Trabajo Práctico N°4: Big Data

Guerrero Ponzo Felicitas Trinidad - Sundblad Alfonso - Vergara Julian

Profesores: Romero Noelia - Oubiña Victoria

Universidad de San Andrés - Primavera 2023

1. PARTE I

1.1. Limpieza base de datos

Para la limpieza de la base de datos realizamos varias acciones. En primer lugar, eliminamos variables con *missings values*. Las variables en cuestión, tenían una gran proporción de valores faltantes y luego para el análisis es necesario darle un tratamiento a estos y lo que consideramos mejor fue directamente eliminar las variables con estos problemas. Eliminar directamente las observaciones con valores faltantes nos hubiese hecho perder una gran cantidad de datos. En segundo lugar, eliminamos las observaciones que contenían como respuesta 9 o 99 que son las categorías que hacen referencia a *no sabe/no responde* que son respuestas a preguntas que no aportan al análisis. En tercer lugar, eliminamos observaciones que identificamos como *outliers*. Consideramos que las variables con potenciales *outliers* podrían ser las referidas a la edad o al ingreso en la que hay una respuesta directa del encuestado y no tiene que elegir entre opciones predeterminadas a las preguntas. El criterio para eliminar *outliers* fue utilizar el rango intercuartil pero solo eliminando las que esten por el límite superior al cuartil. Debido a que, nos interesa el cuartil inferior para luego partir la muestra y ya que se admite la posibilidad de cero tanto en ingresos como en edad. Además, eliminamos variables que tienen el mismo valor para todas las observaciones como región, aglomerado, año, trimestre, entre otras y las variables con observaciones negativas lo cual también es un error en la medición de la encuesta.

1.2. Creación de nuevas variables

Además de las variables que nos pedía la consigna de proporción de niños en el hogar y la situación ocupacional del cónyuge, creamos una variable dummy de hacinamiento y una variable de acceso al sistema financiero de las familias. Para la primera, usamos de referencia la definición de hacinamiento del INDEC [1]. Se considera hacinamiento cuando tres o más personas están en una misma habitación. Consideramos que es una variable relevante ya que el hacinamiento el hogar consideramos que es una característica de los hogares pobres. Con respecto a la segunda, sintetizamos en una única variable aquellas preguntas relacionadas al contacto con entidades financieras de un hogar. En particular, unimos en una dummy que toma valor 1 cuando hay al menos una respuesta positiva a las preguntas de prestamos a bancos/financieras, tarjetas de créditos o inversiones en plazo fijos. Consideramos que podíamos resumir estas tres variables en una sola debido a la correlación entre ellas y también puede servir para describir la pobreza de un hogar. En general, hogares pobres están

[1] Más información en: https://www.indec.gob.ar/ftp/indecinforma/nuevaweb/cuadros/7/sesd$_g$*losario.pdf*

lejos de cumplir las condiciones que se exigen para acceder al sistema financiero y utilizar servicios de financiamiento e inversiones que ofrecen.

1.3. Estadísticas Descriptivas

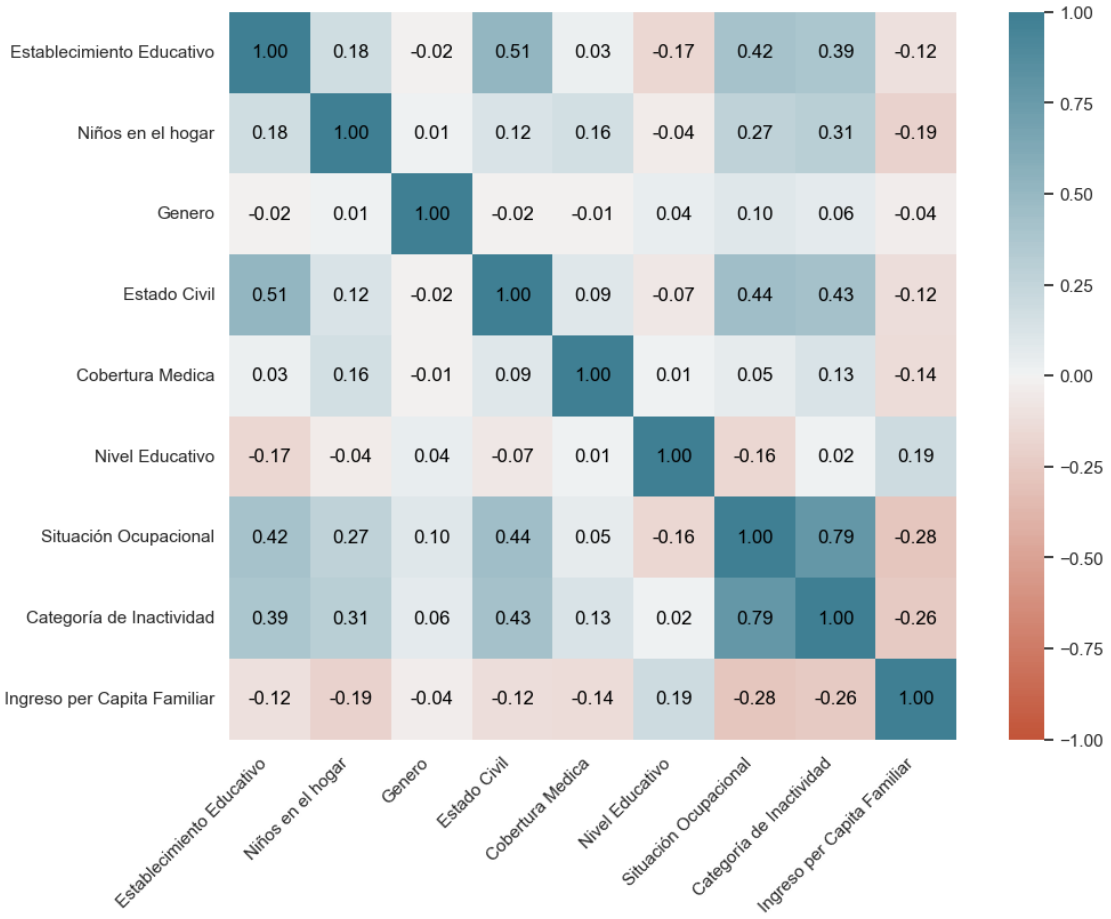

Figura 1: Mapa de correlaciones: variables interesantes

Podemos observar que la variable establecimiento educativo (la cual indica si el establecimiento es privado (2) o público (1)) tiene correlación positiva con la cantidad de niños en el hogar (18 %), el estado civil (51 %), la cobertura médica (aunque la correlación es baja, 3 %), situación ocupacional (0.42 %) y categoría de inactividad (39 %). Mientras que con género (2 %), nivel educativo (17 %) e ingreso per cápita familiar (12 %) tiene relación negativa. Esto nos indica que a mayor ingreso y nivel educativo es más probable que vayan a una institución privada. Niños en el hogar tiene correlación positiva con genero (aunque baja, 1 %), estado civil (12 %), cobertura médica (16 %), situación ocupacional (27 %) y categoría de inactividad (31 %). Tiene correlación negativa con nivel educativo (4 %) e ingreso per cápita familiar (19 %). Género es una variable que tiene correlaciones muy bajas con todas las variables, siendo la más alta la relación con situación ocupacional de 10 %. Estado civil tiene correlaciones positivas y altas con establecimiento educativo (51 %), situación ocupacional (44 %) y categoría de inactividad (43 %). Tiene correlaciones negativas con nivel educativo (7 %) y con ingreso per cápita familiar (12 %). Cobertura médica tiene correlaciones positivas con niños en el hogar (16 %), categoría de inactividad (13 %) y correlación negativa con ingreso per cápita familiar (14 %).

Las variables con mayor correlación son situación ocupacional y categoría de inactividad, con una correlación de 79 %.

Seguidos de estado civil y establecimiento educativo, con una correlación de 51 %.

1.4. Estimación Pobreza

De acuerdo al último informe del INDEC [2], el 40,1 % de la población está bajo la línea de la pobreza y el 29,6 % de los hogares del país son pobres. Nuestra estimación de hogares bajo la línea de pobreza es del 40 %, por lo que estamos sobrestimando el número total de hogares pobres a partir de la comparación entre el ingreso total familiar reportado y la variable de ingreso necesaria creada.

2. PARTE III

2.1. Comparación de modelos (Incisos 3 y 4)

Modelo	Hiperparametro	Accuracy	AUC	Precision	Specificity	Recall	True Negatives	True Positives	False Negatives	False Positive	ECM
CART	19	0.8738	0.8631	0.8449	0.8907	0.8184	660	365	81	67	0.1262
LDA	-	0.8150	0.7883	0.8053	0.8195	0.6771	654	302	144	73	0.185
KNN	1	0.7937	0.7824	0.7257	0.8363	0.7354	603	328	118	124	0.2063
Logit (Ridge reg)	15.85	0.6283	0.5558	0.5231	0.652	0.2533	624	113	333	103	0.3717
Logit (Lasso reg)	1	0.6283	0.5558	0.5231	0.652	0.2533	624	113	333	103	0.3717
RF	41	0.815	0.7792	0.8438	0.8035	0.63	675	281	165	52	0.185
Boosting	14	0.8747	0.8581	0.8691	0.8776	0.7892	674	352	94	53	0.1253
Bagging	-	0.8022	0.7733	0.7908	0.8074	0.6525	650	291	155	77	0.1978

Cuadro 1: Comparación métricas modelo

La tabla 1 resume el desempeño de distintos modelos de acuerdo a diferentes métricas. En primer lugar, cada modelo tiene como hiperparametro aquel que dentro de la lista de opciones de valores minimizaba el error cuadrático medio. En particular, para el modelo Logit, el hiperparametro óptimo, tanto para la regularización Lasso como Ridge fue el parámetro relacionado a la penalidad impuesta a los regresores. Para vecinos cercanos (KNN), buscamos el numero de vecinos cercanos óptimos para una observación. Para Random Forest (RF) buscamos el la cantidad de variables que incluimos en cada árbol mientras que para CART (Classification and Regresion Tree) y Boosting optimizamos la profunidad de cada arbol (cantidad de particiones). Notar que para Bagging y Análisis Discriminante Lineal no optimizamos ningún parámetro.

Ahora analicemos las diferentes métricas. En primer lugar, la accuracy, que indica la cantidad de predicciones que el modelo estimo correctamente. En este caso el modelo con mayor accuracy fue el de Boosting con 0.8747 seguido de muy cerca por CART con 0.8738. Luego analizamos el AUC, el cual indica que tan cerca estamos de la situación ideal de todos verdaderos positivos y verdaderos negativos. En este caso, el modelo con mayor AUC (Area bajo la curva) score es CART con 0.8631, seguido de Boosting con 0.8581. Analizamos la presición de los modelos, de los cuales el que mejor precisión tiene es Boosting con 0.8691 seguido nuevamente por CART con 0.8449. Analizando la specificity la cual indica la proporcion de verdaderos negativos que el modelo clasifica correctamente. El modelo con el mayor valor es CART con 0.8907, seguido de Boosting con 0.8776. Analizamos la cantidad de verdaderos negativos que tiene cada modelo, de los cuales los valores más altos son de RF (674) y Boosting (675). En el caso de los verdaderos positivos son el modelo CART (365) seguido de Boosting (352). En el caso de falsos negativos los valores más bajos son los de CART (81) seguido de Boosting (94), y en el caso de falsos positivos los más bajos son RF (52) y Boosting (53). Finalmente, observamos el error cuadrático medio (ECM) de los modelos. El modelo con el valor más bajo es Boosting con 0.1253, seguido de CART

[2] https://www.indec.gob.ar/uploads/informesdeprensa/eph$_p obreza_0 9_2 326 FC 0901 C 2.pdf$

con 0.1262. En conclusión, podemos obervar que dadas las métricas que usamos los dos modelos que mejor rinden son Boosting y CART, siendo el primero ligeramente superior.

Por otro lado, notar que los modelos Logit tanto con una regularización Lasso como con una Ridge tiene un desempeño muy inferior al de los otros modelos a diferencia de lo que sucedía en trabajos anteriores donde se destacaba como el modelo con mejor desempeño. Notar que también los score son los mismos con ambas regularizaciones lo cual es una señal de un error en el código con respecto a Logit que no pudimos identificar y arruina el desempeño del modelo.

Otro resultado a destacar, es que CART tiene mejor desempeño que Random Forest y Bagging cuando estos últimos son mejores ya que disminuyen el problema de *overfitting* del primero y tienen en general, mayor capacidad predictiva. Nuevamente, esto puede deberse a un problema en el código no detectado o bien que dada la naturaleza de los datos del INDEC la metodología de Bagging y Random Forest no sea la óptima.

Finalmente, comparemos el desempeño de los modelos con respecto al trabajo anterior. Vemos que Vecinos Cercanos y Análisis Discriminante Lineal se desempeñan sistemáticamente peor y todos los métodos relacionados a arboles tienen un desempeño inferior al que tienen los tres modelos utilizados en el trabajo anterior (KNN, Logit, LDA). No obstante, esto no implica que necesariamente sean peor, a priori. Ya que antes pudo existir un problema de *overfitting* que hiciera muy bueno el modelo dentro de la muestra pero muy malo fuera de esta.

2.2. Estimación Pobreza (Inciso 5)

Al igual que en el trabajo anterior, la estimación del numero de pobres es muy mala (0 %) muy lejos de lo reportado por el INDEC y de nuestra primera estimación. Esto marca problemas en nuestro código, especialmente en nuestras funciones que generan un severo problema de *overfitting* que hace muy malo el desempeño del modelo por fuera de la muestra. Creemos que el problema se encuentra en la función de *evalua multiple metodos* y no pudimos detectar.