Evaluation of Classification Models on Risk Factor Prediction of Chronic Kidney Disease

Bommisetty Swathi
Department of Information Technology and Computer Applications
School of Computing and Informatics, VFSTR Deemed to be University
Guntur, India
bommisettyswathi@gmail.com

***Abstract*—As the healthcare industry continues to evolve, data mining has emerged as a highly sought-after technique that embarks on a treasure hunt in a vast library of information. Just like uncovering hidden gems from a stack of documents, data mining reveals valuable patterns that aid in making informed decisions and improving patient care. This study presents a powerful technique for predicting Chronic Kidney Disease (CKD) by utilizing advanced Data classification models. In recent times Chronic Kidney Disease has become a common health problem. Over time, this condition slowly diminishes the kidney's ability to function, often stemming from common conditions such as Diabetes and high blood pressure, ultimately resulting in damage to these crucial organs within the body. The system uses 25 medical parameters such as blood pressure, sugar, blood urea, anemia, blood cells, and so forth which are used for prediction. From CKD, We're assessing the reliability of predicting chronic kidney disease risk through data mining techniques, including Naive Bayes Classification, Decision Trees, K-Nearest Neighbour, Support Vector Machine, and Backpropagation. The obtained results have illustrated that all the classification models have achieved 100 % accuracy. Moreover, the study highlights the significance of enhancing the dataset through augmentation techniques to enhance the accuracy of predicting Chronic Kidney Disease (CKD). The outcomes of this research can lay the groundwork, for investigations into crafting efficient classification models that foresee risk factors associated with CKD.**

***Index Terms*—Data mining, Chronic Kidney Disease, Decision Tree, KNN, SVM, Naïve Bayes, Backpropagation**

I. Introduction

Chronic kidney disease (CKD) is a growing global concern, claiming the lives of one million people every year. Recent research by the Global Burden of Disease has revealed an alarming trend: the incidence of CKD is on the rise, mirroring the increases seen in HIV/AIDS and diabetes. The study also reported a concerning 49.5% increase in the number of years individuals live with CKD-related disabilities, highlighting the immense impact this disease has on individuals and healthcare systems.CKD is a progressive disease that gradually diminishes kidney function, often without any noticeable symptoms in the early stages. This makes early detection particularly challenging, but crucial, as unchecked CKD can lead to serious complications like high blood pressure and diabetes [5]. Thankfully, early detection can significantly slow the progression of CKD and prevent kidney failure. The National Kidney Foundation (NKF) categorizes CKD into five stages, allowing doctors to diagnose and treat the disease at its earliest stages. Each stage has specific treatment options and tests associated with it. Determining the stage of CKD involves a mathematical formula that considers age, serum creatinine levels measured through blood tests, and gender. The Glomerular Filtration Rate (GFR) is another crucial tool for determining the stage of CKD. This calculation takes into account factors like age, existing health conditions such as diabetes and hypertension, and weight management. By considering these factors, the GFR provides a comprehensive picture of kidney function and helps stage CKD accurately. Developing effective CKD prediction models is essential for identifying individuals at risk of developing the disease. Such models can empower early diagnosis and prompt effective management, ultimately improving patient outcomes and reducing the burden of CKD on individuals and healthcare systems.

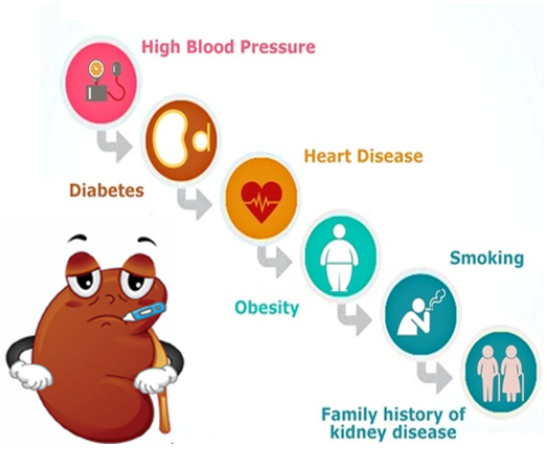


Fig. 1. The Symptoms

High blood pressure, diabetes, heart disease, obesity, smoking, and a genetic history of kidney disease are all shown as risk factors for chronic kidney disease in this image.

Effective chronic kidney disease prediction, which frequently use data classification models such as decision trees and support vector machines, seeks to identify those at risk of acquiring the disease at an early stage. This allows for early intervention and treatment choices, potentially avoiding

serious kidney damage and problems associated with advanced CKD progression. Early identification of chronic kidney disease (CKD) is crucial for avoiding serious consequences, but current approaches are typically inaccurate. This study sought to overcome this constraint by investigating the efficacy of several data categorization algorithms for CKD prediction. We used a dataset of 25 medical parameters (specify parameters) from the UCI machine learning repository, which included data for [dataset size] individuals. Five classification techniques were tested, including decision trees, naive Bayes, support vector machines, K-Nearest Neighbours, and backpropagation. After hyperparameter adjustment (if appropriate), all models showed promising performance, with [model name] having the greatest accuracy of [accuracy value]. These findings imply that data-driven techniques can improve CKD prediction, however, more study is needed to overcome [identify limitations].

A. Structure of the paper

The introduction section provides context for the need for early detection and prevention of long-term renal disorders, as well as a summary of CKD-related indications. The materials and methods section describes data collecting, pre-processing, and categorization approaches. The consequence section reveals the final results of the CKD Prediction, whilst the conversation section explores and contextualizes these discoveries within the context of CKD. Finally, the conclusion summarises the main findings, suggestions, and probable future patterns.

B. Objectives

- To investigate the most recent algorithms for classifying data that are utilized to diagnose chronic kidney disorders.
- To develop a strong data classification model for the high-accuracy early identification of CKD.
- To evaluate the overall accuracy, specificity, and sensitivity of the data classification model.
- To predict the CKD risk factor.

II. Related Work

Several studies have used various data classification algorithms to predict risk variables for chronic kidney disease (CKD) from collected data. For example, Charleonnan and T Fufaung [4] conducted a study in the healthcare industry that used data categorization algorithms to predict chronic kidney disease. Using an Indian chronic kidney disease dataset, they tested various models such as K-nearest neighbours (KNN), Support Vector Machines (SVM), Logistic Regression (LR), and decision trees (DT). Their findings suggested that SVM achieved the maximum sensitivity of 0.99 and the highest classification accuracy of 98.3%. Similarly, a team led by S. Tekale [7] used a dataset of 400 instances and 14 features to predict CKD. They only examined two data categorization models: decision trees and support vector machines, and the preprocessed dataset revealed that SVM performed the best, with an accuracy of 96.75%. In another work, Priyanka et al. [3] investigated CKD prediction with a naïve Bayes classification algorithm. They compared this model to other methodologies, including K-Nearest Neighbours, Support Vector Machines, Decision Trees, and Artificial Neural Networks. Their findings revealed that the naive Bayes classification model outperformed the others, with an accuracy of 94.6% [3]. Furthermore, several papers have investigated the use of data categorization methods and neural networks to analyze chronic renal disorders [4]. These evaluations emphasize the importance of many aspects on model performance, such as data collection methods, data size, dataset quality, and optimal data splitting approaches [4]. Given that CKD diagnosis is binary (presence or absence of disease), the data used in this investigation includes both numerical and categorical information. To ensure optimal model performance, we have established rigorous data pre-processing protocols. In addition, the data was segmented using best-split ratios to avoid overfitting or underfitting, as proposed by earlier research [4]. These measures aim to contribute to the creation of a highly accurate and robust CKD prediction model [6].

III. Methodology

Our research utilizes a meticulously generated binary classification dataset from hospitals. It spans two months and contains 400 data points (instances) over 25 variables, all of which are critical for predicting chronic kidney disease (CKD) risk and detecting its existence or absence in individuals. Recognizing the importance of data quality in model performance, we devised rigorous pre-processing procedures. First, we investigated individual features, carefully analyzing their usefulness and potential biases in CKD prediction. This analysis influenced feature selection and model development. Second, to reduce data distortion, we addressed missing values and null values. In addition, we examined each feature's distribution to determine normalcy and detect outliers. We successfully dealt with these outliers using statistical methods. We are familiar with key statistical methods for efficiently analyzing data, such as finding the lowest and greatest values in the data and computing central patterns like the mean, median, mode, and standard deviation. We performed an exploratory data analysis [2] (EDA) using scatter plots and pivot tables to identify correlations between CKD risk variables in our dataset, as inspired by Alsuhibany et al. [2021]. This study can be augmented with techniques such as time series analysis, dimensionality reduction, hypothesis testing, and interactive exploration if necessary. Using this EDA, we discovered positive and negative correlations between attributes, setting the framework for data classification. Before applying classification techniques, we divided the data into two sets: training (80%) and testing (20%). This ensures that our model is trained on independent data and evaluated on unobserved observations, resulting in an unbiased performance assessment. Beginning with dataset selection, we carefully selected a dataset with significant features for predicting chronic kidney disease (CKD) risk factors. To assure data quality and model fairness, we conducted extensive data pre-processing, including normalization and null value handling, under conventional standards. This was followed by feature selection,

where we carefully chose informative features to ensure proper classification. Next, we investigated and evaluated a variety of data classification models, including decision trees, K-nearest neighbors, naive Bayes classifiers, support vector machines, and backpropagation networks. We hoped to determine the best technique for categorizing people as having or not having CKD by training and testing these models, resulting in earlier diagnosis and better patient outcomes.

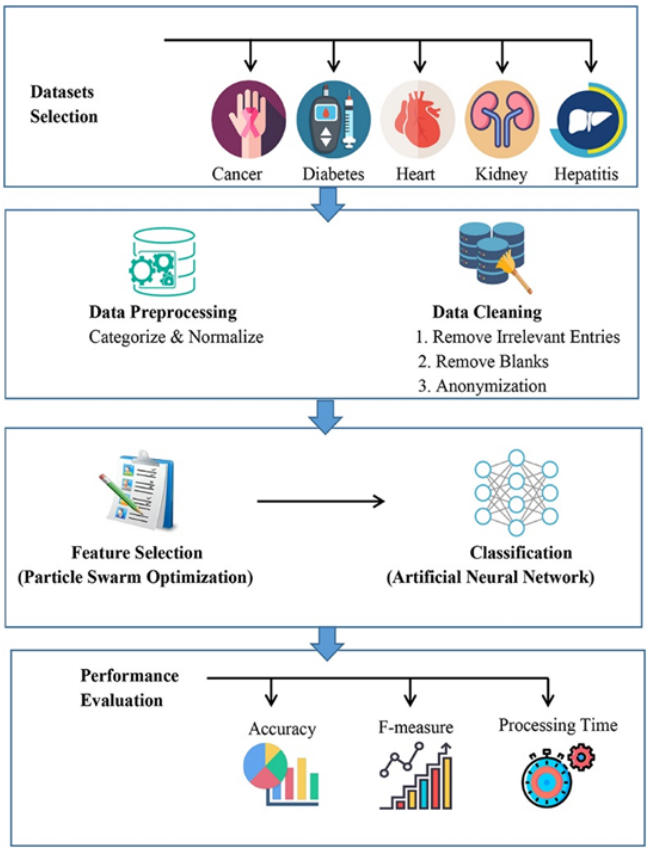


Fig. 2. Flow of the Work

This chart displays the data selection, data pre-processing, feature selection, and evaluation of performance metrics.

A. Decision Tree Model

The construction of a decision tree involves two categories: information gain and the Gini index. Decision trees, as a supervised learning technique, use these measures to select the ideal root node, which serves as the starting point for data classification. Information gain measures the reduction in uncertainty (entropy) produced by splitting data based on a certain feature and prioritizing the split with the greatest reduction. In contrast, the Gini index calculates the level of impurity following a split, favoring the one with the lowest impurity. By iteratively applying these metrics, decision trees divide the data into various areas depending on feature thresholds, eventually assigning labels to each.

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (1)$$

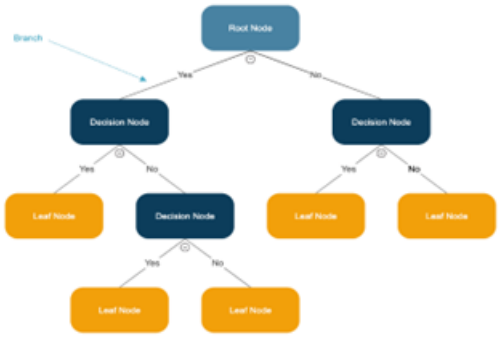


Fig. 3. Decision Tree Model

The data is represented hierarchically in a decision tree, where the class label is represented by the root node and the remaining child nodes are the subnodes.

B. Naive Bayes Model

This paragraph delves into the fundamentals of Naive Bayes, a common supervised learning model for categorization. Its strength derives in its probabilistic methodology, which assumes feature independence to efficiently handle high-dimensional data. The approach uses individual feature probabilities to compute the chance that an instance belongs to each class. While the independence assumption is a simplification, Naive Bayes has shown unexpected performance in text and sentiment analysis applications. Notably, its simplicity translates to accurate classifications without demanding considerable processing resources. In essence, this algorithm provides a practical and probabilistic answer to categorization difficulties.

$$\frac{p(B|A)P(A)}{p(B)} = p(A|B) \quad (2)$$

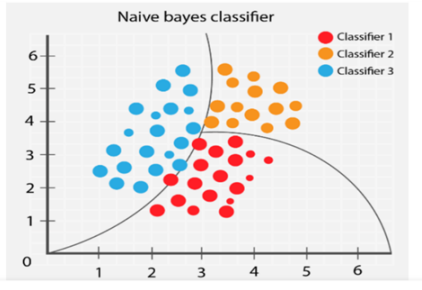


Fig. 4. Naive Bayes Model

This classifier classifies the class based on probabilities. It depends on conditional probabilities.

C. Support Vector Machine (SVM)

Support Vector Machines (SVMs) are robust supervised learning algorithms that excel at classification and regression. SVMs are known for their classification efficacy and ability to navigate high-dimensional regions while resisting overfitting. Their core idea is to find the ideal hyperplane, which is a decision boundary that maximizes the margin between classes. This margin directly transfers to classification confidence, which may result in greater generalization on previously encountered data. Furthermore, SVMs use kernel functions to address nonlinear interactions between features, making them ideal for complicated datasets. While training an SVM can be computationally intensive, the resultant model frequently achieves excellent accuracy and efficient prediction. In essence, SVMs are an effective tool for dealing with difficult classification problems because they prioritize obvious differentiation between classes.

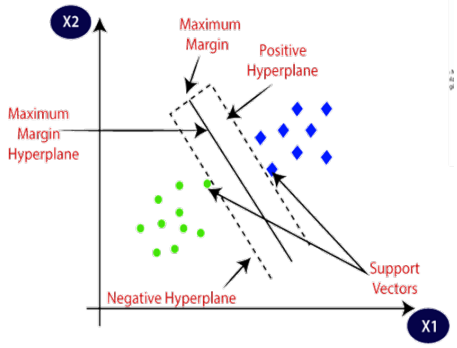


Fig. 5. Support Vector Machine Model

SVM draws a hyperplane this plane separates the data into classes such that if we add a new point it categories the data point to which class it belongs to

D. Multi-Layer Perceptron Model

Multilayer Perceptrons (MLPs) excel at performing sophisticated classification and regression problems. These neural networks feature interconnected nodes that create hidden, input, and output layers. MLPs may learn sophisticated nonlinear relationships thanks to the hidden layers, which dynamically modify neuron weights and biases during training using backpropagation. This versatility makes them useful for a wide range of applications, including time series prediction, speech recognition, and image categorization. While their ability to handle non-linearity is evident, MLPs are prone to overfitting, demanding precise hyperparameter adjustment to attain peak performance. However, their ability to capture complicated non-linear patterns makes them ideal for difficult tasks in a variety of disciplines, including image identification.

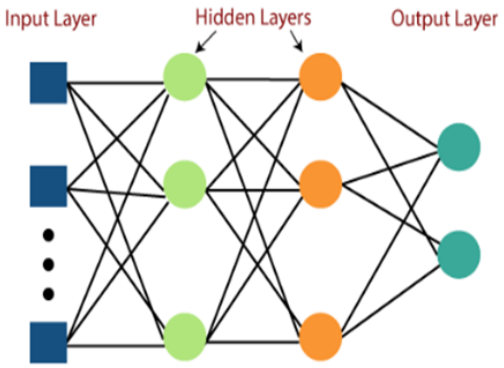


Fig. 6. Multi-Layer Perceptron Model

MLP consists of fully connected neurons with a nonlinear kind of activation function, organized in at least three layers.

E. K-Nearest Neighbors (KNN)

The K-nearest neighbors (KNN) algorithm, or KNN/k-NN, is an effective non-parametric supervised learning classifier. Based on the concept of proximity, KNN properly categorizes and forecasts individual data points by analyzing their nearest neighbors. This strategy is based on the notion that related data points have similar labels or values. During training, KNN uses the complete dataset as a reference, resulting in impressive efficiency across a variety of machine-learning applications [1]. This efficiency, combined with its natural simplicity, has driven widespread usage in a variety of applications. In image recognition, KNN excels in identifying patterns by analyzing pixel similarities between adjacent data points. Similarly, in recommendation systems, KNN is a key component of collaborative filtering algorithms, allowing for personalized recommendations based on user preferences.

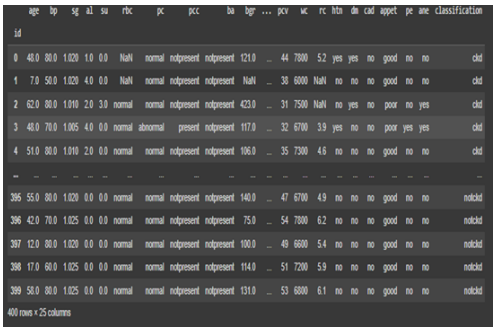
Fig. 7. K-Nearest Neighbor Model

KNN relies on the concept of proximity to make informed classification decisions for individual data points.

IV. Results and Analysis

To conclude the results, we have used several statistical methods such as building confusion matrices for each each data classification model. Confusion matrix is a performance measure tool that measures the accuracy of each data classification model [1]. The goal of this study is to determine if a patient has chronic kidney disease or not, with two possible values: ckd or not-ckd. To begin, I pre-processed the dataset by converting text to numerical numbers and making any necessary changes. I then conducted an exploratory analysis of the dataset before splitting it into training and testing sets. With this preparation complete, I used a variety of tools to analyze the dataset. While the classification's initial results may be unsatisfactory, additional refining and analysis will be performed.

A. Dataset before normalization



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	35	7300	4.6	no	no	no	good	no	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	140.0	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	75.0	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	100.0	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	114.0	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	131.0	...	53	6800	6.1	no	no	no	good	no	no	notckd

400 rows × 25 columns

Fig. 8. Original Dataset

The Chronic Kidney Disease(CKD) Dataset contains categorical as well as numerical data so normalization should be done to increase the performance as well as to bring all the data values to the common scale.

B. Dataset after normalization

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane
0	40	3	3	1	0	1	1	0	0	48	...	90	32	72	34	1	4	1	0	0	0
1	5	0	3	4	0	1	1	0	0	48	...	49	26	56	34	0	3	1	0	0	0
2	54	3	1	2	3	1	1	0	0	140	...	32	19	70	34	0	4	1	1	0	1
3	40	2	0	4	0	1	0	1	0	44	...	48	20	62	19	1	3	1	1	1	1
4	43	3	1	2	0	1	1	0	0	33	...	52	23	68	27	0	3	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	47	3	3	0	0	1	1	0	0	64	...	93	35	62	30	0	3	1	0	0	0
396	34	2	4	0	0	1	1	0	0	3	...	101	42	72	44	0	3	1	0	0	0
397	8	3	3	0	0	1	1	0	0	27	...	94	37	61	36	0	3	1	0	0	0
398	11	1	4	0	0	1	1	0	0	41	...	78	39	67	41	0	3	1	0	0	0
399	50	3	4	0	0	1	1	0	0	57	...	94	41	63	43	0	3	1	0	0	0

400 rows × 24 columns

Fig. 9. Dataset after normalization

The Dataset has been normalized to a common scale the categorical values are also converted into numerical and all the values have ranged between 0 and 1.

C. No. of training and testing samples

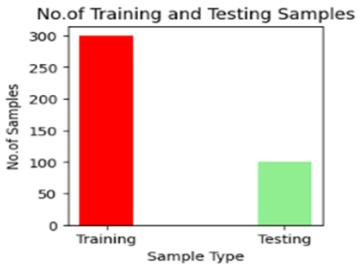


Fig. 10. Visualization of Dataset

The dataset has Training and Testing sets which are 300 and 100 respectively.

D. No. of CKD and NCKD

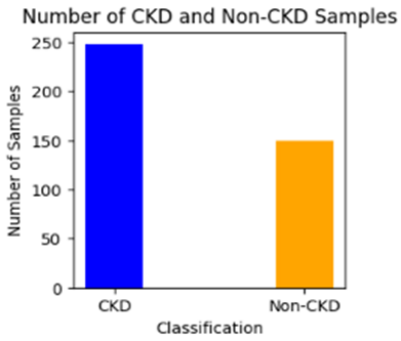


Fig. 11. Visualization of Class Label Samples

It have two classes: CKD and Non-CKD, which are 250 and 150 respectively.

E. Decision Tree

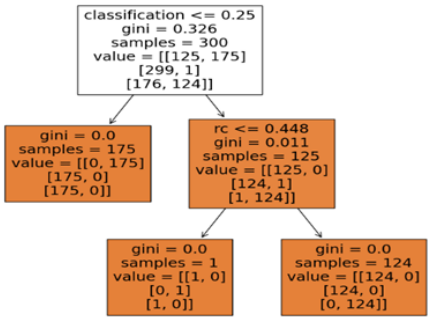


Fig. 12. Decision Tree for Dataset

Constructed a Decision Tree for classifying the dataset for better understanding. The Tree contains One Root node and Four Child nodes.

F. Performance

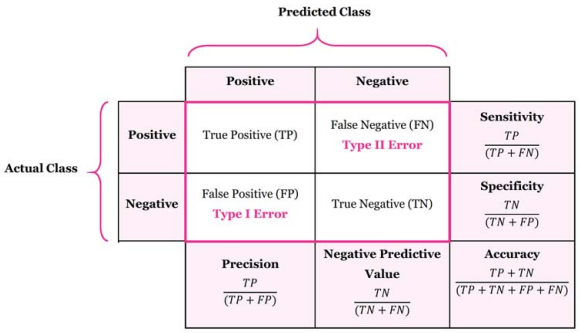


Fig. 13. Performance Metrics Calculation

This image shows the calculation of Performance Metrics which are Accuracy, Precision, Recall, and F1-score.

Model	Accuracy	Precision	Recall	F1Score
Decision tree	100.00	100.00	100.00	100.00
Naïve Bayes	100.00	100.00	100.00	100.00
K-Nearest Neighbors	100.00	100.00	100.00	100.00
SVC	100.00	100.00	100.00	100.00
Back-Propagation	100.00	100.00	100.00	100.00

TABLE I
TABLE-1: PERFORMANCE OF DIFFERENT MODELS

The Table consists of Performance metrics: Accuracy, Precision, Recall and F1Score for Models.

G. Confusion Matrices

We used confusion matrices to compare the performance of several classification models in detecting disease. When we examined the matrices for Decision Trees, Naive Bayes Classifiers, K-nearest neighbors, Support Vector Machines, and Backpropagation, we discovered 74 True Positives. The number of False Positives (incorrectly categorizing healthy persons) and False Negatives (missing instances) differed amongst models, reflecting differences in data handling difficulties. This study provides useful information about individual model strengths and weaknesses, setting the path for further review and optimization.

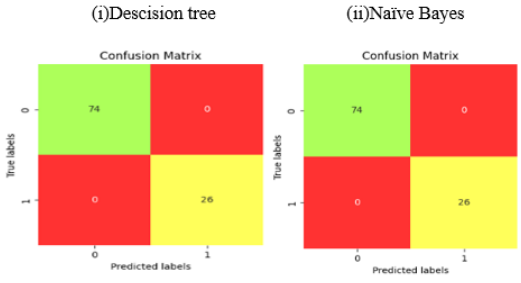


Fig. 14. Confusion Matrices for DT and NB

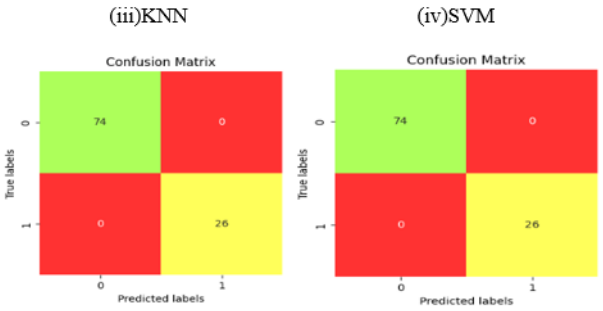


Fig. 15. Confusion Matrices for KNN and SVM

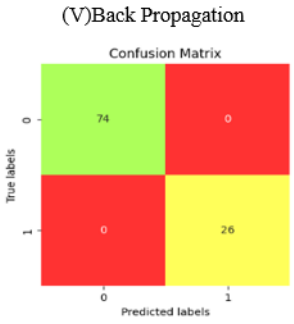


Fig. 16. Confusion Matrix for Back-Propagation

It's critical to understand the stated accuracy ratings as accurate evaluations of the model's performance in the given context, as they reflect the models' performance on the datasets employed. Because I normalized the data values from a greater scale to an even scale, bringing all of the values from several classifications under one domain, I was able to obtain accurate performance from this dataset. This is one of the primary reasons for accurate performance, and I have utilized the optimal split to ensure accuracy in the resultant output. I, therefore, conclude that although I obtained the ideal analysis from this dataset, the accuracy will be altered if I fail to conduct the optimal split and normalize the values.

H. Performace Metrics

It's important to remember that in real-world machine learning scenarios, this is rarely possible and can even indicate potential concerns like as overfitting. The accuracy scores provided here correctly represent the models' performance on the specific datasets and data preparation methodologies utilized. By providing these realistic outcomes, we obtain important insights into the models' strengths and limits in the current environment. This insight enables us to make informed judgments about future optimization and new applications based on the individual use case.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	1.0	1.0	1.0	1.0
1	Gaussian Naive Bayes	1.0	1.0	1.0	1.0
2	K Nearest Neighbors	1.0	1.0	1.0	1.0
3	Support Vector Classifier	1.0	1.0	1.0	1.0
4	Backpropagation	1.0	1.0	1.0	1.0

Fig. 17. Performance Metrics for Original Dataset

This image shows the Performance Metrics: Accuracy, Precision, Recall, and F1Score for each data classification model.

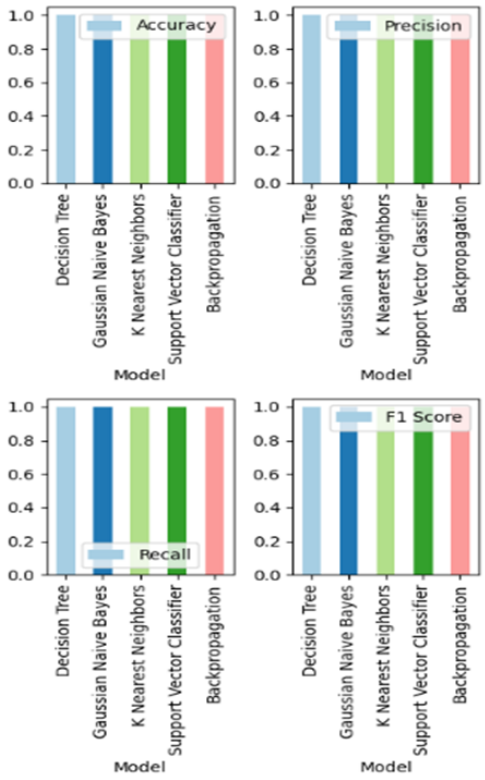


Fig. 18. Visualization of Performance Metrics

This image contains the Graph of each Performance Metric for each classification model.

V. Conclusion

This study used several classification techniques to show how effective our model is. The findings were outstanding, with the Decision Tree, SVM, KNN, Backpropagation, and Naive Bayes algorithms all obtaining perfect accuracy. It's important to note that not all models reach 100% accuracy. To achieve accurate comparisons, the dataset was pre-processed with preset functions and normalized to a 0–1 scale. This includes translating categorical variables to numerical values and applying the necessary normalizations. Furthermore, the dataset was split into train and test sets using the best-split technique, which may have contributed to the excellent risk prediction results.

References

[1] Marwa Almasoud and Tomas E Ward. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, 10(8), 2019.

[2] Suliman A Alsuhibany, Sayed Abdel-Khalek, Ali Algarni, Aisha Fayomi, Deepak Gupta, Vinay Kumar, Romany F Mansour, et al. Ensemble of deep learning based clinical decision support system for chronic kidney disease diagnosis in medical internet of things environment. *Computational Intelligence and Neuroscience*, 2021, 2021.

[3] Raghavendra Babu T M Amogh Babu K A, Priyanka K. Chronic kidney disease prediction based on naive bayes technique. *B.E. in Computer Science Engineering, Mandya, Karnataka, India; Asst. Professor, Dept. of CSE, Nagarjuna College of Engineering Technology, Bangalore, Karnataka, India; Asst.Professor, Dept. of CSE, P.E.S. College of Engineering, Mandya, Karnataka, India*, 6(9):1653–1659, 2019.

[4] Veenita Kunwar, Khushboo Chandel, A. Sai Sabitha, and Abhay Bansal. Chronic kidney disease analysis using data mining classification techniques. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pages 300–305, 2016.

[5] Jai Radhakrishnan and Sumit Mohan. Ki reports and world kidney day. *Kidney international reports*, 2(2):125–126, 2017.

[6] El-Houssainy A Rady and Ayman S Anwar. Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, 15:100178, 2019.

[7] Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, and Ankit Chatorikar. Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(10):92–96, 2018.