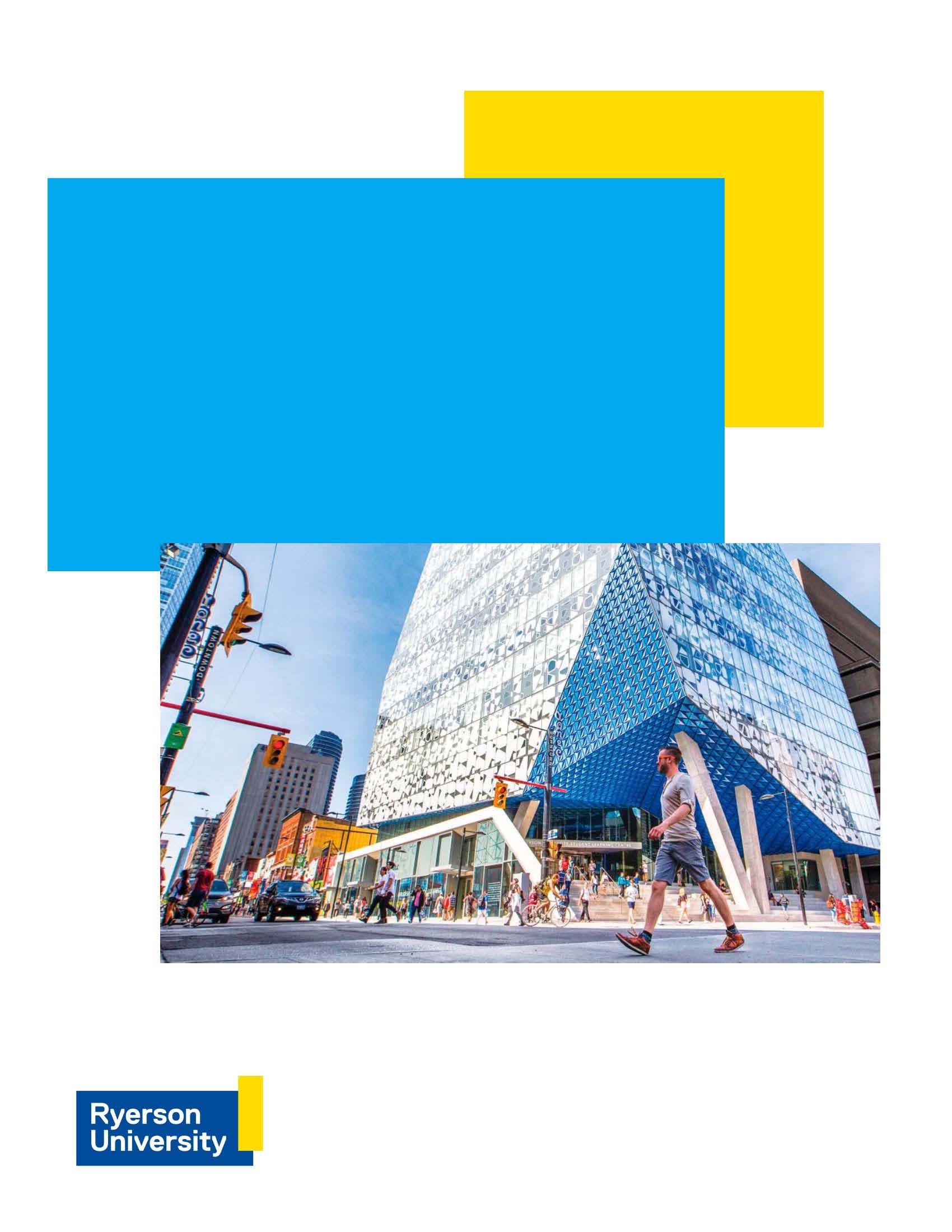
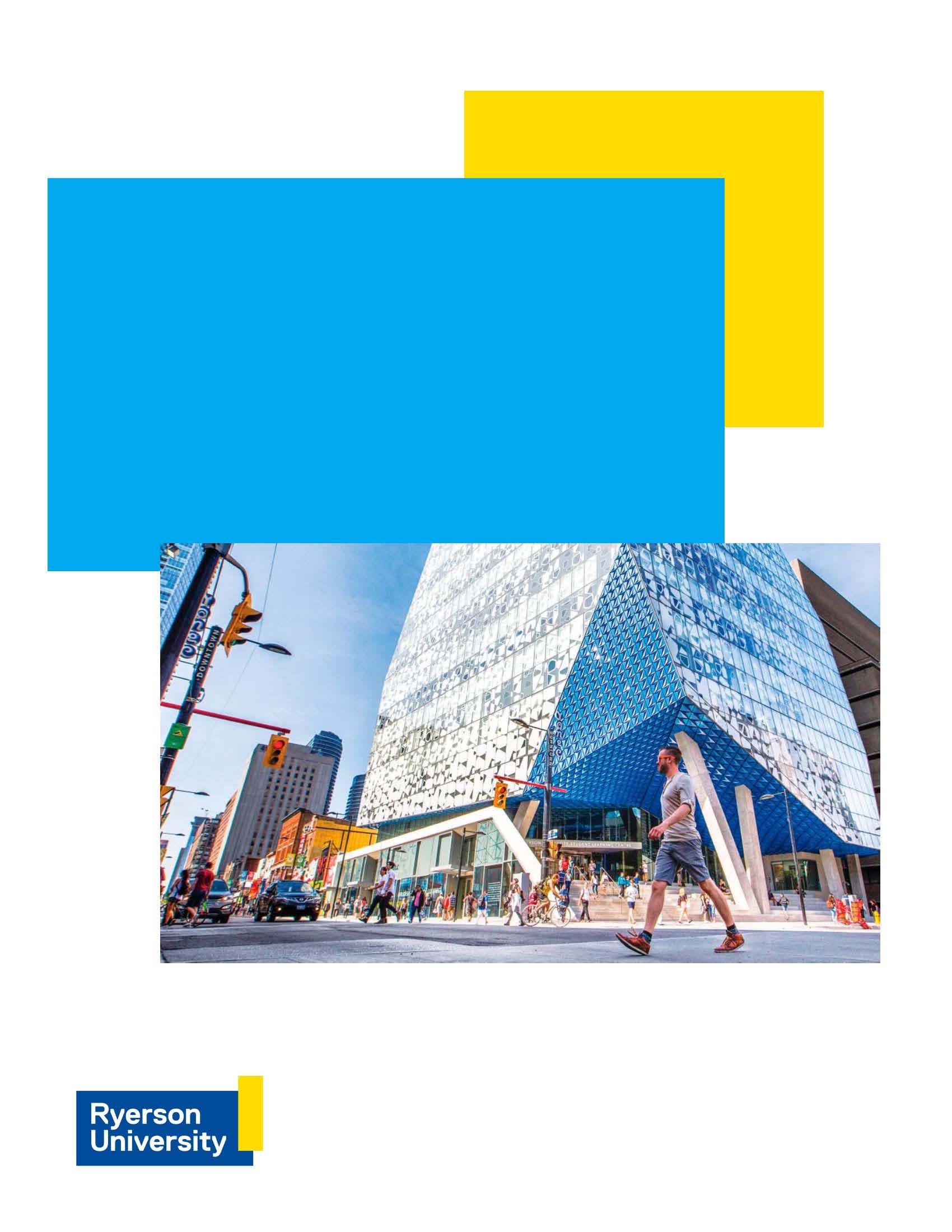
Predicting Chronic Kidney Disease Risk Using Machine Learning: An Analysis of Key Medical and Clinical Indicators

Kshitija (Tija) Murhe

# Abstract

Chronic Kidney Disease (CKD) is a significant global health challenge, often progressing silently until advanced stages. This study aims to analyze risk factors for CKD using the "Risk Factor Prediction of Chronic Kidney Disease" dataset from the UCI Machine Learning Repository ([dataset link](https://archive.ics.uci.edu/dataset/857/risk+factor+prediction+of+chronic+kidney+disease)). By examining lifestyle, demographic, and medical indicators, I seek to identify key correlations and predictive factors associated with CKD onset.

My research seeks to explore how effectively medical and clinical indicators can predict the onset of Chronic Kidney Disease and identify which factors are most strongly linked to CKD risk.

Using machine learning techniques, I will apply classification algorithms such as Naïve bayes, Decision Tree, SVM, Logistic Regression, and K-Nearest Neighbors. to identify patterns in risk factors. Model evaluation involves cross-validation and performance metrics, including accuracy, precision, recall, and F1-score. This study contributes to the field of predictive analytics in healthcare by advancing knowledge of CKD risk factors and supporting efforts to mitigate its progression.

# Similarities and Differences Between my Analysis, the Replicated Study and Past Research

**Overview of the Replicated Study:** I am replicating a study called "Machine Learning Techniques for Chronic Kidney Disease Risk Prediction,”: [Machine Learning Techniques for Chronic Kidney Disease Risk Prediction](https://www.mdpi.com/2504-2289/6/3/98). This study looks at different machine learning methods to predict the risk of Chronic Kidney Disease (CKD) based on various medical and clinical factors.

#### Similarities

1. **Research Questions**: Both my analysis and the original study aim to find out what risk factors are linked to CKD and how well machine learning can predict its onset. We both focus on understanding the causes of CKD risk.
2. **Use of Machine Learning**: Both studies use machine learning techniques to analyze the data. Methods like logistic regression, decision trees, K-NN, Naïve Bayes and SVM are applied in both analyses, making our approaches similar.
3. **Focus on Prediction**: Both studies aim to create models that can predict CKD risk accurately. The goal is to help detect CKD early and improve patient care.

#### Differences

1. **Dataset and Time Frame**: The original study uses a different dataset or cover a different time period. My analysis uses the "Risk Factor Prediction of Chronic Kidney Disease" dataset from the UCI Machine Learning Repository, which includes demographic (age) and medical factors that can give a better overall picture of CKD risk.
2. **Data Attributes**: The factors included in my dataset differ from those in the original study. The dataset that I am using includes additional medical factors, like levels of sugar, blood pressure, presence of heart disease, presence of diabetes, and also has clinical factors such as anemia, pus cell, pus cell clumps, pedal edema and glomerular filtration rate. This helps me gain more insights into how additional medical and clinical factors affect CKD risk.

By replicating this study and using a different dataset, and applying advanced methods, my analysis aims to enhance our understanding of CKD risk factors. Comparing these approaches will not only show how effective machine learning is in predicting CKD but may also uncover new insights that could help in early intervention strategies.

# GitHub Link

A link to the GitHub repository where the code and results are uploaded can be found here:

**GitHub Repository:** <https://github.com/tijamurhe/Big-Data-Analytics-Project.git>

# Applied Methodology and Study Design

The goal is to clean, preprocess, and analyze a Chronic Kidney Disease (CKD) dataset to enable robust insights and conduct predictive analytics. The study involves addressing missing data, transforming categorical and numerical variables, and visualizing relationships between attributes and CKD status.

**Study Design**

The study aims to explore the relationships between various clinical and medical features and the presence of CKD. I hypothesized that certain biomarkers and symptoms would significantly differ between CKD and non-CKD patients.

My analysis proceeded in the following manner:

1. **Data Cleaning:** Ensured data integrity and consistency by handling missing values and standardizing formats.
2. **Feature Transformation:** Standardized variables to enable accurate comparisons.
3. **Exploratory Data Analysis (EDA):** Utilized visualization techniques to identify patterns and insights.
4. **Classification Using Machine Learning Models:** Applied multiple machine learning algorithms to classify CKD vs. Non-CKD cases.

I focused on both categorical features and numerical features to provide a comprehensive overview. I selected visual tools such as histograms and heatmaps for their effectiveness in revealing underlying data structures.

I assumed that missing data were randomly distributed and that mode imputation would not introduce significant bias. However, I acknowledged that this approach might oversimplify the natural variability present in clinical measurements.

**Data Cleaning Methodology**

**Dataset Loading and Initial Examination**

The dataset was loaded using Pandas (pd.read\_csv), a powerful data manipulation library in Python. During this step:

* Metadata rows, such as attribute information, were removed to ensure the dataset's structure aligned with analysis requirements.
* Missing values, represented as "?", were replaced with NaN using pd.NA to leverage Pandas' built-in functionality for handling missing data.

**Handling Missing Values**

* Rows with significant missing data were dropped using the dropna function to eliminate non-informative records.
* Missing numerical values were imputed with the **mode** of the respective column. This imputation method balances the dataset by maintaining distribution properties while minimizing the influence of outliers.

**Cleaning Numerical Columns**

* A custom cleaning function was applied to standardize numerical columns with inconsistencies, such as ranges or non-numeric characters:
  + Missing values within numerical columns were replaced with the mode to maintain consistency.

**Categorical Variable Transformation**

* A predefined dictionary (categorical\_mappings) mapped coded categorical values (e.g., "0", "1") to descriptive labels, improving interpretability.
* These columns were converted to the category data type to optimize memory usage and ensure accurate labeling.

**Attribute Descriptions**

* An attribute dictionary was created to provide metadata for each feature. This step ensures clarity in analysis.

**Data Export**

* The cleaned dataset was exported to a CSV file (to\_csv) for future use, ensuring that subsequent analyses operate on a high-quality dataset.

**Exploratory Data Analysis (EDA) Methodology**

Exploratory data analysis was conducted for categorical and numerical variables. The following provides information on the methods used on each type of data.

**1. Categorical Data Analysis**

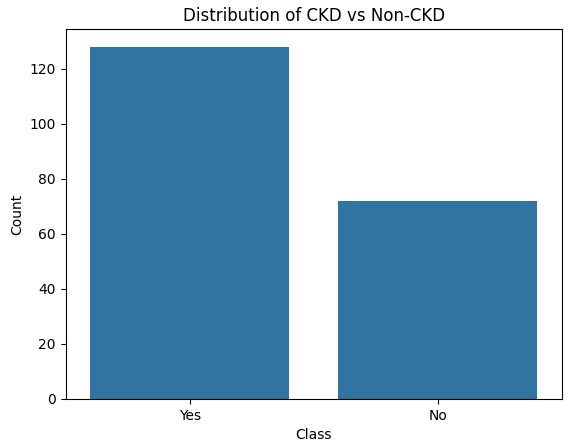
* **Class Distribution:**
  + 

Fig A. Displays the distribution of CKD vs Non-CKD cases in the dataset

* **Feature Relationships:**
  + Count plots (sns.catplot) for each categorical feature, segmented by CKD status, identify key patterns and relationships.
* **Correlation Heatmap:**
  + Categorical features are encoded using one-hot encoding (pd.get\_dummies).
  + A heatmap (sns.heatmap) displays correlations, identifying strong relationships.

**2. Numerical Data Analysis**

* **Distribution Analysis:**
  + Histograms (sns.histplot) show distributions of numerical features, segmented by CKD status.

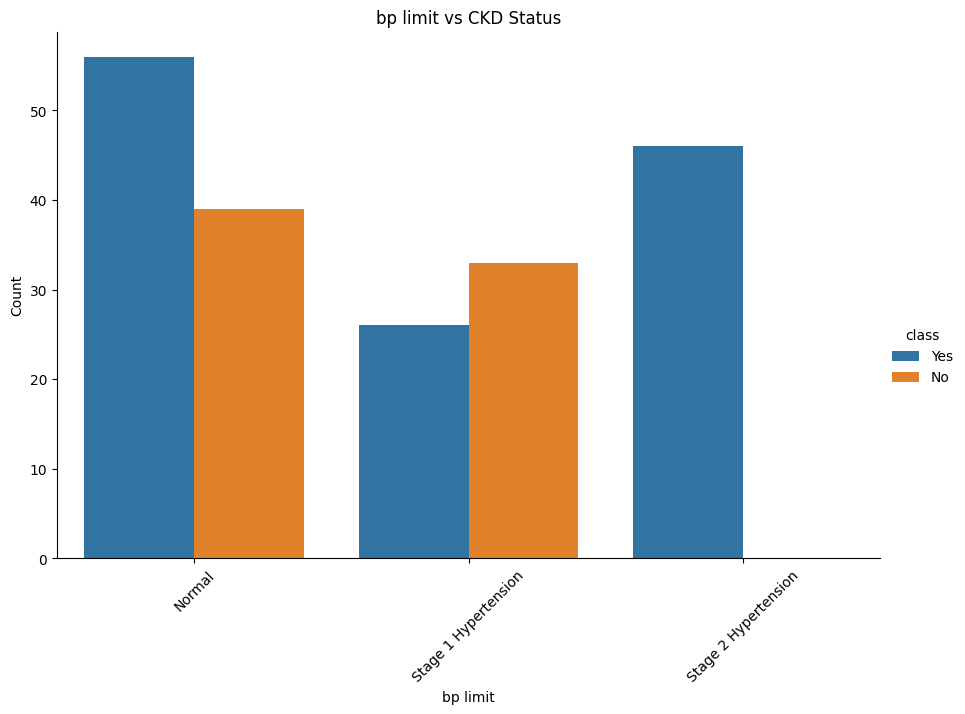
**3. Integration of Categorical and Numerical Data:**

* Patterns from categorical and numerical analysis are integrated to hypothesize feature importance and identify significant variables for CKD diagnosis.

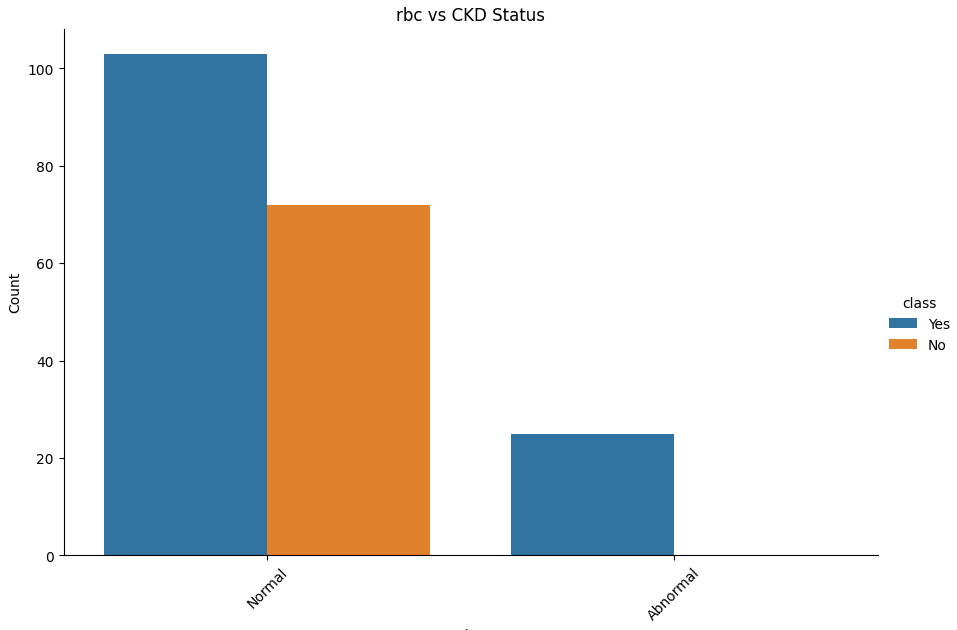
The exploratory data analysis (EDA) aimed to understand patterns in the data and identify important factors for CKD diagnosis. For categorical data, the class distribution chart (Fig A) showed the number of CKD and Non-CKD cases, while count plots helped explore how different features relate to CKD status. A heatmap was used to find strong relationships between features by converting them into numerical form. For numerical data, histograms with smooth curves showed how key features are distributed and how they differ between CKD and Non-CKD groups. Combining insights from both categorical and numerical data helped identify important features and understand their role in diagnosing CKD.

**Outputs of EDA**

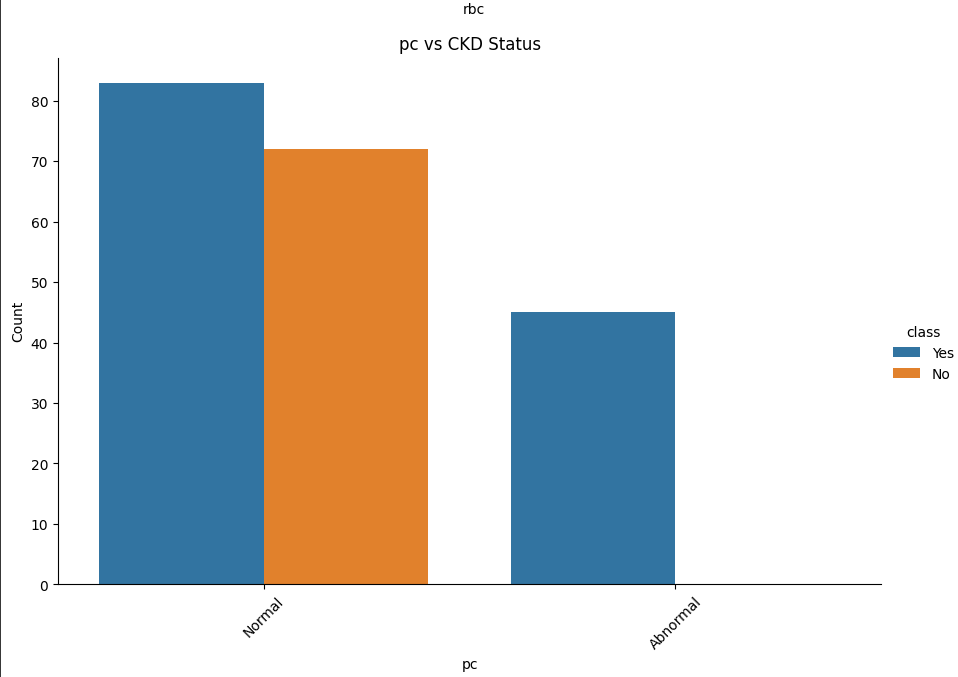
**Categorical Features Analysis**



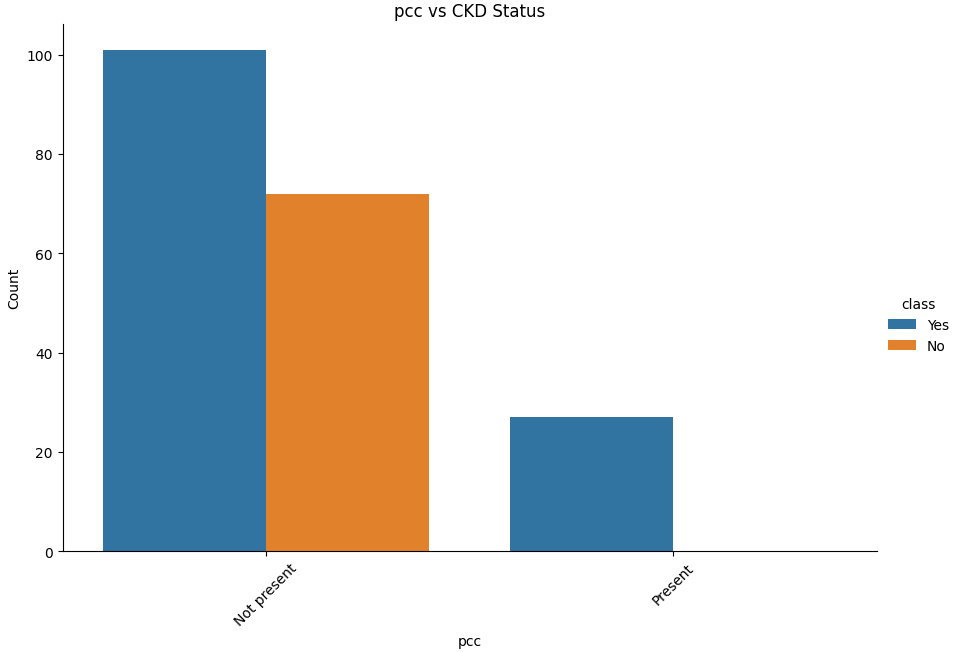
1. Bp (Limit): This attribute represents the severity levels of hypertension, which are closely tied to the progression of CKD. Hypertension accelerates kidney function decline, while CKD can contribute to elevated blood pressure. It is important to note that Stage 2 hypertension is only seen within the patients that have CKD in the study.



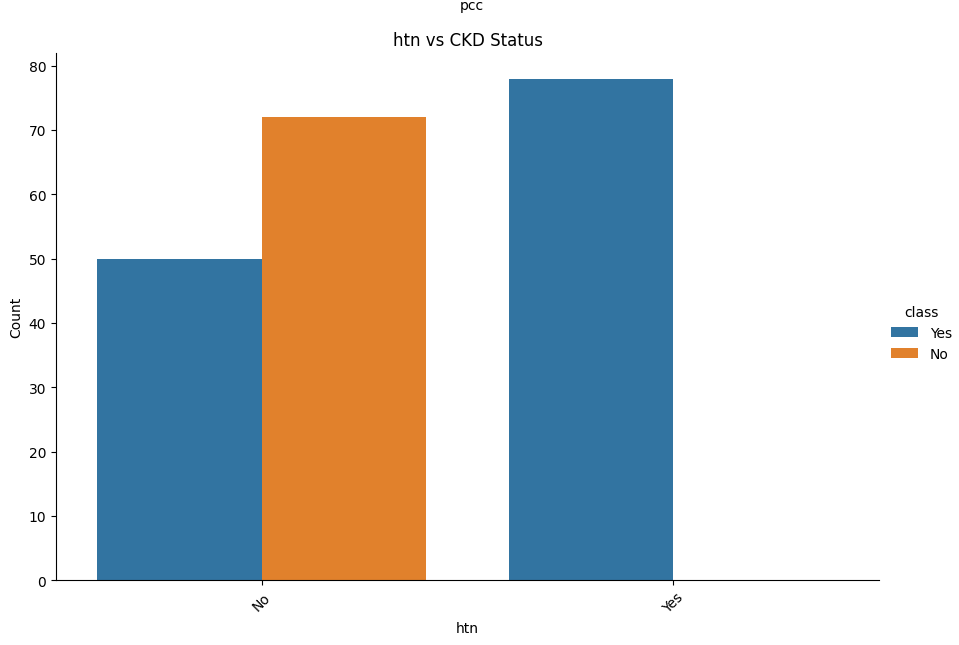
1. Rbc: The presence of red blood cells (RBCs) in urine is classified as either "Normal" or "Abnormal." Abnormal RBC levels, a condition known as hematuria, may signify kidney damage or other kidney-related diseases. As displayed in the graph above, abnormal levels of rbc in urine are only seen amongst cases with CKD in the study.



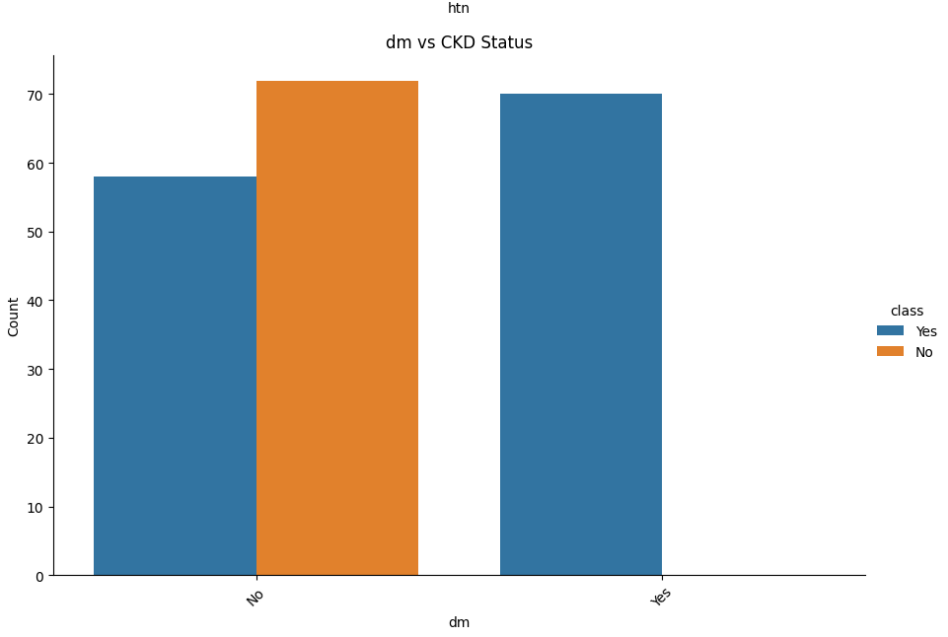
1. Pc: The presence of pus cells (pc) in urine is categorized as either "Normal" or "Abnormal." Elevated levels of pus cells may indicate urinary tract infections (UTIs) or inflammation, which can have significant implications for kidney health. In cases of CKD, recurrent infections or persistent inflammation can exacerbate kidney damage, impairing kidney function and potentially accelerating the progression of the disease. As displayed in the graph above, abnormal levels of pus cells are only seen in cases with CKD in the study.



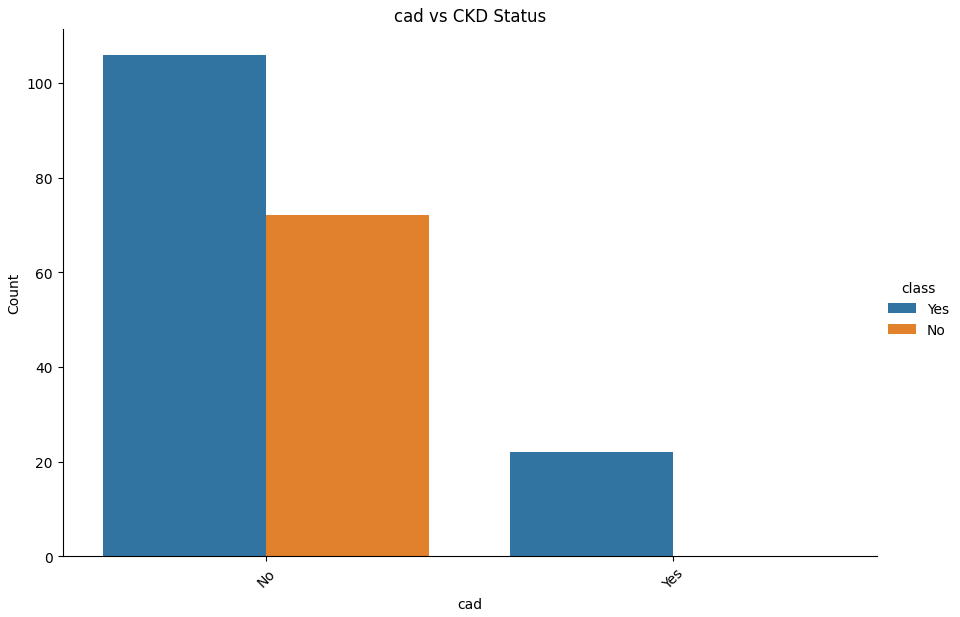
1. The presence of pus cell clumps (PCC) in urine is identified as either "Present" or "Not Present." The detection of pus cell clumps typically indicates severe urinary tract infections (UTIs). Such infections, if left untreated, can lead to kidney damage, further complicating conditions like chronic kidney disease (CKD) by worsening inflammation and impairing kidney function. As displayed in the graph above, presence of pus cell clumps is only seen in cases with CKD in the study.



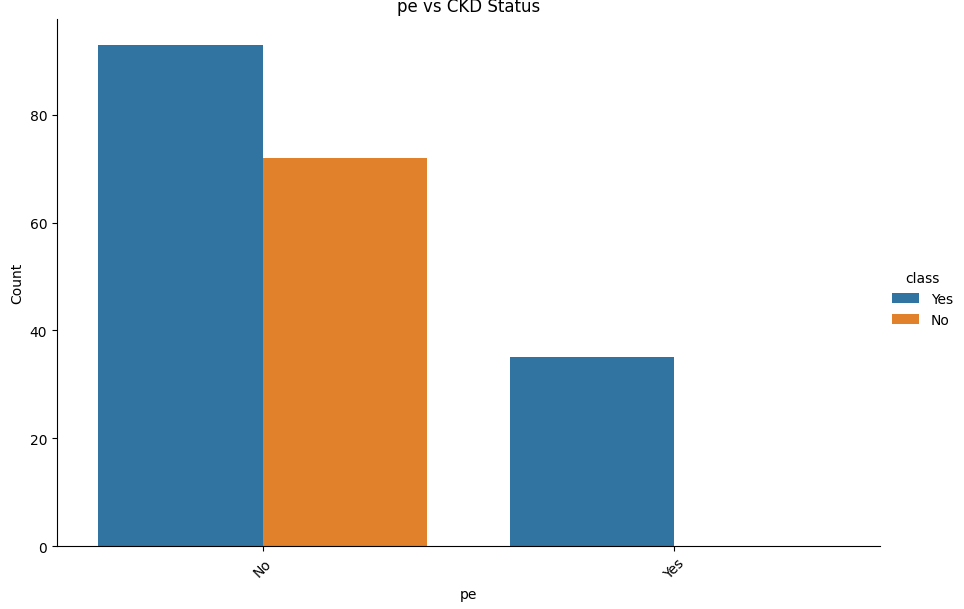
1. Hypertension (Htn) Indicates whether high blood pressure is present ("Yes" or "No").  
   Hypertension is both a primary cause and a consequence of CKD.As displayed in the graph, htn is only present in cases with CKD in the study.



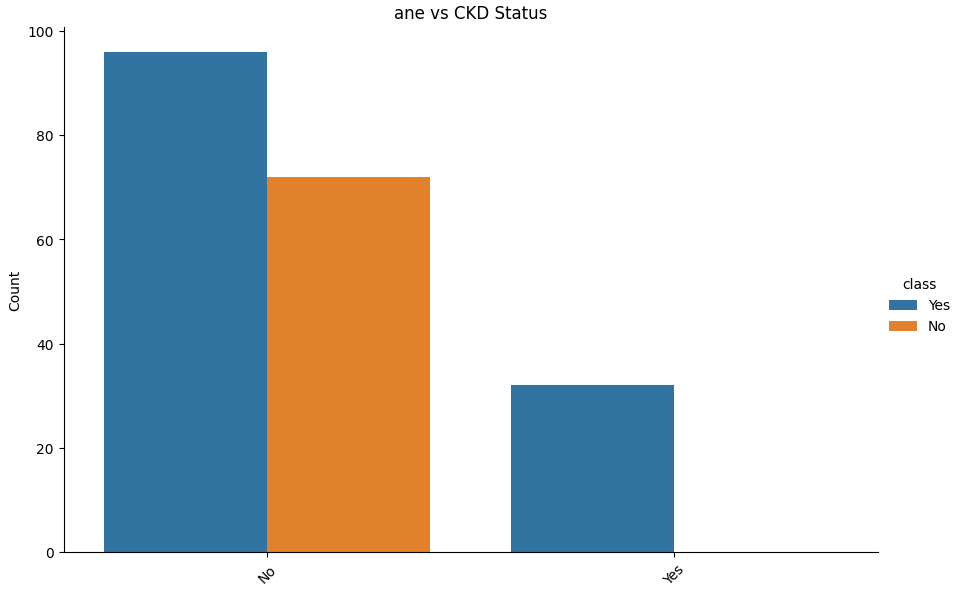
1. DM (Diabetes Mellitus) indicates whether diabetes is present ("Yes" or "No").Chronic hyperglycemia from diabetes damages the kidneys’ blood vessels, making it a leading cause of CKD.As displayed in the graph, dm is only present in cases with CKD in the study.



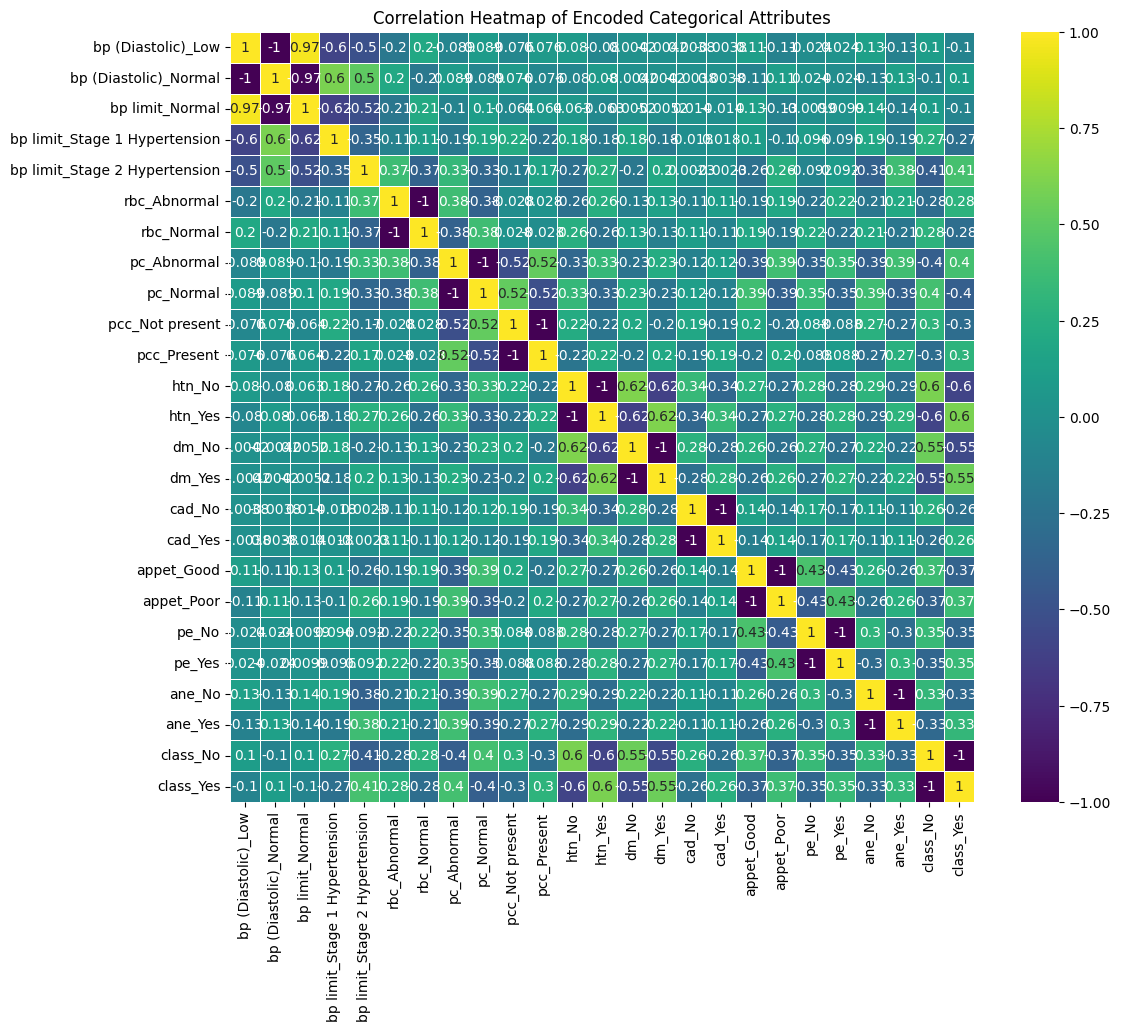
1. CAD (Coronary Artery Disease), indicates whether CAD is present ("Yes" or "No"). CKD and CAD often coexist due to shared risk factors such as hypertension and diabetes. As displayed in the graph, CAD is only present in cases with CKD in the study.



1. PE (Pedal Edema), indicates the presence of fluid retention in the lower limbs ("Yes" or "No"). Edema reflects kidney dysfunction due to impaired fluid regulation. As displayed in the graph, PE is only present in cases with CKD in the study.



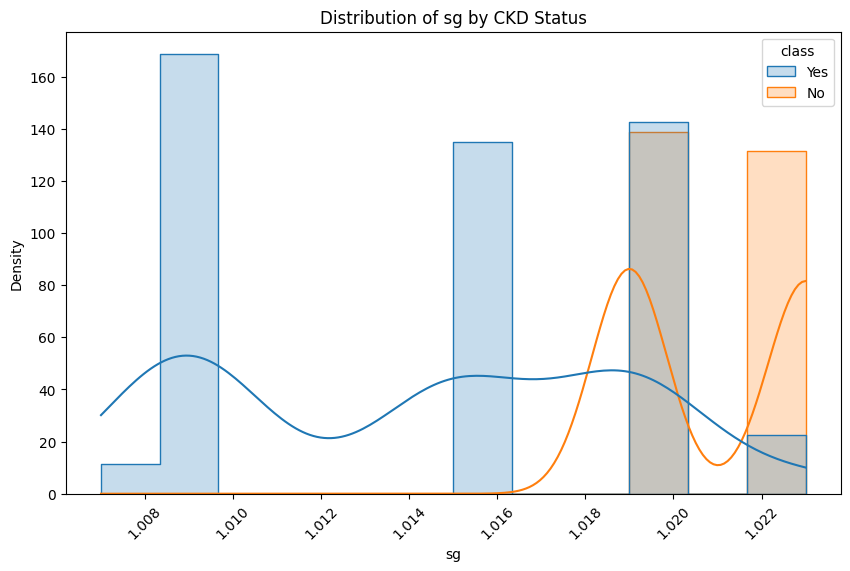
1. Ane (Anemia), indicates whether anemia is present ("Yes" or "No"). Anemia commonly occurs in CKD due to the kidneys’ reduced production of erythropoietin, a hormone essential for red blood cell production. As displayed in the graph, Anemia is only present in cases with CKD in the study.



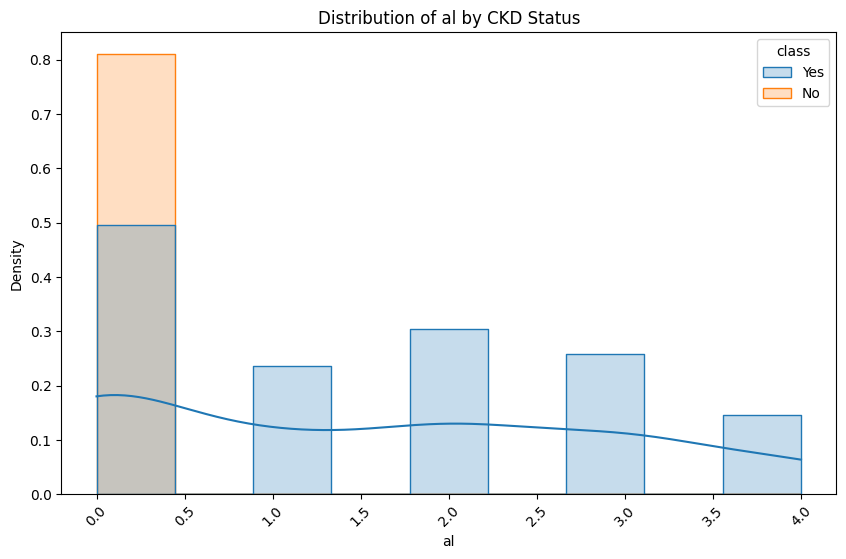
Regarding the correlation with CKD status, variables such as htn\_Yes (Hypertension), dm\_Yes (Diabetes Mellitus), ane\_Yes (Anemia), and pc\_Abnormal (Pus Cells Abnormal) exhibit strong positive correlations with class\_Yes, indicating a strong association with CKD. Conversely, variables like class\_No, appet\_Good (Good Appetite), and htn\_No (No Hypertension) show strong negative correlations with class\_Yes, highlighting that they are healthy indicators against CKD.

Inter-variable correlations reveal meaningful relationships. For instance, htn\_Yes and dm\_Yes are positively correlated (~0.6), reflecting the frequent coexistence of hypertension and diabetes, both of which are significant contributors to CKD. Similarly, ane\_Yes (Anemia) and appet\_Poor (Poor Appetite) show a positive correlation. Anemia is common in advanced CKD, which often coincides with uremia and appetite loss. Additionally, pc\_Abnormal and pcc\_Present demonstrate a moderate positive correlation, emphasizing their shared role as markers of infection or inflammation in the kidneys or urinary tract.

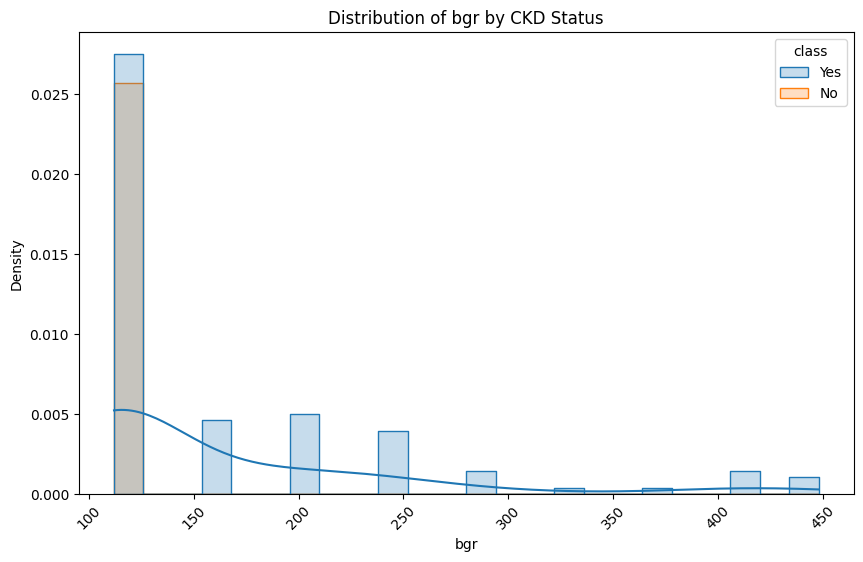
**Numerical Features:**



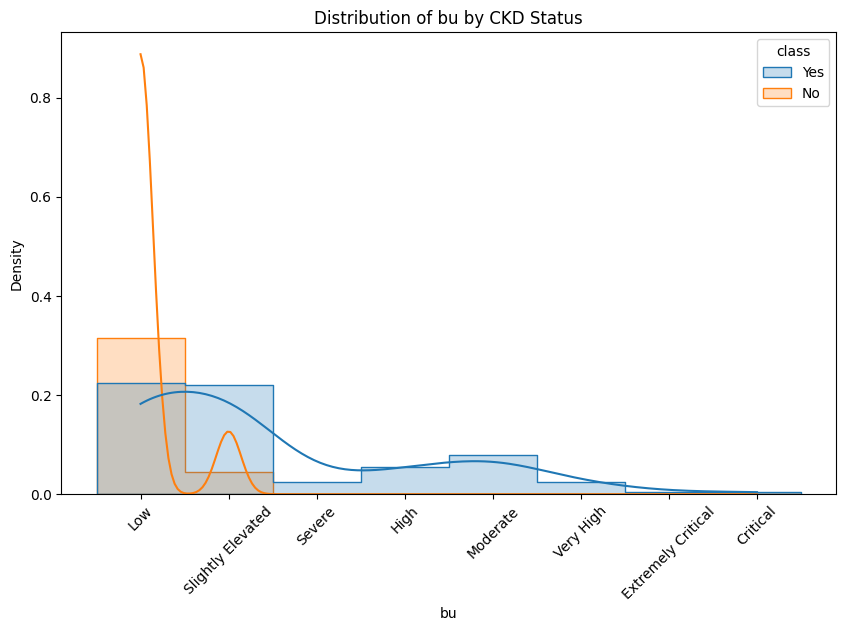
1. Sg (Specific Gravity) measures urine concentration, indicating hydration and kidney filtering ability. In CKD, the kidneys lose their ability to concentrate urine. This leads to dilute urine regardless of the body's hydration status, resulting in low specific gravity. As displayed in the graph, cases with CKD have lower levels of sg compared to cases that do not have CKD.



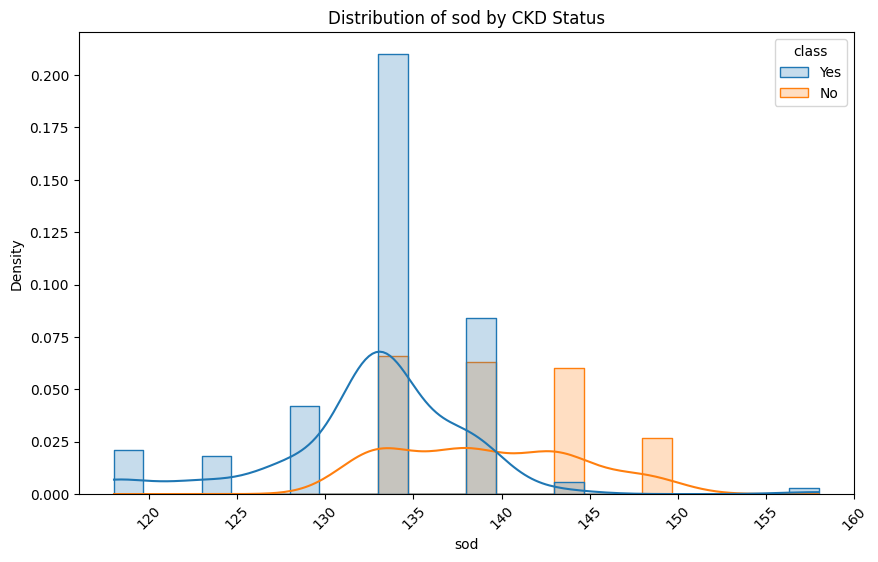
1. Al (Albumin), indicates albumin concentration in urine. High levels (proteinuria) are a hallmark of kidney damage. It is evident in the graph above that high levels of Al are seen in cases with CKD compared to non-CKD.



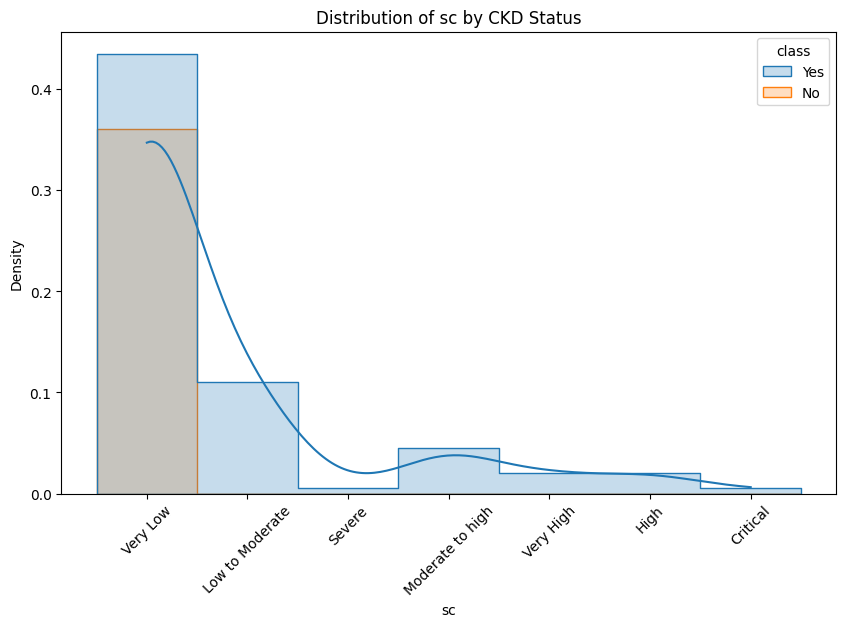
1. Bgr (Blood Glucose Random), reflects random blood sugar levels. High levels indicate diabetes, a leading cause of CKD. As displayed in the graph, high levels of Bgr are only seen in cases with CKD.



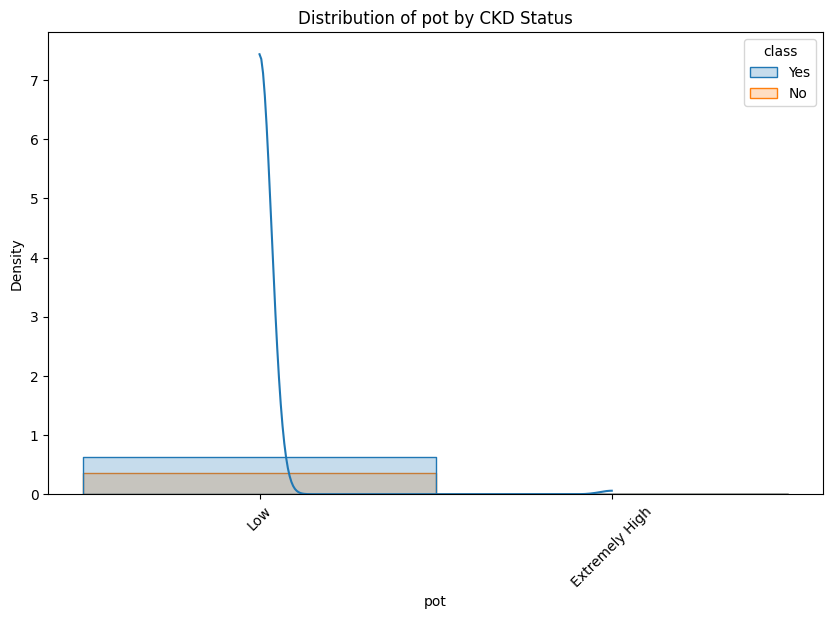
1. Bu (Blood Urea), measures urea in blood. Elevated levels suggest reduced kidney filtering ability. It is evident in the graph that very high and critical levels of Bu are only seen in cases with CKD.



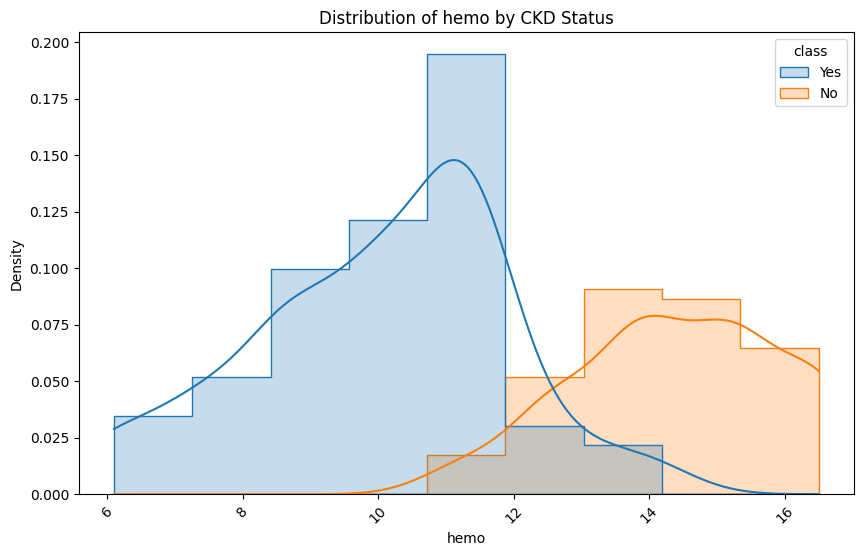
1. Sodium (sod) levels in the blood are an important indicator of kidney function and overall fluid and electrolyte balance. The kidneys play a central role in regulating sodium levels in the body by controlling its reabsorption and excretion. In CKD this regulatory ability can become impaired, leading to abnormal sodium levels. As displayed in the graph, a large number of cases with CKD do not have the normal range of sod levels 135–145 mEq/L when compared to non-CKD cases.



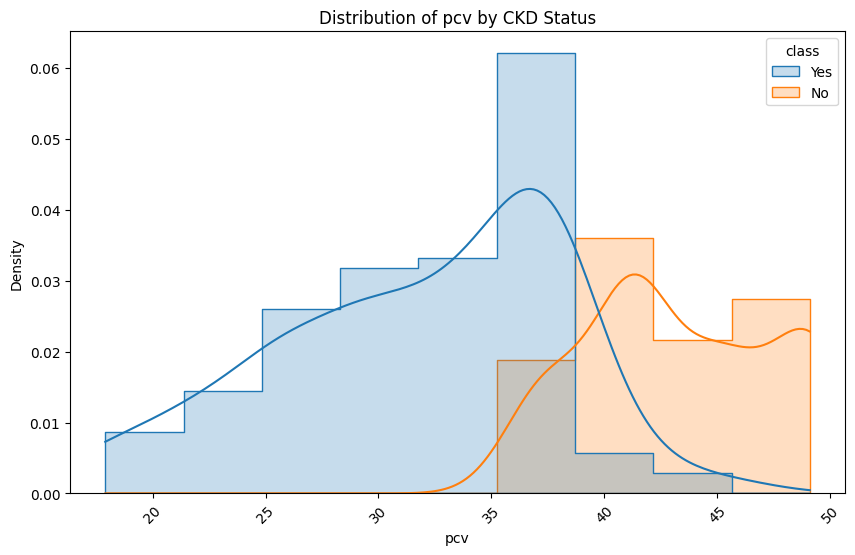
1. Serum Creatinine (sc) is one of the most important markers for assessing kidney function and is a key indicator in diagnosing and monitoring Chronic Kidney Disease (CKD). Elevated levels indicate impaired kidney function, as the kidneys are unable to effectively filter creatinine from the blood. This is seen in the graph as cases with CKD have higher levels of Sc compared to non-CKD cases.



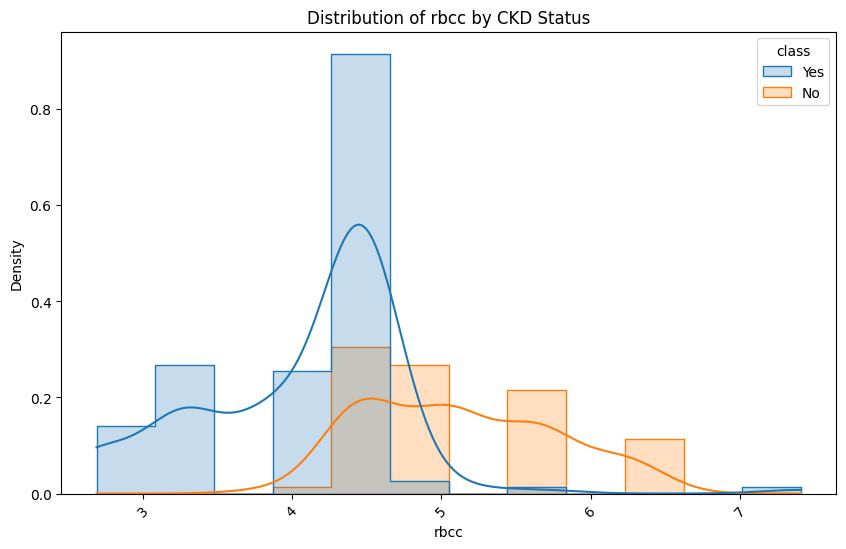
1. Pot (Potassium), measures serum potassium. Elevated levels indicate reduced kidney excretion, posing a risk for cardiac issues. In this study, the levels of Pot are not highly indicative of CKD as both CKD and non-CKD have low pot levels and very few CKD cases have extremely high pot.



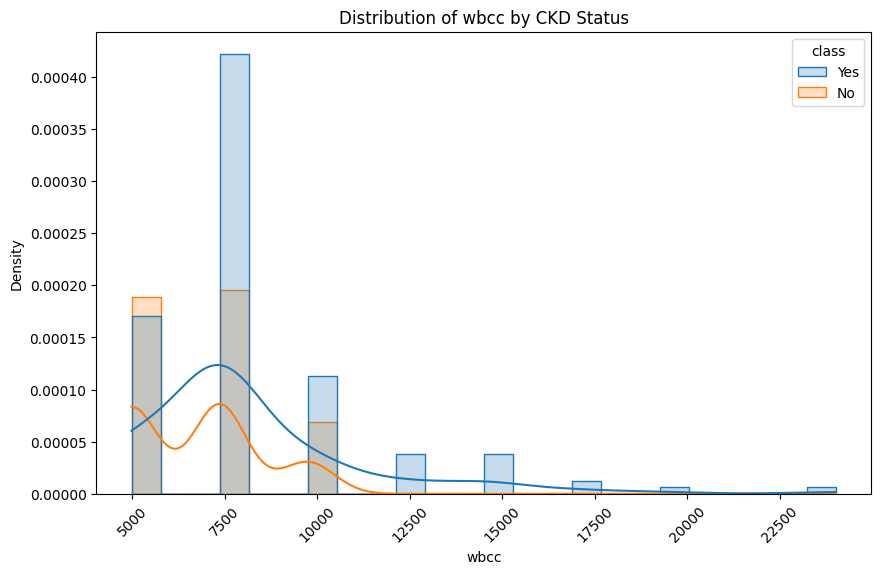
1. Hemo (Hemoglobin) measures blood hemoglobin levels. Low levels indicate anemia which is common in CKD. As displayed in the graph, cases with CKD have significantly lower levels of Hemo compared to non-CKD cases.



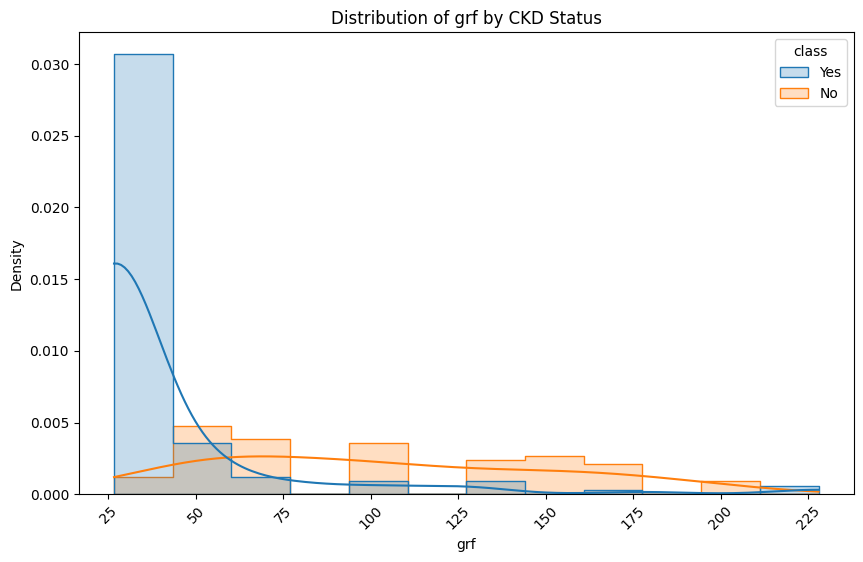
1. Pcv (Packed Cell Volume) reflects the proportion of blood volume occupied by cells. Low values correlate with anemia and CKD severity. As displayed in the graph, cases with CKD have lower PCV compared to non-CKD cases.



1. Rbcc (Red Blood Cell Count) measures RBC concentration. Affected in CKD due to reduced erythropoiesis or loss via urine. As displayed in the graph, cases with CKD have lower Rbcc compared to non-CKD cases.



1. Wbcc (White Blood Cell Count) measures immune cells. Elevated levels suggest infection or inflammation affecting kidneys. As displayed in the graph above, elevated levels of Wbcc are seen in cases with CKD compared to non-CKD.



1. Grf (Glomerular Filtration Rate) estimates filtration efficiency. Low GFR levels are a hallmark of Chronic Kidney Disease (CKD) and indicate reduced kidney function. As displayed in the graph, a high number of cases with CKD have low Grf levels compared to non-CKD cases.

The analysis of categorical and numerical features highlights important patterns related to Chronic Kidney Disease (CKD). Among the categorical features, conditions like high blood pressure (hypertension), abnormal red blood cells (RBCs), abnormal pus cells (PC), the presence of pus cell clumps (PCC), diabetes, coronary artery disease (CAD), pedal edema (PE), and anemia (ANE) were strongly linked to CKD. For example, Stage 2 hypertension, abnormal RBCs, and pus cell clumps were only seen in CKD cases. Some factors, like good appetite and normal pus cells, were more common in non-CKD cases and appeared to protect against CKD.

For numerical features, CKD cases showed clear differences. People with CKD had lower levels of specific gravity (SG), hemoglobin (HEMO), packed cell volume (PCV), and glomerular filtration rate (GFR), indicating poor kidney function. On the other hand, they had higher levels of albumin (AL), blood glucose (BGR), blood urea (BU), serum creatinine (SC), and white blood cell count (WBCC), which signal kidney damage or related health issues. Potassium (POT) levels were less clearly associated with CKD, as they overlapped between CKD and non-CKD cases. Overall, these features provide important clues for identifying CKD and understanding its effects on the body.

#### 

#### Machine Learning Methodology

#### The goal of the analysis is to predict the presence of Chronic Kidney Disease (CKD) using a variety of features. Multiple machine learning models are implemented and evaluated to determine the most effective method for classification.

**Dataset Loading and Preprocessing:**

* The cleaned dataset was loaded
* The pot column was dropped due to its limited relevance.

**Splitting the Dataset:**

* The dataset was split into 70% training data and 30% testing data to ensure a fair evaluation.

**Label Encoding:**

* Categorical features were transformed into numerical values using LabelEncoder.

**Imputation of Missing Values:**

* Missing numerical values were replaced with the mean of the respective column.
* Categorical target variables were filled with their most frequent value.

**One-Hot Encoding:**

* Categorical features were also encoded into binary dummy variables using one-hot encoding to capture more detailed information about categorical attributes.

##### **Classification Models**

I implemented the following machine learning models to classify CKD status:

* **Logistic Regression:** Selected for its simplicity and effectiveness in binary classification problems, making it easy to interpret.
* **Decision Tree Classifier:** Chosen for its ability to capture non-linear relationships and its interpretability through visualizations.
* **K-Nearest Neighbors (KNN):** Evaluated for its simplicity and performance on smaller datasets.
* **Support Vector Machine (SVM):** Applied for its robustness in handling high-dimensional spaces.
* **Naive Bayes:** Tested for its efficiency with categorical and numerical data.

##### **Model Training and Testing**

* The dataset was split into 70% training data and 30% testing data using train\_test\_split
* Each model was trained using the training data to learn patterns and then evaluated on the testing data to assess its performance.

##### **Model Evaluation Metrics**

I used the following metrics to evaluate model performance:

* **Accuracy:** Measures the proportion of correctly classified instances.
* **Precision:** Indicates the proportion of true positive predictions among all positive predictions.
* **Recall (Sensitivity):** Reflected how well the model identified actual positive cases.
* **F1-Score:** Balances precision and recall for a single performance metric, especially useful for imbalanced datasets.

**Machine Learning Outputs**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Class** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
| **Naïve Bayes** | 0 (non-CKD) | 0.87 | 0.87 | 0.87 | 90% |
|  | 1 (CKD) | 0.92 | 0.92 | 0.92 |  |
|  | **Avg** | 0.89 | 0.89 | 0.89 |  |
| **Decision Tree** | 0 (non-CKD) | 1 | 1 | 1 | 100% |
|  | 1 (CKD) | 1 | 1 | 1 |  |
|  | **Avg** | 1 | 1 | 1 |  |
| **Logistic Regression** | 0 (non-CKD) | 1 | 1 | 1 | 100% |
|  | 1 (CKD) | 1 | 1 | 1 |  |
|  | **Avg** | 1 | 1 | 1 |  |
| **KNN** | 0 (non-CKD) | 0.88 | 0.91 | 0.89 | 92% |
|  | 1 (CKD) | 0.94 | 0.92 | 0.93 |  |
|  | **Avg** | 0.91 | 0.92 | 0.91 |  |
| **SVM** | 0 (non-CKD) | 0.85 | 0.74 | 0.79 | 85% |
|  | 1 (CKD) | 0.85 | 0.92 | 0.88 |  |
|  | **Avg** | 0.85 | 0.83 | 0.84 |  |

**Precision**

Precision measures how many of the positive predictions made by the model are actually correct. For example, in the Naïve Bayes model, the precision for CKD (class 1) is 0.92, meaning 92% of the instances predicted as CKD were accurate. High precision indicates that the model is effective at avoiding false positives. However, a lower precision value, such as 0.85 for non-CKD (class 0) in the SVM model, suggests that the model occasionally misclassifies non-CKD cases as CKD.

**Recall**

Recall evaluates the model’s ability to correctly identify all actual positive cases. It reflects how well the model captures true positives from the dataset. For instance, the Decision Tree and Logistic Regression models achieved a perfect recall of 1 for CKD (class 1), correctly identifying all CKD cases. In contrast, SVM’s recall for non-CKD (class 0) is 0.74, indicating that 26% of actual non-CKD cases were missed by the model.

**F1-Score**

The F1-Score is a balance between precision and recall, making it particularly useful when dealing with imbalanced datasets. It provides a single metric that captures both correctness and completeness. In the Naïve Bayes model, the F1-Score for CKD (class 1) is 0.92, showing strong overall performance for this class. On the other hand, the SVM model’s F1-Score of 0.79 for non-CKD (class 0) indicates a less balanced performance for that class.

**Accuracy**

Accuracy represents the overall proportion of correctly classified instances out of the total dataset. Both the Decision Tree and Logistic Regression models achieved 100% accuracy, indicating that every case in the dataset was classified correctly. However, SVM achieved an accuracy of 85%, reflecting its limitations compared to the other models. While accuracy provides a general measure of performance, it may not always be reliable in datasets with class imbalances.

The results show varying levels of effectiveness across models. Decision Tree and Logistic Regression models achieved perfect metrics, demonstrating their strength in this dataset, but they might overfit and require further validation. KNN provided a good balance of precision, recall, and F1-Score, making it a dependable model for classification. The Naïve Bayes model offered solid performance, especially for identifying CKD cases, but fell slightly behind the top performers. SVM struggled with precision and recall for non-CKD cases, indicating potential room for improvement with parameter tuning or feature scaling. These metrics collectively highlight the strengths and weaknesses of each model, providing insights into their suitability for CKD classification.

Best Performance: Logistic Regression and Decision Tree achieve perfect results but may overfit small datasets.

Balanced Trade-off: KNN offers high accuracy and a good balance of precision and recall, making it a robust choice.

Robustness Against Imbalance: Naïve Bayes is dependable, especially in scenarios with more data or noise.

Potential for Improvement: SVM could benefit from fine-tuning or feature scaling to improve recall for class 0.

**Conclusion**

Analysis Limitations:

The analysis had some challenges, mainly related to the quality of the data and assumptions made by the models. The dataset had biases caused by missing values, uneven representation of features, and imbalanced classes. Filling in missing values with averages or the most common values might have oversimplified the data, affecting the reliability of the results. Also, grouping numerical values into ranges instead of using exact numbers could have reduced accuracy. For the models, Decision Trees risked overfitting without proper adjustments, and Naive Bayes assumed that features were independent, which might not be true for this dataset.

Study Implications:

Despite these challenges, the study provides valuable healthcare insights, emphasizing the strong link between CKD and conditions like diabetes and hypertension, which supports integrated management of these comorbidities. Models such as Logistic Regression and Decision Trees offer interpretable predictions that can aid clinicians in early CKD detection and treatment planning. The findings also have policy implications, guiding public health efforts to focus on high-risk groups and allocate resources more effectively.

Ethical Considerations:

Ethical considerations, including data privacy and compliance with regulations like HIPAA, are essential to ensure patient confidentiality. It is equally important to avoid over-reliance on model predictions without thorough clinical validation and to communicate the limitations of the findings to prevent misuse.

Project Continuality:

To ensure project continuity, future work should validate the findings using additional datasets and include data like biomarkers or imaging to improve prediction accuracy. Long-term monitoring models could track CKD progression and adapt to new data for consistent accuracy. Addressing dataset biases and improving generalizability to other populations would enhance the study’s impact. Further, comparing models rigorously, creating intuitive visualizations like decision tree plots, and collaborating with stakeholders, including healthcare professionals and policymakers, would align the analysis with real-world needs, ensuring the work evolves into a comprehensive and impactful tool for CKD prediction and management.Top of Form

Bottom of Form

**Final Remarks**

This study demonstrates the effectiveness of machine learning models in predicting Chronic Kidney Disease (CKD) and identifying its key risk factors. Logistic Regression and Decision Tree models achieved perfect classification metrics, while K-Nearest Neighbors provided a balanced performance. Naïve Bayes was reliable for noisy data, and Support Vector Machines showed potential for improvement. Despite limitations like missing values, class imbalances, and oversimplified numerical ranges, the analysis highlights strong associations between CKD and conditions like diabetes and hypertension, offering valuable insights for early detection and integrated care.

Future work should validate these findings on diverse datasets, incorporate additional data sources, and address dataset biases to improve generalizability. Ethical considerations, including data privacy and transparent communication of limitations, remain critical. By integrating these models into healthcare strategies and collaborating with stakeholders, this research contributes meaningfully to CKD prediction and management.

# References

1. Islam, M., & Akter, S. (2020). Risk Factor Prediction of Chronic Kidney Disease [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5WP64>
2. Alaboudy, A., et al. (2021). Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. Journal of Medical Systems.
3. Ganaie, S. M., et al. (2022). A Machine Learning Driven Nomogram for Predicting Chronic Kidney Disease Stages 3–5. Scientific Reports.
4. Patel, A., et al. (2021). Chronic Kidney Disease Prediction Using Machine Learning. Semantic Scholar.
5. Zhang, Y., et al. (2020). Machine Learning Improves Upon Clinicians' Prediction of End Stage Kidney Disease. PubMed.
6. Kumar, A., et al. (2019). Chronic Kidney Disease (CKD) Prediction by Supervised Machine Learning Techniques. Semantic Scholar.
7. Shafi, T., & Coresh, J. (2021). Chronic Kidney Disease: Definition, Epidemiology, and Overview of Markers. *American Journal of Kidney Diseases.* https://doi.org/10.1053/j.ajkd.2021.05.010