**Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model**

**Alternate Title:**

A Data-Driven and Explainable AI Approach to Early Diagnosis of Chronic Kidney Disease

**Aim**:

 The primary aim of this research is to design, build, and rigorously evaluate an interpretable AI model for the early diagnosis of Chronic Kidney Disease (CKD) by leveraging a diverse dataset of patient information.

**Abstract:**

Chronic Kidney Disease (CKD) is currently experiencing a growing worldwide incidence and can lead to premature mortality if diagnosed late, resulting in rising costs to healthcare systems. Artificial Intelligence (AI) and Machine Learning (ML) offer the possibility of an early diagnosis of CKD that could revert further kidney damage. However, clinicians may be hesitant to adopt AI models if the reasoning behind the predictions is not understandable. Since explainable AI (XAI) addresses the clinicians’ requirement of understanding AI models’ output, this work presents the development and evaluation of an explainable CKD prediction model that provides information about how different patient’s clinical features contribute to CKD early diagnosis. The model was developed using an optimization framework that balances classification accuracy and explainability. The main contribution of the paper lies in an explainable data-driven approach to offer quantitative information about the contribution of certain clinical features in the early diagnosis of CKD. As a result, the optimal explainable prediction model implements with a high accuracy. In addition, an explainability analysis shows that hemoglobin is the most relevant feature that influences the prediction, followed by specific gravity and hypertension. These features are importance for results in a reduced cost of the early diagnosis of CKD implying a promising solution for developing countries.

**Introduction:**

 As we all know that, the Kidney is one of the most important organs for humans and animals as well. The kidney has main functionalities like osmoregulation and excretion. It plays a major role in purifying the blood and removes toxic materials and unwanted substances from the body. Chronic Kidney Disease (CKD) is a severe disease and can be a threat to society since this disease makes the kidney function improperly. Every year, there are approximately 10 lakh cases of Chronic Kidney Disease in India. Chronic Kidney Disease can be detected by regular laboratory tests. There are some treatments to stop the development. This disease can cause permanent kidney failure. Hence it is essential to detect CKD at its early stage but some people have no symptoms. So machine learning can be helpful to predict whether the person has CKD or not. This paper suggests different machine learning methods that are useful for forecasting the liver disease for a person depending on the collected attributes.

**Existing System:**

 The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level. Many medical methods are used for this purpose such as screening, ultrasound method. The datasets are preprocessing after that collected. Then the ordinal encoding, nominal encoding methods are implemented to the datasets. The features selected are denoted by their type (numerical, nominal, and ordinal) as well as the selection method (i.e. ANOVA, Chi-squared, Mutual Information, or Recursive Feature Elimination).

Using the mutual information technique, the features selected are hemo, htn, and sg. Random Forest and Extra Trees (both bagging ensemble trees) achieved similar results. This method is old and slow. We want to detect the kidney disease as soon as possible. So we moved to the proposed system.

**Proposed System**:

 Our Aim is to predict the chronic kidney disease using the machine learning algorithm. Chronic kidney disease (CKD) means your kidneys are damaged and can’t filter blood the way they should. The disease is called “chronic” because the damage to your kidneys happens slowly over a long period of time. This damage can cause wastes to build up in your body. CKD can also cause other health problems.10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year because the doctors are unable diagnose the disease.

Since the aim of this research is to achieve the most balanced CKD prediction model in terms of classification performance and explainability. The system is automation for predicting the CKD. We proposed Random Forest, Logistic Regression, Gradient Boost, XGBoost and SVM machine learning technique for kidney disease prediction of significant features. ML process starts from a pre-processing data phase followed by feature selection based on data cleaning, classification of modeling, performance evaluation, and the results with accuracy. In this process Gradient Boost achieves the results with high accuracy and low rate of error.

**Module Description:**

* Data Pre-Processing
* Algorithm Implementation
* Prediction

**Data Pre-Processing:**

Our Kidney Disease project dataset are collected from kaggle.com. Chronic kidney disease data is pre-processed after collection of various records. The dataset contains a more number of patient records, where some records are with some missing values. Those missing records have been removed from the dataset and the remaining patient records are used in preprocessing. After that we remove some columns based on feature selection.

**Algorithm Implementation:**

 The Classification Algorithms to produce the best results. We are using Random Forest, Logistic regression, Gradient Boost, XGBoost and SVM Algorithms to predict the kidney disease using ML. On an analysis conducted within various algorithms, the Gradient Boost was found to provide highest efficiency.

Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the results based on their low rate of error.

* Random Forest
* Logistic Regression
* Gradient Boost
* XGBoost
* SVM

**Prediction:**

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model. Preprocessed data are trained and input given by the user goes to the trained dataset and predicts the most accurate results.

**Software Requirements:**

* Operating System : Windows 10 (64 bit)
* Software : Python 3.7
* Tools : Anaconda (Jupyter Note Book IDE)

**Hardware Requirements:**

* Hard Disk : 500GB and Above
* RAM : 4GB and Above
* Processor : I3 and Above

**Architecture Diagram:**

Dataset

Preprocessing

User Input

Trained model

Machine Learning

No Chronic Kidney Disease

Chronic Kidney Disease

