Centro de Investigación en Matemáticas (CIMAT).
Unidad Monterrey

Ciencia de Datos Tarea 1

Diego Godinez Bravo

21 de febrero de 2024

1. Problema 1

Considera los datos que se encuentran en el archivo suicide_data.csv, que compila información relacionada con suicidios en 101 países en diferentes años e incluye también algunos indicadores de desarrollo, como el Producto Interno Bruto (GPD) per capita, Índice de Desarrollo Humano (HDI), entre otra información.

a) Formula al menos dos preguntas de interés para analizar este fenómeno en base a estos datos.

b) Realiza un análisis exploratorio y descriptivo de los datos con las herramientas que consideres apropiadas. ¿Qué estructuras o correlaciones encuentras? Comenta tus hallazgos.

c) Trata de responder las preguntas que formulaste en el primer inciso. ¿Puedes obtener conclusiones interesantes?

d) ¿Qué otra información te gustaría tener disponible para analizar este fenómeno y cómo la utilizarías?

El resultado, para éste ejercicio, debe ser un reporte corto donde describas los pasos que seguiste y las conclusiones a las que llegas, incluyendo las gráficas más ilustrativas o informativas que consideres.

Nota: La base de datos de suicidios contiene datos faltantes. Indica en el reporte cómo los manejaste.

1.1. Solución

El conjunto de datos contiene información relevante sobre el número de suicidios en 101 países a lo largo de los últimos años, así como datos socioeconómicos e indicadores de desarollo relacionados con cada uno de ellos. El análisis exploratorio y descriptivo de los datos se llevó a cabo utilizando el lenguaje de programación Python, haciendo uso específicamente de la librería Pandas, la cual está especializada en el manejo y análisis de estructuras de datos.

Con base en los datos proporcionados se formularon las siguientes preguntas:

i. ¿Existe una diferencia significativa en el número de suicidios entre hombres y mujeres? En caso afirmativo, ¿cuáles son los factores que contribuyen a esta disparidad?

ii. ¿Cúal ha sido la tendencia de la tasa de suicidios anual y el Producto Interno Bruto (PIB) en los últimos 10 años en países de América Latina?

Se llevó a cabo un análisis de datos con el fin de abordar las preguntas planteadas. En primera instancia, se exploró el conjunto de datos para comprender su estructura general y visualizar su contenido. De esta manera se encontró que el conjunto de datos estaba conformado por 27, 820 filas y 12 columnas. Se verificó la ausencia de datos duplicados y la presencia de datos faltantes. Finalmente, se examinaron los valores únicos para columnas específicas para obtener panorama general del conjunto de datos, enfocándonos particularmente en la lista de países y los intervalos de edad incluidos en el conjunto de datos.

Una vez que se obtuvo una visión general de los datos, se realizaron diversas tareas preliminares de preparación y limpieza de los datos. Dentro de las cuales se encuentran la redefinición de las columnas para mejorar su comprensión y manejo, así como el tratamiento de los valores faltantes utilizando el método de 'backfill'. Este método permite rellenar los valores nulos o faltantes utilizando la siguiente observación válida, evitando que existan restricciones al momento de aplicar las funciones que nos permitan trabajar con nuestro conjunto de datos.

Para concentrar el análisis en un conjunto de datos más específico, se filtró el conjunto de datos para seleccionar únicamente los países de América Latina. A partir de este se generaron distintos gráficos que proporcionaron un resumen general de las posibles correlaciones y tendencias presentes en el conjunto de datos.

La matriz de correlación es una herramienta estadística que muestra los valores de correlación, los cuales miden el grado de relación lineal entre cada par de variables (**Figura 1.1**). De manera general, se observa que una correlación positiva entre el número de suicidios y la población total, lo que sugiere que a medida que aumenta el tamaño de la población, también lo hace el número de suicidios. Por otro lado, se evidencia una correlación positiva entre el Índice de Desarrollo Humano (HDI) y el Producto Interno Bruto (PIB) per cápita, lo que indica que a medida que mejora el desarrollo humano, también aumenta la riqueza económica. Es importante recordar que la correlación no implica causalidad, por lo que se requiere más información para confirmar cualquier relación causal entre las variables. Además, se debe considerar que a pesar de que se analizaron exclusivamente países de América Latina, no es apropiado generalizar el comportamiento de las variables, dado que cada país tiene su propio contexto socioeconómico. Por lo tanto, se necesita información específica de cada país individualmente para comprender las relaciones entre las variables.

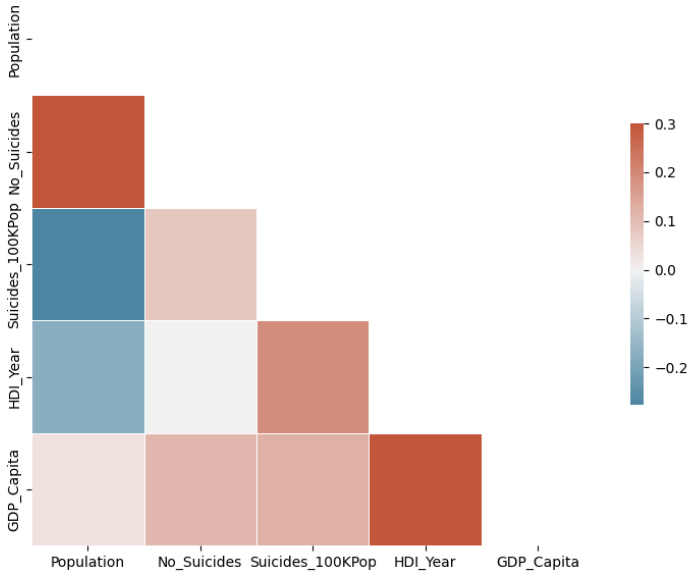

Figura 1.1. Matriz de correlación.

Se generó un gráfico circular el cual ilustra la diferencia en las tasas de suicidios entre hombres y mujeres (**Figura 1.2**). Observamos que existe una diferencia significativa entre la tasa de suicidios de hombres, con un 76.9 %, y la tasa de suicidios de mujeres, con un 23.1 %. Esta notable discrepancia resalta la importancia de promover la salud mental y el bienestar tanto en hombres como en mujeres, identificando y abordando los factores que contribuyan a este fenómeno. Dada la limitación del conjunto de datos, se requieren estudios adicionales que nos ayuden a comprender las causas subyacentes de esta disparidad.

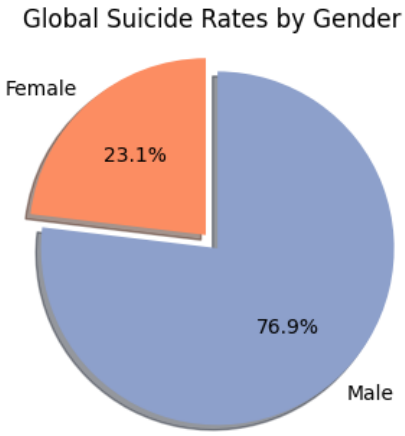

Figura 1.2. Comparación de las tasas de suicidio entre hombres y mujeres.

Para profundizar en la observación previa, se generó un gráfico con el fin de visualizar el número de suicidios a lo largo del tiempo, desglosado por género, para distintos países de América Latina (**Figura 1.3**). En cada caso, se observa una tendencia general donde el número de suicidios entre hombres supera al de mujeres. Esta discrepancia en el número de suicidios entre géneros es consistente con la tendencia global observada en la **Figura 1.2**, resaltando la importancia de comprender los factores específicos que puedan contribuir a esta disparidad en cada país de manera individual.

Con el objetivo de analizar la evolución de las tasas de suicidio en países de América Latina a lo largo del tiempo, se ha generó un gráfico que muestra el número de suicidios para ambos géneros en función del tiempo (**Figura 1.4**). Este gráfico muestra una visión general de cómo ha variado el número de suicidios en la región a lo largo de los años y proporciona una perspectiva sobre la tendencia temporal en la tasas de suicidios. En general, se observa una tendencia creciente en la tasa de suicidios con el paso del tiempo en los países de América Latina. Este hallazgo sugiere la necesidad de una atención continua en la región para abordar los factores subyacentes que pueden contribuir al aumento de las tasas de suicidio e implementar medidas preventivas de manera eficaz.

Es importante destacar que aunque los gráficos presentados proporcionan una visión general de la tendencia en las tasas de suicidio para hombres y mujeres en América Latina, aún falta información y contexto adicional para comprender completamente los factores que puedan explicar este fenómeno. Para profundizar en nuestro análisis, sería útil considerar variables adicionales que pudieran influir en las tasas de suicidio, como el acceso a servicios de salud mental, la prevalencia de trastornos mentales, los niveles socioeconómicos de las personas afectadas, la calidad de vida, así como la accesibilidad y disponibilidad de recursos comunitarios de apoyo. La inclusión de estos datos y variables complementarias podría ayudar a proporcionar una comprensión más completa y precisa de los factores que contribuyen al aumento de las tasas de suicidio en la región.

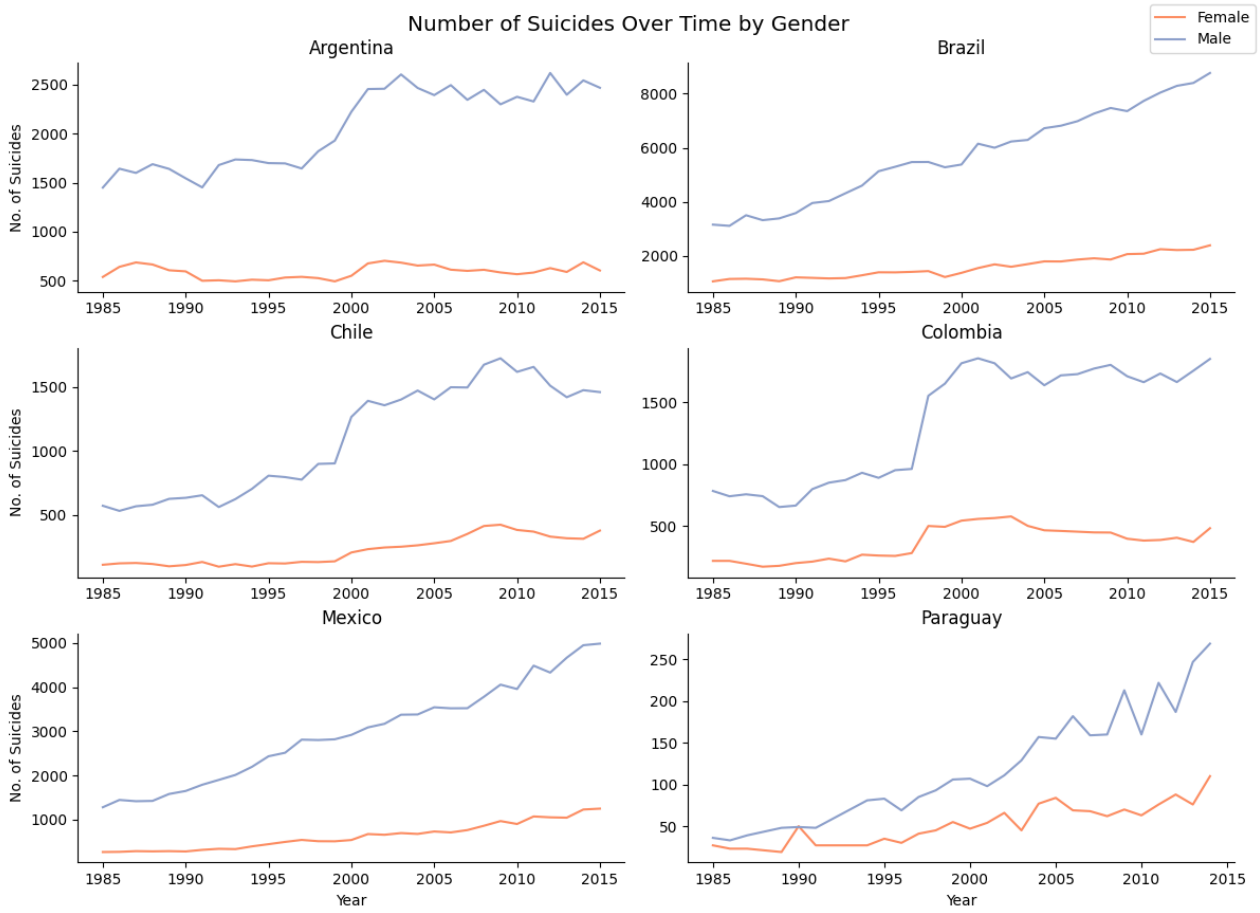

Figura 1.3. Número de suicidios a lo largo del tiempo divididos por género, para distintos países de América Latina.

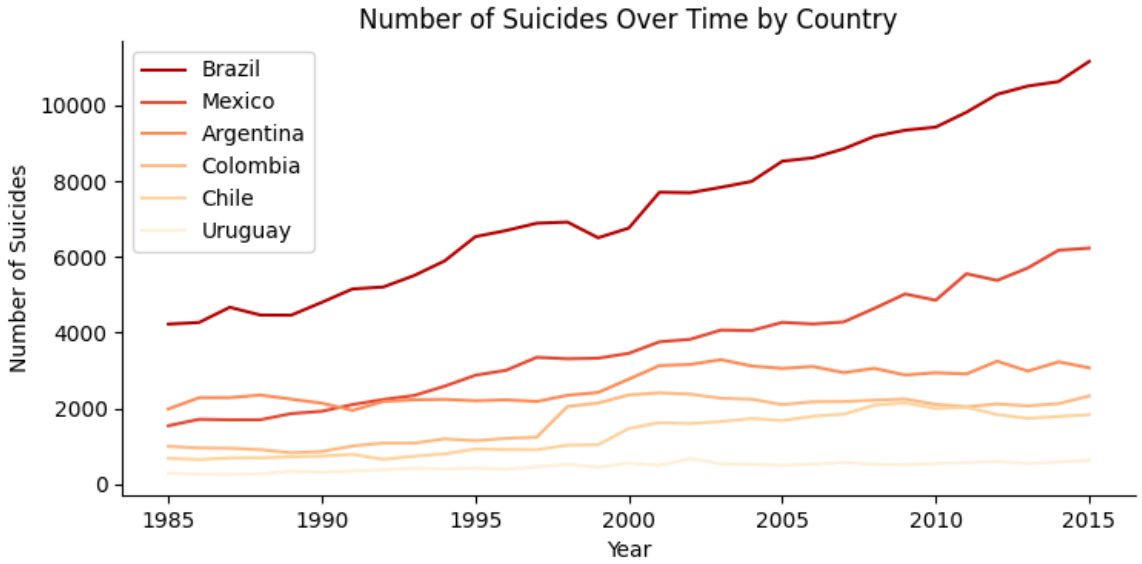

Figura 1.4. Número de suicidios a lo largo del tiempo en países de América Latina.

2. Problema 2

Considera una matriz de datos X_{nxd}. PCA puede formularse también como el problema de encontrar un subespacio (ortonormal) de baja dimensión de forma tal que se minimicen los errores de las proyecciones de los datos en tal subespacio.

Si consideramos una base ortonormal $\{u_j\}, j = 1, \ldots, d$, ya vimos que una observación x_i puede expresarse como una combinación lineal

$$x_i = \sum_{j=1}^{d} \alpha_{ij} u_j.$$

Por la ortogonalidad de u_j, podemos expresar $\alpha_{ij} = x_i' u_j$. Entonces

$$x_i = \sum_{j=1}^{d} (x_i' u_j) u_j.$$

Ahora, considera una aproximación basada en los primeros $p < d$ vectores de la base de acuerdo al modelo lineal:

$$\hat{x_i} = \sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j.$$

Observa que los coeficientes z_{ij} dependen de la observación i, mientras que b_j son constantes para todas las observaciones.

Considera la minimización de la siguiente función de costo:

$$L = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x_i}\|^2. \tag{2.1}$$

a) Muestra que en el mínimo de (2.1):

$$\begin{aligned} z_{ij} &= x_i^{'} u_j, \quad j = 1, ..., p \\ b_j &= \bar{x}^{'} u_j, \quad j = p+1, ..., d \\ x_i - \hat{x_i} &= \sum_{j=p+1}^{d} [(x_i - \bar{x})^{'} u_j] u_j \end{aligned}$$

es decir, la "desviación" está en el espacio ortogonal de los componentes principales.

b) Considerando lo anterior, muestra que (2.1) puede escribirse como:

$$L = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} \left(x_i' u_j - \bar{x}' u_j\right)^2 = \sum_{j=p+1}^{d} u_j' S u_j,$$

es decir, la solución se obtiene resolviendo un problema de valores y vectores propios (restringida), como vimos antes.

Usando el método de Lagrange, es fácil ver (no es necesario demostrarlo) que lo anterior es equivalente a minimizar $L = \sum_{j=p+1}^{d} \lambda_j$, donde λ_j son los valores propios de S, por lo tanto, debemos escoger los vectores propios correspondientes a los valores propios más chicos: $0 \leq \lambda_d \leq \lambda_{d-1} \leq \lambda_{d-p}$, por lo que la mejor aproximación de x (en los componentes principales) está dada por los eigenvectores que corresponden a los eigenvalores más grandes, tal como lo vimos en clase.

2.1. Solución

Desarrollando la expresión 2.1

$$L = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x_i}\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{x_i})'(x_i - \hat{x_i})$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(x_i{}'x_i - x_i{}'\hat{x_i} - \hat{x_i}{}'x_i + \hat{x_i}{}'\hat{x_i}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(x_i{}'x_i - 2x_i{}'\hat{x_i} + \hat{x_i}{}'\hat{x_i}\right)$$

De acuerdo con la condición de optimalidad de primer orden, un punto x^* es un mínimo o máximo local si el gradiente de la función objetivo es cero. Por lo tanto procedemos a calcular la derivada e igualamos a cero.

Derivada con respecto a z_{ij}.

$$\frac{\partial}{\partial z_{ij}} L = \frac{\partial}{\partial z_{ij}} \left(\frac{1}{n}\sum_{i=1}^{n} \left(x_i{}'x_i - 2x_i{}'\hat{x_i} + \hat{x_i}{}'\hat{x_i}\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial}{\partial z_{ij}}(x_i{}'x_i) - \frac{\partial}{\partial z_{ij}}(2x_i{}'\hat{x_i}) + \frac{\partial}{\partial z_{ij}}(\hat{x_i}{}'\hat{x_i})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(- \frac{\partial}{\partial z_{ij}}(2x_i{}'\hat{x_i}) + \frac{\partial}{\partial z_{ij}}(\hat{x_i}{}'\hat{x_i})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(\left(- 2\sum_{j=1}^{p} x_i{}'u_j\right) + \frac{\partial}{\partial z_{ij}}(\hat{x_i}{}'\hat{x_i})\right)$$

Considerando que

$$\hat{x_i}{}'\hat{x_i} = < \hat{x_i}, \hat{x_i} >$$

Entonces

$$= \left\langle \left(\sum_{j=1}^{p} z_{ij}u_j + \sum_{j=p+1}^{d} b_j u_j\right), \left(\sum_{j=1}^{p} z_{ij}u_j + \sum_{j=p+1}^{d} b_j u_j\right)\right\rangle$$

$$= \left(\sum_{j=1}^{p} z_{ij}\right)^2 < u_j, u_j > + \sum_{j=1}^{p} z_{ij} \sum_{j=p+1}^{d} b_j < u_j, u_j > + \sum_{j=1}^{p} z_{ij} \sum_{j=p+1}^{d} b_j < u_j, u_j > + \left(\sum_{j=p+1}^{d} b_j\right)^2 < u_j, u_j >$$

$$= \left(\sum_{j=1}^{p} z_{ij}\right)^2 < u_j, u_j > + 2\sum_{j=1}^{p} z_{ij} \sum_{j=p+1}^{d} b_j < u_j, u_j > + \left(\sum_{j=p+1}^{d} b_j\right)^2 < u_j, u_j >$$

$$= \left(\sum_{j=1}^{p} z_{ij}\right)^2 < u_j, u_j > + \left(\sum_{j=p+1}^{d} b_j\right)^2 < u_j, u_j >$$

Por lo tanto

$$= \frac{1}{n}\sum_{i=1}^{n}\big(\big(-2\sum_{j=1}^{p}x_i{}'u_j\big)+\frac{\partial}{\partial z_{ij}}(\hat{x_i}'\hat{x_i})\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\big(\big(-2\sum_{j=1}^{p}x_i{}'u_j\big)+\frac{\partial}{\partial z_{ij}}(\big(\sum_{j=1}^{p}z_{ij}\big)^2<u_j,u_j>+\big(\sum_{j=p+1}^{d}b_j\big)^2<u_j,u_j>)\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\big(\big(-2\sum_{j=1}^{p}x_i{}'u_j\big)+2\big(\sum_{j=1}^{p}\big)z_{ij}<u_j,u_j>\big)$$

Sabemos que u_j es una base ortonormal, por lo tanto

$$= \frac{1}{n}\sum_{i=1}^{n}\big(\big(-2\sum_{j=1}^{p}x_i{}'u_j\big)+2\big(\sum_{j=1}^{p}z_{ij}\big)\big)$$

Igualando a cero y encontrando el punto de inflexión

$$0 = \frac{1}{n}\sum_{i=1}^{n}\big(\big(-2\sum_{j=1}^{p}x_i{}'u_j\big)+2\big(\sum_{j=1}^{p}z_{ij}\big)\big)$$

$$\frac{n}{2}(0) = \big(\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\big(-x_i{}'u_j\big)+\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\big(z_{ij}\big)\big)\frac{n}{2}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{p}\big(-x_i{}'u_j\big)+\sum_{i=1}^{n}\sum_{j=1}^{p}\big(z_{ij}\big)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\big(x_i{}'u_j\big) = \sum_{i=1}^{n}\sum_{j=1}^{p}\big(z_{ij}\big)$$

De esta manera demostramos que $z_{ij} = x_i^{'}u_j$ con $j = 1, ..., p$.

Derivada con respecto a b_j.

$$\frac{\partial}{\partial b_j}L = \frac{\partial}{\partial b_j}\big(\frac{1}{n}\sum_{i=1}^{n}\big(x_i{}'x_i - 2x_i{}'\hat{x_i}+\hat{x_i}'\hat{x_i}\big)\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\big(-2\sum_{j=p+1}^{d}x_i{}'u_j\big)+2\frac{1}{n}\sum_{i=1}^{n}\sum_{j=p+1}^{d}b_j$$

$$= -\sum_{j=p+1}^{d}\big(\frac{1}{n}\sum_{i=1}^{n}x_i{}'\big)u_j+\sum_{j=p+1}^{d}b_j$$

$$= -\sum_{i=1}^{n}\bar{x}'u_j+\sum_{p+1}^{d}b_j$$

$$\sum_{i=1}^{n}\bar{x}'u_j = \sum_{p+1}^{d}b_j$$

Por lo tanto $b_j = \bar{x}^{'}u_j$ con $j = p+1, ..., d$.

Sea

$$x_i = \sum_{j=1}^{d}(x_i'u_j)u_j$$

y

$$\hat{x_i} = \sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j$$

Entonces

$$x_i - \hat{x_i} = \sum_{j=1}^{d} (x_i' u_j) u_j - \Big(\sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j . \Big)$$

$$= \sum_{j=1}^{d} (x_i' u_j) u_j - \sum_{j=1}^{p} z_{ij} u_j - \sum_{j=p+1}^{d} b_j u_j$$

$$= \sum_{j=1}^{d} (x_i' u_j) u_j - \sum_{j=1}^{p} z_{ij} u_j - \sum_{j=p+1}^{d} b_j u_j$$

Considerando que $z_{ij} = x_i^{'} u_j$ y $b_j = \bar{x}^{'} u_j$, entonces

$$= \sum_{j=1}^{d} (x_i' u_j) u_j - \sum_{j=1}^{p} (x_i^{'} u_j) u_j - \sum_{j=p+1}^{d} (\bar{x}^{'} u_j) u_j$$

$$= \sum_{j=p+1}^{d} (x_i' u_j) u_j - (\bar{x}^{'} u_j) u_j$$

$$= \sum_{j=p+1}^{d} [(x_i{}' - \bar{x}') u_j] u_j$$

Finalmente

$$= \sum_{j=p+1}^{d} [(x_i - \bar{x})' u_j] u_j$$

$$x_i - \hat{x_i} = \sum_{j=p+1}^{d} [(x_i - \bar{x})' u_j] u_j$$

2.2. Solución

Desarrollando la ecuación L

$$L = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} (x_i{}' u_j - \bar{x}' u_j)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} (x_i{}' u_j - \bar{x}' u_j)(x_i{}' u_j - \bar{x}' u_j)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} (u_j{}' x_i - u_j{}' \bar{x})(x_i{}' u_j - \bar{x}' u_j)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} u_j{}' (x_i - \bar{x})(x_i{}' - \bar{x}') u_j$$

$$= \sum_{j=p+1}^{d} u_j{}' \big(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'\big) u_j$$

Sea

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$$

Entonces

$$= \sum_{j=p+1}^{d} u_j{}' S u_j$$

Por lo tanto demostramos que L se puede expresar como

$$L = \sum_{j=p+1}^{d} u_j{}' S u_j$$

3. Problema 3

Los datos que se encuentran en el archivo IMM_2020.xls corresponden al índice de marginación (IM) por cada municipio del país, calculado por el Consejo Nacional de Población (CONAPO) en 2020, y es hasta la fecha, el más reciente. El archivo contiene también los valores de varios indicadores que representan "nueve formas de exclusión de la marginación en las dimensiones: educación, vivienda, distribución de la población e ingresos monetarios", y fueron construidos en base a la información del censo de población y vivienda 2020 realizado INEGI. Estos indicadores se usan para calcular el IM según se describe en la *Nota técnico-metodológica*. El índice de marginación por municipio se muestra de forma categorizada en la Figura 1.

Supón que trabajas como asesor en la Secretaría del Desarrollo Social (o su equivalente) en algún estado del país, y te piden analizar formas alternativas de construir ese índice. Para eso, considera las siguientes actividades.

a) Realiza un análisis de PCA basado en los 9 indicadores de CONAPO. ¿Qué puedes decir respecto al fenómeno de marginación en el país basado en este análisis? ¿Encuentras algún patrón interesante respecto a los municipios?

b) Construye un IM alternativo tomando el primer componente principal que obtuviste con PCA (el indicador debe estar en $[0, 1]$ preferentemente). Compáralo con el de CONAPO y describe tus hallazgos al respecto. ¿Qué tanto cambia el resultado si tomas el segundo o tercer componente principal? ¿tendría algún sentido hacerlo?

c) ¿Qué otra información propondrías que se incluyera dentro de la elaboración de tu índice (ya sea de estadísticas oficiales o de otra fuente)? ¿Estás de acuerdo con la metodología usada? ¿Tienes alguna otra propuesta para la elaboración del índice?

El resultado de éste ejercicio debe ser un reporte ejecutivo con tus hallazgos y conclusiones, como si estuviera dirigido a tu hipotético jefe que no sabe nada (o casi nada) de estos métodos. Cada inciso puede ser un anexo técnico, con las tablas o ilustraciones que creas apropiadas para respaldar tu reporte.

N o t a: Reporte ejecutivo anexado al final del documento.

4. Problema 4

Considera el conjunto de datos Labelled Faces in the Wild (LFW), que consiste en fotografías de rostros recolectados de internet y contenido en sklearn. Algunos rostros identidicados, tienen varias fotos incluídas en el dataset. Vamos a considerar solo aquellas personas que tienen al menos 70 fotografías de su rostro, tambiénm vamos a considerar el tamaño original de la imágen (125 x 94).

```
from sklearn.datasets import fetch_lfw_people
lfw_people = fetch_lfw_people(min_faces_per_person=70, resize=1)
```

Esto resulta en 1299 imágenes que pertenecen a alguna de las etiquetas.

```
>>> for name bluein lfw_people.target_names:
>>>     print(name)
Ariel Sharon
Colin Powell
Donald Rumsfeld
George W. Bush
Gerhard Schroeder
Hugo Chavez
Tony Blair
```

a) Separa un conjunto de entrenamiento (80 %) y prueba (puedes usar la función train_test_split de sklearn.model_selection), por ejemplo:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test, names_train,
names_test = train_test_split(X, y, target_names[y],
test_size=0.2, random_state=42)
```

donde antes, tuviste que declarar X, y, target_names (ve la documentación de fetch_lfw_people). Obtén las eigenfaces del conjunto de entrenamiento. Visualiza los scores de los primeros dos componentes principales. ¿Encuentras patrones interesantes?

b) Proyecta los datos de prueba en los componentes principales. Verifica si se "ubican" en su "individuo" correspondiente al graficarlos en los primeros dos componentes principales.

c) Utiliza el método del vecino más cercano para identificar a un "sujeto" de prueba en las imágenes de entrenamiento. Usa la distancia euclidiana en el espacio de los p componentes principales. Decide qué valor de p usar.

¿Puedes identificar correctamente a los sujetos usando éste criterio? ¿Qúe tanto influye el valor de p?

d) Considera una(s) imágen(es) que no está(n) en la base de datos. ¿Qué se te ocurre para prevenir casos como los que muestran en la Figura 4?

4.1. Solución

El análisis de componentes principales (PCA) es una técnica ampliamente utilizada en el análisis de datos para reducir la dimensionalidad de un conjunto de datos. Transforma un conjunto de variables correlacionadas (p) en un número k ($k < p$) más pequeño de variables no correlacionadas (componentes principales) manteniendo la mayor variación posible en el conjunto de datos original. Los componentes principales son vectores ortogonales que proporcionan una representación de los datos en dimensiones inferiores, al mismo tiempo que preservan una cantidad significativa de la variabilidad presente en el conjunto de datos original. Seleccionar un número reducido de componentes puede resultar en la pérdida de información importante, mientras que elegir un número excesivo puede conducir al sobreajuste o a datos innecesariamente de alta dimensión.

Con el propósito de determinar el número óptimo de componentes principales que capturen la mayor variabilidad en los datos, se generó un 'Scree plot' (**Figura 3.1**). Se observa que los primeros 50 componentes principales capturan aproximadamente el 82 % de la variabilidad total del conjunto de datos.

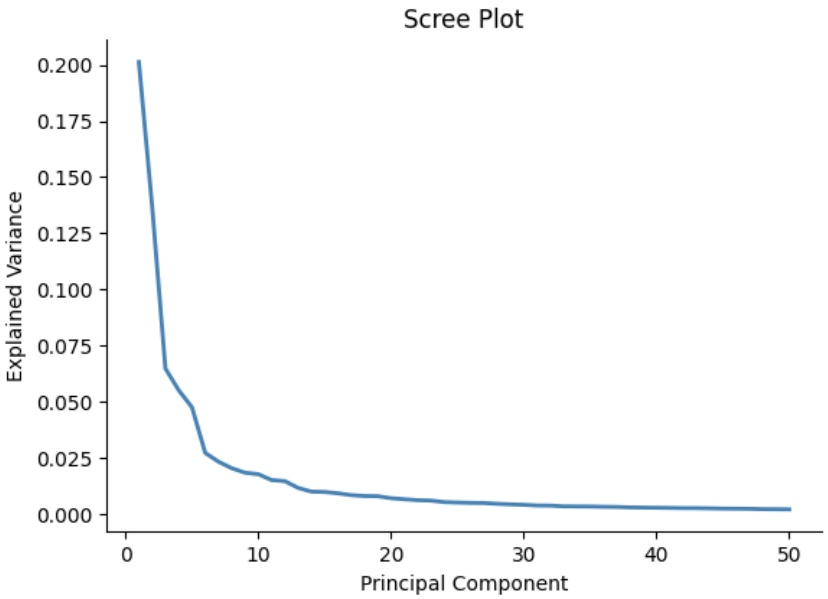

Figura 4.1. Número de componentes principales que capturan la mayor variación en los datos.

La visualización de los scores del primer componente principal en relación con los del segundo componente principal ofrece una visión amplia de la distribución de las observaciones en un espacio de menor dimensión. Cuando estos dos componentes principales capturan la mayor parte de la variación de los datos, el gráfico puede ser de gran utilidad para evaluar la estructura de los datos, así como para identificar posibles agrupaciones, valores atípicos y patrones.

Al observar este gráfico (**Figura 3.2**), se destaca una tendencia notable que relaciona la iluminación de las imágenes con las expresiones faciales. Las imágenes más oscuras tienden a agruparse en el lado derecho del gráfico, exhibiendo rostros más expresivos, mientras que aquellas con mayor claridad en cuanto a iluminación se agrupan en el lado izquierdo, mostrando rostros menos expresivos. Esta asociación sugiere una relación entre la intensidad de la iluminación y la expresividad facial.

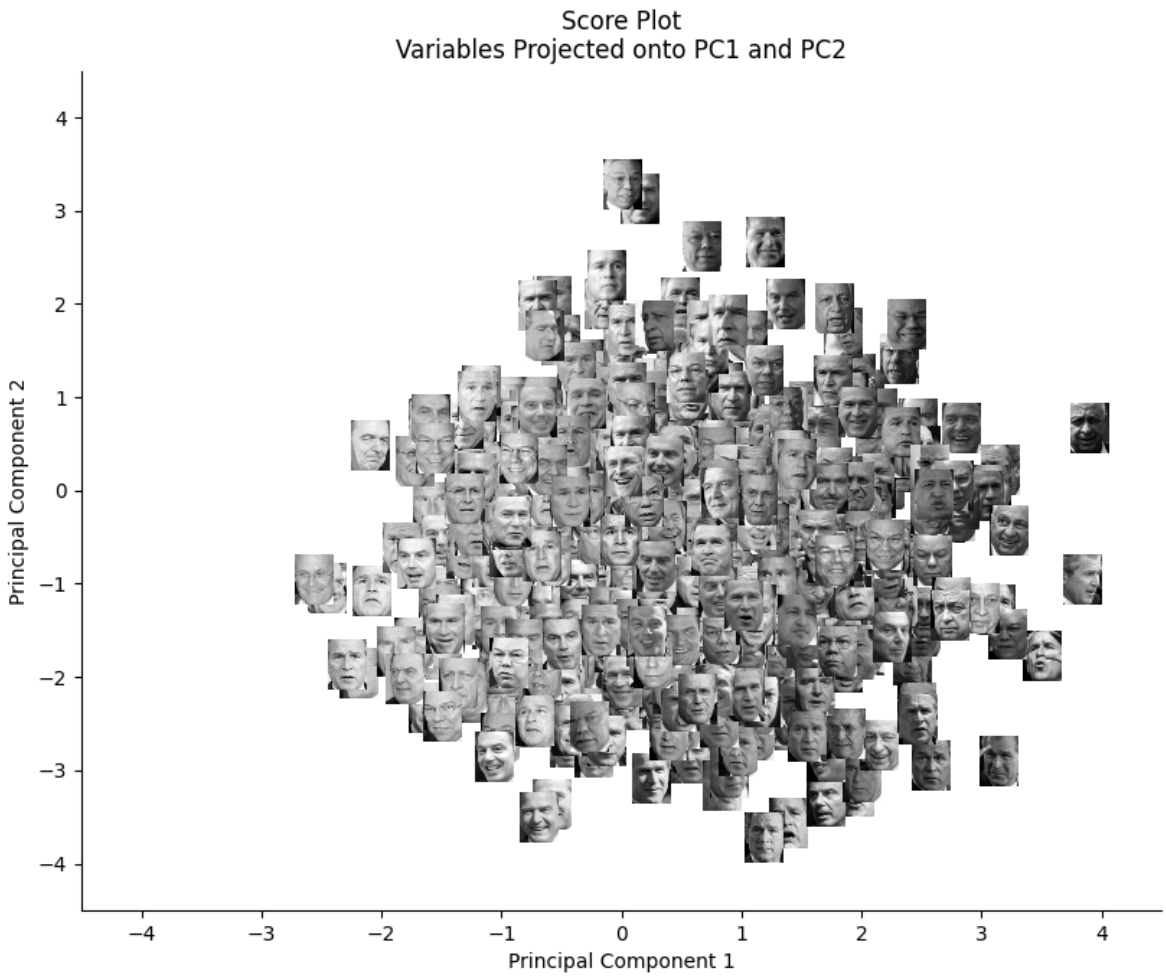

Figura 4.2. Scores del primer componente principal contra los scores del segundo componente principal.

4.2. Solución

Para validar la correspondencia entre los datos de prueba y sus respectivos 'individuos' en el conjunto de datos, realizamos la proyección en las componentes pricipales con el propósito de visualizar la distribución de los datos de prueba en un espacio de menor dimensión y evaluar si se 'ubican' correctamente en relación con sus 'individuos' correspondientes.

Visualizamos las imágenes del conjunto de datos de prueba junto con sus etiquetas correspondientes y las etiquetas predichas (**Figura 3.3**). Observamos que 10 de los 12 rostros coinciden con su respectivo nombre, lo que sugiere que están correctamente representados y mantienen las características importantes de los 'individuos' del conjunto de datos.

Este proceso de validación es fundamental para asegurar la confiabilidad de las proyecciones de los datos de prueba y para validar la utilizad del modelo de PCA en la representación y comprensión de la estructura de los datos.

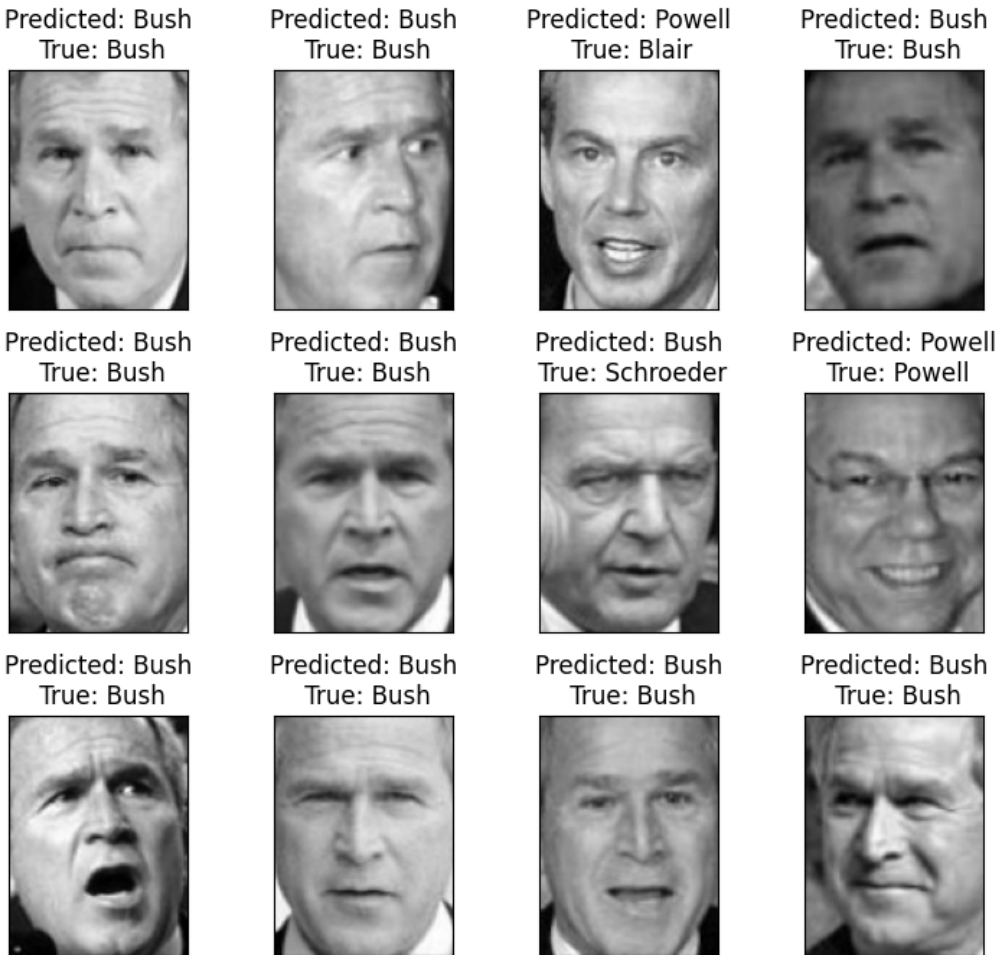

Figura 4.3. Correspondencia entre las imágenes de prueba con la etiqueta predicha.

4.3. Solución

El principio detrás de los métodos del vecino más cercano es encontrar un número predefinido de muestras de entrenamiento cercanas en distancia a un nuevo punto y predecir la etiqueta a partir de ellas. El número de muestras puede ser una constante definida por el usuario o variar según la densidad local de puntos. En general la distancia puede ser cualquier medida métrica, la distancia euclidiana estándar es la opción más común.

En este caso se utilizó dicho método para identificar a un 'sujeto' dentro de las imágenes de entrenamiento, utilizando la distancia euclidiana como medida métrica. Se generó una imagen compuesta que incluye la imagen de prueba, la imagen proyectada en el espacio de las componentes principales y la imagen del vecino más cercano (**Figura 4.4**). Este enfoque nos permitió visualizar y comparar de manera directa la imagen de prueba con su correspondiente más cercano en el espacio de las componentes principales. Vemos que el 'sujeto' en la imagen de prueba coincide con el observado en la imagen del vecino más cercano, tanto en la identidad del personaje como en la expresión facial, lo que resalta la eficacia y la validez de nuestro método de proyección y clasificación.

Figura 4.4. Imagen de prueba con su vecino más cercano correspondiente; valor $p = 25$.

Al variar el número de componentes principales generadas (p) y obtener el vecino más cercano correspondiente, se nota una mejora en la claridad de la imagen proyectada a medida que aumenta el valor de p. Esto se traduce en una aproximación más precisa del vecino más cercano. En ambos casos (**Figuras 4.5** y **4.6**), la imagen del vecino más cercano se asemeja en mayor medida a la imagen de prueba en cuanto a las facciones y expresión facial del sujeto, en comparación con el vecino más cercano obtenido con un valor de $p = 25$ (**Figura 4.4**).

Figura 4.5. Imagen de prueba con su vecino más cercano correspondiente; valor $p = 50$.

Figura 4.6. Imagen de prueba con su vecino más cercano correspondiente; valor $p = 200$.

4.4. Solución

Para prevenir casos como el mostrado en la Figura 4, donde la imagen proyectada es indistinguible y el 'sujeto' en la imagen de prueba no coincide con el observado en la imagen del vecino más cercano, se podrían considerar distintas estrategias:

- Ampliar el conjunto de datos de entrenamiento con una variedad más amplia de imágenes. Esto ayudaría a mejorar la representatividad y eficacia del modelo, reduciendo así la probabilidad de discrepancias entre la imagen de prueba y sus vecinos más cercanos.
- Utilizar técnicas que nos permitan evaluar el rendimiento del modelo con diferentes números de componentes principales. De esta maner,a podríamos encontrar el número de componentes principales que maximice la eficacia del modelo.

5. Problema 5

Dado un conjunto de datos centrados X_{nxd}, vimos que hacer PCA, es realizar la descomposición espectral de la matriz de covarianzas muestral, que puede estimarse como $S = X'X$ (omitimos el coeficiente $n-1$).

Ahora, considera la matriz $K_{nxn} = XX'$.

a) Muestra que es equivalente realizar PCA en S o en K, es decir, que $(\lambda^{-1/2X}u, \lambda)$ es un par eigenvector-eigenvalor normalizado de K, y a su vez, $(\lambda^{-1/2}X^T v, \lambda)$ es un par eigenvector-eigenvalor normalizado de S, donde u y v son vectores propios de S y K, respectivamente.

5.1. Solución

De manera formal, se definen los vectores y valores propios de la siguiente manera:

Sea $A : V \to V$ un operador lineal en un cierto k-espacio vectorial V y v un vector no nulo en V. Si existe un escalar c tal que

$$Av = cv, \qquad v \neq 0,\ c \in K \tag{5.1}$$

En este caso consideramos S y K.

Sea S

$$S_{nxn} = X'X$$

Entonces

$$Su = \lambda u$$

$$(X'X)u = \lambda u$$

$$X(X'X)u = X\lambda u$$

$$(XX')Xu = \lambda(Xu)$$

$$K(Xu) = \lambda(Xu)$$

Sea K

$$K_{nxn} = XX'$$

Entonces

$$Kv = \lambda v$$

$$(XX')v = \lambda v$$

$$X'(XX')v = X'\lambda v$$

$$(X'X)X'v = \lambda(X'v)$$

$$S(X'v) = \lambda(X'v)$$

De esta manera encontramos que los valores propios tanto de S como de K corresponden al valor λ, con vectores propios Xu y $X'v$, respectivamente.

Al calcular la norma de los vectores propios tenemos que

$$Su = \lambda u$$

$$X'Xu = \lambda u$$

$$u'X'Xu = u'\lambda u$$

Dado que u' y u son vectores ortogonales, entonces

$$u'u = 1$$

Por lo tanto

$$u'X'Xu = \lambda u'u$$

$$u'X'Xu = \lambda$$

$$(u'X')(Xu) = \lambda$$

$$(Xu)'(Xu) = \lambda$$

$$||Xu||^2 = \lambda$$

$$||Xu|| = \sqrt{\lambda}$$

norma del vector propio de S.

$$Kv = \lambda v$$

$$XX'v = \lambda v$$

$$v'XX'v = v'\lambda v$$

Dado que v' y v son vectores ortogonales, entonces

$$v'v = 1$$

Por lo tanto

$$v'X'Xv = \lambda v'v$$

$$v'X'Xv = \lambda$$

$$(v'X)(X'v) = \lambda$$

$$(X'v)'(X'v) = \lambda$$

$$||X'v||^2 = \lambda$$

$$||X'v|| = \sqrt{\lambda}$$

norma del vector propio de K.

Tenemos que

$$||Xu|| = ||X'v|| = \sqrt{\lambda}$$

De manera que para K, $(\lambda^{\frac{1}{2}} Xu, \lambda)$ es un par eigenvector-eigenvalor normalizado

$$\frac{Xu}{||Xu||} = \frac{Xu}{\sqrt{\lambda}} = \lambda^{\frac{1}{2}} Xu$$

Y para S, $(\lambda^{\frac{1}{2}} X'v, \lambda)$ es un par eigenvector-eigenvalor normalizado

$$\frac{X'v}{||X'v||} = \frac{X'v}{\sqrt{\lambda}} = \lambda^{\frac{1}{2}} X'v$$

CONAPO
CONSEJO NACIONAL DE POBLACIÓN

Reformulación del Índice de Marginación en México: Propuesta basada en Análisis de Componentes Principales

Febrero de 2024

Resumen

El objetivo de este informe es presentar los resultados de un análisis de componentes principales aplicado a las variables asociadas al índice de marginación propuesto por el Consejo Nacional de Población (CONAPO), según su nota técnico-metodológica publicada en diciembre del año 2021. Este índice se basa en el Censo de Población y Vivienda del año 2020 realizado por el Instituto de Nacional de Estadística y Geografía (INEGI). Se analizan diversos indicadores socioeconómicos relacionados al índice de marginación por entidad federativa y municipio en México, y se propone un nuevo índice de marginación basado en las componentes principales obtenidas del análisis de componentes principales para evaluar la situación de marginación en el país.

Análisis de Componentes Principales

El análisis de componentes principales (PCA) es una técnica utilizada para reducir el número de variables de un conjunto de datos, generando nuevas variables que expliquen la variabilidad contenida en el conjunto de datos minimizando la pérdida de información. En otras palabras, nos permite simplificar conjuntos de datos complejos para comprender mejor las tendencias y patrones que se ocultan en ellos. Esto se logra por medio de la generación de un conjunto más reducido de variables, conocidas como componentes principales. Es importante recalcar que existe un número óptimo de componentes principales, el cual evita la pérdida de información crucial o la generación de datos excesivamente complicados e innecesarios.

Índice de Marginación

El índice de marginación (IM) es una métrica que resume las diferencias entre municipios y estados del país, tomando en cuenta las necesidades básicas que tienen o carecen las distintas entidades, como acceso a la eduación, servicios básicos de vivienda, ingresos monetarios y tamaño poblacional. El CONAPO considera la información del censo realizado por el INEGI para construir este índice.

Considerando lo anterior, se aplicó la técnica de componentes principales para la estimación del índice de marginación utilizando las estimaciones de los indicadores socioeconómicos utilizados por el CONAPO basados en el Censo de Población y Vivienda 2020.

Resultados

Índice de Marginación basado en el Censo de Población y Vivienda 2020

El índice de marginación por entidad federativa y municipio propuesto por el CONAPO considera un total de nueve variables asociadas con la marginación en los rubros de educación, vivienda, distribución de la población e ingresos monetarios.

Se observa que los municipios con las peores condiciones sociales y económicas se encuentran principalmente en los estados de Guerrero, Oaxaca, Chiapas, Durango y Nayarit. En términos generales, los municipios ubicados al sur del país se encuentran principalmente en las categorías 'Muy alto' y 'Alto' de marginación, mientras que una gran proporción de los municipios localizados en el centro y norte del país se encuentran dentro de las categorías 'Muy bajo' y 'Bajo' (**Figura 1**).

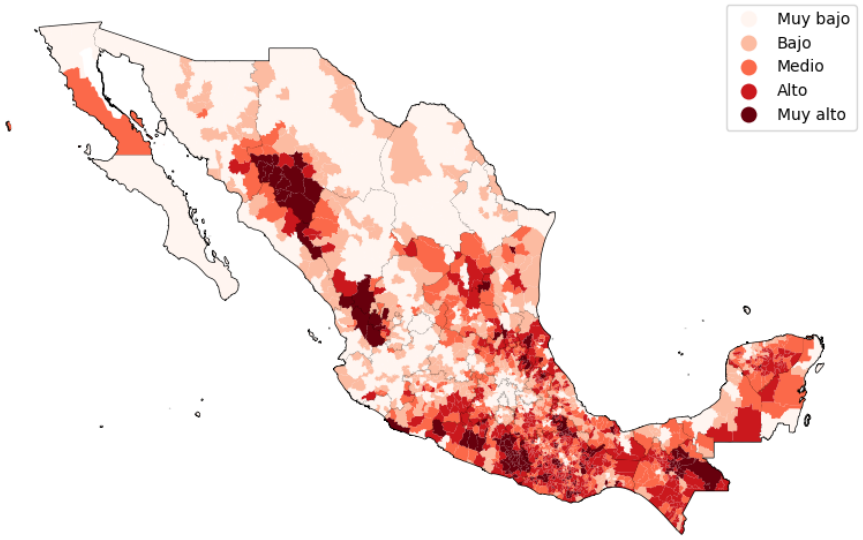

Figura 1. Grado de marginación por entidad federativa y municipio, 2020.

Índices de Marginación basados en los componentes obtenidos del Análisis de Componentes Principales

Basándonos en las nuevas variables obtenidas del análisis de PCA, podemos afirmar que este enfoque proporciona una aproximación sólida al índice propuesto por la CONAPO. Los mapas generados representan el índice de marginación en una escala númerica de 0 a 1, donde los municipios con valores cercanos a 1,0 reflejan zonas con un alto grado de marginación, en contraste con aquellos cercanos a 0,0, que representan zonas con niveles más bajos de marginación.

Al considerar el primer 'componente' del nuevo conjunto de variables para la estimación de la categorización númerica, se observa con claridad que el mapa conserva la estructura general: estados como Guerrero, Oaxaca, Chiapas muestran las peores condiciones socioeconómicas, mientras que se observa una clara disparidad entre los municipios del sur y del norte del país. Donde los del sur presentan valores superiores a 0,5, lo que equivale a categorías de marginación de 'Medio', 'Alto' o 'Muy alto', mientras que los del norte muestran valores inferiores a 0,4, correspondientes a categorías de 'Bajo' o 'Muy bajo' en el índice de marginación (**Figura 2**). Es decir, esta nueva variable conserva la información relevante de cada uno de los indicadores socioeconómicos considerados por la CONAPO en su nota técnico-metodológica.

Por otro lado, al considerar la segunda variable de nuestro conjunto reducido de variables, podemos observar que esta sigue manteniendo cierta información, sin embargo solamente nos ayuda a distinguir los municipios con un nivel de marginación extremadamente alto, mientras que la distinción entre los restantes se pierde en su totalidad. Esto debido a que podemos observar que se dificulta observar la disparidad entre los municipios del sur (**Figura 3**).

Se omiten los mapas generados basados en las variables posteriores debido que se observó una pérdida significativa de información, demostrando que las primeras componentes son las que preservan la mayor parte de información contenida en el conjunto de datos original.

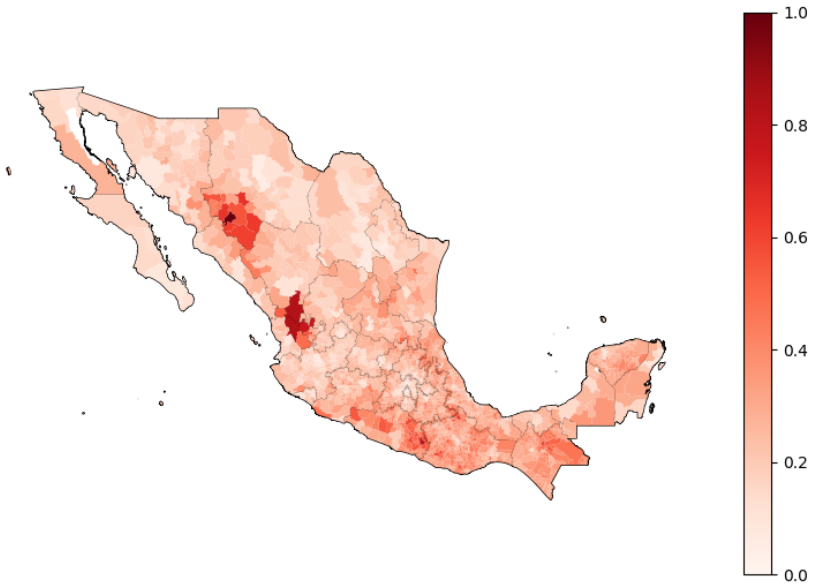

Figura 2. Índice de marginación por entidad federativa y municipio basado en el primer componente principal.

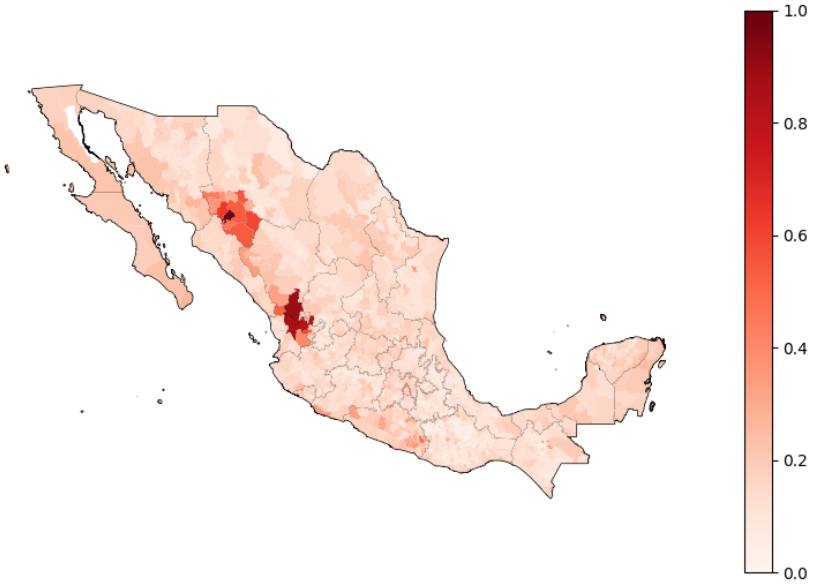

Figura 3. Índice de marginación por entidad federativa y municipio basado en el segundo componente principal.

Conclusiones

El análisis de componentes principales ofrece un enfoque alternativo para la categorización de los niveles de marginación en nuestro país. Esta técnica proporciona una perspectiva amplia y flexible al permitir la inclusión de indicadores adicionales, más allá de los utilizados por el CONAPO, dando como resultado una reducción del número de variables sin perder información relevante de los datos. Esta metodología nos brinda la oportunidad de capturar de manera más completa y precisa la complejidad y diversidad de factores que influyen en los niveles de marginación en los distintos municipios del país.

Recomendaciones

Se sugiere realizar análisis comparativos adicionales para validar la efectividad del índice propuesto por medio del análisis de componentes principales (PCA) en la medición de la marginación en diferentes contextos y regiones de México. Esto implica llevar a cabo estudios que contrasten los resultados obtenidos mediante el método de PCA con aquellos derivados de otros enfoques más tradicionales de categorización de los niveles de marginación. Además, se recomienda realizar futuros análisis de sensibilidad para evaluar cómo pequeños cambios en la selección de variables pueden afectar en los resultados. Estas acciones proporcionarán una mayor confianza en la utilidad y robustez del nuevo índice, todo esto con el fin de proponer estrategias de desarrollo más efectivas y precisas a nivel regional en México.