Utilitas Mathematica

ISSN 0315-3681 Volume 122, 2025

A Machine Learning Methodology for Diagnosing Chronic Kidney Disease

[1] Rohini Priya Gaadasu

PG Scholar, Department of Computer Science & Engineering, Siddhartha Institute of Technology and Sciences, Hyderabad, India.

[2] Mrs. Yamini Chawhan

Assistant Professor, Department of Computer Science & Engineering, Siddhartha Institute of Technology and Sciences,

ABSTRACT

Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it induces other diseases. Since there are no obvious symptoms during the early stages of CKD, patients often fail to notice the disease. Early detection of CKD enables patients to receive timely treatment to ameliorate the progression of this disease. Machine learning models can effectively aid clinicians achieve this goal due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for diagnosing CKD. The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. Random forest imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Missing values are usually seen in real-life medical situations because patients may miss some measurements for various reasons. After effectively filling out the incomplete data set, Five machine learning algorithms (logistic regression, random forest, support vector machine, decision tree, ,naïve bayes, KNN and gradient boost classier) were used to establish models. Also the stage of the disease is also predicted according to the age of the person.

1. INTRODUCTION

CHRONIC kidney disease (CKD) is a global public health problem . This disease is characterised by a slow deterioration in renal function, which eventually causes a complete loss of renal function. CKD does not show obvious symptoms in its early stages. Therefore, the disease may not be detected until the kidney loses about 25% of its function . In addition, CKD has high morbidity and mortality, with a global impact on the human body.Machine learning refers to a computer program, which calculates and deduces the information related to the task and obtains the characteristics of the corresponding

pattern . This technology can achieve accurate and economical diagnoses of diseases; hence, it might be a promising method for diagnosing CKD.We used Random forest imputation to fill in the missing values in the data set, which could be applied to the data set with the diagnostic categories are unknown. Logistic regression (LOG), RF were used to establish CKD diagnostic models on the complete CKD data sets. Random forest imputation is used to fill in the missing values. To our knowledge, this is the first time that KNN imputation has been used for the diagnosis of CKD. In addition, building an integrated model is also a good way to improve the performance of separate individual models.

ALGORITHM:

Random Forest Classifier: Random forest is a most popular and powerful supervised machine learning algorithm capable of performing both classification, regression tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The more trees in a forest the more robust the prediction.

Logistic Regression:

Logistic Regression is a simple yet powerful algorithm used for binary classification. It models the probability that a given input belongs to a certain class. It's often used when the relationship between the features and the target is approximately linear.

2. LITERATURE REVIEW

There are many researchers who work on prediction of CKD with the help of many different classification algorithm. And those researchers get expected output of their model.

Gunarathne W.H.S.D et.al. Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes. S.Ramya and Dr.N. worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD according to its gravity. By analysing different algorithms like Basic Propagation Neural Network, RBF and RF. The analysis results indicates that RBF algorithm gives better results than the other classifiers and produces 85.3% accuracy. \

S.Dilli Arasu and Dr. R. Thirumalaiselvi has worked on missing values in a dataset of chronic Kidney Disease. Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing so they got up with unknown values. They replaced missing values with recalculated values.

Asif salekin and john stankovic they use novel approach to detect CKD using machine learning algorithm. They get result on dataset which having 400 records and 25 attributes which gives result of patient having CKD or not CKD. They use k-nearest neighbours, random forest and neural network to get results. For feature reduction they use wrapper method which detect CKD with high accuracy.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM:

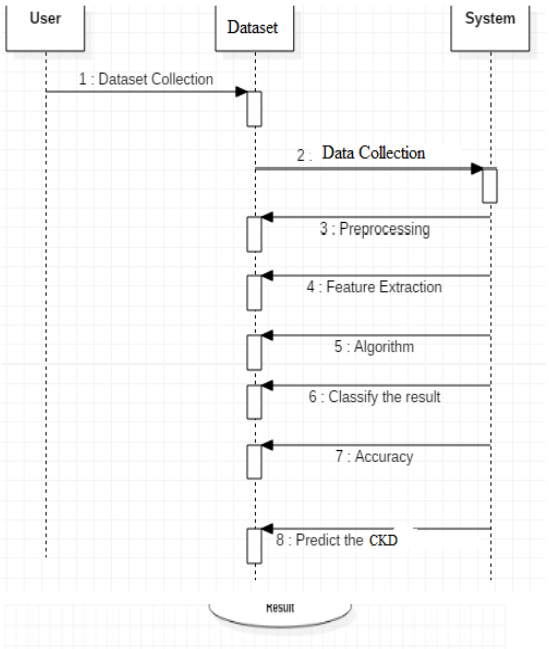
Gunarathne W.H.S.D et.al. Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes. S.Ramya and Dr.N.Radha worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD according to its gravity.

3.2 PROPOSED SYSTEM:

- In this paper CKD dataset is downloaded from UCI repository. This dataset includes 400 patients' records with 25 attributes. All this 25 attributes are main attributes which are related to CKD disease. Out of attributes we only use some attributes to build our predictive model.
- At first, we collected the data and pre-process it. After pre-processing the data, we handled the missing data of the dataset using Anaconda. Feature selection is conducted to extract the most significant features. Then, we apply Random Forest algorithm on the dataset more.

4. SYSTEM DESIGN

4.1 USE CASE DIAGRAM



4.2 SEQUENCE DIAGRAM

5. IMPLEMENTATION

5.1 DATA COLLECTION

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

5.2 DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem.

6. CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. The severity of the disease is found but actual severity calculation requires gender and race of the persons from which e GFR value is calculated that identifies the stage of the disease. In the future, a large number of more complex and representative data will

be collected to train the model to improve the generalization performance while enabling it to detect the exact severity of the disease.

7. REFERENCES

I. 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.

II. 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.

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

IV. 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.

V. 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.