

CHRONIC KIDNEY DISEASE PEDICTION USING MACHINE LEARNING

A Project Report

Submitted by

TRISHALA PANDI S 312321104182

USHARANI T 312321104184

in partial fulfilment for the award of the Degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



St. JOSEPH'S COLLEGE OF ENGINEERING

(An Autonomous Institution)

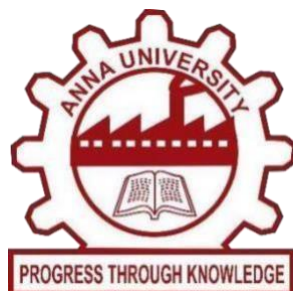
St. Joseph's Group of Institution

Jeppiaar Educational Trust

OMR, Chennai 600

ANNA UNIVERSITY: CHENNAI

SEPTEMBER 2023



BONAFIDE CERTIFICATE

Certified that this project report “**CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING**” is the bonafide work of **TRISHALA PANDI S (312321104182) AND USHARANI T (312321104184)** who carried out the work under my guidance. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

HEAD OF THE DEPARTMENT

Dr. G.Mariakalavathy, M.E., Ph.D.,

Professor & Head of Department
Dept. of Computer Science and Engineering,
St. Joseph's college of Engineering,
OMR, Chennai 600 119

SIGNATURE

SUPERVISOR

Dr. G. Brindha, M.E., Ph.D.,
Dr. E. Ahila Devi, M.E., Ph.D.,
Associate Professor,
Dept. of Computer Science and Engineering,
St. Joseph's college of Engineering,
OMR, Chennai 600 119

ACKNOWLEDGEMENT

At the outset, we would like to express our sincere gratitude to our beloved

Dr. B. Babu Manoharan M.A., M.B.A., Ph.D., *Chairman, St. Joseph's Group of Institutions* for his constant guidance and support to the student community and the Society.

I would like to express my hearty thanks to our respected ***Managing Director Mrs. S. Jessie Priya M.Com.*** for her kind encouragement and blessings.

I wish to express my sincere thanks to the ***Executive Director Mr. B. Shashi Sekar, M.Sc.*** for providing ample facilities in the institution. I express sincere gratitude to our beloved ***Principal Dr. Vaddi Seshagiri Rao M.E., M.B.A., Ph.D., F.I.E.*** for his inspirational ideas during the course of the project.

I express my sincere gratitude to our beloved ***Dean (Student Affairs) Dr. V. Vallinayagam M.Sc., M.Phil., Ph.D., and Dean (Academics) Dr. G. Sreekumar M.Sc., M.Tech., Ph.D.,*** for their inspirational ideas during the course of the project.

I wish to express our sincere thanks to ***Dr. G. Mariyakalavathy M.E., Ph.D., Head of the Department and Dean (Research),*** Department of Computer Science and Engineering, St. Joseph's College of Engineering for his guidance and assistance in solving the various intricacies involved in the project.

I would like to acknowledge my profound gratitude to our ***Supervisor Dr. G. Brindha M.E., Ph.D., E. Ahila Devi M.E., Ph.D.,*** for her expert guidance and suggestion to carry out the study successfully.

Finally, I thank the **Faculty Members** and **my Family**, who helped and encouraged me constantly to complete the project successfully.

ABSTRACT

Kidney is one of the vital organs in a human body while ironically, chronic kidney disease (CKD) is one of the main causes of death in the world. Due to the low rate of loss of kidney function, the disease is often overlooked until it is in a really bad condition. Dysfunctional kidney may lead to accumulation of wastes in blood which would affect several other systems and functions of the body such as blood pressure, red blood cell production, vitamin D and bone health. Machine learning algorithms can help in predicting the patients who have CKD or not. Even though several studies have been made to classify CKD on patients using machine learning tool, not many researchers perform preprocessing and feature selection technique to obtain quality and dependable result. Machine learning used with feature selection techniques are shown to have better and more dependable result. In this study, feature selection methods such as Random Forest feature selection, forward selection, forward exhaustive selection, backward selection and backward exhaustive selection were identified and evaluated. Then, machine learning classifiers such as Random Forest, Linear and Radial SVM, Naïve Bayes and Logistic Regression were implemented. Lastly, the performance of each machine learning model was evaluated in terms of accuracy, sensitivity, specificity and AUC score. The results showed that Random Forest classifier with Random Forest feature selection is the most suitable machine learning model for Prediction of CKD as it has the highest accuracy, sensitivity, specificity and AUC with 98.825%, 98.04%, 100% and 98.9% respectively which outperformed other classifier.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO
1.	INTRODUCTION	7
	1.1 Project Objective	8
	1.2 Project Overview	8
2.	LITERATURE SURVEY	9
3.	SYSTEM ANALYSIS	
	3.1 Existing System	12
	3.2 Proposed System	12
	3.3 Language Specification	13
	3.4 Machine Learning	14
4.	SYSTEM DESIGN	
	4.1 System Architecture	16
5.	DESIGN AND IMPLEMENTATION	17
	5.1 Dataset and Features	20
	5.2 Statistical Description	22

6.	MODULE DESCRIPTION	
	6.1 Logistic Regression Algorithm	23
	6.2 Random Forest Classifier	25
7.	DATA VISUALIZATIONS	27
8.	RESULT	29
9.	LIMITATIONS	32
10.	CONCLUSION	33
11.	REFERENCE	33

CHAPTER 1

INTRODUCTION

Chronic Kidney Disease (CKD) is one of the global medical issue because of extremely high expense for the treatment and also the high death rates. World Health Organization (WHO) announced South East Asia and America witnessed highest rate of population with this illness, from a survey in 2012. Besides, the number of new patients increase yearly, while there are restrictions general medical coverage, for example, no cost or low cost remedy, absence of the vital medicinal gear and therapeutic repayment limit. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, weekboans, poor nutrition health and nerve damage, Decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of your blood creatinine, age, race, gender, and other factors. The earlier disease is detected the better chance of showing or stopping its progression. Based upon GFR the renal damage severity by CKD is categorized into following five stages:

1.1 PROJECT OBJECTIVE

The primary objectives of this chronic kidney disease (CKD) prediction project in machine learning are to collect and preprocess relevant patient data, select informative features, develop and fine tune predictive models, assess their performance, and ensure ethical and clinical applicability. Interpretability and generalizability of models will be emphasized, with a focus on protecting patient privacy and complying with ethical guidelines. The project aims to create a comprehensive documentation package for healthcare professionals, communicate findings effectively, and identify future research directions. Risk assessment and resource allocation will be crucial for successful project completion and evaluation.

1.2 PROJECT OVERVIEW

This systematic review aims to provide a comprehensive, in-depth summary and evaluation of ML-based diagnostic and prognostic tools for CKD development and progression, which will help to better direct future research strategy and methodology in developing ML algorithms in CKD care. This project focuses on developing machine learning models to predict chronic kidney disease (CKD) in its early stages, enabling timely intervention and improved patient outcomes.

It involves collecting and preprocessing patient data, selecting relevant features, training and optimizing predictive models, ensuring ethical data handling, and assessing clinical applicability. By creating accurate and interpretable models, this project seeks to assist healthcare professionals in identifying individuals at risk of CKD, ultimately enhancing healthcare delivery and patient wellbeing.

CHAPTER 2

LITERATURE SURVEY

A literature survey on the topic of chronic kidney disease (CKD) prediction using machine learning can provide valuable insights into the current state of research in this field. Below, I've compiled a list of key research papers and articles that cover various aspects of CKD prediction using machine learning techniques. Please note that the list is not exhaustive, and you may want to explore additional sources for a comprehensive understanding of the topic:

1. Title: "Chronic kidney disease prediction on imbalanced data using deep learning and principal component analysis"

- Authors: V. Prakash, A. Natarajan, and R. Sridharan
- Published: Computer Methods and Programs in Biomedicine, 2021
- Summary: This paper discusses the use of deep learning and principal component analysis to predict CKD, addressing the problem of imbalanced datasets.

2. Title: "Chronic kidney disease prediction using machine learning techniques: A systematic review"

- Authors: N. Bhattacharjee, A. Chaki, and S. Acharjee
- Published: Journal of King Saud University - Computer and Information Sciences, 2021
- Summary: This systematic review provides an overview of machine learning techniques used for CKD prediction, highlighting the strengths and weaknesses of various approaches.

3. Title: "Predicting Chronic Kidney Disease Using Data Mining Methods: A Systematic Review"

- Authors: H. Rajabi, M. Mohseni, and A. R. Khotanlou
- Published: Journal of Medical Systems, 2018
- Summary: This systematic review covers data mining methods for CKD prediction, including machine learning algorithms, and discusses their performance and limitations.

4. Title: "Early detection of chronic kidney disease using ensemble methods"

- Authors: J. Alqudah, N. Iqbal, and A. H. Mir
- Published: Computers in Biology and Medicine, 2019
- Summary: The paper explores the use of ensemble methods for early detection of CKD, emphasizing the importance of feature selection and model selection.

5. Title: "Chronic Kidney Disease Prediction Using Random Forest"

- Authors: K. P. Soman and R. J. Hegde
- Published: Procedia Computer Science, 2015
- Summary: This paper discusses the application of the Random Forest algorithm for predicting CKD and evaluates its performance in comparison to other machine learning methods.

6. Title: "Predicting Chronic Kidney Disease Using Data Mining Techniques: A Systematic Review of Recent Studies"

- Authors: I. A. Hossain, M. S. Alam, and S. M. Lutful Kabir
- Published: Informatics in Medicine Unlocked, 2019
- Summary: This systematic review provides an overview of recent studies that use data mining techniques for CKD prediction and summarizes their findings.

7.Title: "Chronic Kidney Disease Prediction with Artificial Neural Networks"

- Authors: A. S. Malik and M. S. Ali
- Published: Procedia Computer Science, 2015
- Summary: The paper explores the use of artificial neural networks (ANNs) for CKD prediction and discusses the architecture and performance of ANNs in context.

8.Title: "Machine Learning Approaches for Predicting Progression to End-Stage Renal Disease in Patients with Chronic Kidney Disease: A Review"

- Authors: T. Alaa, and M. B. Tahsin
- Published: BMJ Health & Care Informatics, 2021
- Summary: This review focuses on machine learning models for predicting the progression of CKD to end-stage renal disease, emphasizing the clinical implications of such predictions.

These papers and articles should provide you with a solid foundation for understanding the current research landscape in the field of CKD prediction using machine learning. You can use them as references to delve deeper into specific techniques, datasets, and evaluation metrics commonly employed in CKD prediction research

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

In conducting a system analysis for a chronic kidney disease (CKD) prediction project, it's crucial to assess existing data sources, workflows, and models. This involves defining project objectives, evaluating data quality and integration, reviewing model performance, and examining how CKD predictions are integrated into clinical practice. Additionally, focus on user-friendliness, data security, scalability, and regulatory compliance. Regular feedback and cost-benefit analysis should guide ongoing improvements, ensuring the system effectively aids healthcare providers in early CKD detection and patient care.

3.2 PROPOSED SYSTEM

Proposed CKD Prediction System:

1. Objective: Early CKD detection and risk assessment for timely intervention.
2. Components: Data collection, preprocessing, machine learning models, feature selection, model validation, clinical integration, user-friendly interface, data security, scalability, feedback mechanisms.
3. Data Source: Electronic health records (EHRs) with patient demographics, medical history, labs, and medications.
4. Machine Learning: Algorithms like logistic regression, decision trees, or neural networks.
5. Benefits: Improved patient care, cost savings, enhanced decision-making, patient engagement.

6. Compliance: Adherence to data privacy regulations (e.g., HIPAA).
7. Scalability: Designed for handling increased data and users.
8. Feedback: Continuous feedback loops for system refinement and optimization.
9. User Interface: Web-based interface for healthcare providers to input data and receive risk scores.
10. Security: Robust data security measures for patient data protection.

3.3 LANGUAGE SPECIFICATION

Python is an interpreted, object oriented, high level programming language with dynamic semantics. Its high level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.

The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Since there is no compilation step, the edit test debug cycle is incredibly fast.

Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace.

The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit test debug cycle makes this simple approach very effective.

Python uses indentation (whitespace) to define code blocks, which is different from many other programming languages that use braces or other symbols. This indentation is typically four spaces, but it can be configured differently. Python uses colons (":") to denote the start of code blocks. Python supports various data types, including integers, floating-point numbers, strings, lists, tuples, dictionaries, sets, and more. It also supports user-defined classes and objects.

3.4 MACHINE LEARNING

In a project focused on chronic kidney disease (CKD) prediction, various types of machine learning techniques can be employed to build predictive models. Here are some of the common types of machine learning used in such projects:

1. Supervised Learning:

- ❖ Logistic Regression: A straightforward algorithm for binary classification, suitable for predicting CKD presence or absence.
- ❖ Random Forest: Useful for capturing complex relationships between patient attributes and CKD risk.

2. Random Forest:

- ❖ In the chronic kidney disease (CKD) prediction project, Random Forest, a versatile machine learning technique, is employed to develop accurate models. Random Forest creates an ensemble of decision trees by utilizing random subsets of the dataset and features, which collectively make predictions about CKD risk based on patient information. It excels at handling complex relationships in data, reducing overfitting, and identifying influential features. The final prediction is determined by the majority vote of individual trees. Random Forest's robustness and predictive power make it a valuable tool for improving CKD diagnosis accuracy in the project.

3. Decision Tree:

- ❖ The accuracy of decision trees in predicting chronic kidney disease (CKD) can vary depending on several factors, including the quality of the data, the choice of features, and the specific decision tree algorithm used. Decision trees are a popular machine learning algorithm for classification tasks like CKD prediction because they are easy to interpret and can handle both categorical and numerical data. However, their performance can be influenced by the complexity of the problem and the size of the dataset.

4. Model Training:

- ❖ To create an effective CKD prediction model, logistic regression "learns" from historical patient data. During training, it optimizes the values of the coefficients and an intercept term to minimize the difference between predicted probabilities and actual CKD labels in the training dataset. This is typically done using optimization techniques like gradient descent.

5. Making Predictions:

- ❖ Once the model is trained, it can predict the probability of CKD for new, unseen patient data. If the predicted probability exceeds a chosen threshold (usually 0.5), the model classifies the patient as having CKD; otherwise, it classifies them as not having CKD.

6. Developing webpage by Streamlit:

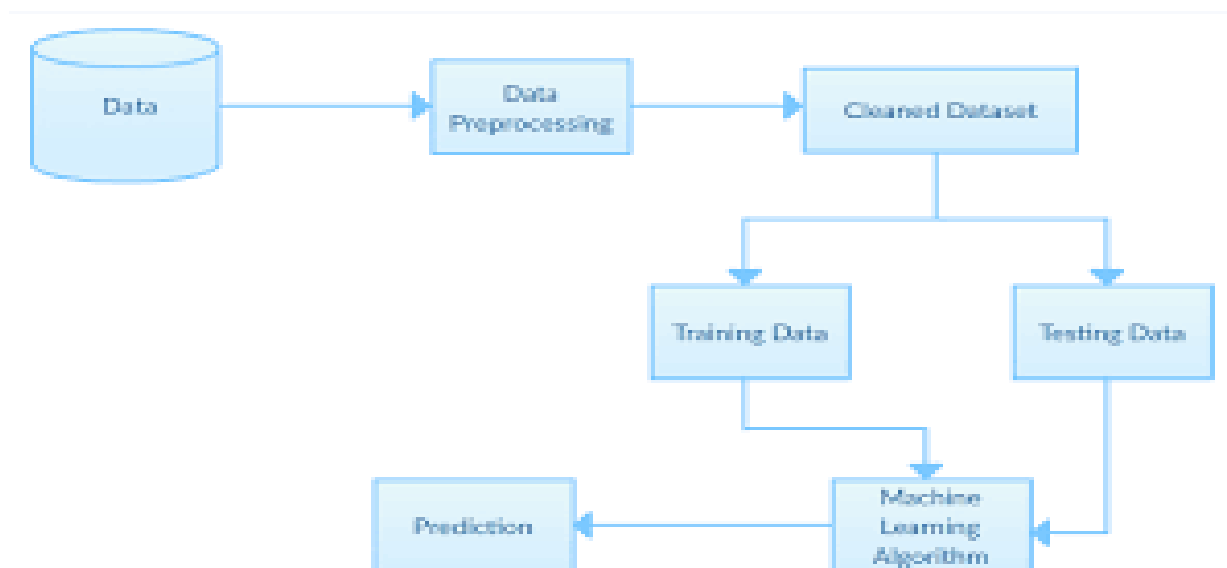
- ❖ Streamlit is a Python library designed to simplify the process of building data-driven web applications. It provides an intuitive and efficient way to create interactive user interfaces for data visualization, exploration, and machine learning. Streamlit eliminates the need for writing HTML, CSS, or JavaScript code, allowing data scientists and developers to focus on the core functionality of their applications.

CHAPTER 4

SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

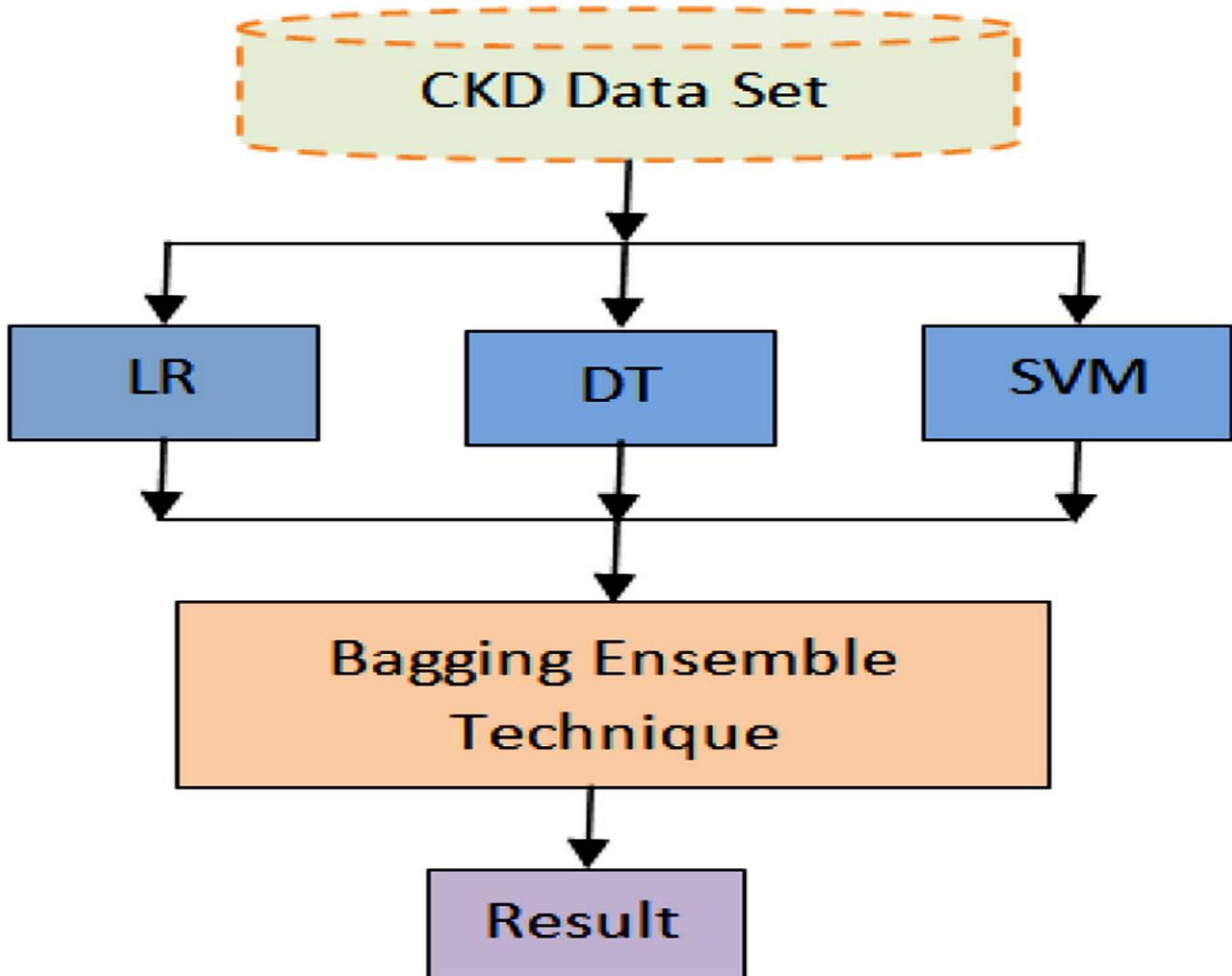
The system architecture for the ChronicKidneyDisease (CKD) prediction project is designed to efficiently and accurately assess CKD risk based on patient data. It begins by collecting data from various sources, including electronic health records (EHRs) and medical databases, and proceeds with data preprocessing to ensure data quality and consistency. Relevant features are then selected or engineered to create a robust feature set for CKD prediction. Machine learning models, such as Random Forest and Logistic Regression, are employed to make predictions, with an optional ensemble approach for improved accuracy. These models are evaluated using standard metrics, and the system is integrated into a healthcare environment with a user friendly interface for healthcare professionals. Stringent security measures are in place to protect patient data and privacy. Optionally, a feedback loop allows practitioners to contribute to model refinement, and continuous monitoring ensures system reliability. Comprehensive documentation outlines the architecture, data sources, algorithms used, and performance metrics, ensuring transparency and scalability for future enhancements.



CHAPTER 5

DESIGN AND IMPLEMENTATION

Designing and implementing a Chronic Kidney Disease (CKD) prediction project involves several key steps and considerations. Here's a high level overview of the design and implementation process:



1. Project Planning and Requirements:

- Define Objectives: Clearly articulate the goals of the CKD prediction project, such as early detection, risk assessment, or treatment recommendations.
- Gather Requirements: Identify data sources, data types, and specific CKD related features required for the project.
- Resource Allocation: Determine the necessary resources, including data, hardware, software, and personnel.

2. Data Collection and Preparation:

- Data Gathering: Collect patient data, including demographics, medical history, laboratory results, and CKD diagnosis labels, from reliable sources like EHR medical databases.
- Data Cleaning: Preprocess the data by handling missing values, outliers, and ensuring data quality.
- Feature Selection and Engineering: Select relevant features and potentially engineer new ones based on domain knowledge.

3. Model Selection and Development:

- Algorithm Selection: Choose appropriate machine learning algorithms for CKD prediction, considering factors like interpretability, model complexity, and accuracy.
- Model Training: Split the dataset into training and testing sets. Train the selected models on the training data and optimize hyperparameters.
- Ensemble Techniques: Consider ensemble methods like Random Forest or Gradient Boosting to improve prediction accuracy.

4. Model Evaluation:

- Performance Metrics: Assess model performance using various metrics, such as accuracy, precision, recall, F1 score, and ROC AUC, on the testing dataset.
- Cross Validation: Employ cross validation techniques to ensure the models' robustness and generalization to unseen data

5. Interpretability and Explainability:

- Feature Importance: Determine feature importance scores to understand which factors contribute most to CKD prediction.

- Model Explainability: Implement techniques to make model predictions interpretable and explainable, especially in a healthcare context

6. Integration and Deployment:

- System Integration: Integrate the CKD prediction models into a healthcare system or application, ensuring seamless data flow and user interaction.
- User Interface (UI): Develop a userfriendly interface for healthcare professionals to input patient data and receive CKD risk assessment

7. Security and Privacy:

- Data Protection: Implement robust security measures to safeguard patient data and ensure compliance with healthcare data protection regulations.
- Ethical Considerations: Adhere to ethical guidelines for handling sensitive medical information and patient privacy.

8. Testing and Validation:

- System Testing: Thoroughly test the integrated system to identify and address any issues or bugs.
- Validation: Validate the CKD prediction models in a clinical setting, ensuring their reliability and clinical relevance.

9. Documentation and Reporting:

- Project Documentation: Create comprehensive documentation covering data sources, methodologies, algorithms, and implementation details.
- Reporting: Prepare a detailed project report summarizing objectives, methods, results, and conclusions for stakeholders.

10. Maintenance and Future Enhancements:

- Ongoing Monitoring: Continuously monitor the system's performance and address any issues promptly.
- Model Updates: Periodically retrain the CKD prediction models with updated data to maintain accuracy.
- Future Directions: Identify opportunities for enhancing the project, such as real time monitoring, personalized treatment recommendations, or integration with electronic health record (EHR) systems.

Throughout the design and implementation process, prioritize data security, ethical considerations, and clinical applicability to ensure the project's success and impact in the healthcare domain.

5.1 DATASET AND FEATURES

ID (numerical)	Patient Id age(numerical)- age in years
Blood Pressure(numerical)	bp in mm/Hg. Blood pressure should be controlled to less than 130/80 if you have Chronic kidney disease.
Specific Gravity	Specific gravity, in the context of clinical pathology, is a urinalysis parameter commonly used in the evaluation of kidney function.
Albumin	Albumin is a protein found in the blood. A healthy kidney doesn't let albumin pass from the blood into the urine.
Sugar	High blood sugar from diabetes can damage blood vessels in the kidneys as well as nephrons so they don't work as well as they should.
Red Blood Cells	Red blood cells are responsible for carrying oxygen from the lungs to the body's tissues and organs.

Pus Cell	The presence of pus cells in the urine may suggest a secondary infection or inflammation of the kidneys or urinary tract.
Bacteria	Bacteria can be present in chronic kidney disease (CKD), but their presence can vary depending on the specific circumstances and underlying causes of CKD.
Blood Glucose Random	Blood glucose random refers to a measurement of your blood glucose levels taken at a random or unspecified time during the day, regardless of when you last ate.
Sodium	In the context of chronic kidney disease (CKD), sodium refers to the element sodium (Na) and its impact on the health of individuals with CKD.
Potassium	In the context of chronic kidney disease (CKD), potassium refers to the mineral potassium that is naturally present in many foods and is a crucial electrolyte in the body.
Haemoglobin	Hemoglobin refers to a protein found in red blood cells that plays a crucial role in transporting oxygen from the lungs to the rest of the body's tissues and organs.
White blood cells	White blood cells, also known as leukocytes, are a crucial part of the immune system. They play a key role in defending the body against infections and foreign invaders.
Hypertension	Hypertension, also known as high blood pressure, is a common medical condition in which the force of the blood against the walls of the arteries is consistently too high.
Diabetes Mellitus	Diabetes Mellitus in Chronic Kidney Disease refers to the presence of diabetes as a contributing factor to the development and progression of chronic kidney disease.
	Coronary artery disease (CAD), also known as coronary heart disease or ischemic heart disease, is a

Coronary Artery Disease	medical condition that involves the narrowing or blockage of the coronary arteries.
Appetite	CKD can affect a person's appetite in several ways, and understanding these changes is important for managing the condition and maintaining overall health.
Anemia	Anemia in chronic kidney disease (CKD) refers to a condition where there is a deficiency of red blood cells (RBCs) or a decrease in the amount of hemoglobin in the blood due to the impaired function of the kidneys.

5.2 THE STATISTICAL DESCRIPTION

Features	CKD			NOT CKD		
	Mean	Standard deviation	Median	Mean	Standard deviation	Median
Age	54.01	17.76	59.00	46.59	15.60	46.00
Blood Pressure	79.32	15.10	80.00	71.33	8.56	70.00
Blood Glucose random	174.46	91.63	143.50	107.61	18.47	107.00
Blood Urea	72.18	58.50	53.00	33.28	14.05	33.00
Serum Creatinine	4.34	6.80	2.25	0.90	0.37	0.90
Sodium	135.16	10.86	136	141.7	4.79	141.00
Potassium	4.68	3.57	4.30	4.35	0.58	4.50
Hemoglobin	10.75	2.24	10.90	15.17	1.25	15.00
Packed Cell Volume	33.14	7.00	33.00	46.28	4.12	46.00
White Blood Cell count	8944.80	3493.61	8500.00	7635.33	1888.57	7450.00
Red Blood Cell count	4.10	0.96	4.00	5.38	0.60	5.30

CHAPTER 6

MODULE DESCRIPTION

For this project, I wanted to compare different machine learning models: Random forests, Logistic Regression and Decision tree. For the purpose of this project, I wanted to compare these models by their accuracy.

6.1 LOGISTIC REGRESSION

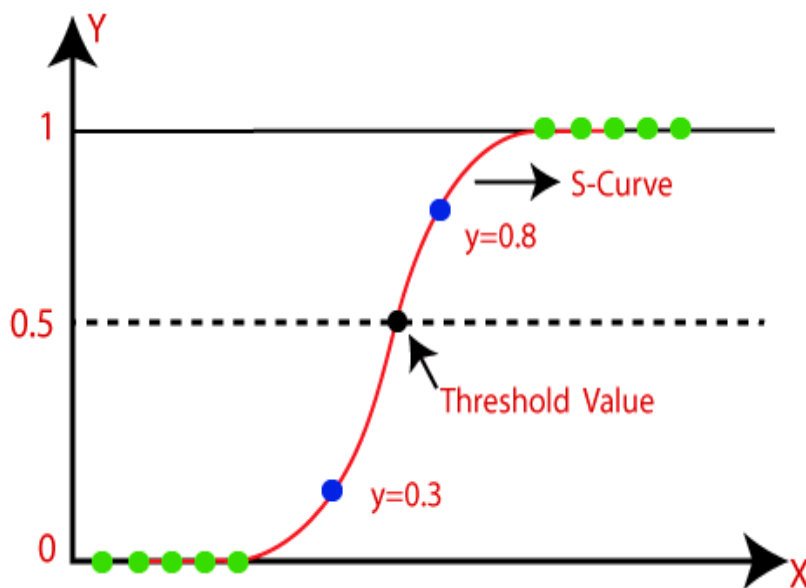
- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

Logistic Function (Sigmoid Function)

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.



Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation.

The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; 0 \text{ for } y=0, \text{ and infinity for } y=1$$

- But we need range between $-\infty$ to $+\infty$, then take logarithm of the equation it will become:

$$\log \left[\frac{y}{1-y} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

```
In [60]: from sklearn.linear_model import LogisticRegression
         clf = LogisticRegression(random_state=0,max_iter=500).fit(x_train, y_train)
         y_pred = clf.predict(x_test)
         from sklearn.metrics import classification_report, confusion_matrix
         print(confusion_matrix(y_test, y_pred))
         print(classification_report(y_test, y_pred))
```

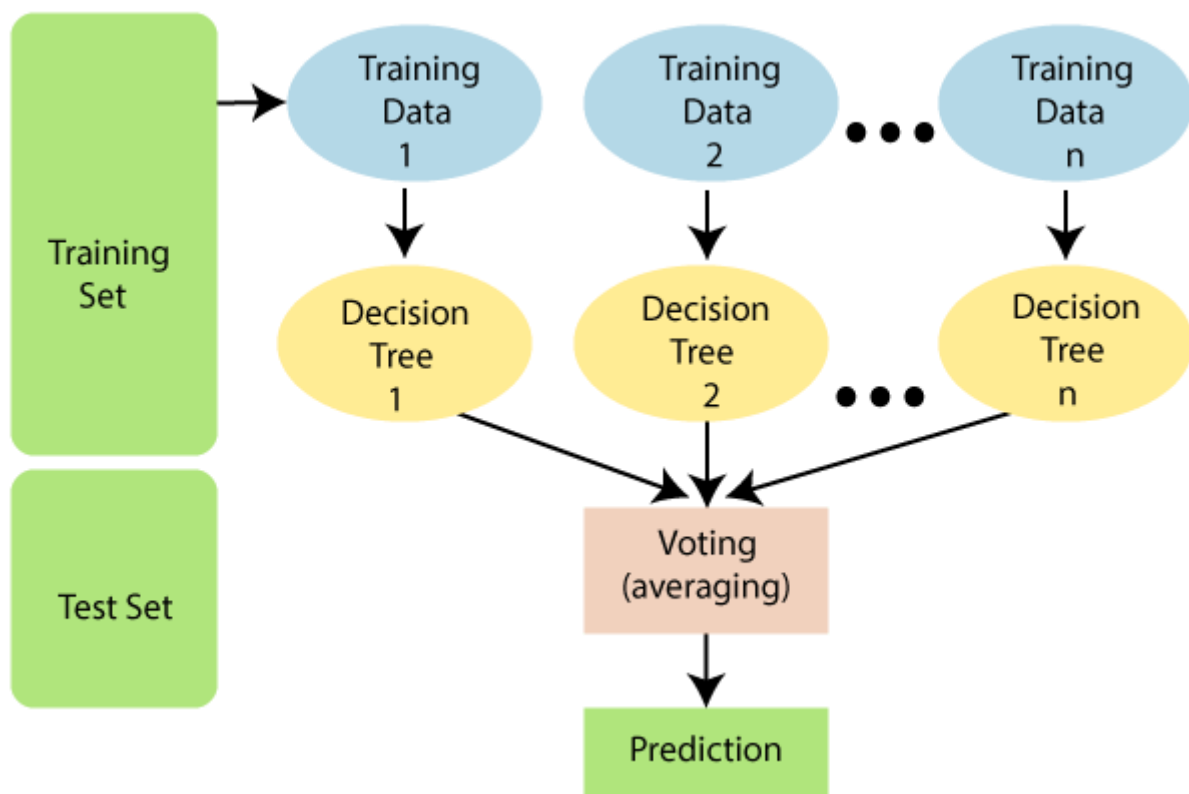
```
[[18  2]
 [ 1 35]]
```

	precision	recall	f1-score	support
0	0.95	0.90	0.92	20
1	0.95	0.97	0.96	36
accuracy			0.95	56
macro avg	0.95	0.94	0.94	56
weighted avg	0.95	0.95	0.95	56

6.2 RANDOM FOREST CLASSIFIER

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. Random forest is an ensemble supervised machine learning algorithm made up of decision trees. It is used for classification and for regression as well. In Random Forest, the dataset is divided into two parts (training and testing). Based on multiple parameters, the decision is taken and the target data is predicted or classified accordingly.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. The below diagram explains the working of the Random Forest algorithm:



```
[130] from sklearn.ensemble import RandomForestClassifier
```

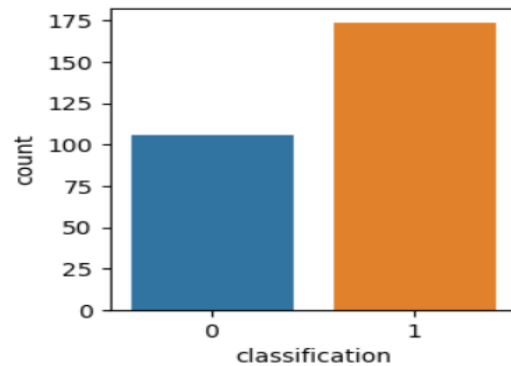
```
[142] from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
rf = RandomForestRegressor(n_estimators = 51, random_state = 1)
model = rf.fit(x_train, y_train)
np.sqrt(mean_squared_error(y_train, model.predict(x_train)))
print("The accuracy is : ",r2_score(y_train, model.predict(x_train)))
```

```
The accuracy is : 0.9933540518603956
```

CHAPTER 7

DATA VISUALIZATIONS

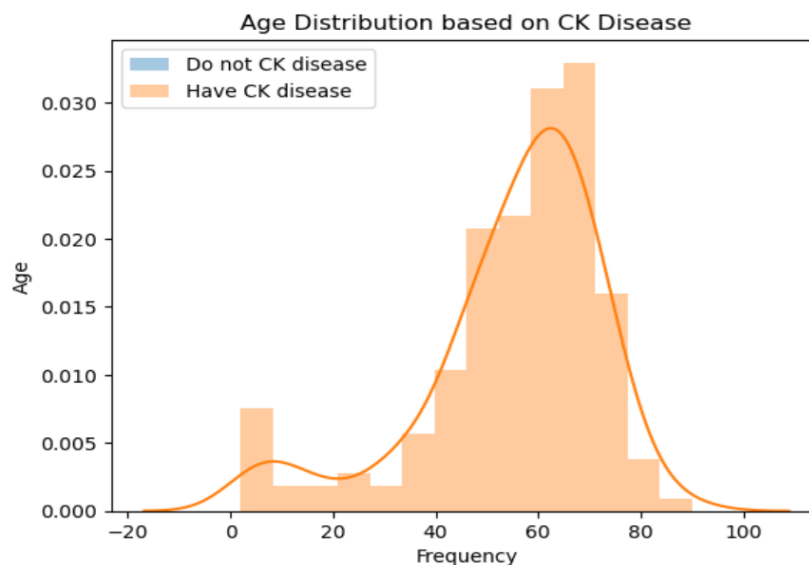
```
In [64]: plt.figure(figsize=(3,3))
sns.countplot(data=df1,x='classification')
plt.show()
```



- This plot clearly depicts that the target variable 1 which represents that the person has affected chronic kidney disease is more than the person who hasn't suffered the disease.

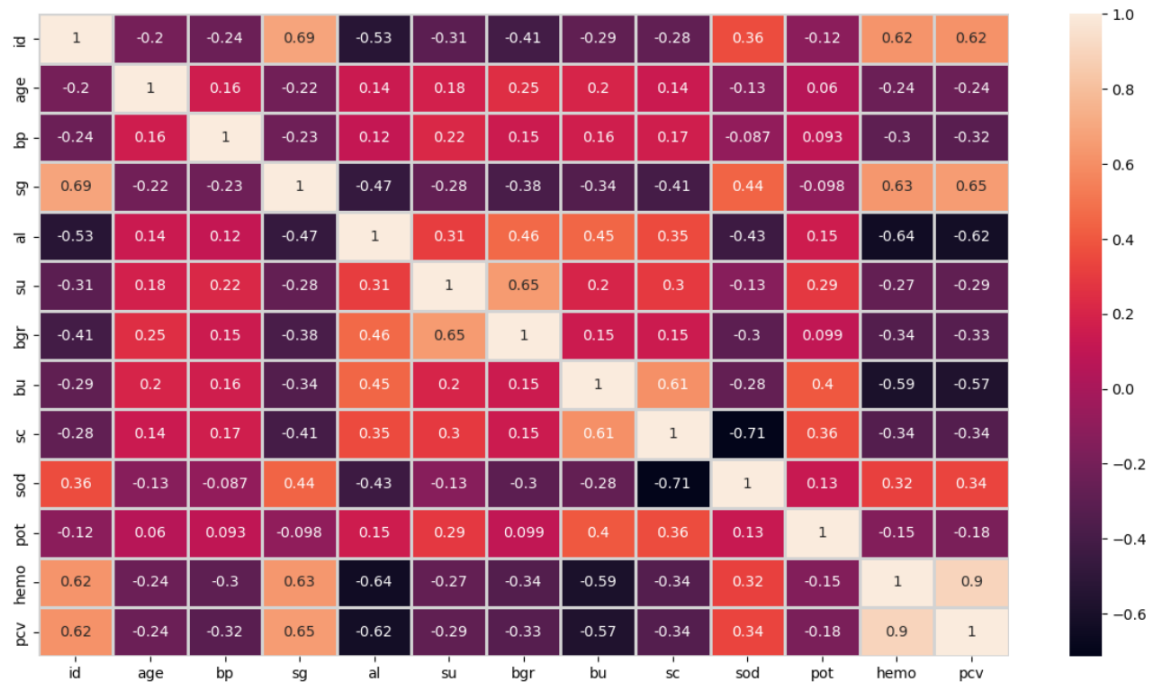
Distplot between age and class to predict the variation of chronic kidney disease among humans

```
In [61]: sns.distplot(df[df['classification'] == 'not ckd']['age'], label='Do not CK disease')
sns.distplot(df[df['classification'] == 'ckd']['age'], label = 'Have CK disease')
plt.xlabel('Frequency')
plt.ylabel('Age')
plt.title('Age Distribution based on CK Disease')
plt.legend()
plt.show()
```

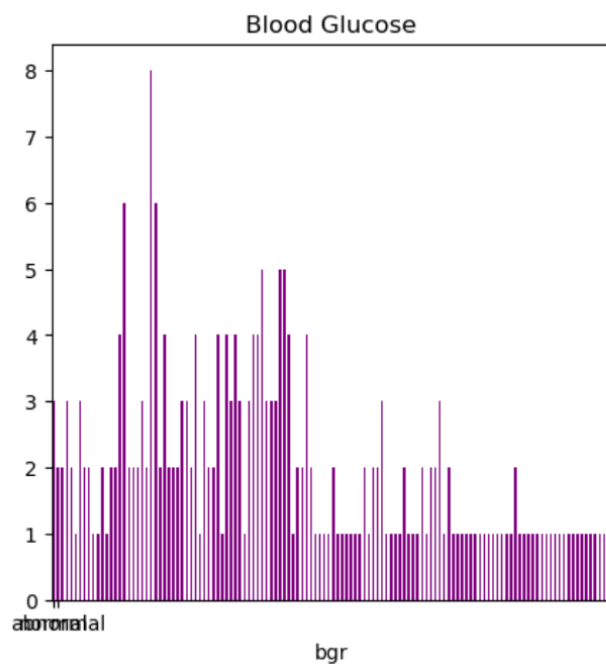


HEAT MAP TO FIND THE CORRELATION

```
In [92]: plt.figure(figsize = (15, 8))
sns.heatmap(df.corr(), annot = True, linewidths = 2, linecolor = 'lightgrey')
plt.show()
```



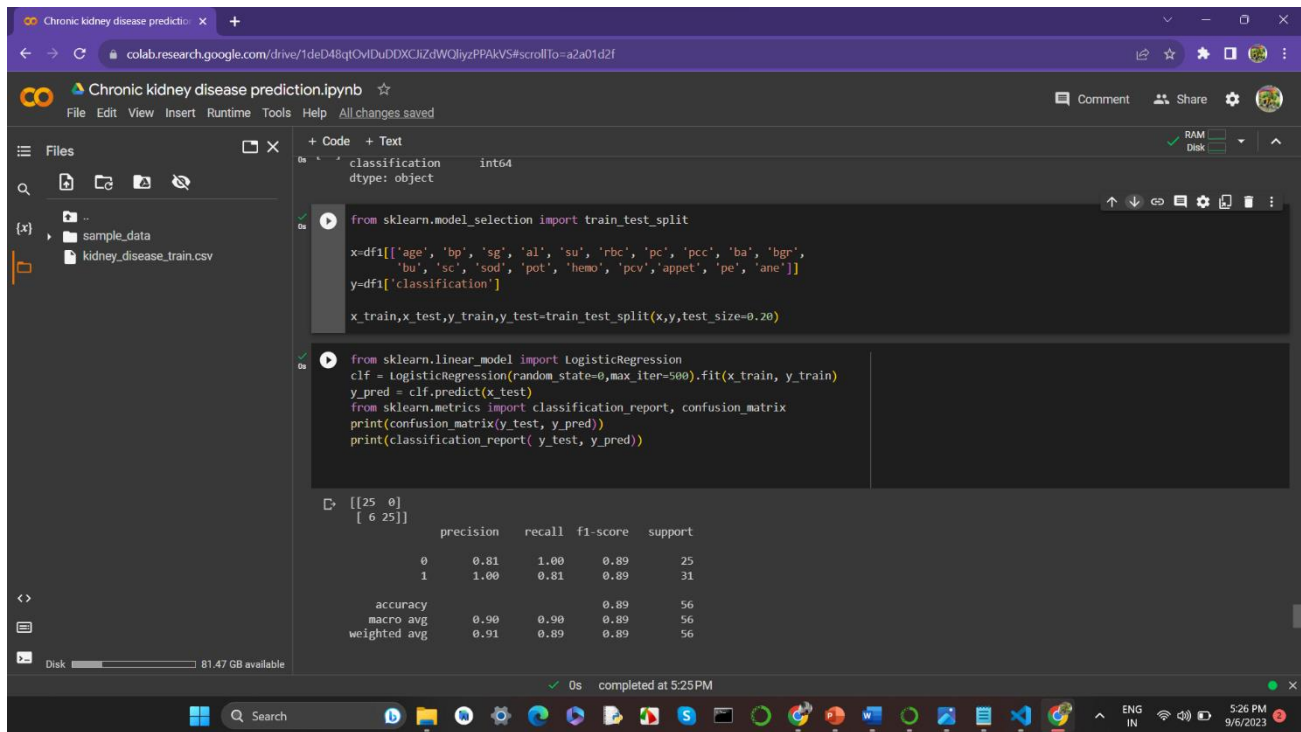
```
In [85]: plt.figure(figsize=(5,5))
df.groupby(df['bgr']).count()['classification'].plot(kind = 'bar', title = 'Blood Glucose',color='purple')
plt.xticks(np.arange(2), ('normal','abnormal'), rotation = 0)
plt.show()
```



CHAPTER 8

RESULT

Accuracy by using Logistic Regression



The screenshot shows a Google Colab notebook titled "Chronic kidney disease prediction.ipynb". The code cell contains the following Python code:

```
classification = int64
dtype: object

from sklearn.model_selection import train_test_split
x=df[['age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr',
      'bu', 'sc', 'sod', 'pot', 'hemo', 'pcv', 'appet', 'pe', 'ane']]
y=df['classification']
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20)

from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(random_state=0,max_iter=500).fit(x_train, y_train)
y_pred = clf.predict(x_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

The output of the code is a confusion matrix and a classification report:

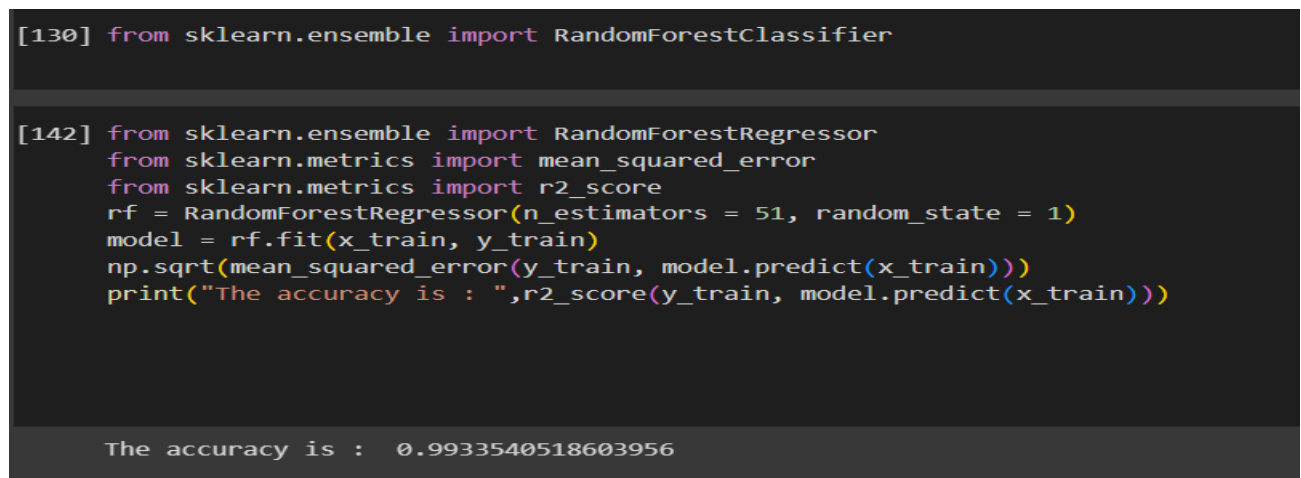
```
[[25  0]
 [ 6 25]]
```

	precision	recall	f1-score	support
0	0.81	1.00	0.89	25
1	1.00	0.81	0.89	31
accuracy			0.89	56
macro avg	0.90	0.90	0.89	56
weighted avg	0.91	0.89	0.89	56

The notebook interface shows the code was completed at 5:25 PM on 9/6/2023.

The Accuracy is 89%

Accuracy by using Random Forest



The screenshot shows a Jupyter notebook cell with the following Python code:

```
[130] from sklearn.ensemble import RandomForestClassifier

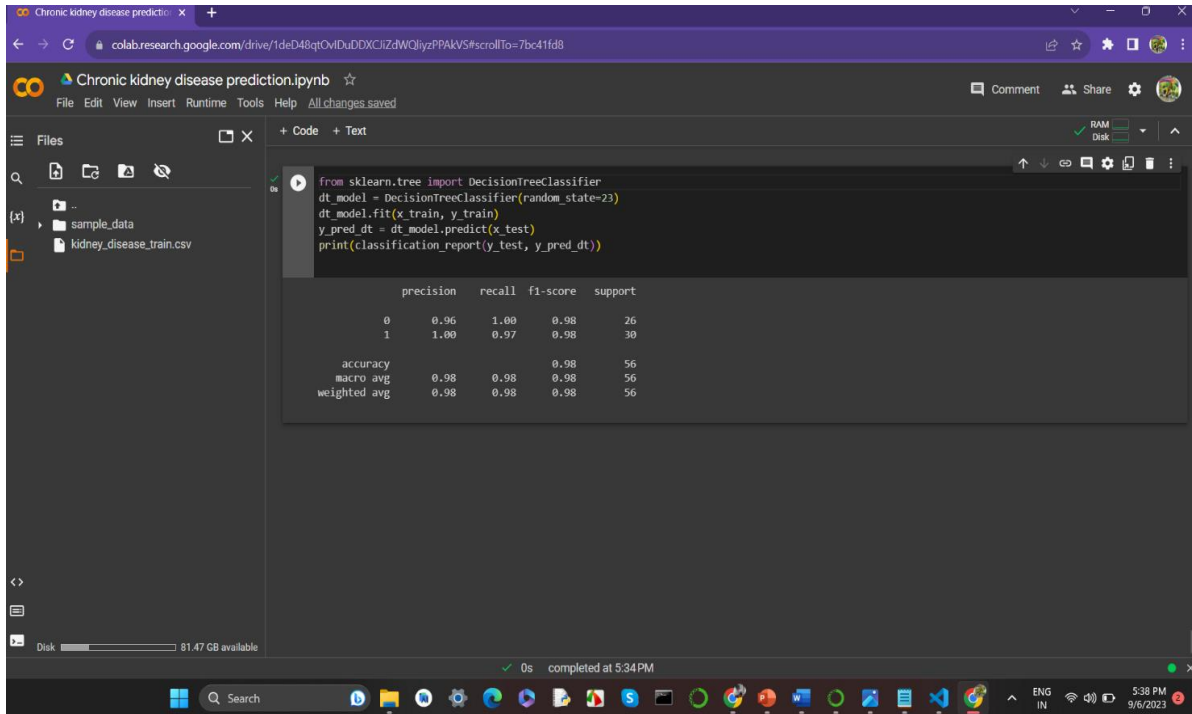
[142] from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
rf = RandomForestRegressor(n_estimators = 51, random_state = 1)
model = rf.fit(x_train, y_train)
np.sqrt(mean_squared_error(y_train, model.predict(x_train)))
print("The accuracy is : ",r2_score(y_train, model.predict(x_train)))
```

The output of the code is:

```
The accuracy is : 0.9933540518603956
```

The Accuracy is 99%

Accuracy by using Decision Tree



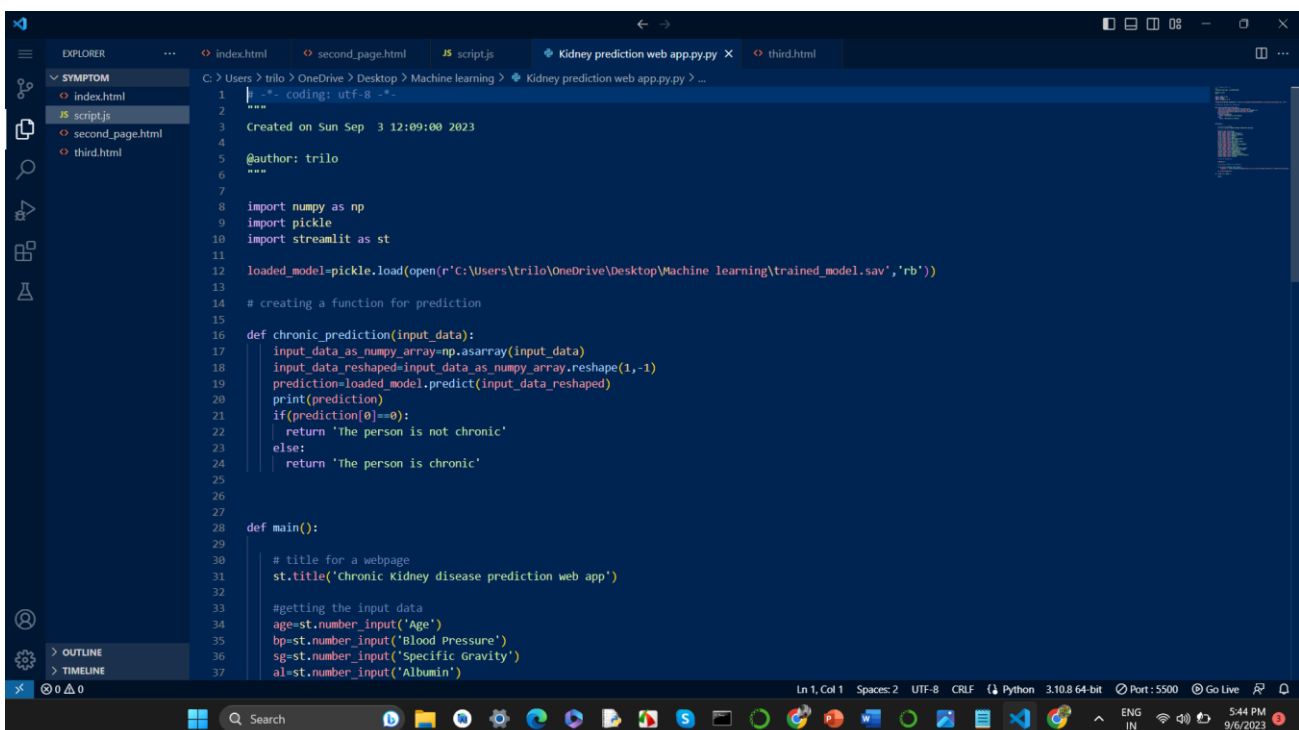
```
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(random_state=23)
dt_model.fit(x_train, y_train)
y_pred_dt = dt_model.predict(x_test)
print(classification_report(y_test, y_pred_dt))
```

	precision	recall	f1-score	support
0	0.96	1.00	0.98	26
1	1.00	0.97	0.98	30
accuracy			0.98	56
macro avg	0.98	0.98	0.98	56
weighted avg	0.98	0.98	0.98	56

The Accuracy is 98%

Creating a Chronic kidney disease web app using Streamlit

Streamlit is an open-source python library that allows swift custom web app building revolving data science, machine learning, and much more. Streamlit aims at building and deploying applications without the necessity of any web development knowledge.



```
#!/usr/bin/env python
# coding: utf-8
"""
Created on Sun Sep 3 12:09:00 2023

@author: trilo
"""

import numpy as np
import pickle
import streamlit as st

loaded_model=pickle.load(open(r'C:\Users\trilo\OneDrive\Desktop\Machine learning\trained_model.sav','rb'))

# creating a function for prediction
def chronic_prediction(input_data):
    input_data_as_numpy_array=np.asarray(input_data)
    input_data_resaped=input_data_as_numpy_array.reshape(1,-1)
    prediction=loaded_model.predict(input_data_resaped)
    print(prediction)
    if(prediction[0]==0):
        return 'The person is not chronic'
    else:
        return 'The person is chronic'

def main():
    # title for a webpage
    st.title('Chronic Kidney disease prediction web app')

    #getting the input data
    age=st.number_input('Age')
    bp=st.number_input('Blood Pressure')
    sg=st.number_input('Specific Gravity')
    al=st.number_input('Albumin')
```

```
32
33
34 #getting the input data
35 age=st.number_input('Age')
36 bp=st.number_input('Blood Pressure')
37 sg=st.number_input('Specific Gravity')
38 al=st.number_input('Albumin')
39 su=st.number_input('Sugar')
40 rbc=st.number_input('Red Blood cells')
41 pc=st.number_input('Pus cells')
42 pcc=st.number_input('Pus cells clone')
43 ba=st.number_input('Bacteria')
44 bgr=st.number_input('Blood Glucose Random')
45 sod=st.number_input('Sodium')
46 pod=st.number_input('Potassium')
47 wc=st.number_input('White blood cell count')
48 rc=st.number_input('Red blood cell count')
49 hemo=st.number_input('Hemoglobin')
50 htn=st.number_input('Hypertension')
51 dm=st.number_input('Diabetes mellitus')
52 cad=st.number_input('Coronary artery disease')
53 ane=st.number_input('Anemia')
54
55 # code for prediction
56 diagnosis=''
57
58 # creating a button for prediction
59
60 if st.button('Chronic test result'):
61 | | diagnosis = chronic_prediction([age,bp,sg,al,su,rbc,pc,pcc,ba,bgr,sod,pod,wc,rc,hemo,htn,dm,cad,ane])
62
63 st.success(diagnosis)
64
65 if __name__ == '__main__':
66 | | main()
67
68
```

CHRONIC KIDNEY DISEASE PREDICTION APP

Chronic Kidney disease prediction web app

Age: 65.00

Blood Pressure: 110.00

Specific Gravity: 3.23

Albumin: 5.20

Sugar: 1.00

Red Blood cells: 422.99

Pus cells: 49.99

Pus cells clone: 6.24

Parameter	Value
	136.00
Potassium	4.69
White blood cell count	7900.00
Red blood cell count	3.18
Hemoglobin	12.59
Hypertension	0.00
Diabetes mellitus	0.00
Coronary artery disease	0.00
Anemia	0.00

Chronic test result

The person is not chronic

CHAPTER 9

LIMITATIONS

Obtaining large, high-quality datasets for kidney disease prediction can be challenging. The available data may be imbalanced, with a disproportionate number of healthy individuals compared to those with kidney disease, making it difficult for the model to learn effectively. Some machine learning models, particularly complex ones like deep learning neural networks, can be challenging to interpret. In the medical field, understanding why a model made a particular prediction is essential for clinical acceptance.

CHAPTER 10

CONCLUSION

In conclusion, this CKD prediction project employs machine learning algorithms, including Random Forest and Logistic Regression, to develop a robust and accurate predictive model for chronic kidney disease. Random Forest, with its ensemble of decision trees, offers high accuracy and resilience against overfitting. Logistic Regression, a simpler yet interpretable model, aids in understanding feature importance. By combining these algorithms, we enhance the project's predictive power and interpretability. Through rigorous data preprocessing, model training, and evaluation, we aim to provide healthcare professionals with effective tools for early CKD detection and risk assessment. This project embodies the fusion of advanced technology and clinical insights to address a critical healthcare challenge, ultimately improving patient care and outcomes.

CHAPTER 11

REFERENCES

- 1.C. Ho, T. Pai, Y. Peng, C. Lee, Y. Chen and Y. Chen, "Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease", *IEEE Complex Intelligent and Software Intensive Systems*, pp. 624 629, July 2012.
- 2.A. Miguel, Estudillo Valderrama, Alejandro Talaminos Barroso, Laura M. Roa, David Naranjo Hernandez, Javier Reina Tosina, et al., "A Distributed Approach to Alarm Management in Chronic Kidney Disease", *IEEE Transl. Biomedical and Health Informatics*, vol. 18, pp. 1796 1803, November 2014.
- 3.A. Rosmani, U. Mazlan, A. Ibrahim and D. Zakaria, "i KS: Composition of Chronic Kidney Disease (CKD) Online Informational Self Care Tool", *Computer Communication and Control Technology IEEE*, pp. 379 383, April 2015.
- 4.Hsieh Jun Wei, C. Hung Lee, Y. Chih Chen, W. Shan Lee and H. Fen Chiang, "Stage Classification in Chronic Kidney Disease by Ultrasound Image", *International Conference on Image and Vision Computing New Zealand ACM*, pp. 271 276, 2014.

5.A. Singh, G. Nadkarni, J. Guttag and E. Bottinger, "Leveraging hierarchy in medical codes for predictive modeling", *Bioinformatics Computational Biology and Health Informatics ACM*, pp. 96 103, 2014.

6.R. Kei Chiu, R. Y. Chen, S. Wang and S. Jian, "Intelligent systems on the cloud for the early detection of chronic kidney disease", *Machine Learning and Cybernetics IEEE*, pp. 1737 1742, July 2012.

7.A. Asuncion and D. J. Newman, *UCI Machine Learning Repository*, 2007, [online] Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

8.J. R. Quinlan, "C4.5: programs for machine learning", *Morgan Kaufmann Publishers Inc.*, 1993.

9.R. G. Brereton and G. R. Lloyd, "Support Vector Machines for classification and regression", *Analyst*, vol. 135, no. 2, pp. 230 267, 2010.

10.S. Galit, R. P. Nitin and C. B. Peter, *Data Mining for Business Intelligence: Concepts Techniques and Applications in Microsoft Office Excel with XLMiner*: Wiley Publishing, 2010.

11.R. Xi, N. Lin and Y. Chen, "Compression and Aggregation for Logistic Regression Analysis in Data Cubes", *IEEE Transl. Knowledge and Data Engineering*, vol. 21, pp. 479 492, April 2009.