Universidad Nacional de San Martín

Escuela de Ciencia y Tecnología

Licenciatura en Ciencia de Datos

Trabajo Práctico Integrador

Estadística e Inferencia II

Modelado Bayesiano

Análisis Bayesiano del dataset Iris
Análisis del World Happiness Report

Estudiante:
Javier Spina
javierspina@gmail.com
jaspina@estudiantes.unsam.edu.ar

Profesora:
Fernanda Micucci

Cuatrimestre:
2do 2025

Noviembre 2025

Índice

1. Introducción

En el presente trabajo vamos a trabajar con modelos bayesianos para estudiar dos conjuntos de datos:

- El clásico dataset **Iris**, que contiene observaciones de flores del género Iris
- Una selección de variables del dataset **Quality of Government**, que compila datos de diversas fuentes de ciencias sociales en clave país-año

Cada dataset se estudia en partes separadas con objetivos distintos. El objetivo que tienen en común es el estudio de datasets con el flujo de trabajo bayesiano, desde modelos simples hasta modelos y aplicaciones más complejas.

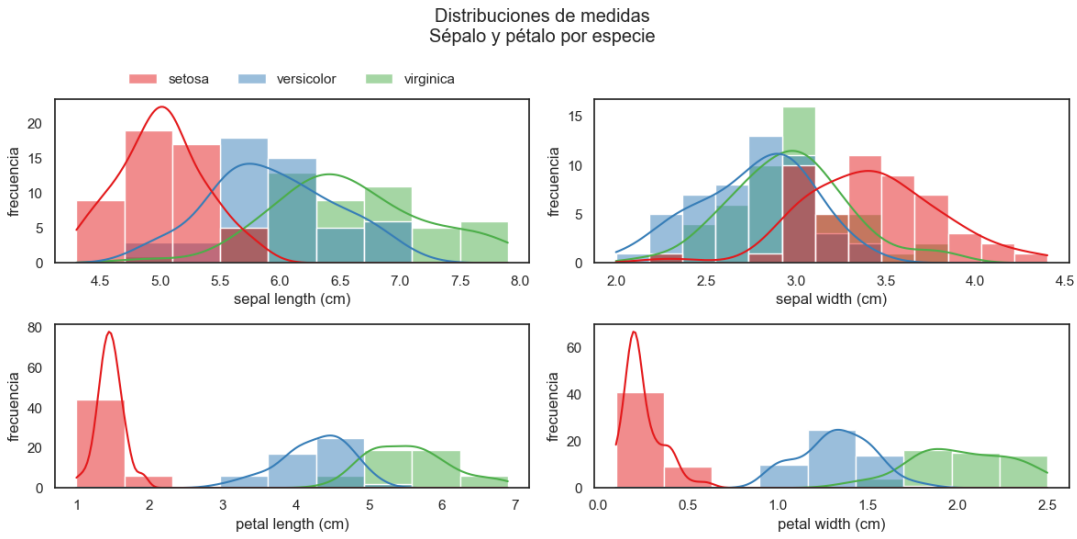
El análisis del dataset Iris busca formalizar y mostrar la estructura de un trabajo bayesiano, que construye desde el análisis exploratorio de las variables hasta un modelo bayesiano jerárquico, usando criterios de información para seleccionar el mejor modelo para hacer inferencia sobre los datos.

Para la segunda parte, vamos a utilizar lo aprendido sobre el flujo de trabajo bayesiano, pero esta vez siguiendo un camino que sale del puro entrenamiento estadístico. Aquí vamos a formular una hipótesis más de "mundo real" (aunque el dataset Iris no es sintético), más allá de la disciplina de Ciencia de Datos.

2. Descripción de los conjuntos de datos

2.1. Dataset Iris

Este conjunto es famoso en las ramas de estadística, datos y análisis numérico. En este caso lo descargamos con ayuda de `scikit-learn`. Cuando cargamos el dataset podemos ver 150 observaciones, sin nulos, con 4 variables numéricas que refieren a las medidas de las flores (pétalo y sépalo) y el target categórico que indica la especie a la que pertenecen.



2.2. Dataset Quality of Government

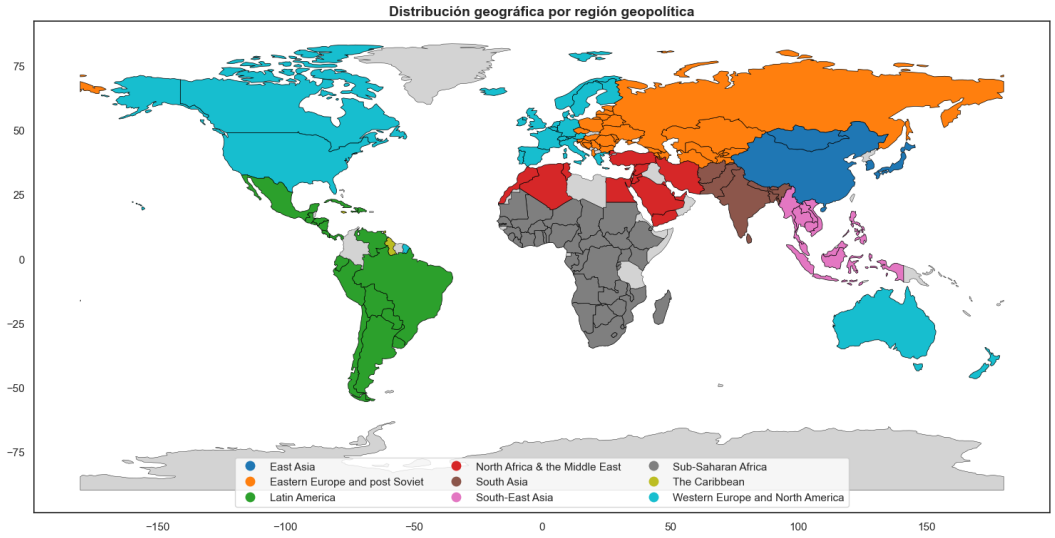
Es un conjunto de datos, creado por la Universidad de Gotemburgo, que registra decenas de miles de entradas de diferentes organizaciones a nivel mundial y las ordena

para cada país en el tiempo. De esta manera, cada observación es un país en determinado año y los distintos índices y métricas de múltiples entes a nivel regional y global.

Algunas de las organizaciones que son compiladas en este dataset son: Naciones Unidas, Banco Mundial, Organización Mundial de la Salud, World Happiness Report, entre muchísimas más.

La versión estándar del dataset tiene 279 columnas en total, y la versión completa tiene una cantidad de variables en la escala de los miles. Para el presente trabajo vamos a tomar un subconjunto de 10 columnas en total.

1. Variables de índice
    - `year`: año al que pertenece la medición. El dataset original comprende desde 1946 hasta 2024, actualizandose anualmente. El subconjunto de datos elegido para este informe usa datos desde 2005 hasta 2022.
    - `cname_qog`: nombre del país estandarizado por el Instituto QoG.
2. Variables categóricas
    - `ht_region`: región geopolítica a la que pertenece. Fuente: The Authoritarian Regime Dataset

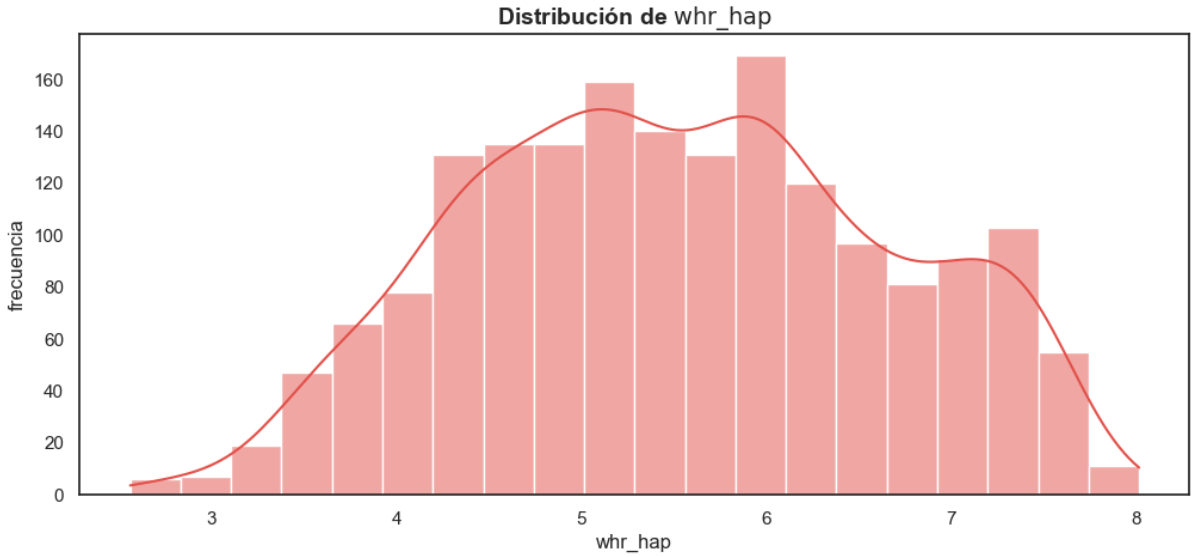


    - `ht_colonial`: bagaje colonial o potencia que dominó al país. Fuente: The Authoritarian Regime Dataset
3. Variables predictoras
    - `undp_hdi`: Índice de Desarrollo Humano. Tiene 3 componentes: **esperanza de vida, tasa de alfabetización e ingreso per cápita**. Fuente: United Nations Development Program
    - `top_top10_income_share`: proporción del ingreso del 10 % más rico de la población del país. Lo vamos a considerar una variable proxy para **desigualdad económica**. Fuente: World Inequality Lab
    - `vdem_gender`: Índice de empoderamiento político de las mujeres. Proxy para **igualdad de género estructural**. Fuente: V-Dem Project
    - `wdi_expedu`: Proporción de **gasto público en educación** sobre el total del PBI del país. Fuente: Banco Mundial

- `wdi_unempilo`: Tasa de **desempleo** en cada país. Fuente: Banco Mundial

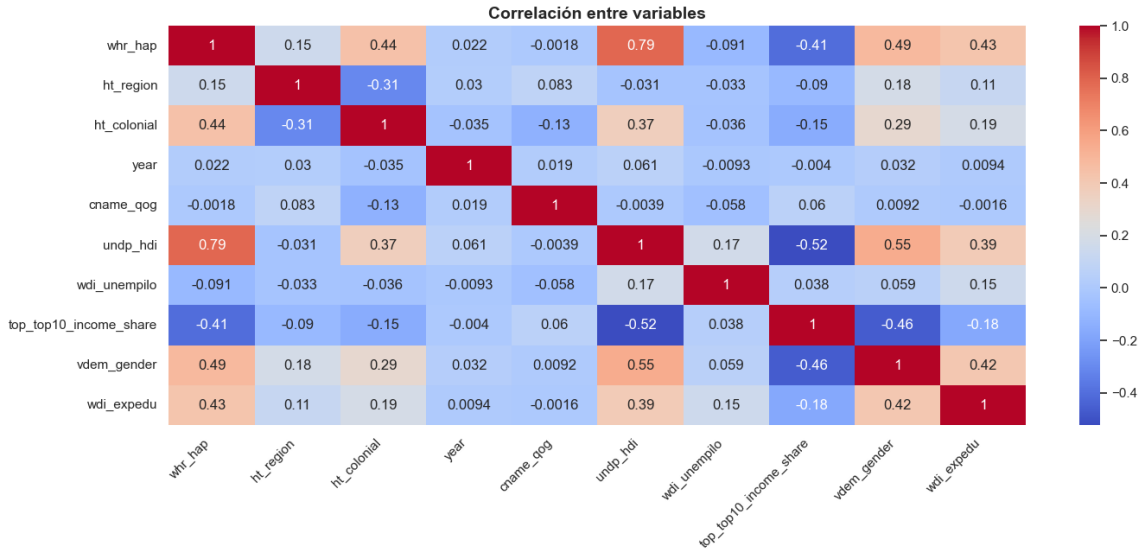
4. Target

   - `whr_hap`: promedio nacional del índice de bienestar subjetivo, también conocido como índice de felicidad. Individuos de un país son encuestados acerca de su percepción en una escala de 0 a 10. Fuente: World Happiness Report



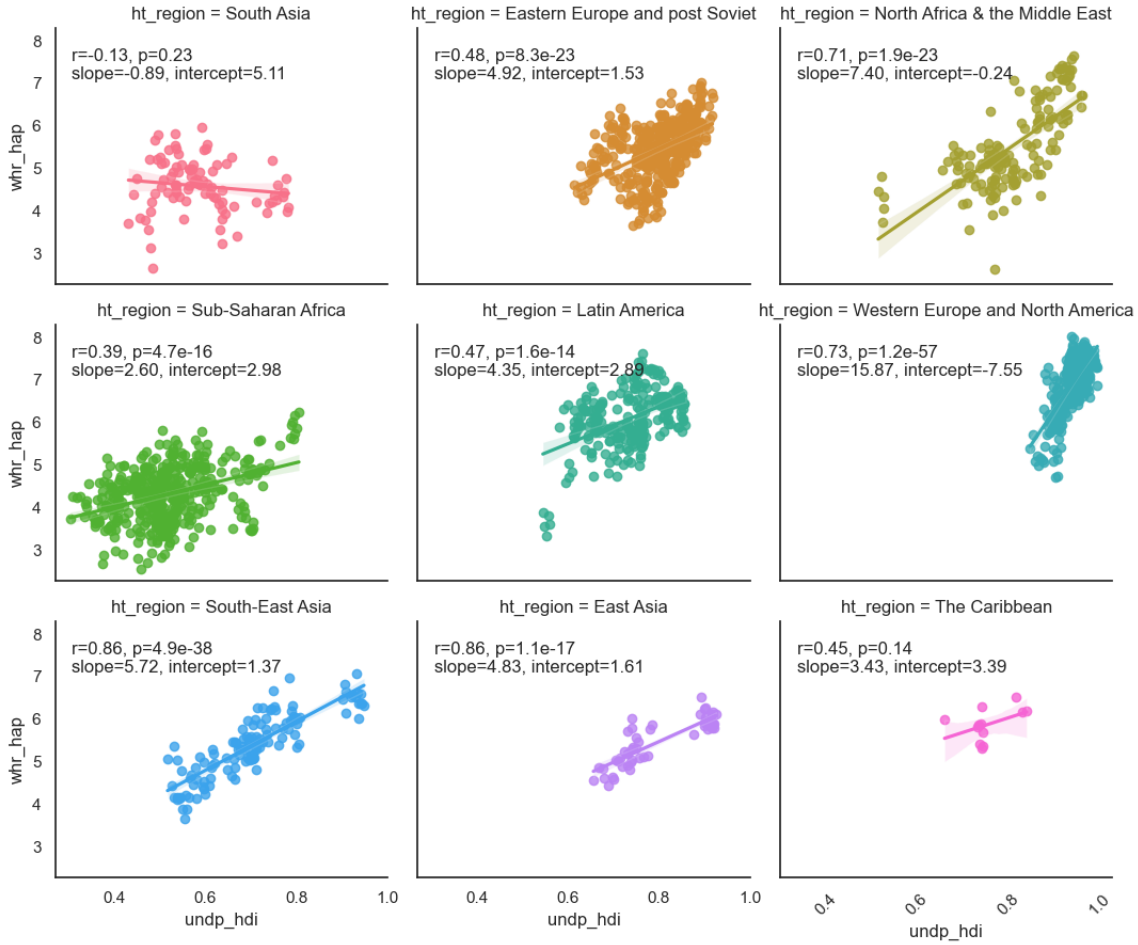
Podemos ver en el histograma de `whr_hap` que se desarrollan 3 picos (modas). Una alrededor de 5, otra cerca de 6 y la otra cerca de 7 y 8. Vamos a estar observando esto más de cerca utilizando un modelo de mezcla.

Veamos también la correlación entre las variables:



- La mejor correlación la llevan `whr_hap` $\sim$ `undp_hdi`, con un $r = 0{,}79$
- `ht_colonial` tiene mayor correlación que `ht_region` con el target, pero los modelos simulados con jerarquías basadas en `ht_colonial` tuvieron un desempeño muy pobre en comparación con `hr_region`
- `wdi_unempilo` es la variable más desacoplada del conjunto.

Veamos un desglose de la relación entre el índice de felicidad y el de desarrollo humano:



Sin ser experto en la materia geopolítica, podría considerar que hay regiones que pueden ser integradas, al menos para este análisis. De todas formas vamos a dejar los datos como los trae la fuente. La agregación no solo se justificaría por proximidad geográfica, sino por desempeños en esta relación para este ejercicio de modelado en particular. A primera vista: este asiático con sudeste asiático, y Latinoamérica con el Caribe. Como mencionamos antes, volveremos a este tema con el modelo de mezcla.

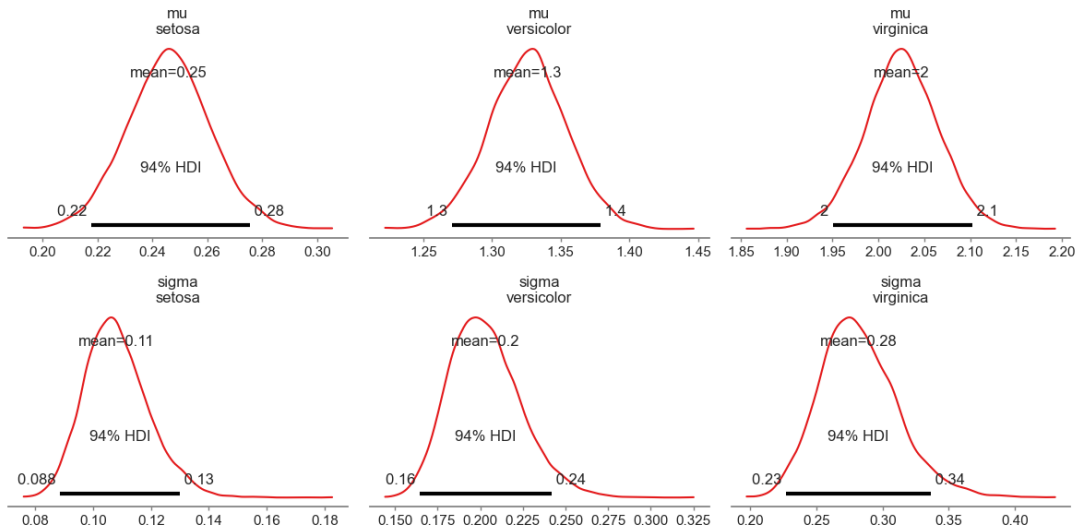
3. Modelos y análisis

3.1. Modelos para Iris

Construimos dos modelos distintos para esta parte, uno jerárquico y el otro no. En ambos casos, decidimos usar priors no informativos. Como variable predictiva elegimos el ancho del pétalo, por la separación natural que ya mostraba en el histograma.

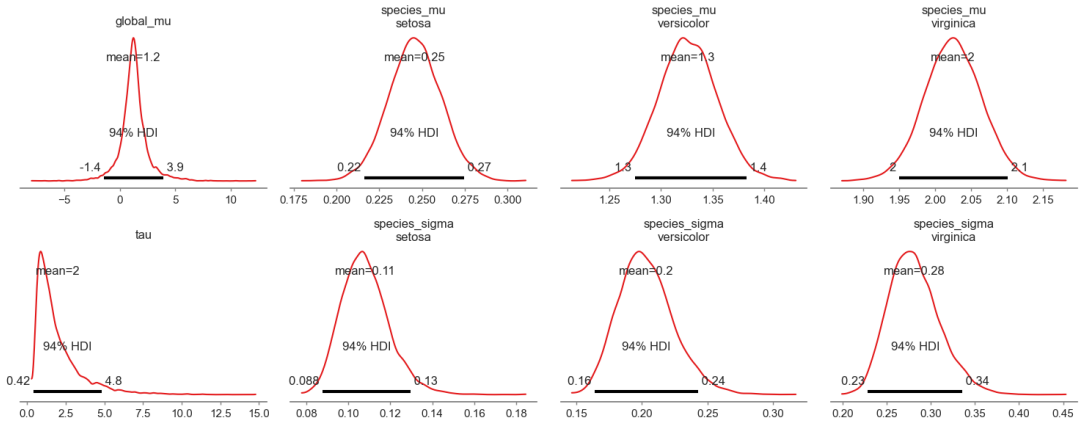
Modelo no jerárquico

$$\begin{aligned} \mu &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{HalfNormal}(0, 5) \\ y_{\text{obs}} &\sim \text{Normal}(f(\mu), f(\sigma)) \end{aligned}$$



Modelo jerárquico

$$
\begin{aligned}
\mu &\sim \text{Normal}(0, 10) \\
\tau &\sim \text{HalfNormal}(0, 5) \\
\mu_{\text{especie}} &\sim \text{Normal}(\mu, \tau) \\
\sigma_{\text{especie}} &\sim \text{HalfNormal}(0, 5) \\
y_{\text{obs}} &\sim \text{Normal}(f(\mu_{\text{especie}}), f(\sigma_{\text{especie}}))
\end{aligned}
$$



3.1.1. Análisis y evaluación

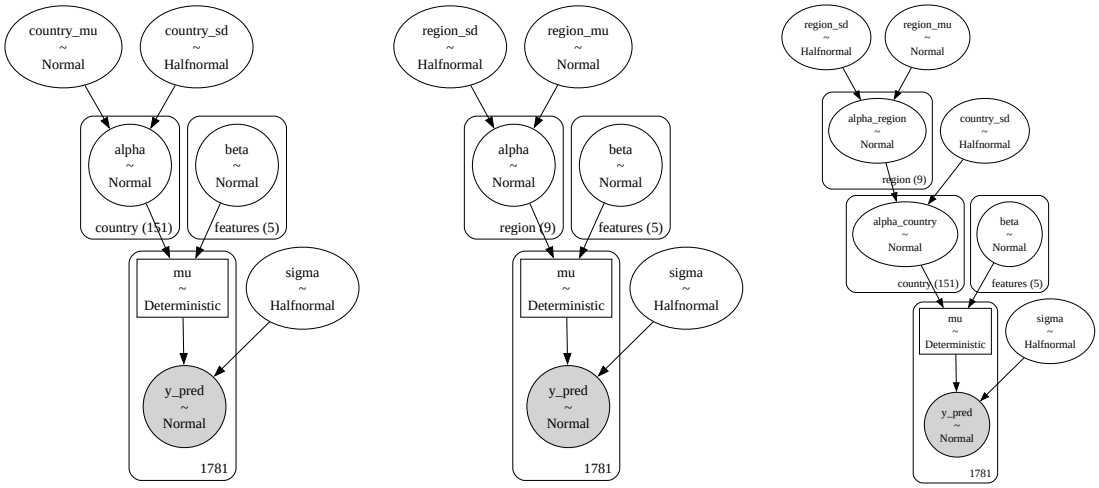
	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse
No jerárquico	0	40.52	5.97	0	1	10.37	0
Jerárquico	1	40.45	6.03	0.07	0	10.38	0.05

- No se observa *shrinkage* o contracción hacia la media en el modelo jerárquico. Los valores de la media μ para cada especie conservan su valor central y rango de confianza.
- El modelo no jerárquico tiene mejor ranking en la comparación de `arviz` usando el criterio de información `LOO`, aunque es una diferencia mínima. El modelo debe estar siendo castigado por su complejidad innecesaria.

3.2. Modelos para Quality of Government

Para esta parte contamos con 3 modelos jerárquicos. Un modelo *unpooled* no es de nuestro interés, ya que nos daría muy poca información para perseguir el objetivo que nos propusimos.

Queremos ver si la percepción de los individuos puede ser explicada por componentes estructurales de cada país. De esta manera, los gobiernos serían capaces de intervenir directamente con políticas de estado para mejorar el bienestar individual de los habitantes de su nación.

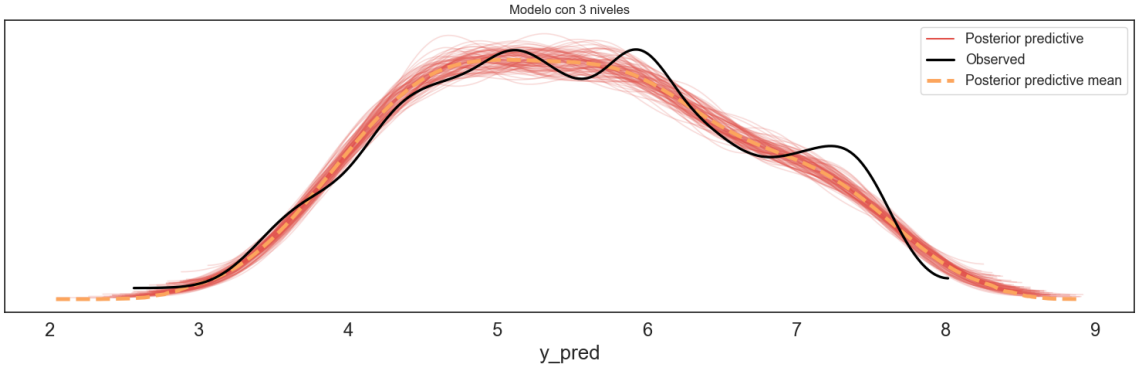


Para todos los modelos se escalaron las variables entre con $\sim \mathcal{N}(0, 1)$. Los beneficios son varios, pero para resaltar los más importantes: comparabilidad de los betas (pesos) de las componentes del modelo y para la convergencia de las cadenas `MCMC`.

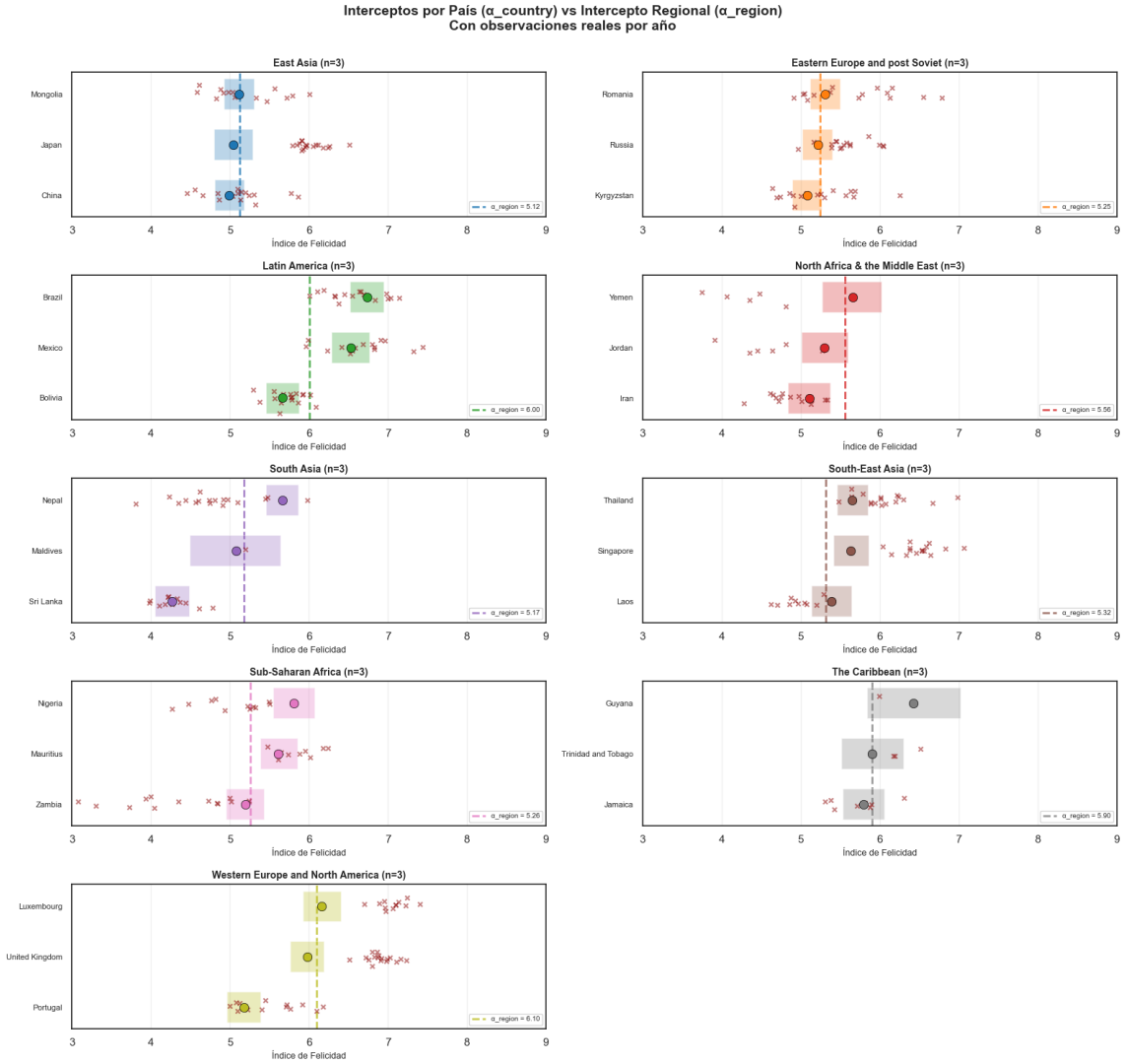
3.2.1. Análisis y evaluación

	rank	elpd_loo	p_loo	elpd_diff	weight	warning
Región → País	0	-865.51	153.48	0	0.59	True
Por País	1	-868.53	154.80	3.02	0.37	True
Por Región	2	-1449.97	16.04	584.46	0.04	False

El modelo con jerarquía por región y país terminó siendo el mejor en esta ocasión. Está muy cerca del modelo jerárquico por país, pero claramente la complejidad agregada suma valor a la inferencia, de lo contrario hubiera sido castigada por la evaluación. Seguiremos el análisis basado mayormente en este modelo.



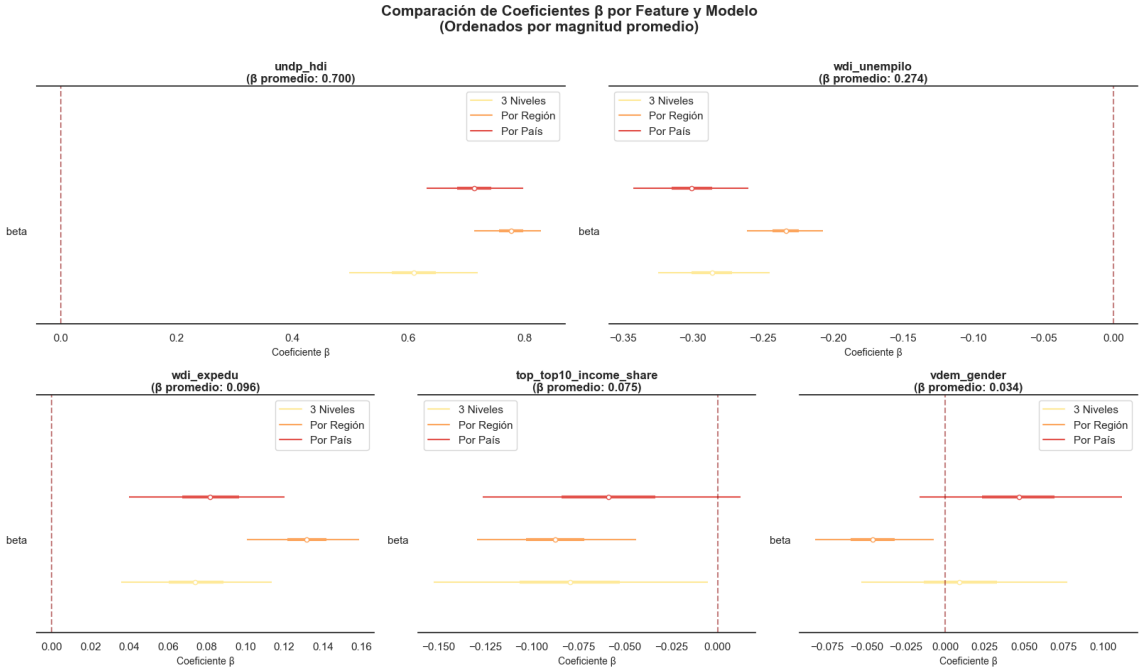
Las muestras de la posterior predictiva trazan de manera aceptable los datos observados. No es capaz de capturar los picos multimodales, pero persigue bien la trayectoria de manera suave. Buen desempeño en capturar la altura de la forma de `y_pred`.



En este caso sí podemos ver contracción de los valores de los interceptos de cada país hacia la media regional. A su vez es interesante destacar otro fenómeno: los *underperformers* y *overperformers*. Son ejemplos de países que según la inferencia del modelo tendrían que tener valores de felicidad más bajos, pero no es el caso y entonces el modelo debe compensar colocándole un intercepto mucho más alto que los valores observados (*overperformers*, ejemplo: Nepal, Nigeria, Zambia, Yemén).

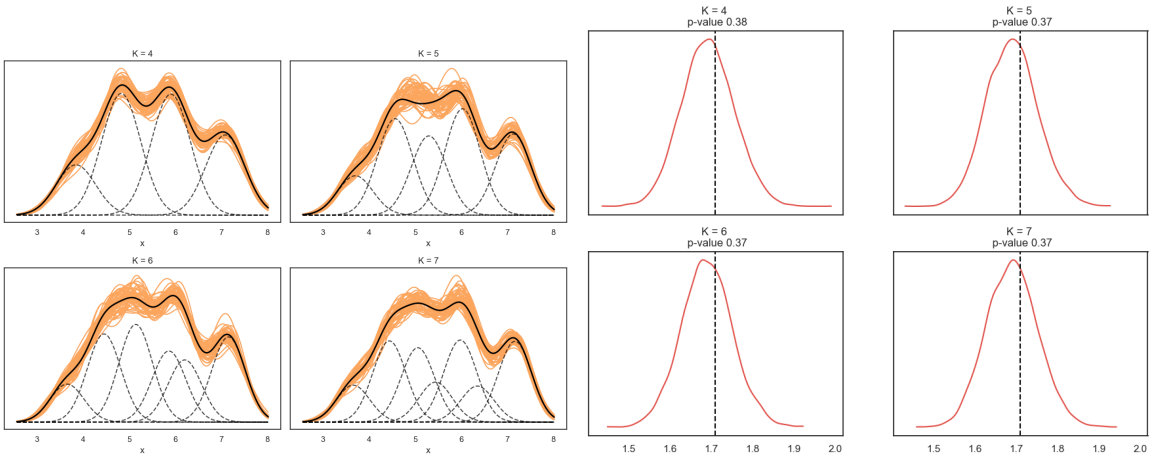
Por el contrario, hay ejemplos donde los predictores indicaban que sus valores de felicidad debían ser muy altos, pero no lo eran. Estos *underperformers* son, por ejemplo: Luxemburgo, Reino Unido, Singapur, Japón, Rumania.

Por último veamos los pesos comparados entre todos los modelos.



Las variables con mayor peso en el modelo son el índice de desarrollo humano y el desempleo. Los que tienen efectos muy reducidos son el gasto público en educación (posible redundancia con índice de desarrollo humano) y la desigualdad. El último caso es el proxy para la igualdad de género, cuyo rango de confianza incluye al cero, lo que equivale a decir que es probable que la variable no tenga peso para que el individuo califique con mayor puntaje su bienestar individual.

3.2.2. Modelo de mezcla



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
7	0	-2682.65	10	0	1	21.454191	0	False	log
6	1	-2682.81	10.27	0.16	0	21.29	0.33	False	log
5	2	-2685.15	10.78	2.50	0	21.14	0.88	False	log
4	3	-2689.39	8.42	6.74	0	21.06	2.55	False	log

El modelo de mezcla lo utilizamos para retomar la pregunta sobre la pertinencia de `ht_region` como buen categorizador de países. Según lo que vemos en estos resultados, tenemos que 7 grupos muestrean mejor que el resto de K_s probados (6, 5 y 4). Nuestra

sugerencia inicial de sumar Latinoamérica y el Caribe por un lado y sudeste asiático con este asiático deja la cantidad de regiones en 7. Merece un análisis más profundo y opinión experta, pero es un indicio de que vamos por buen camino.

4. Resultados y conclusiones

- Los modelos jerárquicos junto con los criterios de evaluación de los mismos son excelentes aliados del trabajo en ciencia de datos.
- Pudimos ver que si sobrecomplicamos un modelo que no necesita jerarquías, la evaluación nos va a informar esto.
- Pero si un modelo es jerarquizado pertinentemente, sumamos información extremadamente valiosa para el análisis.
- En este trabajo creamos muchos puntos de partida para continuar investigando. Deberíamos profundizar el análisis de varianza para poder llegar a conclusiones más potentes.

5. Referencias y fuentes

- Apuntes, slides y notebooks de clase
- Quality of Goverment Institute - Universidad de Gotemburgo
- Algunos de los gráficos fueron generados con ayuda de IA (Claude Sonnet 4.5)