Análisis de la participación femenina en el mercado laboral a través de árboles

En la primera parte se hará un análisis exploratorio de la Base EPH para captar relaciones que se producen en el mercado laboral de Argentina. En la segunda parte se construirán predictores con mediante árboles para entender la participación femenina en el mercado laboral.

PARTE I

Para entender qué mecanismo puede estar detrás de los de missings en los Ingresos de las personas vamos a explorar distintas hipótesis no exhaustivas. Dada la posibilidad de que los missings en ingresos se expliquen por alguna variable no observada, es difícil probar esto de manera concluyente dada las múltiples dimensiones que están fuera del alcance de la EPH.

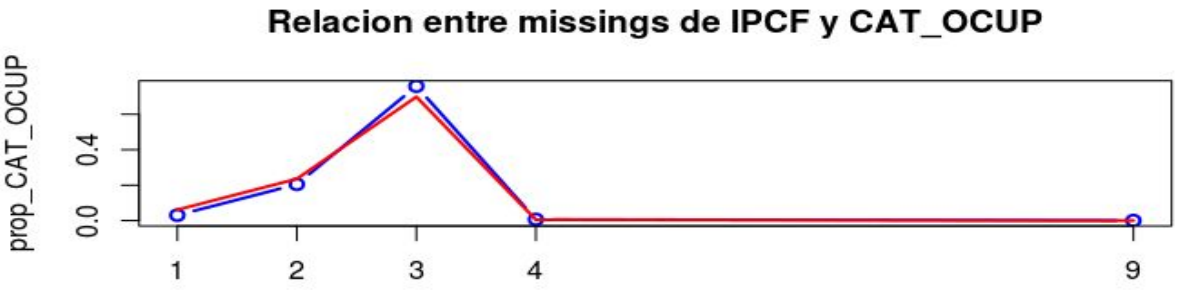
Sin embargo nos proponemos analizar si se verifica alguna relación entre la informalidad (que no es observada en la EPH) y la no respuesta de ingresos. Para eso vamos a utilizar una variables proxy de la situación de informalidad que es la variable PP07G4: "En este trabajo tiene obra social?".

Por lo tanto vamos a comparar cómo se distribuye la variable PP07G4 entre las personas que tienen missings en IPCF (Ingreso per cápita familiar) y en las que no lo tienen.

Como se visualiza en el gráfico, en ambos grupos la distribución entre personas que tiene descuento de obra social es casi idéntica, con lo cual se puede inferir que no hay relación aparente entre situación de informalidad y la no declaración de ingresos. Para poder afirmarlo la línea roja debería haber tenido un a proporción más alta del valor x =2 respecto de la línea azul, o lo que es lo mismo las personas que no tienen obra social (x=2)deberían haber tenido más no respuestas de ingresos que los que tienen obra social (x=1).

Por otra parte, existe evidencia empírica[1] de que la no respuesta en encuestas de ingresos se relaciona con la fuente de ingresos, en donde quienes tienen una actividad lucrativa a través de ganancias de capital o cuenta propia tienen más probabilidad de no dar respuesta a sus ingresos.

Por ello vamos a analizar la relación entre missings de IPCF para las distintas fuentes de ingresos, que se puede ver con la variable "CAT_OCUP" (Categoría Ocupacional)



Como vemos en el gráfico la línea roja se despega para arriba de la línea azul en los valores de 1 (Patrón) y 2 (Cuenta propia) de la variable CAT_OCUP

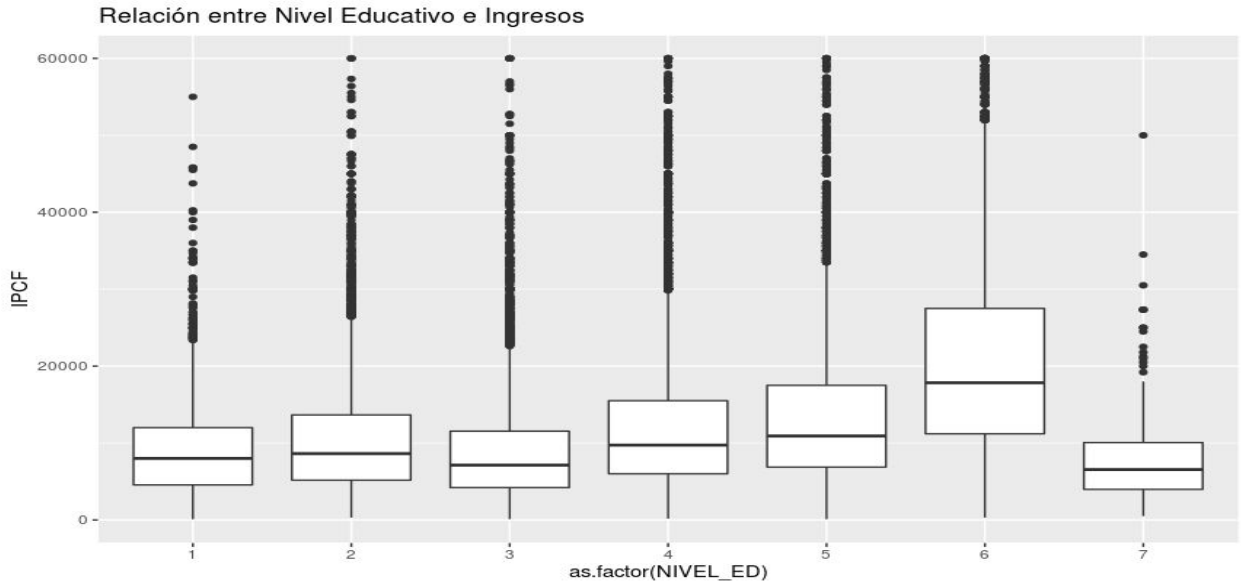
Aquí se ve que entre los missings de IPCF hay una proporción de 6,1% que proviene de personas que declaran ser patrones y en toda la base la proporción de patrones es de 3,0%. Con esto se evidencia que las personas que son patrones tienen una tendencia muy alta a no declarar ingresos. Algo similar pero de menor magnitud sucede con quienes declaran desenvolverse por cuenta propia, que son el 20% del total, pero en el subconjunto de datos que tienen missings por ingresos representan el 23%. Con esto podemos afirmar que los missings de ingresos de los que declaran ser patrones o cuentapropistas no se distribuyen al azar, sino como Missings at Random.

Sin embargo estas dos actividades sumadas tienen una baja participación en el total (23%) y no podríamos encontrar la raíz del problema sólo a partir de este aspecto. Por lo tanto tendríamos que abordar otras variables socioeconómicas que expliquen el fenómeno en su totalidad.

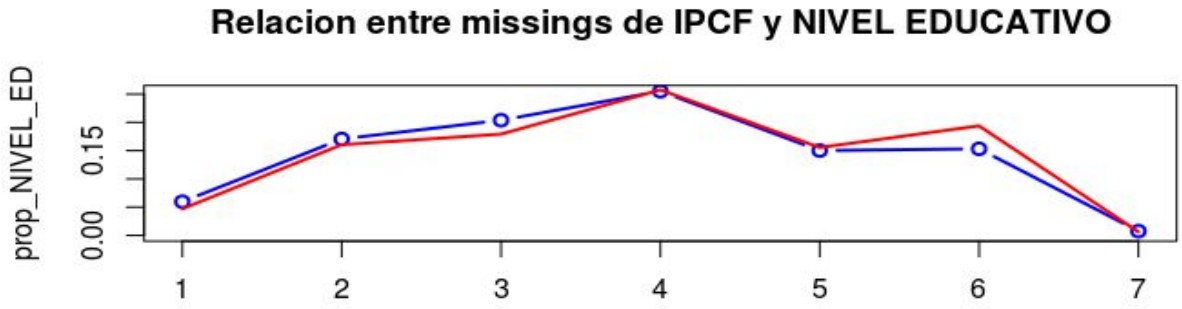
En otro orden de ideas, también existe la posibilidad de que los missings de IPCF tengan que ver con el valor la variable en sí misma, es decir que a determinados valores de IPCF las personas tienden a ocultar sus ingresos. Por lo tanto, vamos a repetir el análisis anterior pero con la variable NIVEL_ED: "Nivel educativo". Elegimos esta variable dado que se verifica que a

[1] Eduardo Donza (2013). Método de imputación de la no respuesta en las preguntas de ingresos en la Encuesta Permanente de Hogares. Gran Buenos Aires 1990-2010. X Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires, Buenos Aires.

mayor nivel educativo se perciben mayores ingresos, sobre todo en quienes tienen universitario completo (x=6), tal como surge del siguiente gráfico.



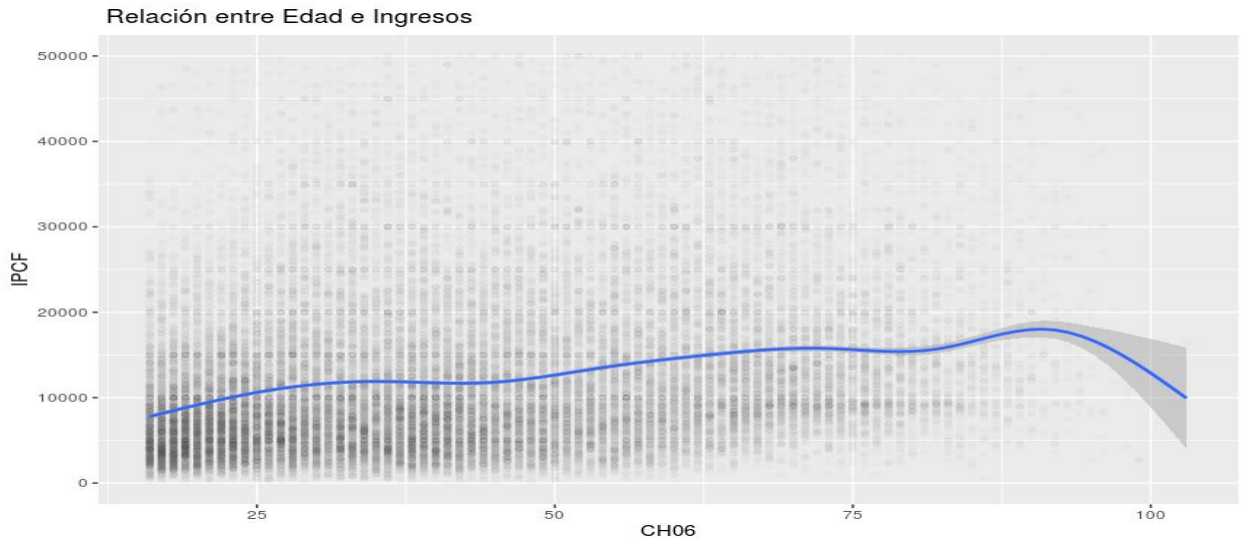
En el siguiente gráfico se puede ver claramente como existe mayor cantidad de no respuesta entre quienes tienen estudios universitarios completos(x=6), con lo cual se confirma la posibilidad de que además de los expuesto anteriormente, el valor mismo de la variable IPCF genera missings.



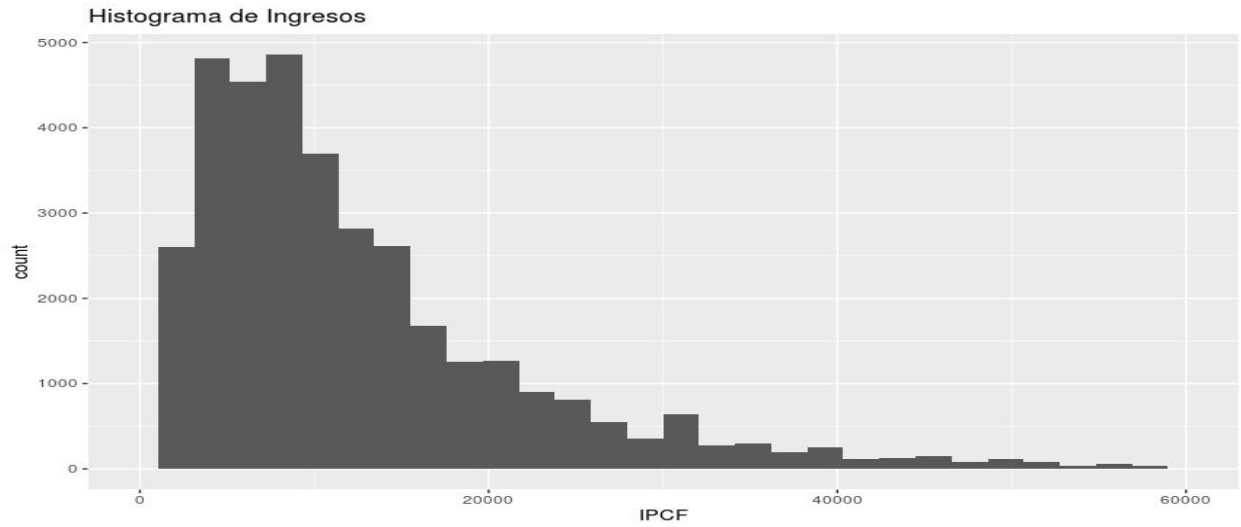
Por lo tanto luego del análisis que hemos realizado podemos decir que operan dos mecanismos de missings, a saber: *Missingness at random* y *Missingness that depends on the missing value itself.*

En un contexto de regresión lineal por OLS lo solucionaríamos seleccionando algunas variables relevantes de tipo sociodemográficas (Por ejemplo edad, sexo, nivel educativo, estado y categoría ocupacional), algunas de las cuales ya hemos detectados que son significativas para explicar las no respuestas. Con esto realizaríamos un modelo teniendo en cuenta 4 aspectos: al ser casi todas las variables categóricas (excepto edad) habría que convertirlas a dummies para que la regresión lineal tenga sentido. Respecto a la variable edad, dado que tiene una

relación no lineal con los ingresos(como se ve en el siguiente gráfico), habría que transformarla en dummy agrupando distintos rangos etarios.



Además como se verifica que la distribución de la variable IPCF es asimétrica hacia la derecha, tal como se ve en el próximo gráfico, se podría hacer una transformación logarítmica de la variable.



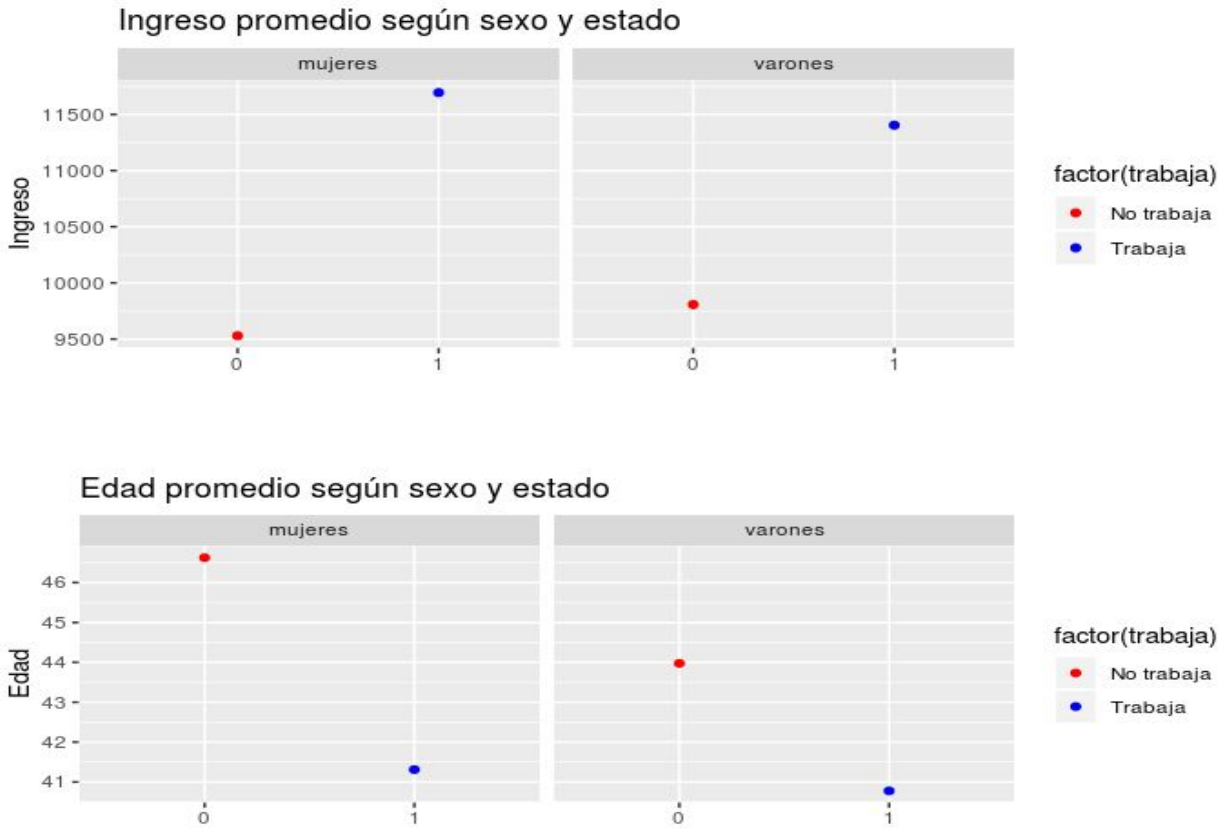
Como último paso habría que eliminar todas las filas que presenten algún missing en las variables que definimos como explicativas.

b)Habiéndose explicado todos los aspectos relevantes para estimar los missings de ingresos a través de OLS, nos vemos ante la dificultad de realizar un modelo para estos fines dado que existen muchos missings en los predictores.
Tal es así que una de las variables más significativas que detectamos para explicar missings de IPCF como lo es CAT_OCUP, contiene un 42% de missings, por ejemplo.

Por lo tanto hemos resuelto reemplazar los missings de IPCF con la media, a sabiendas de que si bien es un método óptimo para el mecanismo *Missingness completely at random , no es lo que encontramos acá (*como se expuso, lo que detectamos fue *Missingness at random* y *Missingness that depends on the missing value itself).*

El costo de esta elección se va a traducir en un mayor sesgo ya que posiblemente se deje de captar un subconjunto de la población asociado a los ingresos por rentas de explotación y/o a ingresos altos, entre otros.

7)



En los gráficos podemos observar los siguientes resultados:

El ingreso promedio según sexo y estado es superior para mujeres que hombres en promedio por unos $500 aproximadamente.No obstante, la edad promedio de las mujeres empleadas en promedio, para cada estado, es de aproximadamente 3 años superior.

Esta discrepancia en ingresos podría ser justificada por el mayor “seniority” en promedio que presentan las mujeres promedio según estado.

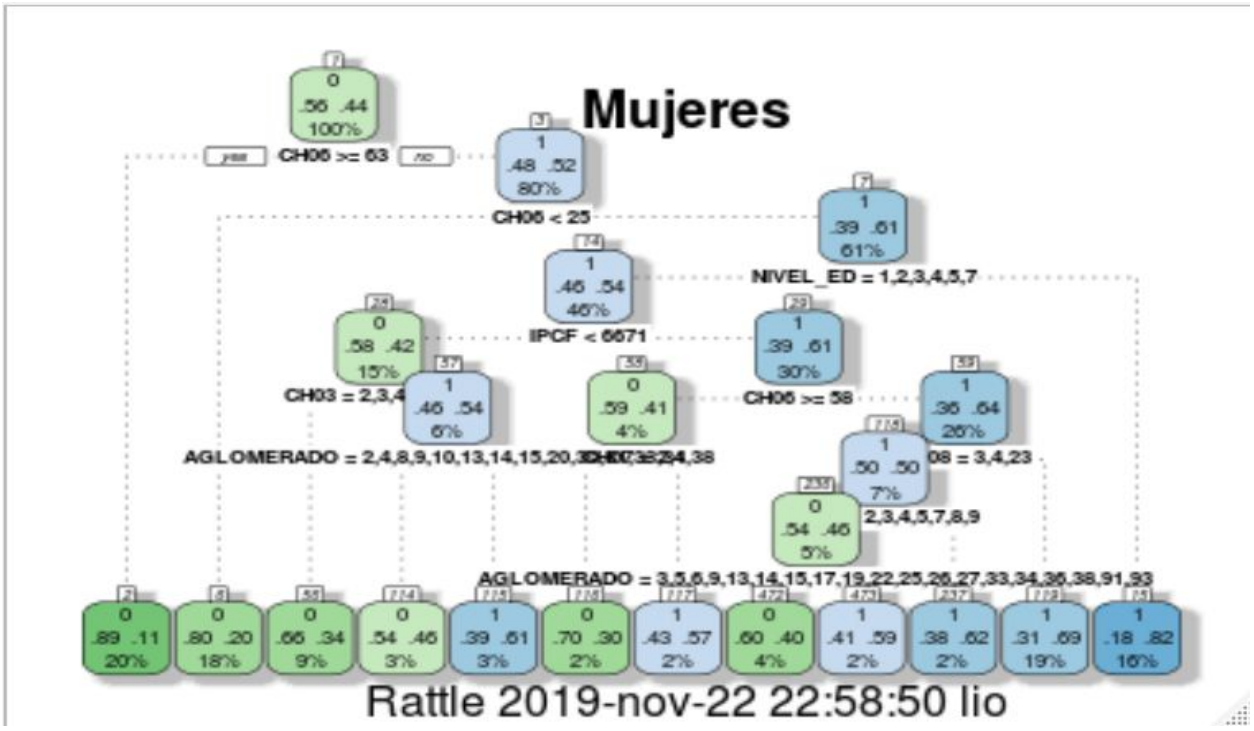
También observamos el promedio de niños donde hay hogares con varones promedio según estado, que trabajan es superior.Sin embargo, esto no es así para las mujeres promedio según estado que no trabajan presentando un promedio de incidencia de 0.6 contra 0.

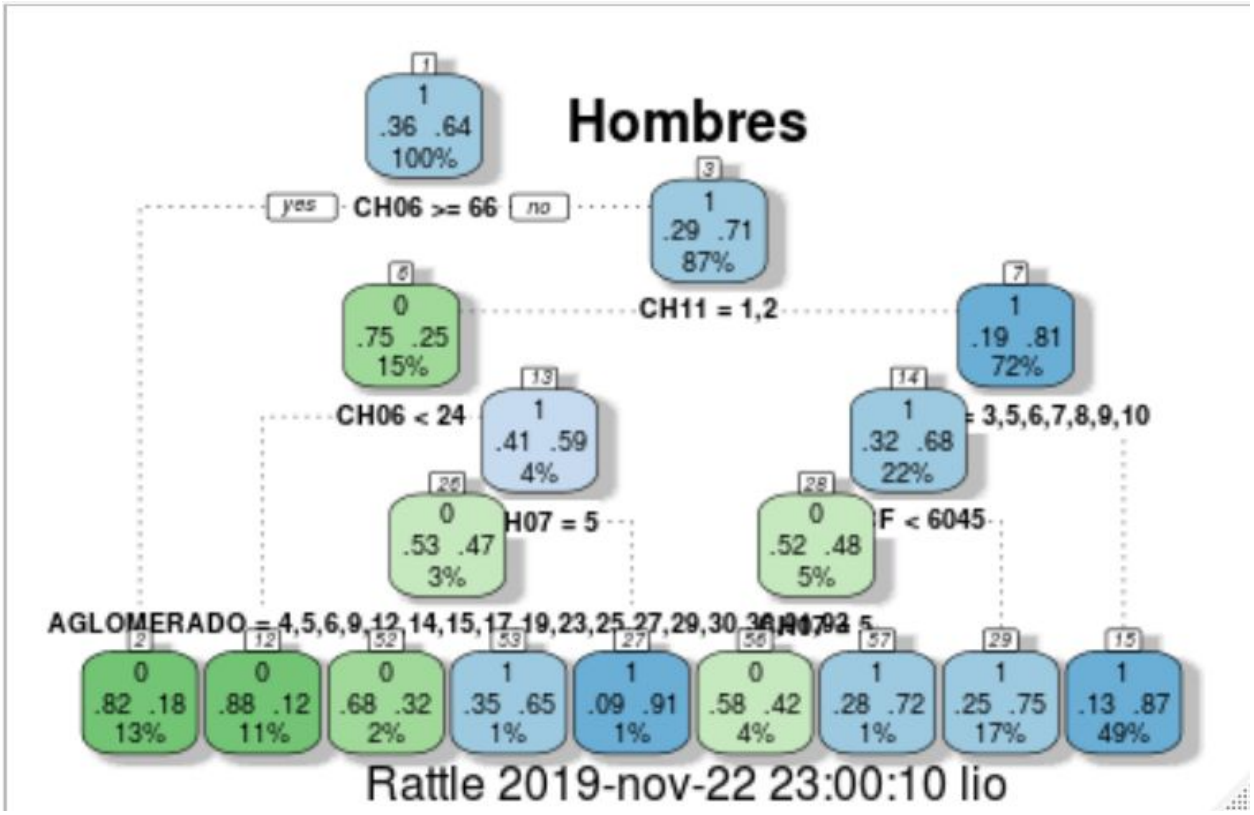
En términos comunes, podríamos decir se sugiere que no hay padres que no trabajen que mantengan custodia de niños en promedio.

Tal como se muestra en el siguiente cuadro, las proporciones de amos de casa es muy dispar a favor de las mujeres, algo que da la pauta de la predominancias de los hombres en el mercado laboral. Por su parte, la distribución de los jubilados según el sexo es un poco más pareja aunque las mujeres son más que los hombres , algo que posiblemente se explique por la edad jubilatoria menor y por una esperanza de vida mayor.

	amos_de_casa	jubilados	Sexo
1	0.08749382	0.3687767	varones
2	0.91250618	0.6312233	mujeres

PARTE II





En el árbol de mujeres aparecen las variables: CH06, CH07, NIVEL_ED, CH03, CH10, CH11, CH12 y IPCF.

En el árbol de hombres aparecen las variables: CH06, CH10, CH11, CH14, CH07, CH03, NIVEL_ED, CH13 y AGLOMERADO.

Es decir que si bien ambos modelos son distintos tienen tienen variables en común, a saber: NIVEL_ED, CH03, CH06, CH07 y CH10. Y a su vez los dos modelos tienen como primer variable para hacer la división a CH06 que es la edad de las personas.

Una de las cosas que diferencia al árbol de mujeres del de los hombres es que en este último es importante la ciudad donde se encuentra la persona, no así en el caso de las mujeres.

La primer variable suele ser un predictor importante porque abarca a una gran proporción de las observaciones. En este caso se ve que la edad es fundamental para identificar la participación en el trabajo de las personas, algo que está asociado a las condiciones de la fuerza de trabajo (poco productivas para los jóvenes o los ancianos)

Las matrices de confusión son medidas útiles para determinar el accuracy de un modelo. Reflejan el número de falsos y verdaderos positivos así como falsos y verdaderos negativos para las categorías sobre las cuales se quiere efectuar predicciones. En este caso el accuracy es una buena medida dado que la población que trabaja y no trabaja, es decir las clases, está bien distribuida.

Encontramos que los dos árboles tienen una buena precisión de 72% para mujeres y 82% para hombres. Por lo tanto el de hombres tiene mayor precisión que el de mujeres a la hora de predecir la participación en el trabajo

Observamos el ratio de falsos negativos para mujeres es significativamente superior al del caso de hombres, pero no hay tanta diferencia respecto de falsos positivos(teniendo en cuenta que el total de mujeres es mayor al total de hombres).

```
       pred_test_mujeres
y_test    1   0
     1  2217 881
     0  1054  2985        accuracy_mujeres=0.72

        pred_test_varones
y_test    1  0
     1  3726 411
     0  738 1520          accuray_varones= 0.82
```

Habiendo construido un modelo Lasso y usando Cross Validation nos quedan las siguientes matriz que tiene un accuracy menor en ambos casos.

```
# pred_test_mujeres
# y_test    0    1
#      0 1858 1271
#      1  909 3099    accuracy= 0.69

# pred_test_varones
# y_test    0    1
#      0 3720  440
#      1  930 1305    accuracy=0.78
```

Con lo cual vemos que este modelo no logra ser mejor a los árboles estimados previamente para ninguno de los dos sexos.

Por último, utilizando la base completa nos quedan las siguientes variables relevantes: CH0, CH04, CH10,CH11,CH07, NIVEL_ED, CH03, CH12, IPCF, CH14.

Donde vemos que el sexo (CH04) es una de las más importantes.