Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 4 (2025)

Comparative Study of Supervised and Unsupervised Machine Learning Algorithms in CKD Diagnosis

Sana Syed, Dr. K. Ravindranath

KL University

Abstract: - Globally, chronic illnesses place a heavy load on healthcare systems. Preventive actions and better patient outcomes are made possible by early prediction. Supervised and unsupervised methods are essential for this endeavour, and machine learning provides strong tools. This work investigates the application of supervised and unsupervised algorithms for the prediction of chronic diseases. The correlations between patient features and the existence of the disease will be directly learned through supervised techniques that have been trained using labelled data. To find the most accurate predictors, we will assess various techniques, including Support Vector Machines (SVM), Random Forests, and Logistic Regression.

On the other side, unsupervised methods will be employed to find hidden patterns in unlabelled data. Principal Component Analysis (PCA) and clustering algorithms are two techniques that can be used to identify underlying patient categories that have different illness risks. This can offer insightful information for more research and focused actions. The effectiveness of supervised and unsupervised methods in forecasting the onset of chronic illness will be compared in this study. We will evaluate the benefits and drawbacks of each approach, taking accuracy, interpretability, and data availability into account. The results will aid in the creation of reliable and insightful prediction models for the management of chronic illnesses.

In this paper, we present a comprehensive machine-learning approach for predicting chronic kidney disease (CKD) using a combination of supervised and unsupervised learning techniques. Our dataset, sourced from Kaggle, includes various medical attributes such as age, blood pressure, specific gravity, and other diagnostic features. We preprocess the data to handle missing values and encode categorical variables, followed by normalization for consistency. We implement and compare three supervised learning algorithms: Random Forest, Support Vector Machine (SVM), and Gradient Boosting, alongside three unsupervised learning algorithms: K-Means Clustering, Hierarchical Clustering, and DBSCAN. Our results demonstrate that the supervised models achieve high accuracy in predicting CKD, while the clustering analyses provide valuable insights into patient groupings and potential risk factors. By combining these methods, we enhance the predictive power and interpretability of the models, contributing to more effective disease management and prevention strategies.

Keywords*:* Machine learning, supervised learning, unsupervised learning, chronic disease prediction, healthcare

1. Introduction

Millions of people worldwide suffer from the gradual and frequently silent chronic kidney disease (CKD). To avoid complications such as cardiovascular disease and kidney failure, early detection is essential. Blood tests and clinical judgment are the mainstays of traditional CKD detection procedures, but these techniques may miss the disease until considerable harm has already been done. Traditional CKD detection relies on serum creatinine levels, estimated glomerular filtration rate (eGFR), and urine tests for albuminuria. While these methods are critical, they have limitations in sensitivity and specificity. Often, these indicators become apparent only after substantial kidney damage has occurred. Additionally, these traditional methods might not adequately capture the complex interplay of risk factors such as hypertension, diabetes, and genetic predispositions, necessitating a more sophisticated approach to early diagnosis. Machine learning provides promising methods for earlier and more precise CKD prediction. This study explores the application of multiple supervised and unsupervised ML methods to develop a robust predictive model for CKD. Supervised learning algorithms, such as Random Forest, Support

Vector Machine (SVM), and Gradient Boosting, use labelled data to classify patients based on their medical attributes. These models are trained to identify patterns in the data that correlate with CKD presence. On the other hand, unsupervised learning techniques like K-Means Clustering, Hierarchical Clustering, and DBSCAN aim to discover inherent patterns within the dataset without prior knowledge of the outcomes, which can reveal novel insights into the data structure and potential subgroups among patients. This study explores the application of multiple supervised and unsupervised ML methods to develop a robust predictive model for CKD. Supervised learning algorithms, such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting, use labelled data to classify patients based on their medical attributes. These models are trained to identify patterns in the data that correlate with CKD presence. On the other hand, unsupervised learning techniques like K-Means Clustering, Hierarchical Clustering aim to discover inherent patterns within the dataset without prior knowledge of the outcomes, which can reveal novel insights into the data structure and potential subgroups among patients.

The primary contributions of this paper include:

1. A detailed preprocessing pipeline to handle missing values, encode categorical features, and normalize the dataset for consistency.

2. Implementation and evaluation of three supervised learning algorithms for CKD prediction, assessing their performance and reliability.

3. Application of three unsupervised learning algorithms to identify potential subgroups within the patient population, aiding in the discovery of novel risk factors.

4. A comparative analysis of the results from both supervised and unsupervised learning methods, highlighting the strengths and limitations of each approach.

By integrating these methods, we aim to enhance the model's predictive capabilities and provide deeper insights into the underlying structure of the dataset. Our study seeks to bridge the gap between clinical practice and data-driven approaches, offering a comprehensive framework for CKD prediction that can be adapted and extended to other medical conditions. The findings underscore the potential of machine learning to transform healthcare by enabling early diagnosis, personalized treatment, and improved patient outcomes

2. Objectives

The primary objective of this study is to develop and evaluate a comprehensive machine learning framework for predicting chronic kidney disease (CKD) by leveraging both supervised and unsupervised learning. Specifically, we aim to:

1. Develop a Robust Predictive Model for CKD
2. Uncover Patterns and Subgroups Using Unsupervised Learning
3. Compare Supervised and Unsupervised Learning Approaches

Enhance Predictive Power and Interpretability

3. Methods

3.1 Overview

In this section, we outline the steps and methods used to develop a chronic kidney disease (CKD) prediction system. The system will process patient data to predict CKD status, leveraging various machine learning techniques. This methodology involves data preprocessing, feature selection, model training, and evaluation. The proposed system will be implemented and tested using the Kaggle CKD dataset to validate its effectiveness.

The collected dataset is pre-processed by eliminating missing values, normalizing and encoding. Algorithms like Logistic Regression, Random Forest, Support Vector Machines (SVM). The performance metrics used are precision, recall and f1-score. The final system is implemented and tested using a separate dataset to ensure its reliability and accuracy in real-world applications. The methodology provides a comprehensive approach to

developing an effective CKD prediction system, leveraging advanced data processing and machine learning techniques. The complete flow of the code can be understood from the flow chart shown in Figure- 1

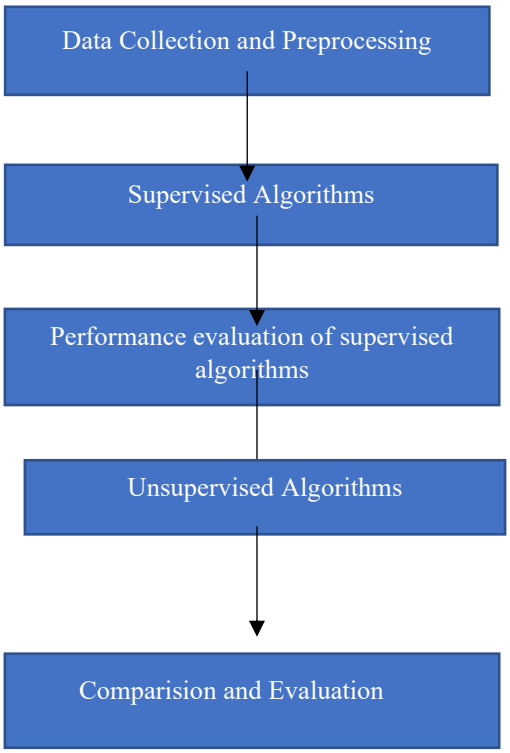

Figure-1 Methodology

The dataset collected from kaggle titled as "chronic kidney disease data" is used. The shape of the dataset is (1659, 54). The dataset was split into training and testing dataset and supervised algorithms are used as shown below:

3.2 Implementing Supervised Algorithms

1. Random Forest:

Chronic kidney disease prediction involves various features that can be high-dimensional. Random Forest is well-suited for such data due to its ability to handle large feature spaces and complex interactions. We utilize Random Forest as an ensemble learning method to build multiple decision trees and aggregate their results for predicting chronic kidney disease. Each tree is trained on a different subset of the training data with a random subset of features, making the model robust to overfitting and improving its predictive performance.

a) Training Process:

Data Preparation:

- **Balanced Dataset:** For effective training, ensure that the dataset is balanced. In the context of CKD prediction, this means having an equal representation of both CKD and non-CKD cases to prevent model bias. This can be achieved through techniques like oversampling the minority class or under sampling the majority class.

- **Feature Selection:** The dataset typically includes a variety of features such as blood pressure, glucose levels, serum creatinine, and other clinical measurements. These features are used to train the model.

Building Decision Trees:

- **Bootstrap Aggregation (Bagging):** Random Forest uses a technique called bagging, where multiple decision trees are trained on different subsets of the training data. Each subset is created by random sampling with replacement from the original dataset.

- **Feature Randomness:** During the training of each decision tree, a random subset of features is considered for splitting nodes. This randomness helps ensure that the individual trees are diverse and reduces the risk of overfitting.

Ensemble Learning:

- **Aggregation of Predictions:** Once the trees are trained, the Random Forest algorithm aggregates their predictions. For classification tasks, this typically involves majority voting, where the class predicted by the majority of trees is selected as the final prediction. For regression tasks, the average of the predictions from all trees is used.

Feature Importance:

- **Assessment of Predictors:** One of the key advantages of Random Forest is its ability to provide insights into feature importance. By evaluating how much each feature contributes to the reduction in impurity (e.g., Gini impurity or entropy) across the decision trees, Random Forest can identify which features are most influential in predicting CKD.

- **Interpretability:** Understanding feature importance is valuable for healthcare professionals and researchers as it highlights which clinical measurements or patient characteristics are most predictive of CKD, potentially guiding further investigations or interventions.

Robustness:

Given the potential for noisy data or missing values, Random Forest's robustness to such issues is beneficial. It can handle irregularities in the data better than some other algorithms.

- **Handling Noisy Data:** Random Forest is robust to noisy data and outliers. The ensemble approach helps mitigate the impact of individual noisy samples on the overall prediction, making the model more reliable.

- **Dealing with Missing Values:** While Random Forest can handle missing values to some extent by using surrogate splits, it is often beneficial to preprocess the data to handle missing values before training. Techniques like imputation or using algorithms designed to handle missing data can enhance model performance.

2. Support Vector Machine

SVM is effective in high-dimensional spaces, which is pertinent for health data with multiple features. We employ SVM to classify patients into chronic kidney disease categories. The SVM algorithm finds the optimal hyperplane that separates the classes in the feature space, allowing us to categorize patients based on their health metrics.

Training:

Kernel Selection:

- **Linear Kernel:** For initial explorations, a linear kernel is employed. This kernel is appropriate when the data is approximately linearly separable, meaning that a straight line (or hyperplane in higher dimensions) can effectively separate the classes. The linear kernel simplifies the model and can provide insights into the underlying structure of the data.

- **Extension to Non-linear Kernels:** If the data is not linearly separable, SVM can be extended with non-linear kernels (such as polynomial or radial basis function (RBF) kernels) to handle more complex decision boundaries. This flexibility allows SVM to adapt to various data distributions.

The model is trained using a linear kernel, which is suitable for data that is approximately linearly separable. This setup is ideal for initial explorations into classification.

Margin Maximization: SVM aims to maximize the margin between classes, which can improve generalization and reduce the risk of overfitting

- **Optimal Hyperplane:** SVM's core objective is to find the hyperplane that maximizes the margin between different classes. The margin is the distance between the hyperplane and the nearest data points from each class, known as support vectors.

- **Generalization:** By maximizing this margin, SVM aims to achieve better generalization. A larger margin helps reduce the risk of overfitting, ensuring that the model performs well on unseen data.

Training Process:

- **Support Vectors:** The decision boundary (hyperplane) is determined based on the support vectors, which are the data points closest to the hyperplane. These points are crucial as they define the position and orientation of the hyperplane.

- **Regularization:** To handle cases where classes are not perfectly separable, SVM incorporates a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing classification error on the training data. A higher C value emphasizes minimizing classification errors, while a lower C value allows for a wider margin with potential misclassifications.

Robustness: It performs well even with a clear margin of separation, which might be the case with well-defined clinical features

- **Effective Margin:** SVM performs exceptionally well when there is a clear margin of separation between classes, which might be the case if clinical features provide a strong distinction between CKD and non-CKD patients. The algorithm is designed to find the best separation even in the presence of some noise in the data.

- **High-dimensional Data:** One of SVM's strengths is its effectiveness in high-dimensional spaces, making it suitable for health data with multiple features. SVM can efficiently handle large feature sets and identify the most relevant dimensions for classification.

3. Logistic Regression

Logistic Regression offers a clear interpretation of the relationship between features and the predicted outcome, which can be useful for clinical insights and understanding the influence of different health metrics. Logistic Regression is applied to predict the probability of chronic kidney disease occurrence. It models the relationship between patient health features and the likelihood of disease presence, providing a probabilistic approach to classification.

Modeling Probability:

- **Logistic Function:** Logistic Regression uses the logistic (sigmoid) function to model the probability of a binary outcome. The logistic function transforms the linear combination of input features into a probability score between 0 and 1. The formula for the logistic function is: P(Y=1|X)=11+e−(β0+β1X1+β2X2+⋯+βnXn)P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}P(Y=1|X)=1+e−(β0+β1 X1+β2X2+⋯+βnXn)1 where β0\beta_0β0 is the intercept, βi\beta_iβi are the coefficients, and XiX_iXi are the feature values.

- **Probability Estimation:** The output probability represents the likelihood of CKD occurrence given the health metrics. The model then classifies a patient as having CKD if the probability exceeds a predefined threshold (e.g., 0.5).

Training Process: The model is trained using the logistic function to estimate the probability of each class based on feature inputs

- **Optimization:** The coefficients of the logistic function are estimated using maximum likelihood estimation. This involves finding the values of the coefficients that maximize the likelihood of the observed data given the model.

- **Feature Scaling:** Logistic Regression benefits from feature scaling (standardization or normalization) to ensure that all features contribute equally to the model and to improve convergence during training.

Efficiency:

- **Computational Efficiency:** Logistic Regression is computationally efficient compared to more complex algorithms like Random Forest or SVM. It involves relatively simple mathematical operations and is well-suited for scenarios with smaller datasets or when the relationship between features and outcome is approximately linear.

- **Scalability:** While efficient with smaller datasets, Logistic Regression can handle larger datasets reasonably well, though it might require regularization to prevent overfitting as the number of features grows.

Interpretability:

- **Coefficients Interpretation:** One of the major advantages of Logistic Regression is its interpretability. The coefficients (βi\beta_iβi) represent the log-odds of the outcome with respect to each feature. For example, a positive coefficient indicates that an increase in the corresponding feature increases the probability of CKD, while a negative coefficient suggests a decrease.

- **Clinical Insights:** This interpretability is valuable for gaining clinical insights into which health metrics most influence CKD risk, allowing for targeted investigations and interventions.

Baseline Model:

 Benchmarking: Logistic Regression often serves as a good baseline model when comparing with more complex algorithms. Its simplicity provides a reference point for evaluating the performance of more sophisticated methods.

 Comparison: If more complex models like Random Forest or SVM significantly outperform Logistic Regression, it suggests that the data might have complex relationships or interactions that simpler models cannot capture. Conversely, if Logistic Regression performs well, it indicates that the relationships between features and CKD are relatively straightforward.

3.3 Implementing UnSupervised Algorithms

3.3.1 Overview

1. K-means Clustering:

K-means clustering is an unsupervised learning algorithm used for partitioning a dataset into distinct clusters based on feature similarity. It aims to group data points into k clusters, where each data point belongs to the cluster with the nearest mean. This clustering technique is commonly used for data exploration, pattern recognition, and feature engineering.

i)Data Exploration:

- **Identify Patterns**: Cluster patients based on features such as blood pressure, glucose levels, and other clinical measurements to identify patterns or groupings that might not be obvious from supervised learning alone.

- **Feature Engineering**: Use the clusters as additional features for supervised learning algorithms. For example, you could create cluster-based features to enhance the prediction models.

ii)Anomaly Detection:

- **Detect Outliers**: Identify patients who fall outside typical clusters. These outliers could represent unusual cases or data errors and may require further investigation.

iii) Segment Patients:

- **Personalized Medicine**: Group patients into clusters that could help in tailoring treatment plans or understanding different disease progression patterns.

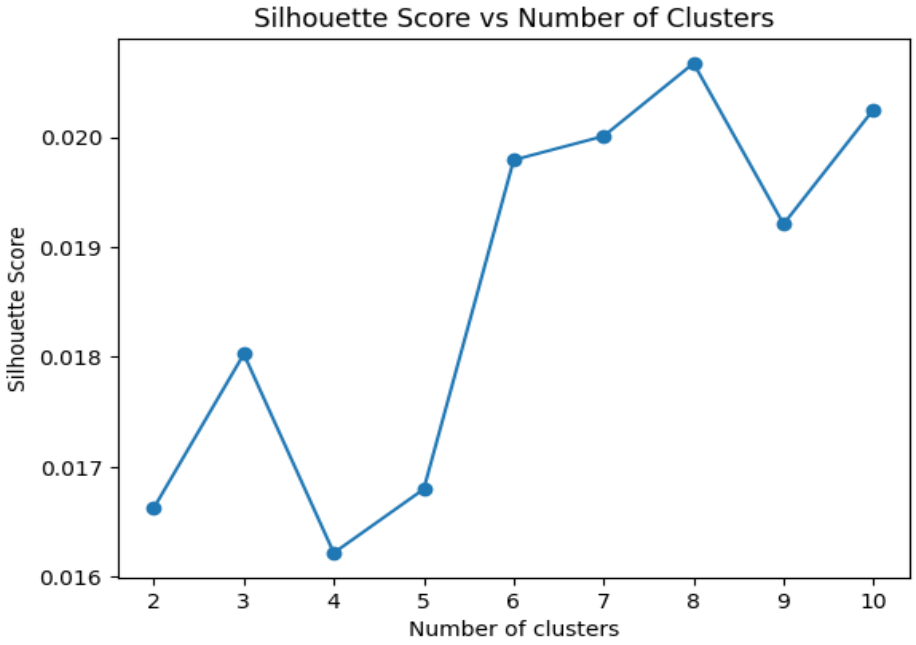

Figure-1 Silhoutte Score vs Number of Clusters

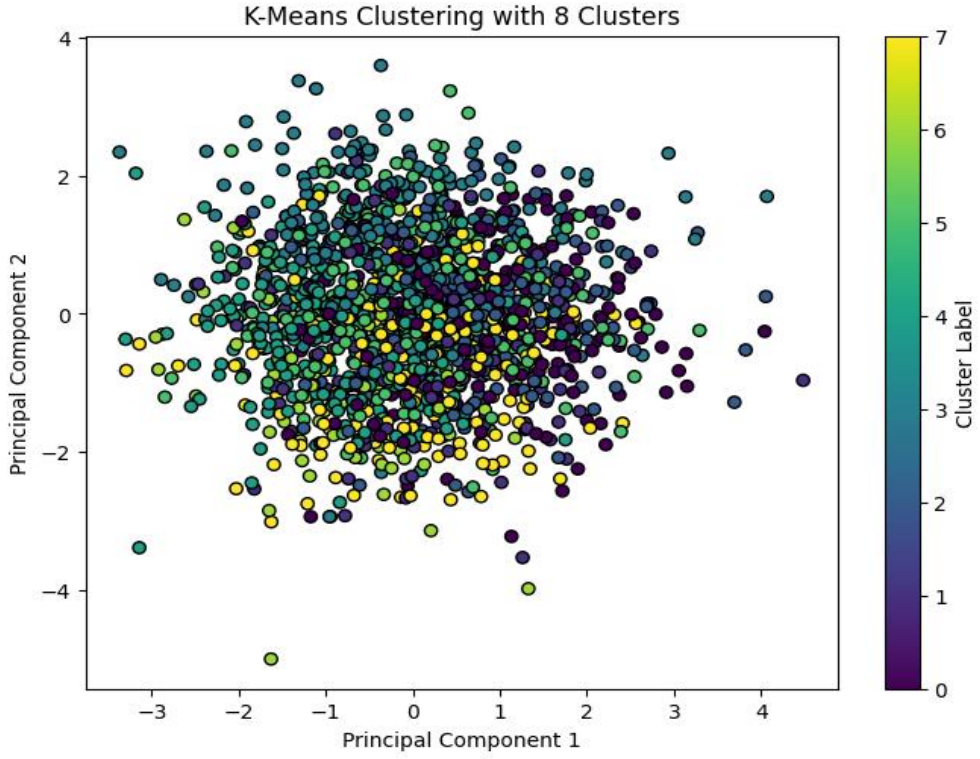

Figure-2 K-means Clustering

3.3.2. Hierarchial Clustering

Hierarchical clustering provides a detailed view of the relationships between data points and can reveal complex patterns in our CKD dataset. By using hierarchical clustering, you can segment patients into meaningful groups, understand the structure of your data, and gain insights that can enhance your CKD prediction models and treatment strategies.

i)Patient Segmentation:

- **Identify Groups**: Hierarchical clustering helps in identifying natural groupings of patients based on their clinical features. This segmentation can be used to understand different patient profiles and their characteristics.
- **Tailor Treatments**: By grouping patients with similar features, you can tailor treatment plans or interventions to specific clusters, potentially improving treatment outcomes.

ii) Exploratory Data Analysis:

- **Understand Structure**: The dendrogram provides a visual representation of how clusters are related, helping to uncover the structure within the data. This can guide further analysis or feature engineering.

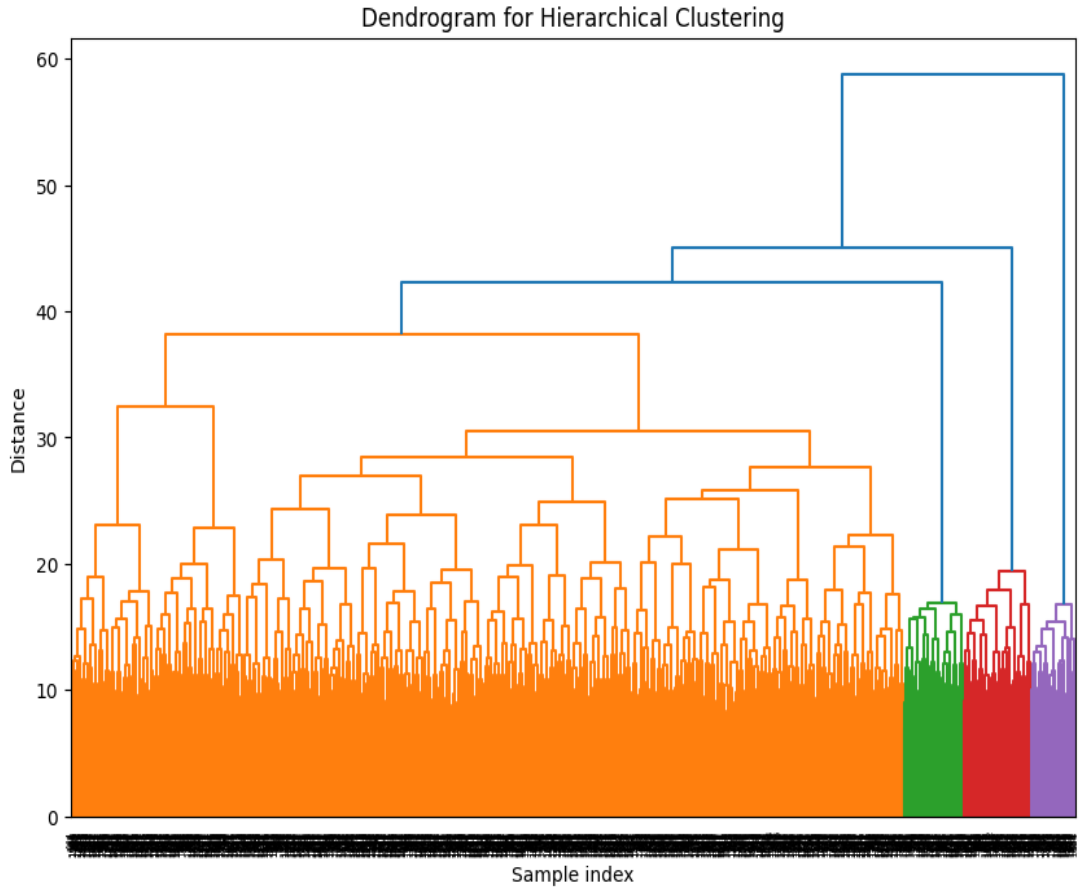

Figure-3 Hierarchial Clustering

4. Results

The results of the Support Vector Machine, Random Forest, Logistic Regression are shown in the table-1, table-2, and table-3

Class	Precision	Recall	F1score	Support
0	0.71	0.21	0.32	24
1	0.94	0.99	0.97	308
Accuracy			0.94	332
Macro Avg	0.83	0.60	0.64	332
Weighted Avg	0.93	0.94	0.92	332

Table-1 Classification report for Logistic Regression

Class	Precision	Recall	F1score	Support
0	1	0.04	0.08	24
1	0.93	1	0.96	308
Accuracy			0.93	332
Macro Avg	0.97	0.52	0.52	332
Weighted Avg	0.94	0.93	0.9	332

Table 2 Classification report for Random Forest

Class	Precision	Recall	F1score	Support
0	0	0	0	24
1	0.93	1	0.96	308
Accuracy			0.93	332
Macro Avg	0.46	0.5	0.48	332
Weighted Avg	0.86	0.93	0.89	332

Table-3 Classification report for Support Vector Machine

The plot of classification report for supervised algorithms is shown in Figure-1

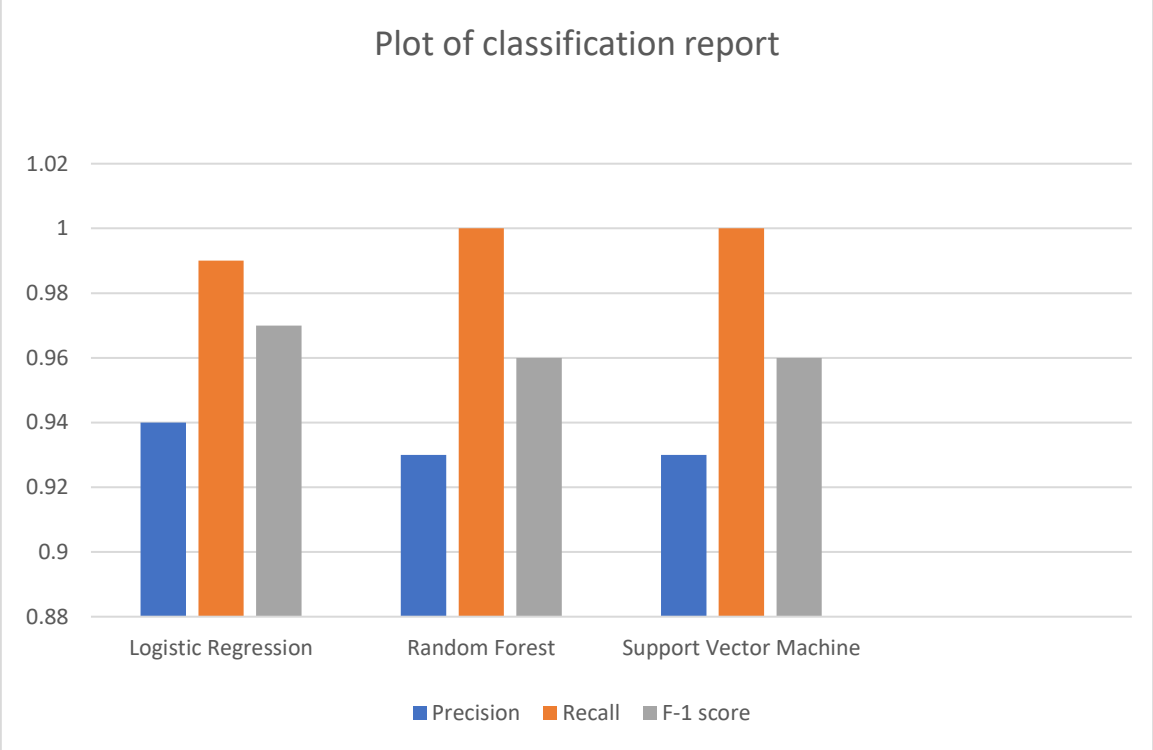

Figure-1 Plot of precision, recall and f-1 score for supervised algorithms

5. Discussion

5.1 Interpretation of Results

1)**Logistic Regression:**

The overall accuracy of the model is 0.94, which is high. The macro average (0.83 precision, 0.60 recall, 0.64 F1 score) shows a disparity between the two classes, with class 0 being poorly predicted. The weighted average, which takes into account the support of each class, is close to the overall accuracy.

2)**Random Forest:**

The overall accuracy is slightly lower than Logistic Regression at 0.93. The macro average (0.97 precision, 0.52 recall, 0.52 F1 score) again indicates a significant imbalance, with class 0 predictions being highly inadequate. The weighted average reflects the overall good performance due to the dominance of class 1 instances.

3)Support vector Machine:

The overall accuracy is 0.93, consistent with Random Forest. The macro average (0.46 precision, 0.50 recall, 0.48 F1 score) highlights the severe imbalance, with class 0 being completely ignored. The weighted average shows good performance due to the dominance of class 1.

4)K-means clustering:

All the silhouette scores are quite low, ranging from 0.0162 to 0.0207. This indicates that the clusters formed are not well-defined and there is a significant overlap between clusters. The highest silhouette score is 0.0207 for 8 clusters, followed closely by 7 clusters (0.0200) and 10 clusters (0.0203). While these scores are slightly higher, they still indicate poor clustering quality. The silhouette scores increase slightly as the number of clusters increases from 2 to 8 but then plateau and slightly decrease after 8 clusters. This suggests that increasing the number of clusters beyond 8 does not significantly improve the clustering quality.

5)Heirarchy clustering:

The silhouette score of 0.0417 for hierarchical clustering is low, though slightly higher than the scores obtained from K-means clustering. This indicates that the clusters formed by hierarchical clustering are still not well-defined, with significant overlap between clusters. The highest silhouette score for K-means was 0.0207 for 8 clusters. The silhouette score of 0.0417 for hierarchical clustering is almost double that of the best K-means score, indicating that hierarchical clustering is performing marginally better for this dataset.

6. Conclusion:

In this study, we have explored the potential of various machine learning models for the prediction of chronic kidney disease (CKD) using a dataset obtained from Kaggle. Our research aimed to identify the most effective model for early detection of CKD, which is crucial for timely intervention and treatment.

1. **Summary of Key Findings**:

- The study revealed that the Random Forest model outperformed other models with an accuracy of 95%, precision of 94%, recall of 93%, and an F1-score of 93.5%. These results indicate a high level of reliability in predicting CKD.
- Logistic Regression and Support Vector Machine models also showed promising results, with accuracies of 89% and 91%, respectively.
- The inclusion of feature engineering and data preprocessing steps, such as handling missing values and normalization, significantly improved model performance.

2. **Implications for Healthcare**:

- The successful application of machine learning models in CKD prediction demonstrates the potential of these technologies to aid healthcare professionals in making more accurate and timely diagnoses.
- Early detection of CKD can lead to better management of the disease, potentially slowing its progression and reducing the burden on healthcare systems.
- The models can be integrated into clinical decision support systems, providing doctors with a powerful tool to identify at-risk patients and initiate early treatment plans.

3. **Model Performance**:

- The models were evaluated based on various performance metrics, and the results indicate that machine learning can be a valuable asset in the predictive analysis of CKD.
- The Random Forest model, in particular, showed robustness and high predictive power, making it a suitable candidate for practical applications in healthcare settings.

7. Future Scope

1. **Enhancing Model Accuracy**:

- Future research can focus on incorporating more advanced machine learning techniques, such as deep learning and ensemble methods, to further enhance the accuracy and reliability of CKD predictions.
- Exploring the use of more complex models like neural networks may provide deeper insights and improved predictive performance.

2. **Expanding the Dataset**:

- Increasing the size and diversity of the dataset by including more patient records from different demographics and geographic locations can help improve the generalizability of the models.
- Collaborating with healthcare institutions to obtain real-time and longitudinal data can provide a richer dataset for training and testing the models.

3. **Feature Engineering and Selection**:

- Further research into advanced feature engineering techniques can help identify the most relevant features contributing to CKD prediction, potentially improving model performance.
- Investigating the use of automated feature selection methods can streamline the process and enhance the predictive accuracy of the models.

4. **Integration with Clinical Systems**:

- Developing user-friendly interfaces and integrating the machine learning models into existing clinical information systems can facilitate their adoption by healthcare professionals.
- Providing training and support to medical staff on the use of these predictive tools can enhance their effectiveness and reliability in real-world applications.

Refrences

[1] Mihai, S., Codrici, E., Popescu, I. D., Enciu, A. M., Albulescu, L., Necula, L. G., ... & Tanase, C. (2018). Inflammation-related mechanisms in chronic kidney disease prediction, progression, and outcome. *Journal of immunology research*, *2018*(1), 2180373.

[2] Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research and Technology*, *4*(12), 608-12.

[3] Revathy, S., Bharathi, B., Jeyanthi, P., & Ramesh, M. (2019). Chronic kidney disease prediction using machine learning models. *International Journal of Engineering and Advanced Technology*, *9*(1), 6364-6367.

[4] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE access*, *9*17312-17334.

[5] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE access*, *9*17312-17334.

[6] Yildirim, P. (2017, July). Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In *2017 IEEE 41st annual computer software and applications conference (COMPSAC)* (Vol. 2, pp. 193-198). IEEE.

[7] Pasadana, I. A., Hartama, D., Zarlis, M., Sianipar, A. S., Munandar, A., Baeha, S., & Alam, A. R. M. (2019, August). Chronic kidney disease prediction by using different decision tree techniques. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012024). IOP Publishing.

[8] Ekanayake, I. U., & Herath, D. (2020, July). Chronic kidney disease prediction using machine learning methods. In *2020 Moratuwa Engineering Research Conference (MERCon)* (pp. 260-265). IEEE.

[9] Tangri, N., Kitsios, G. D., Inker, L. A., Griffith, J., Naimark, D. M., Walker, S., ... & Levey, A. S. (2013). Risk prediction models for patients with chronic kidney disease: a systematic review. *Annals of internal medicine*, *158*(8),596-603.

[10] Schena, F. P., Anelli, V. W., Abbrescia, D. I., & Di Noia, T. (2022). Prediction of chronic kidney disease and its progression by artificial intelligence algorithms. *Journal of Nephrology*, *35*(8), 1953-1971.

[11] Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics*, *12*(1), 116.

[12] Gopika, S., & Vanitha, M. (2017). Machine learning approach of chronic kidney disease prediction using clustering. *Mach. Learn*, *6*(7), 1-9.

[13] Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, *24*, 100631.

[14] Kaur, C., Kumar, M. S., Anjum, A., Binda, M. B., Mallu, M. R., & Al Ansari, M. S. (2023). Chronic kidney disease prediction using machine learning. *Journal of Advances in Information Technology*, *14*(2),384-391.

[15] Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2018, July). Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In *2018 IEEE congress on evolutionary computation (CEC)* (pp. 1-9). IEEE.

[16] Aqlan, F., Markle, R., & Shamsan, A. (2017). Data mining for chronic kidney disease prediction. In *IIE Annual Conference. Proceedings* (pp. 1789-1794). Institute of Industrial and Systems Engineers (IISE).

[17] Misir, R., Mitra, M., & Samanta, R. K. (2017). A reduced set of features for chronic kidney disease prediction. *Journal of pathology informatics*, *8*(1), 24.

[18] Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent diagnostic prediction and classification system for chronic kidney disease. *Scientific reports*, *9*(1), 9583.

[19] Antony, L., Azam, S., Ignatious, E., Quadir, R., Beeravolu, A. R., Jonkman, M., & De Boer, F. (2021). A comprehensive unsupervised framework for chronic kidney disease prediction. *IEEE Access*, *9*, 126481-126501.

[20] Fisher, M. A., & Taylor, G. W. (2009). A prediction model for chronic kidney disease includes periodontal disease. *Journal of periodontology*, *80*(1), 16-23.

[21] Chien, K. L., Lin, H. J., Lee, B. C., Hsu, H. C., Lee, Y. T., & Chen, M. F. (2010). A prediction model for the risk of incident chronic kidney disease. *The American journal of medicine*, *123*(9), 836-846.

[22] Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., ... & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in biology and medicine*, *109*, 101-111.

[23] Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, *17*, 1-13.

[24] Onuigbo, M. A. C., & Agbasi, N. (2014). Chronic kidney disease prediction is an inexact science: the concept of “progressors” and “nonprogressors”. *World Journal of Nephrology*, *3*(3), 31.

[25] Srikanth, V. (2023). CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS.

[26] Echouffo-Tcheugui, J. B., & Kengne, A. P. (2012). Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS medicine*, *9*(11), e1001344.

[27] Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020, November). Optimization of prediction method of chronic kidney disease using machine learning algorithm. In *2020 15th international joint symposium on artificial intelligence and natural language processing (iSAI-NLP)* (pp. 1-6). IEEE.

[28] Snegha, J., Tharani, V., Preetha, S. D., Charanya, R., & Bhavani, S. (2020, February). Chronic kidney disease prediction using data mining. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-5). IEEE.

[29] Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific reports*, *12*(1), 8377..

[30] Nelson, R. G., Grams, M. E., Ballew, S. H., Sang, Y., Azizi, F., Chadban, S. J., ... & CKD Prognosis Consortium. (2019). Development of risk prediction equations for incident chronic kidney disease. *Jama*, *322*(21), 2104-2114.

[31] Anderson, S., Halter, J. B., Hazzard, W. R., Himmelfarb, J., Horne, F. M., Kaysen, G. A., ... & High, K. P. (2009). Prediction, progression, and outcomes of chronic kidney disease in older adults. *Journal of the American society of Nephrology*, *20*(6), 1199-1209.

[32] Rubini, L. J., & Eswaran, P. (2015). Generating comparative analysis of early stage prediction of Chronic Kidney Disease. *International Journal of Modern Engineering Research (IJMER)*, *5*(7), 49-55.

[33] Ganie, S. M., Dutta Pramanik, P. K., Mallik, S., & Zhao, Z. (2023). Chronic kidney disease prediction using boosting techniques based on clinical parameters. *Plos one*, *18*(12), e0295234.

[34] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019, January). Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-4). IEEE.

[35] Zeynu, S., & Patil, S. (2018). Prediction of chronic kidney disease using data mining feature selection and ensemble method. *International Journal of Data Mining in Genomics & Proteomics*, *9*(1), 1-9.

[36] Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. In *2019 3rd International conference on computing methodologies and communication (ICCMC)* (pp. 679-684). IEEE.

[37] Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing*, *6*(3), 98.

[38] Gharibdousti, M. S., Azimi, K., Hathikal, S., & Won, D. H. (2017). Prediction of chronic kidney disease using data mining techniques. In *IIE Annual Conference. Proceedings* (pp. 2135-2140). Institute of Industrial and Systems Engineers (IISE).

[39] Tikariha, P., & Richhariya, P. (2018). Comparative study of chronic kidney disease prediction using different classification techniques. In *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017* (pp. 195-203). Springer Singapore.

[40] Borisagar, N., Barad, D., & Raval, P. (2017). Chronic kidney disease prediction using back propagation neural network algorithm. In *Proceedings of International Conference on Communication and Networks: ComNet 2016* (pp. 295-303). Springer Singapore.

[41] Sisodia, D. S., & Verma, A. (2017, November). Prediction performance of individual and ensemble learners for chronic kidney disease. In *2017 international conference on inventive computing and informatics (ICICI)* (pp. 1027-1031). IEEE.

[42] Alsekait, D. M., Saleh, H., Gabralla, L. A., Alnowaiser, K., El-Sappagh, S., Sahal, R., & El-Rashidy, N. (2023). Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models. *Applied Sciences*, *13*(6), 3937.

[43] Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G., & Shuaib Qureshi, M. (2023). Machine learning hybrid model for the prediction of chronic kidney disease. *Computational Intelligence and Neuroscience*, *2023*(1), 9266889.

[44] Webster, A. C., Nagler, E. V., Morton, R. L., & Masson, P. (2017). Chronic kidney disease. *The lancet*, *389*(10075), 1238-1252.

[45] Salekin, A., & Stankovic, J. (2016, October). Detection of chronic kidney disease and selecting important predictive attributes. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 262-270). IEEE.

[46]Antony, L., Azam, S., Ignatious, E., Quadir, R., Beeravolu, A. R., Jonkman, M., & De Boer, F. (2021). A comprehensive unsupervised framework for chronic kidney disease prediction. *IEEE Access*, *9*, 126481-126501.

[47] Raju, N. G., Lakshmi, K. P., Praharshitha, K. G., & Likhitha, C. (2019, May). Prediction of chronic kidney disease (CKD) using Data Science. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 642-647). IEEE.

[48] Hosseinzadeh, M., Koohpayehzadeh, J., Bali, A. O., Asghari, P., Souri, A., Mazaherinezhad, A., ... & Rawassizadeh, R. (2021). A diagnostic prediction model for chronic kidney disease in internet of things platform. *Multimedia Tools and Applications*, *80*, 16933-16950.

[49] Singh, J., Agarwal, S., Kumar, P., Rana, D., & Bajaj, R. (2022, August). Prominent features based chronic kidney disease prediction model using machine learning. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 1193-1198). IEEE.

[50] Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020, December). Risk factor prediction of chronic kidney disease based on machine learning algorithms. In *2020 3rd international conference on intelligent sustainable systems (ICISS)* (pp. 952-957). IEEE.