Objetivo del trabajo: predecir la pobreza a nivel *hogar*

La data esta dividida en 4: hogares e individuos; train y test. Variable id matchea.

Lo podemos encarar como un problema de clasificación (entre pobre y no pobre) utilizando por ejemplo arboles, o como un problema de predecir ingreso y luego con eso ver si es pobre o no.

**INTRODUCCION:**

Como antecedente podemos mencionar el trabajo de 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 soliciones. 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%.

Sin ser exhaustiva la revisión de la literatura planteada supra, se busca evidenciar la presencia de los métodos de Machine Learning en los temas que plantea el problem set. Vale aclarar, estos trabajos citados utilizan también datos de encuestas de hogares, como es el caso que debemos resolver en el práctico.

**DATA:**

**La base test\_hogares cuenta con las siguientes variables:**

Clase: 1 cabecera, 2 resto (categorica)

Dominio: Cada una de las divisiones (character)

p50000: Incluyendo sala-comedor ¿de cuántos cuartos en total dispone este hogar? (entero)

p5010: ¿En cuántos de esos cuartos duermen las personas de este hogar? (entero)

p5090: La vivienda ocupada por este hogar es (6 categorias: propia totalmente pagada, propia la estan pagando, en arriendo, en usufructo, posesión sin titulo, otro)

p5100: ¿cuanto pagan mensualmente por cuota de amortización? (numerico)

p5130: Si tuviera que pagar arriendo por esta vivienda, ¿cuánto estima que tendría que pagar mensualmente? (numerico)

p5140: ¿Cuanto pagan mensualmente por arriendo? (numerico)

Nper: personas en el hogar (entero)

Npersug: numero de personas en la unidad de gasto (entero)

Li: linea de indigencia (numerico)

Lp: linea de pobreza (numerico)

Fex\_c: Factor de expansión anualizado (numerico)

Depto: departamento (character)

Fex\_dpto: factor de expansión departamental (numerico)

**La base test\_personas tiene las siguientes variables:**

**Base test personas:**

Orden: nro que identifica a cada persona dentro del hogar (entero)

Clase: 1 cabecera, 2 resto (categorica)

Dominio: Cada una de las divisiones (character)

p6020: sexo (categorica)

p6040: edad (entero)

p6050: parentesco con el jefe del hogar (categorica)

p6090: Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud? (categorica)

p6100: A cual de los siguientes regímenes de seguridad social en salud está afiliado (categorica)

p6210: ¿cual es el nivel educativo más alto alcanzado y el ultimo año aprobado en este nivel? (categorica)

p6210s1: Grado escolar aprobado (numerico)

p6240: ¿En que actividad ocupo la mayor parte del tiempo la semana pasada? (categorica)

oficio: ¿que hace en este trabajo? (character)

p6426: ¿Cuanto tiempo lleva trabajando en este lugar de manera continua?

p6430: posicion ocupacional de la primera actividad (categorica)

p6510: el mes pasado recibió ingresos por horas extras (categorica)

p6545: el mes pasado recibió primas? (categórica.)

p6580: el mes pasado recibió bonificaciones? (categorica)

p6585s1: el mes pasado recibió auxilio o subsidio de alimentación?

p6585s2: el mes pasado recibió auxilio o subsidio de transporte?

P6585s3: el mes pasado recibió subsidio familiar?

P6585s4: el mes pasado recibió subsidio educativo?

P6590: ademas del salario, recibió alimentos como parte de pago en su trabajo?

P6600: ademas del salario, recibió vivienda como parte de pago en su trabajo?

P6610: utiliza normalmente transporte de la empresa para desplazarse a su trabajo?

P6620: ademas del salario, recibio otros ingresos en especie por el trabajo?

P6630s1: En los ultimos 12 meses cuanto recibio de prima de servicios

P6630s2: cuanto de prima de navidad

P6630s3: prima de vacaciones

P6630s4: viaticos permanentes

P6630s6: bonificaciones anuales

P6800: cuantas horas a la semana trabaja normalmente

P6870: cuantas personas en total tiene el lugar donde trabaja

P6920: está cotizando a un fondo de pensiones?

P7040: ademas de la ocupacion principal, ¿trabajó la semana pasada en otro trabajo o negocio?

P7045: Cuantas horas trabajo en ese segundo trabajo?

P7050: ocupacion de la segunda actividad (categorica: obrero, empleado, etc.)

P7090: ademas de las horas que trabaja, ¿quiere trabajar mas horas?

P7110: Durante las ultimas 4 semanas, hizo diligencias para trabajar mas horas?

P7120: si la semana pasada tenía la posibilidad de trabajar más horas, ¿estaba disponible para hacerlo?

P7150: durante las ult. 4 sem. ¿hizo diligencias para cambiar de trabajo?

P7160: Si le resultara un nuevo trabajo o empleo a...¿podría empezar a desempeñarlo antes de un mes?

P7310: ha buscado trabajo por primera vez o había trabajado antes por lo menos durante dos semanas consecutivas?

P7350: En este último trabajo era: (tipo de trabajador)… (Desocupados)

P7422: ¿Recibió o ganó el mes pasado ingresos por concepto de trabajo?. (Desocupados) (numerico)

P7472: ¿Recibió o ganó el mes pasado ingresos por concepto de trabajo?. (Desocupados) (categorico)

P7495: el mes pasado, ¿recibió pagos por concepto de arriendos y/o

pensiones?

P7500s2: ¿El mes pasado recibió pagos por b. pensiones o jubilaciones por vejez,invalidez o sustitución pensional ?

P7500s3: ¿El mes pasado recibió pagos por c. pensión alimenticia por

paternidad, divorcio o separación?

P7505: Durante los últimos doce meses, ¿recibió dinero de otros hogares, personas o instituciones no gubernamentales; dinero por intereses, dividendos, utilidades o por cesantias?

P7510s1: Durante los últimos 12 meses, ¿recibió a. dinero de otros hogares o personas residentes en el país?

P7510s2: Durante los últimos 12 meses, ¿recibió b. dinero de otros hogares o personas residentes fuera del país?

P7510s3: Durante los últimos 12 meses, ¿recibió c. ayudas en dinero de instituciones del país?

P7510s5: Durante los últimos 12 meses, ¿recibió d. dinero por intereses de prestamos o CDT´s, depositos de ahorros, utilidades, ganancias o dividendos por inversiones?

P7510s6: Durante los últimos 12 meses, ¿recibió e. dinero por concepto de cesantías y/o intereses a las cesantías?

P7510s7: Durante los últimos 12 meses, ¿recibió f. dinero de otras fuentes diferentes a las anteriores?

Pet: Población en edad de trabajar 1: sí 0: no

Oc: Ocupado. 1: si

Des: Desocupado. 1: si

Ina: inactivo. 1: si

Fex\_c: Factor de expansion anualizado

Depto: departamento. (character)

Fex\_dpto: factor de expansion anualizado departamento