Chronic Kidney Disease Classification

Introduction

Chronic Kidney Disease (CKD) is a progressive condition characterized by the gradual deterioration of kidney function over time. This silent killer often exhibits minimal symptoms in its early stages, making early detection a significant challenge. Delayed diagnosis can lead to severe complications, including cardiovascular disease, anemia, and ultimately, kidney failure.

Project overviews

To develop and optimize a machine learning model capable of accurately classifying individuals as having Chronic Kidney Disease (CKD) based on relevant medical data.

Scope:

The project encompasses the following phases:

- Data acquisition and preprocessing: Collection, cleaning, and preparation of medical datasets for model training.
- Model development: Selection, training, and evaluation of machine learning algorithms for CKD classification.
- Model optimization: Fine-tuning model parameters and hyperparameters to improve performance.
- Performance evaluation: Assessment of the model's accuracy, precision, recall, and other relevant metrics.

Expected Outcomes:

- A well-performing machine learning model for CKD classification.
- Comprehensive documentation of the project methodology and results.
- Insights into the factors influencing CKD development.
- By successfully completing this project, we aim to contribute to early detection and management of CKD, potentially improving patient outcomes

Project Initialization and Planning Phase

Define Problem Statement

The absence of early warning signs coupled with the complexity of traditional diagnostic methods hinders the timely identification of CKD. Current diagnostic procedures are often time-consuming, resource-intensive, and inaccessible to many. This project aims to address these challenges by developing a machine learning-based solution for early CKD detection.

Project Proposal (Proposed Solution)

To develop a robust machine learning model capable of accurately predicting the likelihood of chronic kidney disease based on patient medical data. The model will be integrated into a user-friendly web application to facilitate accessible and early-stage CKD screening.

By providing a rapid and reliable screening tool, this project seeks to improve early detection rates, enabling timely interventions and potentially reducing the burden of CKD on public health.

Initial Project Planning

The initial project planning phase is crucial for outlining the project's scope, defining key milestones, and allocating necessary resources. This section provides a detailed breakdown of the planning process for the chronic kidney disease (CKD) classification project.

Data Collection and Preprocessing Phase

- Understanding & Loading Data: Gain a comprehensive understanding of the dataset structure and contents.
- Load the data into the working environment for analysis.
- Exploratory Data Analysis (EDA): Perform EDA to uncover patterns, trends, and relationships within the data. Visualize data distributions and correlations.
- Handling Null Values: Identify and address missing values in the dataset using
- appropriate imputation techniques or removing records if necessary.
- Handling Outliers: Detect and manage outliers that may skew the model by applying methods such as z-score, IQR, or transformation techniques.
- Handling Categorical Values: Encode categorical variables into numerical format
- using techniques like one-hot encoding or label encoding.

Model Building

- Training the Model: Train multiple machine learning models, including Decision
- Tree, Random Forest, Logistic Regression, using the preprocessed data.
- Comparing Models: Compare the trained models based on performance metrics
- such as accuracy, precision, recall, and F1 score to identify the most effective
- model.
- Evaluating and Saving the Model: Evaluate the best-performing model using the
- test dataset and save the model for future use.

Web Integration and Deployment

- Building HTML Pages: Develop the front-end interface of the web application,
- including pages for home, about, prediction input, and results.
- Local Deployment:
  Deploy the web application locally to test its functionality and ensure smooth integration with the predictive model.

3. Data Collection and Preprocessing Phase

Data Collection Plan and Raw Data Sources Identified

Data Collection Plan :

Extract medical records from hospital databases which contains

Blood cell records , hemoglobin levels , whether the patient suffers

from pedal edema or not

Data Sources :

The raw data source for this project is obtained from Kaggle, a

popular platform for data science competitions and repositories,

The dataset is accessible at Kaggle CKD analysis , it has varialbes such

as age , hemoglobin levels, whilte blood cell count which are critical

for prediction of Chronic Kidney Disease in patients

- Location/URL: Kaggle CKD analysis
- Format: CSV

- Size: Approximately: 10kb
- Access Permissions: Public

3.2. Data Quality Report

- Missing values in all 24 columns except target variable ‘class’
- Categorical data in the dataset in string

3.3. Data Exploration and Preprocessing

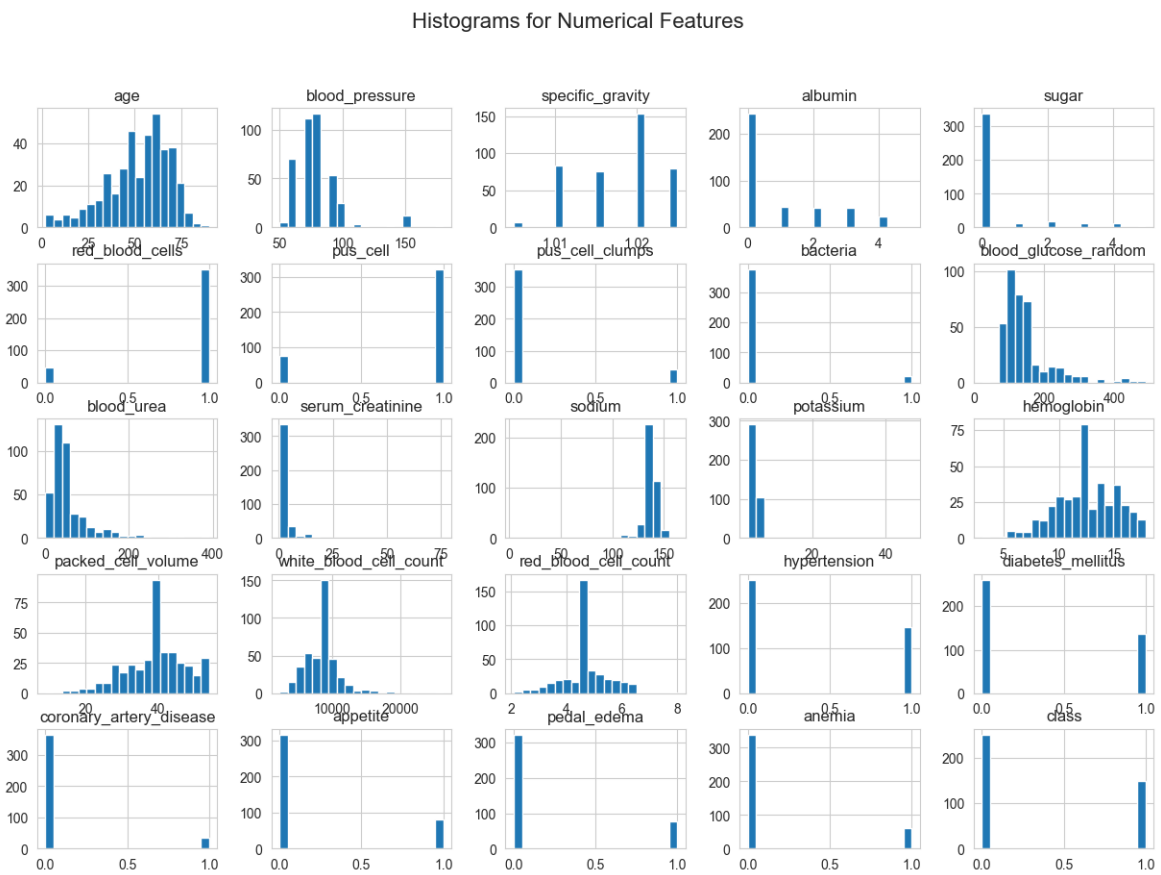
Data Overview:

Dimensions: 400 rows × 26 columns

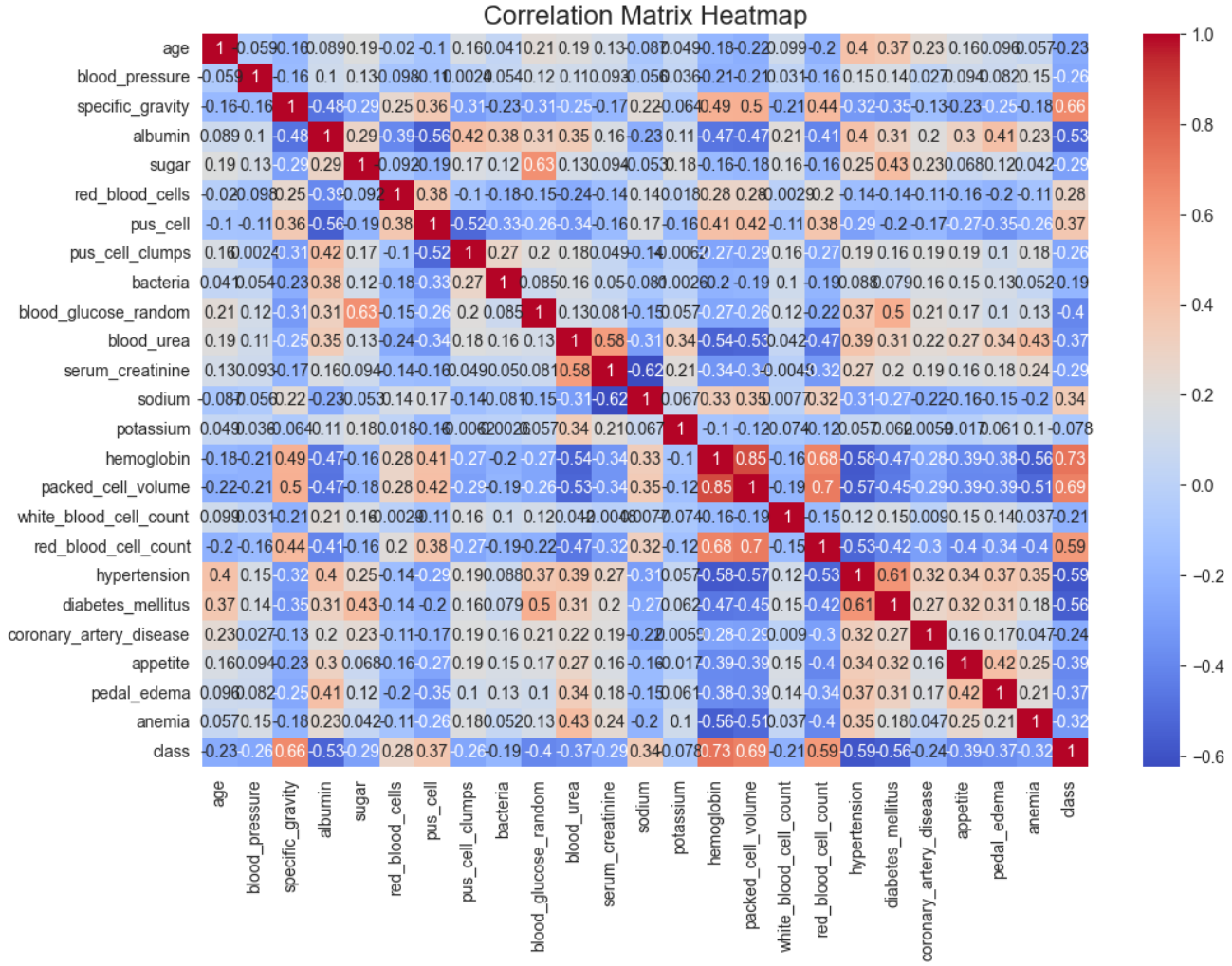
	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

400 rows × 26 columns

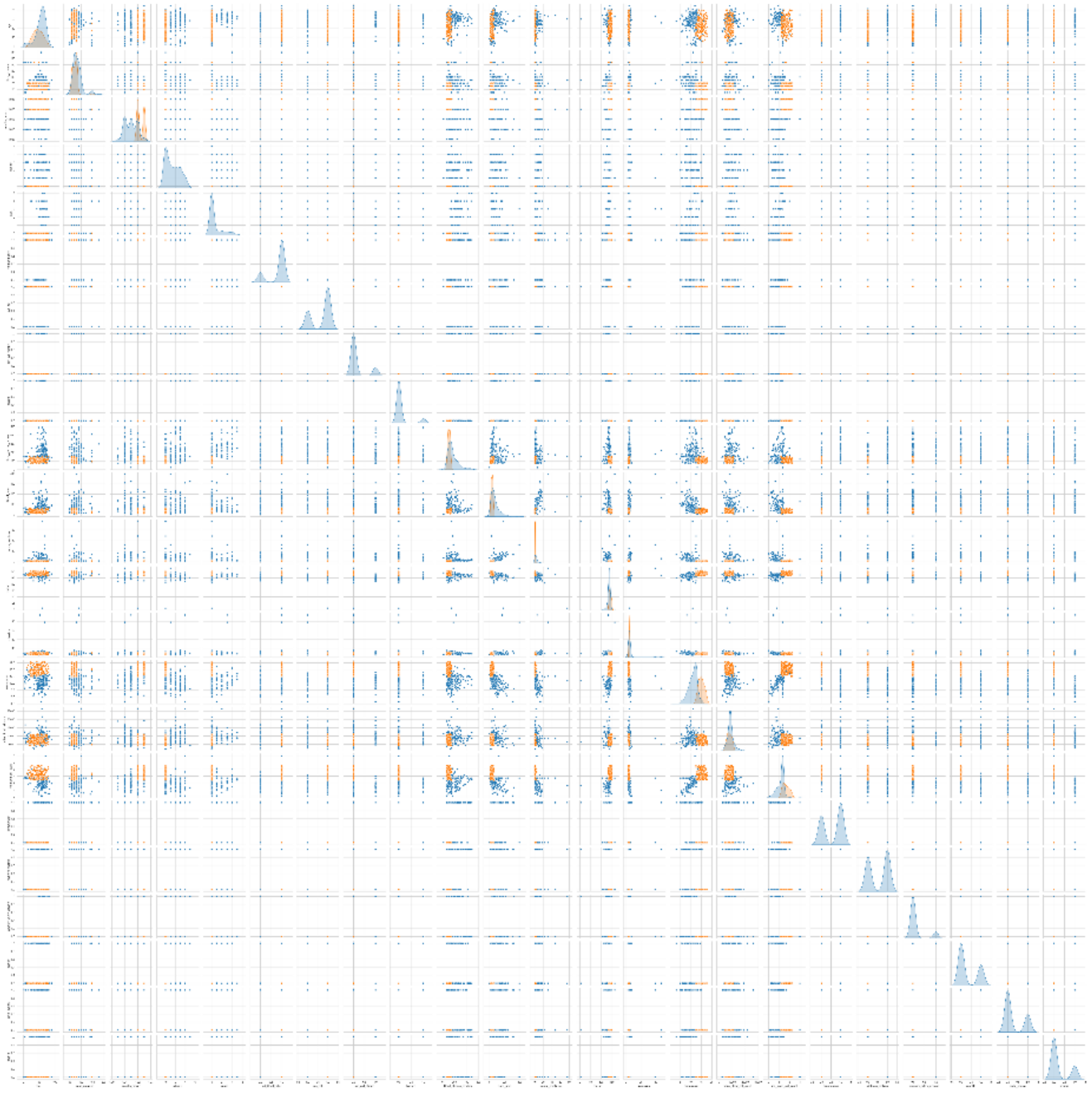
Univariate Analysis:



Bivariate Analysis:



Multivariate Analysis:



Loading Data:

```
#loading the dataset
data = pd.read_csv('chronickidneydisease.csv')
data
```

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

400 rows × 26 columns

Handling Missing Data:

```
data['blood_glucose_random'].fillna(data['blood_glucose_random'].mean(),inplace=True)
data['blood_pressure'].fillna(data['blood_glucose_random'].mean(),inplace=True)
data['blood_urea'].fillna(data['blood_urea'].mean(),inplace=True)
data['hemoglobin'].fillna(data['hemoglobin'].mean(),inplace=True)
data['packed_cell_volume'].fillna(data['packed_cell_volume'].mean(),inplace=True)
data['potassium'].fillna(data['potassium'].mean(),inplace=True)
data['red_blood_cell_count'].fillna(data['red_blood_cell_count'].mean(),inplace=True)
data['serum_creatinine'].fillna(data['serum_creatinine'].mean(),inplace=True)
data['sodium'].fillna(data['sodium'].mean(),inplace=True)
data['white_blood_cell_count'].fillna(data['white_blood_cell_count'].mean(),inplace=True)
```

✓ 0.0s Python

```
data['age'].fillna(data['age'].mode()[0],inplace=True)
data['hypertension'].fillna(data['hypertension'].mode()[0],inplace=True)
data['pus_cell_clumps'].fillna(data['pus_cell_clumps'].mode()[0],inplace=True)
data['appetite'].fillna(data['appetite'].mode()[0],inplace=True)
data['albumin'].fillna(data['albumin'].mode()[0],inplace=True)
data['pus_cell'].fillna(data['pus_cell'].mode()[0],inplace=True)
data['red_blood_cells'].fillna(data['red_blood_cells'].mode()[0],inplace=True)
data['bacteria'].fillna(data['bacteria'].mode()[0],inplace=True)
data['anemia'].fillna(data['anemia'].mode()[0],inplace=True)
data['sugar'].fillna(data['sugar'].mode()[0],inplace=True)
data['diabetes_mellitus'].fillna(data['diabetes_mellitus'].mode()[0],inplace=True)
data['pedal_edema'].fillna(data['pedal_edema'].mode()[0],inplace=True)
data[ 'specific_gravity'].fillna(data[ 'specific_gravity'].mode()[0],inplace=True)
```

✓ 0.0s Python

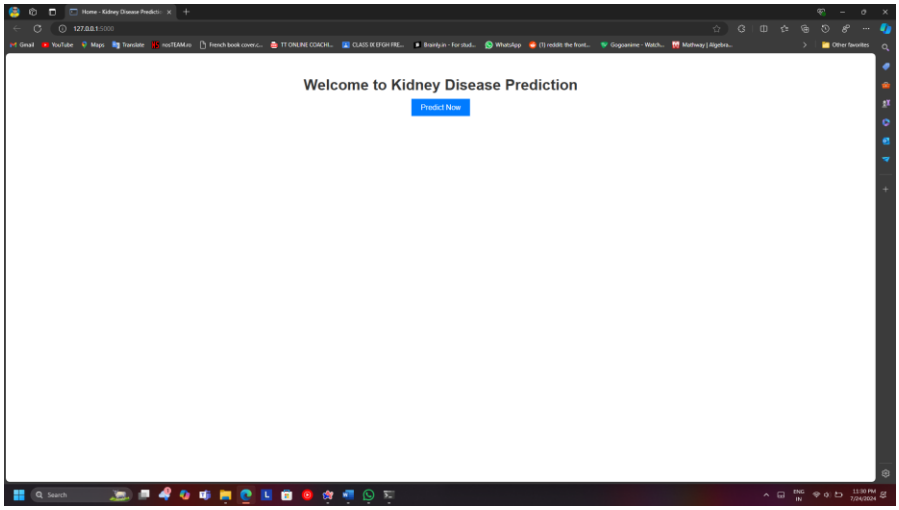
4. Model development Phase:

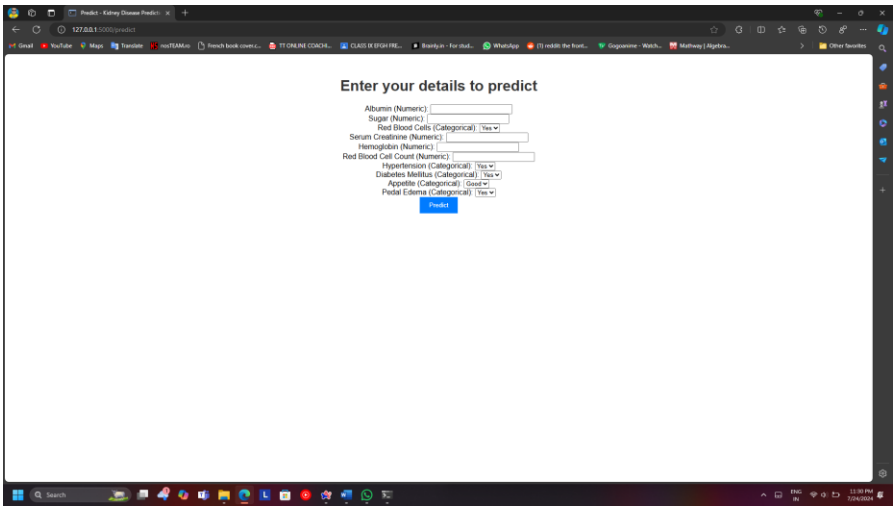
Feature Selection:

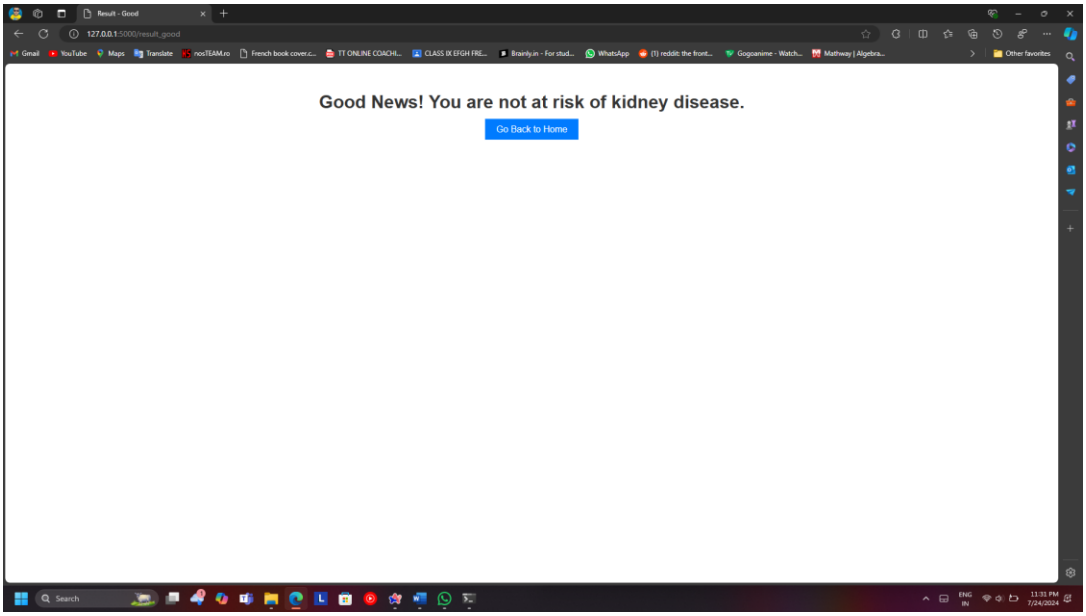
Feature	Description	Selected (Yes/No)	Reasoning
id	Patient id	No	**Irrelevant to the Project**
age	Age of the patient	No	**While age is a general risk factor, it is not as specific or direct an indicator of CKD as some biochemical markers.**
blood pressure	Blood pressure measurement.	No	**CKD can also be caused by other factors such as genetic disorders, glomerulonephritis, infections, or exposure to toxins. In such cases, blood pressure might not be the primary or initial indicator of kidney damage**
Specific gravity	Measure of urine density.	**No**	**This measure of urine concentration can be affected by hydration status and other factors, making it less specific for CKD.**
albumin	Protein levels in urine.	**yes**	**Proteinuria (high levels of albumin in urine) is a key marker for kidney disease.**
sugar	Sugar levels in urine.	**yes**	**Glycosuria (presence of sugar in urine) is primarily associated with diabetes mellitus**
Red blood cells	Red blood cell count.	**yes**	**The presence of red blood cells in urine (hematuria) can indicate underlying kidney issues**
Pus cell	Presence of pus cells in urine.	**no**	**Presence of pus cells might be more indicative of infections rather than CKD.**
Pus cell clumps	Presence of clumps of pus cells.	**no**	**Similar to pus cells, pus cell clumps may indicate urinary tract infections rather than chronic kidney issues.**
bacteria	Presence of bacteria in urine.	**no**	**Presence of bacteria in urine is usually a sign of infection and not directly indicative of CKD.**
Blood glucose random	Random blood glucose level.	**no**	**While it indicates diabetes, it does not directly measure kidney function or damage.**
Blood urea	Level of urea in blood.	**no**	**It is a secondary consequence of reduced kidney function, not a direct measure of the kidney's structural integrity.**

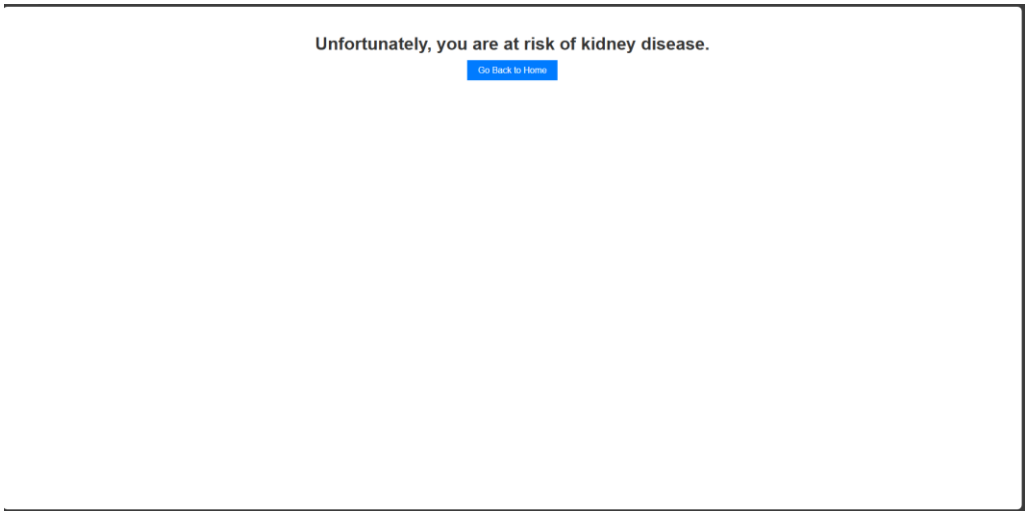
```
Logistic Regression Accuracy: 0.975
Logistic Regression F1 Score: 0.9752060439560438
Logistic Regression Confusion Matrix:
[[51  2]
 [ 0 27]]

Naive Bayes Accuracy: 1.0
Naive Bayes F1 Score: 1.0
Naive Bayes Confusion Matrix:
[[53  0]
 [ 0 27]]

Random Forest Accuracy: 1.0
Random Forest F1 Score: 1.0
Random Forest Confusion Matrix:
[[53  0]
 [ 0 27]]
```

Website Screenshots:




Predict - Kidney Disease Predict
127.0.0.1:5000/predict
Enter your details to predict
Albumin (Numeric):
Sugar (Numeric):
Red Blood Cells (Categorical): Yes
Serum Creatinine (Numeric):
Hemoglobin (Numeric):
Red Blood Cell Count (Numeric):
Hypertension (Categorical): Yes
Diabetes Mellitus (Categorical): Yes
Appetite (Categorical): Good
Pedal Edema (Categorical): Yes
Predict


Result - Good
127.0.0.1:5000/result_good
Good News! You are not at risk of kidney disease.
Go Back to Home



7. Advantages & Disadvantages

Advantages

- **Early detection:** Enables early identification of CKD, allowing for timely interventions and prevention of complications.
- **Improved patient outcomes:** Early treatment can significantly improve the quality of life and longevity of patients with CKD.
- **Reduced healthcare costs:** Early intervention can potentially reduce the overall cost of managing CKD.
- **Efficient screening:** Provides a rapid and non-invasive method for CKD screening.

Disadvantages

- **Model limitations:** The model's accuracy may vary depending on the quality and completeness of the input data.
- **False positives and negatives:** The model may produce incorrect predictions, leading to unnecessary anxiety or delayed diagnosis.
- **Data privacy concerns:** Handling sensitive patient data requires strict adherence to privacy regulations.
- **Model maintenance:** The model may require periodic updates to maintain accuracy as medical knowledge and data evolve.

8. Conclusion

This project successfully developed a machine learning model capable of predicting the likelihood of Chronic Kidney Disease (CKD). The model, when integrated into a user-friendly web application, has the potential to significantly improve early

detection rates and patient outcomes. By providing a rapid and accessible screening tool, this project contributes to better CKD management and prevention.

9. Future Scope

- **Expand dataset:** Incorporate additional medical parameters and patient demographics to enhance model performance.
- **Incorporate longitudinal data:** Utilize patient data over time to track disease progression and refine predictions.
- **Develop risk stratification models:** Create models to predict the risk of CKD progression to different stages.
- **Integration with electronic health records (EHR):** Seamlessly integrate the model into EHR systems for efficient clinical workflow.
- **Mobile application development:** Create a mobile app for convenient CKD risk assessment.
- **Explainable AI:** Develop techniques to understand the model's decision-making process and improve transparency.

10. Appendix

Source Code

GitHub & Project Demo:

https://github.com/YK1218/Chronic-kidney-disease