Investigating Adolescent Risk Behavior with Hierarchical Clustering

Claire Kelly, Casey MacGibbon, and Lily Ruddy

CSC 293: Machine Learning Final Project

INTRODUCTION

- The National High School Youth Risk Behavior Surveillance Survey (YRBS) is a series of measures that capture information about youth demographics, health and risk behaviors, substance use, and student experiences at school community and home.
- The survey contains 87 questions, making it difficult to understand which are important in determining youth risk behaviors to stakeholders
- The objective of the present study was to identify underlying potential clusters of adolescent experience to better understand patterns of youth risk behavior using a combination of unsupervised and supervised machine learning tools.

SURVEY DESIGN

- The survey is administered every two years to a representative sample of American youth. This analysis focused on the most recent data collected in 2023.
- Each of the questions asked in the YRBS has its own column in the dataset.
- Each row represents a participant in the survey, so their response associated with each question is a single observation.
- The dataset mostly consists of integers that correlate to each of the possible multiple choice answers.
  - EX. If for question 1 (q1), when asked how old they were, the observation is 1. That would mean that person responded with A), they were 12 years or younger.

METHODS

1. Data cleaning, including list-wise deletion for missing values, was done prior to analyses. Also removed unnecessary variables to the analysis like height and weight.
2. Principal Component Analysis (PCA) was used to identify which predictors moved together.
3. Scree plots were used to identify the PCs that captured the most variance, and were used to obtain a PC score vector for each observation in the dataset (Figure 2).
4. Hierarchical clustering identified 3 clusters (Figure 3), with each observation being labeled as belonging to a specific cluster. Cluster 1 appeared to be composed of multiple clusters, however, the model was unable to capture them due to heterogeneity in the data and the use of a Euclidean distance function.
5. PCA and hierarchical clustering was conducted a second time only on data originally identified as being in Cluster 1. Observations in Cluster 1 were relabeled as belonging to these 3 additional clusters (1a, 1b, 1c) (Figure 3).
6. Linear discriminant analysis (LDA) was used to reclassify the data within the identified five clusters to determine degree of cluster separation

RESULTS

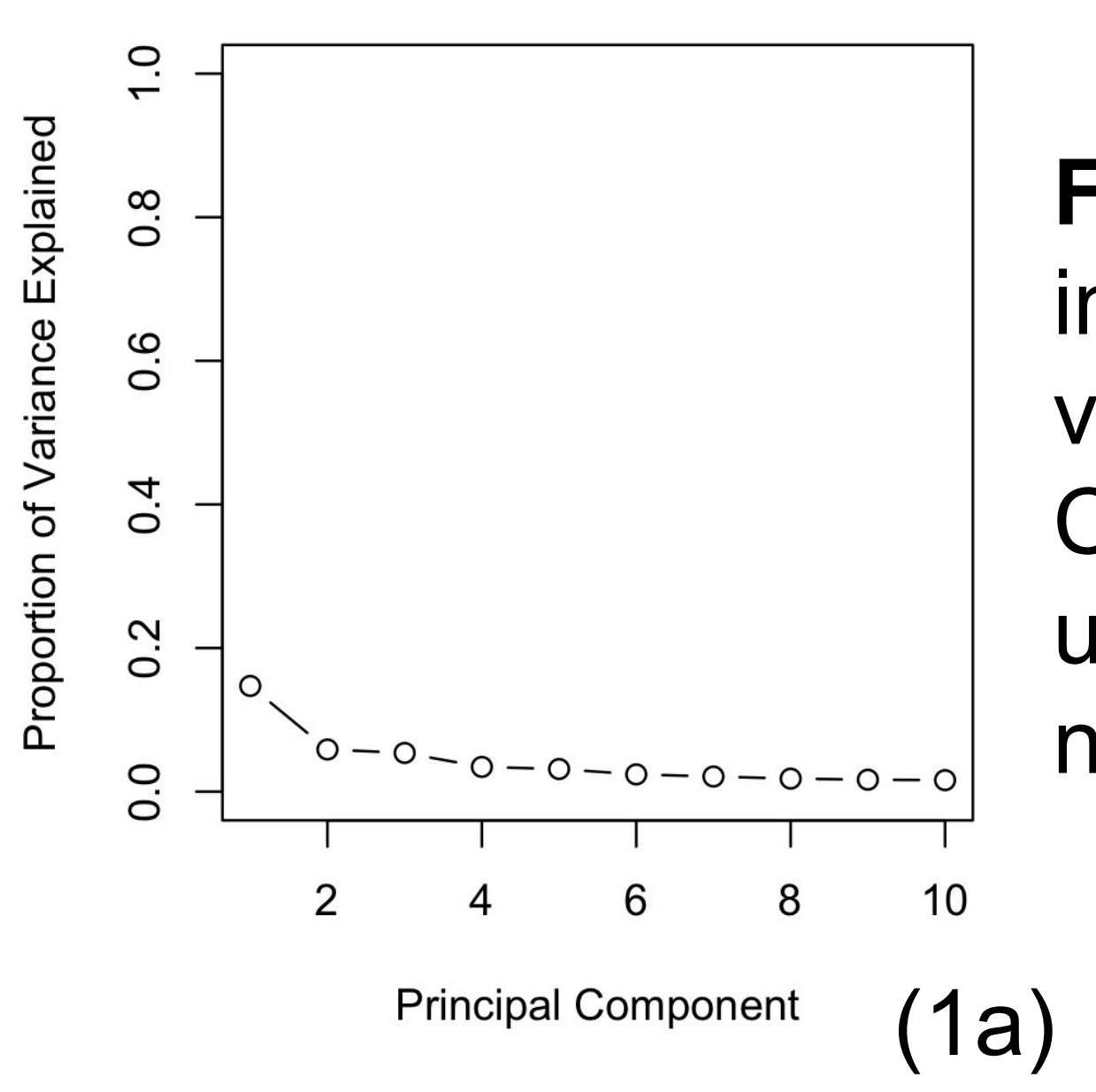
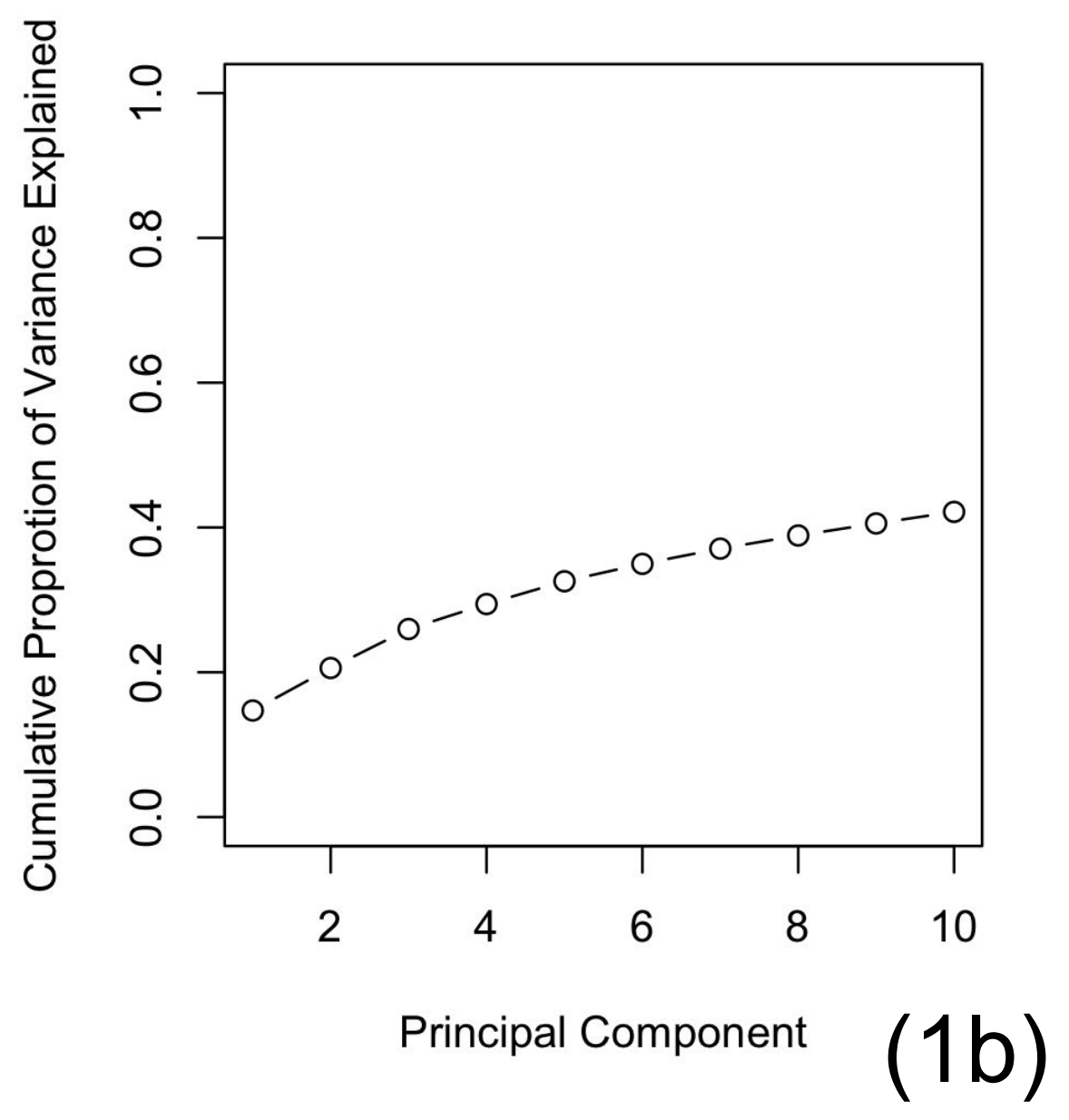


Figure 1: Scree plots of initial PCA. Proportion variance explained (1a) and Cumulative Proportion (1b) used to determine optimal number of PCs



- PCA revealed that only the first 3 principal components held any notable explanatory power albeit all small, suggesting that any resulting model would be characterized by high variance. Further research may benefit from initial dimension reduction techniques to help increase explanatory power.
- Secondary PCA and cluster analysis on cluster 1 indicates strong heterogeneity within the data. For this reason, investigation into non-Euclidean distance functions or other clustering methods.
- Adolescent risk behaviors clustered into 5 groups, with the vast majority of individuals in close proximity (1a, b, c) with the other clusters (particularly 3) exhibiting more extreme response patterns.

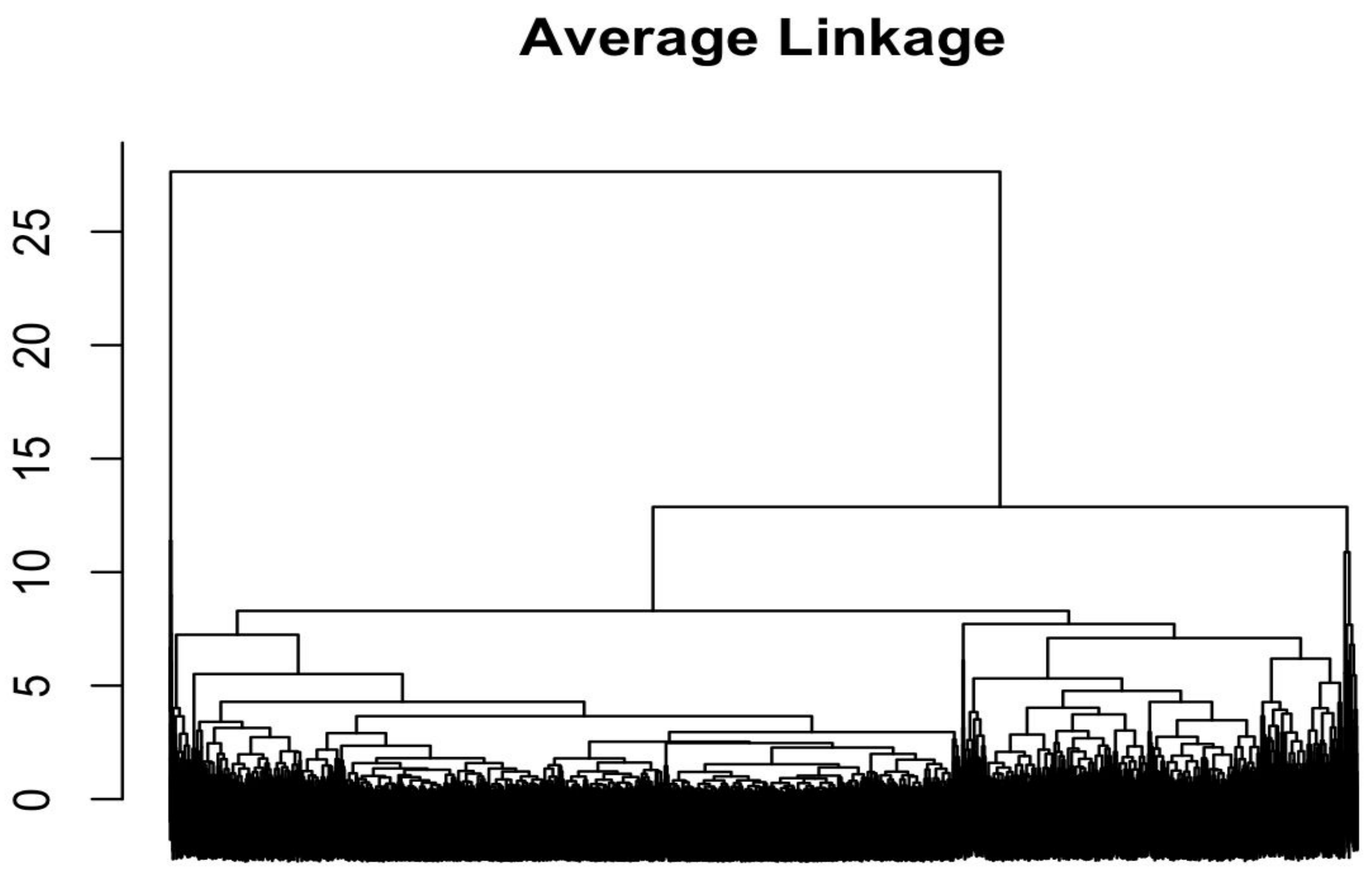


Figure 2: Dendrogram created using average linkage with three primary principal components as predictors

RESULTS

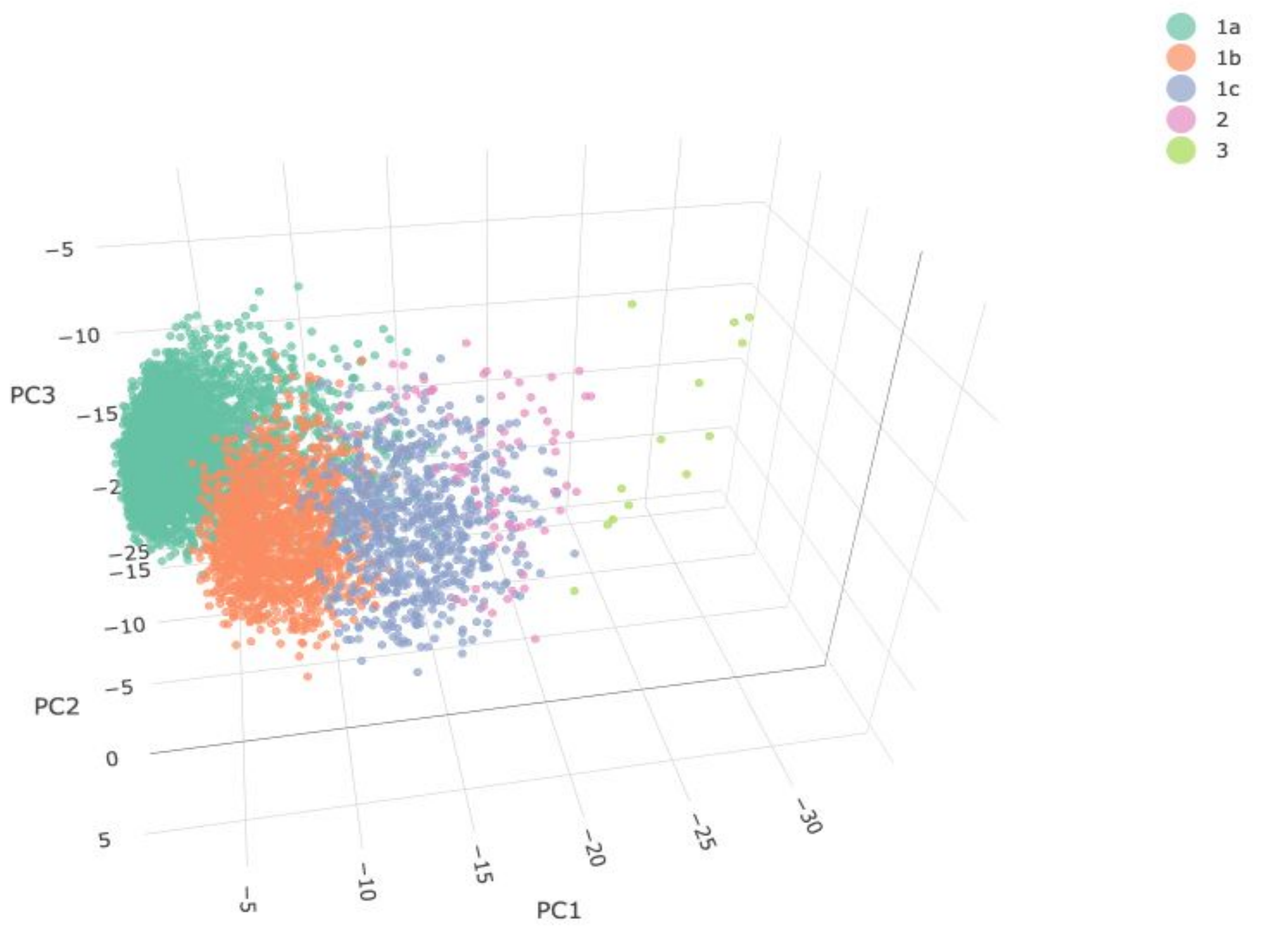


Figure 3: Full data PCA results plotted using the first three principal components in three dimensions.

CONCLUSIONS

- Clustered youth behaviors into five groups using predictors from the Youth Risk Behavior Survey (YRBS).
- PCA-regularized linear discriminant analysis achieved a 96.4% test accuracy, indicating strong separation between clusters.
- Questions 56 (sexual intercourse), 52 (marijuana use), and 55 (injection drug use) were the top three most important predictors, based on squared loadings summed across discriminant functions.
- Questions 60 (alcohol/drug use before sexual intercourse), 54 (ecstasy use), and 53 (methamphetamine use) were also important separators.
- Findings highlight a strong divide in youth behaviors related to drug use and sexual activity.