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

La identificación de hogares pobres constituye uno de los principales desafíos de las economías latinoamericanas, donde persisten elevadas tasas de pobreza. Este fenómeno puede abordarse tanto desde un enfoque monetario como desde una perspectiva multidimensional, que considere aspectos relacionados con la calidad de vida de los miembros del hogar. A su vez, estos países enfrentan limitados recursos fiscales para la recolección de datos. En este contexto, la identificación de variables clave y el diseño de un modelo que permita predecir la condición de pobreza de un hogar con el menor número de variables posibles se presentan como una oportunidad para mejorar la focalización y la evaluación de políticas públicas orientadas a reducir la pobreza.

En este sentido, la literatura ha puesto su atención en la estimación de modelos Probit (Anaya Narváez et al, 2015), aunque recientemente autores como Caballero (2021) y Rincón (2021) han comenzado a explorar el uso de técnicas de machine learning para la predicción de la pobreza en países latinoamericanos.

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 datos de entrenamiento 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 muestran que el modelo de clasificación que predice la pobreza en los hogares resultó ser muy efectivo. Esto indica que el modelo es muy bueno para identificar correctamente a los hogares pobres. La proporción de pobreza predicha por este modelo fue del 19.11 %, un valor bastante cercano a las estimaciones oficiales (obtenidas del DANE), lo que muestra que el modelo tiene un buen rendimiento y podría usarse con confianza para identificar hogares en riesgo de pobreza. Por otro lado, en el modelo de predicción de ingresos individuales, el **Decision Tree Regressor (CART)** fue el que mejor desempeño tuvo. La proporción de hogares pobres predicha por este modelo fue del 18.77 %, también consistente con las estimaciones disponibles.

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 6 variables. Es importante aclarar que se va a estar midiendo pobreza multidimensional, por ello se agregaron variables proxys de posibles dimensiones que afecten en la determinación de si una persona es pobre o no, como educación, salud, condiciones de vida, hacinamiento, etc. La variable numérica incluída es la edad; mientras que las variables categóricas son la educación máxima alcanzada por los miembros del hogar, el tipo de ocupación de los individuos, los regímenes de seguridad social en salud al que están afiliados los miembros del hogar (subsidiado u otros), el número de personas que duermen en una misma habitación y el valor de la línea de pobreza.

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. Niveles bajos de educación pueden señalar hogares que se encuentren en una situación de pobreza. Como proxy de acceso a la salud, se contempló a cuáles regímenes de seguridad social en salud estaban 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). Dadas las características del sistema de seguridad social en salud colombiano, en el régimen subsidiado se encuentran todas las personas más pobres y vulnerables.

Como proxy de condición de vida, se utilizaron distintas variables como la edad de los individuos y su tipo de ocupación. La primera, dado que influye en la etapa del ciclo de vida en que se encuentra una persona, lo que termina afectando su situación económica, pero principalmente porque es una proxy de la experiencia laboral; y la segunda, debido a que algunas ocupaciones ofrecen más estabilidad laboral y beneficios que otras, entonces ocupaciones precarias o informales están más asociadas con la pobreza. Como proxy de hacinamiento crítico se utilizó el número de personas por cuarto. Ésta se creó en función de otras dos variables referentes a la cantidad de cuartos en el hogar (incluyendo sala-comedor) y el número de habitantes de ese hogar. De acuerdo con la DANE (2012), son considerados hogares con hacinamiento crítico aquellos en donde el número de personas por cuarto es superior a tres, contando la sala y el comedor pero excluyendo los baños, garajes y los cuartos utilizados para negocio. Por último, se utilizó la línea de pobreza dado que indica la ponderación de cada región.

Para la segunda parte se seleccionaron 7 variables. Las variables numéricas consideradas son la edad, la cantidad de horas trabajadas[1], la antigüedad laboral y la línea de pobreza. También se han incluido variables categóricas como el máximo nivel de educación y el tipo de ocupación tanto en el primer como en el segundo trabajo.

Estas variables se consideran pertinentes por varias razones:

- Edad: influye en la experiencia laboral y el nivel salarial.
- Horas trabajadas: afecta directamente los ingresos.
- Antigüedad laboral: puede influir en los ingresos a través de aumentos salariales y promociones basadas en la experiencia y el tiempo en la empresa.

[1]Se considera el total de las horas trabajadas normalmente.

- 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.
- Tipo de ocupación: como algunas ocupaciones ofrecen más estabilidad laboral y beneficios que otras, las ocupaciones precarias o informales están más asociadas con la pobreza.
- Línea de pobreza: sirve de ponderación para las regiones, ya que cada una tiene una línea de pobreza distinta.

La base de datos final se obtuvo mediante cinco simples pasos. Primero, se crearon las variables de máximo nivel educativo alcanzado por los miembros del hogar, de régimenes de seguridad social en salud y de hacinamiento. Segundo, se transformaron los datos a formato numérico entero. Tercero, se renombraron las columnas con nombres más indicativos de cada variable. Cuarto, se reemplazaron los datos faltantes por los valores predichos mediante un modelo de regresión. Finalmente, se agruparon los datos por id de hogar, para obtener una base a un nivel más agregado.

Tabla 1: Estadísticas descriptivas de la submuestra seleccionada de entrenamiento

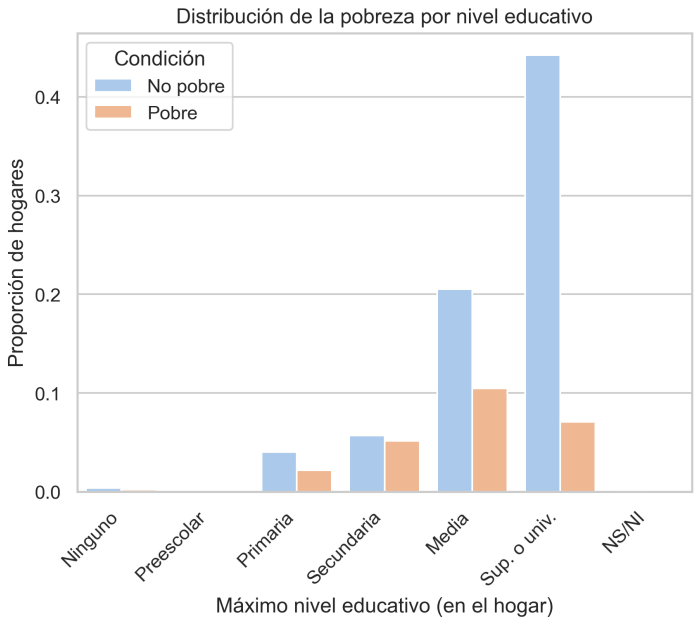
Variable	n	Media	S.E.	Mín	25 %	50 %	75 %	Máx
Ingreso total	447659	774791.90	1373891	0	0	435000	995711	85833330
Edad	543109	33.55	21.65	0	16	31	50	110
Hs de trabajo sem.	248075	44.80	15.71	1	40	48.00	50	130
Antigüedad (meses)	248075	85.78	113.59	0	11	36	120	948

Elaboración propia en base GEIH 2018 (DANE)

El porcentaje de hogares hacinados asciende a 2,5 % de la muestra de entrenamiento, mientras que el porcentaje de hogares con salud en régimen subsidiado alcanza un 36 %.

La Figura 1 muestra la cantidad de hogares de la muestra de entrenamiento clasificados como pobres y no pobres por el DANE según el máximo nivel educativo alcanzado por todos los miembros del hogar. De allí resulta evidente que la categoría educativa con mayor número de pobres es la media, donde aproximadamente 1/3 de los hogares cuyos miembros han alcanzado cómo máximo un nivel medio (10º - 13º grado) son pobres. Por otra parte, resulta llamativa la cantidad de hogares con al menos un miembro con educación superior o universitaria que caen por debajo de la línea de pobreza. Si bien representan una pequeña porción de los universitarios de la muestra, en términos relativos representan más del 5 % de los hogares del total muestral.

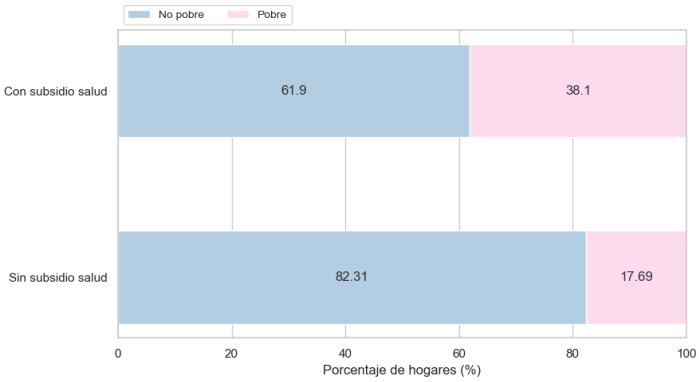
Figura 1: Distribución de la pobreza por Nivel educativo



Elaboración propia en base GEIH 2018 (DANE)

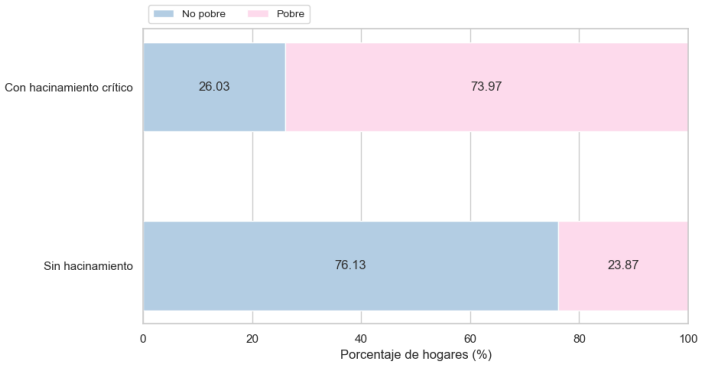
Por otra parte, la Figura 2 muestra la proporción de hogares afiliados al sistema de salud mediante el régimen subsidiado según su condición de pobreza. Dado que la proporción de pobres con subsidio es mayor a la proporción de pobres sin subsidio, sería esperable que esta variable ayude a predecir la condición de un hogar con respecto a la pobreza. De manera similar, la Figura 3 muestra la proporción de hogares pobres y no pobres con hacinamiento crítico. De acuerdo a lo esperado, una gran proporción de los hogares pobres (más del 70 %) habitan viviendas con menos ambientes de los necesarios para una calidad de vida adecuada.

Figura 2: Distribución de la pobreza por Regímenes de Seguridad Social en salud



Elaboración propia en base GEIH 2018 (DANE)

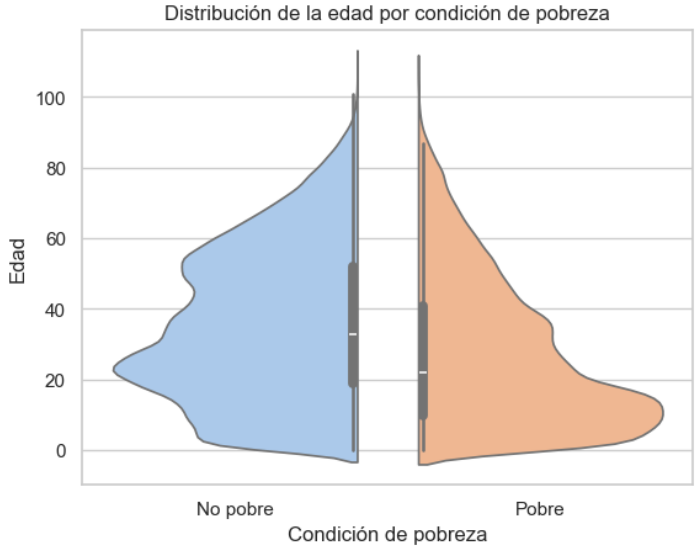
Figura 3: Distribución de la pobreza por Hacinamiento



Elaboración propia en base GEIH 2018 (DANE)

Finalmente, la Figura 4 muestra la distribución etaria de la pobreza. Es posible notar que la pobreza está concentrada en los grupos más jóvenes, especialmente los menores de 20 años. Esto podría ser consistente con una mayor prevalencia de la pobreza en hogares con mayor número de hijos.

Figura 4: Distribución de la pobreza por Edad



Elaboración propia en base GEIH 2018 (DANE)

MODELO

PRIMER MODELO

El objetivo del primer modelo (modelo de clasificación) es predecir si un hogar es pobre o no, utilizando diversas características socioeconómicas de sus miembros. Para esto, se ha creado un dataset a nivel de hogar, agrupando la información individual de los miembros del hogar. La variable objetivo (*Pobre*) indica si el hogar es clasificado como pobre.

Las variables independientes utilizadas en el modelo son:

- **Edad (P6040)**
- **Hacinamiento (hacin)**
- **Máximo nivel de educación (educ_max)**
- **Proxy de acceso a la salud (salud_subsi)**
- **Tipo de ocupación (P6430)**.
- **Línea de pobreza (Lp)**

Anteriormente, se seleccionaron y construyeron las variables relevantes. Para aquellas observaciones con datos faltantes, se construyó un modelo de regresión que predice el valor de aquellas observaciones que no tienen ningún valor. Los datos se agruparon a nivel de hogar, calculando la media de las características y asignando la etiqueta de pobreza al hogar si algún miembro era clasificado como pobre. No se utilizó ninguna estrategia de submuestreo (*sub-sampling*) en los datos.

Se entrenaron varios modelos de clasificación para predecir la pobreza de los hogares:

- **Regresión Logística**
- **Elastic Net**
- **Decision Tree**
- **Random Forest**
- **Gradient Boosting**

A continuación se detallan los hiperparámetros seleccionados para cada modelo:

Regresión Logística: Se utilizó un máximo de iteraciones (*max_iter*) de 1000 para asegurar la convergencia del modelo.

Elastic Net: Se seleccionaron *alpha* = 0.1 y *l1_ratio* = 0.5 para balancear la regularización entre los términos L1 y L2.

Decision Tree: Se entrenó un modelo con los hiperparámetros por defecto y otro con una profundidad máxima (*max_depth*) de 10 para evitar el sobreajuste.

Random Forest: Se entrenaron dos modelos, uno con 100 estimadores y otro con 200 estimadores y una profundidad máxima (*max_depth*) de 10 para mejorar la generalización.

Gradient Boosting: Se entrenaron dos modelos, uno con 100 estimadores y una tasa de aprendizaje (*learning_rate*) de 0.1, y otro con 200 estimadores y una tasa de aprendizaje de 0.05 para una mejor precisión.

Cada modelo fue entrenado utilizando el conjunto de entrenamiento. Las predicciones fueron evaluadas en términos de exactitud (*accuracy*) y F1 Score en los datos de entrenamiento. A continuación, se presentan los resultados de los modelos:

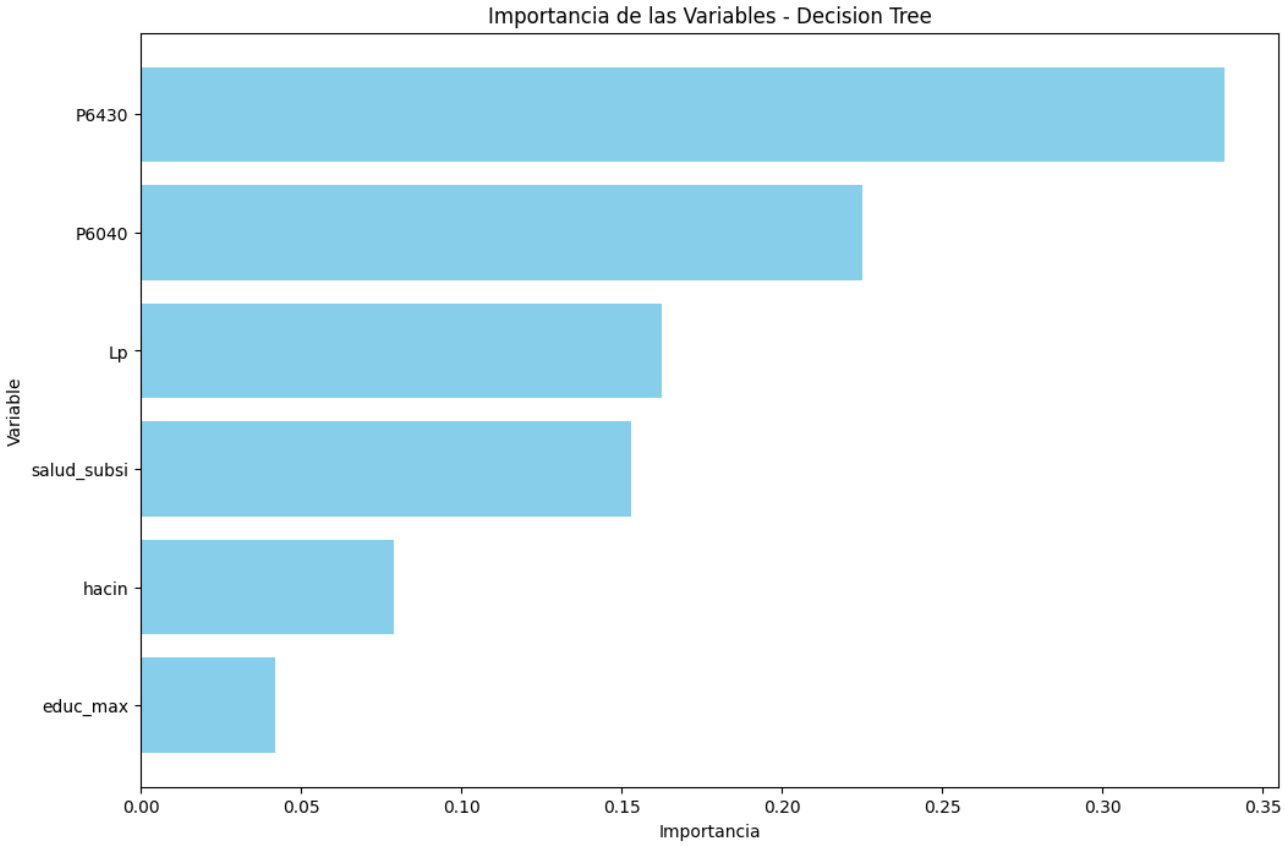
Modelo	Exactitud	F1 Score
Regresión Logística	0.82	0.43
Elastic Net	0.80	0.04
Decision Tree	0.999	0.998
Decision Tree (profundidad máxima = 10)	0.87	0.62
Random Forest	0.999	0.998
Random Forest (200 estimadores, profundidad máxima = 10)	0.87	0.60
Gradient Boosting	0.86	0.57
Gradient Boosting (200 estimadores, tasa de aprendizaje = 0.05)	0.86	0.57

Tabla 2: Resultados de los modelos

El mejor modelo basado en el F1 Score fue el **Decision Tree (CART)**.

A este análisis, se le debe incluir una discusión de la importancia relativa de cada variable en el modelo, lo cual sale de este gráfico:

Figura 5: Importancia relativa de las variables del mejor modelo



Tipo de ocupación (P6430) - Importancia: 0.338

Ocupaciones estables y mejor remuneradas proporcionan una base económica más sólida, reduciendo la probabilidad de que el hogar caiga por debajo de la línea de pobreza.

Edad (P6040) - Importancia: 0.225

Personas de mayor edad y experiencia suelen tener mejores empleos y salarios. Hogares con más adultos en edad laboral tienen mejor capacidad económica para cubrir necesidades básicas.

Línea de pobreza (Lp) - Importancia: 0.163

La línea de pobreza contextualiza el ingreso del hogar respecto al costo de vida local, siendo crucial para clasificar a los hogares como pobres o no.

Proxy de acceso a la salud (salud_subsi) - Importancia: 0.153

El sistema de subisidios a la salud en Colombia esta dividido en dos: un seguro social contributivo y otro subsidiado. El subsidiado está destinado a personas de menores recursos, por lo que ver si una persona accede o no a este recurso es importante para la predicción de la pobreza.

Hacinamiento (hacin) - Importancia: 0.079

Refleja la calidad de vida; altos niveles de hacinamiento indican bajos ingresos y falta de acceso a viviendas adecuadas. No tiene demasiada importancia en comparación al resto de variables, pero sigue siendo significativa.

Máximo nivel de educación (educ_max) - Importancia: 0.042

La educación es clave para acceder a mejores oportunidades laborales y salarios. Sin embargo, tiene la menor importancia relativa.

Se creó un *pipeline* final que incluye la normalización de los datos con *StandardScaler* y el mejor modelo identificado (Decision Tree). El *pipeline* se entrenó con el conjunto de entrenamiento y se usó para realizar las predicciones en el conjunto de prueba. La proporción de hogares pobres predicha fue del 19.11 %.

SEGUNDO MODELO

El segundo modelo es uno de predicción de ingresos individuales que se agrupan por hogar para asi luego poder medir la pobreza. Las variables utilizadas en el modelo son las siguientes:

- **Edad (P6040)**
- **Horas trabajadas (P6800)**

- **Antigüedad laboral (P6436)**
- **Máximo nivel de educación (educ_max)**
- **Tipo de ocupación (P430)**
- **Línea de pobreza (Lp)**
- **Ocupación segunda actividada (P7050)**

Al igual que para el primer modelo, se realizó el preprocesamiento de los datos ajustando los valores faltantes y asegurando que las variables estuvieran en el formato adecuado. Los datos fueron divididos en conjunto de entrenamiento y prueba, utilizando las dos bases de datos proporcionadas. Al estar prediciendo ingresos, se configuró el modelo para que le otorgue valor o a aquellos ingresos predichos negativos.

Se utilizaron los siguientes modelos para la predicción de ingresos:

- **Regresión Lineal**
- **Elastic Net**
- **Random Forest Regressor**
- **Decision Tree Regressor** (CART)
- **AdaBoost Regressor**
- **Gradient Boosting Regressor**

A continuación se detallan los hiperparámetros seleccionados para cada modelo:

Regresión Lineal: Se utilizaron los parámetros por defecto.

Elastic Net: Se seleccionaron *alpha* = 0.1 y *l1_ratio* = 0.5 para balancear la regularización entre los términos L1 y L2.

Elastic Net (alpha=1): Se seleccionaron *alpha* = 1 y *l1_ratio* = 0.7 para una mayor regularización L1.

Random Forest Regressor: Se entrenaron dos modelos, uno con $n\_estimators = 100$ y $max\_depth = 5$, y otro con $n\_estimators = 200$ y $max\_depth = 10$ para mejorar la generalización.

Decision Tree Regressor (CART): Se entrenó un modelo con los hiperparámetros por defecto.

AdaBoost Regressor: Se utilizó $random\_state = 42$ para asegurar la reproducibilidad.

Gradient Boosting Regressor: Se entrenaron dos modelos, uno con $learning\_rate = 0{,}1$ y otro con $learning\_rate = 0{,}05$ para una mejor precisión.

Se entrenaron los modelos en el conjunto de entrenamiento y se evaluaron usando el error cuadrático medio (MSE) en el conjunto de entrenamiento. Los resultados obtenidos fueron los siguientes:

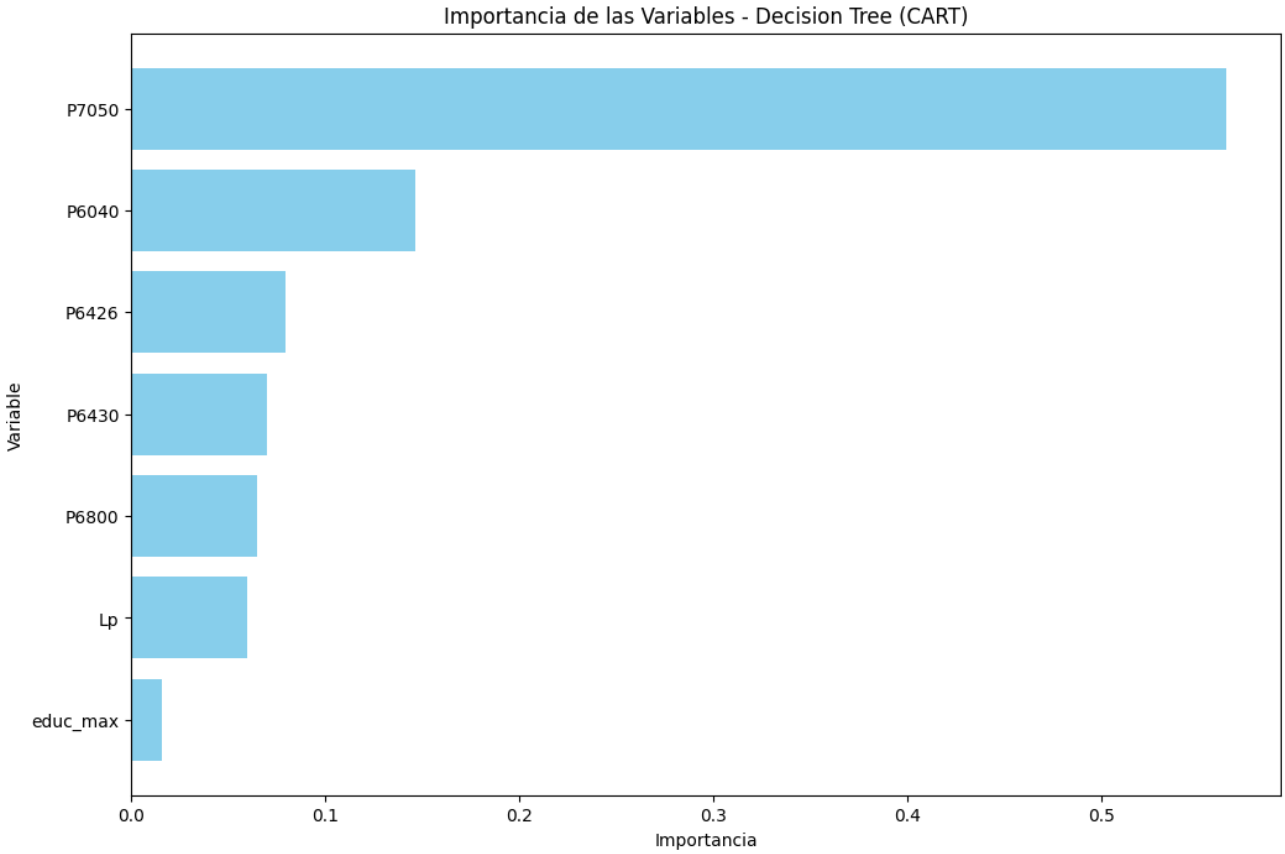
Modelo	**MSE en entrenamiento**
Linear Regression	1225090547409.19
Elastic Net	1226613536149.51
Elastic Net (alpha=1)	1254275893161.64
Random Forest	358826076178.07
Decision Tree (CART)	98861.48
AdaBoost Regressor	7259407886295.95
Gradient Boosting (learning_rate=0.1)	460560372596.55
Gradient Boosting (learning_rate=0.05)	553598644279.69

Tabla 3: Resultados de MSE en entrenamiento para diferentes modelos

El mejor modelo fue el **Decision Tree Regressor** (CART) con un MSE de 98861.48.

A continuación, hacemos un analisis de la importancia relativa de las variables del mejor modelo.

Figura 6: Importancia relativa de las variables del mejor modelo



Ocupación segundo trabajo (P7050) - Importancia: 0.564

El tener un segundo trabajo puede indicar una mayor necesidad económica, ya que los ingresos del empleo principal no son suficientes para cubrir las necesidades del hogar. Las personas con un segundo empleo tienden a tener mayores ingresos en comparación con aquellos que solo trabajan en un empleo.

Edad (P6040) - Importancia: 0.146

Las personas con mayor edad y experiencia suelen tener mejores empleos y salarios. La edad es un indicador importante de la experiencia laboral, lo que generalmente se traduce en mayores ingresos.

Antigüedad laboral (P6426) - Importancia: 0.079

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. Sin embargo, su impacto no es tan grande como el de la edad o el tener un segundo empleo.

Tipo de ocupación (P6430) - Importancia: 0.070

El tipo de ocupación está relacionado con la estabilidad laboral y los beneficios. Las ocupaciones más estables, como las de tiempo completo en sectores formales, suelen ofrecer mayores ingresos que las ocupaciones informales o precarias.

Horas trabajadas (P6800) - Importancia: 0.065

El número de horas trabajadas directamente influye en los ingresos. Las personas que trabajan más horas tienden a generar mayores ingresos, especialmente si se consideran horas adicionales fuera de una jornada laboral estándar.

Línea de pobreza (Lp) - Importancia: 0.060

La línea de pobreza se utiliza para contextualizar los ingresos del hogar con respecto al costo de vida en una región. Aunque tiene un impacto relativo menor en la predicción de ingresos, sigue siendo una variable importante, ya que en regiones con líneas de pobreza más altas, los ingresos necesarios para satisfacer las necesidades básicas también son mayores.

Máximo nivel de educación (educ_max) - Importancia: 0.016

A mayor nivel educativo, mayores son las oportunidades de empleo y los salarios. Sin embargo, en este modelo, su impacto es menor en comparación con las otras variables lo que sugiere que, aunque importante, no es tan determinante.

Este modelo se usó para predecir los ingresos individuales en el conjunto de prueba. Posteriormente, se agruparon los ingresos por hogar para calcular la proporción de hogares por debajo de la línea de pobreza. La proporción de hogares pobres predicha fue del 18.77 %.

CONCLUSIÓN

En este trabajo, se abordó el desafío de predecir la pobreza en Colombia usando modelos de clasificación y regresión, con el objetivo de mejorar la medición de la pobreza a nivel de hogar, basándose en datos del Departamento Administrativo Nacional de Estadística (DANE) y la misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE). La idea central que motivó el trabajo fue la de usar modelos predictivos para optimizar las encuestas y generar intervenciones más efectivas y económicas.

El modelo de clasificación que predice la pobreza en los hogares resultó ser muy efectivo, destacando el **Decision Tree**, que alcanzó un F1 Score de 0.998 y una exactitud del 99.9 %. Esto indica que el modelo es muy bueno para identificar correctamente a los hogares pobres. La proporción de pobreza predicha por este modelo fue del 19.11 %, un valor bastante cercano a las estimaciones oficiales (obtenidas del DANE), lo que muestra que el modelo tiene un buen rendimiento y podría usarse con confianza para identificar hogares en riesgo de pobreza.

Por otro lado, en el modelo de predicción de ingresos individuales, el **Decision Tree Regressor (CART)** fue el que mejor desempeño tuvo, con un error cuadrático medio (MSE) de 98,861.48. Este valor es bajo en comparación, por ejemplo, con el valor promedio de los ingresos en la muestra de entrenamiento (774.813 pesos), lo cual podría indicar que el modelo predice de manera óptima. La proporción de hogares pobres predicha por este modelo fue del 18.77 %, también consistente con las estimaciones disponibles.

Con estos resultados, las políticas públicas pueden ser mucho más eficientes. Al usar estos modelos para identificar con mayor precisión a los hogares pobres, se pueden diseñar intervenciones más focalizadas y reducir el gasto en encuestas exhaustivas. Uno de los logros de este trabajo, es la poca cantidad de variables que se usaron para construir los modelos (6 en el modelo de clasificación y 7 en el de regresión). Esto no solo ahorra recursos, sino que también permite implementar programas que lleguen de manera más directa a las familias que realmente los necesitan. Por ejemplo, programas educativos, de salud o de empleo, dirigidos específicamente a hogares con mayor riesgo de pobreza.

A largo plazo, la implementación de modelos predictivos en lugar de las encuestas tradicionales podría reducir considerablemente los costos de recolección de datos, lo que también hace más fácil realizar un seguimiento constante de la pobreza. Además, esto facilitaría la focalización de políticas públicas en las variables que realmente importan, como el acceso a la educación o al empleo formal, áreas clave para reducir la pobreza de forma efectiva.

REFERENCIAS

- Anaya Narváez, A. R.; Buelvas Parra, J.; Valencia Burgos, L. C. (2015). Modelo Probit para la medición de la pobreza en Montería, Colombia. Opción, vol. 31, núm. 78, pp. 42-64 *Universidad del Zulia*. https://www.redalyc.org/pdf/310/31044046004.pdf

- Departamento Administrativo Nacional de Estadística (DANE) (2012). Atlas Estadístico 2012.

- Galvis Caballero, Angel (2021). ¿Cómo puede contribuir el machine learning a la focalización de programas sociales? Modelo XGBoost para la determinación de pobreza monetaria interpretado mediante Shap Values: caso Colombia 2019-2020.

- Rincón, Ratzanyel (2021). ESTIMACIONES TRIMESTRALES DE POBREZA MULTIDIMENSIONAL EN MEXICO MEDIANTE ALGORITMOS DE APRENDIZAJE DE MAQUINA. Estudios Económicos, vol. 38, num. 1, páginas 3-68.