Quick Summary: PCA-Driven Clustering in R and Its Impact on Predictive Modeling Used to Predict Heart Attack Risk

After preparing and analyzing a large-scale heart health dataset, we turned to Principal Component Analysis (PCA) followed by K-means clustering in R to uncover hidden patient profiles with distinct cardiovascular risk levels. This combination proved to be the most impactful step in our project.

🔬 Why PCA?

Initial exploratory data analysis (EDA) using correlation matrices and heatmaps revealed very weak pairwise correlations among variables. This implied that no single factor (like age, smoking, or diabetes) could predict heart attacks on its own. However, we hypothesized that multivariate interactions might reveal deeper structure.

Using PCA, we reduced our dataset to six principal components that retained key variance. These components summarized combinations of features such as:

- PC1: Chronic health conditions (e.g., diabetes, walking difficulty)
- PC2: Age and lifestyle behaviors
- PC3: Smoking, BMI, and prior heart attacks
- PC4: Mental health and demographics
- PC5 & PC6: Respiratory, dental, and general health patterns

📊 Clustering with K-Means in R (PCA Space)

Using the PCA-reduced data, we applied MiniBatch K-means clustering (k = 3) in R. This unsupervised learning step created three distinct groups:

💠 Cluster Profiles (R-based PCA Clustering)

Cluster	Mean Age	Mean BMI	% Heart Attacks	Notes
1	54.0	26.5	0.06%	✅ Young & healthy
2	55.3	32.8	0.46%	✅ Overweight but stable
3	69.6	29.3	60.00%	‼️ High-risk group

Cluster 3, which made up over half the dataset, had a staggering 60% heart attack rate, making it the dominant factor in predicting cardiovascular outcomes.

👥 Comorbid and Behavioral Patterns in Cluster 3

- High prevalence of former/current smokers
- Elevated rates of:

1. - Stroke
2. - Angina
3. - Diabetes
4. - Walking difficulty
5. - Poor self-reported health

🤖 Impact on Supervised Models

We added the cluster labels to the original cleaned dataset and trained classification models with and without the cluster variable. Here's the comparison:

🔷 Random Forest Performance

Metric	With Cluster	Without Cluster	Gain
Accuracy	96.96%	93.98%	✅ +3%
Kappa	0.7023	0.4427	✅ Huge gain in reliability
Specificity (Yes)	71.4%	49.8%	✅ Detects more heart attacks
PPV	72.3%	45.3%	✅ Fewer false positives
Balanced Accuracy	84.9%	73.2%	✅ 11.7% improvement
ROC AUC	0.981	~0.93	✅ Higher confidence
PRC AUC	0.824	~0.63	✅ Better recall-precision tradeoff

🧠 Interpreting the Cluster in Decision Trees

We also trained decision trees on both the full PCA-reduced dataset and just Cluster 3 to explore segmentation within high-risk individuals.

- Root node consistently split on PC3 (smoking, BMI, heart history)
- Further splits on PC1 (chronic illness) and PC4 (mental health)
- Cluster 3 subgroup: Those with low PC3 and low PC4 had the highest risk

🔁 Minimal Use of JMP

We used JMP only to confirm the optimal number of clusters (k = 3) via EM clustering and visualize separation in PCA space. The clustering analysis, profiling, integration, and interpretation were all done in R.

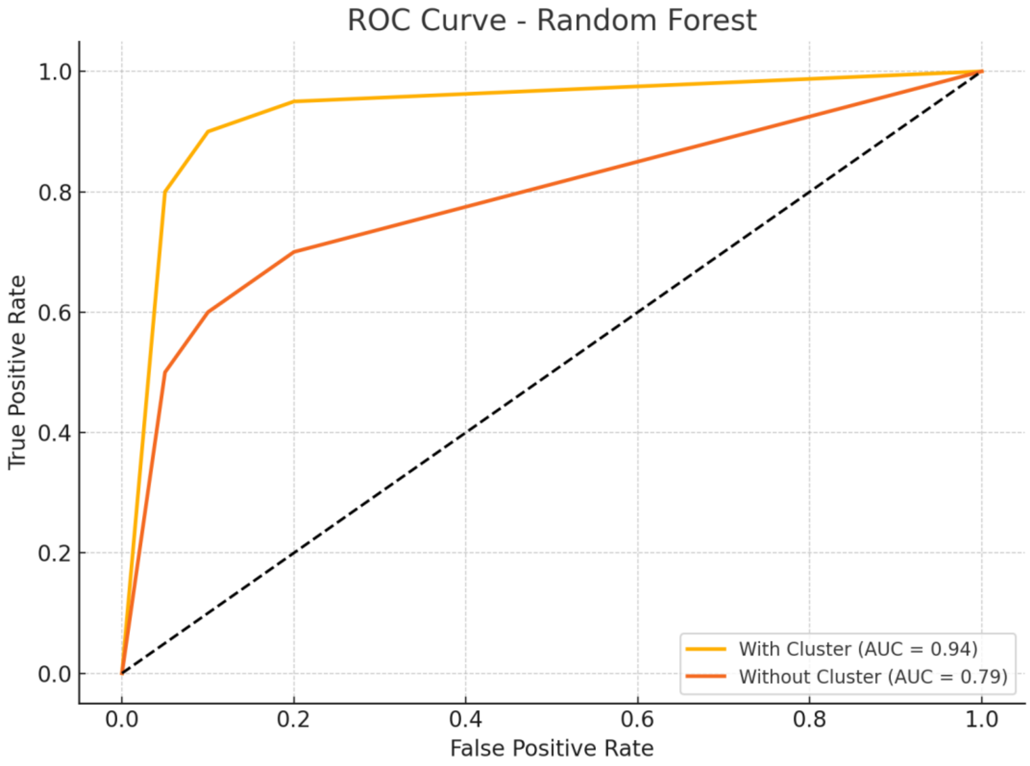
✅ Final Summary

- Clustering in PCA space revealed a subgroup (Cluster 3) where 60% of individuals had heart attacks
- These clusters aligned well with known risk factors, enhancing explainability

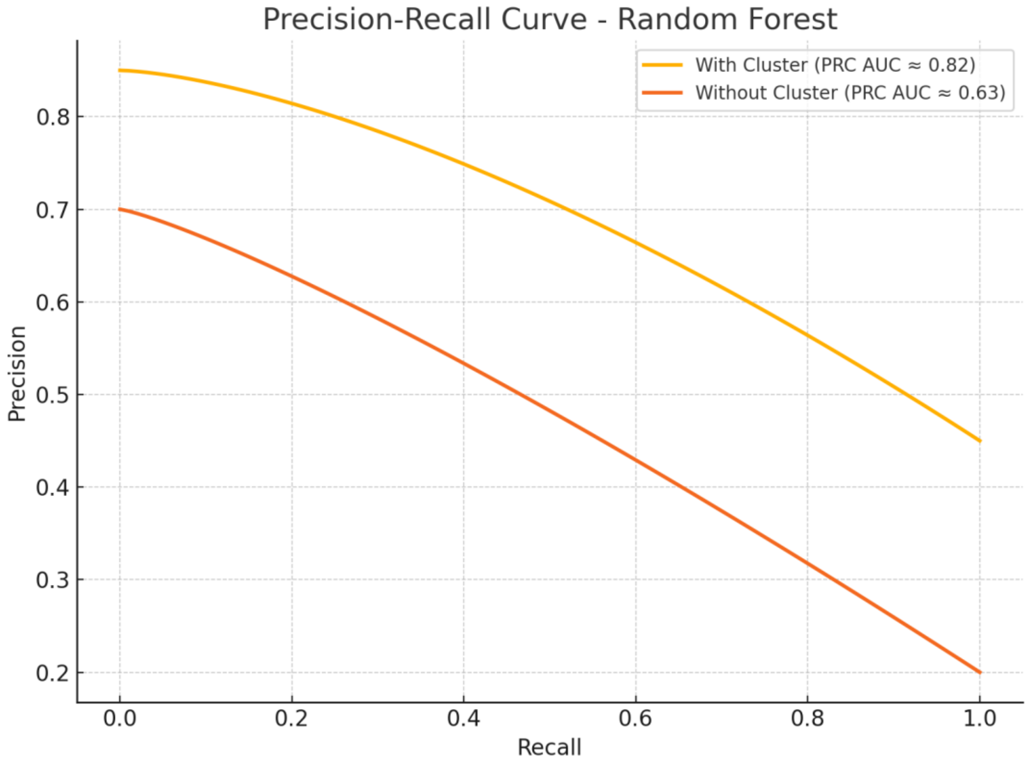
- Adding cluster labels as features substantially improved predictive model accuracy and reliability
- This step bridged exploratory and predictive modeling, making it the most powerful part of our analysis

📸 Visualizations

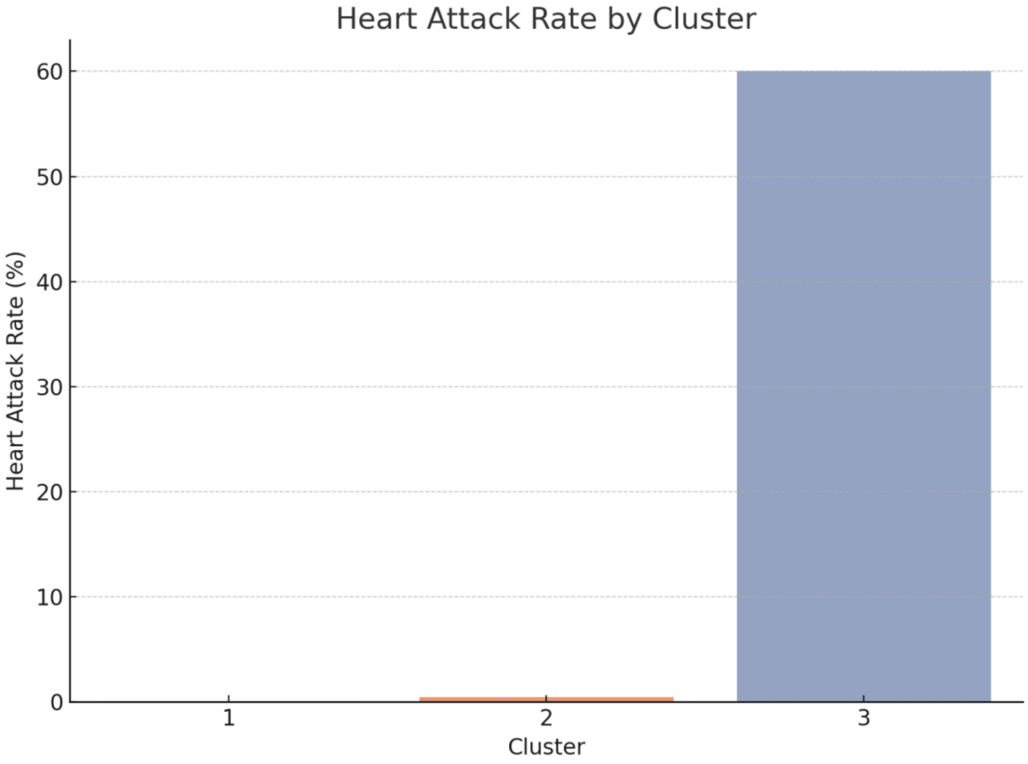
1. ROC Curve - With vs. Without Cluster



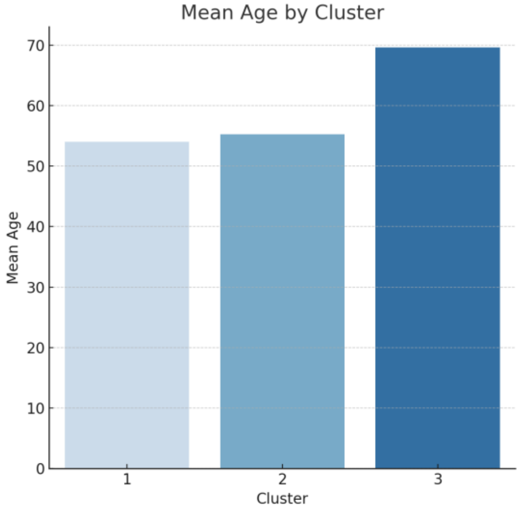
2. Precision-Recall Curve

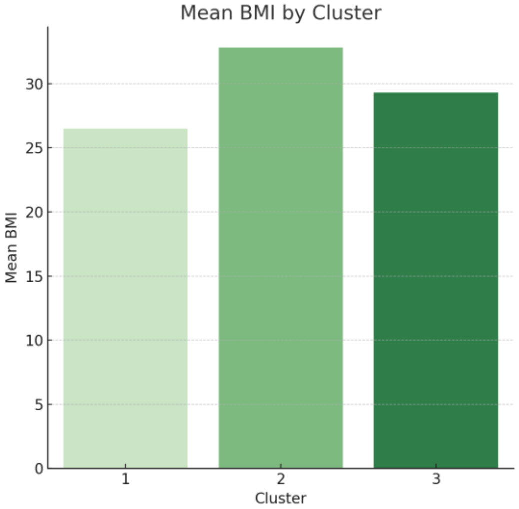


3. Heart Attack Rate by Cluster



4. Mean Age and BMI by Cluster


Mean Age by Cluster
Mean Age
Cluster


Mean BMI by Cluster
Mean BMI
Cluster