ML Project using R

Machine Learning Project on:

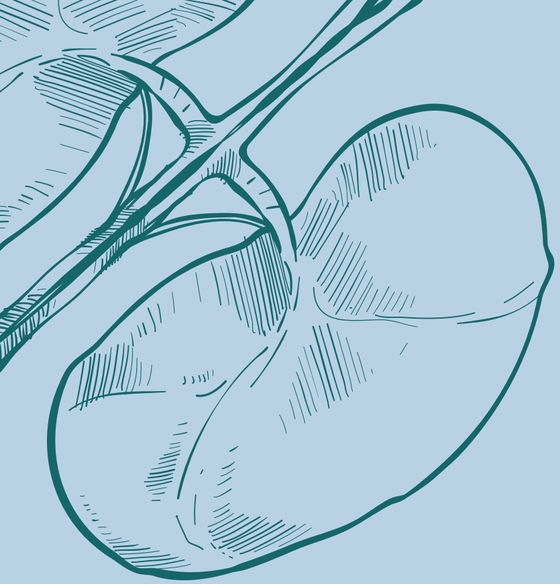
Chronic Kidney Disease Prediction

Aim of Project

To support Singapore's healthcare institutions in addressing the growing challenge of **Chronic Kidney Disease (CKD)** through **descriptive and predictive analytics**

1. Introduction to the Business Problem

- Recognising rise in CKD cases nationally & its economic and social implications on healthcare system and workforce
- Acknowledging the public health issue of CKD prevalence as a significant business opportunity



Rise in Prevalence of Chronic Kidney Disease (CKD)

“Singapore ranks

2nd globally

in Prevalence of CKD”

“**1/4**

of Singapore's adult population will suffer from CKD by **2035**”

- National Kidney Foundation Singapore, 2024

Impacts of Rise in CKD

Private

Reduced productivity levels

Public

Increased resource allocation for healthcare sector

Economic losses

Reduced quality of life

Justification for Business Opportunity

descriptive & predictive analytics

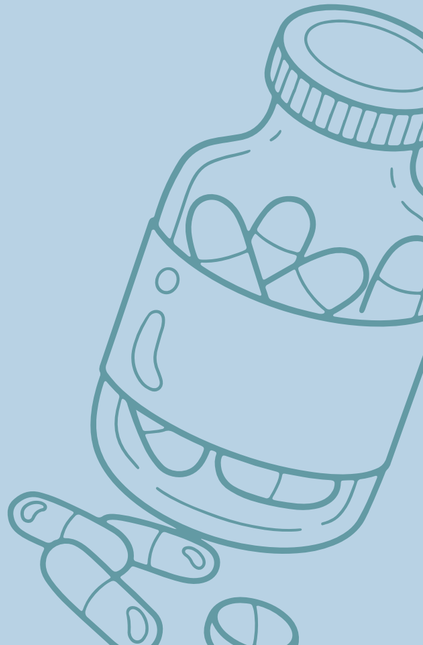
↙ ↓ ↘

rise in CKD

as a national healthcare challenge

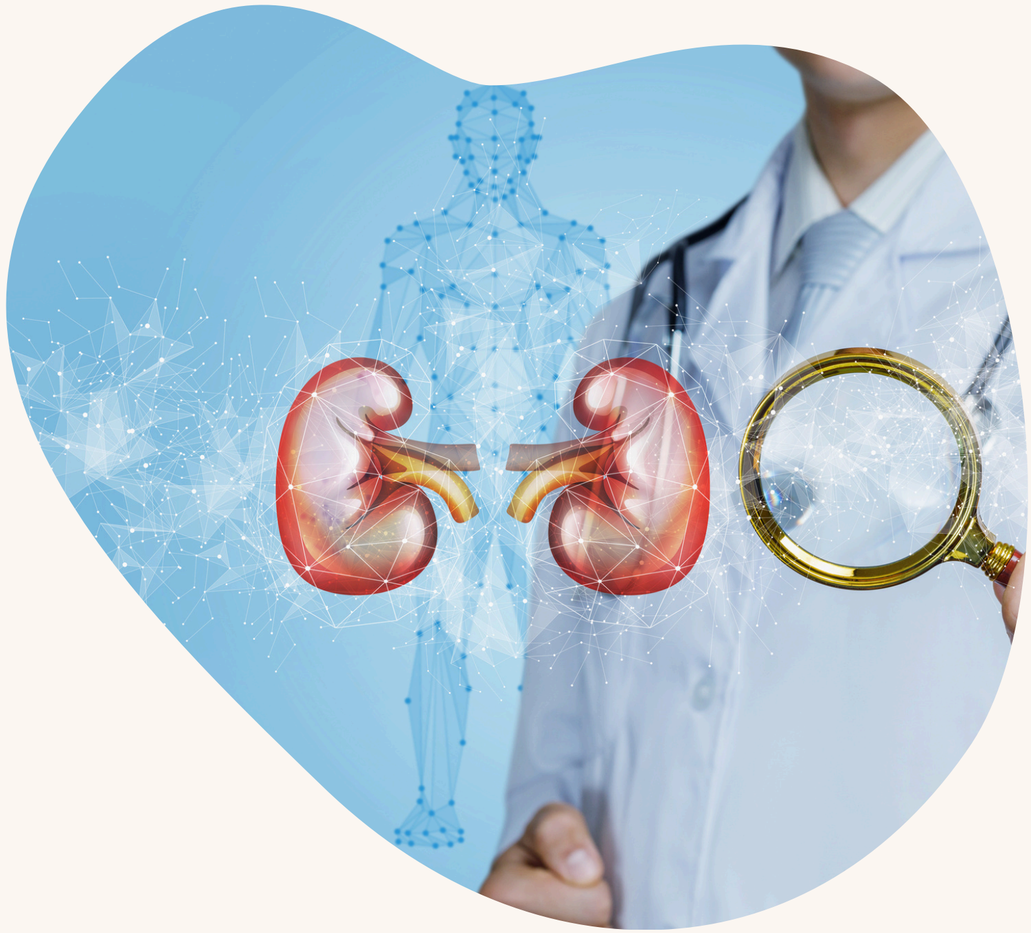
Aligning with Healthier SG initiative:

- to **mitigate spread** of CKD
- to **improve** population's **health**



2. Analysing the Current Solutions

Exploring and analysing the existing methods for diagnosing Chronic Kidney Disease (CKD), determining their effectiveness and their limitations that create a business opportunity



What are the current solutions?

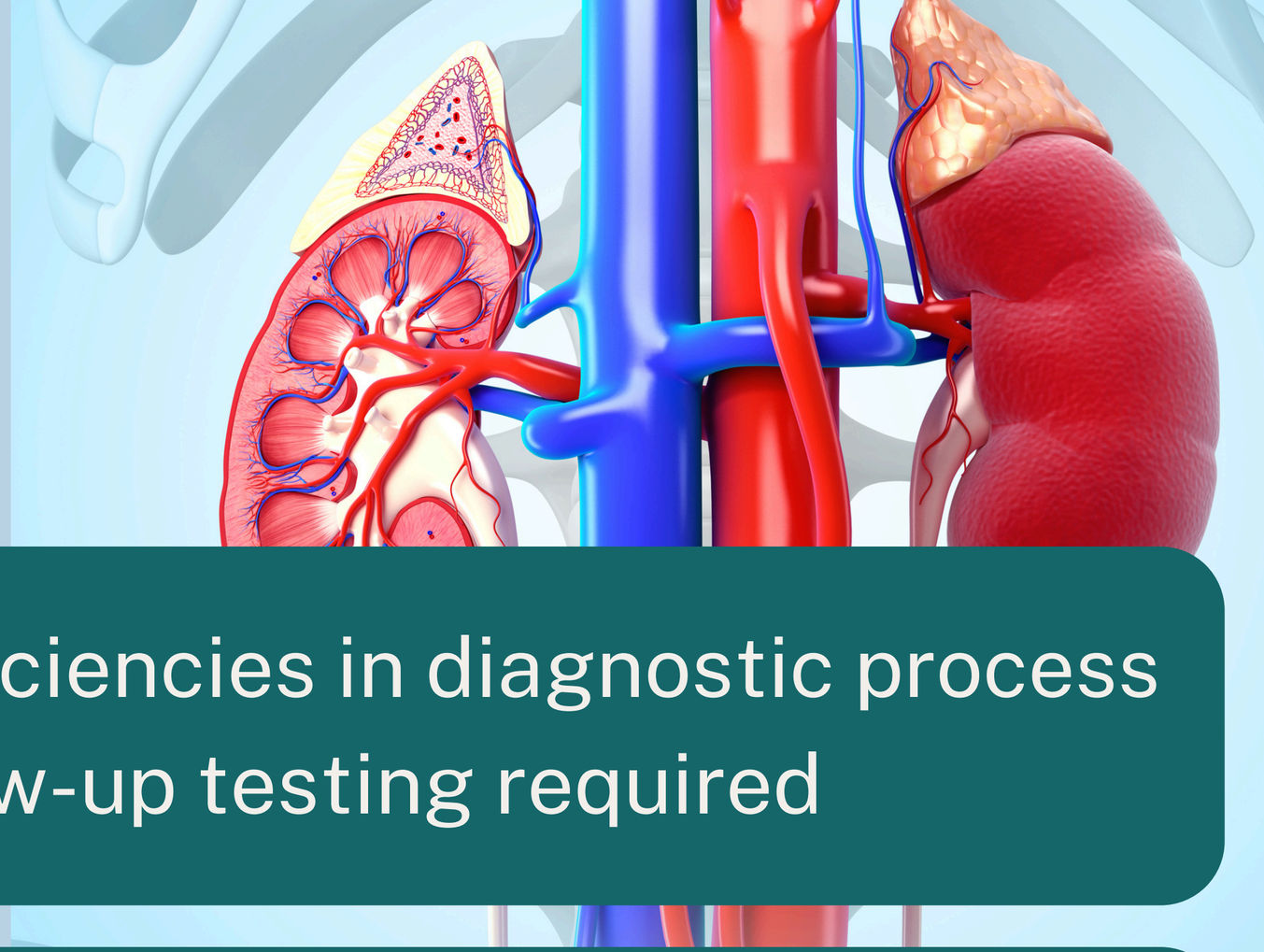
Urine test

Measures protein leakage in the urine

Blood test

Measures amount of serum creatinine in the blood

Limitations of Current Solutions



Urine test

- Inefficiencies in diagnostic process
- Follow-up testing required

Blood test

- Invasive procedure
- Results in non-definitive diagnosis

Capitalising on Business Opportunity

More conclusive and non-invasive daignosis procedure

Global integration

Reduce healthcare spendings

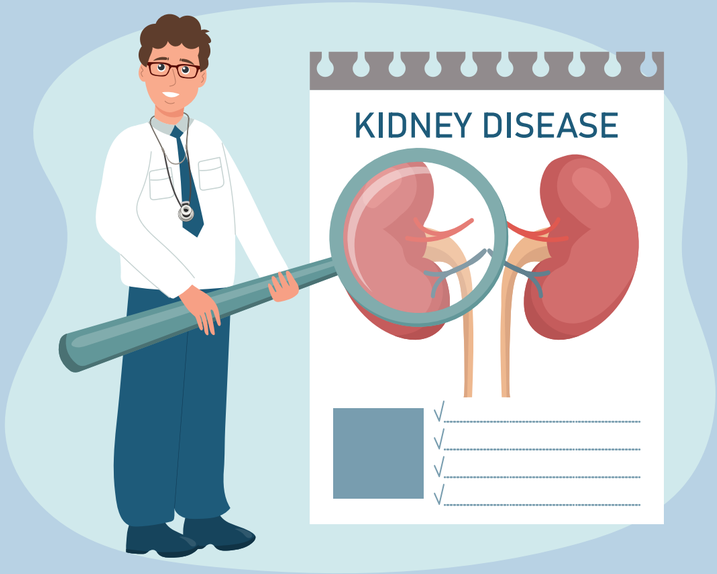
In light of Budget 2024, Singapore aims to enhance the use of predictive analytics to strengthen preventive care for patients.

3. Data Cleaning & Data Exploration

Before building our model, data cleaning and preparation were conducted to ensure greater analysis and accuracy. Followed by exploration of our dataset before diving into the solutions.

Data Cleaning

Our initial data set has 27 columns and 400 rows



01. Removing irrelevant columns

removed column 1 and 27

02. Convert missing values

converted missing values to “NA”

Data Preparation

Our final cleaned data set has
25 columns and 400 rows

01. Renaming of variables

for easier readability and understanding

02. Converting variables

converted some variables to numeric for better computing

03. Addressing mistakes

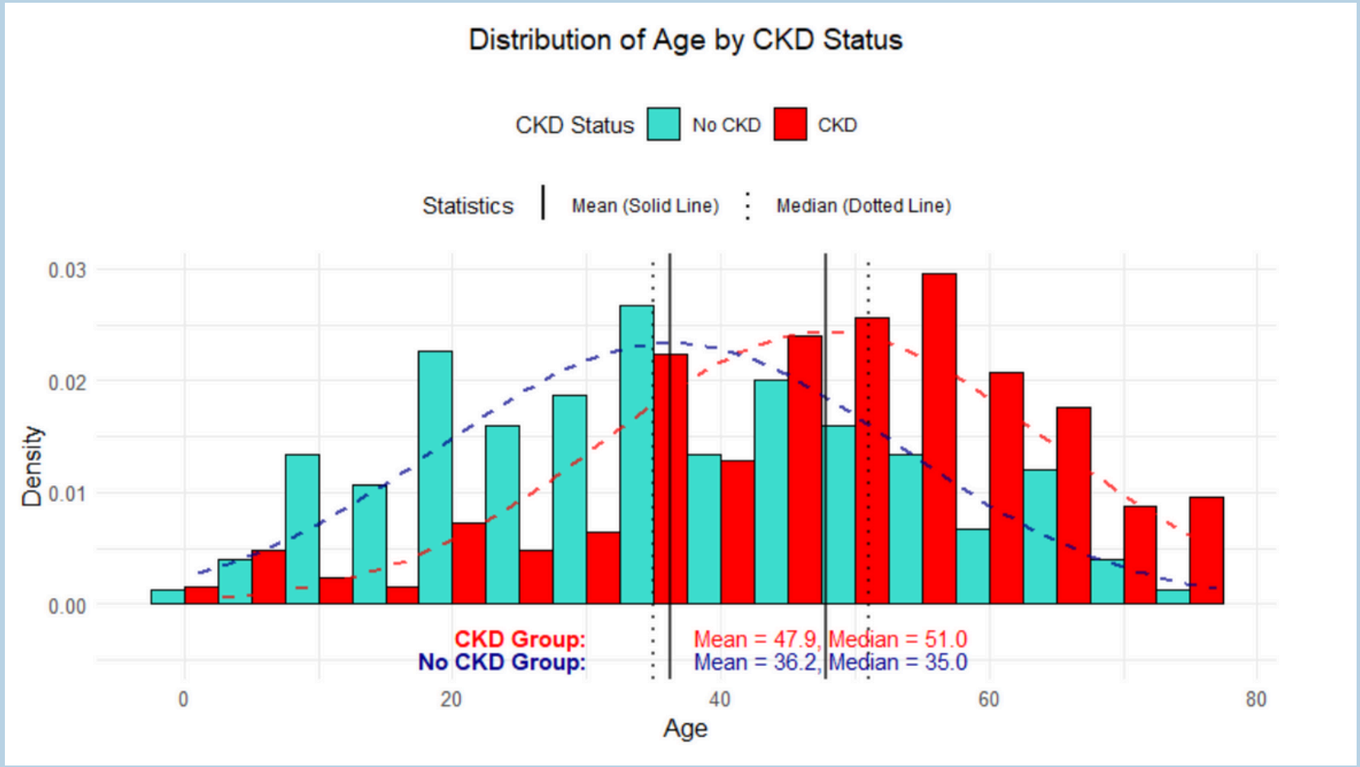
1. Wrong labelled data
2. Empty string in ‘diabetes_mellitus’
3. Removed ‘?’ from factor columns

Histogram: Age (age) by Class (CKD VS Non-CKD)

To see the distribution of different variables between CKD and nonCKD patients

Non-CKD : Lower median and mean age
CKD : Higher median and mean age

Risk of CKD increases with age

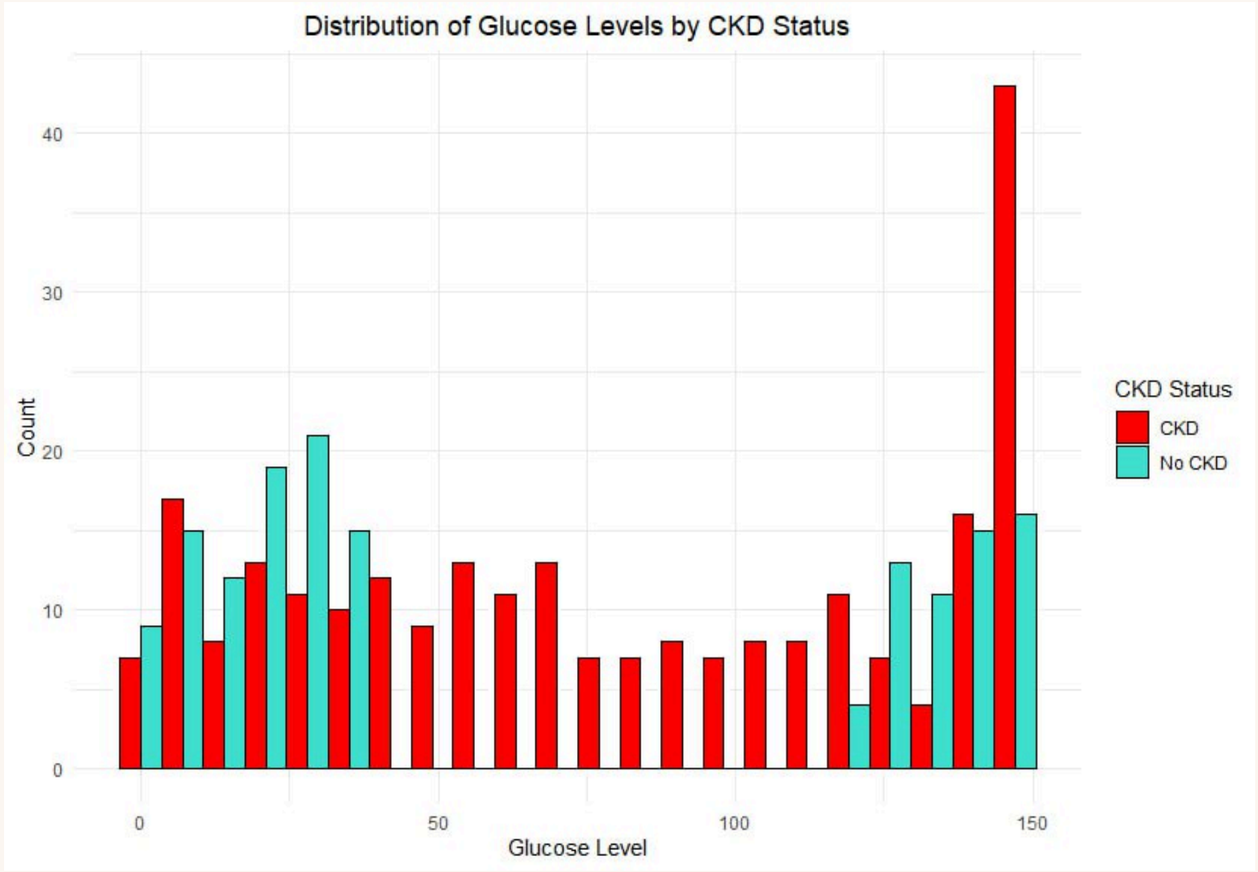


Histogram: Glucose Levels (bgr)

by Class (CKD VS Non-CKD)

Non-CKD : More spread at lower glucose levels
CKD : Significant spike at high glucose levels

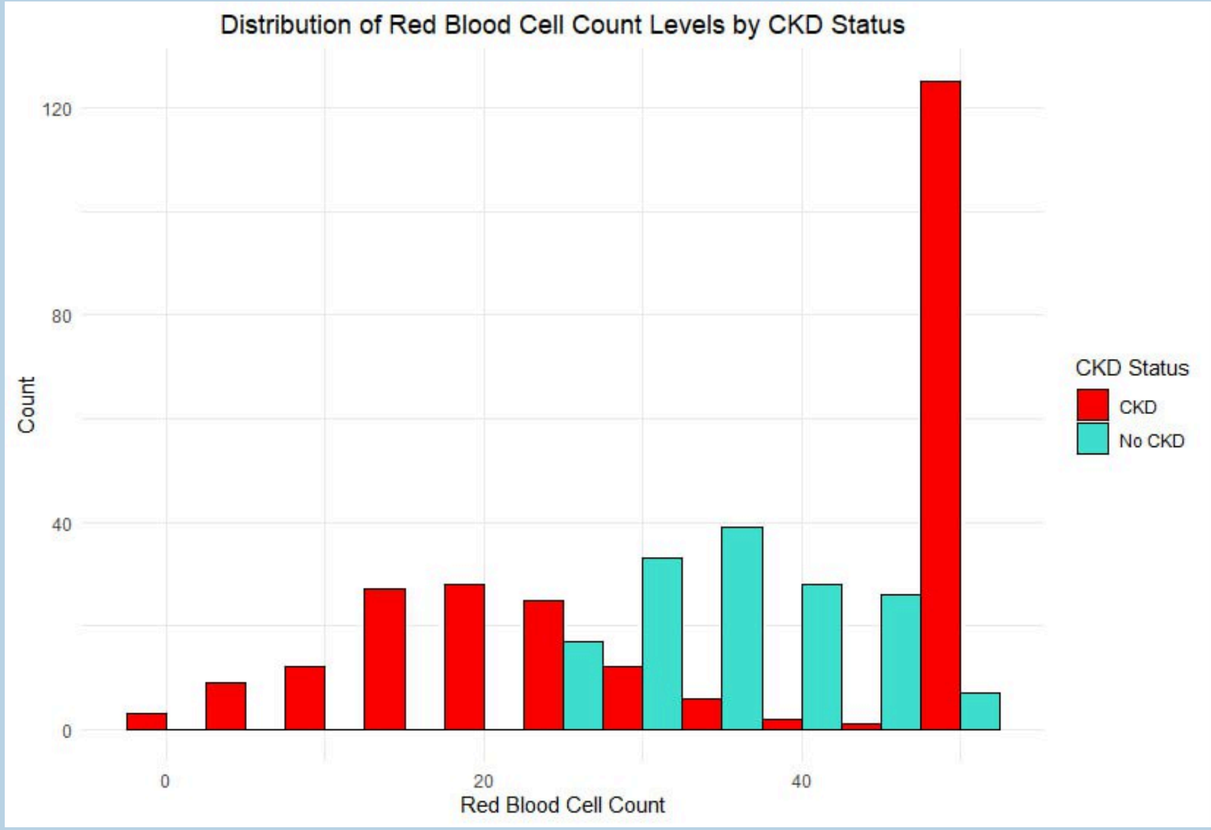
CKD is often linked with higher glucose levels



Histogram:
RBC Count (rbcc)
by Class (CKD VS Non-CKD)

Non-CKD : Even spread of RBC level
CKD : Significant spike at high RBC level, below 20 RBC level only has CKD patients

CKD is often linked with **very high** or **very low** red blood cell count



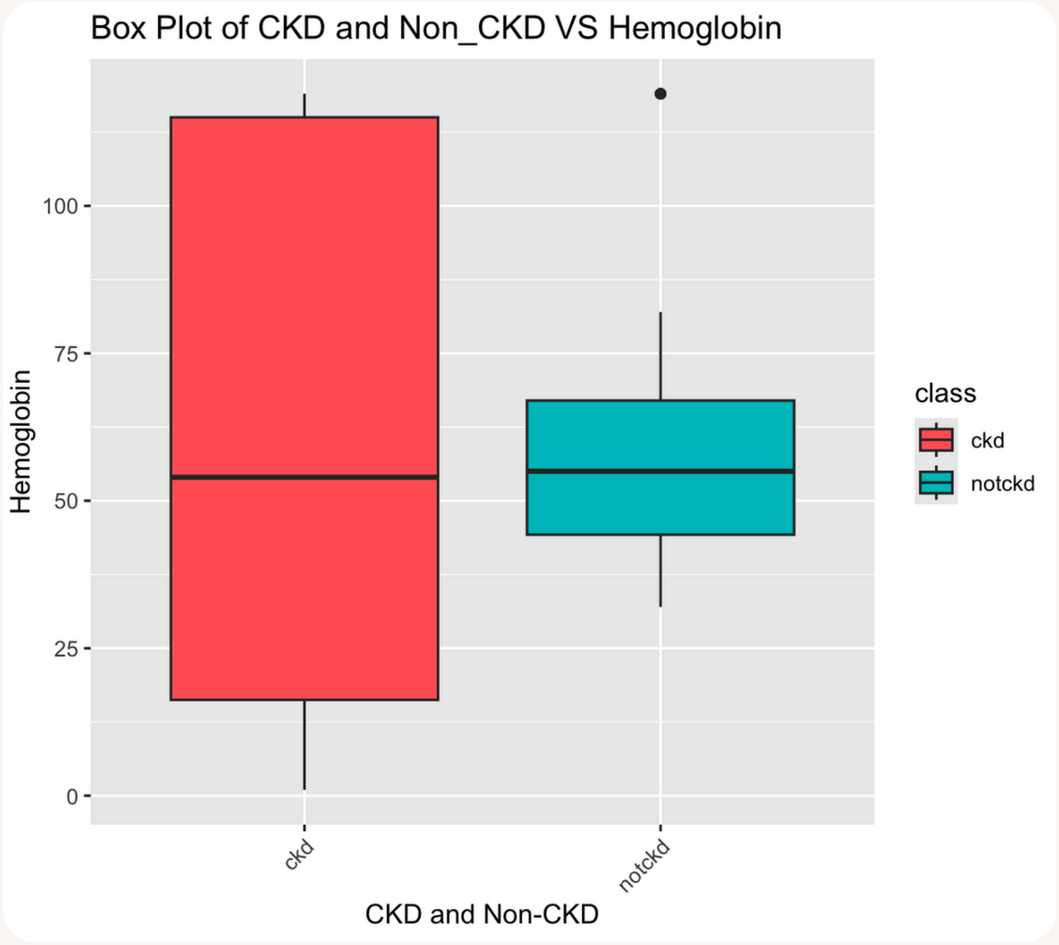
Boxplot: Hemoglobin

by Class (CKD VS Non-CKD)

CKD: Wide spread, as indicated by the large IQR of 98.8.

Non-CKD: More concentrated, with a smaller IQR of 22.8

The wider spread suggests that hemoglobin levels is highly associated with CKD.



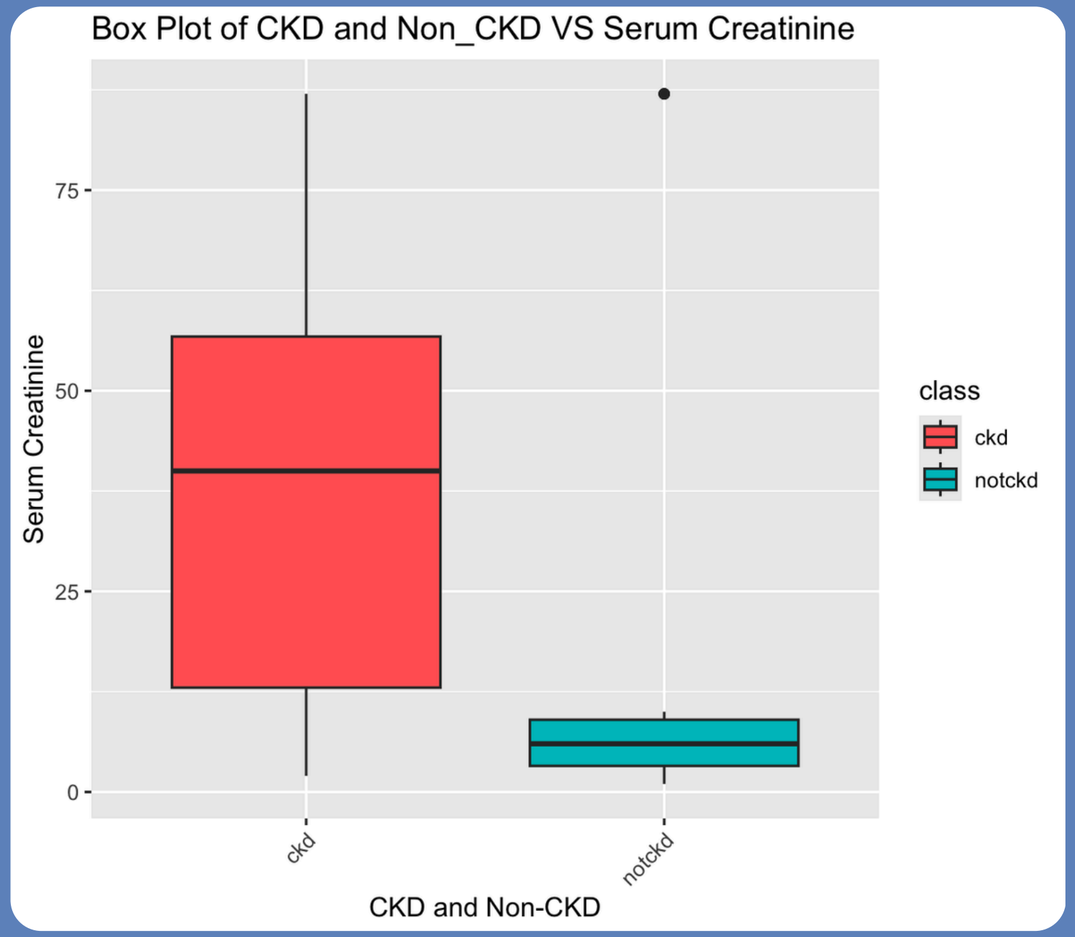

```
# A tibble: 2 × 7
  class    IQR Quartile1 Quartile3 Median   Max   Min
  <fct>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>
1 ckd     98.8      16.2       115     54   119     1
2 notckd  22.8      44.2        67     55   119    32
```

Boxplot: Serum Creatinine (sc)

by Class (CKD VS Non-CKD)

- **CKD:** Significant wide spread, IQR 43.8
- **Non-CKD:** IQR 5.75
- The large difference in IQRs and medians between the two groups provides a clear distinction

A typical physiological markers:
Higher serum creatinine often correlates with impaired kidney function




```
# A tibble: 2 × 7
  class     IQR Quartile1 Quartile3 Median   Max   Min
  <fct>   <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>
1 ckd     43.8       13        56.8     40    87     2
2 notckd   5.75       3.25      9        6    87     1
```

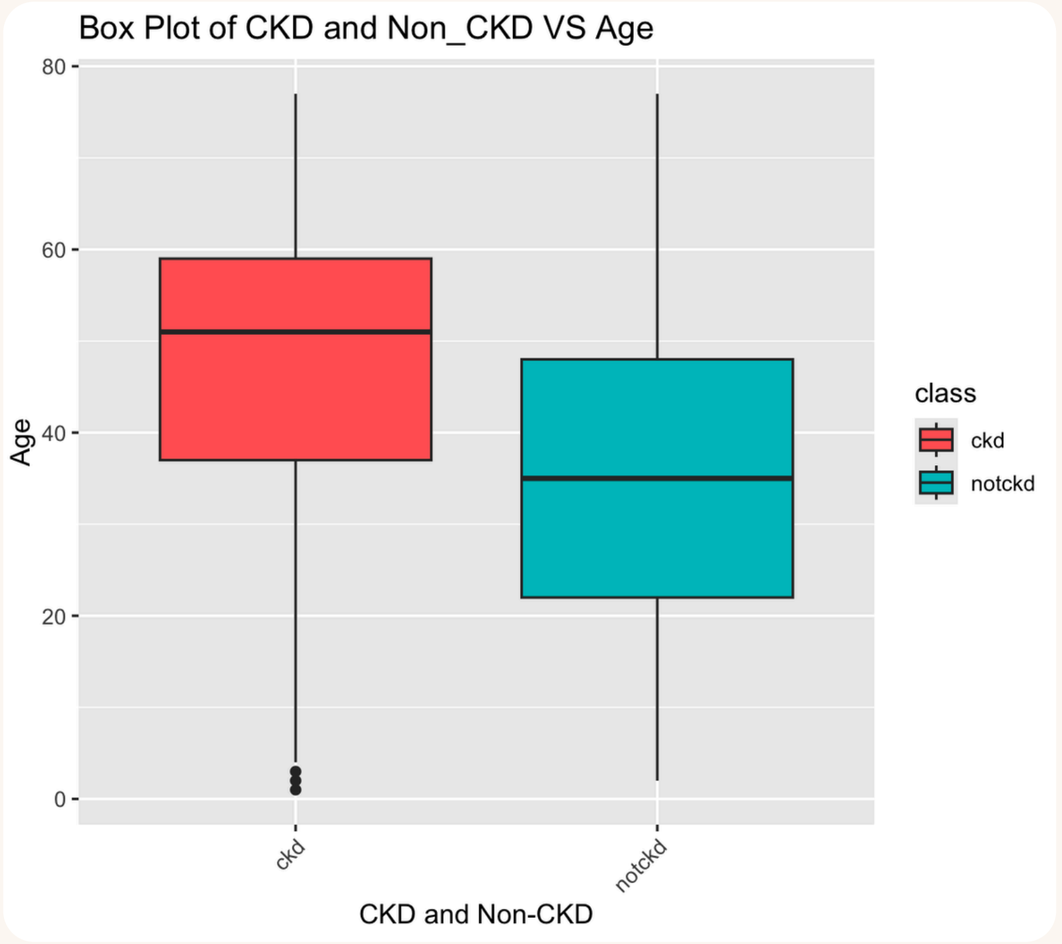
Boxplot: Age

by Class (CKD VS Non-CKD)

CKD: Higher median range of age 51

Non-CKD: Lower median age of 35

CKD is more prevalent in older age groups, aligning with the typical understanding that CKD risk increases with age.

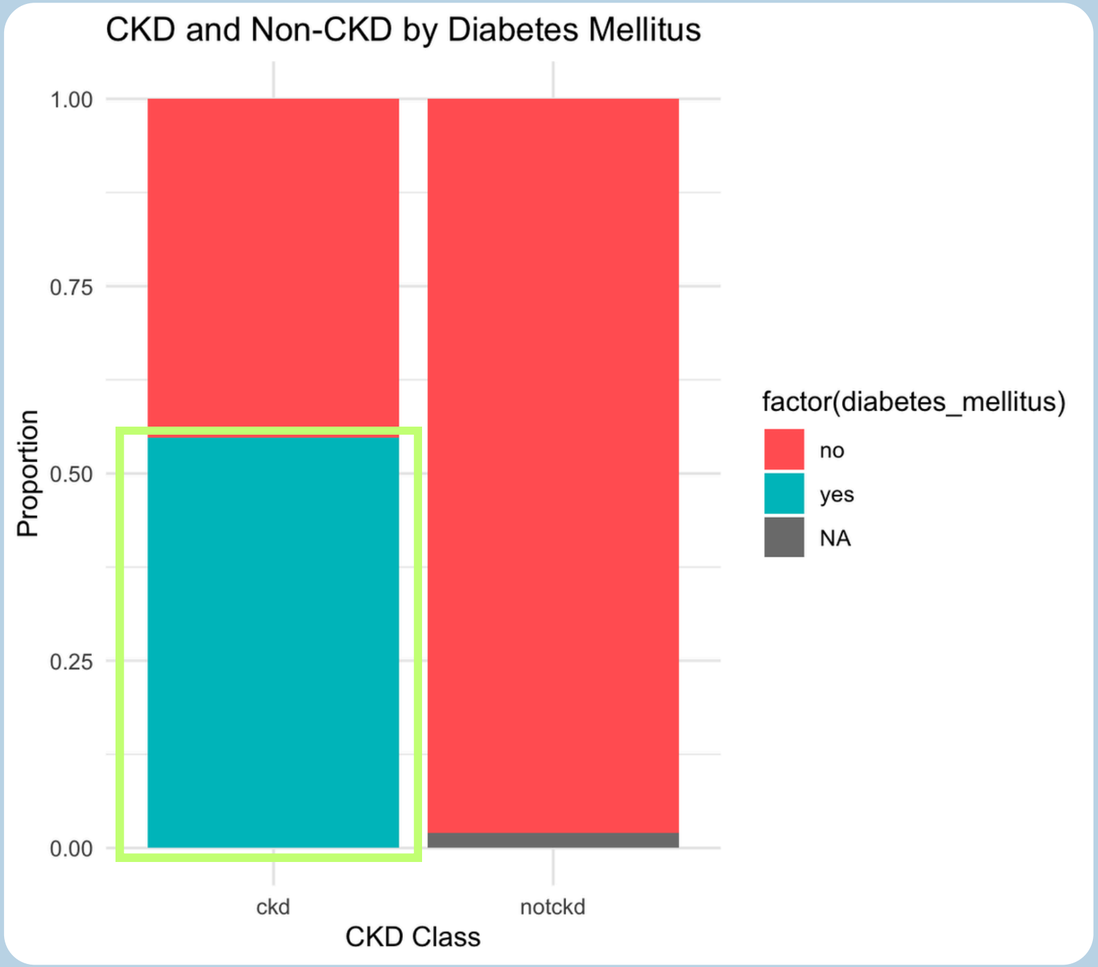



```
# A tibble: 2 × 7
  class    IQR Quartile1 Quartile3 Median   Max   Min
  <fct>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>
1 ckd       22        37        59     51    77     1
2 notckd    26        22        48     35    77     2
```

Boxplot: Diabetes Mellitus (dm)

by Class (CKD VS Non-CKD)

- **> 50%** of CKD patients are diabetic
- **All non-CKD individuals are not diabetic**
- This suggests that diabetes may be a common comorbidity or risk factor associated with CKD



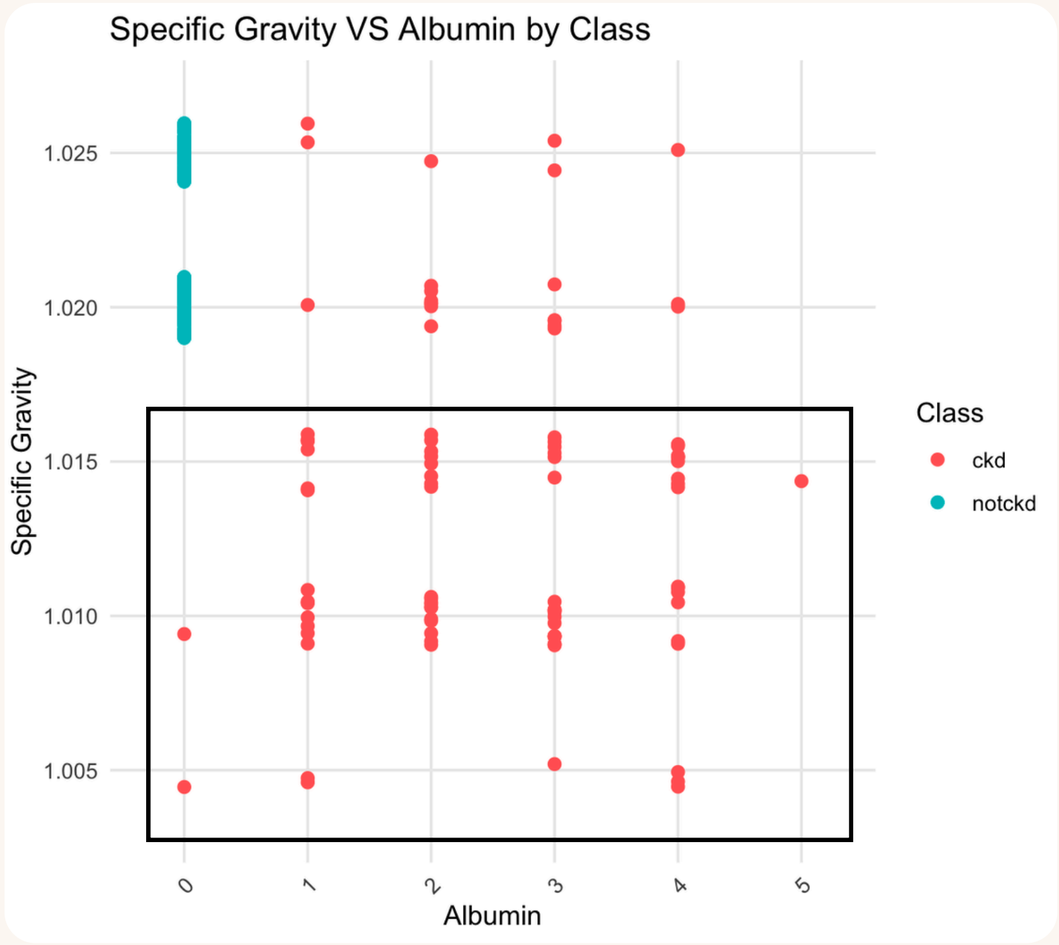
Diabetes Mellitus is highly associated with CKD which supports the CART Decision tree whereby it is one of the split indicators

```
# A tibble: 2 × 7
  class    IQR Quartile1 Quartile3 Median   Max   Min
  <fct>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>
1 ckd    43.8       13        56.8     40    87     2
2 notckd  5.75       3.25      9        6    87     1
```

Jitterplot: Specifc Gravity VS Albumin

by Class (CKD VS Non-CKD)

- Clustering of CKD cases at **lower specific gravity values**
- Non-CKD cases appear mostly at the **left top end** of the specific gravity range with lower albumin levels

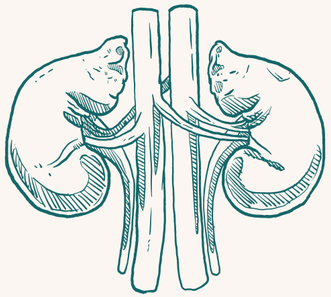


Lower specific gravity and higher albumin levels are observed more frequently among CKD patients

```
# A tibble: 2 × 7
  class    IQR Quartile1 Quartile3 Median   Max   Min
  <fct>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>
1 ckd       22        37        59     51    77     1
2 notckd    26        22        48     35    77     2
```

Correlation Heatmap Between Numeric Features

with p-value



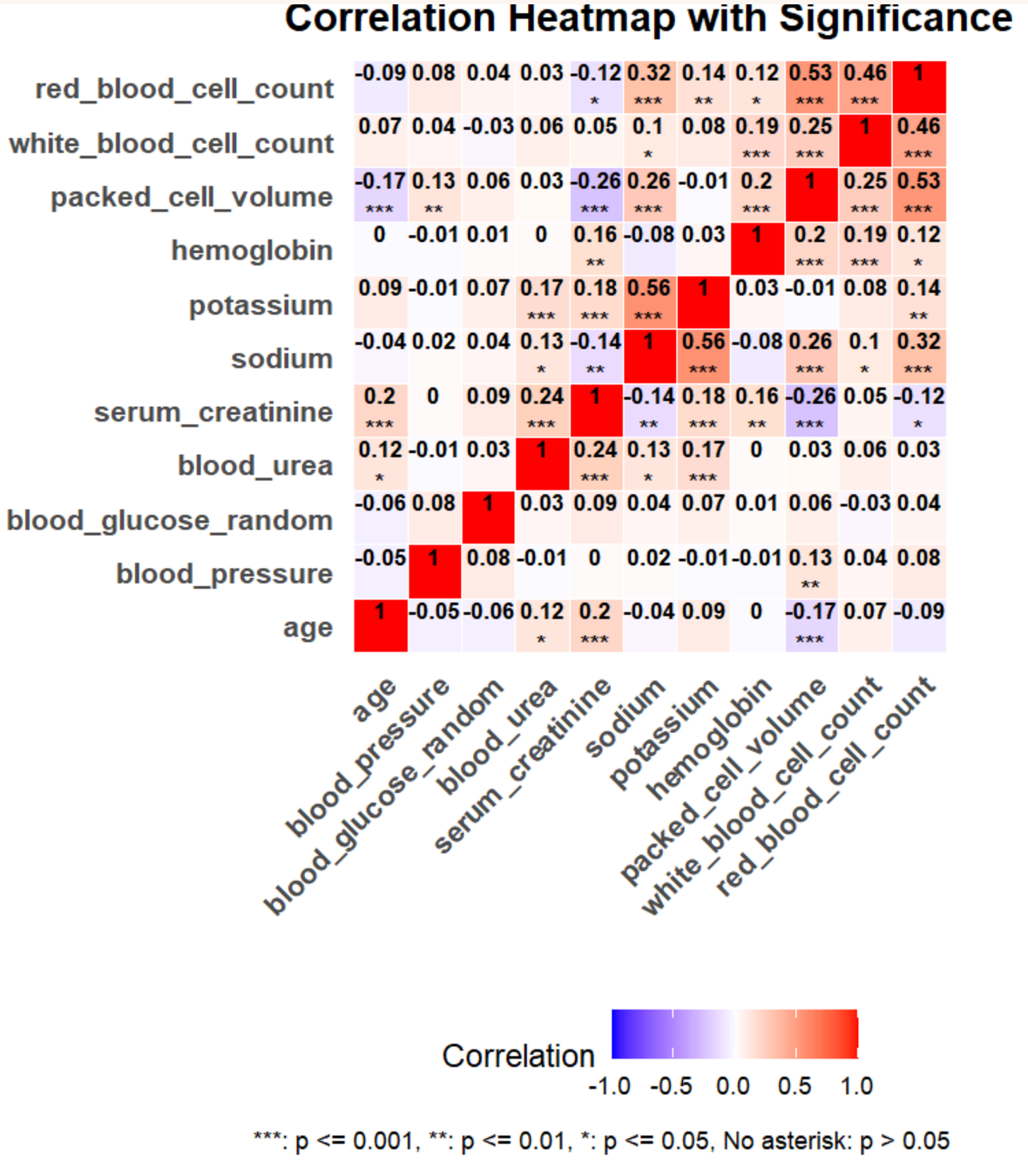
- Visualise relationships between numerical factors
- Calculated p-value to determine statistical significance

Red blood cell & Packed Cell Volume

- **Strong positive correlation** (***)

Serum Creatinine & Packed Cell Volume

- **Strong negative correlation** (***)



4. Analytics Approach

Logistic Regression & CART

To identify the possible predictors that will cause Coronary Kidney Disease (CKD) and help hospitals to identify CKD patients more effectively.
We will be using **2 main machine learning models: Logistic Regression** and **Classification and Regression Trees** (**CART**) **model.**

Analytics Approach

Proposed Solutions

01.

Analyze past patient records and data to train our model

02.

Build a model by identifying high-risk factors that increase chances of CKD

03.

NKF and local hospitals to direct diagnosis efforts towards these high-risk patients, reducing the overall cost of diagnosis.

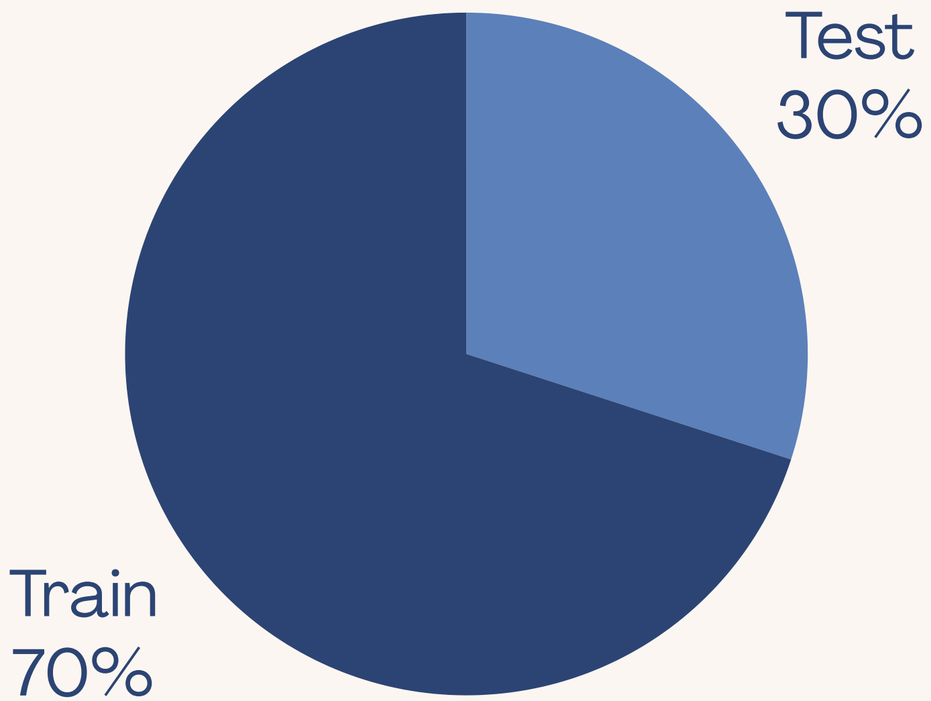
Logistic Regression

Data Splitting
&
Model Building

Result
Interpretation

Metrics
Optimization

70-30 Train-Test Split



Two Main Classes

250 == ‘ckd’

150 == ‘notckd’

70% of each

30% of each

Train Set

175 == ‘ckd’

105 == ‘notckd’

Test Set

75 == ‘ckd’

45 == ‘notckd’

Variable Selection

```
Call:
glm(formula = class ~ age + serum_creatinine + potassium + packed_cell_volume +
    red_blood_cell_count, family = binomial, data = train_set)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.41661    1.05220  -3.247  0.00117 **
age                  0.03080    0.01108   2.781  0.00541 **
serum_creatinine     0.11850    0.02326   5.094  3.5e-07 ***
potassium            0.06237    0.01621   3.847  0.00012 ***
packed_cell_volume  -0.08810    0.02786  -3.163  0.00156 **
red_blood_cell_count 0.05938    0.02140   2.775  0.00552 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.48  on 279  degrees of freedom
Residual deviance: 200.19  on 274  degrees of freedom
AIC: 212.19

Number of Fisher Scoring iterations: 7
```

Five Significant Variables Selected:

- x1 Age
- x2 Serum Creatinine
- x3 Potassium
- x4 Packed Cell Colume
- x5 Red Blood Cell Count

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_5 x_5$$

$$\frac{P}{1-P}(odds) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_5 x_5}$$

Every 1 Unit Increase of Variable X = Odds Increase by e^β

Variable Selection

Every 1 Unit Increase of Variable X = Odds Increase by **e^β**

- 1 mg/dl Increase in Serum Creatinine = **12.58% Increase in Odds** to have **CKD**

Variable Name	Coefficients	Odds of CKD (per 1 unit)
(Intercept)	-3.41661	-
Age	0.03080	+3.13%
Serum_Creatinine	0.11850	+12.58%
Potassium	0.06237	+6.44%
Packed_Cell_Volume	-0.08810	-9.21%
Red_Blood_Cell_Count	0.05938	+6.12%

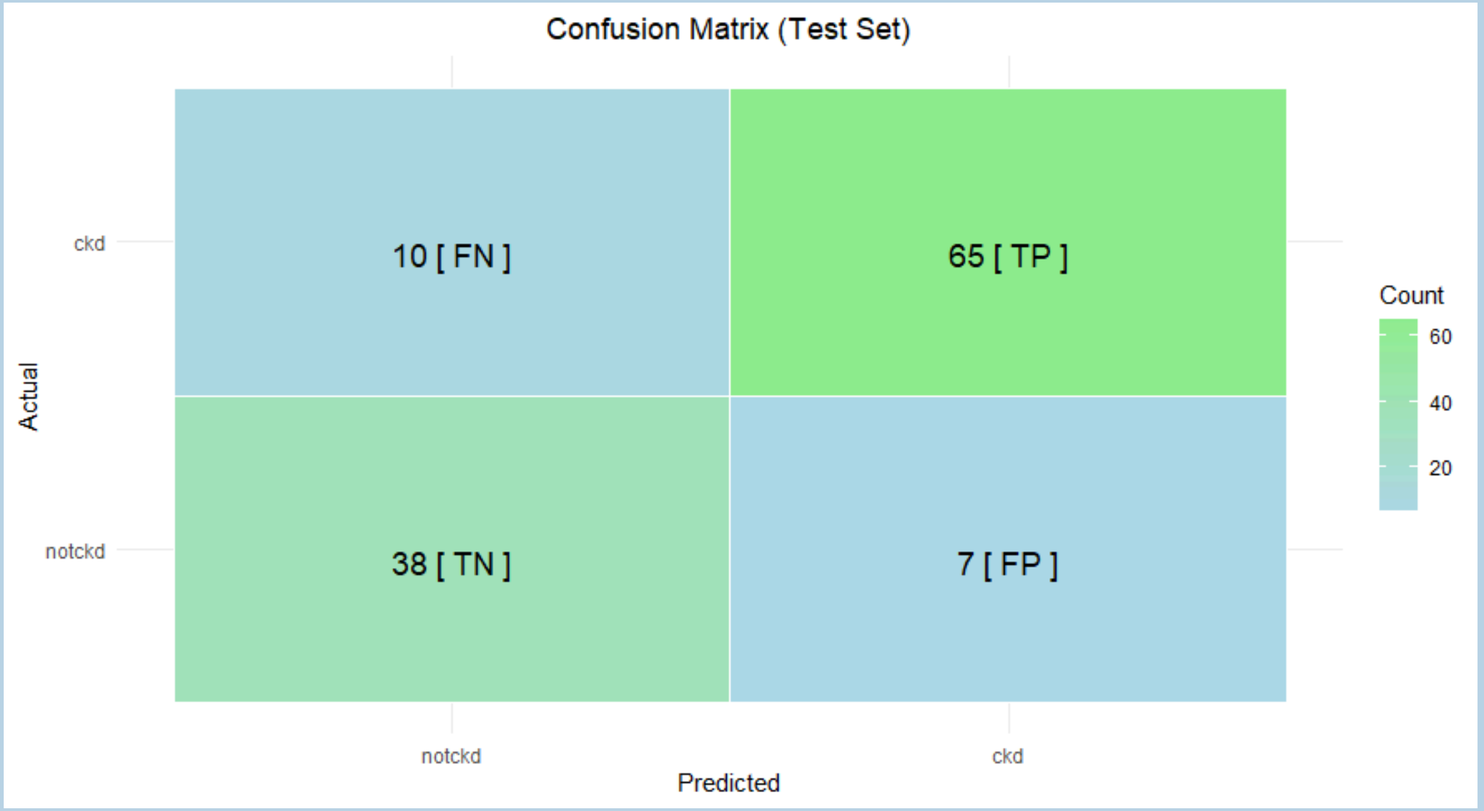
*Multicollinearity Check:

VIF < 10

```
> vif(lm1)
                       age
                  1.032003
           serum_creatinine
                  1.066074
                 potassium
                  1.062777
       packed_cell_volume
                  1.575120
red_blood_cell_count
                  1.687345
```

Result Interpretation

Confusion Matrix: Recall Optimization



Model Metrics:

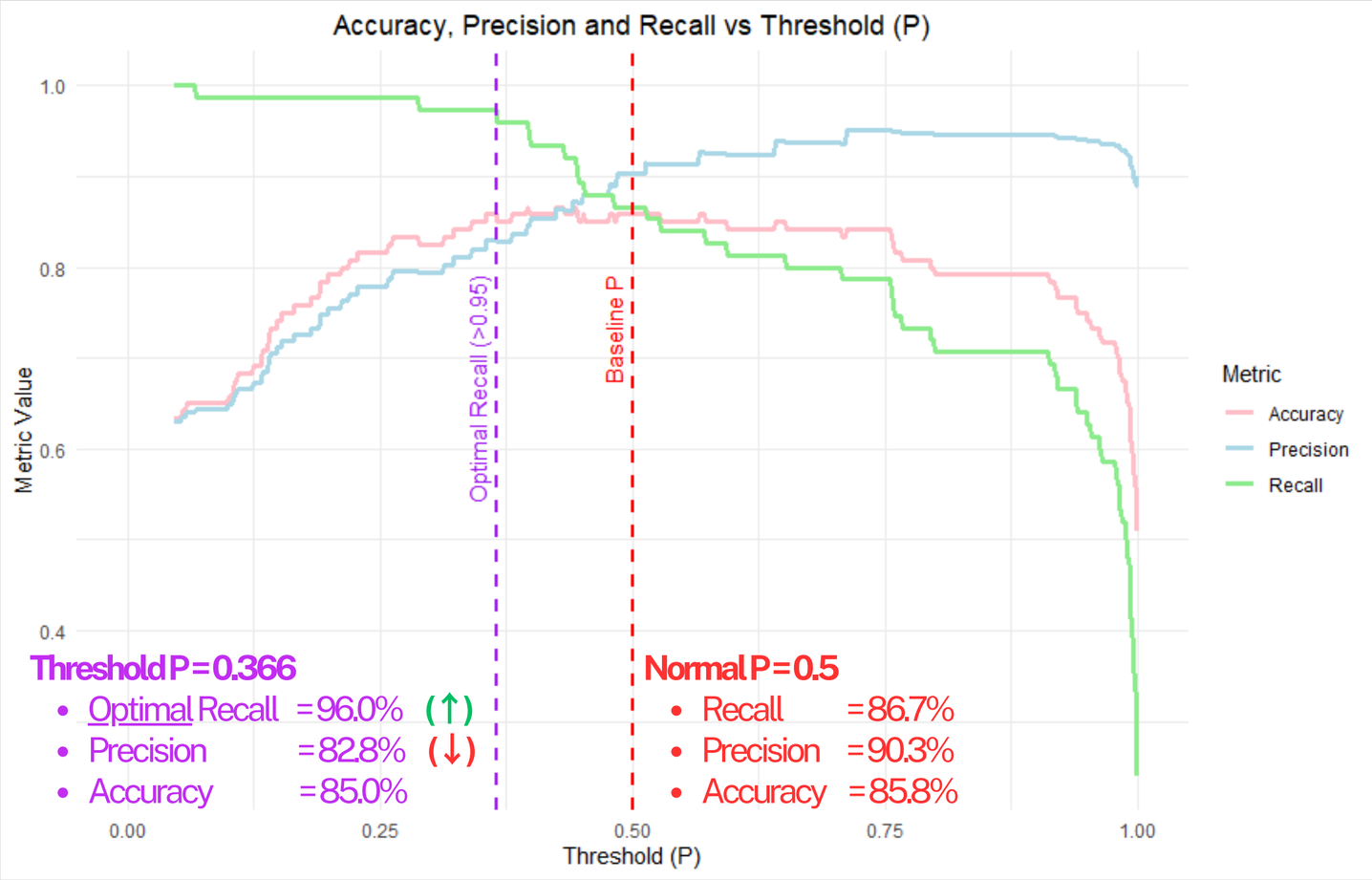
Accuracy $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$	$\frac{65+38}{65+38+7+10}$ **= 85.83%**
Precision $\left(\frac{TP}{TP+FP}\right)$	$\frac{65}{65+7}$ **= 90.28%**
Recall $\left(\frac{TP}{TP+FN}\right)$	$\frac{65}{65+10}$ **= 86.67%**

Recall Metric:
Out of All **Actual Positive Cases**,
How many are **Predicted Correctly**?

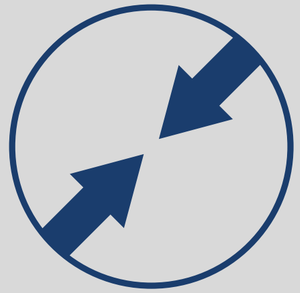
Every missed diagnosis (FN) results in costly, long-term treatments for CKD

Metrics Optimizations

Precision vs Recall Tradeoff



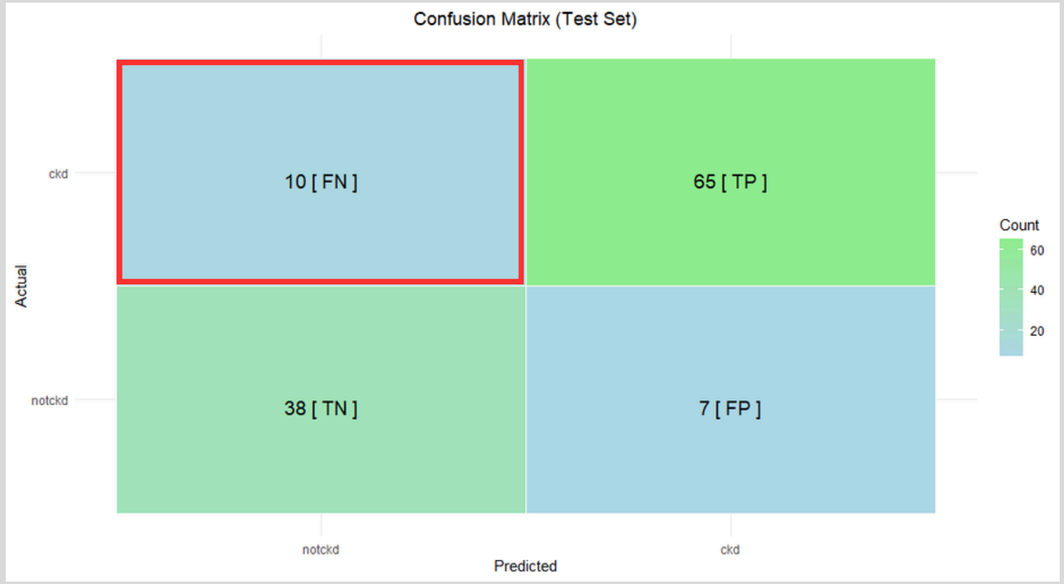
- Usually, Threshold P = 0.5
- Decreasing Threshold P means lower P required to be classified as 'CKD'
  - Higher Chance to be 'CKD'
  - **Conservative** Model

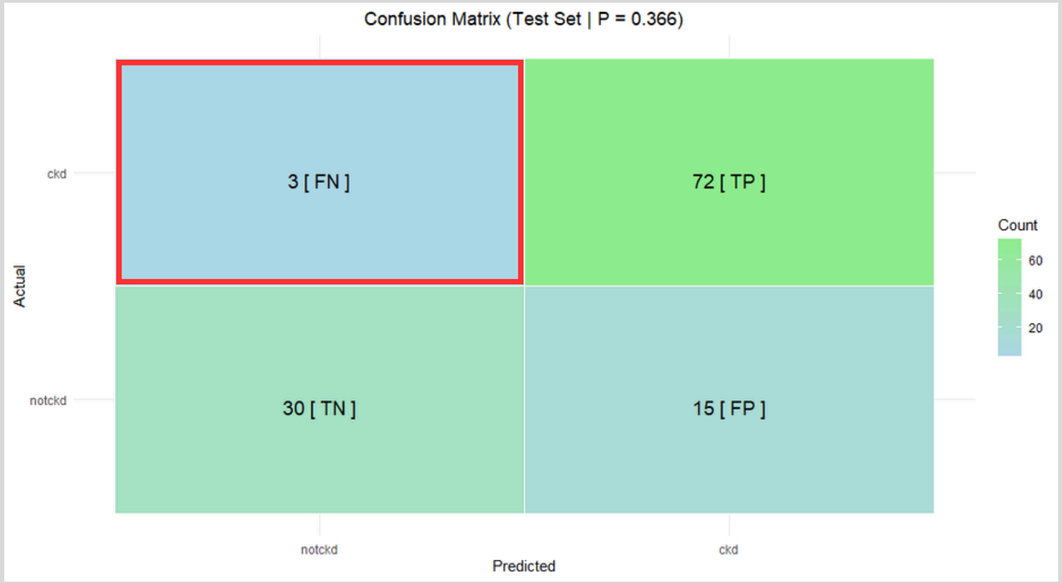


- **Conservative** = Higher Misclassification
  - **Decreased Precision & Accuracy**
- **Lower FN** cases
  - **Increased Recall**

Metrics Optimizations

Threshold Adjustment





P = 0.5

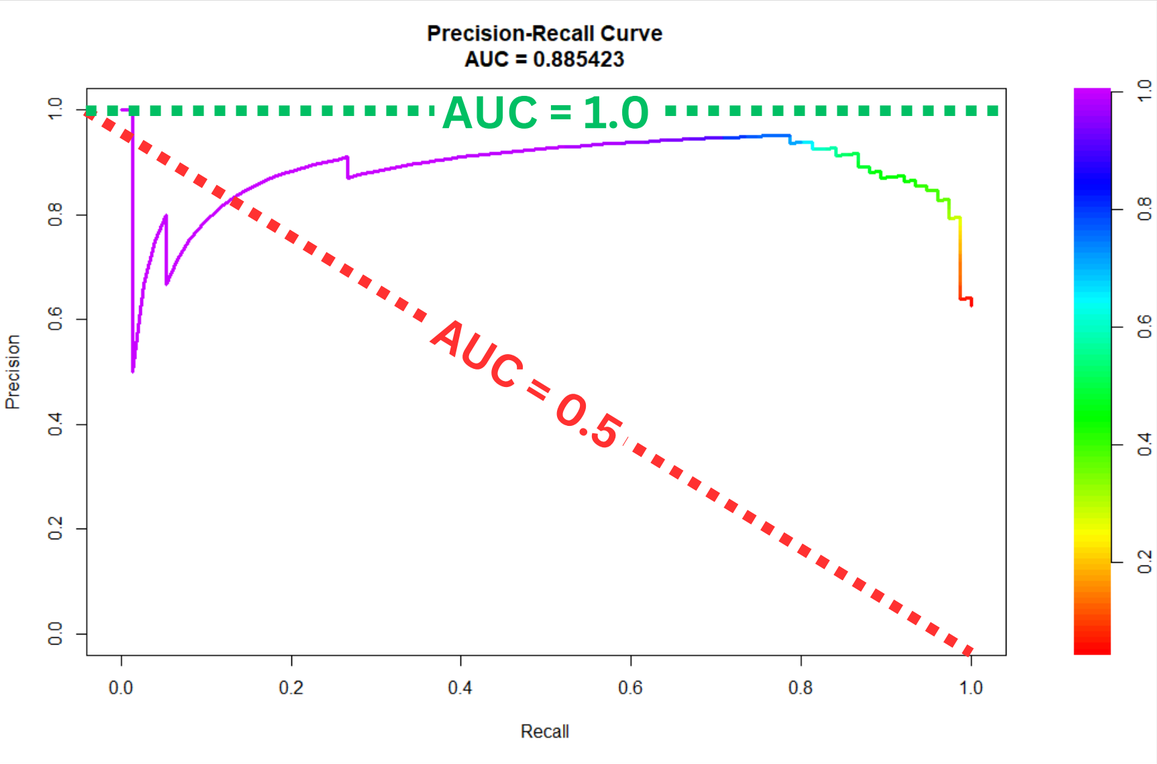
P = 0.366

FN count **reduced** from
10 to 3

Reliable Model that has High Recall with
Reasonable Accuracy & Precision

Metrics Optimizations

Area Under the Curve (Precision-Recall Curve)



Evaluating Model Ability to Distinguish Classes

Used for Imbalanced Datasets

- **250 CKD vs 150 notCKD**

AUC = 1.0

[Perfect Model, All Correct Classifications]

AUC = 0.5

[No-Skill Model, 'Coin-Flip' Guesses]

AUC of Logistic Regression Model = 0.885

Correctly Identifying CKD cases 88.5% of the time

CART Model

Data Splitting & Tree Construction

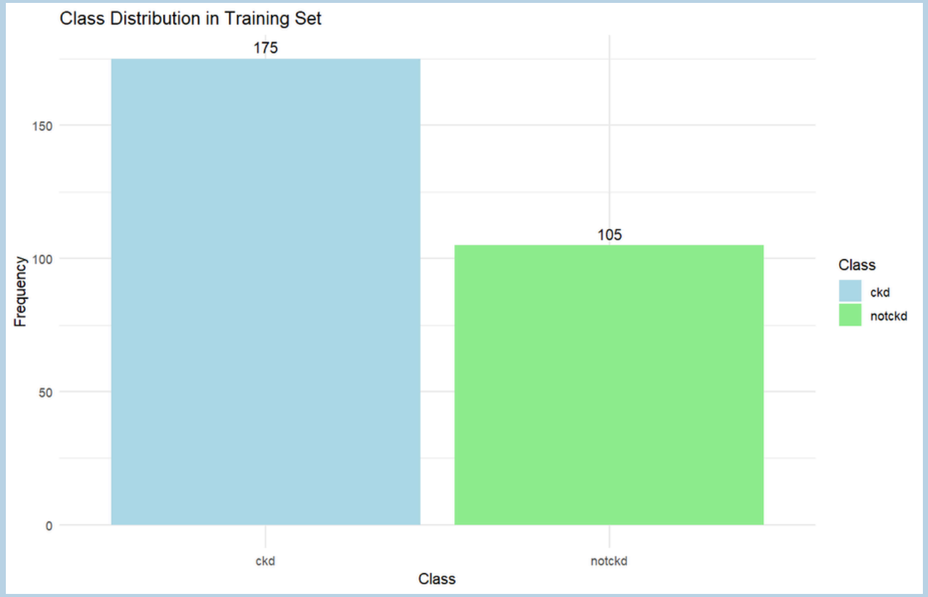
Optimal Pruning for Simplification

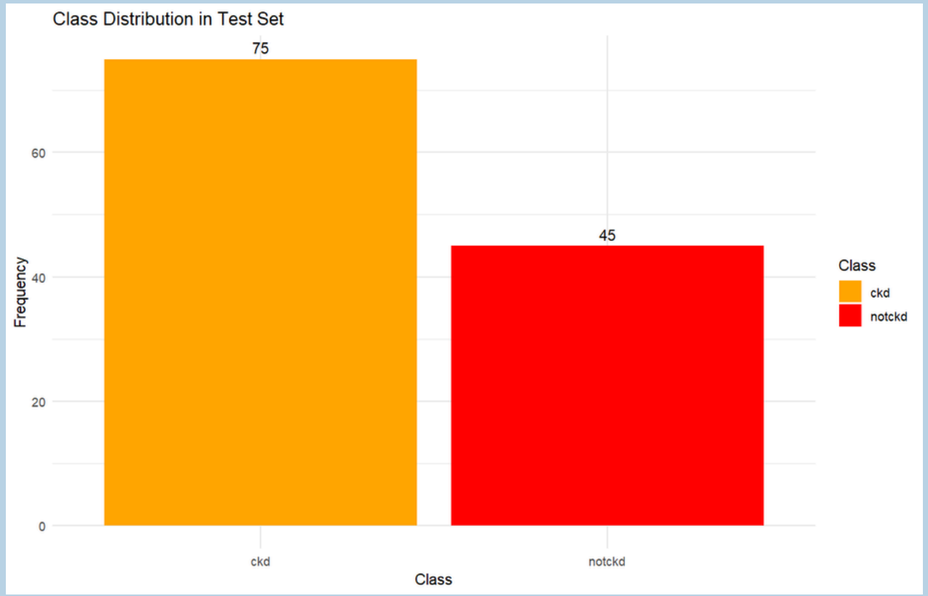
Interpreting Key Variables

Model Performance & Metrics

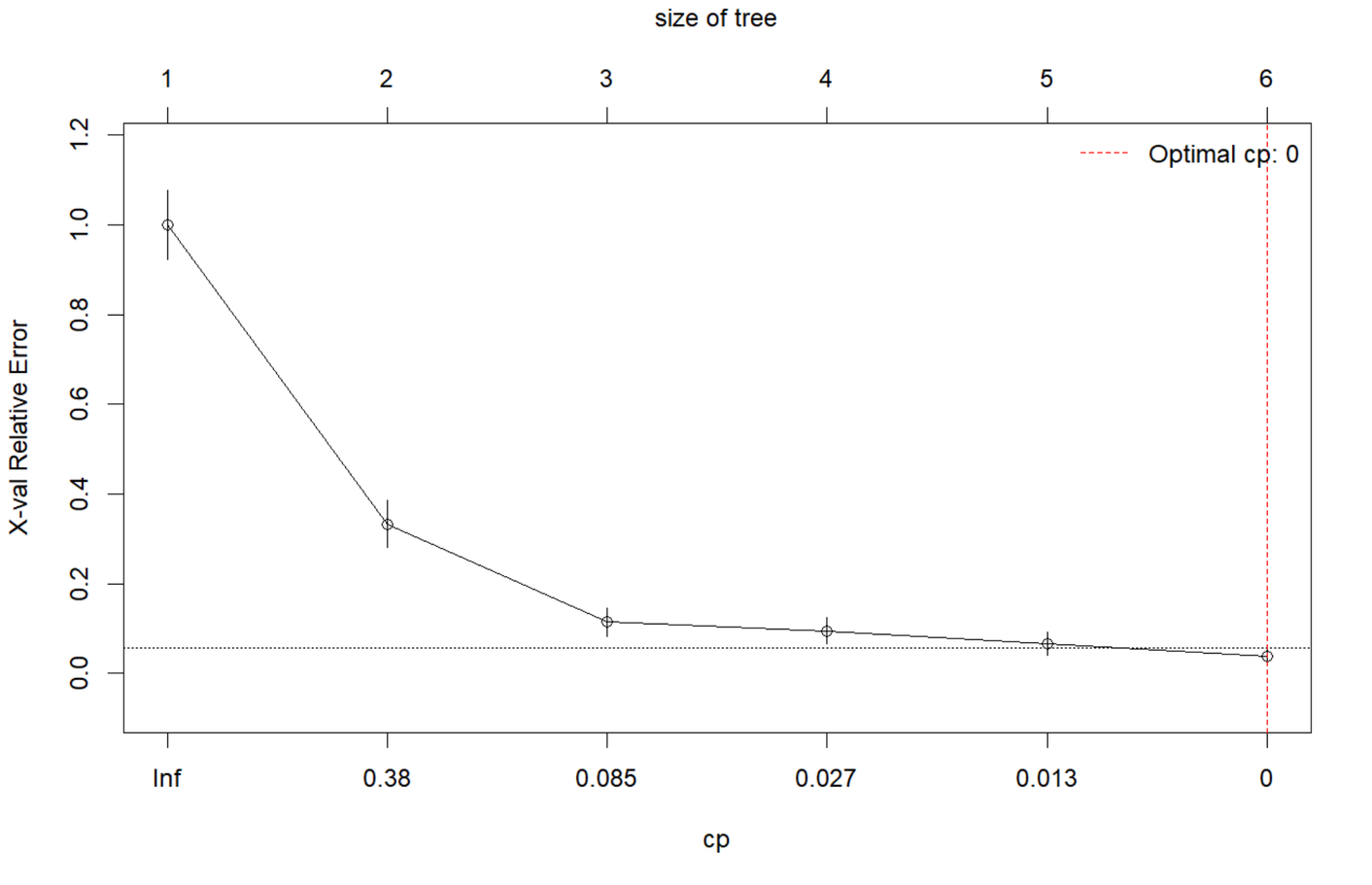
70-30 Train and test data splits

```
[1] "Class Distribution:"
> trainData[,.N,by=class]
    class     N
   <fctr> <int>
1:    ckd   175
2: notckd   105
> testData[,.N,by=class]
    class     N
   <fctr> <int>
1:    ckd    75
2: notckd    45
```

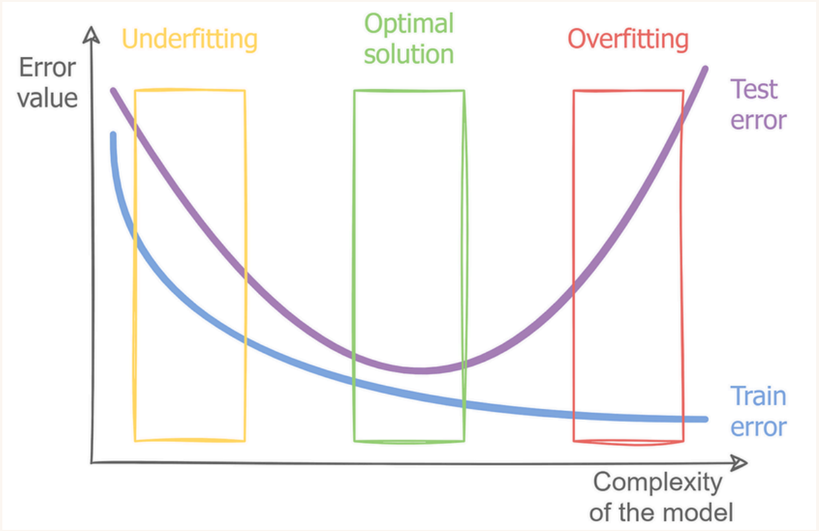




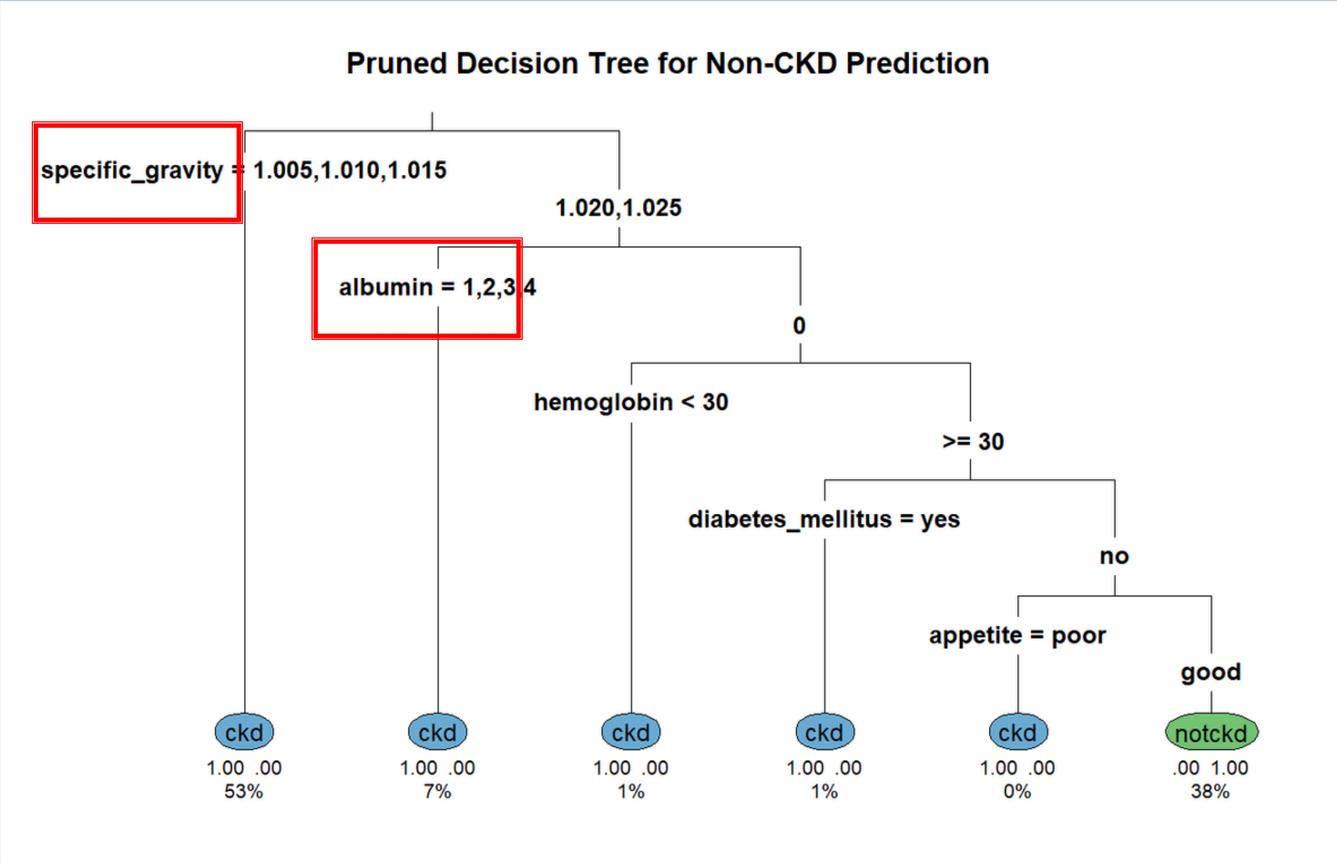
Optimal Pruning for Simplification



- Pruning: Reduces tree complexity, preventing overfitting
- Complexity Parameter (cp): Controls tree size and error rate
- Optimal cp: Found where error is minimized without overfitting (close to zero in this case)



Pruned Decision Tree



	Node	Gini	Entropy
1	1	0.47992195	1.0492402
2	2	0.03321313	0.1487741
3	3	0.35927575	0.9245659
4	4	0.17477823	0.5601194
5	5	0.16806265	0.5612342
6	6	0.50236683	1.2727860
7	7	0.11023673	0.4158601
8	8	0.62633253	1.5161734
9	9	0.07794825	0.3149904
10	10	0.66745749	1.5963452
11	11	0.06087624	0.2411787

Decision Nodes

- Each split based on important health indicators
- Final nodes (leaves) show CKD or Non-CKD prediction with confidence

Gini and Entropy

- Measure purity or confidence at each node
- Lower values indicate high certainty in classifications

Interpreting Key Variables & Insights

```
> print(importance)
  specific_gravity           albumin      hypertension packed_cell_volume  serum_creatinine
         77.710710         69.402702         48.360587         48.051953         47.018245
 diabetes_mellitus        hemoglobin          appetite    blood_pressure         potassium
         38.462854         16.216426         12.393172         11.899679          8.924759
>
```

Variable Selection in Decision Tree

- Split Selection: Prioritizes variables with highest impurity reduction (information gain) at each level.
- High-Importance Variables Not Used:
  - May not appear in splits if other variables provide better immediate separation.
  - Still contribute predictive power and improve overall accuracy.
- Support for Model Accuracy:
  - Enhance prediction accuracy for cases outside main splits.
  - Improve model generalization across diverse patient profiles.

Model Performance & Confusion Matrix

4.2.5 Confusion Matrix and Performance Metrics for Test Set

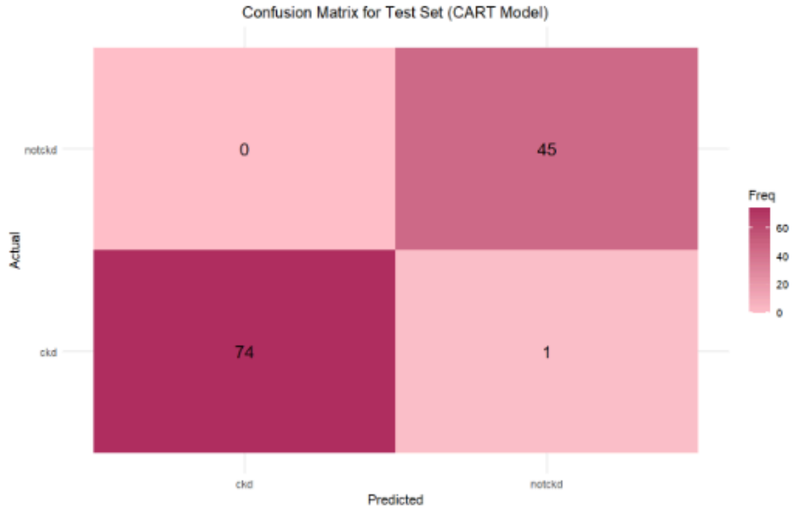


Fig <>. Confusion Matrix on Test Set (CART)

TP	45
TN	74
FP	1
FN	0
Accuracy	$\frac{45+74}{45+74+1+0}$ **= 99.17%**
Precision	$\frac{45}{45+1}$ **= 97.82%**
Recall	$\frac{45}{45+0}$ **= 100%**

Clinical Relevance

- High accuracy and recall are crucial for avoiding misdiagnosis of non-CKD
- Supports early intervention by accurately identifying non-CKD individuals

True Positives (notckd, notckd):

- High rate for non-CKD detection, demonstrating accuracy in identifying non-CKD cases

High Recall & Precision:

- Recall: 100% — captures all non-CKD cases
- Precision: 97.82% — low false positives, enhancing reliability

Overall Accuracy: 99.17%

- Robust in distinguishing CKD vs. non-CKD
- Effective for early detection, reducing diagnostic errors

Evaluation of Solution

01.High Recall with Logistic Regression

Optimized logistic regression model achieves **96% recall** at a **0.604 threshold**, effectively identifying CKD cases, which is crucial for early intervention in progressive diseases like CKD

02. Interpretability with CART Model

CART model offers a **clear, interpretable** decision tree **reduces complexity** in the decision-making process that aids in clinician-patient communication about CKD risk factors

03. Application in Preventive Healthcare

Both models support **screening high-risk individuals** in routine check-ups enabling early follow-up and nephrologist consultations, aligning with preventive healthcare objectives

Limitation of Solution

01. Dataset

- Size
- Class imbalance
- Missing values

02. Models

- CART
- Logistic regression

03. Application

- Severity stages
- No time prediction

Future Directions : Potential Improvements

Enhance and diversify patients **data set** to improve accuracy and reduce biasedness

Explore models that can capture complex patterns and predict CKD severity stages and time prediction for deeper clinical insights

Validate our models with actual clinical data to assess reliability, across different healthcare settings

BC2406 – Analytics I:
Visual & Predictive Techniques

Thank you very much!