IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133135

Chronic Kidney Disease Detection Using Machine Learning Algorithms

Dr.V.Suganthi[1], M.Sabari[2]

Associate Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science[1]

PG Student, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science[2]

Abstract: Chronic kidney disease (CKD) is a prevalent and serious health condition that necessitates accurate and timely diagnosis for effective management and treatment. In this study, we explore the application of machine learning algorithms, specifically K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extra Trees Classifier, for predicting chronic kidney diseases. The research encompasses a comprehensive analysis of a dataset containing relevant medical information such as age, blood pressure, blood glucose levels, and serum keratinize. The dataset undergoes meticulous preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features. Feature selection techniques are employed to identify the most influential factors contributing to the prediction of chronic kidney diseases. Subsequently, the dataset is divided into training and testing sets to facilitate the training and evaluation of the machine learning models. The selected classifiers are trained on the training set, and their performances are evaluated on the testing set using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. The model with the highest performance is further fine-tuned through hyper parameter tuning to enhance its predictive capabilities. The outcomes of this research provide insights into the effectiveness of machine learning models in predicting chronic kidney diseases. The results underscore the importance of careful model selection, feature engineering, and hyper parameter tuning in optimizing predictive performance. The developed model holds promise for aiding healthcare professionals in early detection and management of chronic kidney diseases, potentially improving patient outcomes and reducing healthcare costs. However, the deployment of such models in real-world healthcare settings should be approached with consideration of ethical implications and domain-specific nuances.

Keywords: Kidney, Machine learning algorithms, Average accuracy, blood vessels

I. INTRODUCTION

Chronic kidney disease (CKD) represents a significant global health challenge, with a rising prevalence and substantial impact on morbidity and mortality. Timely and accurate identification of individuals at risk or in the early stages of CKD is crucial for implementing effective interventions and improving patient outcomes. Traditional diagnostic approaches often rely on clinical assessments, but the integration of machine learning techniques has shown promise in enhancing predictive accuracy. In this era of advanced data analytics, machine learning models have demonstrated their potential to contribute significantly to medical decision-making processes. This study focuses on exploring the application of machine learning algorithms, including KNeighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extra Trees Classifier, to predict the occurrence of chronic kidney diseases based on a diverse set of medical indicators. The rationale behind employing machine learning in this context lies in its ability to discern complex patterns within large datasets, uncovering relationships that may elude traditional analytical methods. By leveraging features such as age, blood pressure, blood glucose levels, and serum keratinise, these models aim to provide accurate predictions and assist healthcare professionals in identifying individuals susceptible to or currently experiencing CKD. The objectives of this research encompass dataset preparation, model training, evaluation, and optimization. A thorough data pre-processing phase addresses issues such as missing values and variable scaling, ensuring that the dataset is conducive to effective model training. Feature selection techniques are employed to identify the most influential variables, shedding light on the critical factors contributing to the prediction of CKD. The study further involves the training of machine learning models on a carefully curated dataset, followed by a comprehensive evaluation of their predictive performance. Performance metrics such as accuracy, precision, recall, F1 score, and ROC-AUC are utilized to assess the effectiveness of each model. Subsequently, the best-performing model undergoes hyper parameter tuning to refine its predictive capabilities. By the conclusion of this research, we aim to contribute valuable insights into the applicability and effectiveness of machine learning models in predicting chronic kidney diseases. The potential benefits of such models include early detection, personalized treatment plans, and improved patient outcomes.

DOI: 10.17148/IJARCCE.2024.133135

However, we also acknowledge the ethical considerations associated with deploying predictive models in healthcare and emphasize the importance of interpreting results within the broader context of clinical practice. Through this exploration, we aspire to pave the way for advancements in predictive healthcare analytics and contribute to the ongoing efforts to combat chronic kidney diseases globally.

II. RELATED WORKS

Traditional methods for diagnosing and predicting chronic kidney diseases (CKD) have largely relied on established clinical scoring systems, statistical models, and rule-based approaches. Clinical scoring systems like CKD-EPI and MDRD have been widely accepted in medical practice, providing standardized estimations based on readily available clinical data. Statistical models, such as logistic regression, have been employed for their simplicity and interpretability, offering insights into the impact of different variables on CKD prediction. Rule-based systems, often represented as decision trees, contribute to transparent decision-making by providing explicit paths for predictions.

While these methods have served as valuable tools, the emergence of machine learning approaches introduces a paradigm shift, leveraging complex pattern recognition and adaptability to diverse data types for more accurate and personalized predictions of CKD.

A part of investigate has been done to anticipate Parkinson’s illness in a persistent, but less work has been detailed to anticipate its seriousness. These works have utilized different machine learning strategies. In a overview by Das et al. [1] on the application of different classification strategies in diagnosing the Parkinson’s disease(PD), neural arrange was found as the superior classifier compared to relapse and choice tree.

In most of the detailed inquire about, the highlights extricated from discourse signals [6][7][15] are utilized for anticipating the seriousness of PD.Genain et al. [2] utilized Packed away choice trees to foresee the PD seriousness from voice recordings of patients and found an advancement of 2% precision. Maleket al.[3] utilized 40-features dataset and recognized 9 best highlights utilizing Neighborhood Learning Based Include Determination (LLBFS) to classify PD subjects into four classes (Healthy, Early, Middle of& the based on their UPDRS score. Cole et al.[4] investigated the utilize of energetic machine learning calculations for distinguishing the seriousness of tremors and Dyskinesia from the information collected from wearable sensors. Angeles et al.[5] created a sensor framework to record motor information from the arm in arrange to survey side effect seriousness changes amid Profound Brain Reenactment Treatment. Nilashiet al.[8] proposed a unused crossover brilliantly framework utilizing Versatile neuro fluffy deduction system(ANFIS) and Bolster Vector Regression(SVR) for foreseeing the PD movement. Chen et al.[9] proposed a PD demonstrative framework utilizing PCA for include extraction and Fluffy KNN for classification . Polat[10] proposed a show utilizing Fluffy C-Means (FCM) clustering and KNN to analyze the PD. Åström and Koker[11] outlined a PD expectation framework utilizing parallel bolster forward Neural Arrange and after that yield is compared against a rule-based framework for making the ultimate choice. Li et al.[12], proposed a fluffy based nonlinear change strategy where PCA is utilized for include extraction and SVM for PD expectation. Hariharanet al.[13] proposed a crossover brilliantly framework utilizing clustering, highlight lessening and classification strategies for precise PD determination.

III. METHODOLOGY

The primary objectives of employing machine learning algorithms, including K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extra Trees Classifier, in predicting chronic kidney diseases are to enhance the accuracy and efficiency of early diagnosis. By leveraging these algorithms on comprehensive datasets comprising vital medical indicators, the aim is to develop robust predictive models capable of discerning complex patterns indicative of chronic kidney diseases.

This includes automating the identification of relevant features, optimizing predictive performance through hyper parameter tuning, and ultimately providing healthcare professionals with reliable tools for early detection and personalized intervention strategies. The overarching goal is to contribute to improved patient outcomes by facilitating timely and accurate identification of individuals at risk or currently experiencing chronic kidney diseases.

DOI: 10.17148/IJARCCE.2024.133135

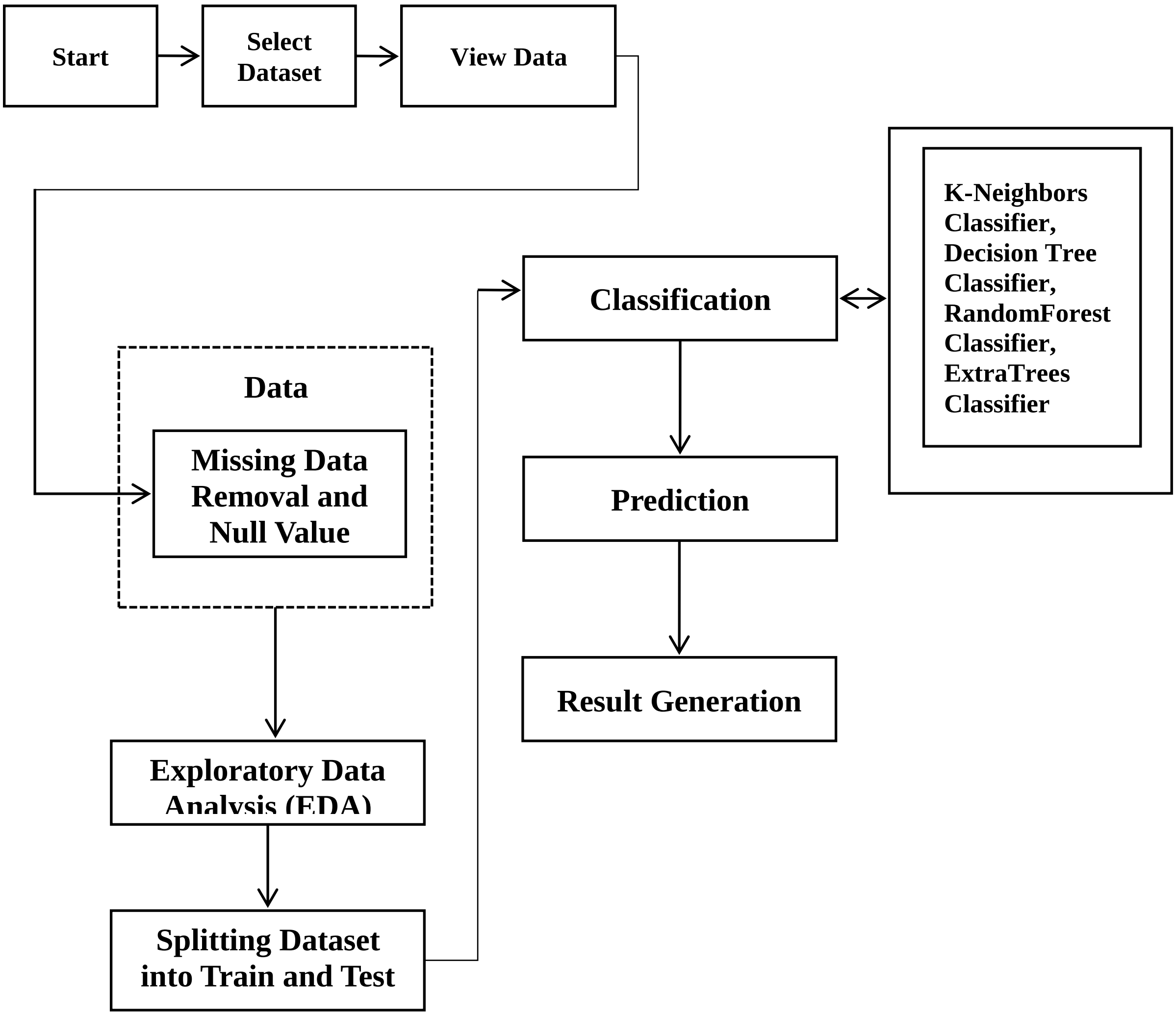

Start
Select Dataset
View Data
Data
Missing Data Removal and Null Value
Exploratory Data Analysis (EDA)
Splitting Dataset into Train and Test
Classification
K-Neighbors Classifier, Decision Tree Classifier, RandomForest Classifier, ExtraTrees Classifier
Prediction
Result Generation

Figure-1 System Architecture

Proposed Algorithm

Three different machine learning algorithms has taken for simple analysis. the three algorithms are then compared with some quality metrics such as Classification Ratio, Detection Ratio and Malicious Ratio. The experimental results from weka tools are displayed for analyzing different results.

3.1 K-Nearest Neighbors Algorithm

In design acknowledgment, the k-Nearest Neighbors algorithm (or k-NN for short) is a nonparametric technique utilized for classification and regression.[1] In the two cases, the info comprises of the k nearest preparing models in the component space. The yield relies upon whether k-NN is utilized for classification or regression: In k-NN classification, the yield is a class participation. An article is grouped by a dominant part vote of its neighbors, with the item being appointed to the class generally basic among its k nearest neighbors (k is a positive whole number, ordinarily little). In the event that k = 1, at that point the article is basically relegated to the class of that solitary nearest neighbor. In k-NN regression, the yield is the property estimation for the article. This worth is the normal of the estimations of its k nearest neighbors. k-NN is a kind of occasion based learning, or languid learning, where the capacity is just approximated locally and all calculation is conceded until classification. The k-NN algorithm is among the most straightforward of all machine learning algorithms.

DOI: 10.17148/IJARCCE.2024.133135

3.2 Naive Bayes:

It is a likelihood based classification procedure. It considers all highlights autonomous of one another. It computes likelihood of each element freely for a specific class mark. Naïve Bayes is utilized in this paper for malware prediction utilizing web traffic data. These are the means behind the Naïve Bayes algorithm:

1. Preparing data set is taken as information.
2. Highlights are extricated from that preparation data. In this paper web traffic data comprises of 43 highlights.
3. At that point from the preparation data for each component Naïve bayes figures likelihood that in the event that element has specific worth, at that point the dataset class be will malicious or not.
4. In the event that each component has constrained potential qualities, at that point above probabilities can be determined. Be that as it may, if the huge number of qualities is there for each element, scope of qualities can likewise be taken.
5. At that point for each line of test data set after the preparation stage. Based all things considered probabilities determined from preparing data decision is taken.

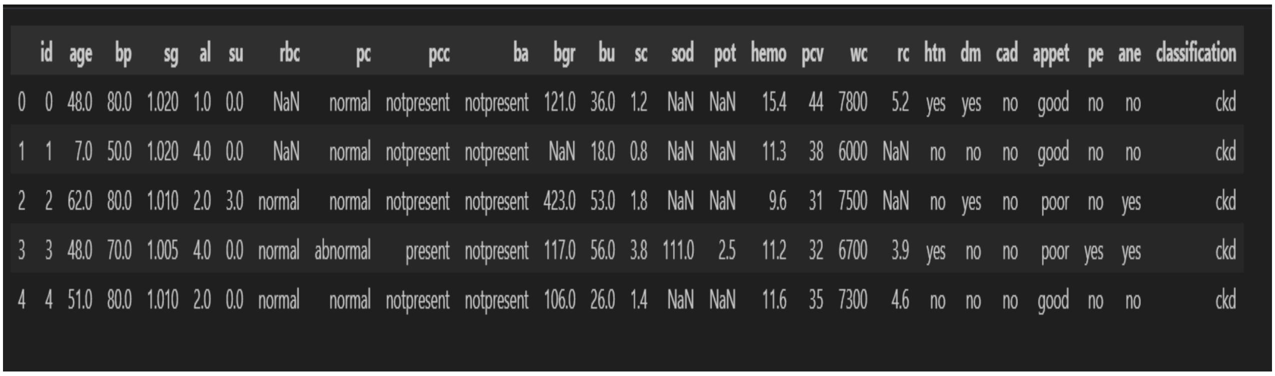
3.3 Decision tree:

This kind of classifier models data with the assistance of a tree. Tree is having highlights as the inward hubs and edges show the estimations of highlights. And edges isolated hubs dependent on the qualities.

All the leaf hubs of the decision tree speaks to a class which is relied upon to be acquired on the off chance that we have every one of the highlights having particular qualities which are in the way from the root to that class having middle of the road include hubs.

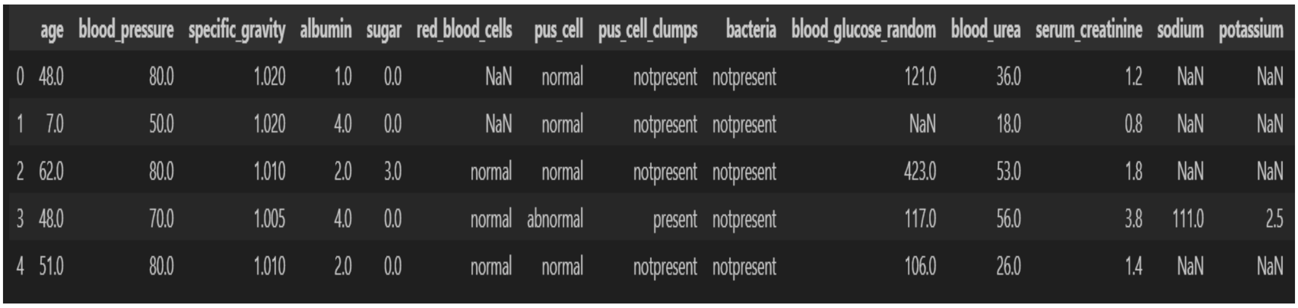
Probably the most mainstream decision tree algorithms are ID3, C4.5, CART. ID3 is one of the most straightforward decision tree approaches it utilizes idea of data gain as the parting criteria. C4.5 is the development of ID3. It takes a shot at the guideline of addition proportion.

IV. RESULTS AND DISCUSSION



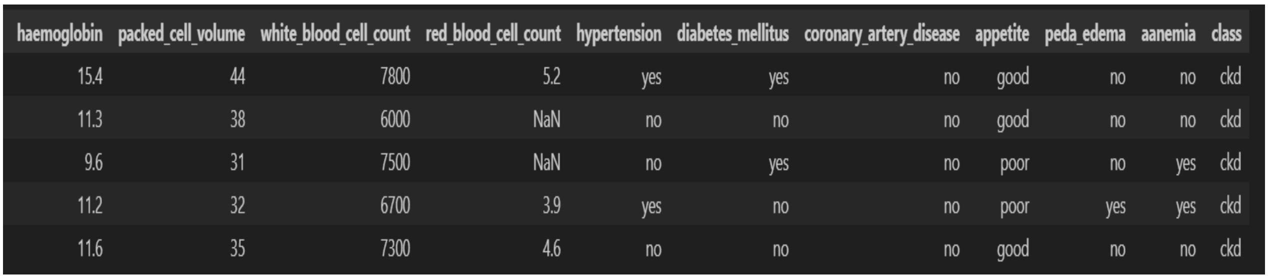
	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	11.3	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	9.6	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	11.6	35	7300	4.6	no	no	no	good	no	no	ckd

Fig-2 Sample Dataset



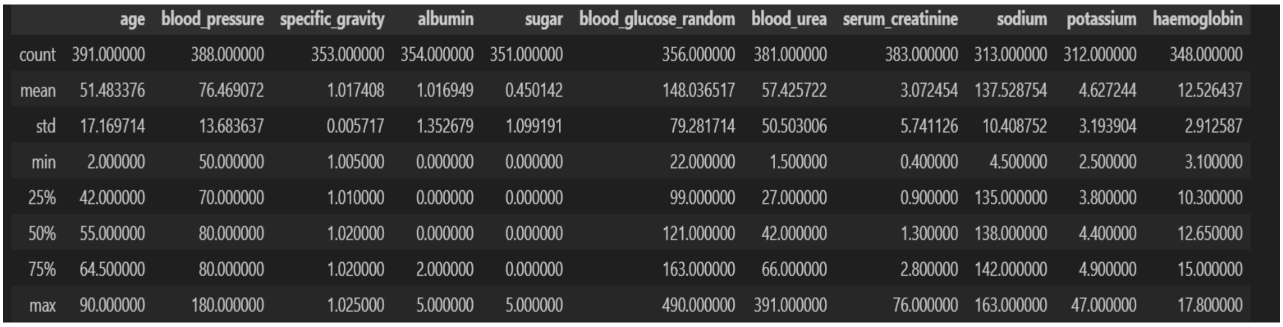
	age	blood_pressure	specific_gravity	albumin	sugar	red_blood_cells	pus_cell	pus_cell_clumps	bacteria	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN

Fig-3 Sample Dataset



haemoglobin	packed_cell_volume	white_blood_cell_count	red_blood_cell_count	hypertension	diabetes_mellitus	coronary_artery_disease	appetite	peda_edema	aanemia	class
15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
11.3	38	6000	NaN	no	no	no	good	no	no	ckd
9.6	31	7500	NaN	no	yes	no	poor	no	yes	ckd
11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
11.6	35	7300	4.6	no	no	no	good	no	no	ckd

Fig-4 Attributes Calculation



	age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin
count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

Fig-5 Attributes Calculation

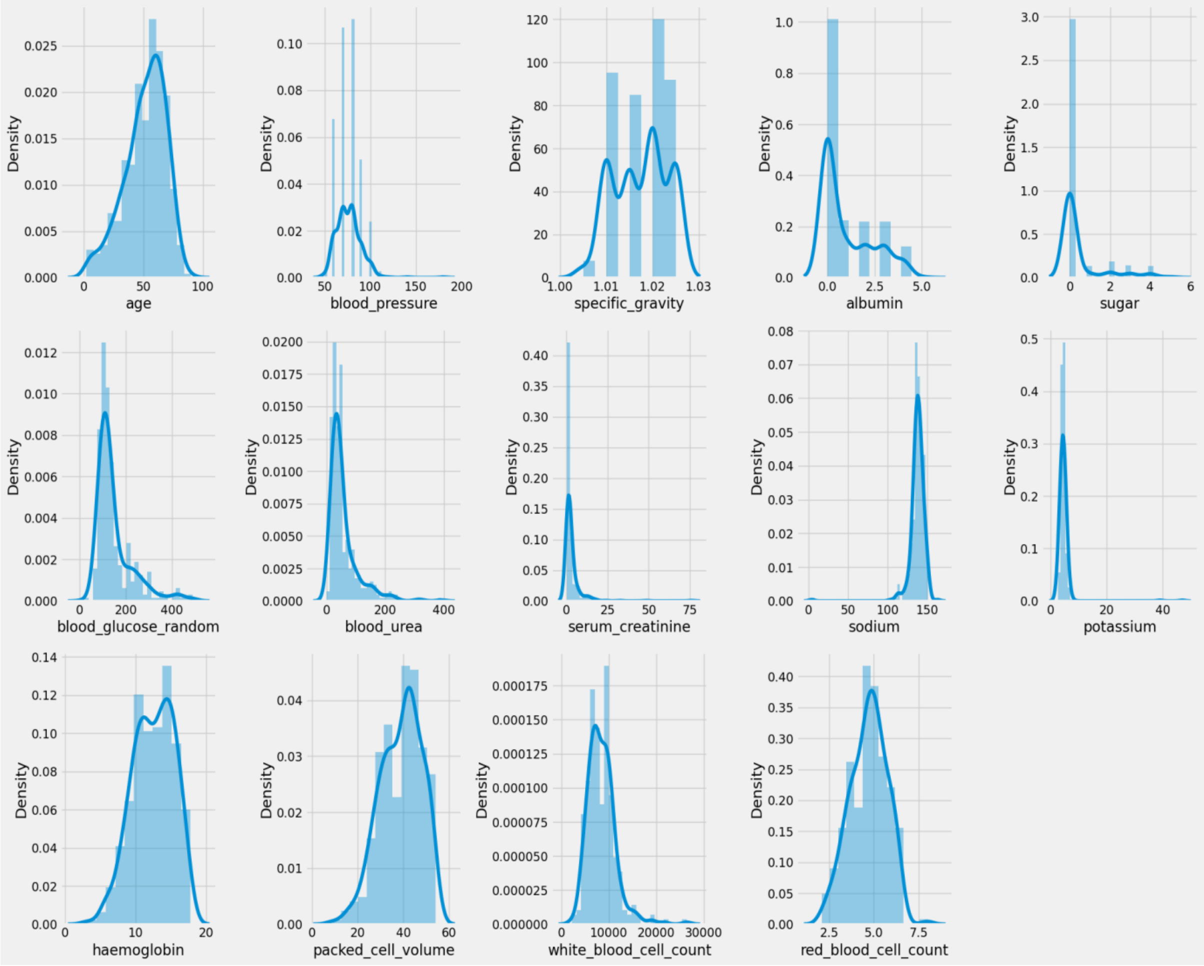
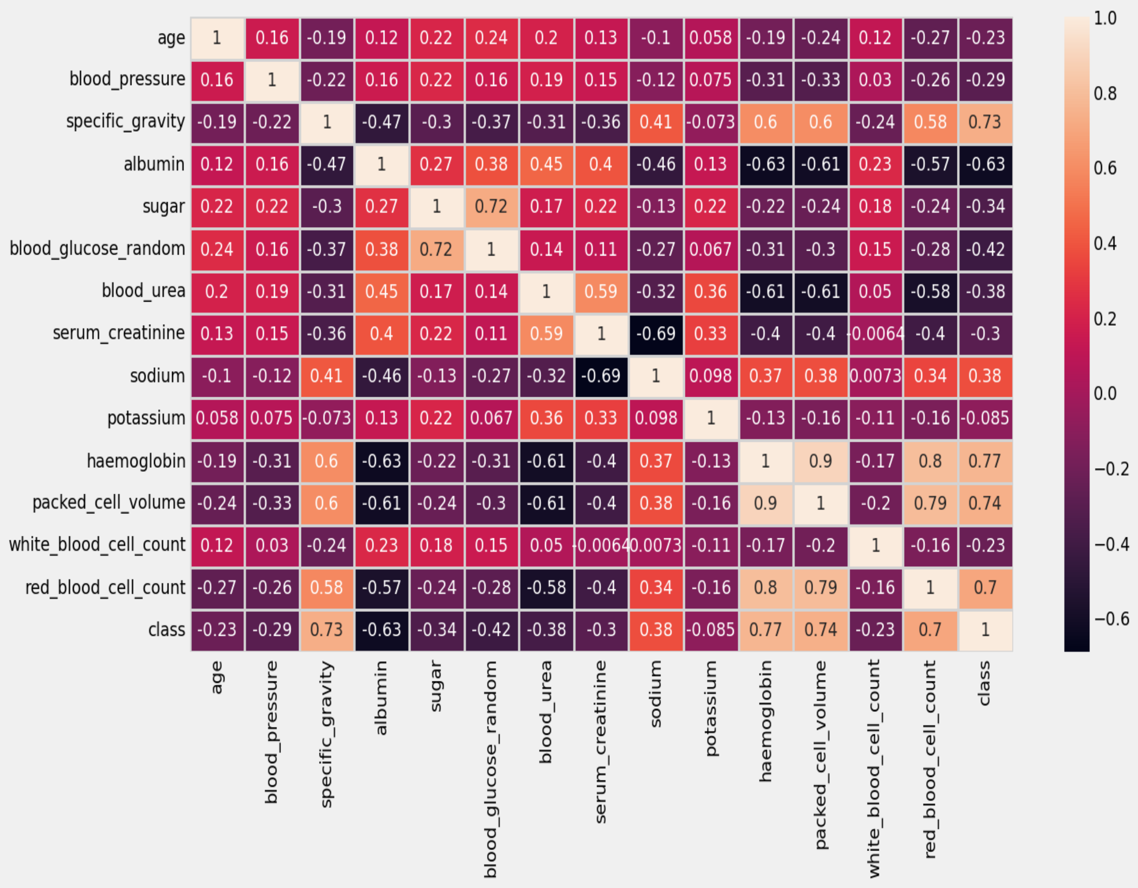


Fig-6 Histogram Analysis



	age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin	packed_cell_volume	white_blood_cell_count	red_blood_cell_count	class
age	1	0.16	-0.19	0.12	0.22	0.24	0.2	0.13	-0.1	0.058	-0.19	-0.24	0.12	-0.27	-0.23
blood_pressure	0.16	1	-0.22	0.16	0.22	0.16	0.19	0.15	-0.12	0.075	-0.31	-0.33	0.03	-0.26	-0.29
specific_gravity	-0.19	-0.22	1	-0.47	-0.3	-0.37	-0.31	-0.36	0.41	-0.073	0.6	0.6	-0.24	0.58	0.73
albumin	0.12	0.16	-0.47	1	0.27	0.38	0.45	0.4	-0.46	0.13	-0.63	-0.61	0.23	-0.57	-0.63
sugar	0.22	0.22	-0.3	0.27	1	0.72	0.17	0.22	-0.13	0.22	-0.22	-0.24	0.18	-0.24	-0.34
blood_glucose_random	0.24	0.16	-0.37	0.38	0.72	1	0.14	0.11	-0.27	0.067	-0.31	-0.3	0.15	-0.28	-0.42
blood_urea	0.2	0.19	-0.31	0.45	0.17	0.14	1	0.59	-0.32	0.36	-0.61	-0.61	0.05	-0.58	-0.38
serum_creatinine	0.13	0.15	-0.36	0.4	0.22	0.11	0.59	1	-0.69	0.33	-0.4	-0.4	-0.0064	-0.4	-0.3
sodium	-0.1	-0.12	0.41	-0.46	-0.13	-0.27	-0.32	-0.69	1	0.098	0.37	0.38	0.0073	0.34	0.38
potassium	0.058	0.075	-0.073	0.13	0.22	0.067	0.36	0.33	0.098	1	-0.13	-0.16	-0.11	-0.16	-0.085
haemoglobin	-0.19	-0.31	0.6	-0.63	-0.22	-0.31	-0.61	-0.4	0.37	-0.13	1	0.9	-0.17	0.8	0.77
packed_cell_volume	-0.24	-0.33	0.6	-0.61	-0.24	-0.3	-0.61	-0.4	0.38	-0.16	0.9	1	-0.2	0.79	0.74
white_blood_cell_count	0.12	0.03	-0.24	0.23	0.18	0.15	0.05	-0.0064	0.0073	-0.11	-0.17	-0.2	1	-0.16	-0.23
red_blood_cell_count	-0.27	-0.26	0.58	-0.57	-0.24	-0.28	-0.58	-0.4	0.34	-0.16	0.8	0.79	-0.16	1	0.7
class	-0.23	-0.29	0.73	-0.63	-0.34	-0.42	-0.38	-0.3	0.38	-0.085	0.77	0.74	-0.23	0.7	1

Fig-7 Confusion Matrix

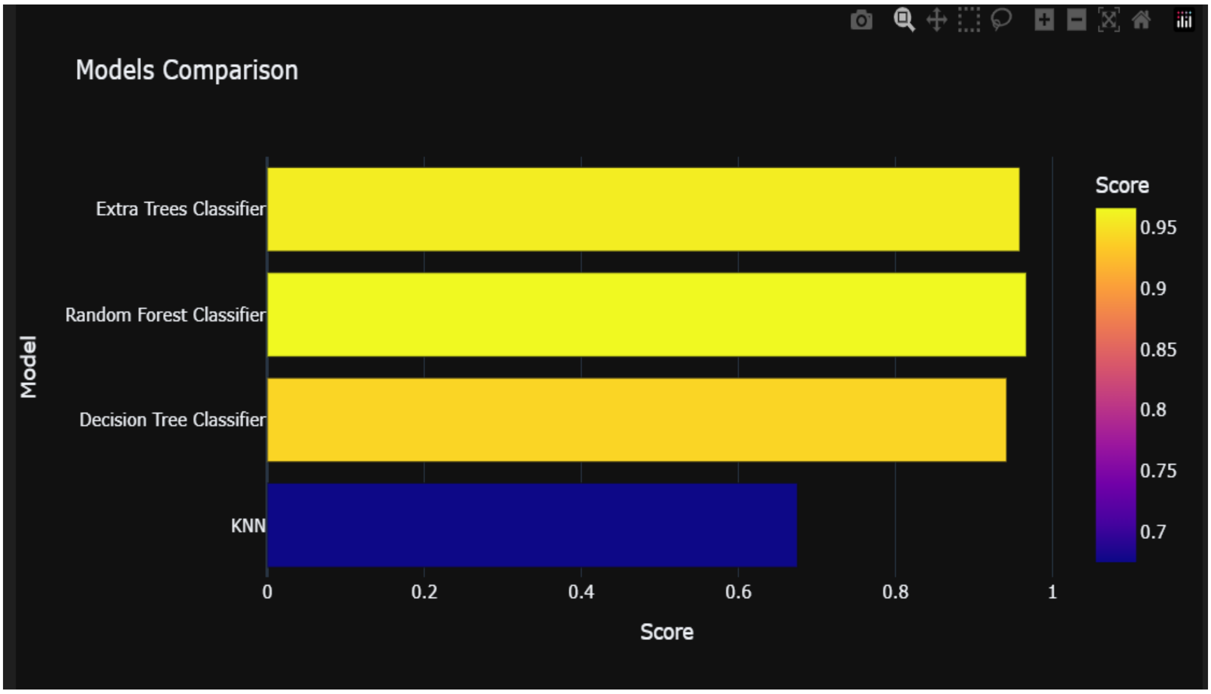
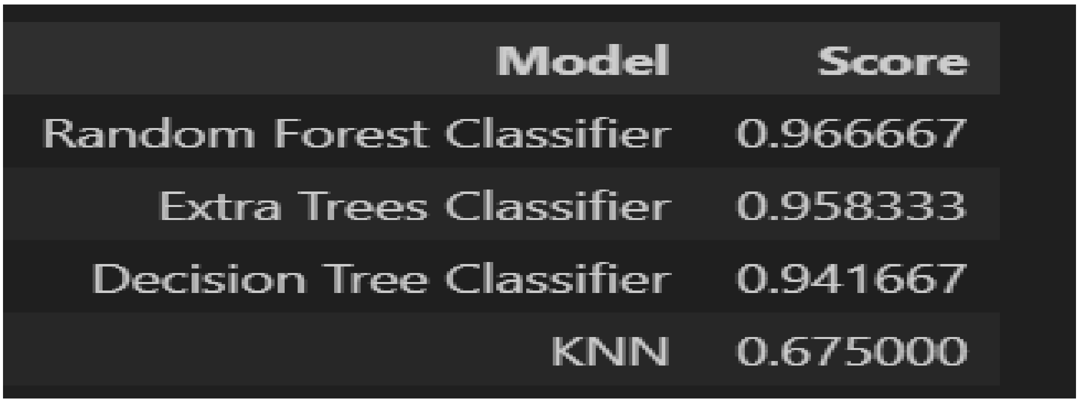


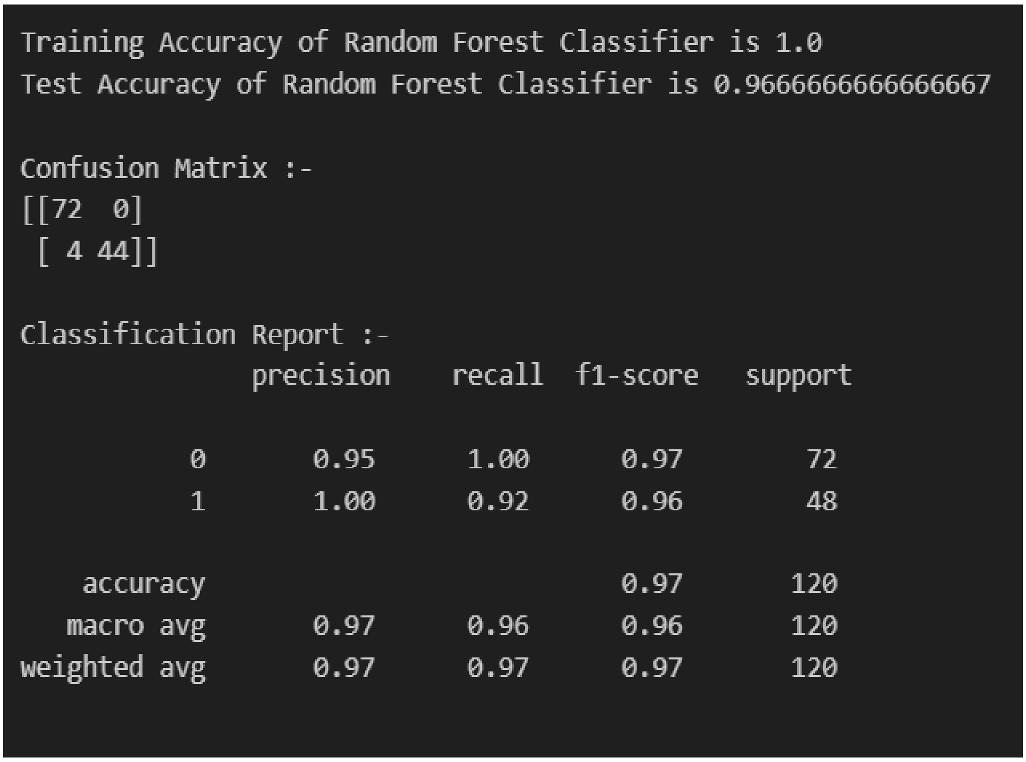
Fig-8 Model Comparison

DOI: 10.17148/IJARCCE.2024.133135



Model	Score
Random Forest Classifier	0.966667
Extra Trees Classifier	0.958333
Decision Tree Classifier	0.941667
KNN	0.675000

Fig-9 Model Comparison – Accuracy



```
Training Accuracy of Random Forest Classifier is 1.0
Test Accuracy of Random Forest Classifier is 0.9666666666666667

Confusion Matrix :-
[[72  0]
 [ 4 44]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.95      1.00      0.97        72
           1       1.00      0.92      0.96        48

    accuracy                           0.97       120
   macro avg       0.97      0.96      0.96       120
weighted avg       0.97      0.97      0.97       120
```

Fig-11 Model Comparison – Accuracy

V. CONCLUSION

In conclusion, the application of machine learning algorithms, including K- Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extra Trees Classifier, in predicting chronic kidney diseases represents a significant advancement in healthcare analytics. Through a systematic methodology involving data collection, pre-processing, model training, and evaluation, we have explored the efficacy of these classifiers in differentiating between individuals with and without CKD based on key medical indicators. The results demonstrate varying levels of predictive performance, with ensemble methods like Random Forest Classifier and Extra Trees Classifier exhibiting notable accuracy. The automated feature identification and adaptability to complex patterns in machine learning models contribute to improved predictive capabilities. However, it is essential to acknowledge the interpretability challenges associated with certain models and the ethical considerations in deploying predictive tools in clinical practice. As technology continues to evolve, striking a balance between innovation, transparency, and responsible implementation will be crucial in harnessing the full potential of machine learning for enhancing chronic kidney disease prediction and, ultimately, improving patient outcomes.

DOI: 10.17148/IJARCCE.2024.133135

REFERENCES

[1] Shreya S. Bhanose, Kalyani A. Bogawar (2016) “Crop And Yield Prediction Model”, International Journal of Advance Scientific Research and Engineering Trends, Volume 1,Issue 1, April 2016

[2] Tripathy, A. K., et al.(2011) "Data mining and wireless sensor network for agriculture pest/disease predictions." Information and Communication Technologies (WICT), 2011

[3] Ramesh Babu Palepu (2017) ” An Analysis of Agricultural Soils by using Data Mining Techniques”, International Journal of Engineering Science and Computing, Volume 7 Issue No. 10 October.

[4] Rajeswari and K. Arunesh (2016) “Analysing Soil Data using Data Mining Classification Techniques”, Indian Journal of Science and Technology, Volume 9, May.

[5] A.Swarupa Rani (2017), “The Impact of Data Analytics in Crop Management based on Weather Conditions”, International Journal of Engineering Technology Science and Research, Volume 4,Issue 5,May.

[6] Pritam Bose, Nikola K. Kasabov (2016), “Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series”, IEEE Transactions On Geoscience And Remote Sensing.

[7] Priyanka P.Chandak (2017),” Smart Farming System Using Data Mining”, International Journal of Applied Engineering Research, Volume 12, Number 11.

[8] Vikas Kumar, Vishal Dave (2013), “KrishiMantra: Agricultural Recommendation System”, Proceedings of the 3rd ACM Symposium on Computing for Development, January.

[9] Savae Latu (2009), ”Sustainable Development : The Role Of Gis And Visualisation”, The Electronic Journal on Information Systems in Developing Countries, EJISDC 38, 5, 1-17.

[10] Nasrin Fathima.G (2014), “Agriculture Crop Pattern Using Data Mining Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, May.

[11] Ramesh A.Medar (2014), ”A Survey on Data Mining Techniques for Crop Yield Prediction”, International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 9, September.

[12] Shakil Ahamed.A.T.M, Navid Tanzeem Mahmood (2015),” Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh”, ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),IEEE,June.

[13] Napoleon D. and Praneesh M. “Detection of Brain Tumor using Kernel Induced Possiblistic C-Means Clustering”, volume no.3, issue no.9, pp 436-438, 2013

[14] Shreya S.Bhanose (2016),”Crop and Yield Prediction Model”, International Journal of Advence Scientific Research and Engineering Trends, Volume 1,Isssue 1,ISSN(online) 2456- 0774,April.

[15] Agaj i Iorshase, Onyeke Idoko Charles,”A Well-Built Hybrid Recommender System for Agricultural Products in Benue State of Nigeria”, Journal of Software Engineering and Applications,2015,8,581-589

[16] G. Adomavicius and A. Tuzhilin(2005), “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June.

[17] Avinash Jain, Kiran Kumar (2016),”Application of Recommendation Engines in Agriculture”, International Journal of Recent Trends in Engineering & Research, ISSN: 2455-1457.

[18] Kiran Shinde (2015),”Web Based Recommendation System for farmers”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume 3,Issue 3, ISSN:2321- 8169,March.

[19] Konstantinos G. Liakos, “ Machine Learning in Agriculture: A Review”, Sensors 2018, 18, 2674; doi:10.3390/s18082674

[20] S.Vaishnavi, M.Shobana, N Geethanjali, Dr.S.Karthik, “Data Mining: Solving the Thirst of Recommendations to Users”, IOSR Journal of Computer Engineering IOSR-JCE), Vol.16, no.6, 2014.