Enfoque en PCA y Aprendizaje No Supervisado

Aliaga Yugra Edwin

Ticona Laura Yoel

Jurado Ever Emerson

INF 354, Inteligencia Artificial, Informe N° 1, INF - FCPN - UMSA

11/06/2025

Resumen

Este estudio explora el uso de PCA (Análisis de Componentes Principales) y K-Means clustering como herramientas de aprendizaje no supervisado para el análisis de datos educativos. Aplicamos PCA para reducir la dimensionalidad de un conjunto de datos sobre estudiantes universitarios y usamos el algoritmo de K-Means para identificar agrupaciones significativas sin la necesidad de etiquetas predefinidas. Los resultados muestran que la reducción de dimensionalidad con PCA preserva una parte significativa de la información y facilita la identificación de patrones en los datos. Además, el clustering revela dos grupos de estudiantes, lo que sugiere posibles intervenciones para mejorar la retención académica.

Palabras clave: Análisis de Componentes Principales, K-Means, Clustering, Aprendizaje no supervisado

Abstract

This study explores the use of PCA (Principal Component Analysis) and K-Means clustering as unsupervised learning tools for analyzing educational data. We apply PCA to reduce the dimensionality of a dataset on university students and use the K-Means algorithm to identify meaningful clusters without the need for predefined labels. The results show that dimensionality reduction with PCA preserves a significant portion of information and facilitates the identification of patterns in the data. Furthermore, clustering reveals two groups of students, suggesting potential interventions to improve academic retention.

Keywords: Principal Component Analysis, K-Means, Clustering, Unsupervised Learning

Introducción

Contexto

La identificación temprana de estudiantes en riesgo de deserción o bajo rendimiento académico es un desafío clave para las universidades. El análisis de grandes volúmenes de datos,

que incluyen variables académicas, socioeconómicas y demográficas, puede proporcionar información valiosa para desarrollar estrategias de apoyo adecuadas. Sin embargo, el manejo de conjuntos de datos de alta dimensionalidad puede resultar complejo.

Justificación del Uso de PCA y Clustering

PCA es una técnica de reducción de dimensionalidad que transforma las variables correlacionadas en un conjunto más pequeño de variables no correlacionadas, denominadas componentes principales, lo que facilita la interpretación y análisis de los datos. K-Means clustering, por otro lado, es un algoritmo de aprendizaje no supervisado que agrupa los datos en clusters basados en similitudes. Estos métodos son útiles cuando no se tiene acceso a etiquetas claras o cuando se desea explorar patrones en los datos sin suposiciones previas.

Objetivo del Artículo

El objetivo de este artículo es aplicar PCA y K-Means clustering a un conjunto de datos de estudiantes para descubrir patrones ocultos y agrupar a los estudiantes sin necesidad de etiquetas previas. El análisis tiene como fin mejorar la comprensión de las relaciones entre las características de los estudiantes y su rendimiento académico, lo que puede contribuir a estrategias de intervención más efectivas.

Metodología

Dataset y sus Características

El conjunto de datos utilizado en este estudio proviene de un proyecto cuyo objetivo es predecir la deserción y el éxito académico de los estudiantes. El dataset incluye 4424 registros de estudiantes con 37 variables, entre las que se encuentran variables académicas, socioeconómicas, demográficas y macroeconómicas. La variable objetivo original clasifica a los estudiantes en tres categorías: Dropout (abandono), Enrolled (matriculado) y Graduate (graduado).

Preprocesamiento

El preprocesamiento de los datos fue esencial para el análisis, e incluyó los siguientes pasos:

1. Imputación de valores faltantes: Se utilizaron la media para imputar los valores faltantes en las variables numéricas.
2. Codificación de variables categóricas: Se utilizó LabelEncoder para transformar las categorías en variables numéricas.
3. Escalado y normalización: Las características numéricas fueron estandarizadas utilizando StandardScaler, ya que el PCA es sensible a la escala de los datos.

Aplicación de PCA

El Análisis de Componentes Principales (PCA) es una técnica que reduce la dimensionalidad de un conjunto de datos conservando la mayor cantidad de varianza posible. El objetivo de PCA es encontrar un conjunto de componentes ortogonales que explican la mayor parte de la varianza en los datos.

Fórmulas para PCA

1. Cálculo de la media: Para cada columna i (dimensión), calculamos la media:

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

2. Centro de los datos: Restamos la media de cada variable:

$$X_{centrado} = X - \mu$$

3. Matriz de covarianza: Se calcula la matriz de covarianza Σ para los datos centrados $X_{centrado}$

$$\Sigma = \frac{1}{n-1} * X_{centrado} X_{centrado}$$

4. Autovalores y autovectores: Para encontrar los componentes principales, resolvemos el sistema característico de la matriz de covarianza Σ:

$$\det det\,(\Sigma - \lambda I) = 0$$

Donde λ son los autovalores y v los autovectores.

Resultados de PCA

Los resultados de la aplicación de PCA en el conjunto de datos mostraron que los primeros 10 componentes principales explican aproximadamente 64.22% de la varianza total, y los 14 componentes explican un 75.43% de la varianza. Esto permite una reducción significativa de dimensionalidad sin perder información clave.

Aplicación de K-Means

El algoritmo de K-Means clustering agrupa los datos en kkk clusters, basándose en las similitudes entre las instancias. El número óptimo de clusters fue determinado utilizando el Silhouette Score, que mide la calidad del clustering.

Fórmula para el Silhouette Score

El Silhouette Score s(i) para un punto i se calcula como:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Donde:

- $a(i)$ es la distancia promedio de i a todos los demás puntos en el mismo cluster.
- $b(i)$es la distancia promedio de i a todos los puntos en el cluster más cercano.

El Silhouette Score varía entre -1 (clusters muy separados) y +1 (buen agrupamiento).

Resultados de K-Means

Se determinaron que 2 clusters son los óptimos según el Silhouette Score, con un valor de 0.63, lo que indica una buena separación entre los clusters. Los estudiantes fueron agrupados en dos clusters principales:

- Cluster 1: Contiene una mezcla de estudiantes Dropout, Enrolled y Graduate.
- Cluster 2: Predomina el grupo de estudiantes Graduate, con una buena proporción de estudiantes Enrolled.

Resultados

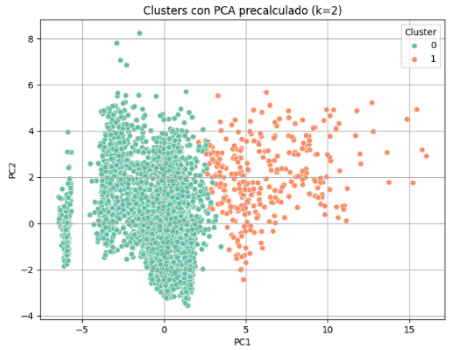
Gráficas de Varianza Explicada en PCA

A continuación, se muestran los gráficos de varianza explicada por cada componente y la varianza acumulada. Estos gráficos ilustran cómo los primeros componentes capturan la mayor parte de la información del dataset.

```
=== Análisis de Componentes Principales (PCA) con 14 componentes ===
Varianza explicada por componente: [0.16177652 0.09453631 0.06906959 0.05785592
 0.05354558 0.04592502 0.04471776 0.04315065
 0.03739322 0.03420326 0.02916098 0.02859319
 0.02803587 0.02635924]
Varianza explicada acumulada: [0.16177652 0.25631283 0.32538242 0.38323834 0.43678392
 0.48270894 0.5274267  0.57057735 0.60797057 0.64217384 0.67133482 0.69992802
 0.72796389 0.75432313]
```

Visualización 2D de los Clusters con PCA

Usando las primeras dos componentes principales, la visualización en 2D de los 2 clusters obtenidos con K-Means muestra cómo se agrupan los estudiantes:



Silhouette Score y Evaluación

El Silhouette Score obtenido para los dos clusters es 0.63, lo que indica una separación moderada entre los grupos generados.

Patrones Encontrados

El análisis mostró que el algoritmo de K-Means logró identificar dos clusters de estudiantes, aunque con algo de solapamiento entre las clases reales de Dropout, Enrolled y Graduate. Esto sugiere que los estudiantes pueden compartir características comunes que no se reflejan completamente en las etiquetas predefinidas.

Utilidad y Confiabilidad de los Clusters

Aunque los clusters tienen cierta variabilidad, pueden ser útiles para segmentar estudiantes y diseñar intervenciones específicas. Sin embargo, la confiabilidad de los clusters debe ser interpretada con precaución debido a la falta de separación clara entre los grupos.

Limitaciones de K-Means

Una limitación importante del K-Means es que el número de clusters debe ser especificado previamente, lo que puede no ser ideal en todos los casos. Además, el algoritmo asume que los clusters son esféricos y de igual tamaño, lo que no siempre es el caso en datasets complejos.

Aplicación en una Universidad

El uso de clustering para identificar grupos de estudiantes con características similares puede ser útil para las universidades. A través de este análisis, las instituciones pueden diseñar programas de apoyo que estén más dirigidos a las necesidades de cada grupo, mejorando la retención y el rendimiento académico.

Discusió

Referencias

- Bishop, C. M. (2006). **Pattern Recognition and Machine Learning**. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning** (2nd ed.). Springer.
- Breiman, L. (2001). **Random Forests**. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324