Maestría en Economía – Universidad Nacional de La Plata

Problem set 2 - Machine Learning

Alumnos: Ara Portentoso Francisco, Corradi Valentin, Di Placido Pedro

Año: 2024

Se puede acceder al repositorio público donde se encuentran los códigos y demás elementos asociados al trabajo a partir del siguiente link: <https://github.com/dipla70/ML-TP2.git>

INTRODUCCIÓN:

La medición de la pobreza por ingreso consiste en establecer un umbral de pobreza y una función indicadora. Si el ingreso es mayor a ese umbral, no se considera pobre al hogar (o persona) y si es menor, la función le asigna el valor 1 y se considera pobre. Sin embargo, este enfoque monetario tiene algunos problemas prácticos. Sin ser exaustivos con estos problemas, podemos mencionar a Fitzpatrick et al. (2018) sostienen que las encuestas nacionales de hogares que recolectan datos de consumo o de ingreso en pos de generar indicadores de pobreza suelen ser complejas. A su vez McBridey Nichols (2018) también sugieren que las encuestas de ingresos y gastos para identificar individuos en condición de pobreza suelen ser más costosas y que su realización demanda mucho tiempo.

Estos inconvenientes en la medición convencional de la pobreza por ingreso plantean el interés en tener una alternativa que sea más rápida y menos costosa, que a su vez cumpla con el objetivo de predecir la pobreza por ingreso observando otras variables. Esto se puede lograr utilizando técnicas de machine learning. Como algunos antecedentes de trabajos con este objetivo podemos mencionar a Solis-Salazar y Madrigal-Sanabria (2022) “*A machine learning proposal to predict poverty*”. En este trabajo se analizan diferentes perspectivas para predecir hogares en la pobreza utilizando un modelo de aprendizaje automático basado en XGBoost. En el concluyen que al menos uno de los métodos usados dan mejor balance entre errores de inclusión y exclusión. Sostienen que el mejor modelo utiliza un enfoque de clasificación.

También está el trabajo de Quin Li et. Al. (2022) “*Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan*”. Allí aplican también el algoritmo XGBoost para predecir pobreza a nivel hogar. Comparan esta metodología con GLM (Generalized Linear Model), donde encuentran que el primero hace un mejor trabajo que el segundo en la mayoría de los casos. Exponen que ML es útil para la selección de variables, mientras que XGBoost es especialmente preferible para variables a priori.

Luego, está el trabajo de Hamzan et. Al (2022) “*Poverty prediction using machine learning approach*” donde postula también como alternativa utilizar el enfoque de machine learning como solución para estimar de manera rápida y eficiente la pobreza, y brindar soluciones. El modelo planteado por ellos utiliza diferentes modelos de ML en donde encuentran que el Random Forest Regressor es el mejor preductor, con un score de 0.9462.

Otro trabajo más reciente que va en este camino es el de Hasssan et. Al. (2024) “*Machine learning study using 2020 SDHS data to determine poverty determinants in Somalia*”. En el buscan aplicar métodos de ML para predecir la pobreza en Somalía. Utilizan Random Forest, Decision Tree, Support vector machine y logistic regression. Encuentran que la mejor performance la tiene el modelo de Random Forest, con un accuracy del 98,36%.

**DATOS**

El objetivo del trabajo es estimar la pobreza a nivel hogar. Para ello se utilizan dataframes procesados a partir de la base de microdatos provista por el Departamento Administrativo nacional de Estadística (DANE) de Colombia. Esos microdatos provienen de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018”. Esta base existe para personas y para hogares. Si bien el objetivo del trabajo es estimar la pobreza para estos últimos, como podemos emparejar las personas con el hogar al que pertenecen, se puede extraer información valiosa de la base de personas que puede ayudarnos a predecir la pobreza del hogar.

Dado que la muestra utilizada ya está dividida en train set y test set, el primer filtro que se le hace a los datos es que las variables seleccionadas estén en ambas muestras. Las base de entrenamiento original tenía más variables que la base de test. No podríamos entrenar el modelo con variables con las que luego no contemos para testearlo.

**BIBLIOGRAFÍA**

Fitzpatrick, C., Bull, P., y Dupriez, O. (2018). Machine learning for poverty prediction: A comparative assessment of classification algorithms.

McBride, L. y Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*

Solís-Salazar, M. y Madrigal-Sanabria, J. (2022). A machine learning proposal to predict poverty. *Tecnología en Marcha. 35*(4).

Li, Q., Yu, S., Échevin, D. y Fan, M (2022). Is poverty predictable with machine learnin? A study of DHS data from Kyrgyztan. *Socio-Economic Planning Sciences (81).*

Hamzah, S., Min, P., Gan, Y., Ong, T. y Sayeed, S. (2022). Poverty prediction using machine learning approach. *Journal of Southwest Jiaotong University.* 57(1).

Hassan, A.A., Muse, A.H. y Chesneau, C. (2024). Machine learning study using 2020 SDHS data to determine poverty determinants in Somalia. *Nature.*