DTSA 5510 Unsupervised Algorithms

in Machine Learning

Chronic Kidney Disease

Clustering

OUTLINE

01 Introduction

02 EDA

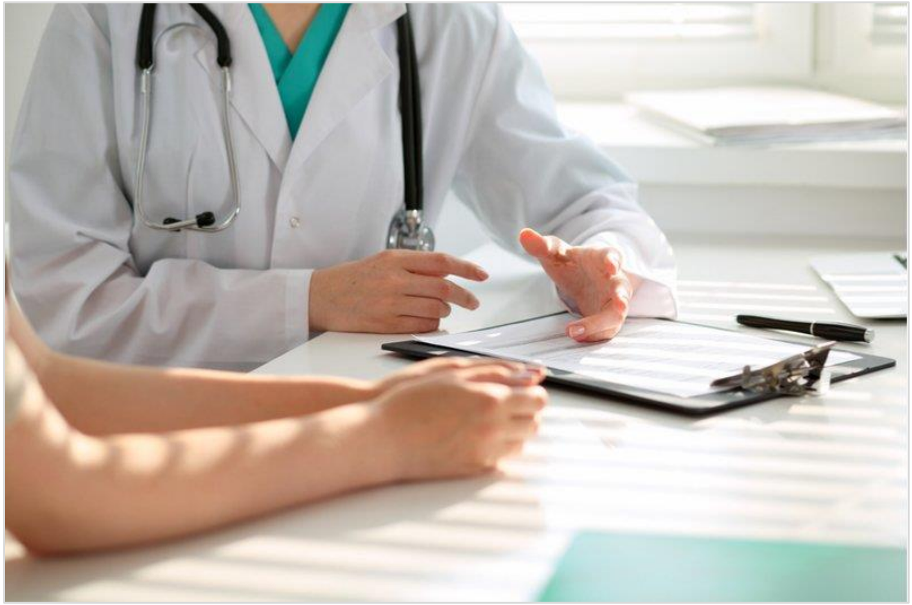
03 Models

04 Results

05 Discussion

06 Conclusion

01 Introduction



Motivation

Exploring clustering methods in binary classification tasks addresses real-world scenarios where labeled data is scarce.

Problem

Diseases like CKD, often classified as binary (CKD/no CKD). Can unsupervised learning reveal meaningful subgroups within CKD beyond this binary framework?

Approach

Apply unsupervised learning algorithms to uncover patterns in unlabeled CKD data and compare these results with supervised models.

02 EDA

400 samples 24 features

Data Cleaning

Typos

Renaming:
\t? to NaN, \t43 to '43',
\tno to no, ···

Mistyped Features

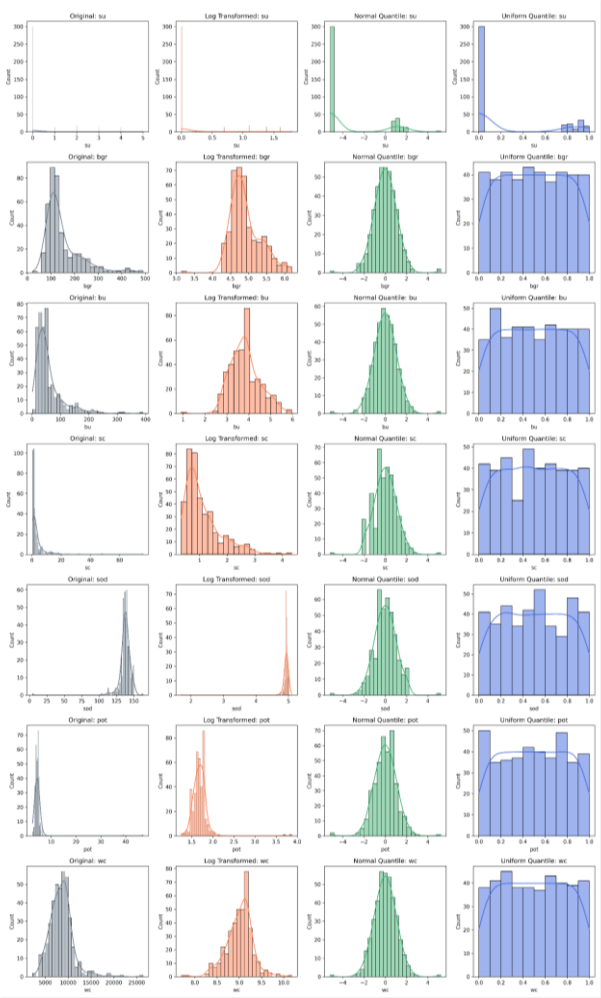
Converting:
'pcv', 'wc', and 'rc'
from object to float64

Missing Values

One-Hot Encoding.
KNNImputer:
with n_neighbors=8

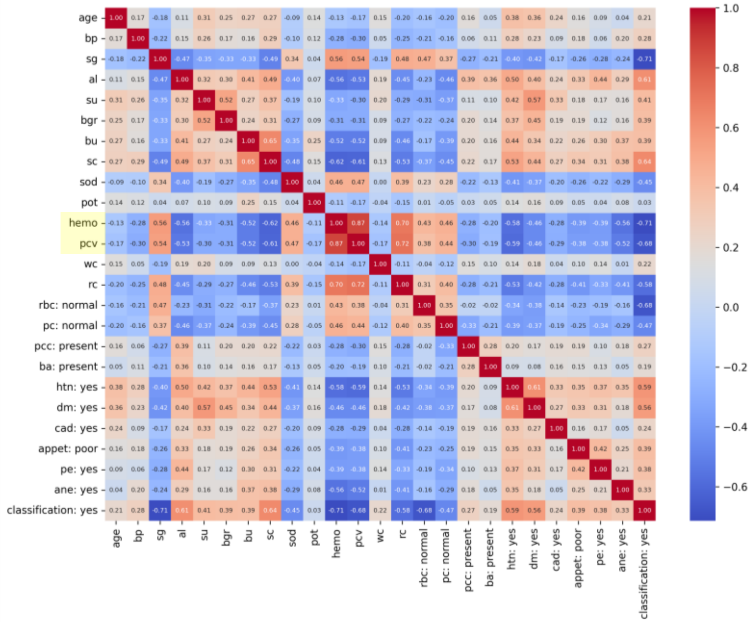
Transform

Skewness Analysis.
QuantileTransformer:
normal distribution

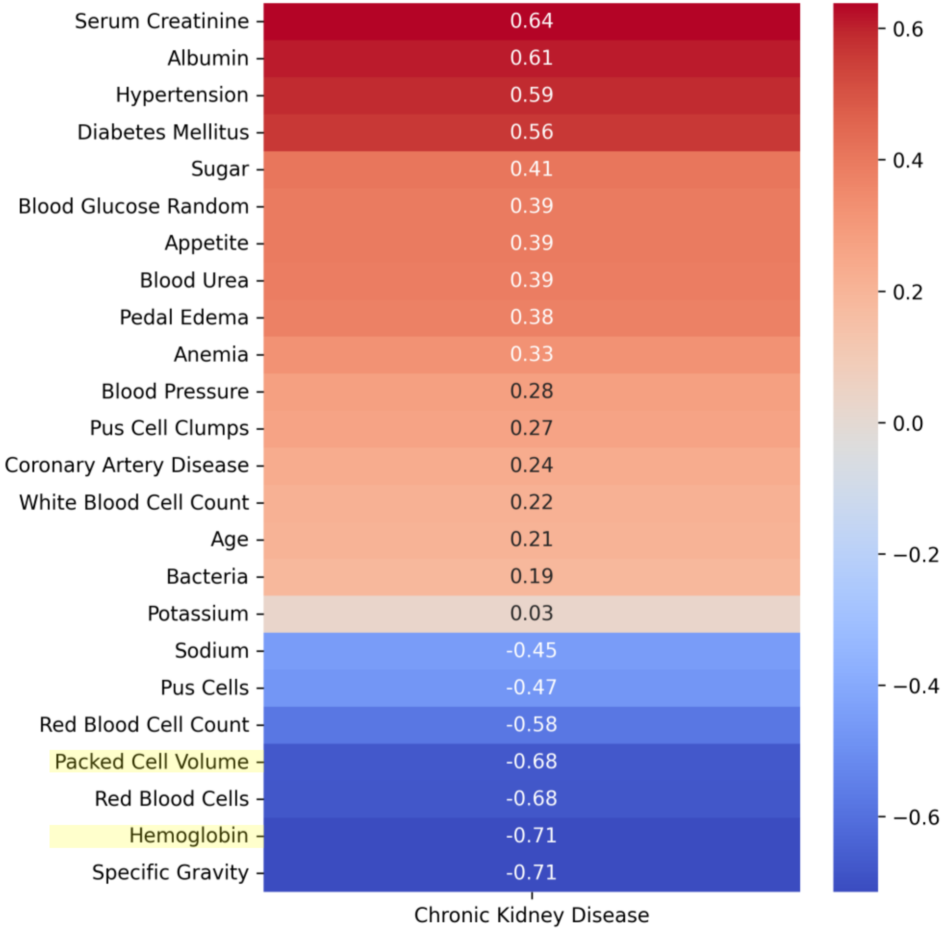


02 EDA

Correlation Matrix



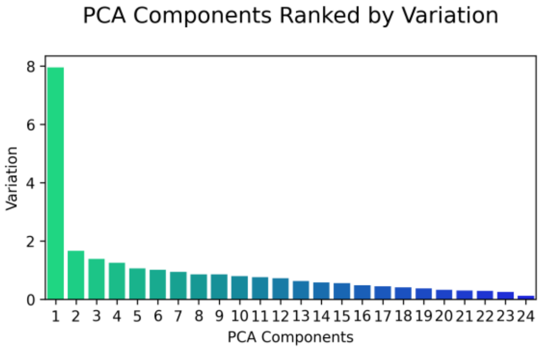
Target Correlation

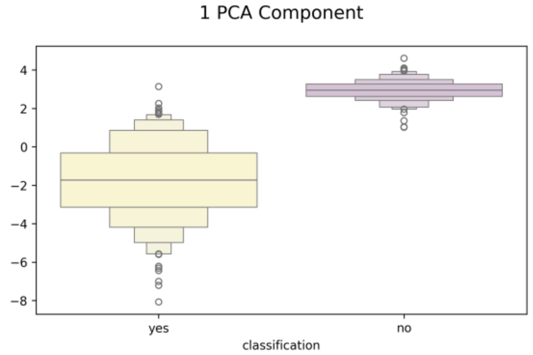


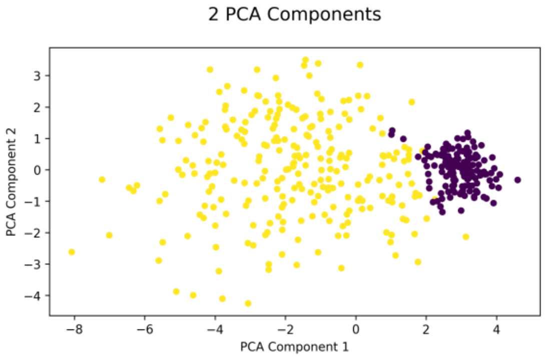
Correlation Analysis

Visualizations

03 Models

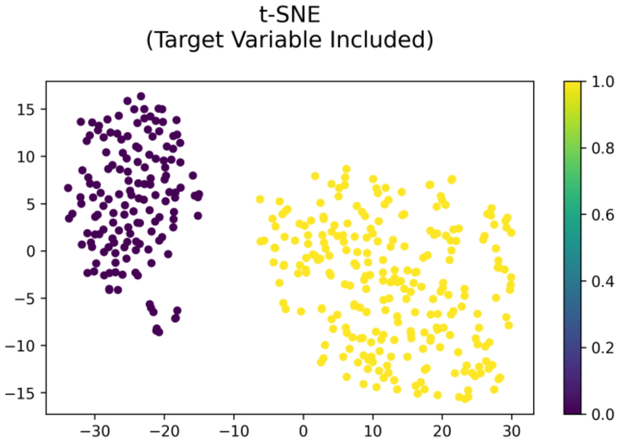

PCA Components Ranked by Variation

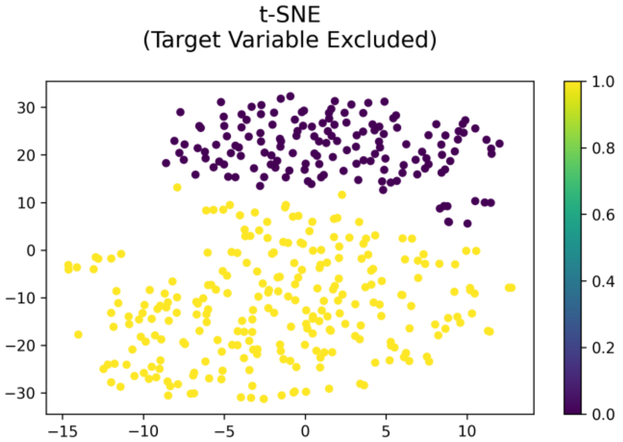

1 PCA Component


2 PCA Components

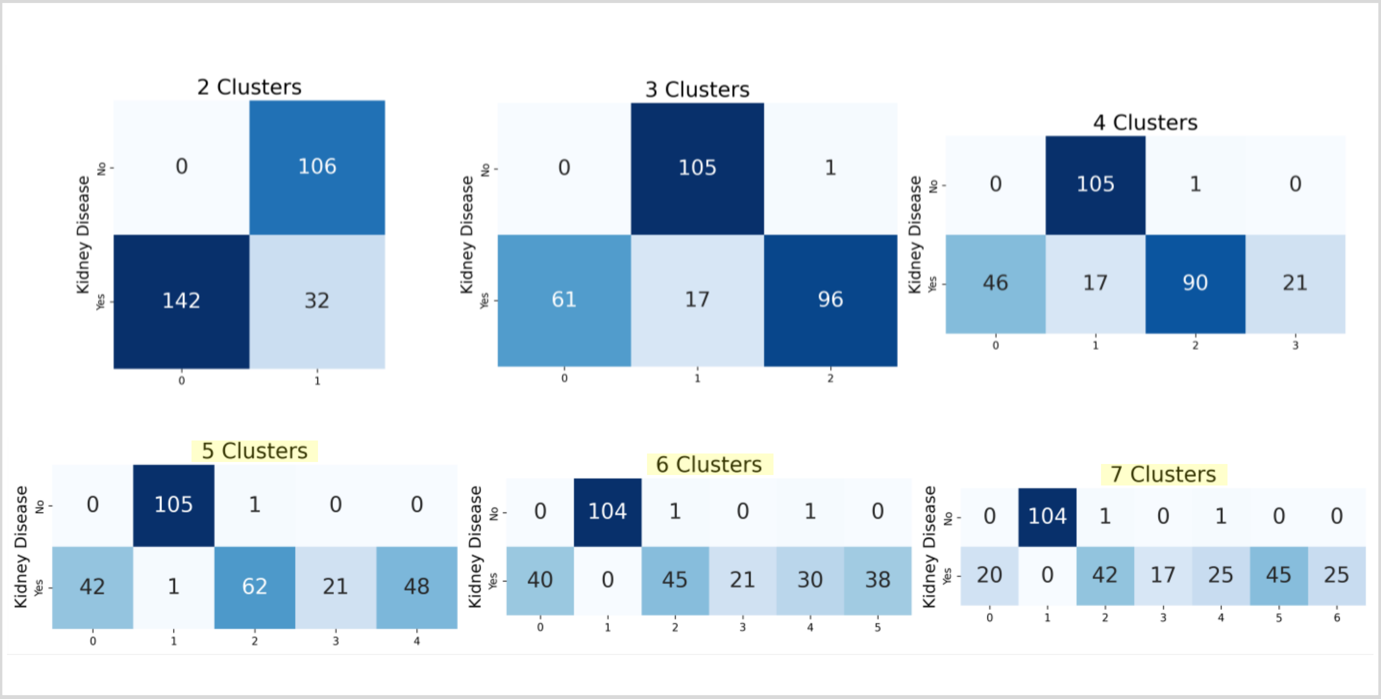
Unsupervised Learning

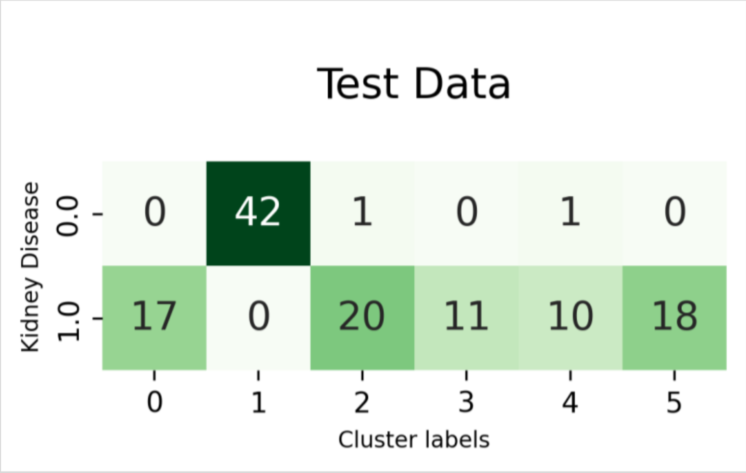
Dimensionality Reduction


t-SNE
(Target Variable Included)


t-SNE
(Target Variable Excluded)

03 Models





Unsupervised Learning

Kmeans Clustering

Elbow Method: k=2 is optimal, aligning with the binary nature of classification.
Training Data: Best result is 5-7 clusters. **(Training Accuracy = 278/280 = 99%)**
Test Data: Setting with n_clusters=6. **(Test Accuracy = 118/120 = 98%)**

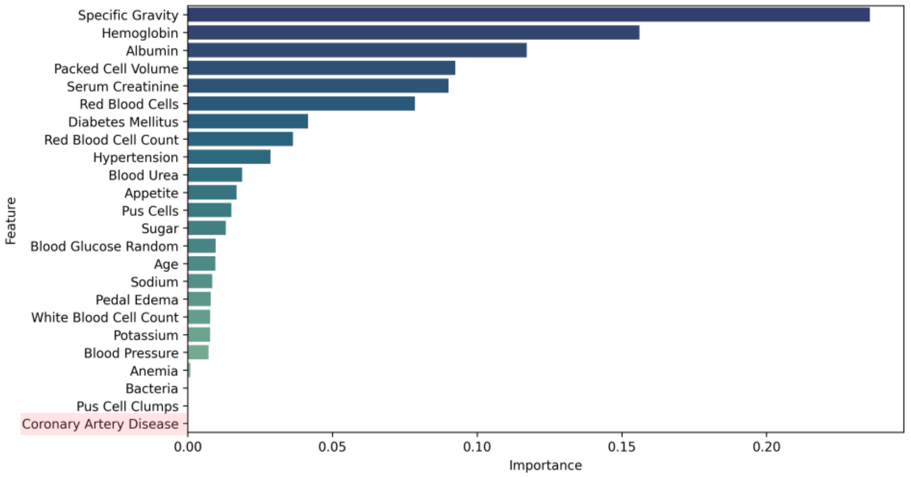
Random Forest
Gradient Boosting
Neural Network

04 Results

Model	Accuracy	F1-Score	Confusion Matrix	ROC-AUC	Misclassified Data Points
KMeans	0.98	0.99	[42 2] [0 76]	0.98	[1 67]
RF	0.98	0.99	[42 2] [0 76]	0.98	[1 67]
GB	0.97	0.98	[42 2] [1 75]	0.97	[1 24 67]
NN	**1.00**	**1.00**	[44 **0**] [**0** 76]	**1.00**	[ ]

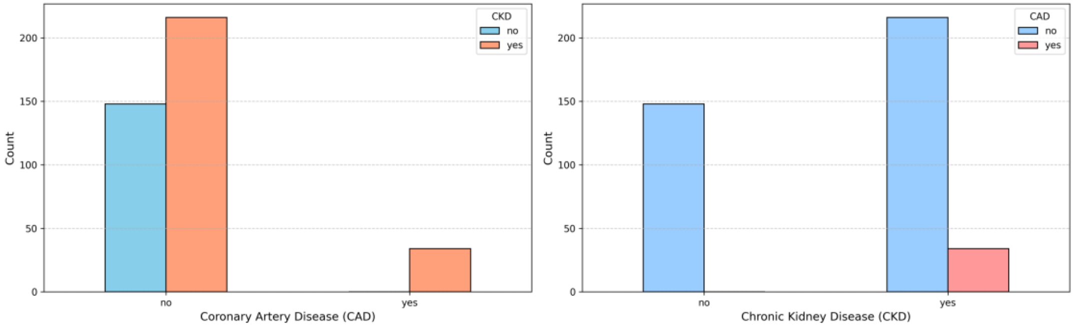
05 Discussion

Feature Importance (RF)



Coronary Artery Disease (CAD)

Data imbalance may cause RF to ignore CAD, giving it a score of 0.



06 Conclusion

Project Summary

This project explored clustering CKD data using Kmeans and used supervised models (RF, GB, and NN) to provide a benchmark for comparison.

Key Findings

KMeans effectively captured the dataset's structure, achieving 98% accuracy and demonstrating the potential of unsupervised learning to uncover patterns without labels.

Future Work

Future work will focus on analyzing subgroups within CKD to uncover meaningful patterns and investigating outliers to gain deeper insights into anomalies in the dataset.

https://github.com/d93xup60126/Unsupervised_Learning_CKD_Clustering

GitHub Repository Link

THANKS