Identifying Patients at a High-Risk of Chronic Kidney Disease

Group 5

Ramon Gonzalez and Alanis Perez

Master of Science in Applied Data Science, University of San Diego

ADS-502B: Applied Data Mining

August 11, 2025

Abstract

Chronic kidney disease (CKD) is a progressive medical condition characterised by kidney damage that may lead to more severe complications if not treated early. The purpose of this project is to classify patients as high- or low/moderate-risk of CKD using a synthetic dataset obtained from Kaggle created for machine learning research and practice. This clean dataset did not contain missing values or inconsistency, allowing for smooth data analysis and modeling. The two primary classification models applied are logistic regression, to estimate the probability of CKD, and random forest, to identify key predictors. These models were evaluated by looking at evaluation metrics (accuracy, precision, recall), ROC curves, and confusion matrices. Analyses of feature importance and coefficients were run to interpret results. An initial accuracy of 100% for random forest provoked a sanity check. Correlation analysis revealed strong correlations with GFR which was removed to make for a more realistic and challenging model. The results demonstrate the robustness of these models in identifying high-risk for CKD to pave the way for preventative healthcare interventions.

Keywords: Chronic kidney disease, high-risk classification, risk prediction, logistic regression, random forest, feature importance, predictive modeling, machine learning

Introduction

Chronic Kidney Disease (CKD) as defined by the National Kidney Foundation is a progressive disease of the kidneys that develop after damage which inhibits its ability to perform essential functions such as waste removal, maintaining blood pressure, and producing healthy red blood cells (2023). Early detection of kidney disease is imperative to avoid morbidity or mortality, as kidney malfunction can lead to severe complications such as heart disease and high blood pressure. The objective of this project is to develop a classification model that can identify patients at a high risk of CKD based on demographic and clinical data. Identifying a patient that may be high risk can allow for early intervention to reduce the need for painful and stressful procedures like dialysis or transplants. A classification model can help to enable healthcare professionals to extend the quality of care needed for a sick patient before the disease progresses.

The dataset that will be used in this project mimics real-world patient records based on relevant features that are commonly associated with kidney disease and function. To satisfy the objective, logistic regression and random forest will be implemented as primary models as they are well suited for interpretable and strong predictive capabilities. A binary logistic regression model estimates the probability of CKD by modeling the log of the odds for the outcome, reflecting the magnitude and direction (positive or negative) of each feature's influence (IBM, 2025). A random forest classification model will give insight into the predictive capabilities of each feature by determining feature importance, ranking features by their contribution to accuracy (IBM). Depending on classification results and accuracy measures, XGBoost analysis will be conducted to evaluate more complex patterns and relationships that the primary models may not capture.

Method

Data Source

The dataset that was used for this project was obtained from Kaggle and consisted of synthetic “but medically realistic” data created for the purpose of helping data scientists and healthcare professionals approach CKD through machine learning models (Miah, 2025). It consists of 2,304 rows of data and a total of 9 features with two target variables. The features include demographic data and clinical measurements of kidney function. The demographic data consists of age (numeric, measured in years), hypertension (categorical, labeled 1 for patients with high blood pressure, 0 otherwise), and diabetes (categorical, labeled 1 for patients with diabetes, 0 otherwise). The clinical data consists of glomerular filtration rate (numeric, measured in ml/min/1.73m^2), creatinine blood level (numeric, measured in mg/dL), blood urea nitrogen level (numeric, measured in mg/dL), and urine output (numeric, measured in ml of urine produced in a day). The two target variables included in this dataset are CKD status (1 for patients with CKD, 0 otherwise) and dialysis needed (1 for patients that will require dialysis based on CKD progression, 0 otherwise). The column for ‘dialysis needed’ was excluded for this project because the focus was to determine if a patient is considered high risk for CKD and dialysis is aimed at determining post-diagnosis data. Including this variable could have improperly and prematurely influenced the risk level of CKD.

Preprocessing

Preliminary exploratory data analysis was performed to give an overview of variable types and data structure using ‘.info()’ and ‘.describe()’ methods. Due to the cleanliness of the raw dataset, all variables were properly typed, no encoding was necessary, there were no duplicate entries, and there were no missing values, making preprocessing smooth and quick. The dataset was split into a traditional 80/20 split with a stratified split to preserve class balance given the small sample size of the data. During initial model training, however, the random forest classifier resulted in 100% accuracy on the test set, which raised some red flags. A sanity check was performed to check for suspected data leakage. A feature correlation analysis revealed that glomerular filtration rate (GFR) had a significantly higher correlation with CKD status than other variables at 0.60. Running a feature importance analysis confirmed that GFR dominated over the other predictors, making for a trivial classification task. Notably, GFR is one of the primary measurements taken in clinical settings to examine CKD and overall kidney function, so keeping it as a predictor variable leads to an unrealistic exploratory analysis project. For this reason, GFR was dropped from the feature set for modeling.

Model Training

Logistic regression (LR) was used as a baseline model to provide insights on linear relationships between the coefficients of features that measure the effect on the target variable. Random forest (RF) was used to investigate more complex and/or non-linear relationships and provided feature importance scores to identify which variables contributed the most to prediction accuracy. After removing GFR to reduce data leakage and create more realistic performances, both models demonstrated to be robust predictors as evaluated by strong performance metrics and high AUC scores. As an attempt to increase accuracy, an XGBoost analysis was performed. However, it was unsuccessful in outperforming the RF model, even with tuning and the inclusion of cross-validation, indicating that a higher-complexity model was not necessary. XGBoost may have been a more suitable model for a messier or more complex dataset, but given the cleanliness and synthetic nature of this data, it was not helpful. For this reason, further interpretations and the creation of predictions on a positive class and risk labels were used with LR and RF as primary modes.

Results

Model Evaluation

As shown in table 1, LR resulted in a weighted accuracy of predicting the presence of CKD at 0.62, precision of 0.63, and recall of 0.62. The ROC curve for LR is shown in figure 3, indicating a stronger performance than random chance but only at moderate strength of discriminating between class separation with an AUC score of 0.68. RF yielded higher accuracy and stronger performance also shown in table 1, with a weight accuracy of 0.71, precision of 0.77, and recall of 0.57. The ROC curve for RF is shown in figure 4, indicating a stronger performance than LR at a slightly higher strength of discriminating between classes with an AUC score of 0.75. The XGBoost model resulted in an accuracy of 0.70, just 0.1% shy of outperforming RF.

Confusion Matrices

Confusion matrices were created to assess predictive capability and determine the reliability of appropriate classifications. Figure 1 shows a confusion matrix for LR. This model showed strength in correctly predicting positive (146) and negative (142) classes, but missed the mark with a few false positives (84) and false negatives (89). Figure 2 shows a confusion matrix for RF. Unfortunately, some precision was lost as a higher number of false negatives (100) were miscalculated and true positives (135) were weaker. It did, however, excel as calculating true positives (186) and only accounted for a few false positives (40).

Feature Importance

Looking at the RF model, a feature importance barplot was created to visualize which features held the most weight in predicting CKD, shown in figure 6. Blood urea nitrogen level (BUN) was the most important with an importance score of ~0.34, followed by creatinine level at ~0.26, then urine output at ~0.20, and age at ~0.17. However, both hypertension (~0.02) and diabetes (~0.02) suddenly tapered off from the graph indicating little to no importance on determining CKD risk.

Coefficients and Log-Odds

Using the LR model, a coefficient plot gave an overview of the direction of influence by the feature set on the target variable, shown in figure 5. Blood urea nitrogen level (BUN) was unsurprisingly the most influential with a positive log-odd effect score of approximately 0.7. This was followed by creatinine level with a positive log-odd effect score of approximately 0.5. Urine output was the third most influential with a negative effect but at a much lower strength than expected at -0.1. Hypertension was also negatively associated with CKD at a small strength smaller than -0.1. Diabetes (< 0.1) and age (< 0.1) were both miniscule in comparison, indicating little to no influence in the direction of predicting CKD risk. BUN and creatinine level were the strongest influential predictors on determining the risk of CKD.

Classifications and Risk Labels

Using predictions and probabilities on the positive class (CKD = 1) for RF with a threshold of 0.7 allowed for the insertion of a CKD probability column with a probability score of disease on the testing set. This probability score was used to create a risk label for patients which labeled ‘high risk’ for those with a probability score above the threshold and ‘low/moderate risk’ for those below the threshold. The test set consisted of 461 patients (taken from the 20 split of the original set). This threshold determined that 112 out of the 461 patients were at high risk of CKD and 349 were at a low/moderate risk.

Discussion

The results of the preliminary models demonstrate that both were robust in predicting patients at a high risk of CKD, each giving insights into feature influences. Logistic regression offered interpretable feature coefficients, describing direction and strength of influence. Random forest provided an overview of non-linear relationships, feature interactions, and high predictive performance. There were some hiccups with initial testing due to the GFR variable which resulted in unrealistic model performance that does not appropriately reflect real-world clinical measurements and complexities. As a dominant predictor because of its direct correlation in measuring primary kidney function, it was removed to properly mirror a more realistic and challenging model.

Limitations

There are a few limitations of this study. To reiterate, this is a synthetic dataset that does not fully capture medical data, variability, or complexity of clinical measurements and demographic data. The exclusion of a very strong predictor (GFR) reduced model accuracy significantly. This dataset is also relatively small, and although measures were taken to balance sample size and classes, a larger dataset can provide greater understanding.

Next Steps

Replication studies can explore more features, including GFR, and test more complex models. Additionally, shifting the perspective to look at time-series models to evaluate CKD progression over time can advance intervention strategies in healthcare. In the future, these models could be applied to real datasets. Reintroducing a variable such as ‘dialysis_needed’ as found in the original dataset can guide research to develop models that can predict advanced stages of disease that can or will require more intense procedures.

References

IBM. (2025, May 14). *What Is Logistic Regression?* IBM. ibm.com/think/topics/logistic-regression

IBM. *What Is Random Forest?* IBM. ibm.com/think/topics/random-forest

Miah, A. (2025). *Kidney Disease Risk Dataset* (Version 1). [Data set]. Kaggle. kaggle.com/datasets/miadul/kidney-disease-risk-dataset

National Kidney Foundation Patient Education Team. (2023, September 11). *Chronic Kidney Disease (CKD)*. National Kidney Foundation. kidney.org/kidney-topics/chronic-kidney-disease-ckd

Table 1

Model Performance Summary

Model	**Weighted Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.62	0.63	0.62	0.63
Random Forest	0.71	0.77	0.57	0.66
XGBoost	0.70	-	-	-

Figure 1

Confusion Matrix for Logistic Regression

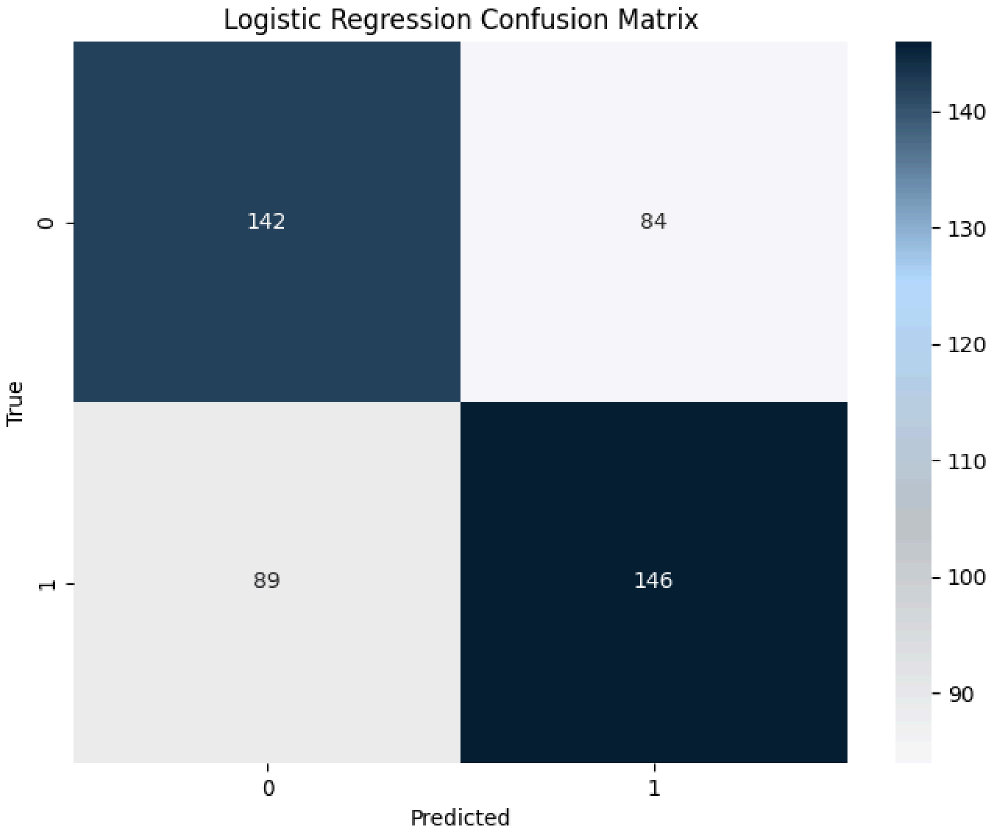


Figure 2

Confusion Matrix for Random Forest

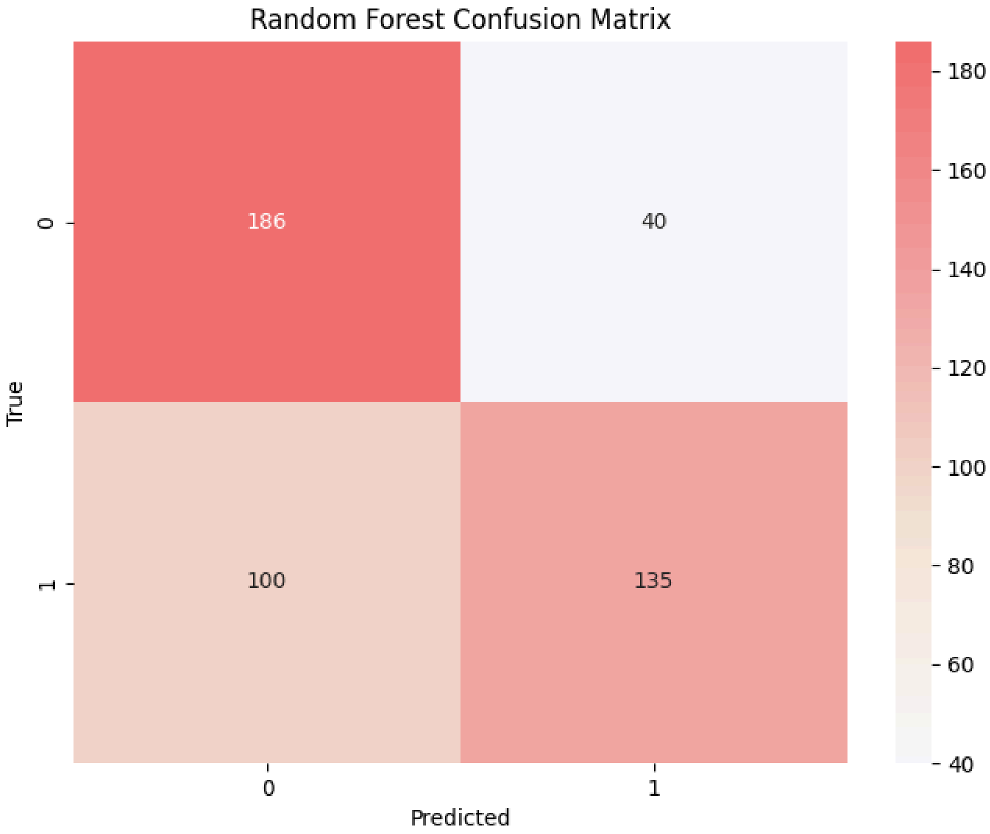


Figure 3

ROC Curve for Logistic Regression

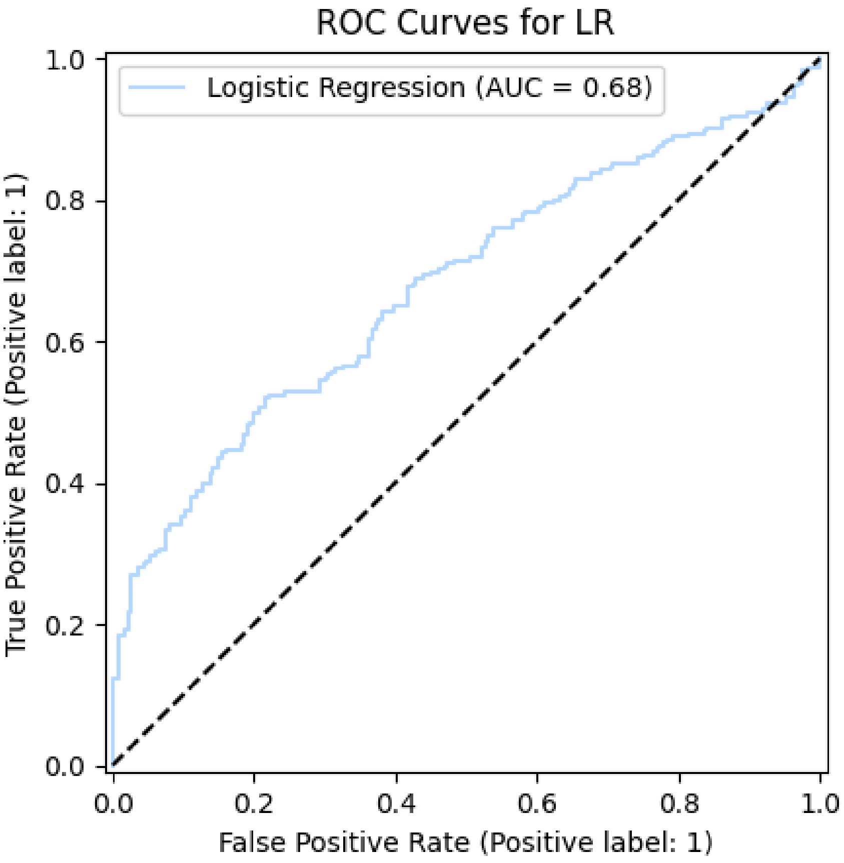


Figure 4

ROC Curve for Random Forest

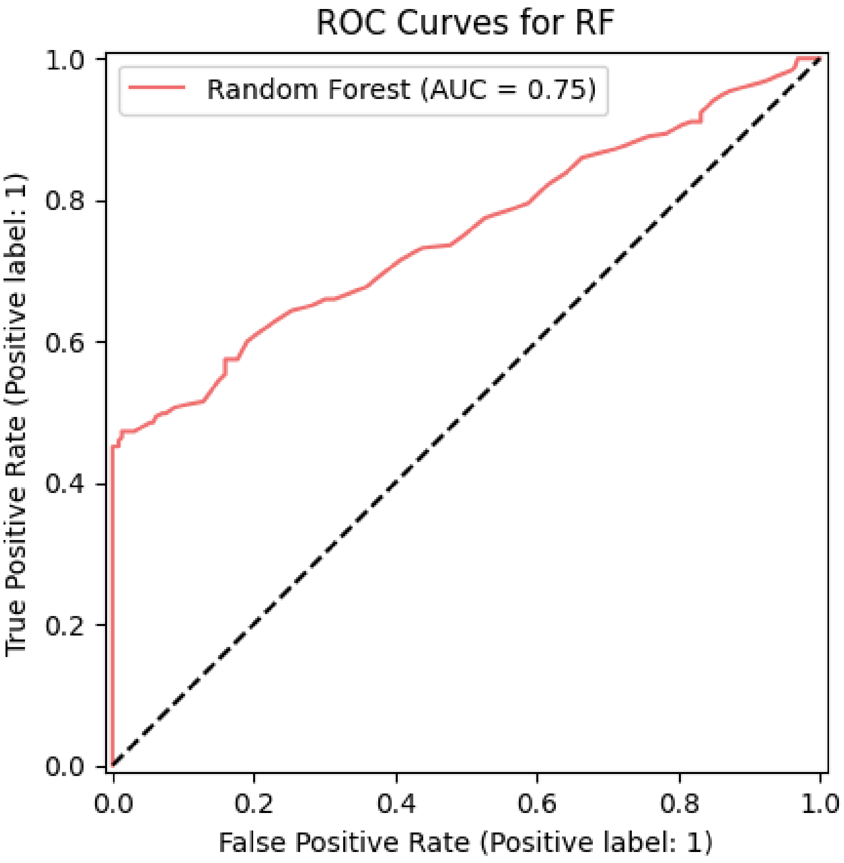


Figure 5

Coefficients of Logistic Regression

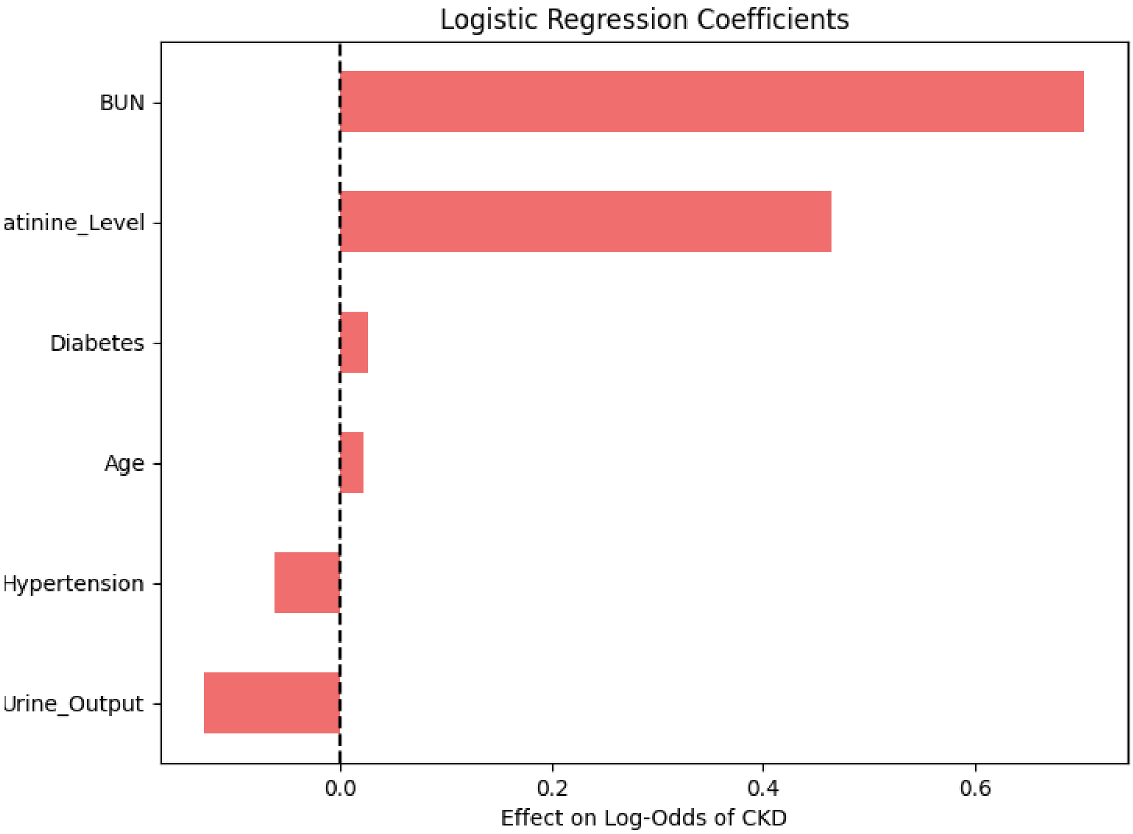
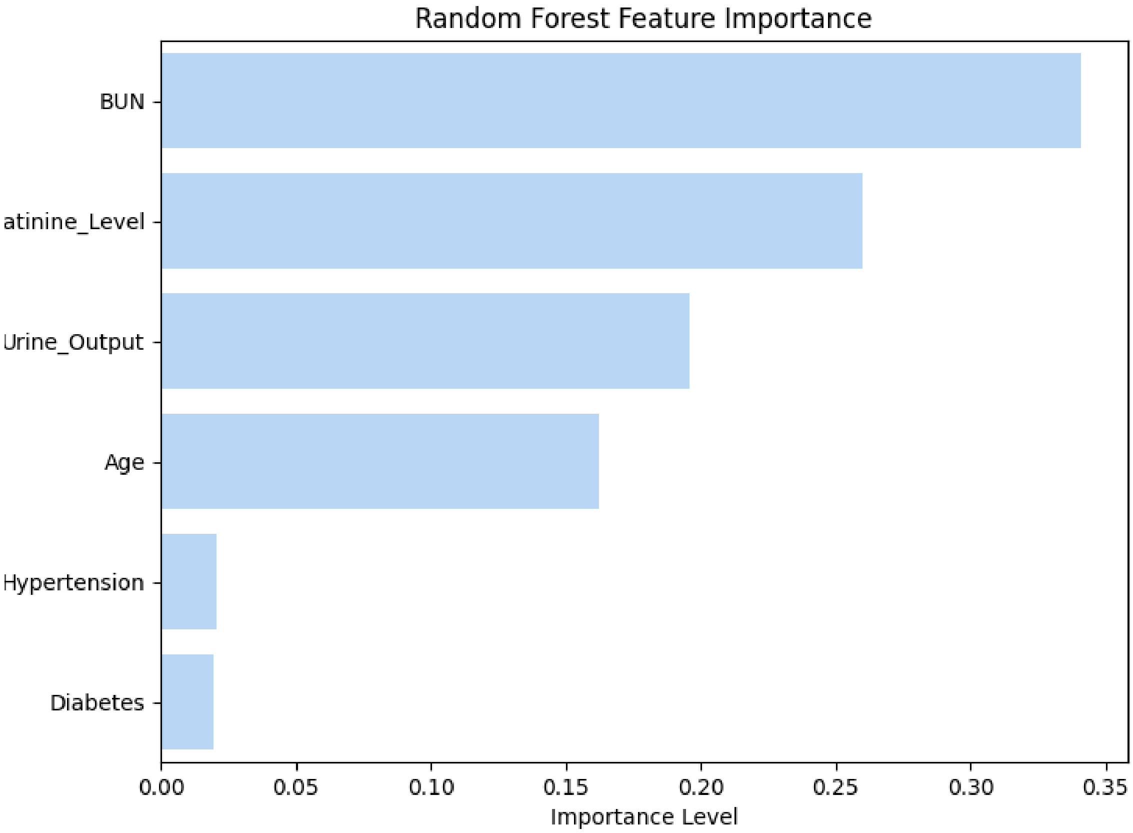


Figure 6

Feature Importance of Random Forest


Random Forest Feature Importance
BUN
atinine_Level
Urine_Output
Age
Hypertension
Diabetes
Importance Level