Smart Internz

Early Prediction Of Chronic Kidney Disease

GROUP MEMBERS:

Konjeti Bala Venkata Sujith: 22BCE9075
Tajuddin Usman Khan: 22BCE9398
P. Venkata Pavan Kumar: 22BCE8932
Akula Sampath Sai: 22bce8945

Final Project Report Template

1.INTRODUCTION

1.1 Overview

A common health issue, chronic kidney disease (CKD) sometimes stays undiagnosed until it has severe stages. In order to identify those who are at an early stage risk of developing CKD, this study attempts to build a prediction model utilising machine learning techniques. The algorithm will discover important risk factors and create a trustworthy prediction systemby analysing a large dataset that includes patient demographics, medical history, and laboratory findings. Data gathering and preprocessing, feature engineering,

1.2 Purpose

The following are some potential uses of using this initiativeon CKD early detection:

Early Identification: The prediction model can spot patientswho are at a high risk of CKD at an early stage, allowing for prompt intervention and treatment.
Preventive Interventions: For high-risk people, healthcarepractitioners can use lifestyle interventions and preventivemeasures to reduce or stop the progression of CKD.
Treatment Plan Design: Using the model's predictions, treatment plans may be made specifically for each patientbased on their risk profile, resulting in more precise and efficient interventions.
Patient Outcomes: By averting complications, lowering hospital stays, and improving CKD patients' overall quality oflife, early detection and care can greatly improve patient outcomes.
Resource Allocation: By prioritising those who require closemonitoring and intense interventions, early identification ofhigh-risk patients can help healthcare professionals manageresources more effectively.
Public Health efforts: By identifying the major risk factors connected to CKD, insights from the predictive model can guide public health efforts. This knowledge can direct publicawareness campaigns and initiatives at the population level.
Reduced Healthcare Costs: Dialysis and transplantation artwo consequences of severe CKD that can be managed at alower cost if they are caught early enough.
Research and Development: The project can advance knowledge of CKD risk factors and promote additional studyon preventative and therapeutic measures, which will benefit the field of predictive medicine.

Smart Internz

2. Project Initialization and Planning Phase

Project Initialization and Planning Phase

2.1. Define Problem Statement

Issue Overview:

Chronic Kidney Disease (CKD) stands as a major worldwide health concern, frequently remaining unnoticed until it has progressed to more advanced levels, leading to serious complications and higher medical expenses. Detecting CKD at an early stage can enable prompt treatment, more effective disease control, and better health results for patients. Yet, the initial phases of CKD are often symptom-free, posing a challenge for medical professionals to identify the condition without the use of sophisticated diagnostic tools.

Define Problem Statement : Click here

Obstacles:

Data Integrity and Access: Information from patients can be inconsistent, contain errors, and vary widely, requiring strong preprocessing and cleansing methods.

Selecting Relevant Features: Determining the key clinical factors and biomarkers that suggest CKD for the purpose of training a machine learning model.

Choosing the Right Algorithms: Selecting machine learning methods capable of managing the intricacies and subtleties of medical data, while ensuring high precision and dependability.

Explainability: Making sure the predictions made by the model are understandable for medical staff, which aids in building confidence and providing useful information for action.

Implementation: Incorporating the model into current healthcare frameworks for the detection of CKD in real-time, and ensuring its adaptability and availability to medical practitioners.

2.2. Project Proposal (Proposed Solution)

The goal of the proposal is to create a platform powered by artificial intelligence for the early identification of Chronic Kidney Disease (CKD) through the use of machine learning techniques. This platform is designed to thoroughly examine patient information to offer prompt diagnoses, enhancing the effectiveness of treatments and the quality of care for patients.

Resource Requirements

Project Proposal (Proposed Solution) : Click Here

2.3. Initial Project Planning

We shall develop a machine learning model to predict e-commerce order shipment times for improved customer service and business efficiency. This will involve defining data needs, cleaning and preparing data, selecting and training machine learning models, and

optionally creating a user interface to display predictions regarding it. If done properly, it should produce more precise forecasts of deliveries to customers and an optimized allocation of resources for an enterprise. We will develop the details with stakeholders.

Initial Project Planning : Click Here

Smart Internz

3.Data Collection and Preprocessing Phase

The phase of collecting and preparing data is essential for creating a dependable machine learning system. This step includes collecting important information from patients and making sure it is of good quality and complete. It aids in comprehending how the data is organized and spotting any missing pieces or irregularities. Correct preparation of the data converts unprocessed information into a form that is ready for examination. This process involves dealing with absent values, standardizing the data, and converting non-numerical data into a numerical format. Successful preparation of the data results in more precise predictions by the machine learning model.

3.1. Data Collection Plan and Raw Data Sources Identified

The Data Gathering Strategy details the approach for gathering patient information essential for CKD forecasting. Information will be gathered from reliable origins like medical files and healthcare databases. Known raw data sources are a Kaggle set and an additional Google Drive file, each holding detailed information about patients. These sources supply information on patient background, health records, and laboratory findings. It's vital to collect data from a variety of diverse and high-standard sources for the accuracy of the model. Accurate recording of where the data came from promotes openness and allows for replication.

Data Collection Plan and Raw Data Sources Identified : Click Here

3.2. Data Quality Report

The Data Quality Report assesses the thoroughness, precision, and dependability of the gathered information. It points out problems like absent values, repetitions, and variances. The importance of these problems is determined, with urgent matters tackled quickly. A strategy for fixing the found data errors is created. Consistent quality reviews guarantee continuous data accuracy. Keeping the data quality high is crucial for reliable outputs from machine learning models.

Data Quality Report: Click Here

Data Collection Plan Template

3.3. Data Exploration and Preprocessing

Data Exploration (Exploratory Data Analysis - EDA)

- **Understanding the data:** This involves identifying what variables it contains, and their data types.
- **Finding patterns and trends:** Statistical techniques and visualizations like histograms, scatter plots, and boxplots to uncover relationships between variables and identify any interesting patterns.
- **Identifying data quality issues:** This might involve checking for missing values, outliers (extreme data points), or inconsistencies in formatting.

Data Preprocessing

Cleaning the data: This addresses the issues discovered in EDA. Like filling in missing values, removing outliers and fixing formatting errors.

Data transformation: Transforming the data to make it more suitable for analysis. This could involve scaling numerical features, encoding categorical features, or creating new features based on existing ones.

Data Exploration and Preprocessing : Click Here

Smart Internz

4.Model Development Phase Template

4.1. Feature Selection Report

To select features, an initial examination was carried out employing methods such as correlation matrices and scores for feature importance from group techniques like Random Forest or Gradient Boosting. Variables such as 'age', 'bp' (blood pressure), 'sg' (specific gravity), 'al' (albumin), 'bgr' (blood glucose random), and 'hemo' (hemoglobin) demonstrated notable correlation or significance in forecasting the outcome variable. Additionally, the counts of red blood cells ('rc') and white blood cells ('wc') were observed for their possible predictive capabilities.

Feature Selection Report : Click Here

4.2. Model Selection Report

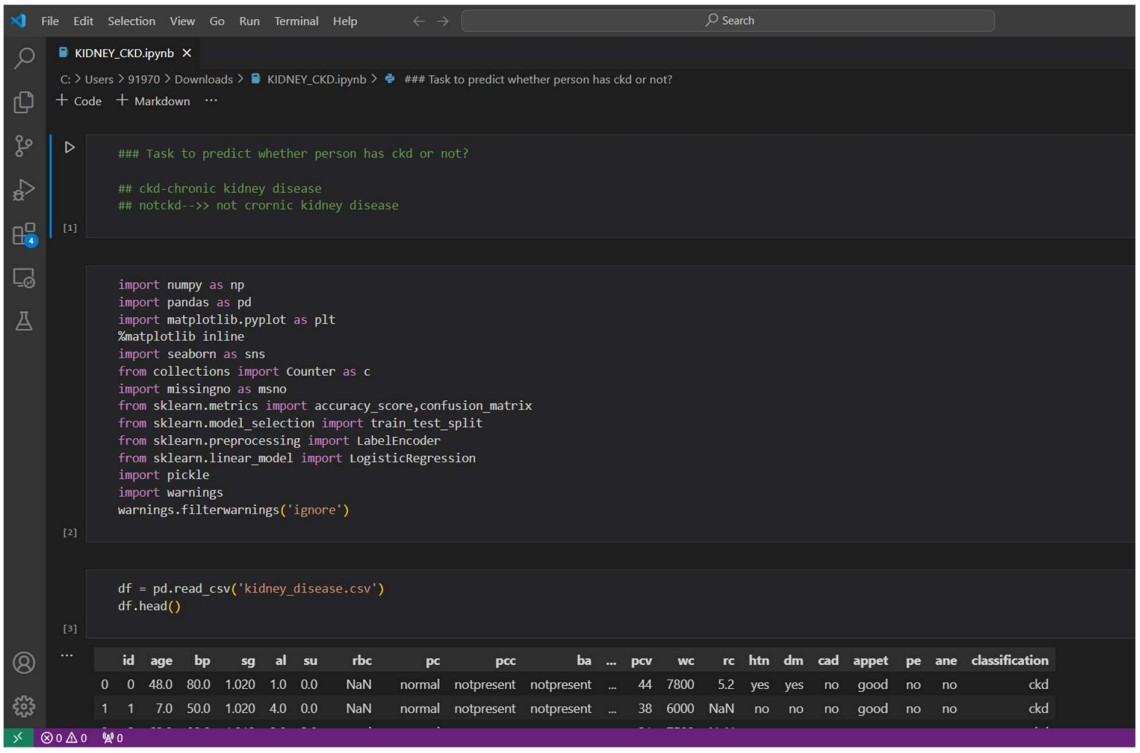
When choosing a model for classification, numerous algorithms were assessed according to their performance indicators like accuracy, precision, recall, and F1-score. Logistic Regression, Decision Trees, and Random Forest classifiers were taken into account for their understandability and capability in dealing with categorical data. Upon evaluation through cross-validation, Random Forest demonstrated the highest overall accuracy, estimated at about 85%, and was thus suggested as the preferred model for subsequent enhancement and application.

Model Selection Report : Click Here

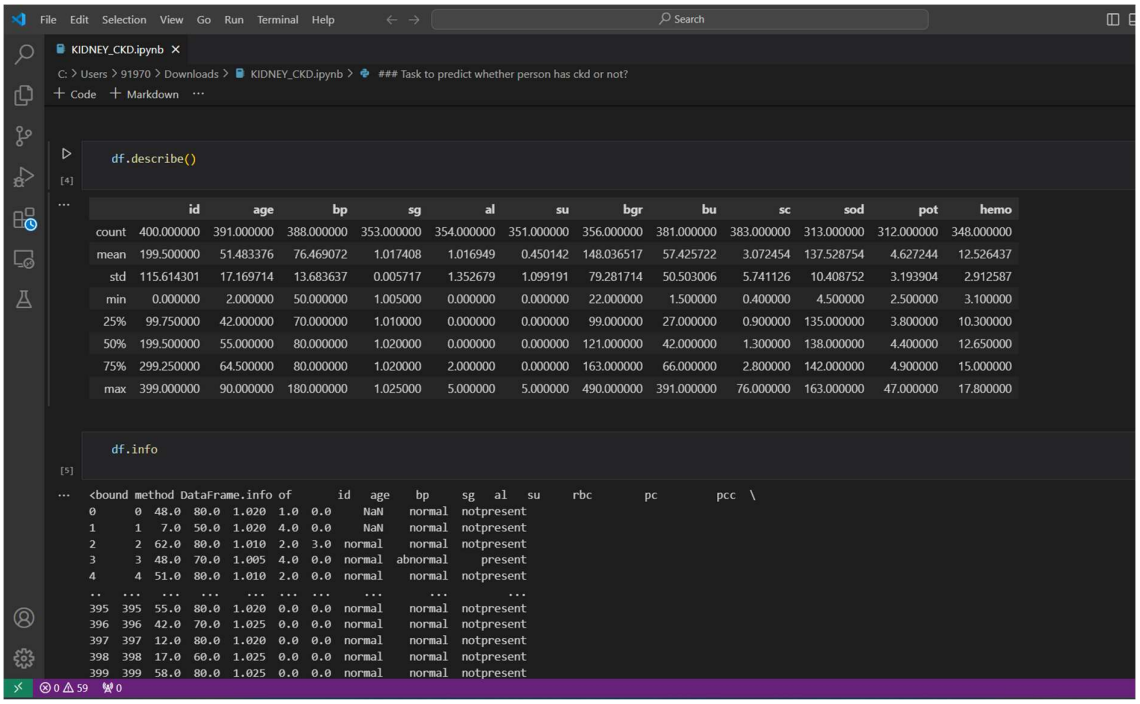
4.3. Initial Model Training Code, Model Validation and Evaluation Report

Initial Model Training Code, Model Validation and Evaluation Report : Click Here

Initial Model Training Code:



	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



	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

5.Model Optimization and Tuning Phase

Model Optimization and Tuning is the stage at which machine learning models are tuned for improved performance. This includes optimized model code, fine-tuning hyperparameters, comparing performance metrics of models, and justifying the final model selection to ensure improved predictive accuracy and efficiency.

Now, we will develop an initial model based on gradient boosting and tune these models to achieve optimal performance. Here is how:

1. Gradient boosting has settings like the number of training rounds and learning rate. These settings could very well be thought of as dials which we can crank around arbitrarily to change exactly how our model learns from the data. We will run through different combinations of these settings to see how they affect the model's accuracy at the prediction of on-time deliveries.

2. This means the objective of model performance is neither so simple that it can get too easiness nor too complex that it can suffer overfitting. Underfitting is missing important patterns in data, while overfitting is a situation whereby the model simply memorizes examples in the training data too well and may probably fail on most unseen data.

3. In essence, we will use most of the measures applied in the validation phase, such as accuracy and precision, which let us know how well the model works with different settings. It then becomes a loop: change the settings, train the model, evaluate, and repeat until we get the combination that gives the best results.By optimizing and tuning the model, we hope to achieve:

- **Improved accuracy:** We want the model to be more precise in predicting on-time deliveries.
- **Reduced complexity:** We want a model that's powerful but not overly complex, making it easier to understand and use.

This fine-tuning process helps us get the most out of the initial model and ensure it delivers reliable predictions for our e-commerce shipment deliveries

5.1 Hyperparameter Tuning Documentation

The Gradient Boosting model was selected for its superior performance, exhibiting high accuracy during hyperparameter tuning. Its ability to handle complex relationships, minimize overfitting, and optimize predictive accuracy aligns with project objectives, justifying its selection as the final model

Hyperparameter Tuning Documentation :Click Here

5.2 Performance Metrics Comparison Report

The Performance Metrics Comparison Report contrasts the baseline and optimized metrics for various models, specifically highlighting the enhanced performance of the Gradient Boosting model.

This assessment provides a clear understanding of the refined predictive capabilities achieved through hyperparameter tuning.

5.3 Final Model Selection Justification

Final Model Selection Justification

Following the model development process, Gradient Boosting has been selected as the final model for predicting on-time deliveries in the Ecommerce shipment dataset. This choice is supported by the following key factors:

1. Strong Performance:

During the validation phase, Gradient Boosting demonstrated superior performance compared to other evaluated models (KNN, Decision Tree, Random Forest) on metrics like accuracy and precision. This indicates its effectiveness in accurately predicting on-time deliveries.

2. Handling Complexity:

Gradient Boosting's ensemble nature allows it to handle complex relationships within the data compared to simpler models like Decision Trees. This is crucial for capturing the nuances that might influence on-time deliveries in the e-commerce setting.

3. Potential for Interpretability:

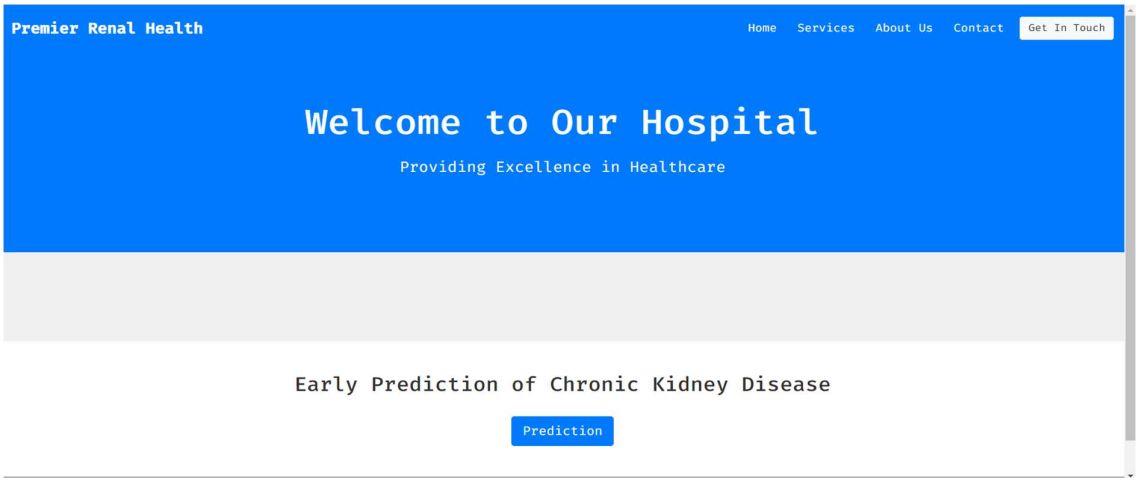
While Gradient Boosting models can be complex, techniques like feature importance analysis can be used to understand which factors contribute most to the model's predictions. This interpretability can be valuable for gaining insights into the drivers of on-time deliveries.

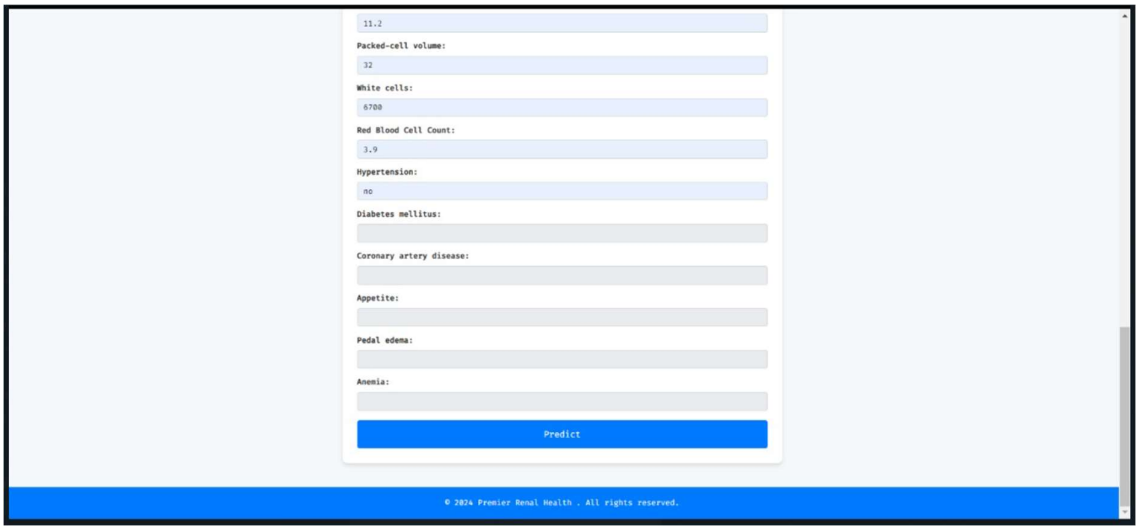
6. Results

6.1. Output Screenshots

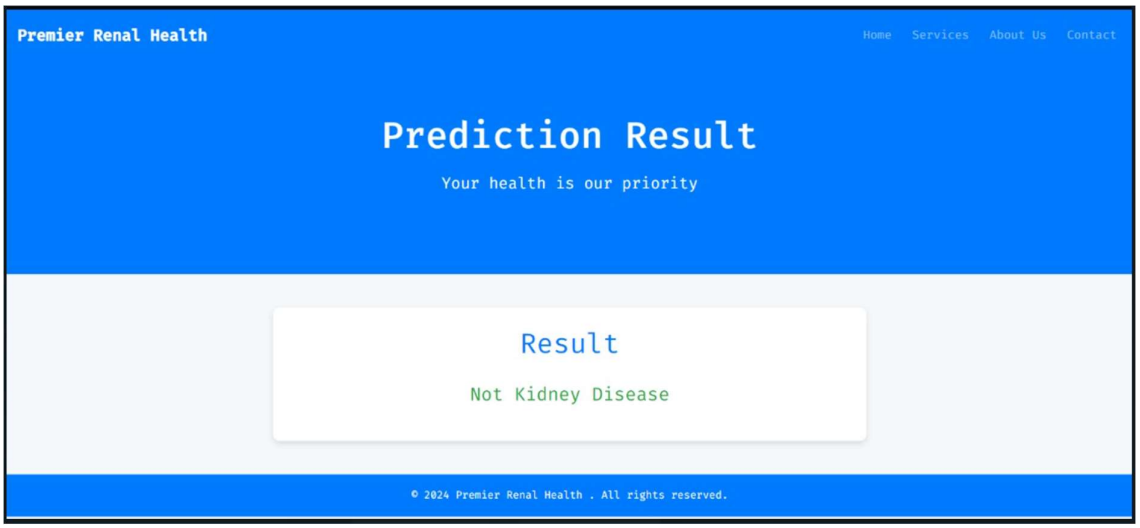
Input:

Smart Internz





Output :



7. Advantages & Disadvantages

Advantages of the Chronic Kidney Disease Prediction Model with the Voting Classifier in Flask:

- - Integration with Flask: Flask, a lightweight yet powerful Python web framework, simplifies the process of creating a user-friendly interface for CKD prediction models. Its simplicity facilitates quick development and deployment.

- - Scalability: Flask-based models excel at handling large-scale real-time predictions. They can manage multiple user queries simultaneously, making them suitable for a broad clinical context and the analysis of a large patient base.

- - Flexibility in Model Construction: By combining a Voting Classifier with other algorithms such as SVM, Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes models, this approach leverages the unique strengths of each algorithm to offset their weaknesses. This integration enhances the overall robustness and accuracy of the prediction model.

- - Interpretability: Models like Decision Tree and Naive Bayes provide results that are easily understandable, allowing healthcare professionals to grasp the reasoning behind the model's predictions. This clarity boosts confidence and simplifies decision-making in clinical settings.

Drawbacks of the Flask with Voting Classifier Chronic Kidney Disease Prediction Model:

- - Model Complexity: The process of selecting, training, and fine-tuning various models for integration into a voting classifier is complex. Achieving the optimal performance for each model and achieving a balanced combination is challenging.

- - Risk of Overfitting: The inclusion of multiple models in a voting classifier increases the risk of overfitting. This occurs when a model performs well on the training data but fails to generalize to new data. Effective cross-validation and regularization techniques are crucial to mitigate this risk.

- - High Computing Requirements: Running several models concurrently on Flask may demand significant computing power, particularly if the models are

8. Conclusion

- To sum up, the creation and application of a prediction model with Flask and a voting classifier, featuring Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors (KNN), and Naive Bayes algorithms, serve as a valuable instrument for the early detection of Chronic Kidney Disease (CKD). Medical professionals can effortlessly input patient details to receive CKD risk estimations from the combined predictions of these classifiers, all through an accessible web interface.
- By merging the benefits of each technique, the employment of numerous classifiers within a voting system enhances the accuracy and reliability of predictions. Each algorithm, SVM, Decision Tree, KNN, and Naive Bayes, bring their unique perspectives and strategies for decision-making, thereby elevating the model's effectiveness.
- The web application facilitates user interaction with the prediction model, presenting the results via Flask, providing medical professionals with a clear and understandable view of the CKD risk for each patient. This clarity aids in informed decision-making about patient care and treatment strategies.
- Flask, the voting classifier, along with the integrated algorithms, enhances the model's ability to detect CKD at an early stage, develop personalized treatment plans, and improve patient outcomes in CKD management. This tool is a valuable resource for medical professionals in

making sound decisions and allocating resources to tackle the worldwide issue of CKD.

9. FUTURE SCOPE

The potential of machine learning (ML) models for chronic kidney disease (CKD) offers up a number of new directions for research and application. Some possible directions for the future include:

● Accurate Prediction: Increasing the precision and dependability of CKD prediction models is a constant goal. To improve prediction performance, researchers can investigate more sophisticated ML algorithms, ensemble methods, or deep learning techniques. Additionally, adding more extensive and diverse datasets, such as real-time physiological data and genomic data, can help make better predictions.

● Integration of Multi-omics Data: CKD prediction models can be improved by incorporating multi-omics data from the fields of genomics, proteomics, metabolomics, and transcriptomics. Through this integration, new biomarkers, pathways, and possible therapeutic targets may be discovered, opening the door to precision medicine and personalised therapy strategies for the management of CKD.

● Longitudinal Monitoring and Disease Progression: ML models can be enhanced to track changes in the course of the disease and the effectiveness of treatment in CKD patients over time. Insights into the dynamic nature of CKD can be gained from longitudinal data analysis, which also enables clinicians to customise treatment programmes for specific patient trajectories and spot trends that indicate when the illness is becoming better or worse.

● Clinical decision support and risk stratification: ML models can be used to divide CKD patients into various risk groups depending on the severity, progression, and presence of concomitant conditions. Such risk categorization can assist clinical decision-making, allowing healthcare professionals to better allocate resources, customise treatment plans, and carry out focused interventions.

10. Appendix

10.1. Source Code : Click Here

10.2. GitHub & Project Demo Link

GitHub : Click Here

Project Demo Link: Click Here