Universidad de San Andrés
Departamento de Economía
Buenos Aires
noviembre, 2023

Trabajo Práctico N° 4

Big Data, Aprendizaje y Minería de Datos

Camila Riancho

Diego Fasan

Ronny M. Condor

Profesora: Noelia Romero

Asistente: Victoria Oubiña

Parte I: Análisis de la base de hogares y cálculo de pobreza

Este análisis de la base de datos no es tan distinto a lo que ya venimos haciendo durante el curso. Gracias al feedback del TP pasado, pudimos identificar el problema en nuestra limpieza de datos. El error en el TP pasado fue cuando hicimos el merge, lo que llevó a quedarnos principalmente con variables individuales, y las del hogar quedaron como missing. En el presente trabajo arreglamos esto. Tomamos mucho tiempo en la limpieza de datos, incluso hicimos una parte en Stata (hay experiencia aquí), como una especie de double check.

Parte II: Construcción de funciones

Las funciones fueron muy similares a las que ya tuvimos en el TP pasado. Para optimizar la selección de hiperparámetros, usamos la función `GridSearchCV`. Obviamente, en el inciso 4, agregamos los nuevos modelos aprendidos en clase: árboles de decisión, bagging, random forest y boosting.

Parte III: Clasificación y regularización

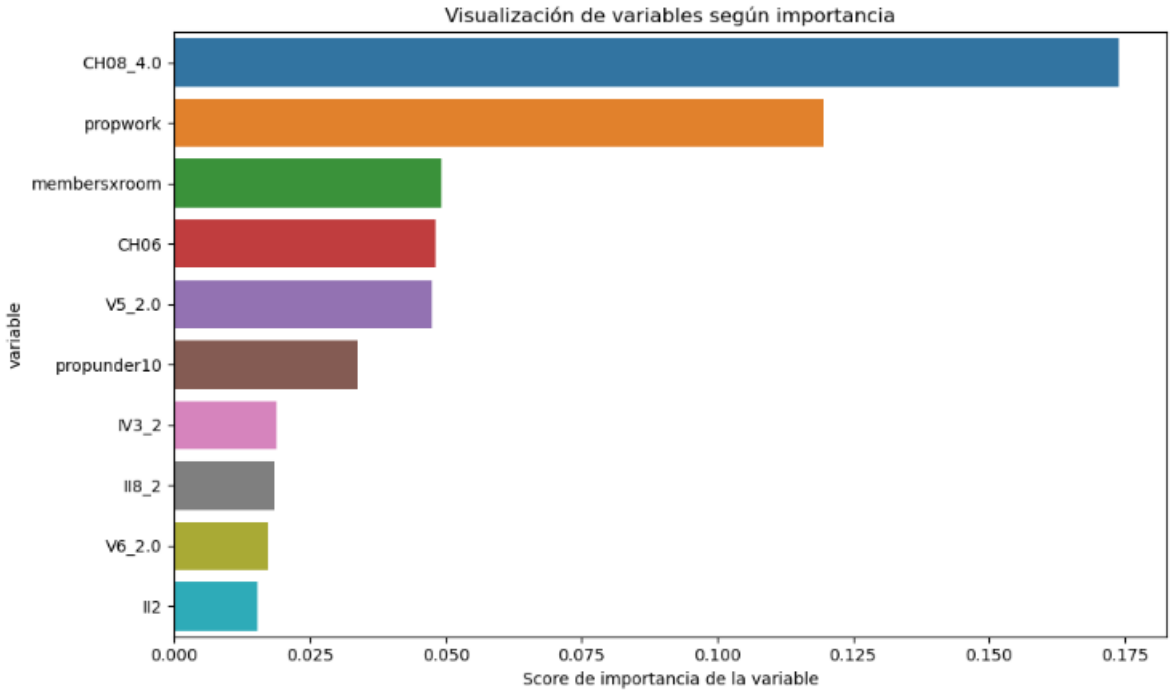
Vale aclarar que estandarizamos las variables de la EPH. Definimos la función `evalua_multiples_metodos()` con una única diferencia al TP anterior: no incluimos el modelo KNN. ¿A qué se debe esto? Bueno, cuando probamos la función con la data ficticia, todo funcionó sin problemas. Sin embargo, cuando queremos hacer lo mismo para los datos de la EPH, tenemos un error. Investigamos detalladamente el origen del error y esto se da cuando usamos `evalua_config()` para KNN. A pesar de las horas dedicadas a resolver este problema, no pudimos solucionarlo, lamentablemente. Tomamos la decisión de seguir adelante sin considerar este modelo. Como mostramos más adelante, las métricas de los otros modelos son ridículamente buenas (al menos, considerablemente mejores que las que tuvimos en el anterior TP). Quedaremos con la duda de si KNN podría haber superado estas métricas.

El modelo que mejor predice es Bagging que da como resultado un accuracy de 0.92 y un ECM de 0.07. El segundo mejor modelo es Random Forest que nos da un accuracy de 0.90.

Por curiosidad, miremos las variables más relevantes del Random forest.

Definitivamente nuestras predicciones mejoraron. Como mostramos en la figura 1, las tres variables que creamos están en el top de variables relevantes para la predicción. Esto nos da la idea de que Big Data no es una caja negra que te da buenas predicciones, por el contrario, para mejorar las

Figura 1: Variables más relevantes del Random Forest (segundo mejor modelo)



predicciones, es necesario incluir variables que creemos ex-ante pueden ser relevantes, como las que incluimos en este trabajo.

Por otro lado, dado el feedback del TP anterior, revisamos minuciosamente la limpieza de datos, la cual mejoramos. Además, incluimos variables relevantes. Este conjunto de estrategias hicieron que mejoráramos considerablemente nuestra predicción.

Finalmente, el porcentaje de hogares pobres predicho en la muestra de los que no respondieron es: 27.62 %, el cual es cercano al de la muestra de los que sí respondieron y a las cifras oficiales.