Data Description of the Chronic Kidney Disease Risk Factor Dataset

Aishani Patnaik. Cody Liddle. Meher Kalsi.

Data Description

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. These features together can provide insight into the relationships between underlying health conditions, lifestyle-related factors, and kidney function.

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 (grf), 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) and blood pressure (bp) are coded as a binary variable for presence or absence. Appetite (appet) is included as a subjective clinical measure, coded as normal or poor.

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, grf, 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’ve 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.

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.

Data Description of the Chronic Kidney Disease Risk Factor Dataset

Aishani Patnaik. Cody Liddle. Meher Kalsi.

Preliminary Analysis

Methods

1. Overview of the 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 will focus primarily on features that are well-defined and clinically interpretable, as discussed in the data description milestone.

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 will emphasize 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 will follow a structured pipeline: data cleaning and preprocessing, exploratory data analysis (EDA), model development, validation, and interpretability.

2. Data Wrangling and Preprocessing

Handling Missing and Ambiguous Data

Data Description of the Chronic Kidney Disease Risk Factor Dataset

Aishani Patnaik. Cody Liddle. Meher Kalsi.

Missing and ambiguous data will be 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 will be 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 will be 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) will be excluded from the baseline model but may be reintroduced in sensitivity analyses to assess their impact on performance.

Standardization and Encoding

Continuous variables will be standardized using z-score normalization to place them on a common scale, improving model stability and convergence. Categorical variables will be 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 will be 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.

3. Model Selection and Justification

Given the clinical nature of the dataset and the presence of both categorical and continuous features, multiple algorithms will be 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, 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.

4. Model Training and Validation

Model training will be conducted in Python using scikit-learn, focusing on developing models that accurately identify patients at risk for CKD while minimizing missed diagnoses. Hyperparameter tuning will be performed with Grid Search Cross-Validation to find the best

combination of model settings that maximize 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 will be 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 will be tested on the held-out 20% test set to evaluate real-world performance. Evaluation metrics will 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.

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 will be examined. This comparison will help 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 will be used to illustrate relationships between predictors and CKD status, revealing how certain lab values or comorbidities differ across patient groups. Outliers will be 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.

6. Replication and Documentation

All data cleaning, preprocessing, and modeling steps will be 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, will be explicitly recorded. Code will be structured to allow other researchers to replicate the results from the original dataset and verify findings.

Summary

This preliminary methods plan outlines a structured and clinically informed approach to predicting chronic kidney disease using demographic, clinical, and laboratory data. The plan focuses on data quality, interpretability, and reproducibility, making sure that each step of the analysis is transparent and based on clinical reasoning. By addressing issues like missing data and unclear variables, the project aims to create models that are both accurate and meaningful in a medical context. The final stage will refine model parameters through cross-validation, compare the performance of Logistic Regression and Random Forest models, and identify the key predictors of CKD risk to support early detection and better clinical decision-making.