Big Data and Machine Learning Based Early Chronic Kidney Disease Prediction

Asra Fatima¹, Shireen Fatima², Ayesha Kiran³

¹Assistant Professor, Department of Computer Science and Engineering Faculty of Engineering and Technology, KBN University, Gulbarga, Karnataka, India. asra@kbn.university
²Assistant Professor, Department of Computer Science and Engineering Faculty of Engineering and Technology, KBN University, Gulbarga, Karnataka, India. shireen1@kbn.university
³Assistant Professor, Department of Computer Science and Engineering Faculty of Engineering and Technology, KBN University, Gulbarga, Karnataka, India. kiran@kbn.university

ABSTRACT

A chronic kidney disease, sometimes called a chronic renal disease, is characterized by a gradual decline in kidney purpose or abnormal kidney purpose which continues for months or even years. Patients with a domestic past of chronic kidney disease (CKD), high BP, or other kidney-related conditions are often the first to have chronic kidney disease (CKD) identified during screenings. Consequently, effective illness prevention and therapy rely on early prediction. Methods from the field of machine learning, including XGBoost, KNN, Decision Tree, and Random Forest, are being considered for use in this CKD project. The final product uses the fewest characteristics possible to determine whether the patient has chronic kidney disease (CKD).

Keywords- Random forest, Decision tree, XGBoot, KNN

I. INTRODUCTION

Current medical consensus is that chronic kidney disease (CKD) stances a important risk to public fitness. Laboratory tests can detect chronic kidney disease on a regular basis, and there are treatments that can stop the disease from progressing, slow it down, lessen danger of cardiovascular disease, its complications, increase subsistence, improve eminence of life.

Here, we look at the pros and cons of using ML for early CKD detection, as well as its practicality. Our goal is to create a machine learning model that uses split ratio, optimum parameters, data imputation, and data scaling approaches to assess classifiers' performance in classification tasks. The objective is to find a way to use ML techniques like DT, k-nearest to diagnose CKD efficiently. We use iterative imputation to manage lost digits, offer a new consecutive data scaling method. To find the most relevant features, we use Boruta feature selection, and then we modify the hyperparameters using grid-search CV. We compare our planned work's testing accuracy to those of existing research to see how well it holds up.

II. RELATED WORK

J. Snegha, integrated a number of data mining tools, including a Backpropagation Neural Network and the Random Forest method, into their suggested system. Because it employs a supervised learning network known as a feedforward neural network, the Back Propagation technique outperforms the other approach in this comparison.[1].

Mohamed Elhoseny, K. Shankar & J. Uthayakumar, detailed a CKD system that employs ACO-based density-based feature selection. When choosing features, the system employs wrapper techniques.[2].

Baisakhi Chakraborty, usage of machine learning algorithms including K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest Multi-Layer Perceptron Algorithm for the suggested construction of a system to detect chronic kidney disease. After applying them, we compare their performance to the outcomes of our recall, precision, and accuracy tests. And lastly, this system's implementation is based on Random Forest.[3].
International Conference on ECCE, developed a method for illness prediction using Boosting Classifiers, Ant-Miner, J48 Decision Tree. Both assessing efficacy of boosting algorithms in CKD detection and drawing rules to show correlations between CKD characteristics are the goals of this work. Based on the findings of the experiments, AdaBoost performed somewhat worse than LogitBoost.[4].

Siddeshwar Tekale, detailed a machine learning system which use support vector machine (SVM) decision-tree approaches. After evaluating two methods, we found that SVM produces the best outcome. So that physicians may examine patients in less time, its prediction procedure is less time-consuming. [5].

III. OBJECTIVES

1. Use the given information to make a prediction about the likelihood of chronic renal disease in a given person.
2. In order to create and verify a model that may predict the onset of chronic renal disease. Predicting and classifying illnesses is a common use case for machine learning algorithms in medicine.

IV. EXISTING SYSTEM

We analyzed the current system using three machine learning classifiers—Logistic Regression (LR), Decision Tree (DT), and SVM—upgraded the model's performance using the bagging ensemble approach. Clusters in the chronic renal illness dataset were used to train machine learning classifiers. We next use nonlinear characteristics and categories to construct the Kidney Disease Collection. The decision tree outperforms the others and has a 95% success rate. One more thing: bagging ensemble method gets us to highest possible correctness of 97%.

V. PROPOSED SYSTEM

Machine learning algorithms including XGBoost, Decision Tree, Random Forest, and K-Nearest Neighbor are to be used in proposed CKD prediction system. After that, we compare their efficacy to the outcomes of the accuracy, precision, and recall tests.

A. ADVANTAGES OF PROPOSED SYSTEM

1. Predicts, prevents and detects the disease
2. Needs less time
3. Accuracy is more

B. ARCHITECTURE

Fig-1: System Architecture
VI. METHODOLOGY

A. Decision Tree (DT):

The design of the decision tree classifier is reminiscent of a tree or a flowchart. A parent node, branches, and child nodes make it up. Subdivisions represent the result of each check on each node, whereas inner nodes include the features. Due to its ability to function without a large amount of field data or restrictions, decision trees are among the most popular classifiers used for categorization determination.

B. K-Nearest Neighbor

This supervised ML technique is popular because of its simplicity. It uses a distance measure like the Manhattan, Minkowski, or Euclidean to forecast class of an remark by looking at classes of its k closest neighbors. Here are the formulas for various distance measurements:

\[ d_{\text{Manhattan}}(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{kj}| \]

\[ d_{\text{Minkowski}}(x_i, x_j) = \left( \sum_{k=1}^{d} |x_{ik} - x_{kj}|^p \right)^{1/p} \]

Here, xik and xij stand for the kth and d-th characteristics of xi, xj, respectively, in a space of d dimensions. By these distance measurements, it finds a data point's k closest neighbors and uses the class that is most prevalent among them to control data point's class. Algorithm's ease of use makes it suitable for a wide range of categorization jobs.

Algorithm provides the pseudocode that outlines step-by-step approach, crucial hyperparameter selections for building, Gaussian NB model in this study. Model construction makes use of hyperparameters including training data, leaf size, parameter, weight function, method, number of neighbors, distance metric. Taking into account the majority class among a test instance's k closest neighbors, the method predicts its class label. The method uses a majority vote procedure to determine the predicted class label after computing detachments among test instance, training examples, picking k closest neighbors, so on.

Fig-2: Pseudo Code
VII. IMPLEMENTATION

Fig-3: Home Page

Fig-4: Dataset
Fig-5: Graphical Dataset

Fig-6: Classification
Analysis of lasting kidney disease in humans is focus of this program. In this study, we compared the accuracy of 14 patient-related characteristics of chronic kidney disease (CKD) using several machine learning methods, like decision tree, random forest. One benefit of this method is the time savings it provides throughout the forecast process. Medical professionals will be able to diagnose more patients in less time and begin treating CKD patients earlier with this technology.

**REFERENCES**