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 major public health concern, and early risk prediction could support more timely intervention. In this project, we analyze the Risk Factor Prediction of Chronic Kidney Disease dataset from the UCI Machine Learning Repository to build models that distinguish CKD from non–CKD patients using demographic, clinical, and laboratory variables. We emphasize careful data cleaning and the removal of diagnostic variables that would cause data leakage, focusing instead on clinically interpretable predictors such as serum creatinine, blood urea, hemoglobin, and comorbid conditions. Using logistic regression as an interpretable baseline and a random forest classifier as a more flexible model, we achieve high predictive performance while maintaining clinical plausibility in feature effects. Our results highlight both the promise of CKD risk modeling with routinely collected data and the importance of rigorous preprocessing and validation to obtain medically credible estimates of performance.

1. Data Description of the Chronic Kidney Disease Risk Factor Dataset

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. 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. Together, these features can provide insight into relationships between underlying health conditions, lifestyle-related factors, and kidney function.

1.1. 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.

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) is coded as a binary variable for presence or absence. Appetite (appet) is included as a subjective clinical measure, coded as *normal* or *poor*.

1.2. 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.

1.3. Preparing for Analysis

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. 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.

2. Methods

2.1. Research Question and Overall 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. The project aims to identify which clinical factors are most predictive of CKD, compare models to balance clinical transparency and predictive performance, and assess how well these models generalize to real-world data.

2.2. Data Wrangling and Preprocessing

Handling Missing and Ambiguous Data. Missing and ambiguous data are handled through targeted imputation and selective exclusion. 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 (e.g., 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 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, with an emphasis on accurately identifying 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 maximize recall and F1-score on the training data. 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.

These metrics together 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. Model Performance

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. Its confusion matrix contained zero misclassifications, though this result must be interpreted cautiously due to the small test size.

3.2. Feature Importance and Alignment with EDA

The random forest feature importance rankings 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. Overall, the research strategy was carried out effectively: 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, one interpretable and one flexible, were tested. 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.

4. Discussion

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