

Modelado de Variables Demográficas Mediante Algoritmos de Aprendizaje Supervisado

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 Bagging, Random Forest y XGBoost.

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.

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

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.

Selección y 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, Haití, Guatemala y República Dominicana.
- Europa del este: Albania y Armenia.
- Norte y Sur de África: Etiopía, Ghana, Sudáfrica y Angola.

En este caso, se dispone de un total de aproximadamente 150,000 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 el proceso de descarga si se seleccionan demasiadas variables puede ser muy lento. Por esta razón se realizó un filtrado previo de las variables a utilizar. Como resultado, se trabaja principalmente con 17 variables que están disponibles para todos los países seleccionados y que, a priori, se consideran las más relevantes para el análisis. Algunas de estas variables no se emplean como predictores, pero resultan útiles en otros aspectos, como el peso muestral de las mujeres en la encuesta, que puede ser incorporado en el modelo.

Tabla 1: Descripción de las variables disponibles en el trabajo

Código DHS	Nombre	Descripción	Tipo	Recorrido
v000	pais	País donde se realizó la encuesta	Categórica	10 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 Entera	2008-2018
v012	edad	Edad de la mujer encuestada	Numérica Entera	15-64
v102	urbano	Indica si la mujer es de área urbana o rural	Categórica	1:Urbano, 2:Rural
v119	electricidad	Indica si el hogar tiene acceso a electricidad	Categórica	0:No, 1:Si
v121	tv	El hogar dispone de una televisión	Categórica	0:No, 1:Si
v133	anio_educ	Total años de educación de la mujer, se calcula a partir del nivel máximo de educación alcanzado, es un valor comparable entre países.	Numérica Entera	0-25

v136	tot_fam	Total de personas que viven en el hogar	Numérica Entera	1-25
v157	leer	Frecuencia con que la mujer lee un periódico, revista o libro	Categórica	0:Nada, 1:Ocasional, 2:Todos los días
v191	ind_riqueza	Indicador de riqueza del hogar estandarizado, se calcula a partir de distintas medidas como materiales de construcción de la casa y servicios disponibles, su calculo puede variar entre países.	Numérica	-4.992-3.46833
v212	edad_phijo	Edad de la mujer al primer hijo	Numérica Entera	3-47
v301	con_antic	Conocimiento de algún método anticonceptivo, donde se distingue entre modernos (ej: Condon, pastillas, parches, etc.), tradicionales (ej: Coito interrumpido, calendario, etc.) y folklóricos (ej: Amuletos, rituales, etc.)	Categórica	0: Ninguno, 1:Folklórico, 2:Tradicional, 3:Moderno
v525	edad_psexo	Edad de la mujer al primer acto sexual	Numérica Entera	0-49
v714	trabajo	Indica si la mujer actualmente trabaja	Categórica	0: No, 1: Si
v836	n_parejas	Cantidad de parejas sexuales que ha tenido la mujer	Numérica Entera	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.

En cuanto a las variables de interés, como la edad de la mujer al primer hijo, se filtran las observaciones con valores extremadamente atípicos o que no tienen sentido desde el punto de vista biológico, como un valor de 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.

Respecto a la otra variable de interés, la edad a la primera relación sexual, se decide filtrar a las mujeres que reportaron valores menores a 10 años, con el fin de evitar casos extremadamente atípicos o errores en el registro de 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 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 viendo de países con contextos muy distintos. En países como Angola, 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, patrón que usualmente se observa 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.

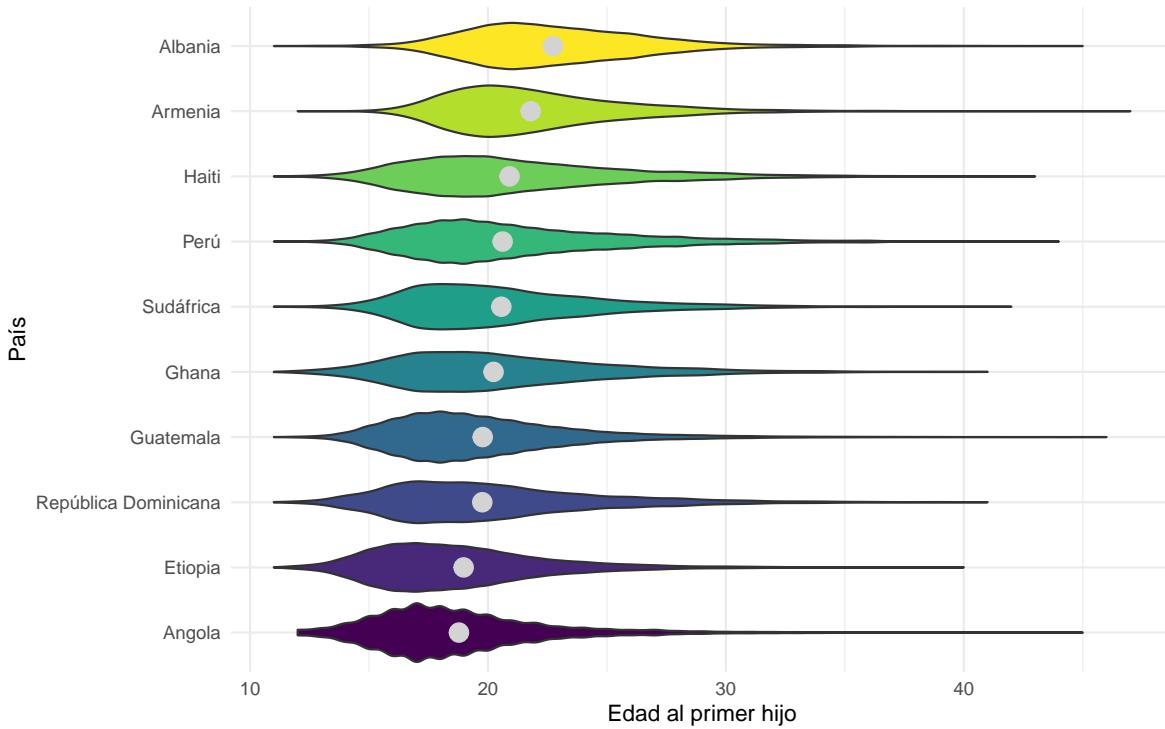


Figura 1: Gráfico violín, distribución de la edad al primer hijo por país, donde el punto indica la edad media.

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 debería ser útil para predecir la variable de interés.

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 mas de 15 años de educación. 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.

Otro factor que posiblemente afecte es el nivel de riqueza y la zona del hogar de pertenencia de la mujer, en la Figura 4, se observa que existe un efecto de atrasamiento en la edad al

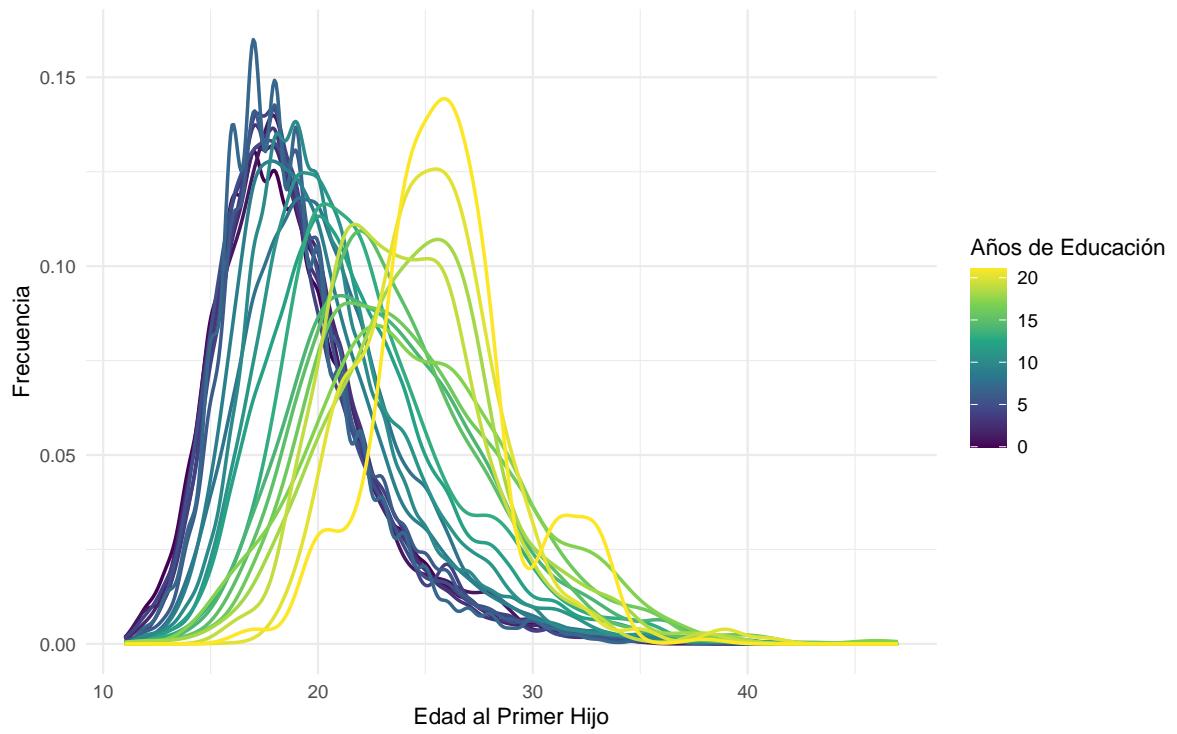


Figura 2: Gráfico Líneas, distribución de la edad al primer hijo por Años de Educación de la Madre.

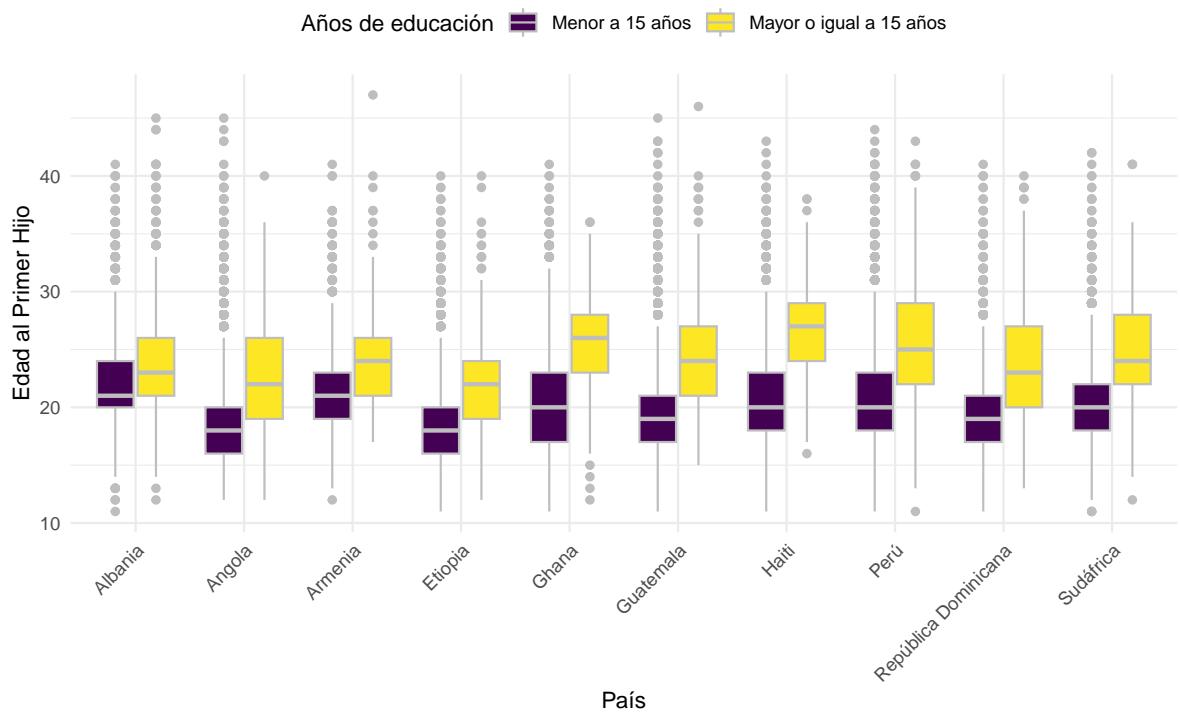


Figura 3: Gráfico Caja, distribución de la edad al primer hijo por país y si la mujer tiene más de 15 años de educación.

primer hijo a medida que el índice de riqueza aumenta, aunque el efecto no es tan marcado y presenta mayor dispersión para mujeres con mayores niveles de la variable, donde el efecto parece no estar influenciado por si la zona es urbana o rural siendo muy similar en distribución. Es destacable que en este caso no se está midiendo la riqueza al momento de tener el hijo, por lo que el contexto de la mujer puede haber cambiado en el proceso aunque sigue pareciendo un indicador posiblemente relevante en la predicción.

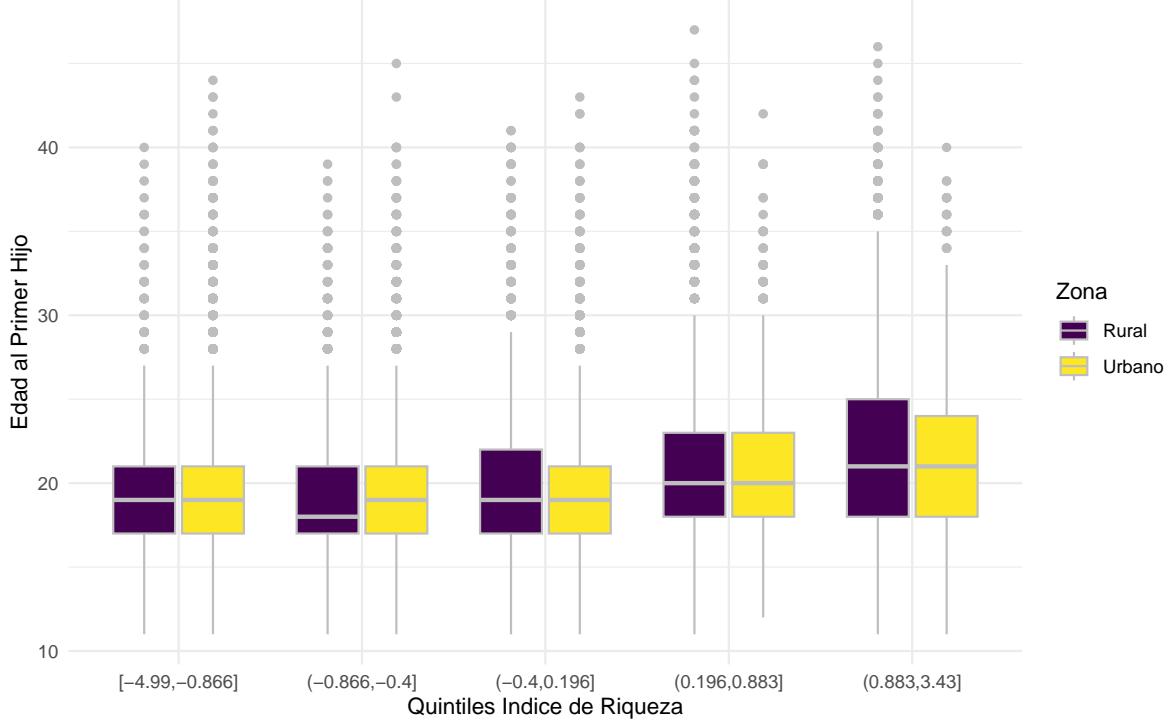


Figura 4: Gráfico Caja, distribución de la edad al primer hijo por país y si la mujer tiene más de 15 años de educación.

Por otro lado variables como el comportamiento o la educación sexual de la mujer también pueden influir en la edad al primer hijo. En la Figura 5 se observa la distribución de la edad al primer hijo diferenciando por la cantidad de parejas sexuales que han tenido las mujeres y si la misma conoce algún método anticonceptivo, tanto tradicional como moderno. En este caso, se observa que las mujeres que han tenido más parejas sexuales tienden a tener su primer hijo a una edad más temprana, aunque el efecto no es tan marcado. Por otro lado, las mujeres que conocen algún método anticonceptivo tanto tradicional como moderno presentan una edad al primer hijo más avanzada, algo esperable para cualquiera de las cantidades de parejas que hayan tenido.

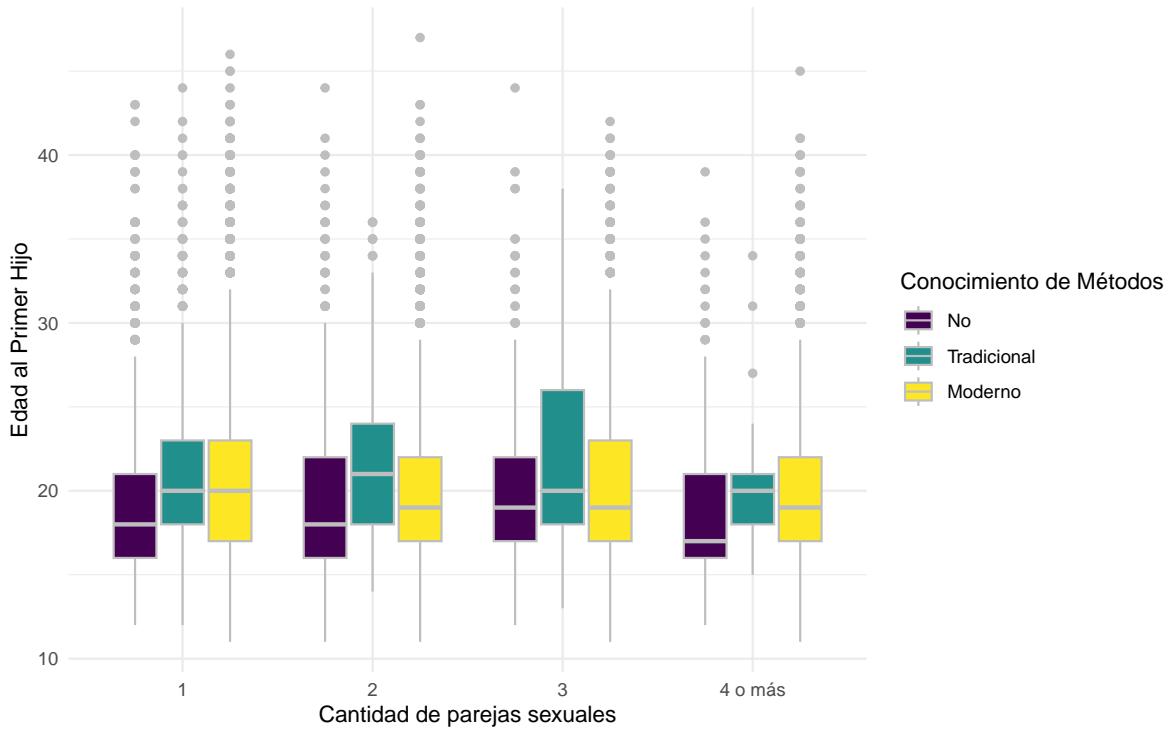


Figura 5: Gráfico Caja, distribución de la edad al primer hijo por Años de Educación de la Madre y por comienzo de relaciones sexuales.

Modelado de la edad al primer hijo

Una vez visualizados los datos y posibles relaciones en las variables se procede a armar los modelos.

Para incorporar información sobre el contexto social y económico se toman en cuenta estas 5 variables, algunas de ellas hemos visualizado su efecto, siendo este relevante:

- País de procedencia de la mujer.
- Índice de riqueza del hogar.
- Si el hogar dispone de televisión.
- Si la mujer se encuentra en una área urbana o rural.
- Si la mujer trabaja en la actualidad.

Por lado del nivel educativo de la mujer se toman en cuenta las siguientes variables:

- Años de educación de la mujer.
- Si la mujer lee periódicos, revistas o libros.

Por último, se consideran las siguientes variables relacionadas con el comportamiento sexual de la mujer:

- Cantidad de parejas sexuales que ha tenido la mujer.
- Si la mujer conoce algún método anticonceptivo.

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 entrena al modelo con 80359 observaciones de mujeres donde se utilizara un 85% de los datos para entrenar y el 15% restante para evaluar el modelo.

En este caso se trabaja integralmente 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 CatBoost, donde para cada uno se realiza su correspondiente tuneo de hiperparámetros.

Bagging

Búsqueda de Hiperparámetros

Para evaluar la eficiencia de los modelos de bagging, se realiza mediante la técnica de validación cruzada la búsqueda de hiperparámetros óptimos. En este caso se utiliza ‘k-folds cross validation’ con 10 folds, donde se busca como hiperparámetros el número mínimo de observaciones por nodo (*min_n*) y la cantidad de árboles a utilizar (*trees*). Como la cantidad de

observaciones es muy grande se opta por armar una grilla con valores de 50 a 5000 para el número mínimo de observaciones, para el número de árboles que no es tan importante como el anterior se opta por valores de 10 a 1000 pero más distanciados. Como engine en este caso se utiliza ‘ranger’.

Tabla 2: Grilla de hiperparámetros para modelos de Bagging

min_n	50	100	150	200	250	300	350	400	450	500	1000	1500	2000	2500	5000
trees	10	50	100	250	500	1000									

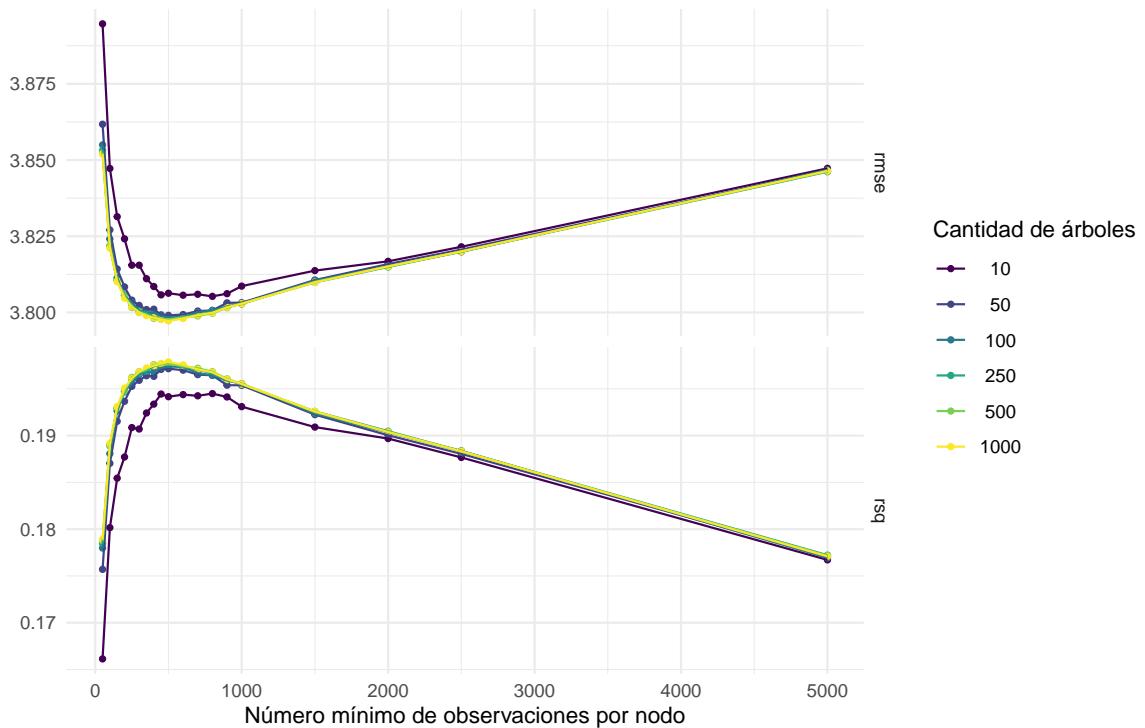


Figura 6: Gráfico de líneas, métrica RMSE y RSQ en función del número mínimo de observaciones por nodo, donde cada línea corresponde a la cantidad de árboles.

Evaluación del modelo

Una de las primeras cosas que salen a la vista al evaluar la performance del modelo es que el mismo presenta un rango de predicción con los datos entrenados desde 17 a 30 años, no contemplando los casos no tan comunes de mujeres que tienen su primer hijo a edades mayores a 30 años, a su vez es extraño que el mismo no presente valores menores a 17 años siendo no tan fuera de lo usual observaciones en los datos con 16 o 15 años.

Observando el análisis descriptivo puede ser que la causa sea que ninguna de las covariables consideradas permita captar y distinguir este tipo de casos de mujeres con hijos en altas edades. También se probó la transformación de la variable, tanto utilizando el logaritmo como la raíz cuadrada, pero sin obtener resultados satisfactorios, obteniéndose un rango de edades similar predicho.

En la Figura 7 se puede visualizar la distribución de los valores predichos con los datos de entrenamiento donde se observa lo que se mencionó, el rango se encuentra acotado, la distribución presenta una moda en los 19 años, lo que es bastante coherente con lo observado en los datos.

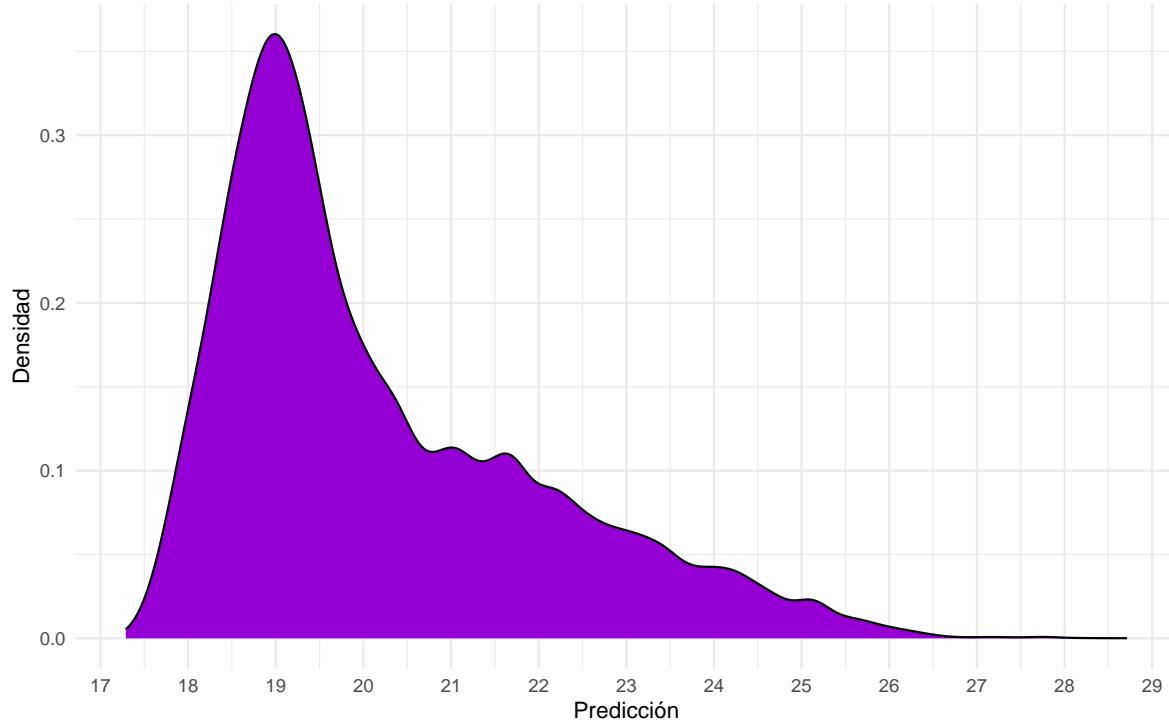


Figura 7: Gráfico distribución, valores predichos por el modelo de Bagging con los datos de entrenamiento.

En la Figura 8 se observa la relación entre las predicciones del modelo Bagging y los residuos de dichas predicciones, junto con una curva ajustada mediante el método de suavizado LOESS. En general, los residuos parecen estar distribuidos de manera homogénea a lo largo del rango de predicciones, sin una tendencia clara, aunque en su mayoría se presentan de forma aleatoria, se nota que el modelo aún muestra una dispersión considerable en alguna zona. Esto indica que, aunque el desempeño general sea bueno, hay casos específicos donde el modelo podría estar cometiendo errores más grandes en sus predicciones.

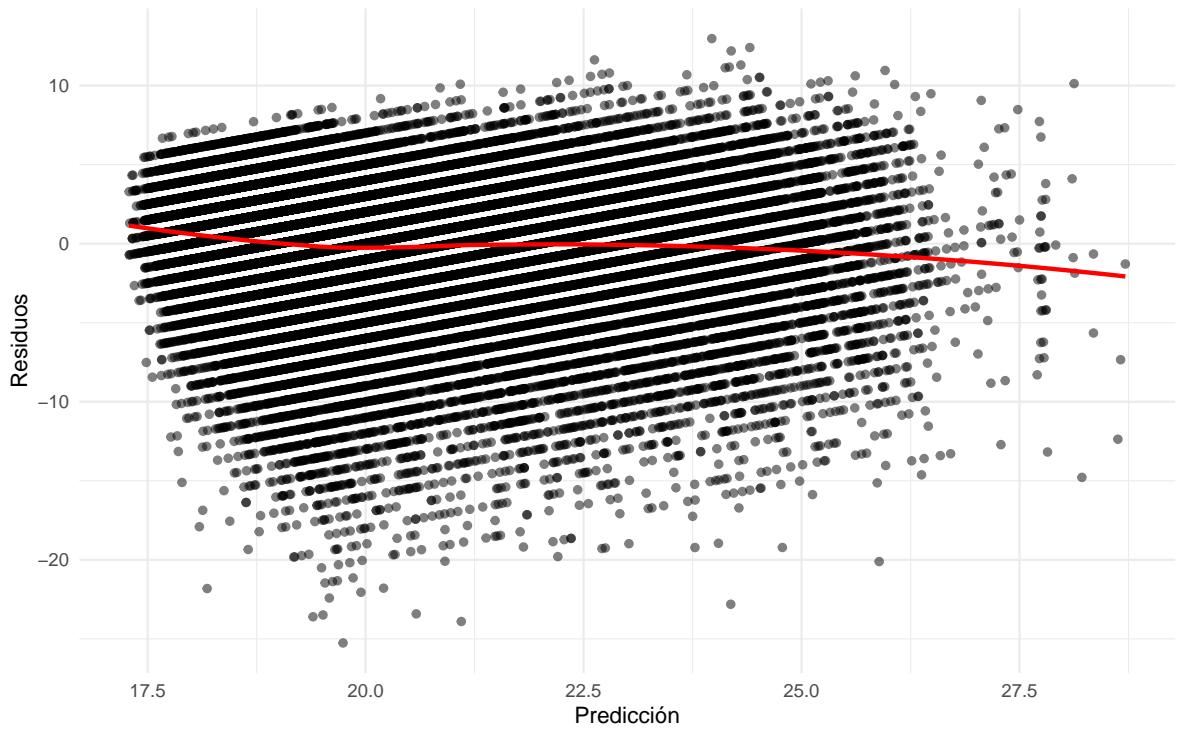


Figura 8: Gráfico de puntos, predicción vs residuos del modelo de Bagging.

Interpretación e importancia de las variables

En esta sección se presentarán gráficos que ayudarán a la visualización de la importancia de las variables, tanto para ver a nivel general cuales son las que tienen mayor importancia en este modelo de Bagging, como además se buscará interpretar algunas de estas variables, mediante gráficos de dependencia parcial (PDP) para apreciar el tipo de relación entre la respuesta y la variable explicativa, buscando ver si es lineal, monótona o más compleja.

En la Figura 9 vemos un gráfico de barras que nos muestra a las variables explicativas ordenadas de mayor a menor importancia dentro del modelo de Bagging. Vemos entonces que la variable de mayor importancia en el modelo es la que indica los años de educación que alcanzó la mujer, seguida por el país de presedencia y el índice de riqueza de la mujer. Llama la atención, y cabe mencionar que en este caso la variable de menor importancia sería la que hace referencia a si la mujer trabaja o no hoy en día.

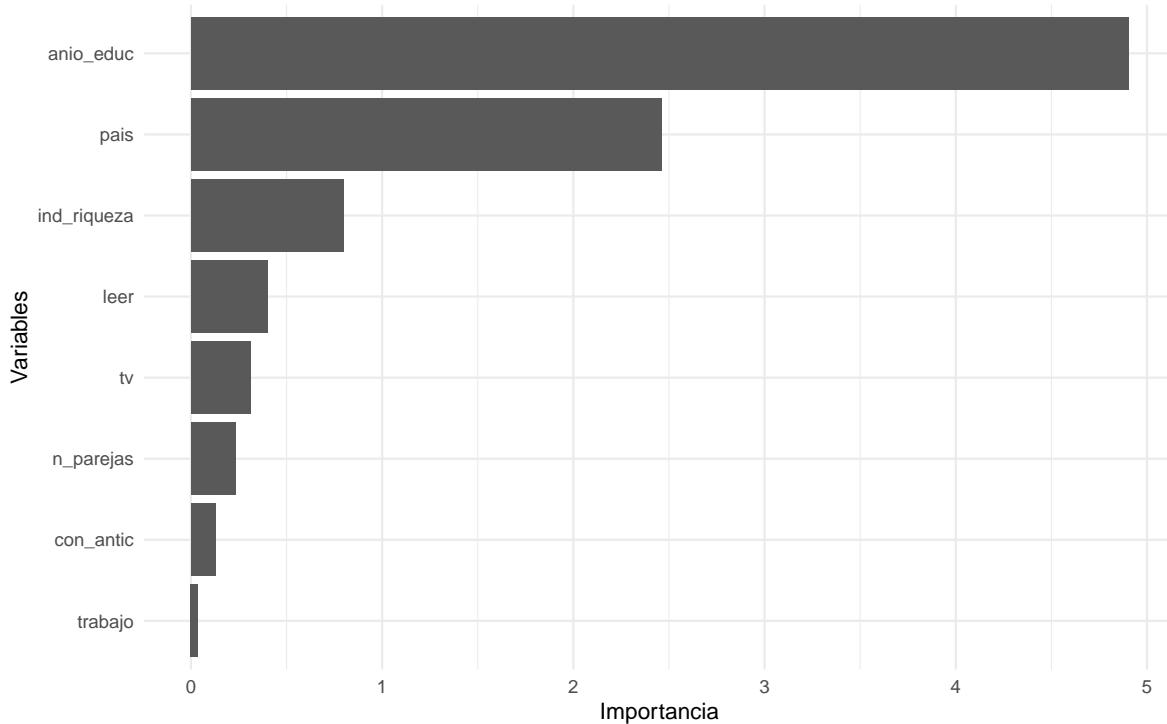


Figura 9: Gráfico barras, importancia de las variables en el modelo de Bagging.

Ahora intentaremos describir la relación entre la respuesta y algunos de los predictores de manera individual, para un mejor entendimiento, mediante gráficos PDP que permiten ver el efecto marginal de las variables explicativas en el valor predicho del modelo. En la Figura 10, observamos el efecto de los años de educación. Se puede ver que, para valores muy bajos de los años de educación, la edad media de las mujeres está entre los 19 y 20 años, hasta llegar

a los 9 años de educación. A partir de ese punto, la edad media empieza a subir de manera muy marcada, alcanzando casi los 24 años para las mujeres con 16 años de educación. Luego vemos cómo sigue subiendo, pero es importante tener en cuenta la cantidad de observaciones. Puede estar ocurriendo que haya años de educación con pocas observaciones, lo que haría difícil afirmar que esto sigue subiendo de manera tan pronunciada.

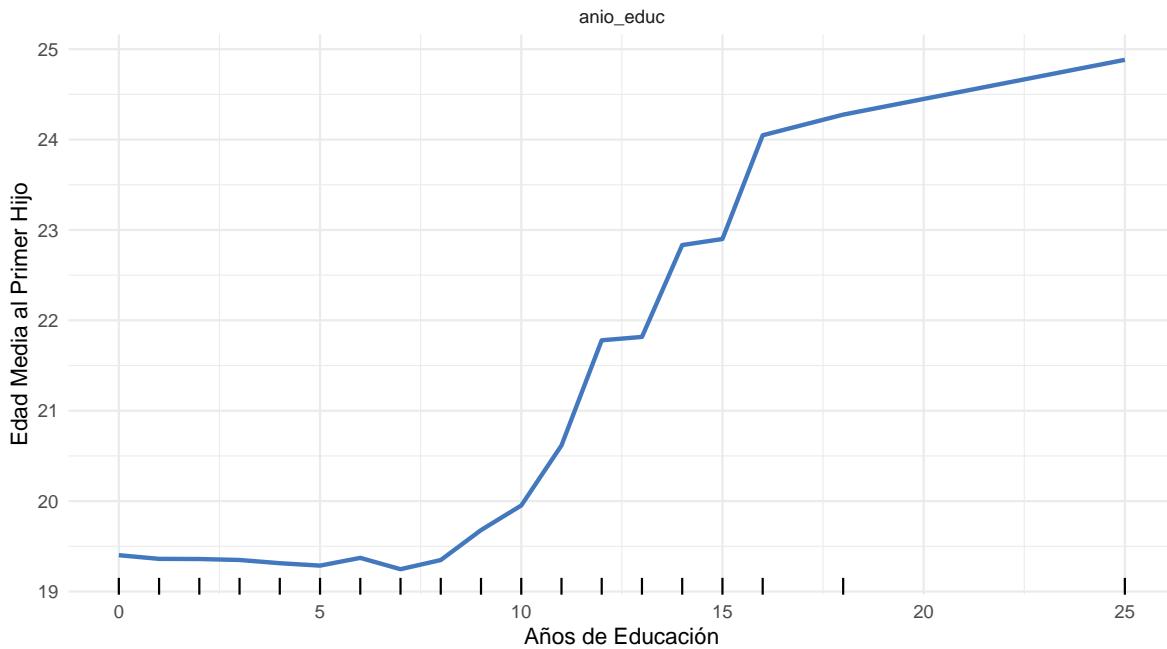


Figura 10: Gráfico PDP, efecto de los años de educación en la edad al primer hijo.

En la Figura 11 vemos el efecto del índice de riqueza en la edad al primer hijo. En esta figura se nota de forma muy clara que hay pocas observaciones en valores extremos del índice de riqueza, por lo que las edades medias en estos extremos no son tan creíbles. Sin embargo, si nos centramos en valores del índice entre -1 y 2, que es donde se encuentran la mayor cantidad de observaciones, podemos decir que la edad media de las mujeres al tener el primer hijo oscila en torno a los 20 años hasta que se alcanza el 0 en el índice. A partir de ese valor, se da un incremento casi monótono en la edad media, con algunas excepciones en ciertos valores que provocan pequeñas caídas, pero llegando, en valores cercanos a 2 del índice, a edades medias de más de 21 años.

Por último presentamos el gráfico PDP para el país de nacimiento de las mujeres en la Figura 12. En este vemos la edad media al primer hijo en forma de barra debido a que la variable país es categórica. Los países con mayor edad media son Albania y Haití con valores por encima de los 21 años, mientras que los países con valores más chicos son Angola y Dominicana con edad

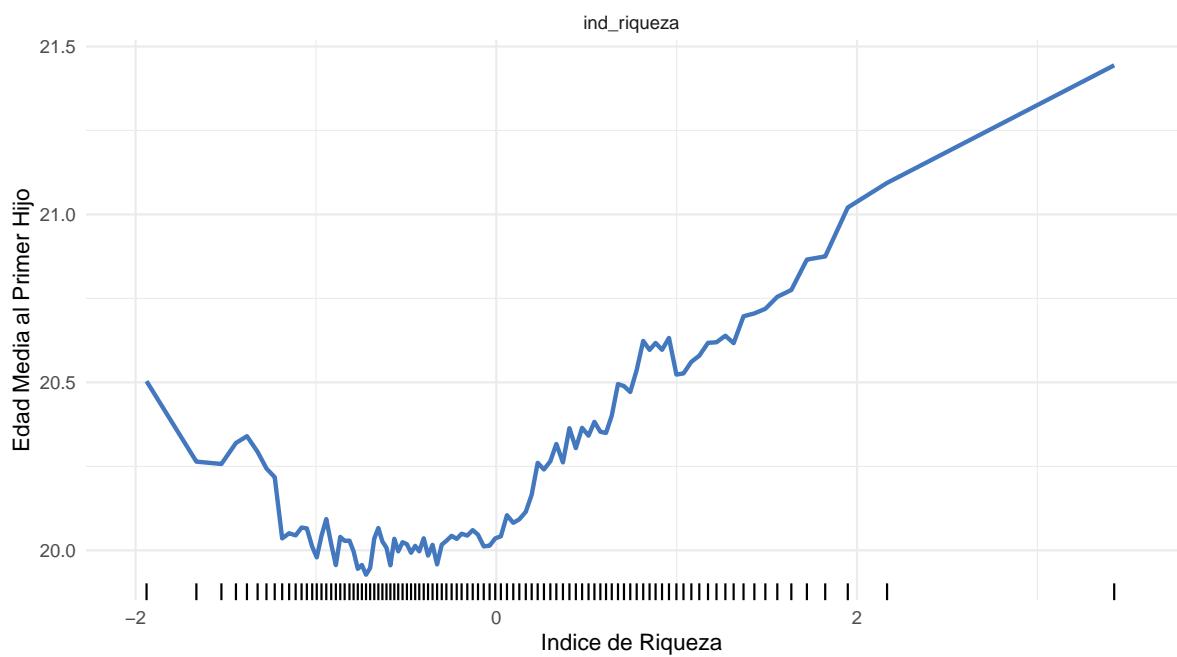


Figura 11: Gráfico PDP, efecto del indice de riqueza en la edad al primer hijo.

media de la mujer al tener el primer hijo por debajo de los 19 años.

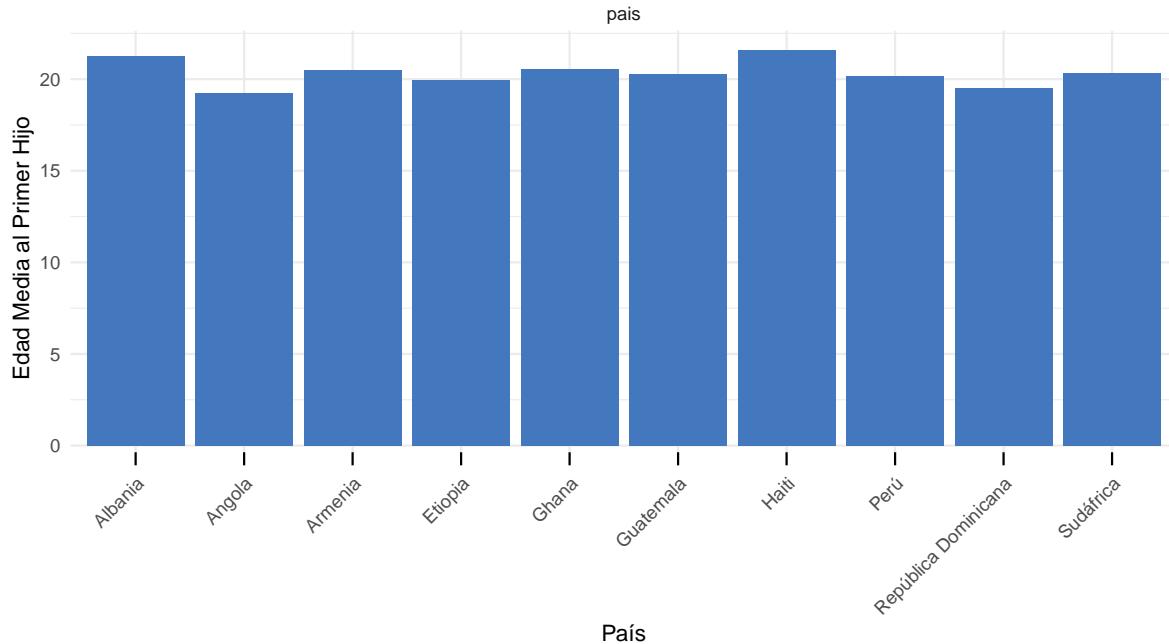


Figura 12: Gráfico PDP, efecto del país en la edad al primer hijo.

Random Forest

Busqueda de Hiperparámetros

Para evaluar la eficiencia de los modelos de Random Forest, se realiza mediante la técnica de validación cruzada la búsqueda de hiperparámetros óptimos, de la misma manera que se hizo en la sección de Bagging. Aquí se vuelve a utilizar 'k-folds cross validation' con 10 folds, donde se busca como hiperparámetros el (*min_n*) y (*trees*), como antes, pero además se busca el número de predictores que se seleccionarán aleatoriamente en cada división (*mtry*). En este caso la grilla se arma con los mismos valores para el número mínimo de observaciones y para el número de árboles, y para el número de predictores se usa una grilla que tome valores de 2 a 8 de 2 en 2. Como engine se vuelve a utilizar 'ranger'.

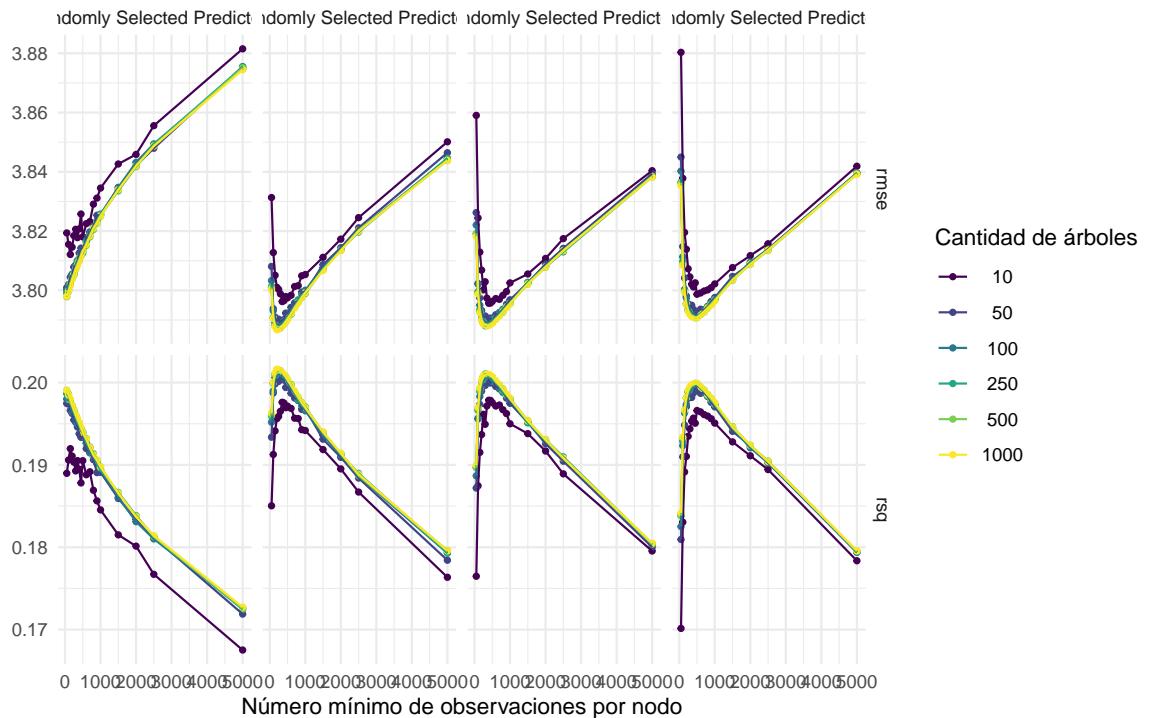


Figura 13: Gráfico de líneas, métrica RMSE y RSQ en función del número mínimo de observaciones por nodo, donde cada línea corresponde a la cantidad de árboles.

Evaluación del modelo

Al igual que en Bagging lo primero que llama la atención al evaluar la performance del modelo es que el mismo presenta un rango de predicción con los datos entrenados desde 17 a 30 años, lo cual ya discutimos en la sección anterior.

En la Figura 14 se puede visualizar la distribución de los valores predichos con los datos de entrenamiento donde se observa lo que se mencionó, el rango se encuentra acotado, la distribución presenta una moda en los 19 años, al igual que sucedían en Bagging, la diferencia en este caso es que los datos están un poco más dispersos cerca de esa moda.

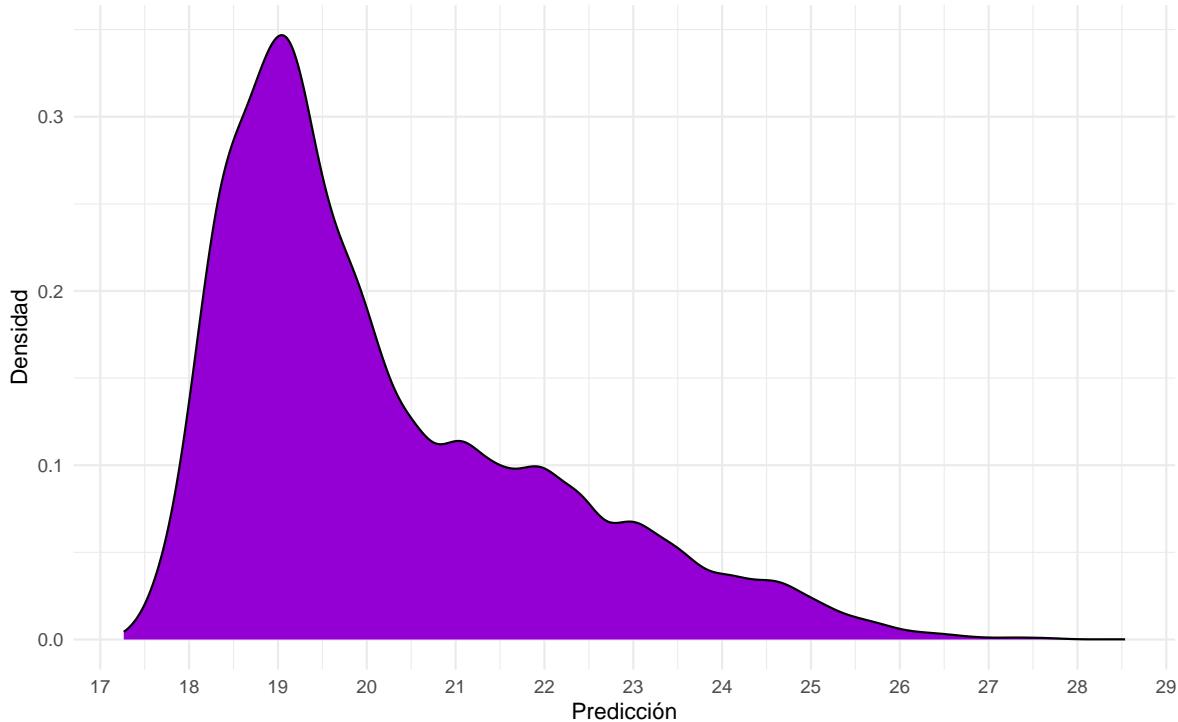


Figura 14: Gráfico distribución, valores predichos por el modelo de rf con los datos de entrenamiento.

En la Figura 15 observamos la misma relación entre las predicciones del modelo RF y los residuos de dichas predicciones, junto con la curva ajustada mediante LOESS, que observamos usando Bagging.

Interpretación e importancia de las variables

En esta sección se presentarán los mismos gráficos que ayudan a la visualización de la importancia de las variables y que sirven para apreciar el tipo de relación entre la respuesta y la

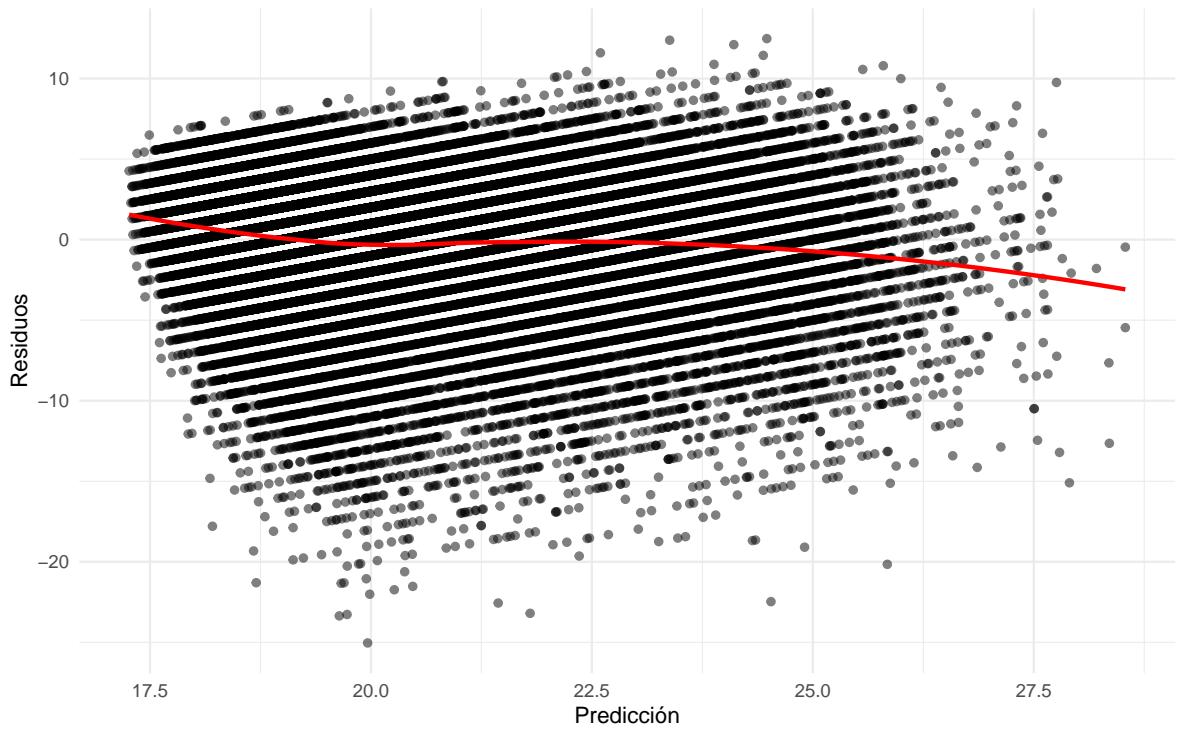


Figura 15: Gráfico de puntos, predicción vs residuos del modelo de rf.

variable explicativa usados anteriormente.

En la Figura 16 vemos un gráfico de barras que nos muestra a las variables explicativas ordenadas de mayor a menor importancia dentro del modelo de Random Forest. Vemos entonces que la variable de mayor importancia en el modelo sigue siendo la que indica los años de educación que alcanzó la mujer, y el orden sigue de la misma manera que en Bagging, pero ahora vemos algunas variables con mayor importancia que la que tenían en la otra sección, como por ejemplo el país y el índice de riqueza.

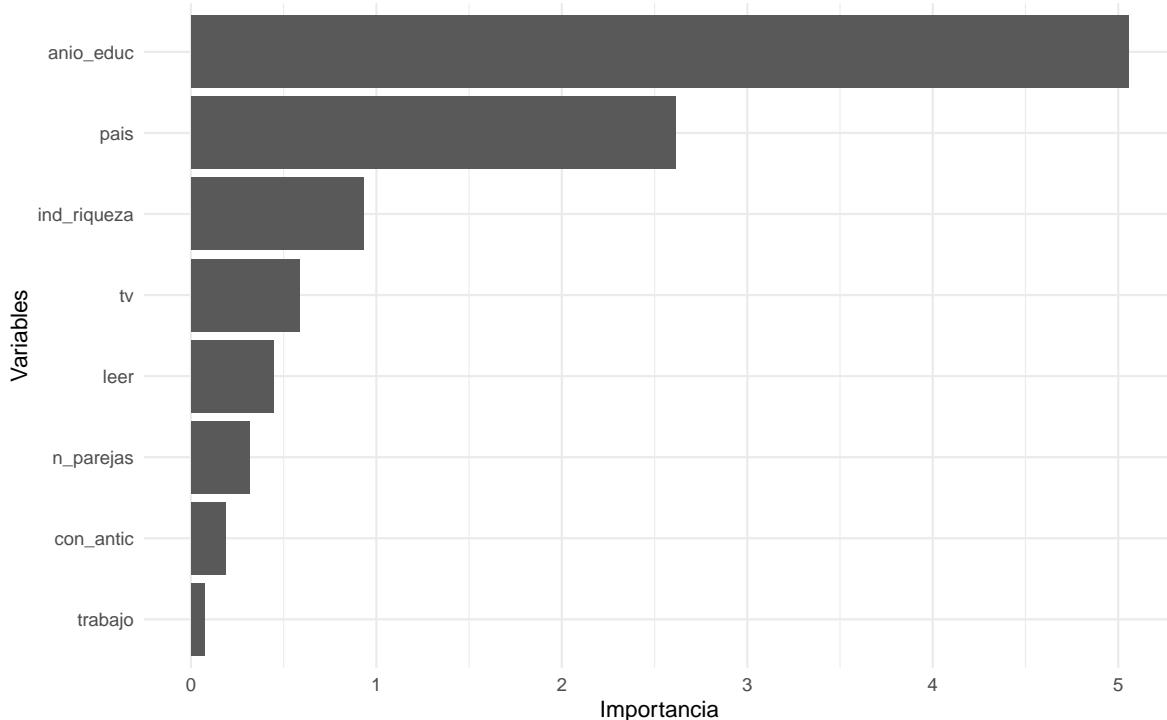


Figura 16: Gráfico barras, importancia de las variables en el modelo de rf.

Ahora intentaremos describir la relación entre la respuesta y algunos de los predictores de manera individual, para un mejor entendimiento, mediante gráficos PDP. En la Figura 17, observamos el efecto de los años de educación. Se pueden ver resultados muy similares a los de la sección pasada, pero debido al problema ya mencionado de tener pocas observaciones para algunos años de educación, vemos como de los 18 a los 25 años de educación no se mueve casi la edad media de las mujeres al tener su primer hijo, parece estabilizarse en los 24 años.

En la Figura 18 vemos el efecto del índice de riqueza en la edad al primer hijo. En esta figura se nota de forma muy clara que hay pocas observaciones en valores extremos del índice de riqueza, al igual que antes, por lo que las edades medias en estos extremos no son tan creíbles. En los valores entre -1 y 2 se aprecian resultados muy similares a los observados en Bagging,

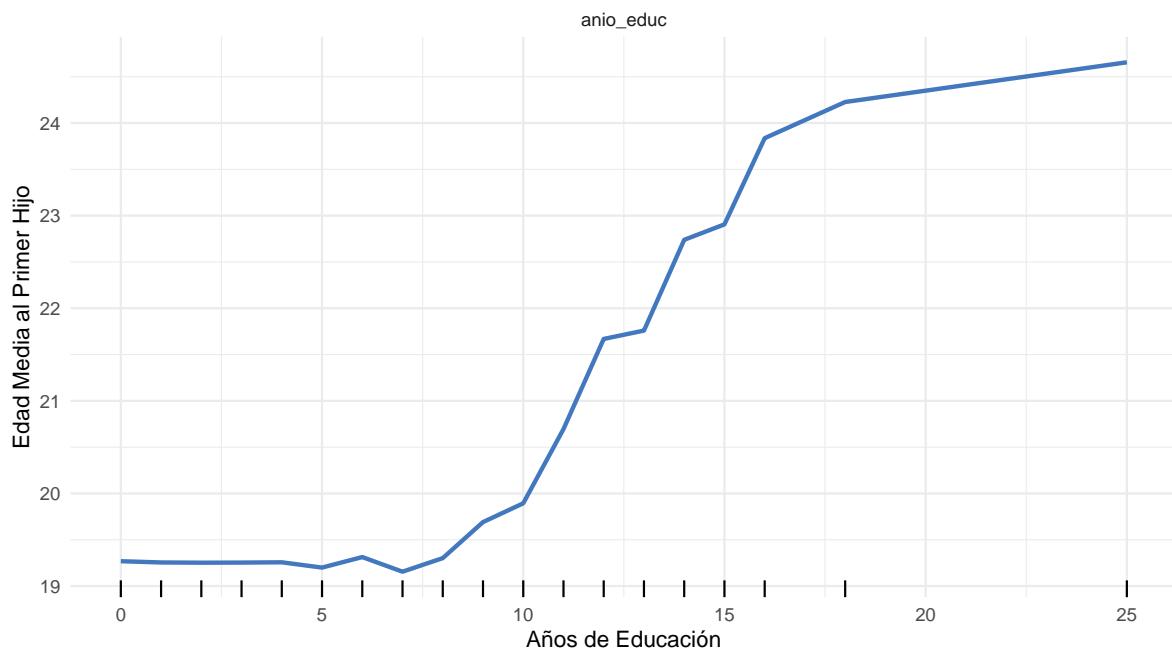


Figura 17: Gráfico PDP, efecto de los años de educación en la edad al primer hijo.

primero oscila un poco en valores cercanos a los 20 años y a partir del 0 empieza a subir la edad media de las mujeres hasta pasar los 21 años para un valor de 2 del indice de riqueza.

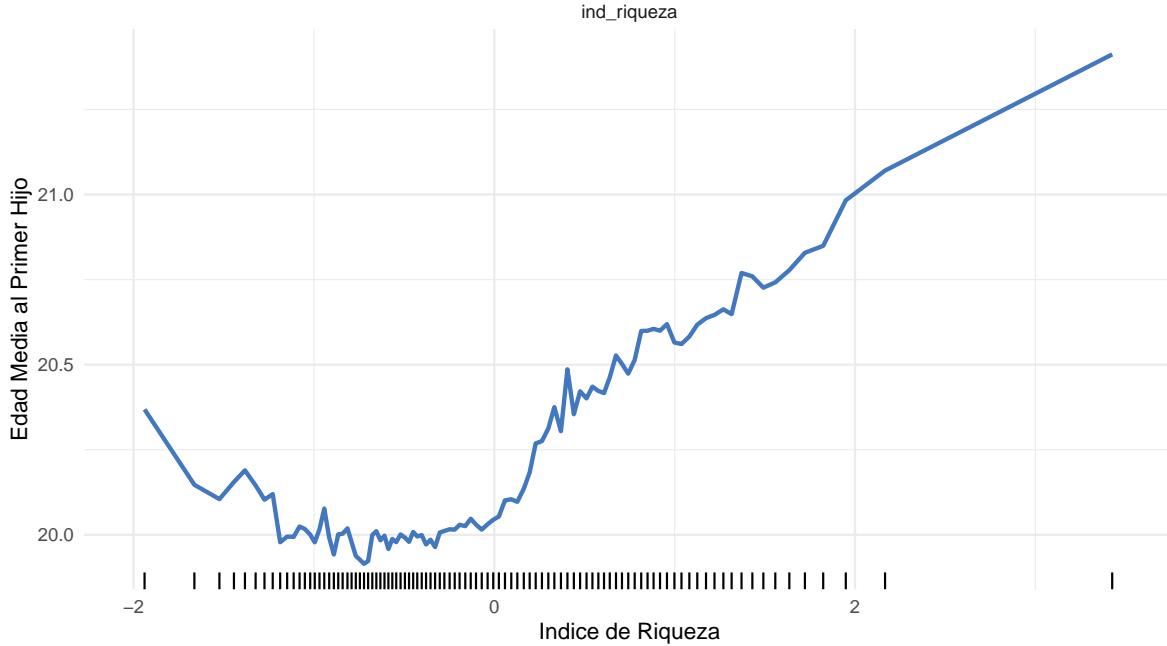


Figura 18: Gráfico PDP, efecto del indice de riqueza en la edad al primer hijo.

Por último presentamos el gráfico PDP para el país de nacimiento de las mujeres en la Figura 19. La única diferencia que vale la pena destacar en comparación con la Figura 12 es que en Haití la edad media de las mujeres al tener el primer hijo es un poco más baja y en Dominicana sucede lo contrario, ahora sube un poco la edad media.

Modelado de la edad a la primera relación sexual

Regresión Logística

```
# A tibble: 13 x 3
  .metric      .estimator .estimate
  <chr>        <chr>       <dbl>
1 accuracy    binary     0.691
2 kap         binary     0.366
3 sens        binary     0.801
```

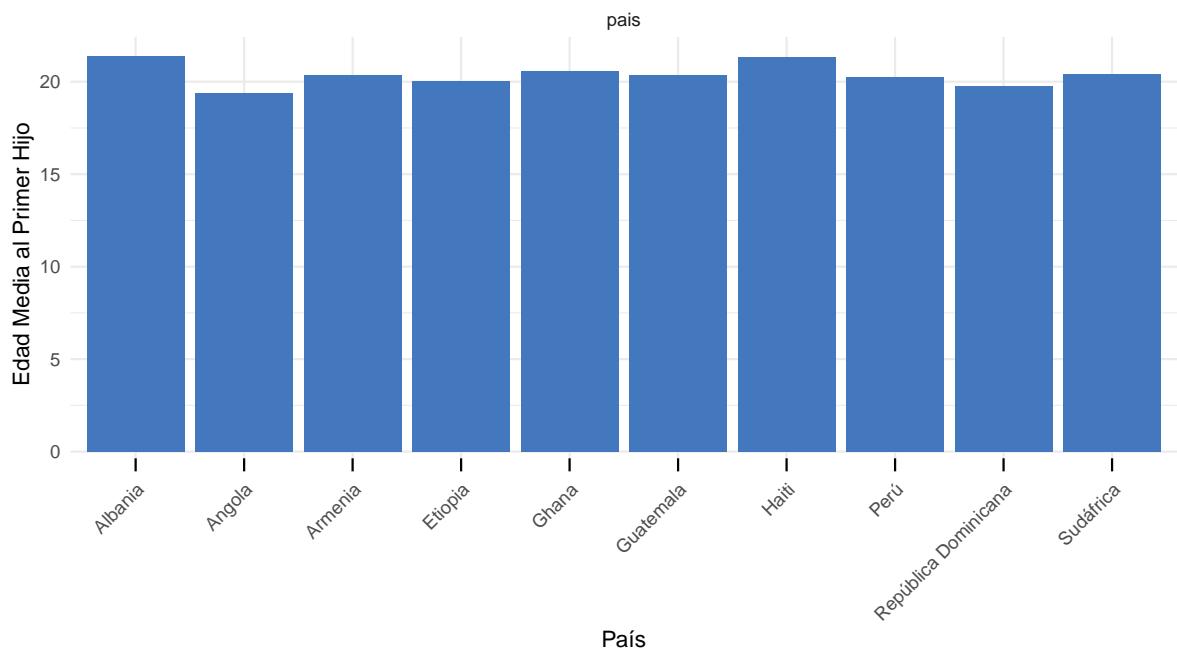


Figura 19: Gráfico PDP, efecto del país en la edad al primer hijo.

4 spec	binary	0.559
5 ppv	binary	0.687
6 npv	binary	0.699
7 mcc	binary	0.373
8 j_index	binary	0.360
9 bal_accuracy	binary	0.680
10 detection_prevalence	binary	0.638
11 precision	binary	0.687
12 recall	binary	0.801
13 f_meas	binary	0.739

Random Forest

Busqueda de Hiperparámetros

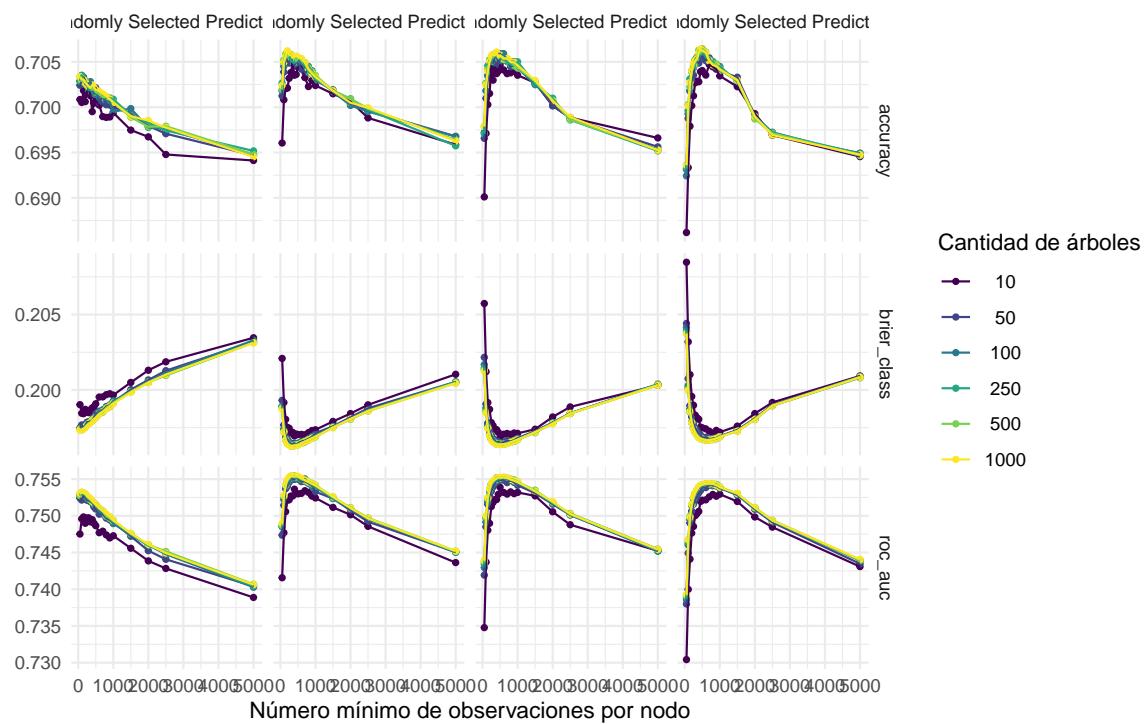
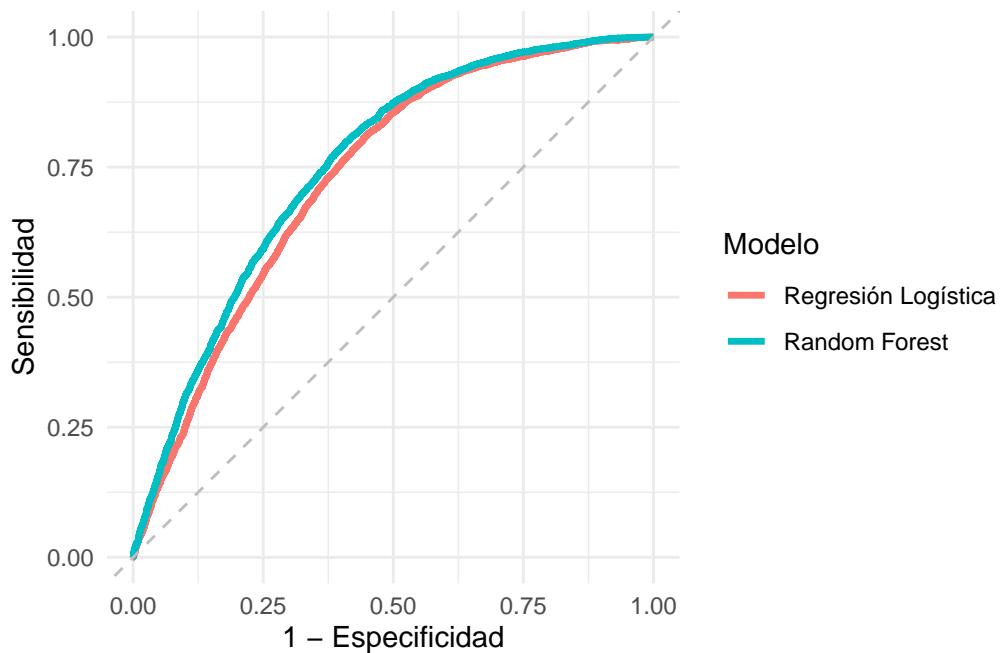


Figura 20: Gráfico de líneas, métrica RMSE y RSQ en función del número mínimo de observaciones por nodo, donde cada línea corresponde a la cantidad de árboles.

Evaluación del modelo

```
# A tibble: 13 x 3
  .metric      .estimator .estimate
  <chr>        <chr>       <dbl>
1 accuracy    binary     0.703
2 kap          binary     0.387
3 sens         binary     0.838
4 spec         binary     0.541
5 ppv          binary     0.688
6 npv          binary     0.734
7 mcc          binary     0.400
8 j_index      binary     0.379
9 bal_accuracy binary     0.689
10 detection_prevalence binary 0.666
11 precision   binary     0.688
12 recall      binary     0.838
13 f_meas      binary     0.755
```



Interpretación e importancia de las variables

