Relación de variables macroeconómicas: 2021

Francisco Blas · Antonio Huerta

Introducción a la ciencia de datos

Primavera, 2022

Resumen

El objetivo de este trabajo es analizar el índice de calidad de vida, el Índice de Gini, el Producto Interno Bruto (PIB), la tasa de inflación, la expectativa de vida y la población de distintos países para evidenciar relaciones en las características macroeconómicas y entender de mejor manera el contexto económico actual. Para ello, se emplean métodos de aprendizaje automatizado, como regresiones lineales para explicar el PIB con variables de expectativa de vida, calidad de vida e inflación. Por otro lado, a través de los algoritmos KNN y K-Medias++, se obtienen el agrupamiento de las variables de expectativa y calidad de vida segmentadas por el nivel de PIB, donde se muestra que los niveles óptimos para el PIB coinciden con la propuesta.

1. Introducción

Para evaluar el desempeño de un país en aspectos como el económico, político, social, demográfico, etcétera, existen múltiples índices y medidas que tratan dar un acercamiento puntual de la realidad. Sin embargo, no es sencillo conseguir que un sólo número dé la suficiente información para un único aspecto de interés, por lo que existen varios índices o parámetros asociados a un ámbito de las naciones.

Desde la perspectiva económica, se pueden enumerar bastos indicadores que tratan de medir el desempeño de los países para mejorar su bienestar. Por ejemplo, el bien conocido PIB, indicadores de consumo, inflación, producción, confianza, inversión y un largo etcétera. De manera general, esos parámetros suelen mantener algunas relaciones.

Si bien, estos indicadores son fuente de variadas investigaciones, el reciente choque de la pandemia ocasionada por la propagación del virus SARS-CoV-2 que provoca la enfermedad COVID-19 tuvo fuertes repercusiones tanto sociales como económicas que deben ser evaluadas con base en la información proporcionada por los indicadores.

En el presente trabajo, se estudian las relaciones entre las variables asociadas al bienestar socioeconómico por medio de métodos de aprendizaje automatizado, como técnicas de agrupamiento y regresiones lineales. Con ello, se contrastan resultados de años anteriores y se sienta una base para futuras investigaciones.

Finalmente, el estudio se divide en cuatro secciones: introducción, marco teórico, análisis de datos y conclusiones. En el marco teórico se ofrece un panorama de las relaciones que se estudian y su fundamento. El análisis de datos, punto medular de la investigación, consta de dos subsecciones que buscan limpiar, ordenar, visualizar y explotar los datos con las diversas técnicas. Por último, se enuncia una serie de conclusiones y resultados.

2. Marco teórico

Si bien, hay variables que intuitivamente se espera que estén correlacionadas, hay evidencia empírica de que algunas de nuestras variables de interés están correlacionadas como es el caso del PIB y la expectativa de vida, cuya relación analítica está dada por la siguiente ecuación.

$$Expectativa\ de\ vida = \beta_0 + \beta_1 \ln(PIB) + u_t,$$

donde los errores u_t están idénticamente distribuidos con una distribución normal estándar. A pesar de que, esta relación se aproxima mejor para países con un PIB bajo, frecuentemente se toma para todos los países, puesto que la correlación tiende a ser significativa.

Por otra parte, aunque el PIB es utilizado como una medida del bienestar, no necesariamente implica que un país con un PIB elevado tenga un alto índice de calidad de vida. Sin embargo, es válido preguntarse si estas variables están correlacionadas significativamente para la mayoría de los países; en particular, veremos si existe alguna relación significativa entre esta y otras variables.

Análogamente, las variables que están suficientemente correlacionadas son regresadas entre sí mientras no haya problemas de multicolinealidad, aunque hay variables que se esperan estar correlacionadas como el PIB y la tasa de inflación. Además, también existen posibles relaciones significativas, como una relación entre la expectativa de vida

y la calidad de vida, exceptuando a países que son contraejemplo de este hecho, como China. Cabe mencionar que no se descarta que existan variables correlacionadas que no tengan una relación causal o explícita más allá de los datos, pero que sirva para trabajos comparativos futuros.

Por otra parte, se aplicarán técnicas de agrupamiento para detectar la cantidad de grupos representativos óptimos por medio del algoritmo de $K-means++$. Este algoritmo está basado en el método usual de $K-means$, con la diferencia de que éste resuelve la problemática $NP-duro\ k-means$, donde el algoritmo usual arroja centros de los grupos que satisfacen la condición de mínima distancia, pero que visualmente se aprecia que los grupos están incorrectamente definidos.

Así pues, en este trabajo no se utiliza un conjunto de etiquetas discretas, sino un conjunto continuo, el cual será mapeado a un gradiente de color que nos permita visualizar tendencias o grupos representativos a partir de aplicar un umbral para determinar el número de grupos necesarios para cada caso.

A fin de lograr un correcto funcionamiento del algoritmo de clasificación, se omiten los valores extremos presentes en la muestra, a excepción de la población. Para identificarlos se considera el método a partir del cálculo de los cuartiles: filtrar fuera los datos que no estén dentro del segundo y tercer cuartil más 1.5 veces el rango intercuartílico de los datos.

Los datos utilizados para el análisis corresponden a una ventana de tiempo con cierre en el año 2021, a fin de evitar sesgos en los resultados debido a la falta de datos para algunas variables.

3. Análisis de datos

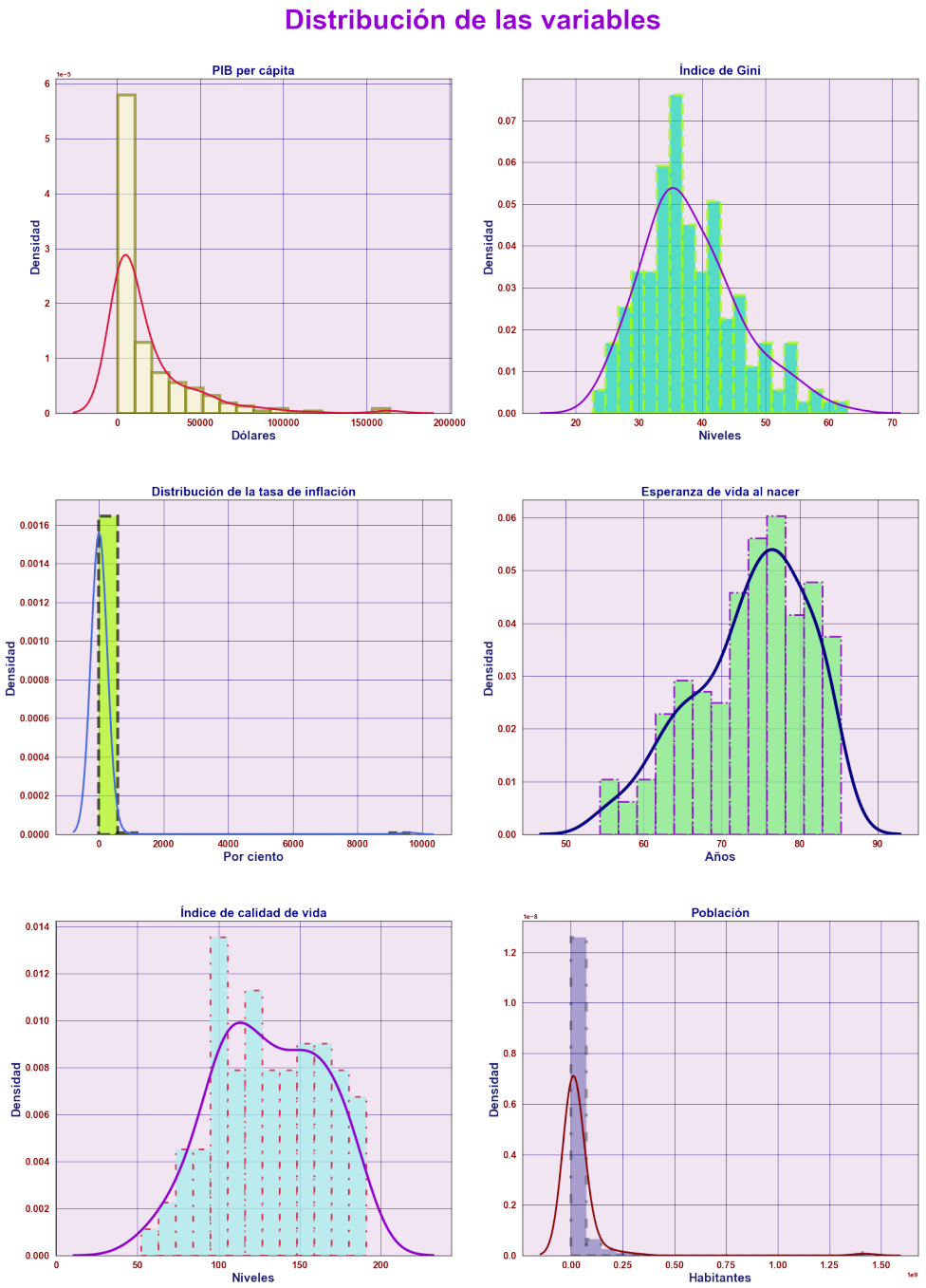
3.1. Limpieza de datos

En primera instancia, de los bases de datos recolectadas con datos sobre el PIB, población, esperanza de vida al nacer, calidad de vida, inflación e Índice de Gini por país, se toman los datos necesarios para realizar el análisis. Además, se considera al nombre del país como la llave primaria dada su unicidad.

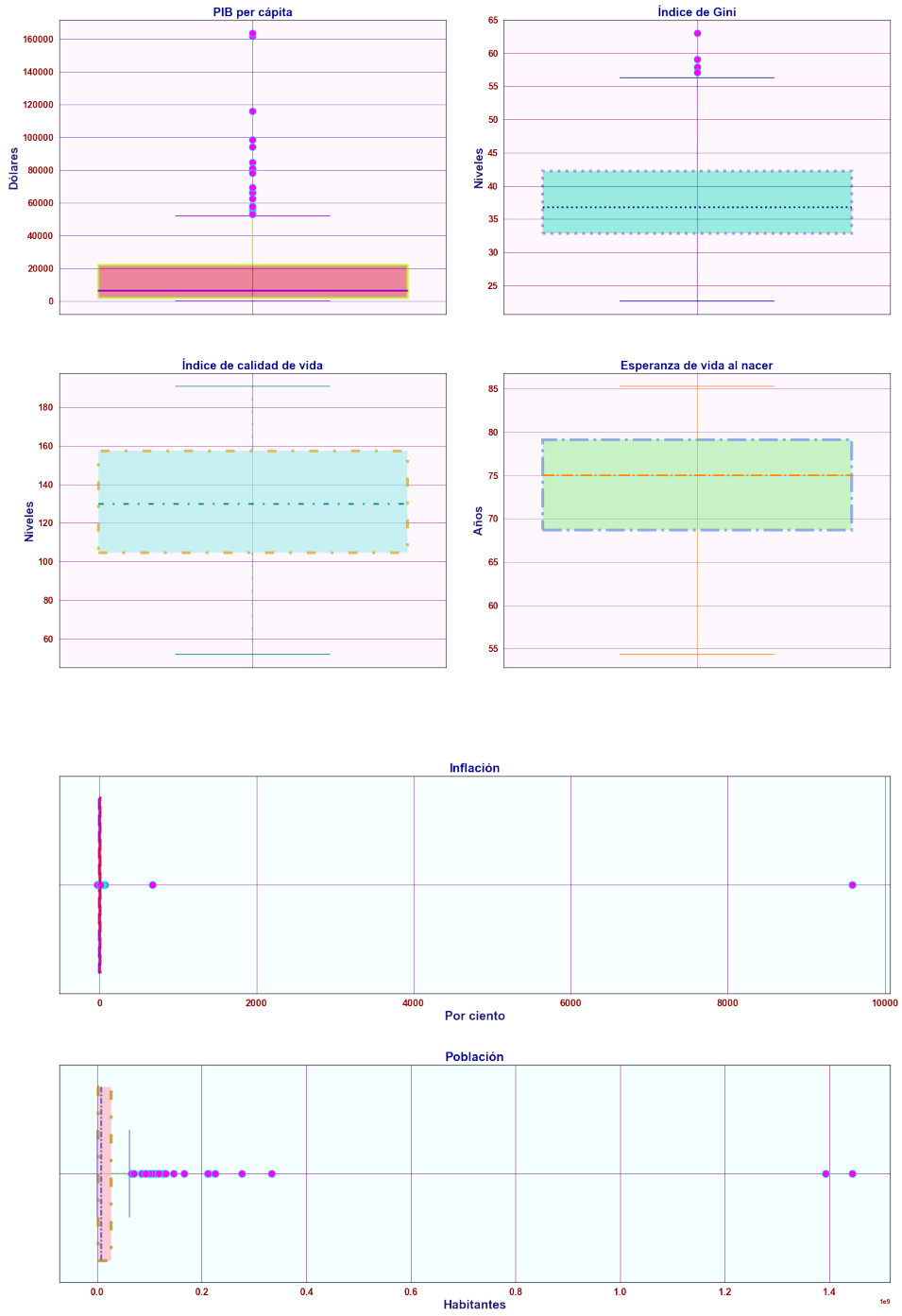
En segundo lugar, a fin de evitar sesgos en los algoritmos, se toma en consideración, para cada método, la importancia de eliminar o sustituir los valores faltantes en las bases de datos. Finalmente, con el propósito de no hacer complejo el análisis, se prefiere tomar variables que engloben más datos o que sean únicas en el sentido de no estar

correlacionadas con otras variables.

A continuación, se muestra la distribución de las variables:

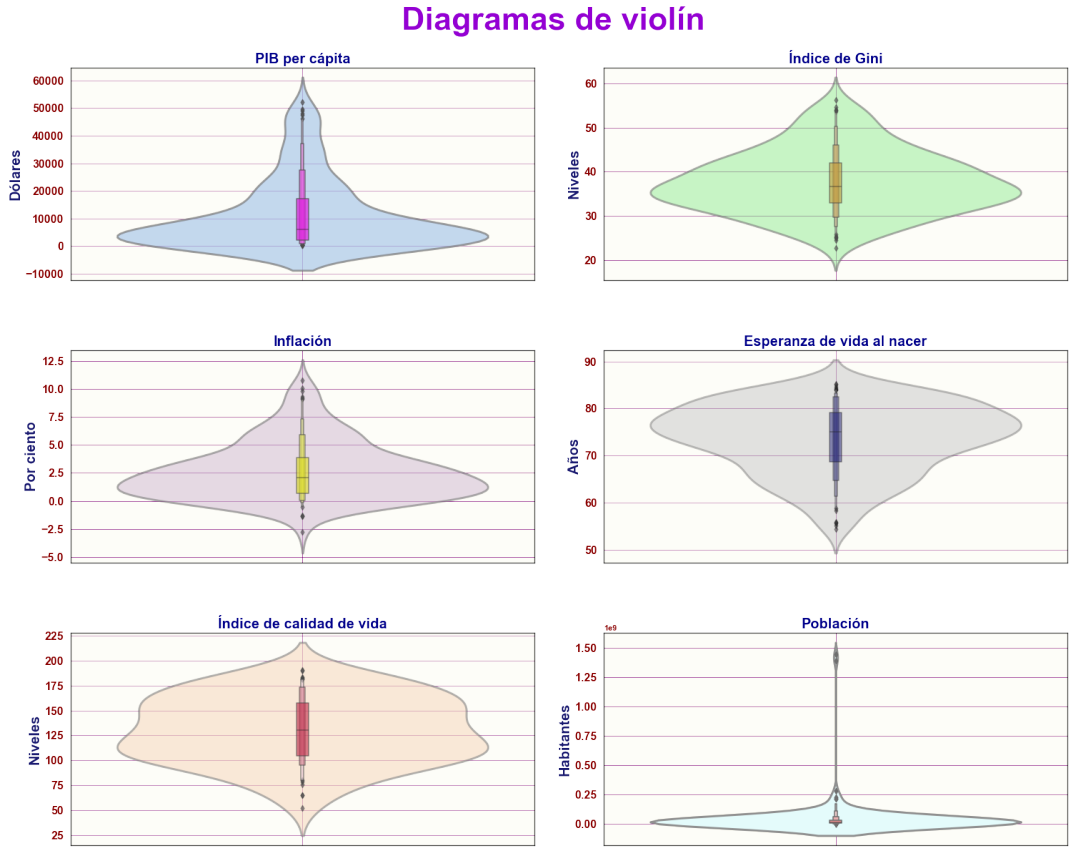


De la figura anterior, se observan algunas anomalías en las distribuciones, como leptocurtosis, sesgos y colas pesadas. De esto, se deduce la existencia de valores extremos (*outliers*), por lo que es necesario apartarlos de la base de datos. Para ello, se toma como auxiliar los gráficos de caja y brazos para analizar cada variable.



De manera particular, es claro el valor extremo de la tasa de inflación debido a la situación económica de las últimas décadas en Venezuela; asimismo, también se esperan *outliers* para el *Índice de Gini* puesto que hay países con una desigualdad marcada y, viceversa, países con una distribución equitativa del ingreso, como República Checa.

Adicionalmente, se aprecia que en las variables del PIB, la tasa de inflación, la población y el Índice de Gini hay outliers, lo cual era de esperarse para alunos casos. Como se mencionó, son removidos con el método descrito. Sin embargo, dadas las regresiones de interés, no se eliminan los outliers de la población, además, esto podría ocasionar una reducción de más del 40 % de los datos si se aplican los filtros a esta variable.

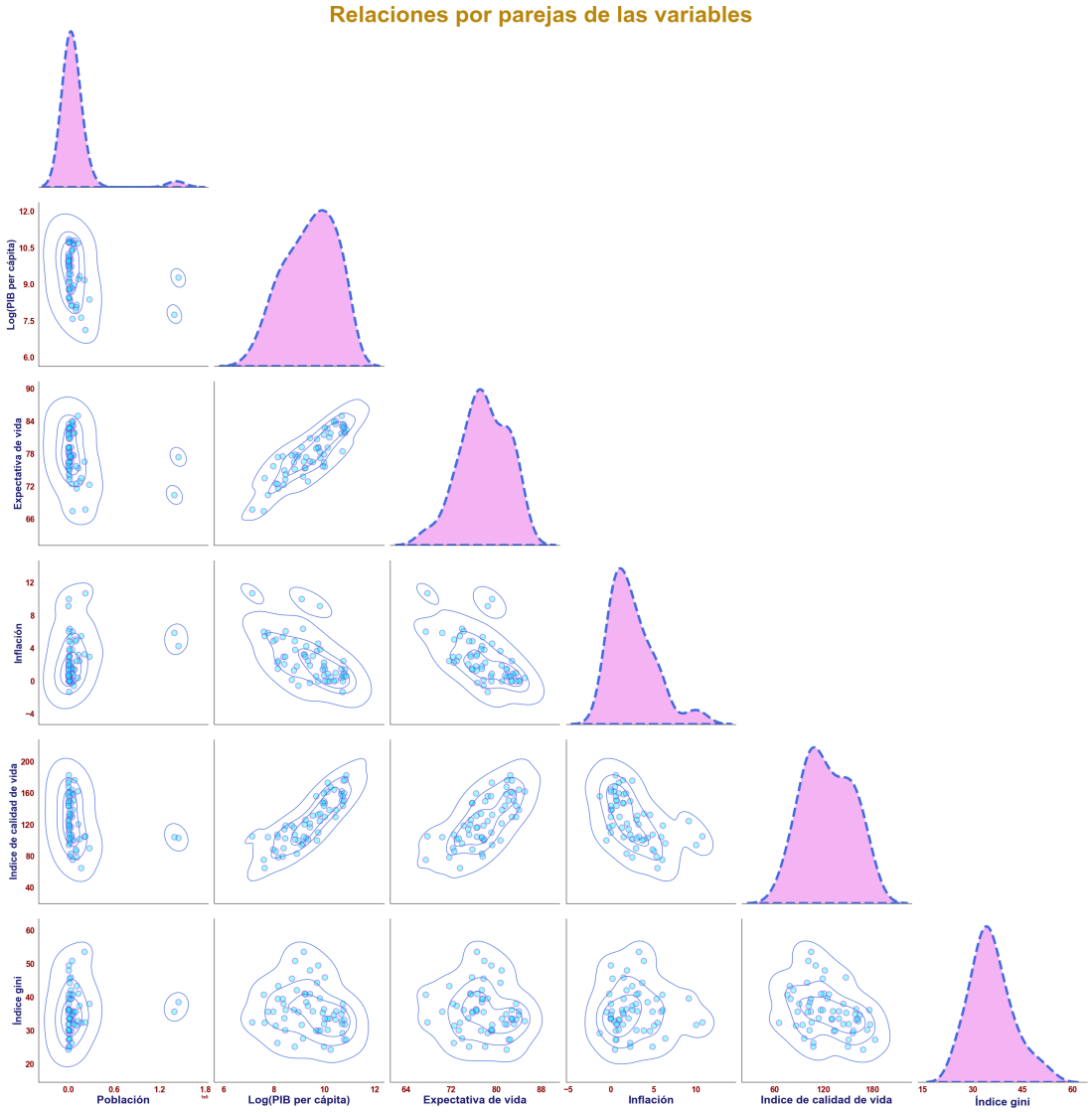


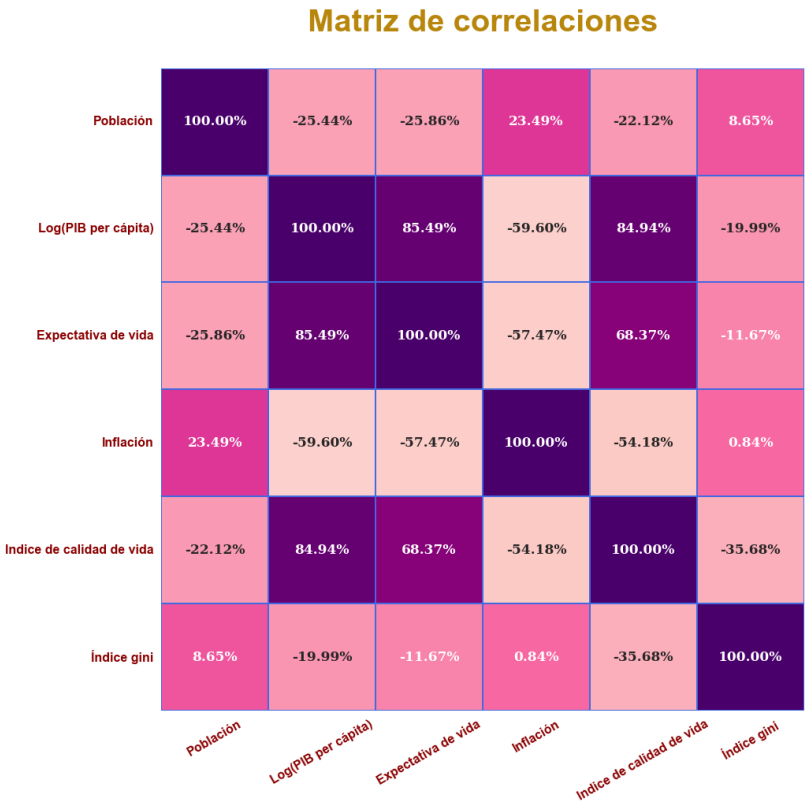
Se puede apreciar que se han removido la mayoría de los datos extremos, a excepción de los datos de la población, dadas las consideraciones antes mencionadas. Con esto, se construye el conjunto de datos para hallar correlaciones y visualizaciones que permitan estudiar el comportamiento entre las distintas variables. En la tabla de abajo, se muestra un resumen de las estadísticas:

	Población	ln($PIBpc$)	Esperanza de vida	Inflación	Calidad de vida	Índice de Gini
Cuenta			59			
Media	93'645,140	9.4	77.9	2.6	125.4	35.7
SD	257'500,600	1.0	4.1	2.6	29.6	6.5
Mínimo	1'215,584	7.1	67.5	-1.3	65.3	24.4
25 %	7'797,106	8.7	75.5	0.6	103.	31.9
50 %	21'497,310	9.5	77.7	1.9	121.7	35.4
75 %	66'816,650	10.2	81.6	3.9	150.2	39.5
Máximo	1'444,216,000	10.9	85.0	10.7	182.8	53.7

De lo anterior, se observa que la limpieza de datos se realizó de manera correcta. Sin embargo, el número de observaciones con las que se cuenta es de 59 países, aunque aún son suficientes para lograr los propósitos de este estudio.

En la siguiente figura se muestran las relaciones de las variables. Se puede apreciar que hay algunas que se agrupan, pero este análisis se hará detalladamente en la sección posterior con el método de $K-means++$. Adicionalmente, se aprecia que hay correlaciones entre distintas variables, por lo que también se considera la matriz de correlaciones, donde se impone un $R^2 \geq 0{,}5$ como umbral para tomar en cuenta las variables en la regresión.





De las figuras anteriores, sólo se tiene evidencia de una relación fuerte entre el logaritmo del PIB per cápita, la expectativa de vida y el Índice de calidad de vida. Además, aunque no muestran una alta correlación, también se muestra una relación entre el logaritmo del PIB per cápita y la tasa de inflación. Sin embargo, dados los problemas de multicolinealidad entre las variables, no es posible hacer una regresión múltiple entre éstas, por lo que las regresiones se realizan por separado entre las pareja que presenten mayor correlación. La multicolinealidad está dada por el criterio de selección con la referencia $R^2 \geq 0{,}5$.

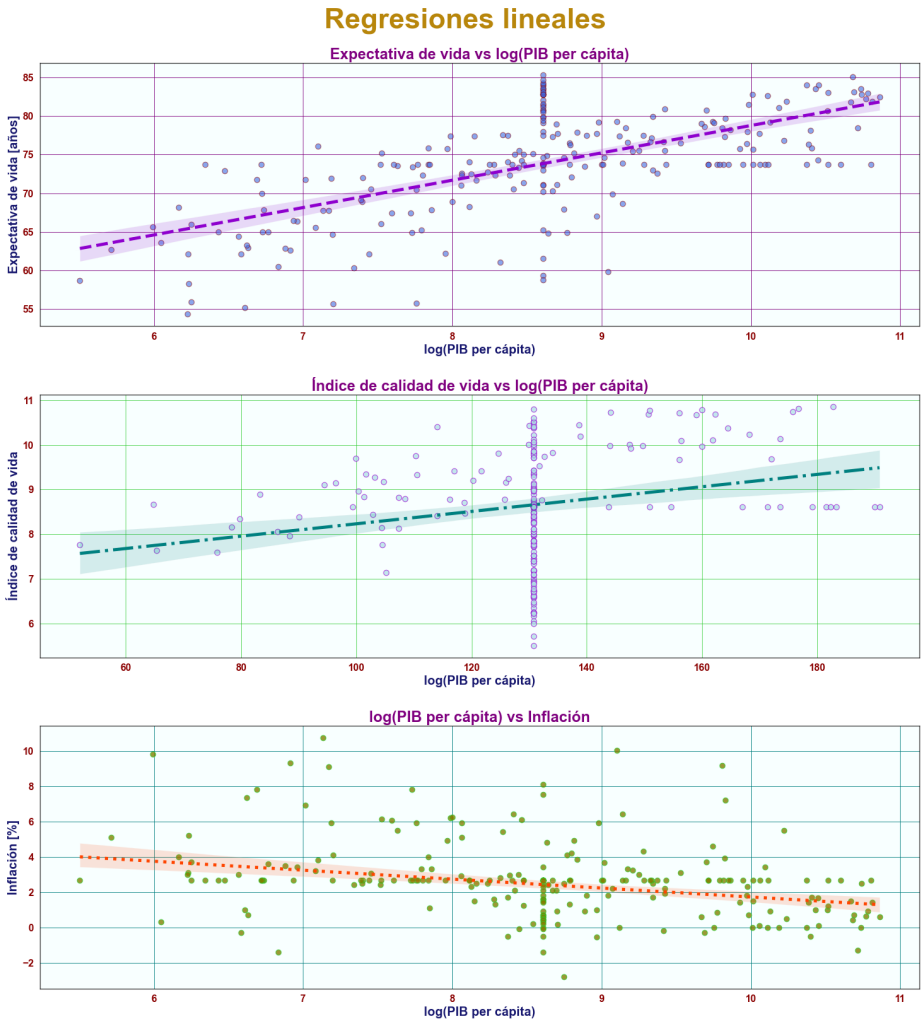
3.2. Regresión lineal

En la figura siguiente se exhiben las relaciones entre el PIB per cápita y la expectativa de vida, calidad de vida e inflación, tal y como lo sugería la matriz de correlaciones. Además, como se observa, los datos faltantes que fueron sustituidos con el promedio no afectan a los estimadores.

En la primeras dos gráficas se muestra la relación con la esperanza de vida al nacer y el Índice de calidad de vida; en línea con lo esperado, ambas tienen una pendiente positiva, es decir, el incremento del PIB per cápita se refleja en un aumento de las dos variables. No obstante, la interpretación se debe manejar con cuidado ya que, por ejemplo, un aumento

excesivo en el PIB per cápita no incrementará en muchos años la expectativa de vida.

En contraste, la gráfica de abajo señala una relación negativa entre el PIB per cápita y la tasa de inflación.



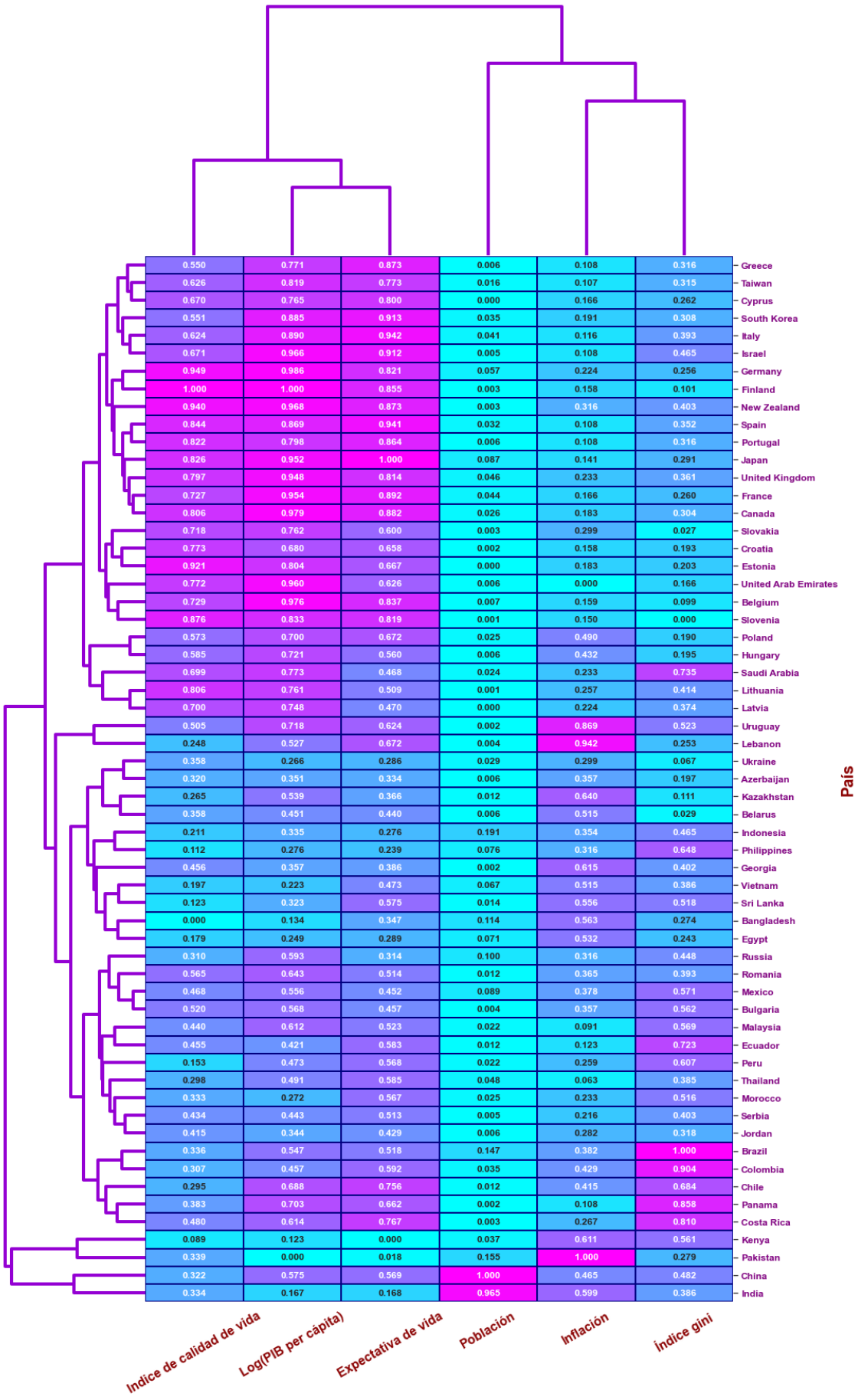
3.3. Clústeres

Para el algoritmo de agrupamiento, se consideran los siguientes criterios:

PIB	Esperanza de vida	Índice de calidad de vida	Inflación	Índice de Gini	Población
$PIB_{alto} \geq \$13,000K$	$EV_{alto} \geq 73$	$ICV_{alto} \geq 140$	$\pi_{alto} \geq 3{,}59\,\%$	$IG_{alto} \geq 45$	$P_{Muy_{a}lta} \geq 100'000,000$
$PIB_{bajo} \leq \$1,000K$	$EV_{bajo} \leq 70$	$ICV_{bajo} \leq 95$	$\pi_{alto} \leq 1{,}50\,\%$	$IG_{bajo} \leq 35$	$P_{Alta} \geq 55'000,000$
$PIB_{regular}$ e.o.c	$EV_{regular}$ e.o.c.	$ICV_{regular}$ e.o.c.	$\pi_{regular}$ e.o.c.	$IG_{regular}$ e.o.c.	$P_{Baja} \leq 15'000,000$
					$P_{Muy_{b}aja} \leq 100,000$
					$P_{regular}$ e.o.c.

Con base en las categorías propuestas para cada columna, se tienen los elementos necesarios para aplicar el algoritmo de KNN. En la siguiente figura se presenta un diagrama con las clasificaciones obtenidas.

Clustermap de las variables



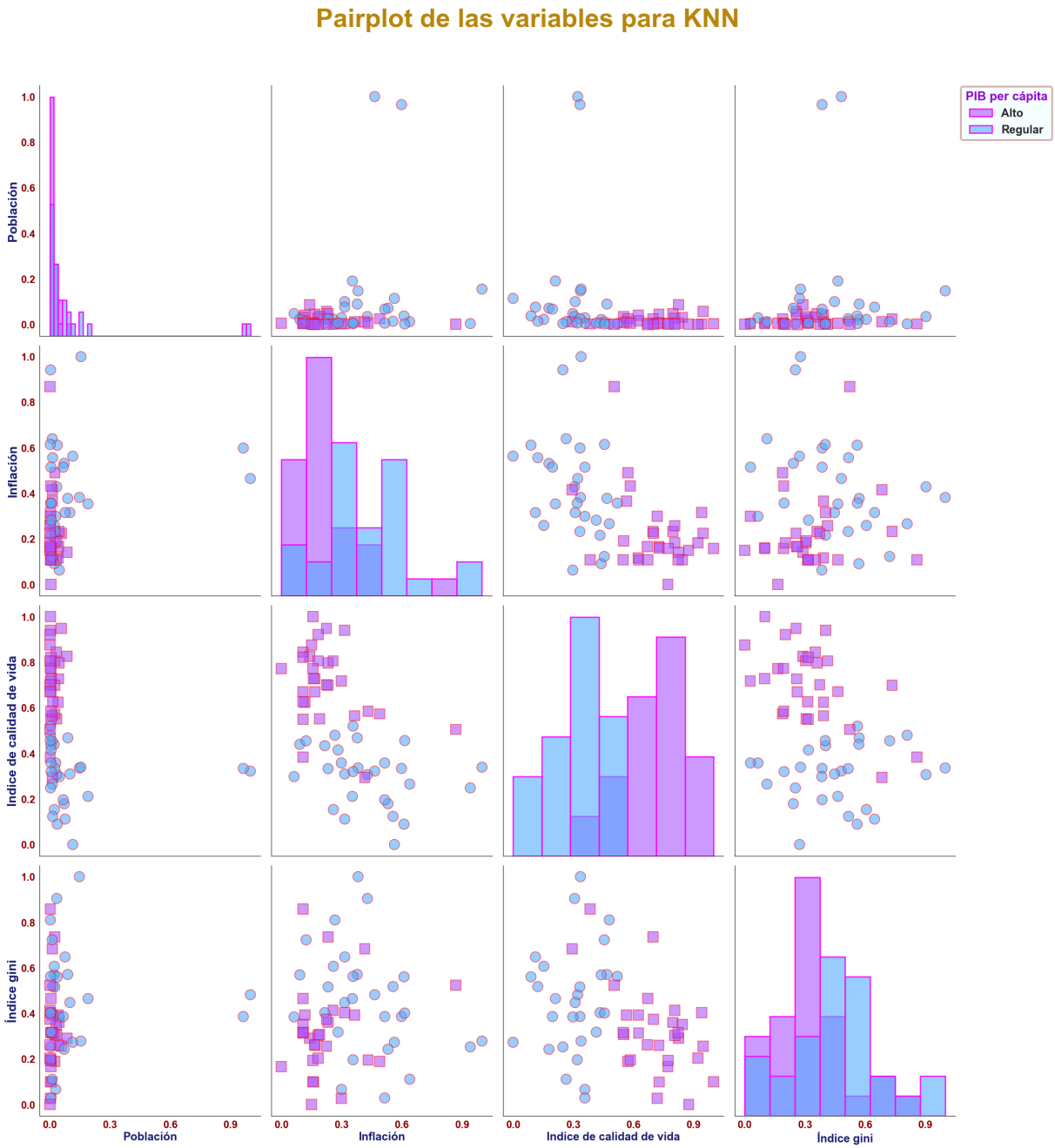
Las variables que están más cercanas entre sí, bajo la métrica de Minkowski, uniéndolas directamente por pares en la raíz (dichas líneas de unión forman una Y), donde la altura de las líneas formas indican qué tan distantes están entre sí bajo dicha métrica, es decir, entre menos altura, más cercanos, tal como el logaritmo del PIB per cápita y la expectativa de vida. Por otro lado, este gráfico indica que entre más próximos estén las variables es más probable que haya clústeres.

Para explotar en mayor medida el gráfico anterior, es útil guiarse con las alturas de las uniones para designar ciertas variables como etiqueta y crear un gráfico de dispersión con otras dos variables. Para nuestro caso, nos interesa segmentar a través del nivel de PIB per cápita, puesto que su distancia con otras variables de interés no es muy grande. Así, basándose en las proximidades de las variables y de la estrategia mencionada se considerarán los siguientes casos, segmentados por el nivel del PIB per cápita, para aplicar KNN:

- Expectativa de vida - Índice de calidad de vida.
- Población - Índice de calidad de vida.
- Índice de Gini - Índice de calidad de vida.
- Inflación - Índice de calidad de vida.

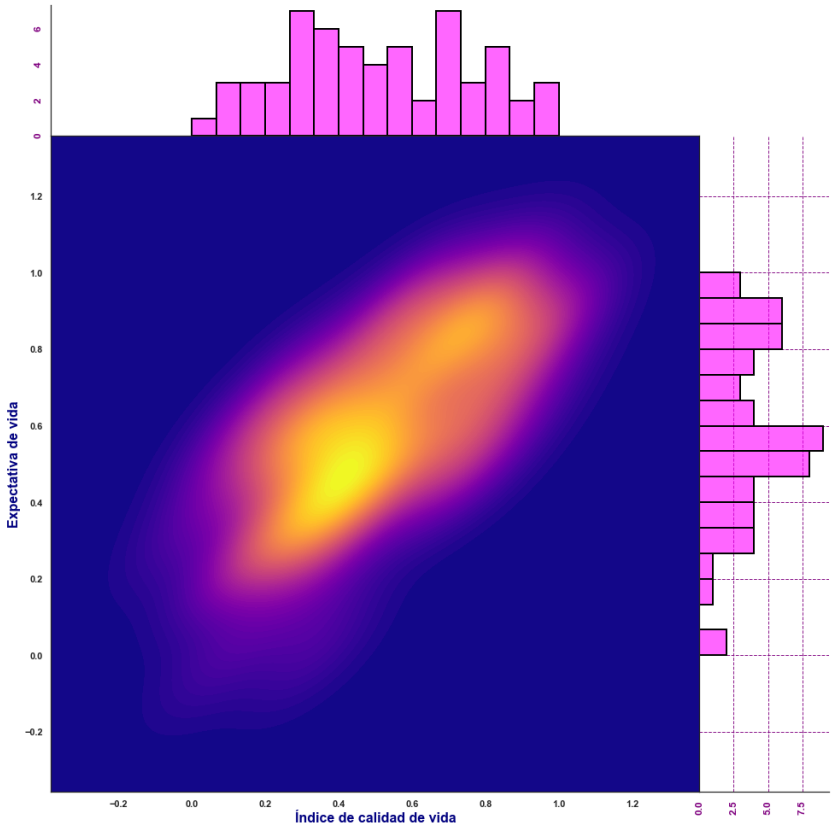
Bajo esta métrica, los países más parecidos con respecto a estas variables son Francia y Canadá. A continuación, se grafican estas relaciones con el PIB per cápita como elemento segmentador.

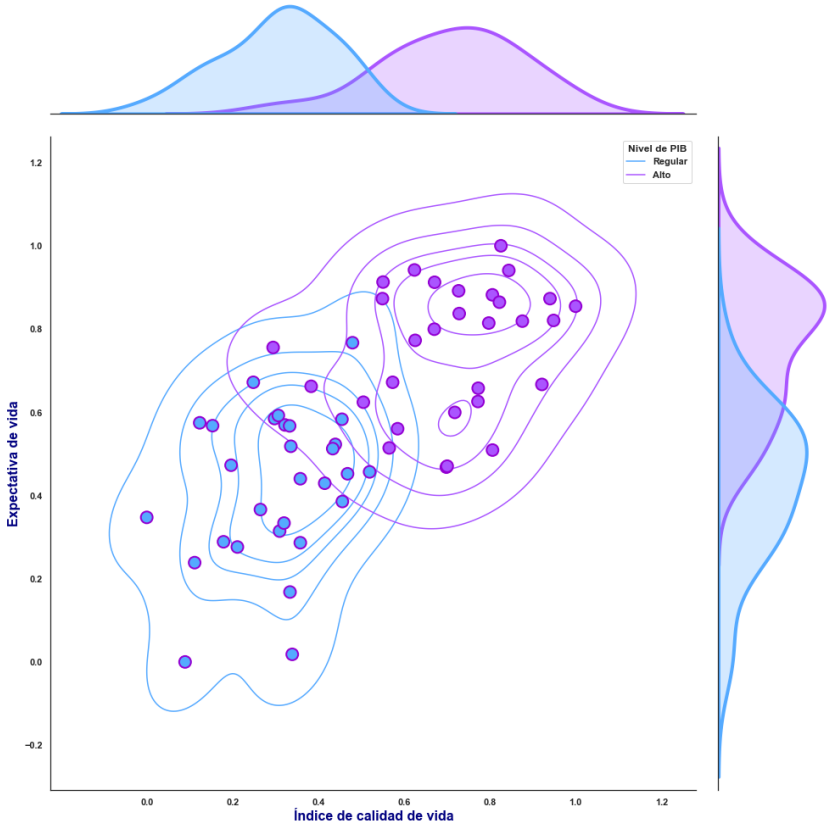
En efecto, se logra apreciar que, en general, los datos tienden a segmentarse de acuerdo al nivel del PIB per cápita. Sin embargo, también se observa que, en el caso del índice de calidad de vida e inflación, el algoritmo por *RandomTree* podría tener un mejor desempeño; motivo por el cual se aplicará también este algoritmo. Además, consideremos el caso especial de la expectativa de vida y calidad de vida segmentados por el nivel de PIB per cápita, ya que fue entre estas tres variables donde se encontró la mayor correlación.



En las siguientes figuras se presenta la relación entre la esperanza de vida al nacer y el Índice de calidad de vida. En el primero, se observa la presencia de dos cúmulos o clústeres principales, en los cuales se tienden a agrupar los datos.

En el segundo gráfico se corresponde el nivel de PIB per cápita donde se aprecia que, en efecto, el nivel de PIB es una buena etiqueta para poder describir los clústeres que se forman en los datos.


Expectativa de vida
Índice de calidad de vida


Nivel de PIB
Regular
Alto
Expectativa de vida
Índice de calidad de vida

4. Aprendizaje automatizado

Esta sección está dedicada a aplicar los algoritmos de aprendizaje automatizado sobre las variables de interés determinadas en la sección anterior.

Se obtienen modelos de regresión simple señalados previamente. Luego, se ajusta el algoritmo de KNN para obtener los clústeres deseados. Adicionalmente a las gráficas en el conjunto de prueba, se evaluará el desempeño de ambos modelos por medio de validación cruzada y matrices de confusión, respectivamente.

Por otro lado, para el caso de segmentación de interés por el nivel del PIB per cápita, se aplica el método del codo para corroborar que el número de clústeres propuestos es correcto, este método usará K–Means++, ya que se especializa en agrupación de manera precisa y eficaz.

4.1. Regresión lineal

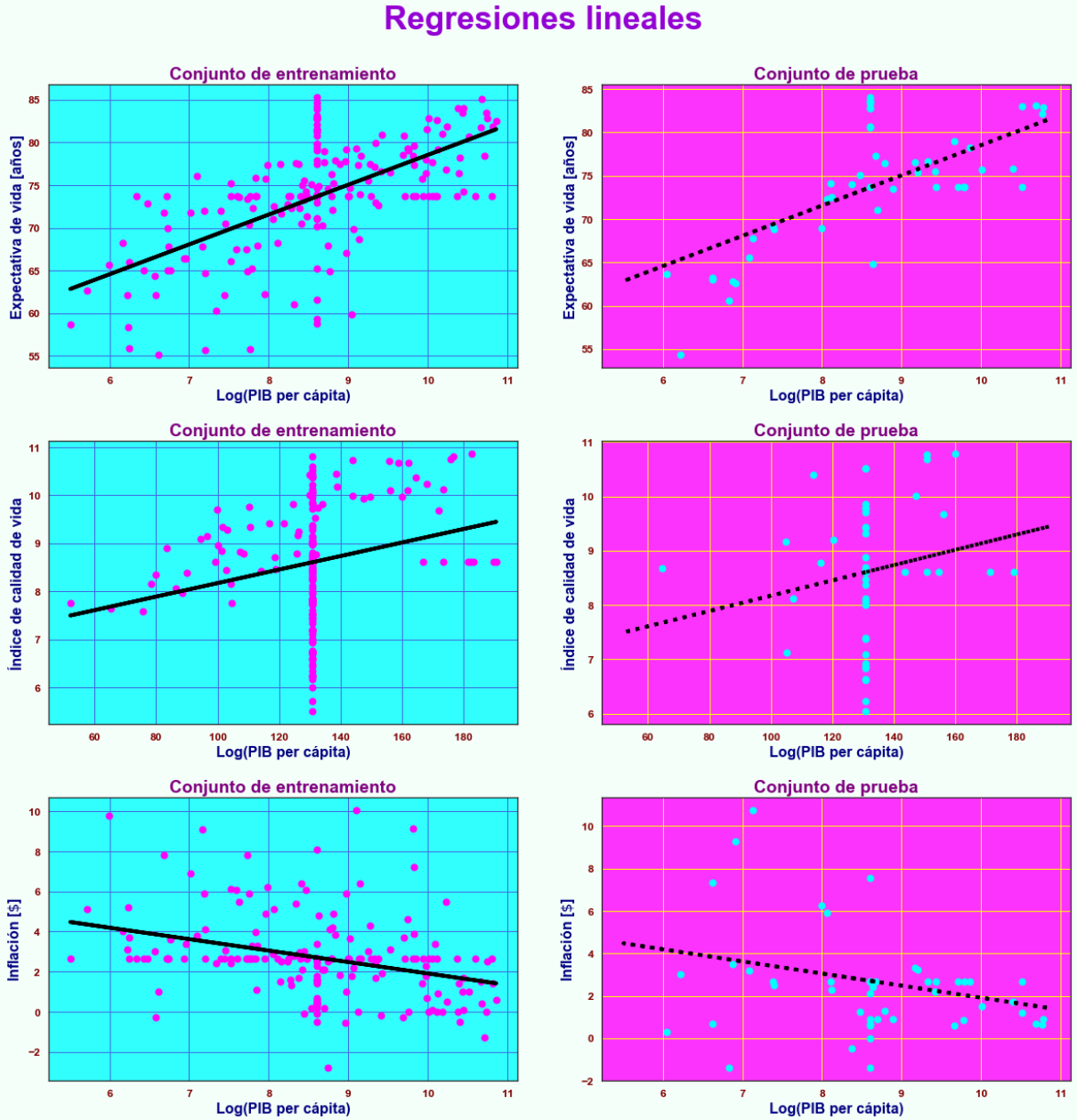
Como se había establecido, los modelos de regresión lineal para cada variable se presentan en la siguiente figura. Se divide por el conjunto de entrenamiento y el de prueba.

En la tabla de abajo se muestran los niveles de exactitud de cada uno de los modelos en ambas partes de los datos.

	Modelo 1	Modelo 2	Modelo 3
Entrenamiento	0.40	0.06	0.11
Prueba	0.52	0.03	0.08

De manera previa se estableció que el modelo de mayor interés es el primero, donde vemos que es el único de los modelos que no muestra evidencia directa de $overfitting$, en comparación con los otros dos modelos, los cuales se desempeñan mejor en el conjunto de entrenamiento que en el de prueba.

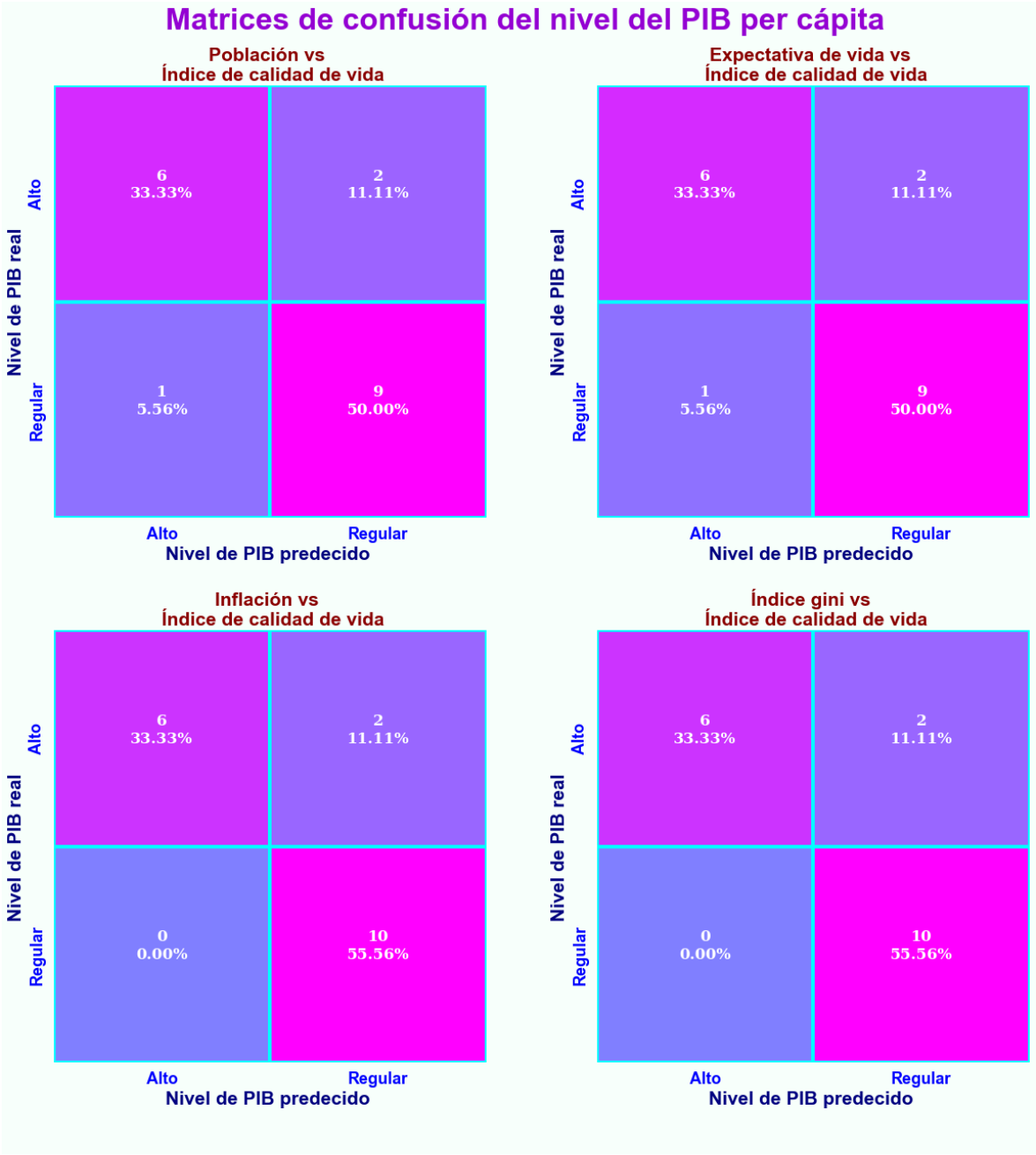
Por otro lado, es importante señalar que los *scores* tienden a ser bajos, pero esto se debe al método que utilizamos para rellenar los datos faltantes: se usó la media de las observaciones de cada variable para rellenar los datos, puesto que no afecta al estimador, pero sí para medir los errores cuadráticos medios.



4.2. Agrupamiento

Al igual que en el caso de regresiones lineales, se sigue el mismo procedimiento para obtener el clasificador por KNN.

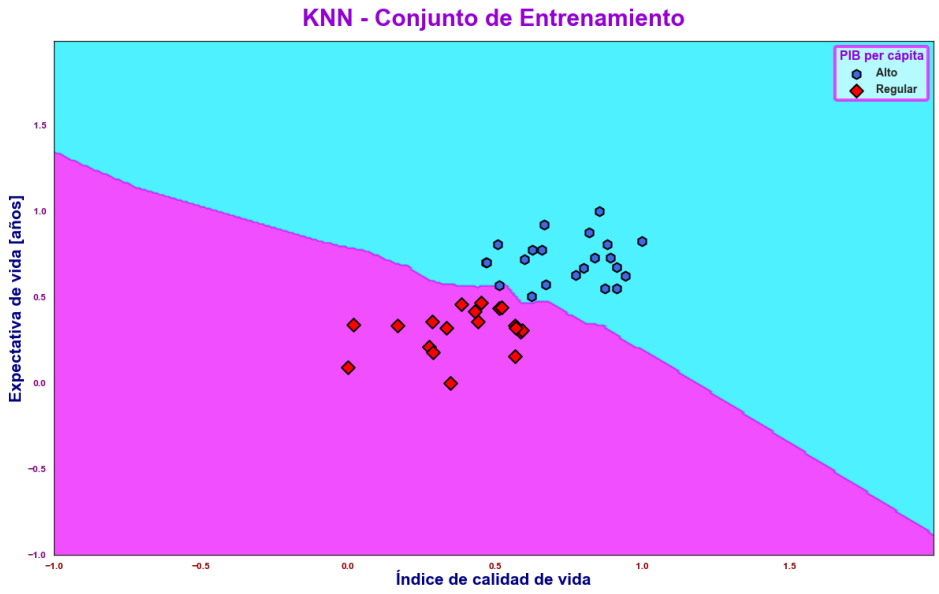
Ahora, se presenta el desempeño del clasificador tanto en el conjunto de entrenamiento como en el de prueba. Para ello, primero se utiliza una visualización cuantitativa a través de las matrices de confusión:

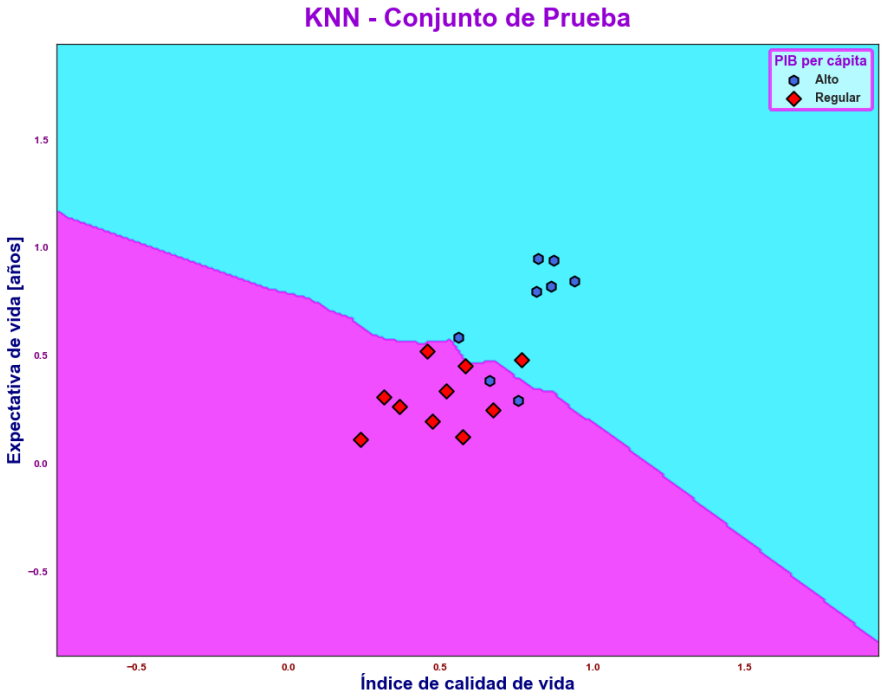


Se aprecia que en todos los casos, el porcentaje de acierto es del 88.89 %, pero la diferencia radica en qué variable se predijo erróneamente. Notemos que en las predicciones correctas, todas obtuvieron el mismo porcentaje para predecir los países con un nivel de PIB per cápita alto, pero la mitad de los modelos difieren del otro par en una predicción correcta para los países con un nivel regular. En general, se considera que el desempeño de los modelos es lo suficientemente aceptable; el modelo podría mejorar considerando una base de datos más grande que contenga los datos completos a diferencia de la información que se está considerando.

Enseguida, se presentan los resultados de la aplicación del algoritmo KNN. La figura de arriba corresponde al conjunto de entrenamiento y la de abajo al de prueba.

A pesar de no haber muchos datos en el conjunto de entrenamiento, el desempeño en la prueba es aceptable, como se vio en las matrices de confusión ya que sólo hubo errores en tres predicciones. Sin embargo, el porcentaje de acierto es alrededor del 90 %, por lo que el desempeño es satisfactorio.





Por otro lado, si se analizan las fronteras de los conjuntos de entrenamiento y prueba

y cómo están distribuidos los datos segmentados por el nivel de PIB per cápita, da indicios de que un mejor algoritmo de clasificación sería un árbol aleatorio. No obstante, un mejor desempeño se obtiene de aplicar el bosque aleatorio, el cual se ajusta para las necesidades planteadas, dado que no se busca predecir datos fuera de la muestra puesto que se pretende dar un enfoque descriptivo.

Ahora bien, se obtienen los modelos tanto por *Random Forest* como por *Gradient Boosting Decision* (GBD). En este caso, se toman todas las variables numéricas como características y el nivel de PIB per cápita como la variable objetivo.

Para el algoritmo GBD se consideran dos modelos para distintos parámetros. En la tabla de abajo se muestra la exactitud de los tres algoritmos:

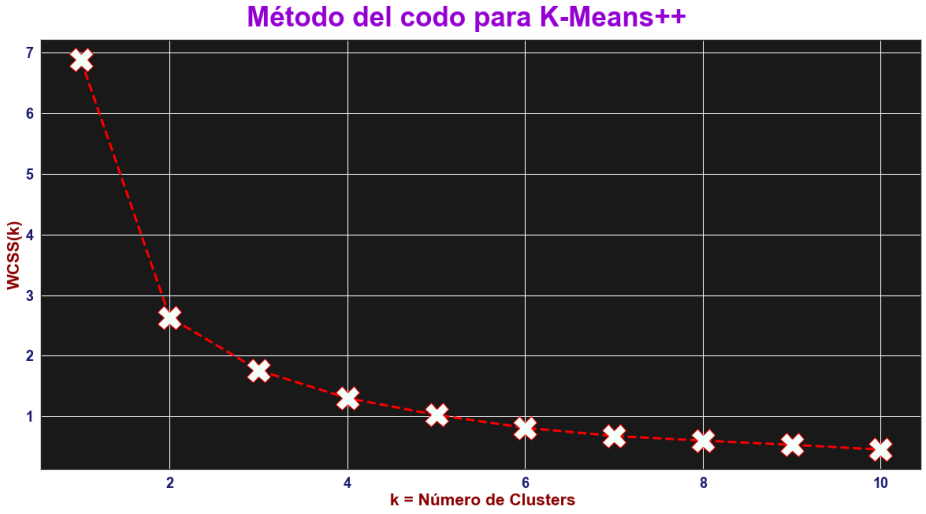
Exactitud	Random Forest	GBD 1 (profundidad=1)	GBD 2 (aprendizaje=1 %)
Entrenamiento	0.96	1.00	1.00
Prueba	0.93	0.87	0.87

A pesar de que hay señales de sobre-ajuste, el desempeño del modelo es mejor en el caso del Random forest en comparación con los modelos previos que tienen un desempeño por debajo del 90 % . Además, contra todo pronóstico, el algoritmo de GDB no obtuvo tan buenos resultados como $RandomForest$. Para mejorarlo es cuestión de determinar qué parámetros conviene cambiar puesto que ofreciendo dos distintos parámetros se obtuvo el mismo resultado, por lo cual inferimos que tiene que ser otro parámetros en el algoritmo.

4.3. Desempeño

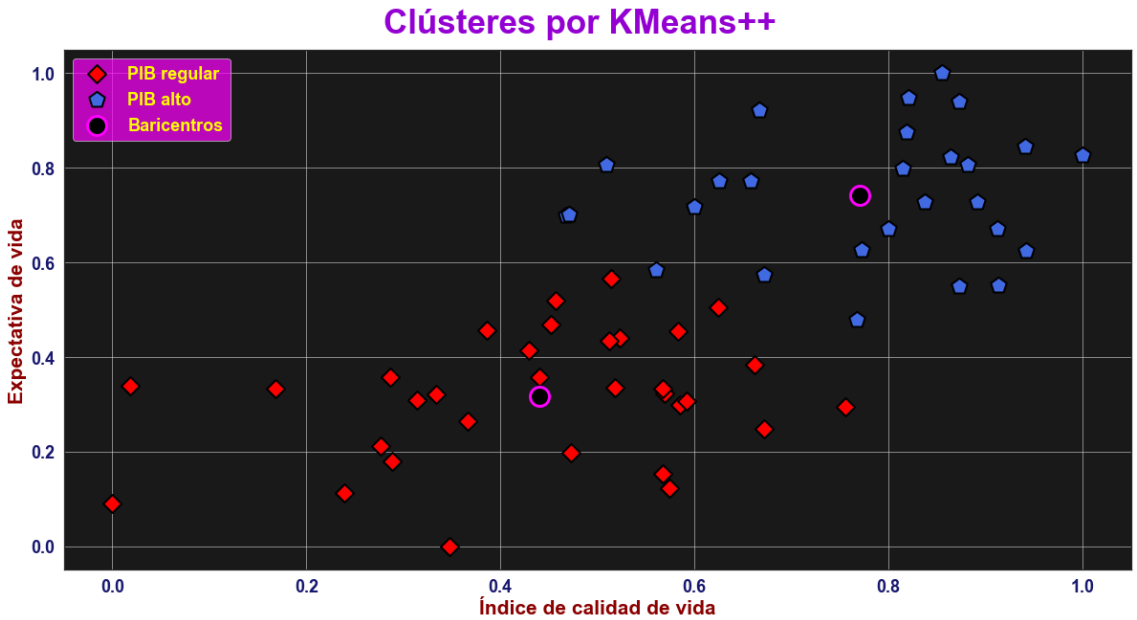
Una pregunta que debe plantearse es si la propuesta de segmentar los datos en dos de tres grupos de acuerdo con el nivel del PIB per cápita es adecuada o acertada con los clústeres que se tienden a formar en el caso de la expectativa de vida en función del logaritmo del PIB per cápita, puesto que sólo se intuyó de manera cualitativa.

Por el contrario, tengamos en cuenta que, debido al método de cómo se decidió el criterio para clasificar, sólo se cuenta con dos niveles observados en los datos. Para lograr encontrar una respuesta, se utiliza el método del codo para el algoritmo K-Means++, debido a que este algoritmo calcula clústeres independientemente de que estén segmentados o no.



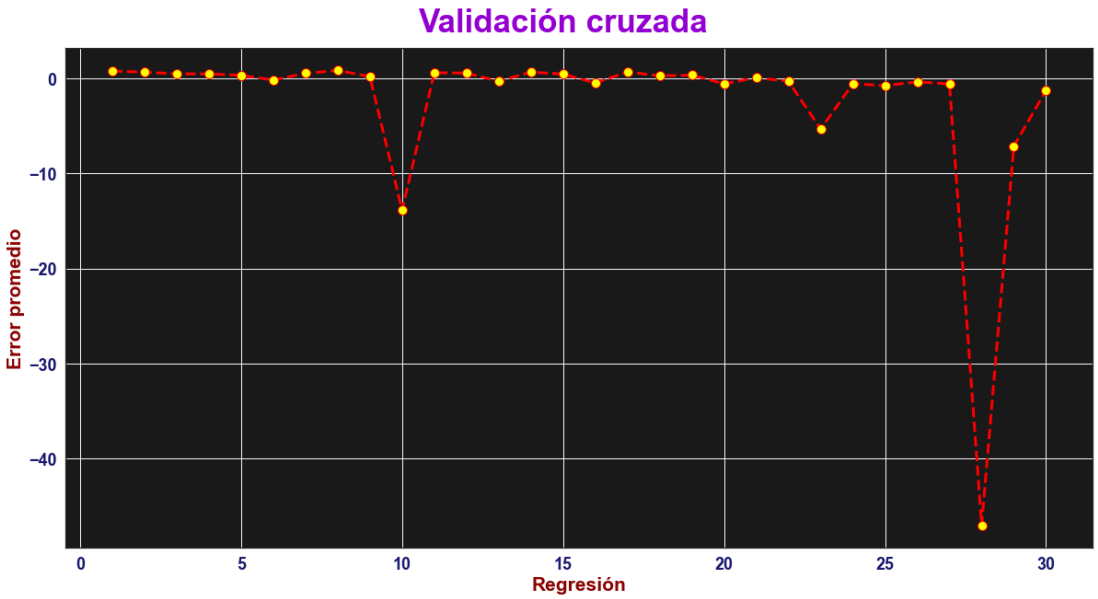
En efecto, como lo muestra la figura de arriba, el número óptimo de clústeres óptimo es $K = 2$, por lo que se asegura que segmentar los datos a dichos niveles de PIB fue adecuado, sin embargo, téngase en cuenta que esto no necesariamente tiene valor explicativo, pero sí descriptivo.

La figura de abajo muestra los grupos obtenidos por el método K-Means++, así como sus respectivos baricentros.



En general, el desempeño del modelo no varía mucho entre las distintas configuraciones del conjunto de entrenamiento y de prueba, salvo tres casos que son los que afectan mucho la media de los *scores*, así como la desviación estándar de éstos. No obstante, se aprecia que el modelo fue calculado de manera adecuada y se puede describir de manera decente

la mayoría de los datos. Asimismo, para mejorar el desempeño, se puede optar por otro método para rellenar los datos faltantes.



5. Conclusiones

Se obtuvo un modelo para describir adecuadamente la relación lineal entre la expectativa de vida y el logaritmo del PIB per cápita, que no exhibiera evidencia de sobre-ajuste. Aunque, debido al método de rellenado para los datos faltantes, el desempeño de este modelo no es tan bueno, obteniendo un *score* alrededor del 50 % , pero un desempeño similar para distintas configuraciones en el conjunto de entrenamiento con una media aproximada de -2 años.

Adicionalmente, los modelos obtenidos a partir de la tasa de inflación y el índice de calidad de vida en función del logaritmo del PIB per cápita mostraron comportamiento con $overfitting$ con un desempeño deficiente –$score$ menor al 30 % en ambos casos–, probablemente debido a los mismos problemas que en el primer caso de interés o por selección del conjunto de entrenamiento.

Finalmente, se mostró la evidencia de clústeres entre todas las distintas variables económicas segmentadas por el nivel del PIB per cápita, obteniendo clasificadores con un desempeño de acierto cercano al 90 % para todos los casos. Sin embargo, también se mostró, por medio del método K-Means++, que la propuesta de segmentación por el nivel de PIB per cápita es adecuada para describir los clústeres en el caso de interés de la expectativa de vida y el logaritmo del PIB per cápita. Además, se logró mejorar el modelo de clasificación a través de un algoritmo $RandomForest$, a pesar de haber señales de $overfitting$.