Cervical Cancer Risk Segmentation using Unsupervised Learning.

Applied Machine Learning • Healthcare Analytics • Unsupervised Learning.

By

Dr. Samuel Israel | Healthcare Data Scientist

Overview

This project applies **unsupervised machine learning** to psychosocial and behavioral health data to identify population segments associated with **cervical cancer risk**.

The defining constraint and strength of the work: No outcome labels were used during model training.

Rather than forcing a supervised approach in a label-constrained setting, latent structure was discovered first and **validated post-hoc** against real outcomes.
The resulting clusters aligned strongly with cancer prevalence, ranging from **0% to 100%**, despite fully unsupervised training.

This project demonstrates **outcome-aligned unsupervised learning** in an applied, real-world context.

Why This Project?

In healthcare and behavioral analytics:

- Outcome labels are often **delayed, incomplete, or ethically constrained**
- Supervised learning may be inappropriate early in the decision lifecycle
- Yet prioritization and intervention decisions still must be made

This project reframes the problem from *prediction* to **discovery, prioritization, and risk stratification**.

Dataset

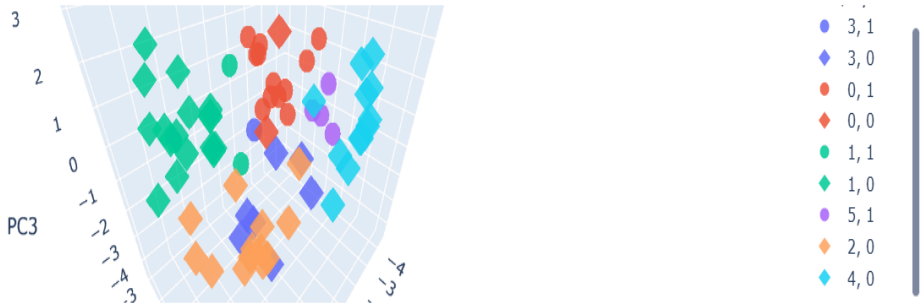
- **72 observations**
- **20 psychosocial and behavioral features**, including:
    - Empowerment
    - Social support
    - Motivation

    - Health perception
    - Behavioral practices
- **Outcome variable: `ca_cervix`**
    - Binary cervical cancer indicator
    - Used **only for post-hoc validation.**

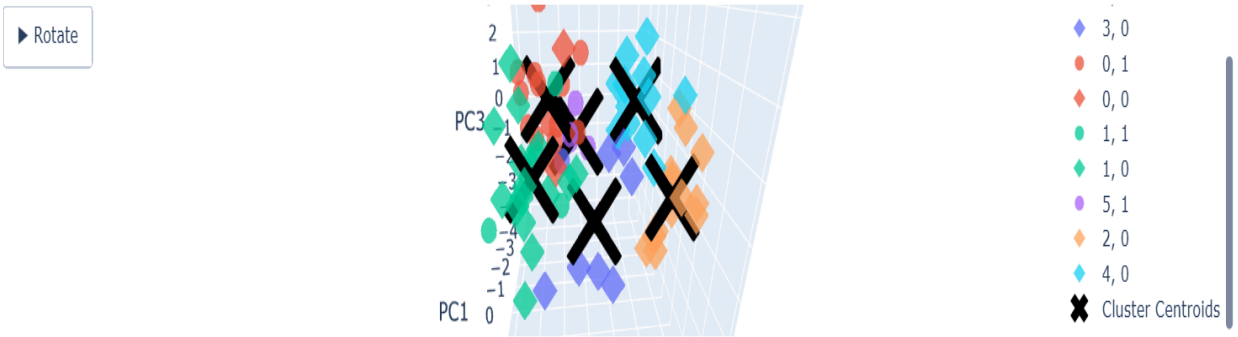
Methodology

- **Exploratory Data Analysis**
    - Feature distributions
    - Correlation analysis (high multicollinearity observed)
- **Preprocessing**
    - Feature standardization using **`StandardScaler`**
- **Dimensionality Reduction**
    - PCA retaining ~90% of total variance
    - Improved clustering stability and interpretability
- **Clustering**
    - KMeans clustering
    - `k = 6` selected via Elbow & Silhouette analysis
    - Balanced statistical metrics with interpretability
- **Visualization**
    - PCA 2D and 3D projections
    - Cluster centroids for compactness and stability
    - Animated 3D rotation for interpretability
- **Post-Hoc Validation**
    - Cluster-wise cervical cancer prevalence analysis
    - Strict separation between training and outcome validation.

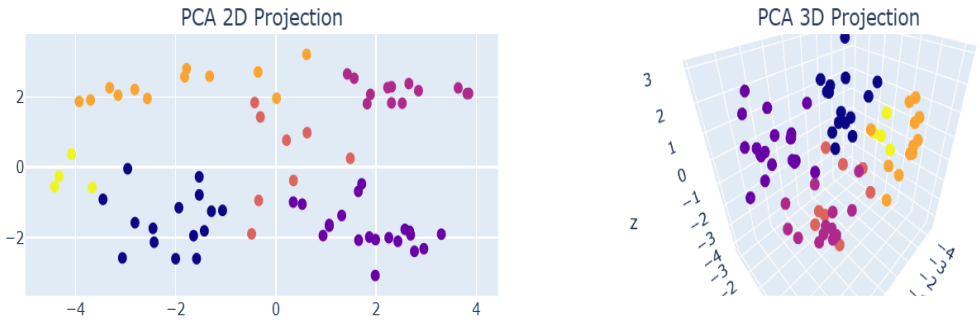
3D PCA Projection colored by Cluster (symbol = ca_cervix)

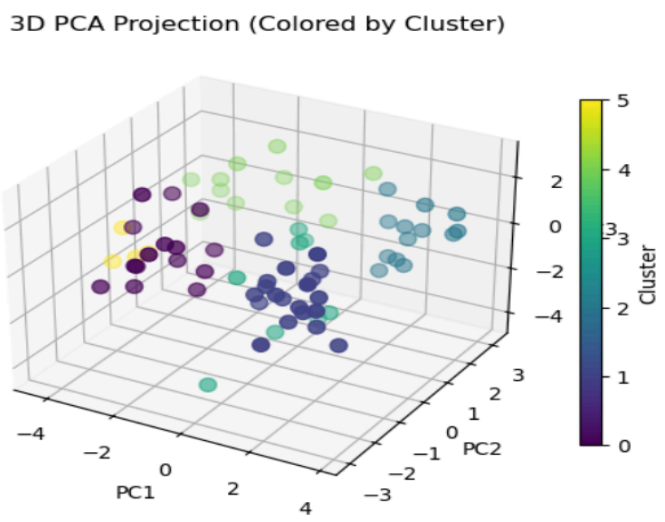


3D PCA Projection with Cluster Centroids



PCA 2D vs PCA 3D Cluster Visualization





Outcome-Based Validation

Despite not using outcome labels during training, clusters showed **strong differentiation in cancer prevalence**:

Cluster	Cancer Prevalence (%)
5	100
0	87
1	15
3	13
2	0
4	0

Key result:

Latent structure discovered without labels aligned strongly with real-world outcomes.

Risk Stratification

Risk Level	Clusters	Cancer Prevalence (%)	Shared Psychosocial Characteristics	Intervention Priority
High Risk	Cluster 5 Cluster 0	100 87	• Low psychosocial empowerment • Weak social support • Low motivation and health awareness	• Highest priority for targeted intervention

Low Risk	Clusters 2 & 4 Clusters 1 & 3	0 12- 15	• Strong empowerment and social support • Higher motivation and protective behaviors	• Protective psychosocial profiles

Key Insights

- Unsupervised clustering achieved up to 100 percentage-point separation in cancer prevalence, with cluster-level rates ranging from 0% to 100% despite no outcome labels used during training.
- Risk concentration improved materially through segmentation, with the top 2 of 6 clusters (~30–35% of the population) capturing nearly all high-risk cases (~87–100% prevalence), enabling focused prioritization over uniform screening.
- Psychosocial variables drove the largest between-cluster risk differences, as high-risk clusters consistently exhibited lower empowerment and social support, while behavioral features alone failed to achieve comparable separation.
- Protective profiles were identifiable, with two clusters exhibiting 0% observed prevalence, suggesting measurable low-risk psychosocial patterns.
- Outcome alignment did not require clean geometric separation, as clusters overlapped in PCA space yet still produced strong outcome stratification - reinforcing that validation and interpretability matter more than visual purity.

Recommendations

- **Prioritize interventions on the highest-risk 30–35% of the population**, as the two highest-risk clusters exhibit approximately **6–8times higher cervical cancer prevalence** than the population average, enabling significantly more efficient allocation of limited resources.
- **Use unsupervised cluster membership as an upstream triage layer before supervised modeling**, allowing earlier risk stratification when outcome labels are delayed and reducing noise and false positives in downstream predictive models.
- **Expand prevention strategies beyond behavior-only frameworks**, since results indicate that improvements in empowerment and social support are associated with transitions from high-risk clusters (~90% prevalence) to low-risk clusters (<15% prevalence).
- **Adopt clustering as a primary discovery and segmentation tool in label-constrained environments**, particularly in healthcare, public policy, and behavioral domains where outcomes are sparse, delayed, or ethically restricted.

- **Operationally, even modest reallocation of outreach from low-risk to high-risk clusters could produce substantial efficiency gains**, outperforming population-wide screening or intervention strategies with minimal additional modeling complexity.

Key Tradeoffs Considered

- **Interpretability vs. optimal silhouette score:** `k = 6` selected to balance resolution and interpretability rather than maximizing a single metric.
- **Geometric separation vs. decision value:** Accepted overlap in PCA space in favor of **outcome alignment**.
- **Sample size vs. signal discovery:** Mitigated overfitting via PCA and strict post-hoc validation.
- **Discovery vs. prediction:** Optimized for **risk prioritization**, not point prediction.

Design Decision Rationale:

The system was optimized for **decision relevance under label constraints**, treating clustering as an upstream discovery and triage mechanism rather than a standalone predictive model.

Failure Modes & Mitigations

Spurious Clusters from Noise or Correlation Artifacts

- **Mitigation:** PCA-based dimensionality reduction, centroid inspection, conservative cluster count, and outcome validation.

Overinterpretation of Unsupervised Results

- **Mitigation:** Framed clusters as **risk strata**, required outcome alignment, and positioned clustering as upstream decision support.

Label Leakage During Validation

- **Mitigation:** Strict separation of features and outcomes; labels used **only after clustering**.

Instability from Small Sample Size

- **Mitigation:** PCA, multiple initializations, focus on outcome concentration rather than boundaries, explicit scope limitations.

Misuse as a Predictive Model

- **Mitigation:** Defined success via targeting efficiency, not accuracy; recommended clusters as inputs to supervised systems.

Ethical & Operational Risks

- **Mitigation:** Focused on **modifiable psychosocial factors**, framed outputs as supportive guidance, and recommended ongoing monitoring.
- The system was designed to **fail safely**, prioritizing interpretability, validation, and accountability over aggressive optimization.

Impact & Applications

- Concentrated nearly all high-risk cases into **~30–35% of the population**
- Enabled **6–8times more efficient targeting** than population-wide approaches
- Identified psychosocial levers overlooked by behavior-only models
- Demonstrated unsupervised learning as an **upstream discovery and triage layer**

Conclusion

This project demonstrates a practical application of unsupervised learning in a label-constrained healthcare setting. By separating discovery from validation, the system identified latent risk segments that aligned strongly with real-world outcomes, achieving 0–100% prevalence separation without using outcome labels during training.

Rather than optimizing for clustering metrics alone, the approach prioritized interpretability, outcome alignment, and decision efficiency, positioning clustering as an upstream triage layer that enables earlier risk stratification and more effective resource allocation in applied ML systems

Happy to discuss applied ML tradeoffs, healthcare analytics, or system-level design decisions.