

Introducción a Machine Learning

Una Aplicación de Miss Forest a la EPH

Miriam Malament

Economía Laboral, UCEMA

Introducción

- Los datos faltantes representan un problema muy habitual al trabajar con bases de datos.
- En este trabajo, busco analizar y explicar el uso de *Miss Forest* como metodología para la imputación de ingresos faltantes en la Encuesta Permanente de Hogares.
- Si se asume, como indica la literatura, que el nivel educativo es determinante del ingreso, entonces puede resultar una metodología válida para la imputación de datos faltantes.
- Vamos a encontrar que puede ser una metodología apropiada para determinadas circunstancias.

¿Qué es la EPH?

→ La Encuesta Permanente de Hogares (EPH) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las características sociodemográficas y socioeconómicas de la población

→ A partir del año 2003, se mejoró el **diseño** de la muestra (dos etapas de selección), se cambió la **periodicidad** de la encuesta (a trimestral), se amplió la **cobertura** (a 31 aglomerados), se generó un **sistema de rotación** (no todos los trimestres se encuesta a los mismos hogares).

El Problema: Datos Faltantes

Poder Estadístico

→ Alrededor de 60,000 personas contestan la encuesta cada trimestre de las cuales tan solo 24,000 indican cuál es su ingreso laboral. Si filtramos por alguna característica adicional como educación o informalidad, la muestra se reduce a pocos miles.

Inflación

→ Trabajar con varias encuestas en simultáneo nos obliga a tener que hacer una corrección que no es fácil ni representativa.

Mi idea

→ Encontrar una manera de conservar el poder estadístico imputando los ingresos que faltan a partir de la educación.

Distribución del Ingreso y Educación

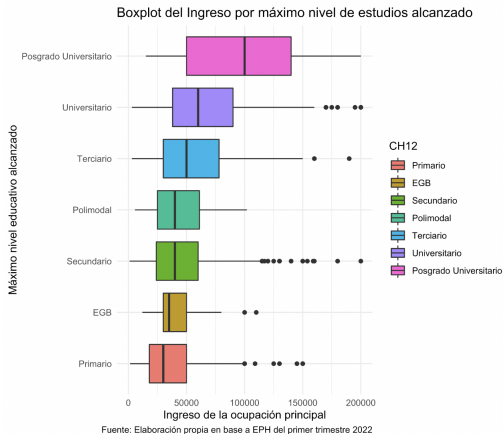


Figure 1: Ingresos por Nivel Educativo

El problema: ¿qué hacemos con los ingresos faltantes?

¿Los datos que faltan son *at random*? La respuesta, a priori sería que no, pero a los efectos del trabajo sí.

Alternativas:

→ Computar la media: claramente no sirve a nuestro análisis. Nos interesa justamente conservar la *heterogeneidad* de la muestra.

→ Computarle la media a cada característica (media por género, nivel educativo, edad, condición, etc).

Machine Learning:

→ A través de herramientas como Árboles de Decisión y Miss Forest (un tipo de Random Forest), podemos conseguir estas "medias específicas" pero sin supuestos sobre la forma funcional.

La Solución: Machine Learning

Variables a Utilizar

Para el análisis utilicé las variables: P21 (ingreso: lo que quiero determinar), CH12 (nivel educativo), CH06 (edad), CH04 (género) y REGION. Es relevante también tener en cuenta ESTADO (si la persona está ocupada o desocupada).







	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
base.P21	396	65	55444.9	50219.6	200.0	45000.0	2000000.0	
base.CH06	104	1	36.3	22.3	1.0	34.0	105.0	
base.CH12	11	5	4.1	2.2	1.0	4.0	99.0	
base.CH04	2	0	1.5	0.5	1.0	2.0	2.0	
base.REGION	6	0	36.5	14.1	1.0	42.0	44.0	
base.ESTADO	5	0	2.2	1.1	1.0	3.0	4.0	

Figure 2: Estadísticas Descriptivas de Variables a Utilizar

Árboles de Decisión: ¿qué está determinando el ingreso?

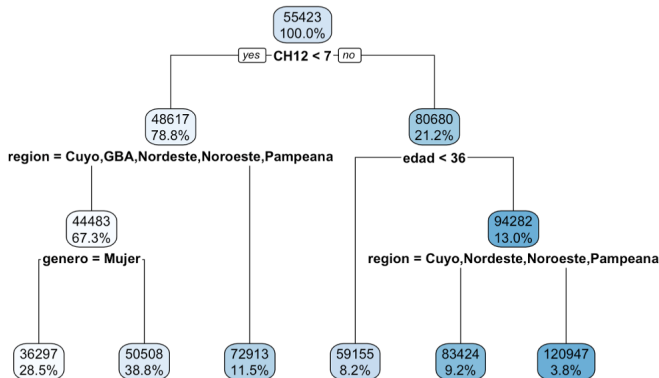


Figure 3: Árbol de Decisión a partir de región, edad, nivel educativo y género para los ingresos del primer trimestre 2022 (EPH-INDEC).

¿Qué es un Árbol de Decisión?

→ Un árbol de decisión es un algoritmo de **aprendizaje supervisado no paramétrico**, que se utiliza tanto para tareas de clasificación como de regresión.

→ El aprendizaje del árbol de decisiones emplea una estrategia de "divide y vencerás" identificando los puntos de división óptimos dentro de un árbol. Este proceso de división se repite de forma *recursiva* de arriba hacia abajo hasta que la mayoría de los registros se hayan clasificado bajo etiquetas de clase específicas.

→ Es fácil de interpretar, pero propenso al *overfitting*

Random Forest

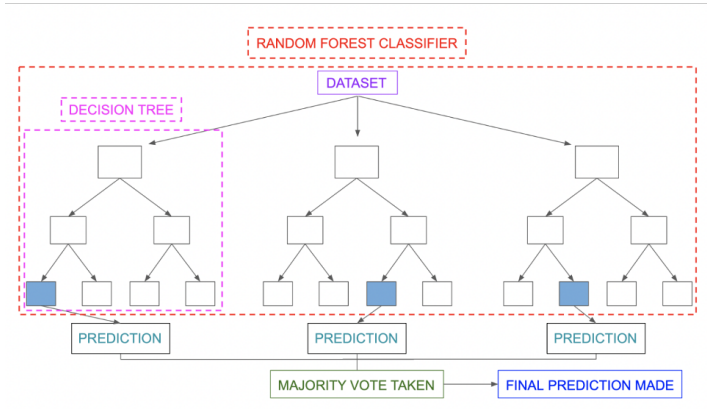


Figure 4: Construcción de Random Forest

¿Qué es un Random Forest?

→ Se trata de una técnica de **aprendizaje supervisado** que genera múltiples árboles de decisión sobre un conjunto de datos entrenados.

→ Un modelo Random Forest está formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping. Esto implica que **cada árbol se entrena con unos datos ligeramente distintos**.

Miss Forest

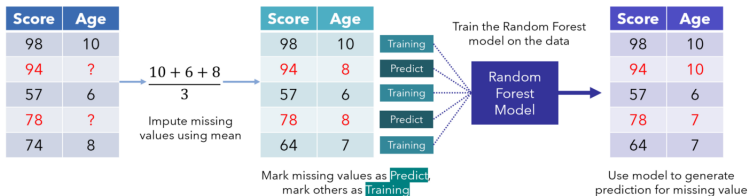


Figure 5: Ejemplo de Miss Forest

→ Método no paramétrico que no toma supuestos explícitos de la forma funcional de f . En cambio, trata de estimar f de manera tal que se acerque tanto a los datos sin volverse impráctico.

Algoritmo en el [► Anexo](#)

Los Resultados

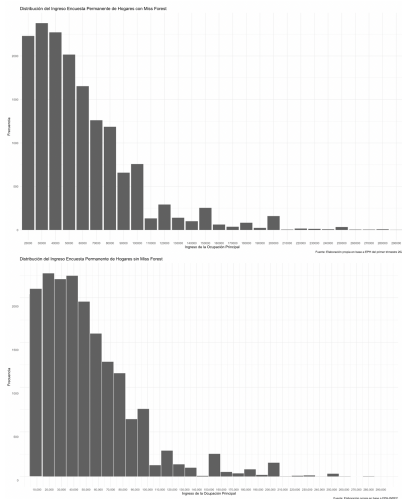
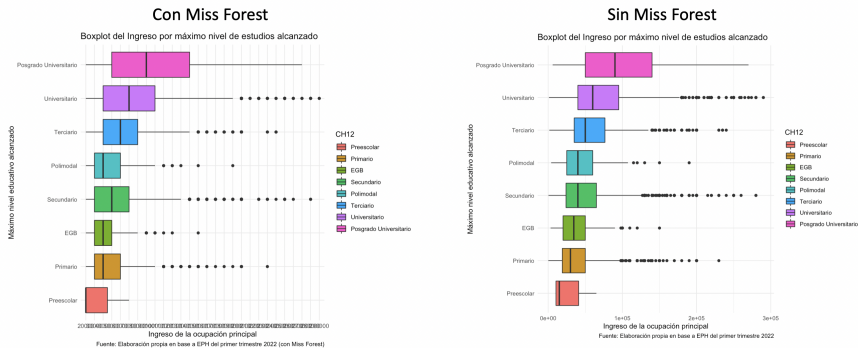


Figure 6: Comparación de Histogramas: el de arriba es con Miss Forest y tiene 49.706 observaciones y el de abajo 15.795



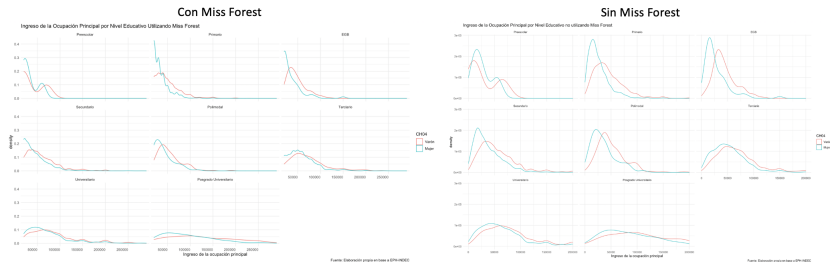


Figure 8: Comparación de Distribuciones por Nivel Educativo

Conclusión

- Efectivamente se consiguió un mayor poder estadístico, aunque a cambio de más varianza.
- Los resultados son un poco ruidosos. Esto puede tener que ver con los pocos datos que tenemos (se está estimando el 65% de los datos).
- Sería bueno incluir ponderadores.
- ¿Es útil como metodología para imputar ingresos? Bastante. En Argentina no hay muchos datos y ajustar por inflación tiene sus complicaciones. Este puede resultar un mecanismo alternativo interesante.

¡Gracias por la atención!

¿Preguntas?

Anexo

Existen dos maneras de pensar a la educación: como señal y como inversión en capital humano.

Es una combinación de ambas, pero suele predominar uno en particular. En Estados Unidos tiende a predominar el *signaling* que provee la educación (fui a Harvard, soy inteligente y estoy conectado con gente inteligente) y en Argentina es más una cuestión de *Capital Humano* (estudié economía que me provee de todas estas herramientas y conocimientos que puedo aplicar).

Este trabajo parte del supuesto (comprobado en la literatura del tema) que existe una fuerte correlación entre ingreso y nivel educativo: A mayor nivel educativo, mayores ingresos.

Retomemos: ◀ ¿Qué hacemos con los datos faltantes?

Miss Forest: el algoritmo

Let X be our $n \times p$ matrix of predictors that requires imputation:

$$X = (X_1, X_2, \dots, X_p) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \ddots & & \\ \vdots & & \ddots & \\ x_{n1} & & & x_{np} \end{bmatrix}$$

An arbitrary variable X_s contains missing values at entries $i_{mis}^{(s)} \subseteq \{1, 2, \dots, n\}$

For every variable X_s that contains missing values, we can separate the dataset into 4 categories:

1. The non-missing values of variable X_s , denoted by $y_{obs}^{(s)}$.
2. The missing values of variable X_s , denoted by $y_{mis}^{(s)}$.
3. The variables other than X_s , with observations $i_{obs}^{(s)} = \{1, 2, \dots, n\} \setminus i_{mis}^{(s)}$, denoted by $x_{obs}^{(s)}$
4. The variables other than X_s , with observations $i_{mis}^{(s)}$, denoted by $x_{mis}^{(s)}$

Miss Forest: el algoritmo

Algorithm:

1. Make an **initial guess** for all missing categorical/numeric values (e.g. mean, mode)
2. $k \leftarrow$ vector of column indices in X , sorted in **ascending order of % missing**
3. **while** not γ **do**:
4. $X_{old}^{imp} \leftarrow$ store previous imputed matrix
5. **for** s in k **do**:
6. Fit a random forest predicting the non-missing values of X_s : $y_{obs}^{(s)} \sim x_{obs}^{(s)}$
7. Use this to predict the missing values of X_s : predict $y_{mis}^{(s)}$ using $x_{mis}^{(s)}$
8. $X_{new}^{imp} \leftarrow$ update imputed matrix, using the predicted $y_{mis}^{(s)}$
9. **end for**
10. update γ
11. **end while**
12. **return** the final imputed matrix X^{imp}

Retomando: ◀ Miss Forest