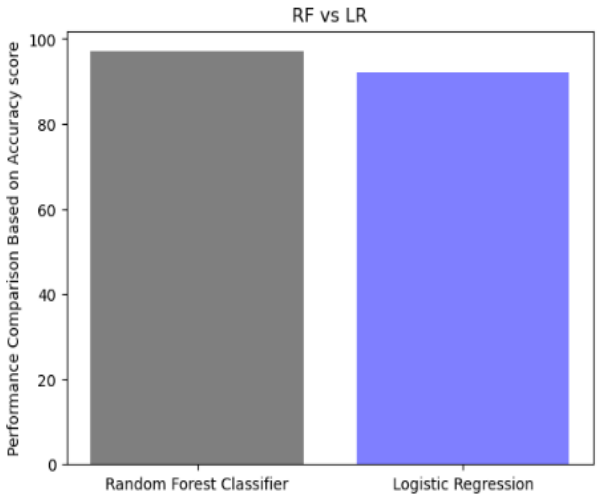
The data preprocessing involves cleaning, extracting, filling missing values and transformation to suitable formats. The cleaning is done through some standard tools available like WEKA. Below Table show how the missing values are filled. We are using the label encoder and min-max algorithm to convert the nominal value to numeric value and bring all the attributes to the same scale respectively for better training.

DATA ANALYSIS

One of the first steps we perform during implementation is an analysis of the data. This was done by us in an attempt to find the presence of any relationships between the various attributes present in the dataset. Acquisition of Training Dataset: The Training data set was acquired from the Kaggle Repository for kidney disease. The dataset was collected from a number of hospitals from Tamil Nadu and the values are actual test results values that were obtained. We have a total of 24 attributes which make up the dataset but on pre-processing it was found that only 6 of the 24 are important in determining that relationship. There are a total of 400 samples. This way the predictions can be used to correctly determine early detection of Chronic Kidney Disease.

EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.



IV. CONCLUSION

This system presented the best prediction algorithm to predict CKD at an early stage. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters. Logistic Regression and Random Forest Algorithm models are constructed to carry out the diagnosis of CKD. The performance of the models is evaluated based on a variety of

comparison metrics are being used, Accuracy. The results of the research showed that Random Forest Algorithm model better predicts CKD in comparison to the Logistic Regression model taking all the metrics under consideration. This system would help detect the chances of a person having CKD further on in his life which would be really helpful and cost-effective people. This model could be integrated with normal blood report generation, which could automatically flag out if there is a person at risk. Patients would not have to go to a doctor unless they are flagged by the algorithms. This would make it cheaper and easier for the modern busy person.

REFERENCES

[1] Z. Chen, Z. Zhang, R. Zhu, Y. Xiang and P. B. Harrington, "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers", Chemometrics Intell. Lab. Syst., vol. 153, pp. 140-145, Apr. 2016.

[2] A. Subasi, E. Alickovic and J. Kevric, "Diagnosis of chronic kidney disease by using random forest", Proc. Int. Conf. Med. Biol. Eng., pp. 589-594, Mar. 2017.

[3] L. Zhang, "Prevalence of chronic kidney disease in China: A cross-sectional survey", Lancet, vol. 379, pp. 815-822, Mar. 2012.

[4] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration", J. Biomed. Informat., vol. 53, pp. 220-228, Feb. 2015.

[5] A. M. Cueto-Manzano, L. Cortés-Sanabria, H. R. Martínez-Ramírez, E. Rojas-Campos, B. Gómez-Navarro and M. Castillero-Manzano, "Prevalence of chronic kidney disease in an adult population", Arch. Med. Res., vol. 45, pp. 507-513, Aug. 2014.

[6] H. Polat, H. D. Mehr and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods", J. Med. Syst., vol. 41, no. 4, pp. 55, Apr. 2017.

[7] C. Barbieri, F. Mari, A. Stopper, E. Gatti, P. Escandell-Montero, J. M. Martínez-Martínez, et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis", Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

[8] V. Papademetriou, E. S. Nylen, M. Doumas, J. Probstfield, J. F. Mann, R. E. Gilbert, et al., "Chronic kidney disease basal insulin glargine and health outcomes in people with dysglycemia: The ORIGIN Study", Amer. J. Med., vol. 130, no. 12, pp. 1465.e27-1465.e39, Dec. 2017.

[9] N. R. Hill, "Global prevalence of chronic kidney disease—A systematic review and meta-analysis", PLoS ONE, vol. 11, no. 7, Jul. 2016.

[10] M. M. Hossain, R. K. Detwiler, E. H. Chang, M. C. Caughey, M. W. Fisher, T. C. Nichols, et al., " Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts ", IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 66, no. 3, pp. 551-562, Mar. 2019.

[11] M. Alloghani, D. Al-Jumeily, T. Baker, A. Hussain, J. Mustafina and A. J. Aljaaf, "Applications of machine learning techniques for software engineering learning and early prediction of students' performance", Proc. Int. Conf. Soft Comput. Data Sci., pp. 246-258, Dec. 2018.

[12] D. Gupta, S. Khare and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques", Proc. Int. Conf. Comput. Commun. Autom. (ICCCA), pp. 281-287, Apr. 2016.

[13] L. Du, C. Xia, Z. Deng, G. Lu, S. Xia and J. Ma, "A machine learning based approach to identify protected health information in Chinese clinical text", Int. J. Med. Informat., vol. 116, pp. 24-32, Aug. 2018.

[14] R. Abbas, A. J. Hussain, D. Al-Jumeily, T. Baker and A. Khattak, "Classification of foetal distress and hypoxia using machine learning approaches", Proc. Int. Conf. Intell. Comput., pp. 767-776, Jul. 2018.

[15] M. Mahyoub, M. Randles, T. Baker and P. Yang, "Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance", Proc. 11th Int. Conf. Develop. eSyst. Eng. (DeSE), pp. 1-11, Sep. 2018.