*PROBLEM SET 2*

*“Las guerras de las naciones se libran para cambiar los mapas. Pero las guerras de la pobreza se libran para mapear el cambio” M. Ali*

***Introducción***

expone brevemente el problema y sus antecedentes, una descripción de los datos y su idoneidad para el problema en cuestión. Contiene una vista previa de los resultados y las conclusiones principales)

El siguiente trabajo está inspirado en una competencia reciente organizada por el banco mundial: Pover-Tests: Predicting Poverty, el cual consistía en predecir la pobreza de forma eficiente. Como afirma la competencia,

“*medir la pobreza es difícil, requiere mucho tiempo y es costoso. Al construir mejores modelos, podemos realizar encuestas con menos preguntas y más específicas que miden de manera rápida y económica la efectividad de las nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, con mayor precisión podremos orientar las intervenciones e iterar las políticas, maximizando el impacto y la rentabilidad de estas estrategia*s”.

El objetivo de este trabajo es predecir la pobreza a nivel de los hogares para Colombia, a través del uso de datos proporcionados a nivel de hogar e individual, con el propósito de que se pueda medir la pobreza de una manera rápida y económica, al construir un modelo que use la menor cantidad de variables para su predicción.

Los datos utilizados para este análisis contienen cuatro conjuntos de datos de la Gran Encuesta Integrada de Hogares (GEIH) para el año 2018. Dicha muestra está dividida entre la muestra de entrenamiento y la de prueba. La descripción de las variables está disponible en los anexos de este trabajo.

El orden de este trabajo será primero esta introducción, seguida de la presentación de los datos utilizados para el ejercicio de predicción y finalizando con la presentación de las especificaciones y modelos utilizados para las tareas predictiva y los resultados.

***Datos***

Trate esta sección como una oportunidad para presentar una narrativa convincente para justificar o defender sus opciones de datos y ayudar al lector a comprender sus datos y su variación. Utilice su conocimiento profesional para agregar valor a esta sección. No la presente como una lista “seca” de ingredientes.

Datos de empleo

Condiciones e indicadores sociales

Ingresos, propiedad, ahorro

***Modelo***

*Esta sección presenta las especificaciones y modelos utilizados para las tareas predictivas.*

Dado que el objetivo principal es construir un modelo predictivo de la pobreza en los hogares colombianos, entonces los hogares serán clasificados como

donde es una función indicadora que toma valores de 1 si el ingreso del hogar está por debajo de cierta línea de pobreza.

Esto sugiere dos formas de predecir la pobreza, en primer lugar, a partir del enfoque como un problema de clasificación: donde se debe asignar ceros para hogares no pobres y unos para hogares pobres. En segundo lugar, como un problema de predicción de ingresos, donde a partir de los ingresos predichos, se pueda obtener la clasificación haciendo uso de la línea de pobreza. En este trabajo se explorarán ambas rutas.

Existen diferentes métodos para llevar a cabo la identificación de la población pobre. Sen (1979) diferencia entre dos métodos generales. El primero de ellos es el método directo. A través de éste se identifica el conjunto de hogares (o personas) que no satisfacen un **grupo específico de necesidades previamente establecidas (condiciones de la vivienda, educación, composición demográfica del hogar, tenencia de activos, etc**.). El segundo se denomina método indirecto. Es este caso la identificación se lleva a cabo mediante el **cálculo de un umbral mínimo, por lo general asociado a un nivel de ingreso o gasto, por debajo del cual se considera que una persona no puede satisfacer sus necesidades básicas**. Mientras que el primero se basa en la satisfacción efectiva de un conjunto de necesidades, el segundo hace referencia a la posibilidad de satisfacerlas (considerando para ello un nivel mínimo de ingresos).

Como ejemplos del método directo se cuentan el Índice de Necesidades Básicas Insatisfechas (NBI) y el nuevo Índice de Pobreza Multidimensional (IPM). La medición de pobreza monetaria a través de la definición de unas líneas de pobreza e indigencia se enmarca dentro del método indirecto, y fue producto de actualización y revisión metodológica por parte de la MESEP.

Para la medición de la pobreza monetaria se requieren dos elementos principalmente: el valor de las líneas de pobreza y el ingreso per-cápita disponible de los hogares. La línea de indigencia corresponde al valor de una canasta básica de alimentos y la de pobreza al valor de una canasta que incluye además de alimentos otros bienes básicos. Esta medición consiste en determinar cuántos hogares carecen de ingresos suficientes para comprar dichas canastas.

Dentro del método indirecto de medición de pobreza existen dos alternativas metodológicas para la definición del nivel mínimo de recursos a partir del cual se supone la satisfacción de las necesidades básicas: el método del Consumo Calórico y el método del Costo de las Necesidades Básicas. A continuación se presenta una breve descripción de estos dos métodos, y adicionalmente, se reseña la metodología aplicada por la CEPAL para la definición de líneas de pobreza de carácter presupuestal.

**MÉTODO DEL CONSUMO CALÓRICO (FOOD ENERGY INTAKE)**

El método del consumo calórico o costo energético tiene por objetivo la especificación del nivel de gasto/ingreso total que permitiría la consecución de los requerimientos nutricionales mínimos. El supuesto básico es que a medida que crece el ingreso (gasto) también aumenta el consumo energético, aunque a una tasa decreciente. Para obtener la línea de pobreza se establece el punto de la función ingreso calórica en el cual se satisfacen los requerimientos nutricionales. Dicho punto tendrá asociado un monto de ingreso (gasto) equivalente al valor de la línea de pobreza (Z). Es importante mencionar que el método considera implícitamente el consumo alimentario y no alimentario. Esto debido a que el eje x representa el gasto total. Este aspecto resulta de gran importancia al analizar la consistencia de las líneas obtenidas a partir de este método.

A pesar de la ventaja que este método ofrece al requerir menor información (no es necesario diferenciar espacialmente los precios ni distinguir el componente no alimentario de la línea de pobreza), las líneas estimadas pueden resultar inconsistentes debido a que el nivel de gasto/ingreso no es el único determinante del consumo energético (Ravallion, 1998). Otros aspectos como las preferencias, el nivel de actividad, los precios relativos o los bienes de provisión pública también influyen en el consumo de calorías.

- Las preferencias: Los hogares urbanos tienden a consumir alimentos cuyo precio por caloría es más elevado (proteínas en lugar de carbohidratos). No se puede concluir, al comparar dos personas con gasto real similar, que aquella que consume alimentos cuyo precio por caloría es más alto sea más pobre.

- Los precios relativos: A diferencia de las zonas rurales, en las zonas urbanas los precios de algunos bienes no alimentarios tienden a ser relativamente menores que los de los alimentos. Puede ocurrir, que para un determinado nivel de ingreso, la demanda por alimentos en la zona urbana sea inferior a la de la zona rural. No significa sin embargo, que los hogares urbanos sean más pobres que los rurales.

- Nivel de actividad: Comúnmente los trabajos que se realizan en la zona urbana demandan menos calorías que los de la zona rural. Por tal razón el consumo de alimentos para un determinado nivel de ingreso tiende a ser menor en la zona urbana, sin que esto guarde relación con un síntoma de pobreza.

Cuando en la definición de las líneas de pobreza no se tienen en cuenta aspectos como los señalados anteriormente, pueden presentarse resultados inconsistentes que impiden una adecuada comparación del bienestar entre individuos. Por ejemplo, que la incidencia de la pobreza en la zona urbana supere la de la zona rural.

**MÉTODO DEL COSTO DE LAS NECESIDADES BÁSICAS (COST OF BASIC NEEDS)**

Este enfoque tiene su origen en los trabajos seminales de Rowntree (1901, 1936 y 1941). En su trabajo de 1901 el autor hace referencia a dos tipologías de pobreza, primaria y secundaria. La primera la define como la situación en la que viven las familias cuyos ingresos no son suficientes para cubrir los requerimientos mínimos para mantener la eficiencia física. La pobreza secundaria la describe como la situación de aquellas familias cuyos ingresos habrían sido suficientes para cubrir los requerimientos mínimos, de no ser porque una porción de estos se destinó a otro tipo de consumo. Ambos conceptos guardan estrecha relación con el enfoque presupuestario a partir del cual se definen las líneas de pobreza extrema y pobreza. El método del costo de las necesidades básicas, como su nombre lo indica, consiste en calcular el costo de una canasta básica de consumo. Para ello se siguen, en términos generales, los siguientes pasos:

- Definir el requerimiento nutricional del componente alimentario. Por lo general se toma como referencia el valor estimado por la FAO (cerca de 2.100 calorías diarias por persona), considerando diferencias por edad, sexo o actividad.

- Estimar el costo asociado a la consecución del requerimiento nutricional, garantizando una dieta que respete los hábitos de consumo del estrato de referencia (por ejemplo, la población del primer quintil de ingreso, o los dos primeros quintiles, o aquellos que consumen entre 2.000 o 2.200 calorías). Este será el componente alimentario igual al valor de la línea de extrema pobreza.

- Obtener el valor de la línea de pobreza. Para establecer este valor se utiliza el Coeficiente de Orshansky que refleja el inverso de la participación del gasto alimentario dentro del gasto total. Así, el valor de la línea de pobreza será la suma correspondiente al componente alimentario y no alimentario.

El método de CNB informa sobre el ingreso mínimo que ofrece la posibilidad de adquirir la canasta básica, pero sin asegurar el consumo efectivo de la misma. Indica si las personas cuentan con los recursos suficientes para comprarla independientemente de si la consumen o no. Por esta razón, en este análisis resulta indispensable contar con información acerca de los precios de los bienes de consumo de la canasta. En ausencia de éstos, los analistas recurren a la metodología del consumo calórico.

El método del costo de las necesidades básicas no está exento de las limitaciones mencionadas en la descripción del método de consumo calórico. Las diferencias atribuibles a los precios relativos, los niveles de actividad y las preferencias también pueden generar inconsistencias en los resultados. Sin embargo, a diferencia del método de consumo calórico, el método del costo de necesidades básicas sí incluye información sobre el contenido de la canasta de alimentos. Gracias a esto, es posible establecer el patrón de consumo de la población que satisface los requerimientos.

**3. ALTERNATIVAS PARA CORREGIR LAS INCONSISTENCIAS**

Ravallion propone algunas alternativas para subsanar las limitaciones anteriormente descritas. De manera particular, para corregir las inconsistencias asociadas a los precios relativos, el autor propone la construcción de un Deflactor Espacial. Con este procedimiento se puede conformar un gasto ajustado por costo de vida, expresando los precios de los diferentes dominios de análisis en función de un dominio base (por ejemplo, los precios rurales en términos de los urbanos). Una vez se tiene el gasto ajustado por costo de vida se puede seleccionar el estrato de referencia para observar su pauta de consumo. Una población de referencia definida de esta manera, en cierta medida, no estará afectada por las diferencias en precios.

Como se mencionó, el método del consumo calórico no cuenta con información de precios recolectada directamente en la fuente de información para construir la línea. En ese caso se puede recurrir a una fuente secundaria para deflactar el gasto total. Para el método del costo de las necesidades básicas la información de precios está disponible directamente. En ambos casos las diferencias en precios relativos se pueden reducir en lo posible a través del Deflactor Espacial de Precios.

De otra parte, como lo reconocen algunos estudios, las diferencias atribuibles a la edad, el sexo, el clima o el nivel de actividad también influyen de manera importante en la pauta de consumo de los hogares: "El ciclo de vida familiar afecta de manera significativa la proporción de gasto destinada, por ejemplo, a transporte, educación, salud y vivienda".Las propuestas en este sentido pueden ser: i) considerar requerimientos nutricionales de acuerdo con el tipo de labor que desarrollan las personas y según otras características demográficas (edad y sexo) y/o ii) utilizar escalas de equivalencia.

Una escala de equivalencia es un índice que permite expresar el costo relativo de adquisición de un conjunto de bienes y servicios en términos de un hogar tipo (según su tamaño y composición). "El concepto de escalas de equivalencia agrupa dos elementos simultáneamente. Por un lado, la escala considera las diferentes necesidades de los miembros del hogar, según edad, género, u otras características demográficas o de tipo de actividad. Por otro, este índice permite tomar en cuenta la existencia de economías de escala, caracterizadas por costos marginales decrecientes para alcanzar un mismo nivel de bienestar ante la adición de un nuevo miembro al hogar. Este último punto está relacionado con la existencia de bienes públicos en el hogar, que pueden ser compartidos sin que haya una reducción del bienestar de las personas".

En relación con las preferencias que caracterizan el consumo de diferentes regiones (urbano/rural), se pueden establecer pautas de consumo geográficamente diferenciadas. "Un método simple es establecer un conjunto de bienes en cada región (por ejemplo) que es el consumo medio de un grupo de referencia establecido a nivel nacional en términos de sus ingresos o gastos". Esta propuesta establece una definición única de la población de referencia (estrato de referencia nacional) y a partir de éste se selecciona el conjunto de bienes que caracterizan las preferencias de cada región (con base en la pauta de

* ***Modelos de clasificación***

- Esta subsección describe el enfoque de clasificación, es decir, su intento de predecir directamente ceros (no pobre) y unos (pobre)

- Una comparación con al menos otras 5 especificaciones/modelos en términos de poder predictivo para la tarea en cuestión.

– Una descripción de las variables utilizadas en el modelo y discutir su importancia relativa en la predicción.

– Una descripción de cualquier estrategia de submuestreo utilizada para abordar el desequilibrio de clases

* ***Modelos de predicción de ingresos***

– Una explicación detallada del modelo final elegido. La explicación debe incluir cómo se entrenó el modelo, la selección de hiperparámetros y otra información relevante.

– Una comparación con al menos otras 5 especificaciones/modelos en términos de poder predictivo para la tarea en cuestión.

– Una descripción de las variables utilizadas en el modelo y discutir su importancia relativa en la predicción.

– Una descripción de cualquier estrategia de submuestreo utilizada para abordar el desequilibrio de clases

***Resultados***

Juzgaré las predicciones en función de las tasas de falsos positivos, las tasas de falsos negativos y la escasez del modelo. En la puntuación final tendrán más peso los falsos negativos (75%), es decir, familias pobres clasificadas como no pobres, y cuantas más variables uses, menor será tu puntuación.

Cuantas más variables uses, menor será tu puntuación

Las transformaciones de la misma variable no cuentan como variables adicionales.

Las tablas, figuras y escritos deben ser lo más prolijos posible. Etiquete todas las variables incluidas. Si tiene algo en sus figuras o tablas, espero que se aborden en el texto.

***Conclusiones y recomedaciones***