Informe Trabajo Práctico 2

Parte I: Analizando la base

1- Para el INDEC, la identificación de las personas pobres o indigentes se realiza a través de un método de medición indirecta. A través de este, se fija una "línea" a partir de la cual los hogares, según sus ingresos registrados en la Encuesta Permanente de Hogares, son clasificados como pobres o indigentes. En el caso de la Línea de Indigencia, esta se establece mediante la Canasta Básica Alimentaria, la cual incluye los alimentos necesarios para las necesidades energéticas y proteicas básicas. Para el caso de la Línea de Pobreza se extiende esta definición, conformando la Canasta Básica Total, que incluye también otros consumos básicos no alimentarios.

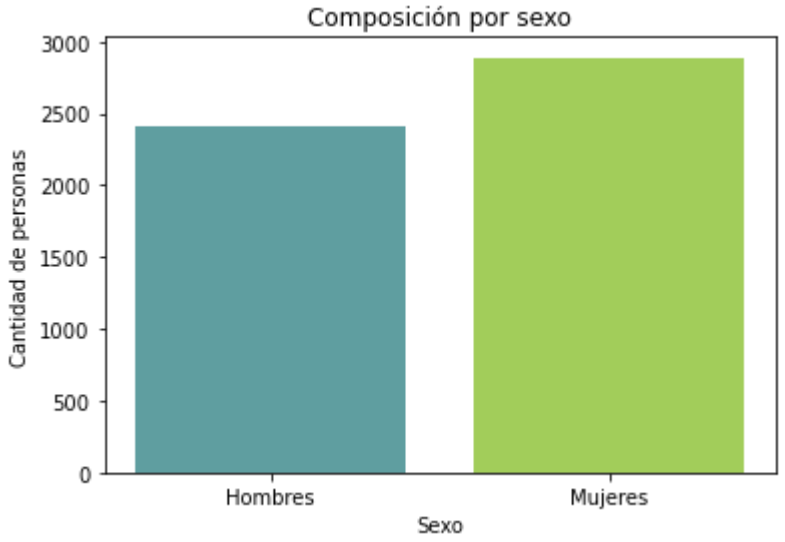
Referencias: Instituto Nacional de Estadística y Censos (2016) *La medición de la pobreza y la indigencia en la Argentina*. INDEC. https://www.indec.gob.ar/ftp/cuadros/sociedad/EPH_metodologia_22_pobreza.pdfLa cantidad de personas que no respondieron y predecimos que son pobres es: 697

La cantidad de personas que no respondieron y predecimos que son pobres es: 697

2-

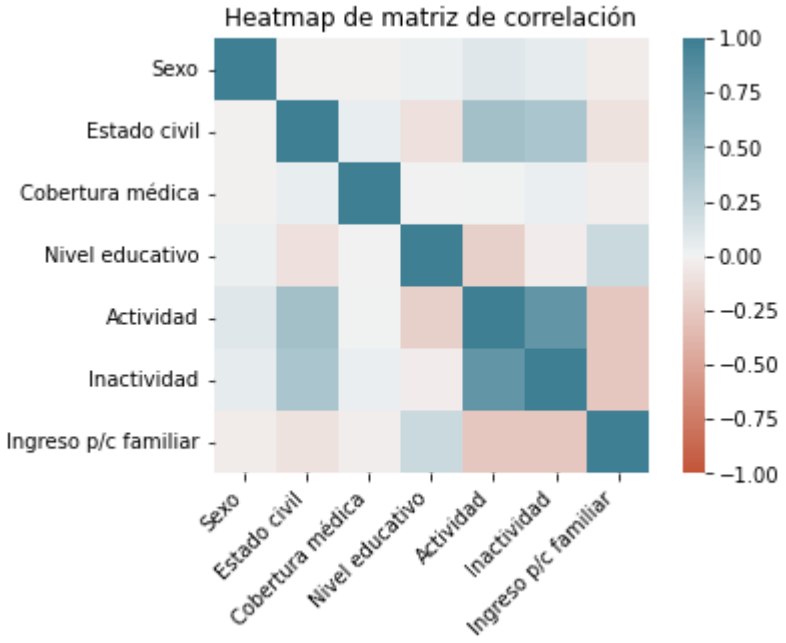
a. En Jupyter Notebook.

b. En Jupyter Notebook.



c. Observamos que la muestra se compone en su mayoría por mujeres.

d.



Observamos que en la diagonal principal la correlación es perfecta, dado que se trata de la correlación de la variable con si misma. Hay una correlación alta entre "actividad" e "inactividad", esto se debe a que los valores más altos de la primera variable se definen para las personas "desocupadas", "inactivas" o "menores". Esto coincide con las especificaciones de las categorías de inactividad. También hay correlación positiva, pero menor, para el "estado civil" con "actividad" e "inactividad". Similarmente sucede con el "nivel educativo" y el "nivel de ingreso familiar".
Por otra parte, podemos señalar un grupo de correlaciones cercanas a cero, estas son:

- "sexo" con todas las otras variables
- "cobertura médica" con todas las otras variables
- "nivel educativo" con "inactividad"

Por último, encontramos las siguientes correlaciones negativas bajas y moderadas:

- "estado civil" con "nivel educativo" y con "ingreso familiar"
- "nivel educativo" con "actividad"
- "actividad" e "inactividad" con "ingreso per cápita familiar"

e. En la muestra hay 216 personas desocupadas y 2421 personas inactivas. La media del ingreso familiar per cápita para las personas ocupadas es de $47629.05, para las personas desocupadas es de $15850.85 y para las personas inactivas es de $24879.70.
f. En Jupyter Notebook.

3- Hay 1551 personas que no respondieron cuál es su ingreso total familiar.

4- En Jupyter Notebook.

5- En la base hay 1190 personas pobres.

Parte II: Clasificación

1- En Jupyter Notebook.

2- En Jupyter Notebook.

3- En Jupyter Notebook.

4- De los tres métodos de predicción, nos quedaríamos con el de análisis discriminante lineal. Esto se debe a que, a pesar de que tiene un accuracy score, AUC y curva de ROC muy similar a la regresión logística, al observar la matriz de confusión vemos que comete menos falsos negativos. En nuestro problema, este es el tipo de error que más nos preocuparía cometer, dado que implica predecir que una persona no es pobre cuando en verdad lo es.

5- La proporción de personas que no respondieron cuál es su ingreso y son pobres es: 45%.

6- En primer lugar, una gran cantidad de variables sólo garantiza un sesgo pequeño dado que se maximiza el R cuadrado y tendrá una buena estimación hacia adentro de la muestra. Sin embargo, un R cuadrado muy alto no genera buenas predicciones fuera de la muestra. Además, la inclusión de muchas variables aumenta la varianza de la estimación.
En cambio, incluir menos variables, si bien puede implicar un menor R cuadrado, disminuye la varianza del estimador y lo hace más eficiente. Esto implica que nos encontramos ante un *trade off* entre el sesgo y la varianza del estimador. Asimismo, incluir demasiadas variables nos puede llevar a un problema de *overfitting*. Por lo tanto, existe una cantidad eficiente de variables a incluir en el modelo para lograr predecir eficientemente fuera de la muestra.

Es por ello que no nos parece una buena decisión incluir todas las variables si el objetivo es de clasificación. En cambio, creemos que una mejor alternativa sería incluir una selección relevante de ellas. Consideramos que esta selección relevante, en este caso, se compone de las variables demográficas y ocupacionales. Esto se debe a que son las variables que de acuerdo a la literatura correlaciona con el nivel de ingresos y por lo tanto permiten explicar este fenómeno.

Observamos que al volver a correr el modelo con estas variables, la precisión disminuye ligeramente (pasa a ser de 0.69). Si bien conceptualmente, esperaríamos que este valor aumente, por disminuir la cantidad excesiva de variables, no observamos ese resultado que esperábamos. Esto podría deberse a una selección no óptima de variables. Dado que el concepto de que muchas variables generan un modelo poco eficiente en la predicción es válido en todos los casos, consideramos que sería necesario interiorizarse más en la problemática para poder plantear un modelo con mejor precisión.