iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53395

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Chronic Kidney Disease Prediction System Using Machine Learning

Rohit More[1], Subodh Shinde[2], Pratik Shah[3], Riya Raut[4], Prof. Payal Mahajan[5]
Zeal College of Engineering and Research

Abstract: ***Chronic kidney disease (CKD) is a life-threatening condition that can be difficult to diagnose early because there are no symptoms. The purpose of the proposed study is to develop and validate a predictive model for the prediction of chronic kidney disease. Machine learning algorithms are often used in medicine to predict and classify diseases. Medical records are often skewed. Chronic Kidney Disease (CKD) or chronic renal disease has become a significant issue with a steady growth rate. A person can only survive without kidneys for an average of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for the early prediction of CKD. Machine learning methods are effective in CKD prediction. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attribute selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. The project also considers the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.***

I. INTRODUCTION

Engineers and medical researchers are trying to develop machine-learning algorithms and models that can identify chronic kidney disease at an early stage. The problem is that the data generated in the health industry is large and complex, making data analysis difficult. However, we can process this data into a data format using data mining technology, and then this data can be translated into machine learning algorithms. A combination of estimated glomerular filtration rate (GFR), age, diet, existing medical conditions, and albuminuria can be used to assess the severity of kidney disease but requires more accurate information about the risk to the kidney is required to make clinical decisions about diagnosis, treatment, and referral. This model aims to develop and validate predictive models for chronic kidney disease. The main goal will be to evaluate kidney failure, which means the need for kidney dialysis or kidney transplant first.

These models also teach the patient how to live a healthy life, help the doctor see the risk and severity of the disease, and how to proceed with the treatment in the future. It may be possible to identify patterns of data collection using ANN, and mining methods and the future occurrence of certain diseases that may cause harm can be predicted in advance.

II. LITERATURE REVIEW

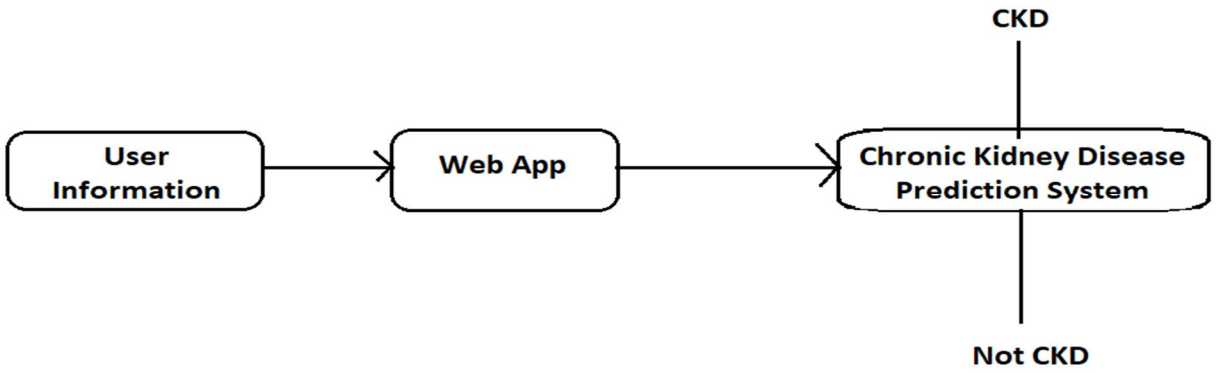
Machine learning techniques can be used to ascertain the existence of chronic kidney disease by imposing various classification algorithms on the patient's medical record. Empirical work is performed on algorithms like Support Vector Machine, Random Forest, XGBoost, Logistic Regression, Neural Networks, and Naive Bayes Classifier. The work is primarily concentrated on finding the best suitable classification algorithm which can be used for the diagnosis of CKD based on the classification report and performance factors. The experimental results show that Random Forest and XGBoost give better results when compared to other classification algorithms and generate 99.29% accuracy. The classification techniques, tree-based decision trees, random forests, and logistic regression, have been analyzed. The different measures have been used for comparison between algorithms for the dataset collected from the standard UCI repository.

Another study proposes and evaluates a Kernel-based Extreme Learning Machine to predict Chronic Kidney Disease. Subsequently, various kernel-based ELM were evaluated. The performance of four kernels-based ELM, namely RBF-ELM, Linear-ELM, Polynomial-ELM, Wavelet-ELM, and of standard ELM were compared. The result showed that the radial basis function extreme learning machine (RBF -ELM) was higher than those from the other tested and give the best prediction sensitivity and specificity of 99.38% and 100% respectively.

Another study demonstrates the early prediction model of kidney diseases using an adaptive neuro-fuzzy logic system (ANFIS). This model diagnoses the stages of kidney diseases so that treatment can be provided according to the disease condition. Mat lab-based ANFIS CKD stage prediction model is presented with an accuracy of 94 percent in terms of actual output to estimated output. Another study extracts the features which are responsible for CKD, then the machine learning process can automate the classification of chronic kidney disease in different stages according to its severity. The objective is to use a machine learning algorithm and suggest suitable diet plans for CKD patients using a classification algorithm on medical test records. Diet recommendations for the patient will be given according to potassium zone which is calculated using blood potassium level to slow down the progression of CKD.

III. PROPOSED SYSTEM

The proposed system aims to develop a Chronic Kidney Disease Prediction system using machine learning algorithms. The system plans to use Random Forest Classifier and XGBoost Classifier to predict the CKD. The model will be evaluated using various evaluation metrics such as confusion matrix, and accuracy score. The system will be developed as a web app. The front end of the web app will be created using the Python Flask framework.



3.1 UML Diagram of the CKD Prediction System

IV. REQUIREMENTS SPECIFICATIONS

The requirements of the proposed system are simple and easily available.

A. Software Requirements

OS - Windows
Python 3.7+
Python Jupyter Notebook

B. Hardware Requirements

RAM: 8 GB min.
Processor with 4 Cores

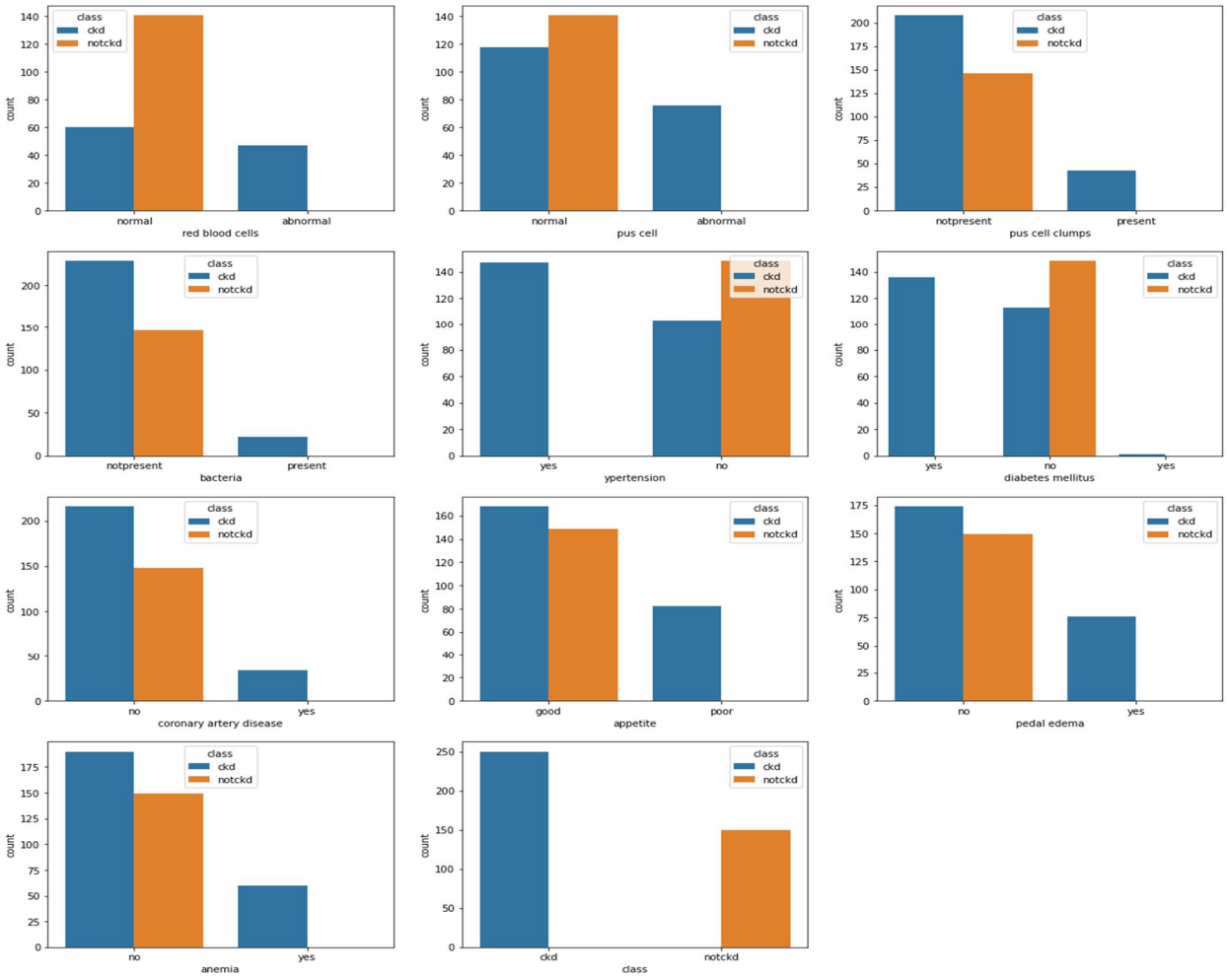
V. METHODOLOGY

Extracting the raw data and performing EDA and data preprocessing on it forms the first step of the design approach. Handling missing data and removing invalid data forms the second step of the design approach. Feature encoding, as well as Feature selection based on the statistical, medical importance, and availability of test features, forms the third step in the design approach. Training the machine learning models, model selection, and model evaluation form the fourth step in the design approach. After the system is tested and verified, the Python pickle module is used to serialize and deserialize the model. The front end is created using the Python Flask framework.

VI. IMPLEMENTATION

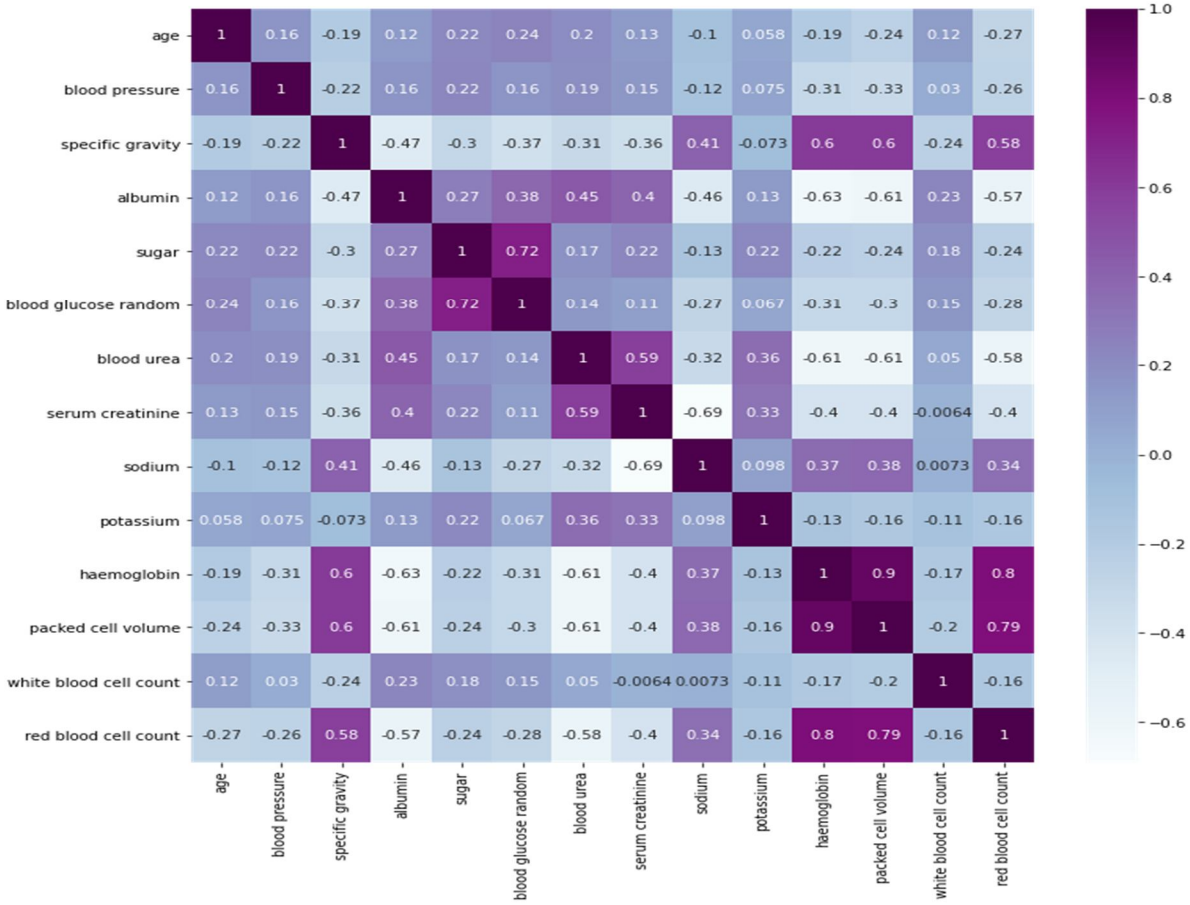
The dataset required for model training, named 'Chronic Kidney Disease', was fetched from Kaggle. Data Cleaning, Data Pre-processing, and Exploratory Data Analysis were performed on the dataset. The missing values of numerical attributes were filled with a suitable statistical measure of central tendencies such as mean, mode, or median. The missing values of the categorical attributes were filled using the 'KNN Imputer algorithm. The attributes with more than 20% missing values were discarded. After extracting the CKD dataset, data cleaning is performed. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. The invalid values of categorical data are either removed or replaced with valid values.

Extensive EDA is performed on the dataset. Data analysis is the process of collecting, modeling, and analyzing data using various statistical and logical methods and techniques. Analysis of data distribution of each and every column is performed. Then label distribution of categorical data is checked. This analysis is required to investigate which features have the most and least impact on the predictions. The analysis is also required when selecting the feature set. Analysis of data distribution of each and every column is performed. Then label distribution of categorical data is checked. This analysis is required to investigate which features have the most and least impact on the predictions. The analysis is also required when selecting the feature set. By plotting the count plot we can observe the count of abnormal and normal features in the dataset as well as their class i.e either they belong to the 'CKD' class or 'Not CKD' class.



6.1 Label Distribution of categorical data with CKD class

Then we check and understand the correlation between different features. The correlation is plotted using a heatmap. Heat Maps are graphical representations of data that utilize color-coded systems. The primary purpose of Heat Maps is to better visualize the volume of locations/events within a dataset and assist in directing viewers toward areas on data visualizations that matter most.



6.2 Correlation between features

We can draw the following conclusions from the correlation between features

1) Rbc count is positively correlated with specific gravity, hemoglobin, and packed cell volume
2) Rbc count is negatively correlated with albumin, blood urea
3) Packed cell volume and hemoglobin are highly positively correlated
4) Packed cell volume is negatively correlated with albumin and blood urea
5) Hemoglobin and albumin are negatively correlated

Checking the outliers in the dataset to identify the correct method of filling in missing values. It helps us to choose the correct type of central statistical measures from mean, mode, median, etc.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

By using data visualization following we can derive the following conclusions.

a) There is a linear relationship between hemoglobin and packed cell volume
b) People with lower RBC count have a high chance of having a chronic disease
c) Whenever hemoglobin is below 13-14 he is positive for chronic disease, Whenever hemoglobin is near 18 he is negative

Missing data can cause serious problems. First, most statistical procedures automatically eliminate cases with missing data. This means that in the end, you may not have enough data to perform the analysis, The dataset has many outliers therefore the numerical missing data is filled with the median of the feature. It was more important to find the missing values and need to clean those missing values by using different methods. (I've dropped the NULL Values). Missing Values lead to False Output and sometimes cause many Problems while Evaluating our Model. The missing data of categorical values are filled with random values using the KNN imputer algorithm.

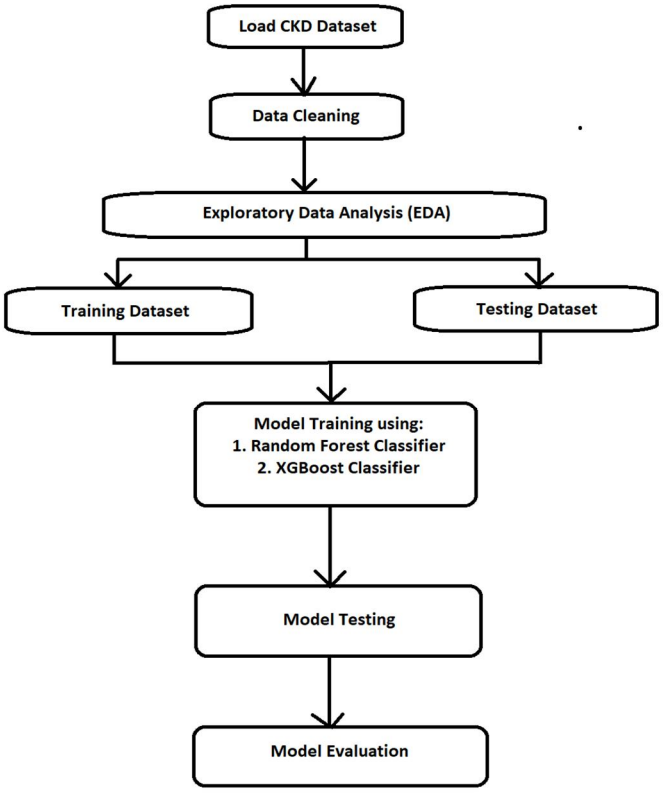
KNNimputer is a scikit-learn class used to fill out or predict the missing values in a dataset. It is a more useful method that works on the basic approach of the KNN algorithm rather than the naive approach of filling all the values with the mean or the median. In this approach, we specify a distance from the missing values which is also known as the K parameter. The missing value will be predicted in reference to the mean of the neighbors.

Label Encoding refers to converting the labels into a numeric form so as to convert them into a machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. Using a label encoder we perform feature encoding on the dataset and make it ready to train the machine learning model.

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine-learning model based on the type of problem you are trying to solve. We 'SelectKBest' and chi-square test for feature selection. Based on the chi-square test scores the features are ranked and the top 10 features are selected as the feature set for training the machine learning model.

Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model. We have used the Random Forest algorithm to train the model to predict Chronic Kidney Disease. We are using the XGBoost classifier algorithm for our machine-learning model. XGBoost (eXtreme Gradient Boosting) is a popular supervised-learning algorithm used for regression and classification on large datasets. It uses sequentially-built shallow decision trees to provide accurate results and a highly-scalable training method that avoids overfitting. 11 Models are trained and tuned for hyperparameters. Based on accuracy and feature importance distribution the best model is selected.

Evaluation of the model is performed using a confusion matrix. A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.



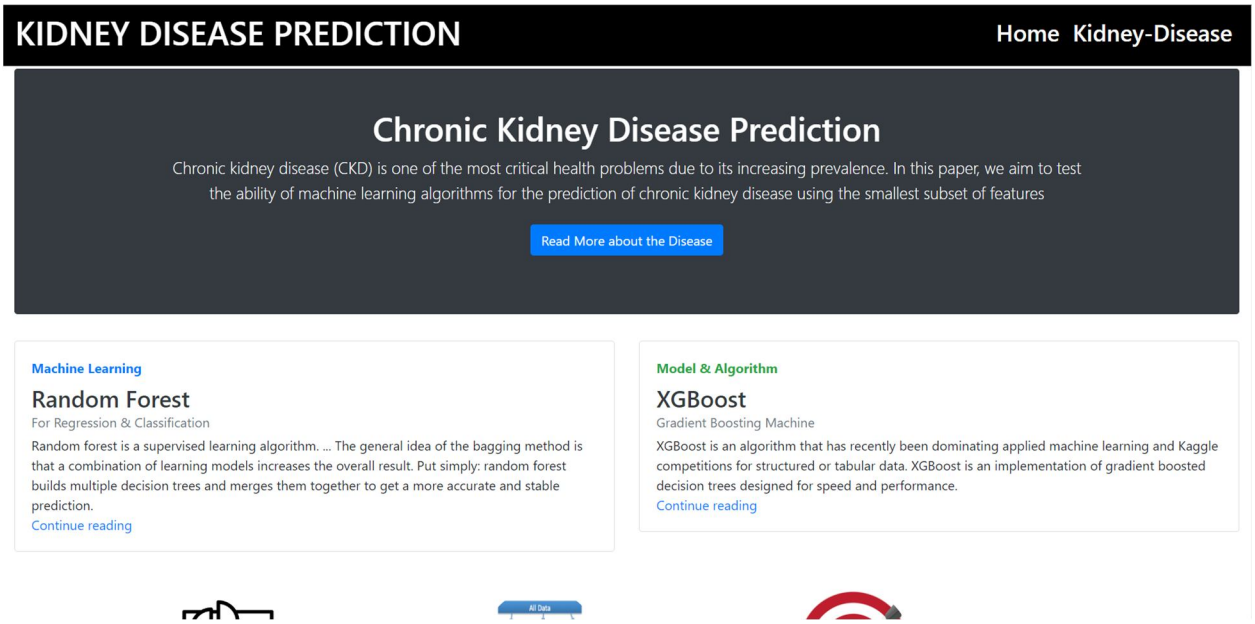
6.3 System Architecture

The serialization and deserialization process between the front end and the back end is handled by the Python Pickle module.

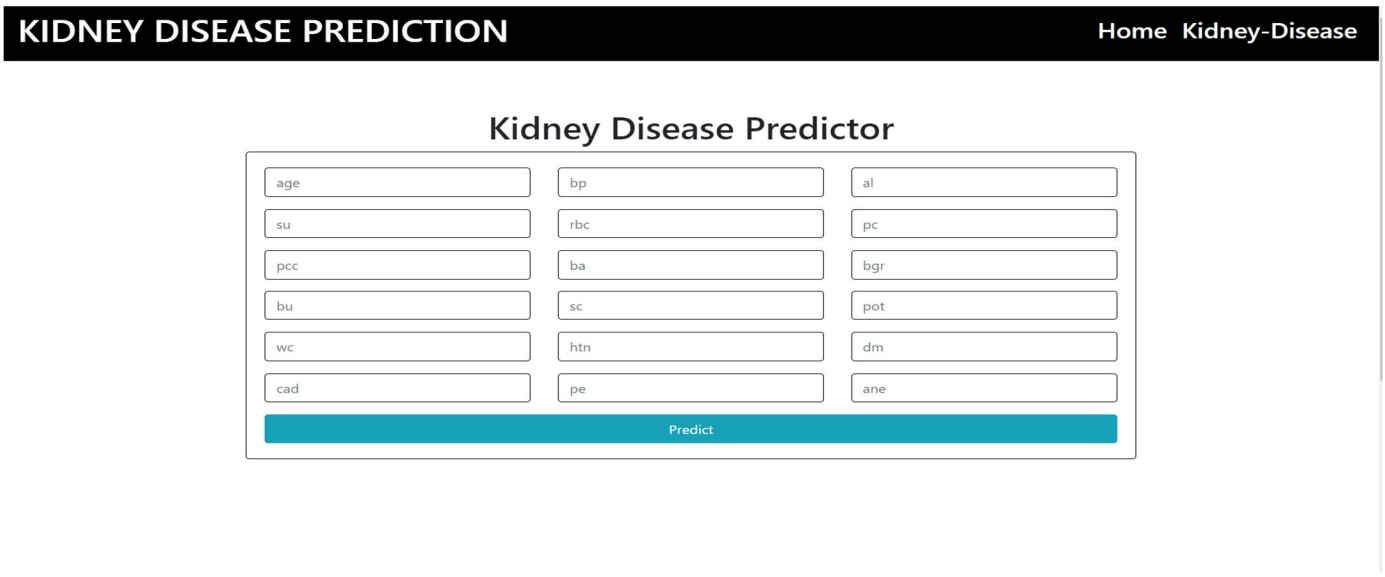
Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. A web app using Flask is built to provide a neat user interface for the users of the Chronic Kidney Disease Prediction System. A Home page is created to provide basic information about Chronic Kidney Disease. It also provides information about the dataset and the algorithms used in training the machine learning model. The Navbar helps to navigate to the 'CKD prediction page'. This page helps the users know whether they have CKD or not based on the medical data provided.

VII. RESULTS

The model is trained using the Random Forest algorithm and XGBoost Classifier algorithm. The accuracy for Random Forest was 95.83% and for XGBoost it was 98.33%. The user-friendly web app for accessing the 'CKD Prediction system' is created. The system, both front-end and back-end is tested rigorously.



7.1 Home page



7.2 Kidney Prediction page

VIII. CONCLUSION

In this project, we have used the chronic kidney disease dataset collected from the Kaggle repository. We have developed a chronic kidney disease prediction model using two machine learning classifiers Random Forest and XGBoost Classifier to measure the performance of the prediction model. The performance of the model depends upon the confusion matrix. The developed chronic kidney disease prediction model has been trained by categorical and non_categorical chronic kidney disease dataset attributes. After applying the base classifiers we find that the Random Forest classifier got an accuracy of 95.83% and XGBoost Classifier got an accuracy of 98.33%. The XGBoost Classifier classifier performed better than Random Forest Classifier. This can help medical practitioners and patients in the early prediction of chronic kidney disease to save a life. In the future, the model can be further tuned by applying feature selection methods to increase the performance of the prediction.

REFERENCES

[1] H. A. Wibawa, I. Malik and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), 2018, pp. 1-4, DOI: 10.1109/ICICOS.2018.8621762.

[2] A. Maurya, R. Wable, R. Shinde, S. John, R. Jadhav, and R. Dakshayani, "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), 2019, pp. 1-4, DOI: 10.1109/ICNTE44896.2019.8946029.

[3] N. V. Ganapathi Raju, K. Prasanna Lakshmi, K. G. Praharshitha and C. Likhitha, "Prediction of chronic kidney disease (CKD) using Data Science," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 642-647, DOI: 10.1109/ICCS45141.2019.9065309.

[4] R. Gupta, N. Koli, N. Mahor and N. Tejashri, "Performance Analysis of Machine Learning Classifier for predicting Chronic Kidney Disease 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, DOI: 10.1109/INCET49848.2020.9154147.

[5] K. Damodara and A. Thakur, "Adaptive Neuro Fuzzy Inference System based Prediction of Chronic Kidney Disease," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 973-976, DOI: 10.1109/ICACCS51430.2021.9441989.

[6] A. Maurya, R. Wable, R. Shinde, S. John, R. Jadhav and R. Dakshayani, "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), 2019, pp. 1-4, doi: 10.1109/ICNTE44896.2019.8946029.

[7] P. Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 193-198, doi: 10.1109/COMPSAC.2017.84.

[8] H. Zhang, C. -L. Hung, W. C. -C. Chu, P. -F. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 1351-1356, doi: 10.1109/BIBM.2018.8621294.

[9] G. Chen et al., "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 100497-100508, 2020, doi: 10.1109/ACCESS.2020.2995310.

[10] A. A. Johari, M. H. Abd Wahab and A. Mustapha, "Two-Class Classification: Comparative Experiments for Chronic Kidney Disease," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 789-792, doi: 10.1109/ISCON47742.2019.9036306.

[11] I. U. Ekanayake and D. Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods," 2020 Moratuwa Engineering Research Conference (MERCon), 2020, pp. 260-265, doi: 10.1109/MERCon50084.2020.9185249.

[12] L. Jena and R. Swain, "Work-in-Progress: Chronic Disease Risk Prediction Using Distributed Machine Learning Classifiers," 2017 International Conference on Information Technology (ICIT), 2017, pp. 170-173, doi: 10.1109/ICIT.2017.46.

[13] C. P. Kashyap, G. S. Dayakar Reddy and M. Balamurugan, "Prediction of Chronic Disease in Kidneys Using Machine Learning Classifiers," 2022 1st International Conference on Computational Science and Technology (ICCST), CHENNAI, India, 2022, pp. 562-567, doi: 10.1109/ICCST55948.2022.10040329.

[14] Rajeshwari and H. K. Yogish, "Prediction of Chronic Kidney Disease Using Machine Learning Technique," 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP), Bengaluru, India, 2022, pp. 1-6, doi: 10.1109/CCIP57447.2022.10058678.

[15] A. Farjana et al., "Predicting Chronic Kidney Disease Using Machine Learning Algorithms," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 1267-1271, doi: 10.1109/CCWC57344.2023.10099221.

[16] S. Kumari and S. K. Singh, "An ensemble learning-based model for effective chronic kidney disease prediction," 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 162-168, doi: 10.1109/ICCCIS56430.2022.10037698.

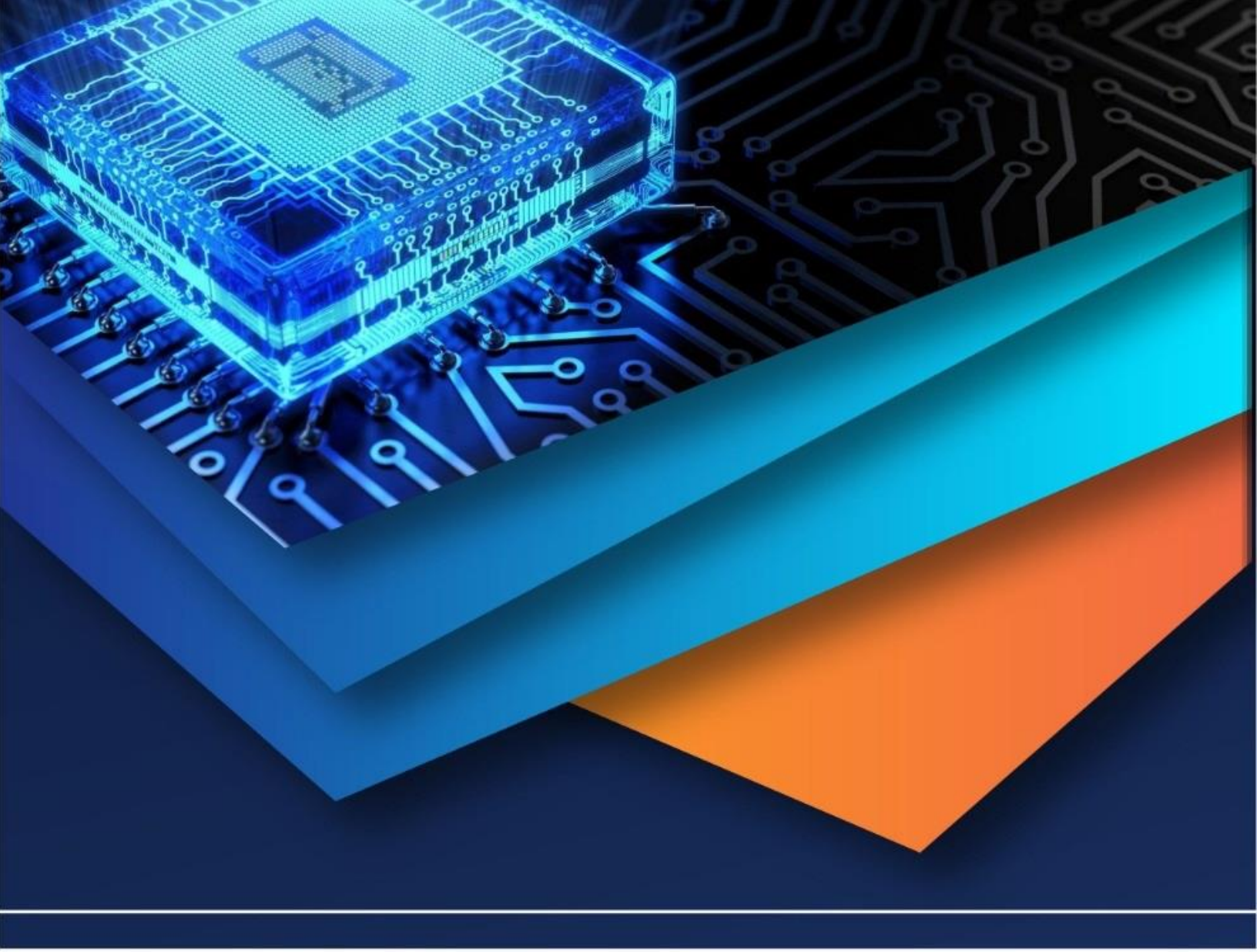
[17] B. Gudeti, S. Mishra, S. Malik, T. F. Fernandez, A. K. Tyagi and S. Kumari, "A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1630-1635, doi: 10.1109/ICECA49313.2020.9297392.

[18] N. I. Mahbub, M. I. Hasan, M. M. Ahamad, S. Aktar and M. A. Moni, "Machine Learning Approaches to Identify Significant Features for the Diagnosis and Prognosis of Chronic Kidney Disease," 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2022, pp. 312-317, doi: 10.1109/ICISET54810.2022.9775827.

[19] M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, Australia, 2017, pp. 300-303, doi: 10.1109/LSC.2017.8268202.

[20] D. Swain, H. Patel, K. Patel, V. Sakariya and N. Chaudhari, "An Intelligent Clinical Support System For The Early Diagnosis Of The Chronic Kidney Disease," 2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur, Odisha, India, 2022, pp. 1-5, doi: 10.1109/iSSSC56467.2022.10051517.

[21] A. Vijayalakshmi and V. Sumalatha, "Survey on Diagnosis of Chronic Kidney Disease UsingMachine Learning Algorithms," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 590-595, doi: 10.1109/ICISS49785.2020.9315880.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY