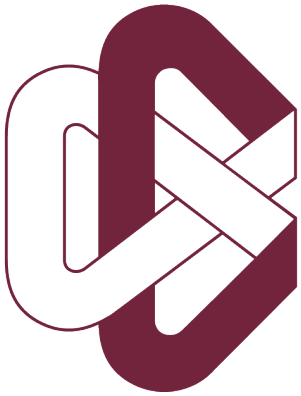

CIMAT

Centro de Investigación en Matemáticas, A.C.

Estadística Multivariada

"Análisis de la calidad de vida a través de diferentes herramientas estadísticas".

Alumnas:

Alondra Elizabeth Matos Mendoza
María Guadalupe López Salomón

01 de Junio de 2023

Índice

1. Resumen

El objetivo de este proyecto es analizar y predecir el índice de felicidad para determinar la calidad de vida, utilizando técnicas estadísticas avanzadas. Se aplicaron regresión multivariada, escalamiento multidimensional (MDS) y k-means a un conjunto de datos que contenía variables relacionadas con la calidad de vida en diferentes países.

En la introducción, se destaca la importancia de comprender y predecir la calidad de vida, así como el papel de las técnicas estadísticas avanzadas en este campo. Se menciona que la Regresión Multivariada permite identificar y cuantificar la influencia relativa de diferentes factores en la calidad de vida, mientras que el Escalamiento Multidimensional captura la complejidad y diversidad de la calidad de vida de manera más precisa y completa.

En los antecedentes, se señala que estudios anteriores han utilizado la Regresión multivariada y el Escalamiento Multidimensional para analizar la calidad de vida en diferentes contextos, y se resalta la importancia de adoptar enfoques multidimensionales en lugar de enfoques unidimensionales tradicionales.

En la metodología, se describe el proceso de ajuste del modelo de Regresión Multivariada y se explican las pruebas realizadas para verificar la adecuación del modelo. También se detalla el uso del Escalamiento Multidimensional para obtener una representación en dos dimensiones de los países y la posterior aplicación de K-means para agrupar los países según su similitud en la calidad de vida.

En los resultados, se muestran los hallazgos obtenidos mediante la Regresión Multivariada. Se informa que el modelo ajustado explica el 77.9 % de la variabilidad del índice de felicidad y que la precisión del modelo es aceptable. Además, se presentan los coeficientes del modelo y se realiza una Prueba de la Razón de Verosimilitud para evaluar la relevancia de ciertas variables predictoras. Se concluye que las variables de generosidad y percepciones de corrupción no tenían un efecto significativo en el índice de felicidad.

Finalmente, se discute la adecuación del modelo ajustado, se exhiben algunas gráficas de residuos y se verifican los supuestos del modelo de regresión. Se concluye que el modelo es válido y que las estimaciones de los coeficientes son confiables.

En resumen, este proyecto aplicó técnicas estadísticas avanzadas para analizar y predecir el índice de felicidad para determinar la calidad de vida. Los resultados proporcionaron información valiosa sobre los factores que influyen en la calidad de vida y pueden ser utilizados para informar políticas públicas y programas de desarrollo social.

2. Introducción

Existen diferentes enfoques para definir la calidad de vida, la cual es un concepto multidimensional que puede abarcar diversos aspectos que influyen en el bienestar y la satisfacción de los seres humanos. A medida que la sociedad avanza y se enfrenta a nuevos retos, se vuelve imperativo contar con herramientas y enfoques que permitan un análisis exhaustivo y preciso de la calidad de vida. Aunque el tipo de medidas de calidad de vida apropiadas para su uso dependen del campo de estudio y la pregunta de investigación que se aborda, se ha demostrado que el uso de una estructura multidimensional es ventajoso al medir y predecir la calidad de vida. La comprensión y predicción de la calidad de vida se han convertido en áreas de interés clave en la investigación social y el diseño de políticas públicas. En este contexto, el uso de técnicas estadísticas avanzadas como la regresión multivariada y el escalamiento multidimensional ha demostrado ser prometedor para analizar y predecir la calidad de vida de manera más precisa y completa.

La regresión multivariada es una técnica estadística que permite examinar las relaciones entre múltiples variables independientes y una variable dependiente. En el contexto del análisis de la calidad de vida, la regresión multivariada puede utilizarse para identificar y cuantificar la influencia relativa de diferentes factores en la calidad de vida de las personas. Variables como el PIB per cápita, el apoyo social percibido, la esperanza de vida saludable, la libertad para tomar decisiones vitales, la generosidad y las percepciones de

corrupción pueden ser consideradas como predictores potenciales de la calidad de vida. Mediante el análisis de regresión multivariada, se puede determinar cómo estas variables se relacionan y contribuyen al bienestar general de las personas.

El escalamiento multidimensional se fundamenta en la idea de que las diferentes dimensiones de la calidad de vida son interdependientes y se influyen mutuamente. Estas dimensiones pueden incluir aspectos como la salud física y mental, el nivel de ingresos, la libre capacidad de tomar decisiones, el acceso a la educación, la vivienda, percecpción de la corrupción, país de origen, el entorno social y cultural, entre otros. Al considerar estas múltiples dimensiones, el escalamiento dimensional *(MDS)* permite capturar la complejidad y la diversidad de la calidad de vida de manera más precisa y completa.

A través de este análisis, que combina la regresión multivariada y el escalamiento multidimensional, se espera contribuir al avance del conocimiento en el campo de la calidad de vida, así como proporcionar información valiosa para la toma de decisiones informadas. La regresión multivariada permitirá identificar y cuantificar la influencia relativa de las variables consideradas, como el PIB per cápita, el apoyo social percibido, la esperanza de vida saludable, la libertad para tomar decisiones vitales, la generosidad y las percepciones de corrupción, en la calidad de vida de las personas. Por otro lado, el escalamiento multidimensional se presenta como una herramienta eficiente y poderosa en el análisis de la calidad de vida, proporcionando una visión detallada de los factores que influyen en el bienestar de las personas. Al combinar ambas técnicas, se espera obtener una comprensión más completa y precisa de la calidad de vida, lo que puede contribuir al desarrollo de enfoques más efectivos para evaluar y mejorar la calidad de vida en diversos contextos. Los hallazgos de este estudio pueden ser de gran utilidad para informar políticas públicas y programas de desarrollo social, al proporcionar evidencia empírica sobre los factores que impactan significativamente en la calidad de vida y orientar la asignación de recursos hacia áreas prioritarias.

3. Antecedentes

A través de los años se han llevado a cabo estudios que han utilizado tanto la regresión multivariada como el escalamiento multidimensional para analizar la calidad de vida en diferentes contextos. Estos estudios han demostrado la utilidad y eficacia de combinar ambas técnicas en el análisis de la calidad de vida, permitiendo una comprensión más profunda de los factores que influyen en el bienestar de las personas.

Tradicionalmente, se han utilizado enfoques unidimensionales para medir la calidad de vida, como el uso exclusivo del PIB per cápita. Sin embargo, se ha reconocido la necesidad de adoptar enfoques multidimensionales que consideren diversas variables y dimensiones para capturar de manera más completa y precisa el concepto de calidad de vida.

La regresión multivariada es una técnica estadística ampliamente utilizada para analizar las relaciones entre múltiples variables independientes y una variable dependiente. En el contexto de este proyecto, se presenta como una herramienta útil para identificar los predictores más influyentes y cuantificar su impacto relativo en el bienestar general de las personas.

Por otro lado, el escalamiento multidimensional (MDS) se introduce como un método gráfico que complementa los métodos descriptivos en la validación y análisis de información y datos sobre la calidad de vida. Proporciona un enfoque visual y estructural para comprender las relaciones y patrones subyacentes en un conjunto de datos multidimensionales.

En el ámbito de la investigación de la calidad de vida, se ha aplicado el escalamiento multidimensional como una técnica estadística que permite representar de manera gráfica las dimensiones y estructuras latentes, brindando una visión más detallada y comprensiva del fenómeno.

En resumen, los estudios previos han evidenciado la importancia de combinar la regresión multivariada y el escalamiento multidimensional en el análisis de la calidad de vida. Estas técnicas han permitido una comprensión más profunda de los factores influyentes en el bienestar de las personas y han superado las limitaciones de los enfoques unidimensionales tradicionales. Su aplicación ha contribuido al desarrollo de

enfoques más completos y precisos para evaluar y mejorar la calidad de vida en diversos contextos.

4. Planteamiento del problema

El análisis de la calidad de vida es un desafío complejo debido a su naturaleza multidimensional y a la interacción de diversos factores que influyen en el bienestar de las personas. Aunque existen enfoques tradicionales, como el uso del PIB per cápita como indicador único, estos enfoques unidimensionales no capturan la complejidad y diversidad de la calidad de vida de manera adecuada. Además, la consideración de múltiples variables, como el PIB per cápita, el apoyo social percibido, la esperanza de vida saludable, la libertad para tomar decisiones vitales, la generosidad y las percepciones de corrupción, plantea desafíos en términos de integración y ponderación de estas dimensiones.

El problema a resolver radica en la necesidad de desarrollar un enfoque metodológico que permita analizar y predecir de manera precisa y completa la calidad de vida, considerando la interacción entre múltiples variables y superando las limitaciones de los enfoques unidimensionales tradicionales. Es fundamental encontrar un equilibrio entre la consideración de las diferentes dimensiones y la identificación de las variables más relevantes para comprender y mejorar la calidad de vida en diversos contextos.

Además, se requiere abordar los desafíos asociados a la integración de datos y la reducción de dimensionalidad en el escalamiento multidimensional. En este contexto, el objetivo principal de este proyecto es desarrollar un enfoque que combine la regresión multivariada y el escalamiento multidimensional para analizar y predecir de manera más precisa y completa la calidad de vida. Se busca superar las limitaciones de los enfoques tradicionales, integrando múltiples dimensiones y considerando las interacciones entre las variables clave mencionadas. Asimismo, se pretende proporcionar herramientas y conocimientos que sean aplicables en diferentes contextos y que puedan respaldar la toma de decisiones informadas en políticas públicas y programas de desarrollo social.

De manera que, el problema a resolver se centra en desarrollar un enfoque metodológico que permita analizar y predecir de manera precisa y completa la calidad de vida, considerando la interacción entre múltiples variables. Este enfoque busca superar las limitaciones de los enfoques unidimensionales tradicionales y proporcionar una base sólida para la toma de decisiones en políticas públicas y programas de desarrollo social.

5. Pregunta de investigación

¿Cómo se puede analizar y predecir de manera precisa y completa la calidad de vida, considerando múltiples variables como el PIB per cápita, el apoyo social percibido, la esperanza de vida saludable, la libertad para tomar decisiones vitales, la generosidad y las percepciones de corrupción, mediante la combinación de la regresión multivariada y el escalamiento multidimensional?

6. Objetivos

- Analizar la calidad de vida utilizando un enfoque multidimensional que incorpore variables como el PIB per cápita (GDP), el apoyo social percibido, la esperanza de vida saludable, la libertad para tomar decisiones vitales, la generosidad y las proporciones de corrupción.
- Identificar y cuantificar la influencia relativa de cada una de las variables consideradas en la calidad de vida de las personas a través del análisis de regresión multivariada.
- Explorar el potencial del escalamiento multidimensional clásico (MDS) en el análisis de la calidad de vida.

- Identificar los desafíos y limitaciones asociados con estas técnicas, así como proponer posibles estrategias para abordarlos.

7. Marco Teórico

A continuación se presentan las técnicas estadísticas implementadas en la metodología de este proyecto.

7.1. Regresión multivariada

Como hemos visto, el análisis de regresión es la metodología estadística para predecir valores de una o más variables de respuesta (dependientes) a partir de una colección de valores de variables predictoras (independientes). También se puede usar para valorar el efecto que tienen las variables predictoras (independientes) sobre las respuestas, es decir, las variables que tienen mayor influencia en la respuesta.

El modelo de regresión multivariado está dado por:

$$\mathbf{Y}_{n\times m} = \mathbf{Z}_{(n\times(r+1))}\boldsymbol{\beta}_{((r+1)\times m)} + \boldsymbol{\varepsilon}_{(n\times m)} \tag{1}$$

$$\boldsymbol{E(\varepsilon) = 0_{(n\times 1)}, Cov(\varepsilon) = \sigma^2_{(n\times n)} I} \tag{2}$$

donde se modelan las m respuestas $Y_1, ..., Y_m$ y un conjunto de r variables predictoras $z_1, ..., z_r$ que forman la matriz de diseño $\mathbf{Z}$, donde además $\boldsymbol{\beta}$ y σ^2 son parámetros desconocidos.

Se asume que cada una de las respuestas sigue su propio modelo de regresión, es decir:

$$Y_1 = B_{01} + B_{11}z_1 + B_{21}z_2 + ... + B_{r1}z_r + \varepsilon_1 \tag{3}$$
$$Y_2 = B_{02} + B_{12}z_1 + B_{22}z_2 + ... + B_{r2}z_r + \varepsilon_2 \tag{4}$$
$$\tag{5}$$

$$Y_m = B_{0m} + B_{1m}z_1 + B_{2m}z_2 + ... + B_{rm}z_r + \varepsilon_m \tag{6}$$

y donde se asume que los términos de error tienen las siguientes propiedades:

1. $\boldsymbol{E(\varepsilon) = 0}$
2. $\boldsymbol{Cov(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I}$

Dados los valores de la matriz respuesta $\mathbf{Y}$ y los valores de las variables predictoras $\mathbf{Z}$, determinamos los estimadores de mínimos cuadrados $\hat{\beta}_{(i)}$ exclusivamente de las observaciones de la i-ésima respuesta, $\mathbf{Y}_i$. Entonces

$$\hat{\beta}_{(i)} = (Z'Z)^{-1}Z'Y_{(i)}$$

coleccionando los estimadores univariados de mínimos cuadrados obtenemos $\hat{\beta} = \boldsymbol{(Z'Z)^{-1}Z'Y}$. Sea $\hat{y} = \boldsymbol{Z\hat{\beta}}$ entonces la matriz de residuales es de la forma: $\boldsymbol{\hat{\varepsilon} = y - \hat{y}}$.

Donde se cumple que $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ y $\boldsymbol{Cov(\hat{\beta}) = \sigma^2(Z'Z)^{-1}}$. Los residuales $\hat{\varepsilon}$ tiene las propiedades: $E(\hat{\varepsilon}) = 0$ y $Cov(\hat{\varepsilon}) = \sigma^2\left[I - Z(Z'Z)^{-1}Z'\right]$. Más aún $\hat{\beta}$ y $\hat{\varepsilon}$ no están correlacionados.

7.1.1. Inferencia sobre el Modelo de Regresión

Antes de que podamos evaluar la importancia de variables particulares en la función de regresión $E(Y) = \beta_0 + \beta_{1z_1} + ... + \beta_r z_r$ debemos determinar las distribuciones muestrales de $\hat{\beta}$ y la suma cuadrada de residuales $\hat{\varepsilon}'\hat{\varepsilon}$. Para hacer lo anterior, asumiremos que los errores ε tienen una distribución normal.

Sea $\boldsymbol{Y = Z\beta + \varepsilon}$, donde $\boldsymbol{Z}$ tiene rango completo $r+1$ y ε se distribuye $\boldsymbol{N_n(0, \sigma^2 I)}$. Entonces el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ es el mismo que el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$, más aún:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(Z'Z)^{-1}Z'Y}$$

se distribuye $\boldsymbol{N_{r+1}(\beta, \sigma^2(Z'Z)^{-1})}$ y está distribuida de manera independiente de los residuales $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y - Z\beta}$. Más aún, $\boldsymbol{n\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}}$ se distribuye $\boldsymbol{\sigma^2\chi^2_{n-r-1}}$, donde $\hat{\sigma}^2$ es el estimador de máxima verosimilitud de σ^2.

Intervalos de confianza simultáneos.
Sea $Y = Z\beta + \epsilon$, donde Z tiene rango completo $r+1$ y $\epsilon \sim N_n(0, \sigma^2 I)$. Entonces los intervalos de confianza simultáneos del $100(1-\alpha)\,\%$ para las β_i están dados por:

$$\widehat{\beta}_i \pm \sqrt{\widehat{Var}(\widehat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1(\alpha)}}, \quad i = 0, 1, ..., r$$

donde $\widehat{Var}(\widehat{\beta}_i)$ es el i-ésimo elemento de la diagonal de $s^2(Z'Z)^{-1}$ y $s^2 = \frac{\widehat{\epsilon}\widehat{\epsilon}}{n-(r+1)} = \frac{y'[I-Z(Z'Z)^{-1}Z']y}{n-r-1} = \frac{y'[I-H]y}{n-r-1}$.

Intervalos de confianza individuales

Los intervalos de confianza del $100(1-\alpha)\,\%$ para cada β_i obtenidos de manera univariada son:

$$\widehat{\beta}_i \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{\widehat{Var}(\widehat{\beta}_i)} \quad i = 0, 1, ..., r$$

Pruebas de razón de verosimilitud para los parámetros de regresión.

Parte del análisis de regresión se ocupa de evaluar los efectos de las variables predictoras particulares sobre la variable de respuesta. Una hipótesis nula de interés establece que algunos de los Z_i no influyen en la respuesta Y. Estos predictores se etiquetarán como $z_{q+1}, z_{q+2}, ..., z_r$. la afirmación de que $z_{q+1}, z_{q+2}, ..., z_r$ no influyen en Y se traduce en la hipótesis estadística:

$$H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_r = 0 \tag{7}$$

o

$$H_0 : \beta_{(2)} = 0 \tag{8}$$

donde $\beta_2' = [\beta_{q+1}, \beta_{q+2}, ..., \beta_r]$

Bajo la hipótesis nula $H_0 : \beta_{(2)} = 0, Y = Z_1\beta_{(1)} + \varepsilon$. La prueba de razón de verosimilitud de H_0 está basada en la suma adicional de cuadrados

$$SS_{res}(Z_1) - SS_{res}(Z) = \boldsymbol{(y - Z_1\hat{\beta}_{(1)})'(y - Z_1\hat{\beta}_{(1)}) - (y - Z\hat{\beta})'(y - Z\hat{\beta})} \tag{9}$$

donde $\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}} = \boldsymbol{(Z_1'Z_1)^{-1}Z_1'y}$.

Sea $\mathbf{Z}$ con rango completo $r+1$ y ε distribuido como $N_n(0, \sigma^2 I)$. La prueba de la razón de verosimilitud de $H_0 : \boldsymbol{\beta_{(2)} = 0}$ es equivalente a una prueba f H_0 basada en la suma adicional de cuadrados de (9) y $s^2 = \boldsymbol{(y - Z\hat{\beta})'(y - Z\hat{\beta})}/(n - r - 1)$. En particular, la prueba de razón de verosimilitud rechaza H_0 si:

$$\frac{SS_{res}(Z_1) - SS_{res}(Z))\,/(r-q)}{s^2} > F_{r-q,n-r-1}(\alpha) \tag{10}$$

donde $F_{r-q,n-r-1}(\alpha)$ es el (100α)ésimo percentil superior de una distribución F- con $r-q$ y $n-r-1$ grados de libertad

7.2. MDS: Modelos de Escalamiento multidimensional

El escalamiento multidimensional se define como una técnica de análisis multidimensional cuyo objetivo es representar las *proximidades* entre un conjunto de objetos o estímulos como distancias en un espacio de baja dimensión (generalmente de dos o tres dimensiones). Dicha técnica tiene como objetivos:

- Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales con la mínima pérdida de información.
- Encontrar grupos en los datos, si existen.
- Clasificar nuevas observaciones en grupos definidos y relacionar conjuntos de variables.
- Encontrar posibles datos atípicos (outliers).

De manera general, el término de proximidad indica cercanía en el espacio, tiempo o cualquier otro contexto. Matemáticamente, ese término hace referencia al concepto de disimilaridad o similaridad.

7.2.1. Modelo Clásico de MDS.

El Modelo Clásico de Escalamiento multidimensional, también conocido como *MDS métrico*, asume que las relaciones de distancia entre los objetos son conocidas y se expresan en una matriz de proximidad. Esta matriz de proximidad contiene las medidas de similitud o distancia entre todos los pares de objetos en el conjunto de datos. De manera formal, tenemos que:

Teorema de MDS clásico.

Sea $\boldsymbol{\Delta}_{(n\times n)} = \{\delta_{ij}\}$ una matriz de distancias entre n puntos en un espacio de configuración de dimensión K y sea $\boldsymbol{B}_{n\times n}$ la matriz dada por $\mathbf{B} = \mathbf{HAH}$, siendo $\mathbf{H}_{n\times n}$ dada por $\mathbf{H} = \mathbf{I} - \mathbf{n}^{-1}\mathbf{1}\mathbf{1}^{\boldsymbol{t}}$ y $\boldsymbol{A_{n\times n}}$ la matriz cuyos elementos vienen dados a través de $a_{rs} = -\frac{1}{2}\delta_{rs}^2$. Entonces, Δ es una matriz de distancias Euclideanas *sii* $\mathbf{B}$ es semidefinida positiva. Además se tiene:

- Si Δ es la matriz de distancias Euclideanas para una configuración dada por $\boldsymbol{Z_{(n\times K)}} = \boldsymbol{(z_1, ..., z_n)^t}$, entonces $\mathbf{B}$ se puede representar como $\boldsymbol{B} = \boldsymbol{(HZ)(HZ)^t}$, es decir,

  $$b_{ij} = \boldsymbol{(z_i - \bar{z})^t(z_j - \bar{z})}\forall i, j = 1, ..., n.$$

  de donde $\boldsymbol{B} \geq 0$. $\mathbf{B}$ será la matriz centrada de productos escalares de $\mathbf{Z}$.
- Inversamente, si $\mathbf{B}$ es positiva semidefinida de rango K, entonces se puede construir una configuración $\boldsymbol{X}$ asociada a $\boldsymbol{B}$ de la siguiente forma:

Sean $\lambda_1 >, ..., > \lambda_K$ los K valores propios positivos de $\mathbf{B}$ y sus vectores propios asociados $\boldsymbol{e_1, e_2, ..., e_K}$ se organizan como columnas formando $\boldsymbol{X_{(n\times K)}} = \left(x_{(1)}, ..., x_{(K)}\right)$ normalizados según la condición

$$\boldsymbol{X'_{(i)}x_{(i)}} = \lambda_i, \forall i = 1, ...K.$$

Entonces $\boldsymbol{d(x_i, x_j) = \delta_{ij}, \forall x_i, x_j} \in \mathbb{R}^K$. Además esa configuración está centrada en $\bar{\boldsymbol{x}} = \mathbf{0}$ y $\mathbf{B}$ es la matriz de productos escalares de esa configuración, es decir $\mathbf{B = XX'}$.

El objetivo real de MDS es que dado $O = \{o_1, ..., o_n\}$ y $\Delta = \{\delta_{ij}\}$ entre ellos, construir una configuración $X = (x'_1, ..., x'_n)$, $x_i \in \mathbb{R}^M, M \ll n$, tal que $d(x_i, x_j) \approx \delta_{ij}$.

- Al igual que en el caso de componentes principales, si se asume que los primeros m valores propios de $\mathbf{B}$ son positivos y muy grandes y los restantes $K - m$ valores propios son mucho menores, podemos obtener una representación aproximada de las disimilaridades utilizando solo los m valores propios positivos más grandes y sus vectores propios asociados.
- El procedimiento para obtener la solución clásica de MDS mediante los vectores y valores propios de $\mathbf{B}$ se conoce como *coordenadas principales*.

7.2.2. Construcción de la configuración MDS mediante coordenadas principales

Supongamos que tenemos una matriz de distancias o disimilaridades al cuadrado Δ. El procedimiento para obtener las coordenadas principales es el siguiente:

- Se construye la matriz $\boldsymbol{B = -\frac{1}{2}H\Delta H}$, de productos cruzados, donde $\boldsymbol{H = I - n^{-1}11^t}$.
- Se obtienen los valores propios de $\mathbf{B}$, tomándose los m valores propios positivos más grandes, de tal forma que los restantes $K - m$ valores propios sean cercanos a cero.
- Si $\mathbf{B}$ no es positiva semidefinida, se podrían ignorar los valores propios negativos y tomar solo los m valores propios positivos más grandes y seguir el procedimiento.
- Considerando los m valores propios positivos y sus vectores propios asociados, por la descomposición espectral podemos aproximar la matriz $\mathbf{B}$ por:

  $$\boldsymbol{B \approx U_m\Lambda_m U'_m = (U_m\Lambda^{1/2})(\Lambda^{1/2}U'_m) = (U_m\Lambda_m^{1/2})(U_m\Lambda_m^{1/2})' = X_m X'_m}$$

  entonces las coordenadas principales de los puntos están dadas por:

  $$\boldsymbol{X_m = (U_m\Lambda_m^{1/2}) = \sqrt{\lambda_1}e_1|...|\sqrt{\lambda_m}e_m\Big]}$$

  Al igual que en componentes principales se puede elegir un número apropiado de dimensiones m, usando el criterio de proporción de varianza explicada por las primeras m dimensiones, dado por:

  $$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}|\lambda_i|}.$$

7.3. Clustering: K-means

Clustering es el proceso de examinar una colección de puntos y agrupar los puntos en *clusters* de acuerdo a alguna medida de distancia. El ánalisis de clusters construye buenas agrupaciones cuando los miembros de un cluster tienen un alto grado de similaridad entre sí y no son como miembros de otros clusters. En este proyecto, se usó a los algoritmos de asignación de puntos. La conocida familia de algoritmos de agrupamiento de este tipo se conoce como k-means (Hartigan Wong, 1979).

7.3.1. Algoritmos K-mean.

Clustering *K-means* es un método de análisis de cluster no jerárquico que busca particionar n observaciones en k clusters en los que cada observación pertenezca al cluster con la media más cercana. El número de clusters es fijo. El método *k-means* puede agrupar grandes cantidades de datos rápida y eficientemente. En general, la distancia empleada es la euclídea. Dado un conjunto de observaciones $(x_1, x_2, ...x_n,)$ donde cada observación es un vector real d-dimensional, k-means busca particionar las n observaciones en $k(\leq n)$ conjuntos $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ para minimizar la suma de cuadrados dentro del grupo (WCSS), es decir, la varianza. Formalmente el objetivo es encontrar:

$$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} ||\boldsymbol{x - \mu_i}||^{\boldsymbol{2}} = \boldsymbol{argmin_S} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} |S_i| Var S_i \tag{11}$$

donde μ_i es la media o también llamado centroide de los puntos en S_i, $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x_i$.

$|S_i|$ es el tamaño de S_i y $|| * ||$ es la norma L^2 usual.

El algoritmo de *K-means* sigue los siguientes pasos:

1. Se toman al azar k clusters iniciales.
2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clusters y se reasignan a los que estén más próximos. Se vuelvean a recalcular los centroides de los k clusters después de las reasignaciones de los elementos.
3. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

Usualmente, se espcifican k centroides iniciales y se procede al segundo paso y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

En particular, con respecto al Escalamiento Multidimensional (MDS) el Análisis de Cluster, comparte con ellas las siguientes características: investigan la estructura de un conjunto de variables, el punto de partida es una matriz de proximidades y en la representación gráfica que se obtiene se pueden interpretar las distancias.

8. Metodología

8.1. Regresión lineal

Para predecir el índice de felicidad, se ajustó un modelo de regresión lineal. Posteriormente, se verificó la adecuación del modelo de manera visual y se realizaron las siguientes pruebas de hipótesis para ciertos supuestos: prueba de normalidad de Anderson Darling y la prueba de autocorrelación de Durbin-Watson.

Luego de verificar la validez del modelo, se calcularon los intervalos de confianza para las estimaciones de los coeficientes del modelo. Bajo sospecha de la relevancia de alguna variable predictora, se realizó una prueba de la razón de verosimilitud. Como resultado, el modelo se reajustó sin algunas variables predictoras.

Cabe mencionar que, para obtener los resultados de las pruebas y del ajuste, se utilizó Python en la versión 3.0.

8.2. MDS

Para obtener una representación en dos dimensiones de los países, primero se calcularon las distancias euclidianas de cada uno (considerando las variables GDP per cápita, Apoyo social, Esperanza de vida, Libertad, Generosidad, Percepciones de corrupción). Posteriormente se construyó una configuración MDS de las poblaciones mediante la solución clásica. Eligiendo una dimensión de la representación euclidiana de acuerdo al porcentaje de la variabilidad explicada (mayor al 70 %).

Posteriormente, se aplicó K-means con el objetivo de agrupar a los países de acuerdo a su similaridad en la calidad de vida.

Cabe mencionar que, para la generación y análisis de resultados, se utilizó RStudio en su última versión.

9. Descripción de los datos

Los datos seleccionados para llevar a cabo este proyecto corresponden a la base de datos denominada “Happiness Index 2018-2019”, obtenida de `https://www.kaggle.com/datasets/sougatapramanick/happiness-index-2018-2019`. De acuerdo al sitio web donde se obtuvo la base de datos, estos se obtuvieron del Informe Mundial de la Felicidad de los años 2018 y 2019.

Esta base de datos contiene información valiosa para analizar y describir la calidad de vida en diferentes países. Incluye variables relevantes como ’Posición general’, que indica la clasificación general del país en términos de felicidad; ’PIB per cápita’, que mide el nivel económico promedio de los habitantes; ’Apoyo social’, que refleja el nivel de apoyo y redes sociales disponibles; ’Esperanza de vida saludable’, que indica la esperanza de vida libre de enfermedades; ’Libertad para hacer elecciones’, que evalúa el grado de libertad individual en la toma de decisiones; ’Generosidad’, que captura la propensión a ayudar y donar; y ’Percepciones de corrupción’, que refleja la percepción de corrupción en el país.

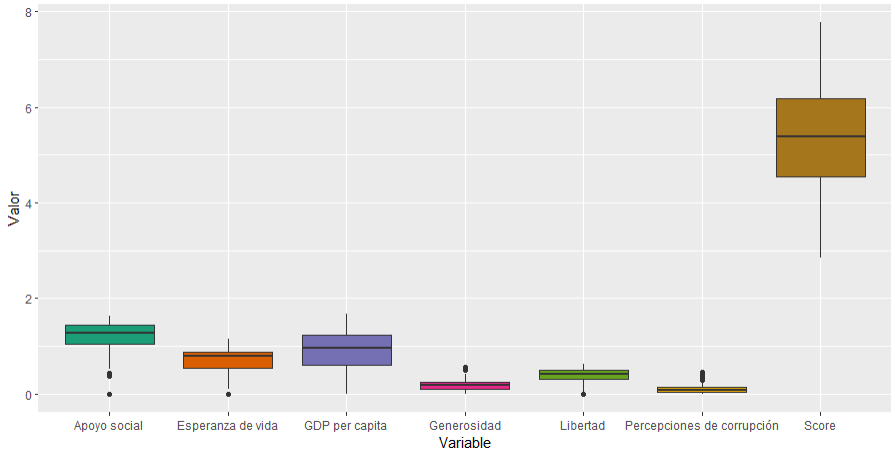

Figura 1: Diagramas de caja y bigote

De la Figura 1 notamos que que para la variable de apoyo social existen algunos valores atípicos, al igual que para la Esperanza de Vida, la Generosidad, la Libertad y para la variable de Percepciones de corrupción. Observamos mayor variabilidad entre los valores de las variables de Esperanza de Vida, Libertad de Decisiones, Generosidad y Percepción de Corrupción.

El Score (Índice de felicidad) por otra parte, mostró ser simétrico, lo que sugiere una distribución más equilibrada. Esto indica que la mayoría de los países tienen puntajes cercanos a la mediana y a la media. Los

valores obtenidos para el GDP per cápita nos indican que el 50 % de los países tienen un GDP per cápita por encima de 0.960. De manera que estos valores nos indican que no existe variabilidad significativa en los niveles económicos de los países analizados. Notamos que para el Apoyo Social la mayoría de los países tienen niveles de apoyo social cercanos a la media y a la mediana, de manera que la mayoría de los países tienen niveles similares de apoyo social.

De la Figura 2, se observa que las variables GDP per cápita, Apoyo social y Esperanza de vida saludable muestran una alta correlación positiva con respecto al índice de felicidad (Score), sugiriendo la existencia de una dependencia lineal. Con respecto a las variables restantes, se observa generalmente una correlación positiva media. Por lo tanto, se espera que haya una buena calidad de ajuste del modelo mediante regresión, debido a la posible asociación lineal entre las variables.

Por otro lado, también se observa cierta correlación entre variables predictoras. Esto podría causar una inflación de la estimación del error de los coeficientes del modelo de regresión.

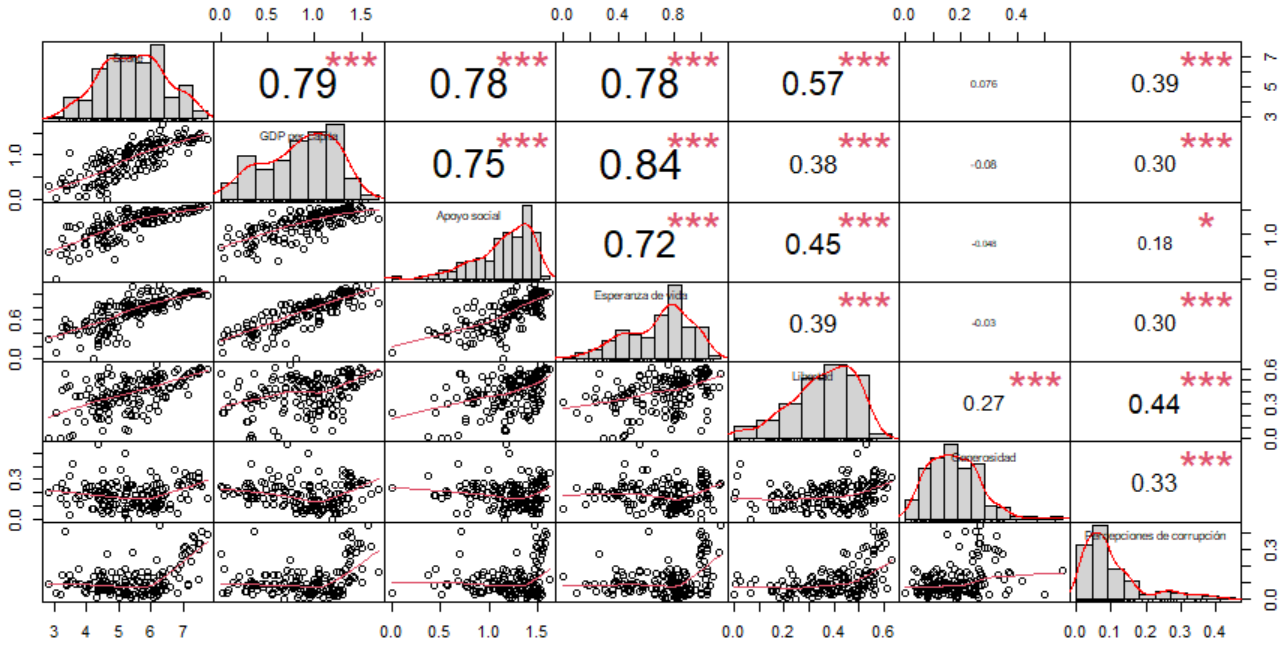

Figura 2: Matriz de correlación

10. Resultados

10.1. Regresión multivariada

Con respecto a la calidad de ajuste del modelo a los datos observados, se obtuvo un coefiente de determinación de $R^2 = 0.779$. En consecuencia, el modelo de regresión explica el 77.9 % de la variabilidad del índice de felicidad, lo que indica un buen ajuste del modelo a los datos.

En cuanto a la precisión del modelo, la raíz del error cuadrático medio fue de 0.5; por lo tanto, se esperaría que el modelo se equivoque en promedio 0.5 unidades del valor real. En relación con la escala de la variable objetivo, puede considerarse que la precisión del modelo es aceptable.

La Tabla 1 muestra los coefientes del modelo ajustado, así como los respectivos intervalos de confianza para un nivel de significancia de $\alpha = 0.05$. Se observa que los intervalos individuales y simultáneos de los coeficientes de Generosidad y Percepciones de corrupción contienen el cero, lo que sugiere que el índice de felicidad no depende de dichas variables. Para evaluar su relevancia en el modelo, se llevó a cabo la prueba de la razón de verosimilitud.

					Intervalos de confianza de 95 %			
					Individuales		Simultáneos	
Variable	coef	std err	t	P>|t|	[0.025	0.975]	[0.025	0.975]
Constante	1.7952	0.211	8.505	0.000	1.378	2.212	0.9915	2.599
GDP per cápita	0.7754	0.218	3.553	0.001	0.344	1.207	-0.0556	1.6064
Apoyo social	1.1242	0.237	4.745	0.000	0.656	1.592	0.2221	2.0263
Esperanza de vida saludable	1.0781	0.335	3.223	0.002	0.417	1.739	-0.1958	2.3521
Libertad de decidir	1.4548	0.375	3.876	0.000	0.713	2.197	0.0255	2.8841
Generosidad	0.4898	0.498	0.984	0.327	-0.494	1.473	-1.4056	2.3852
Percepciones de corrupción	0.9723	0.542	1.793	0.075	-0.099	2.044	-1.093	3.0376

Cuadro 1: Estimaciones acerca de los coeficientes del modelo

Considerando la hipótesis nula que indica que los coeficientes de las variables Generosidad y Percepciones de corrupción son iguales a cero, en la prueba de verosimilitud se obtuvo que $Fratio : 2.77 < F_{2,149}(0.05) = 3.05$. En consecuencia, a un nivel de confianza del 95 %, no hay suficiente evidencia estadística para indicar que la generosidad y las percepciones de corrupción tienen un efecto sobre el índice de felicidad; por ende, tales términos son prescindibles.

Eliminando Generosidad y Percepciones de corrupción como variables predictoras, se obtuvo un coeficiente de determinación $R^2 = 0.771$, mientras que la raíz del error cuadrático medio fue de 0.53. Por lo tanto, sin dichas variables, se mantuvo la buena calidad del ajuste y la precisión aceptable del modelo.

Los coefientes del modelo reajustado se presentan en la Tabla 2. Se observa que ningún intervalo incluye el cero, lo que sugiere la relevancia de las variables en el modelo.

					Intervalos de confianza del 95			
					Individuales		Simultáneos	
Variable	coef	std err	t	P>|t|	[0.025	0.975]	[0.025	0.975]
Constante	1.8921	0.199	9.491	0.000	1.4982	2.286	1.2198	2.5644
GDP per cápita	0.8105	0.216	3.745	0.000	0.3829	1.2382	0.0807	1.5404
Apoyo social	1.0166	0.235	4.331	0.000	0.5528	1.4804	0.225	1.8082
Esperanza de vida saludable	1.1414	0.337	3.384	0.001	0.475	1.8079	0.004	2.2789
Libertad de decidir	1.8458	0.340	5.423	0.000	1.1733	2.5184	0.698	2.9936

Cuadro 2: Estimaciones acerca de los coeficientes del modelo reajustado

10.1.1. Adecuación del modelo ajustado con todas las variables

Se calcularon los puntos de apalancamiento (Leverage points) asociados, los cuales cuantifican el potencial de un punto para ejercer una fuerte influencia en el análisis de regresión. En particular, los puntos de apalancamiento indican la presencia de tres observaciones inusuales, por lo que tres países representan datos atípicos.

Para identificar patrones o tendencias no capturados por el modelo y para evaluar la adecuación del ajuste, se graficaron los residuos contra cada una de las variables predictoras. El resultado se muestra en la

Figura 3, en donde se observa que no hay patrones sistemáticos en la dispersión de los residuos, por lo que no existe la necesidad de añadir más terminos al modelo, pues captura completamente la relación entre las variables.

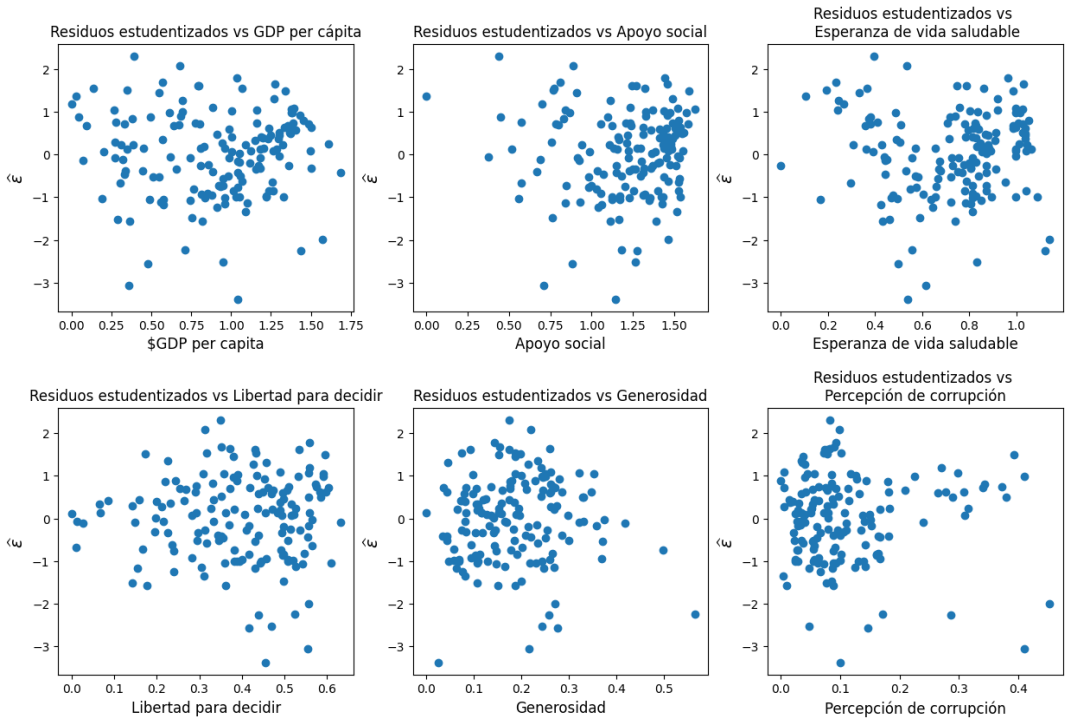

Figura 3: Gráfica de los residuales contra una variable predictora

Verificación de supuestos

Es válido realizar inferencias sobre el modelo si se cumple que:

- Los errores se distribuyen normal.
- Los errores son independientes.
- La varianza de los errores es constante.

Con respecto al supuesto de normalidad, en la gráfica de probabilidad de la Figura 4 se observa que, a pesar que algunos puntos cerca de las colas no caen exactamente a lo largo de la línea recta, la mayoría de los errores parecen estar distribuidos normalmente. Para verificar esta sospecha, se realizó la prueba de normalidad de Anderson-Darling. De acuerdo con el resultado de dicha prueba, a un nivel de confianza del 95 %, no hay suficiente evidencia para indicar que los errores no tengan una distribución normal ($Valor\ p = 0.27 > 0.05$).

Con respecto a la homocedasticidad de los errores, en la gráfica de residuales contra predichos no se observa alguna tendendencia, sugiriendo así que los errores tienen una variación constante.

Por último, la gráfica de residuales en el tiempo parace indicar que no hay correlación entre los errores del modelo en distintos periodos (autocorrelación), ya que no visualiza un patrón evidente. Sin embargo, para verificar de manera formal la independencia, se realizó la prueba de Durbin Watson para autocorrelación, en donde se obtuvo que, a un nivel de confianza del 95 %, no hay suficiente evidencia para indicar la presencia de autocorrelación entre los errores ($1.5 < Durbin - Watson : 1.64 < 2.5$).

Al cumplirse todos los supuestos del modelo de regresión, se asegura la validez de las estimaciones de los coeficientes y las pruebas de hipótesis asociadas.

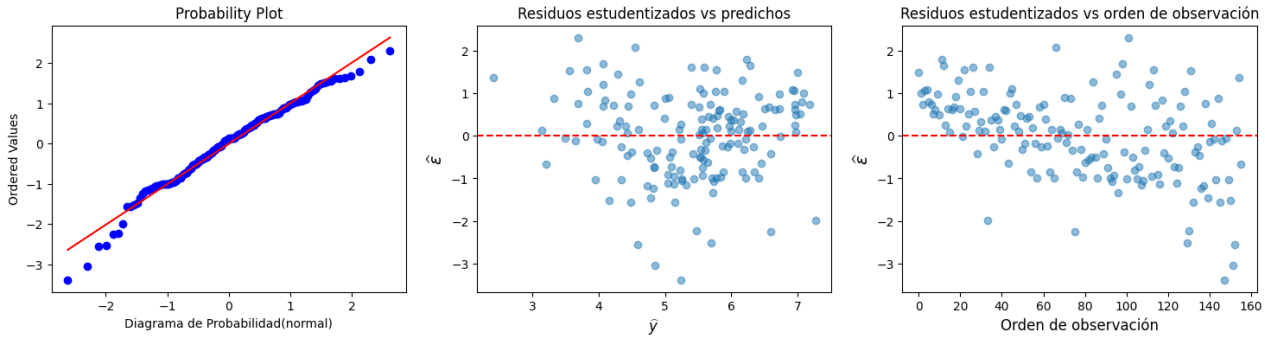

Figura 4: Verificación de supuestos

Para el modelo ajustado sin las variables Generosidad y Percepciones de corrupción, también se cumplen los supuestos, por lo que las estimaciones de los coeficientes son insesgadas y eficientes.

10.2. MDS

Con un 86.64 % de la varianza total explicada, la Figura 5 muestra la representación de los países en dos coordenadas principales. Se observa que los países que se encuentran en la derecha del plano pertenecen en su mayoría al continente africano, que se caracteriza por su pobreza y hambruna. Asimismo, se encuentran algunos países de Asia, que se han visto marcados por la guerra, desastres naturales y los conflictos internos.

Observando los países que conforman los grupos obtenidos mediante K-means con $K = 2$, se puede inferir que el grupo azul son los que presentan una buena calidad de vida, mientras que el grupo amarillo son los que tienen una mala calidad de vida.

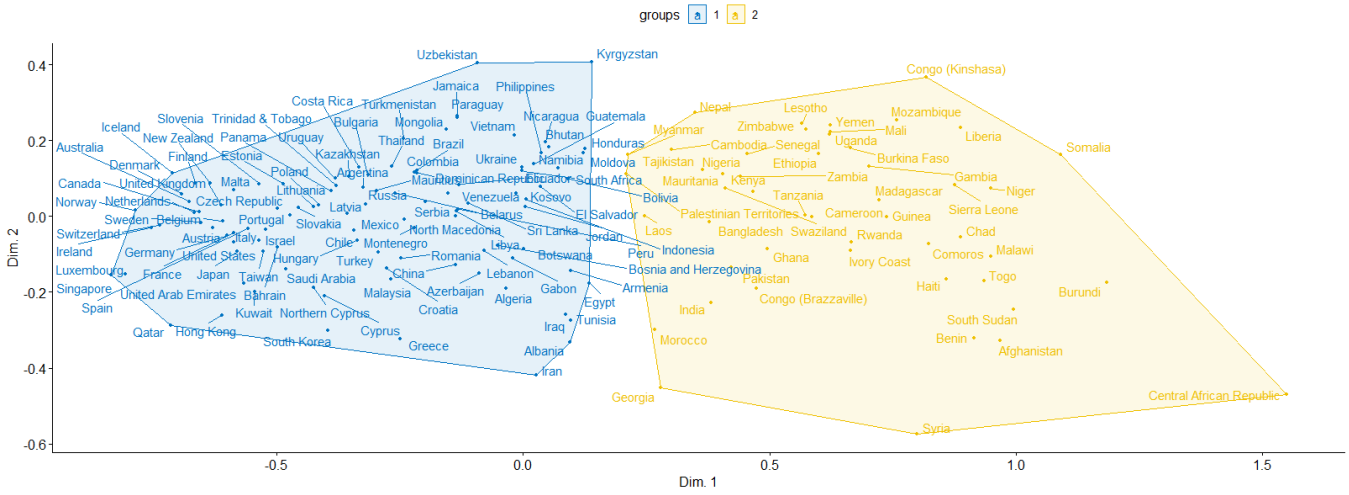

Figura 5: Representación en dos dimensiones obtenida mediante MDS

]

11. Conclusiones

Se puede predecir el índice de felicidad (con una presición aceptable) mediante el siguiente modelo lineal múltiple:

$$\begin{aligned} \textit{Índice de felicidad} = 1.89 + 0.81 * \textit{GDP per cápita} + 1.01 * \textit{Apoyo social} \\ + 1.14 * \textit{Esperanza de vida saludable} + 1.84 * \textit{Libertad de decidir} \end{aligned}$$

Sin embargo, hay ciertos pares de variables predictoras que presentan correlación alta entre ellas, por lo que la varianza del estimador de los coeficientes puede agrandarse, afectando el tamaño del intervalo de confianza.

Por otro lado, mediante MDS se puede representar las relaciones entre los los países en un espacio de dos dimensiones, en donde se observa la similitud de los países de acuerdo a su calidad de vida, determinado por las siguientes variables: GDP per cápita, Apoyo social,Esperanza de vida, Libertad, Generosidad y Percepciones de corrupción.

No obstante, los clústeres obtenidos mediante K-means en la configuración MDS de 2 dimensiones se pueden superponer, indicando que existen países que comparten características comunes y se encuentran en regiones que se superponen entre los dos clústeres. Esto puede dificultar la distinción o asignación precisa de países a un clúster específico.

Referencias

[1] Kemmler, G., Holzner, B., Kopp, M., Dünser, M., Greil, R., Hann, E., & Sperner-Unterweger, B. *Multidimensional Scaling as a Tool for Analysing Quality of Life Data.* [Rev. Quality of Life Research, 11, 3.] (2002). Disponible en: `https://www.jstor.org/stable/4038041?seq=1&cid=pdf-reference#references_tab_contents`

[2] Potter, J., Cantarero, R., & Wood, H. *The Multi-Dimensional Nature of Predicting Quality of Life.* [Procedia- Social and Behavioral Sciences, 1877-0428.] (2012). Disponible en: `http://digitalcommons.unl.edu/arch_facultyschol/31`

[3] Johnson, R. A., & Wichern, D. W. *Applied Multivariate Statistical Analysis. Prentice Hall.* (2007).

[4] Moreno, A. B., & Chávez, A. G. *100 Problemas Resueltos de Estadística Multivariante Implementados en Matlab.*[Delta Publicaciones.](2007).

[5] Montgomery, D. C., Peck, E. A., & Vining, G. G. *Introduction to Linear Regression Analysis. Wiley-Interscience.*(2001).

[6] P. Hryoruk, S. Grygoruk, N. Khrushch, T. Hovorushchenko. *Using non-metric multidimensional scaling for assessment of regions' economy in the context of their sustainable development*[Khmelnytskyi National University, 11 Instytutska Str. Khmenlnytskyi] (2020).