Universidad de Los Andes

Integrantes del grupo

Luis Olegario Borda Silva

Julian Santiago Muñoz Garrido

Joan Sebastián Potosí Hoyos

Juan Felipe Vargas Guacheta

Contenido

1. INTRODUCCIÓN

La comprensión de las variables económicas que inciden en la pobreza de un hogar es esencial para diseñar políticas efectivas que aborden este desafío global. Diversos teóricos económicos han contribuido a esta área, brindando perspectivas valiosas que arrojan luz sobre la complejidad de los factores que perpetúan la pobreza. En este contexto, autores destacados como Amartya Sen, John Maynard Keynes y Milton Friedman han ofrecido análisis profundos que han influido en la comprensión contemporánea de la relación entre economía y pobreza.

Amartya Sen, premio Nobel de Economía, ha subrayado la importancia de evaluar la pobreza más allá de la mera falta de ingresos. Su enfoque de las "capacidades" destaca que la pobreza no solo se trata de carencias económicas, sino también de la falta de acceso a oportunidades educativas, servicios de salud y participación en la toma de decisiones. Este enfoque ampliado proporciona una base conceptual sólida para analizar las múltiples dimensiones que componen la pobreza.

Por otro lado, las teorías keynesianas, avanzadas por John Maynard Keynes, resaltan la importancia de la demanda agregada y la intervención del gobierno para estimular el crecimiento económico y reducir la pobreza. Su énfasis en la inversión pública como motor de desarrollo económico ofrece una perspectiva crucial en la evaluación de políticas económicas que buscan mejorar las condiciones de vida de los hogares en situación de pobreza.

En contraste, la perspectiva de Milton Friedman, defensor del liberalismo económico, destaca la importancia de la libertad individual y la reducción de la intervención gubernamental para fomentar el crecimiento económico. Aunque sus ideas han sido objeto de debates, su enfoque en la eficiencia del mercado y la minimización de la burocracia ofrece un contrapunto valioso en el análisis de políticas orientadas a combatir la pobreza.

La relevancia de realizar un análisis exhaustivo de las variables económicas que afectan la pobreza de un hogar es innegable. En un mundo interconectado, donde la desigualdad persiste como uno de los mayores desafíos, comprender cómo las fuerzas económicas impactan en la distribución de recursos es esencial para informar decisiones políticas y estrategias de desarrollo. Además, un enfoque multidimensional, inspirado en las ideas de Sen, permite una comprensión más holística de la pobreza, reconociendo que las soluciones efectivas deben abordar no solo los aspectos económicos, sino también las barreras sociales y estructurales que perpetúan la privación.

En conclusión, la amalgama de teorías económicas proporciona un marco sólido para evaluar las variables económicas que impactan en la pobreza de un hogar. Este análisis es esencial para diseñar estrategias efectivas que aborden las complejidades de la pobreza y promuevan un desarrollo inclusivo y sostenible. La comprensión de estas variables no solo informa la toma de decisiones políticas, sino que también contribuye al avance del conocimiento en la lucha contra la pobreza en un mundo en constante cambio.

En este contexto, el presente documento desarrolla un ejercicio para predecir si un individuo es pobre o no de acuerdo con la línea de ingresos tanto predicha como establecida que determina desde que nivel un individuo es pobre haciendo uso de diferentes técnicas de aprendizaje de máquinas. La información utilizada proviene del informe de Medición de Pobreza Monetaria y Desigualdad del año 2018, realizado por el DANE de la Gran Encuesta Integrada de Hogares (GEIH). Esta encuesta proporciona información estadística sobre el tamaño y estructura de la fuerza de trabajo (empleo, desempleo y población fuera de la fuerza de trabajo), los ingresos laborales y no laborales de los hogares, la pobreza monetaria y la pobreza monetaria extrema de la población residente en el país (DANE, 2023). Las temáticas por las cuales se indagan en la GEIH permiten caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, nivel educativo, afiliación al sistema de seguridad social en salud, grupos poblacionales como etnias, campesinos, LGBT o con algún tipo de discapacidad, otras formas de trabajo como producción de bienes y servicios para autoconsumo, trabajo en formación y voluntariado, entre otras. Actualmente, la GEIH cuenta con una muestra anual aproximada de 315.000 hogares a nivel nacional, lo que hace que sea la

encuesta de mayor cobertura a nivel nacional. De modo que permite obtener indicadores confiables y series continuas para analizar la fuerza de trabajo del país y los principales indicadores del mercado laboral, considerados como información fundamental para la toma de decisiones de política pública.

2. DATOS

En primer lugar, se realiza una limpieza de outliers en las bases de datos, esto con el fín de mejorar el poder de la predicción que se va a modelar. Las variables continuas fueron imputadas con la media, esto se hizo para: Edad, tiempo que llevan trabajando en la empresa y horas trabajadas la semana pasada. Luego, para realizar la limpieza de outliers se utilizó la metodología de rangos intercuartílicos.

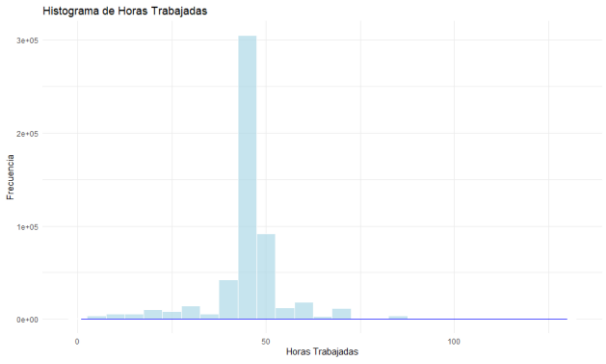
Luego, se crea la función "is_outlier" para a partir de rangos intercuartílicos y así eliminar los valores extremos presentes en las variables continuas. El rango intercuartílico es la diferencia entre el percentil 75 (Q3) y el percentil 25 (Q1) en un conjunto de datos, con lo que se mide la dispersión del 50% medio de los valores.

En el caso de las variables discretas o binarias, se imputaron por la moda, con lo cual el sexo, si es cotizante de seguridad social, el nivel educativo más alto alcanzado, el subsidio de alimentación, auxilio de transporte, el subsidio familiar, si cotiza pensiones, si está ocupado, el número de cuartos que dispone el hogar y el número de cuartos en los que duermen las personas del hogar.

Variables utilizadas

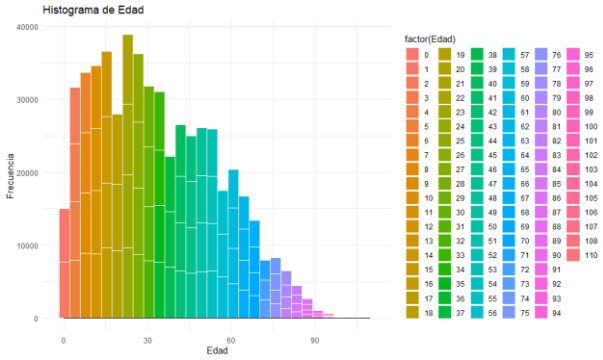
En primer lugar, se cree que la pobreza y el trabajo están estrechamente relacionados. Algunos autores han argumentado que la pobreza se debe en parte al hecho de que las personas pobres trabajan largas horas por salarios bajos, lo que les impide salir de la pobreza. Otros han argumentado que la pobreza se debe a la falta de oportunidades de empleo, lo que lleva a las personas a trabajar en empleos mal remunerados y precarios. Al respecto, se puede observar que una variable importante como las horas trabajadas tiene un alto número de datos en la cifra cercana a las 48 horas que tiene el país en su momento.

Figura 1. Horas trabajadas



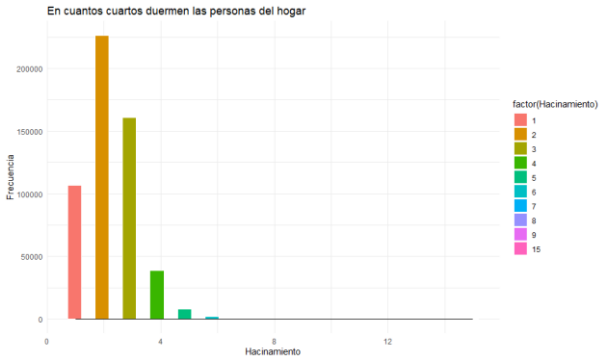
Otra variable definitiva como la edad, tienen un gran número de datos entre los 18 y los 60 años, lo cual está acorde con la edad productiva de una población típica analizada.

Figura 2. Histograma de edad



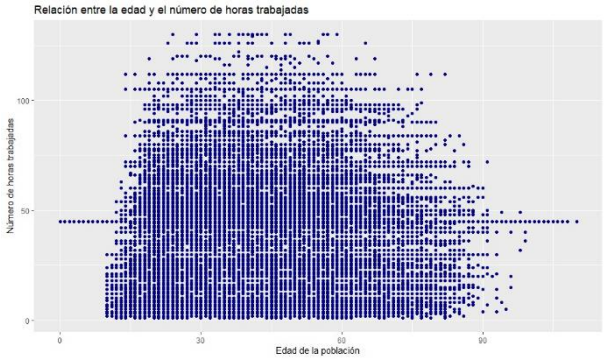
FInalmente, algunos estudios han encontrado que la calidad de la vivienda está relacionada con la pobreza, y que las personas que viven en viviendas de mala calidad tienen más probabilidades de estar en situación de pobreza. Por esta razón, se consideró importante tener en cuenta el número de habitaciones de una vivienda. Al respecto, la distribución del número de cuartos en los que duermen está acorde con lo que se encuentra en las estadísticas oficiales, un gran número de hogares que habitan viviendas que tienen entre 1 y 4 cuartos.

Figura 3. Distribución de los datos: En cuantos cuartos duermen las personas del hogar



Por otra parte, midiendo relaciones entre las variables, no se ve una relación clara entre el número de horas trabajadas la semana anterior y la edad de las personas, esto puede obedecer a que el número de horas que trabaja a la semana una persona puede estar ligada más al tipo de trabajo que desempeña, al nivel educativo que tiene o a otras dimensiones de ese estilo que pueden favorecer o no la carga laboral de un individuo.

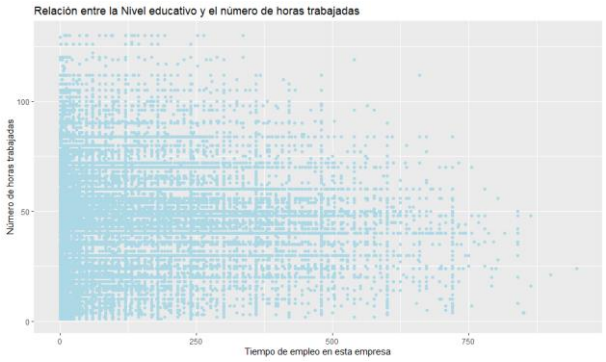
Figura 4. Relación entre la edad y el número de horas trabajadas



Por otro lado, cuando se ve la relación entre el numero de horas que trabajó la semana anterior y el tiempo que lleva en esa empresa, se ve que al principio hay una alta dispersión de la información, pero en la medida

en que el individuo tiene más tiempo en la empresa, se puede ver una baja en el tiempo que trabaja cada semana, lo que puede mostrar una mejora en las condiciones laborales.

Figura 5. Relación de Tiempo de empleo en esta empresa y horas trabajadas



3. MODELOS Y RESULTADOS

En esta sección se exponen las variables utilizadas, el entrenamiento del modelo, selección de los hiperparámetros y finalmente se hace un análisis comparativo.

3.1 Variables utilizadas

Como se mencionó en la sección de datos, se seleccionaron un total de 16 variables para el modelo. Estas variables guardan una relación con los niveles de pobreza, como sucede con los subsidios, la cotización de seguridad social, a pensiones o variables importantes como el número de cuartos que tiene la vivienda, como parte de las cualidades de la vivienda que muestra habitabilidad para un hogar típico.

3.2 Modelos de Clasificación

En primera instancia empleamos modelos de clasificación debido que son herramientas fundamentales en la predicción de la pobreza, ya que permiten identificar patrones y relaciones en los datos que pueden ser utilizados para predecir la condición de pobreza de un individuo o una comunidad. Estos modelos utilizan algoritmos de aprendizaje automático para analizar variables como ingresos, educación, salud, vivienda, entre otras, y asignar a cada observación a una categoría o clase, en este caso, pobreza o no pobreza. En resumen, los modelos de clasificación son una herramienta poderosa para comprender y abordar la pobreza, contribuyendo a la toma de decisiones informadas y a la mejora de la calidad de vida de las personas, por tal motivo se emplearon los siguientes modelos de clasificación:

Para todos los siguientes casos tener en cuenta:

$\boldsymbol{\beta_j}$ = Coeficientes de regresión de todas las variables

$$\boldsymbol{x_{ij}} = Sexo_i\beta_1, Edad_i\beta_2, CotizanteSeguridad_i\beta_3, NivelEducativo_i\beta_4, GradoEstudio_i\beta_5,$$

$$TiempoTrabajando_i\beta_6, SubsidioTransporte\beta_7, SubsidioAlimentacion_i\beta_8, HorasTrabajadas_i\beta_9,$$

$$CotizaPension_i\beta_{10}, Ocupado_i\beta_{11}, CuartosHogar\beta_{12}, CuartosDuerme_i\beta_{15}, Privación_i\beta_{14}$$

$$Formalidad\beta_{15}$$

3.2.1 OLS:

En primera instancia se empleó un modelo OLS (Ordinary Least Squares) el cual es un método de regresión lineal que se utiliza para predecir una variable continua a partir de una o más variables predictoras. Este modelo se basa en la minimización de la suma de los errores cuadráticos entre los valores observados y los valores predichos. En nuestro contexto en particular para la predicción de la pobreza teniendo una variable Dummy como dependiente, se emplean modelos de regresión logística que se basan en una función logística para predecir la probabilidad de pertenecer ser pobre o no. Para lo cual se planteó la siguiente ecuación:

$$\min_b E(\beta) = \sum(Pobre_i - \beta_0 - x_{ij})\text{^}2$$

En este caso buscábamos minimizar el error fuera de la muestra, seleccionando un β que cumpla la función anterior, en este caso como la variable independiente (Ser pobre o no) es una Dummy se emplea un método logístico.

3.2.2 Ridge:

La regresión ridge es una técnica utilizada en estadística para abordar el problema de la multicolinealidad en modelos de regresión lineal, como el anterior. La multicolinealidad ocurre debido que tenemos variables predictoras que están altamente correlacionadas, lo que puede conducir a estimaciones inestables de los coeficientes de regresión en el modelo. La regresión ridge aborda este problema al agregar un término de penalización a la función de mínimos cuadrados ordinarios (MCO), lo que ayuda a reducir la varianza de las estimaciones de los coeficientes. Matemáticamente, el estimador de regresión ridge se obtiene al minimizar la siguiente función de costo:

$$\sum_{i=1}\left(Pobre_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j\right)^2 + \gamma\sum_{j=1}\beta_j^2$$

Aquí estamos penalizando las funciones que crean varianza para lograr el trade-off entre sesgo y varianza.

3.2.3 Lasso:

Por otro lado, ahora en la regresión Lasso, se aplica una penalización a los coeficientes absolutos mientras que a diferencia de la de Ridge que aplica una penalización a los coeficientes de regresión. Esto significa que la regresión Ridge busca reducir la varianza de las estimaciones de los coeficientes, mientras que la regresión Lasso busca seleccionar un subconjunto de variables predictoras relevantes y evitar el sobreajuste. Es decir que la regresión Lasso utiliza una penalización que lleva a la selección de variables, ya que los coeficientes de las variables menos relevantes tienden a ser estimados como cero, bajo la siguiente intuición:

$$\sum_{i=1}\left(Pobre_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j\right)^2 + \sum_{j=1}|\beta_j|$$

Es decir que bajo este método seleccionamos automáticamente los predictores que van en el modelo (βj ≠0) y los que no (βj = 0).

3.2.4 Elastic Net:

En este caso combinamos elementos de los modelos de regresión Ridge y Lasso buscando la minimización de la función de pérdida de la regresión Ridge, pero con una penalización similar a la de la regresión Lasso, que es la suma absoluta de los coeficientes. Esto permitio al modelo Elastic Net seleccionar las variables relevantes y reducir la varianza de las estimaciones de los coeficientes, al tiempo que evita los problemas de multicolinealidad.

La función de nuestro modelo Elastic Net para predicción de la pobreza se puede expresar como:

$$\sum_{i=1}\left(Pobre_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j\right)^2 + \gamma\sum_{j=1}|\beta_j|$$

Donde tenemos ahora tanto el parámetro γ que penaliza la varianza de los estimadores como la penalización de los coeficientes absolutos, de modo que el modelo nos esta ayudando a evitar la multicolinealidad, seleccionar variables relevantes y reducir la varianza de las estimaciones de los coeficientes. A diferencia de la regresión Ridge, que aplica una penalización a los coeficientes de regresión, y la regresión Lasso, que aplica una penalización a los coeficientes absolutos, el Elastic Net combinamos ambos enfoques para obtener un modelo más robusto y con mejor generalización.

3.2.5 Arboles de Decisión:

Para realizar nuestros arboles de decisión realizamos un proceso basado en la división recursiva de nuestro conjunto de datos de entrenamiento en subconjuntos cada vez más homogéneos en términos de la variable de respuesta, si el individuo es pobre o no. El objetivo era encontrar cuales eran las variables de decisión que permitan predecir la variable de mejor manera.

En el proceso de construcción del árbol seleccionamos la variables de pobre (Si o no) y el punto de corte que mejor divide los datos en cada nodo (Aspecto que ahondáremos más adelante), de modo que se maximice la homogeneidad de los subconjuntos resultantes. Este proceso lo repetimos de forma recursiva hasta que se cumple un cierto criterio de parada, como el tamaño mínimo de la muestra en los nodos terminales o la máxima profundidad del árbol. Una vez construido el árbol, se pueden podar algunas ramas para evitar el sobreajuste y mejorar la generalización del modelo.

3.2.6 Bagging y Random Forest:

El último modelo empleado fue un modelo de clasificación de Random Forest, debido que permite mejorar en sobre manera el rendimiento de los árboles basado en el principio de la agregación, es decir, buscamos la agregación de varios árboles de decisión, este modelo lo explicaremos en la siguiente sección ya que como veremos a continuación fue el que obtuvo el mejor resultado.

Para este proceso en primera instancia obtuvimos las muestras aleatorias, dividiendo la muestra en entrenamiento y testo (70% Y 30% respectivamente), posteriormente para cada muestra se ajusta un árbol de regresión con la siguiente función:

$$f^b(x)$$

Para posteriormente promedia las muestras del Bootstrap, obteniendo la siguiente función:

$$f_{bag} = \frac{1}{B}\sum_{b=1} f^b(x)$$

Para así en cada partición utilizar un subconjunto de predictores elegidos al azar.

Finalmente, tras describir el enfoque de los seis modelos de clasificación, realizar una descripción clara y explicar la estrategia empleada con sus respectivas ecuaciones correctas y explicaciones, observamos cuales fueron los MAE (Mean Absolute Error) de los seis modelos encontramos los siguientes resultados:

Modelo Empleado	MAE
OLS	0.4082
Ridge	0.4337
Lasso	0.4339
Elastic Net	0.4479
Arboles de Decisión	0.501083
Random Forest	0.650032

De manera que el mejor modelo que se pudo seleccionar fue el de Random Forest, por tal motivo a continuación explicaremos los procesos empleados detrás del cálculo de este valor.

3.2.7 MODELO SELECCIONADO

Como se observó anteriormente el modelo seleccionado fue el modelo de Random Forest, como el modelo con el mejor MAE, de tal modo que para este caso vamos a explicar a detalle lo realizado en el modelo:

Detalles de la Estimación:

Para realizar la estimación se realizó en primera instancia un proceso de obtener repetidamente las muestras aleatorias de la muestra observada, para esto como estrategia de submuestreo hicimos fue seleccionar el 70% de la base de entrenamiento para entrenar a la muestra y posteriormente testearla con nuestra base *Test*. Tras realizar este proceso continuamos con el desarrollo del modelo empleando una partición de la muestra en 10 partes a través del método de Cross – Validation:

1. Divide los datos en 10 partes iguales (fold).
2. En cada iteración de validación cruzada, 9 de las 10 partes se utilizan para entrenar el modelo, y la última parte se utiliza para validar el modelo.
3. El proceso se repite 10 veces, y el modelo obtiene una medida promedio de su rendimiento en la validación.

Esta función nos fue útil para evitar el sobreajuste y mejorar la generalización del modelo al asegurar que se evalúa su rendimiento en una muestra separada de la que se entrena. Por consecuente creamos un conjunto de configuraciones de hiperparámetros para ajustar un modelo de Random Forest, con el siguiente proceso:

1. Creamos un marco de datos con todas las combinaciones posibles de los valores proporcionados.
2. Definimos el parámetro que representa el número de variables predictoras que se deben seleccionar aleatoriamente en cada división del árbol. En este caso, se están considerando los valores 2, 3, 4, 5 y 8.
3. El siguiente parámetro para definir era para que especificara la regla utilizada para dividir los nodos del árbol. En este caso, se ha fijado en "variance", lo que significa que se está utilizando la varianza para decidir cómo dividir los nodos.
4. Finalmente, un último parámetro que establece el tamaño mínimo permitido para un nodo terminal (hoja) del árbol. Se están considerando los valores 1, 2, 3 y 6 para

Posteriormente, realizamos la ejecución del modelo especificando las variables a tener en cuenta las cuales fueron:

Sexo, edad, cotizante seguridad, nivel educativo, grado de estudio, tiempo trabajado, subsidio de transporte, subsidio de alimentacion, horas trabajadas, cotiza pensión, ocupado, cuartos hogar, cuartos duerme, privación, formalidad.

Destacando adicionalmente, que los datos fueron tomados de la muestra de entrenamiento, con el cross-validation en 10 partes (Explicado previamente), la métrica de RMSE y con el método ranger el cual optimiza el funcionamiento de Random Forest, con este procedimiento ejecutamos el modelo y posteriormente predecimos con los coeficientes hallados para tanto dentro como fuera de la muestra, en este caso para fuera de la muestra, que es lo que nos interesa estimamos el MAE y llegamos al valor de 0.65

4. CONCLUSIONES

En conclusión, el documento analiza el impacto de variables económicas en la pobreza de los hogares y propone diferentes técnicas de aprendizaje de máquinas para predecir si un individuo es pobre o no. Se utilizaron modelos de clasificación como OLS, Ridge, Lasso, Elastic Net, Árboles de Decisión y Random Forest. Tras evaluar los resultados, se determinó que el modelo de Random Forest fue el más efectivo, con un MAE de 0.65. Este modelo se basa en la división recursiva de los datos de entrenamiento en subconjuntos homogéneos para predecir la pobreza. En resumen, el análisis de estas variables económicas es esencial para diseñar estrategias efectivas de reducción de la pobreza y promover un desarrollo inclusivo y sostenible.

Finalmente, observamos como en este estudio, se examinaron diversas variables con el propósito de medir la pobreza, y los resultados revelaron la utilidad significativa de las variables relacionadas con la educación, las características de la vivienda y las condiciones laborales. La educación emergió como un indicador crucial, destacando su influencia en el estatus económico de los individuos. Aquellos con mayores niveles

educativos exhibieron una tendencia a experimentar una menor incidencia de pobreza. Además, las características de la vivienda desempeñaron un papel esencial, evidenciando la importancia de condiciones habitacionales adecuadas en la mitigación de la pobreza. Asimismo, las características del empleo demostraron ser predictores valiosos, subrayando la necesidad de considerar la estabilidad y calidad del empleo al evaluar la situación económica. En resumen, este estudio respalda la relevancia de las variables educativas, de vivienda y laborales como instrumentos fundamentales para la medición precisa de la pobreza, proporcionando insights valiosos para el diseño de políticas dirigidas a mejorar las condiciones socioeconómicas de la población.

5. LINK REPOSITORIO

A este documento se anexa el repositorio GitHub el cual contiene la documentación de la base de datos, los scripts y los outputs correspondientes de cada punto. El repositorio se encuentra en el siguiente link:

https://github.com/Luis-Borda/G10_TALLER_3.git