RISK-BASED PATIENT SEGMENTATION USING R: A DATA MINING CASE STUDY FOR A REGIONAL U.S. HOSPITAL NETWORK

1. Background

A regional hospital group in the Midwest U.S. wanted to identify chronic disease risk clusters among its patient population to prioritize preventive care. The existing manual screening based on age and BMI was inefficient and failed to capture multi-factor risk. The client requested a data mining model in R to uncover hidden patterns and segment patients based on likelihood of developing preventable chronic conditions.

2. Objective

- To group patients into clinically meaningful risk clusters using unsupervised learning
- To build a classification model for predicting high-risk individuals from demographic and medical data
- To generate a report identifying key risk drivers for each cluster to support targeted interventions

3. Data Used

Source: Electronic Health Records (EHR) across 7 hospitals

Structure:

- 38,000 anonymized patient records
- Fields: Patient_ID, Age, Gender, BMI, Smoker_Flag, A1C, Blood_Pressure, Medication_Count, Physical_Activity_Level, Chronic_Disease_Diagnosis (binary target)

4. Modeling Methodology

4.1 Data Preparation

- Handled missing values using mice package
- Scaled numeric variables and encoded flags as binary

- Created Risk_Score index combining A1C, BP, and BMI using weighted z-scores

4.2 Unsupervised Clustering (Patient Grouping)

- Used kmeans() and fviz_nbclust() to determine optimal clusters (k = 3)
- Segmented patients into:
    - Cluster 1: Young & Healthy
    - Cluster 2: Middle-aged, Pre-Chronic Indicators
    - Cluster 3: High-Risk (older, sedentary, high A1C and BP)

```
clusters <- kmeans(scaled_features, centers = 3, nstart = 25)
```

4.3 Predictive Modeling (Chronic Disease Risk)

- Built a logistic regression model using glm() with Chronic_Disease_Diagnosis as target
- Also tested randomForest() and xgboost::xgb.train() models
- Logistic regression chosen for interpretability

```
model <- glm(Chronic_Disease_Diagnosis ~ Age + BMI + Smoker_Flag + A1C +
Activity_Level + Medication_Count, data = patients, family = "binomial")
```

4.4 Risk Scorecard Design

- Developed rules from model output (coefficients + odds ratios)
- Designed a visual scorecard to classify incoming patients as Low, Medium, or High Risk

5. Results

Cluster	Share	Avg A1C	Avg Age	% Chronic Cases	Intervention Suggested
C1	41%	5.1	32	4.2%	No action
C2	37%	6.4	48	21.3%	Regular follow-ups
C3	22%	8.1	61	57.9%	Priority for chronic care

- Logistic regression AUC = **0.83**
- Model identified **BMI, A1C**, and **smoking status** as strongest predictors ($p < 0.001$)
- Random forest showed similar results but less interpretability for healthcare staff

6. Interpretation and Recommendations

- Majority of **Chronic Diagnosed Patients** were concentrated in Cluster 3
- Young sedentary patients with normal vitals in Cluster 2 were showing early risk indicators
- Recommended Cluster 3 for proactive calls, insurance follow-ups, and dietitian support
- Developed **'Watch List'** for Cluster 2 with automated email triggers for missed checkups
- Suggested integrating scorecard into EHR system for **real-time patient risk flagging**

7. Reporting Output

- **R Markdown PDF Report (28 pages)**
  - Cluster summaries
  - ROC curves and confusion matrices
  - Variable importance heatmaps
- **Interactive R Shiny Dashboard**
  - Input: Patient attributes
  - Output: Predicted Risk Level and Suggested Intervention
- **Reusable Code Modules**
  - risk_clustering.R, predict_risk_model.R, generate_scorecard.Rmd, shiny_risk_app.R

8. Business Outcome

- Identified **~8,200 high-risk patients** for priority enrollment in the care program
- **Reduced average first diagnosis delay by 2.6 months** (pilot area)
- Supported grant funding applications for preventive care budget based on cluster analytics
- The model is now used quarterly for **regional chronic risk screening reports**