Grupo 301

Tecnológico de Monterrey

Análisis de ciencia de datos

Documento final del Reto

Docentes:

Alder López Cerda

Julio Antonio Juárez Jiménez

Gerardo Ibarra Vázquez

Autores:

Alfredo André Durán Treviño - A01286222

Eliani González Laguna - A00836712

Fedra Fernanda Mandujano López - A00835797

Juan Marco Castro Trinidad - A01742821

Miranda Isabel Rada Chau - A01285243

14 de junio del 2024

Introducción

Los censos son herramientas muy útiles para tener un mayor conocimiento sobre los habitantes de un país. Estos nos ayudan a tener una mejor comprensión de la cantidad de habitantes que hay, las condiciones en las que viven, entre otros. Estas herramientas también son útiles, porque nos dan acceso a muchos datos que se pueden analizar para saber más sobre la calidad de vida y ciertas características de los habitantes de una zona. En el caso de México, todos estos datos son almacenados y recopilados por el INEGI, quien nos dio acceso a ellos para hacer un análisis en cuanto a los ingresos de los mexicanos.

Se pueden utilizar modelos con inteligencia artificial para hacer este tipo de análisis de datos, ya que se pueden integrar algoritmos de inteligencia artificial para poder identificar patrones y correlaciones dentro de nuestros datos que es difícil de ver en un análisis humano. Por lo que, la inteligencia artificial representa una nueva rama de herramientas de las cuales podemos sacar provecho en nuestros análisis y creación de modelos, pero, también hay que tener en cuenta que estos no deben de hacer todo el proyecto, es necesario un equilibrio entre el humano y la herramienta para que se pueda trabajar correctamente y para formular conclusiones efectivas.

En este caso, vamos a utilizar esta herramienta para nuestro análisis y la creación de modelos para nuestro proyecto, en el que vamos a usar y analizar diferentes modelos de predicción para determinar si alguien pertenece o no a una categoría correspondiente al nivel de ingresos.

El objetivo del proyecto es utilizar una base de datos proporcionada por el INEGI para el desarrollo de un modelo predictivo que permita determinar si una persona se encuentra por encima o por debajo del umbral de ingresos propuesto de 25,000 pesos mexicanos.

Para esto, se redujo la dimensionalidad eliminando variables que no fueran de nuestro interés para este caso. Se realizó esto en primera instancia para que a través de un análisis exploratorio de datos se volvieran a analizar las variables restantes de interés con gráficos, matrices de correlación y mapas de calor que muestren de manera visual e intuitiva las correlaciones. Con todo esto se establecerán las bases metodológicas para el desarrollo del modelo predictivo.

El objetivo del proyecto a nivel social es la identificación de los factores que influyen en el nivel de ingresos de las personas y que este efecto se vea plasmado en un algoritmo predictivo que permita el desarrollo de programas sociales que permitan contribuir a la reducción de la brecha económica y mejorar la distribución de ingresos en la población. El análisis incluirá variables demográficas, educativas y laborales para el desarrollo de un modelo eficiente.

Sin embargo, existen algunos factores que pueden hacer que el algoritmo no sea tan eficiente y robusto como se espera y es necesario tomar en cuenta que la calidad de los datos no pueden ser del todo precisos y completos. Este análisis tiene que ser acompañado de métodos estadísticos bien pensados para poder traducir los resultados del modelo en acciones

concretas y comprensibles para los tomadores de decisiones, asegurando siempre que el modelo sea flexible y se pueda ajustar a cambios en las variables socioeconómicas.

Comprensión del negocio

El INEGI es el Sistema Nacional de Estadística y Geografía, el cual es un organismo público autónomo responsable de normar, coordinar, ayudar, captar y difundir la información mexicana del área generando a su vez estadística básica y recursos a través de censos, encuestas y registros administrativos en torno al territorio, los recursos, la población y economía de la comunidad. La información recopilada y almacenada puede consultarse en la página oficial del INEGI, que contiene múltiples datasets. Estos datasets nos permiten conocer las características de nuestro país y poder realizar análisis para la toma de decisiones. (INEGI, 2024) Para la realización del siguiente proyecto, nos estamos basando en datos recopilados a través de encuestas realizadas por el INEGI. Este tipo de encuestas son de gran valor, ya que nos dan la oportunidad de conocer las características de los habitantes de nuestro país y así también se puede identificar si hay algún área de oportunidad en cuanto a las condiciones bajo las que viven los mexicanos.

Exploración y comprensión de los datos

Análisis Exploratorio de Datos (EDA):
La base de datos original tenía 192 variables y 309682 registros, decidimos reducir la dimensionalidad de la base de datos eliminando variables que a primera vista no tuvieran un impacto evidente para realizar la predicción que queremos lograr o variables que contaban con una escasa cantidad de datos, por lo que, terminarían generando ruido en los análisis posteriores. Esto con el objetivo de hacer que el problema fuera menos complejo y extenuante, evitando así variables que no contaban con tanto impacto para el análisis.

En general, el tipo de variables que se eliminaron fueron:

- Identificadores de Relaciones Familiares:
  Variables que especifican la relación de la persona con otros miembros del hogar (si es padre, madre, hijo, etc.).
  Debido a las diferentes circunstancias que puede tener cada persona, no consideramos esto lo más importante, más adelante declaramos que solo nos quedamos con gente mayor de 18 años, por lo que, saber su relación familiar no es tan relevante al contar sus ingresos.
- Causas de Discapacidad:
  Variables que detallan las razones específicas por las que una persona tiene una discapacidad.
  Lo que nos importa es si tienen alguna discapacidad, la causa no necesariamente se va a relacionar con sus condiciones de vida o posibilidad de recibir ingresos, es por ello, que posteriormente englobamos estas variables dentro de una sola, la cual distingue a las personas con alguna discapacidad.
- Tiempos Específicos de Actividades:

Variables que miden el tiempo exacto que una persona dedica a diversas actividades diarias, como minutos dedicados al trabajo, transporte, ocio, etc.
Esto al ser valores muy variados y tras no presentar información que tenga que ver con el trabajo, decidimos revocarlo.

- Tipo de Afiliación Médica:
  Variables que describen el tipo de seguro de salud o afiliación médica que una persona tiene.
  Esto sí depende de su trabajo, pero, no consideramos que este puede tener un impacto en su sueldo, por lo que terminaría alterando nuestro modelo.
- Atención Médica Recibida:
  Variables que indican si una persona ha recibido atención médica, qué tipo de atención ha recibido, y con qué frecuencia.
  Esto depende de si ha sufrido algo recientemente, pero, no podemos considerarla porque esto puede ser por la gravedad de la situación u otros factores, por lo que, no consideramos importante o beneficioso el considerarlo para el modelo.
- Uso de Lengua Indígena:
  Variables que indican el tipo de lengua indígena que habla.
  Esto es debido a que son características de la persona, pero, no necesariamente, nos habla mucho sobre la capacidad de trabajar o de conseguir un sueldo alto o bajo, por lo que, se optó en revocarla.
- Experiencia de Discriminación:
  Variables que miden si una persona ha experimentado discriminación y de qué tipo.
  Cosas como la discriminación se pueden experimentar en cualquier clase social, por lo que, considerarla solo podría perturbar los resultados del modelo.
- Voluntariado Médico:
  Variables que indican si una persona participa como voluntario en actividades médicas.
  Como lo anterior, no se relaciona el voluntariado con el sueldo, ya que, cualquiera puede tener la intención de ser voluntario.
- Becas o Créditos Educativos:
  Variables que muestran si una persona recibe algún tipo de beca o crédito educativo.
  Se puede tener beca en cualquier clase, puesto que, a final de cuentas, también importa la escuela, por lo que, el hecho de tener beca no nos habla del estado económico.

También hubo variables que pensábamos que eran importantes, pero, debido a la falta de datos (con menos del 10% de datos llenos) se tuvieron que eliminar. Siendo el caso de las siguientes variables:

- El número de hijos nacidos vivos.
- El número de hijos muertos.
- El número de hijos vivos en la actualidad.

Además, se juntaron a las personas con algún tipo de discapacidad haciendo los datos binarios, 0 para las personas con discapacidad y 1 para las personas que no tienen,

sumándose todas las columnas generando la nueva variable ‘disc’ la cual indica el número de personas que cuentan con alguna discapacidad.

Posterior a la eliminación y la mezcla de columnas nos quedamos con 25 variables. Siendo estos los datos y variables idóneas que llegamos a considerar como los más importantes y eficientes para poder predecir los ingresos de la persona de la mejor manera. Estos siendo datos que, en su mayoría, tienen una relación con la clase social o con el entorno educativo y social, por lo que, aparentan ser de gran utilidad y que podrían ser efectivos para la resolución de nuestro modelo. Estos datos fueron obtenidos del dataset original, por medio de la eliminación de las categorías innecesarias, siendo esto lo que previamente ya se desarrolló, resultando en la permanencia de solo 25 variables, que después se volvieron 26.

Las 25 variables restantes antes de ser procesadas con métodos de limpieza tienen los siguientes atributos:

Columna	**Tipo de dato (python)**	**Tipo de variable**	**Descripción**
folioviv	int64	numérica	Identificador de folio de vivienda.
foliohog	int64	categórica {1: hogar principal, 2-5:hogares adicionales}	Identificador de folio de hogar.
numren	int64	numérica	Identificador de número consecutivo en el registro de personas del hogar.
sexo	int64	categórica {1: hombre, 2:mujer}	Género biológico de la persona.
edad	int64	numérica	Edad de la persona.
disc	int64	categórica {0: Sí, 1: No}	Si la persona cuenta con alguna discapacidad.
hablaind	int64	categórica {1: Sí, 2: No}	Si la persona habla alguna lengua indígena.
hablaesp	int64	categórica {1: Sí, 2:No}	Si la persona habla español.
etnia	int64	categórica {1: Sí, 2:No}	Si la persona se considera indígena.
alfabetism	int64	categórica {1: Sí, 2:No}	Si la persona puede leer.

asis_esc	int64	categórica {1: Sí, 2:No}	Si la persona asiste a la escuela.
nivelaprob	int64	categórica {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}	Nivel de instrucción aprobado.
edo_conyug	int64	categórica {1: Vive con su pareja en unión libre, 2: Está casado(a), 3: Está separado(a), 4: Está divorciado(a), 5: Es viudo(a), 6: Está soltero(a)}	Estado civil
segsoc	int64	categórica {1: Sí, 2:No}	Contribución a la seguridad social.
hor_1	int64	numérica	Horas de trabajo de la semana pasada.
usotiempo1	int64	categórica {8: No recuerda, 9:No lo hizo}	Horas de trabajo de la semana pasada.
atemed	int64	categórica {1: Sí, 2:No}	Afiliación para la atención médica.
pagoaten_1	int64	categórica {0: No, 1:Sí}	Si la persona pagó por consulta.
diabetes	int64	categórica {1: Sí, 2:No}	Si la persona tiene diabetes.
pres_alta	int64	categórica {1: Sí, 2:No}	Si la persona tiene presión alta.
trabajo_mp	int64	categórica {1: Trabajó el mes pasado,	Si la persona trabajó el mes pasado.

		2:No trabajó el mes pasado}	
num_trabaj	int64	categórica {1: Solo 1, 2: Dos o más}	Número de trabajos.
entidad	int64	categórica El desglose completo de estas categorías, se muestra en el anexo 1.	Estado de residencia de la persona.
ing_tri_total	float64	numérica	Ingreso trimestral total.
ing_tri_max	float64	numérica	Ingreso trimestral máximo

Las columnas ing_tri_total e ing_tri_max serán eliminadas, ya que nuestra variable objetivo se basa en estas dos. Si se añadieran no tendría sentido, puesto que sería como calcular la variable objetivo con la misma a predecir.

Posterior a esto, se agregó la variable 'ingreso_prom' la cual indica si la persona tiene un ingreso mayor o menor a 25,000 pesos mexicanos.

1. ingreso_prom (object). Es una variable categórica {0: <=25k y 1: >25k}.

Limpieza y preparación de los datos

Tras haber hecho un análisis inicial de los datos y las columnas incluidas en el dataframe. Comenzamos a decidir cuáles creíamos relevantes para el estudio. Después, de este análisis inicial, comenzamos a hacer una limpieza de los datos para crear el dataframe final a partir del cual se iba a hacer el análisis estadístico y el que se iba a modelar más adelante.

El primer paso de nuestra limpieza de datos fue eliminar todas las filas correspondientes a menores de edad, esto implica que se borraron todas las filas que tenían un valor menor a 18 en la columna de "edad". Decidimos eliminar estas filas, ya que, solo nos interesa analizar cuántos adultos ganaban más de 25,000 pesos, ya habiendo eliminado a todos los menores de edad incluidos en el dataset, comenzamos a filtrar los valores nulos presentes en el dataframe. La manera en que tratamos a valores nulos dependía de cuántos espacios vacíos había en cada columna. En el caso de la columna de "ing_tri_max", decidimos solamente borrar las filas vacías, ya que, no había muchos valores vacíos y además, esta es una columna importante para la predicción y el modelado de los resultados, lo cual, implica que no sería apropiado rellenarla bajo nuestros criterios.

Después comenzamos con el proceso de combinar todas las columnas correspondientes a personas con discapacidades en una sola columna que solo mencione si cada persona tiene

una discapacidad o no. Esto se hizo a partir de una serie de replace con las 8 columnas correspondientes a discapacidades, combinándolas en una sola llamada “disc”, la cual se agrega al nuevo dataframe. Habiendo terminado esta nueva columna, se crea un nuevo dataframe que contiene todas las columnas seleccionadas durante el filtrado realizado anteriormente, incluyendo la columna creada para las discapacidades. Cabe mencionar que hasta este punto todavía no se ha obtenido el dataframe limpio para alimentar a los modelos.

Se hizo un análisis de la información básica del dataframe para ver cuáles de las columnas tenían valores vacíos y a partir de ahí comenzamos a rellenar los valores nulos. Los valores nulos en las columnas categóricas, se rellenaron con la moda de la columna, esto sucedió con las columnas correspondientes a si las personas hablan español, si pagan una contribución al servicio social, la cantidad de horas trabajadas en la última semana, entre otras. En el caso de las columnas numéricas, los valores nulos se rellenaron con la media de los valores de la columna.

En el análisis inicial también se pudo ver que los valores ‘&’ son los casos en los que la pregunta no se respondió y estos también se rellenaron con la moda de la columna. En el dataframe original, la mayor parte de las columnas eran object, pero, para poder hacer los análisis gráficos y las transformaciones, era necesario convertir a todas las columnas con datos tipo object en columnas con datos numéricos.

Data Engineering

Después de realizar la limpieza de datos explicada anteriormente, fue necesario llevar a cabo algunas transformaciones para poder seguir con el análisis y para poder desarrollar modelos más adelante. Para poder realizar el análisis estadístico de las variables incluidas en el dataframe fue necesario transformar a todas las columnas categóricas en variables numéricas, pero, antes de hacer esta transformación tuvimos que reemplazar todos los espacios vacíos con un número para que se pudiera transformar la columna en una columna numérica. Estos espacios en blanco nos causaron varias dificultades, ya que no se lograban transformar en tipo numéricos, por lo que, para solucionar este problema, se decidió en cambiar los espacios vacíos por la moda de su columna correspondiente. Al terminar este proceso, fue posible llevar a cabo la transformación de las columnas, para esto fue necesario utilizar la función de astype, esta transformación fue posible debido a que todos los valores en estas columnas eran números y esto implicó que se pudieron transformar directamente.

En general, no se ocupó hacer una manipulación muy grande de los datos debido a que al filtrar las columnas, escogimos columnas que no tuvieran tantos valores nulos. Consideramos que era importante filtrar las columnas, ya que al tener muchos espacios nulos, era más fácil sesgar la información a través del uso de modas y promedios y esto podría tener un gran impacto en la tendencia identificada al final del análisis. Por esto mismo, seguimos los criterios de selección expuestos anteriormente.

Análisis univariado

En el análisis univariado, se busca explorar las variables de manera independiente. Ninguna de estas variables tiene valores nulos, entonces podemos pasar directamente al análisis. Por lo que, para este análisis, se obtuvieron las estadísticas descriptivas, esto mediante el cálculo de los datos unidimensionales:

Mode: Moda, dato que más se repitió en la categoría.
min: valor mínimo en la categoría.
max: valor máximo en la categoría.

Variables categóricas

Descripción de los datos	sexo	disc	hablaind	etnia
Mode	1	1	2	2
min	1	0	1	1
max	2	1	2	2
Descripción	Esta variable indica el género biológico de la persona. Esta tiene los valores 1 y 2. 1 siendo masculino y 2 femenino. En este caso, se puede ver que tenemos más mujeres que hombres.	Esta variable indica si las personas cuentan con alguna discapacidad. 0 indica que sí, 1 indica que no. Por la moda, vemos que la mayoría no cuenta con una discapacidad.	Esta variable muestra si una persona habla una lengua indígena o no. 1, significa que sí habla una lengua indígena. 2, significa que no. Viendo la moda, se puede ver que la mayoría no hablan una lengua indígena.	Esta variable indica si se considera indígena o no. 1, indica que sí. 2, indica que no. Según la moda, la mayoría de las personas encuestadas dijeron que no.

Descripción de los datos	alfabetism	asis_esc	nivelaprob	edo_conyug
mode	1	2	3	2
min	1	1	0	1

max	2	2	9	6
Descripción	Esta variable indica si la persona sabe leer: 1 significa que sí, 2 significa que no. Viendo la moda se puede ver que la mayor parte de las personas sí saben leer.	Esta variable indica si la persona asiste a la escuela. 1 es que sí asiste a la escuela, 2 es que no asiste. Al ver la moda se puede ver que la mayoría de las personas no asisten.	Esta variable indica el máximo nivel de educación aprobado. {0, ninguno 1, preescolar 2, primaria 3, secundaria 4, preparatoria 5, normal 6,carrera técnica 7, profesional 8, maestría 9, doctorado} Tras ver la moda, vemos qué gran parte de los encuestados solo terminaron la secundaria.	Esta variable nos comunica el estado civil de los encuestados. {1, unión libre 2, casado 3, separado 4, divorciado 5, viudo 6, soltero} Con la moda vemos que una gran parte está casada.

Descripción de los datos	**'segsoc'**	**'entidad'**	**hablaesp**	**'atemed'**
mode	1	8	1	1
min	1	1	1	1
max	2	32	2	2
Descripción	Con esto vemos si el encuestado ha contribuido o no a algún seguro social 1, sí. 2, no. Con la moda se puede ver que la mayoría sí.	Esta variable indica los estados organizados de 1 a 32 en orden alfabético para la entidad de la persona. El índice de las categorías, se puede ver en el anexo 1. A través de la	Esta variable indica si el encuestado habla español o no. 1, habla español, 2, no habla español. La mayoría sí habla español.	Esta variable indica si el encuestado está afiliado a recibir atención médica. 1, si 2, no La mayoría sí está afiliado.

		moda, se puede ver que la entidad más mencionada es: Chihuahua.		

Descripción de los datos	**'pagoaten_1'**	**'diabetes'**	**'pres_alta'**	**'trabajo_mp'**
mode	0	2	1	1
min	0	1	1	1
max	1	2	2	2
Descripción	Esta variable indica si el encuestado pagó por una consulta. 0: Sí, 1:No La mayoría de los encuestados no pagan por consultas.	Esta variable indica si la persona padece de diabetes. 1 sí tiene diabetes 2, no tiene diabetes. La mayoría no tiene diabetes.	Esta variable indica si la persona tiene presión alta o no. 1 sí, 2 no. Por la moda se puede ver que la mayoría sí tiene.	Esto indica si la persona trabajó el mes pasado. 1 sí 2 no En este caso, se puede ver que la mayor parte de los encuestados sí trabajaron en el último mes.

Descripción de los datos	**'num_trabaj'**	**'usotiempo1'**	**‘ingreso_prom’**
mode	1	9	0
min	1	8	0
max	2	9	1
Descripción	Número de trabajos que tiene la persona 1, 1 trabajo 2, 2 o más La mayoría tiene un solo trabajo.	Tiempo que se dedicó a trabajar. 8, no recuerda 9, no lo hizo La mayoría de los encuestados no lo hizo.	Indica si las personas ganan más o menos/igual del umbral de 25000. {0: <=25k, 1: >25k} A través de la moda se puede ver que la

			mayoría de los encuestados ganan menos de 25000 al trimestre.

Variables numéricas

En el caso de las variables numéricas, el análisis estadístico es diferente. Para estas, no se toma en cuenta la moda, pero se agregan la media y la desviación estándar.

Mean: Promedio, el promedio de los datos de la categoría.
std: Desviación estándar, indica cuánto se alejan los valores individuales del promedio.

Descripción de los datos	**edad**	**'hor_1'**
mean	43.5498516	23.3889408
std	17.4774303	25.606215
min	18	0
max	109	168
Descripción	Tenemos un alto rango de edad, con un promedio de 43. A través del mínimo se puede ver que sí se eliminaron correctamente los menores de edad. Se identificó un número reducido de personas con 100 años o más de edad (21 datos). Aunque no fueron eliminadas para los fines del proyecto, como buena práctica deberían considerarse para eliminación debido a su escasa representación en el conjunto de datos.	Se detectaron 35 casos en los que personas reportaron trabajar más de 125 horas a la semana, lo cual equivale a trabajar sin descanso. Este hallazgo indica una situación laboral extrema que requiere una revisión detallada y un análisis más profundo de la situación de cada quien.

La siguiente tabla es de las variables que vamos a guardar, pero no se van a contar para los modelos de predicción, vamos a guardarlas para el análisis posterior, ya que estas se relacionan directamente con el ingreso promedio.

Descripción de los datos	'ing_tri_max'	'ing_tri_total'
mean	22440.6605	21913.1425
std	45432.9538	49854.4596
min	1.46	0
max	6854754.09	6854754.09
Descripción	Esta variable indica el ingreso total de la persona por trimestre.	Esta variable indica el ingreso máximo por trimestre de la persona.

Valores atípicos

La variable edad y la variable de horas trabajadas a la semana eran las únicas variables relevantes, entonces, se graficó una caja bigote para cada una para poder observar los valores atípicos.

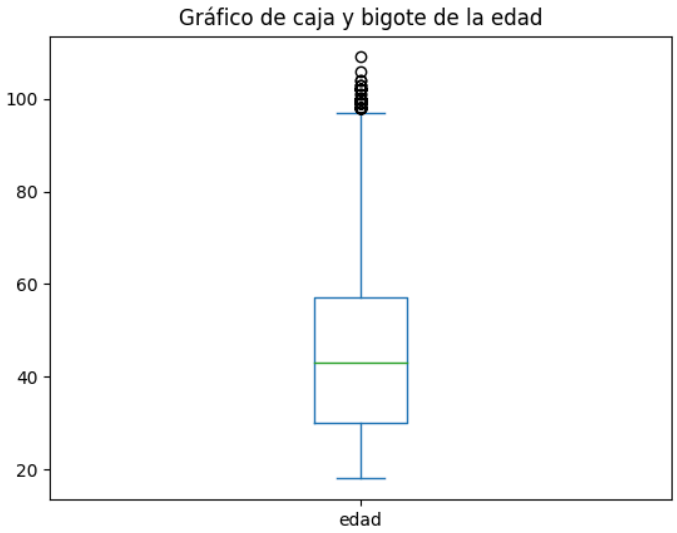


Fig. 1. Caja de bigotes de la variable edad.

Con base en este gráfico, se determinó que como en la variable edad, las personas con una edad por arriba de 100 años se consideran como outliers. En esta columna hay 21 datos mayores a 100, por lo tanto, hay 21 outliers. Estos se podrían eliminar de nuestra base de datos con el fin de no causar ruido en los análisis posteriores, en especial cuando estos son escasos, sin embargo, para propósitos del proyecto se conservarán estos registros.

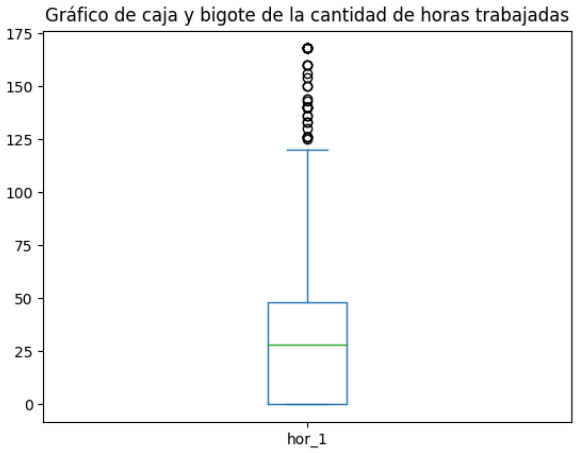


Fig. 2. Caja de bigotes de la variable hor_1.

En este gráfico se puede observar el comportamiento de los datos en la variable correspondiente a las horas trabajadas por semana. Se puede ver que hay 35 valores atípicos que son mayores a 125, indicando que trabajan más de 17 horas al día. Este fue un dato interesante y hasta preocupante porque indica que hay varias personas que están siendo explotadas en su trabajo. Es importante retomar estos casos para revisar la situación de una manera más profunda.

Análisis bivariado

Transformación de variables

En nuestro dataset, se le aplicó una transformación a las variables tipo objeto, y se convirtieron a tipo int o enteros, esto para poder realizar un mejor análisis y trabajar solamente con números para entrenamientos más eficientes.

Variables relevantes

De las 25 variables escogidas, consideramos que cuatro de estas son las más relevantes para observar, las cuales son: Sexo, Edad, Estado Civil, y Nivel de Educación. En lo siguiente se graficaron estas categorías tomando en cuenta el umbral de 25000 pesos para visualizar qué grupo tiende a ganar más.

Para el valor de ingreso promedio (nuestra variable objetivo) el valor de 0 se considera como falso (≤25000), no cumple con el umbral establecido. El valor de 1 se considera como verdadero (>25000) cumple con un ingreso mayor al umbral.

Sexo

Se realizó un gráfico de barras para identificar la distribución de los datos de la variable sexo acorde a los grupos de género en relación con el umbral de 25000 pesos mexicanos.

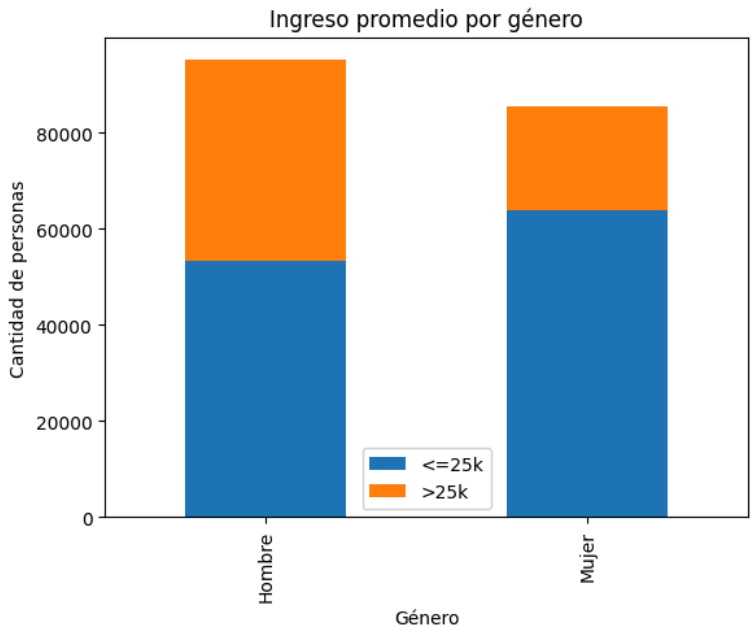


Fig. 3. Gráfica comparativa de ambos sexos dice si cumplen o no con los 25 mil pesos de ingresos.

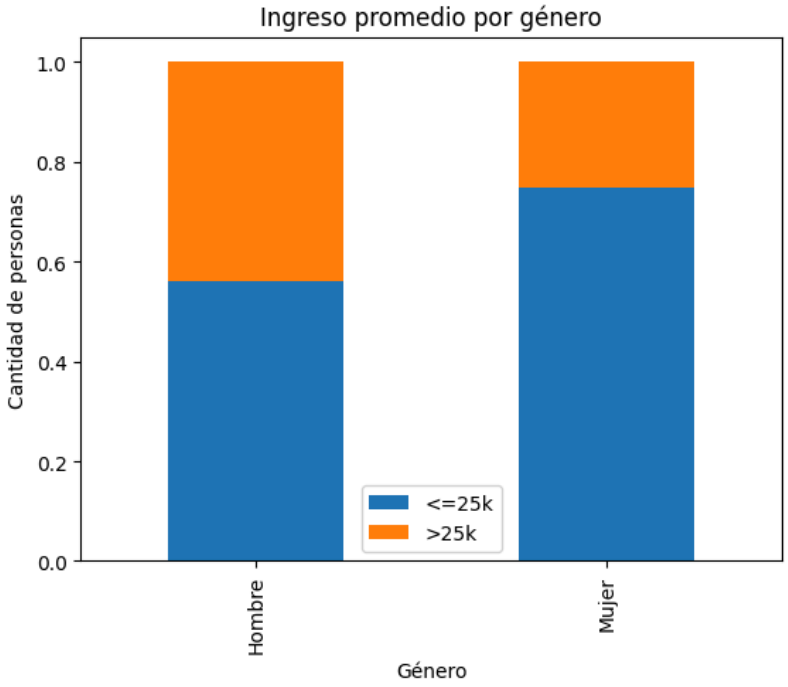


Fig. 4. Gráfica de la Fig.3 normalizada para comparar las proporciones.

Como podemos ver en las Figuras 3 y 4, parece haber cierto tipo de relación con el umbral de ingresos con el sexo de la persona, por lo que, vemos que sí podríamos usar estos datos.

Edad

Se creó un gráfico de densidad el cual muestra la distribución de la variable edad separada por los diferentes valores del umbral de 25000 pesos mexicanos.

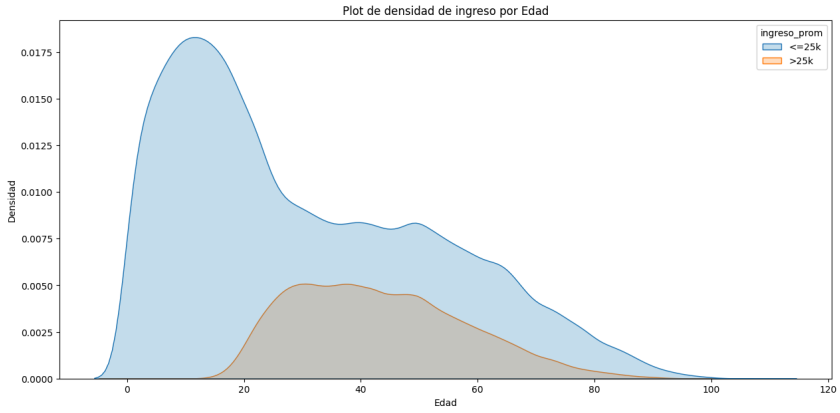


Fig. 5. Gráfica de densidad comparando si cumplen con el ingreso de 25000 pesos en relación con su edad.

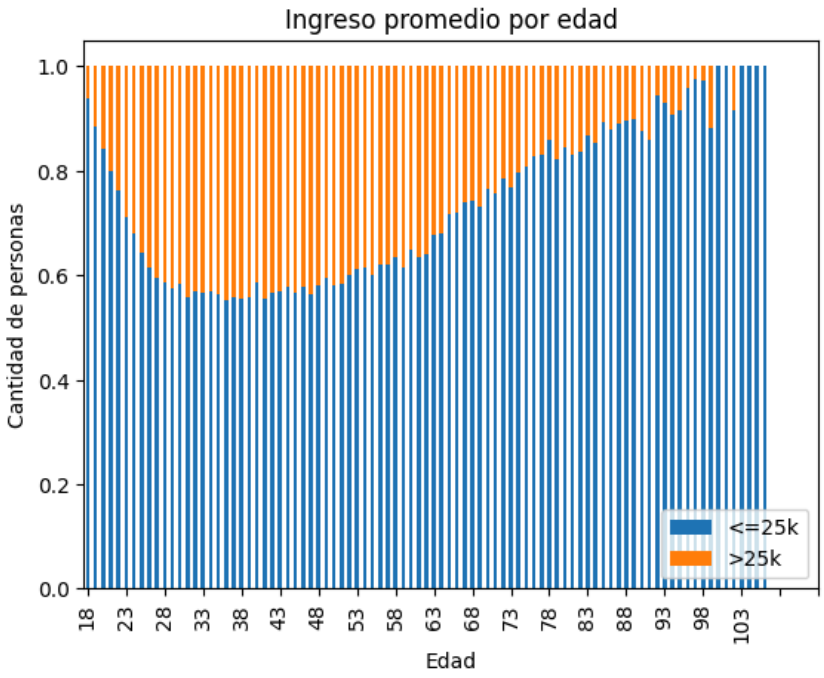


Fig. 6. Gráfica de barras normalizada, mostrando la proporción de la gente que cumple con el umbral de 25000 pesos en comparativa con su edad.

Como observamos en las figuras 5 y 6, hay una clara correlación entre la edad y los ingresos, viendo que se hace una curva muy clara en la comparación, por lo que, podemos pensar que sí tiene mucha relación el ingreso con la edad. En este gráfico se puede apreciar qué en el rango entre 20 y 65 años hay una mayor cantidad de personas que sobrepasan el umbral de ingresos.

Estado Civil

Se graficaron los grupos de estado civil tomando en cuenta el umbral de los 25000 pesos para visualizar qué grupo tiende a ganar más.

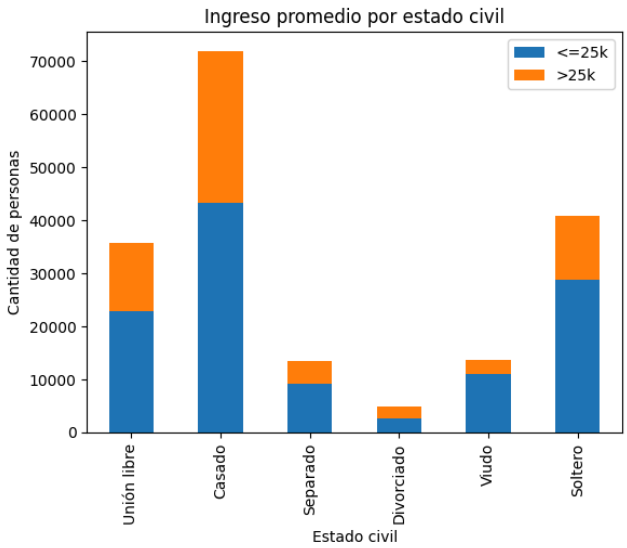


Fig. 7. Gráfica de barras que compara los diferentes estados civiles con la cantidad de gente que gana más de 25000 pesos.

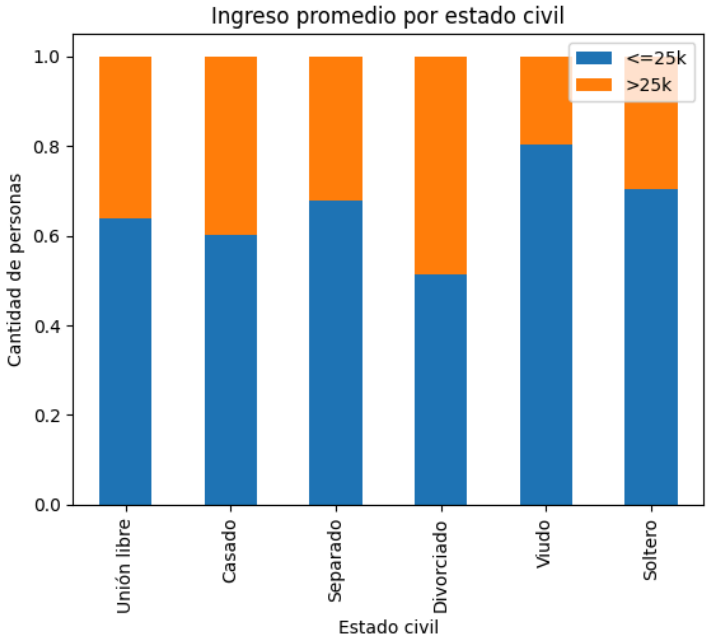


Fig. 8. Gráfica normalizada de la Fig. 7.

Como se visualiza, existe una cierta relación entre los ingresos con el estado civil, a pesar de que no parece ser tan clara comparativamente.

Nivel de Educación

Las siguientes gráficas muestran una comparación entre los niveles educativos aprobados por la población con respecto al nivel de ingresos, tomando en cuenta que estos datos no están normalizados.

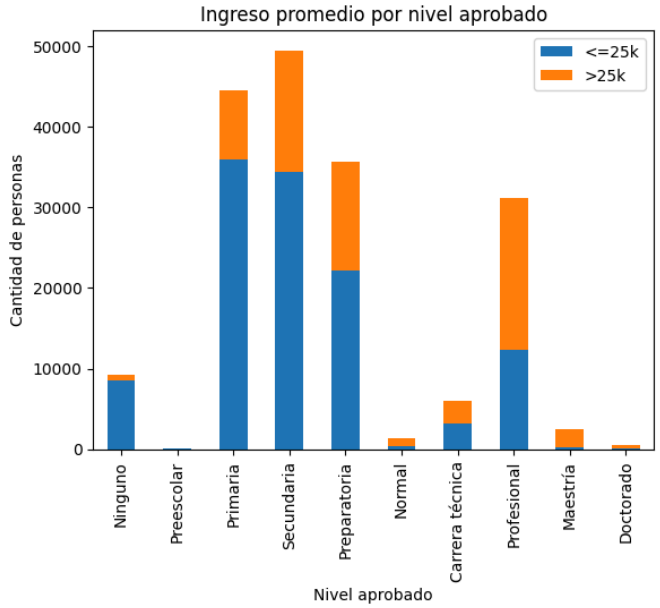


Fig. 9. Gráfica de barras comparando el nivel aprobado dice si cumplen o no con el umbral de 25000 de ingresos.

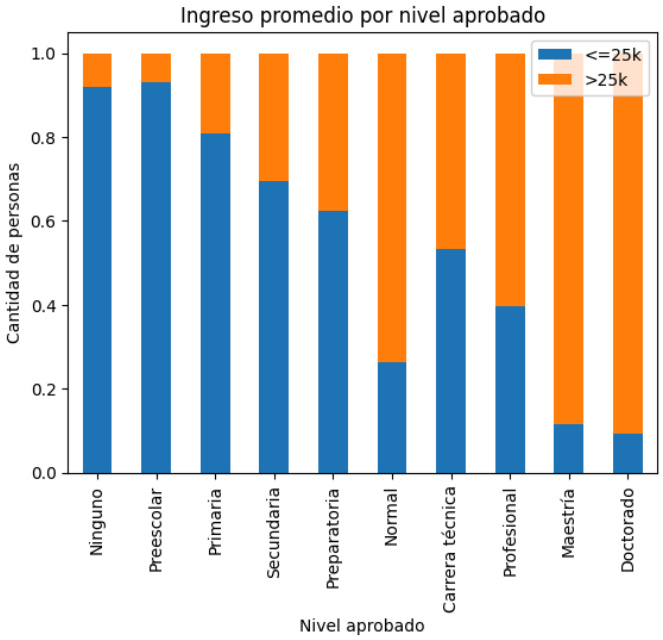


Fig. 10. Gráfica normalizada de la Fig. 9.

Como vemos, es fácil apreciar que de manera ordenada se muestra que la población que ha tenido un mayor nivel educativo tiende a tener mayores ingresos, superando el umbral de los 25000.

Datos útiles y no útiles

Realizando la gráfica de importancia de variables nos dimos cuenta de que las columnas 'ing_tri_total' e 'ing_tri_max' eran las "variables" con mayor importancia para predecir si la persona se encontraba por arriba o por debajo del umbral, esto debido a que se relacionan directamente con el ingreso promedio. Por lo que, no son pensadas para usarse en la clasificación, las vamos a usar más adelante para una interpretación y un análisis a futuro de los resultados que obtengamos de nuestros modelos.

Gráfico con las columnas de ingreso trimestral total y máximo.

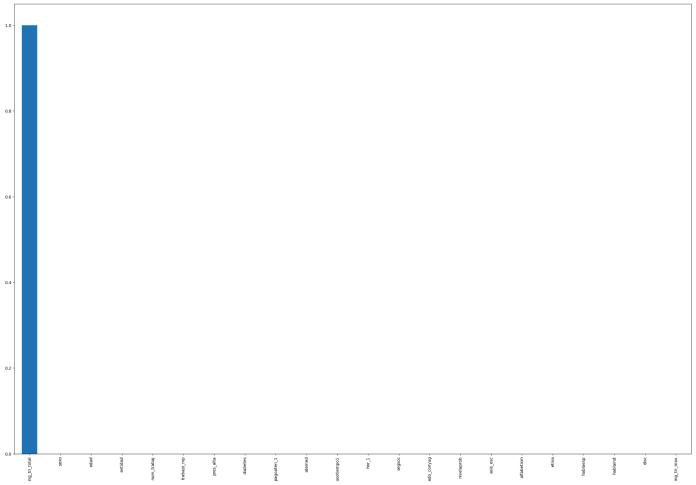


Fig. 11. Gráfica de las columnas ing_tri_total e ing_tri_max.

Gráfico sin las columnas de ingreso trimestral total y máximo.

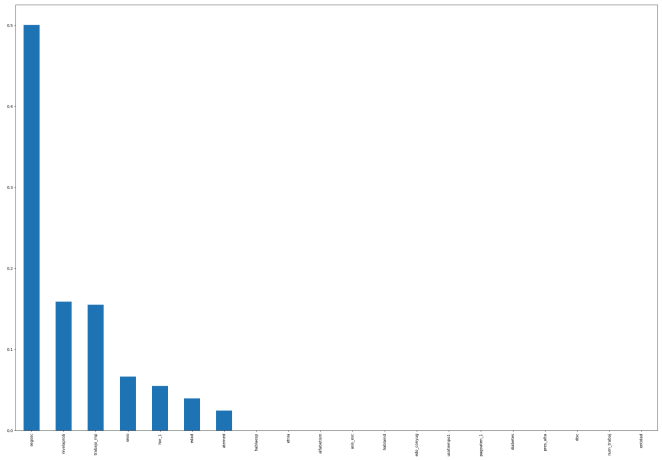


Fig. 12. Gráfica sin las columnas ing_tri_total e ing_tri_max.

En estos gráficos se puede visualizar que las columnas que impactan con mayor fuerza en los ingresos resultan ser:

- segsoc: Contribución a la seguridad social.
- nivelaprob: Nivel de instrucción aprobado.
- trabajo_mp: Si la persona trabajó o no el mes pasado.
- sexo: Género biológico de la persona.
- hor_1: Horas de trabajo.
- edad: Edad de la persona.
- atemed: Afiliación para atención médica.

Después de hacer la selección final de columnas, decidimos realizar un heatmap de correlación, el cual nos muestra que tanta correlación o similitud hay entre cada una de las variables. Si la correlación es de 1, significa una relación total lo cual sería una exactitud de las variables, como se observa en el heatmap, esto sucede principalmente en el eje diagonal, ya que, son las mismas variables. Sin embargo, sí existen unas variables que tienen una correlación alta con otras variables, esto pudiera causar ruido al momento de hacer modelos de predicciones. A continuación, en la Fig. 13 se puede observar la correlación de variables realizada.

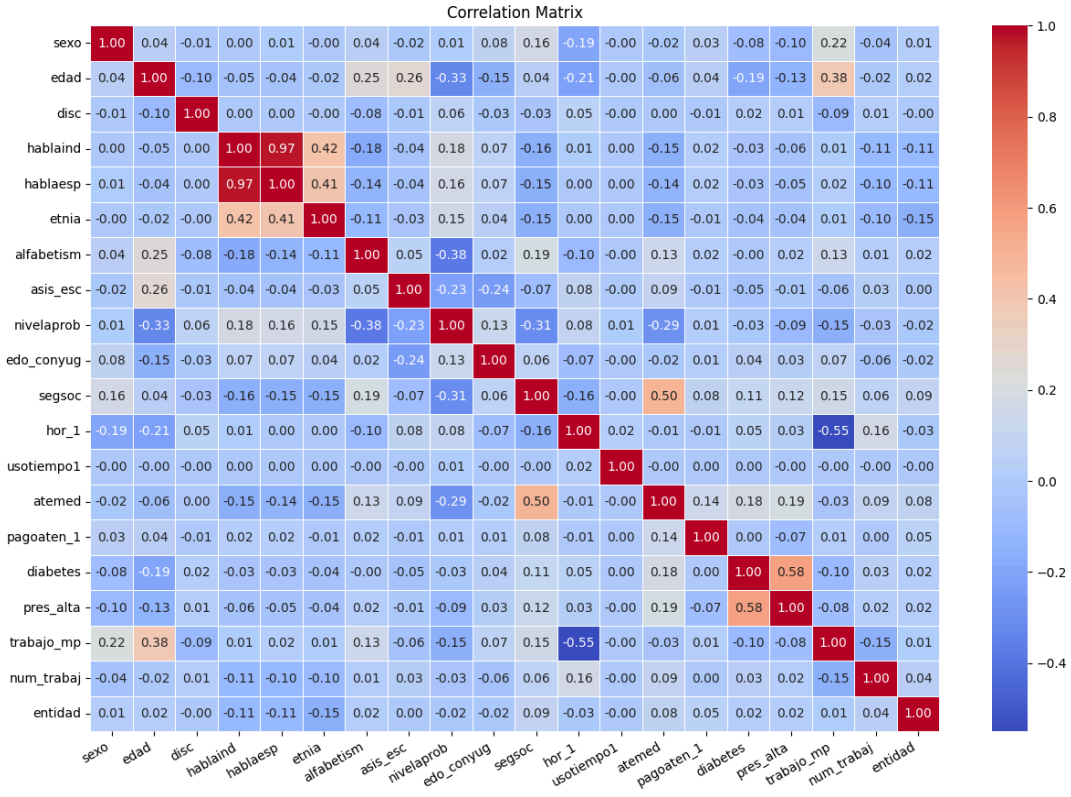


Fig. 13. Heatmap de correlación de las variables

En esta sección se puede apreciar la matriz de correlaciones de las variables seleccionadas. En este caso, se puede ver que hay varias variables que están relacionadas. Se puede ver que las variables más correlacionadas son las relacionadas con el idioma, ya que “hablaesp” y “hablaind” tienen una correlación del 0.97. Esto indica que tienen una correlación casi perfecta. También se puede ver que existe una gran correlación entre la contribución al seguro social y la afiliación de la atención médica. Esto tiene sentido, por el hecho de que las personas que cuentan con seguro social, es más probable que vayan a buscar atención médica. También se puede ver una correlación entre la diabetes y la presión alta, y esto muestra una relación interesante, ya que ambas son factores de salud importantes.

Conclusiones preliminares del procesamiento de datos

Consideramos que los datos siguen manteniendo su integridad y su calidad, debido a que, intentamos mantener una manipulación mínima para no causar un sesgo en el comportamiento de los datos, puesto que esto podría tener un gran impacto en el análisis y las conclusiones que se van a formular más adelante. Por esta razón consideramos primordial eliminar por completo las categorías que no nos sirven al ver los ingresos. A futuro vamos a empezar a ver el posible modelo para predecir si la persona cumple con los 25000 o no, en esta entrega resultaba ser únicamente acorde al análisis exploratorio y a la limpieza de los datos para empezar a trabajar correctamente con el dataset proporcionado por el socio formador. Quedándonos con un total de 217375 registros en nuestro dataset, sin tener ningún valor nulo, aparte de quitar y acomodar las categorías a conforme consideramos fue más conveniente, quedándonos con 20 categorías a ingresar al modelo, siendo estas las que describimos en el análisis univariado. También nos quedamos con dos variables para un futuro análisis, estás siendo 'ing_tri_max' e 'ing_tri_total' que no están en el modelo, ya que ingreso promedio se calcula a través de estas, pero las guardamos para el análisis futuro. Finalmente, nos quedamos con una que sería nuestra variable objetivo ‘ingreso_prom’, por lo que, nuestro dataset está limpio y listo para ser utilizado en un modelo.

Como advertimos, la simple limpieza y análisis de los datos resulta ser un trabajo complejo y de mucho tiempo, pero, es uno de los pasos más importantes para generar un modelo eficiente, ya que, un modelo solo puede ser igual o peor que la calidad de los datos. Por esto mismo, es de suma importancia el contar y entender con lo que se está trabajando y la importancia de solo usar lo que nos interesa y realizar el acomodamiento de manera eficiente.

Modelado

Modelo árboles de decisión

Para el modelado de nuestra información, probamos varios modelos diferentes, pero el primer modelo de los que creamos fue un árbol de decisión. Consideramos que este modelo era un buen punto de partida, ya que se pueden crear muy rápido y en general, pueden tener buenos rendimientos. Basándonos en los resultados que fuimos obteniendo, conforme fuimos generando diferentes árboles, fuimos variando en ciertos aspectos y funciones para intentar mejorar el rendimiento de cada uno de los modelos que fuimos probando. A continuación sé

muestra un desglose del proceso que se llevó a cabo para la creación de los modelos implementados.

Metodología

División de datos

Antes de poder comenzar con la creación de cada uno de los modelos que íbamos a generar fue necesario realizar un proceso de división de datos. Esta división es necesaria, ya que a la hora de generar modelos no se puede usar toda la base de datos porque si se usara no habría manera de probar si el modelo está funcionando correctamente. El primer paso de esta división de datos implicó la eliminación de la columna que se busca predecir, puesto que esta no se puede incluir en los datos que se utilizan. Ya habiendo eliminado esta columna se utilizó una función para hacer una división aleatoria del dataset en una muestra de entrenamiento y una muestra de prueba. En este caso, decidimos utilizar el 80% de los datos para el entrenamiento y dejamos el otro 20% para la validación y las pruebas. Después de realizar este proceso, se comenzó a crear los diferentes modelos.

Experimentos

Para nuestros experimentos, realizamos 7 modelos, los cuales se indican a continuación con sus parámetros utilizados y sus resultados.

Árbol de decisión

Iniciamos nuestra etapa de experimentación, utilizando la biblioteca sklearn, para entrenar un modelo de árbol de decisión como nuestro modelo simple, y así evaluar su rendimiento. Este modelo divide iterativamente los datos en subconjuntos basándose en el valor de la variable 'ingreso_prom'. Primero divide los datos en subconjuntos, un 80% de los datos para entrenamiento y un 20% para la prueba. Se limita el modelo a una profundidad de 4 árboles y entrena los datos de entrenamiento.

Para la evaluación del modelo, se calculan las siguientes métricas con sus resultados:

- La exactitud (0.7468) es la proporción de predicciones correctas sobre todas las predicciones realizadas, lo que nos indica que el 74.68% de los casos el modelo predijo correctamente.
- La precisión (0.6543) es la proporción de predicciones positivas correctas, lo que nos indica que cuando el modelo predice que un caso es positivo, es correcto el 65.43% de las veces.
- La sensibilidad (0.5958) es la proporción de casos positivos reales que el modelo pudo identificar correctamente, el cual el modelo tuvo un rendimiento de 59.58% correcto de los casos positivos reales.
- El puntaje F1 (0.6237) es la medida que combina la precisión y la sensibilidad, el cual obtuvo un valor de 62.37% lo que nos dice que el modelo tiene un rendimiento equilibrado.
- El área debajo de la curva ROC (71.239) significa que el modelo tiene una capacidad del 71.24% para distinguir entre clases.

Estos resultados muestran que el modelo tiene un rendimiento a nuestro parecer eficiente, donde resalta más al predecir valores negativos. Para analizar los valores de una manera más interpretable, se obtuvo una matriz de confusión que nos brinda la cantidad de predicciones erróneas o correctas, positivas y negativas que el modelo predijo. En este caso se obtuvieron:

- Verdaderos positivos: 19395 predicciones correctas de la clase positiva.
- Verdaderos negativos: 7579 predicciones correctas de la clase negativa.
- Falsos positivos: 4003 predicciones incorrectas de la clase positiva.
- Falsos negativos: 5140 predicciones incorrectas de la clase negativa.

Esto muestra resultados positivos, ya que los falsos valores son menores a los verdaderos y muestra una relación a favor de las métricas positivas. Es decir, los errores que son los valores falsos son menores que los resultados verdaderos, lo cual es bueno.

Árbol de decisión con Grid Search

Para optimizar nuestro modelo, tomamos la decisión de implementar la GridSearch. Esta es una técnica de optimización que busca exhaustivamente a fuerza bruta por medio de un conjunto de hiperparámetros. Con este modelo optimizado, se busca encontrar la combinación de hiperparámetros que produzca el mejor resultado posible.

Los parámetros definidos al realizar un modelo optimizado por medio de GridSearch, son los siguientes:

- Se prueban los criterios de ‘gini’ y ‘entropy’ para la división de los nodos del árbol.
- Se prueban las profundidades máximas de 3, 5, 7, y ninguna restricción para el árbol.
- Se prueban los valores mínimos 2, 5, y 10 para el número de muestras requeridas para dividir un nodo interno.
- Se prueban los valores mínimos 1, 2, y 4 para el número de muestras requeridas para ser una hoja del árbol.

Para la evaluación del modelo optimizado, se calculan las mismas métricas:

- La exactitud es de 0.7468.
- La precisión es de 0.6543.
- La sensibilidad es de 0.5958.
- El puntaje F1 es de 0.6237.
- El área debajo de la curva ROC es de 71.239.

Para la matriz de confusión:

- Verdaderos positivos: 19395 predicciones correctas de la clase positiva.
- Verdaderos negativos: 7579 predicciones correctas de la clase negativa.
- Falsos positivos: 4003 predicciones incorrectas de la clase positiva.
- Falsos negativos: 5140 predicciones incorrectas de la clase negativa.

Se puede observar que obtuvimos los mismos resultados con el GridSearch al igual que sin la optimización. Esto nos brinda las teorías de que los hiperparámetros por defecto del modelo ya

son los óptimos para nuestros datos o que simplemente no incluimos valores que mejoren el rendimiento.

Sampling
Son métodos utilizados para tratar con bases de datos no balanceadas, las cuales se caracterizan por tener más de la mitad de las entradas pertenecientes a una sola clase. Para resolver este problema necesitamos usar métricas que consideren este desequilibrio.

Esta técnica busca pre procesar la información de entrenamiento para minimizar la brecha entre las clases, modificando las distribuciones en dicho set.

Árbol de decisión con Over-sampling
Over-sampling incrementa el número de instancias de la clase minoritaria replicándose.

No hay pérdida de información, pero al crear nuevos datos puede conducir a costos computacionales altos. Además, si algunos de los datos a replicar contienen errores, al añadirlos deteriorará el rendimiento de la clasificación en la clase minoritaria.

En el dataset, hasta el momento hay 180583 registros, los cuales como ya se mencionó se dividen en 20% de datos de testeo y 80% de datos de entrenamiento. El modelo se concentra en los datos de entrenamiento que constan de 144466 registros que a su vez se dividen en 0 (≤25000) y 1(>25000).

En la siguiente tabla se aprecia la forma que tenían los datos antes y después de la aplicación del modelo. Se observa que la clase minoritaria es la 1, esa es la que el modelo incrementó hasta obtener la misma cantidad de registros que la clase dominante.

	Antes	**Después**
0	93993	93993
1	50473	93993

Al implementar el modelo, se calculó:

- La exactitud (0.6879) nos indica que cerca del 69% de las predicciones que se hicieron en total fueron correctas.
- La precisión (0.5564) nos muestra que aproximadamente el 56% de las instancias positivas que el modelo predijo fueron correctas.
- La sensibilidad (0.5621) que mide la capacidad del modelo para identificar las verdaderas instancias positivas tuvo un rendimiento del 56%.
- El puntaje F1 (0.5592) nos ayuda a equilibrar la visión que se tiene al conocer la precisión y la sensibilidad, considera ambos aspectos. El valor brindado por el modelo señala que el modelo está moderadamente balanceado.

También se obtuvo la matriz de confusión que nos brinda la cantidad de predicciones erróneas o correctas, positivas y negativas que el modelo predijo, en este caso se obtuvieron:

- Verdaderos positivos: 7149 predicciones correctas de la clase positiva.
- Verdaderos negativos: 17698 predicciones correctas de la clase negativa.
- Falsos positivos: 5700 predicciones incorrectas de la clase positiva.
- Falsos negativos: 5570 predicciones incorrectas de la clase negativa.

Debido a la cantidad en los falsos positivos y falsos negativos, podemos decir que hay un margen para mejorar la capacidad del modelo para clasificar las instancias. Y esta idea se ve reforzada con la métrica AUC (0.6592) que señala que el modelo tiene una capacidad moderada para clasificar las instancias de manera correcta.

Grid search Over-sampling
Los parámetros utilizados para encontrar las mejores métricas de este modelo son los mismos que se emplearon en el modelo optimizado previo.

Las mejores métricas encontradas fueron:

- La exactitud pasó de indicar que el modelo predijo correctamente aproximadamente el 69% (0.6879) de los registros a señalar un 70% (0.6960), lo cual representa una mejora cercana al 1%
- La precisión ha experimentado una mejora significativa, aumentando de 0.5564 a 0.568, lo cual representa un incremento de aproximadamente 0.01.
- La sensibilidad que medía 0.5621 incrementó a 0.5712.
- El puntaje F1 aumentó de 0.5592 a 0.5696.

También se obtuvo una nueva matriz de confusión:

- Verdaderos positivos: 7265 predicciones correctas de la clase positiva.
- Verdaderos negativos: 17873 predicciones correctas de la clase negativa.
- Falsos positivos: 5525 predicciones incorrectas de la clase positiva.
- Falsos negativos: 5454 predicciones incorrectas de la clase negativa.

Aunque hubo una mejora, no fue muy significativa, el modelo sigue mostrando un AUC moderado (0.6673) que indica una capacidad razonable de clasificación, pero aun así con un gran margen de mejora.

Árbol de decisión con Under-sampling
Under-sampling extrae un set pequeño de instancias de la clase dominante mientras conserva todas las instancias de la clase minoritaria. Esto es posible en bases de datos grandes donde el número de instancias en la clase dominante es demasiado alta.

Esta técnica puede llevar a una pérdida de información que degrade el rendimiento del clasificador.

El modelo se concentra en los datos de entrenamiento que constan de 144466 registros que a su vez se dividen en 0 (≤25000) y 1 (>25000).

En la siguiente tabla se aprecia la forma que tenían los datos antes y después de la aplicación del modelo. Se observa que la clase minoritaria es la 1, esa es la que el modelo tomó como base para extraer esa misma cantidad de registros de la clase dominante.

	Antes	**Después**
0	93993	50473
1	50473	50473

Al implementar el modelo, se calculó:

- La exactitud (0.6680) nos indica que cerca del 67% de las predicciones que se hicieron en total fueron correctas.
- La precisión (0.5223) nos muestra que aproximadamente el 52% de las instancias positivas que el modelo predijo fueron correctas.
- La sensibilidad (0.6684) que mide la capacidad del modelo para identificar las verdaderas instancias positivas tuvo un rendimiento del 67%.
- El puntaje F1 (0.5864) brindado por el modelo señala que el modelo está moderadamente balanceado.

También se obtuvo la matriz de confusión que nos brinda la cantidad de predicciones erróneas o correctas, positivas y negativas que el modelo predijo:

- Verdaderos positivos: 8501 predicciones correctas de la clase positiva.
- Verdaderos negativos: 15624 predicciones correctas de la clase negativa.
- Falsos positivos: 7774 predicciones incorrectas de la clase positiva.
- Falsos negativos: 4218 predicciones incorrectas de la clase negativa.

Debido a la cantidad en los falsos positivos y falsos negativos, podemos decir que hay un margen para mejorar la capacidad del modelo para clasificar las instancias. Y esta idea se ve reforzada con la métrica AUC (0.6681) que señala que el modelo tiene una capacidad moderada para clasificar las instancias de manera correcta.

Grid Search Under-sampling

Los parámetros utilizados para encontrar las mejores métricas de este modelo son los mismos que se emplearon en los modelos optimizados previamente.

Las mejores métricas encontradas fueron:

- La exactitud pasó de indicar que el modelo predijo correctamente aproximadamente el 67% (0.6680) de los registros a señalar un 73% (0.7369), lo cual representa una mejora cercana al 6%.

- La precisión ha experimentado una mejora significativa, aumentando de 0.5223 a 0.6, lo cual representa un incremento de aproximadamente 0.07.
- La sensibilidad que medía 0.6671 incrementó a 0.762.
- El puntaje F1 aumentó de 0.5863 a 0.6711.

También se obtuvo una nueva matriz de confusión:

- Verdaderos positivos: 9691 predicciones correctas de la clase positiva.
- Verdaderos negativos: 16925 predicciones correctas de la clase negativa.
- Falsos positivos: 6473 predicciones incorrectas de la clase positiva.
- Falsos negativos: 3028 predicciones incorrectas de la clase negativa.

Hubo una mejora buena, el modelo muestra un AUC (0.7426) que indica una capacidad satisfactoria de clasificación, sin embargo, todavía se tiene un margen de mejora.

Comparación entre Over-sampling y Under-sampling

Los resultados indican que el método de Under-sampling optimizado mediante Grid Search mostró las mejores métricas y un rendimiento general superior en comparación con el Over-sampling. Este hallazgo sugiere que, para este conjunto de datos específico, la estrategia de reducir la cantidad de muestras de la clase mayoritaria mediante Under-sampling resulta más eficaz que la estrategia de aumentar las muestras de la clase minoritaria mediante Over-sampling.

Sin embargo, aunque el rendimiento fue satisfactorio en general, aún se identifica un margen de mejora para optimizar aún más la precisión y la eficiencia de los modelos.

Modelo Random Forest

Este modelo combina varios clasificadores de árboles de decisión en diferentes submuestras y se ajusta con estos para mayor precisión y controlar el sobre ajuste. Este es uno de los modelos más certeros, por lo que podría ser efectivo el utilizar este método.
Primero se decidió usar el método estándar de random forest. Obteniendo los siguientes resultados en la matriz de confusión y sus métricas:

Estos son muy buenos resultados para el modelo, en relación con los demás.
Lo que podemos entender de estas métricas:
Exactitud: 0.755, vemos que el 75% de las predicciones son correctas, siendo de las mejores hasta ahora.
Precisión: 0.728, nos dice la proporción de verdaderos positivos entre todas las predicciones positivas, por lo que vemos que hay menos falsos positivos que en los demás modelos.
Sensibilidad (recall): 0.496, un recall bajo nos dice que logra predecir mejor los negativos, por lo que vemos que es mejor para estas categorías, lo que le quita credibilidad a sus predicciones positivas en comparación.

Puntaje F1: 0.590, combina precisión y recall, un valor alto dice que hay equilibrio para predecir la clase positiva y verdaderos positivos, por lo que tiene un ligero desequilibrio entre estos dos.

Este modelo nos da buenos resultados, pero no es lo ideal y se le dificulta predecir a la gente que gana más de 25000 pesos al trimestre, esto puede ser porque el modelo entienda mejor a los negativos tras tener mayor cantidad, para ver los resultados de manera más interpretativa:

- Verdaderos positivos: 6310 predicciones correctas de la clase positiva.
- Verdaderos negativos: 21052 predicciones correctas de la clase negativa.
- Falsos positivos: 2346 predicciones incorrectas de la clase negativa.
- Falsos negativos: 6409 predicciones incorrectas de la clase positiva.

Optimizando el modelo con Grid Search
En esto se van buscando los diferentes parámetros que mejor convengan para el modelo, investigando uno por uno cuál es la mejor configuración y nos muestra el mejor resultado. Obteniendo los siguientes resultados en la matriz de confusión y sus métricas:

Estos son muy buenos resultados para el modelo, en relación con los demás.
Lo que podemos entender de estas métricas:
Exactitud: 0.755, vemos que el 75% de las predicciones son correctas, siendo de las mejores hasta ahora.
Precisión: 0.734, nos dice la proporción de verdaderos positivos entre todas las predicciones positivas, por lo que vemos que hay menos falsos positivos que en los demás modelos.
Sensibilidad: 0.476, vemos que es mejor para los negativos, no es muy efectivo con los positivos.
Puntaje F1: 0.578, combina precisión y recall, un valor alto dice que hay equilibrio para predecir la clase positiva y verdaderos positivos, por lo que tiene un ligero desequilibrio entre estas dos clases.

Como vemos, nos da resultados similares, lo que podría indicar que el modelo original ya tenía la configuración adecuada para nuestros datos, para ver los resultados de manera más interpretativa:

- Verdaderos positivos: 6063 predicciones correctas de la clase positiva.
- Verdaderos negativos: 21207 predicciones correctas de la clase negativa.
- Falsos positivos: 2191 predicciones incorrectas de la clase negativa.
- Falsos negativos: 6656 predicciones incorrectas de la clase positiva.

Es un buen modelo con valores muy similares al modelo original de Random Forest.

Random forest con Grid Search y undersampling

Under-sampling es una técnica para abordar el desequilibrio de clases en conjuntos de datos, al usarla en random forest implica seleccionar de manera aleatoria ejemplos de la clase con más muestras y eliminarlas del conjunto de entrenamiento. Esto con el objetivo de que tengan casi la misma cantidad de datos de cada clase, así equilibrando la distribución, aparte de que se combina con el Grid search, buscando la mejor configuración para este nuevo set de datos. Obteniendo los siguientes resultados en la matriz de confusión y sus métricas:

Estos son muy buenos resultados para el modelo, en relación con los otros modelos generados.

Lo que podemos entender de estas métricas:

Exactitud: 0.739, vemos que el 73% de las predicciones son correctas, bajando ligeramente de los demás random forests.

Precisión: 0.602, nos dice la proporción de verdaderos positivos entre todas las predicciones positivas, por lo que vemos que hay un poco más de falsos positivos que en los demás modelos de random forest.

Sensibilidad: 0.764, que no tiene tanto sesgo a la hora de hacer sus predicciones, siendo una buena mejora a los demás.

Puntaje F1: 0.673, combina precisión y sensibilidad, un valor alto dice que hay equilibrio para predecir la clase positiva y verdaderos positivos, por lo que tiene un menor desequilibrio entre estos dos.

Como vemos, a pesar de que bajaron ligeramente algunas de las métricas, son unos buenos resultados, ya que todas las métricas son relativamente altas, todas siendo mayores a 0.6, y en los demás teníamos algunos 0.4, para ver los resultados de manera más interpretativa:

- Verdaderos positivos: 9725 predicciones correctas de la clase positiva.
- Verdaderos negativos: 16970 predicciones correctas de la clase negativa.
- Falsos positivos: 6428 predicciones incorrectas de la clase negativa.
- Falsos negativos: 2994 predicciones incorrectas de la clase positiva.

Este modelo como vemos, ha sido de los mejores, ya que muestra un buen rendimiento en todas las métricas y proporciones aceptables al ver los resultados, a comparación de los demás este modelo ha sido el más efectivo en ambas clases.

Métricas de desempeño

Modelo	Verdaderos positivos	Falsos positivos	Verdaderos negativos	Falsos negativos
Árbol de decisión	7579	4003	19395	5140
Árbol de decisión con grid search	7579	4003	19395	5140
Árbol de decisión con Grid search y under-sampling	9691	6473	16925	3028

Árbol de decisión con grid search y over-sampling	7265	5525	17873	5454
Random forest	6063	2191	21207	6656
Random forest con grid search	6063	2191	21207	6656
Random forest con grid search y under-sampling	9725	6428	16970	2994

Con respecto a todos los modelos realizados y evaluados, podemos analizar de esta forma en conjunto los resultados obtenidos por la matriz de correlación por cada modelo, siendo el caso del modelo que mayor cantidad de datos idóneamente analizados e identificados en la categoría de verdaderos positivos fue el modelo random forest con grid search y under-sampling, mientras que, en la categoría de verdaderos negativos fueron los modelos de random forest y random forest con grid search los que mejores resultados tuvieron.

Siendo una señal referente a que el mejor modelo realizado se encuentra dentro de la variedad de los random forest, puesto que, esta alta cantidad de datos correctamente analizados detecta que estos modelos cuentan con un alto rendimiento, en referencia a su alta precisión en cuanto a sus predicciones, una excelente capacidad de generalización para los datos del conjunto de evaluación y su bajo error dentro de estas categorías.

Tabla Comparativa de resultados generales.

Modelo	Precisión	Recall	Accuracy	F1 score	AUC
Árbol de decisión	0.6544	0.5959	0.7469	0.6238	0.7123
Árbol de decisión con grid search	0.6544	0.5959	0.7469	0.6238	0.7123
Árbol de decisión con grid search y under-sampling	0.5995	0.7619	0.7369	0.6711	0.7426
Árbol de decisión con grid search y over-sampling	0.5680	0.5712	0.6960	0.5696	0.6675
Random forest	0.7346	0.4767	0.7550	0.5782	0.6915
Random forest con grid search	0.7346	0.4767	0.7550	0.5782	0.6915
Random forest con grid search y under-sampling	0.6021	0.7646	0.7391	0.6737	0.7449

Posteriormente, se realizó un conjunto de las demás métricas de desempeño aplicadas para cada modelo utilizado y evaluado durante este proyecto. De modo a como se precisó anteriormente, existe una mayor certeza que el mejor modelo aplicado dentro de este extenso

análisis sea de la categoría de los random forest, pues, es en esta comparación donde se estudian las métricas de precisión, recall, accuracy, F1 score y la curva AUC-ROC.

Considerando los resultados presentes se puede inferir que el modelo con mejor precisión ante la dispersión de los conjuntos de datos por mediciones repetidas de una magnitud fue de los modelos random forest y random forest con grid search, ya que, cuentan con una buena precisión, por ende, los valores del modelo cuentan con una menor dispersión entre ellos. Por parte de la métrica recall, fue el modelo de random forest con grid search y under-sampling el que obtuvo un mejor ratio de valores en la categoría de verdaderos positivos, indicando que la cantidad de valores positivos son correctamente clasificados haciendo de esta forma al modelo uno de los más precisos y sensibles.

Por otro lado, el análisis de accuracy o la precisión de los datos nos permite identificar la fracción de predicciones que el modelo realizó, siendo los modelos random forest y random forest con grid search los que cuentan con una métrica idónea en la exactitud global, así como, un conjunto de datos bien balanceados. En adición, resulta ser que el modelo de random forest con grid search y under-sampling el que mejores resultados ha obtenido tanto para la métrica de F1-score como para la curva AUC-ROC, lo cual nos demuestra el estimado de la capacidad de clasificación del modelo de la media armónica de la precisión y el recall, siendo esta más precisa ante las demás afirmando su bajo desequilibrio en el análisis de los datos, así como, la curva AUC-ROC nos decreta el área bajo la curva de las características de funcionamiento del receptor, la cual, mide la capacidad de diferenciar entre los datos del modelo siendo del caso de un certero y sensible ante una clasificación binaria y su umbral de discriminación.

Resultados

Selección

Siendo evidente que el mejor modelo aplicado fue el de random forest con grid search y under-sampling. Acorde a este modelo podemos identificar que en la matriz de confusión se determinaron 16,970 datos como negativos verdaderos y 9725 datos como positivos verdaderos, contando con solo, 9422 errores de categorización de los datos. La dispersión de los conjuntos por mediciones de una magnitud representa una precisión considerable, por ende, los valores tienden a contar con cierta dispersión, aunque esta resulta ser una de las más mínimas entre todos los modelos.

Así como el ratio de verdaderos positivos indica una buena clasificación al modelo siendo más preciso y sensible ante nuevos datos a evaluar, al igual que, la fracción de predicciones en la precisión del modelo demostrando un balance bien aplicado como también de una métrica idónea en la exactitud global. Englobándose en la capacidad de clasificación de la media armónica, siendo el modelo más preciso y de bajo desequilibrio, como se ve aplicado en el gráfico de la curva AUC-ROC.

Validación Cruzada

Posteriormente, se realizó un modelo de validación cruzada para el mejor modelo, del cual se obtuvo el promedio de la precisión del conjunto de entrenamiento, siendo de 0.7671, así como

el promedio de la precisión del conjunto de prueba de 0.7649, siendo esta afortunadamente de una diferencia mínima, lo cual representa la precisión y habilidades con las que cuenta el modelo con base en la técnica de evaluación por medio del conjunto de entrenamiento y de validación para predecir resultados en un conjunto de datos externa o nueva. A su vez, estos buenos resultados indican que se evita el sobre ajuste y se proporciona una estimación más eficiente en el rendimiento del modelo ante su mejor robustez de generalización ante nuevos datos.

Evaluación (del modelo* seleccionado)

Análisis de desempeño

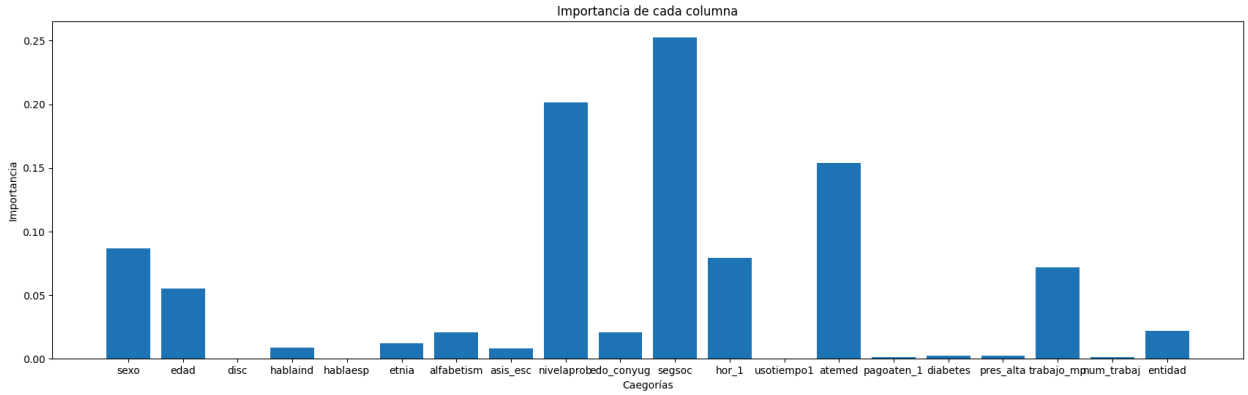
El mejor modelo, Random Forest con Grid search y Under-sampling, es eficiente en identificar la mayoría de los casos negativos. Esto es importante, ya que puede detectar muy bien si un ciudadano está por debajo del umbral de $25,000 que se definió como la variable de predicción. Aunque la precisión es más baja, la alta sensibilidad resulta en un buen F1 score, lo que indica un balance razonable entre la precisión y sensibilidad.

Esta precisión baja sugiere que una proporción significativa de las predicciones positivas son falsos positivos. Este porcentaje puede ser problemático al tratar de predecir si un ciudadano está arriba del umbral cuando en realidad no es así. El desempeño también fue menor que otros modelos, pero por diferencias insignificantes. Se puede apreciar que este modelo se ajusta para maximizar la sensibilidad a costo de una buena precisión.

Donde mejor resalta el modelo a comparación de los demás, es en el score AUC. Es bueno para diferenciar entre los casos, lo que es crucial para un modelo efectivo. Con esto y un F1 score decente, muestra que el modelo logra un buen equilibrio entre precisión y sensibilidad, proporcionando una clasificación general efectiva.

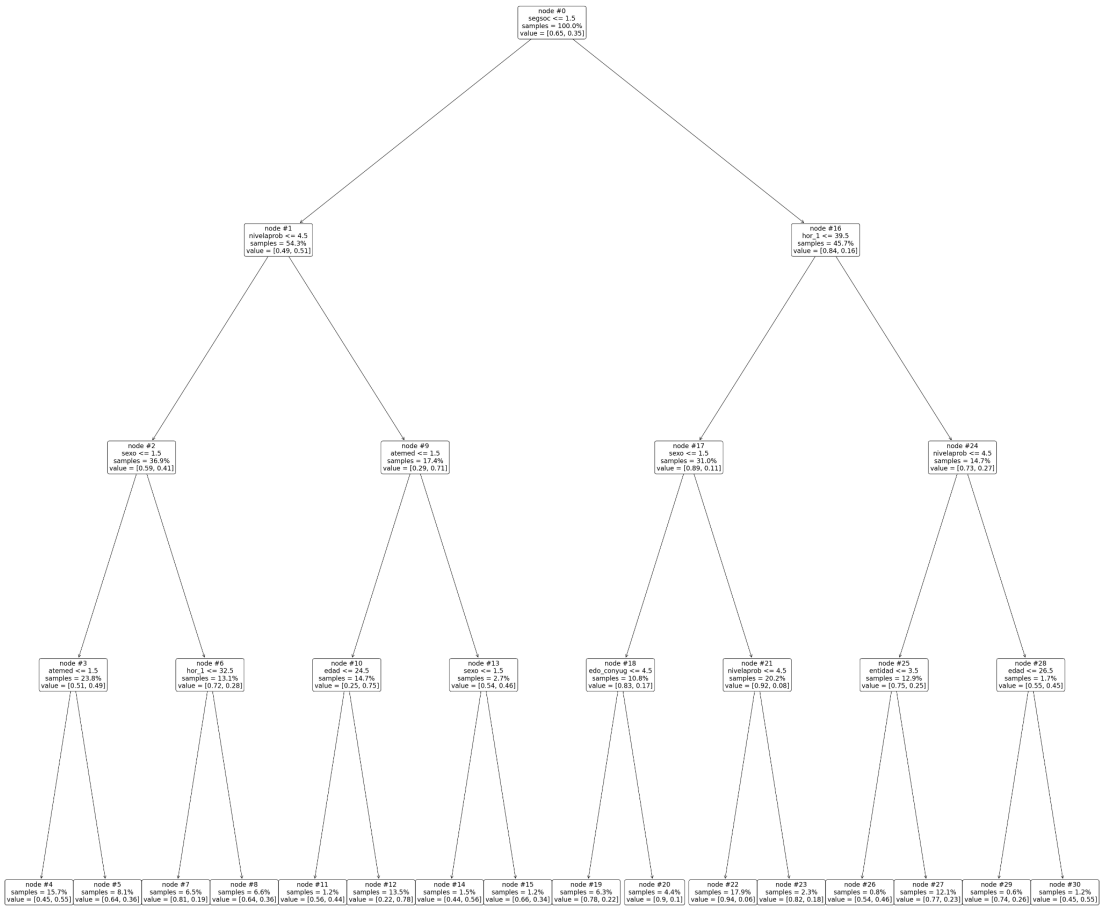
Análisis de características importantes

En este análisis vemos que tanto atribuye cada categoría en la predicción del ingreso con respecto al modelo que obtuvo los mejores resultados, siendo el Random Forest GS under-sampling, para así ver qué es lo que más afecta en el modelo y cuál es la importancia de cada categoría. Una característica con mayor importancia afecta más a los resultados y una con menor importancia afecta menos a los resultados obtenidos. Vemos los resultados de nuestra comparación.

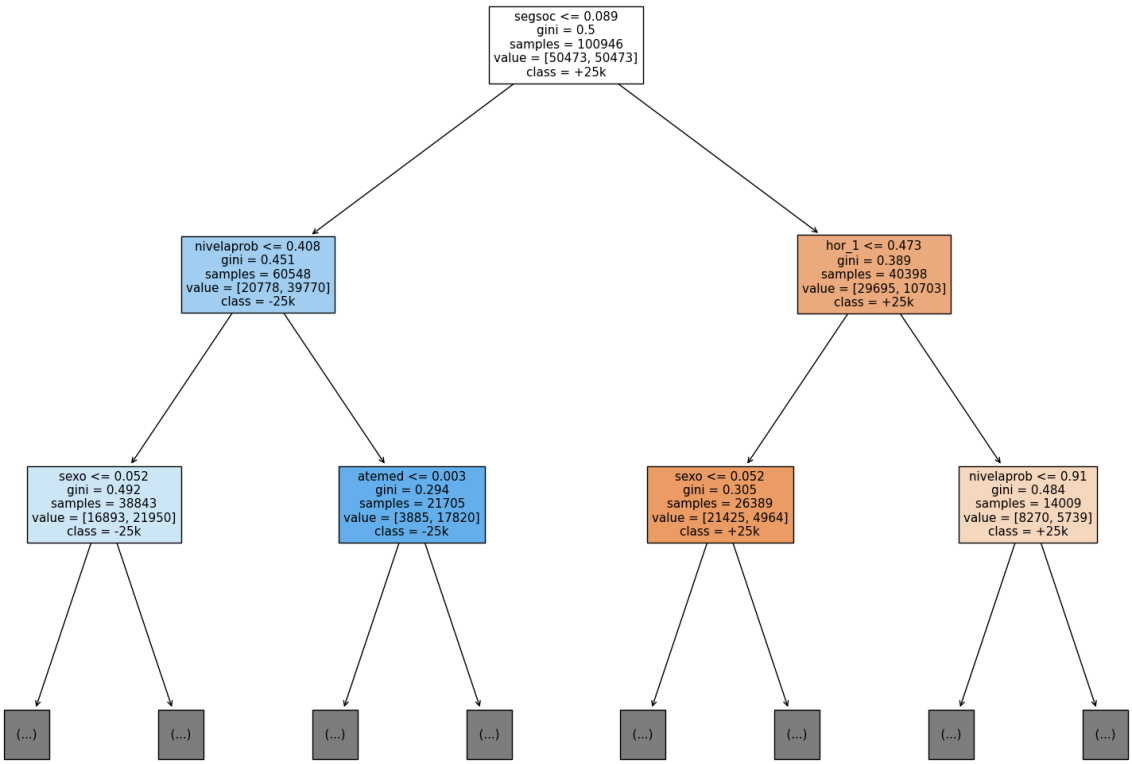


Como vemos en la gráfica tenemos algunas categorías que afectan en gran escala nuestros resultados en este modelo, la que más tiene efecto en el modelo es ‘segsoc’, siendo si la persona ha contribuido a un seguro social, después está ‘nivelaprob’, el nivel educativo al que llegó la persona y ‘atemed’, si está afiliado a recibir o no atención médica.
Tras ver esta gráfica podemos tener una mayor idea de la relación que tienen las variables con el ingreso y cuáles no tienen tanta relación a comparación de las demás, lo que nos puede dar un mayor entendimiento del ingreso con la situación de la gente.

Visualización e interpretación de los árboles



El árbol de decisión muestra cómo se toman decisiones en función de diferentes características y valores, subdividiendo los datos hasta llegar a una predicción final en los nodos hoja. Los nodos hoja contienen la proporción de clases que se usa para hacer la predicción final. Se puede observar que se inicia con la característica ‘segsoc’ y se va subdividiendo en distintas características del dataset.



Para la creación de este árbol, se limitó la cantidad de nodos y ramas en las que se puede extender. Poner un límite en su extensión nos ayuda a poder apreciar mejor los valores y las decisiones que va tomando el árbol en cada uno de los nodos de decisión. Se puede ver que el último nivel de este árbol muestra puros nodos con tres puntos para indicar que el árbol sigue.

Implicaciones y limitaciones del modelo y su uso

El modelo elegido es Random Forest con Under-sampling y Grid search para encontrar las mejores métricas.

Este modelo nos ayuda a manejar los datos desbalanceados del dataframe, tomando en cuenta que hay una mayor cantidad de una clase que de otra con el objetivo de mejorar el rendimiento del modelo y evitando un sesgo para la clase mayoritaria. Con todo, al eliminar registros de la clase mayoritaria puede resultar en una pérdida de patrones importantes, afectando directamente la capacidad del modelo para generalizar.

Con respecto a la aplicación de Grid search para dicho modelo, podemos decir que es muy útil para encontrar el mejor rendimiento del modelo, y personalizándolo con los hiperparámetros seleccionados, obteniendo un modelo preciso y robusto. Sin embargo, debido a la elección manual de los hiperparámetros, se corre el riesgo de un sobre ajuste al dataset de entrenamiento, ocasionando que el modelo no pueda generalizar correctamente los datos, aunando el aumento en el costo computacional que implica su ejecución.

En cuanto al uso del Random Forest, este nos ayuda a evitar un sobre ajuste a través de la reducción de la varianza en un dataset con ruido en los datos (método bagging) además de que es un método robusto a valores atípicos. No obstante, el incremento en el número de árboles genera un costo en términos de memoria y tiempo de cómputo.

Conclusiones

Al final de este análisis, hemos comprendido la importancia y las ventajas que presenta cada tipo de modelo. Este estudio también nos permitió observar que, en ocasiones, los modelos más precisos no son siempre la mejor opción debido al tiempo de procesamiento que requieren. Durante el análisis, enfrentamos algunos problemas con los modelos de Random Forest y sus optimizaciones, ya que en algunos casos llegaron a tardar más de 40 minutos en ejecutarse.

Además, nos dimos cuenta de que el preprocesamiento de la información es fundamental. Al limpiar y analizar los datos, pudimos identificar características importantes de las variables y detectar datos atípicos curiosos, como los que aparecieron en las columnas de edad y horas de trabajo. Este proceso nos permitió mejorar la calidad de los datos y, por ende, la precisión de los modelos.

Adicionalmente, aprendimos que al aplicar los modelos es esencial codificar las variables para que el modelo pueda comprender la información correctamente. También descubrimos que algunas columnas, aunque parecían contener datos numéricos, en realidad eran de tipo objeto. Esto nos obligó a analizar los datos con más detenimiento para asegurarnos de que todos los tipos de datos fueran correctos y coherentes.

El análisis gráfico y numérico también fue crucial, ya que nos permitió identificar y entender las relaciones entre diferentes variables. Estas visualizaciones y estadísticas nos proporcionaron una visión más clara y profunda de los datos, facilitando la interpretación y el desarrollo de estrategias basadas en los resultados.

Consideramos que este proyecto ha sido muy interesante para nosotros, por el hecho de que nos permitió aplicar los conocimientos adquiridos en el curso y desarrollar habilidades de resolución de problemas. En resumen, hemos aprendido que tanto el preprocesamiento de

datos como el análisis detallado y la correcta codificación de las variables son esenciales para obtener resultados precisos y valiosos en cualquier estudio de datos.

Referencias

INEGI. (2024) Quiénes somos por Nacional de Container: Inegi.org.mx URL: https://www.inegi.org.mx/inegi/quienes_somos.html

INEGI. (2024). *Instituto Nacional de Estadística y Geografía (INEGI)*. Inegi.org.mx. https://www.inegi.org.mx/

Anexo 1.
1: Aguascalientes
2: Baja California,
3: Baja California Sur,
4: Campeche,
5: Coahuila de Zaragoza,
6: Colima,
7: Chiapas,
8:Chihuahua,
9: Cd. 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

Anexo 2.
Hipervínculo del código de Python
https://colab.research.google.com/drive/1RpQgPXRe9opG03bntfOlxHBEOU9M5zbd?usp=sharing