CKD

Chronic Kidney Disease — Exploratory Data Analysis

Summary of dataset, clinical findings, and model insights

Student Name: Muhammad Maheem
Class: BSAI-3A
Registration No: SU92-BSAIM-F24-054

50+ Features

Random Forest Model

1. Dataset Overview

This analysis examines patient demographics, clinical measurements, lifestyle factors, and their relationship with CKD diagnosis.

Total Records

—

Replace with dataset count

Total Features

50+

Includes demographics, labs, meds

Target Variable

Diagnosis (0/1)

0 = No CKD, 1 = CKD

1.2 Selected Model Features

The predictive model uses 12 critical clinical indicators.

Age
Years

BMI
Body Mass Index

SystolicBP
mmHg

DiastolicBP
mmHg

FastingBloodSugar
mg/dL

HbA1c
%

SerumCreatinine
mg/dL

BUNLevels
mg/dL

GFR
mL/min/1.73m²

ProteinInUrine
Proteinuria

HemoglobinLevels
g/dL

CholesterolTotal
mg/dL

2. Data Quality Assessment

2.1 Missing Values

Rows with missing values were removed during preprocessing to ensure model reliability. Non-predictive IDs were dropped.

2.2 Data Types

Type	Examples
Numeric	BP, Creatinine, GFR, Cholesterol
Categorical	Gender, Smoking, Education
Target	Diagnosis (0/1)

3. Univariate Analysis

3.1 Target Distribution

Visualize CKD vs No CKD to check class balance. Consider stratified sampling or resampling if imbalanced.

3.2 Age Distribution

Age strongly correlates with CKD — higher prevalence among 60+.

3.3 BMI

Higher BMI categories (overweight/obese) show greater CKD risk.

3.4 Blood Pressure

Hypertension is both a cause and a consequence of CKD; BP control is essential.

3.5 Kidney Function Indicators

Serum creatinine, BUN and GFR are primary kidney function markers with established clinical ranges.

4. Bivariate Analysis

Examining pairwise relationships between features and CKD.

- **Positive Correlations:** Creatinine, BUN, ProteinInUrine, BP, Age, HbA1c
- **Negative Correlations:** GFR, Hemoglobin
- **Blood sugar:** FastingBloodSugar and HbA1c strongly associate with CKD via diabetes
- **Cardio-renal link:** Higher BP relates to lower GFR and higher creatinine

5. Multivariate Analysis

Random Forest feature importance highlights top predictors and risk profiles.

Rank	Feature	Clinical Role
1	GFR	Direct kidney function
2	Serum Creatinine	Filtration indicator
3	ProteinInUrine	Damage marker
4	Age	Primary risk factor
5	BUNLevels	Waste accumulation

Rank	Feature	Clinical Role
6	HemoglobinLevels	Anemia in CKD
7	Blood Pressure	Cardio-renal link
8	HbA1c	Diabetes marker
9	BMI	Obesity-related risk
10	CholesterolTotal	Cardiovascular health

Risk Profiles

High-risk: elderly with diabetes + hypertension and abnormal kidney markers. Protective: normal GFR, controlled BP, healthy BMI.

6. Statistical Insights

Key distribution shapes: creatinine, BUN and proteinuria are right-skewed; age/BMI/BP are near-normal. Outliers must be clinically reviewed.

Outliers to investigate: creatinine > 4.0 mg/dL

Very low GFR < 15 (Stage 5)

7. Clinical Patterns & Insights

Progression Markers

- Early: microalbuminuria, slight creatinine rise, GFR 60-89
- Advanced: creatinine >2.0, GFR <30, anemia

Comorbidities

Common pairing of Diabetes + Hypertension + CKD; cardiovascular disease and metabolic syndrome are frequent.

8. Model Performance Context

Random Forest with StandardScaler for feature scaling. Typical setup: 80/20 train-test split and evaluation using accuracy, precision, recall, F1.

Step	Notes
Scaling	StandardScaler to normalize ranges

Step	Notes
Model	Random Forest — robust to outliers & non-linearities
Validation	Train-test split (80/20), consider cross-validation

9. Key Findings Summary

- **Primary:** GFR and creatinine strongest predictors.
- **Age:** major non-modifiable risk.
- **Diabetes & Hypertension:** clear contributors (HbA1c, BP).
- **Proteinuria:** early marker—monitor closely.

10. Recommendations

Clinical

Screen patients aged 60+, monitor diabetics/hypertensives, refer to nephrology when GFR < 60.

For Modeling & Data

Collect longitudinal data; include albumin & genetic markers; address class imbalance and perform external validation.

11. Limitations

- Cross-sectional snapshot—no progression analysis.
- Selection bias—clinical population may not generalize.
- 12-feature model might miss interactions.
- Potential class imbalance; needs handling.

12. Conclusion

CKD prediction is driven by direct kidney function measures (GFR, creatinine) plus age, diabetes, and hypertension. Early detection and management of modifiable risks can improve outcomes.

Prepared by: Data Science Team — CKD EDA Report

Tip: Replace KPI values and add interactive charts (Plotly/Chart.js) for richer presentation.

Contents