

Descriptive and Inferential Analysis of Chronic Kidney Disease Using Power BI and Python

1.0 Introduction

This project combines Power BI and Python to analyze risk factors associated with chronic kidney disease (CKD). I developed an interactive Power BI dashboard for descriptive analysis and applied logistic regression to identify variables significantly associated with CKD. The project demonstrates healthcare data visualization, statistical modeling, and applied health informatics analytics using a synthetic clinical dataset. The project was completed in two parts:

- Development of an interactive Power BI dashboard for descriptive analysis of CKD risk factors.
- Application of logistic regression using Python to identify variables significantly associated with CKD.

2.0 Part 1: Development of an Interactive Power BI Dashboard

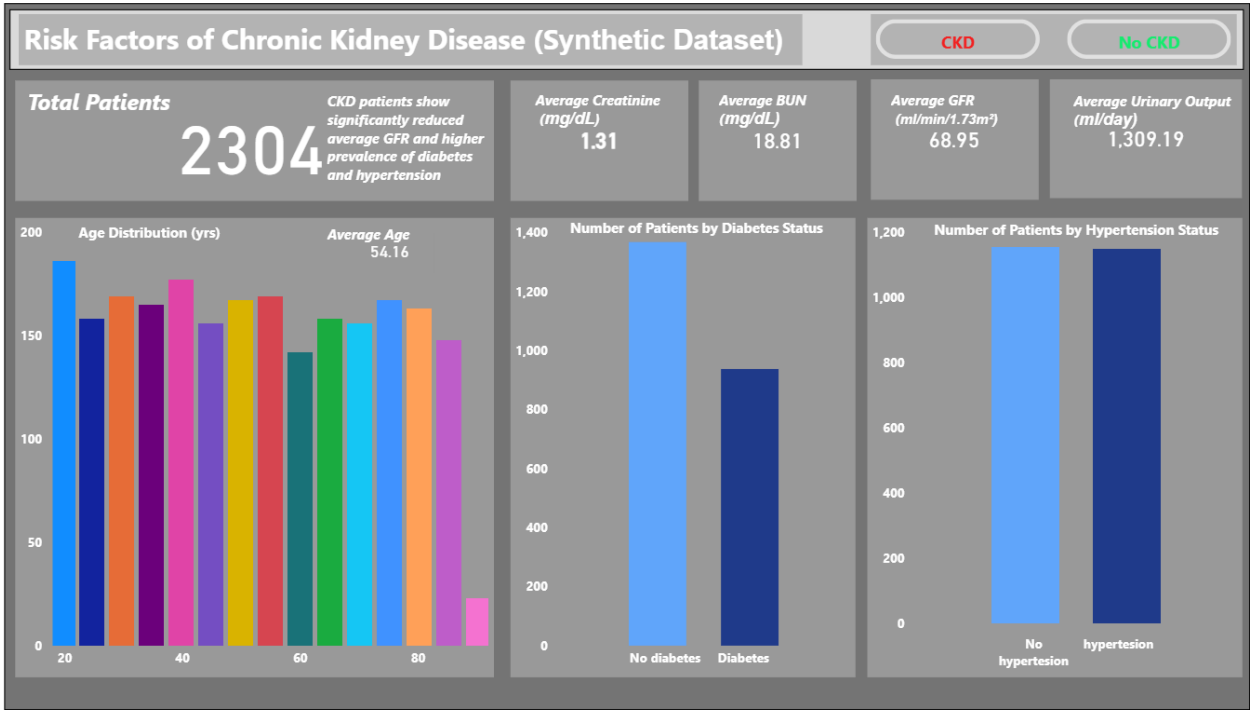
This part of the project shows an interactive Power BI dashboard created to explore and explain risk factors for CKD using a synthetic clinical dataset. The dashboard allows users to easily compare patients with CKD and those without CKD, helping them see differences in kidney function measures and related health conditions through simple visuals.

2.1 Key indicators on the dashboard include:

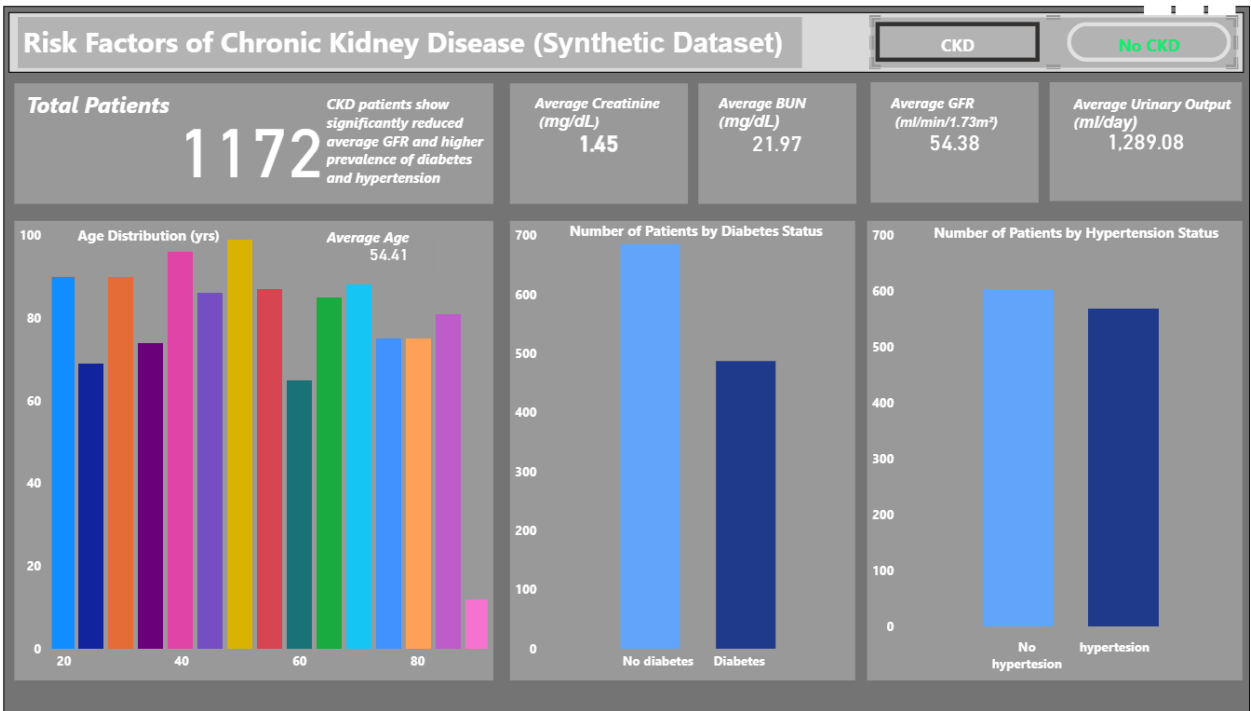
1. Total of patients
2. Average age
3. Average GFR
4. Creatinine
5. BUN,
6. Urinary output
7. Total patients by diabetes status
8. Total patients by hypertension status

Bar charts and charts by category highlight how common diabetes and hypertension are among patients and how these conditions are linked to CKD.

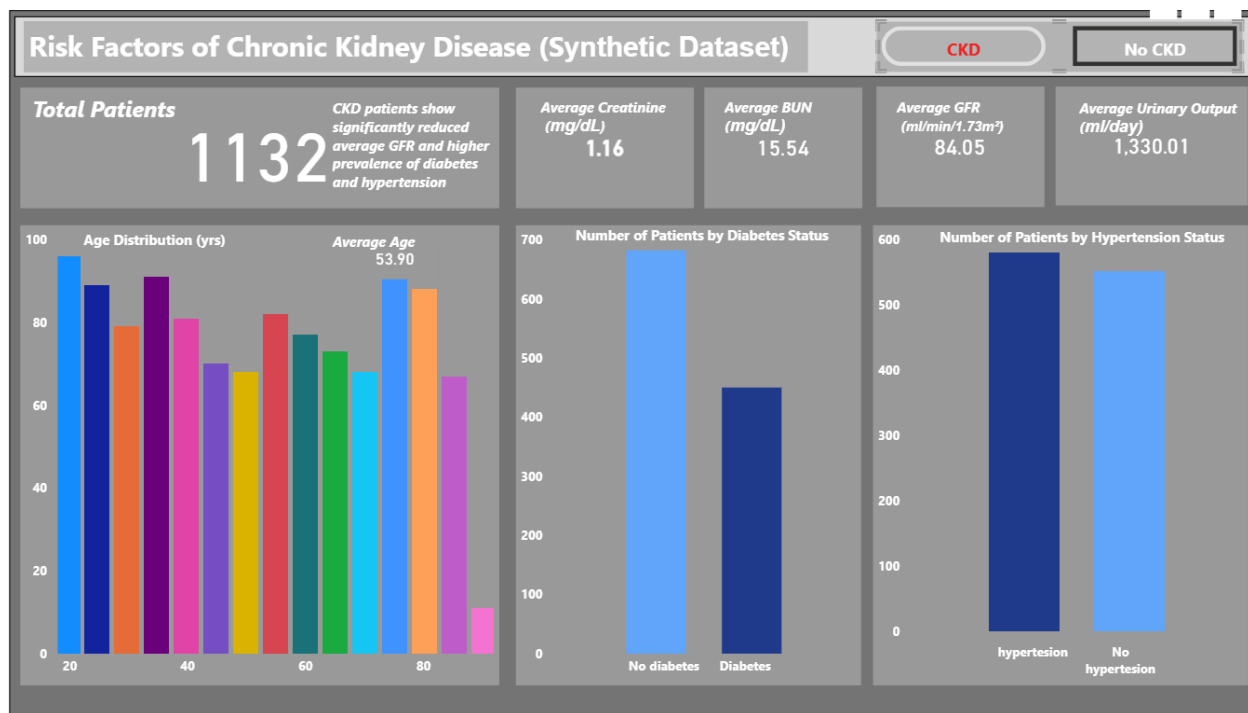
Summary (CKD + non-CKD)



CKD Summary:



Non-CKD Summary



3.0 Part 2: Application of logistic regression using Python

3.1 Methodology

A supervised machine learning framework was employed to develop a logistic regression model to predict the likelihood of chronic kidney disease (CKD). The dataset was randomly partitioned into an 80% training set and a 20% testing set using stratified sampling to preserve outcome proportions. To prevent data leakage, all model development procedures, including cross-validation and coefficient estimation, were conducted exclusively on the training data.

Model stability and generalizability were assessed using 5-fold cross-validation on the training set, with performance evaluated using the area under the receiver operating characteristic curve (AUC) and classification accuracy. Following cross-validation, a final model was fitted using the complete training dataset. Statistical significance of model coefficients was assessed at a predefined alpha level (0.05).

In addition to discrimination metrics, model calibration was evaluated using a reliability (calibration) curve and the Brier score to assess the agreement between predicted probabilities and observed outcomes. Model calibration, predicted probability distributions, and the Brier score were evaluated using the held-out 20% test set to assess probability accuracy on previously unseen data.

3.2 Model Results

3.2.1 Model Fitness

The logistic regression model fitted on the training data demonstrates strong overall explanatory power and statistically significant improvement over the null model. The model converged successfully and included 1,843 observations, with a pseudo-R-squared of 0.487, indicating that nearly half of the variability in CKD status is explained by the included predictors. The likelihood ratio test strongly rejected the null model (LLR p-value < 0.001), confirming that the predictors collectively provide meaningful discriminatory information for CKD status.

3.2.2 Association between Variables and CKD

Variables with p-values greater than 0.05 were considered not to have statistically significant associations with chronic kidney disease (CKD), indicating failure to reject the null hypothesis that their regression coefficients are equal to zero. As shown in Table 1, age, diabetes status, and hypertension status were not significantly associated with CKD in the adjusted model. This lack of statistical significance is further supported by the 95% confidence intervals of their odds ratios, all of which include the null value of one, indicating no independent association with CKD.

In contrast, variables with p-values less than 0.05 were considered to have statistically significant associations with CKD, leading to rejection of the null hypothesis that their coefficients equal zero. Serum creatinine, blood urea nitrogen (BUN), glomerular filtration rate (GFR), and urine output demonstrated significant associations with CKD. The estimated odds ratios indicate that creatinine level and BUN are the strongest risk-increasing factors: each one-unit increase in creatinine was associated with approximately a 2.6-fold increase in the odds of CKD, while each unit increase in BUN corresponded to an approximately 14% increase in CKD odds, holding other variables constant. In contrast, GFR exhibited a strong protective association, with each unit increase reducing the odds of CKD by approximately 9%, reflecting improved renal function. Urine output also showed a modest but statistically significant protective effect, with higher urine output associated with lower odds of CKD.

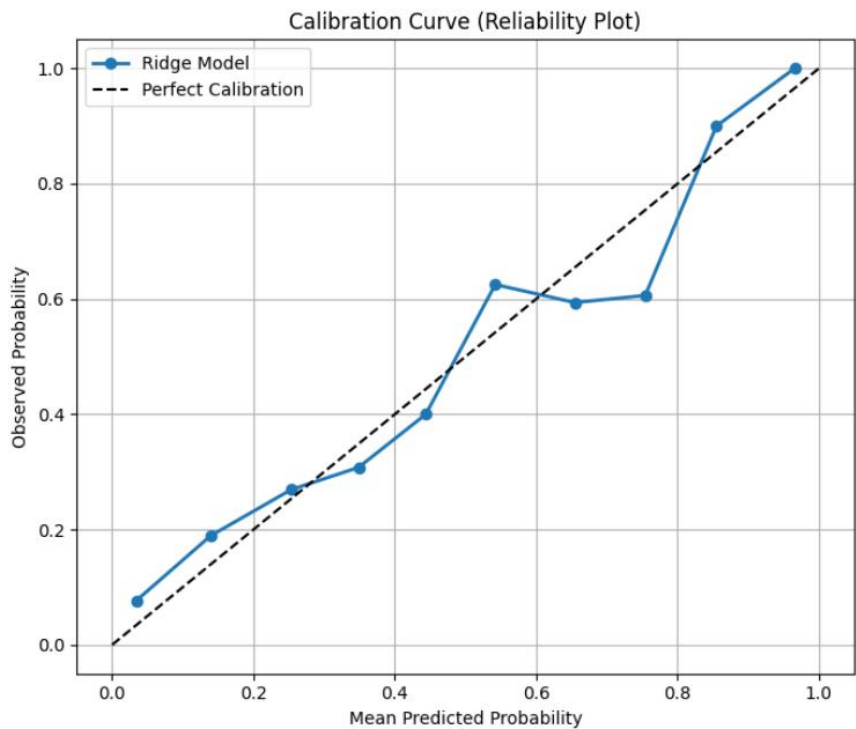
Table 1: Model Summary

Variables	Beta	Std_Error	z_value	p_value	Odds_Ratio (OR)	OR_CI_Lower	OR_CI_Upper
const	4.0353	0.4178	9.6581	0.000	56.5587	24.9378	128.2747
Age	0.001	0.0034	0.2918	0.770	1.0010	0.9943	1.0077
Creatinine_Level	0.944	0.0941	10.0289	0.000	2.5703	2.1373	3.0910
BUN	0.1289	0.0084	15.2840	0.000	1.1376	1.1190	1.1566
Diabetes	0.0476	0.1409	0.3376	0.736	1.0487	0.7957	1.3822
Hypertension	-0.1459	0.1388	-1.0512	0.293	0.8643	0.6584	1.1344
GFR	-0.104	0.0051	-20.4812	0.000	0.9012	0.8923	0.9102
Urine_Output	-0.0003	0.0001	-2.4139	0.016	0.9997	0.9994	0.9999

Total observation 1843, pseudo-R-squared: 0.47, Log-Likelihood (LL): -655.89, LLR p-value: 4.251e-264

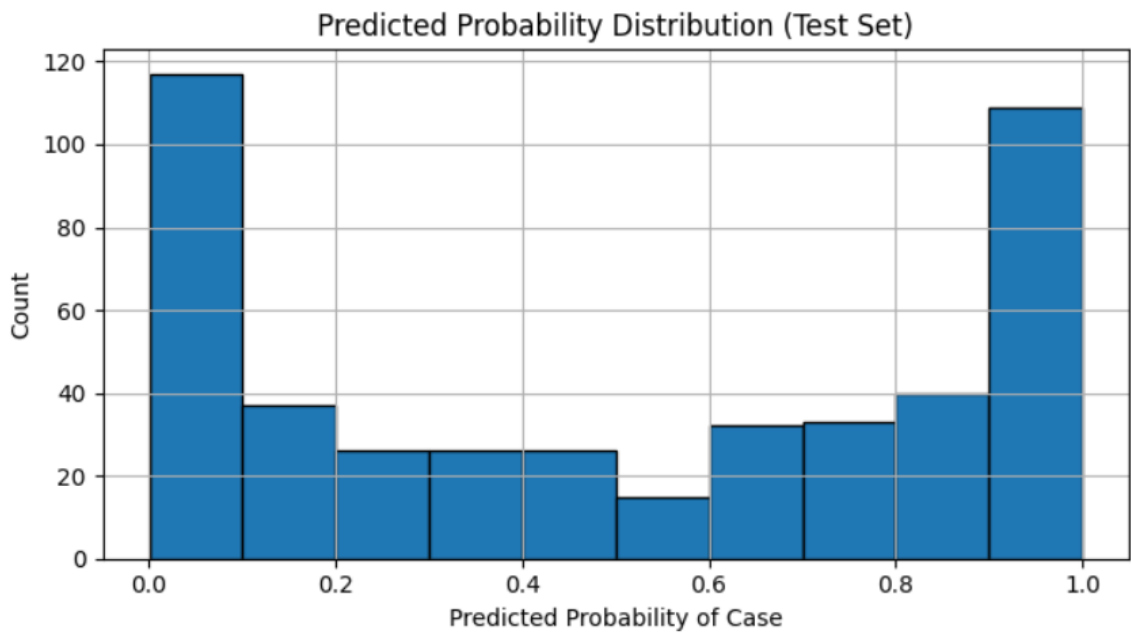
2.2.3 Calibration Curve

The calibration curve shows good agreement between predicted probabilities and observed CKD outcomes across most risk levels, with the model closely following the line of perfect calibration. Minor deviations are observed in the mid-probability range, suggesting slight under- or over-estimation in some bins, likely due to limited sample size. At higher predicted probabilities, the model remains well calibrated, indicating reliable estimation of high CKD risk. Overall, the results suggest that the model provides reasonably accurate and clinically meaningful probability predictions.



2.2.4 Predicted Probability Distribution

The predicted probability distribution shows a bimodal pattern, with most observations clustered at low and high probability ranges. This indicate that the model confidently distinguishes between CKD and non-CKD cases. Relatively fewer predictions fall in the mid-probability range, suggesting good separation of risk groups. The Brier score of 0.1204 indicates good to acceptable calibration, reflecting reasonably accurate probability estimates. Overall, the distribution supports strong discriminatory performance with clinically meaningful risk stratification.



Brier Score: 0.1204
Interpretation: Good/acceptable calibration.