**LAB 3 Case Study**

**Code:**

# Load required libraries

library(stats)

# Define the function to simulate student features

simulate\_student\_features <- function(n = 100) {

# Set the random seed

set.seed(260923)

# Generate unique student IDs

student\_ids <- seq(1, n)

# Simulate student engagement

student\_engagement <- rnorm(n, mean = 50, sd = 10)

# Simulate student performance

student\_performance <- rnorm(n, mean = 60, sd = 15)

# Combine the data into a data frame

student\_features <- data.frame(

student\_id = student\_ids,

student\_engagement = student\_engagement,

student\_performance = student\_performance

)

# Return the data frame

return(student\_features)

}

# Simulate student data

student\_data <- simulate\_student\_features()

# Perform dimensionality reduction using PCA

pca\_result <- prcomp(student\_data[, -1], scale. = TRUE)

# Plot scree plot to decide on the number of principal components to keep

plot(1:length(pca\_result$sdev), pca\_result$sdev^2, type = "b", xlab = "Principal Component", ylab = "Variance Explained", main = "Scree Plot")

# Select the number of principal components based on the scree plot

num\_components <- 2

# Extract the scores for the selected number of principal components

pca\_scores <- as.data.frame(pca\_result$x[, 1:num\_components])

# Cluster the data using KMeans

kmeans\_clusters <- kmeans(pca\_scores, centers = 3, nstart = 25)

# Plot the clusters

plot(pca\_scores, col = kmeans\_clusters$cluster, main = "Clusters by KMeans", xlab = "PC1", ylab = "PC2")

points(kmeans\_clusters$centers[,1:2], col = 1:3, pch = 8, cex = 2)

Report:

- # Approach to Dimensionality reduction and clustering.

Dimensionality reduction was accomplished as part of the data analysis procedure by using the Principal Component Analysis (PCA) approach. In data science, this technique is frequently employed to decrease the high dimensionality of data while retaining the majority of its variation. The purpose of principal component analysis (PCA) is to create a new collection of orthogonal components, or principal components, from the original attributes of a dataset. The majority of the volatility in the data is explained by these components, which are linear combinations of the original attributes. We can decrease the dimensionality of the data while keeping the majority of its significant information by choosing fewer primary components.

We used the KMeans clustering technique to find clusters within the reduced data after it had been reduced using PCA to provide room for features. A well-liked unsupervised learning approach called KMeans divides data into a predetermined number of clusters according to how similar they are, with the centroid of each cluster serving as a representative. Data points are repeatedly assigned to the closest centroid by the algorithm, which updates the centroids until convergence. We can find comparable sets of data points using this technique, and we can learn from the patterns that show up. All things considered, the combination of PCA with KMeans clustering is a potent data analysis method that can be utilized to find hidden correlations and patterns in complicated datasets.

- # The results of your analysis, including the number of clusters identified, the characteristics of each cluster.

In the analysis conducted using PCA for dimensionality reduction and KMeans clustering, three clusters were identified based on the characteristics of student engagement and performance.

A graph of clusters

Description automatically generated with medium confidence

Cluster 1: High Engagement, High Performance

Features: This student body is renowned for its extraordinarily high degree of engagement with the educational process. They actively participate in class activities and debates because they have a strong sense of curiosity and a want to learn. These pupils also regularly demonstrate exceptional academic success in a variety of courses, demonstrating their remarkable cerebral prowess. Both their professors and classmates have acknowledged and applauded their ability to understand difficult topics, analyze critically, and apply information in practical situations.

Findings: This cluster consists of driven and accomplished kids who might gain from extracurricular activities or chances to pursue further education. Acknowledging and fostering these kids' skills might improve their school experience and results even more.

Cluster 2: Low Engagement, Moderate Performance

Characteristics: Although their academic achievement is modest, the students in this cluster show a lack of interest. Even if their academic performance is generally excellent, they might not be motivated or interested in learning activities.

Findings: This cluster identifies a set of students who can benefit from focused interventions to raise their levels of engagement and enrich their educational experiences. To meet the requirements of these kids, strategies like tailored education, individualized support, or motivation-boosting activities might be helpful.

Cluster 3: Moderate Engagement, Low Performance

Characteristics: Students in this cluster are moderately engaged, but they struggle academically and receive worse marks than their peers.

Findings: Despite their modest levels of involvement, this cluster reflects students who could encounter difficulties or obstacles in their educational journey. These kids may benefit from interventions that are designed to address their academic challenges, provide them more academic assistance, or deal with any underlying issues that are affecting their ability to study.

 -# Interpretation of results:

- The scree plot helps in determining the number of principal components to retain.

A graph with a line

Description automatically generated

- The distribution of data points according to their grouping is visually displayed by the cluster plot.

- Analyzing the traits of each student inside a cluster, such as their performance, engagement levels, and any other pertinent information, can help to understand the clusters.

- Comprehending these clusters can aid in discerning distinct cohorts of pupils possessing comparable characteristics, proving beneficial for focused treatments or customized methods inside learning environments.

-# implications of your findings for learning analytics.

The analysis's conclusions have significant implications for educational interventions as well as learning analytics. Teachers can differentiate interventions and support techniques to address the unique requirements of each group of students by categorizing them according to their performance and degree of involvement. While kids in low-engagement and low-performance clusters may need specialized help to get over difficulties in their learning process, high-performing students might benefit from enrichment activities to further develop their skills. Furthermore, the discovery of these clusters can act as a roadmap for the creation of focused interventions and instructional strategies meant to enhance student outcomes and foster academic achievement. With this knowledge, educators may design a more accommodating and successful learning environment for the various needs of students, ultimately leading to a more successful academic path for everyone.

References:

1. Dhavamany, N., & Pavalakodi, S. (2011). An effective method of dimensionality reduction for high dimensional datasets using PCA.*International Journal of Advanced Research in Computer Science, 2*(2) Retrieved from <https://www.proquest.com/scholarly-journals/effective-method-dimensionality-reduction-high/docview/1443705854/se-2>

2. Charalambides, N. (2020). *Dimensionality reduction for \(k\)-means clustering*. Ithaca: Retrieved from https://www.proquest.com/working-papers/dimensionality-reduction-k-means-clustering/docview/2427932015/se-2