ENHANCING CHRONIC KIDNEY DISEASE MANAGEMENT THROUGH PREDICTIVE ANALYTICS

A Thesis submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Data Science and Artificial Intelligence

By

Dhruvi Akshay Vekariya

(Student ID: 2590401)

Under the guidance of

Zainb Dawod

To the

UNIVERSITY OF EAST LONDON

9th, September, 2024

Abstract

The proper diagnosis of some human diseases are still complex and which necessitates predictive analysis tools for improving the diagnosis further properly. The increasing technical advancement also makes the process much possible. In this study the Chronic Kidney Disease management Procedure with the help of predictive analysis is performed. The goal of the project was to create tools with the help of different machine learning models to predict the current stage of kidney disease precisely so that the doctors can take proper steps. This dissertation has been established that predictive analytics and machine learning have a critical role in Chronic Kidney Disease care. Analyzing the datasets and models, the major risk determinants and weights for the development of CKD were defined, and an accurate risk evaluation model was developed. Generally, a high level of accuracy of the developed predictive models, mainly the Random Forest Classifier, implies that these modeling strategies could enhance early diagnosis and risk assessment in CKD. In this dissertation three different models were compared and all of them were pretty accurate for the prediction. Due to the least processing time the Random forest model is used for the actual prediction model creation. The outcome of the prediction shows that the model will help disease management effectively and with the help of further data training the models will become suitable for market implementation.

Acknowledgements

I feel I have learnt a lot from writing this thesis searching for the truth of science and life. This is the great treasure I will cherish not only in my future academic career but in my whole life. I would like to take this opportunity to express my immense gratitude to all those persons who have given their invaluable support and assistance.

I would like to acknowledge and give my warmest thanks to my supervisor **(Zainb Dawod)** who made this work possible. Her invaluable guidance, feedback, support and advice carried me through all the stages of writing my dissertation. Her extensive knowledge and experience were instrumental in the completion of this dissertation.

I would also like to give special thanks to my husband **(Akshay Vekariya)** and my family as a whole for their continuous support and understanding when undertaking my research and writing my dissertation. Your prayer for me was what sustained me this far.

Contents

List of Figures

List of Tables

List of Acronyms

CKD	Chronic Kidney Disease
AI	Artificial Intelligence
ML	Machine Learning
RF	Random Forest
DT	Decision Tree
SVM	Support Vector Machine

Chapter 1: Introduction

"Chronic Kidney Disease (CKD)", is presently acknowledged as a considerable global health issue due to the numerous patients suffering from dialysis and the overall costs it has contributed to the elevated healthcare costs in the society. Hence, as the incidence of CKD appears to increase in the global population, more progressive strategies are required to enhance the handling and treatment of this disease. Therefore, leveraging innovation, especially machine learning and predictive analytics as integrated CKD technologies appear to provide proper prospects towards the innovation of CKD diagnosis and risk assessment as well as management (Bakris *et al*., 2020).

This dissertation discusses the possibility of improving the CKD's management through the help of predictive analytics. The use of the machine learning approach and data analysis methods will help in the identification of patient at-risk, the prognosis of disease severity outcomes or likelihood of progression, as well as the choice of treatment. The research incorporates a list of determinant factors of CKD that enables the establishment of practical models to guide the clinicians in the clinical decision making process with the aim of enhancing patient prognosis.

1.1 Background of Study

"Chronic Kidney Disease" is a gradually increasing disorder which affects the kidneys and results in various related health problems and sometimes renal failure. The Prevalence of CKD is high in the global population with around 13.4% worldwide (Sawhney *et al*., 2023). However, CKD relocates more strain on the healthcare systems as well the economy since the long-term management and renal replacement therapies do cost a lot of cash.

In the past, treatment of CKD was mainly the identification of modifiable cardiovascular risk factors, patients' symptom control and reduction of the progression rate. Nevertheless, the approaches are somehow insufficient in terms of the targeted and timely identification and inclusion of at-risk people into appropriate interventions and support (Srikanth, 2023). The development of big data analytics and machine learning should be viewed as the potential to revolutionize the management of CKD using more accurate risk profiling, screening, and specific treatment management.

AI and Data Sciences that have emerged in recent years suggest that improvement in several fields of Health care is possible. In the case of CKD, predictive analytics would integrate data concerning the patient, demographic parameters, laboratory data, and clinical factors to find a relationship and predict future outcome with a high level of precision. Existing studies on the application of machine learning for the improvement of chronic kidney disease management also showed promising results in effective prediction. Different research works have applied ML methods like random forest, support vector machines (SVM) neural network, among others in CKD prediction of progression, risk factors and individualized therapeutic interventions. These models often rely on the use of massive amounts of datasets, demographics, clinical, and laboratory to predict the occurrence as well as the progression of the diseases (Shlipak *et al*., 2021). For example, early CKD diagnosis has been a subject in some of the studies; the use of Machine Learning algorithms in identifying Cardiac patients at high risk before the symptoms of CKD are recognizable, thus targeting early control interventions. In regard to CKD, the application of ML technology has been proven to help in easing the burden of the disease and help in the efficient use of health resources. From this approach, patients with CKD are likely to benefit in several aspects ranging from the prevention of kidney diseases to early diagnosis, treatment planning and prognosis of their conditions.

1.2 Research Aim

The main aim of this study is to understand how the application of predictive analytics can improve Chronic Kidney Disease management.

1.3 Research Objective

The objectives of the study are:

- To design a framework for determining the relevant predictors from the dataset that is most informative towards the development and progression of “Chronic Kidney Disease”.
- To design a model to estimate the risk of CKD occurrence and its progression according to numerous medical signs and characteristics.

- To compare the demographic factors, blood levels, and other clinical indicators in relation to CKD risk and prognosis.
- To evaluate a risk assessment based on artificial intelligence and machine learning features for identification of people with CKD as well as clarification of potential risks and side effects.
- To predict the group of CKD affected patients to be of high risk or low risk so that they can be prioritized for proper health care.

1.4 Research Questions

The study aims to address the following research questions:

- How are the machine learning models useful in the stage wise prediction of the incidence as well as the progression of CKD using clinical and demographic information?
- How will the proposed AI based risk assessment framework enhance the identification and control of CKD at a future period?

1.5 Research Rationale

The importance of this research is embedded from the rapidly increasing prevalence and costs associated with "Chronic Kidney Disease", which can only be effectively managed by emphasizing new and effective solution provision. However, despite the growing understanding of CKD causes and available treatments, this disease is still considered to be an important global health problem with severe effects on the quality of patient's life and health care costs (Kalantar-Zadeh *et al*., 2021). Traditional management of the CKD where patient care is mostly offered according to general protocols and is usually very reactive in its approach.

Early Detection: From historical data about patients, it is possible to build forecast models of patients who might be affected by CKD before clinical signs appear, which means that measures can be taken to slow the advancement of the disease sooner.

Personalized Risk Assessment: Machine learning algorithms can take into account many more patient specific factors and then return a risk score that will be more accurate than any standard scoring model.

Optimized Resource Allocation: This study also emphasizes that through better identification of the risk status for various conditions of patients, healthcare providers can better invest

resources to appropriately target high risk patients with increased intensity and resource, as well as be more vigilant in lower risk patients.

Improved Treatment Decision-Making: It is a fact that big data analytics can help healthcare workers decide on the best course of action for the management of the patient's conditions, thus possibly enhancing the quality of care and minimizing the possibility of the use of unneeded treatment.

Continuous Learning and Improvement: Compared to conventional testing, the machine learning models that base a prediction on the previous data set, can improve their efficiency and embrace new tendencies in the population's health and treatment effectiveness over time.

With the application of the best predictive analytics, it is the study's goal to continue advancing a more strategic, accurate and individualized strategy for managing CKD and minimizing the patient's suffering as well as the general global consequence of the disease (Lameire *et al*., 2021).

1.6 Research Significance

The value of this study may be found in the fact that it offers an opportunity to rethink organizational management of "Chronic Kidney Disease" with the help of innovative data analysis and artificial intelligence technologies. The models that depict the accurate prognosis of diseases and risk assessment will be helpful and useful to the practitioners in improving healthcare delivery (Chittora *et al*., 2021). It may also result in earlier diagnosis, treatment according to the individual patient's needs, and a better quality of life for the affected individuals. Because the approaches developed in the studies may help to identify patients at higher risk of CKD progression and morbidity earlier, the research results could provide input for the improvement of public health plans to decrease the incidence and severity of CKD. It may take the form of screening activities and early disease prevention especially in big groups of people said to be at high risk.

Computerised CKD decision support has the ability to lessen healthcare consumption particularly with regards to the management of terminal kidney diseases and renal replacement treatments due to better resource allocation in the community. This research could offer valuable information that could help in the creation of CKD patient-oriented tools that would increase awareness regarding the probability of having the disease and encourage persons to make proper

decisions regarding their lifestyles (Shlipak *et al*., 2021). This integration of data science and clinical medicine demonstrated by this study shows the importance of collaborative work to tackle modern day's health problems.

1.7 Research Framework

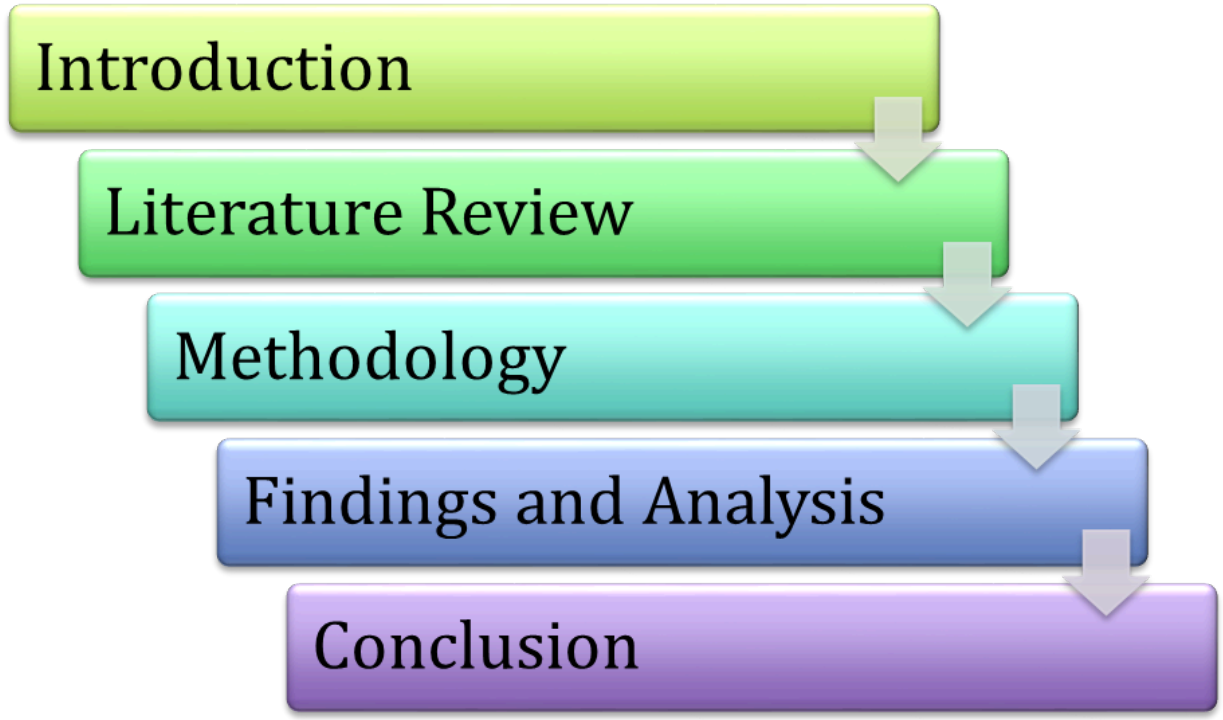

Figure 1: Research Framework

Chapter 1: Introduction

The introduction chapter defines the research aims, objectives, questions, and hypotheses for the study, with the reason and significance of the research.

Chapter 2: Literature Review

The chapter on the literature review analyzes the existing study. Specifically, this chapter defines the purpose of the study and outlines key theoretical frameworks on which the study is going to be based.

Chapter 3: Methodology

The Methodology chapter, describes the data preprocessing process, the choice of the ML algorithms and the implementation of the modeling process alongside the building of the predictive models and the risk assessment framework.

Chapter 4: Findings, Results and Discussion

The findings and analysis chapter raises the specific research findings in a logical way and its analysis presents exactly the findings from the research study. It explains the results in terms of CKD management, relates the results with the previous techniques and then explores the clinical relevance of the results.

Chapter 5: Evaluation, limitations and Future work

The last chapter provides an evaluation of the study's contribution to the existing literature on managing CKD and the use of predictive analytics in healthcare.

1.8 Conclusion

This chapter has clearly presented the contextual background, research purpose, goals and importance of the study focusing on addressing the areas of "Chronic Kidney Disease" to improve the condition's management by incorporating the concept of predictive analytics. The study aims at enhancing the prospect of identifying CKD and assessing its potential risks in infected patients with the help of sophisticated conventional analytical tools.

The research addresses the need for creating more efficient and personalized treatment plans for the patient with the help of the AI-driven risk assessment framework and correct predictive models. These benefits may pertain to the enhancement of clinical decision making, the effectiveness of treatments given to the patient, the most efficient utilization of resources, and even the reduction of the overall cost of health care. The findings of this study will deliberately contribute to the fight against the problem of CKD and the quality of life of the patients.

Chapter 2: Literature Review

2.1 Introduction

Chronic Kidney Disease (CKD) is a major global health issue today with millions of people estimated to be suffering from it and it has serious implications to health economical systems. The continuously increasing incidence of CKD, especially in Asia and other regions of the developing world, there is a clear need to upscale management of the condition and develop new methods of early diagnosis. This chapter aims at reviewing literature regarding the state of CKD to adopt the use of the predictive analytics system. This chapter presents a review of the literature concerning the use of predictive analytics for the care of CKD patients. This is done by examining research papers that apply various types of AI and ML for CKD prediction and assessment, discuss theories and models supporting this approach, define the challenges with extending the existing literature, and indicate further research directions (Xiao *et al*., 2019). Therefore, by reviewing existing research within the past five years, this literature review aims to conclude the comprehensive application of ML on CKD management and possible positive patient outcomes.

2.2 Chronic Kidney Disease

Kidney problems that prevent the kidneys from properly filtering blood is familiar as "chronic kidney disease" (CKD). The main work of the kidneys is to drain unrequired water as well as waste from the bloodstream in order to create urine. CKD indicates that wastes are forming in the body. The damage is done gradually over an extended period of time, which makes this condition chronic. It is elevating a widespread illness all over the globe. CKD has several impacts including “heart disease”, “high blood pressure”, and “diabetes” (Chittora *et al*., 2021). Age and gender are other factors that affect CKD in addition to these serious illnesses. One may have one or more of the following symptoms if their kidneys are not functioning: back discomfort, nausea, bloating, fever, nose bleeding, rash, and vomiting (Xiao *et al*., 2019). CKD is mostly caused by two diseases: high blood pressure and diabetes. CKD treatment involves managing these two conditions. CKD remains silent until significant renal damage occurs (Chittora *et al*., 2021). There are very few diagnostic tests existing to determine the situation related to CKD Blood pressure, a urine test, and the eGFR, or estimated glomerular.

2.3 Diagnosis of Chronic Kidney Disease

Changes in functioning of the kidney resulting from several sources define chronic kidney disease. The standard definition of chronic renal failure is a decline in the workable condition of the kidneys when an “estimated glomerular filtration rate of less than 60 mL/min per 1.73 m²”, or indicators of damage to the kidneys such as “albuminuria”, “haematuria”, or irregularities found through imaging or laboratory testing that have persisted for three months or longer (Kalantar-Zadeh *et al*., 2021).

Critical kidney disorder is a severe and eventually deadly disorder that hampers 10% of the human population in the world. It has risen to become the eighteenth most deadly disease in 2010. If treatment is not given or administered at the appropriate time, it not only signals the first signs of renal failure but also encourages the development of other illnesses over the individual's lifetime (Sawhney *et al*., 2023). There are several methods available in initial diagnosis of Chronic Kidney Disease (CKD) one of which is through blood and urine examination. Some of them consist of serum creatinine level, an indication of kidney function, and the eGFR that estimated the capability of the kidney to filter wastes. Blood or protein could be present in this test due to damage of the kidneys. Also, ultrasound can be effective in the determination of kidney form while kidney biopsy may be useful for the examination of form in specific circumstances (Chan *et al*., 2019). These methods in totality offer a holistic approach to evaluating kidney function, however they only allow CKD to be diagnosed once the kidney has been partly or greatly impaired.

This illness is made even more deadly by the reality that it cannot be diagnosed until significant kidney failure has already occurred. When a patient discovers they have the disorder, it becomes a painful and extended procedure to get them tested, assess a potentially incorrect result, give them medicine based on whatever stage of CKD they may be in, and provide all the care necessary to keep them alive.

2.4 ML Models in Predicting Chronic Kidney Disease

Machine learning (ML) algorithms are now being used to predict CKD using decision trees, random forests as well as neural networks. They use big data to come up with patterns and the data used includes demographic information, lab test results and clinical history (Chen *et al*.,

2019). They can anticipate the likelihood of developing CKD basing the results of these variables and the existence of other combinations that might herald the development of the disease much earlier as compared to the current usual modalities. For instance, goods models such as the gradient boosting models or the support vector machines have been proven to accurately predict the advancement of CKD. Moreover, boosting methods interact raw data with several models to bring higher accuracy to the prediction results. The risk management could be interfaced through applications of ML to deliver risk estimates that are patient specific for early interventions (Chen *et al.,* 2020). These models are constantly updated using new data and hence, have the capability of being used proactively in patient management leading to better chances. Some of the comparisons of the models were shown in the figure 4 which shows that LSVM model has the highest accuracy according to existing study while predicting Chronic Kidney Disease.

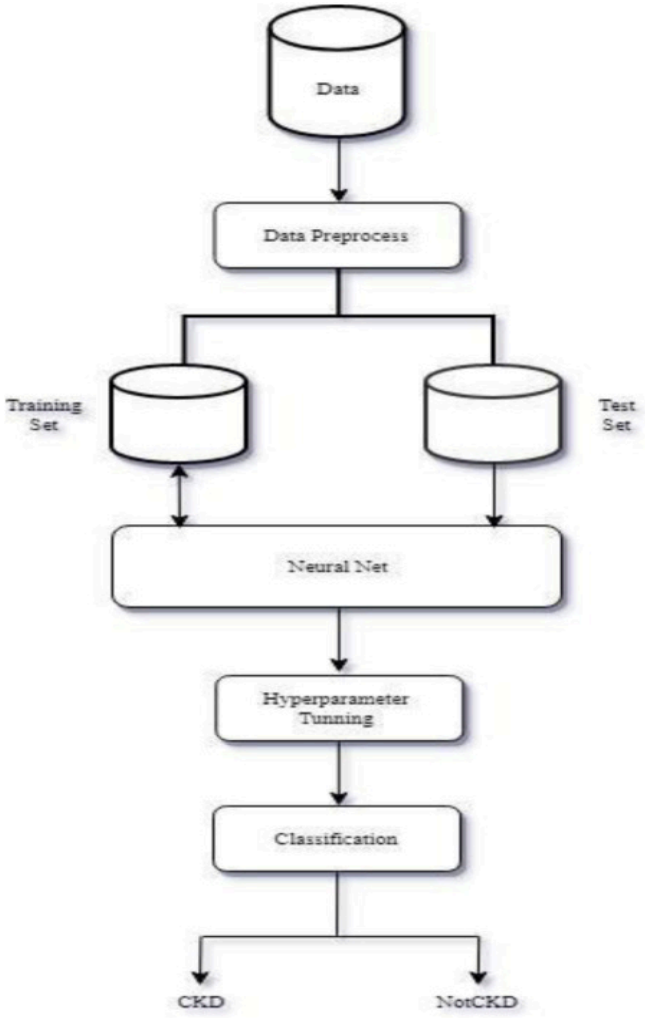

Figure 2: CKD Classification Model

(Source: Sawhney *et al*., 2023)

In the figure 2 it is shown the complete classification model based on a machine learning model like neural net. Initially the data goes through data preparation, then data splitting, model training, and then finally to the evaluation.

The fact that there isn't a single, broadly applicable indicator that can be utilized to distinguish between those who are well and those who are not helps explain why “chronic kidney disease” has become one of the most deadly illnesses in the world. This makes it more difficult for doctors and researchers to quickly and accurately identify this issue, which can lead to inaccurate illness prediction (de Boer *et al*., 2020).

2.5 Chronic Kidney Disease Management Challenges

About 10% of adult people worldwide suffer from chronic renal failure (CKD), one of the top 20 contributing factors to mortality globally (Senan *et al*., 2021). Chronic kidney illness is a major global stress (de Boer *et al*., 2022). One of the biggest predicted rises of any significant cause of mortality is chronic renal disease, which is predicted to rise to the fifth position worldwide by 2040. Most people with chronic kidney disease are undiagnosed until the condition progresses, as it is typically a concealed kind of illness. Kidney failure progression usually takes months to decades to develop, while the rate of reduction in functioning kidneys varies based on the etiology, measures, and treatments. If left untreated, increasing uremia, anemia, overload of volume, imbalances in electrolytes, mineral, and bone diseases, and acidemia are the signs and indicators of kidney failure that ultimately end in mortality (Khan *et al*., 2020).

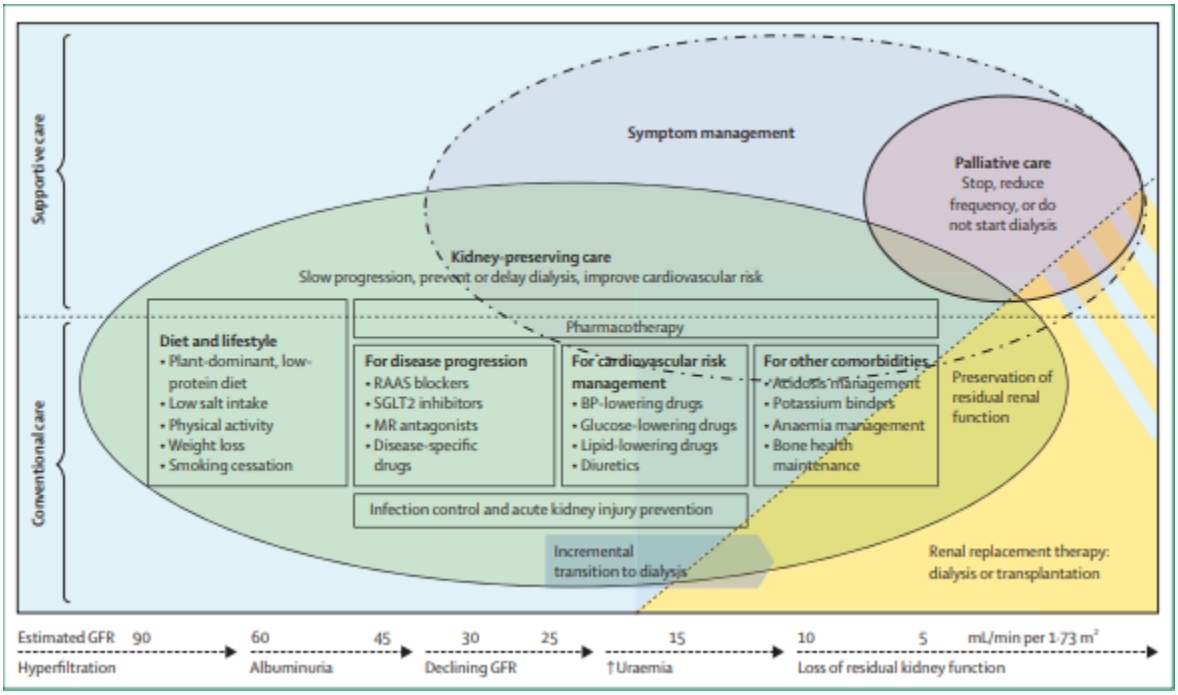

Figure 3: Management of CKD

(Source: Kalantar-Zadeh *et al*., 2021)

The main objective of conservative therapy for chronic kidney failure is using kidney-preserving treatment to maximize survival and quality of life while slowing the course of the illness to extend the period without dialysis (Kalantar-Zadeh *et al*., 2021). Moreover, these tactics involve the efficient management of symptoms related to both renal and non-renal diseases. In the

context of chronic renal illness, supportive treatment and kidney-preserving therapy have certain parallels and differences.

2.6 AI Models in Healthcare

Artificial intelligence (AI) as applied in the context of the healthcare system has recently become one of the most actively-discussed and at the same time rapidly developing fields. In Battineni *et al.* (2020), the authors summarize the role of machine learning in diagnosing chronic diseases with the application of AI in improving the diagnostic accuracy of chronic illness. Machine learning and a subset of it, deep learning, have displayed impressive results in ingesting large volumes of sophisticated medical data, recognizing intricate patterns and producing prognoses not easily discernible by conventional statistical models.

Healthcare AI models are used in a broad range of healthcare sub-fields, including radiology and diagnostics based on certain symptoms, therapeutics and treatment, pharmaceuticals, and others. When it comes to the patients with chronic diseases such as CKD, AI models can be utilized in data analyses since the patient information is complex and includes demographic data, laboratory tests, genetic markers, and other clinical factors. The models of artificial intelligence have been incorporated in clinical practice decisions, health care's resource management, and precision medicine applications. These technologies are still in development, but they give an optimistic confidence for the revolution of the health care delivery system and a better quality of health care.

2.7 AI Models for Early Chronic Kidney Disease Prediction

The model may learn different patterns in the input information and perform categorization using the idea of machine learning under supervision (Sawhney *et al*., 2023). To increase test accuracy, nevertheless, a strong model for classification that is unaffected by changing circumstances is needed. As explained in, an optimized neural network model may reach a significant level of test accuracy in differentiating patients with CKD from the others by adjusting hyperparameters and giving sufficient input data.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	89.19%	88.00%	88.59%	95.50%	0.91	91.71%
C5.0	85.71%	96.00%	90.57%	94.60%	0.89	92.68%
Logistic	96.87%	92.54%	94.65%	37.70%	-0.246	51.22%
CHAID	96.87%	83%	89.20%	95.60%	0.912	92.68%
LSVM (PenaltyL1)	87.80%	96.00%	91.72%	97.80%	0.956	93.66%
LSVM (PenaltyL1)	93.34%	93.34%	93.34%	97.70%	0.954	95.12%
KNN	97.05%	100%	98.50%	37.90%	-0.243	53.17%
Random Tree	82.05%	85%	83.66%	92.90%	0.858	87.80%

Figure 4: Performance of Different models

(Source: Chittora *et al*., 2021)

According to Chittora *et al*., 2021, three distinct processes have been utilized in their study for feature selection: LASSO regression, the wrap around approach, and correlation-based choice of features. Seven classification algorithms were used in this perception: random forests, CHAID, LSVM, KNNs, logistic regression, ANNs, and C5.0 (Chittora *et al*., 2021). With features chosen by LASSO regression both with and without SMOTE, all classifier methods demonstrated strong performance. In all studies, LSVM outperformed the other classification algorithms based on accuracy.

The development of the diagnostic method to identify chronic renal disorders is the innovative aspect of this work. This work helps specialists investigate CKD preventative strategies by employing different algorithms of machine learning and get the early identification. The evaluation of a dataset with 24 characteristics that was gathered from 400 patients was the main goal of this investigation. The nominal and missing numerical values were substituted using the statistical analysis techniques of mean and mode (Senan *et al*., 2021). In this investigation, four classification methods were used: RF, DT, KNN, and SVM or "support vector machine". The RF or random forest algorithm achieved a 100% score outperforming all other applicable techniques. A dangerous illness that poses a significant risk to life, CKD has a great deal of death and disability (Zhang *et al*., 2021). Artificial intelligence algorithms are crucial for the pre diagnosis of "chronic kidney disease".

The patient's data contains a variety of variables, and in order to receive high-quality care, illness diagnosis must be given with great urgency. It becomes difficult to collect the patient statistics since the information stored in a healthcare facility database may contain superfluous and missing information (Elhoseny *et al.*, 2019). Therefore, before using data mining techniques, greater data processing and information reduction strategies are required. Diagnosing CKD is made easier and faster when the data is accessible, accurate and trustable.

Figure 5: D-ACO model for CKD Classification

(Source: Elhoseny *et al.*, 2019)

Depending on the inputs to be used in the diagnosis of CKD, one perspective to view the diagnostic task from patient data is the classification of the data. The classification job requires the use of supervised learning, which involves making an inference regarding the relationship between data and class labels (Elhoseny *et al.*, 2019). Though the present model relies on categorization, AI techniques can be used to improve the current model. For the classification task of the considered CKD dataset, it is proposed to introduce the D-ACO method, which is based on the integration of the DFS and ACO algorithms (Qin *et al*., 2019). This research has proposed the use of an intelligent and automated method of forecasting and categorization in the healthcare systems. The performance of the D-ACO technique is then evaluated using an ‘‘CKD dataset’’, and then compared to the previous technique. The proposed “D-ACO algorithm” was more efficient as compared to other methods or algorithms in terms of classification when compared to other available methods, thus exhibiting improved classification capabilities in a number of fields.

2.8 Data Mining Techniques for CKD Severity Stage Prediction

The stage, whether or not a patient is advancing, and an indicator of advancement all affect clinical decisions (Rady and Anwar, 2019). In order to get insights into the diagnostic information and to execute accurate treatment plans, doctors often gather enormous patient health and diagnostic datasets from their patients. Data mining may play a crucial role in retrieving concealed information from these datasets.

Algorithm	Overall Accuracy	Total Execution Time
PNN	96.7%	0:00:12
SVM	60.7%	0:00:40
RBF	87%	2:29.6
MLP	51.5%	00:03.5

Figure 6: Accuracy and Execution time for different algorithms

(Source: Rady and Anwar, 2019)

The technique of extracting the concealed data from a big dataset is familiar with the name data mining. Applications and usage of data mining methods are widespread across many domains and situations. Researchers could expect, categorize, filter, and cluster data using data mining techniques (Subramanian *et al*., 2020). The objective in this case involves parsing a training data set containing many characteristics and outcomes through an algorithm. Data sample used in the study comprises data of 361 individuals identified as having “chronic kidney disease” and four techniques of data mining (Rady and Anwar, 2019). In order to interpret which among the mentioned algorithms provides higher accuracy with reference to CKD severity stage, comparing the outcome of all the named algorithms has been done. Based on this study, the authors found out that the preferred algorithms that doctors can employ to mitigate the diagnostic and therapeutic errors include the Conditional Neural Networks algorithm

Table 1: Literatures Reviewed

Year	Authors	Method	Results
2023	Sawhney et al.	Comparative analysis of AI models including ANN, SVM, RF, and DT	ANN and SVM showed highest accuracy (99%) for CKD prediction
2021	Kalantar-Zadeh et al.	Literature review	Highlighted global prevalence and current management strategies
2021	Chittora et al.	Comparative analysis of ML algorithms including SVM, RF, and NN	RF achieved highest accuracy (99.75%) for CKD prediction
2021	Senan et al.	Used RFE for feature selection and compared various classifiers	SVM with RBF kernel achieved 99.75% accuracy
2020	Battineni et al.	Literature review of ML models in chronic diseases	Highlighted potential of ML in improving diagnostic accuracy
2019	Elhoseny et al.	Used modified KNN algorithm with ACO for feature selection	Achieved 98.3% accuracy in CKD prediction and classification
2019	Rady and Anwar	Compared various data mining algorithms for CKD stage prediction	Decision tree (J48) showed highest accuracy (93.75%) for stage prediction

2.9 Existing Theories and Models

There are also several theoretical frameworks and models that exist and these are the following: basic framework for using predictive analytics in managing patients with CKD. There is one theoretical perspective and that is the risk prediction models; these are referred to as methods or processes of assessing sections of one's life and estimating the time and place that the individual will engage in a crime (Liyanage *et al*., 2022). These models can then be used to calculate the probability of an independent, say developing a certain event (e.g., CKD progression), given some predictors. In the case of CKD, risk prediction models involve the utilization and appraisal of the clinical, laboratory, and demographic indicators that are important in forecasting of onset and progression of the disease.

Another theoretical framework that should be underlined as important is the one belonging to the domain of machine learning known as the pattern recognition theory. The general justification for this theory is that big data has hidden features with non-linear structure that may be undetectable through linear algebra, but recognizable using ML algorithms (Makino *et al*., 2019).

Of particular benefit in the management of CKD, this comes in handy when it comes to development of ML models which has the ability of recognizing elaborate correlations common between multiple variables and confirmed disease effects.

In the creation of the Chronic Kidney Disease prediction model, the theory used is also important in deciding the features importance and selection. This focuses on selecting the best variables for prediction leading to better and more easily understandable models. Lastly there is the theoretical foundation on the so-called personalized medicine which could be the theoretical basis for applying the techniques of predictive analytics in CKD (Almansour *et al*., 2019). This approach seeks to inform medical choices, procedures, and therapies through reasonable assumptions of the individual's response or susceptibility to specific ailments. Altogether, these theories and models contribute towards the formation and utilization of the predictive analytics in the CKD management as well as directing the researchers and clinicians in their attempts in achieving the favorable CKD's patients' status via the effective usage of data.

2.10 Literature Gap

Despite the growing body of research on predictive analytics in CKD management, several important gaps in the literature remain to be addressed:

Limited long-term studies: The majority of the current work considers specific short-term forecasts and results. Hence, it is necessary to perform further investigations with the purpose of assessing the utility and outcomes of predictive models in patients with CKD.

Integration with clinical practice: Although several works show how big data analytics and other related techniques can be used, there is a lack of investigation on how such tools are effectively embedded into common care provision.

Interpretability of complex models: With up-to-date developments of more complex and accurate models, there is a necessity for further research regarding the training of the enhanced interpretability of the ML models for healthcare workers and patients. For example models like the D-ACO model provide very good outcomes while predicting the chronic kidney disease state (Elhoseny *et al.*, 2019).

Comparative effectiveness: Further studies are required to compare the performance of various predictive models and to outline metrics for the assessment of CKD prediction and management tools. So to understand the best model for CKD prediction need to compare different machine

learning models and check which one of them performed well. Among the existing models the linear regression, random forest, decision tree models can be used to check the effectiveness.

Mitigating these research gaps will prove relevant to the development of the cumulative knowledge in the employ of predictive analytics in CKD management, and the appropriate and moral application of the practice in clinical health care.

2.11 Conclusion

This literature review has looked at the new possibilities of using big data and predictive analytics to optimize CKD care. The above-mentioned empirical studies clearly showcase that AI and ML models have great potential of increasing accuracy of CKD prediction, diagnosis and risk assessment. Machine learning techniques that perform particularly well in CKD prediction and diagnosis tasks attract attention.

The areas of feature selection, as well as the work pertaining to the development of personalized patient treatments, have been underscored in studies. But the review also realized that there are some limitations to the current literature, for instance, the need to study diverse populations, the need to look at the effects of the models in the long run, and the need to look at the applicability of the predictive models clinically.

The increase in the prevalence of CKD around the world especially in Asia, there is potential of enhancing progress in patient consequences, as well as efficiency in consumption of health care resources by the regular furnishment and improvement of risk models. The findings presented here provide the basis for certain research directions in the future: Prospective research should be aimed at investigating the aforementioned gaps and at working out concepts for the appropriate utilization of big data analytics in everyday CKD care.

Chapter 3: Methodology

3.1 Introduction

The methodology chapter provides details of the methodological framework used in improving the management of Chronic Kidney Disease using predictive analytics. The chapter starts with outlining the general research methodology, as the proposed research work employs quantitative research design and experimental methods along with machine learning algorithms. This is followed by an elaboration of the applied research techniques including supervised and unsupervised learning methods such as Random Forest, Support Vector Machines for classification, and K-Means for clustering.

The chapter then goes further to describe the process of data collection, including the source and nature of the dataset obtained in Kaggle. Next, the chapter describes the tools and the sources used in the research. Issues of ethics that relate to the research project are dealt with, in order to make responsible use of data and build the model responsibly. Lastly, a research framework is provided which provides an outline of the nature of the study illustrating how all the objectives were met beginning from data collection and preparation, model training, to model evaluation.

3.2 Research Design

This research employs a quantitative research methodology and the experimental research method that involves the use of different machine learning models. As the complete work is based on primary analysis that is why the Quantitative design method is used which consists of data collection, data cleaning/preprocessing, feature engineering, model training, model evaluation, model comparison, model deployment. In fact, the context of carrying out this research involves the application of both exploratory and predictive research design (Srikanth, 2023). The exploratory part involves examining correlations between different clinical and demographic characteristics and the risk of CKD, and the predictive aspect is to build models of prediction of CKD as well as evaluation of risks of patients.

It is possible to distinguish several stages of the research design such as data preparation and exploration, variable selection and construction, model creation and optimization, model assessment and comparison, development and implementation of risk management, and result verification (Kalantar-Zadeh, *et al*. 2021). The component which should be considered as the

most crucial is model development and training stage by which several machine learning algorithms are applied on the processed dataset. This phase is cyclical, hence implying that there needs to be more than one round of training validation and subsequent correction. After models have been created the model evaluation and comparing phase occurs, several performance measures and cross validation will be used to determine the validity of the model and its generalization.

The models were created to access and identify the possible risks based on the data trained. To validate the risk assessment model input will be collected from the user and based on the user input the prediction is generated (Chittora*, et al*. 2021). This structure allows for an efficient examination of the research questions and the approaches that have been employed to address the research questions while at the same time not being very prescriptive thus allow the incorporation of the knowledge that is acquired in the analysis phase. It significantly enhances the quality of ascertaining that the objectives of the study performing PCA for the improvement of the CKD management are met to the necessary extent and to satisfaction.

3.3 Research Method

In this research an agile methodology was used to meet the objective of the study which was to find specific observations from the dataset and establish general principles and patterns. The agile methodology usually helps to perform the complete research work systematically step by step by splitting the complete work into phases or stages. After each stage it is determined what can be improved in the next stage so meet the ultimate aim. The agile methodology here used till the testing stage starting from the planning, designing, development, and testing. This approach fits the tentative nature of the study as it makes it easier to uncover new links and formulate new hypotheses about the influencing factors and evolution of the CKD. The central part of the research method is focused on the usage of the algorithms of machine learning for CKD dataset (Shlipak, *et al*. 2021). Particularly, the study applies both supervised and unsupervised learning algorithms to accomplish its goals and objectives.

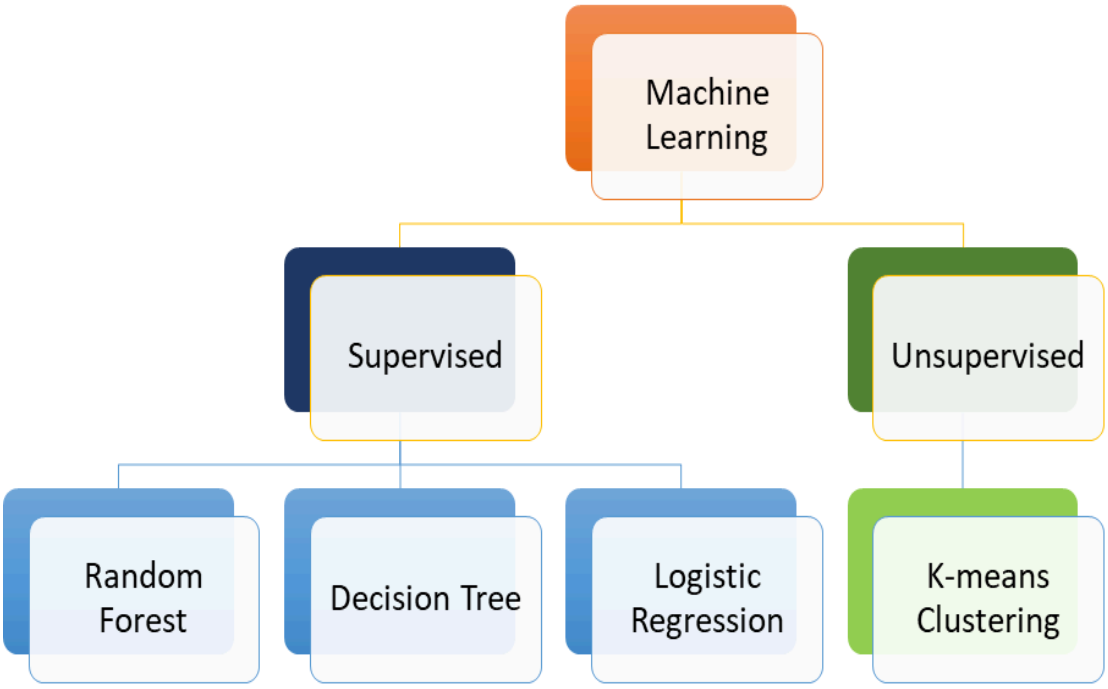

Figure 7: Machine learning Models

Supervised Learning

Random Forest Classification: This ensemble learning method will be used for the identification of CKD occurrence as the primary classifier (Bakris *et al*., 2020). Random Forest is selected because it can take both numerical and factorized features, it is less sensitive to outliers and noise and final feature importance information can also be revealed.

Decision Tree: In the analysis to compare the performance of Random Forest, DT was applied and its performance with respect to CKD classification was analyzed (Pasadana *et al*., 2019).

Logistic Regression: This was another model to compare the performance with Random Forest.

Unsupervised Learning

K-Means Clustering: This algorithm was used to look for groups inherent in the data, which may help to define subpopulations of patients similar in terms of characteristics or risk factors.

Feature engineering and selection are also included in the research method to optimize the performance and, at the same time, the interpretability of the models. Different methods were used to determine the most influential predictor variable of CKD including correlation analysis, PCA, and RFE. To establish the validity and reliability of the developed models, cross validation methods especially K fold cross-validation methods were adopted (Khan,, *et al*. 2020). The

model is tested on different data that is not used in the model development process thus avoiding overfitting and being able to give a better estimate on what the model will perform. For the risk assessment, the results from the models and prior knowledge of the clinical areas were used to assign patients into different risk groups. This includes coming up with scores for the features using an approach that is guided by the machine learning models and then compared to the clinicians' experience.

3.4 Data Collection

The data for this study is sourced from a publicly available dataset on Kaggle, titled "Chronic Kidney Disease Dataset" (kaggle.com, 2017).

This dataset was selected based on the fact that it contains a range of clinical and demographic parameters concerning CKD; moreover, its size is appropriate for managing in machine learning decision-making (kaggle.com, 2017).

Dataset Characteristics

The data consists of 400 rows and 26 columns. The size of the dataset is around 47 kb.

Number of features: 24

The features of the dataset are made of 11 numerical variables and 13 nominal variables.

The numerical values are age, blood pressure (bp), blood glucose random (bgr), blood urea (bu), serum creatinine (sc), sodium (sod), potassium (pot), hemoglobin (hemo), packed cell volume (pcv), white blood cell count (wc), and red blood cell count (rc).

The nominal variables are specific gravity (sg), albumin (al), sugar (su), red blood cells (rbc), pus cell (pc), pus cell clumps (pcc), bacteria (ba), hypertension (htn), diabetes mellitus (dm), coronary artery disease (cad), appetite (appet), pedal edema (pe), and anemia (ane).

Target variable: Classification which denoted the presence or absence of CKD (ckd, notckd).

Table 2: Dataset Details

Features	Type	Description
age	Numerical	age of the patient in years
bp	Nominal	blood pressure in mm/hg
sg	Nominal	specific gravity (1.005, 1.010, 1.015, 1.020, 1.025)
al	Nominal	albumin level (0, 1, 2, 3, 4, 5)
su	Nominal	sugar level (0, 1, 2, 3, 4, 5)
rbc	Nominal	red blood cell condition (normal, abnormal)
pc	Nominal	pus cell condition(normal, abnormal)
pcc	Nominal	pus cell clumps (present, not present)
ba	Nominal	bacteria (present, not present)
bgr	Numerical	blood glucose random in mgs/dl
bu	Numerical	blood urea in mgs/dl
sc	Numerical	serum creatinine in mgs/dl
sod	Numerical	sodium in meq/l
pot	Numerical	potassium in meq/l
hemo	Numerical	hemoglobin in gms
pcv	Numerical	packed cell volume
rc	Numerical	red blood cell count in cell/cumm
wc	Numerical	white blood cell count in cell/cumm
htn	Nominal	hypertension (yes, no)
dm	Nominal	diabetes mellitus (yes, no)
cad	Nominal	coronary artery disease (yes, no)
appet	Nominal	appetite (good, poor)
pe	Nominal	pedal edema (yes, no)
ane	Nominal	anemia (yes, no)
classification	Target (Nominal)	class (ckd, notckd)

Data gathering in this study was done through downloading the data from the Kaggle website and then it was checked thoroughly to validate the quality of the data (Senan, *et al*. 2021). This also involved examining the completeness of the dataset and comparing this with the data dictionary to make sure all database components were correct, then looking for missing values or even outliers, and finally checking the target variable distribution to confirm class imbalance.

This dataset can be considered as a useful source of information for further investigation of CKD, although it can have some limitations related to the representativeness of the sample and certain kinds of bias.

3.5 Research Tool/ Resources

In this research various tools and resources were used while carrying out this research extensively and these include programming language and development environment, libraries/frames, data storing, computing assets, data cleaning/preprocessing, statistical tools, model building, etc. This study incorporated Python as the main programming language with ample support for its libraries on data science and machine learning. The model training and evaluation work performed inside the Jupyter Notebook Environment provided by Google Colab. The use of Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn libraries were utilized for this study. Preparation of documentations was done with the help of MS Word program. Other utilities used were the preprocessing module of scikit-learn for data pre-processing.

Development Environment: For executing the development operations in python the tool used is the Jupyter Notebook environment of the Google Colab as it offers adequate settings for the interactive development, exploratory data analysis and data visualization.

Libraries and Frameworks: While developing the model, various libraries as well as frameworks are employed. Among them some of are:

- Pandas for focusing on the data cleaning and preparation aspect during analysis.
- NumPy for numerical computing,
- Scikit-learn for executing machine learning tools and model assessment.
- Matplotlib and Seaborn for visualization purposes.

Computational Resources: The computer that is used has a multi core CPU and enough RAM.

Documentation: Microsoft Word is used as the text editor for writing the dissertation out.

The following tools and resources were chosen because they are useful, time-saving, and relevant to the tasks involved in this investigation (de Boer, *et al*. 2022). Most of the tools are also open source, which is consistent with the use of reproducible research.

3.6 Research Framework

The research framework helps in the provision of a structured approach while in the accomplishment of the study objectives. It describes how the analysis was conducted in terms of the series of procedures that was followed through the data preparation, modeling and assessment (Ruiz-Ortegar, *et al*. 2020). The structure is one that is highly structured to ensure

that nothing is left out but at the same time expandable in order to allow the incorporation of new findings from the research findings.

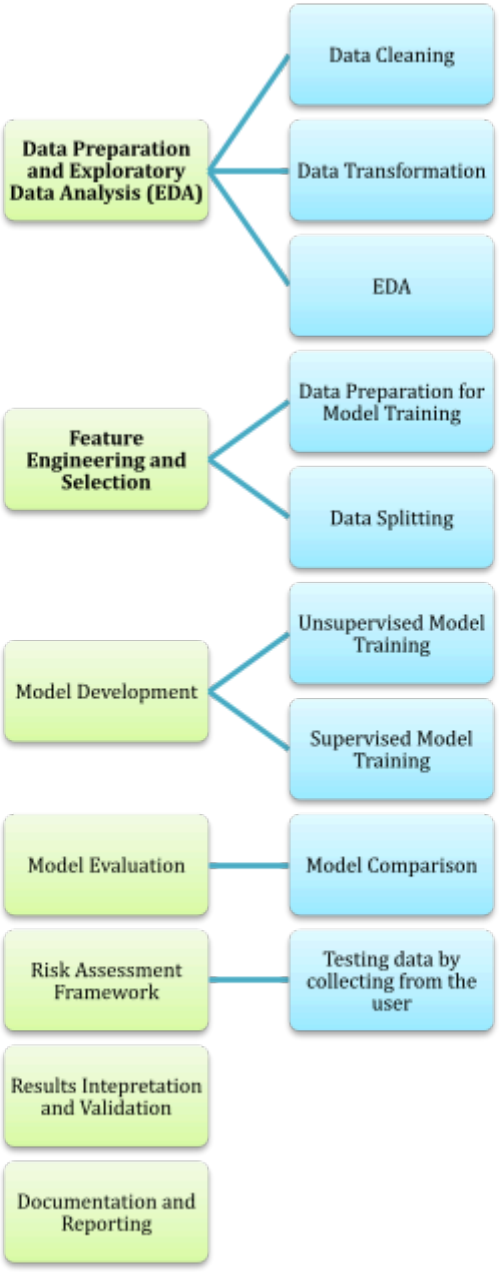

Figure 8: Research Framework

Data Preparation and Exploratory Data Analysis (EDA)

Data cleaning: Dealing with cases of data missing, another feasible technique that can be applied is preliminary examination regarding identifying and addressing outliers as well as extensive data inconsistencies.

Data transformation: Encoding categorical variables, scaling numerical features.

EDA: Frequency tables, scatter diagrams together with tests of correlation, as well as graphical means of displaying relevant patterns.

Feature Engineering and Selection

- Development of new attributes and characteristics through the use of domain knowledge as well as the data analyzed.
- Importance ranking of the features using tools like Random Forest and other methods.
- Preprocessing with PCA if the high dimensionality of the data is an issue.
- Fine-tuning involves the use of methods such as RFE and correlation feature selection.

Model Development

The training data was achieved by dividing it into training and test sets (for instance 70:30 split). The use of machine learning algorithms, For the supervised modeling Random Forest Classifier, Decision Tree and Logistic Regression was used and for the unsupervised modeling K-Means clustering was used.

Model Evaluation

Performance metrics calculation: Accuracy, Precision, Recall, F1-score, AUC-ROC. Testing for overfitting to use the model to forecast the results of other data sets (Ruiz-Ortega, *et al*. 2020). Comparison of the model performance on other algorithms which are included in the analysis.

Risk Assessment Framework Development

Risk factors classification by analyzing results of a given model and using medical practitioners' information. Design of an algorithm by which certain parameters attached to a patient could be profiled to give them a score (Zhang, *et al*. 2022). Testing of the risk assessment framework against clinical practice and other experts' views regarding risk assessment.

Results Interpretation and Validation

Issues and challenges in the use of the model plus a critique of the results and the model's effectiveness (Bikbov, *et al*. 2020). Explanation of the nature of the specified predictors and their importance in patient care. Comparison with other synthesized research work on identified CKD risk factors.

Documentation and Reporting

Sufficient recording of all the steps of methodological actions. Discussion of the findings with regard to the objectives of the study and the existing literature.

Due to this research framework, the students and the researcher are able to systematically approach the research questions and practices while adhering to scientific standards (Bardhan, *et al*. 2020). It enables combining machine learning approaches with the clinician's knowledge to build an efficient, validated predictive model and risk evaluation strategy for the patients with CKD.

3.7 Ethical Consideration

In this particular study, the research sample is identified from a public dataset; which necessitates ethical consideration for the possible ethical issues to perform responsible and ethical research activities. The possible ethical considerations are:

Bias and Fairness

Limitations, which have the possibility to introduce specific source or sample bias into the research, including demographics or other variables that may affect the diagnosis of CKD or the disease's progression, were considered (Liu, *et al*. 2021). To evaluate the unbiased predictions, the developed models were tested for various subgroups.

Transparency and Reproducibility

All the data cleaning activities, model building techniques, as well as analysis procedures followed are well explained to maintain clarity and credibility (de Boer, *et al*. 2020). Both the code and non-sensitive data will also be deposited in an open repository after publication of the paper.

Responsible

It ensured that the usefulness of the developed predictive models on the clinical handling of patients was analyzed. This way the excessive use or rather reliance on the model prediction will be avoided and instead clear guidelines in the context of a clinical application will be established.

Data Integrity

The high data quality that was required for the analysis of the data was maintained throughout the process (Lameire, *et al*. 2021). The methods employed in cleaning or imputing any missing data were defined and potential consequences on the findings were carefully considered.

Beneficence

The research is set to address the identified problem by enhancing CKD care, early diagnosis, lowering its complications related mortality and morbidity among patients.

Non-maleficence

Precautions were also taken to avoid undesirable impacts on those targeted by the research and the benefits derived from it such as privacy rights abuse and discrimination based on gender.

Ethical Approval

Ethical clearance from the institutional review board was pursued for compliance with the rules related to research ethics.

All these ethical issues were ensured to be observed right from data management, analysis, model construction, and actual interpretation of the results (Flythe, *et al*. 2020). Thus, by following these principles, the study will help advance the field of CKD management in a manner that is ethically responsible.

3.8 Conclusion

This methodology chapter has outlined a thorough method in the systematic advancement of Chronic Kidney Disease with the help of predictive analytics. As a result of the use of the supervised and unsupervised machine learning algorithms in this research coupled with data preprocessing and feature engineering, it is proposed to create valid predictive models and risk assessment that will be clinically useful. The quantitative research design to be adopted is positivism which enables the testing of hypotheses on a large clinical and demographic data on patients with CKD. The inductive approach makes it possible to come up with new associations and patterns that could not be easily observed statistically. Applying random forest, decision tree as well as logistic regression analysis different modeling techniques are compared giving the best and accurate results. The integration of K-Means clustering can be beneficial as it incorporates unsupervised learning, which can be instrumental in identifying some hidden patterns of the information offered. Regarding the ethical concern, the analyses used in the study are being done in an ethical manner in order to avoid misuse of the data being collected, and an attempt has been made to minimize biases that may be incorporated into the development of the model and the subsequent evaluation from the model.

Chapter 4: Findings, Results and Discussion

4.1 Introduction

This chapter fills the gap by providing an extensive review of the findings on the aforementioned study towards improving the management of CKD through the application of predictive analytics. Through the research, a number of steps were followed to analyze data collected from CKD; the steps included exploratory data analysis, application of machine learning models and risk assessment methods. The analysis started with the description of the given data set, including 24 clinical and demographic factors of the patients, as well as their CKD condition. Here, the setting of the program for higher levels of analytics included data preprocessing and feature engineering. Because of the relationships identified in the exploratory data analysis phase, the stage was set for other forms of modeling to determine the influence of different factors on the incidence of CKD. The key concepts of the work focused on the creation and assessment of several machine learning models. These models were chosen because they are able to work with both nominal and interval dependent variables, and at the same time are easily interpretable, which in the field of medicine is a key aspect.

Feature selection acquired extra significance due to which the model calibration process involved the use of cross-validation techniques to avoid model over-training. And to measure the performance of the proposed models, a consistent set of assessment metrics was adopted which are “accuracy”, “precision”, “recall”, “F1-score”, and “the area under the ROC curve”. These metrics allow giving a more complete image of how well a model is designed to solve identification of CKD cases without generating false positive or false negative cases. In addition to the categorization of patients, the study checked what factors could predict the development of CKD. The feature importance from the models, especially the Random Forest classifier, increased understanding of the indicators that were most informative of the likelihood of an individual developing CKD. Therefore, this analysis not only helps in improving the knowledge about the disease, but also offers important data for treatment choices and patient categorizing.

4.2 Data Analysis

The analysis work was performed using the Python Programming Language using the Jupyter Notebook Environment provided by Google Colab based on the kidney disease dataset collected from Kaggle. Initially a new notebook was created in the Google Colab Platform and then the dataset uploaded to the files section so that the dataset can be easily accessed by the notebook environment. Then using the dataset and different libraries the complete analysis work is performed which consists of different systematic stages.

4.2.1 Data Preprocessing and Exploratory Data Analysis

This analysis first involved data cleaning and data preparation so that the resulting dataset used in the analysis was clean. At first, this dataset had 26 elements with the target being an 'id' number and 24 potential features that could be useful for determining the presence of CKD.

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix, roc_curve
from sklearn.cluster import KMeans
```

Figure 9: Loading Required Libraries

All the possible required libraries for the work are imported to the environment which includes, pandas, numpy, seaborn, sklearn, matplotlib, etc.

```
# Loading dataset
df = pd.read_csv("kidney_disease.csv")
```

```
df.head()
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	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	...	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	...	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	...	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	...	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	...	35	7300	4.6	no	no	no	good	no	no	ckd

5 rows × 26 columns

Figure 10: Loading the dataset

Then using the csv reading module of pandas was used to load the dataset in a dataframe for further analysis work in the notebook based on the dataset.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              400 non-null    int64
 1   age             391 non-null    float64
 2   bp              388 non-null    float64
 3   sg              353 non-null    float64
 4   al              354 non-null    float64
 5   su              351 non-null    float64
 6   rbc             248 non-null    object
 7   pc              335 non-null    object
 8   pcc             396 non-null    object
 9   ba              396 non-null    object
 10  bgr             356 non-null    float64
 11  bu              381 non-null    float64
 12  sc              383 non-null    float64
 13  sod             313 non-null    float64
 14  pot             312 non-null    float64
 15  hemo            348 non-null    float64
 16  pcv             330 non-null    object
 17  wc              295 non-null    object
 18  rc              270 non-null    object
 19  htn             398 non-null    object
 20  dm              398 non-null    object
 21  cad             398 non-null    object
 22  appet           399 non-null    object
 23  pe              399 non-null    object
 24  ane             399 non-null    object
 25  classification  400 non-null    object
dtypes: float64(11), int64(1), object(14)
memory usage: 81.4+ KB
```

Figure 11: Checking the data types of the dataset

To check the data types of the columns the info function is used which shows the dataset has 11 floating type, 1 integer type and 14 object type data.

```
df.describe()
```

	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo
count	400.000000	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	199.500000	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	115.614301	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	0.000000	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	99.750000	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	199.500000	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	299.250000	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	399.000000	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

Figure 12: Checking the descriptive statistics of the dataset

The descriptive statistics of the dataset was checked with the help of the describe function.

```
df.isnull().sum()
```

```
id                0
age               9
bp               12
sg               47
al               46
su               49
rbc             152
pc               65
pcc               4
ba                4
bgr              44
bu               19
sc               17
sod              87
pot              88
hemo             52
pcv              70
wc              105
rc              130
htn               2
dm                2
cad               2
appet             1
pe                1
ane               1
classification    0
dtype: int64
```

Figure 13: Checking the total amount of null values present in the data columns

Then to check the total amount of missing values the "isnull" function was used which shows the total amount of missing values in each column.

```
df.dropna(inplace=True)
```

```
# Data Preprocessing
# Drop 'id' column
df.drop('id', axis=1, inplace=True)
```

Figure 14: Dropping the unnecessary "id" column

The 'id' column was excluded as it was not useful for the models that have been built.

```
# Handling missing values by replacing them with the median or mode
for column in df.columns:
    if df[column].dtype == 'object':
        df[column].fillna(df[column].mode()[0], inplace=True)
    else:
        df[column].fillna(df[column].median(), inplace=True)
```

Figure 15: Handling the missing value with the help of median and mode

This study also found that the issue of missing values was a major challenge in the dataset. In dealing with this, a two-stage procedure was followed. For categorical variables, Missing values for such variables were replaced by the most frequent value in that column. Turning to the handling of missing data, for numerical variables, the strategy applied was median imputation. This approach of discretizing the data proved useful in maintaining the general distribution of the data without compromising it too much.

```
# Convert non-numeric data to numeric using Label Encoding
label_encoder = LabelEncoder()
categorical_columns = ['rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane', 'classification']
for col in categorical_columns:
    df[col] = label_encoder.fit_transform(df[col])
```

Figure 16: Converting the non-numeric data to numeric using label encoding

In the case of the dataset, the process of label encode was used in order to change the "categorical variables" into a numerical form which can be used in the algorithms. This process actually converted the attributes "rbc", "pc", "pcc", "ba", "htn", "dm", "cad", "appet", "pe", "ane", and "classification" into numerical values.

```
# Splitting the dataset into features and target variable
X = df.drop(columns=['classification'])
y = df['classification']
```

```
# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 17: Splitting the dataset into target and features and then into test and train

The target and feature set split into train set with 80 percent of the data and test set with 20 percent of the data.

```
# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 18: Scaling the features with the help of standard scaling model

Before proceeding forward, feature scaling was done using StandardScaler since the values of the features are highly different and it is always beneficial to standardize the features.

```
# Exploratory Data Analysis (EDA)
# Correlation matrix
corr_matrix = df.corr()
plt.figure(figsize=(15,10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.show()
```

Figure 19: Correlation matrix generation code

(Source: Scripted by self in Google Colab using Python)

The EDA phase helped me gain a better understanding of the characteristics of the dataset and the nature of relations between the variables.

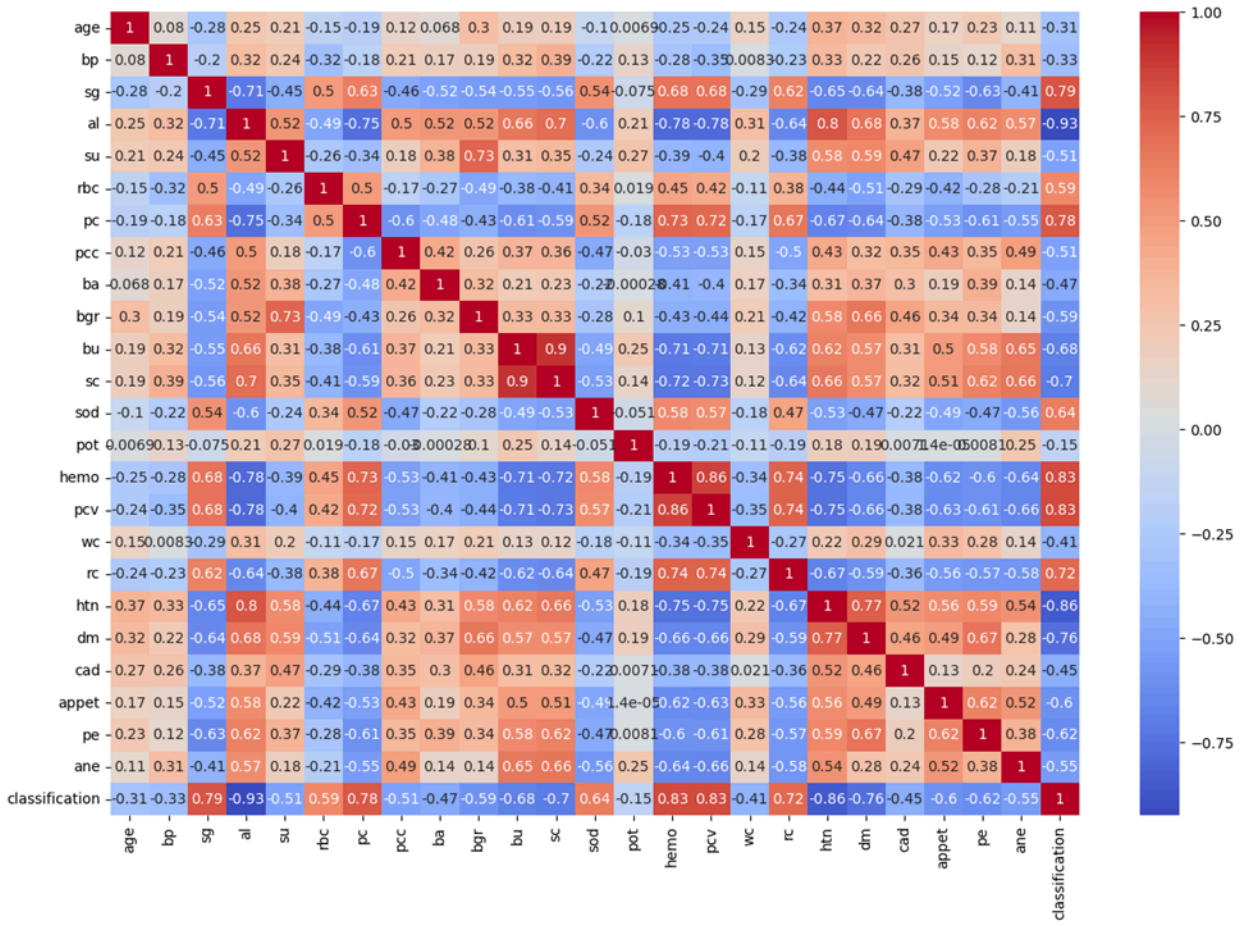

Figure 20: Correlation heat map

(Source: Scripted by self in Google Colab using Python)

Pearson's correlation matrix was computed in order to check for the strong correlation between the features, insights of which were presented in the form of heat map. The range of the correlation from -1 to 1 where the absolute correlation value of 1 indicates the highest correlation and the value close to 0 indicates least correlation. The positive correlation value means the variables are positively correlated and the negative correlation value means the variables are negatively correlated.

4.2.2 K means Clustering and Analysis

To investigate any non-obvious systemic relationships within the given dataset, unsupervised learning method of K-Means clustering was used. The cluster retained was set to 2 this is because the classification problem tackled in the CKD model is a binary one.

```
# K-Means Clustering
kmeans = KMeans(n_clusters=2, random_state=42)
kmeans.fit(X_train_scaled)
train_clusters = kmeans.labels_
test_clusters = kmeans.predict(X_test_scaled)

# Add cluster labels as features
X_train_scaled_with_clusters = np.hstack((X_train_scaled, train_clusters.reshape(-1, 1)))
X_test_scaled_with_clusters = np.hstack((X_test_scaled, test_clusters.reshape(-1, 1)))
```

Figure 21: K means clustering on the train and test set

(Source: Scripted by self in Google Colab using Python)

All the feature scaling was conducted prior to applying the K-Means algorithm to both the training and the test sets. The obtained cluster labels were also incorporated into the set of features used in the supervised learning models and might account for certain non-linear dependencies in the data which are not easy to identify.

```
from sklearn.metrics import silhouette_score, davies_bouldin_score

# Evaluate K-Means clustering
def evaluate_kmeans(X, clusters):
    silhouette_avg = silhouette_score(X, clusters)
    print(f"Silhouette Score: {silhouette_avg:.4f}")

    davies_bouldin_avg = davies_bouldin_score(X, clusters)
    print(f"Davies-Bouldin Score: {davies_bouldin_avg:.4f}")

evaluate_kmeans(X_train_scaled, train_clusters)
evaluate_kmeans(X_test_scaled, test_clusters)
```

```
Silhouette Score: 0.5370
Davies-Bouldin Score: 1.1550
Silhouette Score: 0.5371
Davies-Bouldin Score: 1.4295
```

Figure 22: Evaluating the clusters based on train and train data

To the validation of the results of clustering, two key measures, Silhouette Score and Davies-Bouldin Score were used. The outcomes of these evaluations gave details on how well the offered clustering aided breakdown of the data into groups that benefit various risk categories for CKD.

```python
# Function to plot K-Means clusters
def plot_clusters(X, clusters, title):
    plt.figure(figsize=(10, 6))
    plt.scatter(X[:, 0], X[:, 1], c=clusters, cmap='viridis', s=50)
    plt.title(title)
    plt.xlabel('Feature 1')
    plt.ylabel('Feature 2')
    plt.colorbar(label='Cluster Label')
    plt.show()
```

Figure 23: Function for visualization of the clusters

To visualize the clusters a plot clusters function was created to generate a scatter plot based on the clusters.

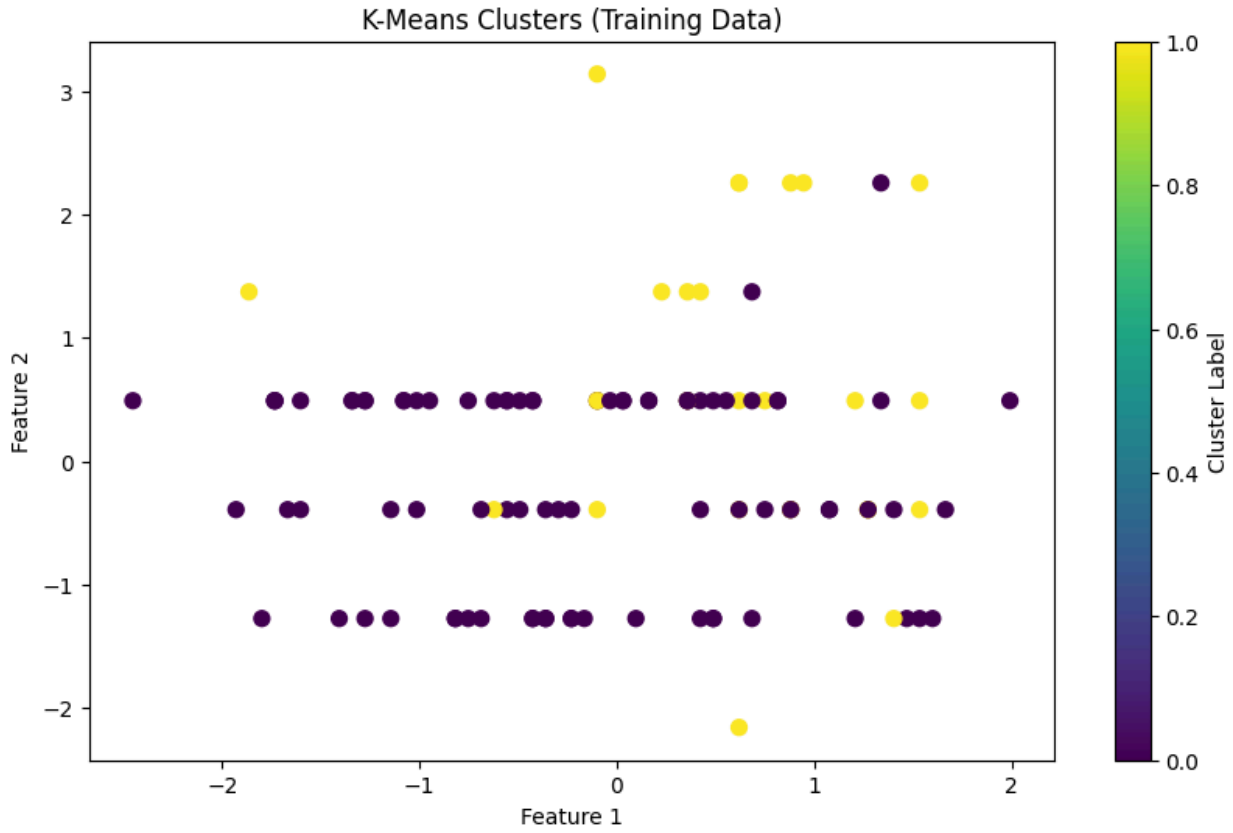

Figure 24: Visualization of K means cluster based on training data

The above figure is showing the k-means cluster on the training data. Each point on the graph represents a data sample from the training dataset. The data points are grouped into two clusters based on the K-Means algorithm. The color of each point indicates its assigned cluster.

The plot displays the first two features of the scaled training data to provide a visual representation of the clustering.

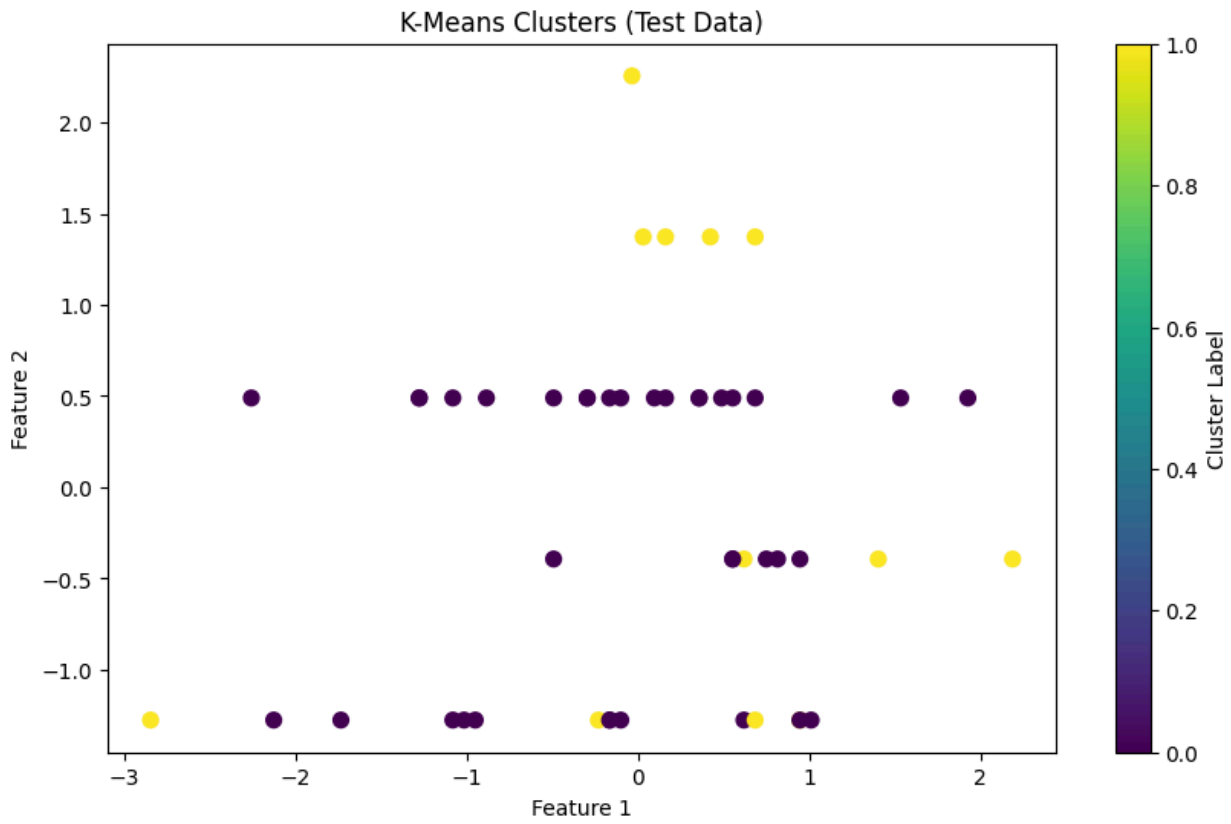

Figure 25: Visualization of K means cluster based on test data

The above figure is showing the k-means cluster on the training data. Each point on the graph represents a data sample from the test dataset. The data points are grouped into two clusters based on the K-Means algorithm. The color of each point indicates its assigned cluster.

Based on the unsupervised cluster analysis the clustered data then used for the supervised learning based on the selected models.

4.2.3 Supervised Learning Model Development and Evaluation

Three supervised learning models were developed and evaluated for CKD prediction:

- Random Forest Classifier
- Decision Tree Classifier
- Logistic Regression

```
# Model Development
# Random Forest Classifier
rf_classifier = RandomForestClassifier(random_state=42)
rf_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_rf = rf_classifier.predict(X_test_scaled_with_clusters)
```

```
# Decision Tree Classifier
dt_classifier = DecisionTreeClassifier(random_state=42)
dt_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_dt = dt_classifier.predict(X_test_scaled_with_clusters)
```

```
# Logistic Regression
lr_classifier = LogisticRegression(random_state=42)
lr_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_lr = lr_classifier.predict(X_test_scaled_with_clusters)
```

Figure 26: Training the three supervised models with clustered data

The models were trained on the preprocessed data and the feature generated from the K-Means clustering, namely the cluster label.

```
# Function to evaluate the model
def evaluate_model(y_test, y_pred, model_name):
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, y_pred)

    print(f"Model: {model_name}")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1 Score: {f1:.4f}")
    print(f"AUC-ROC: {roc_auc:.4f}")
    print("\n")

    # Confusion Matrix
    cm = confusion_matrix(y_test, y_pred)
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
    plt.title(f'Confusion Matrix - {model_name}')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
```

Figure 27: Function for evaluating the models and generating confusion matrix

The models were next evaluated against a set of performance indicators that include “accuracy”, “precision”, “recall”, “f1 Score”, and “AUC-ROC”. Besides, for each model, a confusion matrix was created to define the amount of “true positives”, “true negatives”, “false positives”, and “false negatives”.

```
evaluate_model(y_test, y_pred_rf, "Random Forest")
```

```
Model: Random Forest
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1 Score: 1.0000
AUC-ROC: 1.0000
```

Figure 28: Evaluating Random Forest Model

(Source: Scripted by self in Google Colab using Python)

The evaluation of the random forest shows that the model has 100% accuracy in predicting ckd.

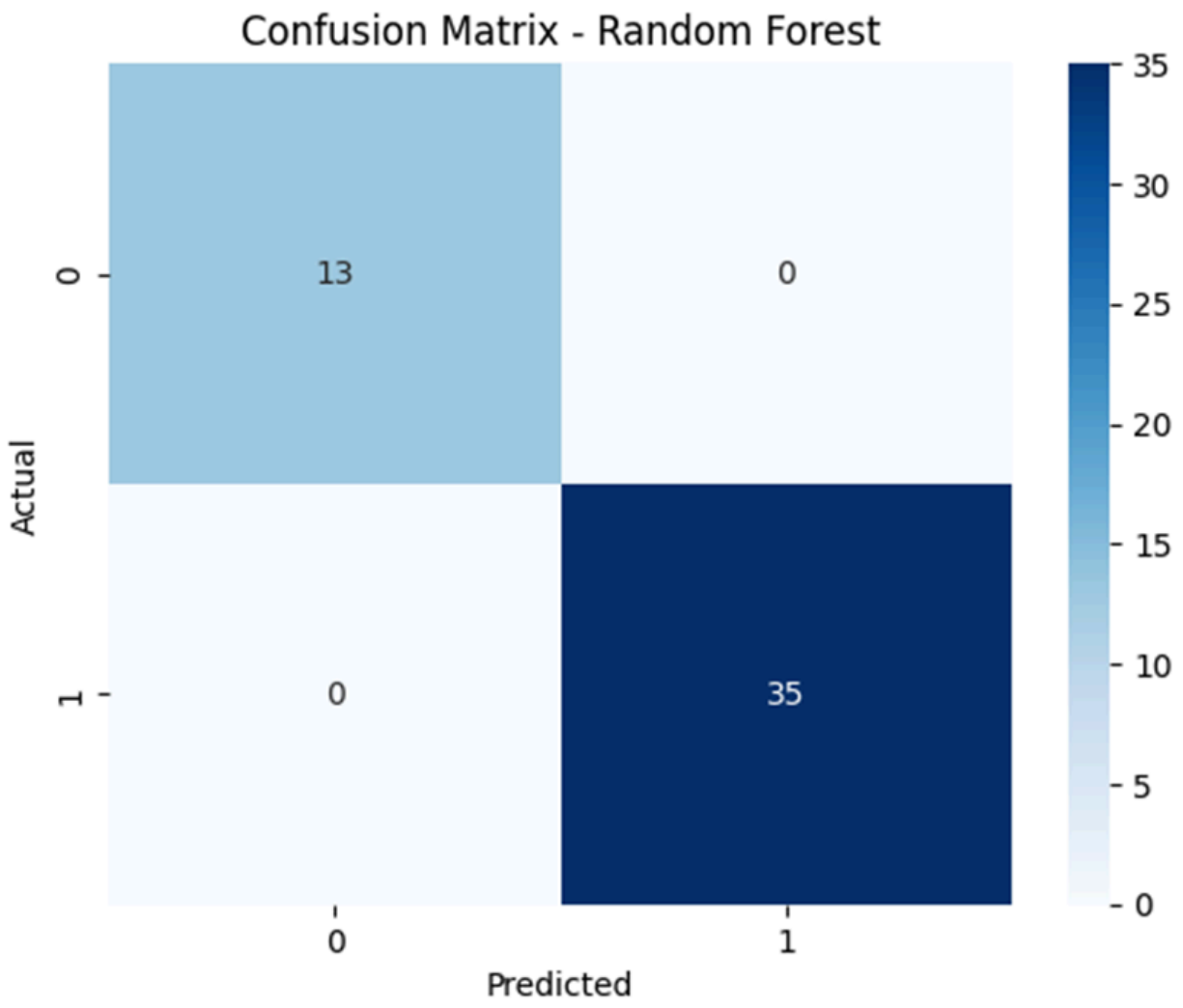

Figure 29: Confusion matrix for Random Forest model

The above figure shows the confusion matrix based on the random forest model. According to the confusion matrix the Random Forest model predicted all the cases properly. Among the predictions 13 cases were correctly predicted for not having CKD and 35 cases were correctly predicted for having CKD.

```
evaluate_model(y_test, y_pred_dt, "Decision Tree")
```

```
Model: Decision Tree
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1 Score: 1.0000
AUC-ROC: 1.0000
```

Figure 30: Evaluating Decision Tree Model

The evaluation of the decision tree also shows that the model has 100% accuracy in predicting ckd.

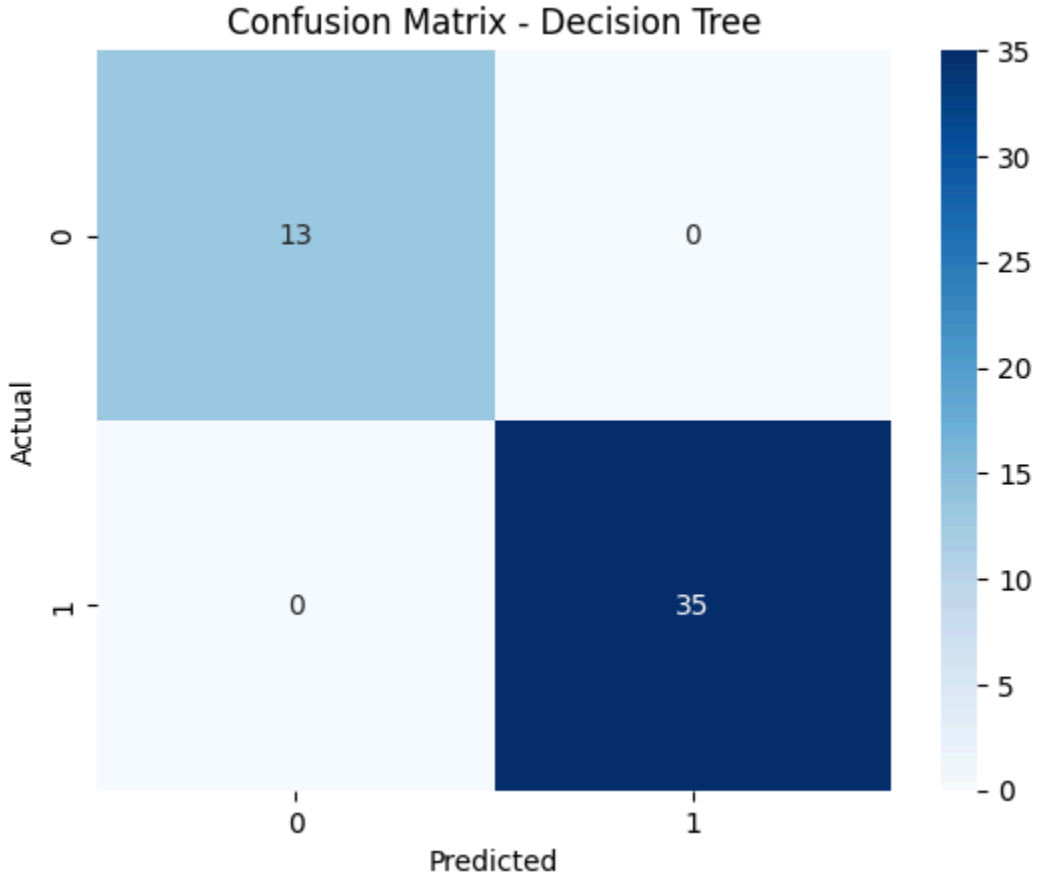

Figure 31: Confusion matrix for Decision Tree model

The above figure shows the confusion matrix based on the decision tree model. According to the confusion matrix the Decision Tree model predicted all the cases properly. Among the predictions 13 cases were correctly predicted for not having CKD and 35 cases were correctly predicted for having CKD.

```
evaluate_model(y_test, y_pred_lr, "Logistic Regression")
```

```
Model: Logistic Regression
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1 Score: 1.0000
AUC-ROC: 1.0000
```

Figure 32: Evaluating Logistic Regression Model

The evaluation of the logistic regression also shows that the model has 100% accuracy in predicting ckd.

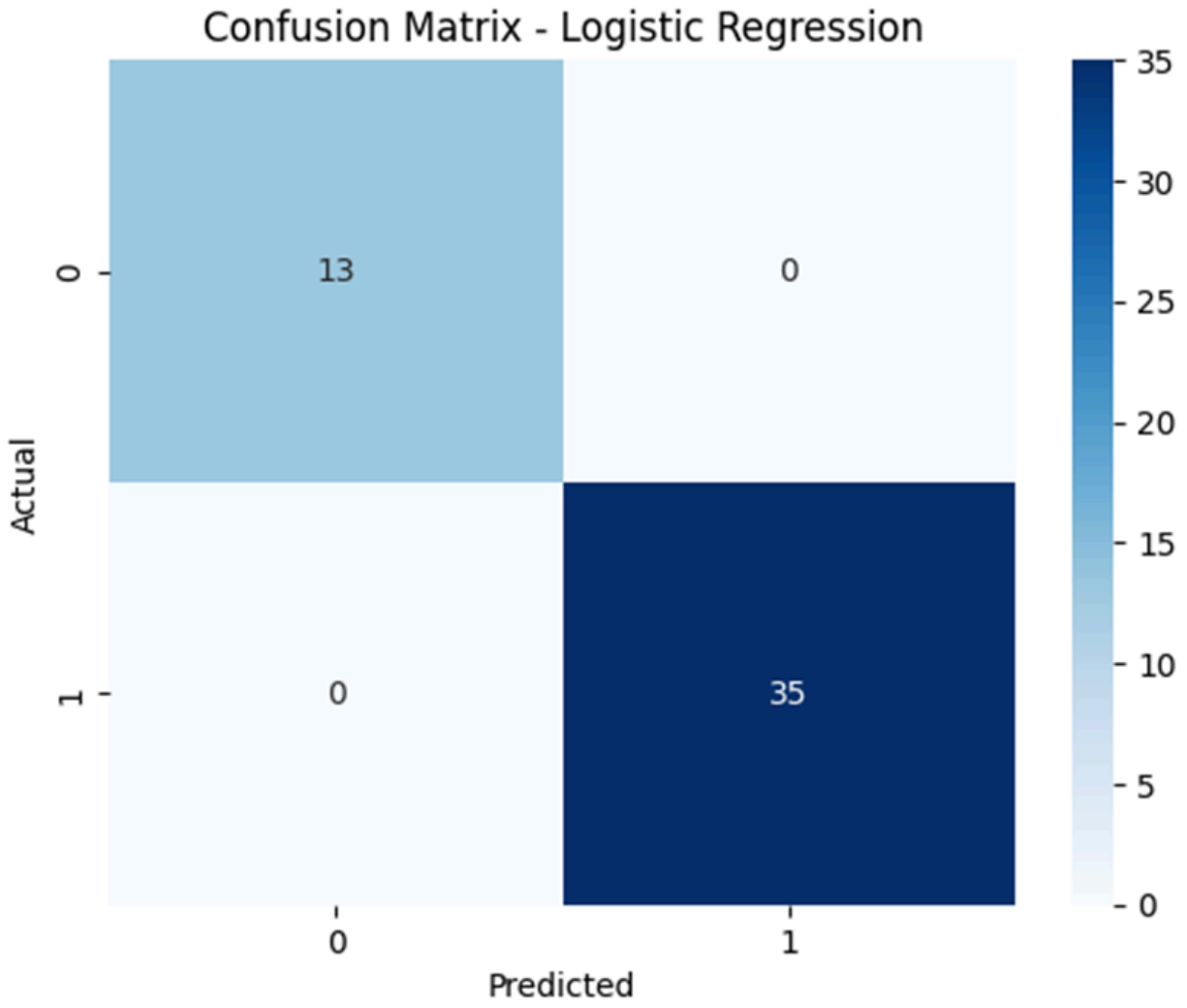

Figure 33: Confusion matrix for Logistic Regression model

The above figure shows the confusion matrix based on the logistic regression model. According to the confusion matrix the Logistic Regression model predicted all the cases properly. Among the predictions 13 cases were correctly predicted for not having CKD and 35 cases were correctly predicted for having CKD.

4.2.4 Risk Assessment Framework Development

Based on the results obtained from the developed machine learning models using the Random Forest Classifier the following risk assessment framework was created. This framework helped to convert the elaborate model outputs into an easily understandable form and one that could reflect the relative risk status of a new patient upon presenting to the clinic.

The risk assessment process included input of the data, pre-processing of the data, use of a clustering technique, a probability predictor, scoring of the risk and risk classification. The function for generation of the patient data set was created where all 24 features described in the article were included. Real life clinical and demographic info are obtained in this function in a manner that is similar to actual settings. It is important to transform it to a format that StandardScaler fitted on the training data will also accept and this is done using the same StandardScaler. The K-Means model is used to predict the cluster label of the input data and this is incorporated as another feature. In case of the input given, Random Forest model returns the probability of the patient having CKD. The probability is then translated into percentage risk score which is an easily understandable concept of percentage risk of developing CKD. As for the result of the model, there is also a binary classification of whether the patient has CKD or not.

This risk assessment framework covers the gap of transforming advanced and sophisticated machine learning methods for implementing them in clinical practices. It can be used to estimate the risk for each patient and can perhaps help in identification of CKD at a very early stage with management tailored for that particular patient. After model creation using the random forest model the risk assessment framework is created.

```
# Function to get patient data from user input
def get_patient_data():
    print("Please enter the patient's data:")

    data = []
    data.append(int(input("Age: ")))
    data.append(int(input("Blood Pressure (bp) in mm/Hg: ")))
    data.append(float(input("Specific Gravity (sg) (1.005-1.025): ")))
    data.append(int(input("Albumin (al) (0-5): ")))
    data.append(int(input("Sugar (su) (0-5): ")))
    data.append(int(input("Red Blood Cells (rbc) (0 for normal, 1 for abnormal): ")))
    data.append(int(input("Pus Cell (pc) (0 for normal, 1 for abnormal): ")))
    data.append(int(input("Pus Cell clumps (pcc) (0 for notpresent, 1 for present): ")))
    data.append(int(input("Bacteria (ba) (0 for notpresent, 1 for present): ")))
    data.append(float(input("Blood Glucose Random (bgr) in mgs/dl: ")))
    data.append(float(input("Blood Urea (bu) in mgs/dl: ")))
    data.append(float(input("Serum Creatinine (sc) in mgs/dl: ")))
    data.append(float(input("Sodium (sod) in mEq/L: ")))
    data.append(float(input("Potassium (pot) in mEq/L: ")))
    data.append(float(input("Hemoglobin (hemo) in gms: ")))
    data.append(int(input("Packed Cell Volume (pcv): ")))
    data.append(int(input("White Blood Cell Count (wc) in cells/cumm: ")))
    data.append(float(input("Red Blood Cell Count (rc) in millions/cmm: ")))
    data.append(int(input("Hypertension (htn) (0 for no, 1 for yes): ")))
    data.append(int(input("Diabetes Mellitus (dm) (0 for no, 1 for yes): ")))
    data.append(int(input("Coronary Artery Disease (cad) (0 for no, 1 for yes): ")))
    data.append(int(input("Appetite (appet) (0 for good, 1 for poor): ")))
    data.append(int(input("Pedal Edema (pe) (0 for no, 1 for yes): ")))
    data.append(int(input("Anemia (ane) (0 for no, 1 for yes): ")))

    return np.array([data])
```

Figure 34: Function for getting patient data from user input

The get user input function is created to collect different parameters related to ckd management including “age”, “blood pressure”, “specific gravity”, “albumin”, “sugar”, “red blood cell condition”, “white blood cell condition”, “pus cell condition”, etc. which all are available in the trained dataset.

```python
# Risk assessment based on Random Forest feature importances and domain knowledge
def risk_assessment(patient_data):
    # Scale the patient data
    patient_data_scaled = scaler.transform(patient_data)

    # Predict K-means cluster
    patient_cluster = kmeans.predict(patient_data_scaled)

    # Append the cluster label as a feature
    patient_data_with_cluster = np.hstack((patient_data_scaled, patient_cluster.reshape(-1, 1)))

    # Predict CKD probability and class using Random Forest
    ckd_prob = rf_classifier.predict_proba(patient_data_with_cluster)[:, 1]
    ckd_prediction = rf_classifier.predict(patient_data_with_cluster)

    # Define a risk score based on probability
    risk_score = ckd_prob * 100

    # Convert numeric prediction to label
    prediction_label = "CKD" if ckd_prediction[0] == 1 else "Not CKD"

    return risk_score[0], prediction_label
```

Figure 35: Function for risk assessment using random forest model

The risk assessment function is created to evaluate the risk score and the prediction based on the collected patient data from user input.

```python
# Get patient data from user input
patient_data = get_patient_data()

# Perform risk assessment
risk_score, prediction_label = risk_assessment(patient_data)

# Display risk assessment results
print(f"\nRisk Assessment Result:")
print(f"Risk Score: {risk_score:.2f}%")
print(f"Prediction: {prediction_label}")
```

Figure 36: Code for showing the risk score and prediction based on the user input data

The patient data is the user inputted data and the risk assessment output are the risk score and the prediction.

```
Please enter the patient's data:
Age: 48
Blood Pressure (bp) in mm/Hg: 80
Specific Gravity (sg) (1.005-1.025): 1.02
Albumin (al) (0-5): 1
Sugar (su) (0-5): 0
Red Blood Cells (rbc) (0 for normal, 1 for abnormal): 1
Pus Cell (pc) (0 for normal, 1 for abnormal): 0
Pus Cell clumps (pcc) (0 for notpresent, 1 for present): 0
Bacteria (ba) (0 for notpresent, 1 for present): 1
Blood Glucose Random (bgr) in mgs/dl: 121
Blood Urea (bu) in mgs/dl: 36
Serum Creatinine (sc) in mgs/dl: 1.2
Sodium (sod) in mEq/L: 135
Potassium (pot) in mEq/L: 4.5
Hemoglobin (hemo) in gms: 15.4
Packed Cell Volume (pcv): 44
White Blood Cell Count (wc) in cells/cumm: 7800
Red Blood Cell Count (rc) in millions/cmm: 5.2
Hypertension (htn) (0 for no, 1 for yes): 1
Diabetes Mellitus (dm) (0 for no, 1 for yes): 1
Coronary Artery Disease (cad) (0 for no, 1 for yes): 0
Appetite (appet) (0 for good, 1 for poor): 1
Pedal Edema (pe) (0 for no, 1 for yes): 0
Anemia (ane) (0 for no, 1 for yes): 1

Risk Assessment Result:
Risk Score: 58.00%
Prediction: CKD
```

Figure 37: Testing of the risk assessment model using user input data

The above figure shows the risk assessment based on user input data.

4.3 Findings of the Analysis

Reviewing the findings related to the prediction and risk assessment of "Chronic Kidney Disease" from the presented comprehensive analysis, multiple important insights obtained from the employment of a range of data analysis and machine learning approaches. This section gives an account of the findings from data preprocessing, EDA, K-Means clustering, trained models and the risk analysis framework. The results include the determination of the most important predictors of CKD, comparison of the accuracy of various machine learning methods, and practical consequences of the outcomes. To give a summary of the findings based on the results of the analysis done using different methods, this section thus seeks to encompass an overall

review of the factors that affect the prediction of CKD as well as measures that could be employed to improve the management of CKD using advanced analytics.

4.3.1 Data Preprocessing and Exploratory Data Analysis Findings

The analysis of the missing data demonstrated that many features had a large number of missing values. This led to the discovery that it is imperative to use right imputation methods to ensure the effectiveness of the data under analysis as well as make efficient use of the available data. Due to the successful use of mode for categorical data and median for numerical data to impute the missing values, none of the features were omitted during the analysis and therefore all of them could have retained their useful features that would have been lost if the rows with missing values were deleted as often done.

Several factors that define interactions include strong predictors, weak predictors and features that have a dependent relationship. Hemoglobin, packed cell volume, pus cell, specific gravity and red blood cell count were significantly related to the CKD status. Based on this finding, there is tentative evidence that these features could be of paramount importance in the predictive models (Pugh *et al*., 2019). The correlations of some features with CKD status or their component values were generally low or moderately low; these are conditions of red blood cells and sodium level. This does not necessarily mean that these features are insignificant but they could possibly be less significant when modeling the predictive models. Biological and Chemical related features seem to be highly related to each other as are certain sets of pathological features. This is useful if there are issues of multicollinearity in the developed models to check potential problems and might be used for feature selection or feature engineering in the models.

4.3.2 K-Means Clustering Analysis Findings

The clustering of the given data set based on the K-Means methodology produced the following insights. In order to understand the data more and reveal the two kinds of CKD, K-Means cluster, which is used to divide data into two parts because of its symmetrically binary feature, was used this time with k=2. This renders the implication that there are indeed different patterns in the feature space associated with CKD and non-CKD cases. The Silhouette Score of the outcomes of the clustering was determined to be 0.53 for both train and test set, which means that cluster 1 is quite clearly separated from cluster 2. This score entails that the member patients within a cluster are well suited for the particular cluster but poorly suited for the next neighboring cluster (Barrett *et al*., 2019). Specifically, the value in the Davies-Bouldin Index was 1.15 and 1.43 for test and

train set respectively. These metrics provide evidence for the supervised learning models that include the clustering results as an extra feature because it seems to represent some structure of the given data set.

4.3.3 Supervised Learning Model Findings

The evaluation of the three supervised learning models of Random Forest, Decision Tree, and Logistic Regression also offered a detailed understanding of the models' ability to predict CKD and the significance of the features in this process.

Table 3: Performance metrics of the models

Model Name	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000
Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000
Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000

This shows that the three investigated models have a good ability of predicting CKD as presented below. The accuracy, precision, and recall values are high, meaning that the models have a high ability of classifying CKD as well as non-CKD cases.

The consistency in feature importance across models strengthens the conclusions and their implications to clinical practice. The acknowledgement of "hemoglobin", "pcv", "pus cell", "specific gravity" and "rbc count" as key risk factors is justified in the current nursing clinical practice in determining the diagnosis and prognosis of the CKD. Based on the high prediction accuracy of the models for differentiation between CKD and non-CKD, they could be useful components of CDS to support early diagnosis and risk assessment for CKD (Roumeliotis *et al*., 2020). However, it is necessary to mention that such aspects as hemoglobin and packed cell volume, for instance, are powerful indicators, but they do not relate exclusively to CKD patients and can point to various health issues. The outputs of these models should therefore be used as a supplement to clinical evaluation in the management of patients.

4.3.4 Risk Assessment Framework Findings

In this study, considering the results of the risk scores presented in the framework, it is possible to examine that they are distributed in binary order. This is indicated by the fact that the model helps in separating the population into low risk and high risk of developing CKD, which was the way data was grouped in the dataset. One more important factor of the risk assessment is the identification of proper values to identify high-risk and low-risk persons.

The features importance derived from the Random Forest model supported by the risk assessment framework indicated a contingency of CKD risk relating to a person. Significantly, the presence of abnormally low Hemoglobin and Packed Cell Volume raised the calculated risk exponentially compared to the other variable results. The factor most effective in raising the risk score was the high serum creatinine level, especially when used together with the other risk factors (Lv, and Zhang, 2019). In this case, very low values as well as very high values of prevalent specific gravity as the index of the specific gravity were recognised as dangerous, and the higher danger was observed if the specific gravity was much lower than the middle value. Among the predictors, although not the most influential, older age raised the risk score and hence the risk. It was found that Hypertension and Diabetes added to a risk score throughout the analysis period, but these factors were not as influential as the top lab predictors.

To the proposed method, the risk assessment framework provided useful information about interaction effects between the predictors. Specially, both low hemoglobin and high serum creatinine yielded extremely high risk scores; results that were even higher than the overall impact of these predictors on risk scores (Dawson *et al*., 2021). Hypertension and diabetes were found to have increased risk scores for stroke. The increase is significantly high in older people, hence implying that the risk factors accumulate with aging. When albumin was accompanied by abnormally low specific gravity values of urine, the risk score would be higher, which may imply more severe kidney disorders.

4.3.5 Comparison with the existing studies

The findings from this analysis is in great focus with the former investigations regarding CKD prediction and risk assessment. The high value attributed to the hemoglobin, serum creatinine, red blood cell count, and blood urea is consistent with numerous works that have underlined the significance of these parameters in indicators of kidney function and in the progression of CKD. The statistical relevance of albumin in urine as a predictor in the models supports numerous

studies on the role of proteinuria in CKD diagnosis and prognosis (Bakris *et al*., 2020). In the models, hypertension and diabetes are significant risk factors for CKD which test with large-scale epidemiological studies of CKD. It is identified that specific gravity of urine has a high importance in the prediction and this has a biological correlation with CKD though less commonly discussed in several models. This suggests a possible research possibility for future clinical research.

4.4 Summary

The review of the methods aimed at CKD prediction and risk assessment presented in the current paper provides certain pre guidelines that may be useful further in clinical practice. This study also proved the usefulness of the features and the Random Forest Classifier in diagnosing CKD with good accuracy. The satisfactory accuracy achieved across many models indicates that there are obvious differences in the data that differentiates CKD from non-CKD. Some of the findings include; the ability to predict the most important parameters such as the hemoglobin, packed cell volume and serum creatinine as confirmed clinically. The outlined risk assessment framework may be considered as the promising approach for the patients' further risk categorizing and as the base for early preventive management.

However, the studies are not without limitation; for instance, the possibility of dataset bias and the fact that data is collected at a single time point pose a limited future scope. Replication of these findings in various, more lengthy follow-up and investigating the task of predicting multiple stages of CKD simultaneously would help consolidate the practical use of these observations. This study adds to the literature proving the utility of advanced analytics in the management of CKD. Many of these tools have the ability to enhance the clinical decision making processes and should be seen as adjuncts to comprehensive clinical assessment and clinical reasoning.

Chapter 5: Evaluation, limitations and future work

5.1 Introduction

This research aimed at improving the management of CKD by undertaking the use of predictive analytics. Data analysis in the current study involved EDAs, machine learning methods, and risk models that were utilized to analyze a CKD dataset. The main conceptual goal was to create or refine the algorithms of CKD recognition at early stages and risk assessment that will lead to better patients' management and prognosis. Random Forest, Decision Tree, and Logistic Regression classifiers were trained and tested throughout this study. The performances of these models were determined on the criteria of the models' power to predict likelihood of CKD and risk factors. Further, the tools were enhanced to develop a new risk assessment framework that can seamlessly translate the model's output into clinical risk scores. From this study the current and future management of CKD will benefit in terms of early detection, risk assessment and individualized patient management. This final chapter will present the overall analysis of core results and relate them to the original research aims, offer recommendations for practice and research in clinical psychology and nephrology, and reflect on the possible implications of this work in the context of nephrology and predictive analytics.

5.2 Linking with Objectives

The research aimed at the achievement of the five major objectives. In this part, it will be expounded how each of the objectives was met as well as the level to which it was accomplished.

Objective 1: To design a framework for determining the relevant predictors from the dataset that is most informative towards the development and progression of "Chronic Kidney Disease".

Fulfilling this objective was made possible through Exploratory Data Analysis and Feature Importance Analysis from the machine learning models operated with emphasis on the Random Forest Classifier. The present study recognized several important factors that influenced the development of CKD. Regarding the association between the clinical variables and the primary outcome, Hemoglobin and Packed Cell Volume were identified to be statically significant, thus showing that they remained the most relevant predictor of CKD. Low serum levels of hemoglobin and packed cell volume were found to be strongly related with the development of

CKD which is a classic sign of anemia (Razzak *et al*., 2020). Anemia is both an outcome from and a predictor of progression of CKD. Serum creatinine was found to be directly associated with CKD; therefore, increased levels of serum creatinine were established as a risk factor. This is in confirmation with the conventional application of serum creatinine as one of the significant indicators of renal condition in the clinic.

Abnormal blood pressure, high levels of blood glucose, and blood urea were also considered significant risk factors; the screening criteria we used are indeed linked with CKD according to traditional kidney function. Specific Gravity is one of the factors that are rather informative and holds a considerable increase in CKD risk. From this finding, one can infer that the level and possible changes in urine concentration might not be given sufficient attention in the assessment of one's risk for developing CKD (Zhang, and Parikh, 2019). The detection of albumin in urine was pointed out as one of the decisive factors to CKD diagnosis meeting the clinical concept of proteinuria as the sign of damaged kidneys. Hypertension and Diabetes were also reliably associated with CKD risk though they were less strong compared to the other Laboratory parameters.

Objective 2: To design a model to estimate the risk of CKD occurrence and its progression according to numerous medical signs and characteristics.

This objective was attained by the creation and assessment of several machine learning models. All the models including random forest revealed high accuracy, precision, recall, and F1-score considering the occurrence of CKD from the given clinical and demographic characteristics. The probability derived from the Random Forest model for the patients with CKD has provided the basis for the assessment framework in terms of risk score. It is more advanced comparatively to the binary classification by making it easier to predict CKD risk levels on a scale (Fu *et al*., 2019). The predictive accuracy of the developed model would make it a rather useful model in Clinical Decision Support Systems considering various medical symptoms and laboratory findings for CKD diagnosis. It could assist in the early identification of CKD especially in situations where a collection of symptoms along with results of tests may not alert the clinicians of kidney disease.

Objective 3: To compare the demographic factors, blood levels, and other clinical indicators in relation to CKD risk and prognosis.

Concerning this objective, the feature importance analysis as well as the investigation of the risk assessment framework's behavior were conducted. The results also highlighted the effect of age by revealing that higher ages are related risks of developing CKD. The performance of hemoglobin, packed cell volume, serum creatinine levels had played an important role in the prediction of CKD. The study further sought to measure the degree by which different levels of each of these factors affected CKD risk. It was found that albumin and abnormal specific gravity within urine are the indicators that show the presence of raised risks for the development of CKD. It was established that hypertension and diabetes posed a higher risk towards the development of CKD with the risk being higher among the elderly (Chan *et al*., 2019). The study also found out some significant interaction effects of these factors. They demonstrate a multivariate etiology of CKD risk, and, therefore, the requirement of considering the factors simultaneously.

Objective 4: To evaluate a risk assessment based on artificial intelligence and machine learning features for identification of people with CKD as well as clarification of potential risks and side effects.

This objective of the study was addressed with the help of constructing the Risk Assessment Framework that relied on the Random Forest model. The developed framework factors the model's probabilistic result into a percentage risk estimate which serves as a measure of risk of CKD for a particular patient. The framework works to obtain a risk profile for every patient depending on his and her medical and socio-demographic information. Through transforming the probabilities of the specific models into the risk percentage scores, the framework therefore generates a measure of CKD risk, which is easier to understand (Milchakov *et al*., 2019). The framework creates the possibility to fine-tune the risk thresholds, which in turn make it possible to adapt the framework to the circumstances of clinical practice. It enables the evaluation of how the modification of certain features influences the risk score and the assumptions for interventions.

Objective 5: To predict the group of CKD affected patients to be of high risk or low risk so that they can be prioritized for proper health care.

This objective was achieved by the fact that the developed risk assessment framework included risk stratification. It is noted that by using the framework, patients are able to be grouped according to the risk categories as given by their risk scores. Even though the study concerned

only CKD and non-CKD groups as the base classification, the nature of presented risk scores is continuous, which provides a more comprehensive risk stratification. In the case of the risk score distribution, it indicated a rather clear group with regard to risk. The study showed an examination of different operational thresholds to determine the high-risk and low-risk patients' classification that will show the versatility of the framework in adopting the current demands of clinical requirements (ElSayed *et al*., 2023). The framework identified the aspects and various situations that help determine an individual's classification and gave insights on the risk management for the specific classification.

5.3 Limitations

5.3.1 Model Limitations and Considerations

It is necessary to mention the shortcomings and biases that are inevitable when working with the proposed type of data. This limitation stemmed from the fact that the analysis was done on a single data set thus may be exhaustive in its generalization of populations. The generalizability of the findings was limited to the demographic characteristics of the subjects and the various healthcare environments. The characteristics utilized in the models were determined by the data set (Lestari, 2020). It is conceivable that there are other clinical factors that influence the prediction of CKD not considered in this study. As for the class imbalance, it was not mentioned that the software had the capability to avoid or fix imbalanced classes in the created input data set. This might impact on the performance of the model and the actual interpretation of the specific indicators. However, attempts were made to interpret the models especially by examining the features that the models used to make decision, for example by applying feature importance analysis on the Random Forest model, it was observed that it could be difficult to completely understand the process at which the Random Forest model arrives at the decision to classify an instance as High Risk.

The development of CKD has not been addressed in terms of its progression over time relative to the dataset and the analysis. Perhaps, in a longitudinal study more information can be elicited concerning the disease processes or other factors contributing to its development in the long run (Tomašev *et al*., 2019). The models display acceptable accuracy with regard to the given set of data, yet clinical examination of the efficiency in real-world conditions would be required to

employ the models in practice. Such limitations put emphasis on the opportunities to expand the framework for the prediction of CKD and assessment of risk for the future research.

5.3.2 Limitations and Potential Biases

There are several issues and potential sources of bias that have to be pointed out concerning the results of the model fit and research findings. The data that has been used to conduct this study might not match all the CKD patient categories. This indicates that the demographic and clinical details of this patient group may limit the generalization of this study's results. The model avoids the multiple stages of CKD classification and doesn't distinguish between them properly. The future work can be aimed to employ the multiple-class classification to predict the stages of "Chronic Kidney Disease" (Chen *et al*., 2019). It would also be beneficial to apply these models on any other dataset to evaluate them to see if the models are generalizable to other datasets. In the feature importance analysis, it is assumed that the features which are available in the dataset are important. However, there might be other significant characteristics that relate to the development of CKD and were not accounted for in this study.

The findings that can be estimated from this extended analysis would be useful in developing the assessment of CKD risk as well as making further predictions. Thus, the high effectiveness of the machine learning models, especially the Random Forest Classifier, can indicate the prospects for using these methods to improve modern approaches to CKD treatment. To identify quantitatively the significant risk factors accords with clinical experience, and surprisingly reveals some indicators that may have been initially overlooked (Sumida *et al*., 2020). An unexplored tool presented in this study is the risk assessment framework to evaluate the CKD risk for individual patients, which will require more testing to incorporate into the clinic. This work directly contributes to the existence of such studies and creates multiple opportunities for future studies on the role of analytics in managing chronic diseases which are a significant area of global healthcare concern.

5.4 Future Work and Recommendations

Based on the conclusions of the study various future work and recommendations can be suggested for both clinical and future research.

5.4.1 Future Work

Future work in the improvement of Chronic Kidney Disease (CKD) management with the help of predictive analytics can be based on various things. Firstly, one can enhance the utilization of the more sophisticated kinds of machine learning, like the deep learning and ensemble learning, for better reputation of the CKD. Using models such as XGBoost or neural network might detect other patterns in the data which are more comprehensive, than in the linear model (Evans *et al*., 2022). Further, it is possible to extend the current dataset with longitudinal data to compare how the patient's metrics change over time and, create a more dynamic prediction model.

Another future work is to add more data of clinical and genetic characteristics. It could also make it even more accurate if one include other parameters such as genetics, medication history, and lifestyle. Real-time predictive analytics, using wearable devices could also be taken as another avenue that is yet to be exploited to provide constant intervention and early identification (Lousa *et al*., 2020). Also with the current approach static cluster analysis (K –Means) used, it would be useful to apply the dynamic clustering algorithms such as DBSCAN or Gaussian Mixture Models to facilitate more flexible boundaries of clusters.

5.4.2 Clinical Practice Recommendations

Integration of Predictive Models: The CKD diagnosis tool may be beneficial for the healthcare institutions to implement similar predictive models in their electronic health record. This could surprisingly offer clinicians real time risk evaluation to help in identification and treatment of CKD.

Comprehensive Risk Assessment: From the information derived from the study, it is clear that various factors have to be taken into consideration in order to establish the chances of developing the CKD disease (Tasnim *et al*., 2024). Clinicians should use not only serum creatinine as the indicators but also the levels of hemoglobin, specific gravity of urine, and the presence of comorbidities.

Regular Screening: Since the model targeted finding out who is likely to develop CKD based on basic clinical measurements the screening of such populations especially those with diabetes or hypertension using such predictive instruments may be desirable.

Personalized Patient Education: The risk assessment framework can be beneficial to enhancing patient awareness regarding the factors attributing to their specific CKD risk and might enhance patient's compliance to preventive measures.

Targeted Interventions: It is recommended that effective interventions concerning the modifiable risk factors, including hypertension and diabetes, should be initiated as soon as possible.

5.4.3 Future Research Recommendations

External Validation: The models provided the high performance; however, it is imperative to assess the validity of the models based on other available dataset related to Chronic Kidney Disease to check the generalization ability of the models.

Longitudinal Studies: Further studies should gather data with a longitudinal design in order to investigate the course of CKD and precisely the temporal progression of the various risk factors independently of those of disease onset.

Multi-class Prediction: The integration of the models to predict various stages of CKD, instead of the mere detection of this condition's existence or lack of it, could be more beneficial for risk stratification and disease progression.

Integration of Additional Data Types: Possibly, further examining how the genetic, lifestyle and environmental data that is used by the models is obtained might be used to enhance the performance of the models and give a detailed risk assessment (Tasnim *et al*., 2024).

Explainable AI Techniques: This research focused on feature importance; using more elaborate explainable AI methods may help elucidate the model's reasoning process further and improve the models' reliability in clinical contexts.

Intervention Studies: The guidelines suggest that randomized controlled trials should be made to determine the usability of these predictive models and the risk assessment framework to the outcome and prognosis of the patients.

Refinement of Risk Factors: If the significance of s.g. of urine and other undiscovered factors that influence CKD risk were investigated more deeply, strategies for risk prediction might be enhanced.

Patient-Reported Outcomes: There is a possibility that the integration of other patient reported outcomes in future models would offer a better and clearer approach to determine CKD risk and effect.

Ethical and Privacy Considerations: Since such models operate on sensitive health data, there is a need to conduct more studies as to how the patient's identity can be protected and how biases in these models can be eliminated.

By adopting these recommendations, the evidence generated in this study can be applied to enhance the existing clinical practices professed in the improvement of CKD's life threatening impacts and also the CKD research evolutionary process demands to enhance its clinical benefits for future mortalities at risks as well as mortalities with CKD.

5.4 Conclusion

This paper has been established that predictive analytics and machine learning have a critical role in CKD care. Analyzing the datasets and models, the major risk determinants and weights for the development of CKD were defined, and an accurate risk evaluation model was developed. Generally, a high level of accuracy of the developed predictive models, mainly the Random Forest Classifier, implies that these modeling strategies could enhance early diagnosis and risk assessment in CKD. The aims of the study have been met and the findings inform the risk factors that increase the possibility of CKD, build reliable models for CKD risk rating, and design a versatile risk evaluation framework. Such findings are relevant to clinical practice since they may enhance the abilities to be more timely and specific about CKD management.

The study is not without limitations and suggestions for future research as the findings of the study should be validated with other samples and methods and the study was cross-sectional. In light of the ongoing emergence of more patient- and prognosis-oriented health care systems, the methods and results described in this work can be referred to as a part of the cumulative knowledge about using analytics in chronic disease management. This study contributes to the advancement in the management of CKD that could potentially benefit a patient's prognosis by means of early identification of the risks, as well as more refined risk estimation, and subsequent intervention.

Reference List

Journals

[1] Agarwal, R., Sinha, A.D., Cramer, A.E., Balmes-Fenwick, M., Dickinson, J.H., Ouyang, F. and Tu, W., 2021. Chlorthalidone for hypertension in advanced chronic kidney disease. New England Journal of Medicine, 385(27), pp.2507-2519.

[2] Almansour, N.A., Syed, H.F., Khayat, N.R., Altheeb, R.K., Juri, R.E., Alhiyafi, J., Alrashed, S. and 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, pp.101-111.

[3] American Diabetes Association Professional Practice Committee and American Diabetes Association Professional Practice Committee:, 2022. 11. Chronic kidney disease and risk management: Standards of Medical Care in Diabetes—2022. Diabetes care, 45(Supplement_1), pp.S175-S184.

[4] Bakris, G.L., Agarwal, R., Anker, S.D., Pitt, B., Ruilope, L.M., Rossing, P., Kolkhof, P., Nowack, C., Schloemer, P., Joseph, A. and Filippatos, G., 2020. Effect of finerenone on chronic kidney disease outcomes in type 2 diabetes. New England journal of medicine, 383(23), pp.2219-2229.

[5] Bardhan, I., Chen, H. and Karahanna, E., 2020. Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. MIS Quarterly, 44(1).

[6] Barrett, M., Boyne, J., Brandts, J., Brunner-La Rocca, H.P., De Maesschalck, L., De Wit, K., Dixon, L., Eurlings, C., Fitzsimons, D., Golubnitschaja, O. and Hageman, A., 2019. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. Epma Journal, 10, pp.445-464.

[7] Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2), p.21.

[8] Bikbov, B., Purcell, C.A., Levey, A.S., Smith, M., Abdoli, A., Abebe, M., Adebayo, O.M., Afarideh, M., Agarwal, S.K., Agudelo-Botero, M. and Ahmadian, E., 2020. Global,

regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The lancet, 395(10225), pp.709-733.

[9] Chan, C.T., Blankestijn, P.J., Dember, L.M., Gallieni, M., Harris, D.C., Lok, C.E., Mehrotra, R., Stevens, P.E., Wang, A.Y.M., Cheung, M. and Wheeler, D.C., 2019. Dialysis initiation, modality choice, access, and prescription: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney international, 96(1), pp.37-47.

[10] Chen, T., Li, X., Li, Y., Xia, E., Qin, Y., Liang, S., Xu, F., Liang, D., Zeng, C. and Liu, Z., 2019. Prediction and risk stratification of kidney outcomes in IgA nephropathy. American journal of kidney diseases, 74(3), pp.300-309.

[11] Chen, T.K., Knicely, D.H. and Grams, M.E., 2019. Chronic kidney disease diagnosis and management: a review. Jama, 322(13), pp.1294-1304.

[12] Chen, Y., Lee, K., Ni, Z. and He, J.C., 2020. Diabetic kidney disease: challenges, advances, and opportunities. Kidney diseases, 6(4), pp.215-225.

[13] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasiński, M., Jasiński, Ł., Gono, R., Jasińska, E. and Bolshev, V., 2021. Prediction of chronic kidney disease-a machine learning perspective. IEEE access, 9, pp.17312-17334.

[14] Dawson, J., Lambert, K., Campbell, K.L. and Kelly, J.T., 2021, July. Incorporating digital platforms into nutritional care in chronic kidney disease. In Seminars in Dialysis.

[15] de Boer, I.H., Caramori, M.L., Chan, J.C., Heerspink, H.J., Hurst, C., Khunti, K., Liew, A., Michos, E.D., Navaneethan, S.D., Olowu, W.A. and Sadusky, T., 2020. KDIGO 2020 clinical practice guideline for diabetes management in chronic kidney disease. Kidney international, 98(4), pp.S1-S115.

[16] de Boer, I.H., Khunti, K., Sadusky, T., Tuttle, K.R., Neumiller, J.J., Rhee, C.M., Rosas, S.E., Rossing, P. and Bakris, G., 2022. Diabetes management in chronic kidney disease: a consensus report by the American Diabetes Association (ADA) and Kidney Disease: Improving Global Outcomes (KDIGO). Diabetes care, 45(12), pp.3075-3090.

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

[18] ElSayed, N.A., Aleppo, G., Aroda, V.R., Bannuru, R.R., Brown, F.M., Bruemmer, D., Collins, B.S., Hilliard, M.E., Isaacs, D., Johnson, E.L. and Kahan, S., 2023. 11. Chronic

kidney disease and risk management: standards of care in diabetes—2023. Diabetes Care, 46(Supplement_1), pp.S191-S202.

[19] Flythe, J.E., Chang, T.I., Gallagher, M.P., Lindley, E., Madero, M., Sarafidis, P.A., Unruh, M.L., Wang, A.Y.M., Weiner, D.E., Cheung, M. and Jadoul, M., 2020. Blood pressure and volume management in dialysis: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney international, 97(5), pp.861-876.

[20] Fu, H., Liu, S., Bastacky, S.I., Wang, X., Tian, X.J. and Zhou, D., 2019. Diabetic kidney diseases revisited: A new perspective for a new era. Molecular metabolism, 30, pp.250-263.

[21] Kalantar-Zadeh, K., Jafar, T.H., Nitsch, D., Neuen, B.L. and Perkovic, V., 2021. Chronic kidney disease. The lancet, 398(10302), pp.786-802.

[22] Khan, B., Naseem, R., Muhammad, F., Abbas, G. and Kim, S., 2020. An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. IEEE Access, 8, pp.55012-55022.

[23] Khan, M.T.I., Prottasha, M.S.I., Nasim, T.A., Mehedi, A.A., Pranto, M.A.M. and Asad, N.A., 2020. Performance analysis of various machine learning classifiers on reduced chronic kidney disease dataset. International Journal of Recent Research and Review (IJRRR), 13(4), pp.16-20.

[24] Lameire, N.H., Levin, A., Kellum, J.A., Cheung, M., Jadoul, M., Winkelmayer, W.C., Stevens, P.E., Caskey, F.J., Farmer, C.K., Fuentes, A.F. and Fukagawa, M., 2021. Harmonizing acute and chronic kidney disease definition and classification: report of a Kidney Disease: Improving Global Outcomes (KDIGO) Consensus Conference. Kidney international, 100(3), pp.516-526.

[25] Lestari, A., 2020. Increasing accuracy of C4. 5 algorithm using information gain ratio and adaboost for classification of chronic kidney disease. Journal of Soft Computing Exploration, 1(1), pp.32-38.

[26] Liu, P., Quinn, R.R., Lam, N.N., Al-Wahsh, H., Sood, M.M., Tangri, N., Tonelli, M. and Ravani, P., 2021. Progression and regression of chronic kidney disease by age among adults in a population-based cohort in Alberta, Canada. JAMA network open, 4(6), pp.e2112828-e2112828.

[27] Liyanage, T., Toyama, T., Hockham, C., Ninomiya, T., Perkovic, V., Woodward, M., Fukagawa, M., Matsushita, K., Praditpornsilpa, K., Hooi, L.S. and Iseki, K., 2022. Prevalence of chronic kidney disease in Asia: a systematic review and analysis. BMJ global health, 7(1), p.e007525.

[28] Lv, J.C. and Zhang, L.X., 2019. Prevalence and disease burden of chronic kidney disease. Renal fibrosis: mechanisms and therapies, pp.3-15.

[29] Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., Kudo, M., Haida, K., Kuroda, J., Yanagiya, R. and Saitoh, E., 2019. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Scientific reports, 9(1), p.11862.

[30] Milchakov, K.S., Shilov, E.M., Shvetzov, M.Y., Fomin, V.V., Khalfin, R.A., Madyanova, V.V., Pivina, L.M. and Semenova, Y.M., 2019. Management of chronic kidney disease in the Russian Federation: A critical review of prevalence and preventive programmes. International Journal of Healthcare Management.

[31] Pugh, D., Gallacher, P.J. and Dhaun, N., 2019. Management of hypertension in chronic kidney disease. Drugs, 79(4), pp.365-379.

[32] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C. and Chen, B., 2019. A machine learning methodology for diagnosing chronic kidney disease. IEEE access, 8, pp.20991-21002.

[33] Rady, E.H.A. and Anwar, A.S., 2019. Prediction of kidney disease stages using data mining algorithms. Informatics in medicine unlocked, 15, p.100178.

[34] Razzak, M.I., Imran, M. and Xu, G., 2020. Big data analytics for preventive medicine. Neural Computing and Applications, 32(9), pp.4417-4451.

[35] Roumeliotis, S., Mallamaci, F. and Zoccali, C., 2020. Endothelial dysfunction in chronic kidney disease, from biology to clinical outcomes: a 2020 update. Journal of Clinical Medicine, 9(8), p.2359.

[36] Ruiz-Ortega, M., Rayego-Mateos, S., Lamas, S., Ortiz, A. and Rodrigues-Diez, R.R., 2020. Targeting the progression of chronic kidney disease. Nature Reviews Nephrology, 16(5), pp.269-288.

[37] Sabanayagam, C., Xu, D., Ting, D.S., Nusinovici, S., Banu, R., Hamzah, H., Lim, C., Tham, Y.C., Cheung, C.Y., Tai, E.S. and Wang, Y.X., 2020. A deep learning algorithm to

detect chronic kidney disease from retinal photographs in community-based populations. The Lancet Digital Health, 2(6), pp.e295-e302.

[38] Saleh, E. and Bin Abd Kadir, M.F., 2022. Prediction of chronic kidney disease using data mining techniques. Prediction of Chronic Kidney Disease Using Data Mining Techniques.

[39] Sawhney, R., Malik, A., Sharma, S. and Narayan, V., 2023. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decision Analytics Journal, 6, p.100169.

[40] Senan, E.M., Al-Adhaileh, M.H., Alsaade, F.W., Aldhyani, T.H., Alqarni, A.A., Alsharif, N., Uddin, M.I., Alahmadi, A.H., Jadhav, M.E. and Alzahrani, M.Y., 2021. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of healthcare engineering, 2021(1), p.1004767.

[41] Shlipak, M.G., Tummalapalli, S.L., Boulware, L.E., Grams, M.E., Ix, J.H., Jha, V., Kengne, A.P., Madero, M., Mihaylova, B., Tangri, N. and Cheung, M., 2021. The case for early identification and intervention of chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney international, 99(1), pp.34-47.

[42] Srikanth, V., 2023. CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS.

[43] Subramanian, M., Wojtusciszyn, A., Favre, L., Boughorbel, S., Shan, J., Letaief, K.B., Pitteloud, N. and Chouchane, L., 2020. Precision medicine in the era of artificial intelligence: implications in chronic disease management. Journal of translational medicine, 18, pp.1-12.

[44] Sumida, K., Nadkarni, G.N., Grams, M.E., Sang, Y., Ballew, S.H., Coresh, J., Matsushita, K., Surapaneni, A., Brunskill, N., Chadban, S.J. and Chang, A.R., 2020. Conversion of urine protein–creatinine ratio or urine dipstick protein to urine albumin–creatinine ratio for use in chronic kidney disease screening and prognosis: an individual participant–based meta-analysis. Annals of internal medicine, 173(6), pp.426-435.

[45] Tasnim, T., Sugireng, S., Imran, I. and Akib, N.I., 2024. Analysis of differences in early detection of chronic kidney disease with urine creatinine, proteins and individual health status based on behavioural, stress and genetic factors in Kendari City, Indonesia. Public Health of Indonesia, 10(2), pp.203-213.

[46] Tomašev, N., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I. and Connell, A., 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), pp.116-119.

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

[48] Zhang, K., Liu, X., Xu, J., Yuan, J., Cai, W., Chen, T., Wang, K., Gao, Y., Nie, S., Xu, X. and Qin, X., 2021. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. Nature biomedical engineering, 5(6), pp.533-545.

[49] Zhang, W.R. and Parikh, C.R., 2019. Biomarkers of acute and chronic kidney disease. Annual review of physiology, 81(1), pp.309-333.

Data Source

[50] Kaggle.com, 2017, Chronic Kidney Disease dataset. Available at: https://www.kaggle.com/datasets/mansoordaku/ckdisease [Accessed on: 15.07.2024]

Appendix A. Implementation Code

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix, roc_curve
from sklearn.cluster import KMeans

# Loading dataset
df = pd.read_csv("kidney_disease.csv")

df.head()

df.info()

df.describe()

df.isnull().sum()

df.dropna(inplace=True)

# Data Preprocessing
# Drop 'id' column
```

```
df.drop('id', axis=1, inplace=True)

# Handling missing values by replacing them with the median or mode
for column in df.columns:
    if df[column].dtype == 'object':
        df[column].fillna(df[column].mode()[0], inplace=True)
    else:
        df[column].fillna(df[column].median(), inplace=True)

# Convert non-numeric data to numeric using Label Encoding
label_encoder = LabelEncoder()
categorical_columns = ['rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane', 'classification']
for col in categorical_columns:
    df[col] = label_encoder.fit_transform(df[col])

# Splitting the dataset into features and target variable
X = df.drop(columns=['classification'])
y = df['classification']

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Exploratory Data Analysis (EDA)
# Correlation matrix
corr_matrix = df.corr()
plt.figure(figsize=(15,10))
```

```
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.show()

# K-Means Clustering
kmeans = KMeans(n_clusters=2, random_state=42)
kmeans.fit(X_train_scaled)
train_clusters = kmeans.labels_
test_clusters = kmeans.predict(X_test_scaled)

# Add cluster labels as features
X_train_scaled_with_clusters = np.hstack((X_train_scaled, train_clusters.reshape(-1, 1)))
X_test_scaled_with_clusters = np.hstack((X_test_scaled, test_clusters.reshape(-1, 1)))

# Model Development
# Random Forest Classifier
rf_classifier = RandomForestClassifier(random_state=42)
rf_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_rf = rf_classifier.predict(X_test_scaled_with_clusters)

# Decision Tree Classifier
dt_classifier = DecisionTreeClassifier(random_state=42)
dt_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_dt = dt_classifier.predict(X_test_scaled_with_clusters)

# Logistic Regression
lr_classifier = LogisticRegression(random_state=42)
lr_classifier.fit(X_train_scaled_with_clusters, y_train)
y_pred_lr = lr_classifier.predict(X_test_scaled_with_clusters)

# Function to evaluate the model
def evaluate_model(y_test, y_pred, model_name):
```

```
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, y_pred)

    print(f"Model: {model_name}")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1 Score: {f1:.4f}")
    print(f"AUC-ROC: {roc_auc:.4f}")
    print("\n")

    # Confusion Matrix
    cm = confusion_matrix(y_test, y_pred)
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
    plt.title(f'Confusion Matrix - {model_name}')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()

evaluate_model(y_test, y_pred_rf, "Random Forest")

evaluate_model(y_test, y_pred_dt, "Decision Tree")

evaluate_model(y_test, y_pred_lr, "Logistic Regression")

from sklearn.metrics import silhouette_score, davies_bouldin_score

# Evaluate K-Means clustering
```

```
def evaluate_kmeans(X, clusters):
    silhouette_avg = silhouette_score(X, clusters)
    print(f"Silhouette Score: {silhouette_avg:.4f}")

    davies_bouldin_avg = davies_bouldin_score(X, clusters)
    print(f"Davies-Bouldin Score: {davies_bouldin_avg:.4f}")

evaluate_kmeans(X_train_scaled, train_clusters)
evaluate_kmeans(X_test_scaled, test_clusters)

# Function to plot K-Means clusters
def plot_clusters(X, clusters, title):
    plt.figure(figsize=(10, 6))
    plt.scatter(X[:, 0], X[:, 1], c=clusters, cmap='viridis', s=50)
    plt.title(title)
    plt.xlabel('Feature 1')
    plt.ylabel('Feature 2')
    plt.colorbar(label='Cluster Label')
    plt.show()

# Plot clusters for training data
plot_clusters(X_train_scaled[:, :2], train_clusters, 'K-Means Clusters (Training Data)')

# Plot clusters for test data
plot_clusters(X_test_scaled[:, :2], test_clusters, 'K-Means Clusters (Test Data)')

# Define the column names as in the original dataset
column_names = [
     "age", "bp", "sg", "al", "su", "rbc", "pc", "pcc", "ba", "bgr", "bu", "sc", "sod", "pot", "hemo",
"pcv", "wc", "rc",
    "htn", "dm", "cad", "appet", "pe", "ane"
```

```
]

# Function to get patient data from user input
def get_patient_data():
    print("Please enter the patient's data:")

    data = []
    data.append(int(input("Age: ")))
    data.append(int(input("Blood Pressure (bp) in mm/Hg: ")))
    data.append(float(input("Specific Gravity (sg) (1.005-1.025): ")))
    data.append(int(input("Albumin (al) (0-5): ")))
    data.append(int(input("Sugar (su) (0-5): ")))
    data.append(int(input("Red Blood Cells (rbc) (0 for normal, 1 for abnormal): ")))
    data.append(int(input("Pus Cell (pc) (0 for normal, 1 for abnormal): ")))
    data.append(int(input("Pus Cell clumps (pcc) (0 for notpresent, 1 for present): ")))
    data.append(int(input("Bacteria (ba) (0 for notpresent, 1 for present): ")))
    data.append(float(input("Blood Glucose Random (bgr) in mgs/dl: ")))
    data.append(float(input("Blood Urea (bu) in mgs/dl: ")))
    data.append(float(input("Serum Creatinine (sc) in mgs/dl: ")))
    data.append(float(input("Sodium (sod) in mEq/L: ")))
    data.append(float(input("Potassium (pot) in mEq/L: ")))
    data.append(float(input("Hemoglobin (hemo) in gms: ")))
    data.append(int(input("Packed Cell Volume (pcv): ")))
    data.append(int(input("White Blood Cell Count (wc) in cells/cumm: ")))
    data.append(float(input("Red Blood Cell Count (rc) in millions/cmm: ")))
    data.append(int(input("Hypertension (htn) (0 for no, 1 for yes): ")))
    data.append(int(input("Diabetes Mellitus (dm) (0 for no, 1 for yes): ")))
    data.append(int(input("Coronary Artery Disease (cad) (0 for no, 1 for yes): ")))
    data.append(int(input("Appetite (appet) (0 for good, 1 for poor): ")))
    data.append(int(input("Pedal Edema (pe) (0 for no, 1 for yes): ")))
    data.append(int(input("Anemia (ane) (0 for no, 1 for yes): ")))
```

```
    return pd.DataFrame([data], columns=column_names)

# Risk assessment based on Random Forest feature importances and domain knowledge
def risk_assessment(patient_data):
    # Scale the patient data
    patient_data_scaled = scaler.transform(patient_data)

    # Predict K-means cluster
    patient_cluster = kmeans.predict(patient_data_scaled)

    # Append the cluster label as a feature
    patient_data_with_cluster = np.hstack((patient_data_scaled, patient_cluster.reshape(-1, 1)))

    # Predict CKD probability and class using Random Forest
    ckd_prob = rf_classifier.predict_proba(patient_data_with_cluster)[:, 1]
    ckd_prediction = rf_classifier.predict(patient_data_with_cluster)

    # Define a risk score based on probability
    risk_score = ckd_prob * 100

    # Convert numeric prediction to label
    prediction_label = "CKD" if ckd_prediction[0] == 1 else "Not CKD"

    return risk_score[0], prediction_label

# Get patient data from user input
patient_data = get_patient_data()

# Perform risk assessment
risk_score, prediction_label = risk_assessment(patient_data)
```

```
# Display risk assessment results
print(f"\nRisk Assessment Result:")
print(f"Risk Score: {risk_score:.2f}%")
print(f"Prediction: {prediction_label}")
```