Análisis de la clasificación de los hogares pobres en Costa Rica, según su grado de severidad de la pobreza

Juan Sebastian Paniagua, Juan Pablo Novoa, Mauricio Roverssi

Resumen

Los niveles de pobreza de Costa Rica han permanecido estancados por más de 4 décadas. Para atacar este problema, el estado costarricense ha implementado una serie de subsidios monetarios a hogares pobres, sin embargo, en febrero del 2022, un informe indicó que cerca de 200.000 personas reciben ayudas estatales, que representa el 26,5% del fondo de pobreza, pese a no calificar en condición de pobreza. Esta situación motiva a revisar si la información que se utiliza para calcular el Índice de Pobreza Multidimensional (IPM), el cual clasifica a un hogar en pobre o no pobre, puede ser objeto de mejora por medio de la aplicación de herramientas de aprendizaje no supervisado para hacer agrupamiento en clusters, permitiendo una mejor clasificación de los hogares y, en consecuencia, una mejor asignación de ayudas estatales.

Introducción

En Costa Rica, el porcentaje de hogares que no tienen el suficiente dinero para hacer frente a sus principales necesidades se ha mantenido alrededor del 20% de la población durante los últimos cuarenta años. Para atacar este problema el Estado costarricense ha implementado una serie de controles para la entrega de ayudas monetarias a los hogares vulnerables. Para el año 2021, se presupuestó destinar 581.500 millones de colones ($913 millones de dólares), para combatir la pobreza a nivel nacional. A pesar de lo anterior, la eficiencia en el uso de los recursos parece no ser la adecuada. El 24 de febrero del 2022, un medio de comunicación nacional publicó que cerca de doscientas mil personas reciben ayudas estatales pese a no calificar en condición de pobreza. El reportaje menciona que estas personas que “no calificaban como vulnerables”, obtuvieron, en 2021, 154.000 millones de colones ($240 millones de dólares), lo que representa el 26,5% del fondo de pobreza.

Por este motivo, se considera que existen puntos de mejora para la clasificación de los hogares, esto con el objetivo de poder dirigir los recursos donde existe mayor vulnerabilidad. Costa Rica ya ha dado pasos en este particular; implementó en 2015 la metodología del Índice de Pobreza Multidimensional (IPM) (Instituto Nacional de Estadística y Censos INEC, 2015). Esta medida es complementaria a la medición tradicional de los ingresos porque contempla las siguientes 5 dimensiones: educación, salud, vivienda, trabajo y protección social.

El IPM, que se mide anualmente en la Encuesta Nacional de Hogares (ENAHO) (INEC-Costa Rica, 2021) por parte del Instituto Nacional de Estadística y Censos (INEC),

cuenta con una gran cantidad de información con la que se hace la clasificación de hogares en las categorías pobre o no pobre. Esta información es objeto de análisis de este proyecto con el propósito de determinar si es posible hacer una mejor clasificación de los hogares, identificando las variables relevantes que puedan permitir una mejor distribución de los subsidios. De esta manera se podría proponer una mejora al IPM.

Materiales y Métodos

Enfoque

Aprendizaje no supervisado.

Unidad de análisis

Dimensiones asociadas al IPM de la Encuesta Nacional de Hogares (ENAHO) de julio 2021 (INEC-Costa Rica, 2021).

Recolección de los datos

El archivo de datos fue descargado desde el sitio web del INEC desde su portal después de completar un formulario de autorización del instituto. El archivo consta de 23.9 MB con extensión .sav (SPSS Statistic Data Document).

Software utilizado en el análisis

Python 3.8.8., librerías numpy, pandas, seaborn, matplotlib, sklearn, kneed, pca y gower.

Limpieza de datos

1. Carga de la base.
2. Selección de los campos relevantes asociados al IPM.
3. Eliminación de filas con datos faltantes.
4. Renombre de las columnas.
5. Configurar variables específicas a una versión adecuada (por ejemplo, edad de string a números enteros)
6. Creación de dataFrame final para el análisis.

Obtención de Estadísticas descriptivas

Generales. Se obtuvieron las estadísticas generales del dataframe como número de variables categóricas y numéricas, así como el número de observaciones.

Variables categóricas. El set de datos tiene 43 variables categóricas.

Variables numéricas. Promedio, deviación estándar, dato mínimo, dato máximo, percentiles 25%, 50% y 75%. Además, se construyó un correlograma para observar las tendencias.

Análisis de los datos

Para el análisis de los datos, primero codificamos con Ordinal Encoder las dimensiones categóricas que indican nivel o prioridad y con Label Encoder las demás dimensiones categóricas. Luego creamos dos variaciones del set de datos:

1. Set de datos Variación 1: todas las dimensiones escaladas. Las variables categóricas se codificaron con etiquetas de números enteros.
2. Set de datos Variación 2: dimensiones numéricas escaladas y dimensiones categorías sin escalar.
3. Set de datos Variación 3: dimensiones numéricas escaladas y dimensiones categorías con las etiquetas de texto del set de datos original

Aplicamos los algoritmos Kmedias, Kmedoides, Clustering Jerarquíco Aglomerativo y DBSCAN a las Variaciones 1 y 2. Repetimos el procedimiento anterior pero esta vez realizamos un PCA para reducir las dimensiones de la Variación 1 y la Variación 2 antes de aplicar los algoritmos. Así mismo, aplicamos SVD, para comparar los resultados con respecto al PCA, y reducir los valores singulares de la variación 1 y 2. Para la Variación 3, se aplicó el algoritmo Kmedoides con distancia Gower. Calculamos la precisión de cada algoritmo asumiendo la calificación pobre/no pobre del IPM como el resultado esperado del agrupamiento y luego se compara con los clusters que arroja cada modelo. Con esta estrategia se espera identificar si hay hogares que pudiesen estar mal clasificados.

Resultados y Discusión

En el Anexo 1 encontramos una tabla con los descriptores completos del set de datos a analizado. En el Anexo 2 Tabla 1 se encuentra un resumen de las principales características de las dimensiones cualitativas.

En el Anexo 3, Tabla 2 se encuentra un resumen de los estadísticos descriptivos de las dimensiones cuantitativas. Es importante notar que y teniendo presente que a la fecha de la encuesta, la pobreza está siendo un problema a futuro, que afecta todas las edades, con un promedio de personas de edad de 36.5 años, edad en que normalmente la persona ya tiene una familia, con una desviación estándar de 21.97 años, valor que coge prácticamente a una muestra muy importante de la población, con una cantidad de habitantes por vivienda de 4 personas, con unas transferencias por hogar muy bajas, y una canasta familiar alta. Por otra parte, es importante resalta que, la dimensionalidad de la medida de cada variable puede

afectar de una manera importante la varianza del ser de datos, por lo que es recomendable escalar los datos para hacer una buena práctica para clusterizar.

Del análisis del coeficiente de Silouethe Figura 1 se concluye que la mejor calidad de agrupamiento se obtiene con 2 clusters, lo cual resulta útil para el desarrollo del proyecto dado que el IPM también hace una clasificación en dos categorías: pobre y no pobre. Con este resultado se decide analizar todos los modelos con 2 clusters. Es importante aclarar que la calificación de “pobre” y “no pobre” se retira del set de datos y no se considera en los modelos de agrupamiento.

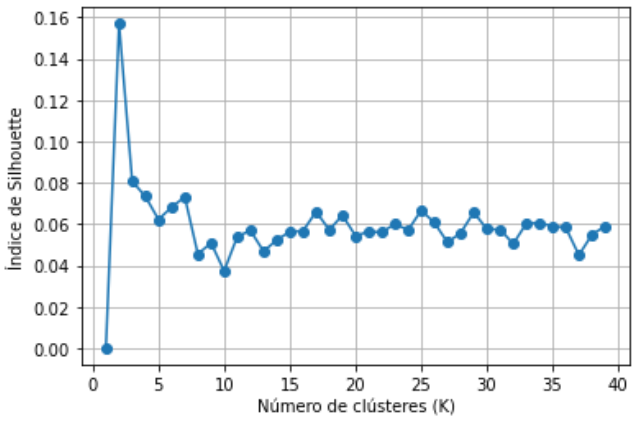

Figura 1: Resultado de coeficiente de Silouethe

En el Anexo 4 Tabla 3 presentamos los principales resultados al aplicar los diferentes algoritmos según las instrucciones descritas en la sección Análisis de los datos de Materiales y Métodos.

Debido a que los resultados obtenidos explorando la varianza con el método PCA nos dio resultados similares a los anteriores, pero a sabiendas de que la naturaleza del problema nos enmarca explorar un poco más en la varianza de los datos, implementamos SVD ya que este método puede aplicarse a cualquier matriz real, dado que en PCA el producto X’X puede generar perdidas de precisión numérica, lo que no sucede con SVD.

Por parte del PCA encontramos que, los componentes principales donde se concentra el 90% de la varianza son 33. Sin embargo, hay dos puntos a revisar, el primero es que se hace un codo en las primeras 4 variables, pero estas representan el 29% de la varianza, lo que a priori nos da indicios de poder trabajar con ellas, sin embargo, y debido a la dispersión que presentan los datos antes mencionada, notamos que en los siguientes 29 componentes principales se distribuye el restante 61% de la varianza, lo que se ve reflejado en los cálculos realizados para PCA, obteniendo valores muy similares a los resultados sin aplicar este método, con Kmedias (precisión 0.818), Jerárquico Aglomerativo con enlace Ward (precisión 0.79), DBSCAN (precisión 0.71).

Con respecto a los modelos desarrollados sobre el set de datos de la Variante 1, se obtuvo el mejor resultado con Kmedias (precisión 0,817). El modelo asocia el Cluster 0 mayoritariamente a la categoría “No pobre” (63,4% de los hogares) y el Cluster 1 a la categoría “Pobre” (18,8% de los hogares). Del 18,2% de los hogares restantes es de interés el 3,3% que están en la categoría “pobre” y que quedaron en el Cluster 0 (mayoritariamente “no pobre”) porque podrían estar mal clasificados y, en consecuencia, estarían recibiendo subsidios de manera errada.

El modelo Kmedoides utilizando el set de datos de la Variante 3 (variables categóricas) y distancia Gower tuvo un desempeño regular (precisión 0,656). El tiempo de procesamiento para calcular las distancias y luego el agrupamiento fue muy alto. Por este motivo, no se exploraron otras alternativas con esta estrategia de análisis.

El modelo de agrupamiento más preciso fue Kmedias utilizando el set de datos de la Variante 1 (ver Tabla 3). No se encontraron características que pertenecieran exclusivamente a la calificación de “pobre” o “no pobre” del IPM con respecto a los dos clusters obtenidos con Kmedias sin embargo, se identificaron características predominantes que pudieran ser las más relevantes para dar un indicio si un hogar es pobre o no pobre (Tabla 4)

Tabla 4

Características predominantes

Hogar pobre	Hogar no pobre
• Estado de techos, pisos y paredes: Malo • Sistema de eliminación de basura: lo queman • Servicio de agua: sin servicio • Seguro de salud: sin seguro	• Estado de techos, pisos y paredes: Bueno • Área de vivienda: mayor que 151 mts2 • Fuente de agua: acueducto municipal • Servicio sanitario: conectado a alcantarillado o cloaca • Nivel de educación: superior o superior con posgrado • Seguro de salud: con seguro • Ingreso per cápita: más de 128.000 • Condición de aseguramiento: asalariado

Por otra parte, de identifica que hay 1.064 hogares (3,3%) en la categoría “pobre” que quedaron en el Cluster 0, que es mayoritariamente de hogares “no pobre”. Estos hogares, en general, tienen niveles de educación intermedios, condiciones de servicio aceptables e ingresos bajos o medios.

Conclusión

Logramos obtener una lista de características que contribuyen más a la separación de los datos en pobres/no pobres (Tabla 4) utilizando el algoritmo K-medias.

Se concluye que estos hogares pudieran ser objeto de revisión en su calificación del IPM porque podrían estar siendo clasificados erróneamente, afectando la correcta distribución de subsidios que da el gobierno de Costa Rica.

Debido a la alta desviación estándar de la muestra está muy alta quiere decir que los datos están muy dispersos, así como la varianza, lo que nos habla de que en Costa Rica hay una desigualdad marcada en su población. La desigualdad de la población habla de la poca concentración de densidad en la muestra, lo que nos da indicios de que se debe indagar por desglosar o revisar la varianza en PCA y SVD.

A pesar de que la porción de varianza explicada para el 90% equivale a 33 valores singulares, se hizo una exploración con iteraciones bajando el número de cantidad de valores singulares hasta encontrar el mejor desempeño con 15 valores singulares, que corresponden al 58.48% de la varianza explicada, donde kmedias (precisión 0.8213) valor que ya supera las anteriores pruebas, sin embargo, no fue el mejor score, en las otras pruebas se obtuvo Kmedoides (precisión 0.6186), seguido por DBSCAN (precisión 0.782), el mejor score fue con Jerárquico Aglomerativo con enlace Ward (precisión 0.8312).

Finalmente, con él algoritmo jerárquico aglomerativo se obtuvo un mejor score de precisión, teniendo en cuenta el contexto de la problemática de la pobreza dimensional y la varianza alta de estas variables, se tomaron un poco más de la mitad y se comenzaron a relacionar desde la base del dendograma con el fin de ir fusionando esas variables, hasta fusionarla y tener n-2 grupos, con lo que podemos aprovechar menos valores singulares para llegar a los dos clúster que se esperan los cuales son pobreza y no pobreza, de tal manera que obtuvimos satisfactoriamente la respuesta a uno de los objetivos principales de esta problemática, realizar una revisión a las variables que correspondían a una clasificación previa de personas pobres o no pobres, con el fin de optimizar el uso de recursos para los subsidios, con el fin de refinar la manera con que esta problemática se calcula en una población.

Bibliografía

INEC-Costa Rica. (2021). *Encuesta Nacional de Hogares.*

Instituto Nacional de Estadística y Censos INEC. (2015). *Índice de Pobreza Multidimensional (IPM) - Metodología.* San José, Costa Rica: Instituto Nacional de Estadística y Censos.

Anexo 1

Descriptores de la Encuesta Nacional de Empleo que se utilizan para calcular el IPM

Descriptores	Tipo	Descripción	Unidad de medición
Zona	Categórica	Se refiere a si vive en zona rural o urbana	Rural /urbano
Est Paredes exterior	Categórica	Estado de las paredes exteriores	bueno /regular /malo
Est techo	Categórica	Estado del techo	bueno /regular /malo
Est piso	Categórica	Estado del Piso	bueno /regular /malo
Mts2 vivienda	Categórica	Metros cuadrados de la vivienda	4 categorías
Abastecimiento agua	Categórica	Como llega el agua a la vivienda	Por tubería /otras
Fuente agua	Categórica	Fuente de agua de la vivienda	Acueducto/ otros
Servicio sanitario	Categórica	Si la vivienda tiene servicio sanitario	Conectado a alcantarilla o tanque / Otros
Exclusividad Serv sanitario	Categórica	Si el servicio sanitario es solo de la vivienda	Si /No
Sistema eliminación basura	Categórica	Como eliminan la basura en la vivienda	Camión recolector /otras
Tiene computador portátil	Categórica	Si dentro de la vivienda hay computador portátil	Si /No
Tiene computador escritorio	Categórica	Si dentro de la vivienda hay computador de escritorio	Si/ No
Tiene internet	Categórica	Si la vivienda tiene servicio de internet	Si / No
Calificación vivienda	Categórica	Si la vivienda es óptima, aceptable, deficiente o insuficiente para vivir	4 niveles
Hacinamiento	Categórica	Relación entre el número de personas en una casa y el espacio en metros cuadrados	Si/ No
Personas en vivienda	Numérica	Cantidad de personas dentro de la vivienda	Cantidad personas
Nivel de instrucción	Categórica	Nivel mayor de educación de cada miembro del hogar	10 niveles
E1 No educación formal	Categórica	Si en el hogar una persona de 5 a 17 años no asiste a la educación formal	Indicador 0/1
E2 Rezago educativo	Categórica	Si en el hogar una persona de 7 a 19 años tiene 2 o más años de rezago educativo	Indicador 0/1
E3 Sin bachillerato	Categórica	Si en el hogar una persona de 18 a 25 años no tiene bachillerato de secundaria	Indicador 0/1
E4 Bajo desarrollo	Categórica	Si en el hogar los miembros entre 25 y 64 años no tienen un nivel de educación mínimo	Indicador 0/1
VUI1 Mal Est techo piso	Categórica	Si el estado del techo o piso es malo	Indicador 0/1
VUI2 Mal Est paredes ext	Categórica	Si el estado de las paredes exteriores es malo	Indicador 0/1
VUI3 Hacinamiento	Categórica	Si existe hacinamiento	Indicador 0/1
VUI3 Sin uso internet	Categórica	Si el hogar no ha usado internet	Indicador 0/1
S1 Sin seguro salud	Categórica	Si alguien mayor de 18 años no tiene seguro de salud	Indicador 0/1
S2 Sin servicio agua	Categórica	Si el hogar no puede abastecerse de agua por tubería o la obtiene de otras fuentes diferentes al acueducto	Indicador 0/1
S3 Sin eliminación excretas	Categórica	Si el hogar no posee servicio sanitario o es compartido	Indicador 0/1
S4 Sin eliminación basura	Categórica	Si el hogar no puede eliminar la basura con un camión recolector	Indicador 0/1
T1 Desempleo larga duración	Categórica	Si existe algún miembro con desempleo mayor a un año	Indicador 0/1
T2 Incumplim Salario Mínimo	Categórica	Si dentro del hogar a alguna persona asalariada se le incumpla el derecho del salario mínimo	Indicador 0/1

T2 Incumplim otros derechos	Categórica	Si dentro del hogar a alguna persona asalariada se le incumpla dos o más derechos laborales	Indicador 0/1
T3 Independiente informal	Categórica	Si existe una persona trabajadora independiente que no esté inscrita en ninguna entidad público	Indicador 0/1
PS1 Infancia sin cuido	Categórica	Si dentro del hogar un infante de 0 a 4 años no tiene cuido	Indicador 0/1
PS2 Adulto mayor sin pension	Categórica	Si dentro del hogar una persona mayor a 65 años no tiene pensión	Indicador 0/1
PS3 Con Discapacidad sin transf.	Categórica	Si en el hogar una persona con discapacidad no recibe transferencias monetarias	Indicador 0/1
PS4 No labora obligación familiar	Categórica	Si en el hogar alguna persona no trabaja por obligaciones familiares	Indicador 0/1
Decil ingreso per cápita hogar	Categórica	Clasificación por deciles del ingreso por persona de los hogares	Deciles
Ingreso neto hogar	Numérica	Ingreso neto en colones del hogar	Colones
Total Transf Monetaria Hogar	Numérica	Cantidad de transferencias monetarias que recibe el hogar	Colones
Transf no monetaria hogar	Numérica	Cantidad de transferencias no monetarias que recibe el hogar	Colones
Sexo	Categórica	Sexo de la persona	Hombre /Mujer
Edad	Numérica	Edad de la persona	
Primera mencion discapacidad	Categórica	Si la persona tiene discapacidad y que tipo es	Tipo de discapacidad
Recibió ayuda IMAS	Categórica	Si la persona recibió subsidios del IMAS y tipo	Tipo de subsidio
Primera mencion Cen-Cinai	Categórica	Si el infante asiste al CEN-CINAI (Centro de educación y nutrición)	Si /No
Condicion aseguramiento	Categórica	Si la persona tiene seguro de salud y por cuál tipo	Tipo de seguro
Asistencia a educación formal	Categórica	Si la persona asiste a la educación formal	Si / No
Canasta Básica Alimentaria	Númerica	Equivalente en colones de la canasta básica alimentaria	Colones
Nivel de pobreza	Categórica	Si la persona se encuentra en pobreza, extrema, pobreza no extrema o no pobre	3 niveles
Índice Pobreza Multidim	Categórica	Si la persona se encuentra en pobreza multidimensional o no	2 niveles

Anexo 2

Tabla 1

Descripción de las dimensiones cualitativas

Variable	Descriptores	Absolutos	Porcentaje
Distribución Rural y Urbano	Rural	21610	0.68
	Urbano	10217	0.32
Est Paredes exterior	Bueno	17421	0.55
	Regular	11190	0.35
	Malo	3216	0.10
Est Techo	Bueno	17179	0.54
	Regular	10584	0.33
	Malo	4064	0.13
Est Piso	Bueno	19045	0.6
	Regular	3132	0.1
	Malo	9650	0.3

Anexo 3

Tabla 2

Estadísticos descriptivos de las dimensiones cuantitativas

Estadístico	Personas en vivienda	Transferencias monetarias por hogar	Transferencias no monetarias	Edad	Canasta básica
Promedio	3.930122	1.370e+05	3245.3	36.57	48598.9
Desviación	1.781404	3.085e+05	23469.7	21.97	3938.5
Valor mínimo	1.0	0	0	0	42871.0
Valor máximo	17.0	9.2628e+06	962000	97.0	51307.0
Percentil 25%	3.0	0	0	18.0	42871.0
Percentil 50%	4.0	3.00e+04	0	35.0	51307.0
Percentil 75%	5.0	1.500e+05	0	54.0	51307.0

Anexo 4

Tabla 3

Resultados cada algoritmo

Algoritmo	PCA/SVD	Codificador	Escalado	Precisión
Set de datos Variante 1				
Kmedias	No	Ordinal encoder y Label encoder con etiquetas de números enteros	Todas las variables (numéricas y codificadas)	0,817
Kmedoides				0,340
Jerárquico Aglomerativo con enlace Ward				0,798
DBSCAN				0,623
Kmedias	No	Ordinal encoder y Label encoder con etiquetas de números enteros	Solo variables numéricas	0,629
Jerárquico Aglomerativo con enlace Ward				0,693
DBSCAN				0,783
Kmedias	15 SVD	Ordinal encoder y Label encoder con etiquetas de números enteros	Todas las variables (numéricas y codificadas)	0.821
Kmedoides				0.618
Jerárquico Aglomerativo con enlace Ward				0.831
DBSCAN				0.782
Set de datos Variante 2				
Kmedias	34 PCA	Ordinal encoder y Label encoder con etiquetas de números enteros	Todas las variables (numéricas y codificadas)	0.818
Kmedoides				---
Jerárquico Aglomerativo con enlace Ward				0.799
DBSCAN				0.71
Set de datos Variante 3				
Kmedoides con distancia Gower	No	Label Encoder	Solo variables	0,656