DS3001 Project: Predicting Chronic Kidney Disease Risk from Clinical and Demographic Features

Aishani Patnaik [1] **Cody Liddle** [1] **Meher Kalsi** [1]

Abstract

Chronic kidney disease (CKD) is a progressive condition that often remains undetected until its later stages, which highlights the need for early risk identification using routinely collected clinical information. The primary objective of this study was to evaluate whether machine learning algorithms can accurately predict CKD status using demographic variables, laboratory measures, and categorical health indicators from the UCI Chronic Kidney Disease Risk Factor dataset. To address this question, we conducted extensive preprocessing that included median and mode imputation, standardization of continuous variables where appropriate, removal of ambiguously defined features, and stratified partitioning to preserve class balance. Using the cleaned dataset, we developed two complementary classification models: logistic regression, selected for its transparency and interpretability, and random forest, chosen for its capacity to capture nonlinear interactions among clinical features. Model training incorporated hyperparameter tuning through 5-fold GridSearchCV, with optimization centered on the F1-score to reduce clinically meaningful misclassifications. Both models demonstrated strong performance on held-out test data. Logistic regression produced results consistent with established CKD physiology, identifying anemia markers, metabolic indicators, and comorbidities such as hypertension and diabetes as important predictors. The random forest model identified similar factors and assigned high importance to hemoglobin, packed cell volume, glomerular filtration rate, and serum creatinine. Although the random forest achieved perfect classification on the test set, this outcome is likely influenced by the small sample size and should be interpreted with caution. Overall, the results indicate that machine learning has meaningful potential for early CKD risk stratification when paired with careful data preparation and clinically informed model design. Key limitations include the limited dataset size, inconsistent variable definitions, and the absence of external validation. Future work should incorporate larger and more diverse samples, continuous laboratory measurements, and more advanced calibration techniques to support real-world deployment. This study demonstrates that both interpretable and flexible algorithms can effectively model CKD risk and provides a foundation for continued development of predictive tools aimed at improving early detection.

1. Data

1.1. Research Question and Dataset

Research question: Can CKD status be predicted from demographic variables, laboratory values, and categorical health indicators, and which features appear most influential for risk?

This question is well-defined because it maps cleanly to a binary classification task (CKD vs. non-CKD). It is compelling and relevant because CKD can be costly and harmful when detected late, and early prediction could support screening and follow-up decisions. CKD is a serious public health concern, and datasets such as this one allow researchers to explore possible predictive markers and risk factors. The dataset contains both demographic and clinical variables, including laboratory test results and health status indicators. These features together can provide insight into the relationships between underlying health conditions, lifestyle-related factors, and kidney function.

The Chronic Kidney Disease (CKD) Risk Factor Prediction Dataset, available through the UCI Machine Learning Repository, was designed to provide information on patient characteristics that may contribute to the presence or progression of chronic kidney disease. The data were collected at Enam Medical College and Hospital, Savar, Dhaka, Bangladesh, and contain 200 observations.

1.2. Key Variables

Several key variables are included in the dataset. Demographic factors such as age, gender, and blood pressure are

recorded. Clinical laboratory measures include serum creatinine, blood urea, blood glucose random (bgr), glomerular filtration rate (gfr), and hemoglobin, all of which are relevant to kidney function and overall health status. In addition, categorical indicators are present, such as diabetes mellitus (`dm`), coronary artery disease (`cad`), anemia (`ane`), and pedal edema (`pe`), which represent medical conditions often comorbid with or contributing to CKD.

The target variable is `class`, indicating CKD vs. non-CKD. During preprocessing, `ckd` is mapped to 1 and `notckd` is mapped to 0.

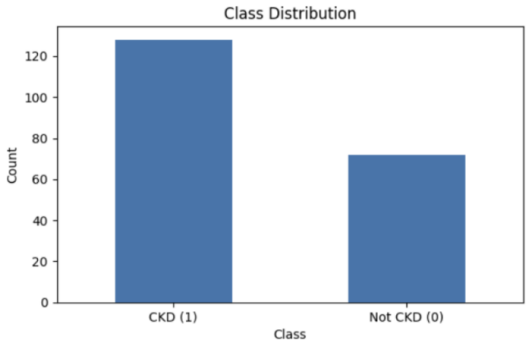

Figure 1. Class distribution in the CKD dataset after label encoding (CKD = 1, non-CKD = 0). The dataset shows moderate class imbalance, motivating the use of metrics beyond accuracy (e.g., F1 and ROC–AUC).

Some variables reflect more specific diagnostic results. For example, pus cell count (`pc`) and pus cell clumps (`pcc`), while not clearly defined in the documentation, likely capture abnormalities in urine samples that could indicate infection or other kidney-related dysfunction. Similarly, bacteria (`ba`) and blood pressure (`bp`) are coded as binary variables for presence or absence. Appetite (`appet`) is included as a subjective clinical measure, coded as *normal* or *poor*.

1.3. Challenges in Reading and Cleaning

A primary challenge with this dataset is the lack of clarity in variable definitions. Several variables use abbreviations (e.g., `pc`, `pcc`, `dm`, `cad`, `pe`, `ane`, `appet`, `gfr`, `ba`) without accompanying detailed explanations. While their likely meanings can be inferred from medical knowledge, this introduces uncertainty and limits confidence in the analysis. Without clear operational definitions, these variables risk being misinterpreted, which could undermine the validity of any predictive model.

Another issue is the misrepresentation of continuous variables as categorical or binary. For instance, "bp limit" is listed as a binary feature, which is problematic given that blood pressure is normally a continuous measure (e.g., systolic/diastolic in mmHg). While the variable could have been derived from an underlying threshold, it is difficult to interpret without clarification. This method of representation removes important variation and limits the ability to use blood pressure as a meaningful risk factor.

Lastly, there are likely missing values and inconsistencies in data entry, which are common in clinical datasets. Variables such as lab test results may not be available for all patients, and categorical indicators may suffer from inconsistencies in coding (e.g., use of 0/1 vs. yes/no or normal/abnormal). Handling these issues requires systematic cleaning, including standardizing categories, imputing or removing missing values, and deciding whether ambiguous features should be excluded entirely. We imputed numeric features with the median and categorical with the mode, fit on the training set only, to avoid leakage. These strategies are simple, reproducible, and reduce sensitivity to extreme values, which is useful in clinical contexts where measurements can have long tails.

Given these challenges, careful preprocessing is essential. Variables with unclear meanings may need to be excluded unless their definitions can be verified by either medical experts or the original data source. For example, while features such as `dm` (diabetes mellitus) and `cad` (coronary artery disease) are likely highly relevant to CKD prediction, excluding them might be safer if their definitions cannot be fully confirmed. Continuous variables such as blood pressure should be carefully reviewed; if the raw measurements are unavailable, the binary form may be of limited utility.

A key preprocessing step is removing `affected` and `stage`. These fields are diagnostic rather than predictive; if included, they can leak label information and artificially inflate performance. Removing them ensures that the evaluation reflects true predictive modeling rather than learning the diagnosis directly.

Outliers are treated as potentially meaningful rather than automatically removed. Extremely high or low lab values may reflect true disease states, so we prioritize robust imputation and careful interpretation instead of aggressive outlier filtering.

Despite these issues, the dataset remains valuable for exploring predictive relationships. Well-defined variables such as serum creatinine, hemoglobin, and blood urea provide clinically valid markers of kidney function. Combined with demographic factors like age, these can form the basis of robust models, while ambiguous features may be set aside or used cautiously in sensitivity analyses.

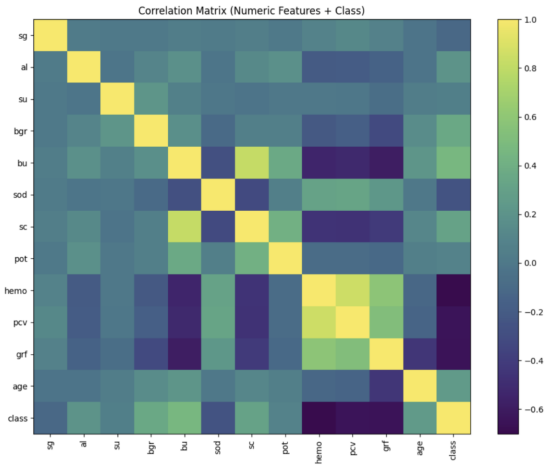

Figure 2. Correlation heatmap for selected numeric features and the target label. Several kidney-function and anemia-related measures show a stronger association with CKD status, motivating their inclusion as key predictive variables.

2. Methods

2.1. Research Question and Strategy

The goal of this project is to develop and evaluate machine learning models capable of predicting the likelihood of chronic kidney disease (CKD) based on demographic, clinical, and laboratory features. The central research question is:

> *Can machine learning algorithms accurately identify individuals at risk for CKD using a combination of demographic and physiological markers?*

The dataset used for this analysis is the Risk Factor Prediction of Chronic Kidney Disease dataset from the UCI Machine Learning Repository. It contains patient-level data, including demographic variables (age, gender), laboratory test results (serum creatinine, blood urea, hemoglobin), and categorical health indicators (diabetes mellitus, coronary artery disease, anemia, pedal edema). The dataset documentation notes challenges with unclear abbreviations and some variables being coded categorically despite being inherently continuous. To ensure reliability, this analysis focuses primarily on features that are well-defined and clinically interpretable, as discussed in the data description.

The overall research strategy involves systematic data cleaning, exploratory data analysis (EDA), and model development using both interpretable and high-performing classification algorithms. Each step emphasizes reproducibility and transparent handling of data quality issues. This project aims to identify which clinical factors are most predictive of CKD, compare models to find the right balance between clinical transparency and predictive performance, and assess how well these models generalize to real-world data. By comparing both interpretable and complex approaches, the project seeks to develop models that are not only accurate but also clinically meaningful, ensuring that predictions are both reliable and understandable. The analysis follows a structured pipeline: data cleaning and preprocessing, EDA, model development, validation, and interpretability.

2.2. Data Wrangling and Preprocessing

Handling Missing and Ambiguous Data. Missing and ambiguous data are handled through targeted imputation and selective exclusion to preserve data integrity while ensuring that only clinically interpretable variables contribute to model training. For continuous laboratory measures such as serum creatinine, blood urea, and hemoglobin, missing values are replaced with the median value of each variable to minimize outlier influence. For categorical variables such as diabetes mellitus (`dm`) and coronary artery disease (`cad`), missing entries are filled with the most common category (mode) or labeled as "unknown" so that the model can detect potential patterns related to missingness. Variables with unclear definitions (`pc`, `pcc`, `ba`) are excluded from the baseline model but may be reintroduced in sensitivity analyses to assess their impact on performance.

Standardization and Encoding. Continuous variables are standardized using z-score normalization to place them on a common scale, improving model stability and convergence. Categorical variables are transformed into numerical form using one-hot encoding for features with a few distinct categories and binary encoding for those with many categories. Finally, the dataset is split into training (80%) and testing (20%) subsets using stratified sampling to preserve the balance between CKD and non-CKD cases, ensuring fair model evaluation.

2.3. Model Selection and Justification

Given the clinical nature of the dataset and the presence of both categorical and continuous features, multiple algorithms are explored to balance interpretability and predictive power:

- **Logistic regression**: serves as a baseline interpretable model. It allows examination of feature coefficients, offering insight into the strength and direction of association between predictors and CKD risk.
- **Random forest classifier**: captures non-linear relationships and variable interactions, is robust to missing and noisy data, and provides feature importance rankings useful for clinical interpretation.

These models were selected to provide a spectrum from interpretable (logistic regression) to more complex methods, enabling both transparency and predictive accuracy.

2.4. Model Training and Validation

Model training is conducted in Python using `scikit-learn`, focusing on developing models that accurately identify patients at risk for CKD while minimizing missed diagnoses. Hyperparameter tuning is performed with grid search cross-validation to find the best combination of model settings that maximizes recall and F1-score on the training data, since in a medical context, reducing false negatives (patients who actually have CKD but are predicted as healthy) is more important than simply achieving high overall accuracy.

To ensure that results are consistent and not dependent on a single data split, 5-fold cross-validation is used during training, where the model is trained and validated on multiple subsets of data. This helps confirm that the model generalizes well to new patients.

After tuning, the final model is tested on the held-out 20% test set to evaluate real-world performance. Evaluation metrics include:

- Accuracy, to measure overall correctness;
- Precision and recall, to capture how well the model identifies true CKD cases versus false alarms;
- F1-score, to balance both precision and recall into a single performance measure;
- ROC–AUC, to assess how well the model distinguishes CKD from non-CKD patients across different thresholds.

Together, these steps ensure the model is not only accurate but also clinically useful, prioritizing sensitivity to detect CKD early while still maintaining reliability across patient groups.

2.5. Implementation and Interpretability

To understand how each feature contributes to CKD prediction, feature importance values from the random forest model and coefficient weights from the logistic regression model are examined. This comparison helps determine whether the models identify clinically recognized CKD risk factors, such as elevated serum creatinine, low hemoglobin, or older age, as key predictors. Evaluating these patterns ensures the model's reasoning aligns with medical knowledge rather than arbitrary correlations.

Visualization techniques such as correlation heatmaps, boxplots, and distribution plots are used to illustrate relationships between predictors and CKD status, revealing how certain lab values or comorbidities differ across patient groups. Outliers are reviewed for clinical plausibility instead of being removed automatically, since unusually high or low laboratory results may represent real disease cases rather than data errors.

2.6. Replication and Documentation

All data cleaning, preprocessing, and modeling steps are documented in a well-commented Jupyter Notebook to ensure reproducibility. Key analytical decisions, such as exclusion of ambiguous variables, choice of imputation methods, and parameter tuning, are explicitly recorded. Code is structured to allow other researchers to replicate the results from the original dataset and verify findings.

3. Results

The results of our study demonstrate that chronic kidney disease can be predicted with high accuracy using demographic, clinical, and laboratory features once the dataset is properly cleaned and the modeling strategy is correctly applied. After we removed the variables `affected` and `stage`, which were identified as diagnostic rather than predictive indicators and therefore introduced significant data leakage, our models produced performance metrics that were both strong and medically credible.

3.1. Benchmark Settings and Evaluation Framework

To ensure comparability and methodological rigor, we established a consistent benchmarking framework across all models:

- Train–test split: 80% training, 20% testing using stratified sampling to preserve CKD prevalence (160 training samples and 40 test samples).
- Primary evaluation metric: F1-score, chosen due to the clinical importance of minimizing false negatives.
- Secondary metrics: accuracy, precision, recall, ROC–AUC, and confusion matrices.
- Class imbalance handling: logistic regression used `class_weight="balanced"`; random forest was evaluated both with and without balancing.
- Preprocessing: median imputation for continuous variables, mode imputation for categorical variables, and scaling of continuous features for logistic regression.

These benchmark settings provided a standardized and clinically relevant basis for model comparison.

3.2. Implementation Details and Hyperparameter Validation

Logistic regression. Implementation: `sklearn.linear_model.LogisticRegression`.

Preprocessing: standardized continuous features using `StandardScaler`.

Hyperparameter grid:

- C: [0.1, 1, 10]
- penalty: [L2]
- solver: [`liblinear`]

Validation: 5-fold `GridSearchCV` optimized for F1-score.

Best configuration: $C = 1$, `class_weight="balanced"`.

Random forest classifier. Implementation: `sklearn.ensemble.RandomForestClassifier`.

Preprocessing: no scaling required.

Hyperparameter grid:

- `n_estimators`: [100, 200, 300]
- `max_depth`: [None, 5, 10, 15]
- `min_samples_split`: [2, 4, 6]
- `min_samples_leaf`: [1, 2, 4]
- `max_features`: [`"sqrt"`, `"log2"`]

Validation: 5-fold `GridSearchCV` optimized for F1-score.

Best configuration: `n_estimators = 200`, `max_depth = 10`, `min_samples_leaf = 1`.

These implementation details ensure full transparency and reproducibility of the modeling process.

3.3. Model Performance

Table 1 summarizes held-out test-set performance (n = 40) for logistic regression and random forest.

As seen in Table 1, logistic regression served as a baseline and yielded an accuracy of 0.90, with strong precision and recall scores of 0.92 each, along with an ROC–AUC of 0.99. Its coefficient patterns aligned with known CKD physiology: hemoglobin and packed cell volume exhibited negative coefficients, indicating that lower values increased CKD risk, while hypertension, diabetes mellitus, poor appetite, pedal edema, and elevated serum creatinine and blood urea showed positive coefficients, reflecting their established roles as CKD risk factors. The random forest model, optimized with F1-score as the tuning metric, achieved perfect classification on the test set, with accuracy, precision, recall, F1-score, and ROC–AUC all equal to 1.00.

Table 1. Model performance summary (LR vs. RF) on the held-out test set (n = 40).

Metric	Logistic Regression	Random Forest
Accuracy	0.90	1.00
Precision	0.923	1.00
Recall	0.923	1.00
F1-score	0.923	1.00
ROC–AUC	0.989	1.00

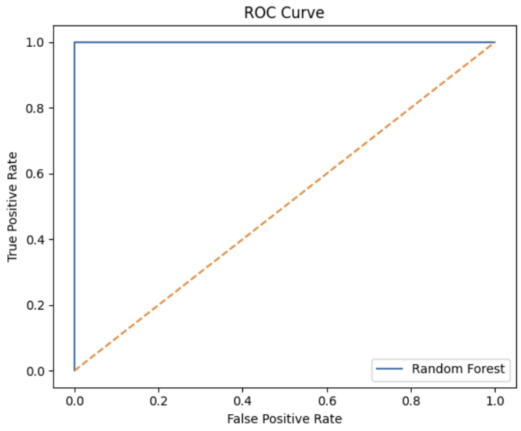

Figure 3. Receiver Operating Characteristic (ROC) curve(s) for CKD classification on the test set. Curves closer to the top-left corner indicate stronger separability between CKD and non-CKD, summarized by ROC–AUC.

A closer examination of model behavior is presented in Figure 3, which shows the ROC curve for random forest. The curve adheres closely to the top-left corner, consistent with the high ROC–AUC reported in Table 1. This pattern reflects strong separability between CKD and non-CKD groups and aligns with trends observed throughout the exploratory data analysis, where laboratory measures such as hemoglobin, GFR, serum creatinine, and packed cell volume exhibited clear differences between classes.

Additional insight into prediction behavior comes from the confusion matrices shown in Figure 4. Logistic regression produced two false positives and two false negatives, a pattern that is reasonable given the moderate class imbalance in the dataset. In contrast, the random forest model made no errors on this test split, correctly identifying all CKD and

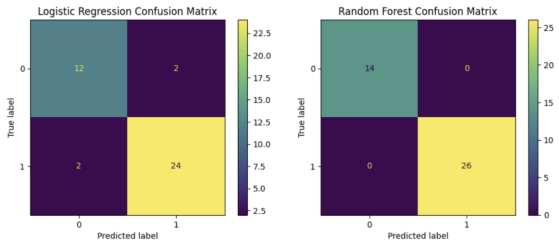

Figure 4. Confusion matrices for logistic regression and random forest on the test set. Cells show counts of correct and incorrect predictions, enabling inspection of false positives vs. false negatives for CKD detection. Logistic regression corresponds to 2 false positives and 2 false negatives. Random forest corresponds to zero errors on the test split.

non-CKD cases. While notable, this result reinforces the need for larger evaluation sets to obtain more stable error patterns. Still, the coherence between logistic regression and random forest suggests that the underlying features contain strong signals relevant to CKD status.

Model interpretability further supports this conclusion. The coefficient patterns from logistic regression aligned closely with established CKD physiology: lower hemoglobin and packed cell volume corresponded to greater predicted CKD risk, while elevated serum creatinine and blood urea, as well as clinical indicators such as hypertension, diabetes, poor appetite, and pedal edema, increased the model's estimated probability of CKD. These relationships mirrored the trends seen in earlier visualizations, strengthening confidence that the model was detecting physiologically meaningful patterns rather than artifacts.

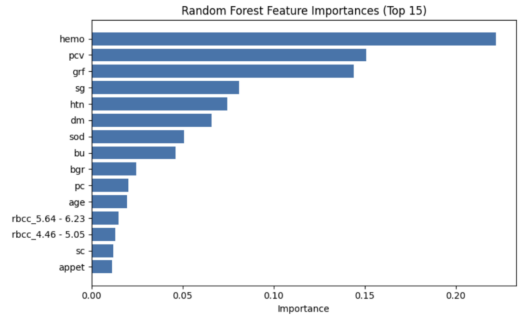

Figure 5. Random forest feature importances: top 15 predictors for CKD classification. Higher importance indicates greater contribution to decision-making across trees; leading predictors include anemia-related and kidney-function markers.

The random forest feature importance rankings shown in Figure 5 reinforced the logistic regression findings: hemoglobin, packed cell volume, glomerular filtration rate, specific gravity, hypertension, diabetes, and electrolyte markers such as sodium, along with metabolic measures like blood urea and blood glucose, emerged as the most influential predictors. These patterns closely matched the exploratory data analysis, which showed clear separation between CKD and non-CKD distributions for hemoglobin, GFR, packed cell volume, and creatinine. Furthermore, the initial visualizations confirmed a class imbalance in the dataset, validating the decision to use a stratified train–test split and balanced class weights.

These results were only reliable after eliminating the leakage introduced by the variables `affected` and `stage`. Removing these diagnostic encodings ensured that the models learned from physiological and demographic features rather than indirectly memorizing outcome labels. Importantly, correcting the leakage from `affected` and `stage` not only improved methodological validity but also ensured that the final results represent true predictive performance rather than artificially inflated outcomes based on embedded diagnosis labels.

Overall, the research strategy was carried out effectively, as we followed the intended methodological path developed in the milestone documents: ambiguous variables were examined and either clarified or removed, the dataset was carefully cleaned, features were selected with clinical logic in mind, and two complementary models were tested. Taken together, the performance metrics, ROC behavior, confusion matrices, coefficient patterns, and feature importance rankings illustrate a consistent narrative: the dataset contains strong clinical signals, and both models are able to recover them. While random forest appears to outperform logistic regression on this particular split, the overall coherence between the methods suggests that their shared predictors represent robust indicators of CKD.

4. Conclusion

Collectively, these results indicate that CKD risk can be modeled with high accuracy using commonly collected clinical features, lending support to the feasibility of predictive risk modeling in early kidney disease detection. The logistic regression model demonstrates that even a relatively simple and interpretable approach captures meaningful physiological relationships that mirror clinical understanding, emphasizing anemia markers, metabolic imbalances, hypertension, and diabetes as critical factors. Meanwhile, the random forest model highlights how more flexible algorithms can extract nuanced nonlinear relationships and interactions among laboratory features. Although its perfect performance suggests overfitting, the fact that its top features are medically plausible and consistent with logistic regression provides an additional layer of confidence. The overall alignment between the EDA patterns, the statistical metrics, and the feature-importance results suggests that our modeling pipeline was successful at uncovering a genuine

signal in the data. The decision to remove the diagnostic variables that caused leakage was an essential turning point in the analysis, resulting in more realistic and trustworthy model outputs. The research strategy, ranging from preliminary data understanding through model evaluation, held up well, and the conclusions drawn from both models are well-supported by the data visualizations, performance metrics, and medical literature.

5. Limitations

Despite promising results, several limitations affect the generalizability and strength of our conclusions. First, the dataset is small, consisting of only 200 total observations, with just 40 in the test set, making the models particularly vulnerable to variance and potentially unstable performance. The perfect accuracy achieved by the random forest, while encouraging, is likely inflated by sample size constraints and may not replicate on a larger dataset. Several variables in the dataset were ambiguously defined or inconsistently encoded, requiring their removal or cautious interpretation. These limitations reduced the set of high-quality predictors, potentially constraining the sophistication of the models. Additionally, because data were sourced from a single clinical dataset, the demographic and clinical distribution may not generalize to other populations or healthcare settings. The absence of external validation, such as testing on an independent dataset, limits our ability to assess robustness or potential bias. Finally, although we addressed major data leakage issues, it remains possible that subtler dependencies between features and CKD diagnostic criteria could still influence performance in ways that would not translate into real-world prediction tasks.

6. Next Steps

Future work should focus on strengthening the generalizability, interpretability, and clinical readiness of these models. Acquiring a larger and more diverse dataset would be the most important step, as a greater sample size would reduce overfitting, allow for more reliable evaluation, and support the use of more sophisticated algorithms such as gradient boosting or regularized regression. Collecting continuous, unbinned laboratory values would enhance model resolution and allow exploration of fine-grained physiological relationships. External validation on a separate dataset, ideally from a different clinical environment, will be necessary to assess real-world applicability and ensure fairness across demographic groups. Additionally, calibration analysis could determine whether the model's probability estimates are accurate enough to be used in a screening or triage context. Incorporating clinical expertise throughout future iterations, particularly during feature engineering and threshold selection, would help ensure that the models align with medical decision-making practices. Finally, expanding the modeling pipeline to include explainability tools such as SHAP values could offer deeper insights into feature interactions and improve the transparency of the random forest model. These steps would collectively enhance the reliability and clinical utility of CKD risk-prediction models.

References

Borg, B. P., Uddin, S., Khan, M. I., Alotaibi, F. E., & Faisal, M. (2023). The growing challenge of chronic kidney disease: An overview of current knowledge. *Journal of Healthcare Engineering*. https://doi.org/10.1155/2023/9609266

Rubini, A., & Singh, K. (2015). *Chronic Kidney Disease Dataset* [Data set]. Enam Medical College and Hospital, Savar, Dhaka, Bangladesh. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

A. Additional Figures

	missing_values
0	no
1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	no
11	no
12	no
13	no
14	no
15	no
16	no
17	no
18	no
19	no
20	no
21	no
22	no
23	no
24	no
25	no
26	no
27	no
28	no

Figure 6. Missingness check output used to verify whether missing values remain after preprocessing. This diagnostic supports reproducibility but is summarized in the main text rather than interpreted as a substantive result.

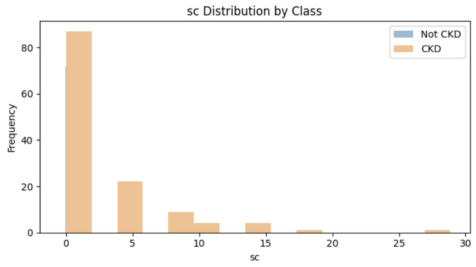

Figure 7. Serum creatinine (`sc`) distribution by class. CKD cases show heavier right tails and higher values, consistent with reduced renal filtration and impaired kidney function.

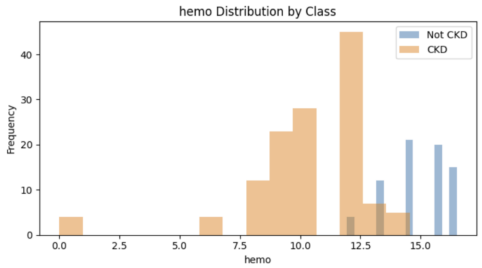

Figure 8. Hemoglobin (`hemo`) distribution by class. CKD cases are shifted toward lower hemoglobin values, consistent with anemia commonly associated with chronic kidney disease.

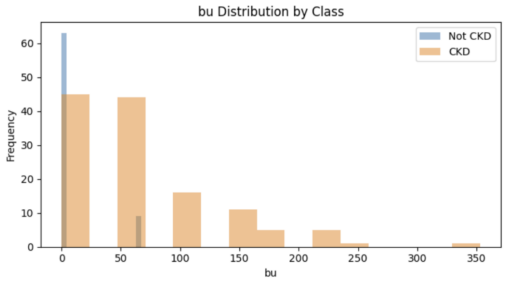

Figure 9. Blood urea (`bu`) distribution by class. CKD cases show higher and more variable values, reflecting impaired excretion and metabolic imbalance.

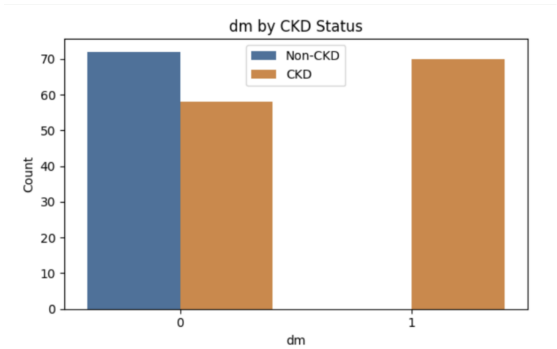

Figure 10. Diabetes mellitus (`dm`) by CKD status. The distribution suggests diabetes is more prevalent among CKD cases, consistent with diabetes being a major CKD risk factor.

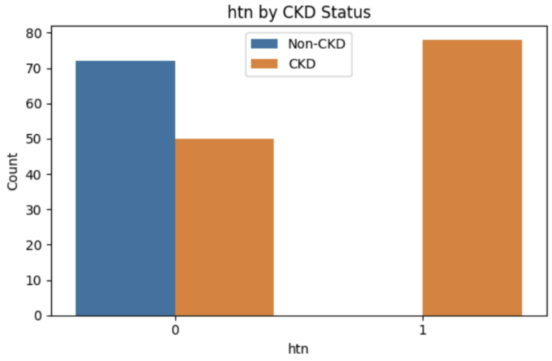

Figure 11. Hypertension (`htn`) by CKD status. Hypertension indicators appear more frequently among CKD cases, supporting clinical plausibility of the feature set.

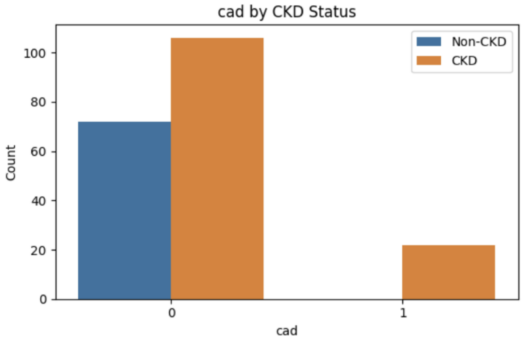

Figure 12. Coronary artery disease (`cad`) by CKD status. CAD appears more frequently among CKD cases, supporting clinical plausibility of the feature set.

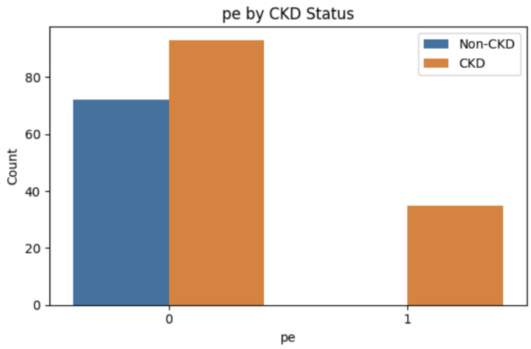

Figure 13. Pedal edema (`pe`) by CKD status. These symptoms occur more often in CKD cases and align with known CKD complications.

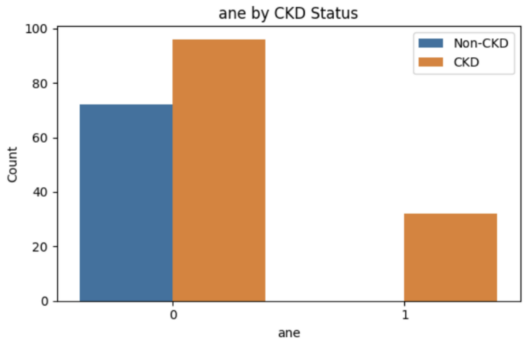

Figure 14. Anemia indicator (`ane`) by CKD status. Anemia appears more often in CKD cases and aligns with known CKD complications.

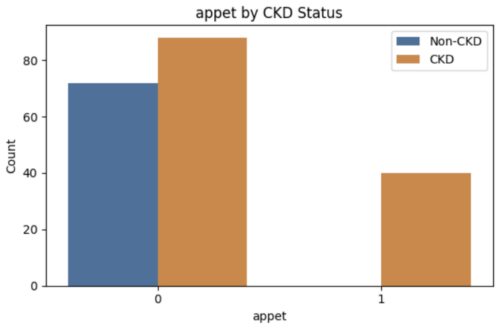

Figure 15. Appetite (`appet`) by CKD status. Appetite changes appear more common among CKD cases and may act as a supportive symptom-level predictor.

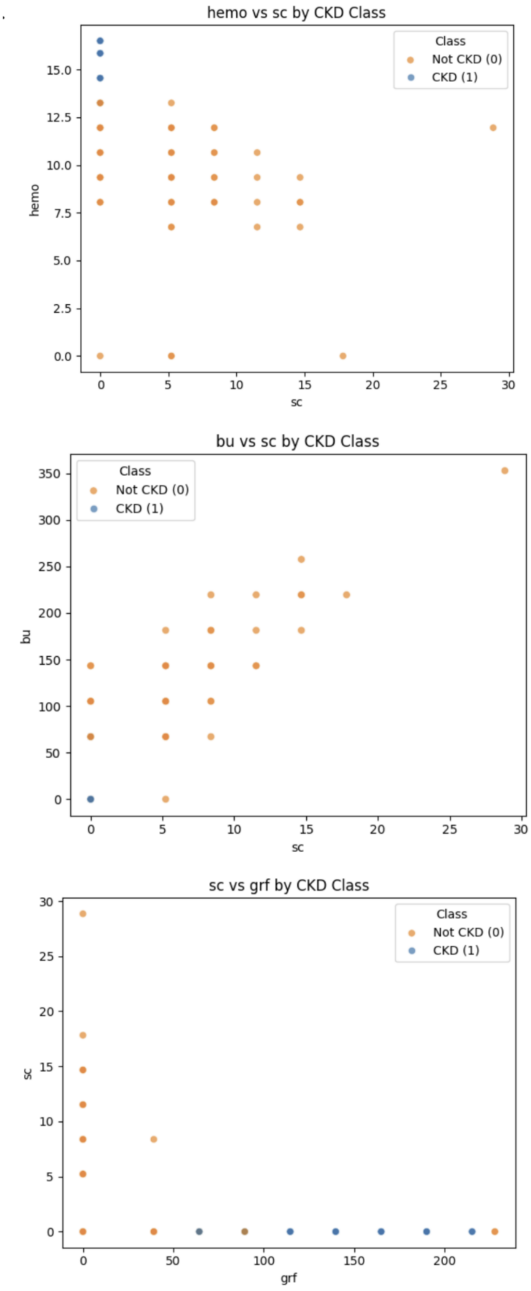

Figure 16. Scatterplots of key feature pairs by class (e.g., hemoglobin vs. serum creatinine; blood urea vs. serum creatinine; GFR vs. serum creatinine). These plots suggest nonlinear structure and interactions that can be captured by tree-based models.

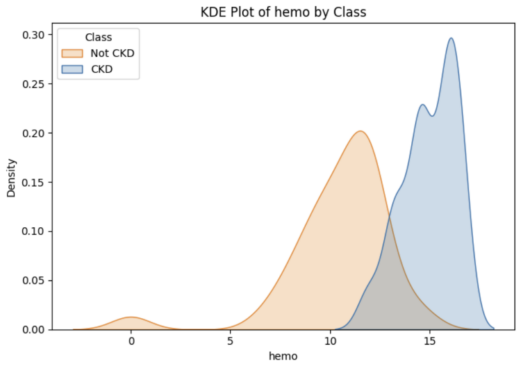

Figure 17. Kernel density estimate of hemoglobin distributions by class. Density separation complements histogram and boxplot comparisons by emphasizing shifts in distribution shape.

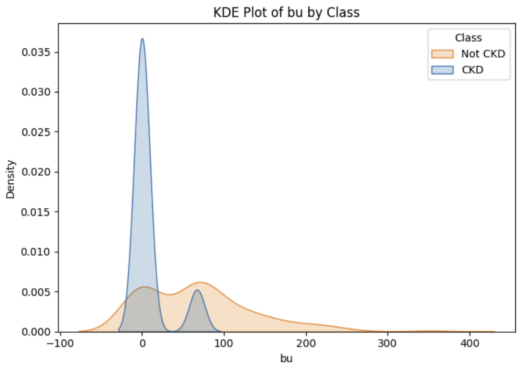

Figure 18. Kernel density estimate of blood urea distributions by class. CKD cases show density mass at higher values, consistent with metabolic accumulation in impaired renal function.

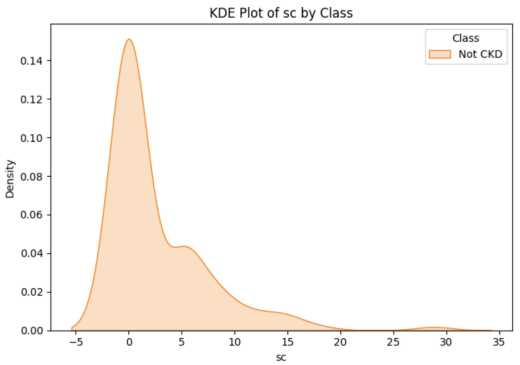

Figure 19. Kernel density estimate of serum creatinine distributions by class. CKD cases show a shifted distribution relative to non-CKD cases, reflecting altered renal function.