UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERIA

Departamento de Ingeniería Informática

Informalidad laboral en Chile: un análisis en base a la Encuesta Nacional de Empleo

Taller de Aprendizaje Automático Aplicado

Autor
Nicolás Torres Hormazábal

Profesor:

Erick González

Diego Machado

Francisco Muñoz

Santiago – Chile

Enero 2025

TABLA DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1. DEFINICIÓN DEL PROBLEMA

La informalidad laboral en Chile abarca tanto a empresas informales como a trabajadores informales. En cuanto al primero, el sector informal incluye a aquellas empresas que no han iniciado actividades en el Servicio de Impuestos Internos (SII), lo que las deja fuera del marco regulatorio, fiscal y de protección laboral. Respecto a los trabajadores informales, estos se definen como aquellos que, entre otras características, carecen de contrato de trabajo y/o no cuentan con seguridad social en su empleo. Esta situación afecta su estabilidad económica y limita su acceso a beneficios sociales y de salud, perpetuando ciclos de vulnerabilidad y desigualdad (INE, 2022, p. 23)

Este proyecto tiene como propósito, a partir de los datos de la Encuesta Nacional de Empleo del Instituto Nacional de Estadísticas (INE), desarrollar un modelo de ciencia de datos que explore y analice las características sociodemográficas, geográficas, nivel educativo y otras características que influyen en la probabilidad de que una persona opte o se vea obligada a trabajar en el sector informal. Entre estas características pueden encontrarse factores como el sexo, la edad, el nivel educativo, la región geográfica, la actividad económica y el tipo de ocupación, que se espera identificar y analizar en detalle a través del uso de técnicas de ciencia de datos.

La informalidad laboral en Chile es producto de múltiples factores estructurales, tales como las desigualdades en el acceso a la educación, la concentración de empleos formales en áreas urbanas, y el bajo nivel de fiscalización en ciertos sectores económicos. Estos factores se ven agravados en el caso de mujeres, jóvenes y personas de menor nivel educativo, quienes enfrentan mayores barreras para ingresar al mercado laboral formal.

La falta de acceso a un empleo formal tiene implicaciones profundas y de largo plazo para los trabajadores. A nivel individual, implica la ausencia de un contrato que regule sus derechos, la falta de seguridad social, y la imposibilidad de acceder a beneficios de salud y pensiones. A nivel macro, el trabajo informal afecta la recaudación fiscal, disminuye la eficiencia del mercado laboral y contribuye a perpetuar la desigualdad socioeconómica en el país (Livert-Aquino et al., 2022, pp. 13-15)

Desde la perspectiva de la ciencia de datos, el análisis de la informalidad laboral representa una oportunidad para identificar patrones complejos y relaciones ocultas en un problema multifactorial.

Este proyecto busca tener un impacto significativo en varios aspectos, siendo algunos:

Identificación de grupos vulnerables: Al identificar factores que predisponen a ciertos individuos a trabajar en el sector informal, el proyecto puede ayudar a focalizar programas de apoyo e incentivos de formalización laboral.

Información para políticas públicas: Los *insights* generados pueden servir como base para desarrollar políticas y programas orientados a la formalización laboral, dirigidos a los sectores de mayor riesgo.

Visualización de datos: La ciencia de datos permite visualizar el problema de la informalidad de una forma que facilite la toma de decisiones, haciendo accesible la información para todos los sectores involucrados.

2. FUENTES DE INFORMACIÓN Y TRABAJOS RELACIONADOS

Las fuentes consultadas para el desarrollo de esta investigación pueden dividirse en dos tipos:

a) Artículos académicos e informes de organismos nacionales e internacionales sobre la informalidad y el mercado laboral en general.
b) Artículos sobre manejo de datos en encuestas, particularmente referidos a manejo e imputación de datos faltantes

Se llegó a la conclusión de usar estos dos tipos de fuentes ya que es un problema de investigación que se encuentra ampliamente estudiado en las fuentes del tipo a). Se puede argumentar que este tipo de fuentes corresponden a "investigaciones sociales sobre informalidad", ya que se encargan, en base a datos como los de la ENE y de otros organismos públicos y privados, hacer una caracterización de causas y efectos de la informalidad, dimensiones donde está presente y quienes están propensos a pertenecer al sector informal.

Por otro lado, las fuentes de información de tipo b) corresponden a fuentes más congruentes con el curso, por ende, escapan del tema principal. En ellos se discuten de métodos para imputación de datos perdidos, métodos de visualización de datos para encuestas, así como estrategias para enfrentarse a encuestas estatales que buscan caracterizar a la población a la cual se le realiza.

Durante el proceso de selección de fuentes de información, se empezó buscando fuentes de información de tipo a). Esto se debe a que lo primero que se debe tener claro cuando se empieza a investigar de un tema es tener una noción de este. Se exploraron múltiples artículos que se detallan en la tabla 1:

Tabla 1

Artículos académicos de tipo a)

Artículo	Año	Autor (es)	Reseña
Estimación de la probabilidad de informalidad laboral a nivel comunal en Chile	2022	Felipe Livert Fidel Miranda Andrés Espejo	Este artículo de la CEPAL indaga más en el fenómeno de la informalidad y genera una probabilidad de informalidad para las comunas. Se acerca bastante a lo que pretendo hacer. Es la parte más teórica de mi proyecto.
Informalidad, productividad y flexibilidad laboral	2020	Nikita Céspedes	Articulo más general, aborda los conceptos claves para entender el fenómeno de la informalidad
Explaining the Shadow Economy in Europe: Size, Causes and Policy Options	2019	B. Kelmanson, K. Kirabaeva, L. Medina, B. Mircheva and J. Weiss	Explora este fenómeno en economías europeas, identifica factores de riesgo y entrega ideas para abordar el problema de la informalidad (o shadow economy) desde las políticas públicas

Estos artículos fueron de gran ayuda para tener una comprensión más amplia sobre la informalidad laboral (en inglés, *shadow economy*), el mercado laboral y su contexto, sobre todo en Latinoamérica y en Chile.

El artículo de la CEPAL hace un muy buen trabajo cuando se trata de caracterizar a las personas que optan por este tipo de sector, y los mapas de calor que generaron que demuestran los sectores de mayor probabilidad de informalidad tienen mucho sentido con la información que se obtiene de la ENE. En palabras de los autores: “La educación también es un factor determinante, ya que a más años de educación formal menor es la informalidad: en el segmento sin estudios la informalidad alcanza al 82,2% mientras que en la población con educación terciaria es del 33,5%. En cuanto a la dimensión territorial y económica, se identifica que la informalidad es mayor en zonas rurales (68,5%) que en zonas urbanas (47%), y es mayor en el sector agricultura (79,2%) que en servicios (49%) o industria (49,1%) (OIT, 2018). Finalmente, los trabajadores por cuenta propia tienen una tasa de empleo informal del 84,1%, mientras que empresas pequeñas (2 a 9 trabajadores) tienen una tasa del 72,4% (Salazar y Chacaltana, 2018).” (p. 10). Este documento ha sido un pilar fundamental para la selección de variables y el enfoque interpretativo que tendrá el presente.

También, como conclusiones de su estudio argumentan que es un fenómeno directamente relacionado con la desigualdad, género, territorio y grupo socioeconómico (Livert-Aquino et al., 2022, p. 65).

Las otras dos fuentes del apartado a) nos sirven para tener un contexto más global acerca del mercado laboral y la informalidad. Destacaré principalmente el artículo en inglés de Kelmanson et al. El artículo examina los factores que contribuyen a la informalidad y estima su tamaño en las economías europeas, proponiendo reformas para incrementar la formalidad, especialmente en economías emergentes.

Para abordar la informalidad, los autores recomiendan un paquete de políticas adaptado a cada país, incluyendo la "reducción de la carga regulatoria y administrativa, promoción de la transparencia y mejora en la efectividad gubernamental" (Kelmanson et al., 2019, p. 19). También enfatizan la importancia de la automatización de procedimientos y la promoción de pagos electrónicos para mejorar la recaudación fiscal y reducir la evasión (Kelmanson et al., 2019, p. 21).

El aspecto mencionado en el párrafo anterior es importante, ya que uno de nuestros objetivos en la investigación es generar conocimiento sobre las causas y efectos de la informalidad laboral, y qué rol debería tomar el Estado frente a este tema.

Como mencionamos, los artículos de tipo b) corresponden a aquellos que son del área de la ciencia de datos y están aplicados a manejar datos de encuestas, es decir, data sets que contienen en su mayoría datos de tipo categórico, así como también el manejo e imputación de datos nulos dentro de estas mismas bases de datos.

El detalle de los artículos revisados se encuentra en la tabla 2.

Tabla 2

Artículos académicos de tipo b)

Artículo	Año	Autor (es)	Reseña
The prevention and handling of the missing data	2013	Hyun Kang	Artículo enfocado a datos del área de salud (anestesiología) entrega información importante sobre los distintos tipos de datos perdidos que existen: MCAR, MAR MNAR, y cómo manejarlos
Multiple Correspondence Analysis and its applications	2017	Nutan Vijay Khangar, Kirtee Kiran Kamalja	Información general y procedimientos para manejar bases de datos con una predominancia categórica, así como también la aplicación del modelo Multiple Correspondance Analysis (MCA)
Métodos de Machine Learning como alternativa para la imputación de datos perdidos. Un ejercicio en base a la Encuesta Permanente de Hogares	2021	Germán Rosati	Este artículo está enfocado en manejar datos perdidos, pero enfocado a una encuesta de hogares e ingresos en Argentina, para eso utiliza métodos como Random Forest, XGboost, y redes neuronales.

Como se observa en la tabla 2, estos artículos abordan principalmente temas ya relacionados y específicos de la ciencia de datos, como son métodos de reducción de dimensionalidad, manejo e imputación de valores faltantes, así como estructuras a utilizar durante el manejo de bases de datos como las utilizadas en este proyecto, con pocas variables numéricas, un predominio de las categóricas, y con gran porcentaje de valores nulos.

3. EXPLICACIÓN Y CONTEXTUALIZACIÓN DE LA BASE DE DATOS

La base de datos utilizada para este análisis corresponde a la edición 2023 de la "Encuesta Nacional de Empleo" (ENE), elaborada por el Instituto Nacional de Estadísticas (INE) de Chile. Esta encuesta se realiza mensualmente y los formularios se incluyen en los anexos de este documento.

Diseño de Muestreo

La ENE emplea un diseño de muestreo basado en dos tipos de unidades:

1. **Unidad Primaria de Muestreo (UPM)**: Consiste en agrupaciones homogéneas de viviendas particulares. La UPM representa la primera etapa de selección en el muestreo. En zonas urbanas, una UPM tiene en promedio 200 viviendas (rango de 160 a 240), mientras que en zonas rurales incluye 90 viviendas (rango de 70 a 110), excluyendo viviendas de temporada.

2. **Unidad Secundaria de Muestreo (USM)**: Comprende las viviendas particulares ocupadas dentro de las UPM seleccionadas, garantizando que la muestra considere únicamente hogares activos y habitados (INE, 2022, p. 43).

Acceso a la Base de Datos

El archivo de datos está disponible en el sitio web del INE en la siguiente ruta: INE -> Bases de datos -> Bases anualizadas -> Formato CSV -> Año 2023. El archivo en formato CSV tiene un tamaño de 119 MB.

Exploración de la Base de Datos

Utilizando el método df.shape de Pandas, se verificó que la base de datos cuenta con 260,978 columnas y 172 filas. De estas columnas, 130 son de tipo float64, 31 de tipo int64 y 11 de tipo object. Aunque la mayoría de las columnas son numéricas, muchas corresponden a respuestas categóricas del cuestionario, por lo que serán consideradas como variables categóricas en el análisis. Las columnas específicas seleccionadas para el análisis se detallan en la sección de Análisis Exploratorio de Datos (EDA).

Análisis de Valores Nulos

La base de datos no presenta valores nulos en las variables clave para caracterizar a las personas encuestadas, tales como edad, tramo de edad, sexo, parentesco, nivel de estudios, estado conyugal, y otras variables relacionadas con el nivel educativo y la ocupación.

Sin embargo, se observó un alto porcentaje de valores nulos (algunos superiores al 90%) en ciertas preguntas del cuestionario que comienzan con una letra en minúscula seguida de un número, identificando la sección del cuestionario a la que pertenecen. Estos valores nulos se deben a la naturaleza excluyente de algunas preguntas; por ejemplo, si se responde a la pregunta "b1", no se responde a la "b3". Este comportamiento genera coincidencias en el número de valores nulos en diversas columnas. Existen métodos de imputación para este tipo de encuestas, que se han revisado en la literatura y se abordarán en la sección del EDA.

Detalle de Valores Nulos

El número de valores nulos en cada columna y el método de tratamiento se discutirán en la sección de descripción del EDA. Además, en los anexos de este documento se proporciona una tabla detallada con los valores nulos por variable.

4. ANALISIS EXPLORATORIO DE DATOS

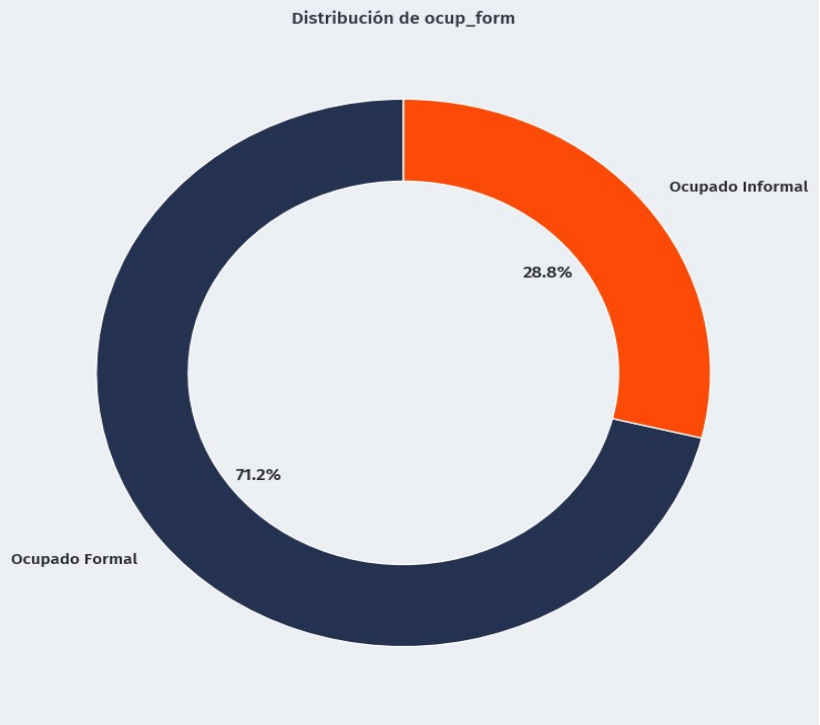
Tal y como se mencionó en la presentación sobre el proyecto de investigación, los objetivos del EDA para este trabajo son:

- Mostrar características del DataFrame
- Análisis univariado para características generales de la encuesta
- Análisis bivariado para relaciones entre variables, como horas de trabajo y sector o sexo
- Análisis multivariado para comprender qué columnas aportan información al DataFrame ya filtrado
- Identificación y aproximación a la imputación de datos faltantes

Considero importante partir este EDA con la cantidad de personas que están en el sector formal e informal, y a que es el principal tema de investigación, como se observa en la figura 1, la mayoría de las personas trabajan en el sector formal, con poco más del 70%, el resto corresponden a ocupados informales, si consideramos que la cantidad de personas que consideré para mi análisis (dejando fuera aquellos fuera de la fuerza de trabajo por razones de edad) son 260.545, el sector informal contiene a 72.956 personas.

Figura 1

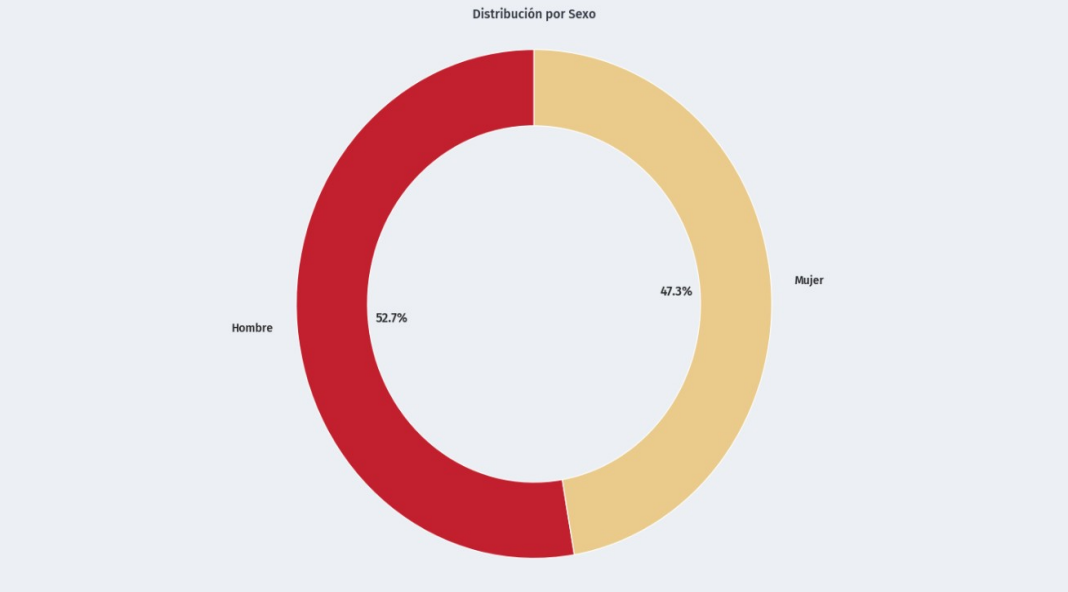
Sector de ocupación de las personas encuestadas



En esta misma línea, y para darnos una idea general del perfil de personas, tenemos el gráfico de dona de distribución de texto, como se representa en la figura 2, es un data set bastante equilibrado para ambos sexos, con casi un 50-50.

Figura 2

Distribución de sexo de los encuestados

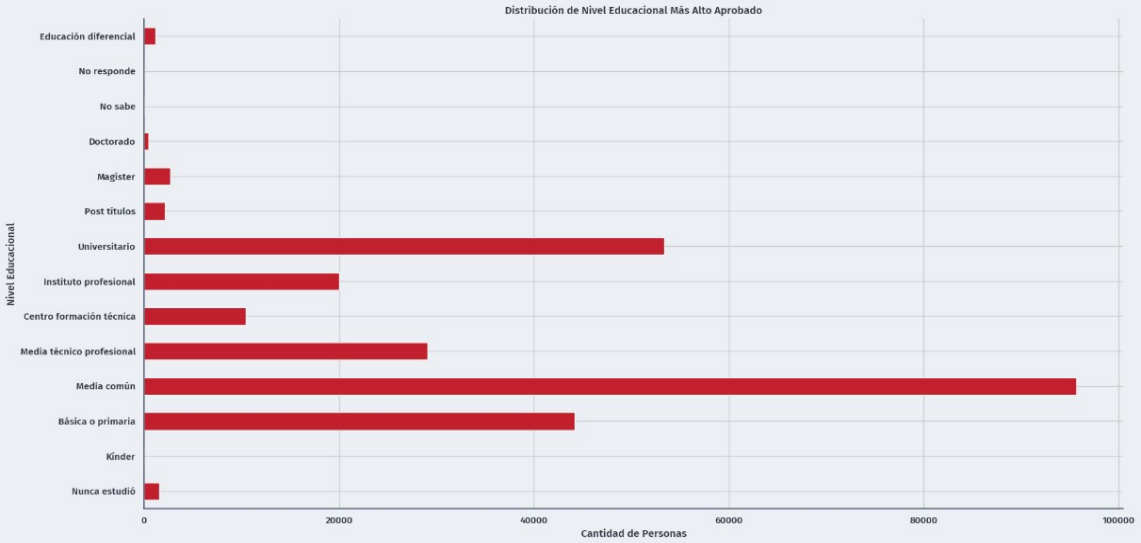


En la figura 3 se observa el nivel de educación de las personas seleccionadas para este trabajo, la mayoría de las personas tienen educación media completa, coloquialmente se le conoce como cuarto medio, luego están los universitarios y CFT. Es interesante mencionar que hay personas con educación básica o primaria, a pesar del filtro de edad. Al ser nivel más alto aprobado, se contabilizan a personas con solamente octavo básico, y personas de 15 a 18 años, que sí son parte de la fuerza de trabajo.

Esto es algo que se mencionó por parte de uno de los profesores durante la presentación, y la única información nueva que puedo presentar es lo anterior.

Figura 3

Cantidad de personas por nivel educativo más alto aprobado



El gráfico 4 muestra un mapa de palabras con las principales plataformas que mencionan los encuestados que sí trabajan en plataformas digitales. Vemos que la mayoría menciona a Facebook, Instagram y WhatsApp, que, si bien son redes sociales, son usadas como plataforma de trabajo, un ejemplo de esto son las tiendas de Instagram, el Marketplace de Facebook y las cuentas de empresas de WhatsApp. Luego está Uber, reconocida aplicación de transporte de pasajeros y su versión de delivery de comida, UberEats.

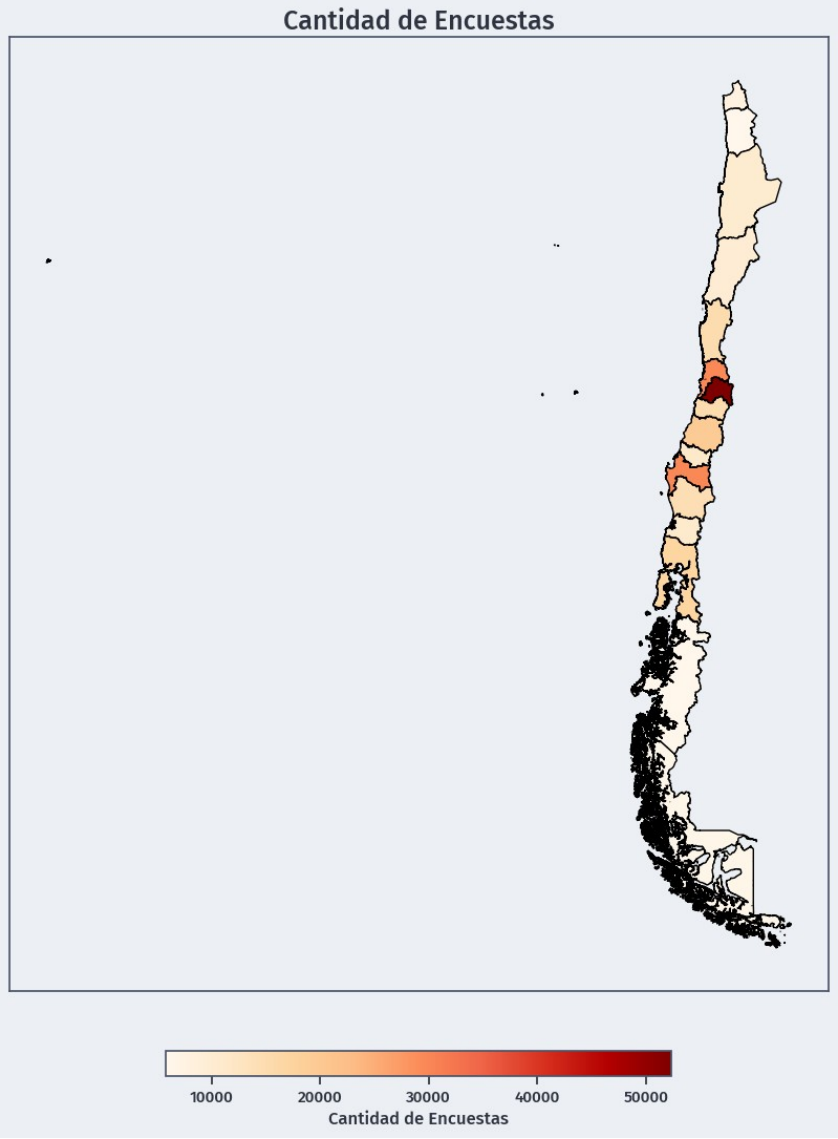
Figura 4

Mapa de palabras de principales plataformas digitales en la variable "pd_especifique"

La figura 5 muestra un mapa de calor de la cantidad de encuestas en el país. Como se observa, la mayoría de las encuestas se concentran en áreas con gran densidad demográfica, como lo son Santiago en la Región Metropolitana, la región de Valparaíso y la Región del Biobío, donde se encuentra Concepción. Las zonas australes y altiplánicas del país están coloreadas con tonalidades más claras, lo que indica que no existe gran cantidad de encuestas realizadas en estas regiones.

Figura 5

Mapa de calor de cantidad de encuestas por región



La figura 6 ilustra un gráfico de QQplot para la variable numérica de “edad”, como se observa, la diagonal roja representa una distribución normal, mientras que los puntos azules representan la distribución de la variable edad. Esto nos demuestra que edad no sigue una distribución normal, y

esto se confirma con la prueba de simetría presente en la tabla 3. Es importante mencionar que el data set que usaremos cuenta con tres variables discretas, estas son edad ya revisada, habituales y efectivas, que representan las horas que trabajan las personas encuestadas. Habituales es un promedio de tres meses anterior a la encuesta y efectivas son de la semana de referencia, partiendo por el domingo anterior al día donde se realizó la encuesta. Estas dos últimas comparten las mismas características que edad; no siguen una distribución normal y están sesgadas negativamente.

Figura 6

QQplot para la variable numérica "edad"

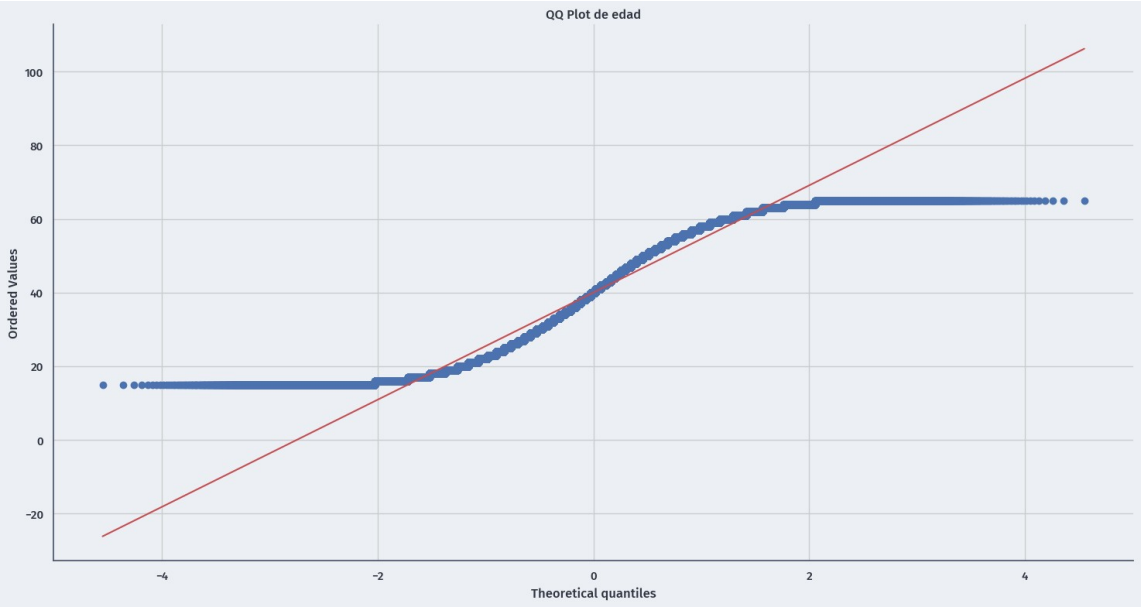


Tabla 3

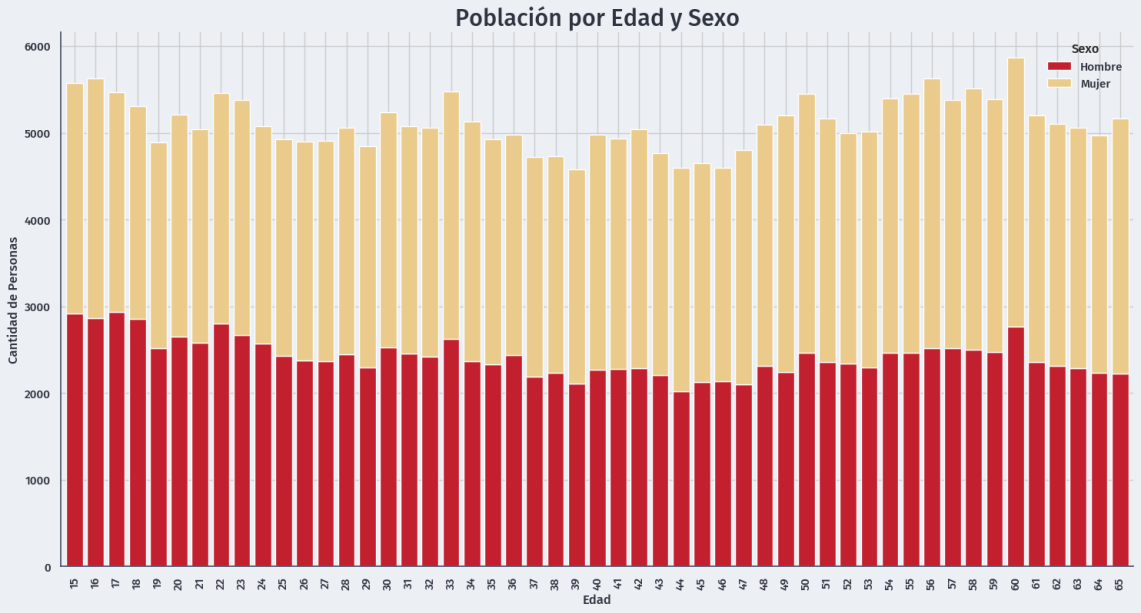
Pruebas estadísticas para la variable "edad"

Prueba	**Estadístico**	**p-valor**	**Resultado**
Kolmogórov-Smirnov (normalidad)	0.2438	0.0	No sigue una distribución normal
Asimetría (skew)	-0.3193	-	Distribución sesgada negativamente (hacia la izquierda)

La figura 7 presenta un análisis bivariado de la edad de las personas y el sexo, como se observa se sigue la tendencia de no-normalidad para la variable edad, y concordante con el data set general, no hay mayor sobre representación de un sexo sobre el otro

Figura 7

Gráfico de barras de población por edad y sexo



La figura 8 también es un análisis entre dos variables, esta vez viendo el estado de la actividad (variable “activ”) por sexo. Como se observa, la mayoría de las personas están ocupadas y existe una cantidad preocupante de personas fuera de la fuerza de trabajo. A su vez, la tabla 4 presenta el resultado de la prueba de chi-cuadrado que se realizó entre ambas variables categóricas. Estos resultados nos indican que existe una asociación importante entre ambas, esto es, el estado laboral de la persona tiene relación con su género.

Figura 8

Distribución de actividad por sexo

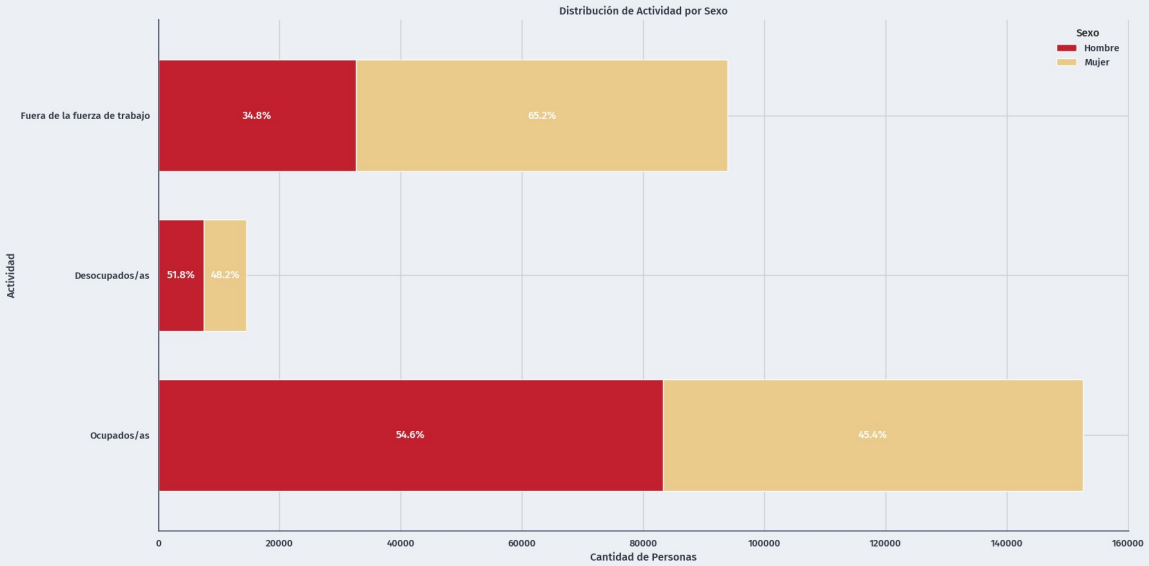


Tabla 4

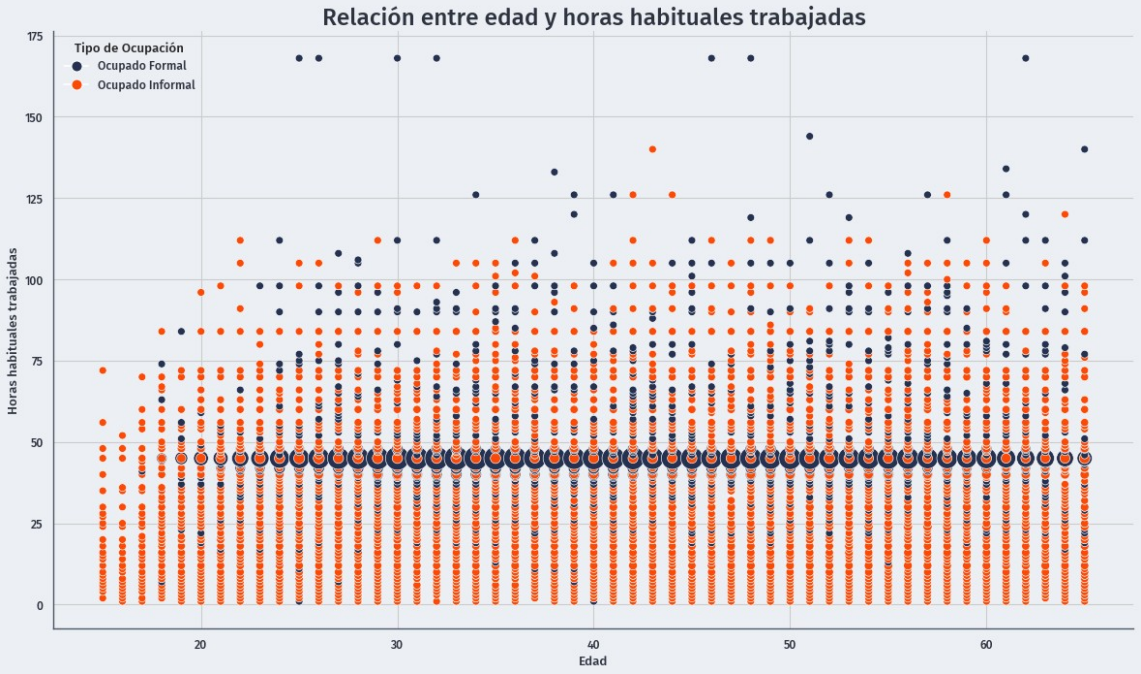
Resultado prueba chi-cuadrado para actividad y sexo

Prueba	**Valor Chi-cuadrado**	**Grados de libertad**	**Valor p**	**Resultado**
Chi-cuadrado	9.266.117	2	0.0	Existe una asociación significativa entre 'activ' y 'sexo' (rechazamos H0)

Por su parte, la figura 9 representa un análisis multivariado entre la edad de las personas, sus horas habituales de trabajo, y está coloreado por ocupación. Como se observa, y se confirma con la figura 10 y su respectiva prueba estadística (tabla 5), existe una diferencia significativa en el área de ocupación y la cantidad de horas que se trabajan, tanto para formales como informales, el sector hogar como empleador (esto es, trabajadores de casas particulares) y la edad. Las conclusiones que saco de estos dos gráficos son que el sector en el que se desarrolle el trabajador y la edad influyen directamente a la cantidad de horas semanales que se trabajan. Si bien esto puede parecer obvio a primera vista, es importante contar con datos empíricos que respaldes estas inferencias.

Figura 9

Gráfico de puntos de edad y horas habituales, por tipo de ocupación



Como se mencionó anteriormente, en la figura 10 aparece un gráfico de caja para la distribución de horas por sector de ocupación.

Figura 10

Gráfico de caja y bigote de horas habituales por sector de ocupación

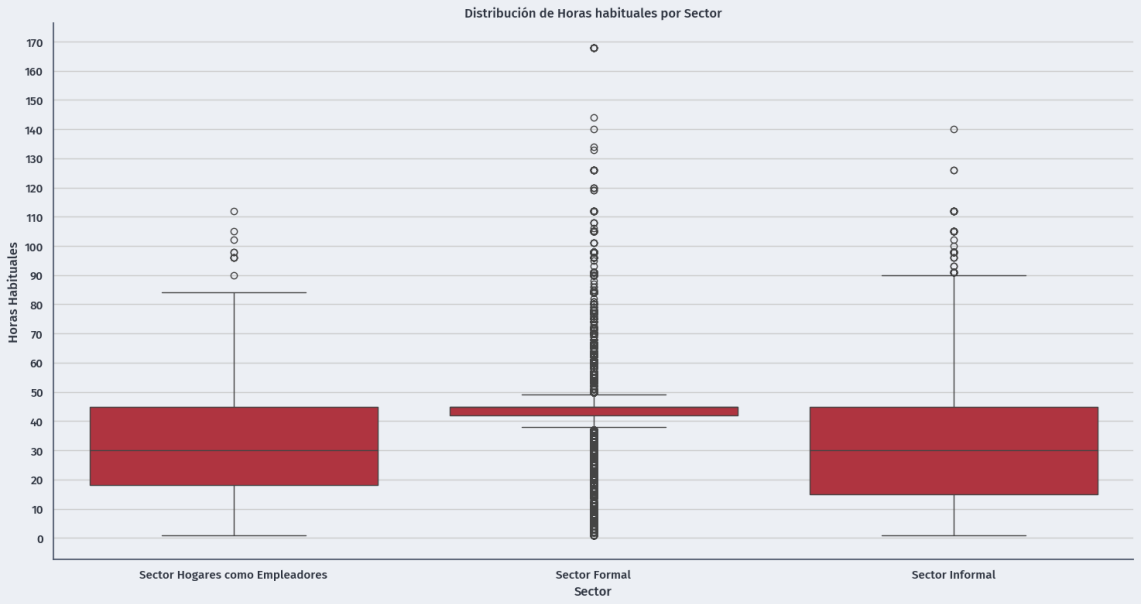


Tabla 5

Resultado prueba de Kruskal-Wallis de horas habituales y sector

Prueba	Estadístico Kruskal-Wallis	Valor p	Resultado
Kruskal-Wallis	16.786.204	0.0	El p-valor es menor que 0.05. Existe una diferencia significativa en 'habituales' entre los grupos de 'sector' (rechazamos H0)

La figura 11 sigue esta línea, pero esta vez con las horas habituales por sexo. Considero que este gráfico es de los más decidores de toda la presentación. Se pueden sacar muchas conjeturas, por ejemplo, parece afirmar que una de las causas de la brecha de remuneraciones entre hombres y mujeres tiene que ver directamente con la cantidad de horas trabajadas por sexo. Pero, otra pregunta quizá para otra investigación es, ¿Cuál es la razón de que las mujeres tiendan a trabajar menos horas que los hombres? Una hipótesis mencionada durante la presentación es que la mayoría de las veces los cuidados de personas mayores o niños recae en las mujeres, esto merma su posibilidad de incorporarse al mercado laboral, cosa que no sucede con los hombres tan a menudo. Este tipo de discusiones son las que busco hacer con este trabajo y considero que

este gráfico hace un buen trabajo para ello. Lo visto en el gráfico se confirma cuando vemos su prueba de Kruskal-Wallis, en la tabla 6.

Figura 11

Gráfico de caja y bigote de horas habituales por sexo

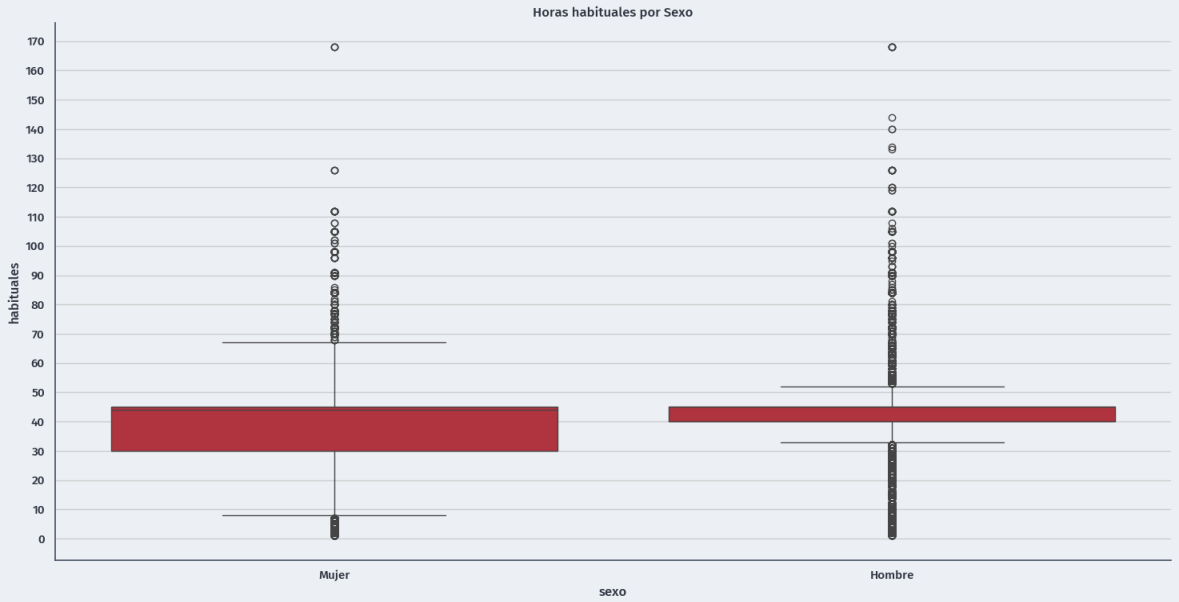


Tabla 6

Resultado de prueba de Kruskal-Wallis de horas habituales y sexo

Prueba	**Estadístico Kruskal-Wallis**	**Valor p**	**Resultado**
Kruskal-Wallis	4.027.745	0.0	El p-valor es menor que 0.05. Existe una diferencia significativa en 'habituales' entre los grupos de 'sexo' (rechazamos H0)

El último análisis de gráficos utilizados para este informe, en la sección de análisis univariado y bivariado, antes de la reducción de dimensionalidad que se aplicó, es el observado en la figura 12. En él se muestra la distribución de informales y formales para ambos sexos. Como se observa, no existen grandes diferencias a rasgos generales, aunque según la prueba estadística de chi-cuadrado contenida en la tabla 7, si parece haber una relación significativa entre el sexo y la ocupación de la persona.

Figura 12

Distribución de ocupación por sexo

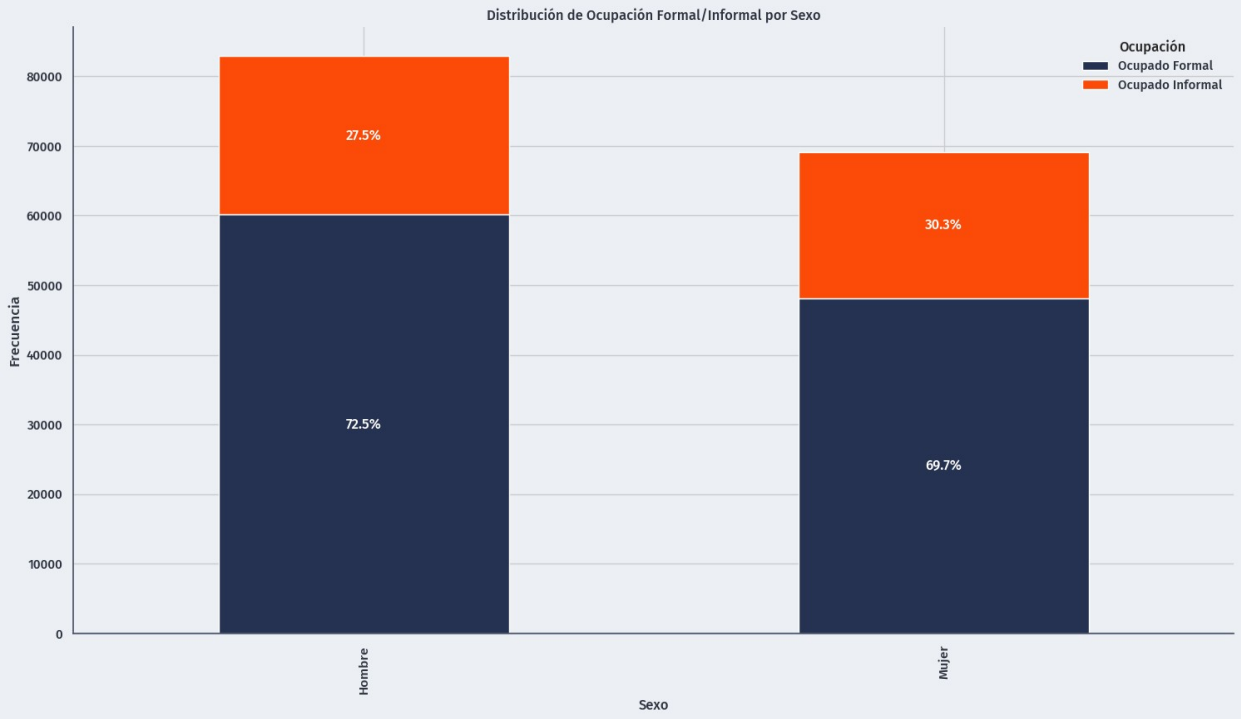


Tabla 7

Resultado de prueba estadística Chi-cuadrado para sexo y ocup_form

Prueba	Estadístico Chi-cuadrado	Grados de libertad	Valor p	Frecuencias esperadas	Resultado
Chi-cuadrado	150.956	1	1.07e-34	59095.38, 23870.62 49246.62, 19892.38	Existe una relación significativa entre 'sexo' y 'ocup_form' (rechazamos H0)

Para finalizar, se adjunta en las tablas 8 y 9 los resultados de la reducción de dimensionalidad MCA, esto se realizó con el fin de explorar las principales variables que aportan a la construcción de nuestro DataFrame, como se observa, para el componente 1 los principales factores son la actividad (trabajando o no) y la cantidad de horas, así como la ocupación y la rama de actividad económica referida a la actividad principal. En el componente 2 (tabla 9) se encuentran las respuestas “no sabe” o “no responde”, que por lo general en la encuesta están tabuladas con 99.

Tabla 8

Principales contribuidores del componente 1

Variable	**Contribución**
activ_1.0_False	0.043745
habituales_42.79421521194719	0.043745
activ_3.0_True	0.039948
a1_2.0_True	0.033494
a1_1.0_False	0.033494
ocup_form_2.0	0.031658
activ_1.0_True	0.031098
a1_2.0_False	0.030287
a1_1.0_True	0.030287

Tabla 9

Principales contribuidores del componente 2

Variable	**Contribución**
b7a_2_99.0	0.078450
b7a_1_99.0	0.073653
b7b_2_99.0	0.072758
b11_proxy_99.0	0.070103
b7b_3_99.0	0.064180
b7a_3_99.0	0.061727
b12_99.0	0.060259

plataformas_digitales_99.0	0.059261
b8_99.0	0.048594
i2_99.0	0.045580

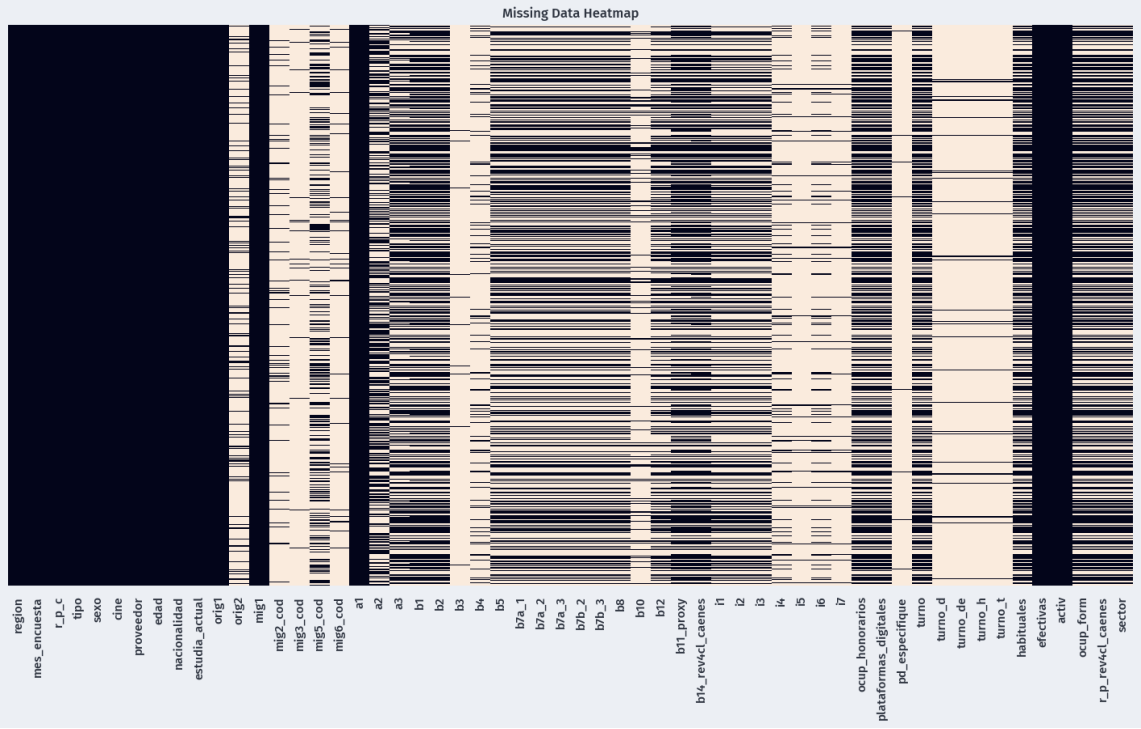
5. PREPARACIÓN DE LA BASE DE DATOS

Las principales variables que se utilizarán para el análisis posterior se pueden encontrar en el anexo C, y siguen los estudios realizados por la CEPAl, son principalmente características de la persona como nivel económico, ‘geográfico, nivel educativo, genero, edad, y otras que serán importantes.

Pero muchas de ellas tenían grandes valores nulos, por lo que se aplicó otor criterio respecto al porcentaje de valores nulos por columna, que se fijó en 60%. Con esto, la figura 13 contiene las variables que se usarán para el análisis y que cumplen con los criterios de valores nulos y pertenecer a variables de interés según el libro de códigos y el informe de la CEPAL.

Figura 13

Mapa de calor de valores nulos de las variables resultantes



Se puede observar que la mayoría de los valores nulos se encuentran en las respuestas a las preguntas, pero no en variables claves como edad, sexo, nacionalidad y sector de ocupación.
Por lo anterior se decidió seguir al libro de códigos, y gracias a la naturaleza estructural de estos valores nulos, se reemplazarán los NA por 99, que para la mayoría de las preguntas corresponde a no contesta.

Con esto, se solucionó el problema de los valores nulos sin inventar categorías ni imputar datos faltantes que artificialmente crean nuevas categorías o sobre estiman una por sobre otra, como se intentó realizar durante las pruebas que se hicieron durante el transcurso de este proyecto.

6. PROCEDIMIENTO DE DIVISIÓN DEL CONJUNTO DE DATOS

Para la realización de este informe, se usó una división de 70% para entrenamiento y 30% para prueba. Este es un estándar dentro del entrenamiento de modelos de machine learning, y este trabajo no será una excepción.

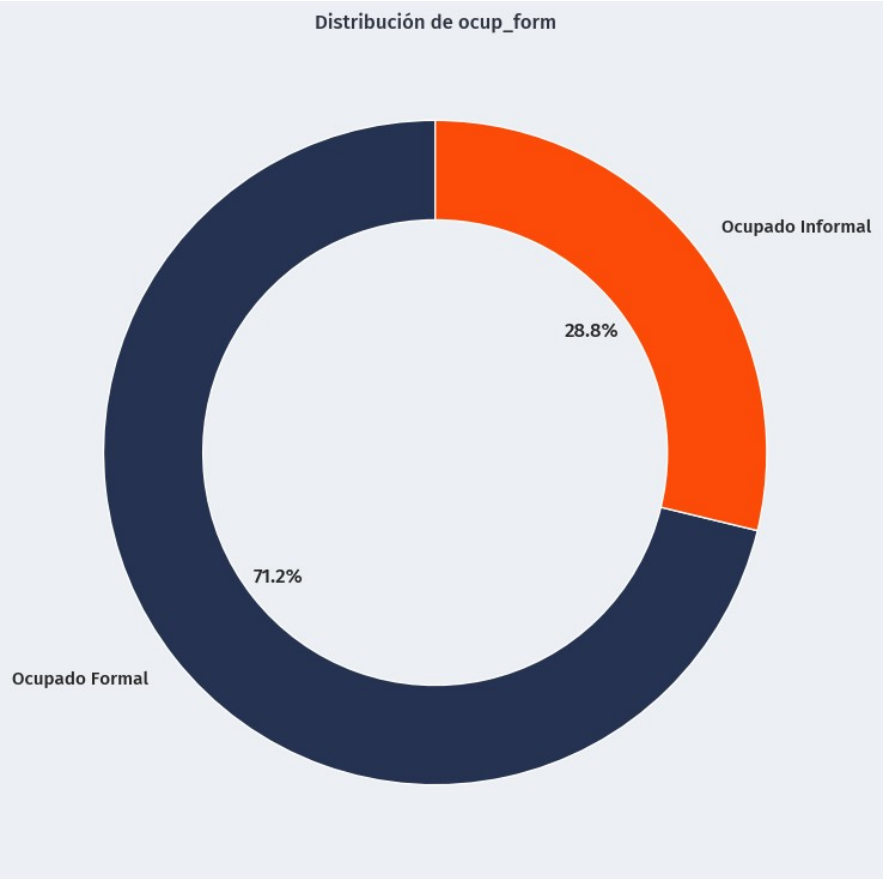
Del total de encuestas que se encuentran en el año 2023 (más de 300mil), en este trabajo se filtraron a las personas menores de 18 y mayores de 70 años, así como aquellas que no se encontraban con un empleo por distintos motivos, como pueden ser embarazo, jubilación anticipada, discapacidad, etc. Estas personas se filtraron gracias a la columna “activ” de la base de datos, en que la categoría 3 corresponde a aquellas personas fuera de la fuerza de trabajo.

Con lo anterior, se trabajó con 152.105 datos, que son personas únicas. De estos, en el conjunto de entrenamiento quedaron 106.473 personas y en el de prueba 45.632, que corresponden al 70 y 30% del total, respectivamente.

Es importante mencionar que nuestra variable objetivo será “ocup_form”, variable binaria que nos dice si la persona pertenece al sector formal o informal. Esta variable se encuentra desbalanceada, con una mayoría de las personas pertenecientes al sector formal. Esta diferencia se puede observar en la figura 13.

Figura 14

Distribución de la variable objetivo



Este factor es importante ya que se usó una estratificación para lidiar con este desbalanceo, también, se usará un modelo especializado en este tipo de datos desbalanceados.

Por lo mismo, usaremos como principal métrica de evaluación la precisión y el recall, aunque métodos más clásicos como la matriz de confusión y la curva ROC-AUC también serán utilizados, aunque se hará énfasis en los primeros.

7. VARIABLES DEL PROYECTO

Las variables seleccionadas para entrenar y probar los modelos de machine learning vistos en la presentación se seleccionaron en base a la cantidad de valores nulos y que tan relevantes son para identificar el sector en el que se desempeña la persona encuestada.

En cuanto a los valores nulos, se seleccionaron aquellas variables que tienen menos de 40% de valores nulos. Como se mencionó, esta base de datos contiene datos nulos del tipo “structural missingness”, esto es, valores nulos que existen debido a que el diseño de la encuesta así lo permite. Por ejemplo, existe una serie de preguntas que se refieren a la condición de migrante del encuestado, estas son “mig”, “mig1” o “mig_cod”. De estas, solamente es obligatorio para el encuestado responder la primera, y según la respuesta de la primera se responden o no las

siguientes. Este criterio nos sirvió para seleccionar aquellas preguntas más o menos relevantes para la encuesta y nuestro trabajo.

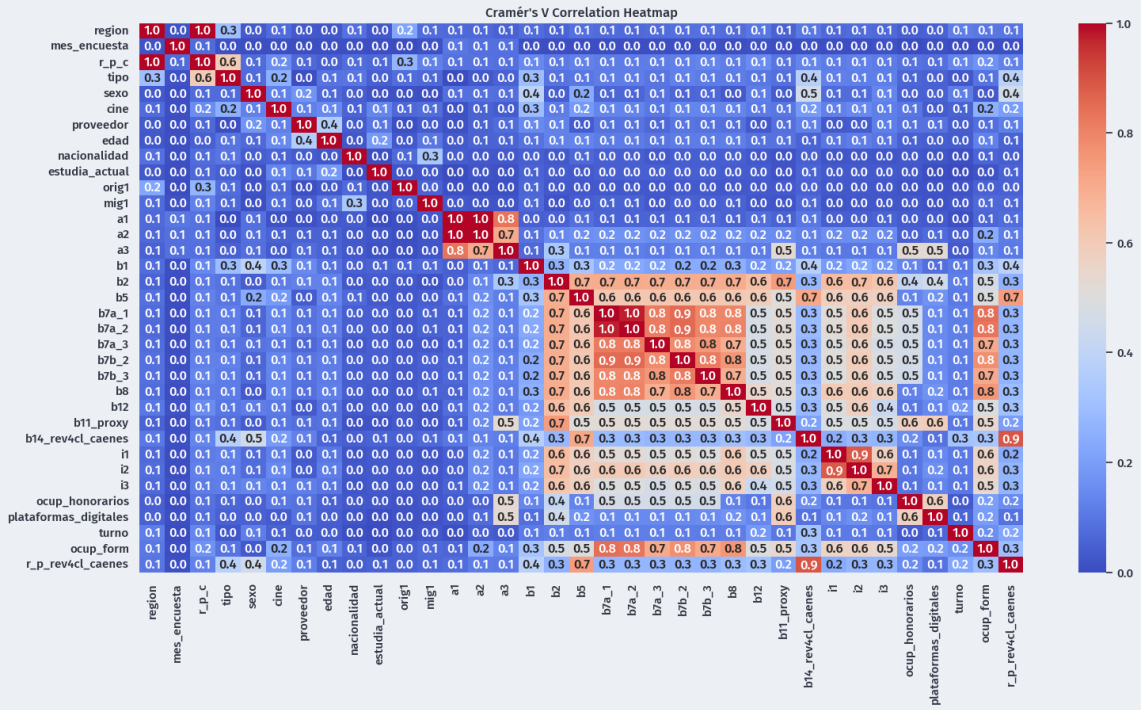
Por otro lado, existen variables que contienen más información para la informalidad laboral, que fueron seleccionadas en base a la literatura revisada en la primera entrega.

Las variables que se usarán para esta entrega se encuentran en el anexo 1, y la variable dependiente será “ocup_form”, binaria que determina si una persona pertenece a la ocupación formal e informal.

También, otro factor importante fue la siguiente tabla de correlaciones que se presenta en la figura 15, en donde podemos ver las relaciones que guardan principalmente las preguntas que son respuestas a preguntas como las iniciadas en con la letra “b” y sus derivados.

Figura 15

Mapa de correlación de variables categóricas con V de Cramer



8. MODELOS UTILIZADOS

Tabla 10

Detalles del modelo de Regresión Logística

Regresión Logística	
Desarrolladores y año de creación	Ronald Fisher (1936), base estadística del modelo actual
Propósito del modelo	Clasificación binaria; predice la probabilidad de pertenecer a una de dos categorías

Estructura del modelo	Basado en una función sigmoide para modelar probabilidades. Asume relaciones lineales entre variables independientes y el logit de la probabilidad
Hiperparámetros más importantes	C: Controla la regularización para evitar sobreajuste. penalty: Especifica el tipo de regularización (L1, L2)
Ventajas	Fácil de interpretar. Escalable para grandes volúmenes de datos. Rápido en entrenamiento
Desventajas	Asume independencia entre variables. Menos eficaz para problemas no lineales

Tabla 11

Detalles del modelo de árboles de decisión

Árbol de Decisión	
Desarrolladores y año de creación	J. Ross Quinlan (1986), introdujo el algoritmo ID3, base de los árboles de decisión modernos
Propósito del modelo	Clasificación y regresión; utiliza un enfoque jerárquico para tomar decisiones basadas en las características de entrada
Estructura del modelo	Árbol jerárquico donde cada nodo interno representa una característica y las hojas representan las clases
Hiperparámetros más importantes	max_depth: Profundidad máxima del árbol. min_samples_split: Número mínimo de muestras para dividir un nodo. criterion: Métrica para medir la calidad del split (e.g., gini o entropy)
Ventajas	Intuitivo y fácil de interpretar. Captura interacciones no lineales entre variables
Desventajas	Propenso al sobreajuste. Menor robustez comparado con métodos ensemble

Tabla 12

Detalle del modelo de bosques aleatorios

Bosques Aleatorios (Random Forest)	
Desarrolladores y año de creación	Leo Breiman (2001), introdujo Random Forest como un método ensemble basado en árboles de decisión
Propósito del modelo	Clasificación y regresión; utiliza un conjunto de árboles de decisión para mejorar precisión y robustez
Estructura del modelo	Ensamble de múltiples árboles de decisión entrenados en subconjuntos aleatorios de datos y características
Hiperparámetros más importantes	n_estimators: Número de árboles en el ensamble. max_depth: Profundidad máxima de cada árbol. max_features: Número de características consideradas por árbol
Ventajas	Reduce el riesgo de sobreajuste comparado con árboles individuales. Robusto ante datos ruidosos o faltantes
Desventajas	Mayor consumo computacional. Menos interpretable que un solo árbol

Tabla 13

Detalle del modelo XGBoost

eXtreme Gradient Boosting (XGBoost)	
Desarrolladores y año de creación	Tianqi Chen (2014), introdujo XGBoost como una versión optimizada del boosting basado en gradientes
Propósito del modelo	Clasificación y regresión; implementa boosting basado en gradientes para optimizar el desempeño
Estructura del modelo	Ensamble de árboles de decisión construidos secuencialmente, donde cada árbol corrige errores del anterior
Hiperparámetros más importantes	n_estimators: Número de árboles.

	learning_rate: Tasa de aprendizaje. max_depth: Profundidad máxima de los árboles. subsample: Fracción de datos utilizada por árbol
Ventajas	Alto desempeño en competiciones y tareas complejas. Maneja valores faltantes de forma eficiente
Desventajas	Configuración más compleja debido a sus numerosos hiperparámetros. Requiere más tiempo de entrenamiento

Tabla 14

Detalle del modelo Balanced Random Forest

Balanced Random Forest Classifier	
Desarrolladores y año de creación	Basado en Random Forest, fue adaptado por Chen y Breiman (2004) para manejar datos desbalanceados
Propósito del modelo	Clasificación; ajustado para problemas con clases desbalanceadas, como en datasets con pocas muestras de una categoría
Estructura del modelo	Modificación de Random Forest que utiliza muestreo balanceado para entrenar árboles en subconjuntos equilibrados
Hiperparámetros más importantes	n_estimators: Número de árboles. max_depth: Profundidad de los árboles. sampling_strategy: Estrategia para balancear las clases
Ventajas	Reduce el sesgo hacia clases mayoritarias. Ideal para problemas de desbalance extremo
Desventajas	Mayor tiempo de entrenamiento comparado con Random Forest. Limitado a la calidad del muestreo utilizado

Como se planteó un problema relativamente sencillo, como lo es los factores que explican que una persona dentro de la base de datos corresponda a la clase 0 o 1, los modelos fueron seleccionados gracias a su simpleza y gran capacidad para clasificar e identificar las determinantes para lograrlo.

Se partió desde un modelo simple y conocido como lo es la regresión logística, hasta modelos como los árboles de decisiones y bosques aleatorios, y también un modelo más complejo pero que suele tener mejor rendimiento como lo es XGBoost.
Esta lógica de pensamiento fue crucial para el resto del trabajo.

9. MÉTRICAS DE EVALUACIÓN

Las métricas de evaluación serán accuracy, precisión, la curva ROC, recall y el F1 score. Las definiciones y fórmulas de cálculo se pueden observar en la tabla 15.

Tabla 15

Detalle de las métricas de evaluación

Métrica	Descripción	Fórmula
Accuracy	Mide la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones. Es útil cuando las clases están balanceadas.	Accuracy = Predicciones correctas / Total de predicciones
ROC-AUC (Área bajo la curva ROC)	Representa la capacidad del modelo para distinguir entre clases. Un valor más cercano a 1 indica un mejor rendimiento. Es especialmente útil para problemas con clases desbalanceadas.	N/A
Precision	Evalúa la proporción de verdaderos positivos sobre todas las instancias predichas como positivas. Es importante cuando los falsos positivos son costosos.	Precision = Verdaderos positivos / (Verdaderos positivos + Falsos positivos)
Recall	También conocido como sensibilidad, mide la proporción de verdaderos positivos que fueron correctamente identificados sobre el total de positivos reales. Es crucial cuando los falsos negativos son más costosos.	Recall = Verdaderos positivos / (Verdaderos positivos + Falsos negativos)
F1-Score	Es la media armónica de Precision y Recall, proporcionando un balance entre ambos. Es útil cuando existe un equilibrio entre falsos positivos y falsos negativos.	F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

10. OPTIMIZACIÓN DE HIPERPARÁMETROS

Balanced Random Forest

Se realizó la optimización de hiperparámetros para el modelo Balanced Random Forest Classifier utilizando optimización bayesiana con la librería Optuna. Los hiperparámetros seleccionados y sus justificaciones fueron:

- n_estimators: Número de árboles en el bosque (rango: 10-100).

- max_depth: Profundidad máxima de los árboles (rango: 3-15).

- min_samples_split: Número mínimo de muestras para dividir un nodo (rango: 5-20).

- min_samples_leaf: Muestras mínimas requeridas en un nodo hoja (rango: 2-10).

- max_features: Fracción de características por división ('sqrt' o 'log2').

La función objetivo optimizó un promedio ponderado de F1-Score y Recall para balancear precisión y sensibilidad. El proceso incluyó 50 iteraciones o 30 minutos como límite. Los mejores hiperparámetros encontrados fueron:

- n_estimators: 87

- max_depth: 12

- min_samples_split: 10

- min_samples_leaf: 4

- max_features: sqrt

El puntaje combinado fue 0.885, destacando el equilibrio entre precisión y sensibilidad.

XGBoost

Se optimizaron los hiperparámetros del modelo XGBClassifier utilizando la misma estrategia de optimización bayesiana. Los hiperparámetros seleccionados y sus justificaciones fueron:

- n_estimators: Número de árboles (rango: 50-100).

- max_depth: Profundidad máxima de los árboles (rango: 3-15).

- learning_rate: Tasa de aprendizaje (rango: 0.01-0.3).

- colsample_bytree: Fracción de columnas utilizadas por árbol (rango: 0.5-1.0).

- subsample: Fracción de datos utilizados por árbol (rango: 0.5-1.0).

- min_child_weight: Peso mínimo requerido para dividir un nodo (rango: 1-10).

- gamma: Reducción mínima de la pérdida para dividir un nodo (rango: 0-5).

- scale_pos_weight: Peso de las clases positivas para datos desbalanceados (rango: 0.5-10.0).

La función objetivo también combinó F1-Score y Recall. Los mejores hiperparámetros fueron:

- n_estimators: 90

- max_depth: 10
- learning_rate: 0.1
- colsample_bytree: 0.8
- subsample: 0.9
- min_child_weight: 4
- gamma: 1.5
- scale_pos_weight: 3.0

El puntaje combinado fue 0.915, mostrando un excelente balance entre precisión y sensibilidad, particularmente en el manejo de clases desbalanceadas.

La optimización bayesiana permitió encontrar configuraciones de hiperparámetros robustas para ambos modelos. El BalancedRandomForestClassifier logró un equilibrio sólido entre precisión y sensibilidad, mientras que el XGBClassifier demostró un rendimiento sobresaliente en datos desbalanceados. Estos resultados validan el uso de Optuna como una herramienta eficiente para la optimización de hiperparámetros en problemas complejos.

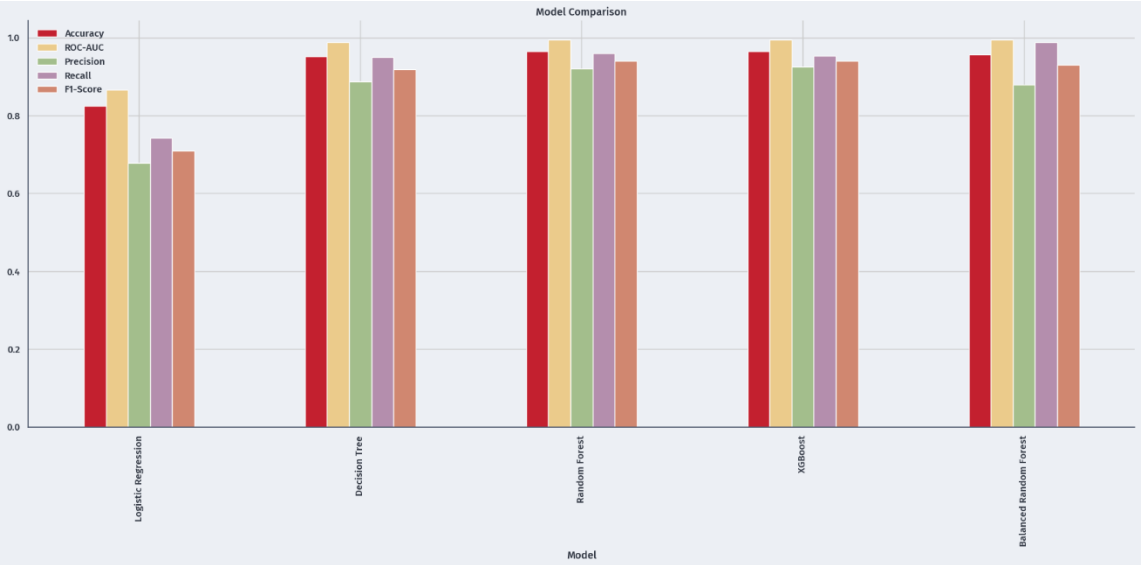
11. ANÁLISIS DE RESULTADOS

En primer lugar, se probaron los modelos presentados en las páginas anteriores, siguiendo la división y estratificación presentada.

Estos modelos dieron muy buenos resultados cuando hablamos de las métricas generales, así como las que más usaremos (F1, recall y precisión). Los resultados se pueden consultar en la figura siguiente:

Figura 16

Principales métricas de evaluación de los modelos



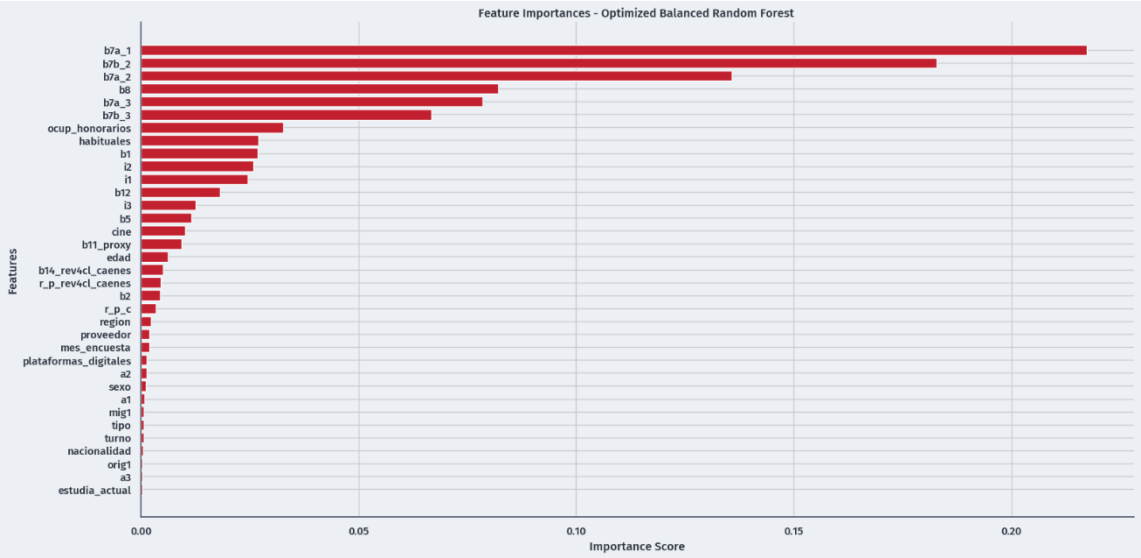
Los modelos más complejos tienden a tener un mejor comportamiento con este problema de clasificación, con un promedio mayor a 80% en todos los indicadores, destacando el caso de XGBoost y Balanced Random Forest, modelo que lidia por defecto con los problemas de clasificación desbalanceados como el nuestro. Estos dos modelos nos servirán de base para el resto del trabajo y serán aquellos que optimizaremos con los parámetros que vimos en la sección anterior.

Una vez realizada la optimización de hiperparámetros, el modelo Balanced Random Forest optimizado obtuvo un F1-Score de 0,93 y recall de 0,99, con los parámetros que se presentaron en la sección anterior.

Al analizar la importancia de cada variable para construcción del modelo, como se indica en el gráfico 6, las mayores variables son la pregunta b7 y sus variantes, así como la b8.

Figura 17

Feature importance del modelo BRFC



La siguiente tabla sintetiza a qué pregunta corresponde cada variable

En la etapa de modelamiento se pueden realizar distintas iteraciones de los modelos utilizados, una de ellas puede estar relacionada con evaluar el desempeño de distintos tipos de modelos, comenzando por los simples hasta llegar a técnicas complejas. Mientras que otras iteraciones de la etapa de modelamiento, pueden estar relacionadas con la evolución del modelo al probar una configuración de hiperparámetros no optimizada con otras optimizadas. Entonces, en esta sección se debe analizar cuantitativa y cualitativamente los resultados obtenidos en las distintas iteraciones de la etapa de modelamiento, para ello establezca algún modelo línea base (base de comparación) que permita comparar críticamente los resultados obtenidos.

Tabla 16

Variables triviales del análisis

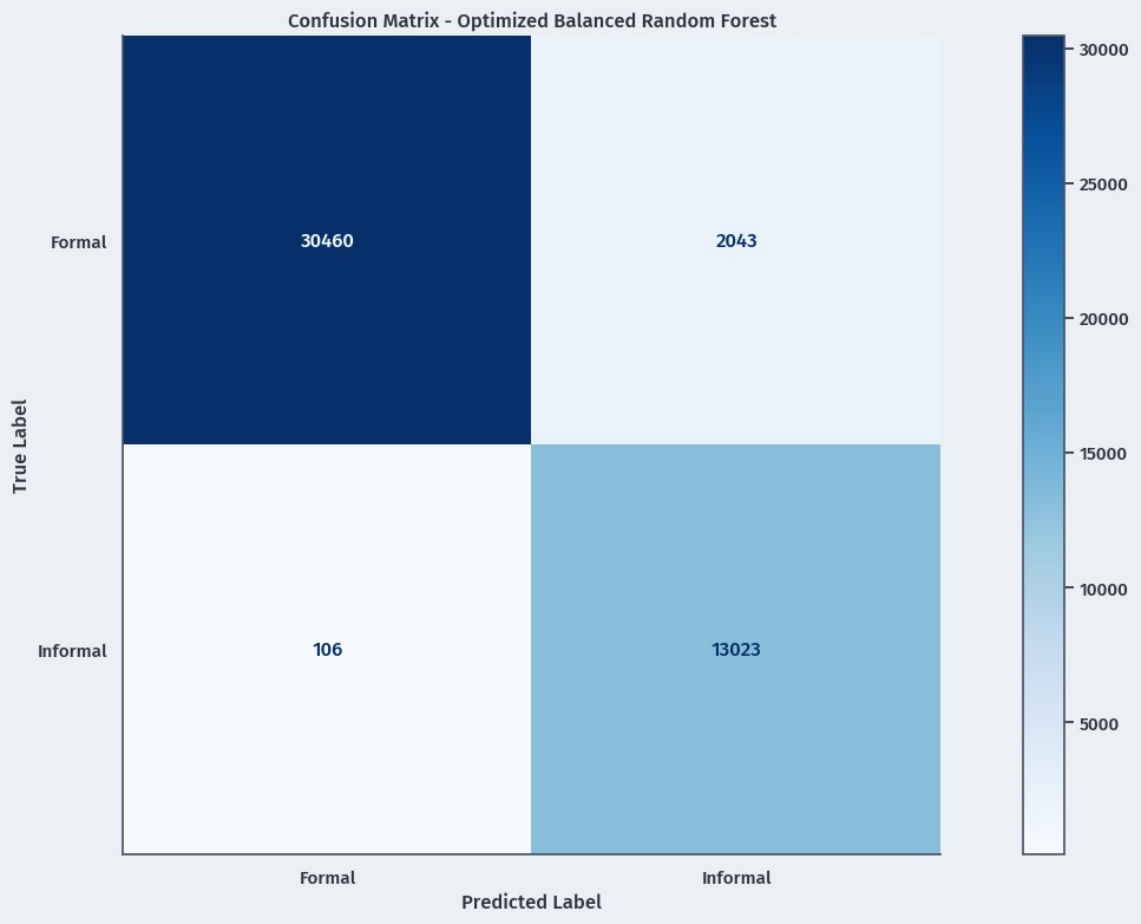
Variable	Descripción	Categorías Observadas
b7a_1	Su empleador, ¿cotiza por usted en el sistema previsional o de pensión?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde
b7a_2	Su empleador, ¿cotiza por usted en el sistema de salud (público o privado)?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde
b7a_3	Su empleador, ¿cotiza por usted en el sistema de seguro de desempleo?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde

b8	En ese empleo, ¿tiene contrato escrito?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde
b7b_2	En este trabajo, ¿tiene derecho, aunque no utilice, a días pagados por enfermedad?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde
b7b_3	En este trabajo, ¿tiene derecho, aunque no lo utilice, a permiso por maternidad o paternidad?	1 sí 2 No 77 No aplica 88 No sabe 99 No responde

Por otro lado, la figura 16 muestra la matriz de confusión del modelo anteriormente descrito:

Figura 18

Matriz de confusión del modelo BRFC



Esto nos demuestra que la optimización de hiperparametros mejoró un modelo que ya era bastante bueno, y una manera de comprobar que el modelo está funcionando de manera correcta

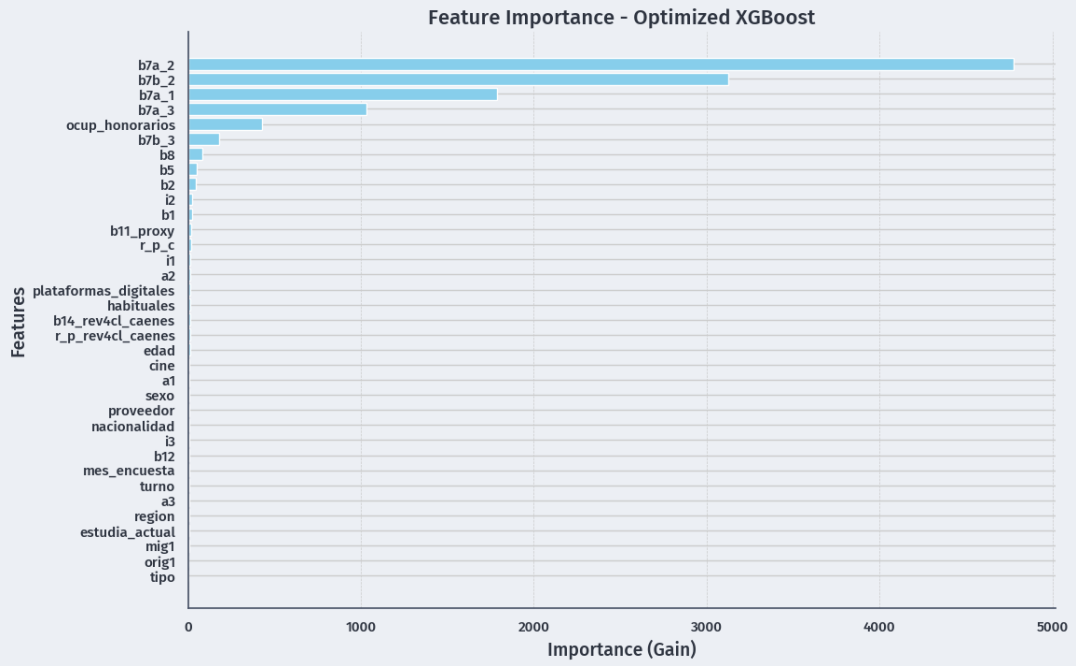
es el hecho de que las principales variables usadas para crear el modelo son preguntas que tienen directa relación con condiciones laborales de las personas encuestadas, como lo son la existencia de contratos, tipo de pago, cotizaciones en seguridad social, entre otras.

Con el otro modelo, XGBoost, se realizó lo mismo, y los resultados son bastante similares. Este modelo obtuvo un 0,94 en F1 Score y 0,98 en Recall.

Las principales variables para la creación del XGBoost son las siguientes:

Figura 19

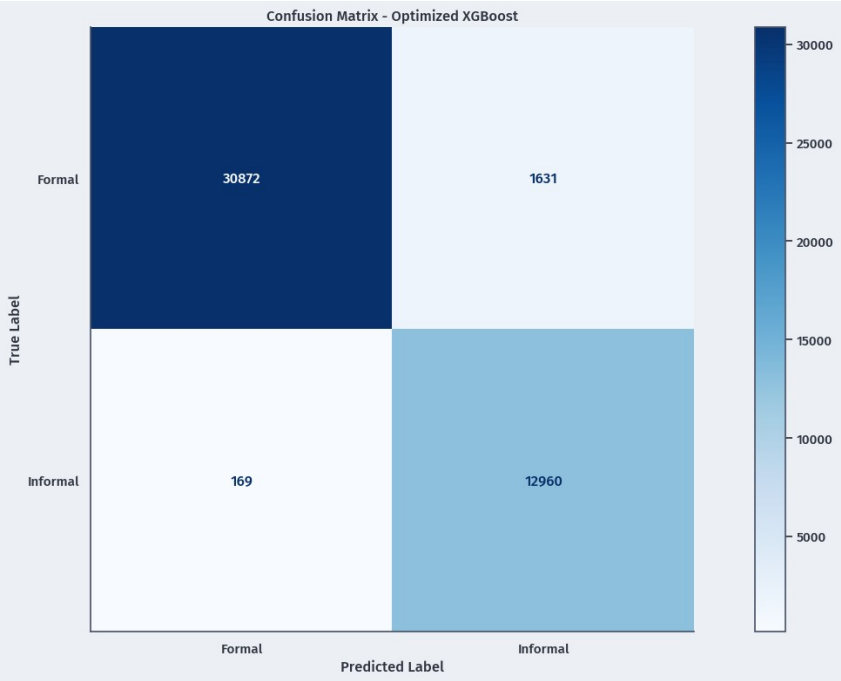
Feature importance del modelo XGBoost optimizado



Y la matriz de correlación es la siguiente:

Figura 20

Matriz de confusión del modelo XGBoost optimizado



En conclusión, se puede observar que la optimización de hiperparametros sirvió para mejorar los resultados obtenidos en comparación a los parámetros por defecto, y que se seleccionaron aquellos modelos que generan mejor rendimiento.

12. CAPACIDAD DE GENERALIZACIÓN DE RESULTADOS

Continuando con el análisis, considero que el modelo obtuvo resultados tan buenos debido a que algunas de las características elegidas son bastante triviales. En este problema, es obvio que preguntas que contengan información acerca de la existencia de contrato, tipo de pago, cotizaciones en seguridad social, tipo de contrato, entre otras, son importantes para la discriminación entre la clase formal e informal. Esto es positivo porque nos confirma que el modelo está haciendo su trabajo, pero a la vez, no deja obtener información relevante en las otras columnas que no pueden parecer tan obvias.

Es por lo anterior que se realizó una nueva prueba de estos dos clasificadores, con los mismos parámetros optimizados, pero esta vez sin estas variables que son más bien triviales, los resultados para ambos se muestran en las figuras 21 y 22.

Figura 21

Feature importance del modelo XGBoost optimizado, segunda iteración

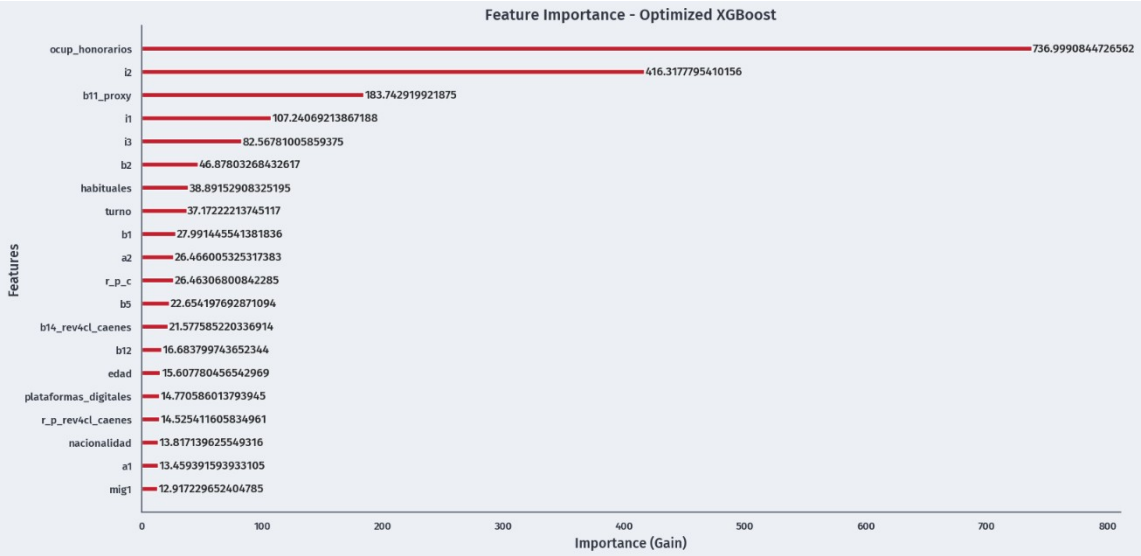
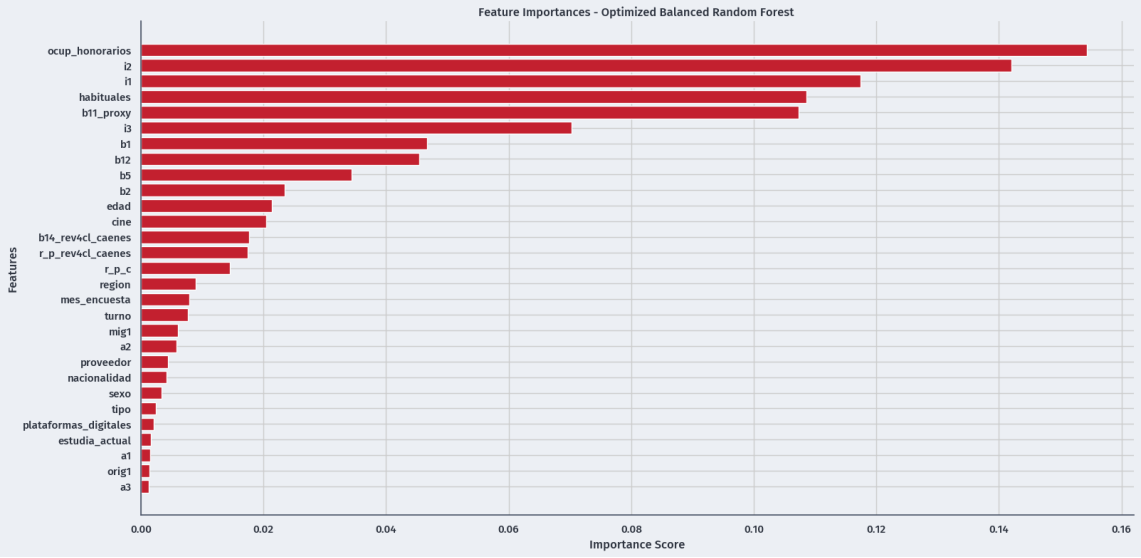


Figura 22

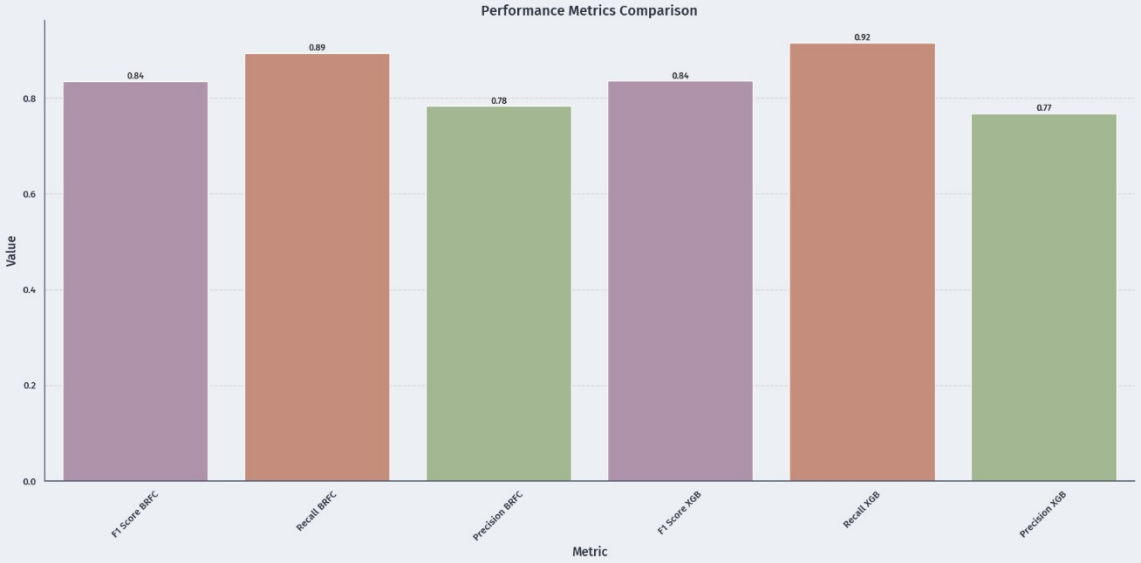
Feature importance del modelo BRFC optimizado, segunda iteración



Y los resultados de F1 Score, recall y precisión:

Figura 23

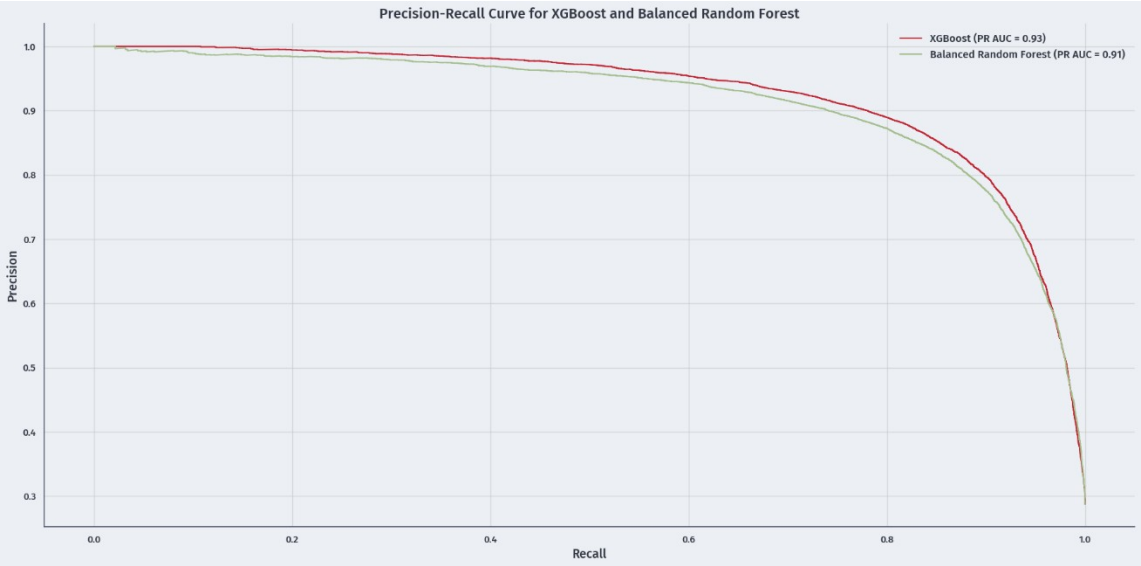
Comparación de los modelos optimizados, segunda iteración



Estos resultados nos permiten ahondar más en el problema, obteniendo la importancia real de las otras variables, con un modelo que sigue teniendo resultados bastante buenos.

Figura 24

Curva precision-recall de los modelos optimizados, segunda iteración



13. EVALUACIÓN DE COSTOS COMPUTACIONALES

Los tiempos de entrenamiento de los modelos son:

a) 1,8 segundos para Logistic Regression
b) 0,4 segundos para Decision Tree Classifier

c) 10,8 segundos para Random Forest
d) 1,5 segundos para XGBoost
e) 8,4 segundos para Balanced Random Forest

Esto con los parámetros predeterminados de cada modelo.

El uso de memoria no encontré manera de medirlo modelo por modelo, pero al usar Visual Studio Code, el proyecto en total usaba entre 1 y 1,5 gb de RAM.

La optimización de los modelos XGBoost y Balanced Random Forest tomó 3 minutos y 25 segundos para el primero y 1 minuto para el segundo.

En este proyecto no fueron utilizados ningún tipo de optimización de recursos adicional, como el uso de GPU a través de torch. Todos los cálculos fueron realizados en mi computadora local, que contiene las siguientes especificaciones de hardware:

CPU: AMD Ryzen 5 2600X, 3,6 GHz velocidad base

RAM: 16GB DDR4, velocidad de 3200 MHz

GPU: AMD Radeon RX 5600 XT, con 6GB de VRAM GDDR6

SO: Windows 11 actualizado

14. GENERACIÓN DE RECOMENDACIONES DEL PROYECTO DE CIENCIA DE DATOS

Considero que el proyecto dio soluciones satisfactorias al problema propuesto, esto es, factores que influyen en que una persona opte o se vea obligada a desempeñarse en el sector informal o formal, todo esto en base a las respuestas a las preguntas realizadas por la ENE. Entre los factores relacionados a respuestas a preguntas, las principales se muestran en la tabla siguiente

Tabla 17

Variables importantes para determinar el sector de ocupación de la persona encuestada

Variable	Descripción	Categorías Observadas
ocup_honorarios	Por el trabajo realizado, ¿entregó una boleta de honorarios?	1 sí, 2 No, 3 No sabe / No responde, 77 No aplica
i2	La empresa, negocio o institución que le paga su sueldo, ¿cuenta con los servicios de un contador o tiene oficina de contabilidad?	1 sí, 2 No, 88 No sabe, 99 No responde
i1	La empresa, negocio o institución que le paga su sueldo, ¿está registrada en el Servicio de Impuestos	1 sí, 2 No sabe, pero la empresa entrega boleta o factura, 3 No, 88 No sabe, 99 No responde

	Internos (SII) o tiene iniciación de actividades?	
habituales	Actividad principal: Total horas semanales trabajadas habitualmente	Valores numéricos
b11_proxy	¿El pago por su actividad principal fue a través de sueldo, salario o jornal?	1 sí, 2 No, 88 No sabe, 99 No responde
i3	¿Cuál es el nombre de la empresa, negocio o institución que le paga su sueldo?	1 informante identifica nombre, 2 El negocio no tiene nombre, 3 Trabaja para un hogar particular, 4 Otro asalariado, 88 No sabe, 99 No responde
b1	Grupo de ocupación según CIUO-08	Ver detalle en Anexo 3
b12	Está contratado o tiene un acuerdo de trabajo...	1 directamente con la empresa, 2 Con un contratista, 3 Con una empresa de servicios temporales, 4 Con un enganchador, 88 No sabe, 99 No responde
b5	El negocio, empresa o institución donde trabajó la semana pasada era...	1 estatal, 2 Privado, 3 Hogar particular

Como se mencionó en las secciones anteriores, estas variables fueron obtenidas durante la segunda iteración y son aquellas que no contienen información tan obvia como las de la tabla 16 del presente, y permiten de manera más compleja captar las relaciones entre las respuestas a la pregunta y el sector de ocupación de la persona.

Considero que estas variables deberían ser observadas con detenimiento si se quieren usar como insumo para generar políticas públicas o programas para reducir la informalidad y sus consecuencias entre la población.

Me detendré en la variable “ocup_honorarios”, ya que esta práctica de contratación está bastante arraigada en instituciones públicas como nuestra universidad y en mi experiencia como practicante del Ministerio del Trabajo también sucedía que muchas personas recibían sueldos acordes al mercado, pero no gozaban de un contrato. Esto muchas veces genera que las

personas tengan que cotizar por su cuenta, que no puedan acreditar antigüedad laboral, y que estén a merced de la renovación del contrato año a año. Siento que es una pregunta bastante potente para generar una idea de que tipo de trabajo estamos creando en nuestro país.

15. CONCLUSIONES DEL PROYECTO

A modo de conclusiones generales, respecto al análisis exploratorio, tenemos una base de datos bastante representativa de la realidad del país, con una proporción casi igualitaria de hombres y mujeres, con énfasis en las grandes ciudades, que es donde se concentra la mayor fuerza de producción del país, y con personas de todo estrato y modo de trabajo. Las pruebas estadísticas nos permitieron concluir y confirmar sospechas respecto a las variables categóricas como sexo y horas trabajadas, sexo y sector económico. En este apartado, vimos que existen relaciones significativas entre el sexo y el sector de ocupación de las personas, diferencias entre horas trabajadas y también diferencias entre las horas trabajadas por sector económico.

También descubrimos la gran cantidad de preguntas con valores nulos que obtuvimos, que, si bien fueron manejados para lograr una respuesta satisfactoria al proyecto, contienen información que puede ser valiosa si no existiera en un primer lugar.

Con respecto a la selección de variables en base a los criterios de valores nulos y relevancia para el problema, considero que fue un enfoque que permitió optimizar el tiempo de trabajo y la precisión del proyecto. El hecho de tener pocos valores nulos hizo que su imputación no generara mayores sesgos, aunque no es un enfoque que me hubiera gustado utilizar en un primer lugar.

Ahora, los modelos que plantee fueron bastante acertados, esto considerando la relativa simpleza del problema planteado. Todos obtuvieron buenos resultados con parámetros predeterminados, y su optimización solamente hizo que fueran mejores.

Las métricas de evaluación, enfocadas en F1 y recall, permitieron obtener mejores resultados considerando la naturaleza de la variable objetivo; al estar desbalanceada era necesario tener un enfoque distinto a otros problemas de clasificación. La optimización de parámetros de XGBoost y de Balanced Random Forest sirvió de sobre manera y permitieron obtener información no tirival acerca de la base de dato y las principales variables a considerar para resolver el problema.

Con todo, considero que el proyecto sirvió para intentar dar una respuesta a la definición del problema que se planteó. Se obtuvo una lista de preguntas que aportan al problema, lo que luego puede ser usado para un estudio más a profundidad del mercado laboral y cómo el Estado puede responder frente a las demandas de mayor formalidad de organismos internacionales, o simplemente para entender el fenómeno.

Los modelos de clasificación funcionaron, dieron buenos resultados y se trabajaron exhaustivamente para intentar evitar el sobreajuste o el data leakage, se probaron y compararon varios y se optimizaron los que funcionarían de mejor manera, se quitaron variables que podrían

provocar leakage, y con esto se puede asegurar que los resultados son genuinos y no overfitting de los modelos.

También creo importante el manejo que se le hizo a las preguntas con valores nulos, ya que son una manera de recolección de datos bastante popular en las ciencias sociales y de distintas entidades estatales, y que son bastante importantes para intentar recolectar información compleja y sistematizarla para poder generar conocimiento.

Evalúo de manera positiva los resultados del proyecto y los conocimientos que obtuve al realizarlo, tanto desde los conocimientos adquiridos en ciencia de datos como del mercado laboral, ya que también me sirvió para mi estadía durante mi práctica laboral en el Ministerio del Trabajo.

16. PROPUESTAS CONCRETAS PARA MEJORAS FUTURAS Y EXTRACCIÓN DE CONOCIMIENTO

Las extensiones que he pensado a este proyecto son las siguientes:

a) Se podría ahondar más en redes neuronales para obtener relaciones más complejas dentro de las preguntas, cosa que escapa a mis conocimientos actuales,
b) Estudiar el problema con series de tiempo, existen datos desde enero de 2010 hasta la fecha, trimestralmente, de la encuesta. Esto es a su vez complicado ya que muchas preguntas quedaron obsoletas, cambiadas o derechamente removidas. Con un estudio de series de tiempo se podría obtener factores externos que la encuesta no identifica que podrían afectar a la condición de formalidad o informalidad del trabajador, como lo podría ser la pandemia,
c) Probar otras maneras de imputar los datos nulos, ya que el reemplazo por 99 puede generar cierto *bias* en los datos, aunque esto también implica que la encuesta debería tener otra manera de presentar los datos,
d) A su vez, probaría modelos más simples, ya que todos los aquí cubiertos son buenos, pero complejo, el hecho de tener modelos más simples nos permitiría dar una respuesta más rápida y obtener conocimiento que quizá los modelos más complicados no captan de la misma manera, y finalmente,
e) Respecto a la encuesta en sí, propondría que existiera una representación más grande de sectores rurales, menos preguntas, ya que muchas se podrían sintetizar en una sola.

17. APORTE AL CONOCIMIENTO DEL FENÓMENO

Este proyecto ha servido para la mejora en la detección de posibles factores de riesgo que hacen que una persona tenga que optar por el sector formal o informal de empleo. Todo esto ha sido posible gracias a la utilización de modelos de clasificación basados en arboles de decisiones como XGBoost y Random Forest.

Se han encontrado variables como la existencia de contrato, tipo de pago, cotización del empleador en el sistema público o privado, horas de trabajo, sector de trabajo, cantidad de horas, entre otras, que sirven como un indicador de en qué sector de ocupación se desempeña la persona.

Estos conocimientos son claves para entender el mercado laboral chileno, y si se usan correctamente, pueden servir de base para la generación de políticas públicas focalizadas en esas áreas en específico para disminuir la informalidad laboral y generar mejores condiciones para los trabajadores del país.

REFERENCIAS

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Céspedes, N. (2020). Informalidad, productividad y flexibilidad laboral. *Revista de Análisis Económico y Financiero*. Recuperado de https://contabilidadyeconomiausmp.edu.pe/OJS2020/index.php/RAEF/article/view/27

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley, Technical Report*, 1–12.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Instituto Nacional de Estadísticas de Chile (INE). (2022). *Documento Metodológico: Encuesta Nacional de Empleo (ENE) (Versión 2.0)*. Instituto Nacional de Estadísticas. Recuperado de https://www.ine.cl

Instituto Nacional de Estadísticas de Chile (INE). (2024). *Libro de Códigos Base de Datos: Encuesta Nacional de Empleo (ENE)*. Instituto Nacional de Estadísticas. Recuperado de https://www.ine.cl

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*(5), 402–406. Recuperado de https://synapse.koreamed.org/articles/1155616

Kelmanson, M. B., Kirabaeva, K., Medina, L., Mircheva, M. B., & Weiss, J. (2019). Explaining the shadow economy in Europe: Size, causes, and policy options. *Fondo Monetario Internacional*. Recuperado de https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3523151

Khangar, N. V., & Kamalja, K. K. (2017). Multiple Correspondence Analysis and its applications. *Electronic Journal of Applied Statistical Analysis, 10*(2), 432–462. https://doi.org/10.1285/i20705948v10n2p432

Livert-Aquino, F., Miranda, F., & Espejo, A. (2022). Estimación de la probabilidad de informalidad laboral a nivel comunal en Chile. *CEPAL*. Recuperado de https://repositorio.cepal.org/handle/11362/47727

OIT. (2008). 18.a Conferencia Internacional de Estadísticos del Trabajo: Resolución I: Resolución sobre la medición del tiempo de trabajo. *Organización Internacional del Trabajo*. Recuperado de https://www.ilo.org/es/media/270671/download

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

Rosatti, G. (2021). Métodos de Machine Learning como alternativa para la imputación de datos perdidos. Un ejercicio en base a la Encuesta Permanente de Hogares. *Estudios del Trabajo, 61*, 122–135. Recuperado de http://ojs.aset.org.ar/revista/article/view/81

ANEXO A: Tabla de valores nulos

Variables	Valores Nulos
ano_trimestre	0
mes_central	0
ano_encuesta	0
mes_encuesta	0
region	0
provincia	0
r_p_c	0
estrato	0
tipo	0
conglomerado	0
id_identificacion	0
hogar	0
idrph	0
nro_linea	0
edad	0
tramo_edad	0
sexo	0
parentesco	0
curso	0
nivel	0
termino_nivel	0
estudia_actual	0
cine	0
est_conyugal	0
proveedor	0
orig1	0
orig2	225038
orig3	260804

nacionalidad	0
mig1	0
mig2_cod	236720
mig3_cod	254966
mig4	0
mig5_cod	163385
mig6_cod	246060
a1	0
a2	137049
a3	119398
a4	259691
a5	141580
a6	248634
a6_otro	260842
a7	258162
a8	258407
b1	108440
b2	108440
b3	257296
b4	225129
b5	146258
b6	256138
b7a_1	146258
b7a_2	146258
b7a_3	146258
b7b_1	146258
b7b_2	146258
b7b_3	146258
b7b_4	146258
b8	146258

b9	146258
b10	228505
b11_proxy	109714
b12	150451
b13_rev4cl_caenes	241207
i1	150451
i2	150451
i3	150451
i3_v	150451
i4	223160
i5	247141
i6	223160
i7	246912
b14_rev4cl_caenes	108440
b15_1	112633
b15_2	204517
b16	112633
b16_otro	260459
b17_mes	108440
b17_ano	108440
b18_codigo	108440
b18_region	108440
b18_varias	108440
ocup_honorarios	108440
plataformas_digitales	108440
pd_especifique	257234
b19	108440
dependencia_segunda	255940
sda_pd	255940
sda_pd_especifique	260491

turno	108440
turno_d	248511
turno_de	248511
turno_h	248511
turno_t	248511
c2_1_1	108440
c2_1_2	108440
c2_1_3	108440
c2_2_1	255940
c2_2_2	255940
c2_2_3	255940
c3_1	146258
c3_2	146258
c3_3	146258
turno_cont_d	248599
turno_cont_de	248599
turno_cont_h	248599
turno_cont_t	248599
c4	259745
c5	108440
c6	256764
c7	112654
c8	237531
c9	233317
c9_otro	260559
adicionales_h	241542
adicionales_d	241542
adicionales_t	241542
c10	108440
c11	240250

c12	240250
habituales	108440
efectivas	108440
e2	0
e3_1	236284
e3_2	236284
e3_3	236284
e3_4	236284
e3_5	236284
e3_6	236284
e3_7	236284
e3_8	236284
e3_9	236284
e3_10	236284
e3_11	236284
e3_12	260827
e3_total	236435
e4	251634
e5	245779
e5_dia	245779
e5_sem	245779
e5_mes	245779
e5_ano	245779
e6_mes	246269
e6_ano	246269
e7	236937
e9	167247
e9_otro	259952
e10	259588
deseo_trabajar	170761

e11	156052
e12	188719
e12_otro	260378
e13	152538
e21_mes	196142
e21_ano	196142
e21_tramo	196142
e22	196142
e23	255111
e23_otro	260634
e24	241435
e24_otro	259420
activ	0
obe	108440
tpi	108440
id	167017
ftp	167017
cae_general	0
cae_especifico	0
categoria_ocupacion	0
ocup_form	108440
r_p_rev4cl_caenes	108440
sector	108440
fact_anual	0

ANEXO B: Formulario de preguntas ENE

Cuestionario de Preguntas INE, enlace.

ANEXO C: Variables seleccionadas en base a informe de la CEPAL sobre informalidad laboral

Variable	Descripción

region	Región donde se realizó la encuesta
mes_encuesta	Mes de recolección de datos
r_p_c	Comuna
tipo	Estrato rural/urbano
sexo	Sexo de la persona
cine	Clasificación Internacional Normalizada de la Educación (CINE)
tramo_edad	Tramo de edad
proveedor	Proveedor principal del hogar (sí/no)
edad	Edad de la persona encuestada
nacionalidad	Preguntas sobre nacionalidad del encuestado
estudia_actual	Si estudia o no
orig1	Si pertenece a un pueblo originario
orig2	Si pertenece a un pueblo originario
mig1	Pregunta sobre personas migrantes
mig2_cod	Pregunta sobre personas migrantes
mig3_cod	Pregunta sobre personas migrantes
mig4_cod	Pregunta sobre personas migrantes
mig5_cod	Pregunta sobre personas migrantes
mig6_cod	Pregunta sobre personas migrantes
a1	¿Trabajó al menos una hora?
a2	Realizó 'pololos' (independiente de la respuesta a1)
a3	¿Recibió o recibirá pago en dinero o especie?
b1	Grupo de ocupación según CIUO-08
b2	Situación en el empleo: ¿por cuenta propia o empleado?
b3	Tipo de compensación por el trabajo
b4	¿Emplea personas en su negocio?
b5	Sector del negocio (estatal o privado)
b7a_1	Cotización en el sistema previsional o de

	pensión
b7a_2	Cotización en el sistema de salud
b7a_3	Cotización en Seguro de Desempleo
b7b_2	Derecho a días pagados por enfermedad
b7b_3	Permisos por maternidad/paternidad
b8	Existencia de contrato escrito
b10	Para saber si es un trabajo temporal
b12	Está contratado o subcontratado
b11_proxy	Pago a través de sueldo, salario o jornal
b14_rev4cl_caenes	Rama de actividad económica según CAENES
i1	Registro en el Servicio de Impuestos Internos (SII)
i2	Servicios de contabilidad en la empresa
i3	Nombre de la empresa o negocio que paga el sueldo
i4	Empresa registrada en SII
i5	Tipo de registro de la empresa
i6	Variable no especificada
i7	Variable no especificada
ocup_honorarios	Si entregó boleta de honorarios
plataformas_digitales	Si el trabajo es realizado a través de una aplicación móvil o web
pd_especifique	Especificación de la aplicación utilizada
turno	Si trabaja por turnos
turno_d	Días trabajados habitualmente en el turno
turno_de	Días de descanso habituales en el turno
turno_h	Horas diarias trabajadas en el turno
turno_t	Total horas trabajadas habitualmente en el turno
habituales	Horas habituales de trabajo
efectivas	Horas efectivas trabajadas en la semana de

	referencia
activ	Condición de actividad (1 ocupado, 2 desocupado, 3 fuera FDT)
ocup_form	Personas ocupadas según formalidad (1 formal, 2 informal)
r_p_rev4cl_caenes	Rama de actividad económica
sector	Ocupados según sector (1 formal, 2 informal, 3 sector hogares)

ANEXO D: Detalle de variables utilizadas para esta entrega, anterior a filtrado

Column	Non-Null Count	Dtype
region	269978 non-null	int64
mes_encuesta	269978 non-null	int64
r_p_c	269978 non-null	int64
tipo	269978 non-null	int64
sexo	269978 non-null	int64
cine	269978 non-null	int64
proveedor	269978 non-null	int64
edad	269978 non-null	int64
nacionalidad	269978 non-null	int64
estudia_actual	269978 non-null	int64
orig1	35948 non-null	float64
orig2	269978 non-null	int64
mig1	24258 non-null	float64
mig2_cod	6012 non-null	float64
mig5_cod	97598 non-null	float64
mig6_cod	269978 non-null	int64
a1	269978 non-null	int64
a2	123989 non-null	float64
a3	134978 non-null	float64
b7a_1	114728 non-null	float64
b7a_2	114728 non-null	float64
b7a_3	114728 non-null	float64
b8	114728 non-null	float64
b7b_2	114728 non-null	float64
b7b_3	114728 non-null	float64
b12	118527 non-null	float64
b11_proxy	152538 non-null	float64
b14_rev4cl_caenes	152538 non-null	float64
s1	112527 non-null	float64
s2	112527 non-null	float64
13	37887 non-null	float64
14	37887 non-null	float64
15	12467 non-null	float64
16	12467 non-null	float64
turno_d	12467 non-null	float64
turno_t	12467 non-null	float64

efectivas	269978 non-null	int64
activ	269978 non-null	float64

ANEXO E: Detalle de variables utilizadas, posterior a filtrado e imputación

Column	Non-Null Count	Dtype
region	152105 non-null	category
mes_encuesta	152105 non-null	category
r_p_c	152105 non-null	category
tipo	152105 non-null	category
sexo	152105 non-null	category
cine	152105 non-null	category
proveedor	152105 non-null	category
edad	152105 non-null	category
nacionalidad	152105 non-null	category
estudia_actual	152105 non-null	category
orig1	152105 non-null	category
mig1	152105 non-null	category
a1	152105 non-null	category
a2	152105 non-null	category
a3	152105 non-null	category
b7a_1	152105 non-null	category
b7a_2	152105 non-null	category
b7a_3	152105 non-null	category
b8	152105 non-null	category
b7b_2	152105 non-null	category
b7b_3	152105 non-null	category
b12	152105 non-null	category
b11_proxy	152105 non-null	category
b14_rev4cl_caenes	152105 non-null	category
ocup_honorarios	152105 non-null	category
plataformas_digitales	152105 non-null	category

turno	152105 non-null	category
ocup_form	152105 non-null	category
r_p_rev4cl_caenes	152105 non-null	Category
habituales	152105 non-null	float64

ANEXO F: Red Neuronal

Se decidió anexar la red neuronal ya que no encuentro que sea un trabajo digno de explicar durante el trabajo sino más bien un intento de probar un modelo mucho más complicado para el modelo. Finalmente, no considero que poseo los conocimientos para que sea una herramienta tan útil como los modelos que vimos en este proyecto, pero considero importante hacerlo notar aunque sea en un anexo. Con esta entrega, también se adjuntará el código utilizado para generar la tabla siguiente.

Resultados de la Optimización con Optuna	
Descriptor	Valor
Hyperparameter: num_units	28
Hyperparameter: learning_rate	0.006358720879026974
Hyperparameter: dropout_rate	0.4908199591525962
Hyperparameter: batch_size	32
Hyperparameter: epochs	21
Final Test F1 Score	1.0
Final Test Recall	1.0
Class	Precision, Recall, F1-Score, Support
0.0	1.00, 1.00, 1.00, 32503
1.0	1.00, 1.00, 1.00, 13129
Accuracy	, , 1.00, 45632
Macro avg	1.00, 1.00, 1.00, 45632
Weighted avg	1.00, 1.00, 1.00, 45632