Chronic Kidney Disease Prediction using Machine Learning

D Venkata Reddy
Computer Science & Engineering
Narasaraopeta Engineering College
Narasaraopet,India
doddavenkatareddy@gmail.com

Jahnavi Chekuri
Computer Science & Engineering
Narasaraopeta Engineering College
Narasaraopet,India
jahnavichekuri999@gmail.com

Sharmila Shaik
Computer Science & Engineering
Narasaraopeta Engineering College
Narasaraopet,India
sharmilashaik@gmail.com

Lakshmi Shravani Challamcharla
Computer Science & Engineering
Narasaraopeta Engineering College
Narasaraopet,India
lakshmisravani706@gmail.com

ABSTRACT- **Chronic kidney disease (CKD) is a significant health condition that can persist throughout an individual's life, resulting from either kidney malignancy or diminished kidney function. In our study, we delve into the potential of various machine learning methods to offer an early diagnosis of CKD. While previous research has extensively explored this area, our aim is to refine our approach by employing predictive modeling techniques. Initially, we considered 25 variables alongside the class property. LOGISTIC REGRESSION classifier gave an accuracy of (0.99), Recall Our research underscores the potential of recent advancements in machine learning, coupled with predictive modeling techniques, to offer a promising avenue for developing novel solutions in accurately predicting kidney disease and potentially other medical conditions in the future.**

KEYWORDS: XGBOOST classifier, Logistic Regression, Chronic Kidney Disease, Decision Tree, Random Forest.

I. INTRODUCTION

Chronic kidney disease (CKD) is a significant public health problem for worldwide, especially for a low and medium-income countries. [1] About 10 percent of the population for a worldwide suffering from (CKD) and millions of people die. In 2020, a study was conducted by International Society of Numerology (ISN) on global burden disease, they reported that CKD has been raised an important cause of the mortality worldwide with the number of deaths increasing by 82.3 percent in the last two decades. The worst possible outcome of the chronic kidney disease and the symptoms causing any reduced kidney functioning would lead to kidney failure and [2] Early detection and treatment of CKD can slow or stop the progression of the kidney disease.
But the CKD, in early stages show no symptoms. The illness known as chronic kidney disease, or CKD, occurs when the kidneys are so severely damaged that they are unable to filter blood [3] as effectively as they should.

The elimination of waste and surplus water from the circulation is the kidneys' primary function. Urine is created in this manner. Waste accumulation in the body is indicated by CKD. The reason this ailment is considered chronic is that the damage develops
gradually over an extended period of time.
As a result of CKD, you may encounter a number of health issues. CKD can result from a wide range of illnesses, diabetes, high blood pressure, and heart disease being only three of them. Apart from these grave health issues, age and gender also

When to visit a Doctor :
Schedule a visit with your physician if you exhibit any signs or symptoms of renal illness. If renal disease is identified early on, it may be able to be treated before kidney failure develops. If you have a medical condition that puts you at risk for renal disease, your doctor may use blood and urine tests to monitor your kidney function and blood pressure during office visits.

Tests for CKD:

Chronic kidney disease is when a disease or condition makes it hard for the kidneys to work, causing the damage to the kidneys to get worse over time. [4] This can occur when the kidneys are affected by another disease or condition. Studies show that the number of people with CKD who are admitted to hospitals is going up by 6.23 percent every year, even though the global death rate has stayed the same. There are just a few diagnostic tests available to check the status of CKD, including: (i) estimated glomerular filtration rate (EGFR) (ii) a urine test; (iii) a blood pressure reading; (iv) tests for CKD. This disease affected 753 million people globally in 2016 in which 417 million are females and 336 million are males. Majority of the time the disease is detected in its final stage and which sometimes leads to kidney failure. The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level [5] many medical methods

II. LITERATURE SURVEY

Siddheshwar Tekale et al.[1] described a system using machine learning which uses Decision tree SVM techniques. By comparing two techniques finally concluded that SVM gives the best result. Its prediction process is less time consuming so that doctors can analyze the patients within a less time period.

Parul et al.[2] ,Performed a classification algorithm has been compared on the basis of accuracy, precision and total execution time for prediction of Chronic Kidney disease. MATLAB was used for this classification model. Performance of K-Nearest Neighbour classifier was 78.75% which was better than Support Vector machine with an accuracy of 73.75%

K. A. Padmanaban et al.[3] aimed in their work to .In their research, they used 600 clinical records collected from a leading Chennaibased diabetes research center. The authors have tested the dataset using the decision tree and Naïve Bayes methods for classification using the WEKA tool. They concluded that the decision tree algorithm outweighs the Naïve Bayes with an accuracy of 91%.

Dr.Uma et al.[4] performed Extraction of action Rules for. Naïve Bayes with One attribute Selector was used for prediction CKD status of a patient . The idea was to select a subset from input data by elimination idle data which carried little or no predictive knowledge. Data sets were taken from UCI ML Repository. The result and analysis proposed Naïve Bayes with One with highest improved accuracy and also reduced number of attributed to 80% which is 05 from total of 25 attributes compared to other attribute evaluators.

Dr.N.Radha et al.[5] performed a diagnosis of like Back Propagation neural network, Random forest, Radial Basis function, ANN. The data for this research was medical reports of patients taken from different labs in South India. They have used 1000 instances with 15 CKD related attributes. Their model is evaluated on different measures like Sensitivity, Accuracy, and Specificity. The experimental results proved that Radial Basis Function performed better than other algorithms and obtained an accuracy of 85.3%.

Dr.S.Vijayarani et al.[6] used MATLAB tool. Their work focuses on serum-levels finding best classification algorithm on basis of accuracy and execution time for prediction of Kidney Disease. In their Prediction model SVM classification algorithm performed better than Naïve Bayes with an accuracy of 76.32

Salekin et al.[7]. evaluated three classifiers: random forest, K-nearest neighbors, and neural network to detect the CKD .They used a dataset with 400 patients form UCI with 24 attributes. By using the wrapper method, a feature reduction analysis has been performed to find the attributes that detect this disease with high accuracy. By considering: albumin, specific gravity, diabetes mellitus, hemoglobin, and hypertension as features, they can predict the CKD with .98 F1 and 0.11 RMSE.

III. PROPOSED SYSTEM

This flow diagram in the “Fig 1” outlines the process for developing and deploying a machine learning-based system for kidney disease prediction. Each step involves various tasks and considerations to ensure the system's success in accurately identifying and managing kidney-related health issues.

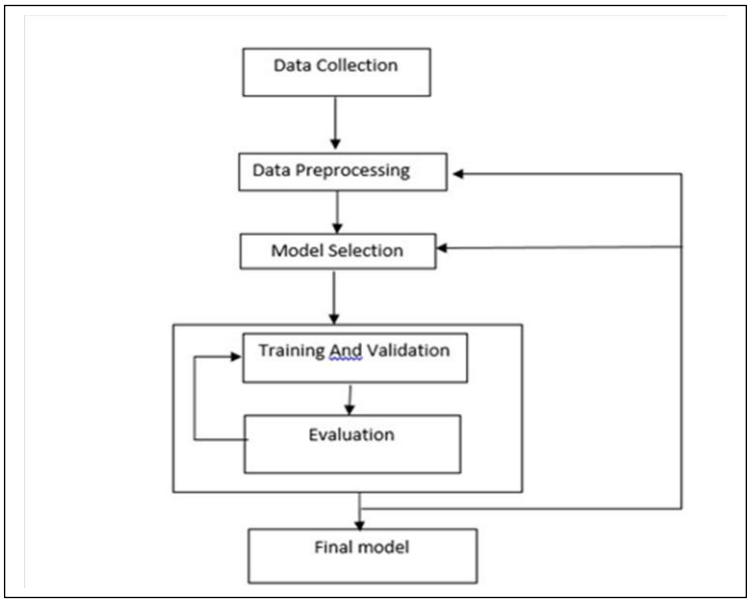

Fig.1. The steps involved in the Model

Dataset Analysis
Data Visualization
Preprocessing Techniques
Model Creation and Evaluation
Accuracy

A.Dataset Analysis:

We have taken the dataset from kaggle[14]. The dataset consists of 25 features and these are descripted as :

The TABLE1 gives the detailed description of the attributes involved in the dataset.

TABLE I. Attributes in the dataset.

Name	Description	Type: unit/ values
Age (age)	Patient's age	Numeric: years
Blood pressure (bp)	Blood pressure of thepatient	Numeric: mm/Hg
Specific gravity (sg)	The ratio of the densityof urine	Nominal: 1.005, 1.010, 1.015, 1.020,1.025
Albumin (al)	Albumin level in theblood	Nominal: 0,1,2,3,4,5
Sugar (su)	Sugar level of the patient	Nominal: 0,1,2,3,4,5
Red blood cells (rbc)	Patients' red blood cells count	Nominal: normal, abnormal
Pus cell (pc)	pus cell count of patient	Nominal: normal, abnormal
Pus cell clumps (pcc)	Presence of pus cell clumps in the blood	Nominal: present, not present
Bacteria (ba)	Presence of bacteria in the blood	Nominal: present, not present
Blood glucose (bgr)	blood glucose random count	Numeric: mgs/dl
Blood urea (bu)	blood urea level of the patient	Numeric: mgs/dl
Serum creatinine (sc)	serum creatinine level in the blood	Numeric: mgs/dl
Sodium (sod)	sodium level in the blood	Numeric: mEq/L
Potassium (pot)	potassium level in the blood	Numeric: mEq/L
Hemoglobin (hemo)	hemoglobin level in the blood	Numeric: gms
Packed cell volume (pcv)	packed cell volume in the blood	Numeric
White blood cell count (wc)	white blood cell count of the patient	Numeric: cells/cumm
Red blood cell count (rc)	red blood cell count of the patient	Numeric millions/cmm
Hypertension (htn)	Does the patient has hypertension on not	Nominal: yes, no
Diabetes mellitus (dm)	Does the patient has diabetes or not	Nominal: yes, no
Coronary artery disease (cad)	Does the patient has coronary artery disease or not	Nominal: yes, no
Appetite (appet)	Patient's appetite	Nominal: good, poor
Pedal Edema (pe)	Does patient has pedal edema or not	Nominal: yes, no
Anemia (ane)	Does patient has anemia or not	Nominal: yes, no
Class	Does the patient has kidney disease or not	Nominal: CKD, not CKD

B .Data Visualization:

A wide range of factors are present in the dataset utilized for CKD prediction, such as clinical measurements blood pressure, serum-creatinine levels, medical history comorbidities, and demographic data age, gender and as described in the "Fig 2","Fig 3" the count plots of the features are described as well. To predict whether CKD would be present or not, machine learning algorithms use these properties as input variables.

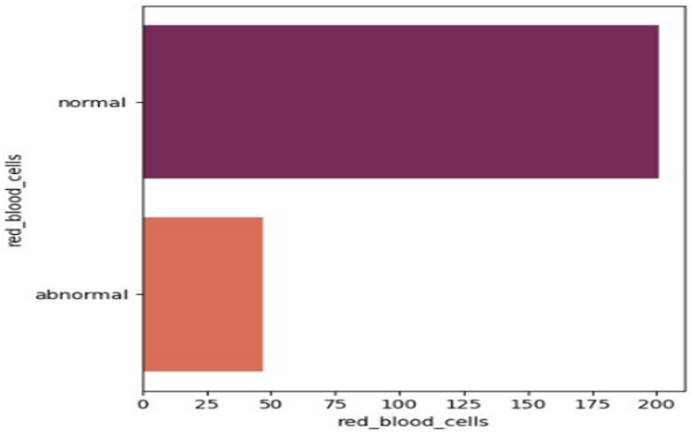

Fig. 2. The count plot of red blood cells in the dataset

Analyzing the dataset's major feature distribution is where we start. The distribution of continuous variables, such as age, blood pressure, and serum creatinine levels, as in "Fig 4"can be seen using histograms and density plots, which shed light on the range and spread of these characteristics among them.

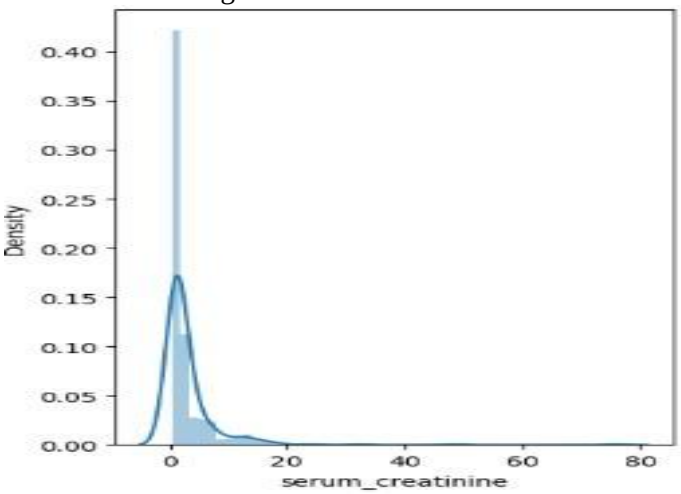

Fig. 3. The density plot for serum_creatinine

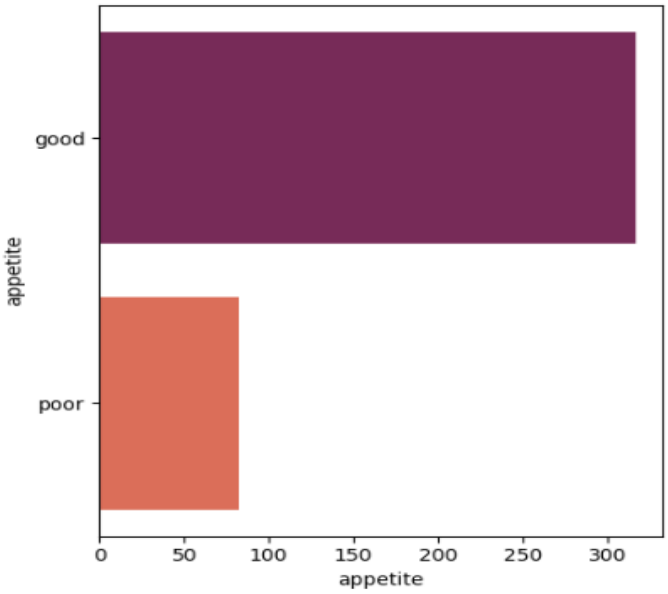

Fig. 4. The count plot of the values in the appetite

We look at the relationship between the target variable (CKD status) and the characteristics. The correlation
matrix is visualized using heat maps, which illustrate the connections between various characteristics and their capacity to predict CKD. Finding pertinent features for feature selection and model training is aided by this analysis.

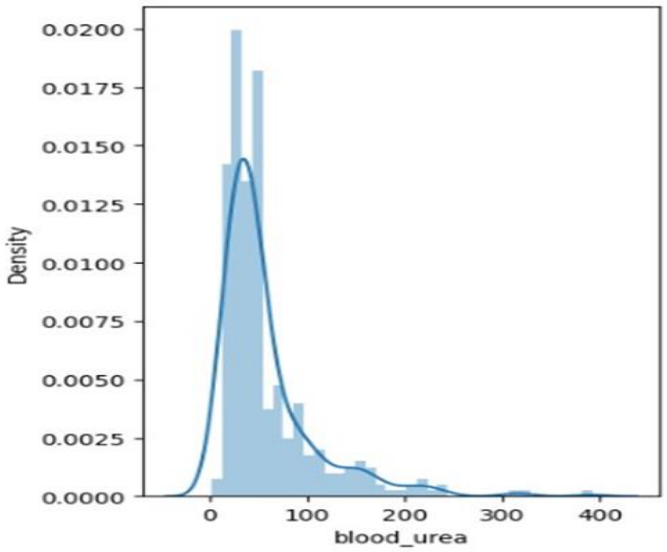

Fig. 5. The density plot for the attribute blood_urea

As described in “Fig 4” ,”Fig 5” ,”Fig 6” Density plots are used to show how numerical features are distributed between instances with and without CKD. The frequency distribution of continuous data is represented by histograms, The probability density function is smoothed out in density plots, which make it easier to spot trends and possible cutoff lines for predictive modeling and in “Fig 7” the scatter plots of age is visually observed and displayed.

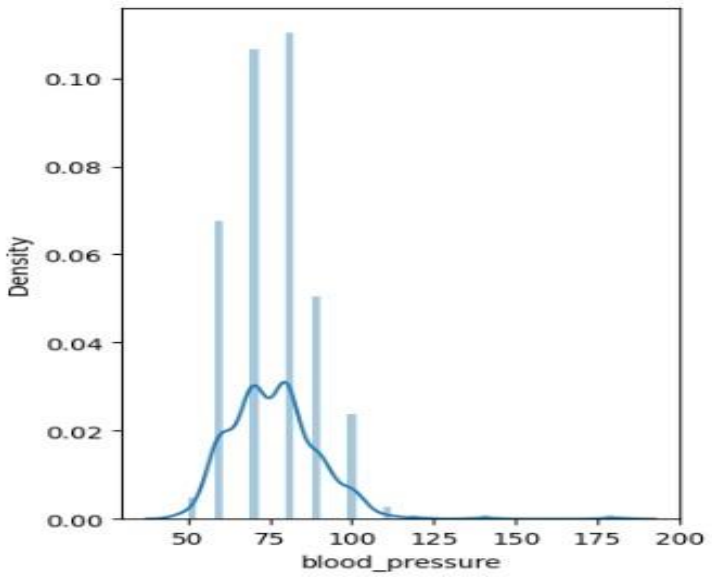

Fig.6. The density plot for the attribute blood_pressure

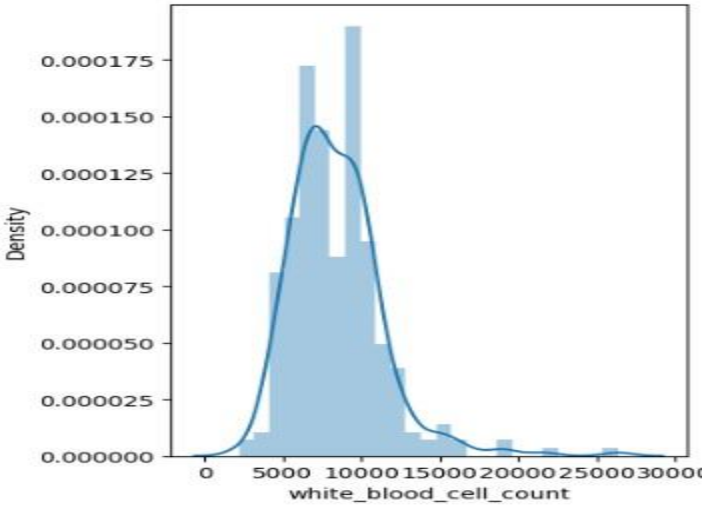

Fig.7. The density plot for the attribute white_blood_cells

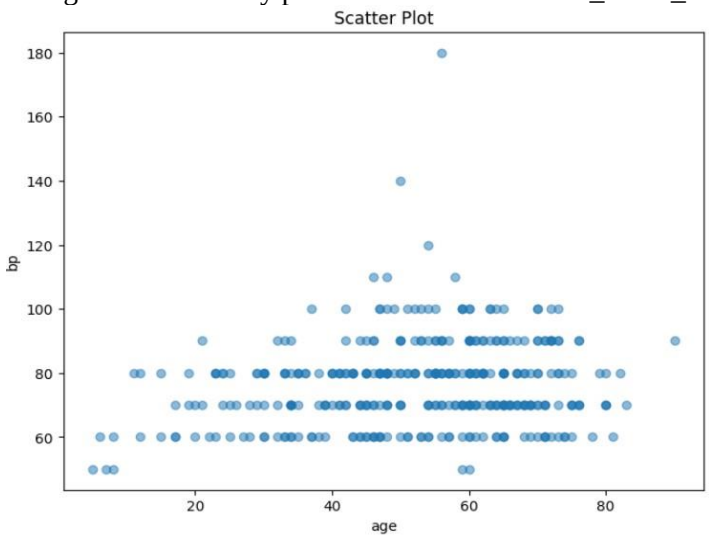

Fig.8. scatter plot for age

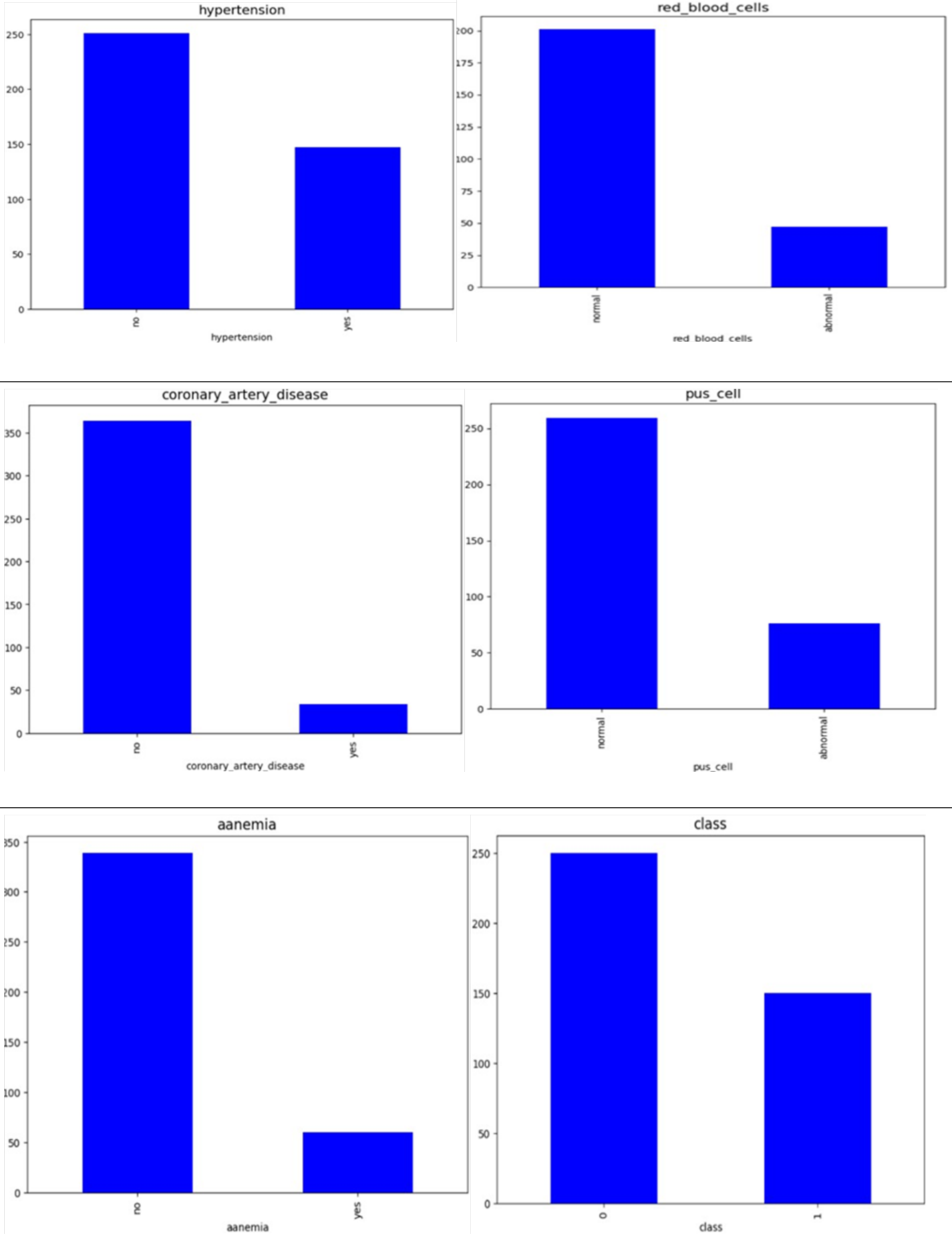

Fig.9. Value Count for features

The value counts of features are used here in “Fig 8” to count the number of occurences of each unique value.

C. Pre Processing Techniques

Pre-Processing starts by importing all the necessary libraries and by reading the dataset.

Handling Missing Values in the “Fig 9” and also removing unwanted and duplicate values from the dataset then removing them in “Fig 10” from the dataset.

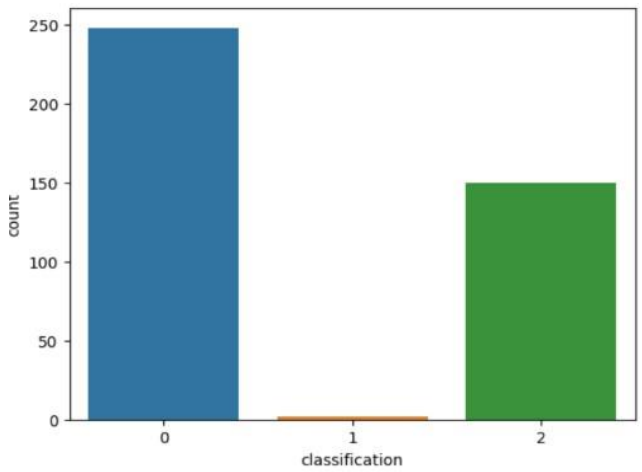

Fig.10 Before removing unwanted value (2) from the target class.

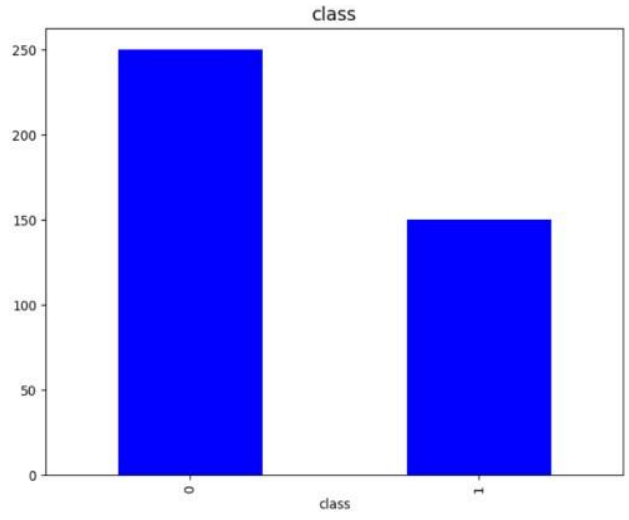

Fig.11. After removing unwanted value from the target class.

Scaling features to have a mean of 0 and a standard deviation of 1 in our CKD dataset. Converted binary variables from 0s and 1s also presenting a correlation matrix for them . Selecting features based on statistical tests like chi-square test.

Data addressed class imbalance using techniques like SMOTE. To reduce the number of features while maintaining a good analytical result. For this purpose, feature selection and features associations or correlation have been studied to remove redundant information And then Splitting the dataset into training, validation, and test sets to evaluate model performance.

Example Code :

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,
classification_report
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
model_l = LogisticRegression()
model_l.fit(X_train, y_train)
y_pred = model_l.predict(X_test)
lr_acc = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print(confusion_matrix(y_test, y_pred))
```

```
[[26  0]
 [ 0 14]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        26
           1       1.00      1.00      1.00        14

    accuracy                           1.00        40
   macro avg       1.00      1.00      1.00        40
weighted avg       1.00      1.00      1.00        40
```

D. Creation and Evaluation of Model :

Decision Tree :

When it comes to solving categorization issues, one of the most effective and widely used strategies for supervised machine learning is known as the decision tree.

Logistic Regression :

It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1.

KNN :

The KNN is a simple supervised learning approach widely applied to resolve classification and regression issues. The value of k is automatically chosen to increase the accuracy of the KNN algorithm .The KNN is a simple supervised learning approach widely applied to resolve classification and regression issues.

XG-BOOST :

XG-Boost is commonly known to offer smart solutions to structured data problems through the implementation of the gradient boosted trees technique.

In this step, a suitable machine learning algorithm is selected and trained on the prepared data. The model is trained by optimizing its parameters to minimize the difference between its predicted output and the true output in the training data. After training the model, it is evaluated on a separate validation dataset to assess its performance.The evaluation metrics used depend on the type of problem and the performance criteria. Common evaluation metrics as in "FIG 12" include accuracy, precision, recall, F1 score, and ROC curve that meant to know the performance of classification model .

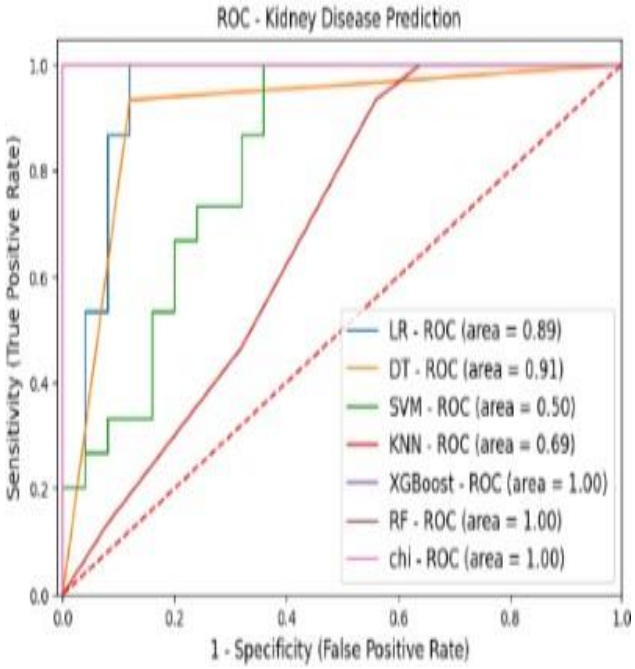

Fig.12. ROC curve for the classifiers

Based on the evaluation results, the model can be further refined by adjusting its parameters or selecting a different algorithm. This step is important for improving the model's performance on new and unseen data. Once the model is trained and evaluated, it can be deployed in a production environment to make predictions or decisions on new data. Overall, the process of creating and evaluating a machine learning model involves several steps that require careful consideration and attention to detail. By following these steps, it is possible to create models that can learn and make accurate predictions on new data.

Model Comparison: Compared the performance of different models based on evaluation metrics to identify the most effective one for kidney disease prediction.
Cross-Validation: Performed cross-validation to assess the robustness of the chosen model and ensure its generalization to unseen data.
Model Deployment: Once the best-performing model is identified, model is deployed it in a real-world setting for kidney disease prediction.

E. Result And Analysis :

The accuracy a common metric used to evaluate the performance of a Machine Learning algorithm In the TABLE2 and TABLE3 all the accuracies are shown accordingly from both existed and the proposed models. It measured the proportion of correctly classified instances among all the instances in the test dataset.The "Fig 13" and the "Fig 14" are the predicted outcomes of the model.

TABLE II. performance of various classifiers on CKD dataset based on existed paper [3]

Classifiers	Accuracy	Precision	Recall	F1-Score
KNN	0.65	0.66	0.65	0.66
XG-BOOST	0.98	0.98	0.98	0.98
Decision Tree	0.96	0.96	0.96	0.96
Random Forest	0.97	0.98	0.97	0.97
SVM	0.93	0.94	0.93	0.93

TABLE III. Accuracy of classifiers after SMOTE based on proposed model :

Classifiers	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.99	1.00	0.96	0.98
KNN	0.99	1.00	0.98	0.99
XG-BOOST	0.99	1.00	0.98	0.99
Decision Tree	0.98	1.00	0.96	0.98
Random Forest	0.98	0.93	1.00	0.98
SVM	0.98	1.00	1.00	1.00
Chi-Square	0.98	1.00	1.00	1.00

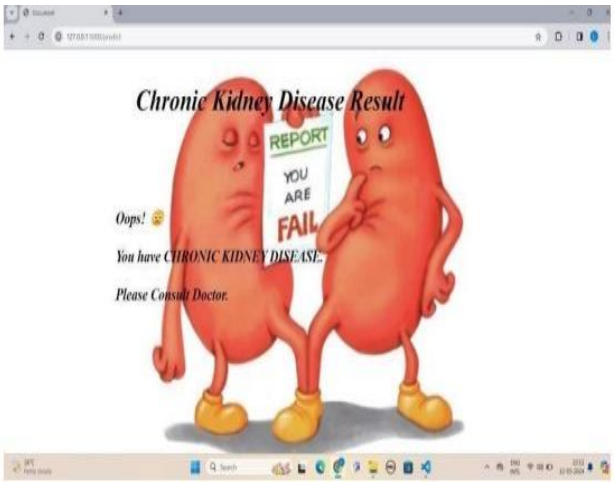

Fig.13. Prediction when CKD is predicted

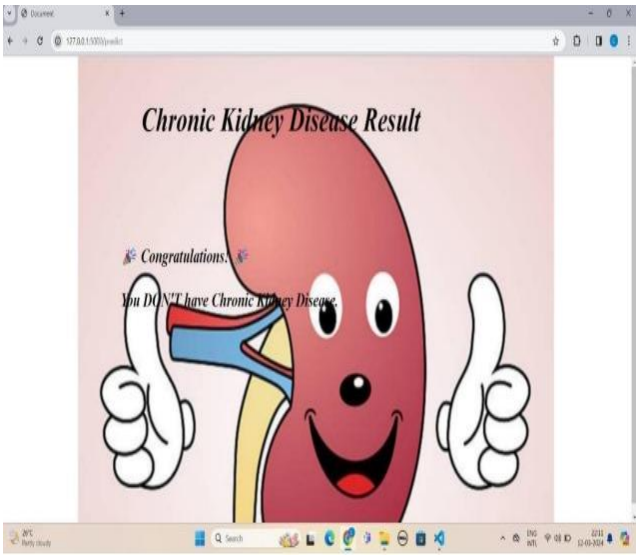

Fig.14. Prediction when CKD is not predicted

IV CONCLUSION AND FUTURE SCOPE

In conclusion the application of seven machine learning algorithms like Logistic Regression, KNN, SVM, Random Forest, Decision Tree, XG-Boost, Chi-Square for predicting the probability of kidney diseases range in a normal human body.

Although XG-Boost and Logistic Regression algorithms predicted similar accuracy scores we considered opting Logistic Regression .By applying SMOTE technique for the class balancing purpose and K-Fold to know the accuracies better at every fold. While Evaluating the Logistic Regression model that involves assessing its performance using various metrics.

These metrics help gauge how well the model distinguishes between patients with and without kidney disease. The application of machine learning techniques for kidney disease prediction shows promising results in leveraging patient data to assist in detection and management of kidney-related conditions and holds considerable promise for revolutionizing the diagnosis and management of kidney-related disorders.

V REFERENCES

[1] Saurabh Pal , Prediction for chronic kidney disease by categorical and non_categorical attributes using diferent machine learning algorithms springer Multimedia Tools and Applications (2023) 82:41253–41266

[2] Md. Ariful Islam a, Md. Ziaul Hasan Majumder b , Md. Alomgeer Hussein c Chronic kidney disease prediction based on machine learning algorithms Journal of Pathology Informatics 14 (2023) 100189 journalhomepage: www.elsevier.com/locate/jpi

[3] Jaber Qezelbash-Chamak a,∗ , Saeid Badamchizadeh b , Kourosh Eshghi c , Yasaman Asadi d A survey of machine learning in kidney disease diagnosis elsvier Machine Learning with Applications 10 (2022) 100418 journalhomepage: www.elsevier.com/locate/mlwa

[4] Sujata Drall, 2 Gurdeep Singh Drall, 3 Sugandha Singh, 4 Bharat Bhushan Naib Chronic Kidney Disease Prediction Using Machine Learning: A New Approach International Journal of Management, Technology And Engineering Volume 8, Issue V, MAY/2018

[5] Marwa Almasoud 1 Tomas E Ward2 Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors(2013)

[6] Peterson DJ, Ostberg NP, Blayney DW et al (2021) Machine learning applied to electronic health records: identification of chemotherapy patients at high risk for preventable emergency department visits and hospital admissions. JCO Clin Cancer Inform 5:1106–1126. https://doi.org/10.1200/CCI.21.00116

[7] Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh Chronic Kidney Disease Prediction using Machine Learning Models International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019

[8] Mohammed Deriche, “Feature Selection using Ant Colony Optimization”, International Multi-Conference on Systems, Signals and Devices, 2009

[9] Baisakhi Chakraborty, “Development of Chronic Kidney Disease Prediction Using Machine Learning”, International Conference on Intelligent Data Communication Technologies, 2019

[10] Rodriguez M, Salmeron MD, Martin-Malo A et al (2016) A new data analysis system to quantify associations between biochemical parameters of chronic kidney disease-mineral bone disease. PLoS ONE11(1):e0146801.
https://doi.org/10.1371/journal.pone.0146801

[11] Razib Hayat Khan, Jonayet Miah, Md Abdur Rakib Rahat, Ashiqul Haque Ahmed, Md Ahnaf Shahriyar, Ehsanur Rashid Lipu, "A Comparative Analysis of Machine Learning Approaches for Chronic Kidney Disease Detection", 2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), pp.1-6, 2023

[12] Yildirim, Pinar. "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction." In Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, vol. 2, pp. 193-198. IEEE, 2017

[13] Gunarathne, W. H. S. D., K. D. M. Perera, and K.A.D. C. P. Kahandawaarachchi. "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)." In Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on, pp. 291-296. IEEE, 2017.

[14]DatasetLink:
https://www.kaggle.com/datasets/mansoordaku/ckdisease