Trabajo Final IVSII

Saire - Roitman - Yudcovsky

2C 2024

1 Enunciado

Elegimos tratar el primer caso del módulo de visualización de datos, que plantea lo siguiente:

A partir del dataset en cuestión, deberán responder dos preguntas:

a) ¿Cuál es la clusterización más efectiva para agrupar los países?

b) A partir de la clusterización identificada en la pregunta anterior, ¿cómo se compone cada cluster? Es decir, deberán caracterizar cada uno de los clusters.

Deberán construir al menos dos visualizaciones (una para cada pregunta, lógicamente, puede ser más). Deberán justificar detalladamente las decisiones de diseño tomadas para la visualización de datos presentada: tamaño, color, escala, etc. A su vez, deberán explicitar los criterios utilizados para organizar y transformar los datos (p. ej. agrupaciones, cálculos, etc.).

2 Análisis del dataset

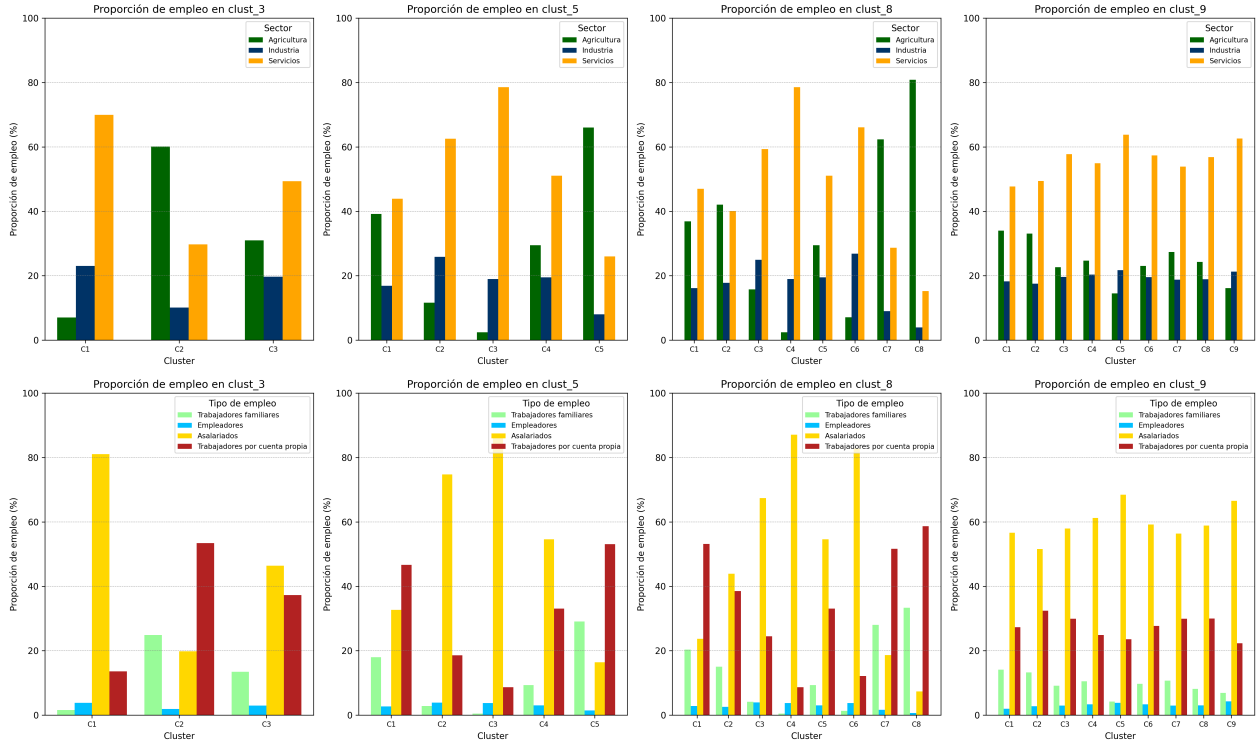
Vamos a detallar el análisis de datos que realizamos a partir del objetivo general del trabajo:
Objetivo: dar una caracterización de cada cluster en función de la clusterización más efectiva para agrupar a los países. Detallamos a continuación los pasos a seguir previo a la elección de la clusterización.

- Importamos el dataset y nos fijamos cuáles son los países dentro del mismo y a qué cluster pertenecen en cada una de las 4 clusterizaciones.
- Juntamos aquellas subclases del dataset que, para cada país, sumaban 100. Es decir, todas las categorías que se llamaban "prop_ :" tenían que pertenecer a una subcategoría tal que la suma de los elementos de tipo "prop_ :" de la misma sumen 100, pues se trata de una proporción. Pensando un poco en términos económicos, identificamos dos subclases del dataset como:
  - Sector productivo: Para cada país, la proporción de los tres sectores productivos económicos. Estos son agricultura, industria y servicios.
  - Sectores laborales: Para cada país, la proporción de los cuatro sectores laborales a los que pertenecen los miembros de su población activa. Estos son empleadores, asalariados, trabajadores familiares y trabajadores por cuenta propia.
  - Luego, chequeamos que la suma de las columnas de la subclase para cada dataset sumen el 100%. Efectivamente (y con un error numérico esperado de 1e-3), las sumas de ambas subclases sumaban aproximadamente 100 para todos los países.
- Luego, chequeamos que la suma de las columnas de la subclase para cada dataset sumen el 100%. Efectivamente (y con un error numérico esperado de 1e-3), las sumas de ambas subclases sumaban aproximadamente 100 para todos los países.

Posteriormente, buscamos la forma adecuada de representación de estas dos subclases para poder sustraer información en pos de responder las preguntas pertinentes a la investigación.

- Como son proporciones, nos pareció una buena elección realizar un gráfico de barras. Son útiles para comparar los porcentajes entre los distintos clusters y sectores.
- Para realizar los gráficos agrupamos por la columna 'clust_n' haciendo un promedio de los valores de cada columna de los países que conforman cada cluster.

- Los colores fueron elegidos según nos pareció que mejor representaban a cada sector. Por ejemplo, el sector agricultura lo representamos con un verde oscuro.



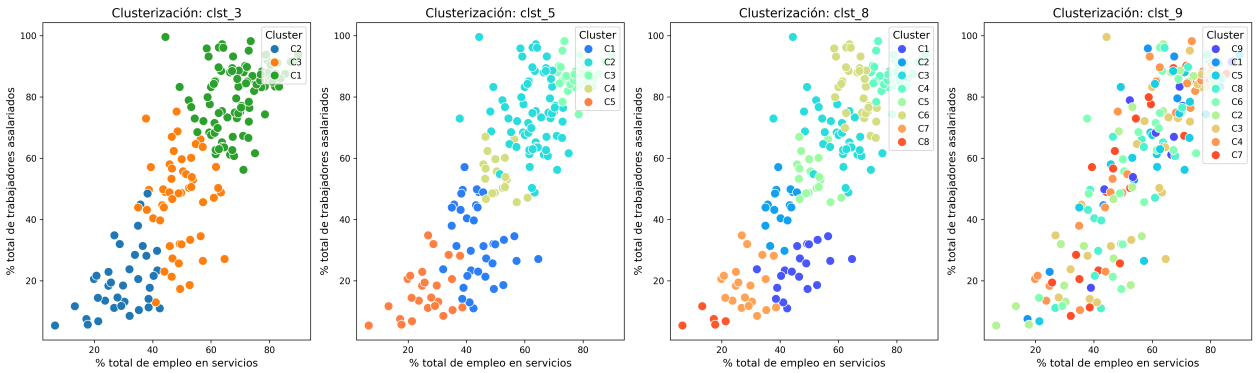
- En el código, también se encuentra una versión del gráfico con las subclases 'apiladas' que ayudan a entender cómo se distribuyen las proporciones dentro de un cluster. Pero para la finalidad del trabajo nos pareció una mejor representación de los porcentajes los gráficos que mostramos en el informe.

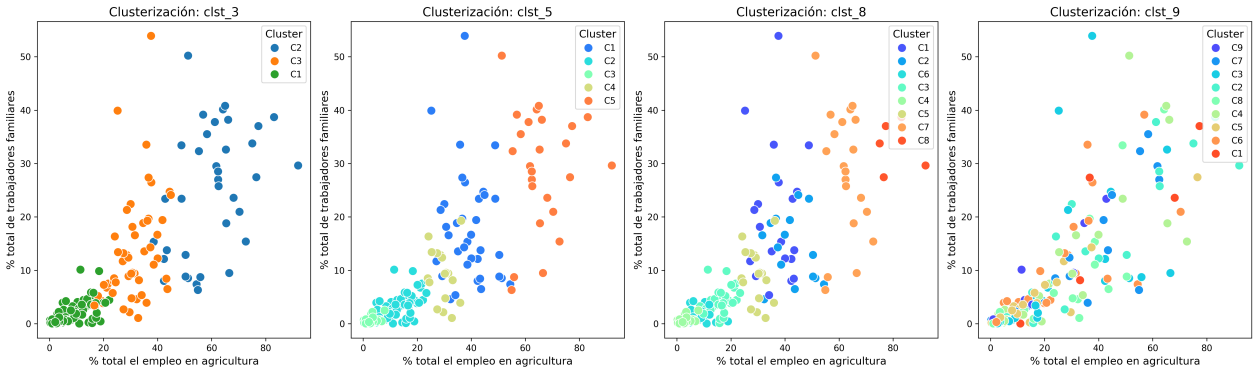
3 Eligiendo la clusterización

Nuestro objetivo es develar cuál es la composición de alguno de los 4 clusters a partir de la visualización de los sectores productivos y las divisiones en clases de actividades de los trabajadores.

Una vez que obtuvimos las representaciones anteriores, que muestran las proporciones en las diferentes categorías de cada uno de los clusters en lo que refiere a sectores de empleo, entes económicos, y la relación entre estás dos subclases del dataset, queremos decidir cuál de las cuatro clusterizaciones "caja negra" es la mejor para identificar las composiciones de los clusters. Para llegar al objetivo final, quisimos generar scatter plots según las variables que visualmente encontrábamos relevantes para su comparación. El inconveniente es qué, según la clusterización que se elija, puede que las nubes de puntos de cada cluster no sean homogéneas, ya que para dos clusterizaciones distintas, dos variables pueden tener mucho o poco que ver.

Analicemos cuáles pueden ser los "vs" de los scatter. Excepto en clust_9, podemos observar que la proporción de sector agrícola con la de empleo de subsistencia (prop_familiar) están íntimamente relacionadas. A su vez, notamos lo mismo si comparamos sector servicios con la proporción de asalariados. Además, los países que tienen mucha parte industrial y servicios, suelen aparecer en los mismos clusters, además de tener muchos asalariados y empleadores. Ambas relaciones parecen ser pseudo-lineales.





3.1 Descartando clust_9 y clust_3

Con esta información, y utilizando nuestros conocimientos generales de la temática, podemos descartar un par de clusterizaciones:

1) clust_9

- En la primer subclase, que refiere a la división de la masa de los trabajadores en sectores productivos, podemos observar que los 9 clusters tienen proporciones muy parecidas en los tres sectores (agricultura, industria y servicios). Tanto el gráfico de barras apiladas como el de separadas muestran esta similitud que no permite distinguir correctamente cuál es el contenido de cada uno.
- En la segunda subclase, hay un comportamiento similar, por lo que sigue siendo una mala elección a la hora de determinar la composición.
- Cuando comparamos prop_asal vs prop_emp, vemos las nubes de puntos correspondientes a cada cluster completamente mezcladas e indistinguibles de las otras.

2) clust_3 - En los scatter plots, podemos visualizar que los tres clusters se dividen de forma distinguible. Sin embargo, para datos que en términos sociales pueden tener muchas diferencias entre sí, como lo son los países y el bienestar económicos. Por tanto, utilizar solo 3 clusters puede volver dificultoso el análisis, caeríamos en una simplificación del problema de categorización.

3.2 ¿Con cuál nos quedamos? ¿clust_5 o clust_8?

En primer lugar, podemos notar que los tres clusters que se agregan del clust_5 al clust_8 parecen ser particiones de un cluster anterior en todos los scatter plots. Por ejemplo, cuando vemos el C5 del clust_5, se parte en C7 y C8 para clust_8. Entonces, si dividen "en casi los mismos grupos", ¿por qué nos quedaríamos con uno sobre otro?

- En el sentido explicativo, quizás más clusters diferencian mejor dos grupos que en la práctica estaban menos relacionados, pero de haber sido así, al agregar tres centroides más, algunos datos se hubiesen ido a un cluster y otros a otro. Por lo que puede observarse en varios de los cruces que hicimos, eso no parece ocurrir, y podría concluirse que más clusters no son necesarios por este mismo motivo.
- En segundo lugar, C6, C4 y C3 están mezclados en los scatter plots, por lo que tal vez clust_8 no es la mejor elección cuando se trata de un clasificador.

Por lo tanto, podemos concluir que en estas condiciones, lo mejor sería optar por el **clust_5**.

4 Develando la caracterización del cluster elegido

Una vez que elegimos el cluster de 5 centroides, quisimos intentar entender qué tenían en común dentro del dataset. Cuando vimos los países que tenían mayor cantidad de sector agrícola y economía familiar en el C5, notamos que correspondían a países muy pobres (Burundi, Rep. Centroafricana, Chad, entre otros), mientras que por el contrario encontrábamos países europeos o países Americanos como Argentina en el C3. Por conocimiento en cultura general, nos pusimos a pensar en qué categorías Argentina podría estar con países europeos.

En ese momento, uno de nosotros (que es medio obsesivo con datos curiosos), se acordó que en los rankings mundiales de desarrollo humano, y que los países africanos que aparecían como C8 en clust_8 (los que tenían menos asalariados, y por lo tanto menos servicios), eran los que recordaba como los peores cinco países en Índice de desarrollo humano.

El Índice de Desarrollo Humano (IDH) es una herramienta que sintetiza el progreso de un país en tres aspectos fundamentales del desarrollo humano: una vida larga y saludable, el acceso a la educación y un nivel de vida adecuado. Se calcula como la media geométrica de índices normalizados que reflejan los logros en estas áreas, utilizando indicadores como la esperanza de vida al nacer, los años promedio y esperados de escolarización, y el ingreso familiar disponible o consumo per cápita. Así, el IDH permite comparar la esperanza de vida, la educación, la alfabetización y el nivel de vida entre países a nivel global.

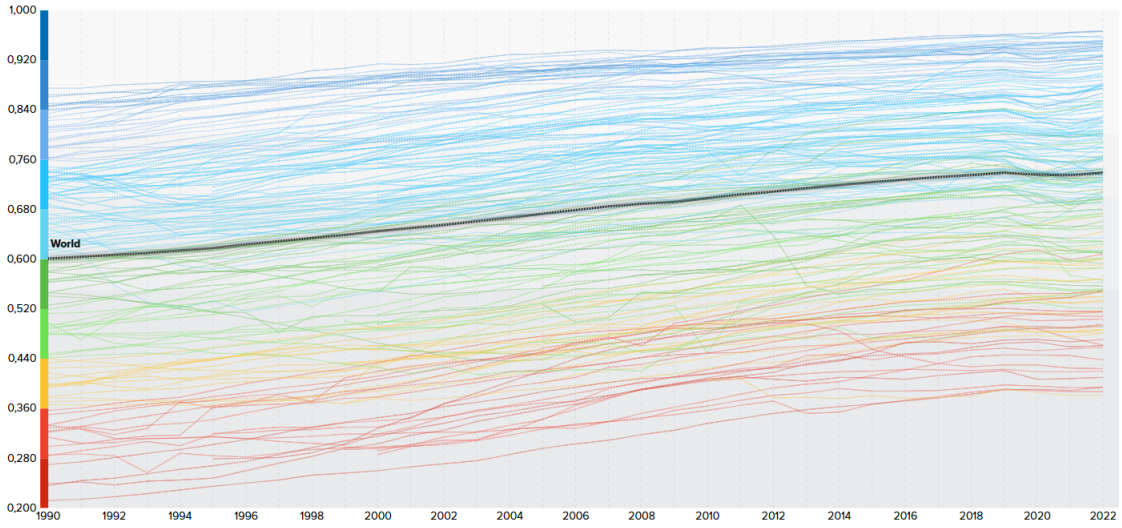
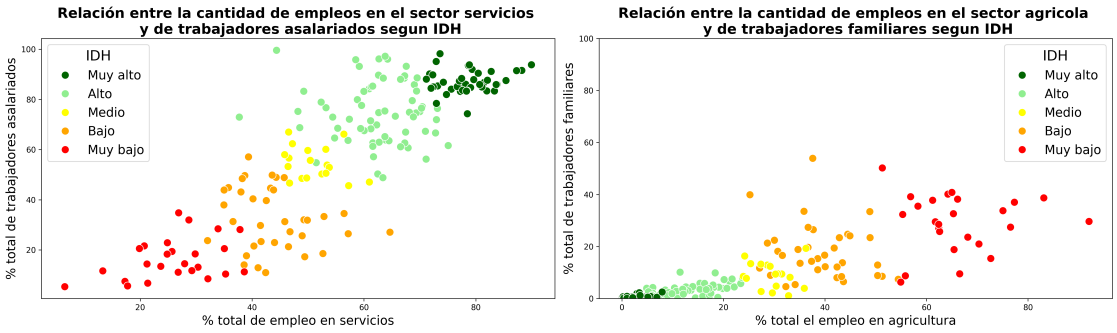


Figure 1: Índice de desarrollo humano por país (y en el mundo) desde 1990

En el código, analizamos el data set del Banco de Datos de la ONU del IDH en 2019 (ya que el data set de este trabajo tenía un nombre con 2019), y observamos que las similitudes con nuestros países eran contundentes.

Lo más sorprendente es que el data-set de la ONU diferencia los países en cinco categorías: very high, high, medium, low, and very low. Si vemos qué países corresponden a cada categoría en este nuevo data set y los cinco clusters que teníamos en 'clust_5', podemos emparejarlos de la siguiente manera: C3 con "very-high", C2 con "high", C4 con "medium", C1 con "low" y C5 con "very low".

Ahora que sabemos que representan los clusters podemos rehacer los scatter plot de antes mejorados.



En conclusión, no sólo encontramos la caracterización correcta para nuestros datos, sino que también la mejor de las 4 clusterizaciones.

5 Bibliografía

- Banco de datos de la ONU: `https://hdr.undp.org/data-center`
- `https://hdr.undp.org/system/files/documents/hdr2019overview-spanish.pdf`
- Anexo: Países por Índice de Desarrollo Humano (Wikipedia)