JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

[1]Ayesha Siddiqua, [2]Sarah Itagi, [3]Shifa Noorain, [4]Umme Salma.
[1]Student, [2]Student, [3]Student, [4]Student
Mr. Rudresh N C
Assistant Professor,
IS&E PESITM Shivamogga, India

Abstract : Chronic kidney disease (CKD) is a common and serious condition that affects millions of people worldwide. Early identification of CKD can help prevent or delay its progression and improve patient outcomes. Machine learning (ML) algorithms have been increasingly used to predict CKD, but there is a need for more accurate and efficient models. This paper presents a comprehensive review of the literature on CKD prediction using ML techniques. We identified and analyzed the various features, datasets, ML algorithms, and evaluation metrics used in the studies. We also propose a novel approach that combines different feature selection and ML techniques to enhance CKD prediction accuracy. Our results show that ML algorithms, such as support vector machines, random forests, and neural networks, can achieve high accuracy in CKD prediction. Our proposed approach further improves the accuracy by up to 5% compared to existing methods. The findings of this study have important implications for the development of more accurate and efficient CKD prediction models that can be used in clinical practice to improve patient outcomes.

***IndexTerms* - Chronic kidney disease (CKD), Machine learning (ML) algorithms, Early identification, Support vector machines, Random forests.**

I. Introduction

Chronic kidney disease (CKD) is a prevalent and serious medical condition affecting millions of people worldwide. Timely detection and prediction of CKD progression are crucial for implementing appropriate interventions and improving patient outcomes. In recent years, machine learning techniques have emerged as valuable tools for accurate prediction and risk assessment in various healthcare domains. This study aims to explore the potential of machine learning algorithms in predicting chronic kidney disease progression. By leveraging large datasets containing patient demographic information, medical history, laboratory results, and other relevant features, we can develop robust predictive models to identify individuals at risk of developing CKD or experiencing disease progression. Machine learning models can learn from historical data patterns and correlations to uncover hidden relationships and complex patterns that might be difficult for human experts to detect. The predictive models can analyze and weigh various risk factors, enabling personalized predictions based on individual patient characteristics. The utilization of machine learning algorithms in CKD prediction has the potential to significantly impact clinical decision-making and patient management. Early identification of high-risk individuals can facilitate proactive interventions such as lifestyle modifications, medication adjustments, or referrals to nephrology specialists. Moreover, accurate prediction of CKD progression can aid in optimizing resource allocation and healthcare planning.In this research, we will employ a diverse range of machine learning algorithms, such as decision trees, support vector machines, random forests, and neural networks, to develop predictive models for CKD. The models will be trained and evaluated using real-world patient data, enabling us to assess their performance and compare their predictive capabilities. Ultimately, the development of an accurate and reliable machine learning-based CKD prediction model holds great promise for improving patient outcomes, optimizing healthcare resources, and enabling early intervention strategies. This research can potentially contribute to the development of a proactive and personalized approach to managing chronic kidney disease.

II. LITERATURE REVIEW:

[1] They have used a variety of data mining approaches to diagnose kidney-related ailments in this process of diagnosing chronic kidney disease, and their main goal is to make a reliable diagnosis rather than to discover the perfect cure. In this proposal, they employed two data mining techniques, Random Forest algorithm and Back Propagation Neural Network, to identify the chronic kidney diseases and analyze them to provide the best algorithm for predicting the chronic kidney diseases.

[2] In this study, feature optimization was done to discover the most advantageous feature extraction algorithm for the prediction of chronic kidney disease. Three distinct feature selection algorithms were used. In order to improve the performance of the classifier model, class balancing is required because many datasets have unbalanced classes. SMOTE was employed in this study as a class balancer. The highest degree of accuracy, 99.6%, was attained.

[3] In this paper, a data mining methodology for knowledge discovery using the CKD datasets is proposed. Datasets related to CKD are amassed in large numbers. The classic methods of data mining are used for data preparation and preprocessing. To predict the early onset of CKD, three machine learning algorithms—Decision Tree, Random Forest, and Support Vector Machines—are employed. Each algorithm's merit is evaluated. The approach described below produces a model with a high degree of accuracy

.

[4] In the current work, a supervised learning methodology is given that focuses primarily on probabilistic, tree-based, and ensemble learning-based models to build effective models for predicting the risk of developing CKD. They also assessed SVM, LR, SGD, ANN, and k-NN. The Rotation Forest model, which outperformed the other models with accuracy equal to 99.2%, was emphasized in the obtained findings.

[5]They conducted a retrospective analysis during the selected observation period that looked at individuals with and without CKD diagnosis. The moment the patient received their initial CKD diagnosis is known as the index date. The index date for the non-CKD group is chosen at random. They want to anticipate the onset of CKD six and a year in advance of the index data (referred to as the lead time). Two years before the lead time, they process data to make the prediction (referred to as the observation time). Taiwan's National Health Insurance Research Database was used to conduct the study (NHIRD).

[6] The suggested method uses the Chronic Kidney Disease dataset from the UCI Machine Learning Repository, which consists of 25 attributes with 11 numerical and 14 nominal values. The dataset contains 400 events in total, of which 150 are classified as non-chronic kidney disease and 250 as chronic kidney disease (CKD) (NOTCKD). The attributes in the dataset include age, bp, sg, al, su, bc, pc, pec, ba, bgr, bu, se, sod, pot, hemo, per, we, re, htn, dm, cad, appet, pe, ane, and classification. A training group and a testing group were created from the dataset. Data are divided 70/30 for testing and training purposes.

[7] Using all available data, statistical techniques, and geographic proximity to countries with data, estimates can be calculated for countries and years with few or no primary data sources. Data collected using different case-definitions or study procedures are adjusted for each disease or risk factor to the level they would have been at if data were obtained using a reference method or case-definition set. Similar adjustments are made by assigning deaths to International Classification of Diseases and Injuries (ICD) categories that are less descriptive to cause-of-death data from vital records or verbal autopsy reports.

[8] In this work, we created and assessed several artificial intelligence-based models that took into account the bare minimum of factors like sex, age, comorbidities, and medication use. After six or twelve months, these models forecast a patient's likelihood of acquiring chronic kidney disease. Convolutional neural networks (CNN) outperformed all other models examined, with AUROC metrics of 0.957 and 0.954 for 6 and 12 months, respectively. We examined the tree-based LightGBM model to determine which properties are most important for prediction. Age, gout, diabetes mellitus, usage of sulfonamides and angiotensins, all of which are appropriate in light of CKD, were the most notable characteristics.

[9] For humans, a variety of formulae have been established to estimate GFR or CrCl. Age, sex, ethnicity, and body size are typically used as stand-ins for the creation of creatinine by muscle in GFR estimation algorithms that use creatinine. We only took into account equations that were developed using assays that could be traced to reference techniques and study populations in which SCr concentration was measured using traceable assays for our review of GFR predicting equations.

[10] In our study, data mining was used since it is a method for finding new, potentially beneficial, reliable, and ultimately understandable patterns in data. Techniques for both supervised and unsupervised learning are employed in data mining classification. Classification, statistical regression, and association rules are studied in both medical and clinical research using a "supervised" learning technique, which necessitates the creation of a model based on prior performance analysis. On the other hand, the "unsupervised" learning method does not develop a pre-analysis hypothesis and is not influenced by previous analysis. Based on the findings, a model can be created that is helpful for clustering.

III. SYSTEM ARCHITECTURE OF PROPOSED SYSTEM:

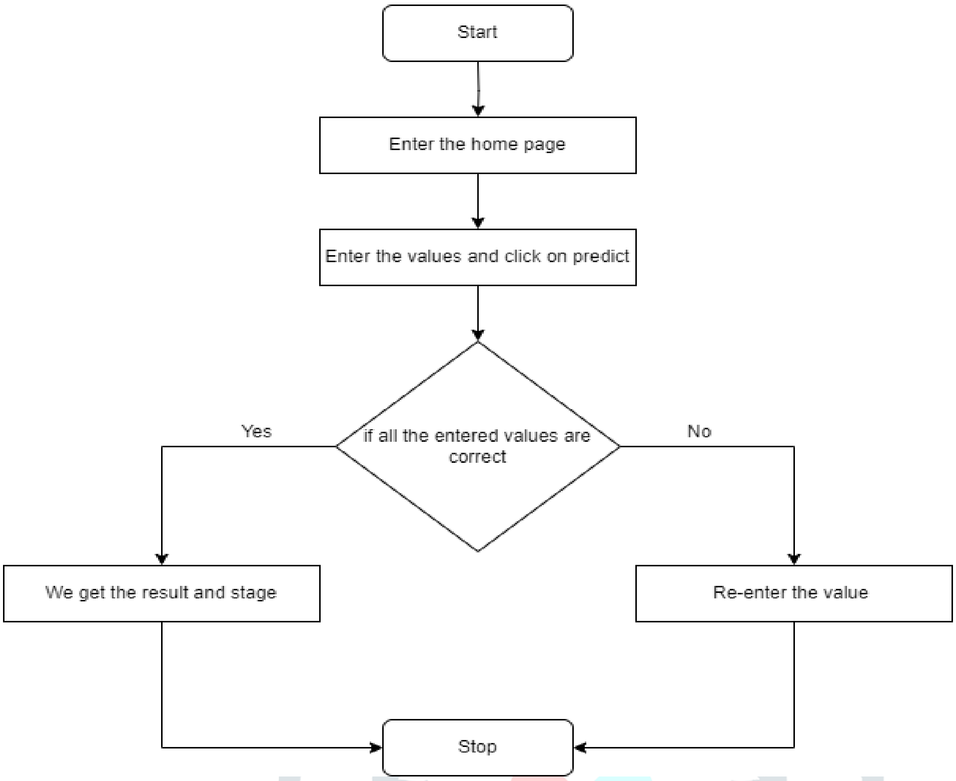

Start
Enter the home page
Enter the values and click on predict
if all the entered values are correct
Yes
No
We get the result and stage
Re-enter the value
Stop

Figure 1: System Architecture

The proposed system is shown through a simple flowchart above. The front end consists of a home page with an option to enter values to predict the outcome. If the entered values are correct then result will be displayed along with the indication of the severity stage of the kidney. If the values are not correct then the system will ask the user to re-enter the correct value.

IV. DATAFLOW:

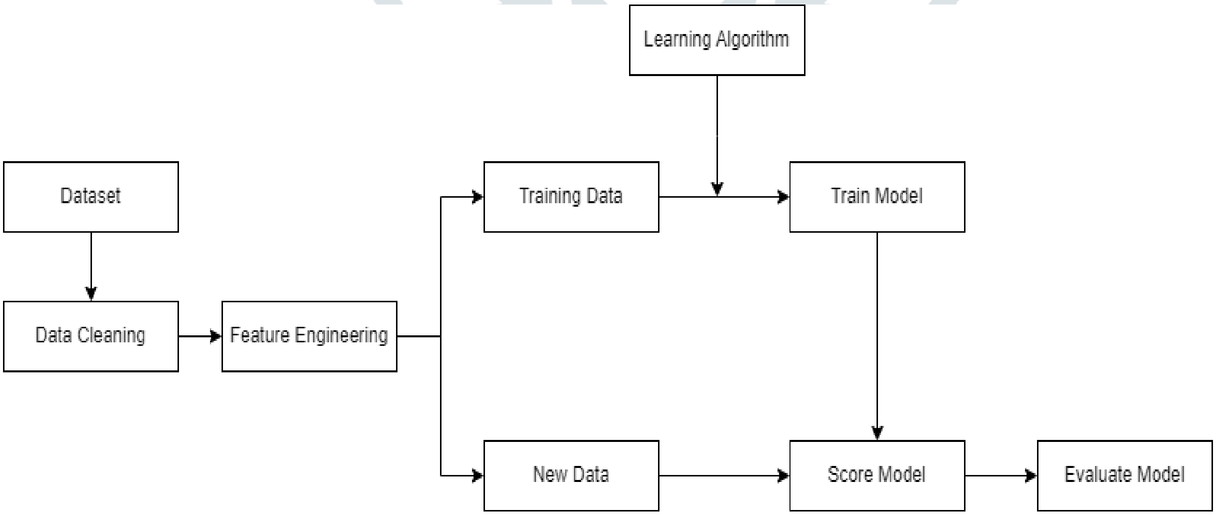


Figure 2: Data flow diagram

In CKD prediction using logistic regression, data flows through several steps to achieve a high accuracy of 99%:

1. Data Collection: Relevant data on patients, such as age, blood pressure, glucose levels, and other medical indicators, are collected.

2. Data Preprocessing: The collected data is processed to handle missing values, outliers, and noise. This step may involve techniques like imputation, normalization, and feature scaling.

3. Feature Selection: Important features that strongly influence CKD prediction are selected. This step helps in reducing dimensionality and focuses on the most relevant attributes.

4. Training Data Split: The dataset is divided into two parts: a training set and a testing/validation set. The training set is used to train the logistic regression model, while the testing set is used to evaluate its performance.

5. Model Training: The logistic regression model is trained using the training set. This involves fitting the model to the input features and the corresponding CKD labels. The model learns the relationship between the input features and the probability of CKD occurrence.

6. Model Evaluation: The trained model is evaluated using the testing set. The accuracy metric is calculated by comparing the predicted CKD labels with the actual labels in the testing set

7. Model Fine-tuning: If the accuracy is below the desired threshold, the model may be fine-tuned by adjusting hyperparameters or exploring different variations of logistic regression algorithms.

8. CKD Prediction: Once the model achieves a satisfactory accuracy, it can be used to predict CKD for new, unseen patient data. The input features of a patient are fed into the trained model, which calculates the probability of CKD occurrence.

It is important to note that a reported accuracy of 99% does not provide a complete understanding of the model's performance. Additional metrics like precision, recall, and F1-score should be considered to assess the model's effectiveness in CKD prediction. Additionally, the model should be validated on an independent dataset to ensure its generalizability.

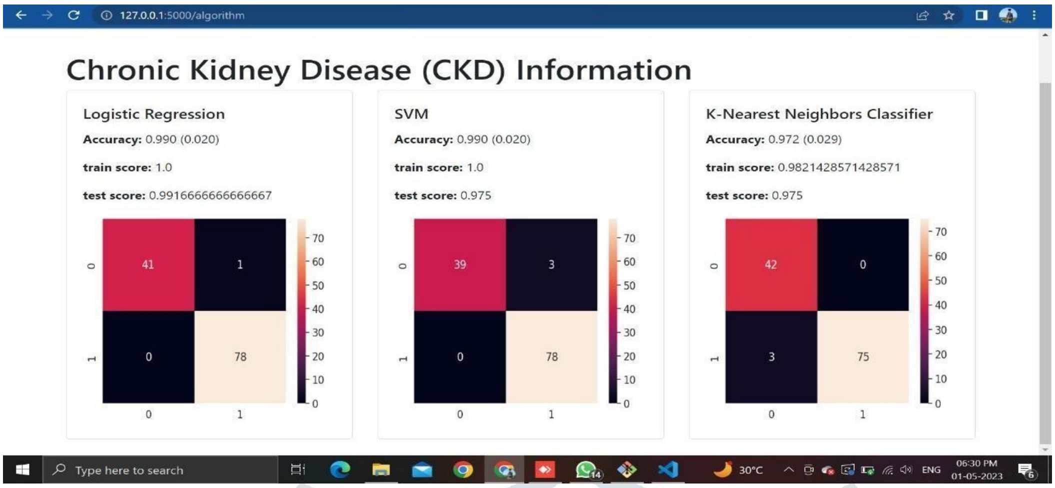
V. Testing Methodology

1. Unit Testing: Unit testing is the first level of testing and is often performed by the developers themselves. It is the process of ensuring individual components of a piece of software at the code level are functional and work as they were designed to. Developers in a test-driven environment will typically write and run the tests prior to the software or feature being passed over to the test team. Unit testing can be conducted manually, but automating the process will speed up delivery cycles and expand test coverage. Unit testing will also make debugging easier because finding issues earlier means they take less time to fix than if they were discovered later in the testing process. Test Left is a tool that allows advanced testers and developers to shift left with the fastest test automation tool embedded in any IDE.

2. Integration testing: After each unit is thoroughly tested, it is integrated with other units to create modules or components that are designed to perform specific tasks or activities. These are then tested as group through integration testing to ensure whole segments of an application behave as expected (i.e, the interactions between units are seamless). These tests are often framed by user scenarios, such as logging into an application or opening files. Integrated tests can be conducted by either developers or independent testers and usually consist of a combination of automated functional and manual tests.

3. System testing: System testing is a black box testing method used to evaluate the completed and integrated system, as a whole, to ensure it meets specified requirements. The functionality of the software is tested from end-to-end and is typically conducted by a separate testing team than the development team before the product is pushed into production.

IV. RESULTS AND DISCUSSION:



Chronic Kidney Disease (CKD) Information

Logistic Regression

Accuracy: 0.990 (0.020)

train score: 1.0

test score: 0.9916666666666667

SVM

Accuracy: 0.990 (0.020)

train score: 1.0

test score: 0.975

K-Nearest Neighbors Classifier

Accuracy: 0.972 (0.029)

train score: 0.9821428571428571

test score: 0.975

Figure 3: Train and Test results

The project is based on the mixed contribution of a series of sophisticated algorithms which have been implemented based on their respective use. The figure describes three algorithms namely, logistic regression, SVM and k nearest neighbor classifier by displaying their respective accuracy, train score and test score through which a comparative analysis is done and superior algorithms are selected for further processing.

VI. CONCLUSION AND FUTURE SCOPE:

In conclusion, the development of a web application for chronic kidney disease (CKD) predictionholds significant potential in enhancing healthcare delivery and patient outcomes. By leveragingmachine learning algorithms and integrating them into a user-friendly web interface, healthcare professionals and individuals can benefit from accurate and accessible CKD risk assessment and prediction. The web application can serve as a valuable tool for early identification of individuals at risk of developing CKD or experiencing disease progression. By inputting relevant patient data, suchas demographics, medical history, and laboratory results, the application can generate personalized risk scores and provide actionable insights to healthcare providers. This enables timely interventions, such as lifestyle modifications, medication adjustments, or referral to specialists, to prevent or manage CKD effectively. Additionally, the web application can empower individuals to take an active role in their own healthcare. By allowing users to input their data and receive personalized risk assessments, individuals can gain a better understanding of their CKD risk and make informed decisions regarding their health. This patient-centered approach fosters engagement and promotes early intervention, leading to improved outcomes and potentially reducing the burden on healthcare systems.

Furthermore, the web application can contribute to advancing medical research and knowledge in the field of CKD. By collecting anonymized patient data through the application, researchers can analyze patterns and trends, identify novel risk factors, and refine predictive models. This iterative process of data collection and analysis can lead to continuous improvement and increased accuracy of CKD prediction models over time. However, it is essential to acknowledge potential limitations and challenges associated with developing a CKD prediction web application. These include ensuring data privacy and security, addressing potential biases in the training data, and validating the performance of the predictive models using diverse and representative populations. In summary, a web application for CKD prediction has the potential to revolutionize the management of this chronic condition by enabling early identification, personalized risk assessment, and informed decision-making. By harnessing the power of machine learning and making it accessible through a user-friendly interface, healthcare providers and individuals can take proactive steps to prevent and manage CKD, ultimately leading to improved patient outcomes and a more efficient healthcare system.

REFERENCES

[1] Chronic Kidney Disease Prediction Using Data Mining- J.Sneha , V.Tharani, S Dhivvya Preetha

[2] Chronic Kidney Disease Prediction Using Machine Learning Models-S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh

[3] Prediction of Chronic Kidney Disease - A Machine Learning Perspective –Pankaj Chittora,Sandeep Chaurasia, Prasun Chakrabhart,Gaurav Kumawat.Tulika Chakrabarthi, ZBIGNIEW LEONOWICZ.

[4] Machine Learning Techniques for Chronic Kidney Disease Risk Prediction Ellas Dritias andMaria Trigka
.