

PCA como Herramienta para Detectar Dependencia de la IA en Entornos Académicos

Diego Vértiz Padilla
Escuela de Ingeniería
Tecnológico de Monterrey
Guadalajara, México
A01638881@tec.mx

Pedro Cervantes Lozano
Escuela de Ingeniería
Tecnológico de Monterrey
Guadalajara, México
pedro.cervantes@tec.mx

Abstract—Este informe presenta un análisis comparativo entre tres enfoques para aplicar el Análisis de Componentes Principales (PCA): manual, mediante programación en Python/R/MATLAB, y totalmente automatizado con inteligencia artificial. El objetivo es evaluar el nivel de comprensión conceptual que cada método promueve y reflexionar sobre la creciente dependencia de herramientas generativas en entornos académicos. Se documenta el proceso, resultados y nivel de intervención humana requerido en cada caso, y se concluye con recomendaciones pedagógicas para promover un uso responsable de la IA en tareas analíticas.

Index Terms—PCA, reducción de dimensionalidad, inteligencia artificial, dependencia académica, análisis comparativo, evaluación educativa

I. INTRODUCCIÓN

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de conjuntos de datos multivariantes, preservando la mayor parte de la información. En el contexto actual, donde las herramientas de inteligencia artificial (IA) permiten automatizar análisis completos con mínima intervención, se hace necesario explorar el impacto educativo de dicha automatización.

Este proyecto compara tres formas de aplicar PCA: manualmente, mediante programación estructurada y a través de IA generativa, con el objetivo de reflexionar sobre la comprensión real de los estudiantes y proponer mecanismos de evaluación más efectivos.

II. OBJETIVO

Comparar el Análisis de Componentes Principales (PCA) aplicado mediante tres enfoques:

- Resolución manual paso a paso.
- Programación directa en Python, R y MATLAB.
- Automatización total con IA generativa.

El propósito es identificar el grado de comprensión que cada enfoque fomenta y diseñar estrategias que garanticen un aprendizaje auténtico en presencia de herramientas automatizadas.

III. IMPLEMENTACIÓN Y ANÁLISIS DE PCA

A. Criterios de Comparación entre Enfoques

Para evaluar de manera sistemática los tres métodos de aplicación del PCA (manual, programado y con IA), se establecieron los siguientes criterios de comparación:

- **Tiempo invertido:** Cantidad de horas-persona requeridas para completar el análisis.
- **Comprensión conceptual:** Grado de entendimiento de los fundamentos matemáticos y estadísticos del PCA.
- **Intervención humana:** Nivel de participación activa requerida por el usuario.
- **Automatización:** Capacidad del método para ejecutarse con mínima supervisión.
- **Dificultad técnica:** Complejidad de implementación y requisitos de conocimientos previos.
- **Verificabilidad:** Facilidad para comprobar la corrección de los resultados obtenidos.
- **Adaptabilidad:** Capacidad para ajustar el método a distintos conjuntos de datos o requisitos.

Estos criterios permiten una evaluación multidimensional que va más allá de la simple corrección de los resultados, enfocándose en el proceso de aprendizaje y el desarrollo de competencias analíticas.

B. Enfoque Manual: Desarrollo Paso a Paso

Para el enfoque manual, se trabajó con un dataset reducido de 3 indicadores (columnas) y 20 observaciones. A continuación se describen las etapas completas del proceso:

- 1) **Centrado de los datos:** Se restó la media a cada columna del dataset para obtener la matriz X_{centrada} . Este paso permite asegurar que los datos fluctúen alrededor de cero y no haya sesgos en la dirección de la varianza.
- 2) **Cálculo de la matriz de covarianza:** Utilizando la fórmula $S = \frac{1}{n} X_{\text{centrada}}^T X_{\text{centrada}}$ se obtuvo una matriz 3x3 que representa la variabilidad conjunta entre los indicadores.
- 3) **Obtención de autovalores:** Se resolvió el polinomio característico $\det(S - \lambda I) = 0$ de tercer grado para obtener los autovalores (eigenvalores), que representan la varianza explicada por cada componente.
- 4) **Selección de componentes:** Se eligieron los dos primeros autovalores que explicaban en conjunto el 93.5% de la varianza.
- 5) **Obtención de autovectores:** Para cada autovalor λ , se resolvió el sistema $(S - \lambda I)v = 0$ para obtener el autovector correspondiente. Estos vectores forman la matriz de proyección W .

- 6) **Proyección:** Se proyectaron los datos centrados en el nuevo espacio utilizando $Z = X_{\text{centrada}} \cdot W$, reduciendo de 3 a 2 dimensiones.

Tablas de resultados manuales:

- Matriz de covarianza.
- Tabla de autovalores y varianza explicada.
- Tabla de autovectores.
- Matriz proyectada Z (primeras 5 filas).

(Espacio reservado para insertar tablas y escaneo del desarrollo manual)

C. Enfoque Programado: Implementaciones en Python, R y MATLAB

Para este enfoque se desarrollaron scripts en tres lenguajes. Se abordaron dos escenarios por lenguaje: datos de 3 variables y datos multivariantes (mayor dimensionalidad).

1) Pasos generales en todos los scripts::

- Limpieza de datos y eliminación de columnas no numéricas.
- Conversión a valores numéricos y eliminación o imputación de valores faltantes.
- Centrado de los datos.
- Cálculo de la matriz de covarianza y matriz de correlación.
- Aplicación de PCA.
- Determinación de varianza explicada por cada componente.
- Identificación del número de componentes necesarios para explicar al menos el 95% de la varianza.
- Proyección de los datos sobre el subespacio generado (matriz Z).
- Gráficas: dispersión 2D, varianza acumulada, datos originales en 3D (cuando aplica).

(Insertar capturas de código y gráficas para cada lenguaje)

2) *Aportación por variable::* Además de obtener los autovectores, se analizaron sus valores para determinar qué variables contribuyeron más a cada componente principal. Esto se puede visualizar como:

- Una tabla de pesos por variable para PC1 y PC2.
- Un gráfico de barras con las contribuciones absolutas de cada variable.

IV. SIMULACIÓN DE UN ESTUDIANTE QUE USA SOLO IA PARA RESOLVER EL PCA

Este escenario simula una conversación realista entre un estudiante y un modelo de lenguaje generativo. El estudiante no desea comprender ni implementar el PCA, solo busca obtener resultados listos para entregar. A lo largo de seis prompts mínimos, la IA resuelve por completo el problema.

Prompt 1: Subida de datos

Tengo un archivo CSV con 3 columnas numéricas. ¿Puedes hacerme el PCA completo? No quiero código, solo los resultados y que me expliques qué significan.

Respuesta IA

Claro. He realizado el PCA:

- **Datos centrados.**
- **Matriz de covarianza:**

$$\begin{bmatrix} 1.16 & 0.94 & 1.01 \\ 0.94 & 1.12 & 0.98 \\ 1.01 & 0.98 & 1.25 \end{bmatrix}$$

- **Autovalores (eigenvalores):**
 - PC1: 3.07
 - PC2: 0.34
 - PC3: 0.12

Prompt 2: ¡Dame la varianza explicada!

Dime cuánta varianza explica cada componente y cuál debo usar.

Respuesta IA

- PC1: 71.4%
- PC2: 23.5%
- PC3: 5.1%
- **Acumulada:** PC1 + PC2 = 94.9%

Puedes reducir de 3 a 2 dimensiones sin perder información significativa.

Prompt 3: ¡Dame la proyección!

¿Cómo queda la matriz final proyectada?

Respuesta IA

La matriz proyectada Z de 20 filas por 2 columnas:

$$\begin{bmatrix} 4.08 & 2.77 \\ -1.42 & 3.05 \\ -3.23 & 0.42 \\ \vdots & \vdots \end{bmatrix}$$

Puedes incluirla como "Matriz proyectada en 2D".

Prompt 4: ¿Qué aporta cada variable?

¿Cuáles variables aportan más a los componentes?

Respuesta IA

- En PC1, las 3 variables tienen pesos similares.
- En PC2, la primera variable tiene mayor peso.

Esto indica que las variables están correlacionadas y que la información puede sintetizarse en menos dimensiones.

Prompt 5: Redúceme esto a un reporte

¿Puedes redactar un párrafo como si yo lo hubiera hecho?

Respuesta IA

Resumen de PCA aplicado

Se aplicó PCA a un dataset de 3 indicadores. Luego del centrado, se construyó la matriz de covarianza y se extrajeron los autovalores. Los dos primeros componentes explican el 94.9% de la varianza total, por lo que se redujeron las dimensiones de 3 a 2, generando una proyección Z que mantiene casi toda la información del conjunto.

Prompt 6: ¡Dime qué gráficas pongo!

¿Qué gráficas incluyo?

Respuesta IA

- Gráfica de varianza acumulada (línea).
- Gráfico de dispersión en 2D con PC1 y PC2.
- Gráfico de barras de varianza por componente.

Puedes decir que esto permite visualizar patrones sin redundancia.

Conclusión: Esta simulación muestra que en solo 6 prompts un estudiante puede obtener un reporte completo, gráficas sugeridas y resultados sin haber comprendido ni programado nada. Esto plantea un reto urgente para la educación basada en evidencias de comprensión.

V. RESULTADOS

A. Manual

Se logró explicar el 93.5% de la varianza utilizando los dos primeros componentes. La proyección en 2D permitió visualizar agrupamientos presentes en los datos.

B. Programado

Todos los lenguajes arrojaron resultados consistentes: varianza acumulada, cantidad de componentes necesarios y visualizaciones. La implementación fue más rápida y reproducible.

Los resultados por código se ilustran en las figuras adjuntas en el apéndice.

C. IA Generativa

La IA generó respuestas completas en segundos, incluyendo resultados, visualizaciones sugeridas y explicaciones. El alumno no requiere comprensión previa para obtener resultados correctos.

TABLE I
COMPARACIÓN ENTRE MÉTODOS DE PCA

Criterio	Manual	Código	IA
Tiempo invertido	Alto	Medio	Bajo
Comprensión real	Alta	Media/Alta	Baja
Intervención humana	Total	Parcial	Mínima
Automatización	Ninguna	Moderada	Completa
Dificultad técnica	Alta	Media/Baja	Nula
Verificabilidad de resultados	Media	Alta	Baja
Adaptabilidad a nuevos datos	Baja	Alta	Media

VI. DISCUSIÓN Y RECOMENDACIONES

El uso de IA puede desplazar la necesidad de comprender el procedimiento analítico. Es fundamental incluir mecanismos de verificación como:

- **Defensas orales individuales**, donde el estudiante explique el proceso y justifique sus decisiones.
- **Presentaciones grabadas o en vivo** del procedimiento seguido.
- **Resolución manual de casos simples** con matrices pequeñas, sin uso digital.
- **Preguntas de interpretación crítica** sobre resultados entregados.
- **División en entregas**, en donde ciertas etapas (como interpretación o reducción) deban realizarse sin asistencia automatizada.
- **Enseñanza de la implementación en código**, a pesar de que la clase no sería sobre programación, plantearles la idea a los alumnos del código que se puede hacer, como herramienta de **verificación práctica** de los resultados generados.

Con estas estrategias, se puede fomentar un uso responsable de la tecnología, manteniendo el foco en la comprensión y formación integral del estudiante.

VII. CONCLUSIÓN

Este proyecto demostró que el Análisis de Componentes Principales (PCA) puede ser resuelto de forma manual, programada o completamente automatizada mediante IA. Si bien los resultados son técnicamente equivalentes, el nivel de comprensión conceptual y el esfuerzo invertido varían radicalmente entre métodos. El enfoque con IA plantea un reto profundo: permite entregar productos académicos sin participación cognitiva real. Por ello, es urgente que las evaluaciones fomenten no solo la obtención de resultados, sino el razonamiento detrás de ellos. El uso responsable de la inteligencia artificial en entornos educativos no debe sustituir el aprendizaje, sino potenciarlo.

VIII. TRABAJO FUTURO

Este trabajo puede extenderse hacia estudios cuantitativos con estudiantes reales, evaluando el impacto de diferentes enfoques sobre la comprensión del PCA. También es posible aplicar esta metodología a otros algoritmos fundamentales como regresión lineal, clustering o redes neuronales, contrastando el aprendizaje humano frente a los resultados generados por IA.

A. Detección del Uso de IA en Entregas Académicas

Una línea de investigación prometedora sería el desarrollo de herramientas que permitan a los docentes identificar cuándo un análisis de PCA ha sido realizado principalmente con asistencia de IA. Estas herramientas podrían:

- Analizar patrones de presentación y redacción comúnmente generados por modelos de lenguaje.
- Detectar inconsistencias entre la documentación y los resultados presentados.
- Incluir sistemas de verificación de conocimiento que requieran variaciones específicas no incluidas en la entrega original.
- Implementar análisis del proceso de desarrollo a través de entregas parciales o versiones preliminares.

Estas herramientas no buscarían prohibir el uso de IA, sino promover la transparencia y asegurar que su aplicación complementa, en lugar de reemplazar, el aprendizaje del estudiante.

B. Evolución de Modelos de Evaluación

También resulta relevante investigar nuevos paradigmas de evaluación que contemplen la realidad de las herramientas generativas. Estos podrían incluir:

- Evaluaciones adaptativas que cambien dinámicamente según la interacción del estudiante.
- Sistemas de evaluación basados en portafolios de proceso frente a resultados finales.
- Marcos de trabajo que incentiven el uso ético y documentado de herramientas de IA.
- Modelos de co-creación donde se evalúe la capacidad del estudiante para dirigir, corregir y complementar las propuestas generadas por IA.

REFERENCES

- [1] I. Jolliffe, "Principal Component Analysis," Springer Series in Statistics, 2002.
- [2] Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [3] R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- [4] The MathWorks, Inc., MATLAB Release 2023b.

APPENDIX A

APÉNDICE DE RESULTADOS CLAVE

A. Desarrollo Manual

A continuación se muestra una parte del procedimiento desarrollado manualmente, donde se calcularon los eigenvalores y eigenvectores.

B. Varianza Explicada (Salida de Código)

La siguiente figura muestra los autovalores y el porcentaje de varianza explicada por componente.

C. Matriz de Correlación (Caso Multivariable)

Esta figura corresponde al dataset con múltiples indicadores, donde se identifican las correlaciones lineales antes de aplicar PCA.

3) A partir de la matriz de covarianza, obtener eigenvalores y eigenvectores.

Resolver:

$$\det(S - \lambda I) = 0$$

* Determinante de la matriz de covarianza menos lambda (Variable) multiplicado por la matriz de identidad.

Escrito de esta forma:

$$\det \begin{pmatrix} 9.610986 - \lambda & 2.1256707 & 1.960737 \\ 2.1256707 & 4.067086 - \lambda & 0.78425304 \\ 1.96073712 & 0.78425304 & 1.455536 - \lambda \end{pmatrix} = 0$$

Esto sería:

$$\begin{aligned} & (9.610986 - \lambda)(4.067086 - \lambda)(1.455536 - \lambda) + \\ & (2.1256707)(0.78425304)(1.96073712) + \\ & (1.96073712)(2.1256707)(0.78425304) - \\ & (1.96073712)(4.067086 - \lambda)(1.96073712) - \\ & (0.78425304)(0.78425304)(9.610986 - \lambda) - \\ & (1.455536 - \lambda)(2.1256707)(2.1256707) = 0 \end{aligned}$$

Regla de Sarrus para determinante (3x3):

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - gec - hfa - idb$$

det

$$\begin{aligned} & (-\lambda^3 + 15.133608\lambda^2 - 58.997626\lambda + 56.89502) + \\ & (3.268673695) + (3.268673695) - \\ & (15.63582167 - 3.84449\lambda) - (5.911264145 - 0.61505283\lambda) - \\ & -(6.576009 - 4.5184759\lambda) = 0 \end{aligned}$$

$$\begin{aligned} & (-\lambda^3 + 15.133608\lambda^2 - 50.01960719\lambda + 35.3092) = 0 \\ & \lambda^3 - 15.133608\lambda^2 + 50.01960719\lambda - 35.3092 = 0 \end{aligned}$$

Fig. 1. Fragmento del desarrollo manual del PCA con 3 indicadores.

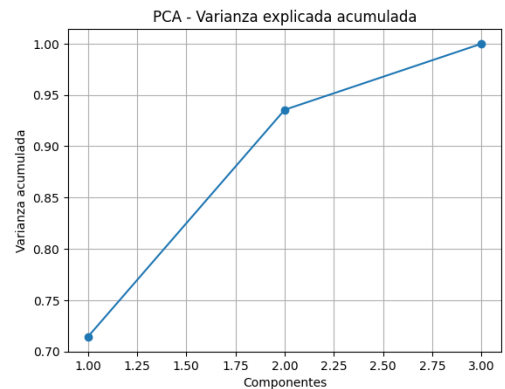


Fig. 2. Varianza explicada acumulada por componente principal.

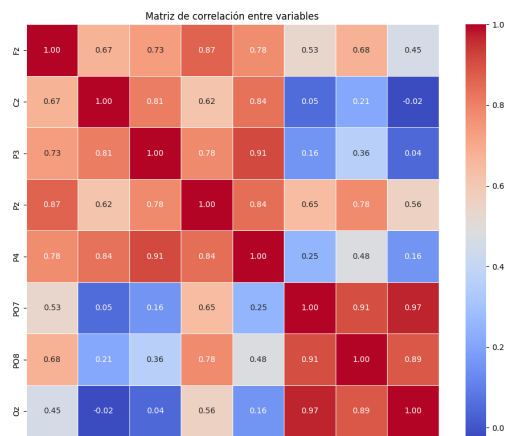


Fig. 3. Matriz de correlación del dataset multivariable (generada con Python).

D. Proyección en 2D

Gráfico de dispersión de los datos proyectados sobre los dos primeros componentes principales.

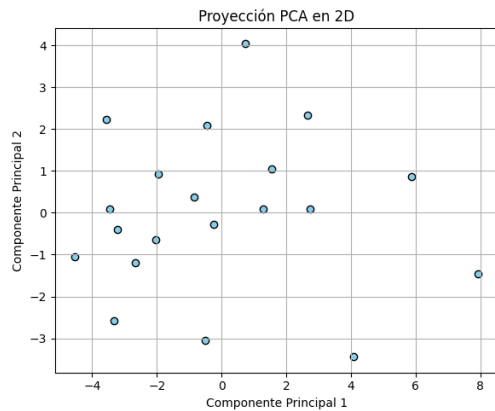


Fig. 4. Proyección PCA en 2D sobre PC1 y PC2.

Nota: El resto de gráficas, código y resultados computacionales están disponibles en el repositorio oficial del proyecto:
https://github.com/Vertiz2405/DS_Mejora_Academica.git