# Problem Set N°2 - Machine Learning
# UNLP

Cantero, Lara Sofía
Temossi, Francisco
Vollmer, Candela

Repositorio de GitHub

## 1)

### CONTEXTO

Este problema se inspiró en una competencia reciente organizada por el Banco Mundial: Pruebas Pover-T: Predicción de la pobreza. La idea es predecir la pobreza en Colombia. Como se afirma en la competencia, "medir la pobreza es difícil, lleva tiempo y es costoso. Al construir mejores modelos, podemos realizar encuestas con menos preguntas y más específicas que midan de manera rápida y económica la efectividad de nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, con mayor precisión podremos orientar las intervenciones e iterar sobre las políticas, maximizando el impacto y la relación costo-efectividad de estas estrategias".

El objetivo principal es construir un modelo predictivo de la pobreza de los hogares. Nótese que un hogar se clasifica como:

$$Pobre = I(Inc < PL)$$

donde $I$ es una función indicadora que toma uno si el ingreso familiar está por debajo de una cierta línea de pobreza.

Esto sugiere dos maneras de predecir la pobreza. Primero, abordarla como un problema de clasificación: predecir ceros (no pobre) y unos (pobre). Segundo, como un problema de predicción de ingresos. Con los ingresos previstos, se puede utilizar la línea de pobreza y obtener la clasificación. Se explorarán ambas rutas en este trabajo.

### INTRODUCCIÓN

VER ANTECEDENTES

Los datos utilizados se obtuvieron del reporte "Medición de Pobreza Monetaria y Desigualdad" que extrae información de las bases de microdatos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 que elabora el Departamento Administrativo Nacional de Estadística (DANE) y de la misión Empalme de las Series de Empleo, Pobreza y Desigualdad - MESEP. Los datos contienen cuatro conjuntos divididos en conjuntos de datos de capacitación y de prueba a nivel de hogar e individuo.

La Gran Encuesta Integrada de Hogares es una encuesta que recolecta información sobre las condiciones de vida de los habitantes de Colombia (condiciones de la vivienda,

educación, nutrición, composición demográfica del hogar, tenencia de activos, etc.), además de las características generales de la población como sexo, edad y estado civil. La GEIH proporciona información a nivel nacional, cabecera - resto, regional, departamental, y para cada una de las capitales de los departamentos.

Las características más importantes de esta encuesta son:

- Unidad de observación: Hogares.
- Muestra (total anual): 240.000 hogares, aproximadamente.
- Periodicidad: trimestral y anual.
- Temas: algunos son ingresos, condiciones sociales, nutrición, propiedad, ahorro/inversión, empleo y desempleo, situación económica, economía rural, vivienda, fertilidad y uso de servicios públicos.
- Cobertura: cobertura nacional que permite obtener resultados por zona urbana y rural, cinco grandes regiones (Región Atlántica, Región Oriental, Región Central, Región Pacífica y Región Bogotá) y total por departamentos: 24 departamentos, las 13 grandes ciudades con sus áreas metropolitanas y 11 ciudades intermedias.
- Universo: población civil no institucional, residente en todo el territorio nacional.

Una de las principales ventajas de esta fuente de datos es la amplitud de los aspectos relevados, que abarcan una variedad de temáticas demográficas y socioeconómicas. Esto la convierte en una herramienta particularmente valiosa para analizar cuestiones relacionadas con la pobreza e indicadores sociales y distribución del ingreso, lo que la hace especialmente adecuada para la construcción de un modelo de predicción de pobreza.

Los resultados obtenidos al usar una muestra de entrenamiento del 30 % de la muestra total indican que el mejor modelo para predecir pobreza .....

## DATOS

En el presente trabajo la predicción de la pobreza se llevará a cabo de dos maneras distintas. Primero, se abordará como un problema de clasificación, es decir, la predicción de la pobreza como variable binaria; y segundo, como un problema de predicción de ingresos, donde con los ingresos previstos y la línea de pobreza se obtendrá la clasificación.

Para la primera parte se seleccionaron 7 variables. Es importante aclarar que se va a estar midiendo pobreza multidimensional, no solo monetaria, por lo que se incluyeron variables proxys de educación, salud y nivel de vida. La variable monetaria considerada es el ingreso total de la unidad de gasto antes de imputación de arriendo a propietarios y usufructuarios; mientras que las variables proxys incluídas son la educación máxima alcanzada por los miembros del hogar, la clase del hogar (cabecera o resto), los regímenes de seguridad social en salud al que están afiliados los miembros del hogar (subsidiado u otros), si reciben subsidio familiar, si reciben subsidio alimentario y el número de personas que duermen en una habitación.

Existe consenso en que en el cálculo de la pobreza monetaria intervienen dos elementos: el valor de la línea de pobreza y el ingreso de la unidad de gasto, donde cada uno de estos debe hacer referencia a los mismos rubros. En este caso, como dentro de los rubros de gasto considerados para la construcción de línea de pobreza se incluye un monto por arriendo imputado, es necesario que dentro del ingreso de los hogares también se considere

la inclusión del monto por imputación de arriendo para propietarios con el fin de garantizar consistencia al momento de estimar.

Ahora, como en el presente trabajo se está estimando pobreza monetaria se agregaron variables proxys de otras posibles dimensiones que afecten en la determinación de si una personas es pobre o no, como por ejemplo educación, salud, condiciones de vida, hacinamiento, etc. Como proxy de educación, se utilizó el máximo nivel educativo alcanzado por los miembros del hogar. Esta variable asigna a todos los miembros del hogar el mismo nivel educativo del individuo más educado. Como proxy de salud, se usó a cuál regímenes de seguridad social en salud están afiliados los miembros del hogar. Esta variable asigna un valor 1 a aquellos individuos que estén bajo un régimen subsidiado y un valor 0 a aquellos que estén bajo cualquier otro régimen (contributivo o especial). Como proxy de condición de vida se utilizaron varias variables categóricas: la tenencia o no de algún subsidio familiar y/o alimentario (valor 1 a los que si tenían-valor 0 a los que no) y la clase a la que pertenecía el hogar. Esta última variable hace referencia a si el hogar se encuentra en la cabecera del departamento/municipio o si se encuentra en cualquier otro lugar. Y por último, como proxy de hacinamiento se utilizó la variable de número de personas que duermen en una habitación. Ésta se creó en función de otras dos variables referentes a la cantidad de habitaciones donde duermen los integrantes del hogar y el número de habitantes de ese hogar.

Para la segunda parte se seleccionaron 7 variables. Las variables numéricas consideradas son la edad, la cantidad de horas trabajadas[1] y la antigüedad laboral. También se han incluido variables categóricas como el sexo, la condición respecto al sistema de pensiones, el máximo nivel de educación y el tipo de ocupación.

Estas variables se consideran pertinentes por varias razones. Con respecto a la edad, influye en la experiencia laboral y el nivel salarial. Con respecto a la cantidad de horas trabajadas, afecta directamente los ingresos. Con respecto a la antigüedad laboral, puede influir en los ingresos a través de aumentos salariales y promociones basadas en la experiencia y el tiempo en la empresa. Con respecto al sexo, es relevante porque existen diferencias salariales significativas entre hombres y mujeres en muchos sectores laborales, por lo que incluirla yuda a capturar estas disparidades en el modelo. Con respecto a la condición respecto al sistema de pensiones, las personas que están inscritas en sistemas de pensiones pueden tener ingresos diferentes a las que no lo están, debido a las diferencias en beneficios y contribuciones. Con respecto al máximo nivel de educación, es uno de los factores más importantes en la predicción de ingresos, ya que generalmente, a mayor nivel educativo, mayores son las oportunidades de empleo y salarios. Y con respecto al tipo de ocupación, se debe a que diferentes ocupaciones tienen diferentes niveles salariales.

[1]Se considera el total de las horas trabajadas la semana previa a la encuesta.