La utilización de aprendizaje automático en ciencias sociales y los resguardos metodológicos necesarios. Cuestionamientos a partir de la imputación de ingresos en la EPH.

Luis Nahuel Fernández

XV Jornadas de la Carrera de Sociología Facultad de Ciencias Sociales - Universidad de Buenos Aires

2023

1. Introducción

La corrección de la no respuesta parcial es una de las grandes problemáticas de las encuestas de hogares. Esto, debido a que es una tarea necesaria para garantizar la representatividad de la muestra y la calidad de las estimaciones.
A partir del gran desarrollo de las ciencias computacionales, en las últimas décadas comenzaron a estar disponibles los algoritmos de aprendizaje automático como herramientas para la imputación de la no respuesta parcial que están siendo estudiadas y puestas a prueba en el campo de la investigación social. Está demostrado en diversos estudios que en términos de error cuadrático medio, o error medio absoluto estos mecanismos tienen un rendimiento superior a las técnicas tradicionales.
En el presente trabajo se buscará comparar el rendimiento de hot deck aleatorio que es una técnica estadística "tradicional" utilizada desde hace décadas con este fin, árboles aleatorios como un ejemplo de ensambles más "novedosos" y regresión lineal múltiple robusta como un tercer ejemplo de control. Pero con ellos se buscará ir más allá de las métricas utilizadas en ciencia de datos y evaluar qué impacto tienen en la elaboración de estadísticas sociales. ¿Alcanza con el ECM y el MAE en cs. sociales?. El ejemplo desde el cual se abordará la cuestión es qué sucede con la distribución del ingreso y el coeficiente de Gini.

2. Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

Los antecedentes más relevantes que serán tenidos en cuenta en la realización de este trabajo, son publicaciones de autores de Sud América sobre la imputación de ingresos. El primero de ellos es el trabajo de Medina y Galván para la Unidad de Estadísticas Sociales de la División de Estadística y Proyecciones Económicas de la Comisión Económica para América Latina y el Caribe (CEPAL). Este es un trabajo que es utilizado como referencia en gran parte de los institutos de estadísticas de Latino América. Este trabajo da un marco general a la problemática de la no respuesta. Allí se plantea que " Está demostrado que la falta de respuesta se asocia, por ejemplo, al estatus económico de la familia, al área geográfica de residencia del hogar, al nivel de estudio de las personas, y en estas situaciones el patrón de datos omitidos no puede ni debe ser ignorado (MNAR)." (pág 20). Por otro lado afirma que "El hot-deck y las variantes que se han comentado se consideran mejores opciones que los procedimientos listwise deletion, pairwise deletion, y es superior los métodos de medias condicionadas y no condicionadas, ya que no introduce sesgos en el estimador y su error estándar. Si se desea preservar la distribución de probabilidad de las variables imputadas, conforme a la opinión de algunos autores se considera que el procedimiento hot-deck es más eficiente que el algoritmo la imputación múltiple y la regresión paramétrica (Durrent, 2005). "
El segundo antecedente importante es el trabajo de Hozsowski titulado "Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares". Este es un trabajo publicado por el Instituto Nacional de Es-

tadísticas y Censos donde se describe el problema de la no respuesta a ingresos entre el año 2004 a 2006 en dicha encuesta. El autor llega a la misma conclusión que en el trabajo de la CEPAL, el Hot-Deck tiene una serie de ventajas sobre la regresión que fue el método que la EPH utilizó previamente. La gran ventaja de este mecanismo es su transparencia y sencillez, atributos en los institutos públicos de estadística son activos muy cotizados. Los algoritmos "de caja negra", aún teniendo ventajas en muchos aspectos, tienen mucha resistencia en este ámbito ya que se prestan a la malinterpretación y requieren muchas explicaciones.
El tercero es el trabajo de G. Rosati. Esta investigación al ser más reciente y provenir de un ámbito distinto se permite plantear discusiones acerca de la posibilidad de utilizar ensambles y modelos de aprendizaje profundo para la imputación de ingresos. Es de destacar que en este paper "se asumió un proceso generador de datos perdidos MCAR o MAR". El autor sostiene que en su trabajo "Se mostró la mayor performance que los ensambles y el MLP tienen en comparación con la técnica habitualmente utilizada en algunas dependencias del Sistema Estadístico Nacional. En efecto, al cuantificar dos indicadores usuales para este tipo de problemas se observó que el RMSE de los modelos basados en Machine Learning oscilaba alrededor de los $3.800 y $4.000, mientras que el RMSE de Hot Deck superaba los $5.900, lo cual implica una mejora de alrededor del 33 %. Valores similares mostraba el indicador MAE". Con respecto a la tasa de no respuesta hace una aseveración que es importante destacar y tener en cuenta acerca de que se dio un crecimiento de la no respuesta a ingresos "del 8 % en 1995 al 24 % en 2010", a esto hay que añadir que si bien disminuyó, se ubicó en 16 % en el cuarto trimestre de 2021, y se encuentra en 17 % al cuarto trimestre de 2022.

2.2. Conceptos y técnicas de ciencia de datos utilizados

En este trabajo se utilizan dos técnicas: la regresión lineal múltiple y árboles aleatorios.
La fórmula de la regresión lineal múltiple es:

$$E(Y\|X1, X2, ...Xn) = \beta0 + \beta1X1 + \beta2X2 + ...\beta nXn$$

Donde los valores β son los coeficientes por los que se ajustan las variables con el fin de predecir un valor. Uno de los defectos de este algoritmo es lo que se denomina colinealidad. Según Szretter "La colinealidad ocurre cuando dos o más variables explicativas están altamente correlacionadas, a tal punto que, esencialmente, guardan la misma información acerca de la variabilidad observada de Y"(Szretter, página 192). Algunos de los problemas que trae aparejada la colinealidad son:
"1. Los coeficientes de regresión estimados se modifican sustancialmente cuando se agregan o se quitan variables del modelo.
2. Los errores estándares de los estimadores de los coeficientes aumentan espuriamente cuando se incluyen covariables muy correlacionadas en el modelo. Esto se denomina inflar la varianza estimada de los estimadores.

3. Los coeficientes pueden ser no significativos aún cuando exista una asociación verdadera entre la variable de respuesta y el conjunto de variables regresoras"(Szretter, pág 220)
Otro de los defectos que posee es que requiere la utilización de variables continuas, y en la base que se está utilizando solo la edad y la cantidad de horas trabajadas cumplen con esta condición. También debe cumplir con supuestos como el de normalidad de los errores, homoscedasticidad, etc. para tener un correcto funcionamiento. Debido a la presencia de valores atípicos se utilizó la función *lmrob* de la librería *Robust*. Esta función reemplaza el método de mínimos cuadrados en el cálculo de los coeficientes por un estimador de tipo MM.
Los árboles aleatorios son una forma de bootstrap agregating ya que Utilizan un conjunto de árboles de decisión, promediando la predicción que proporciona cada árbol para obtener la predicción final. Estos árboles de decisión se construyen a partir de un remuestreo tanto de las variables explicativas como de las unidades de análisis. Aventajan en general a los modelos de regresión cuando hay interacciones complejas y relaciones no lineales. Las variables auxiliares pueden ser continuas o cualitativas, su rendimiento es bueno con ambas y No se requieren supuestos sobre la relación entre la variable a imputar y las explicativas.
El hot-deck aleatorio implica que a partir de un vector de variables auxiliares Xi. Si Yi es un valor perdido, se selecciona aleatoriamente (en forma equiprobable o con algún peso), un elemento j tal que xi = xj. Se asigna a Yi el valor Yj. Es un método no determinista. Se utiliza el paquete *VIM*.
Otro concepto que toma relevancia en el trabajo es el de bias-variance tradeoff, " In fact, it is a very simple idea. A model can be bad for two different reasons. Either it is not accurate and doesn't match the data well, or it is not very precise and there is a lot of variation in the results. The first of these is known as the bias, while the second is the statistical variance. More complex classifiers will tend to improve the bias, but the cost of this is higher variance, while making the model more specific by reducing the variance will increase the bias" (Marsland, página 35).

2.3. Métricas para la evaluación de la imputación

Para la evaluación de los modelos se utilizan distintas métricas. Para el sesgo, el error cuadrático medio, el coeficiente de determinación y el error medio absoluto el paquete *Metrics* de R. Con el objetivo de comparar varianzas se utilizó la función *var* de Rbase se restó la varianza relevada con la varianza predicha. También se utiliza el coeficiente de Gini, el cual es una medida utilizada para cuantificar la desigualdad en la distribución de ingresos o riqueza dentro de una población. Se utiliza comúnmente en economía y sociología para evaluar la disparidad económica entre los individuos o grupos dentro de una sociedad. Se calcula mediante la representación gráfica de la curva de Lorenz, que muestra la acumulación de la riqueza o ingresos en relación con la distribución acumulada de la población. En esta curva, el eje horizontal representa el porcentaje acumulado de la población, mientras que el eje vertical representa el porcentaje acumulado de ingresos o riqueza. El coeficiente de Gini se obtiene comparando el área entre la curva de Lorenz

y la línea de igualdad perfecta (donde todos los individuos tienen la misma cantidad de ingresos o riqueza) con el área total debajo de la línea de igualdad perfecta. Cuanto más cerca esté el coeficiente de Gini de 1, mayor será la desigualdad en la distribución de ingresos o riqueza, mientras que un valor de 0 indica igualdad perfecta. Para tomar este coeficiente se utiliza el paquete *reldist* de R.

El Mean Absolute Error (MAE), o **Error Medio Absoluto** en castellano, se define como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

donde y_i son los valores observados y $\hat{y}_i$ son los valores predichos para $i = 1, 2, \ldots, n$.

El Mean Squared Error (MSE), o **Error Cuadrático Medio** en castellano, se define como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde y_i son los valores observados y $\hat{y}_i$ son los valores predichos para $i = 1, 2, \ldots, n$.

o expresado de otra manera

$$MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) + \Big[\text{Sesgo}(\hat{\theta})\Big]^2 + \text{Error irreducible}$$

donde $\hat{\theta}$ es el estimador y el Error irreducible es el error que no puede ser reducido por el modelo y proviene del ruido inherente en los datos o factores no controlables..

La fórmula de la **varianza** se representa como:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

donde σ^2 es la varianza, x_i son los valores de la muestra y μ es la media de la muestra para $i = 1, 2, \ldots, n$.

El **sesgo** de un estimador se define como:

$$\text{sesgo} = E(y_i - \hat{y}_i)$$

donde y_i es el valor declarado, e $\hat{y}_i$ es el predicho.

La fórmula del **coeficiente de Gini** se representa como:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu}$$

donde G es el coeficiente de Gini, x_i son los valores de la muestra, n es el tamaño de la muestra y μ es la media de la muestra.

La fórmula del **coeficiente de determinación (R cuadrado)** se calcula de la siguiente manera:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

donde y_i es el valor observado, $\hat{y}_i$ es el valor estimado por el modelo y $\bar{y}$ es el promedio de los valores observados.

3. Metodología

3.1. Presentación y descripción de los datos utilizados

El conjunto de datos a utilizar se encuentra disponible en la página web del Instituto Nacional de Estadísticas y Censos de la Argentina. Este instituto realiza un relevamiento trimestral de viviendas en 31 aglomerados de la República Argentina, las personas respondentes al cuestionario individual constituyen los registros del conjunto de datos.
El dataset perteneciente al 4to trimestre de 2022 se compone de 48.545 registros y 177 variables. El diseño del registro para las bases se encuentra también en la página web del instituto[1].
Ya que el dataset seleccionado posee 177 variables es recomendable realizar una selección de las mismas debido a las características de los algoritmos que se van a utilizar, principalmente la regresión lineal múltiple. Otra ventaja de la selección de variables es que acelera los tiempos computacionales de procesamiento para los árboles aleatorios.
Por último, la selección de variables sirve a quitar variables que no poseen relación alguna con el ingreso como las variables del bloque para los desocupados o que son constantes como el año y el trimestre.
Las variables seleccionadas y construidas se encuentran en el anexo.

3.2. Análisis exploratorio de datos

Como se observa en la figura 1, el ingreso de la ocupación principal posee una distribución log-normal. Por ello mismo, para la aplicación del modelo de regresión lineal se utilizará el logaritmo de la P21 con el objetivo de cumplir el supuesto de distribución normal. El mínimo de ingresos declarado en toda la base es de $200, la mediana de $70.000, el tercer cuartil finaliza en $110.000 y el máximo declarado es de $4.000.000. En el caso de los asalariados el mínimo es de $1500 y el máximo de $1.500.000 con una media de $90.600. En los patrones $500, $137.061 y $4.000.000. Si se ponderan las distribuciones la media de ingresos para patrones, utilizando el PONDIIO, es de $143.000 y para asalariados de $100.000. La muestra cuenta con 763 patrones, 4725 cuentapropistas y 16485 asalariados, de los cuales 535, 3495 y 13135 respectivamente respondieron afirmativamente sobre sus ingresos.

En segundo lugar se observa que la tasa de no respuesta a ingresos de la ocupación principal, a nivel muestral, es de 17,3 %, siendo del 28,8 % entre los patrones y del 15,5 % entre los asalariados.
Por otro lado, si se evalúa correlación de variables continúas mediante Spearman, se observa que entre los asalariados, la correlación de P21 con la edad es del 0,219 y con la PP3E_TOT es de 0,398. Mientras que entre los patrones la correlación con la edad es de 0,197 y con la cantidad de horas trabajadas la semana anterior es de 0,060.
En la figura 2 se observa que la distribución de ingresos según sexo por categoría ocupacional posee diferencias leves al interior de asalariados y patrones, siendo la mediana

[1] https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos

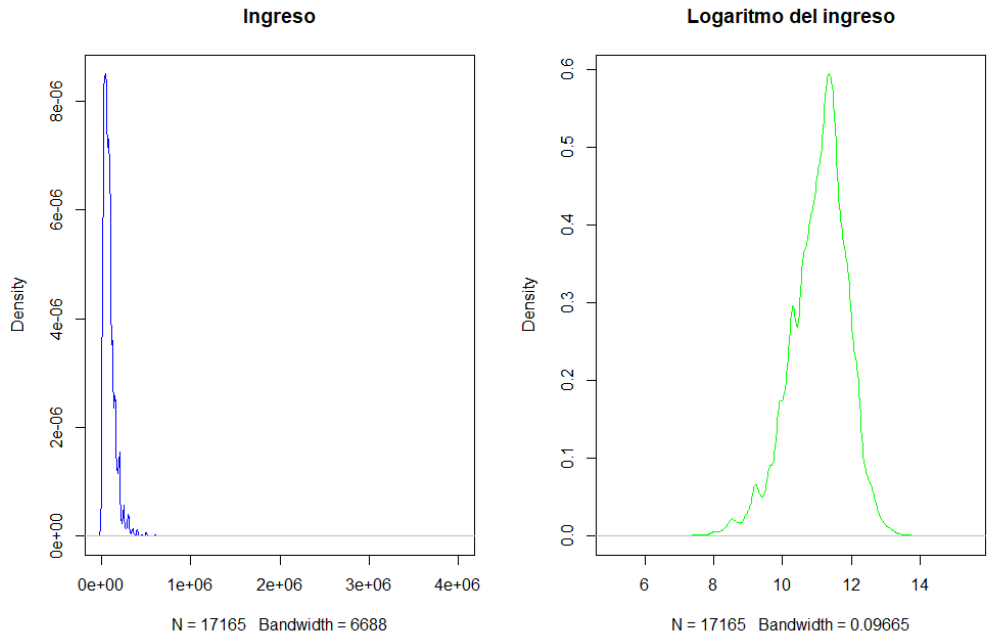

Figura 1: **Distribución del ingreso de la ocupación principal**

levemente más alta en varones que en mujeres en ambas categorías ocupacionales. En cambio en la figura 3 se observa una diferencia mucho más importante en todas las categorías entre aquellos que se encuentran en la informalidad con quienes se encuentran en el sector formal de la economía. De izquierda a derecha se observa la distribución muestral para asalariados formales del sector formal, formales que trabajan en hogares, formales de unidades económicas informales, informales del sector formal, informales de hogares, informales del sector informal, otros asalariados, otros patrones, patrones formales y patrones informales.

Los gráficos de cajas se realizaron excluyendo outliers con el objetivo de resaltar lo que sucede en el rango intercuartílico, sin embargo es necesario destacar la presencia de valores extremos que en el estudio de ingresos son muy importantes. Si bien la mayor concentración de ingresos se muestra en el límite indicado por los gráficos, hay 135 personas que declararon ingresos superiores a $350.000. En el extremo más alto se encuentran cinco casos de declaraciones de ingresos mayores al millón de pesos, el más alto de ellos un patrón del sector hotelero, pero los otros cuatro son asalariados.
Por último, es interesante observar las diferencias en la tasa de no respuesta según las distintas variables. Como se observa en la tabla 1, al desagregar según nivel educativo la tasa de no respuesta casi se duplica de primario incompleto a universitario completo. La media muestral de ingresos por ocupación principal entre ambas categorías es más del doble. Para quienes tienen el nivel universitario completo la media es de $95.600, mientras que para quienes no terminaron el primario es de $42.700.

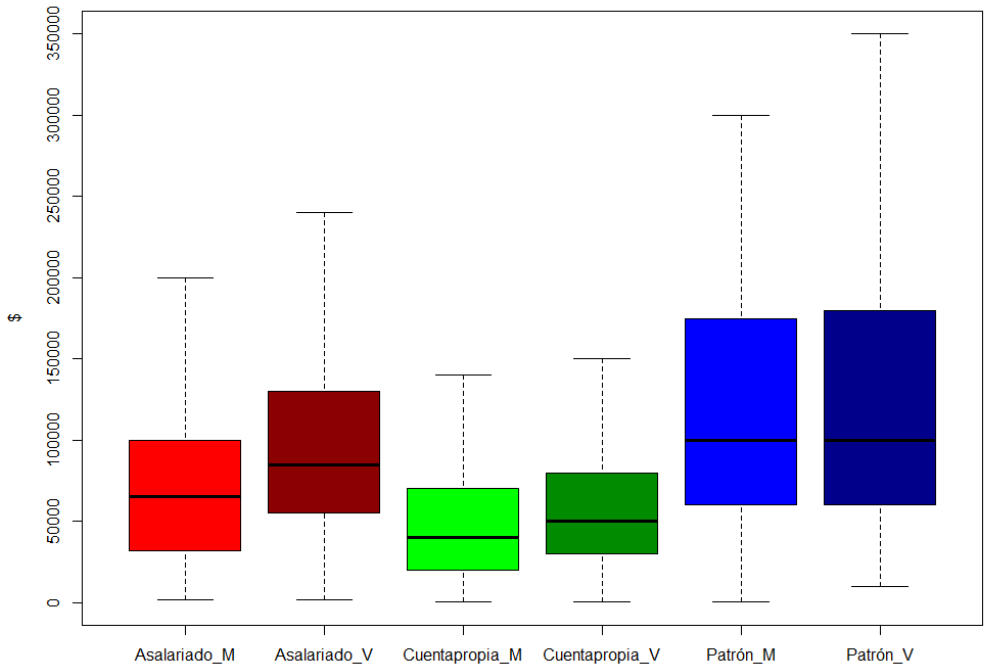

Figura 2: **Ingreso de la ocupación principal según genero y categoría ocupacional**

Cuadro 1: **No respuesta a ingreso de la ocupación principal**

Categoría	Personas respondentes	Personas no respondentes a ingresos	Tasa de no respuesta a ingresos
Primario incompleto	679	81	11,93 %
Primario completo	2486	352	14,16 %
Secundario incompleto	3633	489	13,46 %
Secundario completo	6399	1061	16,58 %
Universitario incompleto	2957	457	15,45 %
Universitario completo	4990	1131	22,67 %
NS/NR	58	12	20,69 %

Si se toma como referencia el *"rol"* se observa un fenómeno semejante ya que la tasa de no respuesta a ingresos de la ocupación principal es del doble en los profesionales con respecto a los obreros no calificados (24,8 % y 12,0 %) siendo de $113.000 la media de ingresos muestrales de los primeros y de $42.000 la de los segundos. Entre varones la tasa es de 18 % y en mujeres del 15,6 %, teniendo también un mayor ingreso los varones. Frente a los datos faltantes, de detectaron 2 casos de datos perdidos en *Jerarquía* y 23 en *PP3E_TOT*. El criterio ante estos casos fue imputar la media en *PP3E_TOT* y poner 9,

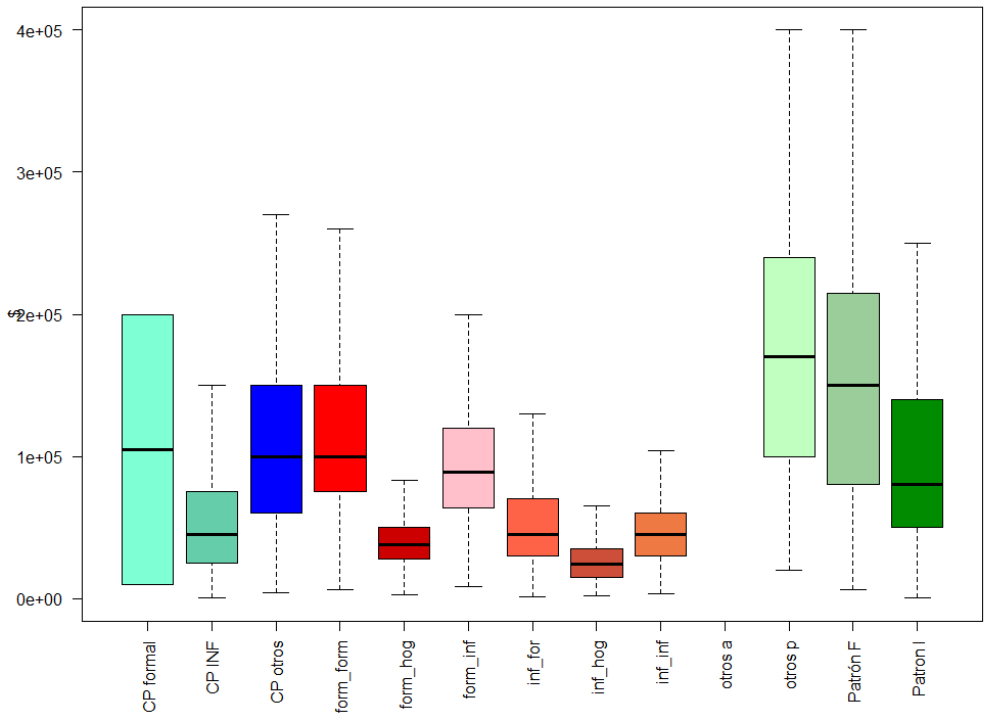

Figura 3: **Ingreso de la ocupación principal según categoría ocupacional, formalidad del empleado y de la unidad económica**

que es el valor correspondiente de no sabe/no responde en la *Jerarquía*.

3.3. Técnicas y algoritmos a utilizar

La metodología del presente trabajo consiste en generar tres metodologías alternativas con el objetivo de imputar. Como ya se dijo previamente las técnicas por utilizar son una regresión lineal múltiple robusta, hot-deck y árboles aleatorios, se busca que tengan las mismas variables explicativas, por lo que para la regresión se utilizaron las seis variables más importantes según la importancia de variables de los árboles aleatorios. Con estos modelos se pretende evaluar métricas de sesgo, varianza y ECM, y por otro lado contrastar cómo la utilización de ellos puede impactar de distinta manera en las estimaciones utilizando como ejemplo el coeficiente de Gini.
El ejercicio se realiza sobre el conjunto de la base de ocupados incluyendo asalariados, cuentapropistas y patrones.
La experimentación se llevaría adelante evaluando los modelos en ambas bases filtrada por los casos respondentes. En un primer lugar se genera una simulación de la no respuesta calculando la probabilidad de no respuesta de cada registro mediante una regresión logística teniendo en cuenta la edad, el nivel educativo, la región y el sexo, para luego generar un vector de números aleatorios con una distribución uniforme. Si el valor aleatorio supera a la probabilidad de respuesta en ese caso se utilizarían esos registros

para comparar la imputación de esos casos con respecto a los valores declarados y así evaluar los resultados de la predicción. Este experimento se repetiría 30 veces (en un principio se pensaba realizar más repeticiones pero debido al gran tiempo requerido en cada iteración y los límites temporales de las presentaciones se concluyó que 30 es un número suficiente). Este procedimiento es una forma de evaluar el modelo sin utilizar validación cruzada, sino teniendo en cuenta la probabilidad de no respuesta de cada respondiente, asumiendo que la no respuesta no es absolutamente aleatoria, sino que existe una mayor probabilidad de no respuesta en algunos casos que en otros.
Para los árboles aleatorios se requirió el paquete *randomForest* de R. Para cualquier consulta sobre el código y los paquetes utilizados consultar el repositorio detallado en el anexo. Como hiper-parámetros del modelo de AA se utilizaron los definidos en el trabajo de Rosati, es decir mtry=23, min.node.size=10. Este ejercicio se realizó sobre los ingresos declarados, sin ninguna transformación, a diferencia de la regresión lineal que se hizo utilizando el logaritmo del ingreso.
Para Hot-deck aleatorio se utilizó la función *hotdeck* del paquete *VIM*, de tal manera que se le pasen por parámetro todas las variables explicativas ordenadas tomando en cuenta la importancia de variables de los árboles aleatorios, con estas variables explicativas se imputan todos los casos posibles. Para los casos que quedan sin imputar se genera una iteración a partir de la cual la variable menos importante se quite del vector y se imputa nuevamente, esto se repite hasta que se logra imputar todo el dataset.
Para la regresión lineal múltiple se utilizó la función *lm()* de Rbase en primera instancia, sin embargo debido a la existencia de valores extremos con alto leverage sobre el modelo se debió reemplazar por un modelo robusto mediante la librería *robustbase*. El modelo que se utilizó fue el siguiente:

$$\begin{aligned}\log(P21) =& \beta_0 + \beta_1 \cdot AGLOMERADO + \beta_2 \cdot INF + \\ & \beta_3 \cdot CH06 + \beta_4 \cdot car1 + \beta_5 \cdot PP3E\_TOT + \\ & \beta_6 \cdot PP04C + \frac{1}{10} \cdot Z\end{aligned}$$

Buscando tener en cuenta el σ de la regresión, se generó la variable aleatoria Z siguiendo una distribución normal con media 0 y desviación estándar igual a 1/10 del valor de σ.

En cada una de las imputaciones para las simulaciones de no respuesta se calculó el ECM y el MAE ya que son las métricas que se suele utilizar en ciencia de datos. Luego, con el objetivo de tener una mayor precisión de lo que estaba sucediendo se calculó por separado el sesgo y la varianza. Estas métricas se calcularon sobre los conjuntos de testeo creados a partir de la regresión logística. El sesgo no suele utilizarse para el análisis de conjuntos de testeo con el fin de mejorar el rendimiento de modelos ya que suele conducir a un sobreajuste, pero en este caso el objetivo de su utilización es meramente descriptivo y no se pretende optimizar ningún modelo.
En un ejercicio aparte se hizo el cálculo del coeficiente de Gini, tomando todos los ocupados que respondieron efectivamente y siendo los registros imputados cerca del 20 % de los mismos en cada iteración.

4. Resultados

4.1. Error cuadrático medio y error medio absoluto

En un primer acercamiento se observa que el ECM y el EMA es mucho más cercano a 0 en los modelos de árboles aleatorios. Esto daría la impresión de un rendimiento mucho más satisfactorio.
La figura 4 demuestra que el EMA es en todos los ejercicios menos de la mitad en AA que en la regresión lineal, y también tiene un mejor rendimiento que el HD. La utilización de AA permite que esta métrica se ubique siempre por debajo de los $40.000, mientras que en la regresión siempre está por encima de $70.000, y en el HD siempre rondando esta cifra.
Al utilizar ECM la diferencia relativa que se observa en el EMA disminuye, seguramente debido a la existencia de valores extremos mal imputados en AA y se observa una mayor cercanía con HD.

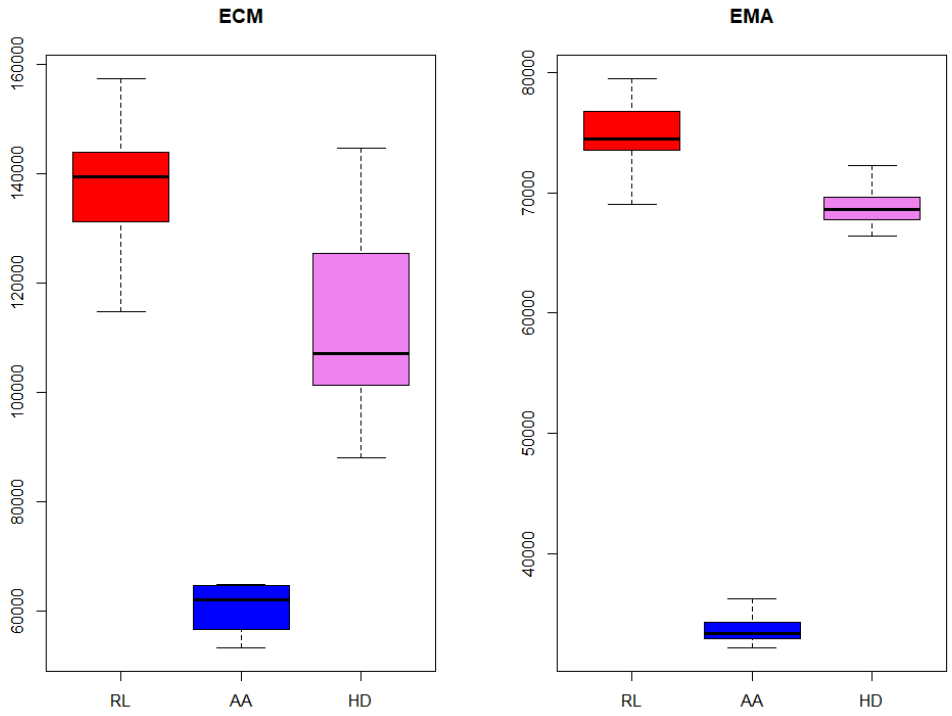

Figura 4: **distribución de ecm y ema en 30 ejercicios según método de imputación**

4.2. Sesgo y varianza

En la figura 5 se observa en distintos gráficos de cajas los resultados obtenidos en los 30 experimentos en relación al sesgo y la varianza. En cuanto al sesgo una vez más detectamos que AA tiene un rendimiento mucho más eficiente, ya que se encuentra mucho

más cerca del cero que la regresión lineal múltiple robusta y también que el Hot-Deck. Pero observando la varianza se encuentran los resultados más llamativos. Los gráficos muestran que la varianza de la variable predicha siempre disminuye con respecto a la variable declarada al utilizarse AA, esto es interesante ya que termina de confirmar cuál es el motivo por el cual este método logra una disminución del ECM y el MAE. Con respecto a la regresión tiende a suceder lo contrario, la varianza de la variable predicha tiende a ser más alta, por eso los gráficos de cajas se ubican por debajo de la linea verde punteada que marca el cero. El HD tiene el mejor rendimiento en este caso y se encuentra muy cerca del 0.

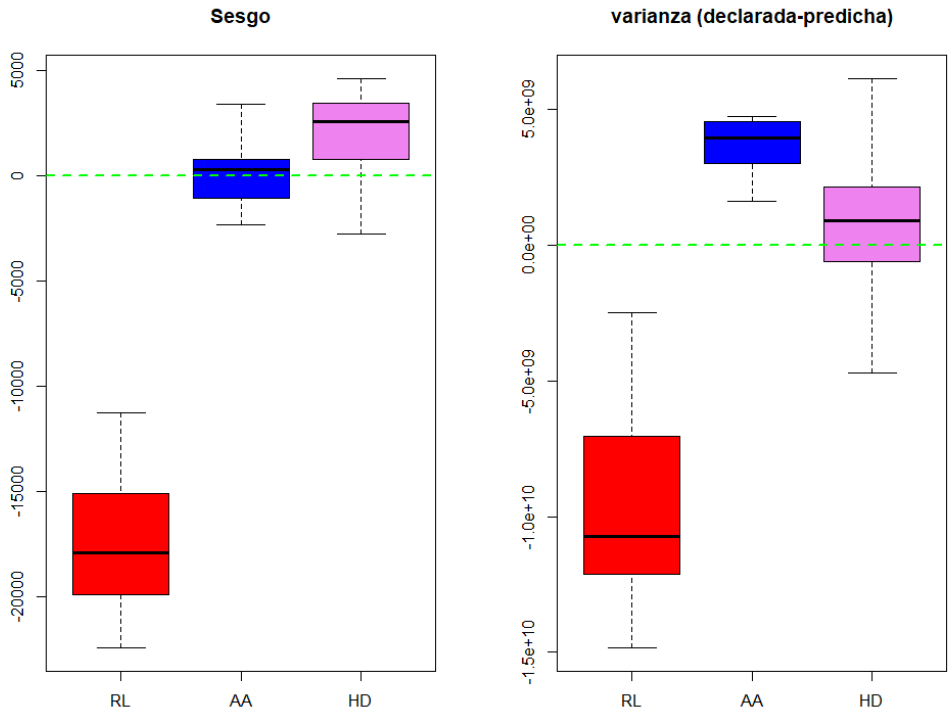

Figura 5: **distribución de sesgo y varianza en 30 ejercicios según método de imputación**

4.3. El coeficiente de gini

En la figura 6 se observa que la elección del método de imputación tiene un impacto directo en el coeficiente de Gini. En esta figura, la linea verde punteada representa el coeficiente de Gini declarado para el ingreso de la ocupación principal de los ocupados en el trimestre 4 del año 2022. Las cajas representan los resultados del cálculo de este coeficiente para la simulación de la imputación con árboles aleatorios en rojo, la regresión lineal en azul y el hot-deck en violeta. La disminución de la varianza utilizando árboles

trae aparejada una reducción de este coeficiente que mide la desigualdad en los ingresos, mientras que el aumento de la varianza utilizando la regresión genera un aumento del coeficiente de Gini en todos los casos. Hot-deck logra un muy buen rendimiento y en todos los ejercicios es el que más se acerca a la estimación según lo declarado.

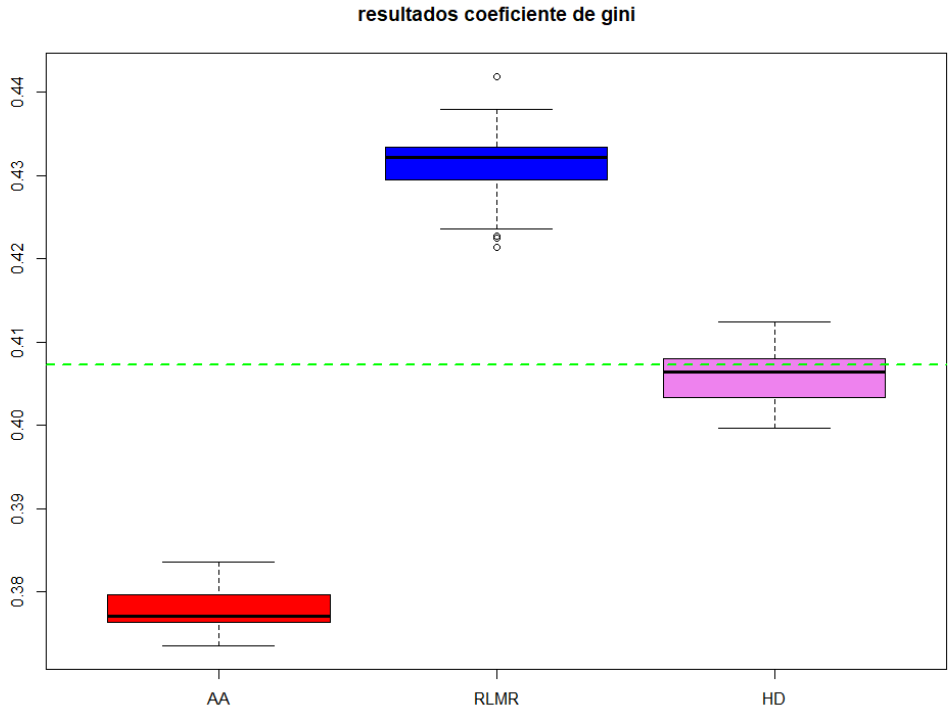

Figura 6: **distribución del coeficiente de Gini en 30 ejercicios según método de imputación**

5. Conclusión y discusiones

En un primer acercamiento a la cuestión, pareciera que la utilización de modelos de árboles aleatorios tiene importantes ventajas tanto si se mide con el ECM o con el MAE. Sin embargo cuando se observa la influencia que tienen estos modelos en la distribución de la variable y cómo ello afecta al coeficiente de Gini la cuestión da un giro. Este algoritmo tiene las mejores métricas en términos de lo que regularmente se observa en ciencia de datos, pero lo logra a partir de los cambios que genera en la distribución de la variable, gracias a una disminución de la varianza. Y esto, que puede ser una ventaja en muchos campos en los que el fin es la predicción de casos individuales, en la elaboración de estadísticas sociales se transforma en un defecto. En los experimentos realizados se observa que los árboles aleatorios generan una distribución que tiende a ser leptocúrtica como se muestra en la figura 7. Esto deriva entre otras cosas, en una disminución de la desigualdad en el ingreso, que es una cuestión fundamental en el estudio de las ciencias sociales. Por fuera del ejemplo actual, es de esperar también que esta disminución de la varianza tenga influencia en las mediciones de pobreza o ingresos por decil. En palabras de Marsland sobre los árboles aleatorios: El remuestreo que se efectúa en los árboles aleatorios, así como introduce la aleatoriedad en la selección de variables predictivas para el entrenamiento de cada árbol, también hace más veloz su procesamiento, ya que hay menos variables en cada etapa. Como es obvio, esto introduce un nuevo parámetro (cuántas variables considerar), pero los árboles aleatorios no son muy sensibles a ello; en la práctica, una submuestra del tamaño de la raíz cuadrada de la cantidad de las variables predictivas es lo común. El efecto de estas dos formas de aleatoriedad es reducir la varianza sin afectar el sesgo. (Marsland, pág 275). Por esto los AA y todos aquellos algoritmos que en el trade-off sesgo varianza prioricen la reducción de la varianza con el objetivo de mejorar el ECM no deberían ser tenidos en cuenta como medios de imputación.

Por otro lado el modelo de la regresión líneal múltiple robusta, con todas sus limitaciones y deficiencias como un ECM mucho más alto que los árboles aleatorios, la colinealidad, que en este caso no se cumple el supuesto de normalidad (como se observa en la figura 8), etc. en la gran mayoría de los experimentos se encuentra más cerca del coeficiente de Gini de la variable relevada. Esto no significa que sea un mejor mecanismo de resolución de la problemática planteada, pero deja al descubierto que un modelo con pésimos resultados en la medición del ECM y el EMA puede a fin de cuentas igualar o incluso superar a otros con resultados óptimos en estas métricas para determinados fines.

El hot deck, aún con un peor rendimiento en términos de EMA y ECM, logra preservar la distribución y con ello la varianza, asumiendo el costo de un mayor sesgo que los arboles. Sin embargo el impacto de este sesgo no sería tan importante a la hora de medir desigualdad o pobreza. Esto hace que sea el más recomendable de los tres.

Las diversas metodologías que utilizan los algoritmos de aprendizaje supervisado o técnicas estadísticas pueden ser ventajosas dependiendo de qué es lo que se busca realizar con ellos. Sin embargo se debe prestar especial atención a que no se generen artificialmente alteraciones que generen impacto en las estimaciones que se pretende realizar. Tal como plantea Hozsowski el hot-deck aleatorio sigue siendo método transparente y sencillo de aplicar que da garantías en cuanto un rendimiento aceptable y que no altera

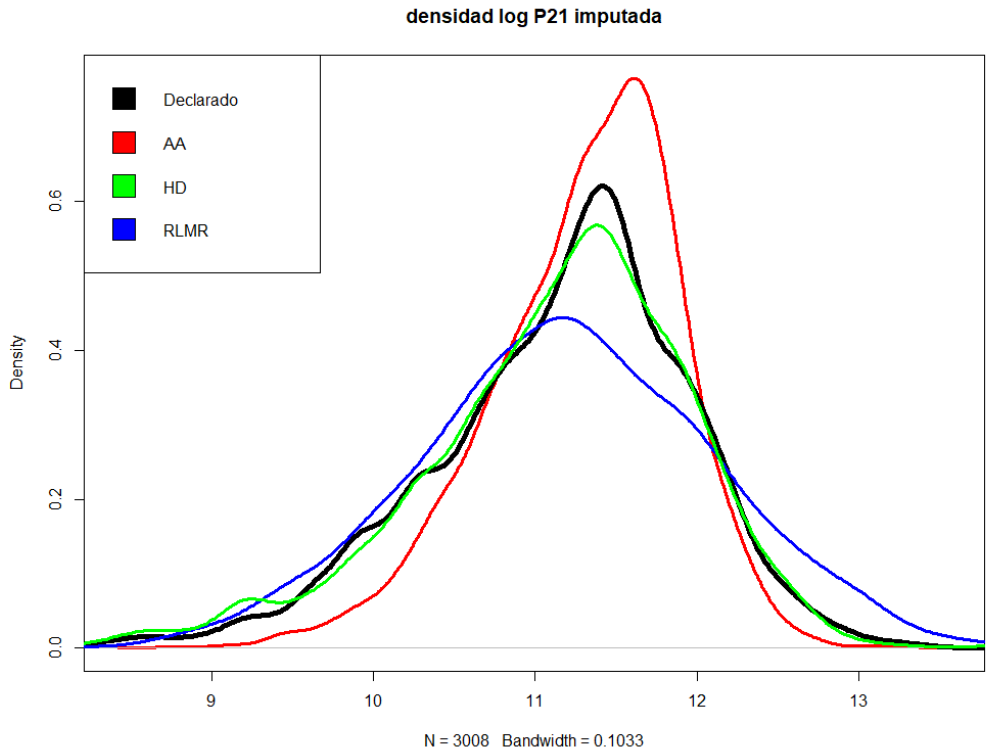

Figura 7: **densidad del logaritmo del ingreso para la ocupación principal**

la distribución de las variables.
En resumen los resultados demuestran que el ECM y el EMA son insuficientes a la hora de evaluar rendimientos cuando lo que se pretende es imputar ingresos. En el tradeoff sesgo-varianza, la varianza propia en la distribución del ingreso es un fenómeno que se pretende estudiar y no debe ser alterada artificialmente. Por otro lado estas métricas toman en cuenta tanto el sesgo como la varianza y no permiten estudiar por separado qué es lo que sucede con cada una de ellas, cuestión que debería ser tenida en cuenta. La utilización de las nuevas herramientas de aprendizaje automático a primera vista son atractivas y demuestran ventajas, pero es necesario estudiar a fondo qué es lo que sucede cuando se pretende utilizarlas en el campo de las ciencias sociales antes de sacar conclusiones definitivas.

6. Anexo

Repositorio con los ejercicios:
https://github.com/fernandezluisn/ponencia2023

Diseño del registro:
https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_4T2022.pdf

Cuadro 2: Variables construidas

Nombre	Descripción
Rama_eph	Se construye en base a la *PP04B_COD*, utilizando la tabla de correspondencias entre CAES Mercosur 1.0 y CAES.
INF	Es una variable categórica que se construye en base a la categoría ocupacional, y si la ocupación y la unidad económica son informales. Para los asalariados, la formalidad se construye utilizando la *PP07H*. Para la unidad económica se elaboro una variable proxy que se construye según la *PP04C* y la calificación del puesto. En el cuestionario actual de la EPH no existe una variable que sirva a distinguir unidades económicas formales de informales con certeza.
Rol	Es una desagregación que se aplica a los asalariados teniendo en cuenta jerarquía y calificación
Región	Es un agrupamiento de los aglomerados según región geográfica.
Jerarquía	Se construye en base al clasificador nacional de ocupaciones con la variable *PP04D_COD* y hace referencia al grado de responsabilidad del ocupado en su trabajo, es categórica con 4 respuestas posibles.
Calificación	Se construye en base al clasificador nacional de ocupaciones con la variable *PP04D_COD* y se refiere a si el puesto es profesional, técnico, operativo o no calificado. Categórica con 4 respuestas posibles.
Car1	Se construye en base al clasificador nacional de ocupaciones con la variable *PP04D_COD* y contiene 10 categorías que actúan como los grandes grupos ocupacionales.
Car2	Se construye en base al clasificador nacional de ocupaciones con la variable *PP04D_COD* y contiene las 51 categorías de ocupaciones.
TECNO	Se construye en base al clasificador nacional de ocupaciones con la variable *PP04D_COD* y detalla si se utilizan maquinarias o sistemas informatizados o electromecánicos. Categórica con tres respuestas posibles.

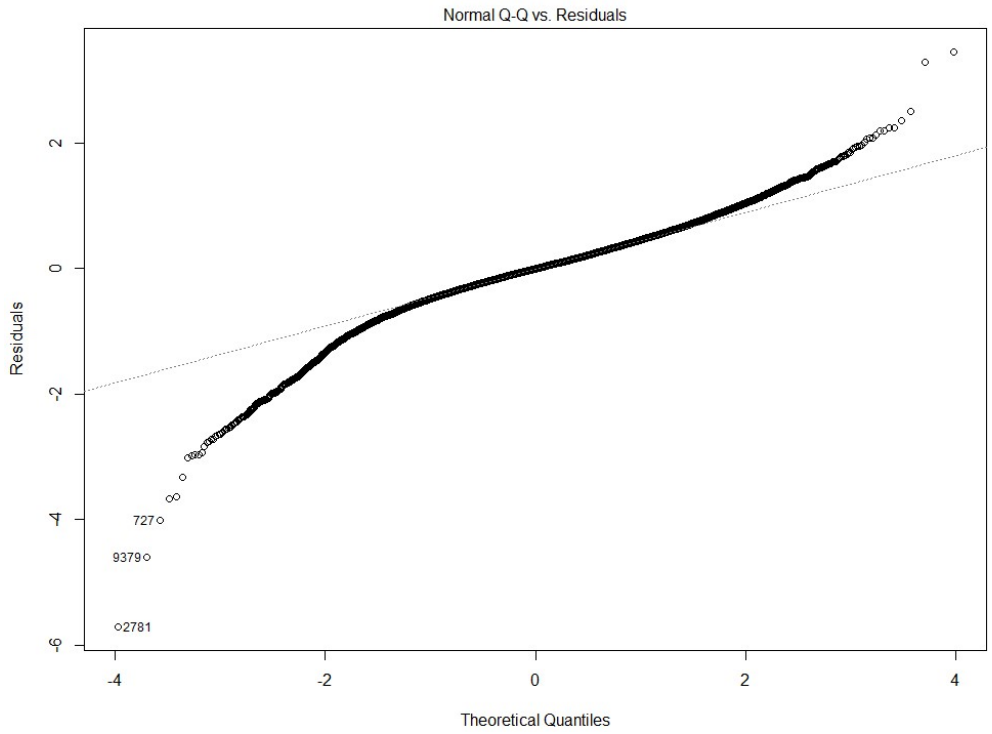

Figura 8: **Gráfico cuantil-cuantil de una de las iteraciones de la RLMR**

7. Bibliografía

Breiman, L. (2001). Random Forest. Machine Learning, 42, 5-32. Recuperado a partir de
https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

Instituto Nacional de Estadística y Censos (INDEC), (2018). Clasificador Nacional de Ocupaciones. Recuperado a partir de
https://www.indec.gob.ar/ftp/cuadros/menusuperior/clasificadores/definiciones_conceptuales_cno.pdf

Instituto Nacional de Estadística y Censos (INDEC), (sin fecha). Metodología nº 15. Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares. Recuperado a partir de
http://www.estadisticasantafe.gob.ar/wp-content/uploads/sites/24/2019/01/EPH-Encuesta-Permanente-de-Hogares-metodologia-N%C2%BA-15.pdf

Hoszowski, A., Messere, M., y Tombolini, L. (2004). Tratamiento de la no respuesta a las variables de ingreso en la Encuesta Permanente de Hogares de Argentina. Trabajo presentado en el XIV Taller Programa para el Mejoramiento de las Encuestas y la Medición de las Condiciones de Vida en América Latina y el Caribe (MECOVI), Buenos Aires.

Medina, F. and Galván, M., 2007. Imputación de datos: teoría y práctica. [online] https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf

Marsland S., 2015. MACHINE LEARNING An Algorithmic Perspective. Taylor & Francis Group. Boca Ratón.

Rosati, G. (2021). Métodos de Machine Learning como alternativa para la imputación de datos perdidos. Estudios Del Trabajo. Revista De La Asociación Argentina De Especialistas En Estudios Del Trabajo (ASET), (61). Recuperado a partir de https://ojs.aset.org.ar/revista/article/

Szretter Noste, María Eugenia (2017). Apunte de Regresión Lineal. Recuperado apartir de https://mate.dm.uba.ar/ meszre/apunte_regresion_lineal_szretter.pdf

Librerías importantes de R:

https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1/topics/randomForest

https://cran.r-project.org/web/packages/reldist/index.html

https://cran.r-project.org/web/packages/robustbase/robustbase.pdf

https://cran.r-project.org/web/packages/VIM/index.html