International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 138-142 www.ijaem.net ISSN: 2395-5252

Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries with Machine Learning Techniques

D.Premchandu[1], Mr. Praveen[2], K. Triveni[3], G. Gnanasri[4], M. Mounisha[5]

[2]*Assistant Professor,Teegala Krishna Reddy Engineering College,Hyderabad, Telangana.*
[1,3,4,5]*Student, Teegala Krishna Reddy Engineering College, Hyderabadd, Telangana*

Submitted: 25-06-2022 Revised: 01-07-2022 Accepted: 06-07-2022

ABSTRACT:

The high incidence and prevalence of chronic kidney disease (CKD), often caused by latediagnoses, is a critical public health problem, especially in developing countries such as Brazil. CKDtreatment therapies, such as dialysis and kidney transplantation, increase the morbidity and mortalityrates, besides the public health costs. This study analyses the usage of machine learning techniques toassistintheearlydiagnosisofCKDindevelopingcountries.Qualitativeandquantitativecomparativeanalyses are, respectively, conducted using a systematic literature review and an experiment withmachinelearningtechniques,withthek-foldcross-validationmethodbasedontheWekasoftwareanda CKD dataset. These analyses enable a discussion on the suitability of machine learning techniquesforscreeningforCKDrisk,focusingonlow-incomeandhard-to-reachsettingsofdevelopingcountries, due to the specific problems faced by them, e.g., inadequate primary health care. The studyresults show that the J48 decision tree is a suitable machine learning technique for such screening indevelopingcountries,duetotheeasyinterpretationof itsclassificationresults,with95.00%accuracy,reaching a nearly perfect agreement with an experienced nephroligists opinion. Conversely, randomforest,naiveBayes,supportvectormachine,multilayerperceptron,andk-nearestneighbortechniques,respectively, yield 93.33%, 88.33%, 76.66%, 75.00%, and 71.67% accuracy, presenting at leastmoderateagreementwiththe nephrologist,atthe costofa moredifficultinterpretationoftheclassificationresult.

I. INTRODUCTION

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

CHRONIC DISEASEINTRODUCTION

CHRONIC kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population. According to another study, this percentage has reached 14.7% in the Mexican adult general population. 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. It can induce the occurrence of cardiovascular disease.CKD is a progressive and irreversible pathologic syndrome. Hence, the prediction and diagnosis of CKD in its early stages is quite essential, it may be able to enable patients to receive timely treatment to ameliorate the progression of the disease. Machine learning refers to a computer program, which calculates and deduces the information related to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 138-142 www.ijaem.net ISSN: 2395-5252

the task and obtains the characteristics of the corresponding pattern.

II. WORKING PRINCIPLE

DataCollection:

As a society, we're generating data at an unprecedented rate. These data can be numeric,(temperature, loan amount, customer retention rate), categorical (gender, color, highest degreeearned), or even free text (think doctor's notes or opinion surveys). Data collection is the processof gathering and measuring information from countless different sources. In order to use the datawe collect to develop practical artificial intelligence (AI) and machine learning solutions, it mustbe collected and stored in a way that makes sense for the business problem at hand. Collectingdataallowsyoutocapturearecordofpasteventssothatwecanusedataanalysistofindrecurringpatterns.Fromthosepatterns,youbuildpredictivemodelsusingmachinelearningalgorithmsthatlook for trends and predict future changes.

DataPreprocessing

Real-worldrawdataandimagesareoftenincomplete,inconsistentandlackingincertainbehaviors or trends. They are also likely to contain many errors. So, once collected, theyare pre- processed into a format the machine learning algorithm can use for the model. Pre-processing includes a number of techniques and actions: Data cleaning. Thesetechniques,manual and automated, remove data incorrectly added or classified.

Datanormalization:

The size of a dataset affects the memory and processing required for iterations duringtraining.Normalizationreducesthesizebyreducingtheorderandmagnitudeofdata.Thosetechniquespointtothe typesof machinelearningavailable tomobileappdevelopers.

Splittingthe data:

Data splitting is commonly used in machine learning to split data into a train, test, orvalidation set. This approach allows us to find the model hyper-parameter and also estimate thegeneralization performance. In this project, we conducted a comparative analysis of differentdatapartitioning algorithms on both real and simulated data.

MachineLearning:

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as trainingdata, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications.

III. SOFTWARE DETAILS

1.Block diagram

2.Decision Tree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Types of Decision Trees

Types of decision trees are based on the type of target variable we have. It can be of two types: • Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable decision tree. • Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 138-142 www.ijaem.net ISSN: 2395-5252

Important Terminology related to Decision Trees:

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets. • Splitting: It is a process of dividing a node into two or more sub-nodes.

- Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.
- Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.
- Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
- Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.
- Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

Working of Decision Trees:

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees. Decision trees use multiple algorithms to decide to split a node into two or more subnodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

ID3 → (extension of D3) C4.5 → (successor of ID3) CART → (Classification And Regression Tree) CHAID → (Chi-square automatic interaction detection Performs multi-level splits when computing classification trees) MARS → (multivariate adaptive regression splines) The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking. A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment.

3. K-Nearest Neighbor (KNN) Algorithm for Machine Learning:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

Working of K-NN: The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors.

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category. Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready. Suppose we have a new data point an

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

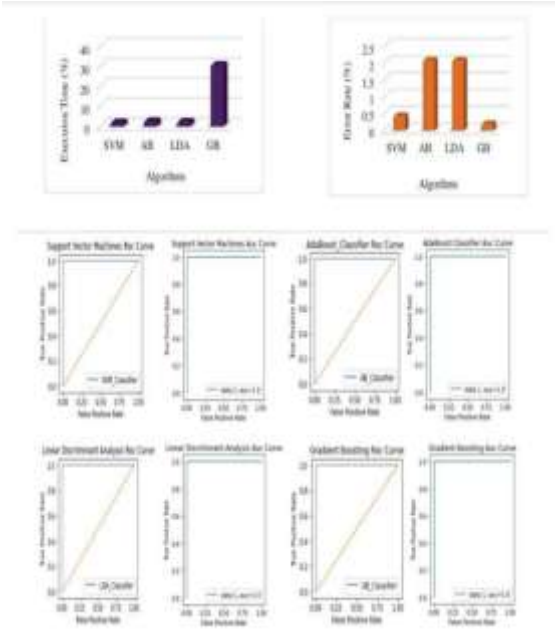
Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points.

IV. RESULTS

STAGE I RESULT:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 138-142 www.ijaem.net ISSN: 2395-5252



ROC curve is diagnosed because the receiver working characteristic curve in which AUC is the vicinity under the ROC curve. If the rating of AUC is excessive, the performance of the version must be excessive, and vice versa. The ratings of Support Vector system, Linear Discriminant Analysis, and Gradient Boosting Classifier provide the best rating all of them in each ROC and AUC curves. Support Vector Machine that represents fourth model in the carve and provides 1.0 score in both curves. The score of Linear Discriminant Analysis (LDA) gives the midlowest score in both ROC and AUC curves. The score of Gradient Boosting Classifier gives the highest predictable score in both ROC and AUC curves.

Best Parameters

```
The best parameters are:
{'criterion': 'gini', 'max_features': 'auto', 'min_samples_leaf': 5, 'splitter': 'best'}

The best model after gridsearch is:
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=5, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
```

Performance metrics

```
Precision: 0.9574468085106383
Accuracy: 0.9583333333333334
Recall: 0.9375
F1-score: 0.9473684210526315
```

Classification Report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98        72
           1       0.98      0.96      0.97        48

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

STAGE II RESULT :
K-Nearest Neighbor
Best Parameters

```
Best parameters :
{'algorithm': 'auto', 'n_jobs': 1, 'n_neighbors': 2, 'weights': 'uniform'}

Best model after gridsearch:
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=1, n_neighbors=2, p=2,
                     weights='uniform')
```

Performance metrics

```
Precision: 0.9787234042553191
Accuracy: 0.975
Recall: 0.9583333333333334
F1-score: 0.968421052631579
```

V. ADVANTAGES AND APPLICATIONS

ADVANTAGES

- CKD diagnoses assisted by software systems may improve confidence in clinical evaluations, which would help address the problem of low-quality primary health care in developing countries.
- Diagnosing CKD in its early stages of development can play a relevant role in helping decrease morbidity and mortality rates, as well as public health costs, in developing countries.
- This study analyses the usage of machine learning techniques to assist in the early diagnosis of CKD in developing countries.

APPLICATIONS

- The number of chronic kidney disease attributes used during chronic kidney disease risk classifications impacts the cost of usasge and the performance of the classifiers.

- The past decade has seen an increasing focus on chronic kidney disease (CKD) and its attendant complications, which has resulted in improved understanding of their impact on health-care resources.
- The early detection of CKD has been facilitated by the implementation of routine reporting of estimated glomerular filtration rates (eGFRs) and by education of primary care physicians on the implications of detecting a decreased eGFR with respect to patient safety as well as to cardiovascular and renal outcomes.

VI. CONCLUSION AND FUTURE SCOPE

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After un supervised 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. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of es tablishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples. Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (ckd and notckd) of data samples in the data set, the model cannot diagnose the severity of CKD. 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 severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.In this project we have studied different machine learning algorithms. We have analysed 24 different attributes related to CKD patients and predicted accuracy for different machine learning algorithms like Decision tree and KNN. From the results analysis, it is observed that the decision tree algorithms give the accuracy of 5.8% and KNN gives accuracy of 97.5%. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.

REFERENCES

[1]. Z. Chen et al., “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.
[2]. A. Subasi, E. Alickovic, J. Kevric, “Diagnosis of chronic kidney disease by using random forest,” in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.
[3]. L. Zhang et al., “Prevalence of chronic kidney disease in china: a cross sectional survey,” Lancet, vol. 379, pp. 815-822, Aug. 2012.
[4]. A. Singh et al., “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015.
[5]. A. M. Cueto-Manzano et al., “Prevalence of chronic kidney disease in an adult population,” Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.
[6]. H. Polat, H.D. Mehr, A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” J. Med. Syst., vol. 41, no. 4, Apr. 2017
[7]. C. Barbieri 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 et al., “Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study,” Am. J. Med., vol. 130, no. 12, Dec. 2017.
[9]. N. R. Hill et al., “Global prevalence of chronic kidney disease - A systematic review and metaanalysis,” Plos One, vol. 11, no. 7, Jul. 2016.
[10]. M. M. Hossain et al., “Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo COMPUTER AIDED DIAGNOSIS OF CHRONIC KIDNEY DISEASE