

Modelado de Variables Demográficas Mediante Algoritmos de Aprendizaje Supervisado

Aprendizaje Estadístico Supervisado 2024

Facundo Morini, Mateo Pérez, Ignacio Tarigo

Introducción

Cuando se trabaja con modelos de micro-simulación en demografía, como el modelado de las tasas específicas de fecundidad por edad (ASFR), uno de los factores usualmente considerados es la edad en que las mujeres comienzan a estar en riesgo de concebir. En este contexto, modelar la edad en que una cohorte de mujeres experimenta su primer nacimiento o inicia su vida sexual puede ser de gran utilidad en este tipo de simulaciones.

En este trabajo se propone el uso de modelos de aprendizaje supervisado para evaluar su desempeño en dos objetivos: por un lado, predecir la edad al primer hijo de las mujeres, y por otro, clasificar si una mujer comienza su vida sexual antes o después de una cierta edad. En ambos casos, se emplearán variables socioeconómicas y comportamentales de las mujeres como predictores. Se priorizará el uso de modelos basados en árboles de decisión, como métodos de *Bagging* o *Boosting*.

Los datos utilizados para este propósito provienen de las encuestas DHS, realizadas principalmente en países en vías de desarrollo. En este caso particular, se analizarán las encuestas de las rondas VI y VII, correspondientes a distintos países y realizadas entre los años 2010 y 2020.

Análisis exploratorio

Fuente de los datos

Como se mencionó en la introducción, los datos provienen de las encuestas DHS¹, el Programa de Encuestas Demográficas y de Salud (DHS) es responsable de recopilar y difundir datos precisos y representativos a nivel nacional sobre salud y población en países en desarrollo.

La idea detrás es ayudar a los distintos países sobre todo del continente africano a dar asistencia técnica para la implementación de encuestas demográficas y de salud, donde se recolecta información sobre la fecundidad, planificación familiar, salud maternal e infantil y otros temas de salud como la malaria o el VIH que son de interés en gran parte de los países que contempla la encuesta.

Descripción de los datos

Como primer paso, debido a la magnitud de la encuesta, que abarca un gran número de países y años, se optó por seleccionar un grupo de países representativos de distintas regiones. El objetivo fue capturar la heterogeneidad en las variables de interés desde diversos contextos económicos y culturales. Asimismo, se procuró que las encuestas seleccionadas correspondieran a un período de tiempo similar (en este caso, la década de 2010 a 2020) y que las variables, tanto de interés como explicativas, estuvieran disponibles en todas las encuestas incluidas. Como resultado de este proceso, se eligieron los siguientes países:

- América Latina y el Caribe: Perú, Guatemala y República Dominicana.
- Europa del este: Albania y Armenia.
- Norte y Sur de África: Etiopía y Sudáfrica.

En este caso, se dispone de un total de aproximadamente 105,000 observaciones correspondientes a mujeres de entre 15 y 64 años. Estas observaciones incluyen variables relacionadas con aspectos económicos, sociales y de salud reproductiva. Una complicación presente en los datos es que la disponibilidad de las distintas variables pueden variar de un país al otro y ciertas variables pueden presentar demasiados valores faltantes para algunos países para ser utilizable en los modelos. Por esta razón se realizó un filtrado previo de las variables a utilizar. Como resultado, se trabaja principalmente con 30 variables que están disponibles para todos los países seleccionados y que, a priori, se consideran posiblemente relevantes para el análisis.

Tabla 1: Descripción de las variables disponibles en el trabajo.

Codigo DHS	Nombre	Descripción	Tipo	Recorrido
caseid	id	Identificador de la mujer encuestada	Catégorica	104487 Mujeres

¹<https://dhsprogram.com/data/available-datasets.cfm>.

v000	pais	País donde se realizó la encuesta	Categórica	7 Países
v005	ponderador	Peso de la mujer en la encuesta, número entero de 8 dígitos, donde 6 dígitos corresponden a decimales	Numérica	
v007	anio	Año en que se realizó la encuesta	Numérica Discreta	2008-2018
v012	edad	Edad de la mujer encuestada	Numérica Discreta	15-59
v021	psu	Unidad primaria de muestreo	Numérica Discreta	
v022	estrato	Estrato del que se obtuvo la observación	Categórica	
v025	zona	Indica si la mujer es de un área urbana o rural	Categórica	1:Urbano, 2:Rural
v119	hogar_elect	Indica si el hogar tiene acceso a electricidad	Categórica	0:No, 1:Si
v121	hogar_tv	Indica si el hogar dispone de una televisión	Categórica	0:No, 1:Si
v133	anio_educ	Total de años de educación de la mujer, calculado a partir del nivel máximo de educación alcanzado; es un valor comparable entre países	Numérica Discreta	0-25
v136	n_fam_hogar	Total de personas que viven en el hogar	Numérica Discreta	1-25
v157	leer	Frecuencia con la que la mujer lee un periódico, revista o libro	Categórica	0:Nada, 1:Ocasional, 2:Todos los días
v158	radio	Frecuencia con la que la mujer escucha la radio	Categórica	0:Nada, 1:Ocasional, 2:Todos los días
v159	noticias	Frecuencia con la que la mujer ve noticias	Categórica	0:Nada, 1:Ocasional, 2:Todos los días
v190	cat_riqueza	Categoría de riqueza del hogar	Categórica	1:Más Pobre, 2:Pobre, 3:Medio, 4:Rico, 5:Más Rico
v191	indice_riqu	Índice de riqueza del hogar estandarizado, calculado a partir de distintas medidas como materiales de construcción de la casa y servicios disponibles; 5 dígitos corresponden a decimales	Numérica	-4.992-3.09962
v212	edad_phijo	Edad de la mujer al primer hijo	Numérica Discreta	3-47
v301	con_antic	Conocimiento de algún método anticonceptivo, donde se distingue entre modernos (ej.: condón, pastillas, parches, etc.), tradicionales (ej.: coito interrumpido, calendario, etc.) y folklóricos (ej.: amuletos, rituales, etc.)	Categórica	0: Ninguno, 1:Folklórico, 2:Tradicional, 3:Moderno
v364	uso_antic	Intención de uso de métodos anticonceptivos	Categórica	1:Usando Método, 2:No Usando Método, 3:Nunca Tiene Relaciones
v503	cant_union	Cantidad de uniones que ha tenido la mujer	Categórica	0: No casada, 1: Una unión, 2: Más de una unión
v531	edad_psexo	Edad de la mujer al primer acto sexual	Numérica Discreta	0-45
v536	rec_sexo	Relaciones sexuales recientes de la mujer	Categórica	0: Nunca, 1: Activa, 2: No activa
v613	nideal_hijos	Número ideal de hijos	Numérica Discreta	0-30
v715	anio_educ_	Total de años de educación de la pareja de la mujer	Numérica Discreta	0-24
v717	ocupacion	Ocupación de la mujer	Categórica	0: No Trabaja, 1: Profesional/Técnico, 2: Agricultura, 3: Servicios/Ventas, 4: Doméstico, 5: Manual
v836	n_parejas	Cantidad de parejas sexuales que ha tenido la mujer	Numérica Discreta	1-95

Antes de realizar el análisis exploratorio, se procede a limpiar los datos. Tomándose la precaución en cada variable de considerar las codificaciones de no respuesta y dato faltante, como el valor 99, que puede afectar la interpretación sobre todo en variables numéricas.

Se debe tener en cuenta que los datos provienen de una muestra compleja, donde se utiliza un diseño muestral por conglomerados estratificado en dos etapas, donde en general se busca que la misma sea representativa tanto a nivel nacional, regional como al contexto urbano y rural.

Debido a esto para una mejor representación a nivel de poblacional el diseño debería ser contemplado en los modelos a utilizar. En modelos como *Random Forest* la bibliografía de como incorporar esta información del diseño no esta ampliamente desarrollada, por lo cual para esta primera aproximación al problema se optó por no incorporar explícitamente en los modelos aunque si se ha tenido en cuenta en el análisis exploratorio, lo cual permite una mejor comprensión de la estructura y distribución de los datos.

Edad al primer hijo

La edad en que las mujeres comienzan a tener hijos está influenciada por una serie de factores en los que se encuentran comportamientos sociales como también por aspectos económicos y culturales. En la bibliografía suelen mencionarse algunos factores que están fuertemente vinculados a los procesos de fecundidad, como la mayor participación en el mercado laboral, el acceso a la educación de las mujeres, la planificación familiar y el uso de métodos anticonceptivos.

En la Figura 1 y tabla posterior se puede observar como, dependiendo de que país sea la mujer encuestada, varía la edad que tenían al momento en que nació su primer hijo, algo esperable viniendo de países con contextos muy distintos. En países como Etiopía, Guatemala o Dominicana, hay una gran parte de las mujeres encuestadas que tiene su primer hijo con veinte años o menos, mientras que en países del Este de Europa algo más desarrollados como Armenia o Albania, la edad al primer hijo tiende a ser mayor con una media por encima de los 20 años. Este patrón es usualmente observado mientras más desarrollo más se atrasan todos los procesos relacionados con la fecundidad. Más allá de las diferencias, existe una gran concentración de mujeres que tienen su primer hijo entre los 18 y 23 años en todos los países, esto está fuertemente relacionado a la edad en que las mujeres comienzan a vivir en pareja o se casan, donde comienzan a tener usualmente riesgo a concebir.

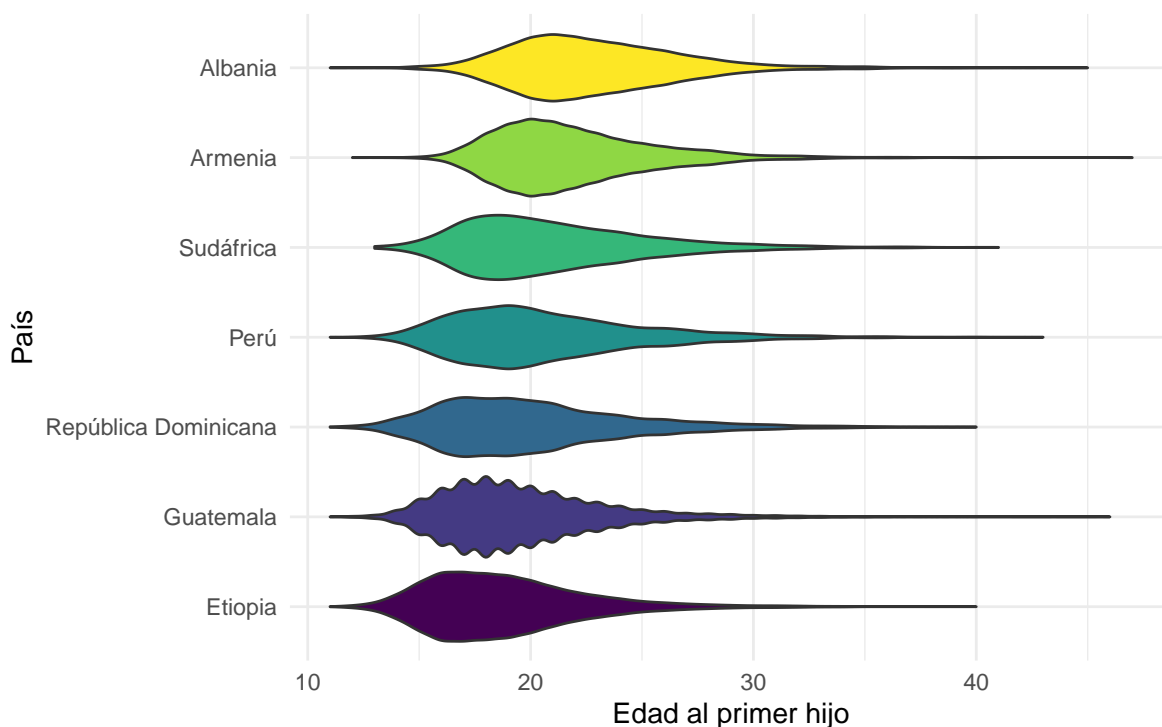


Figura 1: Gráfico violín, distribución de la edad al primer hijo por país.

Tabla 2: Media y Desvio Estimados de la Edad al Primer Hijo por país

País	Media Edad	Desvio
Etiopia	18.84	0.08
Guatemala	19.60	0.04
República Dominicana	20.11	0.10
Perú	20.65	0.09
Sudáfrica	20.95	0.13
Armenia	21.87	0.07
Albania	22.84	0.07

Uno de los factores más importantes en el proceso del desplazamiento de la fecundidad a edades más avanzadas es el logro educativo de las mujeres. Esto es algo que en los datos utilizados en este trabajo debería poder ser captado. En la Figura 2 se observa la distribución de la edad al primer hijo donde cada línea corresponde a la cantidad de años de educación que tuvo la madre. En este caso, se puede observar claramente un corrimiento hacia la derecha de las modas de la distribución a medida que la mujer tiene más años de educación. Esto indica que mientras más estudios tienen las mujeres, más aplazan el momento de tener su primer hijo. Con lo cual en un principio esta variable y las relacionadas deberían ser útiles para predecir la variable de interés.

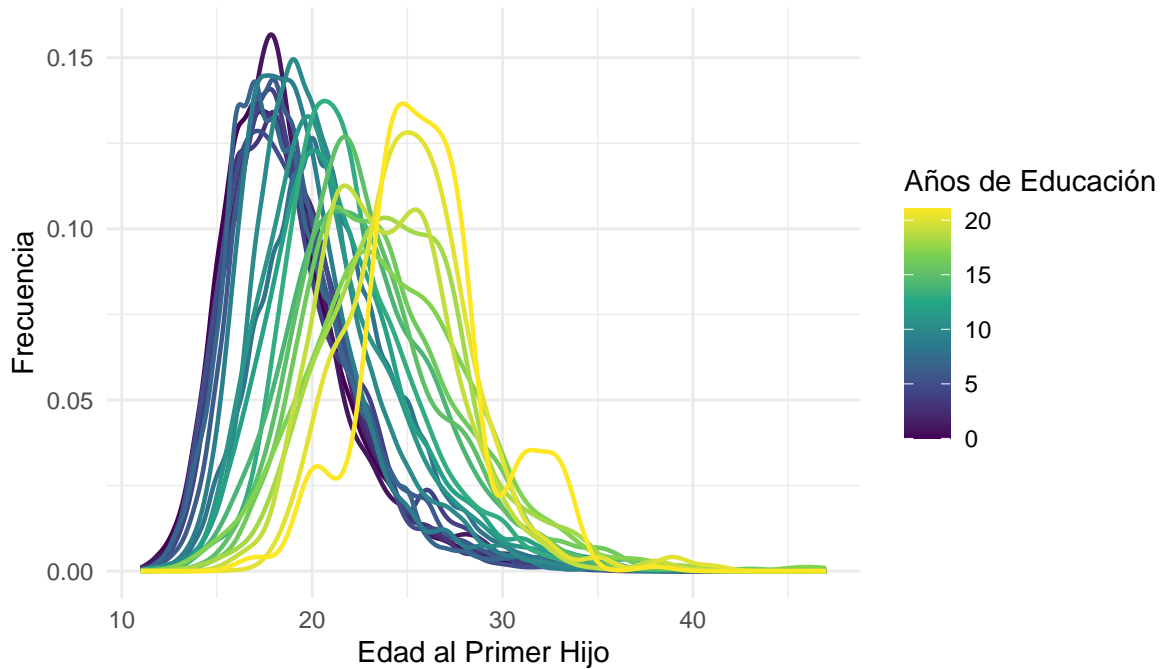


Figura 2: Gráfico Densidad, edad al primer hijo por Años de Educación de la Madre.

Este patrón es repetido por todos los países, aunque su efecto es más marcado en algunos países, se observa la Figura 3 donde se tiene la distribución de la edad al primer hijo por país y si la madre presenta más de 15 años de educación (años usualmente de completado grado secundario). En todos los países, las mujeres con más años de educación tienden a tener su primer hijo a una edad más avanzada, aunque en países como Albania, este efecto parece bastante menos notorio.

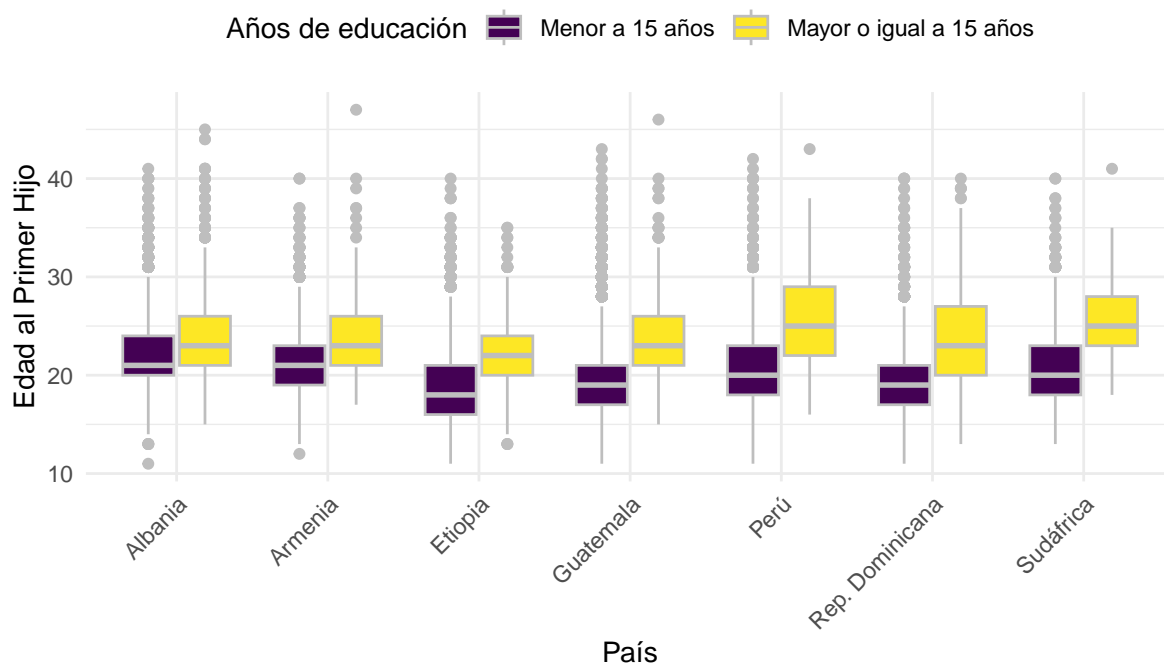


Figura 3: Gráfico Caja, distribución de la edad al primer hijo por país y si la mujer tiene más de 15 años de educación.

En la Figura 4 podemos observar la distribución de la edad al primer hijo para los distintos niveles de riqueza de los hogares, para algunos países, en general en todos los países observamos como se atrasa la edad a medida que el hogar es más rico, sobre todo en Albania, Perú y Sudáfrica. En otros como Dominicana o Armenia vemos que los niveles de riqueza están más concentrados en algún valor, salvo por el nivel más rico que ahí si se muestra un desplazamiento, no muy grande, pero un desplazamiento en fin, hacia edades un poco mayores. Por lo tanto se podría decir que en este caso la riqueza del hogar es una variable importante en este estudio.

Otro aspecto interesante en este tipo de estudio que se suele ver, es la cantidad de uniones y parejas sexuales que tienen las mujeres, por lo que se presenta la Figura 5, donde observamos la distribución de la edad al primer hijo según estas variables. Vemos que las mujeres con más uniones tienden a tener hijos más jóvenes cuando tienen más de una pareja sexual, y además observamos esta tendencia en las mujeres con más parejas sexuales que no están casadas.

También podemos observar que en las mujeres con una sola unión la edad al primer hijo se encuentra concentrada en los 20 años, fuertemente vinculado a la edad a esa unión.

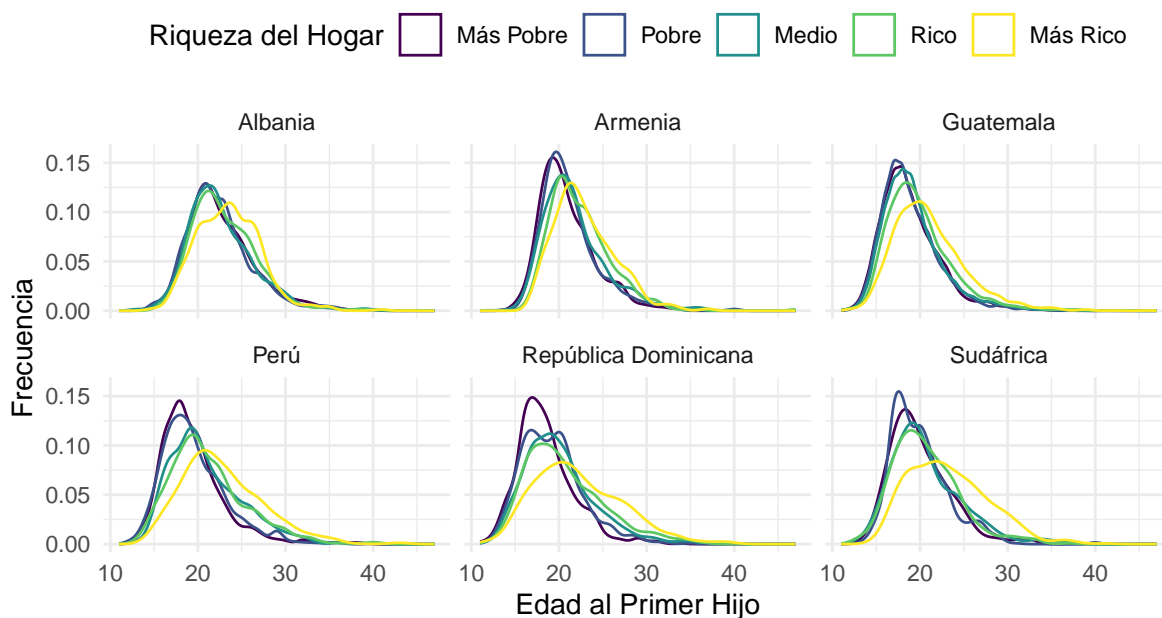


Figura 4: Gráfico Densidad, distribución de la edad al primer hijo por riqueza del hogar.

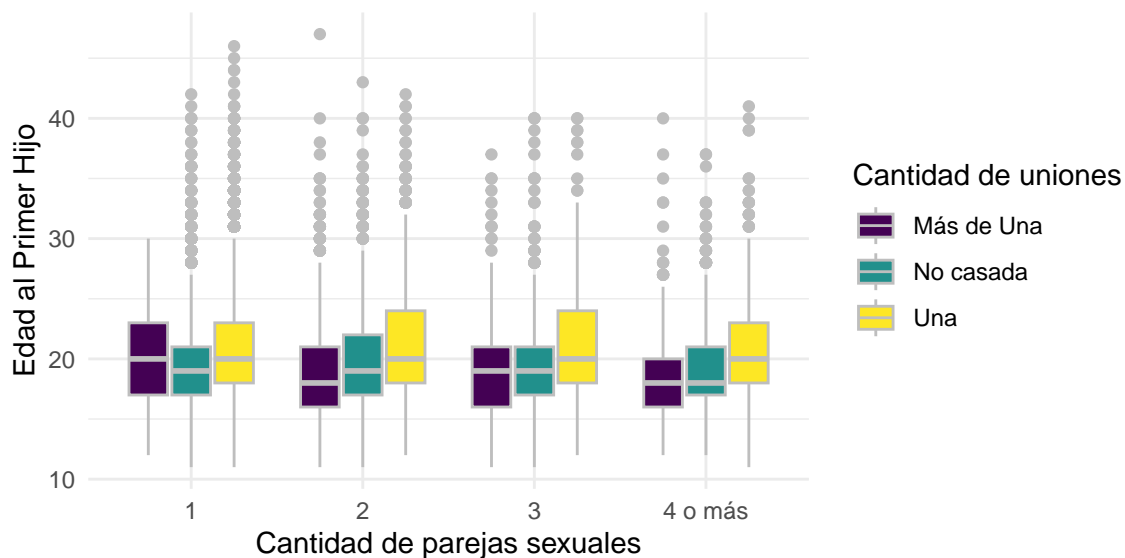


Figura 5: Gráfico Caja, distribución de la edad al primer hijo por cantidad de uniones y cantidad de parejas sexuales.

Primera relación sexual

La edad en que las mujeres comienzan a tener relaciones sexuales marca el inicio del riesgo a concebir y por tanto la misma está fuertemente relacionada con la edad al primer hijo. Gran parte de las variables comportamentales, económicas, educativas de la mujer que se observaron en la edad al primer hijo, también son relevantes para la edad al primer acto sexual. En este caso se trabajará con la variable si la mujer comienza su vida sexual antes o después de los 18 años.

En la Figura 6 se puede observar la proporción estimada de mujeres que comienzan a tener relaciones sexuales antes de los 18 años por país, donde se observa un patrón bastante marcado. Países tanto latinos como africanos tienden a tener una proporción bastante igualada entre mujeres que comienzan a tener relaciones sexuales antes y después de los 18 años, mientras que en los países del este europeo seleccionados que comparten un contexto cultural y religioso similar tienden a tener una proporción mucho mayor de mujeres que comienzan a tener relaciones sexuales después de los 18 años, estando por encima del 80% en ambos casos.

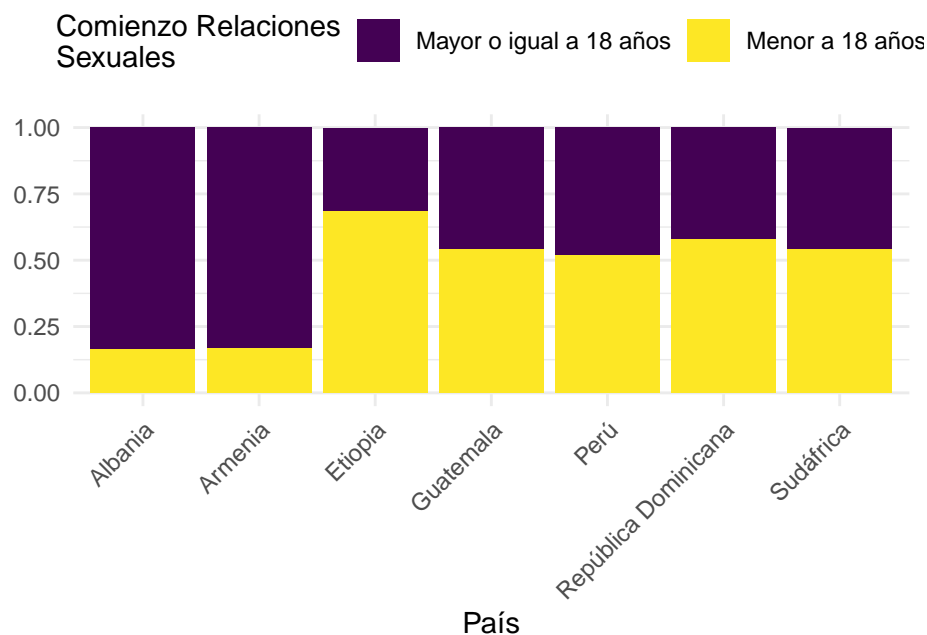


Figura 6: Gráfico barras, Estimación de la proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años por país.

Al igual que en el caso de la edad al primer hijo, la educación de la mujer debería ser un factor muy relevante para esta variable de interés. En la Figura 7 se observa la distribución de los años de educación de la mujer por si comienzan a tener relaciones sexuales antes o después de

los 18 años. En este caso, se observa que las mujeres que comienzan a tener relaciones sexuales después de los 18 años tienden a tener mas años de educación con una gran concentración en los 10 años de educación, mientras que las que comienzan antes de los 18 años tienden a situarse entre los 0 a 10 años de educación.

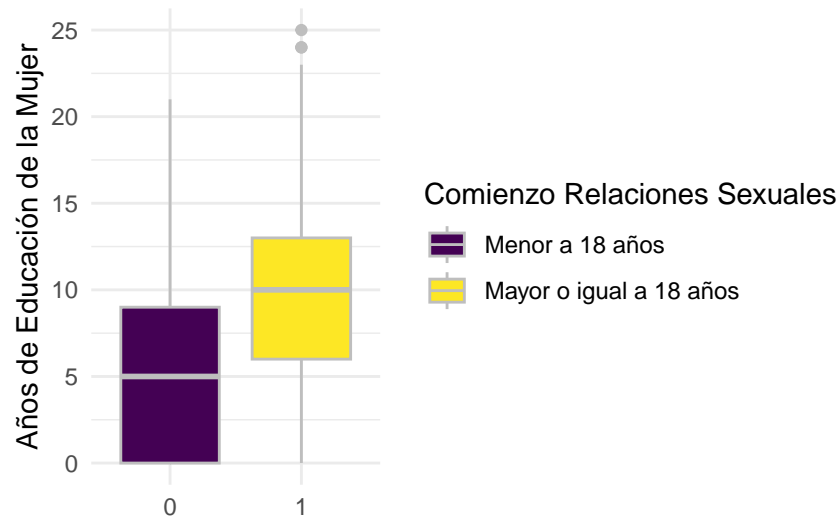


Figura 7: Gráfico de caja, distribución de años de educación de la mujer por comienzo de relaciones sexuales.

En la Figura 8, podemos ver que la proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años es más alta en las categorías socioeconómicas más bajas. Esto puede estar relacionado con un mayor acceso a información sobre salud sexual, por ejemplo. En los hogares categorizados como más pobres vemos una proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años de 0.56, mientras que en los hogares contrarios, es decir los más ricos, esta proporción es de 0.36.

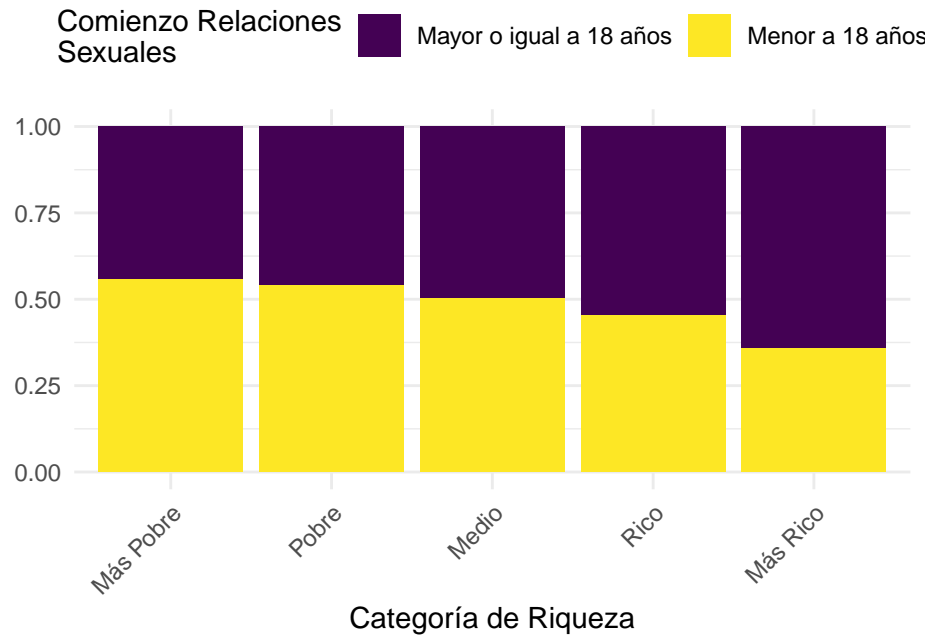


Figura 8: Gráfico barras, Estimación de la proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años por categoría de riqueza.

Este patrón entre el contexto económico y la variable de interés se puede visualizar en muchas de las otras variables referidas a la riqueza del hogar, como se observa en la Figura 9 donde se ve un claro patrón de mayor cantidad de mujeres que comienzan a tener relaciones sexuales antes de los 18 años en hogares que no cuentan con televisión, relacionado a un contexto económico mas desfavorable como a una menor exposición a información.

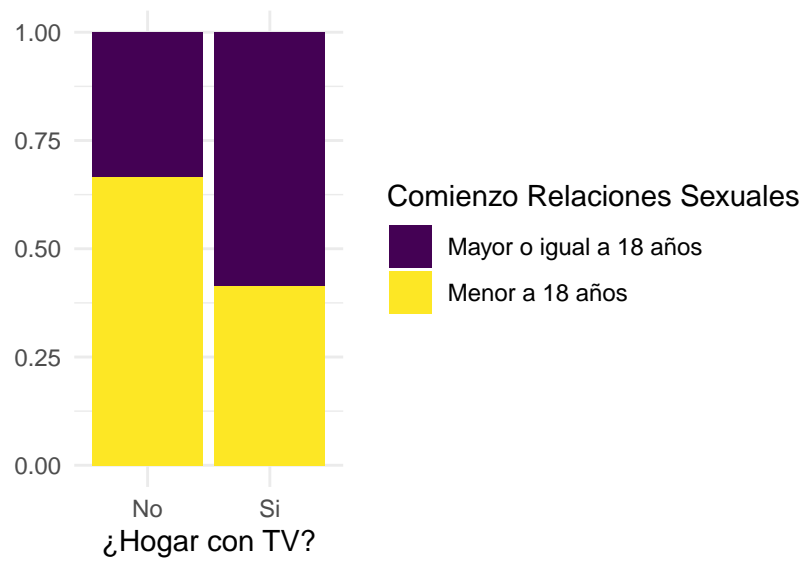


Figura 9: Gráfico barras, Estimación de la proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años por si poseen o no un televisor en su hogar.

Modelado de la edad al primer hijo

Antes de comenzar con el modelado se procede a realizar un filtrado de los datos y a seleccionar las variables más relevantes para el modelo.

En esta primera aproximación al problema se opta por solamente tomar las observaciones que dispongan de todos los datos necesarios para el modelo, no realizándose imputaciones de datos faltantes o análisis de la no respuesta a la encuesta.

En total se utilizan 46029 observaciones de mujeres donde se utilizara un 85% de los datos para entrenar y el 15% restante para evaluar los modelos considerados. En esta división se tendrá en cuenta el país de origen de la mujer para que no haya un desbalance de muchas observaciones de algún país en particular, teniendo en cuenta que algunos países cuentan con más observaciones que otros.

Uno de los problemas que se presentan en datos provenientes de encuestas es la presencia de valores extremos posiblemente vinculados a errores en el rellenado o por falta de conocimiento de la mujer encuestada. Para el caso de la edad al primer hijo, se filtran las observaciones con valores extremadamente atípicos o que no tienen sentido desde el punto de vista biológico, como por ejemplo un valor que indica que una mujer tuvo su primer hijo a los 3 años. En este caso, se consideran únicamente las mujeres que tuvieron su primer hijo a partir de los 11 años, límite establecido con base en criterios biológicos.

Selección de covariables relevantes

Antes de seleccionar las variables para el modelo, se excluyen aquellas que carecen de sentido práctico. Por ejemplo, la variable edad de la mujer no se considera, ya que al trabajar únicamente con mujeres que ya han tenido hijos, incluirla implicaría utilizar una variable estrechamente relacionada con la variable de interés, lo que podría sesgar el modelo, especialmente en el caso de mujeres muy jóvenes. A su vez se excluye también la variable de edad a la primera relación sexual, ya que como se vio en el análisis exploratorio están muy fuertemente relacionadas e interesa más ver el efecto de otros factores desconociendo esta variable.

Debido a que se cuenta con un gran número de variables que posiblemente no sean todas relevantes, como paso previo a cada modelo se realiza un análisis de importancia de las variables. Para esto se utiliza un modelo de *Random Forest* donde se utiliza el método de permutación para calcular la importancia de las variables.

Para el ajuste de este modelo no se procura analizar ni afinar en detalle la grilla de hiperparámetros, sino que se busca obtener un modelo relativamente bueno del que se pueda extraer las variables más importantes. En base a esto se utiliza una grilla de 30 combinaciones posibles de la cantidad de variables a considerar en cada nodo (*mtry*) y la cantidad mínima de observaciones en cada nodo terminal (*min_n*). Se utilizó un *10-fold cross-validation* para evaluar

los modelos y se seleccionó el mejor modelo en base a la métrica de raíz del error cuadrático medio.

En la Figura 10 se observa la importancia de las variables consideradas en el modelo seleccionado. El mismo pareciera ser consistente con lo esperado, donde variables relacionadas con la educación o riqueza del hogar tienden a tener mayor importancia algo que concuerda con el análisis exploratorio previo. Aunque llama la atención que las variables relacionadas con la planificación familiar como el conocimiento de métodos anticonceptivos se encuentren en los últimos lugares de importancia, algo que en la literatura se destaca como un buen predictor de los procesos de fecundidad.

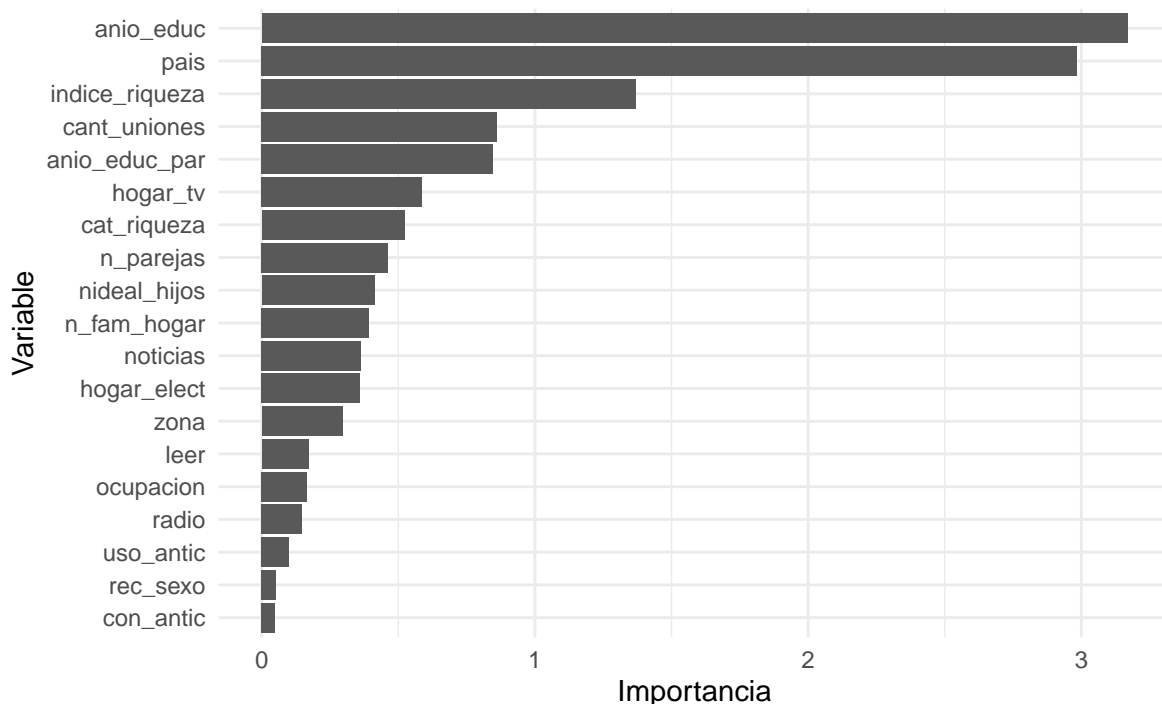


Figura 10: Gráfico barras, importancia de las variables consideradas en el modelo de Random Forest mediante permutación para predecir la edad al primer hijo.

Una vez visualizados los datos y posibles relaciones en las variables se procede a armar los modelos. De lo visto mediante el modelo de *Random Forest* se seleccionan las 10 variables que más importancia presentaron para predecir la edad al primer hijo.

En este caso las variables que quedaron seleccionadas son:

1. Años de educación de la mujer.
2. País de procedencia de la mujer.
3. Índice de riqueza del hogar.

4. Cantidad de uniones que ha tenido la mujer.
5. Años de educación de la pareja de la mujer.
6. Si el hogar dispone de televisión.
7. Categoría de riqueza del hogar.
8. Numero de parejas sexuales que ha tenido la mujer.
9. Numero ideal de hijos.
10. Numero de personas que viven en el hogar.

En este caso se trabaja íntegramente con ‘*tidymodels*’ framework que permite realizar la evaluación de los modelos de manera sencilla. Se opta por evaluar la eficiencia de modelos basados en árboles: *Bagging*, *Random Forest* y *XGBoost*, donde para cada uno se realiza su correspondiente tuneo de hiperparámetros.

Búsqueda de Hiperparámetros

Se procede a la búsqueda de hiperparámetros para los modelos en busca del modelo que presente la mejor *performance*. Para esto se crea una grilla específica para cada uno, evaluando todos los modelos en conjunto mediante *workflow_map*.

Para los modelos de *Bagging* se entrenaron los parámetros de cantidad mínima de observaciones en cada nodo terminal (*min_n*) y la cantidad de árboles (*trees*). En la siguiente tabla se puede observar la grilla utilizada, que luego se visualizará si la misma es la adecuada para este problema.

Tabla 3: Grilla de hiperparámetros para modelos de Bagging

min_n	50	100	150	200	250	300	350	400	450	500	1000	1500	2000	2500
trees	50	100	500	1000										

En cuanto a los modelos de *Random Forest* se ajustan los hiperparámetros de cantidad mínima de observaciones en cada nodo terminal (*min_n*), la cantidad de variables a considerar en cada nodo (*mtry*) y la cantidad de árboles (*trees*). En la siguiente tabla se puede observar la grilla utilizada:

Tabla 4: Grilla de hiperparámetros para modelos de Random Forest

min_n	50	100	150	200	250	300	350	400	450	500	1000	1500	2000	2500
trees	50	100	500	1000										
mtry	2	4	6	8	10									

Para los modelos de *XGBoost* debido a la gran cantidad de hiperparámetros a ajustar se procedió de forma distinta no definiendo una grilla fija. En este caso se utilizará un método de

búsqueda que permite evaluar un espacio de hiperparámetros de forma más eficiente, en particular se optó por utilizar *Latin hypercube sampling*, evaluando un total de 60 combinaciones posibles de hiperparámetros.

En primer lugar se evaluó posibles rangos válidos para algunos de estos hiperparámetros. De esto se optó por limitar el rango de algunos de ellos como el *learn_rate* a un rango más acotado, mientras que a otros simplemente se utilizó los rangos por defecto que vienen en *tidymodels*.

Tabla 5: Grilla rango hiperparámetros para modelos de XGBoost

	Valor min.	Valor max.
tree_depth	1	15
min_n	50	500
loss_reduction	1e-10	31
sample_size	0.1	1.0
mtry	1	10
learn_rate	0.0016	0.1600

La búsqueda se realiza al igual que para el modelo de visualización de importancias mediante *10-fold cross-validation*. En la Figura 11 y Figura 12 se puede visualizar los valores de *RMSE* y *RSQ* para los modelos de *Bagging* y *Random Forest* respectivamente. En ambos gráficos parece bastante marcado una mejor *performance* para valores de *min_n* cercano a los 400, pareciendo la grilla propuesta adecuada para este problema.

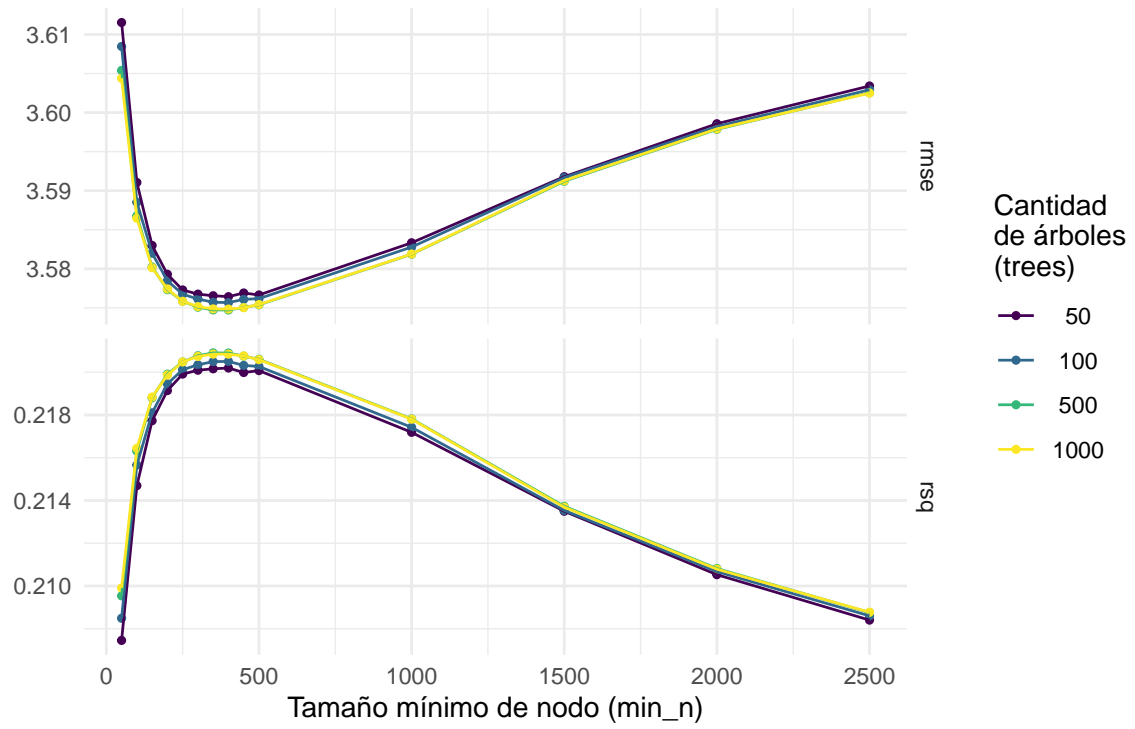


Figura 11: Gráfico, resultados de la búsqueda de hiperparámetros para modelos de Bagging.

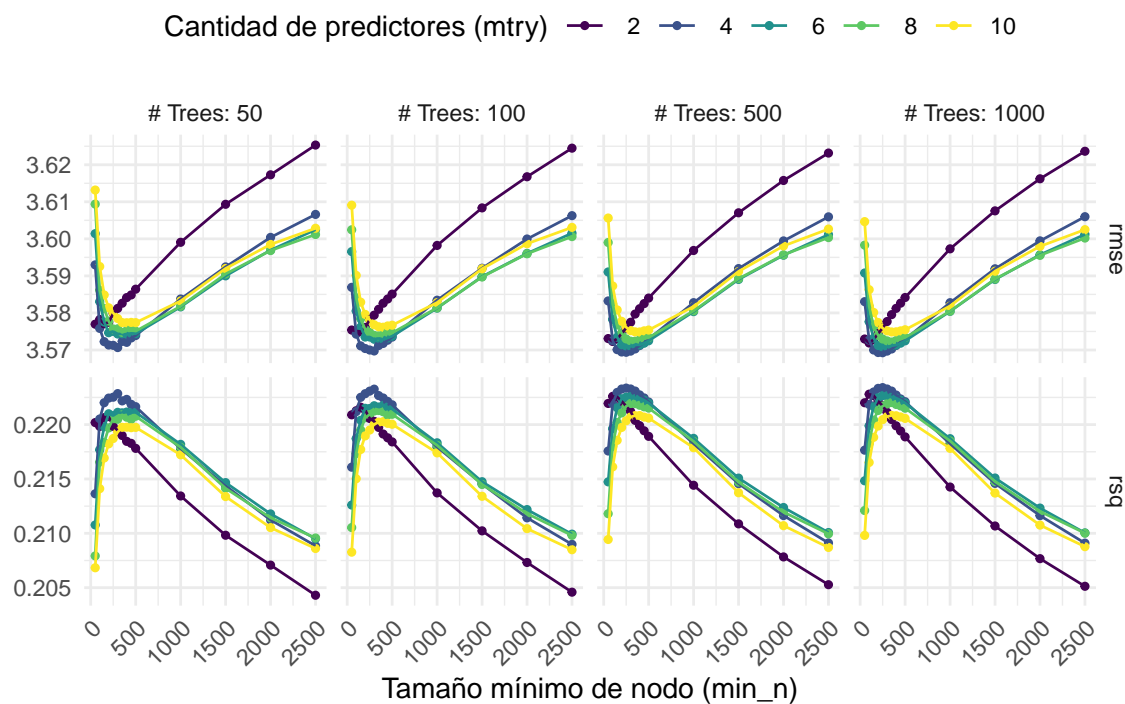


Figura 12: Gráfico, resultados de la búsqueda de hiperparámetros para modelos de Random Forest.

Comparación de modelos

Para comparar los modelos obtenidos, se realizó una comparación entre la métrica RMSE (Raíz del Error Cuadrático Medio) de cada uno. Esta medida representa la raíz cuadrada de la media de la diferencia entre los valores predichos y los observados al cuadrado. Mientras más bajo sea este valor, significa que menor es el error de estimación del modelo; entonces, al comparar los tres modelos realizados, se puede elegir cuál es el mejor para estimar, teniendo en cuenta esta métrica. En este caso, el modelo ajustado por el método de Random Forest es el que tiene un valor más bajo del RMSE, aunque los tres valores son muy cercanos entre sí, por lo cual sugiere que el rendimiento de los modelos es muy similar. Esto se puede deber a que los tres son basados en árboles, entonces, por más que haya diferencias en la metodología, a grandes rasgos deberían predecir de forma similar.

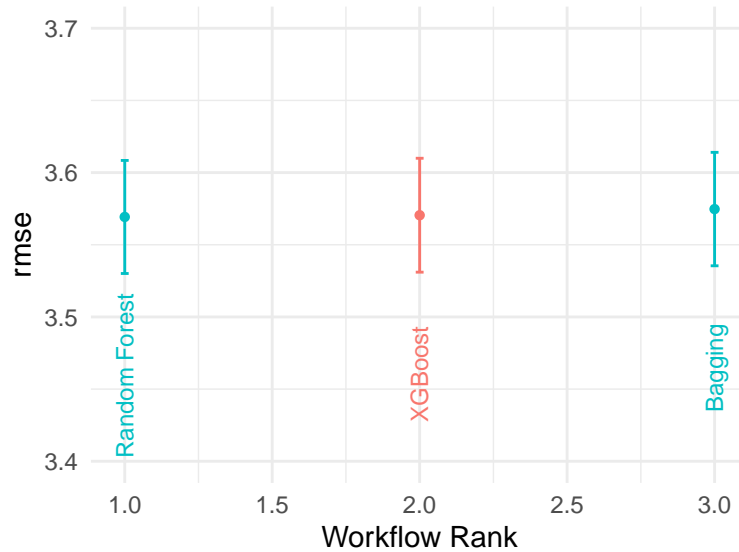


Figura 13: Gráfico comparación de los modelos de Bagging, Random Forest y XGBoost para predecir la edad al primer hijo.

Evaluación modelo final

Una vez realizada la comparación de los modelos, y decidiéndose por el modelo realizado mediante el método de *Random Forest*, se procede a evaluar este modelo como modelo final.

El primer paso a realizar es el análisis de los residuos. Se obtuvo la Figura 14 que compara el valor predicho por el modelo para cada individuo en el eje horizontal y el residuo, que es el valor verdaderamente observado menos el predicho, en el eje vertical. En el gráfico se puede observar cómo los puntos son bastante simétricos alrededor del cero, lo cual nos da un indicio de que el modelo no estaría sesgado. Tampoco se observa un patrón en la nube de puntos que indique que el modelo no haya ajustado correctamente. Lo que sí se puede observar es la aparición de algunos valores atípicos de los residuos en la parte negativa del gráfico. Esto indica que, para los valores observados más altos, en los que una mujer tuvo un hijo, el modelo no logra estimar bien la edad a la que esa mujer realmente tuvo su primer hijo. Esto se puede corroborar observando que en el eje horizontal solo tenemos valores predichos menores a 30 años, mientras que en la base de datos sí existen mujeres que hayan tenido hijos luego de esa edad. Se intentó ajustar el modelo realizando transformaciones y teniendo en cuenta el muestreo en la edad, pero no tuvimos éxito en lograr que el modelo capte mejor a esas mujeres. Tal vez las covariables que están en el modelo no sean suficientemente relevantes para explicar este fenómeno.

Otro punto a considerar es que, observando los puntos segmentados por país, todos presentan aproximadamente el mismo comportamiento, lo cual lleva a considerar que el modelo predice

de buena manera en todos ellos.

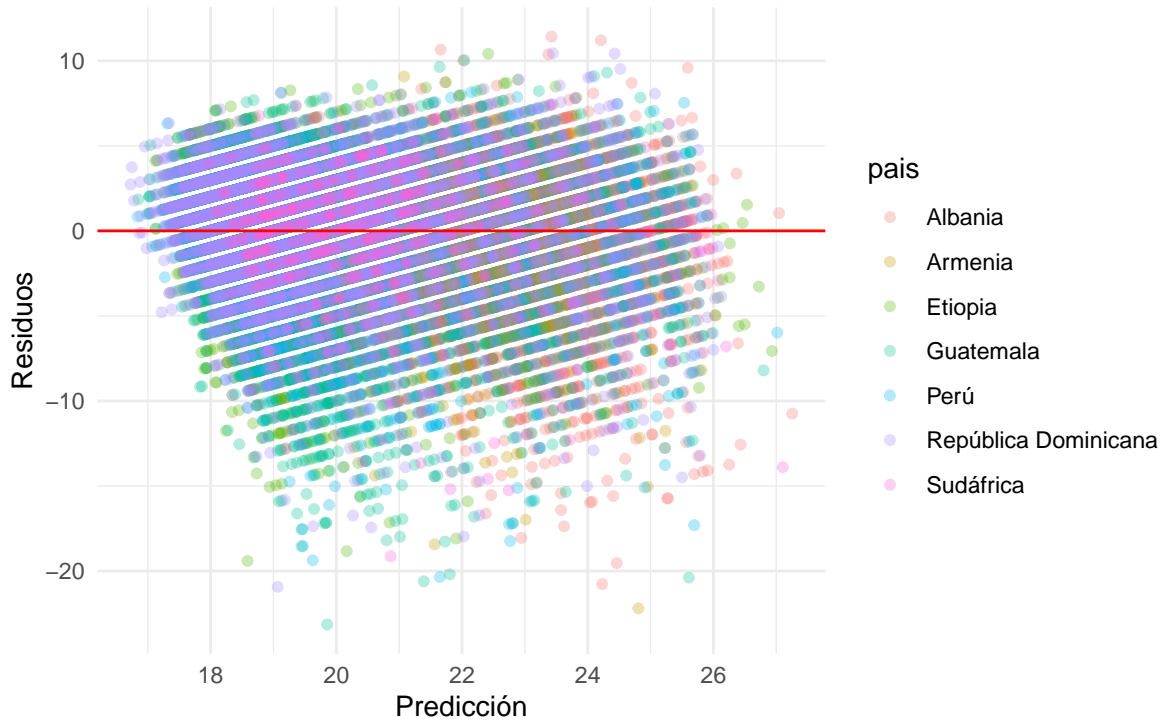


Figura 14: Gráfico residuos vs predichos, conjunto de entrenamiento.

Luego se realizó la Figura 15 para el conjunto de datos de testeo, en donde se observa en el eje horizontal el valor predicho de la edad de la mujer a la hora de tener su primer hijo, y en el eje vertical el valor real. Se pueden considerar los mismos puntos que el gráfico anterior, lo que indica que no existe un sobreajuste hacia el subconjunto de datos de entrenamiento. También, de igual manera que en el anterior gráfico, el modelo no logra predecir bien la edad de las mujeres que tuvieron su primer hijo después de los 30 años.

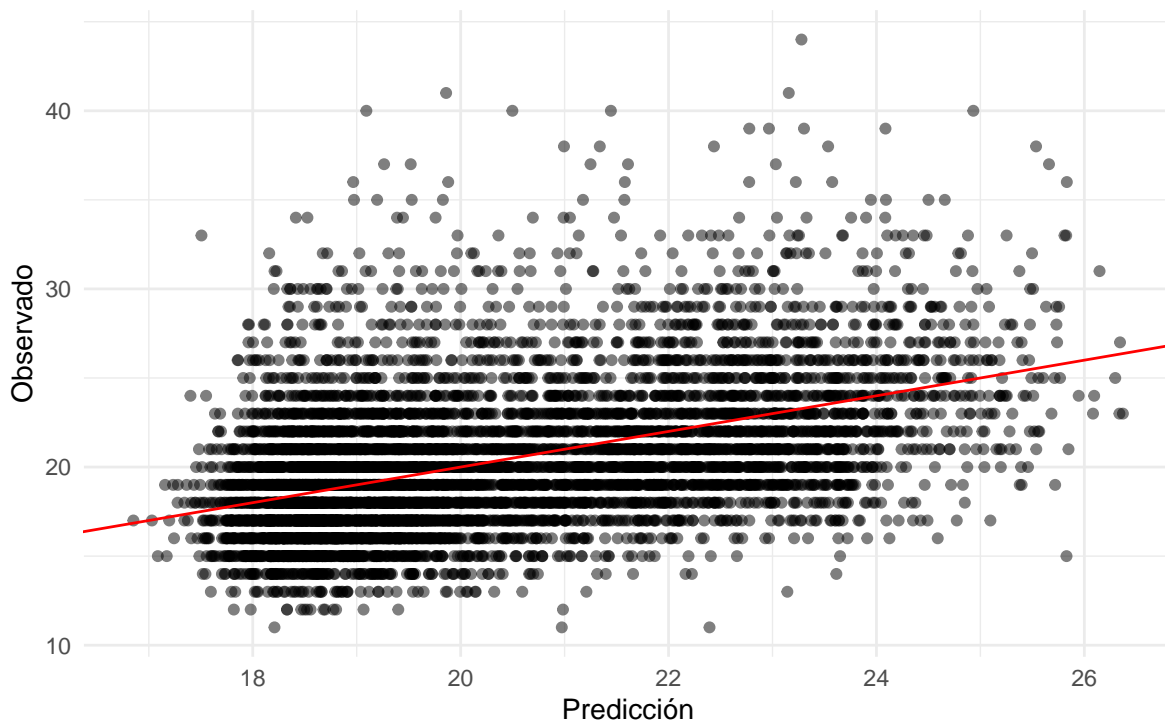


Figura 15: Gráfico predicho vs observado, conjunto de testeo.

En la siguiente tabla se puede observar como el $RMSE$ del modelo para cada país es muy similar al resto, lo cual confirma el punto observado en el gráfico anterior. La variable MAPE (Mean Absolute Percentage Error) indica la media del error porcentual de todas las observaciones de cada país. Los valores de esta métrica son muy parecidos para todos los países, y no indican que ningún país prediga muy mal la edad que tiene una mujer cuando tiene su primer hijo.

Tabla 6: Métricas en el conjunto de testeo para predecir la edad al primer hijo por país

pais	rmse	mape
Albania	3.58	12.02
Armenia	3.31	11.33
Etiopia	3.48	14.17
Guatemala	3.49	13.37
Perú	3.87	14.20
República Dominicana	3.99	15.19
Sudáfrica	3.79	13.44

Interpretación e importancia de las variables

Lo primero a destacar es que en la Figura 16, se obtuvieron resultados muy similares a los obtenidos en la Figura 10 y el análisis exploratorio, siendo las variables años de educación y país de la mujer las variables con más importancia.

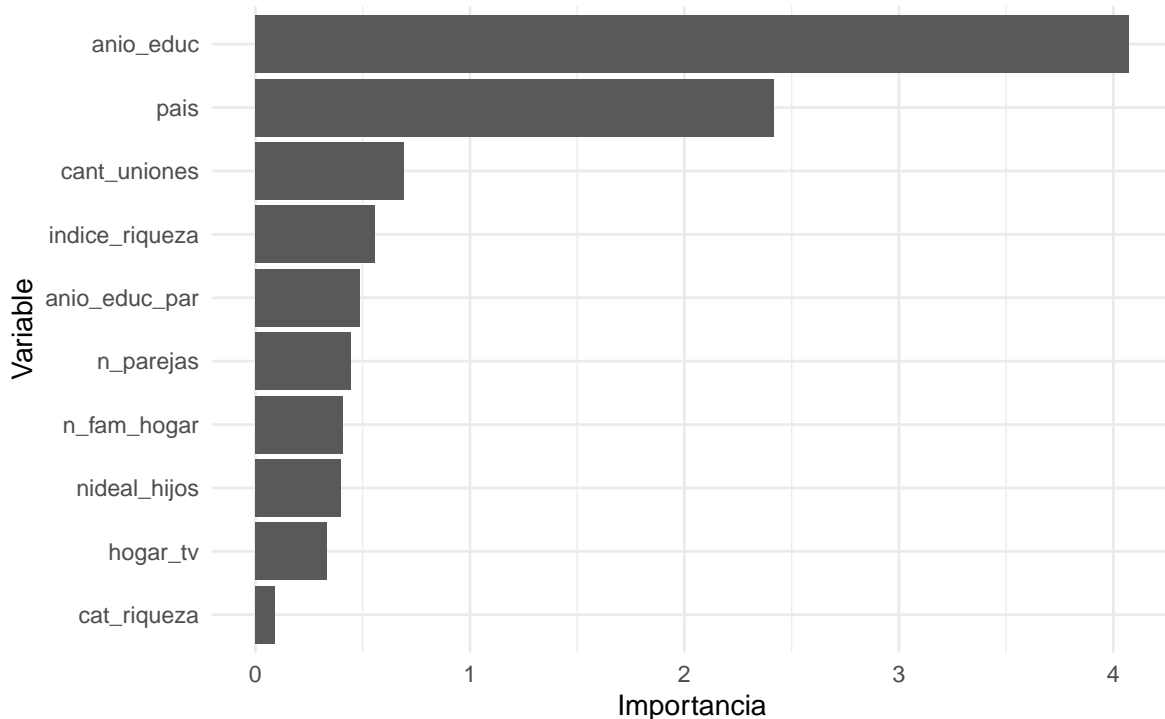


Figura 16: Gráfico barras, importancia de las variables consideradas en el modelo final.

La Figura 17 muestra un gráfico de Dependencia Parcial del efecto marginal que tiene la variable de años de educación de la madre con respecto a la variable de interés del problema planteado. Este gráfico se obtiene mediante la función de dependencia parcial, que es la siguiente:

$$f(x_S) = E_{X_C}[\hat{f}(x_S, x_C)]$$

En la cual X_S sería la variable años de educación y X_C el resto de variables explicativas utilizadas en el modelo.

En el gráfico se puede observar que la edad media de la mujer a la hora de tener su primer hijo en mujeres que tienen 10 o menos años de educación permanece constante entre los 19 y

20 años. Luego de los 10 años de educación, la edad media de la mujer al tener su primer hijo crece de forma abrupta por cada año extra de educación que tiene la mujer.

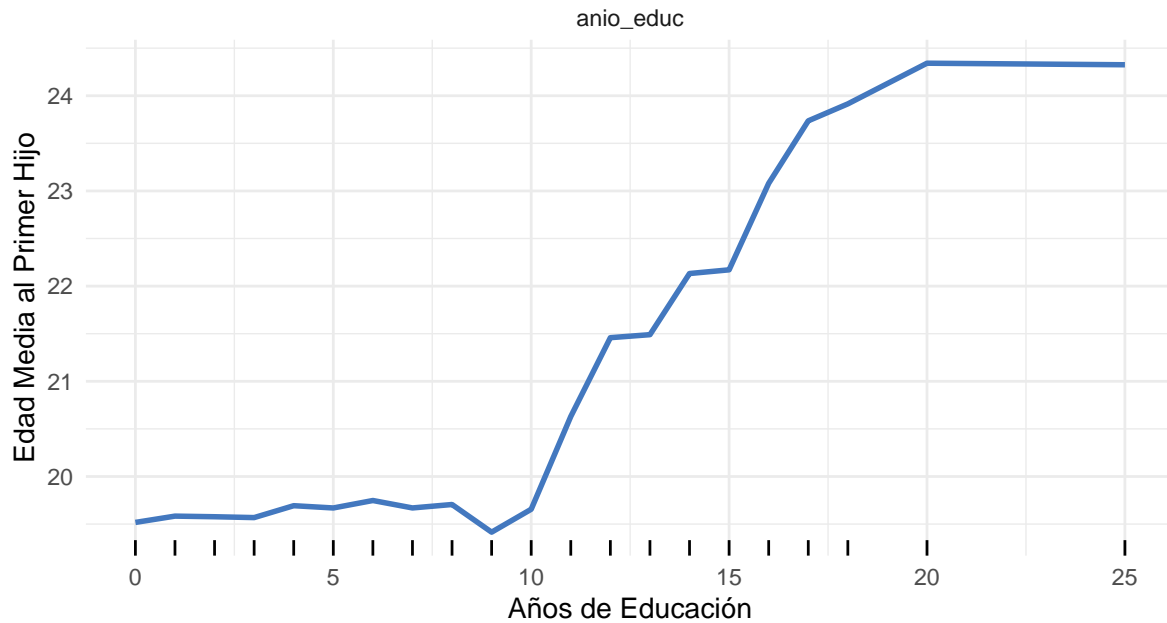


Figura 17: Gráfico PDP, efecto de los años de educación en la edad al primer hijo.

La siguiente Figura 18 también muestra un gráfico PDP para analizar el efecto marginal de la variable de país sobre la variable de interés. En este caso, se observa algo similar a lo observado en el análisis exploratorio en la Figura 1, en donde los países europeos tienen la edad media de la mujer al tener su primer hijo un poco más alta que el resto.

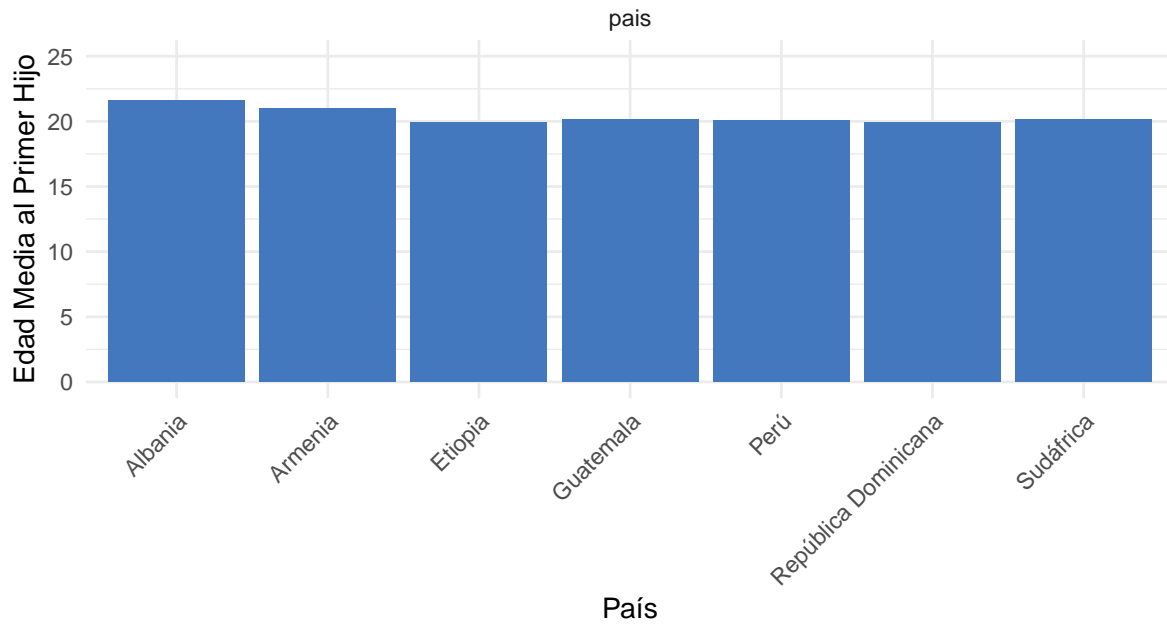


Figura 18: Gráfico PDP, efecto del país en la edad al primer hijo.

Modelado de la edad a la primera relación sexual

Se procede a realizar el modelado de la edad a la primera relación sexual donde el procedimiento a realizar es muy similar al anterior. En primer lugar se realiza un filtrado de los datos, donde nuevamente se trabajará con observaciones que dispongan de todos los datos necesarios para el modelo. En este caso la variable de interés no está presente con lo cual se debe construir a partir de la variable *edad_psexo*.

Como se visualizó en el análisis exploratorio la proporción de mujeres que comienzan a tener relaciones sexuales antes de los 18 años en algunos países es bastante más baja. Con lo cual se decide crear otra variable que combine el país y la variable de interés, para que en el proceso de división de los datos o el remuestreo no haya un desbalance.

Previo a la división de los datos se procede por filtrar observaciones con valores extremadamente atípicos o que podrían carecer de sentido, en este caso se toma como límite inferior de la edad a la primera relación sexual los 9 años. Con lo que se cuentan con 49349 observaciones donde nuevamente se utilizará un 85% de los datos para entrenar y el 15% restante para evaluar los modelos considerados.

Selección de covariables relevantes

Nuevamente al igual que en el caso anterior existen potencialmente variables irrelevantes para el problema, con lo cual se decide realizar nuevamente un modelo de *Random Forest* y visualizar la importancia de las variables. El procedimiento utilizado es el mismo que en el caso anterior, donde se prueba una grilla de 30 combinaciones para *mtry* y *min_n*, seleccionando el mejor modelo en este caso al estar trabajado en un problema de clasificación se utiliza como métrica el área bajo la curva *ROC* (*roc_auc*).

En la Figura 19 se observa la importancia de las variables consideradas en el modelo seleccionado. Al igual que en la Figura 10, variables relacionadas con la educación o riqueza del hogar tienden a tener mayor importancia. De todas formas existen algunas diferencias sobre todo en el orden de importancia, aquí vemos que la cantidad de parejas de la mujer parecería incidir más en la edad a la que comienzan a tener relaciones sexuales. Además cabe destacar el aumento de importancia en este caso de la frecuencia con la que la mujer ve noticias, ésta información suele ser de interés en este tipo de estudios.

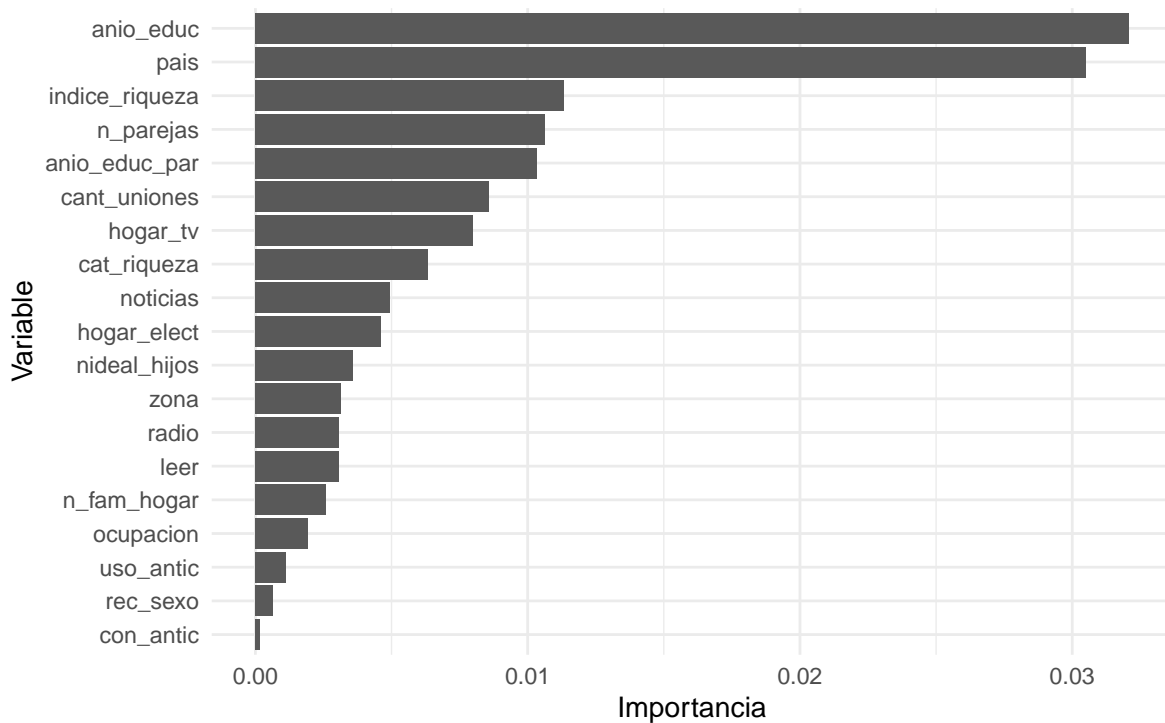


Figura 19: Gráfico barras, importancia de las variables consideradas en el modelo de Random Forest mediante permutación para predecir si comienza a tener relaciones luego de los 18 años.

Una vez visualizados los datos y las posibles relaciones entre las variables se procede a armar los modelos. De lo visto, mediante el modelo de *Random Forest*, se seleccionan las 10 variables que más importancia presentaron para clasificar si una mujer comienza su vida sexual antes o después de los 18 años.

En este caso las variables que quedaron seleccionadas son:

1. Años de educación de la mujer.
2. País de procedencia de la mujer.
3. Índice de riqueza del hogar.
4. Numero de parejas sexuales que ha tenido la mujer.
5. Años de educación de la pareja de la mujer.
6. Cantidad de uniones que ha tenido la mujer.
7. Si el hogar dispone de televisión.
8. Categoría de riqueza del hogar.
9. La frecuencia con la que la mujer ve las noticias.
10. Si el hogar tiene acceso a electricidad.

En esta ocasión se propone el uso de modelos basados en árboles: *Bagging* y *Random Forest*. Además, se ajusto un modelo de regresión logística para evaluar si los resultados obtenidos por métodos de árboles o un método mas clásico como la logística difieren en su *performance* en este problema.

Búsqueda de Hiperparámetros

En este caso se ajustan los hiperparámetros de manera similar a lo realizado en el caso anterior, se ajustan en el caso de *Bagging* la cantidad de árboles y el tamaño mínimo de nodo, mientras que en el caso de *Random Forest* se ajustan la cantidad de variables a considerar en cada nodo (*mtry*), la cantidad mínima de observaciones en cada nodo terminal (*min_n*) y la cantidad de árboles (*trees*). En el caso de la regresión logística no requiere ajustes, ni hiperparámetros aunque si se visualizó la significancia de las covariables del modelo.

Las grillas utilizadas para ambos modelos son las mismas que se utilizaron para la edad al primer hijo, observándose si la misma es adecuada para este problema. En la Figura 20 y Figura 21 se pueden observar los resultados de la búsqueda de hiperparámetros para los modelos de *Bagging* y *Random Forest* respectivamente. Donde se presentan las metricas de *accuracy*, *Brier Score* y área bajo la curva *ROC*.

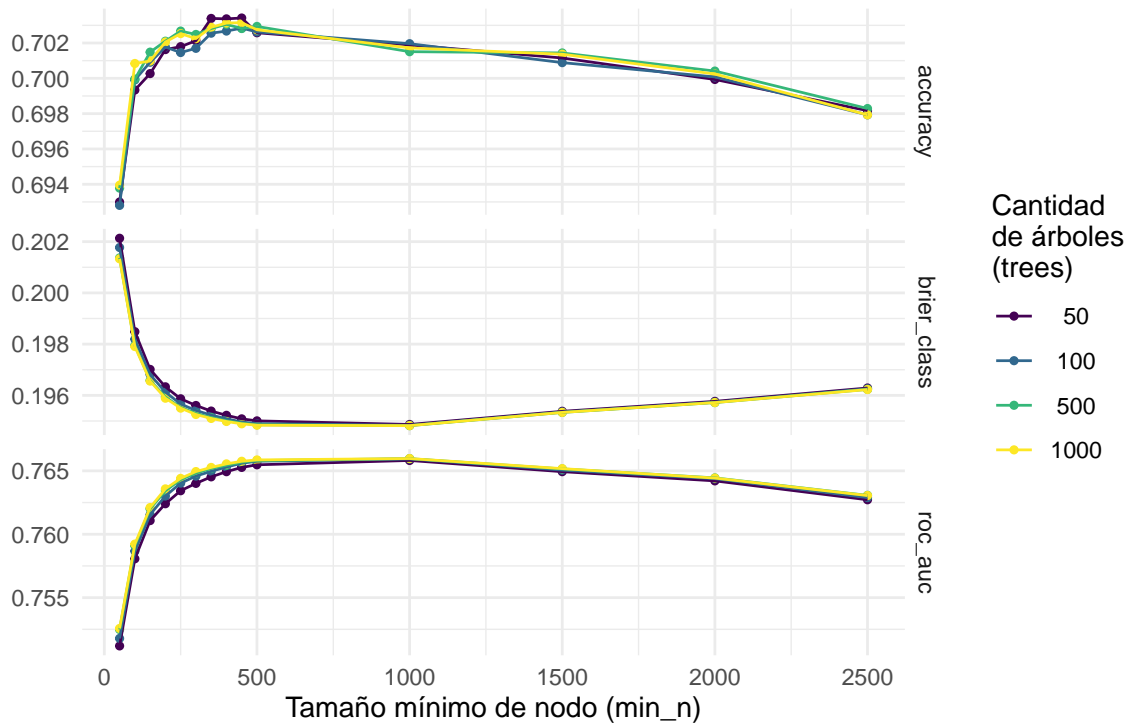


Figura 20: Gráfico, resultados de la búsqueda de hiperparámetros para modelos de Bagging.

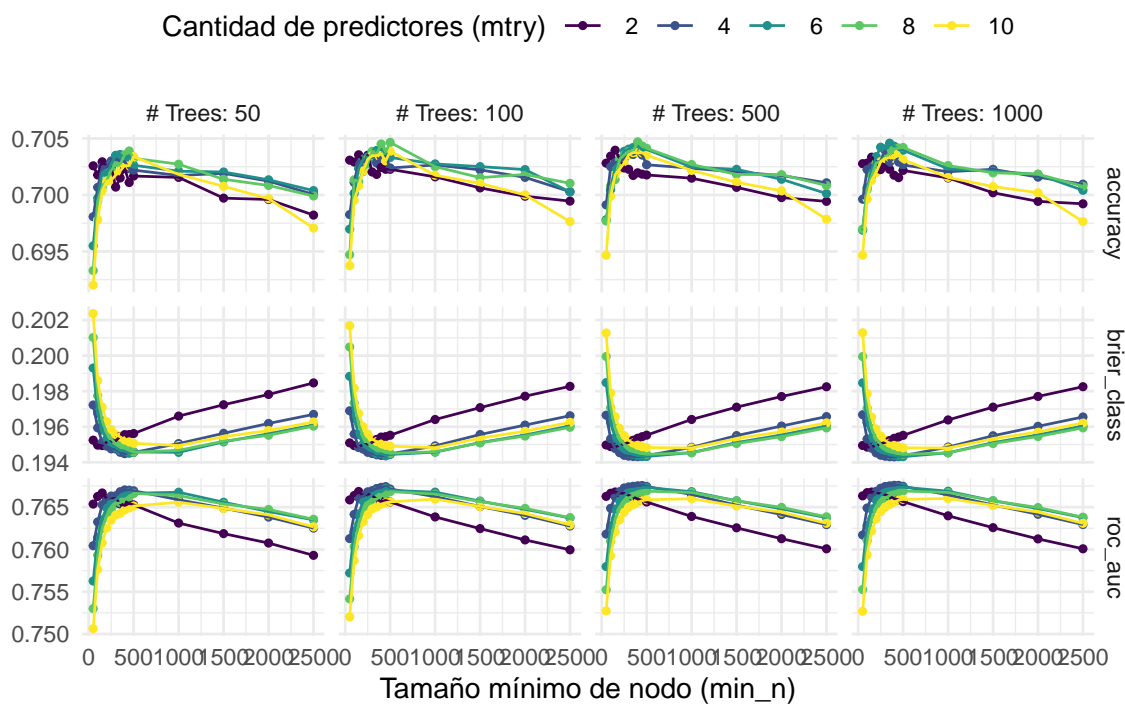


Figura 21: Gráfico, resultados de la búsqueda de hiperparámetros para modelos de Random Forest.

Comparación de modelos

En este caso en la Figura 22 se observa la comparación de los modelos de *Bagging*, *Random Forest* y *Regresión Logística* mediante la métrica de área bajo la curva *ROC*. En este caso el modelo de *Random Forest* es el que presenta un mejor desempeño, aunque al igual que en el primer problema los modelos de *Bagging* difieren muy poco en su *performance*. La regresión logística presenta un desempeño inferior algo esperable dada la gran cantidad de datos y variables aunque presenta un desempeño aceptable.

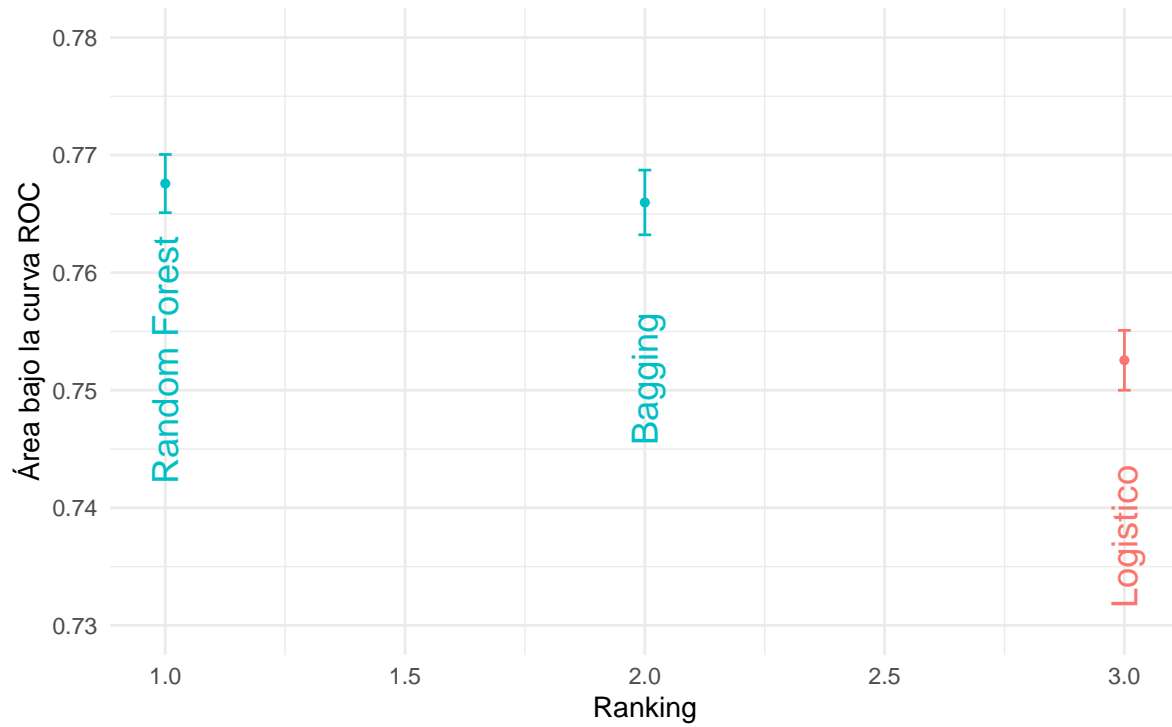


Figura 22: Gráfico comparación de los modelos de Bagging, Random Forest y Regresión Logística, mediante validación cruzada.

También podemos visualizar los rendimientos de los modelos mediante las curvas *ROC* en la Figura 23. En este caso también se observa como los modelos de *RF* y *Bagging* presentan una *performance* algo superior a la regresión logística.

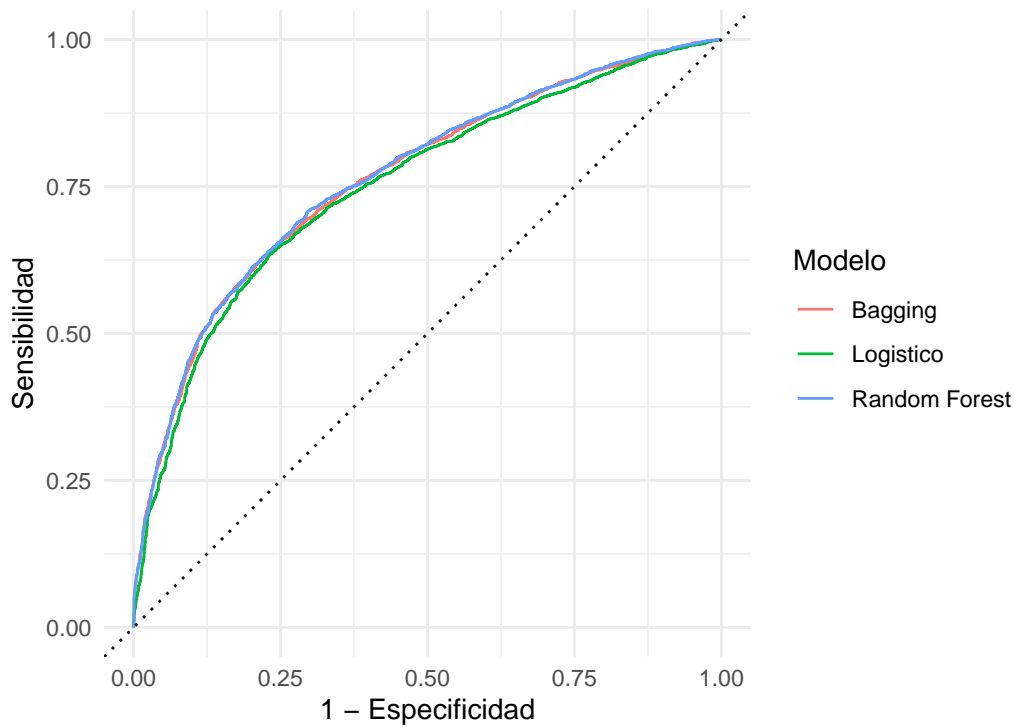


Figura 23: Gráfico, comparación curvas ROC para los modelos considerados en el conjunto de testeo

Evalación modelo final

En la Figura 24 se observa la matriz de confusión para el modelo de *Random Forest* en el conjunto de testeo. En este caso el rendimiento del modelo parece aceptable, donde está ciertamente balanceado en la cantidad de falsos positivos y falsos negativos. Aunque el mismo presenta un desempeño algo inferior en clasificar mujeres que comienzan su vida sexual antes de los 18 años.

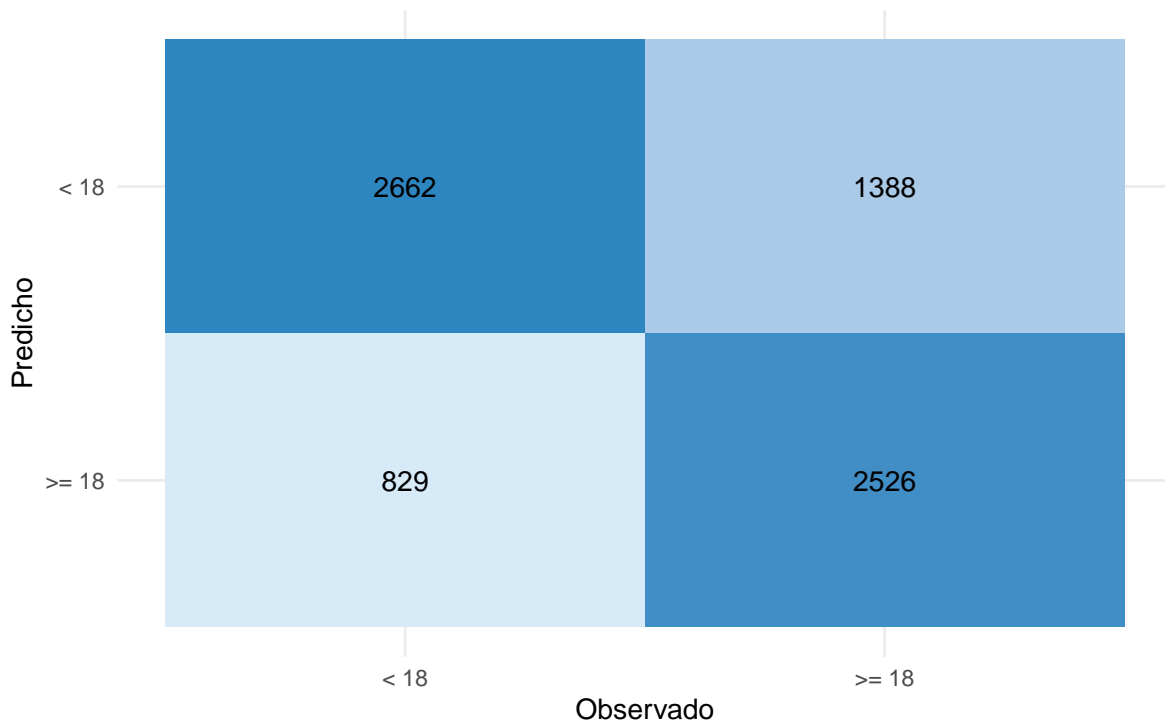


Figura 24: Matriz de confusión para el modelo de Random Forest en el conjunto de testeo.

A partir de la misma es posible calcular las métricas para evaluar el rendimiento del modelo, se puede observar en la siguiente tabla las métricas más importantes. En este caso la precisión del modelo supera por encima del 70% lo cual es un valor aceptable considerando el problema planteado. En cuanto a la sensibilidad y especificidad del modelo se observa que la especificidad es algo inferior a la sensibilidad, al igual que lo observado en la matriz de confusion el modelo tiene mayor dificultad en clasificar correctamente las mujeres que comienzan a tener relaciones sexuales antes de los 18 años.

Tabla 7: Métricas en el conjunto de testeo para predecir si comienza a tener relaciones luego de los 18 años

Metrica	Valor
Accuracy	0.70
Sensibilidad	0.76
Especificidad	0.65

Interpretación de las variables

Al igual que el caso de la edad al primer hijo se procede a visualizar la interpretación de las variables en el modelo mediante los gráficos de dependencia parcial.

Se observa en la Figura 25 que el efecto de los años de educación de la mujer en la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años. Se observa que la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años aumenta de forma abrupta a partir de los 10 años de educación el efecto es similar al observado en el caso de la edad al primer hijo.

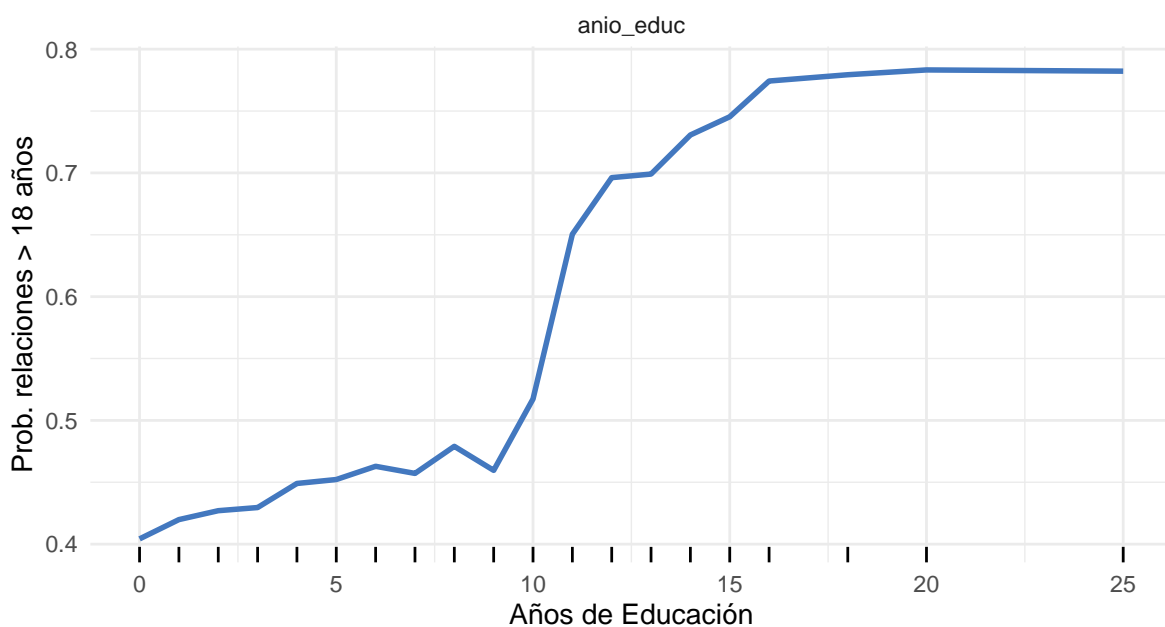


Figura 25: Gráfico PDP, efecto de los años de educación en la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años.

En cuanto al índice de riqueza el efecto parece menos marcado que en el caso anterior, en la Figura 26 se observa que la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años incrementa pero de forma irregular al aumentar el índice de riqueza del hogar.

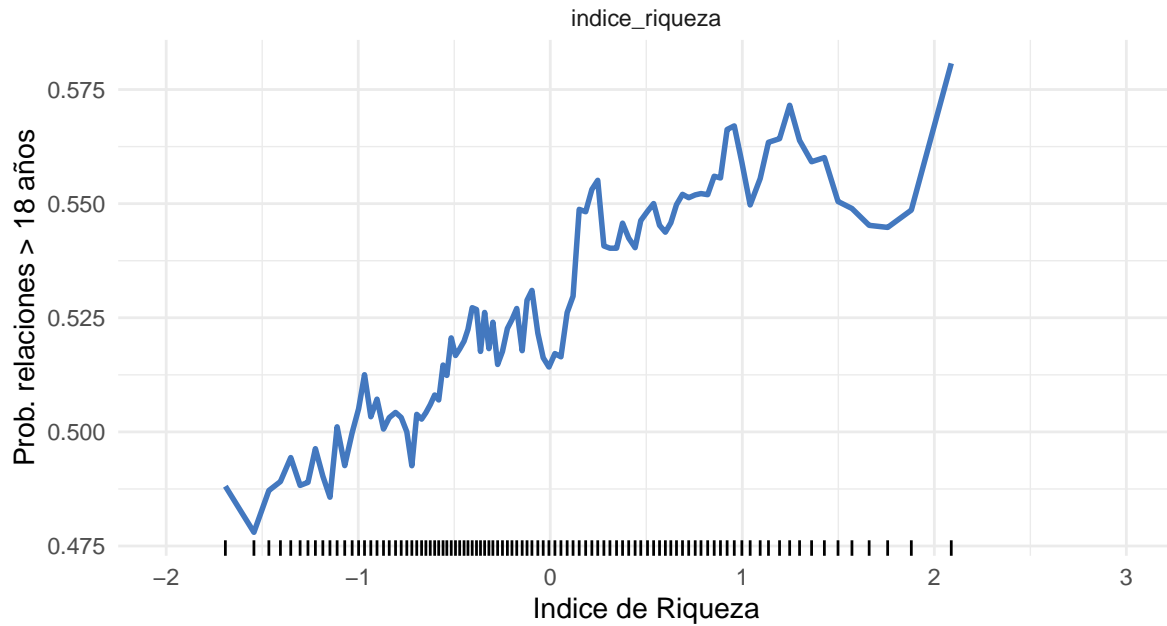


Figura 26: Gráfico PDP, efecto del índice de riqueza en el hogar en la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años.

En cuanto al país de residencia de la mujer, al igual que como se observó en el análisis exploratorio, es esperable que países como Albania o Armenia presenten una mayor probabilidad de comenzar a tener relaciones sexuales luego de los 18 años. En la Figura 27 se observa algo razonable con lo visto donde Armenia y Albania están por encima del resto de los países.

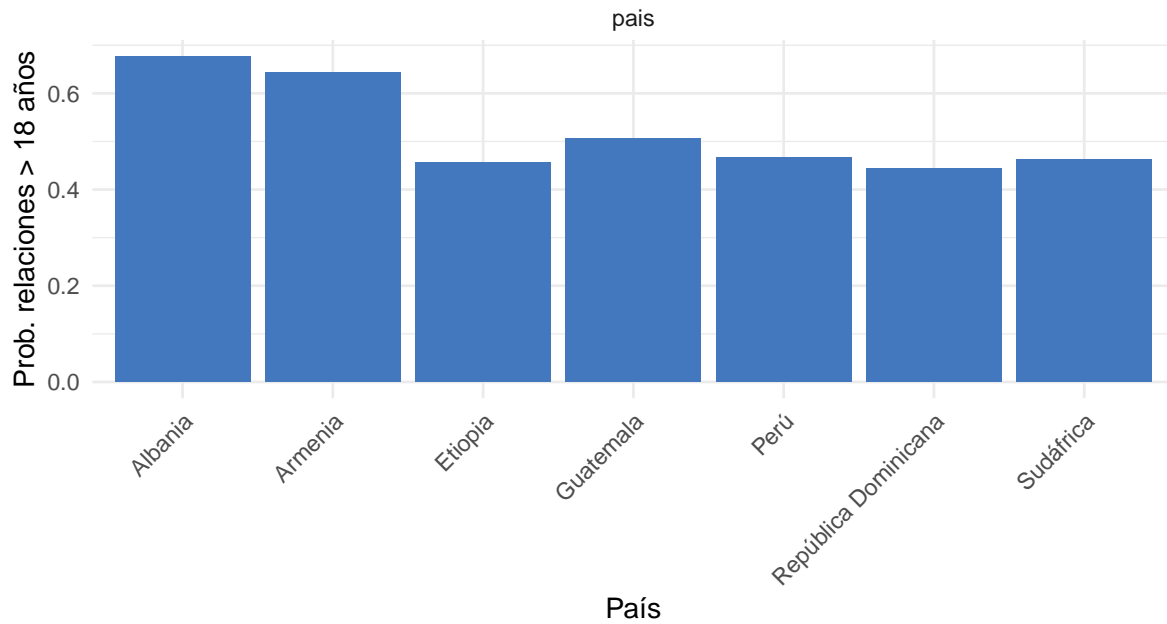


Figura 27: Gráfico PDP, efecto del país de la mujer en la probabilidad de comenzar a tener relaciones sexuales luego de los 18 años.

Comentarios Finales

Como se observó el resultado general de los modelos fue aceptable en ambos casos, teniendo en cuenta que las variables de interés son de naturaleza compleja y el error irreducible en este tipo de caso es elevado. En cuanto a las mejoras posibles se debería considerar formas de mejorar la predicción en el caso de la edad del primer hijo para mujeres de mayores edades algo que no se ha conseguido captar en este caso, algo que posiblemente se deba a que no se tengan covariables óptimas para captar este fenómeno.