Implementación de una técnica de Estadística Multivariada a una base de datos sobre la prueba SABER 11

Andrés Hernando Borda Muñoz

Facultad de Ciencia y Tecnología

Departamento de Matemáticas

Universidad Pedagógica Nacional

Bogotá, Colombia

2025

Implementación de una técnica de Estadística Multivariada a una base de datos sobre la prueba SABER 11

Andrés Hernando Borda Muñoz

Trabajo de grado presentado como requisito parcial para optar al título de:

Licenciado en Matemáticas

Director:
Prof. César Rendón Mayorga, MSc.

Departamento de Matemáticas
Universidad Pedagógica Nacional
Facultad de Ciencia y Tecnología
Bogotá, Colombia
2025

"Si no eres capaz de explicar algo claramente,
es que aún no lo has entendido lo suficiente."
Albert Einstein

Implementación de una técnica de Estadística Multivariada a una base de datos sobre la prueba SABER 11

Declaración

Me permito afirmar que he realizado este trabajo de grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el presente texto. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá, 19 de agosto de 2025

Andrés Hernando Borda Muñoz

Implementación de una técnica de Estadística Multivariada a una base de datos sobre la prueba SABER 11

Agradecimientos

Este trabajo de grado representa la culminación de mi ciclo formativo como Licenciado en Matemáticas. Por ende, es el momento de agradecer y enaltecer el nombre de Dios, quien me ha brindado la salud, el equilibrio espiritual y los recursos materiales necesarios para concluir mi trasegar por este programa académico.

Asimismo, agradezco a mi padre, Hernando Borda, por tener siempre una fe ciega en mis capacidades, así como el amor que me manifiesta día a día. Es un hombre sin igual a quien espero devolverle "al menos" una fracción de todo lo que ha hecho por mí.

Finalmente, agradezco a la Universidad Pedagógica Nacional por acogerme en sus brazos, a su cuerpo profesoral y en especial a mi director de trabajo de grado el profesor César Rendón, quien con su guía y paciencia se configura como un pilar fundamental para el desarrollo de este documento y un ejemplo a seguir.

Resumen

Implementación de una técnica de Estadística Multivariada a una base de datos sobre la prueba SABER 11

En este trabajo de grado se presentan los resultados de la implementación de las técnicas multivariadas análisis de componentes principales (PCA por sus siglas en inglés) y análisis de *clusters* (*k-means*) a una base de datos de las pruebas educativas SABER-11.

Las técnicas implementadas en Python mediante el entorno de desarrollo de Google Colab, identificarón tres *clusters* de estudiantes (Bajo, Medio y Alto rendimiento) definidos por sus puntajes y características socioeconómicas como el acceso a internet y el estrato. Se concluye que estas técnicas son efectivas para segmentar y comprender la estructura de los datos de pruebas estandarizadas, revelando perfiles de rendimiento diferenciados.

Palabras clave: Estadística Multivariada, Prueba SABER 11, Análisis de Componentes Principales, Análisis de *Clusters*.

Abstract

Implementation of Multivariate Statistical Techniques on a SABER 11 Test Database

In this thesis, the results of implementing multivariate techniques of principal component analysis (PCA) and cluster analysis (k-means) on a database of SABER-11 educational tests are presented.

The techniques implemented in Python using the Google Colab development environment, identified three student clusters (Low, Medium, and High performance) defined by their scores and socioeconomic characteristics such as internet access and socioeconomic stratum. It is concluded that these techniques are effective for segmenting and understanding the structure of standardized test data, revealing differentiated performance profiles.

Keywords: Multivariate Statistics, SABER 11 Test, Principal Component Analysis, Cluster Analysis.

Índice de figuras

Índice de tablas

Lista de códigos

Índice general

Capítulo 1

Introducción

El sistema educativo colombiano enfrenta desafíos constantes respecto a los resultados de las pruebas estandarizadas que realiza año tras año (tanto del orden nacional como del orden internacional), ya que se evidencian diferencias con respecto al desempeño y brechas de calidad. Específicamente, la prueba SABER 11 implementada por el ICFES[1] se concibe como "el instrumento de evaluación estandarizada que mide oficialmente la calidad de la educación formal impartida a quienes terminan el nivel de educación media" (ICFES, 2025), y se constituye en un insumo fundamental para la estructuración de políticas públicas del sector educativo nacional.

Específicamente, en una investigación liderada por el Laboratorio de Economía de la Educación de la Pontificia Universidad Javeriana se muestran como los puntajes promedios globales de la prueba SABER 11 entre el 2014 al 2023 han tenido fluctuaciones importantes como se observa en la Figura 1.1.

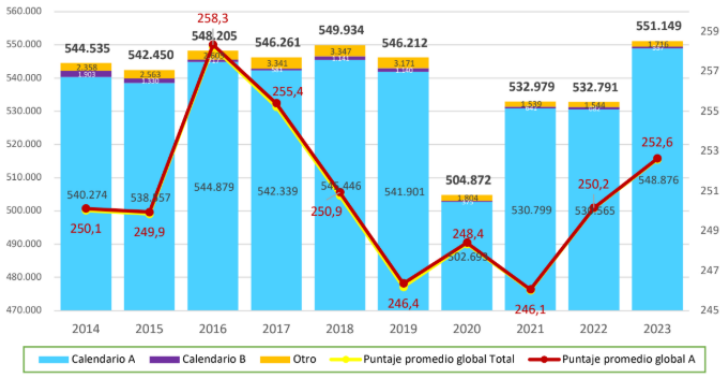


Figura 1.1: N° de evaluados según calendario y puntaje promedio global Saber 11 Nacional 2014-2023 (LEE, 2024)

En particular, en el período comprendido entre 2019 y 2022 se exhiben variaciones

[1]Instituto Colombiano para la Evaluación de la Educación.

importantes, esto en el contexto pre (2019), durante (2020 - 2021) y post pandemia de COVID-19 (2022) hecho histórico que según la UNESCO[2] (2024) provocó una crisis educativa sin precedentes a nivel global, donde el cierre de las escuelas tuvo un impacto directo en el aprendizaje de los estudiantes afectando profundamente su interacción con los docentes, sus pares y su vínculo con la escuela. Además, se presume fundadamente que se produjeron pérdidas de aprendizaje significativas y un aumento de las brechas de resultados, según el nivel socioeconómico de las familias de los estudiantes.

En el caso colombiano, y con base en la Figura 1.2, se observa un incremento significativo en las brechas de los puntajes globales promedio de las pruebas SABER 11 entre colegios oficiales y no oficiales apoyando en cierta medida lo planteado por la UNESCO.

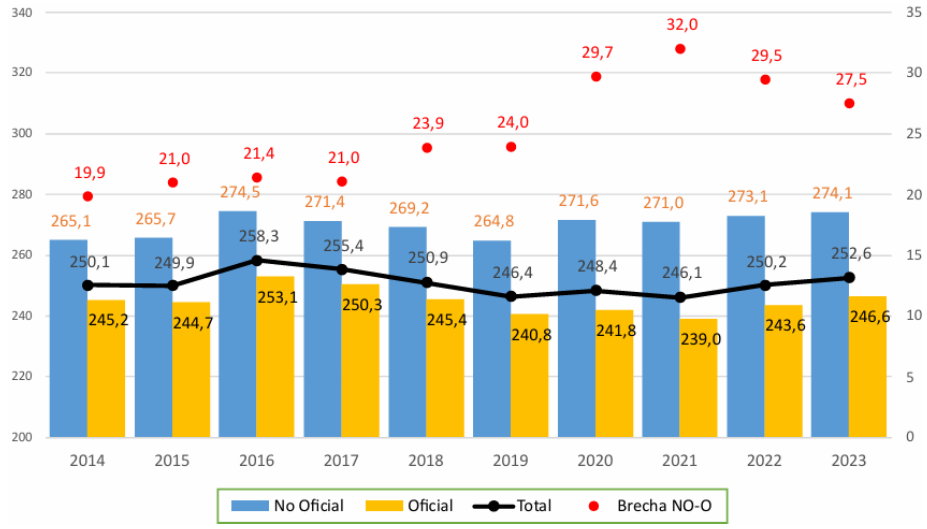


Figura 1.2: Desempeño global: Pruebas Saber 11 Nacional 2014-2023 (LEE, 2024)

Bajo este panorama, surge la inquietud de analizar de manera sistemática y objetiva los diversos factores asociados al desempeño en las pruebas SABER 11. Para lo cual, el presente trabajo de grado expone la implementación de técnicas de estadística multivariada, específicamente el análisis de componentes principales (PCA) y análisis de *clusters*, con el objetivo de identificar patrones o relaciones entre las variables asociadas a la base de datos Resultados únicos SABER 11.ºbtenida del portal web Datos Abiertos una iniciativa del gobierno colombiano para la transparencia y la toma de decisiones basada en datos públicos de libre acceso y sin costo (MIN-TIC, 2025).

[2]United Nations Educational, Scientific, and Cultural Organization

Capítulo 2

Planteamiento del problema

2.1. Descripción del problema

El sistema educativo colombiano está en constante cuestionamiento debido al bajo rendimiento de sus estudiantes en las pruebas estandarizadas. Se cuestiona la idoneidad de la calidad y equidad de la educación que imparte, poniendo en tela de juicio si los estudiantes reciben una "educación de calidad" que les permita adquirir las competencias básicas en áreas fundamentales como matemáticas, lectura crítica, ciencias naturales, ciencias sociales e inglés que evalúan las pruebas SABER 11.

Asimismo, este panorama se recrudece por la inequidad en el acceso a las diversas instituciones educativas del país, lo cual está directamente relacionado al origen socioeconómico, la ubicación geográfica (urbana o rural) y el tipo de institución (pública o privada) a la que asisten los estudiantes.

Estas problemáticas se evidencian en los resultados de las pruebas SABER 11. Por ejemplo, paneles como "¿Crisis en la educación media en Colombia?" alertan sobre la influencia de la calidad y la equidad en la educación, bajo las siguientes premisas (Universidad de los Andes, 2024):

- Aunque la pandemia agudizó el problema, el país cumplirá una década en que los y las jóvenes próximos a graduarse no superan los 260 puntos de 500 posibles en las pruebas SABER 11.
- La brecha en las pruebas entre colegios no oficiales y oficiales es de 33 puntos. Esa misma diferencia aplica en las instituciones educativas urbanas frente a las rurales.
- Si se compara el Índice de Nivel Socioeconómico (INSE), la desigualdad en las pruebas Saber 11 es aún mayor (67 puntos). Mientras el promedio de jóvenes del nivel

más alto es de 292 puntos, el del nivel más bajo es de 225.

Los resultados históricos muestran fluctuaciones importantes en el rendimiento académico. Como se observa en la Figura 1.1, los puntajes promedio globales han experimentado variaciones considerables entre 2014 y 2023, con una tendencia particularmente preocupante durante el período 2019-2022. Esta situación se empeora cuando se consideran las desigualdades regionales y socioeconómicas existentes en nuestro país.

2.2. Formulación del problema

2.2.1. Pregunta principal

Teniendo en cuenta el panorama expuesto en la sección anterior, se pretende con este trabajo realizar un estudio preliminar bajo el siguiente cuestionamiento:

¿Cuáles son algunos de los factores que explican las fluctuaciones en el desempeño académico de los estudiantes en las pruebas SABER 11[1] y cómo estos factores que se identifiquen se relacionan con las características socioeconómicas, institucionales o regionales?

2.3. Justificación

2.3.1. Importancia teórica

El análisis sistemático de los factores asociados al desempeño académico en las pruebas SABER 11 contribuye a dar una mirada al conocimiento sobre evaluación educativa estandarizada. Como señalan diversos estudios, analizar los factores asociados que inciden en los desempeños de los estudiantes (Min-Educación, 2022) es fundamental para comprender la complejidad del proceso educativo.

Cabe resaltar que según Diaz-Monroy (2007): cada respuesta o atributo está asociado con una variable; si tan sólo se registra un atributo por individuo, los datos resultantes son de tipo univariado, mientras que si más de una variable es registrada sobre cada objeto, los datos tienen una estructura multivariada.

Aun más, pueden considerarse grupos de individuos, de los cuales se obtienen muestras de datos multivariados para comparar algunas de sus características o parámetros. En una forma más general, los datos multivariados pueden proceder de varios grupos o poblaciones de objetos; donde el interés se dirige a la exploración de las variables y la búsqueda de su interrelación dentro de los grupos y entre ellos.

[1] Enlace a la base de datos estudiada.

El enfoque del análisis multivariado propuesto permite superar las limitaciones de los análisis tradicionales univariados, proporcionando una visión más integral de las múltiples dimensiones que caracterizan el desempeño académico. Esta mirada es fundamental para avanzar hacia una comprensión más profunda de los procesos educativos y sus factores determinantes.

2.3.2. Relevancia Práctica

Los futuros hallazgos de este estudio podrían ser un insumo basado en evidencia para la discusión y la toma de decisiones sobre políticas educativas, debido a que la identificación de factores relacionados con el rendimiento académico podría sugerir posibles acciones para contribuir a la mejora de la calidad educativa y la reducción de brechas.

En el caso de que los resultados no arrojen hallazgos significativos, esto también sería un aporte valioso, ya que señalaría una ruta metodológica a descartar y justificaría una reevaluación de las técnicas o métodos utilizados.

Además, el uso de un conjuntos de datos proveniente del portal "Datos Abiertos" adscrito al gobierno colombiano contribuye a la transparencia y la democratización del acceso a la información sobre evaluación educativa, permitiendo que cualquier ciudadano pueda acceder, descargar, analizar y utilizar esta información de forma libre y gratuita.

2.4. Objetivos

2.4.1. Objetivo general

Implementar una técnica de estadística multivariada, a través de software estadístico, a la base de datos sobre las pruebas SABER 11, con el fin de reconocer relaciones entre variables que puedan incidir en los resultados de dicha prueba.

2.4.2. Objetivos específicos

1. Realizar una contextualización y descripción de los datos presentes en la base de datos acerca de la prueba SABER 11 que además permita reconocer la técnica multivariada a emplear.

2. Establecer el marco matemático a partir de una revisión bibliográfica que sustente las técnicas estadísticas a utilizar.

3. Identificar y representar posibles relaciones a partir de la implementación de la técnica de estadística multivariada.

4. Sistematizar y analizar los resultados obtenidos del análisis estadístico multivariado.

Capítulo 3

Marco de referencia

La estadística multivariada es una de las ramas fundamentales de las matemáticas aplicadas que permite el análisis simultáneo de múltiples variables medidas sobre un conjunto de datos.

Actualmente, el análisis de datos y la toma de decisiones basada en estos son competencias altamente demandadas por el mercado y la sociedad, razón por la cual las técnicas multivariadas se han convertido en herramientas indispensables para el estudio de grandes volúmenes de datos, reduciendo la dimensionalidad de estos y descubriendo estructuras o patrones propios de la información en estudio. Este marco de referencia presenta los fundamentos matemáticos y metodológicos necesarios para comprender y aplicar dos técnicas multivariadas a saber: el análisis de componentes principales (PCA) y el análisis de *clusters*.

El desarrollo de este capítulo sigue una secuencia la cual inicia con los conceptos estadísticos fundamentales de población, muestra y cálculo de tamaños muestrales, estableciendo las bases para un análisis estadísticamente válido. Luego, se introducen los principios de la estadística multivariada, incluyendo la representación matricial de los datos y las técnicas de visualización que facilitan la interpretación de estructuras de datos multidimensionales. La exposición teórica se centra específicamente en el PCA como método de reducción dimensional y en el análisis de *clusters* como técnica de agrupamiento, proporcionando tanto la fundamentación matemática como un esbozo de los algoritmos computacionales necesarios para su implementación.

La selección de estas técnicas específicas responde a su complementariedad en el análisis exploratorio de datos: mientras que el PCA permite identificar las direcciones de máxima variabilidad en un conjunto de datos y reducir su dimensionalidad preservando la mayor cantidad de información posible, el análisis de *clusters* facilita la identificación de grupos o patrones de similitud entre observaciones.

El desarrollo teórico que se presentará a continuación requiere que el lector posea una base matemática sólida, cimentada en tres pilares fundamentales que se entrelazan entre sí. El primer pilar se centra en el álgebra lineal, donde se asume el dominio de los siguientes conceptos: operaciones matriciales, transformaciones lineales, matrices simétricas, valores y vectores propios, diagonalización, rango matricial y espacios vectoriales constituyendo la base conceptual para comprender la descomposición espectral que fundamenta el Análisis de Componentes Principales. Adicionalmente, para complementar este pilar, la geometría analítica aporta la "intuición visual" indispensable para interpretar proyecciones ortogonales, rotaciones de sistemas de coordenadas y la representación de estructuras de datos en espacios multidimensionales.

El segundo pilar se centra en la estadística, donde se debe tener como prerrequisito conocimiento de los conceptos fundamentales de varianza, covarianza y correlación, interpretación de matrices de covarianza como la interpretación de estadísticos de dependencia lineal entre variables. El conocimiento de la distribución normal multivariada y de los principios de inferencia estadística resulta crucial para comprender los criterios de validación y los fundamentos probabilísticos que sustentan las técnicas multivariadas. Finalmente, el tercer pilar incorpora nociones de optimización multivariada, incluyendo el cálculo de gradientes y la búsqueda de extremos condicionados, elementos esenciales para entender tanto los algoritmos iterativos del análisis de *clusters* como los procesos de maximización de varianza característicos del PCA.

La integración de estos tres pilares (álgebra, estadística y cálculo) permite una aproximación a la conceptualización de las técnicas multivariadas que se desarrollan en las siguientes secciones.

3.1. Tamaño muestral

3.1.1. Definiciones iniciales

Tanto para el lector como para la estructura de este estudio es pertinente recordar las definiciones fundamentales para el desarrollo del muestreo estadístico.

Población y Muestra

Población: Una población $\mathcal{P}$ es el conjunto completo de elementos o unidades de análisis que poseen las características de interés para un estudio particular. Formalmente,

si denotamos cada elemento como u_i, entonces:

$$\mathcal{P} = \{u_1, u_2, ..., u_N\} \tag{3.1}$$

donde N es el tamaño de la población.

Muestra: Una muestra $\mathcal{S}$ es un subconjunto de la población, seleccionado mediante un procedimiento específico. Formalmente:

$$\mathcal{S} = \{u_{i_1}, u_{i_2}, ..., u_{i_n}\} \tag{3.2}$$

donde $\{i_1, i_2, ..., i_n\}$ es un subconjunto de índices de $\{1, 2, ..., N\}$ y n es el tamaño de la muestra, con $n \leq N$.

Proporción poblacional: La proporción poblacional p de elementos que poseen una característica específica A se define como:

$$p = \frac{N_A}{N} \tag{3.3}$$

donde $N_A = |\{u_i \in \mathcal{P} : u_i \text{ posee la característica } A\}|$.

3.1.2. Cálculo del tamaño de muestra para estimación de proporciones

Para una población finita, el tamaño de muestra requerido para estimar una proporción con un nivel de confianza $(1 - \alpha)$ y un margen de error e está dado por:

$$n_0 = \frac{z^2_{\alpha/2} \cdot p \cdot (1 - p)}{e^2} \tag{3.4}$$

Donde:

- $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar
- p es la proporción poblacional (usamos $p = 0{,}5$ para maximizar la varianza)
- e es el margen de error deseado

Para poblaciones finitas, aplicamos la corrección:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{3.5}$$

3.2. Estadística Multivariada

El núcleo de este trabajo de grado es el estudio y la selección de algunas técnicas de estadística multivariada (EM) para analizar una base de datos de las pruebas SABER 11, pero ¿qué es la estadística multivariada? ¿cuáles y cuántas técnicas de EM existen? ¿todas las técnicas de EM son aplicables a cualquier conjunto de datos?

Para dar respuesta a estos interrogantes a continuación se toma como referencia el libro "Análisis estadístico de datos multivariados" una producción del Departamento de Estadística de la Universidad Nacional de Colombia (Diaz-Morales, 2012). En esta obra, la estadística multivariada se define como una rama de la estadística que se ocupa del análisis simultáneo de múltiples variables medidas sobre un conjunto de individuos u objetos.

3.2.1. Datos Multivariados

Variable estadística: Una variable es una característica o atributo que puede ser observado o medido sobre un conjunto de individuos u objetos. En el contexto multivariado, cada variable se refiere a una dimensión específica del espacio de análisis.

- Un **dato univariado** registra un solo atributo por individuo.
- Un **dato multivariado** registra múltiples atributos sobre cada individuo u objeto, generando una estructura multidimensional.

Matriz de Datos: Una **matriz de datos X** es una estructura de tamaño $n \times p$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

donde n representa el número de individuos (filas) y p el número de variables (columnas).

3.3. Representación de datos multivariados

Representación tridimensional: Los datos multivariados pueden conceptualizarse como un prisma tridimensional como se observa en la Figura 3.1 donde:

- **Objetos (O)**: Individuos o unidades de análisis

- **Variables (V)**: Características o atributos medidos
- **Tiempo (T)**: Instante o período de medición

Un punto X_{ijt} representa el valor del atributo j para el individuo i en el tiempo t.

Las diferentes técnicas estadísticas trabajan en alguna región de este prisma. Por ejemplo, las regiones paralelas al plano OV son estudiadas por la mayoría de las técnicas del análisis multivariado; a veces se les llama estudios transversales, de las regiones paralelas a VT se ocupan los métodos de series cronológicas (estudios longitudinales). Es fundamental comprender que en la mayoría de los procedimientos estadísticos se consideran constantes o fijos algunos de los tres componentes.

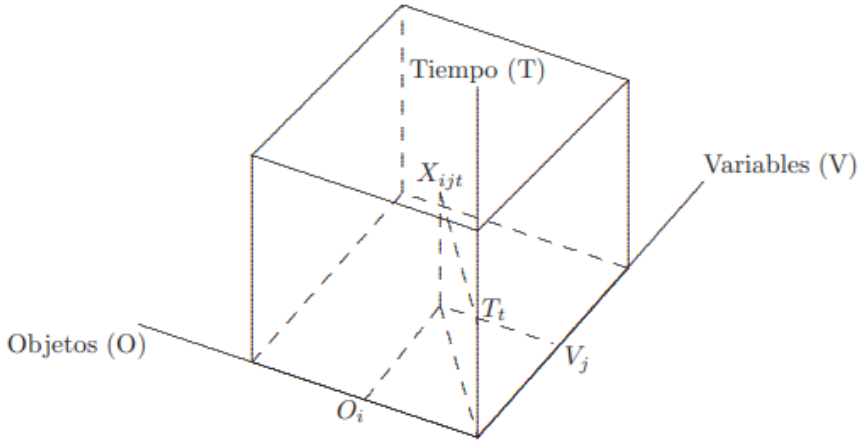


Figura 3.1: Representación multivariada de datos. (Diaz & Morales, 2012)

3.3.1. Tipos de representaciones

El objetivo de estos tipos de representaciones es facilitar la lectura e interpretación acerca de la información contenida en los datos, por ende las gráficas no deben ser más complejas de leer que los datos en bruto. Algunas de estas representaciones son y se observan en las Figuras 3.2 y 3.3 :

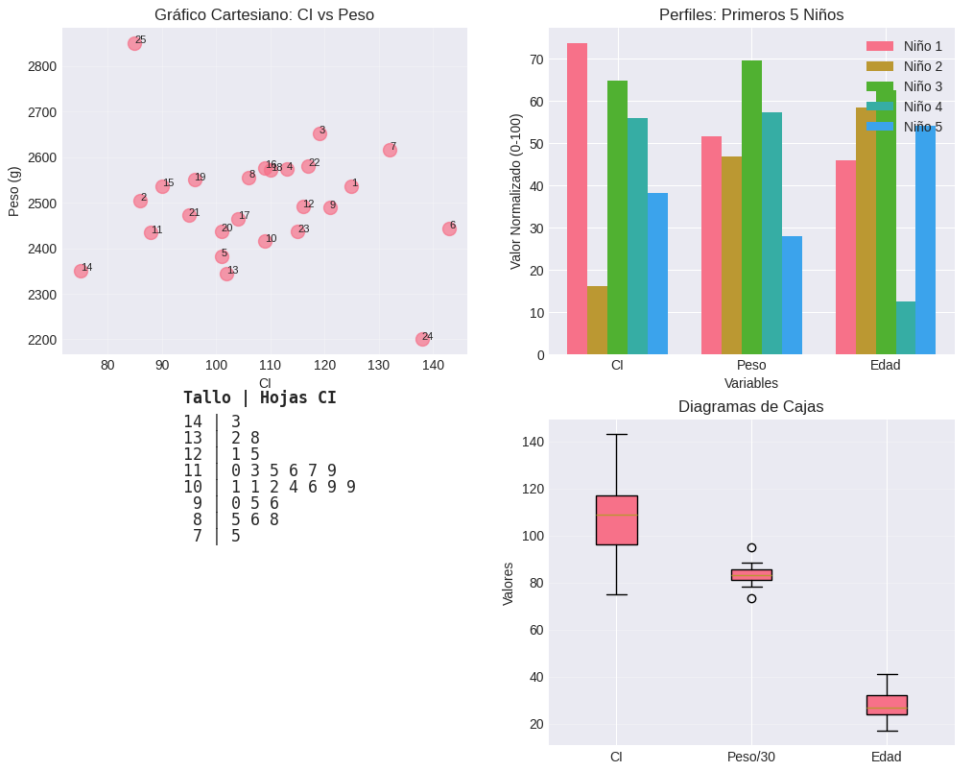
Gráficos Cartesianos: Representación en un plano mediante la elección de variables cuantitativas, donde cada individuo se representa como un punto en el espacio definido por las variables seleccionadas.

Perfiles: Representación tipo histograma donde cada barra corresponde a una variable y su altura al valor de la misma, permitiendo visualizar el patrón de cada objeto.

Diagramas de Tallo y Hojas: Procedimiento gráfico para representar datos cuantitativos que permite visualizar la distribución manteniendo la información de los valores individuales.

Diagramas de Cajas (Box-plots): Representación gráfica que muestra la mediana, cuartiles y valores extremos de una distribución, facilitando la identificación de valores atípicos y la comparación entre variables.

Diagramas de dispersión: Son gráficos en los cuales se representan los individuos u objetos por puntos asociados a cada par de coordenadas (valores de cada par de variables).



Tallo | Hojas CI

Tallo	Hojas
14	3
13	2 8
12	1 5
11	0 3 5 6 7 9
10	1 1 2 4 6 9 9
9	0 5 6
8	5 6 8
7	5

Figura 3.2: Técnicas de visualización. Construcción propia.

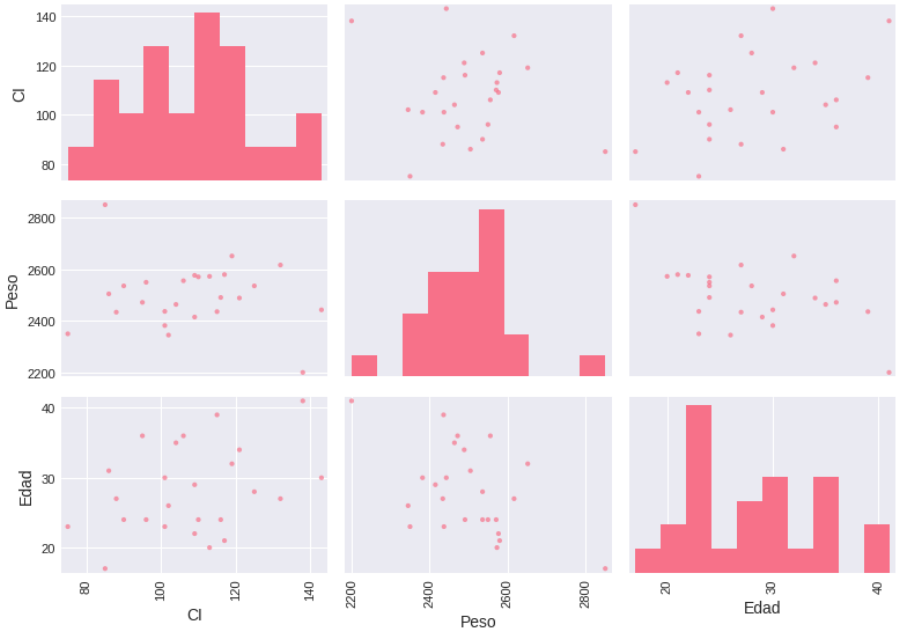


Figura 3.3: Matriz de diagramas de dispersión. Construcción propia.

3.4. Técnicas del Análisis Multivariado

Las técnicas del Análisis Multivariado (AM) tratan con datos asociados a conjuntos de "medidas" sobre un número de individuos u objetos. El enfoque de este estudio se centra en la clasificación de las técnicas por dependencia e interdependencia, las cuales se definen así:

- **Dependencia:** Interesa hallar la asociación entre dos conjuntos de variables, en el cual uno es considerado como la realización de mediciones dependientes de otro conjunto de variables.
- **Interdependencia:** El propósito es estudiar la interdependencia entre las variables. Esta puede examinarse desde la independencia total de las variables hasta la dependencia de alguna con respecto a un subconjunto de variables (colinealidad).

3.4.1. Técnicas asociadas a dependencia o asociadas a interdependencia

A continuación se expone una breve descripción de cada una de las técnicas reportadas en la anterior tabla.

Tabla 3.1: Clasificación de Técnicas de Análisis Multivariado

Análisis de Dependencia	**Análisis de Interdependencia**
Regresión Múltiple	Análisis de Componentes Principales (PCA)
Análisis Discriminante	Análisis Factorial
Análisis de Correlación Canónica	Análisis de Correspondencia
Análisis Logit	Análisis de *clusters*
Análisis de Varianza Multivariado	Escalamiento Multidimensional
Análisis Conjunto	Modelos Log-lineales

3.4.2. Métodos de dependencia

- **Regresión Múltiple:** Técnica que centra su análisis en la dependencia de una variable respuesta respecto a un conjunto de variables regresoras o predictoras, mediante un modelo de regresión que mide el efecto de cada variable regresora sobre la respuesta.
- **Análisis Discriminante:** Técnica que asigna un individuo a uno de varios grupos definidos de antemano, basándose en las características (variables) del individuo y la información disponible sobre los grupos.
- **Análisis de Correlación Canónica:** Busca una relación lineal entre un conjunto de variables predictoras y un conjunto de criterios medidos u observados, inspeccionando combinaciones lineales para las variables predictoras y criterio.
- **Análisis Logit:** Caso especial del modelo de regresión donde el criterio de respuesta es de tipo categórico o discreto, dirigiéndose a investigar los efectos de un conjunto de predictores sobre la respuesta.
- **Análisis de Varianza Multivariado:** Técnica para evaluar múltiples criterios (tratamientos) y determinar su efecto sobre una o más variables respuesta en un experimento, permitiendo comparar vectores de medias asociados a varias poblaciones.
- **Análisis Conjunto:** Técnica que trata la evaluación de un producto o servicio con base en las cualidades que éste requiere o esperan sus consumidores o usuarios, buscando la combinación óptima de atributos.

3.4.3. Métodos de interdependencia

- **Análisis de Componentes Principales**: Técnica de reducción de datos cuyo objetivo es construir combinaciones lineales (componentes principales) de las variables

originales que contengan la mayor proporción de la variabilidad total original.

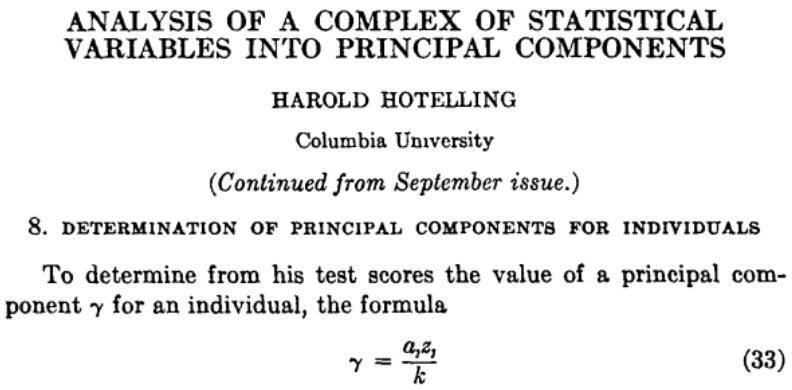
- **Análisis Factorial**: Describe cada variable en términos de una combinación lineal de factores comunes no observables y un factor único para cada variable, buscando los factores que recojan el máximo de información de las variables originales.
- **Análisis de Correspondencias**: Método dirigido al análisis de tablas de contingencia que busca conseguir la mejor representación simultánea de los dos conjuntos de datos contenidos en la tabla (filas y columnas).
- **Análisis de *Clusters***: Técnica de reducción de datos cuyo objetivo es la identificación de grupos similares respecto a sus variables, garantizando cercanía o similitud entre los objetos de un mismo grupo.
- **Escalamiento Multidimensional**: Permite explorar e inferir criterios que la gente utiliza en la formación de percepciones acerca de la similitud y preferencia entre objetos, transformando las similaridades percibidas en distancias para ubicar los objetos en un espacio multidimensional.
- **Modelos Log-lineales**: Permiten investigar la interrelación entre variables categóricas que forman una tabla de contingencia, expresando las probabilidades de las celdas en términos de efectos principales e interacción.

Dado un panorama general y sucinto de la estadística multivariada la siguiente sección centrará su atención en dos técnicas principalmente: Análisis de Componentes Principales (PCA) y Análisis de *Clusters*, ya que son las técnicas escogidas para analizar la base de datos y en el capítulo denominado "Metodología" se justificará su elección.

3.5. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA por sus siglas en inglés) como se observó en la sección anterior es una técnica de estadística multivariada de reducción de dimensionalidad, lo cual permite simplificar la complejidad de grandes conjuntos de datos mientras se conserva la mayor cantidad posible de información relevante (Jolliffe-Cadima, 2016). En esencia, el PCA transforma un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables no correlacionadas denominadas componentes principales, organizadas de manera que las primeras componentes capturan la mayor variabilidad presente en los datos originales.

La técnica fue desarrollada por Karl Pearson en 1901 como un razonamiento análogo del teorema de los ejes principales en mecánica, y posteriormente fue estudiada de forma independiente por Harold Hotelling en la década de los 30's y como curiosidad histórica anexo un recorte del *paper* original en la Figura 3.4.



ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS

HAROLD HOTELLING

Columbia University

(*Continued from September issue.*)

8. DETERMINATION OF PRINCIPAL COMPONENTS FOR INDIVIDUALS

To determine from his test scores the value of a principal component γ for an individual, the formula

$$\gamma = \frac{a_i z_i}{k} \qquad (33)$$

Figura 3.4: Paper de Harold Hotelling.(Hotelling, 1933)

Desde aquel entonces, la técnica se ha convertido en una de las más utilizadas en análisis multivariado, con aplicaciones en diferentes áreas del saber como la biología, la psicología, la economía, la ingeniería, las ciencias sociales y ciencias de la computación.

El PCA es útil cuando se trabaja con conjuntos de datos con las siguientes características:

- **Alta dimensionalidad**: Cuando el número de variables es grande y se desea una representación "manejable".
- **Multicolinealidad**: Cuando existe correlación entre las variables originales.
- **Redundancia de información**: Cuando varias variables miden aspectos similares del fenómeno de interés.
- **Necesidad de visualización**: Cuando se requiere representar gráficamente datos multidimensionales.

El PCA puede aplicarse a variables cuantitativas continuas o discretas, siempre que estén medidas en escalas numéricas. Cabe resaltar que la técnica **asume** relaciones lineales entre las variables y es sensible a las diferencias en las escalas de medición, por lo que frecuentemente se requiere la estandarización de los datos (Jolliffe-Cadima, 2016).

3.6. Interpretación geométrica de las Componentes Principales

Desde una perspectiva geométrica, el PCA puede conceptualizarse como un proceso de rotación del sistema de coordenadas original para alinearlo con las direcciones de máxima variabilidad de los datos (Hastie et al., 2009). Es importante indicar que, para la elaboración de las siguientes secciones relativas al PCA, se ha tomado como referente bibliográfico fundamental el libro "*Methods of Multivariate Analysis*" de Rencher y Christensen (2012).

3.6.1. Representación espacial de los datos

Consideremos un conjunto de datos con p variables que pueden representarse como puntos en un espacio p-dimensional. En este espacio, cada observación corresponde a un punto cuyas coordenadas están determinadas por los valores de las p variables. La nube de puntos resultante tiene una forma particular que refleja la estructura de correlación presente en los datos.

Cuando las variables están correlacionadas, la nube de puntos no se distribuye de manera uniforme en todas las direcciones del espacio. En cambio, tiende a concentrarse a lo largo de ciertas direcciones preferenciales, formando estructuras elongadas que pueden asemejarse a elipses (en dos dimensiones) o elipsoides (en dimensiones superiores).

El PCA identifica iterativamente las direcciones en las cuales los datos presentan la mayor variabilidad. La primera componente principal corresponde a la dirección en la cual la proyección de los datos tiene la varianza máxima. Geométricamente, esto equivale a encontrar la línea recta que mejor se ajusta a la nube de puntos en el sentido de mínimos cuadrados.

La segunda componente principal es la dirección ortogonal a la primera que maximiza la varianza de las proyecciones. Este proceso continúa hasta obtener p componentes principales, todas mutuamente ortogonales entre sí. Un ejemplo de una representación de un PCA con tres componentes principales es la representación de la Figura 3.5.

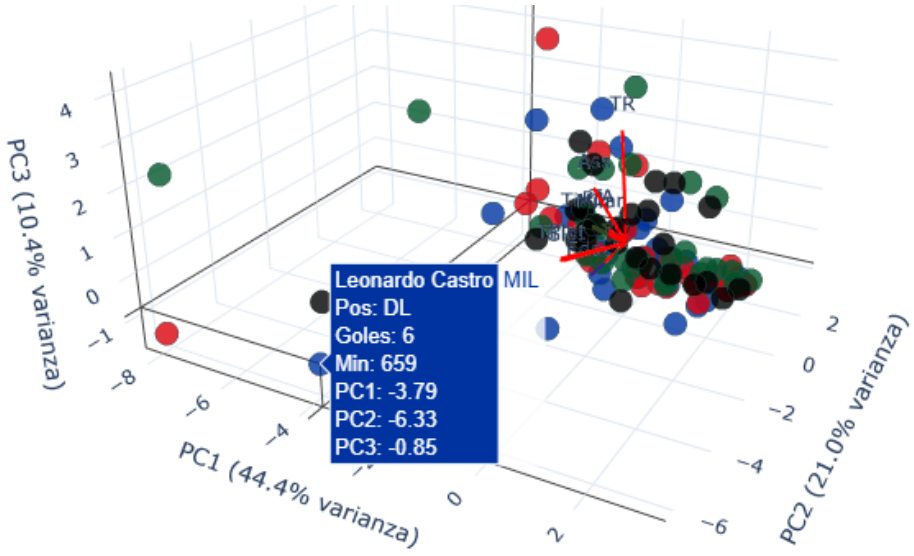


Figura 3.5: Análisis de componentes principales. Tres dimensiones. Construcción propia.

3.6.2. Transformación de coordenadas

Matemáticamente, si tenemos un vector de observaciones $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ centrado en el origen, la transformación a componentes principales se expresa como:

$$\mathbf{z}_i = \mathbf{A}'\mathbf{y}_i \tag{3.6}$$

donde $\mathbf{A}$ es la matriz de coeficientes (vectores propios) y $\mathbf{z}_i$ son las puntuaciones en las componentes principales.

Esta transformación representa una rotación rígida del sistema de coordenadas original, preservando las distancias entre puntos y la estructura geométrica general de los datos. La diferencia fundamental es que el nuevo sistema de coordenadas está alineado con las direcciones de máxima variabilidad.

3.6.3. Elipsoide de concentración

Una interpretación geométrica útil es entender el PCA como el ajuste de un elipsoide p-dimensional a los datos (Mardia et al., 1980). Los ejes de este elipsoide corresponden a las componentes principales, y sus longitudes son proporcionales a las raíces cuadradas de los valores propios correspondientes.

Los ejes más largos del elipsoide indican las direcciones de mayor variabilidad, mientras que los ejes más cortos corresponden a direcciones con menor variabilidad como se observa en la Figura 3.6.

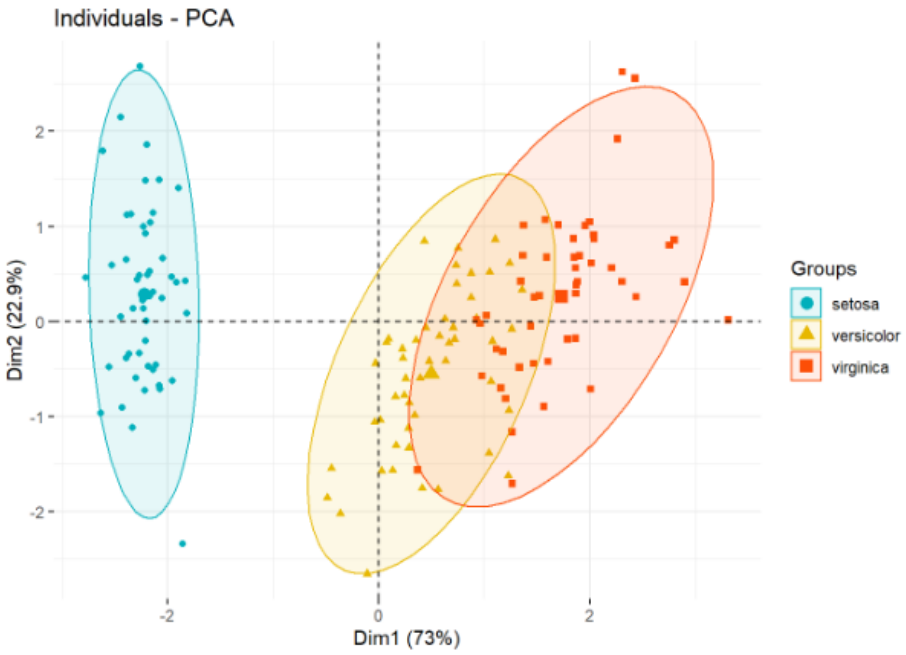


Figura 3.6: Elipse de concentración. (Albornoz et al., 2022)

3.6.4. Proyección y reducción dimensional

La reducción de dimensionalidad se logra proyectando los datos originales sobre un subespacio generado por las primeras k componentes principales (donde $k < p$). Geométricamente, esto equivale a proyectar la nube de puntos p-dimensional sobre un hiperplano k-dimensional que captura la mayor variabilidad posible.

Esta proyección minimiza la suma de los cuadrados de las distancias perpendiculares desde los puntos originales hasta el subespacio de menor dimensión, garantizando que se preserve la máxima cantidad de información posible en la representación de dimensión reducida.

3.7. Interpretación algebraica de las Componentes Principales

La fundamentación algebraica del PCA se basa en la descomposición espectral (resultado fundamental del álgebra lineal que establece que toda matriz simétrica real puede expresarse como el producto de tres matrices específicas) de la matriz de covarianzas en sus valores propios y vectores propios. Los vectores propios representan las direcciones de máxima varianza en el espacio de los datos, es decir, las nuevas dimensiones ortogonales a lo largo de las cuales la información se dispersa más. A su vez, los valores propios

cuantifican la magnitud de esa varianza explicada por cada vector propio correspondiente.

3.7.1. Planteamiento del problema de optimización

El PCA busca encontrar combinaciones lineales de las variables originales que maximicen la varianza. Formalmente, para un conjunto de variables $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ con matriz de covarianzas $\mathbf{S}$, se quiere encontrar un vector de coeficientes $\mathbf{a}_1$ tal que:

$$\max_{\mathbf{a}_1} \text{Var}(\mathbf{a}_1'\mathbf{y}) = \max_{\mathbf{a}_1} \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 \tag{3.7}$$

sujeto a la restricción de normalización $\mathbf{a}_1'\mathbf{a}_1 = 1$.

3.7.2. Valores propios y vectores propios: Claves en el PCA

Los valores propios y vectores propios constituyen los elementos fundamentales que hacen posible el análisis de componentes principales. Estos conceptos del álgebra lineal otorgan la solución computacional al problema de optimización y ofrecen una interpretación nutrida de la estructura de los datos en estudio.

Definición

Para una matriz cuadrada $\mathbf{A}$, un vector no nulo $\mathbf{v}$ es un **vector propio** con **valor propio** correspondiente λ si satisface:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{3.8}$$

De la anterior ecuación se obtiene que cuando la matriz $\mathbf{A}$ actúa sobre el vector propio $\mathbf{v}$, el resultado es simplemente el mismo vector escalado por el factor λ. El vector propio define una "dirección estable" bajo la transformación lineal representada por $\mathbf{A}$, mientras que el valor propio cuantifica el factor de escala a lo largo de esa dirección.

En el PCA, la matriz de interés es la matriz de covarianzas $\mathbf{C}$, que posee las siguientes propiedades:

- **Simetría**: $\mathbf{C} = \mathbf{C}^\top$, lo que garantiza valores propios reales.
- **Semidefinida positiva**: Todos los valores propios son no negativos. ($\lambda_i \geq 0$)
- **Ortogonalidad de vectores propios**: vectores propios correspondientes a valores propios distintos son ortogonales.

Descomposición Espectral

La matriz de covarianzas puede descomponerse completamente en términos de sus valores propios y vectores propios:

$$\mathbf{C} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top = \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i^\top \tag{3.9}$$

donde:

- $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$ es la matriz ortogonal de vectores propios.
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ es la matriz diagonal de valores propios.
- Los valores propios están ordenados: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$

Esta descomposición espectral es fundamental porque muestra cada valor propio obtenido representa la magnitud de la varianza explicada por su vector propio correspondiente. Esto significa que los primeros valores propios, que son los más grandes, identifican las direcciones (las componentes principales) a lo largo de las cuales los datos exhiben la mayor dispersión y, por ende, donde se concentra la mayor parte de la información relevante.

3.8. Algoritmo para el cálculo del Análisis de Componentes Principales (PCA)

Con la fundamentación matemática que se presentó de manera sucinta en las secciones anteriores se presenta a continuación el algoritmo básico para calcular las componentes principales de un conjunto de datos.

3.8.1. Estandarización de los datos

Si las variables tienen diferentes escalas, es recomendable estandarizarlas ya que sus varianzas pueden diferir por órdenes de magnitud. Cada variable se transforma como:

$$z_{ij} = \frac{y_{ij} - \overline{y}_j}{s_j} \tag{3.10}$$

donde y_{ij} es el valor de la variable j en la observación i, $\overline{y}_j$ es la media muestral de la variable j, y s_j es su desviación estándar muestral.

La media muestral de la variable j se calcula como:

$$\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \tag{3.11}$$

y la desviación estándar muestral como:

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2} \tag{3.12}$$

Después de la estandarización, cada variable transformada z_j tiene media cero ($\overline{z}_j = 0$) y desviación estándar unitaria ($s_{z_j} = 1$), lo que garantiza que todas las variables contribuyan equitativamente al análisis, independientemente de sus escalas originales de medición.

3.8.2. Cálculo de la matriz de covarianzas

La matriz de covarianzas es fundamental en el PCA ya que contiene toda la información sobre las relaciones lineales y la variabilidad entre las variables. Esta matriz muestra tanto las varianzas individuales de cada variable (diagonal principal) como las covarianzas entre pares de variables (elementos fuera de la diagonal).

A partir de la matriz de datos centrados $\mathbf{Y}_c$, donde cada fila representa una observación y cada columna una variable, se calcula la matriz de covarianzas muestral $\mathbf{S}$ como:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^\top = \frac{1}{n-1}\mathbf{Y}_c^\top \mathbf{Y}_c \tag{3.13}$$

donde:

- $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})^\top$ es el vector de la i-ésima observación
- $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_p)^\top$ es el vector de medias muestrales
- $\mathbf{Y}_c$ es la matriz de datos centrados de dimensión $n \times p$
- $n - 1$ es el divisor que proporciona un estimador insesgado de la covarianza poblacional

Los elementos individuales de la matriz de covarianzas se calculan como:
Elementos diagonales (varianzas):

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2 \tag{3.14}$$

Elementos fuera de la diagonal (covarianzas):

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)(y_{ik} - \overline{y}_k) \quad \text{para } j \neq k \tag{3.15}$$

La matriz de covarianzas resultante $\mathbf{S}$ es una matriz simétrica de dimensión $p \times p$ con las siguientes propiedades:

- **Simetría**: $s_{jk} = s_{kj}$ para todos los pares (j, k)
- **Semidefinida positiva**: Todos los valores propios son no negativos
- **Diagonal principal**: Contiene las varianzas de cada variable individual
- **Elementos fuera de la diagonal**: Miden la covariabilidad entre pares de variables

Si las variables han sido previamente estandarizadas, la matriz de covarianzas se convierte en la matriz de correlaciones $\mathbf{R}$, donde cada elemento r_{jk} representa el coeficiente de correlación de Pearson entre las variables j y k.

3.8.3. Descomposición espectral (valores y vectores propios):

Se resuelve el problema de valores propios y vectores propios:

$$\mathbf{Sa} = \lambda\mathbf{a} \tag{3.16}$$

donde λ es un valor propio (autovalor) y $\mathbf{a}$ el correspondiente vector propio (autovector). Los autovectores determinan las direcciones de las componentes principales.

3.8.4. Ordenamiento de componentes

Ordenar los pares $(\lambda_j, \mathbf{a}_j)$ de mayor a menor según el valor de λ_j, ya que los valores propios indican la cantidad de varianza explicada por cada componente.

Cálculo de puntuaciones (scores): Las nuevas variables transformadas (componentes principales) se calculan como:

$$z_{ik} = \mathbf{a}_k^T(\mathbf{y}_i - \overline{\mathbf{y}}) \tag{3.17}$$

o en forma matricial:

$$\mathbf{Z} = \mathbf{XA} \tag{3.18}$$

donde $\mathbf{X}$ es la matriz de datos estandarizados y $\mathbf{A}$ es la matriz de vectores propios seleccionados.

Tabla 3.2: Resumen del algoritmo para el cálculo de componentes principales

Paso	Descripción
1	**Estandarización (Opcional):** Transformar cada variable para tener media cero y desviación estándar uno, especialmente si las variables están en diferentes escalas.
2	**Cálculo de la matriz de covarianza (datos sin estandarización) o de la matriz de correlación (datos con estandarización):** Calcular la matriz de covarianza muestral a partir de los datos centrados o calcular la matriz de correlación con los datos estandarizados.
3	**Descomposición espectral:** Obtener los valores propios y vectores propios de la matriz de covarianza o de la matriz de correlación para identificar las direcciones principales de variabilidad.
4	**Ordenamiento:** Clasificar los vectores propios en orden descendente según sus valores propios asociados.
5	**Proyección:** Calcular las nuevas variables (componentes principales) proyectando los datos sobre los vectores propios seleccionados.

3.9. Determinación del número de Componentes Principales

Ejecutado el anterior algoritmo la decisión sobre cuántas componentes principales retener es uno de los aspectos más críticos y, a menudo, más subjetivos del PCA. No existe una regla única y universal que sea óptima en todas las situaciones, por lo que se recomienda utilizar múltiples criterios de forma complementaria (Jolliffe-Cadima, 2016).

3.9.1. Criterios Estadísticos

Criterio de Kaiser-Guttman

El criterio de Kaiser, también conocido como regla de valores propios mayores que uno, es uno de los más utilizados en la práctica (Kaiser, 1960). Este criterio establece que se deben retener únicamente las componentes cuyos valores propios sean mayores que 1.

La justificación teórica de este criterio se basa en que:

- En el PCA de correlaciones, cada variable estandarizada contribuye con una unidad de varianza
- Un valor propio mayor que 1 indica que la componente explica más varianza que una variable individual
- Componentes con valores propios menores que 1 explicarían menos varianza que una variable original

Matemáticamente, se retienen k componentes donde:

$$k = |\{\lambda_i : \lambda_i > 1\}| \tag{3.19}$$

Tabla 3.3: Ventajas y desventajas del criterio de Kaiser

Ventajas	**Desventajas**
Simplicidad de aplicación	Puede sobreestimar el número de componentes en muestras pequeñas
Interpretación intuitiva	Puede subestimar el número de componentes cuando hay muchas variables
Amplia aceptación en la literatura	No considera la estructura específica de los datos

Criterio de Proporción de Varianza Acumulada

Este criterio retiene el número de componentes necesarias para explicar un porcentaje predeterminado de la varianza total. Los umbrales comúnmente utilizados son:

- 70-80 % para análisis exploratorios
- 80-90 % para análisis más precisos
- 90-95 % para aplicaciones que requieren alta fidelidad

Se retienen k componentes tal que:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \alpha \tag{3.20}$$

donde α es el umbral de varianza deseado (por ejemplo, 0.80 para 80 %).

Gráfico de Sedimentación (Scree Plot)

Propuesto por Cattell, el gráfico de sedimentación representa los valores propios en orden decreciente como se muestra en la Figura 3.7. La metáfora proviene de la geología, donde "scree" se refiere a la acumulación de rocas pequeñas al pie de una montaña (Cattell, 1966).

Interpretación del gráfico de sedimentación:

- Se busca un "codo" o punto de inflexión en la curva. Criterio subjetivo donde se evidencie un cambio drástico de pendiente de inclinación.
- Las componentes antes del codo se consideran significativas.

- Las componentes después del codo representan "ruido" o variabilidad residual.

Criterios de identificación del codo:

- Punto donde la pendiente cambia drásticamente
- Transición de una curva pronunciada a una línea relativamente plana
- Inspección visual de la discontinuidad en la pendiente

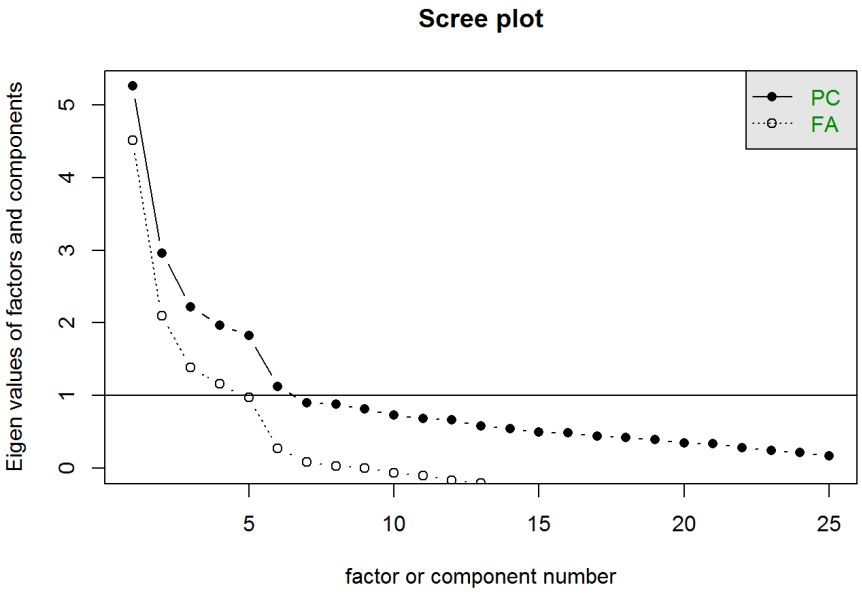


Figura 3.7: Scree Plot. (Bolaños, 2020)

3.10. Interpretación de las Componentes Principales

La interpretación de las componentes principales es un proceso que requiere rigor estadístico como conocimiento y dominio de la técnica. Las componentes principales, al ser combinaciones lineales de las variables originales, no siempre tienen una interpretación directa e intuitiva (Rencher-Christensen, 2012).

3.10.1. Coeficientes de las Componentes (*Loadings*)

Los coeficientes de las componentes principales, también llamados *"loadings"* o cargas, son los elementos de los vectores propios que definen cada componente. Para la k-ésima componente principal:

$$Z_k = a_{k1}Y_1 + a_{k2}Y_2 + \cdots + a_{kp}Y_p = \mathbf{a}_k'\mathbf{Y} \tag{3.21}$$

donde a_{kj} representa el coeficiente (*loading*) de la variable j en la componente k. Para una interpretación idónea se sugiere revisar la siguiente tabla.

Tabla 3.4: Interpretación de los coeficientes de las componentes principales

Interpretación
Magnitud de los coeficientes:
Coeficientes próximos a 0 indican que la variable tiene poca influencia en la componente.
Coeficientes grandes (en valor absoluto) implican una fuerte influencia de la variable.
Umbrales comunes para considerar un coeficiente como "grande" son 0.3, 0.4 o 0.5 (criterio subjetivo).
Signo de los coeficientes:
Coeficientes positivos indican una correlación directa con la componente.
Coeficientes negativos indican una correlación inversa con la componente.

3.11. Análisis de *clusters*

Como se ha puesto de manifiesto a lo largo de este capítulo las técnicas de estadística multivariada son muy diversas pero es de del intéres de este documento adentrarse en el análisis de *clusters*.

Desde los primeros trabajos en el área de la psicología por Tryon (1939), el análisis de *clusters* ha evolucionado hasta convertirse en una herramienta indispensable en la ciencia de datos.

El objetivo central del análisis de *clusters* es descubrir una serie de "estructuras" o "grupos" en un conjunto de datos, agrupando objetos de tal manera que aquellos dentro del mismo grupo compartan características similares, mientras que sean diferentes a los objetos en otros grupos. Esta premisa que se expone tiene una rigurosa fundamentación matemática que se explora a continuación.

A diferencia de las técnicas de clasificación supervisada, donde contamos con etiquetas predefinidas (estimación basada en un conjunto de entrenamiento), el análisis de *clusters*

opera en un entorno no supervisado, permitiendo que los datos revelen sus propias "estructuras internas" (Marden, 2015). Esta característica lo hace particularmente valioso en contextos exploratorios, ya que es propio de este estudio el conocer relaciones "inesperadas" entre las variables de nuestra base de datos.

Para comprender el análisis de *clusters*, debemos explicitar las definiciones, teoremas y métodos porpios de las matemáticas detrás del mismo. Esto permitirá un correcto análisis e interpretación de la técnica.

3.11.1. Espacio de Características y Métricas de Distancia

El *core* o núcleo del análisis de *clusters* se encuentra en el concepto de distancia entre objetos. Consideremos un conjunto de observaciones $X = \{x_1, x_2, ..., x_n\}$ en un espacio p-dimensional $\mathbb{R}^p$. Cada observación $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ representa un punto en este espacio multidimensional, donde cada dimensión corresponde a una característica medida.

La elección de la "medida" de distancia es fundamental, ya que define matemáticamente qué significa que dos objetos sean "similares". Las diferentes distancias entre objetos capturan distintas definiciones de similaridad o proximidad entre objetos:

1. Distancia Euclidiana:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

   Esta medida es la más intuitiva y ampliamente utilizada, representa la distancia en una "línea recta" existente entre dos puntos. Es particularmente efectiva cuando las variables están en la misma escala y son independientes.

2. Distancia de Manhattan:

$$d(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

   También conocida como distancia *city-block*, es más robusta ante valores atípicos y puede ser más apropiada en espacios donde el movimiento está restringido a direcciones específicas.

3. Distancia de Mahalanobis:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}$$

   donde Σ^{-1} es la matriz inversa de la matriz de covarianzas.

Esta sofisticada medida toma en cuenta la estructura de covarianza de los datos, siendo especialmente útil cuando las variables están correlacionadas.

Es preciso señalar que la elección de la medida de distancia debe basarse en una comprensión tanto de la naturaleza de los datos como del objetivo del análisis, ya quee esta decisión influirá significativamente en los resultados del *clustering*.

3.11.2. Métodos de Clustering

Con una mirada sucinta a los fundamentos matemáticos establecidos, ahora se dará una mirada a los principales métodos de *clustering*. Cada método representa un enfoque único para abordar el problema de la agrupación, con sus propias fortalezas y limitaciones.

K-means

El algoritmo *k-means*, introducido por MacQueen (1967), representa quizás el método más conocido y ampliamente utilizado en el análisis de *clusters*. Su popularidad se debe a su simplicidad conceptual y eficiencia algorítmica.

El objetivo fundamental del k-means es minimizar la varianza total dentro de los clusters:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - \mu_j\|^2$$

donde cada componente tiene el siguiente significado:

- J: Función objetivo a minimizar, que representa la suma total de las varianzas entre *clusters* (también conocida como *Within-Cluster Sum of Squares*, WCSS).
- k: Número de *clusters* predefinido por el analista.
- n: Número total de observaciones en el conjunto de datos.
- $x_i^{(j)}$: La i-ésima observación que ha sido asignada al *cluster* j. El superíndice (j) indica la pertenencia al *cluster* j. Es importante notar que no todas las observaciones del conjunto de datos tendrán este superíndice para un j específico, sino únicamente aquellas que han sido clasificadas dentro del *cluster* j.
- μ_j: Centroide del *cluster* j, calculado como el vector promedio de todas las observaciones asignadas al *cluster* j:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{3.22}$$

donde C_j representa el conjunto de observaciones pertenecientes al *cluster* j y $|C_j|$ es su cardinalidad.

- $\|x_i^{(j)} - \mu_j\|^2$: Distancia euclidiana al cuadrado entre la observación $x_i^{(j)}$ y el centroide μ_j de su *cluster* asignado. Esta medida cuantifica qué tan alejada está cada observación de su centro de cluster correspondiente.
- $\sum_{j=1}^{k}$: Primera sumatoria que itera sobre todos los k clusters definidos.
- $\sum_{i=1}^{n}$: Segunda sumatoria que, para cada cluster j, suma sobre todas las observaciones que han sido asignadas específicamente a ese cluster. Aunque el índice va de 1 a n, en la práctica solo se suman las observaciones que pertenecen al cluster j en cuestión.

La siguiente Tabla realiza un breve resumen del algoritmo:

Tabla 3.5: Fases del algoritmo *k-means*

Fase	**Descripción**
Fase de Inicialización	Selección de k centroides iniciales $\{\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}\}$ del conjunto de datos $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$. El método *k-means++* selecciona centroides con probabilidad proporcional a $D^2(x)$, donde $D(x)$ es la distancia al centroide más cercano ya elegido.
Fase de Asignación	Cada observación x_i se asigna al cluster $C_j^{(t)}$ mediante la regla: $C_j^{(t)} = \{x_i : \|x_i - \mu_j^{(t)}\|^2 \leq \|x_i - \mu_l^{(t)}\|^2, \forall l \in \{1, \ldots, k\}\}$, minimizando la distancia euclidiana al cuadrado.
Fase de Actualización	Los centroides se actualizan según: $\mu_j^{(t+1)} = \frac{1}{\lvert C_j^{(t)}\rvert} \sum_{x_i \in C_j^{(t)}} x_i$, calculando el centroide de masa de cada cluster.
Criterio de Convergencia	El algoritmo converge cuando se minimiza la función objetivo $J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$ o cuando $\|\mu^{(t+1)} - \mu^{(t)}\| < \epsilon$ para un umbral ϵ predefinido.

La importancia del k-means radica en su garantía de convergencia, aunque posiblemente a un mínimo local. Esta característica nos lleva naturalmente a considerar múltiples inicializaciones para encontrar una solución más robusta.

Clustering Jerárquico

El *clustering* jerárquico representa una aproximación radicalmente distinta al problema de agrupamiento de datos. Mientras que métodos como *k-means* buscan una agrupación

óptima, el *clustering* jerárquico construye una estructura completa de relaciones entre los datos, descubriendo agrupamientos a diferentes niveles de granularidad simultáneamente, es como tener un zoom a diferentes niveles.

Existen dos estrategias complementarias para construir esta jerarquía:

- **El enfoque aglomerativo**, sigue una filosofía constructiva, partiendo de lo particular hacia lo general. Inicialmente, cada observación constituye su propio cluster individual. El algoritmo procede iterativamente, identificando y fusionando en cada paso los dos clusters más similares según algún criterio de proximidad. Este proceso continúa hasta alcanzar un único cluster que engloba todos los datos, generando así una jerarquía completa de agrupamientos anidados.

- **El enfoque divisivo**, por contraste, adopta una perspectiva deconstructiva. Comienza considerando todos los datos como un único cluster global y procede a identificar divisiones óptimas que separan los datos en grupos cada vez más específicos. El proceso culmina cuando cada observación queda aislada en su propio cluster.

Distancias entre *clusters*

La elección del criterio de enlace (*linkage*) determina fundamentalmente la estructura y características de los clusters resultantes. Cada método captura una noción diferente de "proximidad entre grupos":

El *single linkage* adopta una visión optimista, definiendo la distancia entre clusters como la mínima distancia entre cualquier par de puntos:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

Esta aproximación tiende a formar clusters elongados y puede conectar regiones distantes a través de cadenas de puntos cercanos, siendo particularmente sensible a puntos con posiciones atípicas o *outliers*.

El *complete linkage* toma la perspectiva opuesta, considerando la máxima distancia entre pares de puntos:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

Este criterio conservador produce clusters compactos y bien separados, ofreciendo mayor robustez frente al ruido.

El *average linkage* busca un término medio, promediando todas las distancias entre

pares:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Finalmente, el método de *Ward* adopta un enfoque basado en la varianza, eligiendo en cada paso la fusión que minimiza el incremento en la suma de cuadrados intra-cluster, favoreciendo así la formación de clusters esféricos y homogéneos en tamaño.

3.11.3. Validación de *Clusters*

La validación de *clusters* representa uno de los desafíos fundamentales en estet tipo de análisis. ¿Cómo podemos evaluar la calidad de un análisis de *clusters*?

Métricas Internas

Las métricas internas evalúan la calidad del clustering utilizando solo la información intrínseca de los datos:

1. Índice de Silhouette (Rousseeuw, 1987):

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

Este elegante índice combina cohesión y separación:

- $a(i)$ mide la cohesión como la distancia media dentro del cluster
- $b(i)$ mide la separación como la distancia al cluster más cercano
- El rango [-1, 1] proporciona una interpretación intuitiva

Valores de referencia según la literatura:

- $s(i) > 0{,}71$: Estructura fuerte y bien definida
- $0{,}51 < s(i) \leq 0{,}70$: Estructura razonable
- $0{,}26 < s(i) \leq 0{,}50$: Estructura débil, podría ser artificial
- $s(i) \leq 0{,}25$: No se ha encontrado estructura significativa

El coeficiente de Silhouette promedio del clustering completo se considera aceptable cuando supera 0.5, aunque este umbral puede variar según el dominio de aplicación.

2. Índice Davies-Bouldin (Davies & Bouldin, 1979):

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \right\}$$

Este índice evalúa la relación entre la dispersión dentro de los clusters y la separación entre clusters:

- σ_i representa la dispersión dentro del cluster i
- $d(\mu_i, \mu_j)$ mide la separación entre clusters
- Valores más bajos indican mejor clustering

Interpretación según valores típicos:

- $DB < 0{,}5$: Clustering excelente con clusters bien separados y compactos
- $0{,}5 \leq DB < 1{,}0$: Clustering aceptable
- $DB \geq 1{,}0$: Clustering deficiente, los clusters presentan considerable solapamiento

Es importante notar que estos valores son orientativos y deben interpretarse en el contexto específico de los datos y el problema de análisis.

Determinación del Número Óptimo de Clusters

Un aspecto crucial del análisis de clusters es determinar el número apropiado de grupos:

1. Método del Codo (*Elbow Method*): Este método visual examina la variación de la suma de cuadrados dentro de los clusters (WSS):

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - \mu_j\|^2$$

El "codo" en la curva WSS vs. k sugiere un número óptimo de clusters.

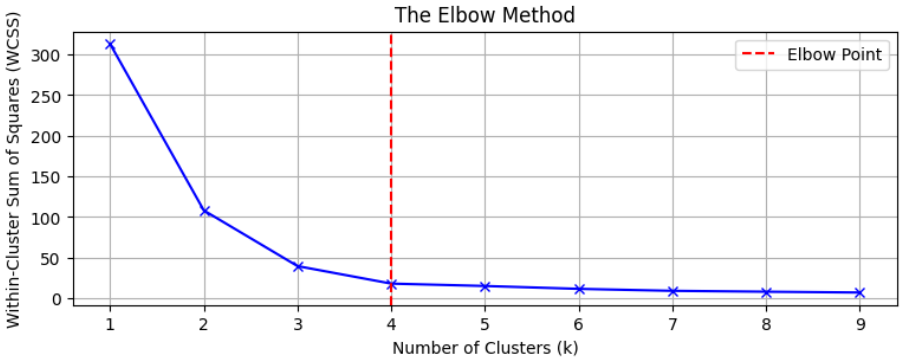


Figura 3.8: *Elbow Method.* (GeeksforGeeks, 2025)

2. **Gap Statistic:**

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k)$$

Este método más sofisticado compara el rendimiento observado con una distribución nula:

- W_k es la dispersión observada
- $E_n^*[\log(W_k)]$ es el valor esperado bajo la hipótesis nula
- El máximo del estadístico Gap sugiere el número óptimo de clusters

Capítulo 4

Aspectos metodológicos

Como se ha resaltado en los capítulos iniciales de este trabajo de grado la evaluación del rendimiento académico y los factores que influyen sobre él son elementos fundamentales para la estructuración de políticas educativas. En Colombia, las pruebas SABER 11 son uno de los instrumentos más importantes de medición estandarizada de la calidad educativa a nivel nacional y para miles de estudiantes representan el último baremo para su acceso a la educación superior. Este examen, impartido por el ICFES, evalúa las competencias en cinco áreas del saber (Lectura Crítica, Matemáticas, Ciencias Sociales y Ciudadanas, Ciencias Naturales e Inglés) de los estudiantes que están por finalizar el grado undécimo, último nivel de la educación media en Colombia.

El presente capítulo detalla la metodología empleada para analizar las relaciones entre múltiples variables que pueden incidir en los resultados de las pruebas SABER 11. Se describe el proceso de obtención y procesamiento de datos, el diseño del estudio, las variables consideradas y las técnicas estadísticas implementadas. La metodología propuesta busca aprovechar el potencial de las técnicas de la estadística multivariada para identificar patrones y relaciones que podrían escaparse con enfoques más tradicionales de análisis.

4.1. Obtención y caracterización de los datos

Los datos utilizados en este estudio provienen del portal "Datos Abiertos Colombia", una iniciativa gubernamental adscrita al Ministerio de Tecnologías de la Información y las Comunicaciones (MIN-TIC) que desde 2011 ha facilitado el acceso a información pública en formatos estandarizados e interoperables. Este portal representa un esfuerzo significativo hacia la transparencia y el acceso democrático a la información pública, permitiendo su uso y reutilización bajo licencias abiertas (MIN-TIC, 2025). La base de datos utilizada en este

estudio se titula "Resultados únicos Saber 11"[1], fue publicada por el ICFES el 7 de junio de 2023 y actualizada el 20 de abril del 2024. Este conjunto de datos comprende los resultados de las pruebas SABER 11 desde 2010 hasta 2022, conteniendo aproximadamente 7.11 millones de registros con 51 variables asociadas, ocupando un espacio de almacenamiento de 2.73 GB.

La elección de esta base de datos se sustenta en varias características claves que garantizan la confiabilidad y relevancia de la información. Su carácter oficial, al provenir directamente del ICFES, asegura la veracidad y precisión de los datos. La actualidad de la información, que incluye resultados recientes hasta 2022, permite un análisis contemporáneo de la situación educativa nacional. Adicionalmente, la amplia cobertura de la base de datos abarca múltiples aspectos del contexto educativo y socioeconómico de los estudiantes, proporcionando un marco para el análisis multivariado.

4.2. Diseño del estudio

4.2.1. Enfoque de la exploración de los datos

Este estudio emplea un enfoque cuantitativo con alcance exploratorio y descriptivo. Esta elección se fundamenta en el tipo de datos, los cuales provienen de mediciones estandarizadas, y en el objetivo principal de identificar y describir relaciones entre las variables que influyen en el rendimiento académico. Este diseño permite un análisis sistemático y estadísticamente riguroso de múltiples variables.

El rendimiento académico, al ser un fenómeno multifactorial, requiere un análisis que aborde diversas variables y sus interrelaciones. El enfoque cuantitativo provee las herramientas para examinar estas dimensiones de forma sistemática y "objetiva". Sin embargo, es importante reconocer que los cuestionarios socioeconómicos pueden introducir cierta subjetividad en los datos.

El carácter exploratorio de esta investigación se justifica por la necesidad de identificar patrones y relaciones en el contexto educativo colombiano, lo que puede enriquecer la comprensión del fenómeno. A su vez, el componente descriptivo permite caracterizar con precisión las relaciones halladas, sentando una base sólida para futuras investigaciones y la toma de decisiones en política educativa.

[1] Enlace a la base de datos estudiada.

4.2.2. Variables consideradas

Las variables seleccionadas para el estudio se pueden categorizar en tres grandes grupos, cuya selección se fundamenta en la literatura existente sobre factores que influyen en el rendimiento académico (Chica et al., 2012). Esta categorización permite un análisis estructurado que considera tanto los resultados académicos como los factores contextuales que pueden incidir en el desempeño académico.

Variables de Rendimiento Académico

Las variables de rendimiento académico constituyen el núcleo central del análisis, representando los diferentes aspectos de las competencias evaluadas en las pruebas SABER 11. Estas incluyen los puntajes obtenidos en cada una de las cinco áreas evaluadas: Inglés (PUNT_INGLES), Matemáticas (PUNT_MATEMATICAS), Ciencias Sociales y Ciudadanas (PUNT_SOCIALES_CIUDADANAS), Ciencias Naturales (PUNT_C_NATURALES) y Lectura Crítica (PUNT_LECTURA_CRITICA), cada uno medido en una escala de 0 a 100 puntos.

La variable PUNT_GLOBAL tiene un método de cálculo especial. El puntaje global se obtiene mediante un promedio ponderado que refleja la importancia relativa de cada área evaluada. Específicamente, los puntajes de Lectura Crítica, Matemáticas, Ciencias Sociales y Ciudadanas, y Ciencias Naturales se multiplican por una ponderación de 3, mientras que el puntaje de Inglés se multiplica por 1. La suma de estos productos se divide entre 13 (suma total de las ponderaciones: $4\times3 + 1\times1 = 13$) y finalmente se multiplica por 5, resultando en una escala de 0 a 500 puntos.

Variables Sociodemográficas

Las variables sociodemográficas proporcionan el contexto social y económico necesario para comprender los factores externos que pueden incidir en el desempeño estudiantil. El género del estudiante (ESTU_GENERO) permite analizar posibles diferencias en el rendimiento académico asociadas al género. El estrato socioeconómico de la vivienda (FAMI_ESTRATOVIVIENDA) constituye un indicador crucial del nivel socioeconómico familiar.

Las características del hogar se capturan mediante el número de personas que lo conforman (FAMI_PERSONASHOGAR) y el número de cuartos destinados para dormir (FAMI_CUARTOSHOGAR), variables que pueden reflejar tanto las condiciones de hacinamiento como la disponibilidad de espacios apropiados para el estudio. La variable de acceso a internet (FAMI_TIENEINTERNET) cobra particular relevancia, ya que las

tecnologías de la información y comunicación se han convertido en herramientas fundamentales para el aprendizaje.

Variables Institucionales

Las variables institucionales capturan características específicas del contexto educativo formal. El período de presentación del examen (PERIODO) permite controlar por efectos temporales y variaciones en las condiciones de aplicación de las pruebas. El indicador de bilingüismo institucional (COLE_BILINGUE) resulta especialmente relevante para el análisis del rendimiento en lenguas extranjeras y puede reflejar diferencias en la calidad y enfoque educativo de las instituciones.

Tabla 4.1: Variables de rendimiento académico

Variable	**Descripción**
PUNT_INGLES	Puntaje en inglés (0–100).
PUNT_MATEMATICAS	Puntaje en matemáticas (0–100).
PUNT_SOCIALES_CIUDADANAS	Puntaje en Ciencias Sociales y Ciudadanas (0–100).
PUNT_C_NATURALES	Puntaje en Ciencias Naturales (0–100).
PUNT_LECTURA_CRITICA	Puntaje en Lectura Crítica (0–100).
PUNT_GLOBAL	Promedio ponderado de los puntajes anteriores (0–500).

Tabla 4.2: Variables sociodemográficas

Variable	**Descripción**
ESTU_GENERO	Género del estudiante.
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de la vivienda.
FAMI_PERSONASHOGAR	Número de personas en el hogar.
FAMI_CUARTOSHOGAR	Número de cuartos destinados a dormir.
FAMI_TIENEINTERNET	Acceso a internet en el hogar.

Tabla 4.3: Variables institucionales

Variable	Descripción
PERIODO	Período de presentación del examen.
COLE_BILINGUE	Indicador de bilingüismo institucional.

4.2.3. Procedimiento de muestreo

Para el análisis se seleccionó una muestra de 500 registros mediante un muestreo aleatorio simple sin reemplazo, aplicando la corrección para poblaciones finitas. El tamaño de muestra se determinó utilizando la fórmula para estimación de proporciones:

Con un nivel de confianza del 95 % ($z_{0,025} = 1{,}96$) y un margen de error del 4.38 %, obtenemos:

$$n_0 = \frac{(1{,}96)^2 \times 0{,}5 \times 0{,}5}{(0{,}0438)^2} = \frac{3{,}8416 \times 0{,}25}{0{,}001918} = 500{,}73 \approx 501 \tag{4.1}$$

Aplicando la corrección por población finita:

$$n = \frac{501}{1 + \frac{500}{7{,}110{,}000}} = \frac{501}{1{,}00007} \approx 500 \tag{4.2}$$

Por lo tanto, una muestra de $n = 500$ estudiantes proporciona un margen de error del 4.38 % con 95 % de confianza.

Este método de muestreo se eligió por las siguientes razones: permite reducir el sesgo potencial en la selección de casos al garantizar que cada registro tenga la misma probabilidad de ser seleccionado y facilita el manejo computacional de los datos manteniendo la representatividad estadística necesaria para las técnicas de análisis multivariado.

4.3. Análisis estadístico

4.3.1. Software estadístico

Para el análisis de los datos se utilizó Google Colab el cual utiliza el lenguaje de programación Python, plataforma elegida principalmente por su acceso gratuito, implementación en la nube, facilidad para compartir y reproducir los análisis realizados. Además,

tiene disponible bibliotecas especializadas para ciencia de datos como scikit-learn, pandas y numpy.

4.3.2. Técnicas de Análisis Multivariado

El análisis de datos se realizó mediante dos técnicas estadísticas multivariadas: análisis de componentes principales (PCA) y análisis de *clusters*.

El PCA se aplicó para reducir la dimensionalidad de los datos e identificar las principales fuentes de variación en el rendimiento académico. Esta técnica resume"la información de múltiples variables correlacionadas en un conjunto reducido expresado en componentes principales, lo que facilita la interpretación y visualización de patrones y revela qué combinaciones de variables explican la mayor parte de la variabilidad total.

Complementariamente, el análisis de *clusters* se empleó para identificar patrones de agrupación entre los estudiantes, creando perfiles basados en su rendimiento y características. Específicamente, se utilizó el algoritmo K-means para formar grupos internamente homogéneos, pero heterogéneos entre sí, lo que permite caracterizar distintos perfiles estudiantiles.

La combinación de PCA y *clustering* ofrece un enfoque integral: mientras el PCA revela las dimensiones clave de variación, el *clustering* muestra cómo estas dimensiones se combinan en perfiles poblacionales, enriqueciendo la interpretación de los resultados y la comprensión del fenómeno educativo.

—

4.4. Limitaciones del estudio

Es importante reconocer las limitaciones inherentes a este estudio para hacer una correcta lectura de los hallazgos y el alcance. El posible sesgo temporal en los datos, derivado del período específico de recolección de la base de datos original de 2010-2022, ya que fue segmentado y se escogieron los períodos de 2020-2022 (debido a que es del interés de este estudio mirar el período durante y después de la pandemia). Esto podría afectar la generalización de los resultados, especialmente considerando los cambios significativos en el sistema educativo y las condiciones socioeconómicas durante este período.

Las limitaciones propias del método de muestreo empleado, aunque minimizadas por el diseño metodológico riguroso, deben ser consideradas en la interpretación de los resultados. El muestreo aleatorio simple, si bien reduce el sesgo de selección, tiene limitaciones en la captura de la diversidad regional y socioeconómica de la población estudiantil colombiana.

Adicionalmente, la existencia de factores no medidos que podrían influir en el rendimiento académico representa una limitación importante que debe ser reconocida en el análisis final. Por ejemplo, variables que miden la calidad específica de la enseñanza, el clima escolar, las metodologías pedagógicas implementadas, o factores familiares más sutiles no están directamente capturados en la base de datos utilizada, pero podrían tener efectos significativos en los resultados observados.

Por último, debe considerarse que las técnicas de análisis multivariado empleadas, aunque potentes y apropiadas para los objetivos planteados, imponen ciertas suposiciones sobre la naturaleza de los datos y las relaciones entre variables que podrían no cumplirse completamente en todos los casos, lo cual debe tenerse en cuenta al interpretar los resultados.

4.5. Depuración y alistamiento de la base de datos

Para la depuración y alistamiento de los datos se utilizó R como lenguaje principal de programación. A continuación se presenta el código utilizado para la depuración de los datos.

```
1
2  *** Install libraries
3
4  //***// install.packages("readr", "plyr", "dplyr")
5  readr --> Read Rectangular Text Data: provide a fast and friendly way to
       read rectangular data (like 'csv', 'tsv', and 'fwf').
6  plyr --> Tools for Splitting, Applying and Combining Data.
7  dplyr --> A Grammar of Data Manipulation: A fast, consistent tool for
      working with data frame like objects, both in memory and out of
      memory.
8
9  *** Call libraries
10 library(readr)
11 library(plyr)
12 library(dplyr)
13
14 *** Read a big dataset
15 train <-- read_csv("ds_saber_11_min.csv") --> 3GB
16
17 *** Show dataset representation (table)
18 view(train)
19
20 *** Filter
```

```
21
22 sample1 <- filter(train, PERIODO != '20101' & PERIODO != '20142')
23
24 *** Delete Columns
25
26 sample8 <- sample7 %>% select(-c(COLE_NOMBRE_SEDE, COLE_MCPIO_UBICACION,
       FAMI_TIENELAVADORA, FAMI_TIENECOMPUTADOR, ...))
27
28 *** Saving
29 write.csv(sample8, "sample54.csv")
30 sample <-- read_csv("sample54.csv")
31
32 *** Sample
33 set.seed(123)
34 randomsample <- sample_n(sample, 500)
35
36 # *** Create a new .csv file with the random sample
37 write_csv(randomsample_00, "tg_upn_500.csv")
```

Listing 4.1: Depuración de la base de datos en R

En resumen, con la base de datos original se eliminaron los registros diferentes a los periodos 2020-2022, a su vez las columnas correspondientes a las variables COLE NOMBRE SEDE, COLE MCPIO UBICACION, FAMI TIENELAVADORA, FAMI TIENECOMPUTADOR, y otras más pasando de 51 columnas a tan solo 14 por decisión conjunta entre director y tesista. Esta reducción obedeció a la necesidad de excluir variables categóricas nominales, variables booleanas (como FAMI_TIENELAVADORA, FAMI_TIENECOMPUTADOR), códigos identificadores geográficos y otras variables de naturaleza cualitativa que resultan incompatibles con los supuestos matemáticos del análisis de componentes principales y las métricas de distancia empleadas en el análisis de *clusters*, las cuales requieren variables cuantitativas continuas para el cálculo de matrices de covarianza y la aplicación de medidas de distancia.

Por último se realizó el muestreo con la selección aleatoria de 500 registros. Esta muestra puede ser consultada en el siguiente enlace.

Capítulo 5

Análisis y discusión de Resultados

El presente capítulo constituye el núcleo práctico de este estudio, donde se materializa la aplicación de las técnicas de estadística multivariada sobre la muestra de 500 estudiantes extraída de la base de datos de las pruebas Saber 11. La implementación se realizó mediante Google Colab por medio del lenguaje de programación Python, ello facilitó tanto la reproducibilidad del análisis como la generación de visualizaciones interactivas. Este entorno permitió la integración fluida de las librerías especializadas de Python, particularmente `pandas` para la manipulación de datos, `numpy` para operaciones matemáticas, `matplotlib` y `seaborn` para visualizaciones, y `scikit-learn` para la implementación de los algoritmos de análisis multivariado.

La estructura del análisis sigue una progresión lógica que permite comprender gradualmente la complejidad inherente a los datos educativos. En primer lugar, se presenta un análisis exploratorio del dataset que distingue entre variables categóricas y numéricas, estableciendo así una comprensión fundamental de la naturaleza y distribución de los datos. Para las variables numéricas, correspondientes a los puntajes en las diferentes áreas evaluadas (Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés y Puntaje Global), se calculan estadísticas descriptivas que incluyen medidas de tendencia central, dispersión y forma de las distribuciones.

Posteriormente, se implementa el Análisis de Componentes Principales (PCA) como técnica de reducción dimensional para la identificación de las relaciones entre las diferentes áreas evaluadas. Cada etapa del análisis está respaldada por visualizaciones 2D o 3D y el código fuente correspondiente, garantizando así la transparencia y reproducibilidad de los resultados obtenidos.

5.1. Análisis exploratorio de los datos

```
# Limpieza basica de datos: Eliminar columnas no utiles
if 'Unnamed: 0' in df.columns:
    df = df.drop('Unnamed: 0', axis=1)
if 'X' in df.columns:
    df = df.drop('X', axis=1)

# Agrupar columnas categoricas
columnas_categoricas = ['COLE_BILINGUE', 'ESTU_GENERO', '
    FAMI_CUARTOSHOGAR', 'FAMI_ESTRATOVIVIENDA', 'FAMI_PERSONASHOGAR', '
    FAMI_TIENEINTERNET']

# Agrupar columnas numericas
columnas_numericas = ['PUNT_INGLES', 'PUNT_MATEMATICAS', 'PUNT_SOCIALES'
    , 'PUNT_C_NATURALES', 'PUNT_LECTURA_CRITICA', 'PUNT_GLOBAL']

for col in columnas_categoricas:
    df[col] = df[col].astype('category')

# Verificar valores faltantes
print("Valores faltantes por columna:")
print(df.isnull().sum())

print("Tipos de datos por columna:")
print(df.dtypes)
print("\nInformaci n general del dataset:")
df.info()

df.head()
```

Listing 5.1: Alistamiento y visualización preliminar del *dataset*.

La Figura 5.1 es producto del código dispuesto en el Listing[1] 5.1 y muestra un resumen general de la estructura del *dataset* utilizado en el análisis, el cual está compuesto por 500 observaciones y 13 variables, de las cuales seis son categóricas y siete son numéricas enteras (tipo int64). Entre las variables categóricas se encuentran aspectos sociodemográficos e institucionales como el género del estudiante, el estrato socioeconómico, y la condición de bilingüismo del colegio. Por otro lado, las variables numéricas corresponden a puntajes obtenidos por los estudiantes en diferentes áreas del conocimiento, así como un puntaje

[1]El paquete listings es el de uso habitual para la adición de código fuente de múltiples lenguajes de programación en un documento LaTex.

global. Todas las variables cuentan con registros completos, es decir, no tienen datos faltantes. Para ilustración del lector se deja una visualización del dataset en la Figura 5.2

```
Información general del dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   PERIODO               500 non-null    int64
 1   COLE_BILINGUE         500 non-null    category
 2   ESTU_GENERO           500 non-null    category
 3   FAMI_CUARTOSHOGAR     500 non-null    category
 4   FAMI_ESTRATOVIVIENDA  500 non-null    category
 5   FAMI_PERSONASHOGAR    500 non-null    category
 6   FAMI_TIENEINTERNET    500 non-null    category
 7   PUNT_INGLES           500 non-null    int64
 8   PUNT_MATEMATICAS      500 non-null    int64
 9   PUNT_SOCIALES         500 non-null    int64
 10  PUNT_C_NATURALES      500 non-null    int64
 11  PUNT_LECTURA_CRITICA  500 non-null    int64
 12  PUNT_GLOBAL           500 non-null    int64
dtypes: category(6), int64(7)
```

Figura 5.1: Output 01.

index	PERIODO	COLE_BILINGUE	ESTU_GENERO	FAMI_CUARTOSHOGAR	FAMI_ESTRATOVIVIENDA	FAMI_PERSONASHOGAR	FAMI_TIENEINTERNET	PUNT_INGLES	PUNT_MATE
0	20201	N	F	Tres	Estrato 4	3 a 4	Si	71	
1	20201	N	F	Tres	Estrato 5	3 a 4	Si	77	
2	20201	N	F	Dos	Estrato 3	3 a 4	Si	61	
3	20201	N	M	Tres	Estrato 5	3 a 4	Si	60	
4	20201	N	M	Cinco	Estrato 3	5 a 6	Si	54	

Figura 5.2: *Dataset.*

5.1.1. Análisis descriptivo de las variables numéricas

El análisis estadístico de los puntajes obtenidos por los 500 estudiantes en las diferentes áreas evaluadas revela patrones importantes sobre el desempeño académico y la variabilidad entre competencias. Como se observa en la Figura 5.3, el puntaje global presenta una media de 252.12 puntos con una desviación estándar de 51.96, lo que indica una dispersión considerable en el rendimiento general de los estudiantes. Entre las áreas específicas, Lectura Crítica y Matemáticas muestran el mejor desempeño promedio (52.67 y 51.42 puntos respectivamente), seguida de cerca por Inglés y Ciencias Naturales (49.98 y 49.87 puntos), mientras que Sociales y Ciudadanas (47.90 puntos) presenta el desempeño promedio más bajo. Los valores mínimos revelan que todas las áreas presentan casos de desempeño muy bajo (24-29 puntos), mientras que los máximos alcanzan 100 puntos en dos áreas específicas, con excepción del puntaje global que llega hasta 377 puntos de 500 posibles. La distribución de los cuartiles indica que, para la mayoría de las áreas, existe una distribución relativamente simétrica, aunque con una ligera tendencia hacia valores superiores en

Lectura Crítica, donde el 75 % de los estudiantes supera los 61 puntos, evidenciando una concentración de estudiantes con mejor desempeño en esta competencia.

Lo anterior, es más evidente mediante la representación *boxplot* de la Figura 5.4 donde deliberadamente se excluye el puntaje global, ya que está en una escala de medición diferente.

	count	mean	std	min	25%	50%	75%	max
PUNT_INGLES	500.0	49.984	13.238514	29.0	39.00	49.0	58.00	100.0
PUNT_MATEMATICAS	500.0	51.418	12.175375	23.0	43.00	51.0	59.00	100.0
PUNT_SOCIALES	500.0	47.900	11.728304	24.0	39.00	47.0	57.25	77.0
PUNT_C_NATURALES	500.0	49.870	10.335121	24.0	42.00	49.0	58.00	75.0
PUNT_LECTURA_CRITICA	500.0	52.666	10.847570	28.0	44.00	53.0	61.25	79.0
PUNT_GLOBAL	500.0	252.118	51.958825	148.0	211.75	247.0	292.00	377.0

Figura 5.3: Estadísticas descriptivas para variables numéricas.

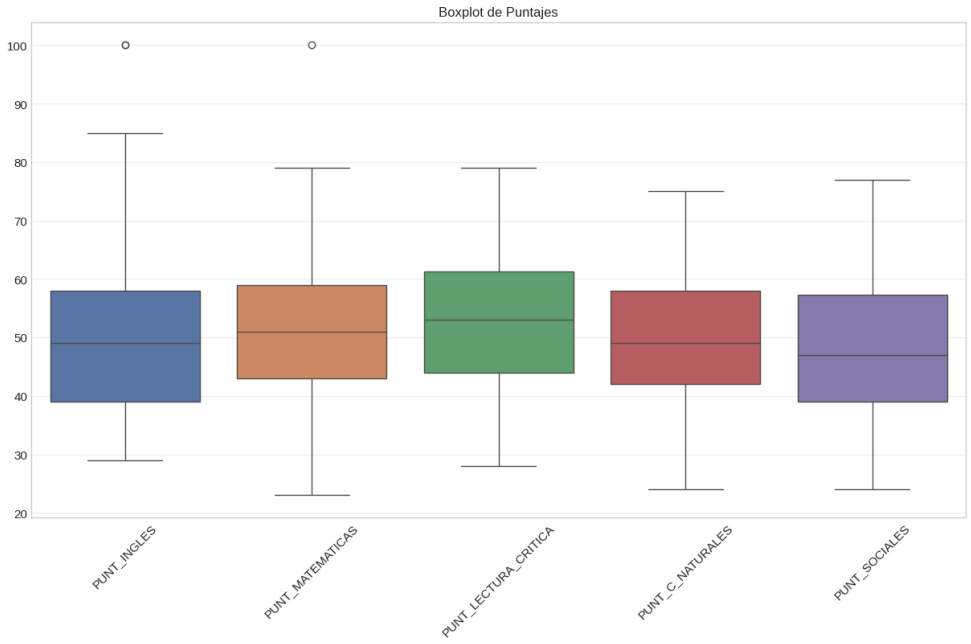


Figura 5.4: *Boxplot* puntajes por competencia.

```
#Visualizacion de distribuciones de puntajes por variables
import matplotlib.pyplot as plt

columnas_a_graficar = ['PUNT_INGLES', 'PUNT_MATEMATICAS','PUNT_SOCIALES'
    ,'PUNT_C_NATURALES','PUNT_LECTURA_CRITICA', 'PUNT_GLOBAL']

# Crear una figura grande para todos los subplots
fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(18, 12))
axes = axes.flatten()

# Iterar sobre las columnas y crear un histplot en cada subplot
for i, col in enumerate(columnas_a_graficar):
    sns.histplot(df[col], kde=True, ax=axes[i])
    axes[i].set_title(f'Distribuci n de {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Frecuencia')
    axes[i].axvline(df[col].mean(), color='r', linestyle='--', label=f'
    Media: {df[col].mean():.2f}')
    axes[i].axvline(df[col].median(), color='g', linestyle='-.', label=f
    'Mediana: {df[col].median():.2f}')
    axes[i].legend()


plt.tight_layout()
plt.show()
```

Listing 5.2: Distribuciones de puntajes por competencias.

Las anteriores líneas de código generan un histograma por variable númerica mediante el uso de la librería `matplotlib` y la función `sns.histplot()`, lo cual genera un acercamiento visual a la forma en la que los datos se distribuyen. A continuación, se visualizan las gráficas.

Al observar las representaciones de la Figura 5.5 se evidencia que las distribuciones de los puntajes académicos muestran comportamientos mayoritariamente "simétricos.º ligeramente sesgados hacia la derecha. En particular, las variables PUNT INGLES, PUNT C NATURALES, y PUNT SOCIALES muestran distribuciones con una ligera asimetría positiva. En contraste, PUNT LECTURA CRITICA y PUNT MATEMATICAS muestran un comportamiento más “equilibrado” donde sus medias y medianas son prácticamente iguales, lo que puede implicar una distribución normal. La variable PUNT GLOBAL, al ser un promedio ponderado de los puntajes, también mantiene la tendencia (cierta simetría), aunque presenta una mayor dispersión y cierta asimetría positiva.

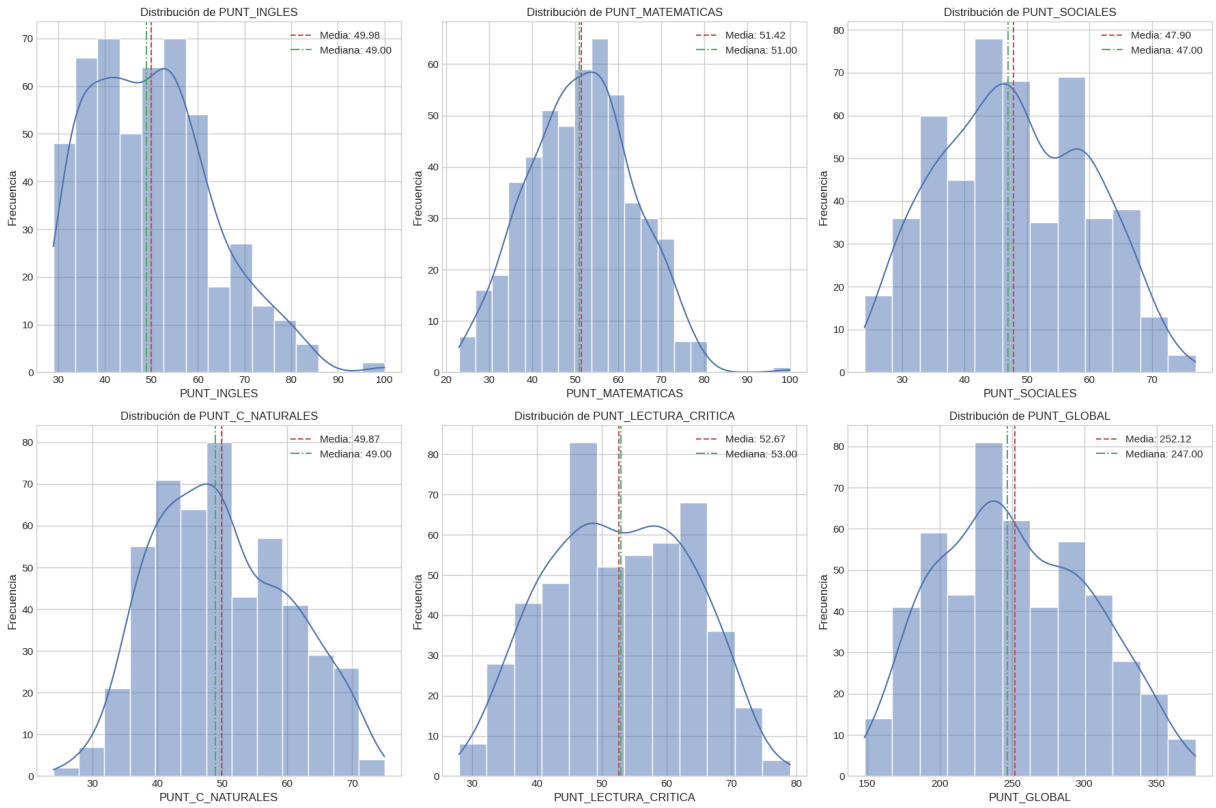


Figura 5.5: Distribución de los puntajes.

Adicionalmente, no se observan distribuciones severamente sesgadas ni presencia evidente de valores atípicos extremos, lo cual es favorable para la aplicación de técnicas estadísticas basadas en supuestos de normalidad, como el Análisis de Componentes Principales (PCA).

5.1.2. Análisis descriptivo de las variables categóricas

```
print("Numero de individuos por acceso a internet:")
print(df['FAMI_TIENEINTERNET'].value_counts())

print("Distribuci n por Acceso a Internet:")
print(df['FAMI_TIENEINTERNET'].value_counts(normalize=True).mul(100).
    round(2))
```

Listing 5.3: Estadísticas descriptivas de las variables categóricas.

Para esta sección se tomo la decisión de no desplegar visualizaciones (histogramas), ya que este tipo de gráficos no aportan información adicional a la que un resumen puede hacer.

```
Número de Colegios:
COLE_BILINGUE
N    489
S     11
Name: count, dtype: int64

Porcentaje de Colegios:
COLE_BILINGUE
N    97.8
S     2.2
Name: proportion, dtype: float64
```

```
Número de  Personas por género:
ESTU_GENERO
F    260
M    240
Name: count, dtype: int64
Distribución por Género:
ESTU_GENERO
F    52.0
M    48.0
Name: proportion, dtype: float64
```

```
Número de individuos por acceso a internet:
FAMI_TIENEINTERNET
Si    359
No    141
Name: count, dtype: int64
Distribución por Acceso a Internet:
FAMI_TIENEINTERNET
Si    71.8
No    28.2
Name: proportion, dtype: float64
```

```
Número de individuos por estrato:
FAMI_ESTRATOVIVIENDA
Estrato 2      175
Estrato 1      139
Estrato 3      113
Estrato 4       32
Sin Estrato     16
Estrato 5       15
Estrato 6       10
Name: count, dtype: int64
Distribución por Estrato Socioeconómico:
FAMI_ESTRATOVIVIENDA
Estrato 2      35.0
Estrato 1      27.8
Estrato 3      22.6
Estrato 4       6.4
Sin Estrato     3.2
Estrato 5       3.0
Estrato 6       2.0
Name: proportion, dtype: float64
```

```
Número de  cuartos por hogar:
FAMI_CUARTOSHOGAR
Tres          197
Dos           190
Cuatro         59
Uno            26
Cinco          15
Seis o mas     13
Name: count, dtype: int64
Distribución de Cuartos en el Hogar:
FAMI_CUARTOSHOGAR
Tres          39.4
Dos           38.0
Cuatro        11.8
Uno            5.2
Cinco          3.0
Seis o mas     2.6
Name: proportion, dtype: float64
```

```
Número de individuos por número de personas en el hogar:
FAMI_PERSONASHOGAR
3 a 4      257
5 a 6      153
1 a 2       43
7 a 8       33
9 o más     14
Name: count, dtype: int64
Distribución por Número de Personas en el Hogar:
FAMI_PERSONASHOGAR
3 a 4      51.4
5 a 6      30.6
1 a 2       8.6
7 a 8       6.6
9 o más     2.8
Name: proportion, dtype: float64
```

Figura 5.6: Número de individuos y distribución por cada variable categórica.

De este resumen estadístico se observa que un 97.8 % de los colegios a los que pertenecen los estudiantes no son bilingües. En cuanto a la distribución por género de los estudiantes, hay cierta paridad, con un 52 % de mujeres y un 48 % de hombres. La conexión a internet la tienen el 71.8 % de los estudiantes, un factor importante para el acceso a recursos educativos.

Si nos detenemos en el estrato socioeconómico, la distribución muestra una concentración significativa en los estratos más bajos: el Estrato 1 representa el 35 % de los estudiantes y el Estrato 2 el 27.8 %, ello implica un porcentaje superior al 60 % de la población estudiantil. Esta distribución sugiere un predominio de estudiantes en contextos socioeconómicos vulnerables. Respecto a las condiciones de vivienda, la mayoría de los hogares tienen entre dos y tres cuartos (38 % y 39.4 % respectivamente) y por último, el tamaño de los hogares es predominantemente mediano a grande, con la mayor proporción (51.4 %) de hogares conformados por 3 a 4 personas, seguido por un 30.6 % con 5 a 6 personas, esto puede sugerir hogares multigeneracionales o con una basta descendencia donde los estudiantes conviven.

5.2. Análisis multivariado de los datos

Hasta aquí se hizo un resumen de estadística descriptiva elemental mediado por el ambiente de desarrollo de Google Colab, pero la intención de este trabajo de grado es aplicar técnicas poco tradicionales en el devenir académico cotidiano, por ende a continuación

se expone el análisis de componentes principales y de *clusters* con el objetivo de recabar información valiosa entre las variables del dataset.

5.2.1. Análisis de Componentes Principales (PCA)

Recurriendo a la Tabla 3.2 se seguirán los pasos del algoritmo para el PCA. La selección de las variables es simple, se escogen aquellas que son númericas, es decir los puntajes por área a excepción del "PUNT GLOBAL", ya que es una ponderación de los demás y según el marco de referencia no aporta significativamente a nuestro estudio. Adicionalmente, se realiza la estandarización de los datos a pesar de tener la misma escala por sugerencia de los referentes bibliográficos citados en el marco teórico.

Estandarización

```
variables_pca = ['PUNT_INGLES', 'PUNT_MATEMATICAS', 'PUNT_SOCIALES', '
    PUNT_C_NATURALES', 'PUNT_LECTURA_CRITICA']

# Crear dataset para PCA
data_pca = df[variables_pca].copy()

#ESTANDARIZACION DE LOS DATOS
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_pca)

# Convertir de vuelta a DataFrame para facilitar manejo
data_scaled_df = pd.DataFrame(data_scaled, columns=variables_pca)

print("Datos estandarizados (media = 0, desviacion estandar = 1):")
print(data_scaled_df.describe())
```

Listing 5.4: Estandarización de los datos.

Cálculo de la matriz de correlación

```
#MATRIZ DE CORRELACION CON MAPA DE CALOR

cov_matrix = np.cov(data_scaled.T)
print("Matriz de covarianzas:")
print(cov_matrix)

# Crear mapa de calor
```

```
plt.figure(figsize=(10, 8))
mask = np.triu(np.ones_like(cov_matrix, dtype=bool))
sns.heatmap(cov_matrix, annot=True, cmap='coolwarm', center=0,
            xticklabels=variables_pca, yticklabels=variables_pca,
            mask=mask, fmt='.3f', square=True)
plt.title('Matriz de Covarianzas - Mapa de Calor')
plt.tight_layout()
plt.show()
```

Listing 5.5: Matriz de correlación.

Tabla 5.1: Matriz de correlación entre puntajes académicos.

	INGLÉS	MATEMÁTICAS	SOCIALES	C. NATURALES	LECTURA CRÍTICA
INGLÉS	1.00	0.68	0.68	0.70	0.69
MATEMÁTICAS	0.68	1.00	0.78	0.84	0.78
SOCIALES	0.68	0.78	1.00	0.83	0.82
C. NATURALES	0.70	0.84	0.83	1.00	0.80
LECTURA CRÍTICA	0.69	0.78	0.82	0.80	1.00

La matriz de correlación calculada por el software muestra relaciones lineales positivas moderadas a fuertes entre los distintos puntajes. Es de destacar que la mayor correlación se da entre Ciencias Naturales y Matemáticas (0.84), lo cual sugiere que los estudiantes que obtienen buenos resultados en una los obtendrán en la otra. Asimismo, Ciencias Naturales presenta correlaciones superiores a 0.69 con las demás materias, evidenciando un posible papel transversal en el desempeño académico. A continuación se dispone el mapa de calor de la matriz de correlación, el cual permite una mejor visualización.

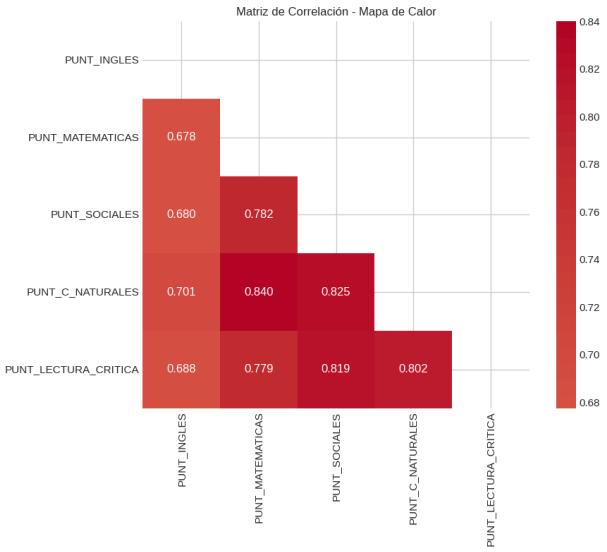


Figura 5.7: Mapa de calor de la matriz de correlación.

Descomposición Espectral y Ordenamiento

```
# Calcular valores propios (eigenvalues) y vectores propios (
    eigenvectors)
eigenvalues, eigenvectors = np.linalg.eig(cov_matrix)

print("Valores propios:")
for i, val in enumerate(eigenvalues):
    print(f"  {i+1}: {val:.4f}")

print("\nVectores propios (cada columna es un vector propio):")
print("Filas: variables | Columnas: componentes")
eigenvectors_df = pd.DataFrame(eigenvectors,
                               index=variables_pca,
                               columns=[f'PC{i+1}' for i in range(len(
    eigenvalues))])
print(eigenvectors_df)

#ORDENAR VECTORES DE MANERA DESCENDENTE SEGUN VALOR PROPIO

sorted_indices = np.argsort(eigenvalues)[::-1]

eigenvalues_sorted = eigenvalues[sorted_indices]
eigenvectors_sorted = eigenvectors[:, sorted_indices]

print("Valores propios ordenados (descendente):")
for i, val in enumerate(eigenvalues_sorted):
    print(f"PC{i+1}: {val:.4f}")
```

Listing 5.6: Cálculo de valores propios y vectores propios.

```
=== VALORES Y VECTORES PROPIOS ===
Valores propios:
λ1: 4.0453
λ2: 0.3830
λ3: 0.1499
λ4: 0.1870
λ5: 0.2448

Vectores propios (cada columna es un vector propio):
Filas: variables | Columnas: componentes
                           PC1       PC2       PC3       PC4       PC5
PUNT_INGLES          -0.411282 -0.910357 -0.034505 -0.027556 -0.012125
PUNT_MATEMATICAS     -0.452147  0.223128 -0.484731  0.274387 -0.659943
PUNT_SOCIALES        -0.455361  0.234533 -0.394420 -0.640368  0.414732
PUNT_C_NATURALES     -0.462351  0.191667  0.769671 -0.265621 -0.294191
PUNT_LECTURA_CRITICA -0.453088  0.172400  0.126038  0.665828  0.552971

=== ORDENAMIENTO POR VALORES PROPIOS ===
Valores propios ordenados (descendente):
PC1: 4.0453
PC2: 0.3830
PC3: 0.2448
PC4: 0.1870
PC5: 0.1499

Varianza explicada por cada componente:
PC1: 80.74% (Acumulado: 80.74%)
PC2: 7.64% (Acumulado: 88.39%)
PC3: 4.89% (Acumulado: 93.27%)
PC4: 3.73% (Acumulado: 97.01%)
PC5: 2.99% (Acumulado: 100.00%)
```

Figura 5.8: Resultados del cálculo de valores propios y vectores propios.

Con respecto a los valores propios obtenidos (Figura 5.8) es evidente como λ_1 tiene un valor muy alto (4.045) con respecto a los demás, por ende, es consecuente que el PC1 explique el 80.74 % de la varianza del conjunto de datos. Además, los valores de las componentes del vector propio asociado a PC1 son negativos (indica la dirección del vector) y muy similares entre sí, luego se puede concluir que PC1 es un indicador global de rendimiento (no se puede hacer una clara diferenciación entre que variable aporta más al componente).

Caso contrario ocurre con PC2, el cual explica el 7.64 % de la varianza. Al observar los coeficientes (*loadings*) asociados al vector propio la componente de inglés tiene un valor negativo alto diferenciándose radicalmente de los otros cuatro puntajes, los cuales son positivos. Ello podría interpretarse como un eje que contrasta las habilidades en inglés con respecto a las demás.

Para realizar este tipo de interpretaciones es ideal generar el siguiente tipo de mapa de calor de la Figura 5.9.

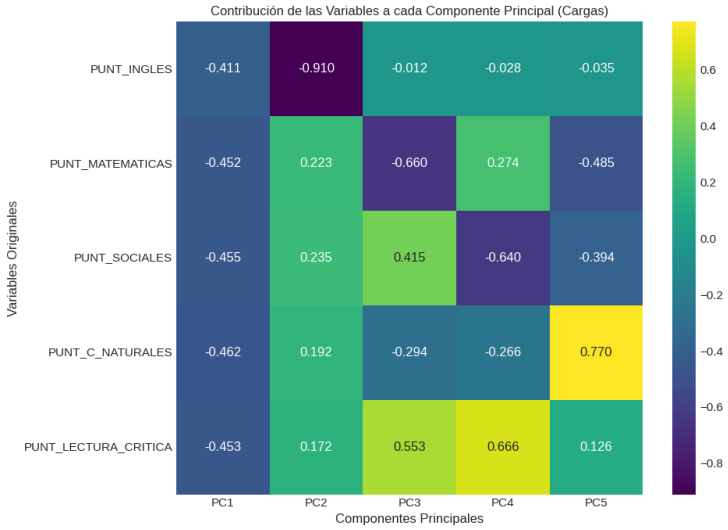


Figura 5.9: *Heat Map Loadings.*

Proyección

Dados los resultados de la sección anterior y recurriendo al criterio de proporción de varianza acumulada, donde para realizar análisis más precisos se debe retener entre el 80-90 % de la varianza total, el número de componentes principales seleccionados es 2, ya que explican el 88.39 % de la varianza total.

5.2.2. Transición del cálculo manual de PCA a la implementación con sklearn

La implementación "manual" del PCA mediante el cálculo directo de valores propios y vectores propios de la matriz de covarianza proporciona una comprensión profunda del proceso matemático en el cual se fundamenta el análisis. Por ello, se calculó explícitamente la matriz de covarianza, se obtuvó el conjunto de valores y vectores propios usando `np.linalg.eig()`, se ordenaron manualmente según su magnitud, y finalmente se obtuvieron las componentes principales. Sin embargo, esta aproximación presenta varias limitaciones prácticas (visualización) que `sklearn`[2] resuelve.

La librería `sklearn` implementa PCA utilizando la Descomposición en Valores Singulares (SVD) en lugar del método tradicional (valores propios y vectores propios), ofreciendo ventajas computacionales significativas, como el manejo eficiente de *datasets* donde el nú-

[2]Librería python de código abierto para aprendizaje automático con soporte para aprendizaje supervisado y no supervisado. scikit-learn

mero de variables excede el número de observaciones o individuos, situación problemática para el cálculo directo de la matriz de covarianza.

Asimismo, la implementación de `sklearn` calcula automáticamente los *loadings* (cargas) correctamente escalados como (components_* sqrt(explained_variance_))[3], mientras que en el método "manual" los vectores propios requieren escalamiento adicional para representaciones gráficas precisas como *biplots*.

Otra diferencia crucial es el manejo de los signos de los componentes. Los vectores propios pueden tener signos arbitrarios, lo que puede ocasionar inconsistencias entre diferentes ejecuciones o métodos. Además, `sklearn` ofrece funcionalidades adicionales integradas como la selección automática del número de componentes con base en la varianza explicada deseada, transformaciones inversas, y métodos especializados como PCA incremental para *datasets* que no caben en memoria.

```
#Verificacion usando sklearn
pca_sklearn = PCA()

#Ajustar el modelo PCA a los datos estandarizados
pca_sklearn.fit(data_scaled)

#Transformar los datos para obtener los componentes principales
principal_components_sklearn = pca_sklearn.transform(data_scaled)

#Crear DataFrame con los componentes principales de sklearn
pc_sklearn_df = pd.DataFrame(principal_components_sklearn,
                             columns=[f'PC{i+1}' for i in range(len(
    variables_pca))])

print("Varianza explicada (sklearn):")
for i, var in enumerate(pca_sklearn.explained_variance_ratio_ * 100):
    print(f"PC{i+1}: {var:.2f}%")

print("=== BIPLOT CON SKLEARN ===")
plt.figure(figsize=(12, 8))

# Scatter plot de las observaciones usando componentes de sklearn
plt.scatter(pc_sklearn_df['PC1'], pc_sklearn_df['PC2'], alpha=0.6, s=50)

# Vectores de las variables (loadings) de sklearn
loadings = pca_sklearn.components_.T * np.sqrt(pca_sklearn.
    explained_variance_)
scale_factor = 3
```

[3]Vectores propios normalizados multiplicados por la raíz cuadrada de la varianza explicada.

```
27
28 for i, var in enumerate(variables_pca):
29     plt.arrow(0, 0,
30               loadings[i, 0] * scale_factor,
31               loadings[i, 1] * scale_factor,
32               head_width=0.1, head_length=0.1, fc='red', ec='red')
33     plt.text(loadings[i, 0] * scale_factor * 1.1,
34               loadings[i, 1] * scale_factor * 1.1,
35               var, fontsize=10, ha='center', va='center')
36
37 plt.xlabel(f'PC1 ({pca_sklearn.explained_variance_ratio_[0]*100:.1f}% de
       varianza)')
38 plt.ylabel(f'PC2 ({pca_sklearn.explained_variance_ratio_[1]*100:.1f}% de
       varianza)')
39 plt.title('Biplot - Primeros Dos Componentes Principales (sklearn)')
40 plt.grid(True, alpha=0.3)
41 plt.axhline(y=0, color='k', linestyle='-', alpha=0.3)
42 plt.axvline(x=0, color='k', linestyle='-', alpha=0.3)
43 plt.tight_layout()
44 plt.show()
```

Listing 5.7: PCA implementando `sklearn`.

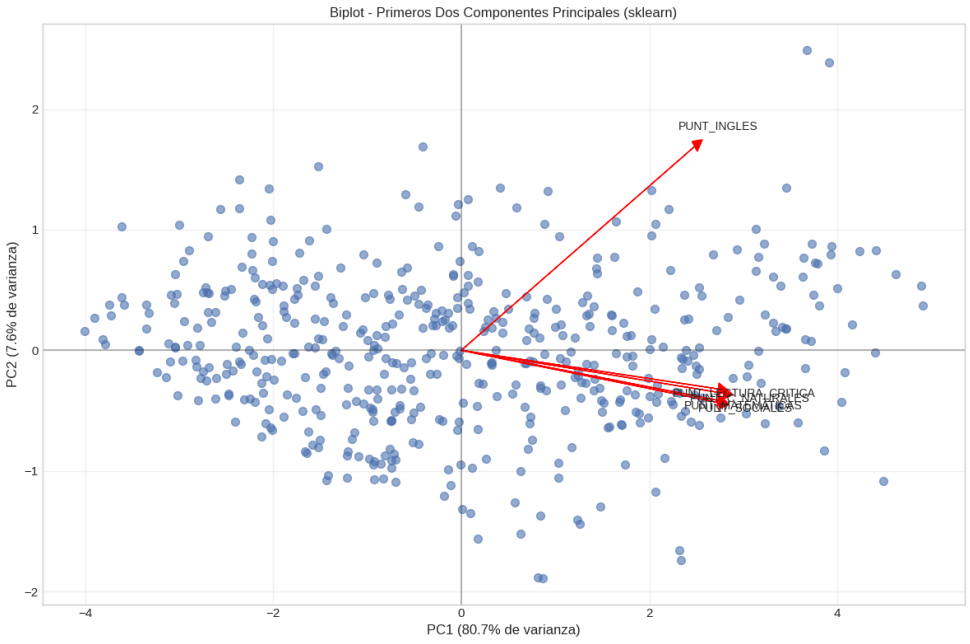


Figura 5.10: Biplot generado con la implementación de `sklearn`.

5.2.3. Interpretación del Análisis de Componentes Principales

El PCA y su visualización mediante el *biplot* revela que las cinco competencias evaluadas en las pruebas SABER 11 están altamente correlacionadas, generando una acumulación alrededor de un eje. El PC1 captura el 80.7 % de la varianza total, mientras que PC2 explica apenas el 7.6 %, implica que existe un factor que impacta en todas las competencias evaluadas. Podríamos llamar a este factor como "estudiante integral", es decir se podría hacer una analogía donde este estudiante es un "habilidoso" en todas las materias.

Lo anterior, lo confirma la dirección de los vectores, ya que todas las competencias apuntan hacia la derecha del gráfico, indicando que los estudiantes que obtienen buenos resultados en una materia tienen buen desempeño en las otras "estudiante habilidoso".

La dispersión de los puntos (estudiantes) forma una nube "aproximadamente simétrica" alrededor del origen, con mayor variabilidad a lo largo de PC1 que de PC2. Esta forma de distribución confirma que la principal fuente de variación entre estudiantes es su desempeño global en las 5 competencias evaluadas en la prueba.

Por último PC2 intenta hacer una diferenciación entre los estudiantes destacados en inglés de las demás competencias pero comparado con PC1 no muestra una relevancia alta.

5.2.4. PCA y variables categóricas

La visualización y análisis de los *biplots* integrando las variables categóricas (ver Figura 5.11) revela ciertos patrones diferenciados, ello permite la identificación de cuáles factores socioeconómicos o institucionales se asocian con lo capturado por PC1.

Inicialmente hay dos variables categóricas que no muestran una asociación clara al PC1 "estudiantes habilidosos" son ESTU_GENERO y COLE_BILINGUE. La primera muestra una distribución similar entre ambos géneros, no hay una acumulación de alguno de los dos grupos en alguna zona de la gráfica y la segunda es dominada por colegios NO bilingües pero los pocos colegios bilingües se distribuyen a lo largo del eje, luego no hay una distribución particular. Por otro lado, las variables categóricas con alguna evidencia de asociación con el rendimiento académico son:

- **FAMI_CUARTOSHOGAR:** Muestra una distribución clara a lo largo de PC1, donde los estudiantes más cuartos, (categorías 5 o más) se ubican hacia valores positivos de PC1 (mejor rendimiento), mientras que aquellos con menor número de cuartos se concentran en valores negativos.

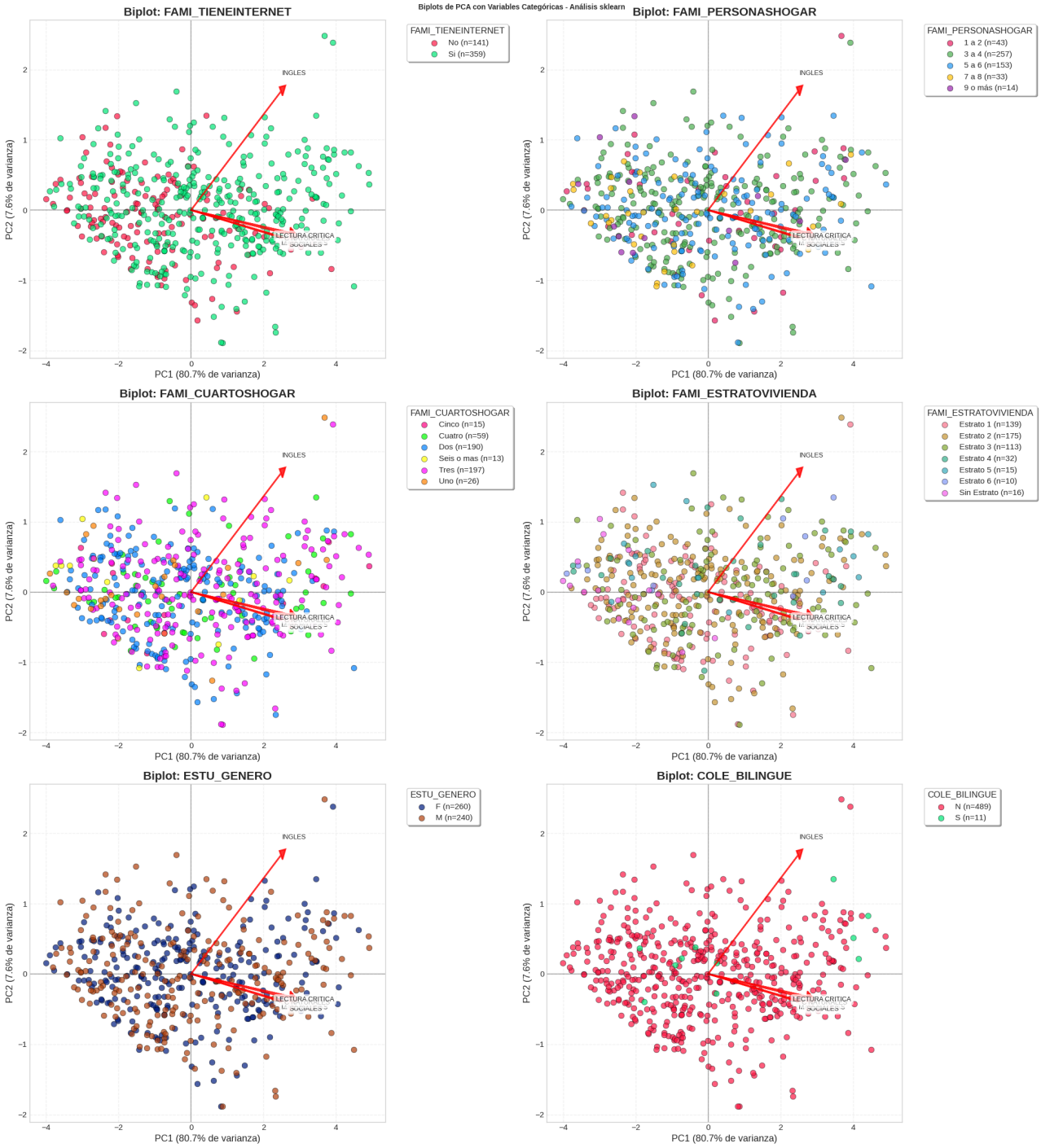


Figura 5.11: Biplot y variables categóricas.

- **FAMI_ESTRATOVIVIENDA:** Se observa como los estratos 5 y 6 están ubicados hacia la derecha (PC1 positivo), mientras que los estratos 1 y 2 se concentran mayoritariamente en el lado izquierdo, estableciendo una relación entre el estrato socioeconómico y el desempeño académico.
- **FAMI_PERSONASHOGAR:** Exhibe un patrón inverso al esperado, donde hogares con menos personas tienden a ubicarse valores positivos de PC1, sugiriendo que el tamaño familiar influye en los recursos disponibles por estudiante.
- **FAMI_TIENEINTERNET**: Marca una división binaria clara (tiene o no acceso a internet), donde los estudiantes con acceso a internet se ubican en valores positivos de PC1.

Las variables relacionadas con recursos disponibles en el hogar (espacio, estrato, internet) muestran las asociaciones más fuertes, mientras que características institucionales como el bilingüismo o la diferenciación por género del estudiante no evidencian efectos discriminantes significativos.

5.3. Análisis de *Clusters*

El procedimiento metodológico para el análisis de *clusters* se expondrá en cuatro apartados así:

- Clustering jerárquico ->Dendograma ->k óptimo
- Validación del número de *clusters*
- K-means con el número de *clusters* sugerido.
- Interpretación de los *clusters*

5.3.1. Clustering jerárquico - Dendograma

Para determinar el número óptimo de *clusters* en el espacio de componentes principales, se implementó un análisis de *clustering* jerárquico utilizando el método de Ward aplicado a todo el *dataset*. Se empleó el método de Ward, ya que en cada etapa, se unen los dos *clusters* para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del *cluster*. El proceso comienza con m *clusters*, cada uno de los cuales está compuesto por

un solo individuo, por lo que cada individuo coincide con el centro del *cluster*. (Gallardo, 2011)

```
#CLUSTERING JER RQUICO Y DENDROGRAMA
#Crear DataFrame con los PC1 y PC2
df_clustering = pd.DataFrame(principal_components_sklearn[:, :2],
                             columns=['PC1', 'PC2'])

print(f"Total de observaciones en el dataset: {len(df_clustering)}")

#Clustering jer rquico con m todo Ward --> (TODOS los datos)
Z = linkage(df_clustering, method='ward', metric='euclidean')

#Dendrograma
plt.figure(figsize=(15, 8))
dendrogram(Z,
           no_labels=True,
           color_threshold=0.7*max(Z[:,2]),
           above_threshold_color='gray')

plt.title('Dendrograma - Clustering Jer rquico (M todo Ward)',
          fontsize=14, fontweight='bold')
plt.xlabel(' ndice  de Observaci n', fontsize=10)
plt.ylabel('Distancia', fontsize=10)
plt.axhline(y=0.7*max(Z[:,2]), color='red', linestyle='--',
            label=f'Corte sugerido (altura={0.7*max(Z[:,2]):.2f})')
plt.legend(fontsize=12)
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```

Listing 5.8: *Clustering* Jerárqico y Dendograma.

El dendrograma resultante de la Figura 5.12 evidenció una estructura jerárquica clara con puntos de unión a diferentes alturas. Al analizar los cortes en distintos niveles arrojó los resultados de la Tabla 5.2

Con estos resultados se observa que para $k = 2$ se obtiene una división entre estudiantes de bajo y alto rendimiento (proporciones 25.2 % y 74.8 %), mientras que $k = 3$ genera una partición más equilibrada (proporciones de 25.2 %, 31.4 % y 43.4 %) y con ello se pueden generar tres *clusters* que impliquen un rendimiento académico bajo, medio y alto. En términos prácticos, facilitará la interpretación futura.

En conclusión, el análisis del dendrograma, complementado con el análisis cuantitativo de los tamaños del *cluster*, indica que $k = 3$ representa el punto óptimo donde se maximiza

Tabla 5.2: Distribución de Clústeres.

K (# de Clústeres)	Clúster	Tamaño (# individuos)	Proporción (%)
2	np.int32(1)	126	25.2 %
	np.int32(2)	374	74.8 %
3	np.int32(1)	126	25.2 %
	np.int32(2)	157	31.4 %
	np.int32(3)	217	43.4 %
4	np.int32(1)	126	25.2 %
	np.int32(2)	157	31.4 %
	np.int32(3)	98	19.6 %
	np.int32(4)	119	23.8 %
5	np.int32(1)	75	15.0 %
	np.int32(2)	51	10.2 %
	np.int32(3)	157	31.4 %
	np.int32(4)	98	19.6 %
	np.int32(5)	119	23.8 %

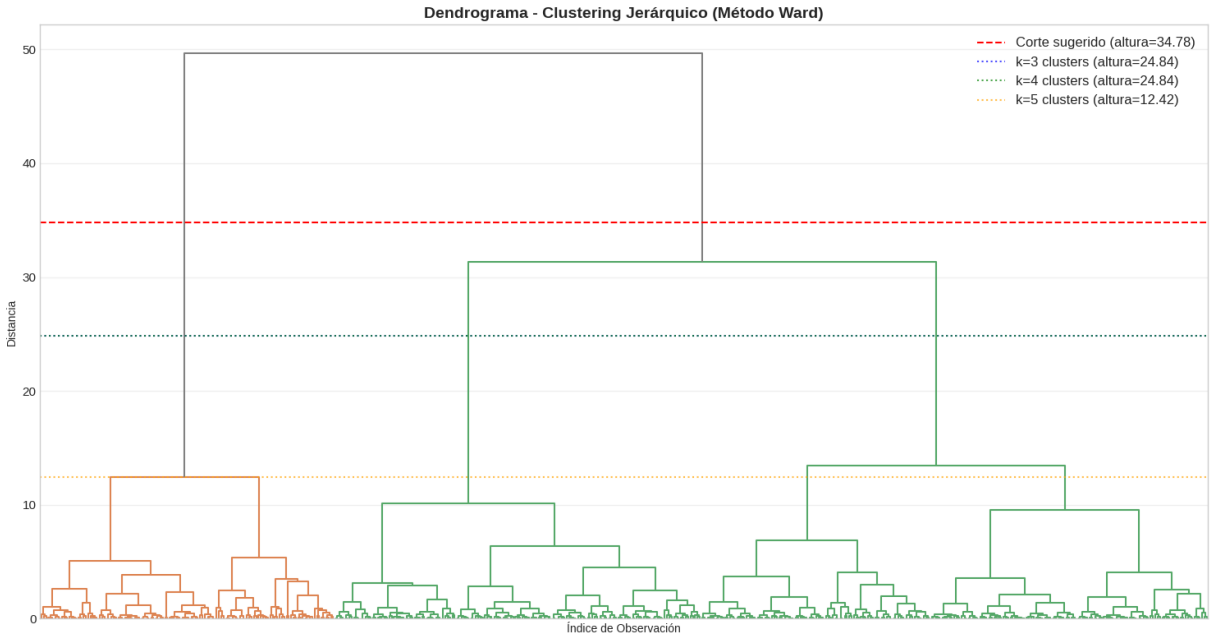


Figura 5.12: Dendograma

la homogeneidad de las proporciones de los grupos.

5.3.2. Validación del número de clústers

La selección del número óptimo de *clusters* (k) es un paso crítico en nuestro estudio, ya que esto influirá directamente en la interpretación y validez de los resultados. Para determinar la cantidad de *clusters* que mejor represente la estructura de los datos, se utilizan diversas herramientas de validación (algunas de ellas se revisaron en el marco teórico). En esta sección, se evalúa el número de *clusters* utilizando dos métodos: el Método del Codo (*Elbow Method*), que examina la inercia entre *clusters* para identificar el punto de inflexión donde la adición de más *clusters* no aporta una reducción significativa de la varianza; y el Análisis de Silhouette, que mide la cohesión de los objetos dentro de sus propios *clusters* y la separación de objetos de *clusters* vecinos, arroja un valor que indica qué tan bien se ha agrupado cada objeto. Por ende, hay una representación gráfica en la Figura 5.13 y las métricas asociadas a los métodos mencionados se presentan en la Tabla 5.3

```
#VALIDACION DEL NUMERO OPTIMO DE CLUSTERS

#Metodo del codo
inertias = []
silhouette_scores = []
K_range = range(2, 11)

for k in K_range:
   kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
   kmeans.fit(df_clustering)
   inertias.append(kmeans.inertia_)
   silhouette_scores.append(silhouette_score(df_clustering, kmeans.
   labels_))

print(inertias)
print(silhouette_scores)
```

Listing 5.9: inertias (Elbow Method) & silhoette scores.

Analizando los resultados de las métricas de validación, la elección de 3 clústeres ($k = 3$) se presenta como la opción viable para nuestro estudio. El Método del Codo muestra un evidente punto de inflexión en $k = 3$, donde la reducción en la inercia entre *clusters* comienza a disminuir de forma menos pronunciada (pasando de una disminución de 310.04 al pasar de $k = 2$ a $k = 3$, a 132.30 al pasar de $k = 3$ a $k = 4$).

Tabla 5.3: Resumen de Métricas de Validación para Diferentes Números de *Clusters* (k)

k	**Inercia**	**Silhouette**	**Δ Inercia**
2	784,38	0,5303	0,00
3	474,33	0,4492	310,04
4	342,03	0,4252	132,30
5	278,56	0,3913	63,47
6	236,52	0,3894	42,04
7	205,44	0,3699	31,08
8	179,49	0,3640	25,96
9	161,93	0,3714	17,55
10	148,14	0,3666	13,80

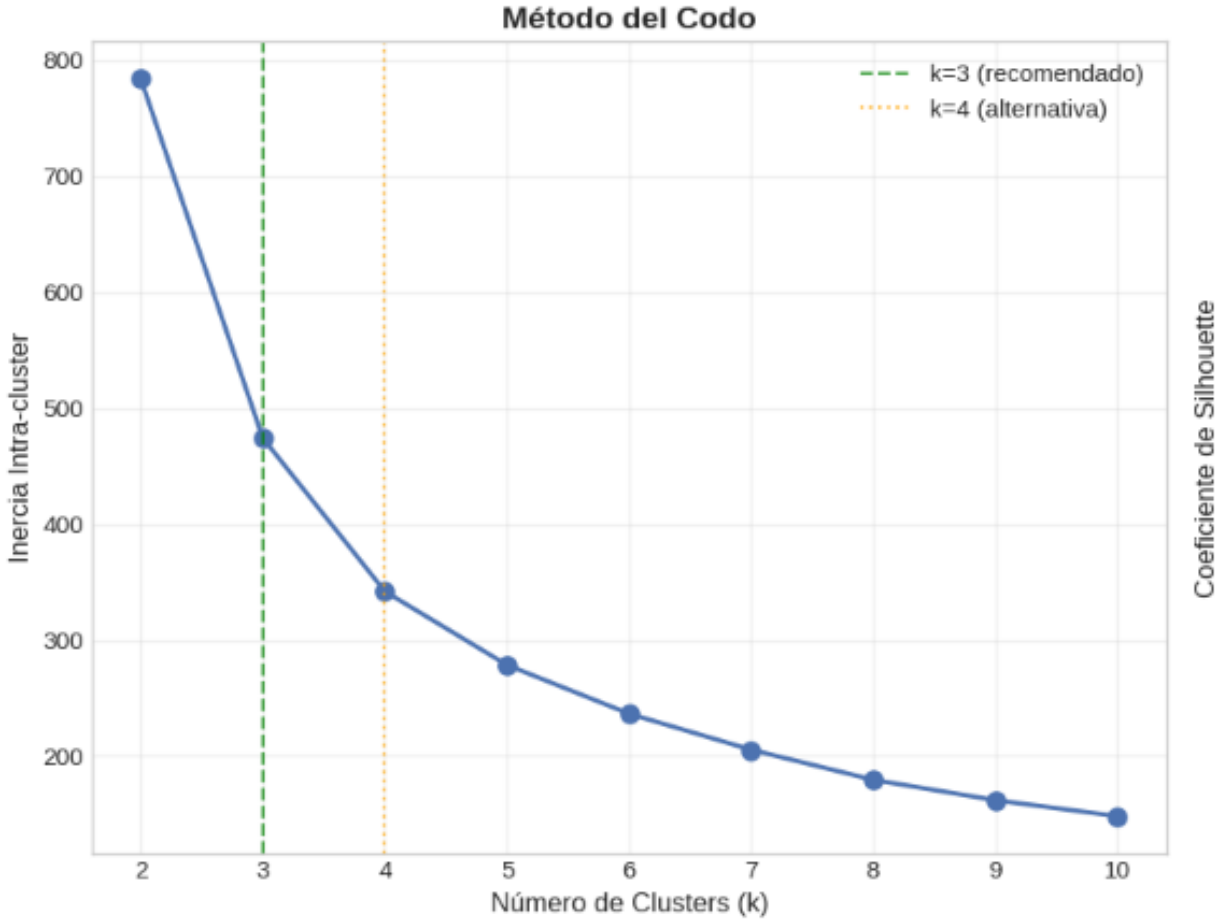


Figura 5.13: Visualización del Método del codo.

Aunque $k = 2$ presenta el mayor coeficiente de Silhouette (0.5303), el Análisis de Silhouette para $k = 3$ (0.4492) sigue siendo considerablemente alto, aunque no es el ideal para nuestro proceso de interpretación resulta evidentemente mejor.

5.3.3. *K-means* con el número de *clusters* sugerido

El algoritmo *K-means*[4] es uno de los métodos de agrupamiento más populares y ampliamente utilizados en el aprendizaje automático no supervisado. Su objetivo es dividir un conjunto de n observaciones en k *clusters*, donde cada observación pertenece al *cluster* cuyo centro (o çentroide") es el más cercano, basándose en una métrica de distancia, la utilizada en este estudio es la distancia euclidiana.

El algoritmo se inicia seleccionando aleatoriamente k centroides iniciales. Luego, se itera en dos pasos: primero, se asigna cada punto de datos al *cluster* del centroide más cercano; segundo, se recalculan las posiciones de los centroides tomando el promedio de todos los puntos asignados a cada *cluster*. Este proceso iterativo continúa hasta que los centroides no cambian significativamente entre iteraciones, o hasta que se alcanza un número máximo de iteraciones predefinido, minimizando así la suma de las distancias cuadradas entre cada punto y su centroide asignado. (Ramírez, 2024)

A continuación, se presentan los resultados de aplicar el algoritmo *K-means* con el número de clústeres establecido ($k = 3$). Tanto las estadísticas en las Tablas 5.4 y 5.5 como su representación gráfica en la Figura 5.14

```
#K-MEANS FINAL CON k=3

# Aplicar k-means con k=3
k_final = 3
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=20)
cluster_labels = kmeans_final.fit_predict(df_clustering)

df_clustering['Cluster'] = cluster_labels
df['Cluster'] = cluster_labels

# Visualizar los clusters
plt.figure(figsize=(12, 8))

# Colores distintivos para cada cluster (solo 3 ahora)
colors = ['#FF1744', '#FFC107', '#00E676']  # Rojo, Amarillo, Verde (
    Bajo, Medio, Alto)
cluster_names = ['Cluster 1', 'Cluster 2', 'Cluster 3']
```

[4] Documentación sobre el algoritmo.

```
17
18 for i in range(k_final):
19     cluster_data = df_clustering[df_clustering['Cluster'] == i]
20     plt.scatter(cluster_data['PC1'], cluster_data['PC2'],
21                 c=colors[i], label=f'{cluster_names[i]} (n={len(
    cluster_data)})',
22                 s=60, alpha=0.7, edgecolors='black', linewidth=0.5)
23
24 #Centroides
25 centroids = kmeans_final.cluster_centers_
26 plt.scatter(centroids[:, 0], centroids[:, 1],
27             c='black', s=400, alpha=1, edgecolors='white', linewidth=3,
28             marker='*', label='Centroides', zorder=5)
29
30 #Etiquetas a los centroides
31 for i, centroid in enumerate(centroids):
32     plt.annotate(f'C{i+1}', xy=(centroid[0], centroid[1]),
33                 xytext=(centroid[0]+0.1, centroid[1]+0.1),
34                 fontsize=12, fontweight='bold', color='black',
35                 bbox=dict(boxstyle='round,pad=0.3', facecolor='yellow',
    alpha=0.7))
36
37 # A adir vectores de loadings
38 loadings = pca_sklearn.components_.T * np.sqrt(pca_sklearn.
    explained_variance_)
39 scale_factor = 3
40
41 for i, var in enumerate(variables_pca):
42     plt.arrow(0, 0,
43               loadings[i, 0] * scale_factor,
44               loadings[i, 1] * scale_factor,
45               head_width=0.15, head_length=0.15,
46               fc='darkblue', ec='darkblue', alpha=0.8, linewidth=2)
47     plt.text(loadings[i, 0] * scale_factor * 1.15,
48              loadings[i, 1] * scale_factor * 1.15,
49              var.replace('PUNT_', '').replace('_', ' '),
50              fontsize=10, ha='center', va='center',
51              bbox=dict(boxstyle='round,pad=0.3', facecolor='white', alpha
    =0.8))
52
53 plt.xlabel(f'PC1 ({pca_sklearn.explained_variance_ratio_[0]*100:.1f}% de
     varianza)', fontsize=13)
54 plt.ylabel(f'PC2 ({pca_sklearn.explained_variance_ratio_[1]*100:.1f}% de
     varianza)', fontsize=13)
```

```
plt.title('Clustering K-means (k=3) en Espacio de Componentes Principales', fontsize=16, fontweight='bold')
plt.legend(loc='best', frameon=True, fancybox=True, shadow=True, fontsize=11)
plt.grid(True, alpha=0.3)
plt.axhline(y=0, color='k', linestyle='-', alpha=0.3)
plt.axvline(x=0, color='k', linestyle='-', alpha=0.3)

#Elipses distintivas
from matplotlib.patches import Ellipse
for i in range(k_final):
    cluster_data = df_clustering[df_clustering['Cluster'] == i]
    cov = np.cov(cluster_data[['PC1', 'PC2']].T)
    eigenvalues, eigenvectors = np.linalg.eig(cov)
    angle = np.degrees(np.arctan2(eigenvectors[1, 0], eigenvectors[0, 0]))
    width, height = 2 * np.sqrt(eigenvalues) * 2  # 95% confidence
    ellipse = Ellipse(centroids[i], width, height, angle=angle,
                      facecolor=colors[i], alpha=0.1, edgecolor=colors[i], linewidth=2)
    plt.gca().add_patch(ellipse)

plt.tight_layout()
plt.show()
```

Listing 5.10: Algoritmo *K-means* y graficación en el espacio PCA.

Tabla 5.4: Estadísticas Descriptivas de los Clústeres (k=3) - Parte 1: Tamaño y Centroides

Clúster	**Tamaño**	**(%)**	**Centroide**	
			PC1	PC2
3 - Rendimiento BAJO	152	30,4	−2,290	0,131
1 - Rendimiento MEDIO	201	40,2	−0,107	−0,142
2 - Rendimiento ALTO	147	29,4	2,514	0,059

Tabla 5.5: Estadísticas Descriptivas de los Clústeres (k=3) - Parte 2: Dispersión y Puntaje Promedio

Clúster	**Dispersión PC1**		**Dispersión PC2**		**Puntaje**
	μ	σ	μ	σ	**Promedio Global**
3 - Rendimiento BAJO	−2,290	0,667	0,131	0,510	38,5
1 - Rendimiento MEDIO	−0,107	0,693	−0,142	0,656	49,8
2 - Rendimiento ALTO	2,514	0,934	0,059	0,633	63,4

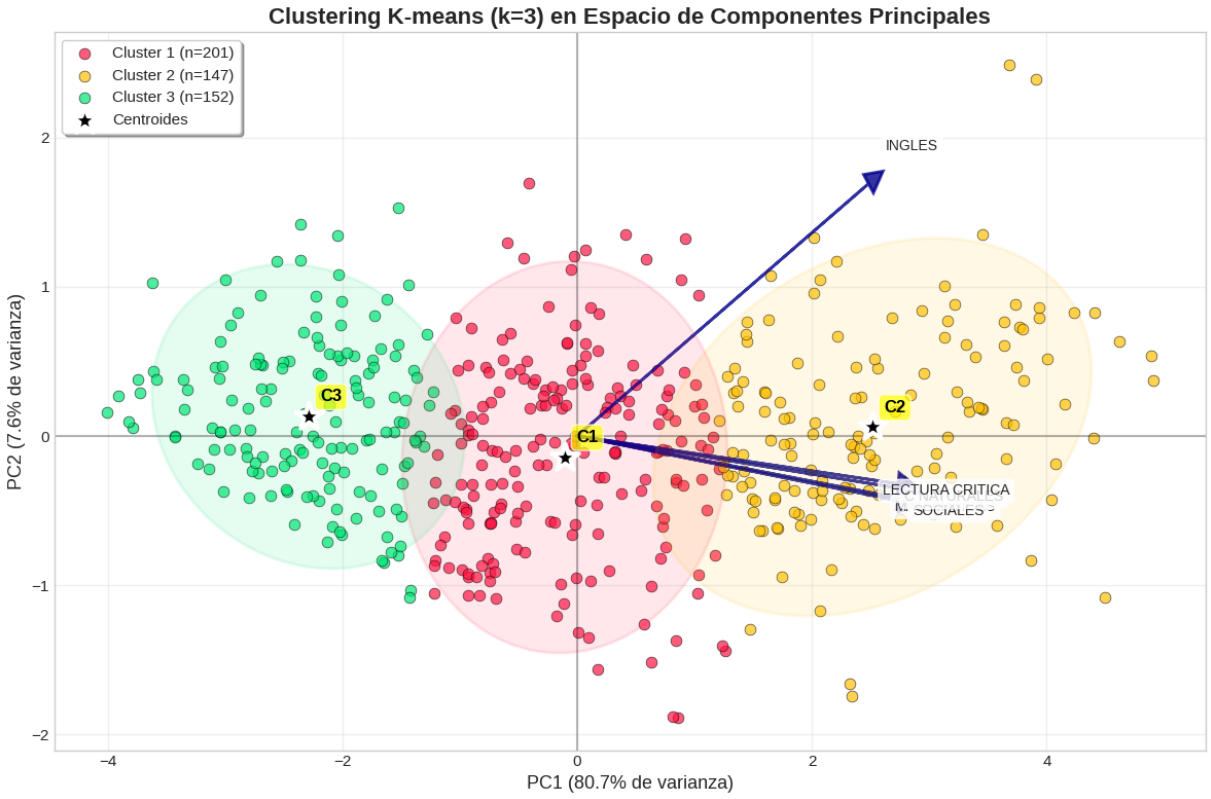


Figura 5.14: *Clustering K-means*

5.3.4. Interpretación de los *clusters*

La aplicación del algoritmo *K-means* con $k = 3$ *clusters* de la sección anterior ha permitido identificar una clara diferenciación de los individuos, fuertemente correlacionada con su nivel de rendimiento académico. A continuación, se realiza una caracterización de cada *cluster*.

Caracterización de *Clusters*.

- ***Cluster* 3 - Rendimiento BAJO:** Este es el grupo más pequeño, con 152 estudiantes (30.4 %).

  Centroide: Ubicado en PC1= −2,290 y PC2= 0,131. Su valor negativo en PC1 indica una baja contribución en la primera componente principal, y esto se asocia con bajo puntaje académico.

  Dispersión: Presenta una dispersión moderada en ambas componentes (PC1 σ = 0,667, PC2 $\sigma = 0{,}510$).

  Puntaje Promedio Global: Con 38.5, este clúster tiene el puntaje promedio más bajo.".

- ***Cluster* 1 - Rendimiento MEDIO:** Es el clúster más grande, con 201 estudiantes (40.2 %).

  Centroide: Localizado cerca del origen en PC1= −0,107 y $PC2 = -0{,}142$. Esto sugiere una contribución "neutralçon respecto a ambas componentes principales.

  Dispersión: Las dispersiones (PC1 σ = 0,693, PC2 σ = 0,656) son ligeramente mayores que las del *cluster* 3, por ende tiene una variabilidad similar.

  Puntaje Promedio Global: Su puntaje de 49.8 es intermedio entre los otros dos *clusters*.

- ***Cluster* 2 - Rendimiento ALTO:** Con 147 estudiantes (29.4 %), es un grupo de tamaño similar al Clúster 3.

  Centroide: Su centroide está en PC1= 2,514 y PC2= 0,059. El valor alto y positivo en PC1 indica que estos individuos tienen una fuerte contribución positiva a la primera componente principal, lo que se implica un alto rendimiento.

  Dispersión: Presenta la mayor dispersión en PC1 ($\sigma = 0{,}934$), lo que sugiere una mayor heterogeneidad en esta dimensión, mientras que en PC2 ($\sigma = 0{,}633$) es similar a los otros.

Puntaje Promedio Global: Con un puntaje significativamente alto de 63.4.

5.3.5. *Clusters* y variables categóricas

Al igual que en el PCA relacionamos las variables categóricas con los *clusters* para nutrir nuestra interpretación mas allá del rendimiento académico. La Figura 5.15 resume las proporciones de individuos por *cluster*, por ejemplo de los individuos del *cluster* 1 el 28.4 % de ellos no tiene acceso a internet y el 71.6 % si lo tiene.

```
Cluster 3 (BAJO):
FAMI_TIENEINTERNET
No    44.7
Si    55.3
----------------------
Cluster 3 (BAJO):
FAMI_PERSONASHOGAR
1 a 2       7.9
3 a 4      48.0
5 a 6      27.0
7 a 8      11.2
9 o más     5.9
----------------------
Cluster 3 (BAJO):
FAMI_CUARTOSHOGAR
Cinco          3.9
Cuatro        10.5
Dos           40.8
Seis o mas     5.3
Tres          30.9
Uno            8.6
----------------------
Cluster 3 (BAJO):
FAMI_ESTRATOVIVIENDA
Estrato 1      28.3
Estrato 2      36.2
Estrato 3      14.5
Estrato 4       3.9
Estrato 5       4.6
Estrato 6       3.3
Sin Estrato     9.2
----------------------
Cluster 3 (BAJO):
ESTU_GENERO
F    49.3
M    50.7
----------------------
Cluster 3 (BAJO):
COLE_BILINGUE
N    98.7
S     1.3
```

```
Cluster 1 (MEDIO):
FAMI_TIENEINTERNET
No    28.4
Si    71.6
----------------------
Cluster 1 (MEDIO):
FAMI_PERSONASHOGAR
1 a 2       8.5
3 a 4      51.7
5 a 6      34.3
7 a 8       4.5
9 o más     1.0
----------------------
Cluster 1 (MEDIO):
FAMI_CUARTOSHOGAR
Cinco          1.5
Cuatro        12.4
Dos           38.8
Seis o mas     1.0
Tres          41.3
Uno            5.0
----------------------
Cluster 1 (MEDIO):
FAMI_ESTRATOVIVIENDA
Estrato 1      31.3
Estrato 2      37.8
Estrato 3      21.9
Estrato 4       6.5
Estrato 5       1.5
Estrato 6       0.0
Sin Estrato     1.0
----------------------
Cluster 1 (MEDIO):
ESTU_GENERO
F    54.7
M    45.3
----------------------
Cluster 1 (MEDIO):
COLE_BILINGUE
N    97.5
S     2.5
```

```
Cluster 2 (ALTO):
FAMI_TIENEINTERNET
No    10.9
Si    89.1
----------------------
Cluster 2 (ALTO):
FAMI_PERSONASHOGAR
1 a 2       9.5
3 a 4      54.4
5 a 6      29.3
7 a 8       4.8
9 o más     2.0
----------------------
Cluster 2 (ALTO):
FAMI_CUARTOSHOGAR
Cinco          4.1
Cuatro        12.2
Dos           34.0
Seis o mas     2.0
Tres          45.6
Uno            2.0
----------------------
Cluster 2 (ALTO):
FAMI_ESTRATOVIVIENDA
Estrato 1      22.4
Estrato 2      29.9
Estrato 3      32.0
Estrato 4       8.8
Estrato 5       3.4
Estrato 6       3.4
Sin Estrato     0.0
----------------------
Cluster 2 (ALTO):
ESTU_GENERO
F    51.0
M    49.0
----------------------
Cluster 2 (ALTO):
COLE_BILINGUE
N    97.3
S     2.7
```

Figura 5.15: *Clusters* y las proporciones por variable categórica.

Entrando en el análisis las variables referentes al género y la distinción de bilingüismo del colegio no revelan mayor información. Las variables FAMI_PERSONASHOGAR y

FAMI_CUARTOSHOGAR en el PCA tenían cierta interpretación (tenian una distribución hacia a un lado u otro del PC1) pero con los *clusters* no se evidencia mayor aporte a alguno de los tres grupos, ya que tienen porcentajes similares, por ejemplo en la variable FAMI_PERSONASHOGAR se puede concluir que en los tres *clusters* las personas por hogar están entre 3 a 6 con una carga porcentual mayor al 75 %. Al igual en la variable FAMI_CUARTOSHOGAR los cuartos por hogar están entre 2 a 4 con una carga porcentual mayor al 80 %, es decir no hay una diferenciación clara.

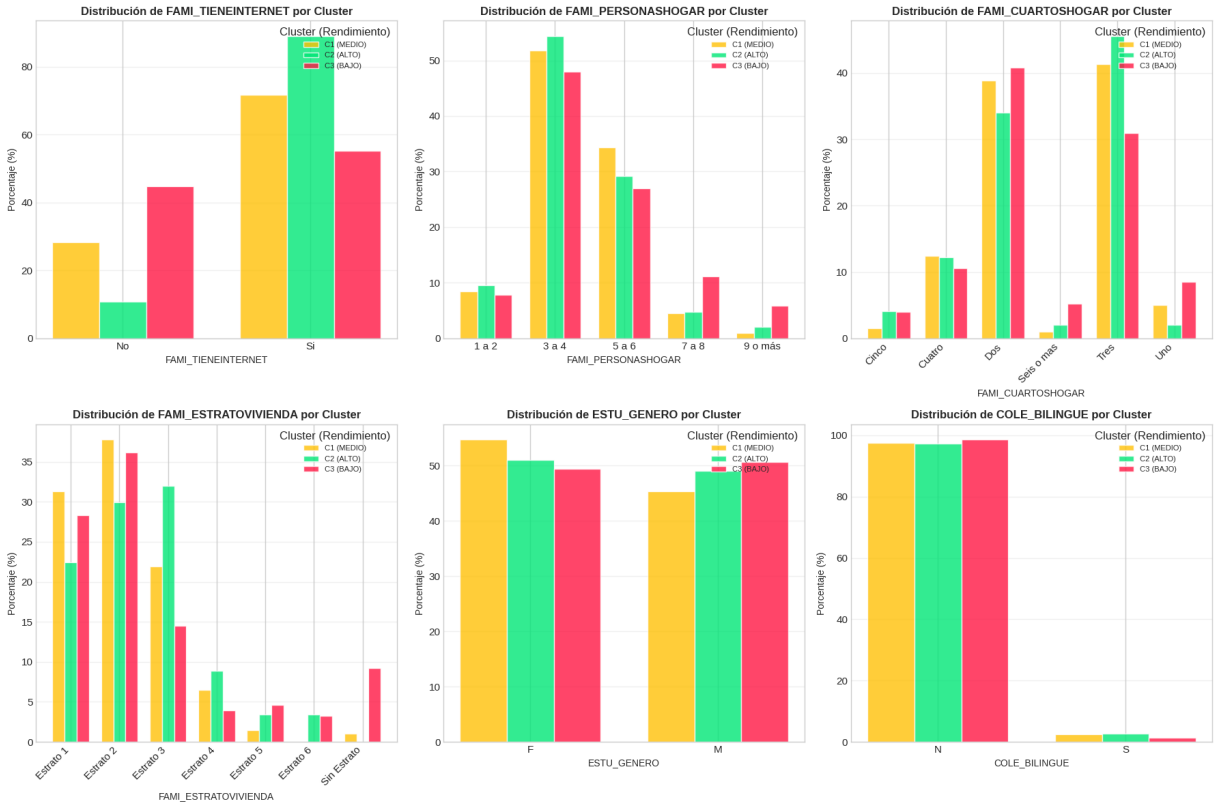


Figura 5.16: Histograma con las cargas porcentuales de las variables categóricas por *cluster*.

Ahora, las variables que si evidencian algún tipo de influencia en los *clusters* son la presencia de conexión a internet y el estrato.

- ***Cluster* 3 (Rendimiento BAJO):** Este clúster, con el puntaje promedio global más bajo, muestra un perfil socioeconómico más vulnerable en comparación con los otros.

  FAMI_TIENEINTERNET: Un porcentaje significativo de hogares (44.7 %) no cuenta con acceso a internet, lo cual es notablemente superior a los otros *clusters*. Aunque la mayoría (55.3 %) sí tiene, la proporción de no conectados es la más alta.

FAMI_ESTRATOVIVIENDA: Predominan los estratos socioeconómicos bajos, con un 28.3 % en Estrato 1 y un 36.2 % en Estrato 2, sumando más del 64 % de este *cluster* en los estratos más bajos. La presencia en estratos altos (4-6) es muy baja.

- ***Cluster* 1 (Rendimiento MEDIO):** Este *cluster* intermedio presenta un perfil socioeconómico "mixto", con mejoras respecto al *cluster* de bajo rendimiento.

  FAMI_TIENEINTERNET: La mayoría de los hogares (71.6 %) sí cuenta con internet, lo que representa una mejora sustancial respecto al Cluster 3.

  FAMI_ESTRATOVIVIENDA: Los estratos 1 (31.3 %) y 2 (37.8 %) siguen siendo predominantes, se observa una mayor proporción en el Estrato 3 (21.9 %) en comparación con el *cluster* 3, indicando una ligera mejora en las condiciones de vivienda. Los estratos más altos (5-6) son prácticamente marginales.

- ***Cluster* 2 (Rendimiento ALTO):** Este *cluster*, asociado al rendimiento más alto, exhibe un perfil socioeconómico más favorecido.

  FAMI_TIENEINTERNET: La conectividad a internet es casi universal (89.1 %). Solo un 10.9 % carece de internet, la proporción más baja de los tres *clusters*.

  FAMI_ESTRATOVIVIENDA: Aunque aún hay presencia en estratos bajos, este *cluster* muestra una distribución más equilibrada y una mayor representación en Estrato 3 (32.0 %) y Estrato 4 (8.8 %) en comparación con los otros grupos. La suma de Estrato 1 y 2 (52.3 %) sigue siendo importante, pero se observa un desplazamiento hacia estratos medios y un poco hacia los altos.

En conclusión, existe una clara correlación entre el rendimiento académico y el acceso a recursos y condiciones socioeconómicas.

El *cluster* 3 (BAJO rendimiento) se caracteriza por la menor proporción de acceso a internet y una fuerte concentración en los estratos socioeconómicos más bajos.

El *cluster* 1 (MEDIO rendimiento) muestra una mejora en la conectividad a internet y una mayor representación en estratos socioeconómicos medios.

Por último, el *cluster* 2 (ALTO rendimiento) destaca por su alta conectividad a internet y una distribución de estratos que tiende más hacia los niveles medios y, en menor medida, altos.

Capítulo 6

Conclusiones

El presente trabajo de grado tuvo como propósito fundamental el estudio de las relaciones a través de la implementación de las técnicas de estadística multivariada entre las variables de la base de datos de las pruebas SABER 11 y su incidencia en los resultados. A lo largo del desarrollo de este estudio, se abordaron los objetivos planteados inicialmente, alcanzando resultados significativos que se describen a continuación.

Logro del objetivo general

El objetivo general del estudio, que consistía en implementar una técnica de estadística multivariada a la base de datos sobre las pruebas SABER 11 con el fin de reconocer relaciones entre variables que pudieran incidir en los resultados de dicha prueba, fue plenamente alcanzado. Se logró la implementación de dos técnicas de análisis multivariado: el análisis de componentes principales (PCA) y el análisis de *clusters (K-means)*. Estas técnicas se aplicaron exitosamente a las variables numéricas (puntajes por área), y posteriormente se estableció su relación con las variables categóricas de índole socioeconómica, permitiendo identificar patrones significativos de rendimiento estudiantil.

Logro de los objetivos específicos

Los objetivos específicos, que guiaron el proceso metodológico del presente trabajo de grado, se cumplieron así:

Contextualización y descripción de los datos

Se realizó una depuración y selección de las variables de la base de datos de las pruebas SABER 11. Este proceso fue fundamental para asegurar la calidad y pertinencia de los

datos utilizados, así como insumo fundamental para la adecuada implementación de las técnicas de análisis multivariado propuestas.

Revisión bibliográfica y marco matemático

El trabajo incluye un marco de referencia sólido que ahonda en la sustentación matemática de las técnicas de PCA y *K-means*, fundamentales para el análisis de los datos. Adicionalmente, se desarrolló un anexo dedicado a Álgebra Lineal, disciplina que proporciona el sustento matemático esencial para la comprensión y aplicación del análisis multivariado, enriqueciendo así la base teórica del estudio.

Implementación de la técnica y análisis de relaciones

La implementación de las técnicas de análisis multivariado se realizó exitosamente en el capítulo de análisis y discusión de resultados. En este apartado, se presentaron las representaciones gráficas resultantes de PCA y *K-means*, y se establecieron las relaciones clave entre las variables numéricas y categóricas. Este paso fue crucial para identificar los perfiles de estudiantes asociados a diferentes niveles de rendimiento.

Sistematización y análisis de resultados

El presente documento, en su totalidad, da cuenta de la sistematización y el análisis de los resultados obtenidos del estudio estadístico multivariado. Desde la preparación de los datos hasta la interpretación de los clústeres, cada etapa del proceso ha sido documentada y analizada rigurosamente, reflejando el cumplimiento integral de este objetivo.

Aportes y trabajos futuros

El presente estudio demuestra que se puede optar por herramientas estadísticas no tradicionales para el análisis de bases de datos del ICFES para la identificación de grupos con respecto al rendimiento estudiantil a partir de datos de pruebas estandarizadas. Los hallazgos aunque no fueron tan ambiciosos ofrecen una visión estructurada y justificada empíricamente de las variables que inciden en los resultados de las pruebas SABER 11, lo cual puede ser de utilidad para la formulación de políticas educativas.

Como trabajo futuro, se podría explorar la aplicación de otras técnicas de agrupamiento más avanzadas, formulación de modelos, *machine learning*, la incorporación de un mayor número de variables socioeconómicas y académicas, o la replicación del estudio con series de tiempo de las pruebas para analizar la evolución de los patrones de rendimiento a lo largo de los años.

Impacto como futuro educador matemático

La realización de este documento ha sido una experiencia formativa retadora que reafirma mi vocación como futuro educador matemático. Este breve estudio no solo reavivó mi interés en ampliar mis conocimientos en el área de la tecnología (programación) y las matemáticas, sino que a la par me proporcionó una comprensión más profunda de cómo estas disciplinas se integran para dar una mirada a las posibles causas de problemas sociales.

Además, resalto como este estudio puso a prueba mis conocimientos teóricos (los cuales debo ampliar y reafirmar constantemente) y me obligó a adquirir nuevas habilidades prácticas. La implementación de técnicas estadísticas multivariadas en Python y el análisis de datos reales me permitieron comprender la relevancia del Álgebra Lineal, la Estadística y la Programación en la investigación de fenómenos educativos. Esta experiencia sin duda me dota de herramientas metodológicas que como educador, podré utilizar para motivar a mis futuros estudiantes mostrándoles la aplicabilidad de las matemáticas en contextos diferentes al salón de clases.

Asimismo, si en un futuro mi trasegar profesional gira a un cargo directivo de cualquier institución educativa cuento con herramientas y rutas metodológicas para la toma de decisiones, habilidades altamente demandadas en el mercado laboral.

Bibliografía

Albornoz, M., Cotes, D., & Rivera, D. (2022). Informe N3 Análisis de componente principal. RPubs. https://rpubs.com/MikeyVega/Informe3

Bolaños, L. (2020). Anális Factorial. RPubs. https://rpubs.com/luis_bolanos/FA

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chica, S., Galvis, D., Ramírez, A. (2012, May 30). Determinantes del rendimiento académico en Colombia. Pruebas ICFES - Saber 11o, 2009*. Revista Universidad EAFIT. https://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/view/754

Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227.

https://doi.org/10.1109/tpami.1979.4766909

Diaz Monroy, L. G. (2007). Estadistica multivariada: Inferencia y metodos. Departamento de Estadistica, Facultad de Ciencias, Universidad Nacional de Colombia.

Diaz Monroy, L. G., Morales Rivera, M. A. (2012). Analisis Estadistico de Datos Multivariados. Universidad Nacional de Colombia.

Facultad de de Ciencias UNAL-MED. (2014). Clase 16. Parte 1. Valores y vectores propios. Universidad Nacional de Colombia - Sede Medellín.

https://ciencias.medellin.unal.edu.co/cursos/algebra-lineal/clases/8-clases/121-clase-16-parte1.html

Gallardo. (2011). Métodos Jerárquicos de Análisis Cluster. Universidad de Granada. http://www.ugr.es/gallardo/pdf/cluster-3.pdf

GeeksforGeeks. (2025, April 2). Elbow method for optimal value of K in kmeans. https://www.geeksforgeeks.org/machine-learning/elbow-method-for-optimal-value-of-k-in-kmeans/

Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning data mining, Inference, and prediction (2nd ed.). Springer New York.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441. https://doi.org/10.1037/h0071325

ICFES. (2025, March 27). Acerca del examen Saber 11°. Instituto Colombiano para la Evaluación de la Educación - ICFES. https://www.icfes.gov.co/

Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaiser, H. F. (1960). The application of electronic computers to Factor Analysis. Educational and Psychological Measurement, 20(1), 141–151.

https://doi.org/10.1177/001316446002000116

LEE. (2024). Informe 92: Pruebas Saber 11: una década de análisis (Abril 2024). Laboratorio de Economía de la Educación (LEE) de la Pontificia Universidad Javeriana. https://lee.javeriana.edu.co/w/lee-informe-92

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Semantic Scholar.

https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed

Marden, J. I. (2015). Multivariate statistics: Old school.

https://people.stat.sc.edu/hansont/stat730/Marden2013.pdf

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1980). Multivariate analysis. Academic Press.

Microsoft. (2025). ¿Qué es la ciencia de datos? cómo convertirte en un científico de datos. Microsoft Azure.

https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-data-science

Min-Educación. (2022). Pruebas saber. Portal MEN. https://www.mineducacion.gov.co/

MIN-TIC. (2025). Datos Abiertos Colombia. Ministerio de Tecnologías de la Información y las Comunicaciones. https://www.datos.gov.co/

Poole, D. (2006). Algebra Lineal - Una Introduccion Moderna. Cengage Learning Editores S.A.

Ramírez, L. (2024, October 30). Algoritmo K-means: ¿Qué es y cómo funciona?. IEBS Biztech School.

https://www.iebschool.com/hub/algoritmo-k-means-que-es-y-como-funciona-big-data/

Rencher, A. C., Christensen, W. F. (2012). Methods of multivariate analysis. Wiley.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Tryon, R. C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor)

Analysis for the Isolation of Unities in Mind and Personality. Edwards brother, Incorporated.

UNESCO. (2024). El impacto de la pandemia en los aprendizajes de los estudiantes de América Latina y el Caribe. UNESCO - UNESDOC Digital Library.

https://unesdoc.unesco.org/ark:/48223/pf0000390609 Corporate author: UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean [810] Latin American Laboratory for the Assessment of Quality in Education [172]

Universidad de los Andes. (2024, July 3). ¿Crisis en la educación media en Colombia?. Universidad de los Andes - Noticias.

https://www.uniandes.edu.co/es/noticias/educacion/las-razones-de-la-crisis-en-la-educacion-media-en-colombia

Anexo de Álgebra Lineal

Dar una mirada al álgebra lineal es un requisito fundamental para el estudio de la estadística multivariada, técnicas propias de esta área del saber como el análisis de componentes principales (PCA), el análisis factorial, el análisis discriminante, la correlación canónica, etc..., tienen su fundamentación teórica en definiciones, propiedades, teoremas, postulados y métodos relacionados con esta rama de las matemáticas.

Para entender y aplicar correctamente estas técnicas, este anexo tiene como objetivo recopilar y recordar conceptos fundamentales del álgebra lineal tomando como referente principal a Poole (2006) y a las clases de Álgebra Lineal de la Facultad de Ciencias de la Universidad Nacional de Colombia - Sede Medellín (2014).

0.1. Vectores en $\mathbb{R}^n$

Definición y notación

Definimos el espacio vectorial euclidiano $\mathbb{R}^n$ como el conjunto de todas las n-tuplas ordenadas de números reales, dotado de las operaciones de suma vectorial y multiplicación por escalares. Formalmente:

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) : x_i \in \mathbb{R},\, i = 1, 2, \ldots, n\} \tag{1}$$

donde cada elemento $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ se denomina **vector** y cada x_i se llama la i-ésima componente o coordenada del vector.

Operaciones fundamentales

- **Suma vectorial:** Para $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ y $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ en $\mathbb{R}^n$:

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n) \tag{2}$$

- **Multiplicación por escalar:** Para $\mathbf{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$ y $\alpha \in \mathbb{R}$ (donde α es

un *escalar*, es decir, un número real que actúa como factor de escalamiento sobre el vector):

$$\alpha\mathbf{v} = (\alpha v_1, \alpha v_2, \ldots, \alpha v_n) \tag{3}$$

Con estas operaciones, $(\mathbb{R}^n, +, \cdot)$ forma un espacio vectorial real de dimensión n sobre el cuerpo de los números reales $\mathbb{R}$.

Un elemento $\mathbf{v}$ de $\mathbb{R}^n$ es llamado vector y puede representarse como una matriz de la forma:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \tag{4}$$

El escalar v_i se denomina la i-ésima componente del vector $\mathbf{v}$.

Por ejemplo, $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}$ es un vector en $\mathbb{R}^3$ cuyas componentes son 1, 2 y 5.

También es común representar un vector como una tupla de la forma:

$$\mathbf{v} = (v_1, v_2, \ldots, v_n) \tag{5}$$

El vector cero, denotado por $\mathbf{0}$, es aquel cuyas componentes son todas iguales a cero:

$$\mathbf{0} = (0, 0, \ldots, 0) \tag{6}$$

0.1.1. Espacios vectoriales

Los vectores son elementos de estructuras algebraicas denominadas espacios vectoriales.

Definición: Un *espacio vectorial* sobre el cuerpo de los números reales $\mathbb{R}$ es una terna $(V, +, \cdot)$ donde V es un conjunto no vacío cuyos elementos se denominan *vectores*, junto con dos operaciones binarias:

- **Suma vectorial:** $+ : V \times V \to V$, que asigna a cada par de vectores $\mathbf{u}, \mathbf{v} \in V$ un vector $\mathbf{u} + \mathbf{v} \in V$
- **Producto por escalar:** $\cdot : \mathbb{R} \times V \to V$, que asigna a cada escalar $c \in \mathbb{R}$ y vector $\mathbf{v} \in V$ un vector $c \cdot \mathbf{v} \in V$

Axiomas del espacio vectorial: Sean $\mathbf{u}$, $\mathbf{v}$ y $\mathbf{w}$ vectores arbitrarios en V y sean $c, d \in \mathbb{R}$ escalares. Entonces las operaciones deben satisfacer los siguientes ocho axiomas:

Axiomas de la suma vectorial:

A1. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (Conmutatividad)

A2. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ (Asociatividad)

A3. Existe un elemento $\mathbf{0} \in V$ tal que $\mathbf{u} + \mathbf{0} = \mathbf{u}$ para todo $\mathbf{u} \in V$ (Elemento neutro)

A4. Para cada $\mathbf{u} \in V$, existe $(-\mathbf{u}) \in V$ tal que $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$ (Inverso aditivo)

Axiomas del producto por escalar:

M1. $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$ (Distributividad respecto a la suma de vectores)

M2. $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$ (Distributividad respecto a la suma de escalares)

M3. $c(d\mathbf{u}) = (cd)\mathbf{u}$ (Asociatividad mixta)

M4. $1 \cdot \mathbf{u} = \mathbf{u}$ para todo $\mathbf{u} \in V$ (Elemento unitario)

Ejemplo fundamental: El conjunto $\mathbb{R}^n$ con las operaciones de suma vectorial y producto por escalar definidas componente a componente forma un espacio vectorial real de dimensión finita n. Específicamente:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n) \tag{7}$$

$$c \cdot (x_1, \ldots, x_n) = (cx_1, \ldots, cx_n) \tag{8}$$

Este ejemplo es fundamental pues todo espacio vectorial real de dimensión finita n es isomorfo a $\mathbb{R}^n$.

0.2. Combinaciones lineales

Una de las operaciones fundamentales en álgebra lineal y especialmente útil en estadística multivariada es la combinación lineal de vectores, que permite generar nuevos vectores a partir de un conjunto dado.

Un vector $\mathbf{v}$ es una combinación lineal de los vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ si existen escalares $c_1, c_2, \ldots, c_k$ tales que:

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k \tag{9}$$

Los escalares $c_1, c_2, \ldots, c_k$ son llamados los coeficientes de la combinación lineal.

Consideremos los vectores $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$ y $\mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ -3 \end{pmatrix}$ en $\mathbb{R}^3$.

El vector $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ puede ser expresado como una combinación lineal de estos vectores:

$$\mathbf{v} = 3\mathbf{v}_1 + 2\mathbf{v}_2$$

Comprobemos:

$$3\mathbf{v}_1 + 2\mathbf{v}_2 = 3 \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} + 2 \begin{pmatrix} -1 \\ 1 \\ -3 \end{pmatrix}$$

$$= \begin{pmatrix} 3 \\ 0 \\ 9 \end{pmatrix} + \begin{pmatrix} -2 \\ 2 \\ -6 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

El resultado coincide con el vector $\mathbf{v}$ propuesto. Esto indica que $\mathbf{v}$ puede expresarse como combinación lineal de $\mathbf{v}_1$ y $\mathbf{v}_2$ con los coeficientes dados.

0.2.1. Independencia lineal

Un concepto estrechamente relacionado con las combinaciones lineales es el de independencia lineal, que resulta fundamental para entender nociones como rango, base y dimensión, que a su vez son vitales en estadística multivariada.

Un conjunto de vectores $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ en un espacio vectorial V se dice linealmente independiente si la ecuación:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \tag{10}$$

tiene como única solución $c_1 = c_2 = \cdots = c_k = 0$.

Si existe otra solución distinta a la trivial, entonces el conjunto se dice linealmente dependiente.

Por ejemplo, consideremos los vectores $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$ y $\mathbf{v}_3 = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$ en $\mathbb{R}^3$.

Verifiquemos si son linealmente independientes resolviendo: $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$

$$c_1 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} + c_3 \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Esto nos da el sistema de ecuaciones:

$$c_1 + 2c_3 = 0$$

$$c_2 + c_3 = 0$$

$$2c_1 - c_2 + 3c_3 = 0$$

De la primera ecuación: $c_1 = -2c_3$ De la segunda ecuación: $c_2 = -c_3$
Sustituyendo en la tercera ecuación:

$$2(-2c_3) - (-c_3) + 3c_3 = 0$$

$$-4c_3 + c_3 + 3c_3 = 0$$

$$0 = 0$$

Obtenemos una identidad $0 = 0$, lo que significa que c_3 puede tomar cualquier valor. Si elegimos $c_3 = 1$, entonces $c_1 = -2$ y $c_2 = -1$.

Por lo tanto, existe una solución no trivial para el sistema, lo que demuestra que los vectores son linealmente dependientes. De hecho, podemos expresar:

$$\mathbf{v}_3 = 2\mathbf{v}_1 + \mathbf{v}_2$$

0.2.2. Subespacio generado

El concepto de subespacio generado (o span) está estrechamente relacionado con las combinaciones lineales y es crucial en la comprensión de las técnicas multivariadas.

El subespacio generado por un conjunto de vectores $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ en un espacio vectorial V, denotado por $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, es el conjunto de todas las combinaciones lineales posibles de estos vectores:

$$\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k : c_i \in \mathbb{R},\ i = 1, 2, \ldots, k\} \qquad (11)$$

Propiedades importantes del subespacio generado:

- El span de cualquier conjunto de vectores es un subespacio vectorial de V

- Contiene al vector cero (tomando todos los coeficientes $c_i = 0$)
- Es el menor subespacio que contiene a todos los vectores generadores
- Si los vectores $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ son linealmente independientes, entonces forman una base del subespacio generado

0.2.3. Proyección ortogonal

Para comprender completamente cómo se relacionan los vectores dentro de un subespacio, es fundamental introducir el concepto de proyección ortogonal, que permite descomponer cualquier vector en componentes paralelas y perpendiculares a una dirección dada.

Conceptos preliminares:

- Dos vectores $\mathbf{u}$ y $\mathbf{v}$ son **ortogonales** si su producto interno es cero: $\mathbf{u} \cdot \mathbf{v} = 0$. Geométricamente, esto significa que forman un ángulo de 90°.
- Un vector $\mathbf{w}$ es **paralelo** a $\mathbf{v}$ si existe un escalar $k \in \mathbb{R}$ tal que $\mathbf{w} = k\mathbf{v}$. Esto significa que ambos vectores tienen la misma dirección (o direcciones opuestas si $k < 0$).

La proyección ortogonal de un vector $\mathbf{u}$ sobre un vector no nulo $\mathbf{v}$ en $\mathbb{R}^n$, denotada por $\text{proy}_{\mathbf{v}}\mathbf{u}$, se define como:

$$\text{proy}_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} \tag{12}$$

donde $\mathbf{u} \cdot \mathbf{v}$ denota el producto interno euclidiano y $\|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$ es el cuadrado de la norma de $\mathbf{v}$.

Propiedades de la proyección ortogonal:

- El vector resultante $\text{proy}_{\mathbf{v}}\mathbf{u}$ es paralelo a $\mathbf{v}$, ya que es un múltiplo escalar de $\mathbf{v}$
- La diferencia $\mathbf{u} - \text{proy}_{\mathbf{v}}\mathbf{u}$ es ortogonal a $\mathbf{v}$, es decir:

$$(\mathbf{u} - \text{proy}_{\mathbf{v}}\mathbf{u}) \cdot \mathbf{v} = 0 \tag{13}$$

- Esta descomposición permite escribir cualquier vector $\mathbf{u}$ como la suma de una componente paralela y una componente ortogonal a $\mathbf{v}$:

$$\mathbf{u} = \text{proy}_{\mathbf{v}}\mathbf{u} + (\mathbf{u} - \text{proy}_{\mathbf{v}}\mathbf{u}) \tag{14}$$

Ejemplo:

Consideremos los vectores $\mathbf{u} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ y $\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ en $\mathbb{R}^3$.

Calculemos la proyección de $\mathbf{u}$ sobre $\mathbf{v}$:

$$\mathbf{u} \cdot \mathbf{v} = 3 \cdot 1 + 2 \cdot 1 + 1 \cdot 0 = 5$$
$$\|\mathbf{v}\|^2 = 1^2 + 1^2 + 0^2 = 2$$

Por lo tanto:

$$\begin{aligned} \text{proy}_{\mathbf{v}}\mathbf{u} &= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} \\ &= \frac{5}{2}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 5/2 \\ 5/2 \\ 0 \end{pmatrix} \end{aligned}$$

El componente de $\mathbf{u}$ ortogonal a $\mathbf{v}$ es:

$$\begin{aligned} \mathbf{u} - \text{proy}_{\mathbf{v}}\mathbf{u} &= \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 5/2 \\ 5/2 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix} \end{aligned}$$

Verificamos que este vector es efectivamente ortogonal a $\mathbf{v}$:

$$\begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{2} \cdot 1 + \left(-\frac{1}{2}\right) \cdot 1 + 1 \cdot 0 = 0$$

Esta descomposición ortogonal es fundamental en el análisis de componentes principales, donde se buscan direcciones ortogonales que maximicen la varianza proyectada de los datos, y cada observación se proyecta sobre estas direcciones principales para obtener sus

coordenadas en el nuevo sistema de referencia.

0.2.4. Proceso de Gram-Schmidt

Definiciones preliminares:

Conjunto ortogonal: Un conjunto de vectores $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ en $\mathbb{R}^n$ se denomina ortogonal si todos los vectores son mutuamente ortogonales, es decir:

$$\mathbf{v}_i \cdot \mathbf{v}_j = 0 \quad \text{para todo } i \neq j$$

Conjunto ortonormal: Un conjunto de vectores $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ en $\mathbb{R}^n$ se denomina ortonormal si:

1. Es ortogonal: $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ para todo $i \neq j$
2. Cada vector tiene norma unitaria: $\|\mathbf{u}_i\| = 1$ para todo $i = 1, 2, \ldots, k$

El proceso de Gram-Schmidt es un método sistemático para convertir un conjunto de vectores linealmente independientes en un conjunto ortogonal (y posteriormente ortonormal). Este proceso es fundamental en muchos algoritmos numéricos utilizados en estadística multivariada, especialmente en la construcción de bases ortonormales necesarias para técnicas como el PCA.

Ejemplo:

Dado un conjunto de vectores linealmente independientes $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, podemos construir un conjunto ortogonal $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ que genera el mismo subespacio mediante el siguiente proceso:

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 \\
\mathbf{u}_2 &= \mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1 \\
\mathbf{u}_3 &= \mathbf{v}_3 - \frac{\mathbf{v}_3 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{v}_3 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2}\mathbf{u}_2 \\
&\vdots
\end{aligned}$$

En general, para $i = 2, 3, \ldots, k$:

$$\mathbf{u}_i = \mathbf{v}_i - \sum_{j=1}^{i-1} \frac{\mathbf{v}_i \cdot \mathbf{u}_j}{\mathbf{u}_j \cdot \mathbf{u}_j}\mathbf{u}_j \tag{15}$$

Para obtener un conjunto ortonormal, simplemente normalizamos cada vector:

$$\mathbf{e}_i = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \tag{16}$$

Consideremos los vectores $\mathbf{v}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}$ y $\mathbf{v}_2 = \begin{pmatrix}1\\1\\0\end{pmatrix}$ en $\mathbb{R}^3$.

Aplicando el proceso de Gram-Schmidt:

$$\mathbf{u}_1 = \mathbf{v}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}$$

$$\mathbf{u}_2 = \mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1$$

$$= \begin{pmatrix}1\\1\\0\end{pmatrix} - \frac{1\cdot 1 + 1\cdot 0 + 0\cdot 1}{1^2+0^2+1^2}\begin{pmatrix}1\\0\\1\end{pmatrix}$$

$$= \begin{pmatrix}1\\1\\0\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\0\\1\end{pmatrix} = \begin{pmatrix}1/2\\1\\-1/2\end{pmatrix}$$

Verificamos que $\mathbf{u}_1$ y $\mathbf{u}_2$ son ortogonales:

$$\mathbf{u}_1 \cdot \mathbf{u}_2 = 1\cdot\frac{1}{2} + 0\cdot 1 + 1\cdot\left(-\frac{1}{2}\right) = \frac{1}{2} - \frac{1}{2} = 0$$

Para obtener vectores ortonormales:

$$\mathbf{e}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\1\end{pmatrix} = \begin{pmatrix}1/\sqrt{2}\\0\\1/\sqrt{2}\end{pmatrix}$$

$$\mathbf{e}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \frac{1}{\sqrt{\frac{1}{4}+1+\frac{1}{4}}}\begin{pmatrix}1/2\\1\\-1/2\end{pmatrix} = \frac{1}{\sqrt{\frac{6}{4}}}\begin{pmatrix}1/2\\1\\-1/2\end{pmatrix} = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\2\\-1\end{pmatrix}$$

0.3. Matrices y transformaciones lineales

Las matrices son fundamentales en álgebra lineal y estadística multivariada, ya que permiten representar datos multidimensionales, transformaciones lineales, sistemas de ecuaciones, y muchas otras estructuras y operaciones.

0.3.1. Definición y notación

Una matriz es una función $A : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{R}$ que asigna a cada par ordenado (i, j) un número real $A(i, j)$, comúnmente denotado como a_{ij}.

Representación: Esta función se representa mediante un arreglo rectangular de números reales organizados en m filas y n columnas de la forma:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{17}$$

donde a_{ij} representa el elemento ubicado en la fila i y columna j de la matriz, correspondiente al valor $A(i, j)$ de la función.

Una matriz de tamaño $m \times n$ tiene m filas y n columnas.

0.3.2. Tipos especiales de matrices

Antes de caracterizar tipos específicos de matrices, es fundamental definir la operación de transposición:

Sea $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ una matriz. La matriz transpuesta de A, denotada por A^T, es la matriz de dimensiones $n \times m$ definida por:

$$A^T = (a_{ji})_{n \times m}$$

Es decir, el elemento en la posición (i, j) de A^T es el elemento en la posición (j, i) de A. Formalmente: $(A^T)_{ij} = a_{ji}$.

Interpretación geométrica: La transposición refleja la matriz respecto a su diagonal principal, intercambiando filas por columnas.

Ejemplo: Si $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, entonces $A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$.

Propiedades de la transposición:

- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- $(cA)^T = cA^T$ para cualquier escalar $c \in \mathbb{R}$
- $(AB)^T = B^T A^T$ (inversión del orden en el producto)

Con esta definición, ahora se puede caracterizar los tipos especiales de matrices:

- **Matriz cuadrada**: Una matriz con igual número de filas y columnas ($m = n$).
- **Matriz diagonal**: Una matriz cuadrada en la que todos los elementos fuera de la diagonal principal son cero, es decir, $a_{ij} = 0$ para $i \neq j$.

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} \tag{18}$$

- **Matriz identidad**: Una matriz diagonal cuyos elementos diagonales son todos iguales a 1, denotada I_n o simplemente I.

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{19}$$

  Propiedad fundamental: $AI_n = I_m A = A$ para cualquier matriz $A \in \mathbb{R}^{m \times n}$ (elemento neutro del producto matricial).

- **Matriz triangular superior**: Una matriz cuadrada donde todos los elementos por debajo de la diagonal principal son cero: $a_{ij} = 0$ para $i > j$.
- **Matriz triangular inferior**: Una matriz cuadrada donde todos los elementos por encima de la diagonal principal son cero: $a_{ij} = 0$ para $i < j$.
- **Matriz simétrica**: Una matriz cuadrada A tal que $A = A^T$, es decir, $a_{ij} = a_{ji}$ para todos los índices i, j.

- **Matriz antisimétrica (o hemisimétrica)**: Una matriz cuadrada A tal que $A = -A^T$, es decir, $a_{ij} = -a_{ji}$ para todos los índices i, j.

0.3.3. Operaciones con matrices

Suma de matrices

Sean $A = (a_{ij})$ y $B = (b_{ij})$ dos matrices de dimensiones $m \times n$. La suma de A y B, denotada por $A + B$, es la matriz $C = (c_{ij})$ de dimensiones $m \times n$ definida por:

$$c_{ij} = a_{ij} + b_{ij} \quad \text{para todo } i \in \{1, 2, \ldots, m\} \text{ y } j \in \{1, 2, \ldots, n\}$$

Propiedades de la suma:

- **Conmutatividad:** $A + B = B + A$
- **Asociatividad:** $(A + B) + C = A + (B + C)$
- **Elemento neutro:** Existe la matriz cero O tal que $A + O = A$
- **Elemento opuesto:** Para cada matriz A existe $-A$ tal que $A + (-A) = O$

Producto por escalar

Sea $A = (a_{ij})$ una matriz de dimensiones $m \times n$ y sea $k \in \mathbb{R}$ un escalar. El producto por escalar $k \cdot A$, denotado simplemente como kA, es la matriz $B = (b_{ij})$ de dimensiones $m \times n$ definida por:

$$b_{ij} = k \cdot a_{ij} \quad \text{para todo } i \in \{1, 2, \ldots, m\} \text{ y } j \in \{1, 2, \ldots, n\}$$

Propiedades del producto por escalar:

- **Distributividad:** $k(A + B) = kA + kB$ y $(k + l)A = kA + lA$
- **Asociatividad:** $k(lA) = (kl)A$
- **Elemento unitario:** $1 \cdot A = A$

Producto de matrices

Sean $A = (a_{ik})$ una matriz de dimensiones $m \times p$ y $B = (b_{kj})$ una matriz de dimensiones $p \times n$. El **producto matricial** AB es la matriz $C = (c_{ij})$ de dimensiones $m \times n$ definida

por:

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \quad \text{para todo } i \in \{1, 2, \ldots, m\} \text{ y } j \in \{1, 2, \ldots, n\}$$

El producto AB está definido si y solo si el número de columnas de A es igual al número de filas de B.

Propiedades del producto matricial:

- **Asociatividad:** $(AB)C = A(BC)$ (cuando los productos están definidos)
- **Distributividad:** $A(B + C) = AB + AC$ y $(A + B)C = AC + BC$
- **No conmutatividad:** En general, $AB \neq BA$ (incluso cuando ambos productos existen)
- **Elemento neutro:** $AI_n = I_m A = A$ para matrices compatibles

0.3.4. Matriz inversa

Sea A una matriz cuadrada de orden n. Una matriz A^{-1} de orden n se denomina **matriz inversa** de A si satisface:

$$AA^{-1} = A^{-1}A = I_n$$

donde I_n es la matriz identidad de orden n.

Existencia y unicidad

- Una matriz cuadrada A es **invertible** (o **no singular**) si y solo si $\det(A) \neq 0$
- Una matriz cuadrada A es **singular** si $\det(A) = 0$ y por tanto no tiene inversa
- Si existe, la matriz inversa es única

Propiedades de la matriz inversa

Para matrices invertibles A y B de orden apropiado:

a. **Unicidad:** Si A es invertible, entonces A^{-1} es única

b. **Involución:** $(A^{-1})^{-1} = A$

c. **Producto:** Si A y B son invertibles, entonces AB es invertible y $(AB)^{-1} = B^{-1}A^{-1}$

d. **Transposición:** Si A es invertible, entonces A^T es invertible y $(A^T)^{-1} = (A^{-1})^T$

e. **Potencias:** $(A^{-1})^n = (A^n)^{-1}$ para cualquier entero positivo n

La matriz inversa es fundamental para resolver sistemas de ecuaciones lineales, calcular la distancia de Mahalanobis, y en la estimación de parámetros mediante mínimos cuadrados.

0.3.5. Determinante

El determinante es una función que asigna a cada matriz cuadrada un escalar y proporciona información sobre si la matriz es invertible.

Para una matriz A de tamaño 2×2:

$$\det(A) = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{20}$$

Para matrices de mayor tamaño, el determinante se puede calcular utilizando expansiones por cofactores o menores.

Propiedades del determinante

a. $\det(I_n) = 1$

b. $\det(AB) = \det(A)\det(B)$

c. $\det(A^T) = \det(A)$

d. Si A es invertible, entonces $\det(A^{-1}) = \frac{1}{\det(A)}$

e. Una matriz A es invertible si y solo si $\det(A) \neq 0$

0.3.6. Rango de una matriz

El rango de una matriz mide la dimensión efectiva de la información que contiene, siendo un concepto fundamental en estadística multivariada para comprender la dimensionalidad de los datos.

El rango de una matriz A, denotado por rango(A), es el número máximo de filas (o columnas) linealmente independientes de A.

Propiedades del rango:

a. $\text{rango}(A) \leq \text{mín}(m, n)$ para una matriz A de tamaño $m \times n$.

b. $\text{rango}(A) = \text{rango}(A^T)$

c. Si A es una matriz cuadrada de tamaño $n \times n$, entonces A es invertible si y solo si $\text{rango}(A) = n$.

0.4. Valores y vectores propios

Los valores y vectores propios constituyen uno de los conceptos más fundamentales del álgebra lineal, con aplicaciones cruciales en estadística multivariada. Estas herramientas son la base teórica de técnicas como el análisis de componentes principales, análisis factorial y análisis discriminante, permitiendo la identificación de direcciones principales de variabilidad en conjuntos de datos multidimensionales.

0.4.1. Definiciones fundamentales

Sea $A \in \mathbb{R}^{n \times n}$ una matriz cuadrada. Un escalar $\lambda \in \mathbb{R}$ se denomina **valor propio** de A si existe un vector no nulo $\mathbf{x} \in \mathbb{R}^n$ tal que:

$$A\mathbf{x} = \lambda\mathbf{x} \tag{21}$$

Un vector $\mathbf{x} \neq \mathbf{0}$ que satisface esta ecuación se llama **vector propio** de A correspondiente al valor propio λ.

Un vector propio de una matriz A es un vector que, al ser transformado por A, resulta en un múltiplo escalar de sí mismo. La matriz A "estira" o "comprime" el vector por un factor λ, pero no cambia su dirección (excepto posiblemente por una reflexión si $\lambda < 0$).

0.4.2. Caracterización algebraica

De la ecuación característica $A\mathbf{x} = \lambda\mathbf{x}$, se obtiene:

$$A\mathbf{x} - \lambda\mathbf{x} = \mathbf{0} \tag{22}$$

$$A\mathbf{x} - \lambda I_n\mathbf{x} = \mathbf{0} \tag{23}$$

$$(A - \lambda I_n)\mathbf{x} = \mathbf{0} \tag{24}$$

Para que exista un vector no nulo $\mathbf{x}$ que satisfaga $(A - \lambda I_n)\mathbf{x} = \mathbf{0}$, es necesario que la matriz $(A - \lambda I_n)$ sea singular, es decir:

$$\det(A - \lambda I_n) = 0 \tag{25}$$

Esta ecuación se conoce como la **ecuación característica** de A.

El polinomio $p(\lambda) = \det(A - \lambda I_n)$ se denomina **polinomio característico** de la matriz A. Sus raíces son los valores propios de A.

0.4.3. Espacios propios

Sea λ un valor propio de la matriz A. El **espacio propio** asociado a λ, denotado por E_λ, se define como:

$$E_\lambda = \text{Nul}(A - \lambda I_n) = \{\mathbf{x} \in \mathbb{R}^n : (A - \lambda I_n)\mathbf{x} = \mathbf{0}\} \tag{26}$$

Propiedades del espacio propio:

- E_λ es un subespacio vectorial de $\mathbb{R}^n$
- Todo vector no nulo en E_λ es un vector propio asociado a λ
- La dimensión de E_λ se llama **multiplicidad geométrica** de λ
- $\dim(E_\lambda) \geq 1$ para todo valor propio λ

Ejemplo:

Consideremos la matriz:

$$A = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$$

Verifiquemos si $\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$ es un vector propio de A.

Calculando $A\mathbf{x}_1$:

$$A\mathbf{x}_1 = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 16 \\ 8 \\ 16 \end{bmatrix} = 8 \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = 8\mathbf{x}_1$$

Por lo tanto, $\mathbf{x}_1$ es un vector propio de A con valor propio asociado $\lambda = 8$.

Para contrastar, consideremos $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$:

$$A\mathbf{x}_2 = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 4 \\ 9 \end{bmatrix}$$

Como no existe un escalar λ tal que $\begin{bmatrix} 9 \\ 4 \\ 9 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, el vector $\mathbf{x}_2$ no es un vector propio de A.

0.4.4. Algoritmo para encontrar valores y vectores propios

1. Plantear la ecuación característica: $\det(A - \lambda I_n) = 0$
2. Resolver el polinomio característico para encontrar los valores propios λ_i
3. Para cada valor propio λ_i, resolver el sistema homogéneo $(A - \lambda_i I_n)\mathbf{x} = \mathbf{0}$
4. Los vectores solución (no nulos) constituyen una base del espacio propio E_{λ_i}

Los valores y vectores propios de la matriz de covarianza de un conjunto de datos proporcionan las direcciones de máxima variabilidad (vectores propios) y la magnitud de esa variabilidad (valores propios), fundamentos del análisis de componentes principales.

0.5. Producto punto y propiedades geométricas

El producto punto es una operación fundamental que permite cuantificar conceptos geométricos como longitud, ángulos y ortogonalidad en espacios vectoriales. Esta herramienta es esencial para comprender las propiedades métricas que sustentan muchas técnicas de estadística multivariada.

Dados los vectores $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$ y $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$ en $\mathbb{R}^n$, el producto punto se define como:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \tag{27}$$

Propiedades del producto punto

El producto punto cumple las siguientes propiedades:

a. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ (Conmutatividad)

b. $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$ (Distributividad respecto a la suma)

c. $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v})$ (Homogeneidad respecto al producto por escalar)

d. $\mathbf{u} \cdot \mathbf{u} \geq 0$ (No negatividad)

e. $\mathbf{u} \cdot \mathbf{u} = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}$ (Positiva definida)

0.6. Longitud o norma de un vector

La longitud o norma de un vector $\mathbf{v}$ en $\mathbb{R}^n$ se define como:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \tag{28}$$

También podemos expresar:

$$\|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v} \tag{29}$$

Propiedades de la longitud de un vector:

1. $\|\mathbf{v}\| = 0 \Leftrightarrow \mathbf{v} = \mathbf{0}$
2. $\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$ para cualquier escalar c
3. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ (Desigualdad triangular)

Consideremos el vector $\mathbf{v} = \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}$ en $\mathbb{R}^3$.

La norma de $\mathbf{v}$ es: $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2} = \sqrt{3^2 + 4^2 + 0^2} = \sqrt{9 + 16 + 0} = \sqrt{25} = 5$

0.6.1. Ángulo entre vectores

El producto punto nos permite calcular el ángulo formado por dos vectores no nulos.

Si $\mathbf{u}$ y $\mathbf{v}$ son vectores no nulos en $\mathbb{R}^n$, el ángulo θ entre ellos es el único valor en el intervalo $[0, \pi]$ que satisface:

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{30}$$

De esta definición se desprenden algunos casos especiales importantes:

1. Si $\mathbf{u} \cdot \mathbf{v} = 0$, entonces $\cos(\theta) = 0$, lo que implica $\theta = \frac{\pi}{2}$. En este caso, decimos que los vectores son ortogonales (perpendiculares).

2. Si $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\|$, entonces $\cos(\theta) = 1$, lo que implica $\theta = 0$. En este caso, los vectores tienen la misma dirección.

3. Si $\mathbf{u} \cdot \mathbf{v} = -\|\mathbf{u}\|\|\mathbf{v}\|$, entonces $\cos(\theta) = -1$, lo que implica $\theta = \pi$. En este caso, los vectores tienen direcciones opuestas.

Consideremos los vectores $\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ y $\mathbf{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ en $\mathbb{R}^3$.

El producto punto es: $\mathbf{u} \cdot \mathbf{v} = 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 = 0$

Las normas son: $\|\mathbf{u}\| = \sqrt{1^2 + 0^2 + 0^2} = 1$ $\|\mathbf{v}\| = \sqrt{0^2 + 1^2 + 0^2} = 1$

Por lo tanto: $\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = 0$

Esto implica que $\theta = \frac{\pi}{2} = 90$, lo que confirma que los vectores son ortogonales, como era de esperar para los vectores unitarios en la dirección de los ejes x e y.

0.6.2. Desigualdad de Cauchy-Schwarz

Una de las desigualdades más importantes en álgebra lineal, que relaciona el producto punto con las normas de los vectores, es la desigualdad de Cauchy-Schwarz.

Para cualesquiera vectores $\mathbf{u}$ y $\mathbf{v}$ en $\mathbb{R}^n$:

$$|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|\|\mathbf{v}\| \tag{31}$$

con igualdad si y solo si uno de los vectores es múltiplo escalar del otro (son linealmente dependientes).

Esta desigualdad es fundamental en muchas demostraciones y aplicaciones, incluidas las relacionadas con estadística multivariada.