TECNOLÓGICO DE MONTERREY

PREDICCIÓN DE INGRESOS

REPORTE

INEGI

ENIGH

Carol Rendon A0142534
Carolina Treviño A00835598
Luis Garza A00836982
Paula Alba A01722262
Rolando Ruiz A00835733

JUNIO 2024

1. Introducción

En un país numerosamente poblado como lo es México con más de 126 millones de habitantes (INEGI, 2020), se presenta una gran diversidad de escenarios y características demográficas que determinan el estilo de vida de los individuos; esto es porque factores como el sexo, el lugar de nacimiento o el origen étnico tienen influencia sobre los trabajos, oportunidades y condiciones a los que es posible acceder. De esta manera, el ingreso económico es variado a lo largo de la República y, tomando sólo a la población mayor de edad que cuenta con algún ingreso, el promedio se sitúa en poco menos de 25 mil pesos mexicanos (mpm) trimestrales. Por ello, con el fin de analizar la diversidad socioeconómica del país y poner en práctica el uso de métodos estadísticos y algorítmicos, el presente trabajo propone la creación de un modelo de *machine learning* capaz de predecir si un individuo mayor de edad cuenta con un ingreso trimestral mayor o menor que la media nacional de 25 mpm aproximadamente. Esto a partir de datos sociodemográficos de la persona como variables de entrada provenientes de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2022 (ENIGH) del INEGI. Específicamente, es posible dividir el presente documento en dos etapas principales: 1) el entendimiento, obtención y manejo de los datos para obtener una base de datos íntegra y 2) el entrenamiento, experimentación y evaluación de los modelos de *machine learning,* específicamente de árboles de decisión.

2. Comprensión del negocio

El Instituto Nacional de Estadística y Geografía (INEGI) es un organismo público de México cuyas responsabilidades residen en garantizar la accesibilidad, transparencia y objetividad e independencia de la información de interés nacional (SNIEG, s. f.). Esto implica utilizar y estandarizar métodos estadísticos para facilitar la comparación de los datos en el tiempo y en el espacio. Por ello, entre los procedimientos más importantes del INEGI se encuentra la realización del censo nacional de población, o la recopilación de datos mediante documentos institucionales como la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) que recopila datos como los ingresos trimestrales y la información sociodemográfica asociada a los individuos encuestados (INEGI, 2023).

Por la fiabilidad de estos datos, la amplia documentación proporcionada por la institución y su fácil acceso, se utilizan los registros captados en la ENIGH 2022 para la elaboración de este trabajo.

3. Comprensión de los datos

Tras tomar los datos de Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2022 proporcionados por el INEGI, se realizó una unión de las bases de datos de ingresos y de población tomando los folios y número de referencia como

identificadores. Con esto, se obtuvo una base combinada original constituida por 309,684 registros y 192 campos para cada uno. Sin embargo, después de realizar la filtración de individuos mayores de edad y de eliminar todos los registros con un ingreso trimestral no especificado, se mantuvo una base de datos de sólo 180,583 observaciones. De estos registros, 117391 (65%) pertenecen a observaciones por debajo de la media de 25 mpm trimestrales y 63192 (35%) representan a individuos

Dado el alto número de características en el conjunto de datos original, se decidió realizar una selección de variables consideradas, bajo criterio propio, como relevantes para el objetivo de predicción de ingreso menor o mayor a 25 mpm por individuo. Las variables elegidas reflejan distintos aspectos sociales (atención médica, padre o madre de hogar, etc.), educativos (nivel máximo de estudios), culturales (habla indígena) y físicos (discapacidades, edad) de la persona en cuestión, estas se muestran en la Tabla 1 y se detallan en mayor medida en el Anexo 1. Estos campos son de naturaleza categórica, a excepción de la edad y el número de horas trabajadas en la semana anterior (hor_1) que son variables numéricas discretas.

Variables elegidas para análisis

parentesco	edo_conyug	disc_ver	ss_aa	inscr_6
sexo	hor_1	disc_brazo	ss_mm	inscr_7
edad	atemed	disc_apren	inscr_1	
madre_hog	num_trabaj	disc_oir	inscr_2	
padre_hog	entidad	disc_vest	inscr_3	
hablaind	clave_max[1]	disc_habla	inscr_4	
nivelaprob	disc_camin	disc_acti	inscr_5	

Tabla 1. Columnas seleccionadas de la base de datos de población de la ENIGH 2022.

Asimismo, se tomó la columna de ingreso trimestral (ing_tri_total[2]) y se convirtió a una columna binaria (ing_binario) para indicar si el individuo cuenta con un ingreso trimestral superior o igual a 25 mpm (valor 1 binario) o si este se encuentra por debajo del umbral (valor 0 binario). Esta columna representa nuestro objetivo a predecir para el modelo de *machine learning.*

Con esto, es posible realizar las etapas posteriores de análisis exploratorio univariado y bivariado, así como la limpieza e ingeniería de características necesarias para construir una base de datos íntegra y apropiada para el modelo de predicciones. Estos procedimientos se presentan a continuación.

[1] En la base de datos de población de la ENIGH esta se encuentra como “clave”.

[2] En la base de datos de ingresos de la ENIGH esta se encuentra como “ing_tri”.

3.1 Análisis Exploratorio de Datos (EDA)

3.1.1 Análisis univariado: exploración de variables individuales

3.1.1.1 Variables de discapacidad

En la Figura 1 se indica el grado de discapacidad o dificultad de una persona para realizar ciertas acciones como caminar, aprender, vestirse, entre otras. En estas gráficas, el valor 1 indica que el individuo no puede realizar la acción, el valor 4 denota la ausencia de cualquier problema o dificultad para realizarla y los valores 2 y 3 son estados intermedios.

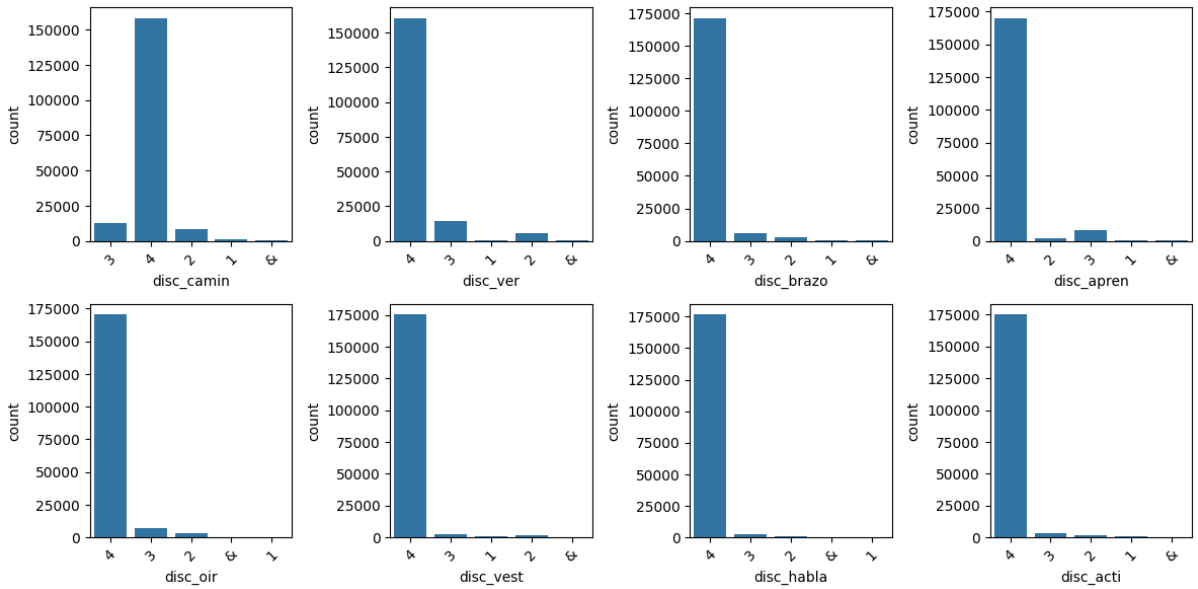

Figura 1: Gráficas de distribución de discapacidades.

Se observa que todas las variables de la Figura 1 están considerablemente desbalanceadas y dominadas fuertemente por la categoría 4. Del total de 180,583 registros, se tiene que más de 150,000 registros toman este valor predominante, lo que indica que la vasta mayoría de individuos no presentan discapacidad alguna. Así, esta variable podría no aportar mucha información para los modelos o causar sesgo. Con el fin de disminuir el desbalance, se aborda una unificación de todas las discapacidades en una sola variable binaria para indicar si una persona presenta o no alguna dificultad en actividades diarias. Este procedimiento se describe dentro de la sección de ingeniería de características (Sección 4.3.1). De manera alternativa, podría optarse por no incluir estas variables en el análisis y experimentación.

3.1.1.2 Variables sexo, madre y padre de hogar

La Figura 2 muestra las distribuciones de las gráficas para las variables de sexo, padre_hog y madre_hog. En “sexo”, los valores 1 y 2 indican hombre y mujer,

respectivamente. Por otra parte, en “padre_hog” y “madre_hog” los valores 1 y 2 denotan la presencia o ausencia del padre o madre en el hogar según sea el caso.

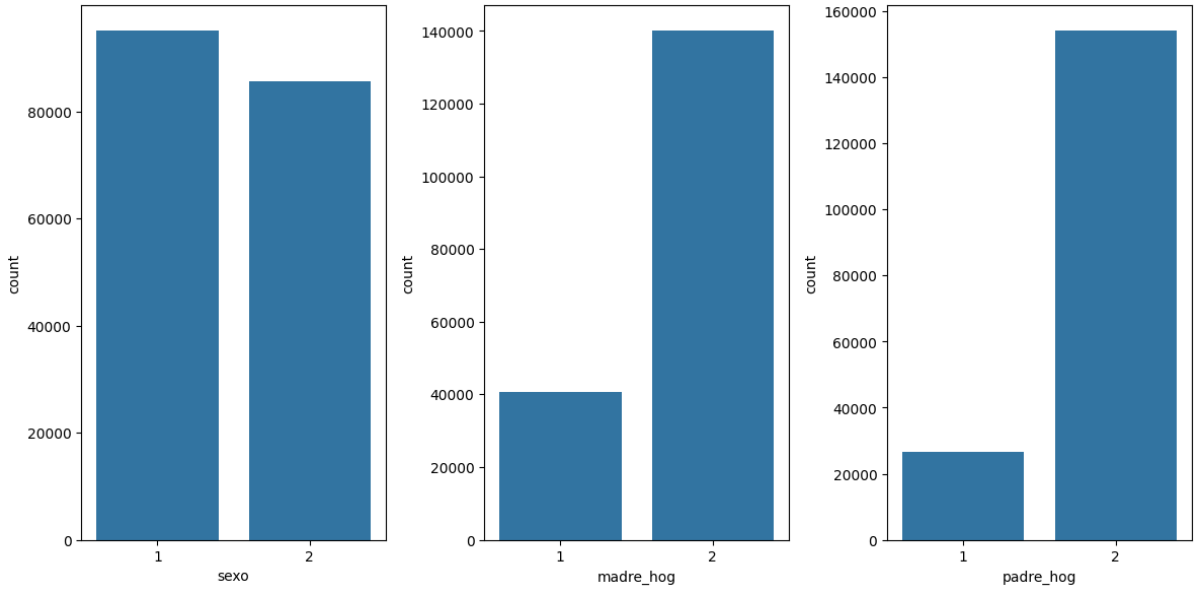

Figura 2: Gráfica Sexo, Madre en hogar y Padre en hogar

Es apreciable para la característica de sexo un ligero desbalance entre los números de hombres y mujeres registrados, sin embargo, se considera poco significativo para generar un efecto de sesgo en el comportamiento de la variable dentro del modelo. En cambio, para las variables de padre y madre en el hogar, se presenta un desequilibrio pronunciado hacia la ausencia de estos miembros en el hogar. Esto es especialmente notorio para el padre de familia, cuyo número de registros positivos se encuentra cerca de la mitad en comparación con aquellos de madre en el hogar.

3.1.1.3 Variable de parentesco

La variable de parentesco representa qué tipo de relación tiene el encuestado con el jefe del hogar. Cada categoría tiene un significado denotado en el diccionario de la ENIGH y se encuentran descritos en el Anexo 1. En la Figura 3 se muestra una gráfica de barras con las frecuencias de las distintas categorías de parentesco encima de cada una.

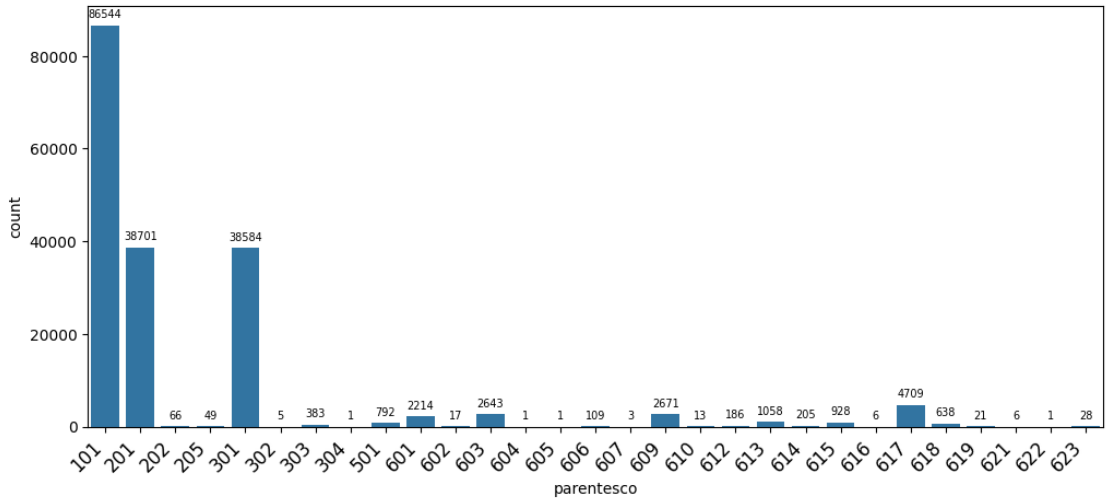

Figura 3: Gráfica de distribución de parentesco.

Es fácilmente identificable la prevalencia de la categoría 101 que denota el parentesco “es jefe” y acapara casi la mitad de los registros. Seguidamente en frecuencia se encuentran los valores 201 y 301 que indican las relaciones “es esposo(a)” y “es hijo(a)” respectivamente; estos cuentan con un conteo muy similar. El resto de parentescos con una menor frecuencia son ‘no hay parentesco’, ‘padre o madre’, ‘cuñado’, entre otros. Debido a su carácter desbalanceado y al alto número de categorías, la variable puede no tener gran importancia en la forma actual, por lo que será modificada posteriormente como parte del proceso de ingeniería de características (ver en Sección 4.3.4).

3.1.1.4 Variable de entidad

La variable de entidad representa, de forma codificada, la entidad en que se encuentra el registro del encuestado. En la Figura 4 se aprecia la frecuencia para cada una de las entidades federativas.

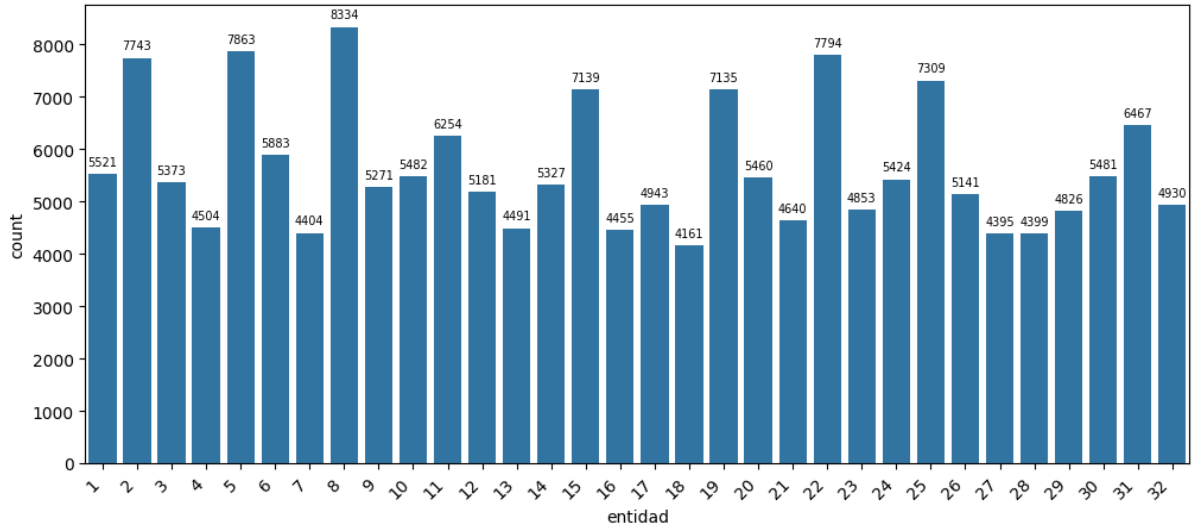

Figura 4: Gráfica de distribución de entidades.

La gráfica muestra algunos estados que cuentan con una mayor cantidad de registros que el resto de entidades, como Baja California, Coahuila, Chihuahua, Estado de México, Nuevo León, Querétaro y Sinaloa. Debido a esta dispersión de frecuencias y al alto número de valores distintos para la variable, se considera poco pertinente trabajar con cada entidad de manera individual. Con el objetivo de simplificar y representar de mejor manera estos valores, se realiza un agrupamiento por zonas dentro de la sección de ingeniería de características (ver en Sección 4.3.2)

3.1.1.5 Variable de clave_max

La columna “clave_max” representa una codificación de acuerdo a la forma en que se obtuvieron ciertos ingresos, no necesariamente de un empleo. Debajo, en la Figura 5, se muestra el conteo de cada una de las categorías para los ingresos.

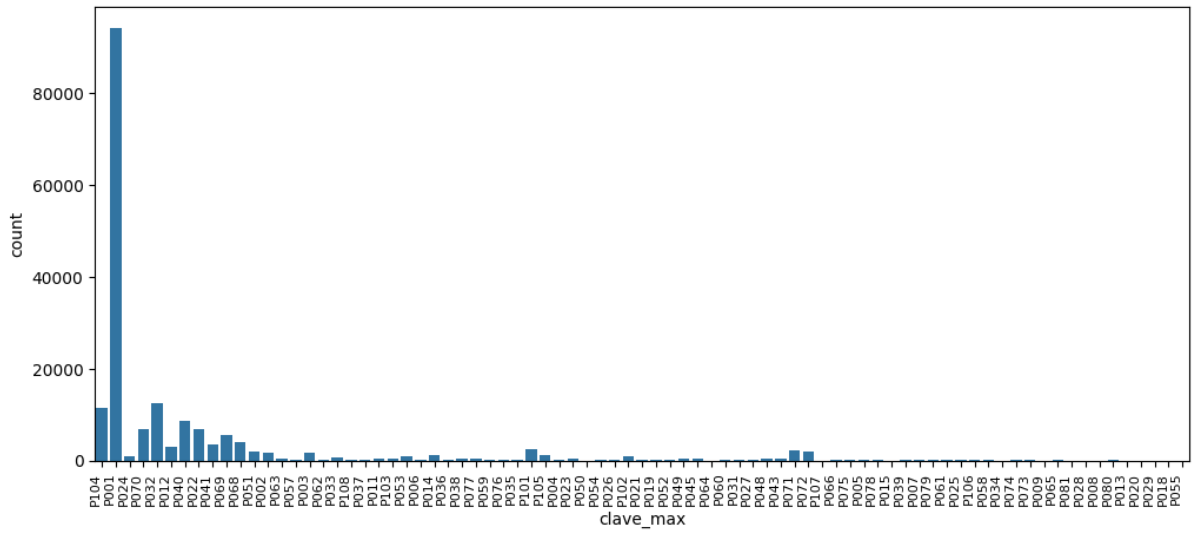

Figura 5: Gráfica de distribución de clave de ingresos

Lo más destacable de la gráfica es el predominio de la clave “P001”, la cual acapara más de la mitad de los registros (94,112). Esta clave representa aquellos ingresos obtenidos a partir de un trabajo subordinado en forma de sueldo o salario. En menor medida, le siguen las claves “P032”, “P104” y “P040”, que representan jubilaciones, ingresos del programa de bienestar para adultos mayores y donativos de otros hogares, respectivamente. Debido a este gran desbalance, esta variable igualmente se transforma a un predictor binario (ver en Sección 4.3.4).

3.1.2 Análisis bivariado: exploración de relaciones entre variables

Tras un análisis de las variables individuales, se realiza una comprensión de la relación de algunas de ellas con la variable objetivo de ingreso binario. Para esto, se emplean gráficos de barras apiladas utilizando las etiquetas 0 y 1, las cuales hacen

referencia a los registros de individuos con ingreso trimestral menor o mayor a 25 mpm respectivamente; asimismo, se utiliza un azul oscuro para denotar a la clase 0, y un azul claro para la clase 1. Es posible entonces observar la relación entre las diferentes variables y su comportamiento con las etiquetas binarias en situaciones sociodemográficas influyentes tales como si la persona es hombre o mujer, su nivel de escolaridad, el estado civil, la edad, si es hablante indígena, entre otras.

3.1.2.1 Distribucion de ingresos binarios por sexo

La Figura 6 presenta un gráfico de barras apiladas etiquetadas con las categorías 1 y 2 que representan a las categorías de hombre y mujer respectivamente, junto con la composición por tipo de ingreso (menor o mayor que 25 mpm).

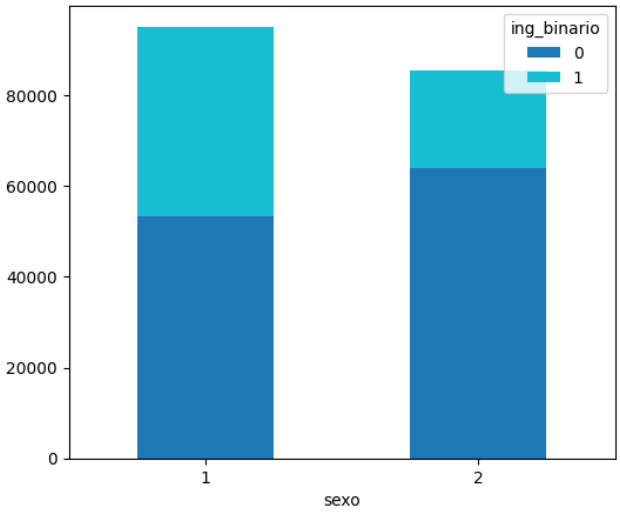

Figura 6: Gráfica Distribucion de Ingresos Binarios por Sexo

La figura mencionada permite observar una composición desigual en la que los hombres cuentan notoriamente con un número mayor de registros por encima de 25 mpm. Esto podria revelar que el factor de sexo puede ser una variable relevante para la predicción del modelo.

3.1.2.2 Distribucion de ingresos binarios por nivel aprobado (nivelaprob)

A continuación se presenta la Figura 7 donde el eje horizontal representa el nivel de estudios máximo alcanzado por cada individuo (donde 0 es ninguno, 1 es preescolar, 2 es primaria, 3 es secundaria, 4 es preparatoria, 5 es normal, 6 es carrera técnica, 7 es profesional, 8 es maestría y 9 es doctorado), mientras que el eje vertical representa la frecuencia de personas en cada categoría de ingresos (menor o mayor a 25 mpm). Similarmente, en la Figura 8 se presenta la misma variable pero esta vez normalizada para mostrar la proporción de las etiquetas de ingreso en el eje vertical.

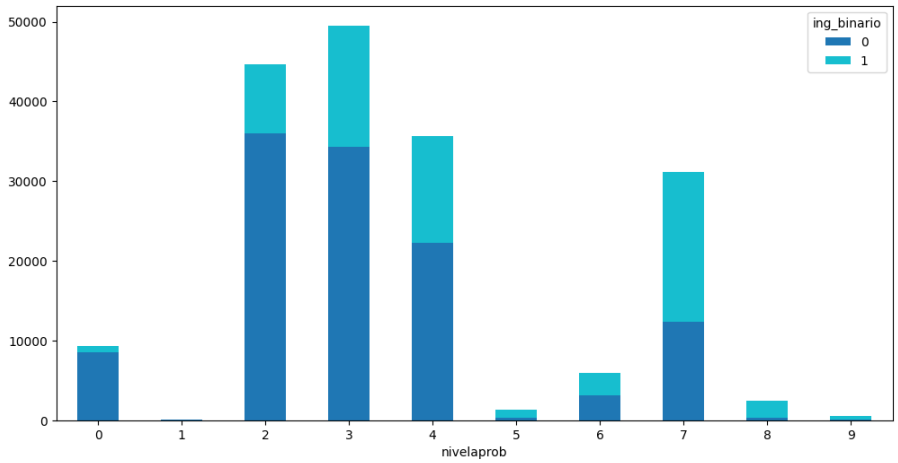

Figura 7: Gráfica de nivel máximo de estudios

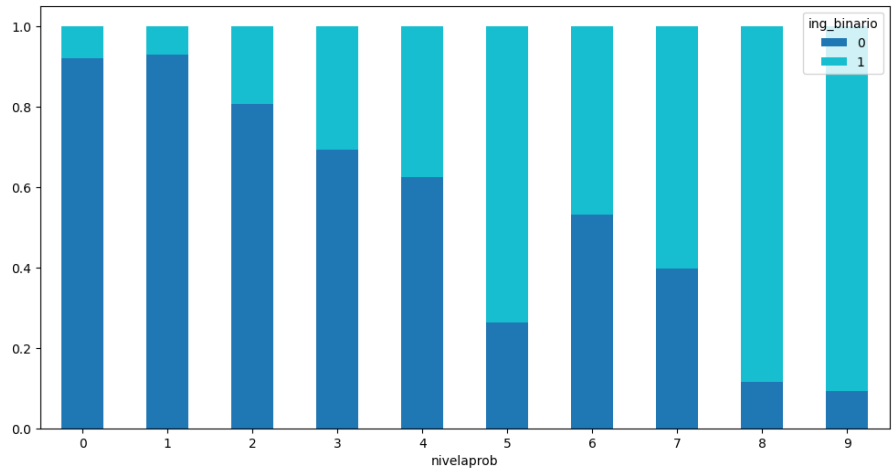

Figura 8: Gráfica de nivel máximo de estudios normalizada

Al observar la figura 7 es posible notar la prevalencia de categorías pertenecientes a una etapa inicial o intermedia de educación tales como la primaria, secundaria y preparatoria (categorías 2 a 4). A estas le sigue en frecuencia la categoría de estudios profesionales (grupo 7). En menor medida se registran individuos con estudios nulos o de preescolar, así como estudios avanzados de maestría y doctorado.

Por otra parte, en cuanto a la distribución de ingresos, es apreciable en la Figura 8 un aumento gradual en la proporción de personas con una remuneración por encima de 25 mpm trimestrales conforme se incrementa el nivel de aprobación. Esto indica que la variable en cuestión podría ser un buen predictor para el modelo de clasificación. Sin embargo, debido al desbalance significativo observado en las frecuencias de las categorías, se propone un agrupamiento distinto de los registros en la Sección 4.3.5.

3.1.2.4 Distribucion de ingresos binarios por estado conyugal (edo_conyug)

Posteriormente se presenta la Figura 9 para la variable “edo_conyug” donde el eje de las abscisas representa los diferentes estados conyugales y el eje de las ordenadas representa el número de personas en cada categoría de ingresos. Similarmente, en la Figura 10 se muestran estos mismos datos, pero aplicando una normalización al eje vertical para mostrar la proporción de registros de cada etiqueta de ingreso.

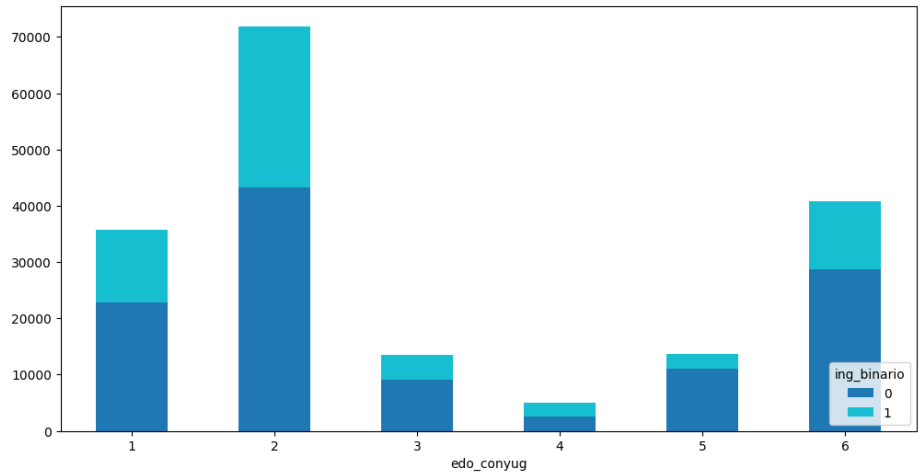

Figura 9: Gráfica estado conyugal

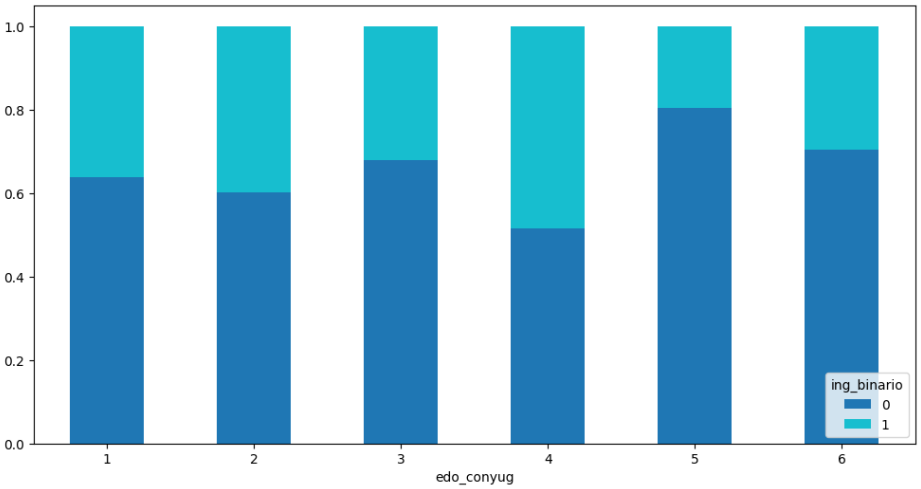

Figura 10: Gráfica estado conyugal normalizada

La Figura 9 muestra una mayor frecuencia de registros para las personas pertenecientes a la categoría 2 que denota a los individuos casados, seguidos de los registros de categoría 6 de soltero y los de categoría 1 que hace referencia a la unión libre. Los estados conyugales 3, 4 y 5 (separado, divorciado y viudo) cuentan con una cantidad considerablemente menor de individuos. Asimismo, es apreciable que las personas casadas cuentan con el mayor número de casos con un ingreso superior a 25 mpm trimestrales. Proporcionalmente, esto sigue manteniéndose al observar en la

Figura 10 que la categoría 2 cuenta con un 40% de individuos con una remuneración por encima del umbral y es el segundo estado conyugal con mayor proporción, sólo por debajo de la categoría 4 de divorciados, aunque en frecuencia esta última registra muy pocas observaciones. Se aprecia igualmente que la categoría 1 es similar en proporción a la categoría 2.Por otra parte, las categorías 5 y 6 de personas viudas y solteras cuentan con la menor proporción de retribuciones altas.

Debido al desbalance existente en las frecuencias de las categorías, esta variable se transforma a un predictor binario en la Sección 4.3.3 para aprovechar la distinción que existe en el ingreso para el grupo de personas en algún tipo de relación y aquellas que se encuentran fuera de estas.

3.1.2.6 Distribucion de ingresos binarios por edad

A continuación se presenta la Figura 11 donde el eje horizontal representa la edad, desde los 18 hasta los 109 años, mientras que el eje vertical representa la frecuencia de registros asociados; por otra parte, en la Figura 12, el eje vertical representa la proporción de individuos para las categorías de ingreso.

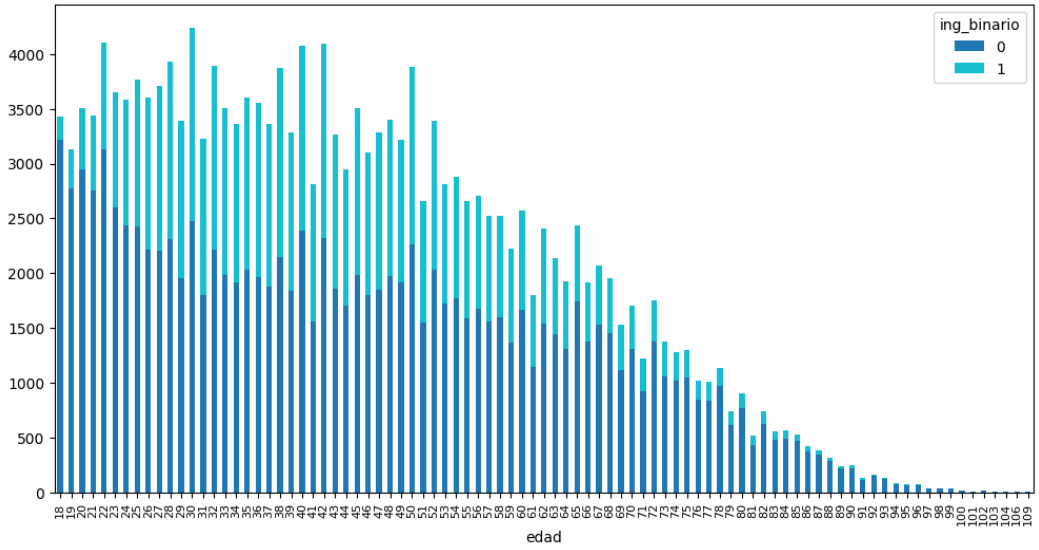

Figura 11: Gráfica de edad

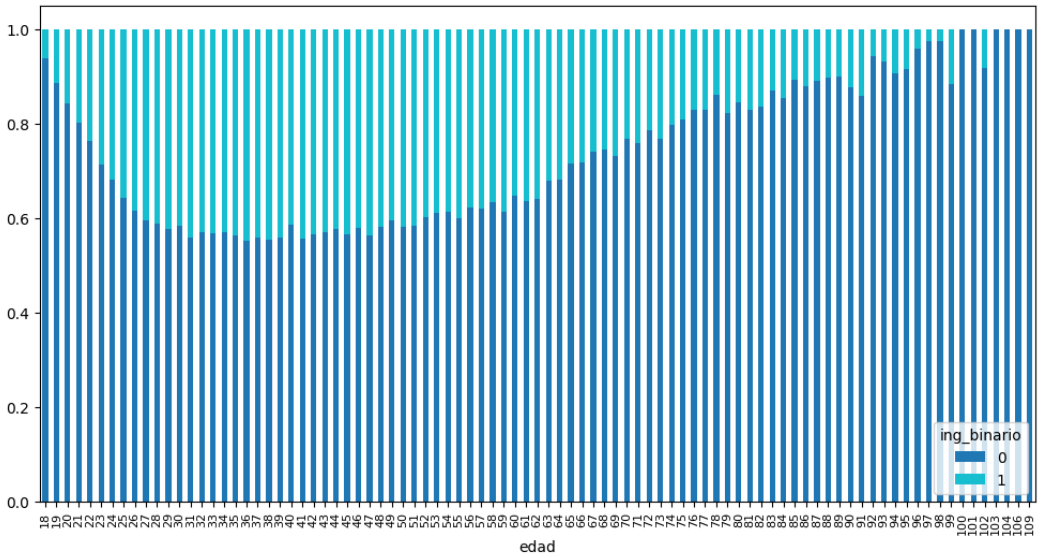

Figura 12: Gráfica de edad normalizada

En la Figura 11 es posible observar la prevalencia de personas con ingreso superior a 25 mpm al trimestre dentro del rango de los 31 a 50 años. Además, es apreciable un declive significativo en la frecuencia general de individuos a partir de esta última edad y una disminución aún más rápida en la frecuencia de personas pertenecientes a la parte superior del umbral monetario.

Este mismo fenómeno es observable en la Figura 12, en la que existe un incremento en proporción de individuos con ingreso mayor a 25 mpm entre los 18 y 31 años. Posteriormente, el porcentaje se mantiene constante en un 40% del total de registros de cada edad hasta aproximadamente los 52 años; después de este punto, la proporción comienza a descender.

3.1.2.8 Distribucion de ingresos binarios por hablante indigena (hablaind)

A continuación se presenta en la Figura 13 un gráfico de barras cuyo eje horizontal representa con las etiquetas 1 y 2 a las personas que hablan o no alguna lengua indígena respectivamente. Estos datos se muestran nuevamente en la Figura 14 de forma normalizada.

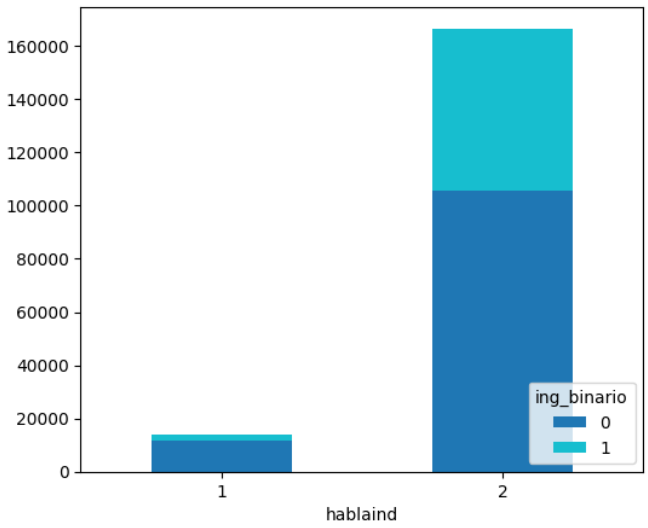

Figura 13: Gráfica de hablantes indígenas

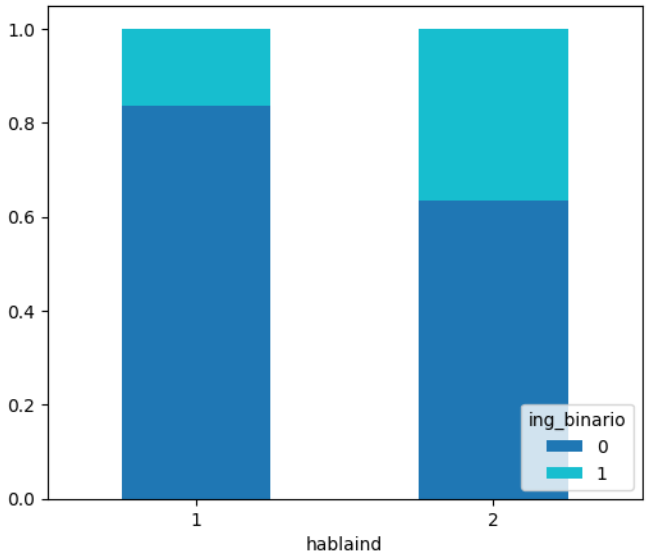

Figura 14: Gráfica de hablantes indígenas normalizada

Es evidente en la Figura 13 la baja frecuencia de hablantes indígenas frente al resto de la población. Además, se tiene que muy pocos de ellos cuentan con un ingreso superior al umbral de 25 mpm. En efecto, al observar la Figura 14 se muestra que más del 80% de los hablantes indígenas cuenta con un ingreso inferior, frente al 60% del resto de los individuos. Esto indica un claro contraste cultural ante las oportunidades económicas.

4. Preparación de los datos

Con los procedimientos de exploración realizados a lo largo del análisis univariado y bivariado se observaron oportunidades para modificar datos faltantes, simplificar variables y categorías, así como renombrar atributos que podrían llevar a un modelo de aprendizaje automático con mayor desempeño e interpretabilidad. Por ello, en

esta sección se lleva a cabo la limpieza de datos mediante la corrección de inconsistencias en ciertas variables y la imputación de valores faltantes. Asimismo, se realizan procedimientos de ingeniería de características para combinar variables y atender los puntos vistos previamente en la Sección 3.

4.1 Limpieza de datos

4.1.1 Valores faltantes y errores de datos

Durante el proceso de exploración se encontró la aparición de valores faltantes representados como caracteres de espacio y signos *ampersands,* específicamente dentro de las variables de discapacidad. El listado de variables y sus respectivos conteos del número de registros sin valores se encuentra en la Tabla 2. Estos campos vacíos fueron sustituidos por el valor NaN[3] para una posterior imputación.

Variable	***Valores encontrados***	***Conteo de faltantes***
hor_1	Caracter de espacio	64372
num_trabaj	Caracter de espacio	40841
*disc_*[4]	&	238
ss_aa	Caracter de espacio	82320
ss_mm	Caracter de espacio	82320
insc_1	Caracter de espacio	127168
insc_2	Caracter de espacio	169095
insc_3	Caracter de espacio	166133
insc_4	Caracter de espacio	178288
insc_5	Caracter de espacio	177715
insc_6	Caracter de espacio	179017
insc_7	Caracter de espacio	174965
insc_8	Caracter de espacio	180361

Tabla 2: Valores faltantes encontrados en variables

[3] Estas son las siglas de “Not a number” y se utiliza el valor para un manejo sencillo de imputación en la librería de Pandas.

[4] Todas las variables de discapacidad con el prefijo “disc_” cuentan con los mismos registros y conteo de valores faltantes.

Asimismo, se observaron inconsistencias en los tipos de datos contenidos en la variable de atención médica “atemed”. Por un lado, el campo presentaba valores numéricos 1 y 2 para ciertos registros, mientras que para otros, los valores se encontraban codificados como cadenas de texto (“1” y “2”). Estos últimos datos fueron convertidos a su representación numérica correspondiente para obtener una variable íntegra y coherente.

4.1.2 Imputación de valores faltantes

Tras la identificación de los valores faltantes en cada variable se llevaron a cabo distintos procesos de imputación con el fin de obtener un *dataset* completo. Notablemente, para la variable de número de trabajos “num_trabaj” se realizó una deducción lógica a partir del valor registrado en el campo de horas trabajadas la semana anterior “hor_1”, así como la edad del individuo. De manera general, resulta razonable asumir que si las horas trabajadas son mayores a cero, entonces el individuo debe contar más probablemente con un único trabajo, debido a que esta es la moda para la variable. Por otra parte, en las situaciones en las que no se registra algún número de horas, resultó observable que la mayoría de estos individuos pertenece a los grupos etarios mayores, que usualmente cuentan con algún tipo de pensión o apoyo monetario; por ello, la variable se establece como cero trabajos para la población por encima de 40 años. Un razonamiento similar ocurre para las personas jóvenes para las que es más probable su involucramiento en alguna actividad laboral remunerada. De esta forma, la asignación completa de valores para “num_trabaj” dependiendo de las características del registro se visualiza en la Figura 15.

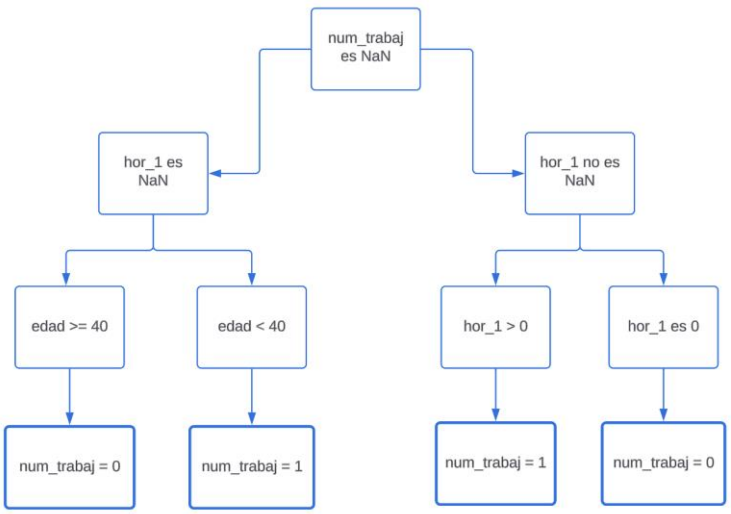

Figura 15: Proceso de imputación para “num_trabaj”

Por otro lado, debido a la naturaleza dispersa de la variable de horas trabajadas (“hor_1”) y al alto número de valores faltantes en ella, se decidió eliminarla del conjunto de datos finales para el entrenamiento del modelo. De otro modo, realizar una imputación con la media habría implicado introducir un gran desbalance en la distribución con un valor modificado por valores atípicos por encima de cien horas.

Finalmente, para el resto de variables de discapacidad, tiempo de contribución a la seguridad social (“ss_aa” y “ss_mm”), y la causa de afiliación de un seguro médico (“inscr_”) fueron imputadas utilizando la moda, la media, o el valor 0 respectivamente. La síntesis del proceso de asignación de valores a registros vacíos se muestra en la Tabla 3.

Variable	***Estrategia de imputación***
hor_1	Eliminada por dispersión y cantidad de valores faltantes.
num_trabaj	Imputada siguiendo el esquema de Figura 15.
*disc_*[5]	Imputada con el valor 4 (sin dificultad de realizar acciones), el cual es la moda de la variable.
ss_aa	Imputada con la media.
ss_mm	Imputada con la media.
*inscr_*[6]	Imputada con el valor 0 para indicar desconocimiento o falta de afiliación.

Tabla 3: Estrategias para imputar valores faltantes

4.3 Ingeniería de características

De forma general, la ingeniería de características consistió en una reducción de categorías y el balanceo de estas mismas con el fin de obtener un mejor entrenamiento y rendimiento del modelo, esto debido a la disminución del sesgo en ciertas variables. Debajo de cada modificación se muestra el resultado de las distribuciones.

4.3.1 Atributo: discapacidad

Se redujeron las siete columnas que representaban la dificultad que un individuo tenía respecto a ciertos niveles de discapacidad en diferentes áreas, como la visión, el habla, la movilidad, etc. En estas columnas el registro con 1 es una gran dificultad para que la persona ejecute una acción y 4 ninguna dificultad. Esto se aprovechó para

[5] Todas las variables de discapacidad con el prefijo “disc_” fueron imputadas de la misma forma.

[6] Todas las variables de causas de afiliación con el prefijo “inscr_” fueron imputadas de la misma forma.

revisar si un registro contaba con al menos un valor menor a 4 y así identificar a las personas con alguna discapacidad en función de valores menores a 4, es decir, crear una variable binaria que indique la presencia o ausencia de discapacidad en el conjunto de datos, 1 (dificultad en discapacidad), 0 (sin dificultad). Esta nueva columna tiene nombre “discapacidad” y maneja solo los valores binarios 1 y 0, como se puede observar en la Figura 16.

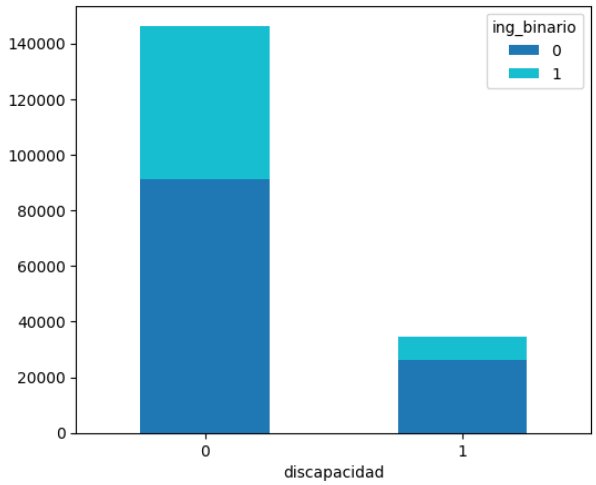

Figura 16: Gráfica de discapacidad

4.3.2 Atributo: zona

Asimismo, se trabajó con la columna ‘entidad’ que indica la entidad federativa a la que el registro pertenece. Para no atender los 32 estados de forma independiente, se optó por juntarlos por zona geográfica en el país, siendo dividida en norte, centro y sur (100, 200 y 300; las claves de identificación por zona respectivamente). La elección de los estados por zona es mostrada debajo en la Tabla 4, así como la distribución de sus registros en la Figura 17 la cual ahora se encuentra mejor balanceada.

Zonas	***Estados que la comprenden***
Norte	Baja California, Baja California Sur, Coahuila de Zaragoza, Chihuahua, Durango, Nuevo León, San Luis Potosí, Sinaloa, Sonora, Tamaulipas y Zacatecas.
Centro	Aguascalientes, Ciudad de México, Guanajuato, Hidalgo, Jalisco, México, Michoacán de Ocampo, Morelos, Nayarit, Querétaro y Tlaxcala.
Sur	Campeche, Colima, Chiapas, Guerrero, Oaxaca, Puebla, Quintana Roo, Tabasco, Veracruz de Ignacio de la Llave y Yucatán.

Tabla 4: Composición de zonas

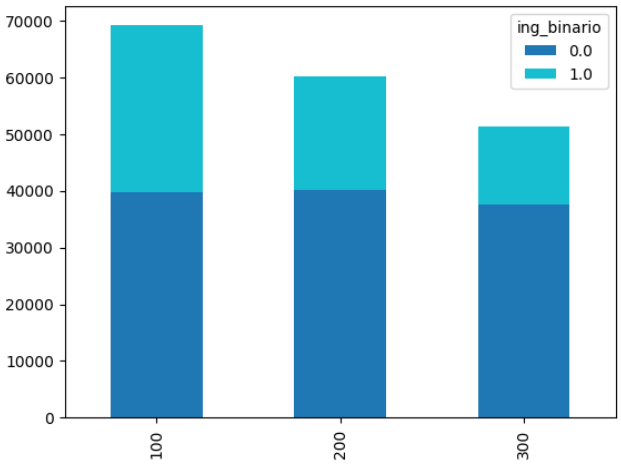

Figura 17: Gráfica mejor balanceada de “zona”

4.3.3 Atributo: ‘en_pareja’

De igual manera, se agruparon las categorías de “edo_conyug” que representa el estado conyugal de un registro. Anteriormente se tenían 6 categorías, pero ahora fueron reducidas a si hay una situación de pareja (casado, unión libre) o no (soltero, divorciado, separado, viudo); se representan por 1 y 2 respectivamente en la columna “en_pareja”, como se puede apreciar debajo en la Figura 18.

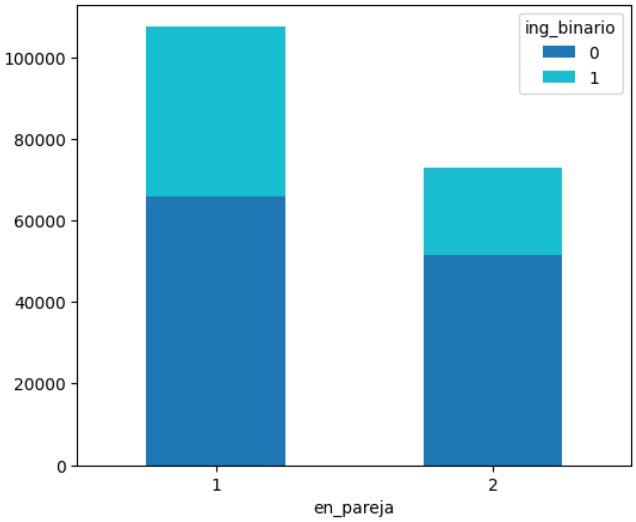

Figura 18: Gráfica balanceada de ‘en_pareja’

4.3.4 Atributos: ‘es_jefe’ y ‘asalariado’

Para ‘parentesco’ y ‘clave_max’ se realizó una reducción a un sentido binario de la columna principalmente para balancear la cantidad de valores en las categorías. En parentesco se opuso el valor 101, que representa al jefe de familia, con el resto de etiquetados, renombrando así la columna a ‘es_jefe’ a 1 y 2 (sí y no). Para clave_max se emplea la clave ‘P001’, que representa al trabajo asalariado, con el resto de

categorías para renombrar la columna ‘asalariado’. Las gráficas finales se muestran en la Figura 19.

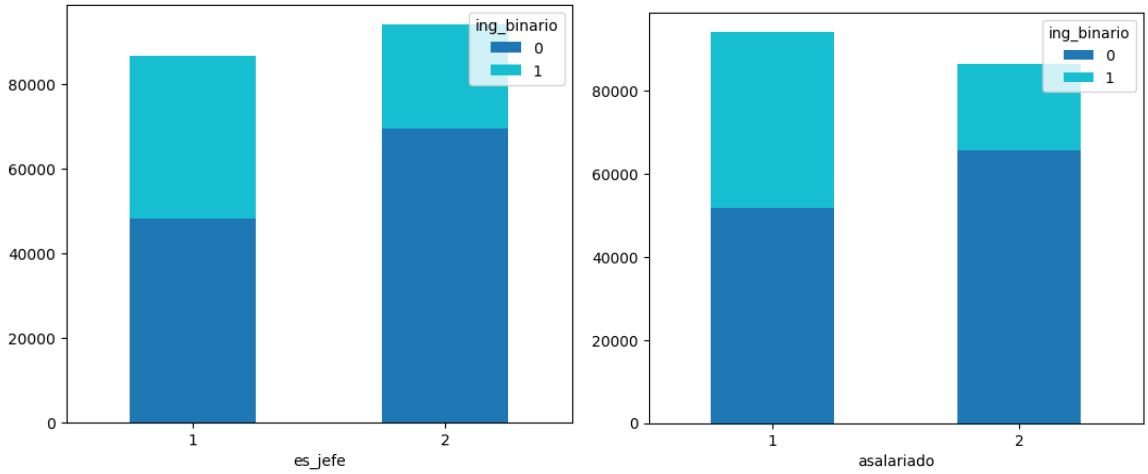

Figura 19: Gráficas balanceadas de ‘es_jefe’ y ‘asalariado’

4.3.5 Atributo: ‘nivelaprob’

Para esta característica se consideró una reducción de categorías de tal forma que se unificaran los niveles de escolaridad en tres categorías generales: de 0 a 2 agrupados en la categoría 0; de 3 a 5 agrupados en 1; y niveles 6 a 9 agrupados en 2. De esta manera, se reduce la cardinalidad de la variable a un sentido de nivel bajo (o nulo), medio y alto de escolaridad. En la Figura 20 se expone la distribución de los registros.

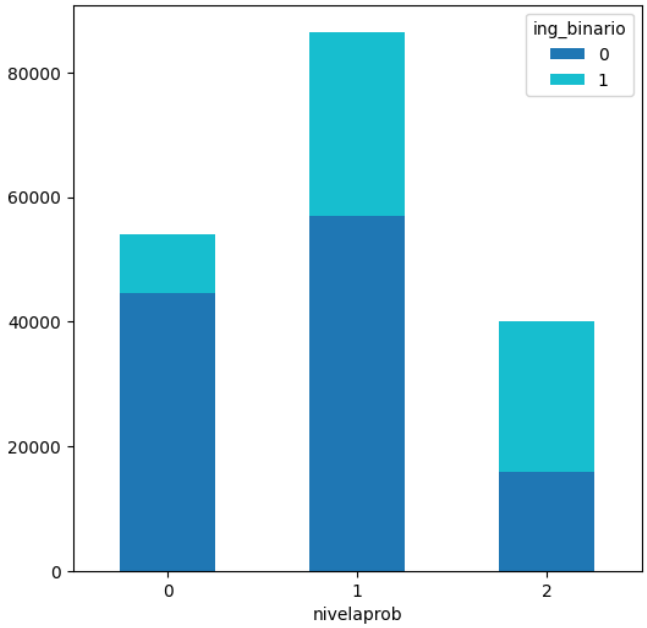

Figura 20: Gráfica balanceada de ‘nivelaprob’

4.3.6 Atributo: ‘seg_meses_tot’

Para esta característica se consideró una reducción de categorías de tal forma que se unificaran dos columnas: ‘ss_aa’ y ‘ss_mm’ que indican el tiempo de contribución en años y meses respectivamente a la seguridad social. Para hacer esto solo se convierte la columna de años en meses, multiplicando ‘ss_aa’ por 12 y sumando la columna de meses se obtiene la nueva columna **‘seg_meses_tot’** que indica el número de meses totales de contribución a la seguridad social.

4.3.7 Atributo: ‘inscr’:

De igual manera, se agruparon las columnas de ‘inscr_1’, ‘inscr_2’, ‘inscr_3’, ‘inscr_4’, ‘inscr_5’, ‘inscr_6’, ‘inscr_7’, que son los motivos de afiliación o inscripción, estas eran columnas binarias donde la mayoría de los registros presentaba solamente una razón por afiliación a la salud por lo que se decidió juntarlas en una sola columna que tomara una de las razones de las 7 categorías y así crear una nueva columna llamada ‘inscr’ que indicaría una razón por la que una persona se afilió a una institución médica.

4.4 Selección de variables finales

Posterior al análisis realizado y a la modificación de las variables elegidas, se realizó una matriz de correlación, donde es posible revisar la correlación para cada par de características. Mediante la representación de la Figura 21 se logra observar qué columnas se encuentran relacionadas de cierta manera con las demás, y se determina si deben eliminarse por presentar información redundante y así reducir la dimensionalidad del problema.

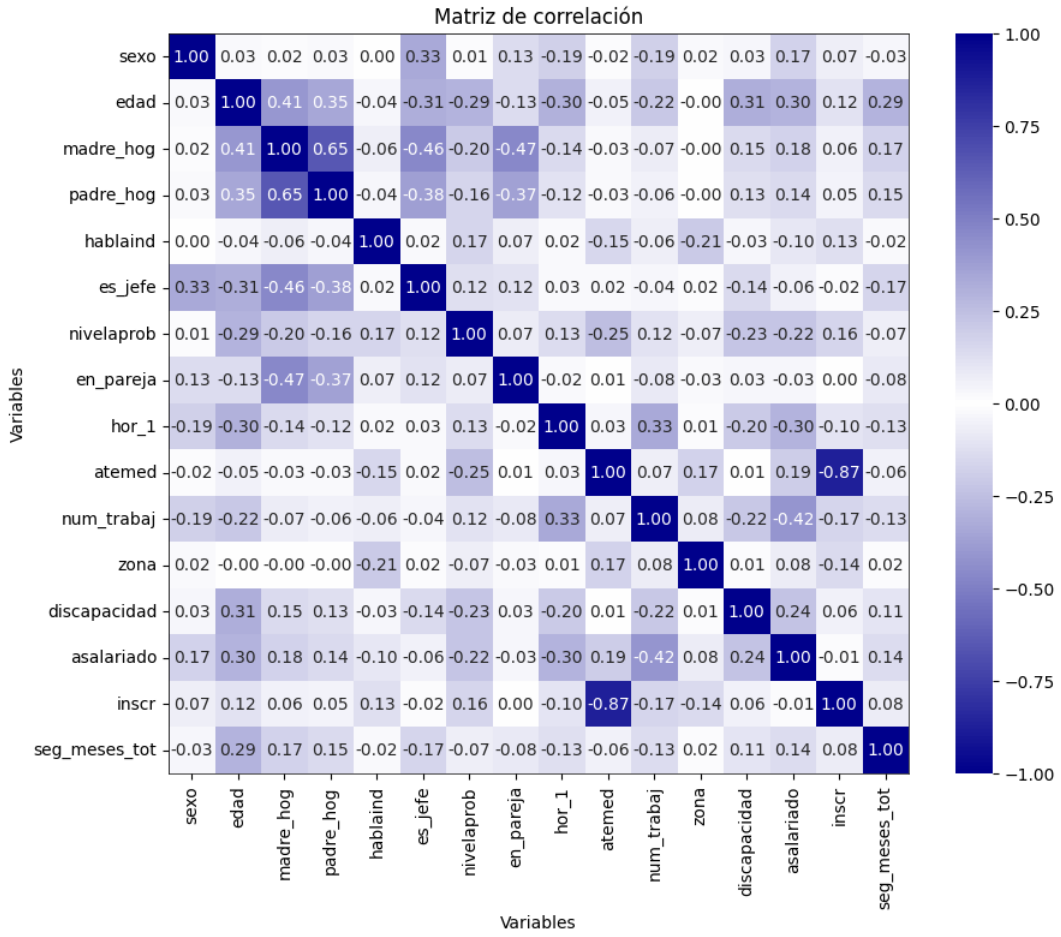

Figura 21: Matriz de correlación entre variables

Tanto madre_hog como padre_hog cuentan con una alta correlación entre ellos mismos, así también en variables como en_pareja, es_jefe y edad. Asimismo, atemed e inscr muestran la mayor correlación de toda la matriz.

A continuación se presenta la Tabla 5, la cual muestra la selección de las variables que serán parte del modelado y su posterior evaluación. Con asterisco se distinguen aquellas variables que fueron ingeniadas a partir de las características originales de la base de datos.

Variables finales

sexo edad hablaind num_trabaj	nivelaprob* en_pareja* inscr* es_jefe*	zona* discapacidad* asalariado* seg_meses_tot*

Tabla 5: Selección de variables finales

5. Modelado

Después de asegurar la integridad de la base de datos mediante la limpieza e ingeniería de características para la selección de variables relevantes y completas es posible comenzar con el modelado de árboles de decisión con el fin de predecir si un individuo mayor de edad cuenta con un ingreso trimestral igual o mayor a 25 mpm o si, por el contrario, se encuentra debajo del umbral. Así, se presenta a continuación la experimentación de modelos a partir del uso de la librería Scikit Learn en el lenguaje de programación de Python. .

5.1 Árbol de decisión

Según la biblioteca de aprendizaje automático Scikit Learn (s.f.), “los árboles de decisión son métodos no paramétricos de aprendizaje supervisado, utilizados para clasificación y regresión”. La meta de ello es crear modelos que predigan mediante reglas de decisión que dividan el dataset y realice la estimación o clasificación.

Algunas ventajas que presentan estos modelos son:

- Facilidad de interpretar y visualizar.
- Poca preparación de datos necesaria.
- Posibilidad de utilizar tanto datos numéricos como categóricos.

Sin embargo, se presentan algunas desventajas:

- Es necesario un manejo de hiperparámetros para evitar el *overfitting*.
- Son inestables al percibir variaciones en los datos que cambian la estructura del árbol.

5.2 Metodología

5.2.1 División de datos

Para el entrenamiento y la prueba de los modelos, se realizó una partición de los registros disponibles en la base de datos, de tal forma que estos sean independientes entre sí. La distribución se conforma de 80% de las instancias para el entrenamiento y el 20% restante para la prueba.

5.2.2 Experimentación

La realización de los experimentos se basó en pruebas sistemáticas que mantuvieran las mismas condiciones para las distintas derivaciones de árboles de decisión. Por ello, permanecen constantes las variables elegidas, así como la partición inicial de entrenamiento y prueba de los datos para cada variación de modelo.

Las variaciones del árbol de decisión se realizaron mediante la implementación de una *Grid Search*, una herramienta que busca exhaustivamente entre las

combinaciones en el espacio de hiperparámetros ya predefinido y mantiene aquella configuración que genera un mejor desempeño (Chan et al., 2015). En la Tabla 6 se muestran los hiperparámetros que fueron incluídos en el espacio de búsqueda para la experimentación y optimización de los modelos, así como los valores que puede tomar cada uno.

Hiperparámetros	***Valores***			
criterion	'gini'	'entropy'		
max_depth	3	5	7	None
min_samples_split	2	5	10	
min_samples_leaf	1	2	4	

Tabla 6: Espacio de búsqueda de Grid Search

Adicionalmente, para pruebas posteriores se empleó un *Random Resampling*, una técnica que modifica el desbalance de la variable objetivo de tal forma que reduce el sesgo hacia una clase y permite una mejor generalización por parte del modelo (Zhu et al., 2020). Para esta técnica, es posible emplear dos estrategias: *Oversampling*, que aumenta los registros de la clase minoritaria; y *Undersampling*, que reduce los registros de la clase mayoritaria.

De esta manera, se implementaron cuatro modelos de árbol de decisión marcados en la Tabla 7 y acompañados con una breve descripción. Los modelos simple y simple optimizado fueron entrenados utilizando datos de salida desbalanceados, con el 65% de las etiquetas pertenecientes a la clase de ingreso trimestral superior a la media nacional. Por otra parte, los modelos optimizados de *Oversampling* y *Undersampling* equilibran la proporción de las etiquetas.

Modelo	***Descripción***
Simple	Decision Tree sin cambios respecto a la implementación predeterminada de la librería Scikit Learn.
Simple Optimizado	Decision Tree con hiperparámetros *criterion, max_depth, min_sample_split* y *min_samples_leaf* optimizados por Grid Search.
Optimizado con Oversampling	Decision Tree con hiperparámetros *criterion, max_depth, min_sample_split* y *min_samples_leaf* optimizados por Grid Search y registros resampleados por Oversampling.
Optimizado con Undersampling	Decision Tree con hiperparámetros *criterion, max_depth, min_sample_split* y *min_samples_leaf* optimizados por Grid Search y registros resampleados por Undersampling.

Tabla 7: Modelos utilizados en experimentación

5.2.3 Métricas de desempeño

Con el fin de evaluar el rendimiento de los modelos y compararlos entre sí, se utilizan cuatro valores principales resultantes de una predicción en particular. Específicamente, se tienen los valores

- Verdadero Positivo (TP): Predicho verdadero y verdadero en realidad. En este contexto, esto implica que el modelo tiene una predicción de ganancia **por encima** de 25 mpm, mientras que el valor real concuerda con este ingreso.
- Verdadero Negativo (TN): Predicho falso y falso en realidad. Es decir, el modelo tiene una predicción de ganancia **por debajo** de 25 mpm, mientras que el valor real concuerda con este ingreso.
- Falso Positivo (FP): Predicción de verdadero y falso en la realidad. Esto implica una predicción de retribución por encima de 25 mpm, pero un ingreso real por debajo del umbral.
- Falso Negativo (FN): Predicción de falso y verdadero en la realidad. Esto implica un valor predicho por debajo de 25 mpm trimestrales, pero un ingreso real por encima de esta cantidad.

Para evaluar cada uno de los modelos anteriores se utilizan las siguientes métricas (Chauhan, 2023):

- Accuracy (Exactitud): la métrica accuracy representa el porcentaje total de valores correctamente clasificados, tanto positivos como negativos. Su valor se calcula de acuerdo a la fórmula: $\frac{TP + TN}{TP + TN + FP + FN}$
- Precision (Precisión): esta métrica representa el porcentaje de valores que el modelo ha clasificado como positivos de entre las instancias positivas clasificadas por el modelo, mide la exactitud de las predicciones positivas del modelo. Su valor se calcula de acuerdo a la fórmula: $\frac{TP}{TP + FP}$
- Recall: esta métrica también conocida como el ratio de verdaderos positivos, es el porcentaje de proporción de las instancias realmente positivas que fueron correctamente identificadas por el modelo. Su valor se calcula de acuerdo a la fórmula: $\frac{TP}{TP + FN}$
- F1- score: es la media armónica de la precisión y el recall. Su valor se calcula de acuerdo a la fórmula: $\frac{2(Precision \times Recall)}{Precision + Recall}$

Cada una de las métricas toma valores entre cero y uno, en donde el cero indica el peor rendimiento posible, y el uno indica el mejor rendimiento posible donde el modelo no tiene equivocaciones. Por lo tanto, mientras más cerca estén las métricas de uno, mejor es el rendimiento del modelo en términos de clasificación y predicciones.

5.3 Resultados

Tras la experimentación realizada con los modelos previamente mencionados en la Tabla 7, se obtienen los resultados de desempeño en *accuracy, precision, recall* y *F1-score* de la Tabla 8. A continuación se ofrece un análisis de estos valores.

Modelo	**TP**	**FP**	**TN**	**FN**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Simple	6955	4456	18942	5764	0.7170307	0.6094996	0.5468197	0.5764608
Optimizado	7259	2816	20582	5460	0.7708558	0.7657429	0.7708558	0.7636931
Optimizado con Oversampling	7642	5415	17983	5077	0.7094996	0.71132042	0.7094996	0.7103475
Optimizado con Undersampling	9364	5691	17707	3355	0.7495362	0.7636829	0.7495362	0.7534892

Tabla 8: Resultados de modelos

- Modelo simple: registra una precisión moderada y un *recall* relativamente bajo, lo que indica que, aunque una parte significativa de las predicciones positivas son correctas, el modelo no identifica todas las instancias positivas. El *F1-score* es relativamente bajo y refleja equilibrio moderado entre precisión y *recall.*
- Modelo optimizado: muestra mejoras significativas en todas las métricas en comparación con el modelo simple. La precisión y el recall son ambos altos, lo que sugiere que el modelo es tanto preciso en sus predicciones positivas como efectivo en la identificación de instancias positivas. El *F1-score* es el más alto entre todos los modelos, indicando un buen equilibrio entre precisión y *recall.*
- Modelo optimizado con oversampling: registra un valor de *precision* y *recall* equilibrados, ambos en torno a 0.71. Sin embargo, su exactitud es ligeramente inferior a la del modelo optimizado. El *F1-score* igualmente es alto, indicando un buen equilibrio, pero ligeramente inferior al del modelo optimizado sin *oversampling.*
- El modelo optimizado con undersampling cuenta con la mayor cantidad de verdaderos positivos (TP) y una precisión y recall equilibrados y altos. La exactitud es también alta y comparable a la del modelo optimizado sin técnicas de sampling. El *F1-score* es el segundo más alto, muy cercano al del modelo optimizado, indicando un excelente balance entre precisión y *recall.*

5.3.1 Selección y validación cruzada del mejor modelo

El modelo optimizado sin técnicas de *sampling* muestra el mejor equilibrio general con el *F1-score* más alto (0.7636), también en cuanto al resto de métricas resulta ser el más alto en cada una de ellas lo que sugiere que es el modelo más adecuado para

clasificar correctamente y predecir si una persona recibe un remuneración por encima o debajo de 25 mpm trimestrales.

Posteriormente, con el fin de verificar el correcto funcionamiento del modelo y descartar algún caso de sobreajuste, se realiza una validación cruzada utilizando cinco *folds.* De esta forma se obtiene una exactitud promedio de 0.77032 para el proceso de entrenamiento, y una exactitud promedio de 0.76865 para el proceso de validación. Esto representa una diferencia baja de 0.00167 e indica que el modelo no cuenta con sobreajuste, por lo que es capaz de generalizar sus predicciones a datos no vistos.

6. Evaluación

Una vez seleccionado el mejor modelo de árbol de decisión, es posible realizar distintas exploraciones para interpretar su comportamiento. En esta sección se analiza el desempeño del modelo específicamente para ciertos grupos sociodemográficos, se describen las características más importantes que considera para la clasificación de los individuos, y se discute su alcance y limitaciones.

6.1 Análisis de desempeño

Con el fin de observar el rendimiento del modelo en contextos sociodemográficos particulares, se obtuvieron los valores de exactitud para grupos etarios (Tabla 9), grupos de sexo (hombre y mujer, en Tabla 10), así como de la población minoritaria de hablantes indígenas (mostrados en Tabla 11). De esta forma, es observable en la Tabla 9 un mejor desempeño del modelo para el grupo de edades mayor de 54 años, el cual supera el 80% de exactitud frente al valor respectivo de 76% dentro de los años 18 a 34. Similarmente, un salto de rendimiento mucho más notorio se aprecia al probar el modelo separadamente por sexo: mientras que se obtiene una exactitud de 72% para los hombres, el rendimiento en el modelo para las mujeres es de 82%, es decir, existe un salto de diez puntos percentiles para estos dos grupos. Un fenómeno similar ocurre para los hablantes indígenas, que cuentan con una tasa de clasificación correcta por encima del 85%, mientras que para el grupo no hablante, el modelo se comporta con un rendimiento cercano al valor original de exactitud del 77% de la Tabla 8 de resultados.

Rango de edad	***Accuracy***
18 a 34	0.7623
34 a 51	0.7473
54 y mayor	0.8004

Tabla 9: Desempeño del modelo para grupos etarios

Sexo	*Accuracy*
Hombre	0.7220
Mujer	0.8250

Tabla 10: Desempeño del modelo por sexo

Hablante indígena	*Accuracy*
Sí	0.8566
No	0.7636

Tabla 11: Desempeño del modelo para hablantes indígenas

Los anteriores valores muestran que el modelo tiende a contar con un mejor rendimiento de clasificación en torno a grupos minoritarios o en desventaja frente a las oportunidades laborales y de remuneración. Por otra parte, para grupos hegemónicos como aquellos constituidos por hombres jóvenes no indígenas, el modelo tiende a desempeñarse de forma regular. Este comportamiento es fácilmente notorio, y podría llevar a una exploración sociodemográfica interesante para futuros estudios.

6.2 Análisis de características importantes

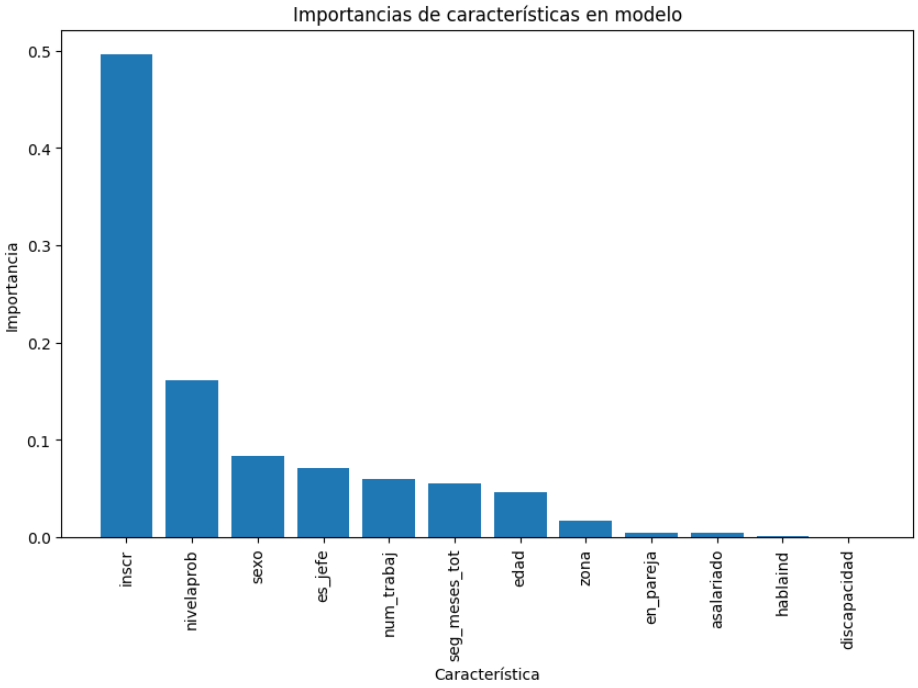

Figura 22: Gráfico de barras de importancia de la variable en el modelo

En la gráfica 22 se muestran la importancia de cada una de las características que componen al modelo árbol de decisión, se indica cuánto contribuye cada característica a las decisiones del modelo. En este caso:

Inscr (Afiliación a la salud):

- **Importancia Alta** : Esta característica es la más importante en el modelo. Esto nos puede indicar que la afiliación a la salud puede estar fuertemente correlacionada con los ingresos, ya que aquellos con acceso a servicios de salud podrían estar empleados en trabajos formales con mayores salarios o tener beneficios adicionales que incrementen sus ingresos y que les permitan afiliarse a una institución médica.

Nivelaprob (Escolaridad):

- **Importancia Moderada**: Esto nos puede indicar que la educación es un factor clave en la determinación de los ingresos. Un mayor nivel educativo generalmente se asocia con mejores oportunidades laborales y, por lo tanto, salarios más altos, que a la vez se asocia con la característica anterior de trabajos formales con mayores salarios que permiten afiliarse a una institución médica o tener más facilidad de acceder a servicios de la salud.

Sexo (Género):

- **Importancia Baja - Moderada** : Aunque tiene una menor importancia en comparación con las dos anteriores, el género puede influir en los ingresos debido a disparidades salariales basadas en el género y estereotipos socio culturales que se han impuesto en cuanto a roles de género.

Es_jefe (Cabeza de familia):

- **Importancia Baja**: nos indica que ser el jefe de la familia puede estar relacionado con responsabilidades económicas adicionales y, por ende, con la necesidad de mayores ingresos.

Num_trabaj (Número de trabajos):

- **Importancia Baja**: Tener múltiples trabajos puede indicar una mayor necesidad de ingresos o, alternativamente, podría reflejar mayores ingresos debido a la suma de varios empleos, sin embargo la importancia es baja y no nos podemos confiar de esta característica del todo para predecir la categoría de ingresos de una persona

6.3 Implicaciones y limitaciones del modelo y su uso

Realizar un modelo de esta naturaleza, que considere características demográficas y socioeconómicas de un país entero, no es una tarea fácil. La gran diversidad de contextos sociales que existe en México deriva en una variabilidad de las características de los grupos poblacionales, lo cual obstaculiza el emprendimiento y generalización de los modelos de aprendizaje automático, lo que posteriormente implica la medición de predicciones correctas de forma limitada.

Por otro lado, la decisión realizada de seleccionar únicamente un conjunto restringido de variables para simplificar la exploración y el entrenamiento del modelo implica la pérdida de información de las características que no fueron consideradas y que, probablemente, podrían haber mejorado el rendimiento obtenido. Asimismo, la elección de un árbol de decisión implica optar por la interpretabilidad del modelo por encima del desempeño de predicción, el cual puede conseguirse en modelos más complejos como redes neuronales a cambio de su entendimiento. Por último, al ser un modelo de clasificación, este es incapaz de predecir valores continuos de ingresos trimestrales; y, dado que la remuneración trimestral de la población puede llegar a ser muy variada en distintos estratos sociales, las predicciones binarias del modelo podrían aportar muy poca información de la situación económica de una persona.

7. Conclusión

En este reporte, se llevó a cabo un análisis exhaustivo de 180,583 registros y un entendimiento del contexto socioeconómico de México para desarrollar un modelo

clasificador de árbol de decisión para predecir si el ingreso de un individuo mayor de edad se encontraba por encima o debajo de 25 mil pesos mexicanos (mpm).

Partiendo de una totalidad de 192 variables, se realizó una selección previo análisis de 30 de ellas, las cuales fueron consideradas importantes para el propósito del proyecto. Con estas características, se trabajó en un análisis exploratorio donde se obtenía información relevante sobre sus composiciones, así como la distribución de los que ganaban más o menos de 25 mpm. Con ello, se dio a conocer las desigualdades de los ingresos en la población según las características que presentaban los encuestados.

Posteriormente, se trabajó en un proceso de limpieza de datos, donde se aplicaron distintas técnicas de imputación, desde utilizar la moda y media hasta la inclusión de un esquemático de decisión con apoyo de otra variable. Asimismo, se aplicó ingeniería de datos a siete características relevantes, como la discapacidad, parentesco o estado conyugal, las cuales adquirían un valor de utilidad para la posterior aplicación en el modelo predictivo.

En el proceso de experimentación, se utilizó una selección final de 12 características, donde se realizaron pruebas con el árbol de decisión y variaciones con Grid Search y Random Resample. De ello resultó que el modelo optimizado con Grid Search tuvo el mejor rendimiento con un F1-score de 0.7636. Asimismo, mediante el análisis de importancia de características, se definió que la variable que indicaba la razón de inscripción al seguro médico era la más importante.

Durante la evaluación del mejor modelo, se aprecia que este tiene un rendimiento mayor en las secciones de datos que representan a sectores de la población en desventaja o grupos vulnerables, como mujeres, adultos mayores o hablantes de alguna lengua indígena. Esto podría darse por la mayoría existente de los que presentan ingresos menores a 25 mpm, lo cual abre una posibilidad de abarcar el tema de la desigualdad en futuras investigaciones.

Sin embargo, implicaciones como la diversidad social y económica de la población mexicana y la sencillez de un modelo clasificatorio como el árbol de decisión limita la posibilidad de obtener un rendimiento más alto para realizar predicciones de ingreso trimestral.

8. Referencias

INEGI (2020). *Población total.* https://www.inegi.org.mx/temas/estructura/

INEGI (2023). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2022 Nueva serie.*

https://www.inegi.org.mx/programas/enigh/nc/2022/

SNIEG (s. f.). *Acerca del INEGI como Unidad Central Coordinadora del SNIEG.* https://www.snieg.mx/ucc_inegi_acerca_de/

https://scikit-learn.org/stable/modules/tree.html

Chan, S., & Treleaven, P. (2015). Continuous Model Selection for Large-Scale Recommender Systems. *Handbook of Statistics*, *33*, 107–124. https://doi.org/10.1016/B978-0-444-63492-4.00005-8

Zhu, W., Mousavi, S. M., & Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks. *Advances in Geophysics*, *61*, 151–177. https://doi.org/10.1016/BS.AGPH.2020.07.003

Chauhan, N. S. (2023). *Métricas de evaluación de modelos en el aprendizaje automático*. DataSource.ai. https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Anexo 1. Variables elegidas para análisis

La ENIGH cuenta con un documento que proporciona todos los detalles relacionados al resto de variables, así como información adicional de las variables aquí mostradas. Esta información adicional puede ser consultada en la siguiente liga.

Variable	Definición	Etiquetas y significado
parentesco	Vínculo o lazo de unión que existe entre el jefe(a) y los integrantes del hogar, ya sea conyugal, por consanguinidad, adopción, afinidad o costumbre.	101: Jefe(a) 201: Esposo(a), compañero(a), cónyuge, pareja, marido, mujer, señor(a), consorte 202: Concubino(a) 203: Amasio(a) 204: Querido(a), amante 205: Pareja del mismo sexo 301: Hijo(a), hijo(a) consanguíneo, hijo(a) reconocido 302: Hijo(a) adoptivo(a) 303: Hijastro(a), entenado(a) 304: Hijo(a) de crianza 305: Hijo(a) recogido(a) 401: Trabajador(a) doméstico(a) 402: Recamarero(a) 403: Cocinero(a) 404: Lavandera(o) 405: Nana, niñera, nodriza 406: Mozo 407: Jardinero(a) 408: Velador, vigilante 409: Portero(a) 410: Chofer 411: Ama de llaves 412: Mayordomo 413: Dama de compañía, acompañante 421: Esposo(a) del(la) trabajador(a) doméstico(a) 431: Hijo(a) del(la) trabajador(a) doméstico(a) 441: Madre, padre del(la) trabajador(a) doméstico(a)

		451: Nieto(a) del(la) trabajador(a) doméstico(a) 461: Otro pariente del(la) trabajador(a) doméstico(a) 501: No tiene parentesco 502: Tutor(a) 503: Tutelado(a), pupilo(a), alumno(a) 601: Madre, padre 602: Padrastro, madrastra 603: Hermano(a) 604: Medio(a) hermano(a) 605: Hermanastro(a) 606: Abuelo(a) 607: Bisabuelo(a) 608: Tatarabuelo(a) 609: Nieto(a) 610: Bisnieto(a) 611: Tataranieto(a) 612: Tío(a) 613: Sobrino(a) 614: Primo(a) 615: Suegro(a) 616: Consuegro(a) 617: Nuera, yerno 618: Cuñado(a) 619: Concuño(a) 620: Padrino, madrina 621: Ahijado(a) 622: Compadre, comadre 623: Familiar, otro parentesco 701: Huésped, abonado(a), pensionista 711: Esposo(a) del(la) huésped 712: Hijo(a) del(la) huésped 713: Madre o padre del(la) huésped 714: Nieto(a) pariente del(la) huésped 715: Otro(a) pariente del(la) huésped 999: Parentesco no especificado
sexo	Distinción biológica que clasifica a las personas en hombres o mujeres.	1: Hombre 2: Mujer
edad	Años transcurridos entre la	No hay etiquetas, los

	fecha de nacimiento de la persona y la fecha de la entrevista.	números que se muestran en las gráficas, son la edad que tienen
madre_hog	Identifican a la madre de las personas del hogar.	1: Sí 2: No
padre_hog	Identifican al padre de las personas del hogar.	1: Sí 2: No
hablaind	Personas de 3 años o más que hablan alguna lengua indígena o dialecto.	1: Sí 2: No
hablaesp	Personas de 3 años o más que hablan alguna lengua indígena o dialecto que también habla español.	1: Sí 2: No
nivelaprob	Año máximo aprobado en la escuela, por el integrante del hogar de 3 o más años dentro del Sistema Educativo Nacional.	0: Ninguno 1: Preescolar 2: Primaria 3: Secundaria 4: Preparatoria o bachillerato 5: Normal 6: Carrera técnica o comercial 7: Profesional 8: Maestría 9: Doctorado
edu_conyug	Estado conyugal del integrante del hogar de 12 o más años.	1: Vive con su pareja o en unión libre 2: Está casado(a) 3: Está separado(a) 4: Está divorciado(a) 5: Es viudo(a) 6: Está soltero(a)
hor_1	El tiempo, en horas, que las personas dedicaron a trabajar.	No hay etiquetas, los números que se muestran en las gráficas, son las horas de trabajo de los individuos
atemed	Personas están o no afiliadas o inscritas a alguna institución que proporciona atención médica.	1: Sí 2: No
num_trabaj	Número de trabajos que los integrantes del hogar realizaron durante el mes pasado.	1: Solo 1 2: Dos o más

entidad	Contiene la clave de la entidad federativa, estas corresponden al Catálogo de claves de entidades federativas, municipios y localidades, que está disponible en el sitio del INEGI.	01: Aguascalientes 02: Baja California 03: Baja California Sur 04: Campeche 05: Coahuila de Zaragoza 06: Colima 07: Chiapas 08: Chihuahua 09: Ciudad de México 10: Durango 11: Guanajuato 12: Guerrero 13: Hidalgo 14: Jalisco 15: México 16: Michoacán de Ocampo 17: Morelos 18: Nayarit 19: Nuevo León 20: Oaxaca 21: Puebla 22: Querétaro 23: Quintana Roo 24: San Luis Potosí 25: Sinaloa 26: Sonora 27: Tabasco 28: Tamaulipas 29: Tlaxcala 30: Veracruz de Ignacio de la Llave 31: Yucatán 32: Zacatecas
clave_max	Producto, bien o servicio adquirido por los integrantes del hogar en los periodos de referencia.	Q001: Depósitos en cuentas de ahorro, tandas, cajas de ahorro, etcétera Q002: Préstamos a personas ajenas al hogar Q003: Pagos a tarjeta de crédito bancaria o comercial (incluye intereses) Q004: Pago de deudas a la empresa donde trabajan y/o a otras personas o instituciones (excluya créditos hipotecarios) Q005: Pago de intereses por préstamos recibidos

		Q006: Compra de monedas nacionales o extranjeras, metales preciosos, alhajas, obras de arte, etcétera Q007: Seguro de vida capitalizable Q008: Herencias, dotes y legados Q009: Compra de casas, condominios, locales o terrenos que no habita el hogar Q010: Compra de terrenos, casas o condominios que habita el hogar Q011: Pago de hipotecas de bienes inmuebles: casas, locales, terrenos, edificios, etcétera Q012: Otras erogaciones no consideradas en las preguntas anteriores Q013: Compra de maquinaria, equipo, animales destinados a la reproducción, utilizados en negocios del hogar Q014: Balance negativo en negocios del hogar agropecuarios y no agropecuarios Q015: Compra de valores: cédulas, acciones y bonos Q016: Compra de marcas, patentes y derechos de autor Q100: Pago de la vivienda propia y que se está pagando
disc_camin	Discapacidad que presenta algún integrante del hogar para caminar, subir o bajar usando sus piernas.	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_ver	Discapacidad que presenta algún integrante del hogar para ver (aunque use lentes).	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_brazo	Discapacidad que presenta	1: No puede hacerlo

	algún integrante del hogar para mover o usar brazos o manos.	2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_apren	Discapacidad que presenta algún integrante del hogar para aprender, recordar o concentrarse.	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_oir	Discapacidad que presenta algún integrante del hogar para escuchar (aunque use aparato auditivo).	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_vest	Discapacidad que presenta algún integrante del hogar para bañarse, vestirse o comer.	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_habla	Discapacidad que presenta algún integrante del hogar para hablar o comunicarse.	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
disc_acti	Discapacidad que presenta algún integrante del hogar para realizar actividades diarias por problemas emocionales o mentales (con autonomía e independencia).	1: No puede hacerlo 2: Lo hace con mucha dificultad 3: Lo hace con poca dificultad 4: No tiene dificultad
ss_aa	Número de años de contribución a la seguridad social.	No hay rango, pero es importante destacar que el valor -1 representa un dato “no especificado”
ss_mm	Número de meses de contribución a la seguridad social.	No hay rango, pero es importante destacar que el valor -1 representa un dato “no especificado”
inscr_	Origen de la afiliación o inscripción de las personas a las instituciones de salud o que les otorgan alguna pensión.	1: Prestación en el trabajo 2: Jubilación o invalidez 3: Algún familiar en el hogar 4: Muerte del asegurado 5: Ser estudiante

		6: Contratación propia 7: Algún familiar de otro hogar 8: No sabe