Proyecto Integrador 2023-II: Predicción de la Modalidad de Ocupación para la Primera Experiencia Laboral en Colombia

Elaborado por:

Andrés Morales Martínez

Giovanny Gómez Convers

Laura Margarita Guerrero Guerra

Sun Yin Huang Huang

Universidad EAFIT

Maestría en Ciencia de los Datos y Analítica

Presentado a:

Paula María Almonacid Hurtado

Juan David Martínez Vargas

Edison Valencia Díaz

Medellín

2 de diciembre de 2023

Tabla de contenido

Introducción 3
Marco teórico 3
Definición de conceptos claves 4
Estudios previos 5
Desarrollo Metodológico 9
Descripción del Problema 9
Análisis exploratorio de datos 10
Selección del modelo 15
Tecnología 19
Desarrollo del proyecto 19
Conclusiones generales del proyecto 25
Referencias bibliográficas 27

Introducción

En Colombia, más del 50% de la población trabaja en el sector informal[1], impactando de manera considerable la calidad de vida de los hogares y la economía del país en general, motivo por el cual se hace prioritario buscar e implementar estrategias que promuevan el acceso al empleo formal. A través de este proyecto, se busca predecir la probabilidad de que una persona acceda a una modalidad de ocupación formal o informal[2] como primera experiencia laboral, con base en las condiciones laborales de sus padres y en otras características o condiciones propias del individuo, tales como sus años de escolaridad y la composición del hogar. Para esto, se utilizan datos provenientes de la Gran Encuesta Integrada de Hogares 2022 (GEIH), elaborada por el Departamento Administrativo Nacional de Estadística (DANE), y se implementan los modelos de regresión logística (logit), árboles de decisión, random forest y XGboost, con el fin de compararlos y definir cuál es el que mejor se ajusta a los datos, minimizando el error en las estimaciones realizadas. A partir de los resultados obtenidos, el objetivo es entender cómo las variables puestas en consideración influyen en el tipo de empleo al que las personas acceden, lo cual permitiría orientar las medidas e iniciativas para llevar a cabo, con el fin de impulsar la disminución de la informalidad en Colombia.

Marco teórico

La formalidad e informalidad laboral son condiciones de gran impacto social en la sociedad colombiana. La informalidad laboral se refiere a la falta de protección social y derechos laborales que enfrentan los trabajadores en empleos no regulados (que son la mayoría en Colombia, más del 56%), mientras que la formalidad implica la adhesión a las normas y regulaciones laborales que ayudan a la seguridad social.

La informalidad laboral puede tener un impacto significativo en la movilidad intergeneracional. Cuando los padres trabajan en empleos informales, con bajos ingresos y sin acceso a beneficios sociales, es más probable que sus hijos enfrenten condiciones similares. La falta de oportunidades y de acceso a educación de calidad pueden perpetuar la informalidad y limitar las posibilidades de movilidad ascendente.

La movilidad intergeneracional es un fenómeno social que ha logrado capturar la atención de investigadores a lo largo del tiempo. Esta movilidad, se refiere a los cambios en la posición social de una generación a otra. La movilidad intergeneracional puede tener implicaciones de gran impacto en la igualdad y la forma en que se acceden a los recursos. Uno de los factores más importante que se ha identificado como variable influyente en la movilidad intergeneracional es la modalidad ocupacional parental. La ocupación de los padres y la forma de emplearse pueden desempeñar un papel fundamental en la formación de las metas profesionales y laborales de los hijos, así como en las oportunidades y recursos a los que tienen acceso.

[1] Para el tercer trimestre de 2023, esta cifra ascendía al 56,1%.

[2] Según el concepto formal de la Organización Internacional del Trabajo.

Se ha observado que el tipo de empleo que las personas obtienen en su primera experiencia laboral puede ser decisivo a largo plazo, es decir, marca un antes y puede llegar a ser determinante para el resto de la vida. Por lo tanto, comprender los factores que influyen en la elección del primer empleo, es fundamental para comprender la movilidad intergeneracional y las oportunidades laborales de los colombianos.

Definición de conceptos claves

- **Trabajo formal:**

En Colombia, el trabajo formal se refiere a aquellas actividades laborales que se realizan dentro del marco legal establecido por el país. Estas actividades se caracterizan por estar registradas ante las autoridades competentes, contar con contratos de trabajo formales, cumplir con las obligaciones fiscales y de seguridad social, y otorgar a los trabajadores los beneficios y protecciones establecidos por la legislación laboral colombiana. El trabajo formal en Colombia se encuentra regulado por el Código Sustantivo del Trabajo y brinda mayor estabilidad, seguridad social y derechos laborales a los trabajadores.[3]

- **Modalidad de ocupación:**

La modalidad de ocupación se refiere a las diferentes formas de empleo y trabajo en una sociedad. Incluye aspectos como el empleo asalariado, trabajo independiente, temporal, a tiempo parcial, por cuenta propia y el trabajo informal. Cada modalidad tiene características distintas en términos de contrato, dependencia y duración del empleo.

- **Movilidad intergeneracional:**

La movilidad intergeneracional se refiere a los cambios en la posición socioeconómica de una persona en relación con la posición de sus padres o generaciones anteriores. Es la capacidad de una persona para mejorar su nivel de vida y oportunidades en comparación con sus padres. La movilidad intergeneracional puede indicar la existencia de igualdad de oportunidades o desigualdades en una sociedad.[4]

- **Movilidad intergeneracional relativa ocupacional:**

La movilidad intergeneracional relativa ocupacional se refiere a la probabilidad de que una persona ocupe una posición ocupacional diferente en comparación con la ocupación de sus padres. Se utiliza para analizar si los hijos logran ascender o descender en la escala ocupacional en relación con sus padres. Es una medida de la movilidad social y de igualdad de oportunidades en una sociedad.

- **Trabajo informal:**

El trabajo informal en Colombia se refiere a actividades laborales que se realizan sin cumplir con los requisitos legales y regulaciones laborales establecidas por el país. Estas

[3] Ministerio del Trabajo de Colombia. (2023). Formalización Laboral. Recuperado de https://www.mintrabajo.gov.co/empleo-y-pensiones/empleo/subdireccion-de-formalizacion-y-proteccion-del-empleo/formalizacion-laboral

[4] Angulo, R., Azevedo, J. P., Gaviria, A., & Paez, G. N. (2012). Movilidad Social en Colombia. Recuperado de https://economia.uniandes.edu.co/sites/default/files/imagenes/eventos/Roberto-Angulo-Movilidad-Social-en-Colombia.pdf

actividades se caracterizan por la falta de registro, contratos formales y protecciones laborales, y suelen llevarse a cabo en el sector informal de la economía.

Estudios previos

Iniciando la investigación, se intentó ver estudios previos donde fuera posible establecer la relación ocupacional de los padres en el primer empleo de sus hijos. Sin embargo, ningún estudio previo, pudo sostener hipótesis directas acerca del problema a resolver. Aun así, la movilidad intergeneracional ha movido intereses desde la década de los 80 y ha sido motivo por varios autores e inclusive gobiernos.

A continuación, se exponen los más relevantes, estudios que indirectamente muestran una relación sólida con lo que es la movilidad intergeneracional, la influencia de los padres en el futuro de sus hijos y la relación en carreras universitarias con ambientes familiares de primer grado:

- **Informalidad laboral en Colombia: características y determinantes**

  - **Características de la informalidad:** Se encontró que la informalidad laboral en Colombia es más común en trabajadores de menor nivel educativo, especialmente aquellos con educación primaria o sin educación formal. También se observó una mayor incidencia de informalidad en trabajadores de sectores agrícolas y en empleos por cuenta propia.

  - **Determinantes de la informalidad:** El estudio identificó varios factores que influyen en la elección de la informalidad por parte de los trabajadores. Entre ellos se encuentran la falta de oportunidades de empleo formal, la rigidez de la regulación laboral, la falta de acceso a servicios de salud y protección social, y la existencia de altos costos asociados con la formalización, como los impuestos y las contribuciones.

  - **Impacto económico y social:** La informalidad laboral tiene efectos negativos tanto a nivel económico como social. En el ámbito económico, contribuye a la baja productividad y a la reducción de la recaudación fiscal. A nivel social, los trabajadores informales enfrentan mayores dificultades para acceder a servicios básicos, como la salud y la seguridad social, lo que aumenta su vulnerabilidad.

  - **Recomendaciones de política:** El estudio sugiere que, para reducir la informalidad laboral en Colombia, es necesario adoptar medidas que fomenten la generación de empleo formal, mejoren la calidad de la educación y promuevan una regulación laboral más flexible. Además, se destaca la importancia de implementar políticas que faciliten la formalización de los trabajadores informales, como reducir los costos asociados con la formalidad y brindar incentivos para la adhesión al sistema formal.

- **La formalización del empleo en Colombia: diagnóstico, determinantes y recomendaciones de política**

  - **Diagnóstico de la informalidad:** El estudio revela que la informalidad laboral en Colombia ha sido históricamente alta, con una gran proporción de trabajadores involucrados en empleos informales o por cuenta propia. Se identificaron varios factores que contribuyen a la persistencia de la informalidad, como la falta de oportunidades de empleo formal, la baja productividad, la complejidad y rigidez de la regulación laboral, y la baja calidad de la educación.

  - **Determinantes de la informalidad:** El análisis de los determinantes de la informalidad revela que existen diferencias significativas entre las regiones de Colombia. Factores como la estructura productiva, la pobreza, la calidad de la infraestructura, la violencia y el acceso a servicios públicos afectan la incidencia de la informalidad en cada región.

  - **Recomendaciones de política:** El estudio propone una serie de recomendaciones de política para fomentar la formalización del empleo en Colombia. Estas incluyen mejorar la calidad de la educación y la capacitación laboral, facilitar el acceso a servicios financieros para emprendedores y pequeñas empresas, simplificar y flexibilizar la regulación laboral, promover la formalización gradual a través de incentivos fiscales y mejorar la coordinación entre entidades gubernamentales para abordar de manera integral la informalidad.

  - **Impacto económico y social:** El estudio destaca que la informalidad laboral tiene un impacto negativo en la economía y la sociedad colombiana. Contribuye a la baja productividad, la evasión fiscal, la falta de protección social para los trabajadores y la desigualdad. Además, la informalidad dificulta la implementación de políticas públicas efectivas y limita el acceso a servicios básicos, como la salud y la seguridad social.

- **Determinantes de la informalidad laboral en Colombia: un enfoque a nivel territorial**

  - **Análisis territorial:** El estudio analiza la informalidad laboral en diferentes regiones de Colombia y busca identificar los determinantes específicos de la informalidad en cada área. Se examinan factores como la densidad empresarial, la educación, la pobreza, la violencia y el acceso a servicios públicos.

  - **Factores determinantes:** Los resultados del estudio muestran que existen diferencias significativas en los determinantes de la informalidad entre las regiones de Colombia. La densidad empresarial y la presencia de sectores económicos específicos tienen una influencia importante en la incidencia de la

informalidad. Además, se encontró que la educación, la pobreza y la violencia también desempeñan un papel significativo en la informalidad laboral en cada región.

- **Conclusiones:** El estudio resalta la importancia de tener en cuenta los factores territoriales en el diseño de políticas para abordar la informalidad laboral en Colombia. Las políticas deben adaptarse a las características específicas de cada región y considerar tanto los factores económicos como los sociales. Además, se destaca la necesidad de mejorar la educación, promover el emprendimiento y fortalecer la presencia empresarial en áreas con alta informalidad laboral.

- **Implicaciones de política:** El estudio sugiere que las políticas para reducir la informalidad laboral en Colombia deben ser diferenciadas y adaptadas a las realidades territoriales. Se recomienda impulsar la creación de empleo formal a través de la promoción del desarrollo empresarial en áreas con baja densidad empresarial. Asimismo, se resalta la importancia de políticas que mejoren la calidad de la educación y reduzcan la pobreza y la violencia, factores que influyen en la elección de la informalidad laboral.

▪ **Parental Influences on Occupational Aspirations: A Comparative Study of Adolescents in China, Japan, and South Korea**

- En este estudio longitudinal, los investigadores se propusieron analizar cómo la influencia de los padres puede moldear las aspiraciones ocupacionales de los jóvenes adolescentes a lo largo del tiempo. Su objetivo era comprender cómo los factores parentales, como el nivel educativo, la ocupación y las expectativas de los padres, pueden influir en las metas profesionales de los adolescentes.

- Los hallazgos de este estudio revelaron que la influencia de los padres es un predictor significativo de las aspiraciones ocupacionales de los jóvenes adolescentes. Se encontró que los jóvenes cuyos padres tienen un nivel educativo más alto y ocupaciones más prestigiosas tienden a tener aspiraciones ocupacionales más altas y diversas.

- Además, el estudio mostró que las expectativas de los padres también desempeñan un papel importante en las aspiraciones ocupacionales de los jóvenes adolescentes. Los jóvenes que perciben que sus padres tienen altas expectativas sobre su éxito y logros profesionales tienden a establecer metas más ambiciosas y orientadas al logro.

- **The Influence of Parental Occupation and Educational Expectations on Vocational Aspirations in Adolescence**

  - El impacto de los antecedentes familiares en las aspiraciones vocacionales de los adolescentes. Uno de los estudios clave en esta área es el de Shanahan y otros autores (2007), titulado "The Influence of Parental Occupation and Educational Expectations on Vocational Aspirations in Adolescence".

  - En este estudio, se intensifica la relación entre la ocupación de los padres y las expectativas educativas, y cómo estas variables influyen en las aspiraciones vocacionales de los adolescentes. Su objetivo era comprender cómo el entorno familiar, específicamente la profesión de los padres y las expectativas educativas puede moldear las metas y aspiraciones profesionales de los adolescentes durante la etapa crucial de la adolescencia.

  - En temas de resultados de este estudio, estos revelaron que la ocupación de los padres y las expectativas educativas tienen una influencia significativa en las aspiraciones vocacionales de los adolescentes. Los adolescentes cuyos padres tienen profesiones de alto prestigio o altas expectativas educativas tienden a tener aspiraciones vocacionales más altas y a buscar carreras que reflejen el estatus y las metas familiares.

Como se puede observar en los estudios mencionados, ninguno tiene una relación entre variables directas que puedan dar una hipótesis previa a la que aquí se quiere demostrar. Sin embargo, de los estudios se obtienen unos datos y hallazgos que nos permiten establecer una relación directa entre la modalidad ocupacional parental con el primer empleo de los hijos.

- El primer elemento de hallazgo es que los padres desempeñan un papel importante en la formación de las metas y aspiraciones profesionales de los adolescentes. Esto sugiere que la modalidad de ocupación de los padres puede tener un impacto en la elección de carrera de sus hijos.

- Se demuestra que los padres tienen un rol crucial en las aspiraciones ocupacionales de los adolescentes. Si los padres tienen trabajos de alto estatus o tienen altas expectativas educativas, es más probable que sus hijos también tengan metas profesionales más elevadas. Los factores parentales, como la educación, la ocupación y las expectativas, tienen una influencia significativa en las aspiraciones ocupacionales de los adolescentes.

Las investigaciones mencionadas proporcionan evidencia concluyente de una relación directa entre la modalidad de ocupación de los padres y el primer empleo de los hijos, lo cual afecta la movilidad intergeneracional. Los factores parentales, como el nivel educativo, la ocupación y las expectativas de los padres, influyen en las aspiraciones profesionales de los hijos. El apoyo instrumental de los padres, la comunicación abierta y el acceso a información sobre diferentes carreras también desempeñan un papel

importante. Estos hallazgos subrayan la importancia de considerar la influencia ocupacional en el estudio de la movilidad intergeneracional y la necesidad de tener en cuenta características y variables que pueden cambiar el rumbo de una generación completa.

Desarrollo Metodológico

Descripción del Problema

Colombia tiene una de las tasas de empleo informal más altas del mundo. Según el Departamento Administrativo Nacional de Estadística (DANE), en el trimestre julio - septiembre del 2023, la proporción de ocupados informales a nivel nacional fue del 56,1%. Esta modalidad de ocupación es muy común entre los jóvenes (quienes no tienen experiencia), y está asociada con trabajos de baja calidad, en los que no existe ninguna regulación de las jornadas laborales, no brindan condiciones para la protección contra el despido arbitrario ni, mucho menos, representación en los sindicatos (OIT, 2002). Tampoco suelen ofrecer ingresos fijos y justos, creando inestabilidad económica en los individuos, en conjunto con ciclos de pobreza y exclusión social (Pérez et al, 2014). Dentro de sus principales consecuencias se encuentra el no acceso al sistema de pensiones y de salud, lo cual tiene efectos considerables sobre el bienestar de las familias, dadas las deficiencias que se generan para ellas en términos de salud. Adicionalmente, teniendo en cuenta que los salarios en el mercado informal tienden a ser mucho más bajos que los del sector formal, contribuyen a que se perpetúe la pobreza en el país. Otras consecuencias igualmente importantes son la reducción de la productividad, la congestión de los servicios públicos sin contribuir a su financiación, la baja cobertura de la seguridad social contributiva y la reducción de los recaudos tributarios, lo cual constituye una amenaza para la sostenibilidad del sistema de seguridad social y disminuye las posibilidades de los trabajadores de obtener una pensión adecuada.

Lo anterior demuestra la importancia de tomar medidas que promuevan el acceso al empleo formal por parte de los colombianos, para lo cual es indispensable identificar los factores que inciden en su materialización, que pueden darse, tanto desde los negocios o empresas que lo ofrecen, como desde las características y condiciones de vida de quienes acceden a dicha modalidad de ocupación. Frente a estos últimos, una variable que puede tener gran relevancia es la movilidad intergeneracional, que hace referencia al cambio de posición social de un individuo respecto a sus padres o antepasados (en diferentes generaciones), dado que se hace necesario prevenir que los estatus laborales (en este caso, la modalidad de ocupación informal) se hereden entre generaciones y, por el contrario, promover transiciones hacia la formalidad. Otras variables como las características propias de la persona (nivel de escolaridad, edad, etc.) y las condiciones del hogar del que hace parte también pueden tener una influencia importante sobre su modalidad de ocupación a nivel laboral.

El objetivo de este proyecto es predecir cuando una persona accede a una modalidad de ocupación formal o informal como primera experiencia laboral, a partir de las condiciones laborales de sus padres (estudio de la movilidad intergeneracional relativa ocupacional), de otras características y/o condiciones propias del individuo, entre ellas, sus años de escolaridad y la composición del hogar y controlando por algunas variables macroeconómicas como la variación del índice de precios al consumidor y la tasa de desempleo. El entendimiento de estas relaciones permitiría orientar las medidas e iniciativas para llevar a cabo, con el fin de impulsar la disminución de la informalidad en Colombia. En el caso de la movilidad intergeneracional, si la modalidad de ocupación de los padres tiene una influencia importante sobre la modalidad de ocupación de los hijos, será conveniente implementar acciones que fomenten la transición a la formalidad desde los mismos padres, o en el caso de los años de escolaridad, se podrían implementar programas de educación a los cuales las personas con menos posibilidades puedan acceder fácilmente. Estos son solo algunos ejemplos del enfoque que podrían tener dichas iniciativas.

Análisis exploratorio de datos

1. Entendimiento de los datos

Los datos empleados para el desarrollo del proyecto provienen de la Gran Encuesta Integrada de Hogares 2022 (GEIH), elaborada por el Departamento Administrativo Nacional de Estadística (DANE), que es la entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia.

La GEIH constituye una base de datos pública y anonimizada, y su objetivo es proporcionar información básica sobre el tamaño y estructura de la fuerza de trabajo del país, así como de las características sociodemográficas que permiten caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, años de escolaridad, ingresos y afiliación al sistema de seguridad social. Esta encuesta es representativa a nivel departamental y se realiza mensualmente a diferentes viviendas de diferentes áreas metropolitanas.

La encuesta se compone por nueve módulos:

- Características generales, seguridad social en salud y educación
- Tipo de investigación
- Fuerza de trabajo
- Ocupados
- No ocupados
- Otras formas de trabajo
- Migración
- Datos del hogar y de la vivienda
- Otros ingresos e impuestos

Finalmente, la GEIH tiene como unidad de observación al hogar, por lo que provee información de cada individuo perteneciente a este. Esto permite que, a cada individuo, según sus características, se le destine un módulo en específico. Dentro de una vivienda pueden existir varios hogares, y dentro del hogar, cada individuo tiene una categoría diferente.

Para el desarrollo del proyecto, se tomó como base los resultados de la GEIH correspondientes al rango enero-diciembre de 2022, específicamente los módulos "Características Generales, Seguridad Social en Salud y Educación" y "Ocupados".

Para tener más información sobre la GEIH, se puede remitir al enlace: https://microdatos.dane.gov.co/index.php/catalog/771/study-description.

Por otro lado, para agregar las variables macroeconómicas mensualizadas se utilizaron la variación del índice de precios al consumidor; cuya fuente es el DANE, la tasa de desempleo y la tasa de interés de las empresas recolectadas desde la plataforma de Bloomberg. La inclusión de las variables macroeconómicas también puede contribuir a mejorar la precisión y validez del modelo al controlar factores externos que podrían afectar la relación entre las variables independientes y la variable dependiente. Estas variables adicionales se conocen como variables de control. Al introducir variables de control, se busca aislar el efecto específico de las variables de interés, permitiendo así realizar observaciones más precisas y comparables, y, por lo tanto, incluirlas en el modelo ayuda a capturar mejor la complejidad de las relaciones económicas.

2. Preparación de los datos

Dado el objetivo del proyecto integrador, la muestra de individuos a considerar para su desarrollo debe cumplir con las siguientes condiciones:

- Que la GEIH tenga información de un jefe de hogar y un hijo
- Que el hijo se encuentre laborando
- Que sea la primera experiencia laboral del hijo

Para ello, se realizó el siguiente procedimiento:

- Unificación de los meses y módulos requeridos, en un data set (en la sección de tecnología se puede encontrar información más detallada).

- Selección de variables requeridas para el estudio, a partir de la revisión de literatura y del marco teórico del problema.

- Implementación de las 3 condiciones mencionadas anteriormente:

  - Información del jefe de hogar y de, al menos, un hijo suyo: La GEIH tiene la variable 'P6050', que hace referencia al parentesco de la persona encuestada con el jefe del hogar en el que reside. Si la persona encuestada es el jefe de hogar, esta variable toma el valor de 1, mientras que, si es hijo/a del jefe del hogar, la variable toma un valor de 3. Para hacer el respectivo filtro, se seleccionaron aquellos hogares que tienen valores de 1 y 3 en esta variable.

  - Hijo ocupado (que se encuentre laborando): La GEIH tiene la variable 'OCI', que hace referencia al módulo de Ocupados, es decir, que actualmente se encuentran laborando. Para hacer el respectivo filtro, se seleccionaron los hogares en donde los hijos tienen la variable dummy.

  - Primera experiencia laboral: La GEIH tiene la variable 'P7020' asociada a la pregunta: Antes del actual trabajo, ¿tuvo otro trabajo?, que toma el valor de 1 si es afirmativa y 2 en caso contrario. Para hacer el respectivo filtro, seleccionamos aquellos hijos que tengan esta variable igual a 2.

- Creación de las variables para el modelo: A cada hijo se le asignan variables binarias por cada característica propia, de sus padres, del hogar y de contexto. Ejemplo:

	mujer	joven	raza	educ	j_formal	j_mujer	zona	…
Hijo3122	1	1	0	0	1	0	1	

  Para más información acerca de las variables, remitirse al código adjunto en la sección 5.2.

- Eliminación de duplicados, es decir, de los demás miembros del hogar. Hasta el momento se mantuvieron todos los miembros del hogar para crear variables como el logaritmo del ingreso de los padres, la cantidad de individuos, si es un hogar nuclear, entre otros.

- Ahora bien, dado que necesitamos un único hijo por hogar, finalizamos la preparación de los datos con la eliminación de duplicados de hijos. Para ello, filtramos en el siguiente orden: el hijo de mayor edad, si seguían duplicados procedemos con el hijo que lleva mayor tiempo laborando y si seguían con duplicados, eliminamos aleatoriamente. Con estos últimos se revisaron otras variables que nos permitiera seleccionar a uno de ellos, no obstante, ninguna variable fue decisoria. Esta eliminación aleatoria representa únicamente el 0,09% de la submuestra.

- Finalmente, identificamos que algunas de estas observaciones tenían edades atípicas. Para ello, utilizando la regla de 3 sigma, se eliminan aquellas que se encuentren por encima del límite superior del bigote. Esto corresponde al 3.85% de la submuestra. La principal razón por la cual se puede encontrar estas edades es porque la población continúa trabajando en el primer trabajo que obtuvo, y por lo tanto, nos permite validar la importancia del objetivo de este estudio.

3. Análisis descriptivo e insights importantes

Iniciamos con una muestra de 919.459 observaciones y después de la preparación, terminamos con una submuestra de 26.741 individuos.

Las variables seleccionadas son:

Variable Dependiente	
Ocupación: Se utiliza el módulo de "Ocupados" de la GEIH.	
Formal: Se define como aquellas personas que se encuentran cotizando al sistema de salud y pensión.	Informal: Se define como aquellas personas que se encuentran ocupadas y no entran en la categoría de formal

Esta variable 'formal' se crea de una combinación de las variables 'salud' y 'pensión' que hacen referencia a que en el momento de la encuesta cotizan a estos sistemas.

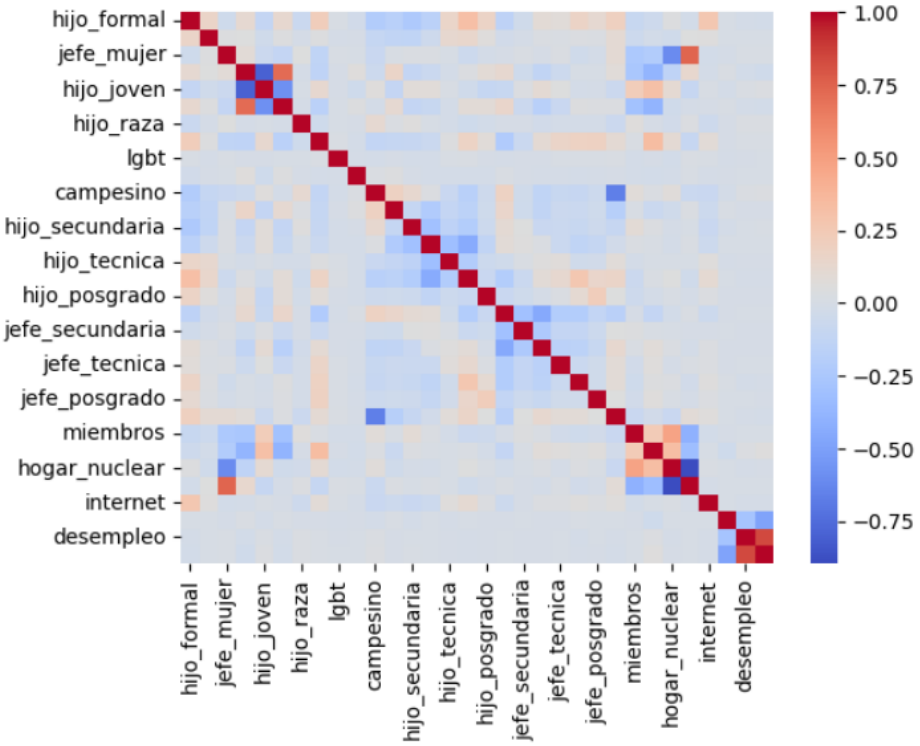
Variables Independientes			
Se utiliza el módulo de "Características generales, seguridad social en salud y educación" y "Ocupados" de la GEIH			
Del hijo	**Del padre**	**Del hogar**	**Macroeconómicas**
Edad (joven) Sexo Nivel educativo Raza Primer empleo (filtro) LGBT Campesino Discapacidad	Edad Sexo Nivel educativo Modalidad de ocupación (formal/informal) Raza	Log. Ingreso padres Número de miembros Hogar nuclear Zona (urbana/rural) Departamento Internet	Variación del índice de precios al consumidor Tasa de desempleo Tasa de interés a las empresas

La submuestra de 26.741 individuos se compone principalmente por hombres (59,9%) y población joven (57,3%) con unos ingresos promedio de $992.303 y un nivel educativo medio (10° - 11°) seguido de pregrado. En cambio, el jefe de hogar es principalmente una mujer con una edad promedio de 55 años, un ingreso 2% menor ($973.385) y un nivel educativo primario seguido de educación media.

Dentro de la población encontramos un 11,4% que se reconoce como indígena, gitano, raizal del archipiélago de San Andrés, Providencia y Santa Catalina, palenquero (a) de San Basilio, negro, mulato o afrocolombiano, un 1,03% de la comunidad LGBT, un 18,2% campesina y un 1,4% en situación de discapacidad.

En cuanto al hogar, la mayoría de los hijos se encuentra en un hogar nuclear, es decir, compuesto por un jefe de hogar y una pareja (53,6%). Por otro lado, frente a los hogares monoparentales, la mayoría tienen una jefe de hogar mujer (40,1%). Adicionalmente el 86,9% de los hogares se encuentran en la zona urbana y tienen en promedio 3,56 miembros.

Finalmente, encontramos 1.358 individuos (5,1%) que dentro del hogar utilizan activamente el internet para la búsqueda de trabajo.



Para el análisis de correlación, varias variables estuvieron correlacionadas en más de un 70%:

```
hogar_madre y jefe_mujer tienen una correlación de 0.74
hijo_joven y edad_hijo tienen una correlación de -0.80
edad_jefe y edad_hijo tienen una correlación de 0.72
hogar_madre y hogar_nuclear tienen una correlación de -0.89
tasa y desempleo tienen una correlación de 0.84
```

Bajo estos resultados se seleccionaron las siguientes variables para el modelo.

```
var_x = ['jefe_formal', 'hijo_mujer', 'jefe_mujer', 'hijo_joven', 'edad_jefe', 'hijo_raza', 'jefe_raza', 'lgbt', 'discapacidad', 'campesino',
         'hijo_primaria', 'hijo_secundaria', 'hijo_media', 'hijo_tecnica', 'hijo_pregrado', 'hijo_posgrado',
         'jefe_primaria', 'jefe_secundaria', 'jefe_media', 'jefe_tecnica', 'jefe_pregrado', 'jefe_posgrado',
         'urbana', 'miembros', 'Ling_padres', 'hogar_nuclear', 'internet', 'antioquia', 'atlantico', 'bogota', 'bolivar', 'boyaca', 'caldas',
         'caqueta', 'cauca', 'cesar', 'cordoba', 'cundinamarca', 'choco', 'huila', 'guajira', 'magdalena', 'meta', 'narino', 'norte',
         'quindio', 'risaralda', 'santander', 'sucre', 'tolima', 'valle', 'VariacionIPC', 'desempleo']
```

Finalmente, la proporción de formales e informales para los hijos y jefes del hogar se presentan en la siguiente tabla.

Hijo	**Jefe**	
	Formal	*Informal*
Formal	2.763	6.804
Informal	1.953	15.221

Selección del modelo

1. Modelos

La selección de un modelo supervisado se define por la existencia de la etiqueta, que en este caso es la modalidad de ocupación. Ahora bien, esta etiqueta es una variable binaria que toma el valor de 1 cuando el individuo es un ocupado formal y el contrafactual seria ocupado informal. En la literatura, dentro de los modelos tradicionales para este tipo de etiqueta están los modelos de respuesta binaria como la regresión logística (logit), por otro lado, también se destacan modelos de aprendizaje automático como los árboles de decisión, random forest y XGBoost.

Estos últimos algoritmos ofrecen ventajas significativas en comparación con modelos convencionales como la regresión logística (logit). En primer lugar, los árboles de decisión permiten capturar patrones complejos y no lineales en los datos, lo que los hace especialmente efectivos cuando las relaciones entre variables son intrincadas y no se ajustan bien a modelos lineales.

Random Forest, por su parte, mejora aún más la robustez y precisión al combinar múltiples árboles de decisión, reduciendo así el riesgo de sobreajuste y mejorando la generalización del modelo. Esta capacidad de manejar conjuntos de datos complejos y grandes contribuye a un rendimiento más sólido y a la capacidad de abordar problemas de alta dimensionalidad.

XGBoost, un algoritmo de refuerzo destaca por su eficiencia y velocidad en la construcción de modelos. Al utilizar técnicas de refuerzo secuencial, XGBoost mejora la precisión del modelo en cada iteración, permitiendo la identificación de patrones más sutiles y la optimización continua del rendimiento.

Comparados con modelos convencionales como logit, los mencionados algoritmos de aprendizaje automático tienden a ofrecer una flexibilidad superior y a adaptarse mejor a la complejidad inherente a muchos conjuntos de datos del mundo real. Además, al poder capturar interacciones no lineales y manejar datos heterogéneos de manera más efectiva, estos modelos brindan una visión más precisa y completa, lo que resulta crucial en situaciones donde la relación entre variables es intrincada y no puede ser plenamente capturada por modelos lineales más simples. En resumen, la adopción de árboles de decisión, Random Forest y XGBoost representa un avance significativo en la capacidad de los modelos predictivos, permitiendo abordar problemas complejos con mayor precisión y flexibilidad que los modelos tradicionales como logit.

Para efectos de este ejercicio, los cuatro modelos son puestos a prueba realizando el siguiente procedimiento:

- Se crea un pipeline para la transformación de los datos numéricos y categóricos.
- Se divide la submuestra en datos de entrenamiento (70%), validación (15%) y testeo (15%) manteniendo la estratificación de la etiqueta.
- Entrenar el modelo con los datos de testeo
- Encontrar los mejores parámetros de los modelos. Para ello, utilizamos GridSearch y RandomizedSearch.
- Ajustar los parámetros al modelo.
- Entrenar el modelo con los datos de validación.
- Se predice la etiqueta con los datos de testeo.
- Se calcula la precisión, recall y F1.
- Se grafica la curva ROC.

Es crucial resaltar un aspecto esencial del proceso experimental llevado a cabo, que involucra la realización de dos experimentos distintos para abordar la estandarización de las variables numéricas. En la primera iteración, se optó por no estandarizar las variables numéricas utilizando la función de min-máx.. Sin embargo, los resultados obtenidos no se alinearon con las expectativas iniciales, lo cual generó interrogantes y motivó la realización de un segundo experimento más exhaustivo.

En el segundo experimento, se decidió llevar a cabo una estandarización al aplicar la función de min-máx. a todas las variables numéricas. Esta decisión estratégica surgió como respuesta a las discrepancias observadas en el primer experimento y tuvo como objetivo evaluar de manera más integral el impacto de la estandarización en el rendimiento del modelo. La inclusión de todas las variables numéricas en este proceso de estandarización proporcionó una visión más holística y detallada, permitiendo una comprensión más profunda de cómo la transformación afecta el comportamiento del modelo en su conjunto.

A partir de estos resultados se define que el modelo escogido es el XGBoost según el F1 Score. Para nuestro caso, el F1 Score es la métrica ideal ya que no tenemos una proporción balanceada entre formales e informales.

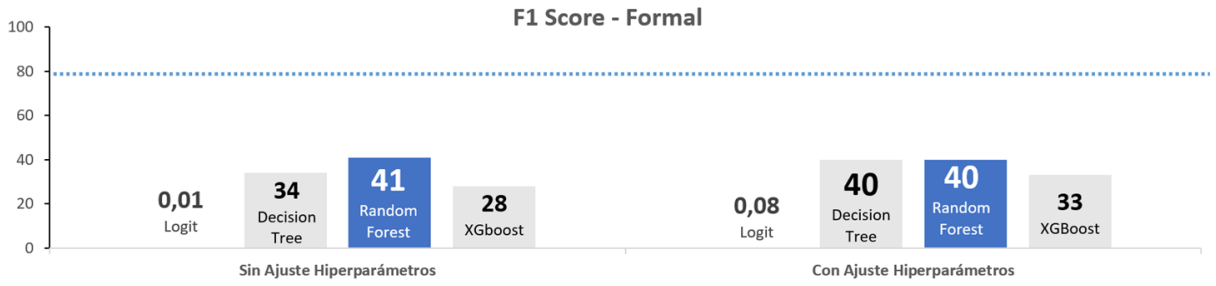
En la siguiente imagen se pueden encontrar los resultados del experimento 1.

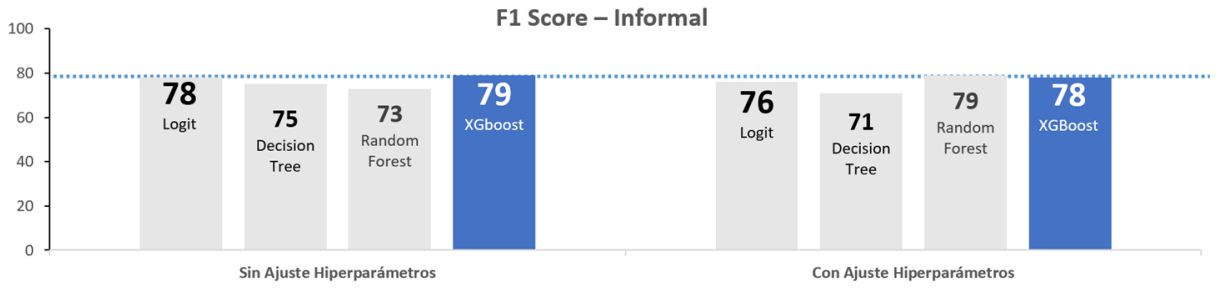
80

Modelo predice Formal	Sin ajuste hiperparámetros	Con ajuste hiperparámetros	Diferencia al F1 Score
Logit	0,01	0,08	**-79,92**
Decision Tree	34	40	-40
Random Forest	**41**	40	-39
XGboost	28	33	-47

80

Modelo predice Informal	Sin ajuste hiperparámetros	Con ajuste hiperparámetros	Diferencia al F1 Score
Logit	**78**	76	-2
Decision Tree	**75**	71	**-5**
Random Forest	73	**79**	-1
XGboost	**79**	78	-1





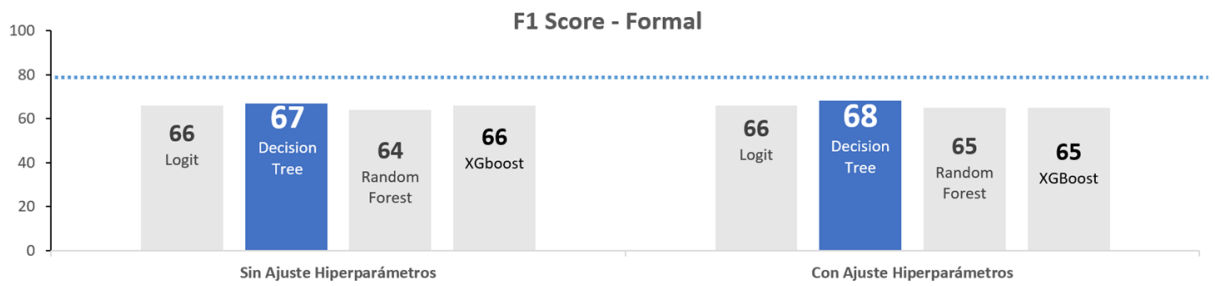
No obstante, los mejores resultados se presentan en el experimento 2.

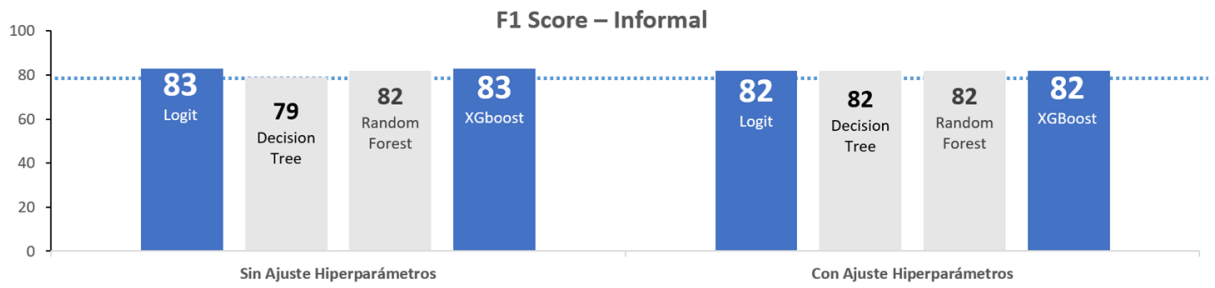
80

Modelo predice Formal	Sin ajuste hiperparámetros	Con ajuste hiperparámetros	Diferencia al F1 al Score
Logit	66	66	-14
Decision Tree	67	**68**	**-12**
Random Forest	64	65	-15
Xgboost	66	65	-15

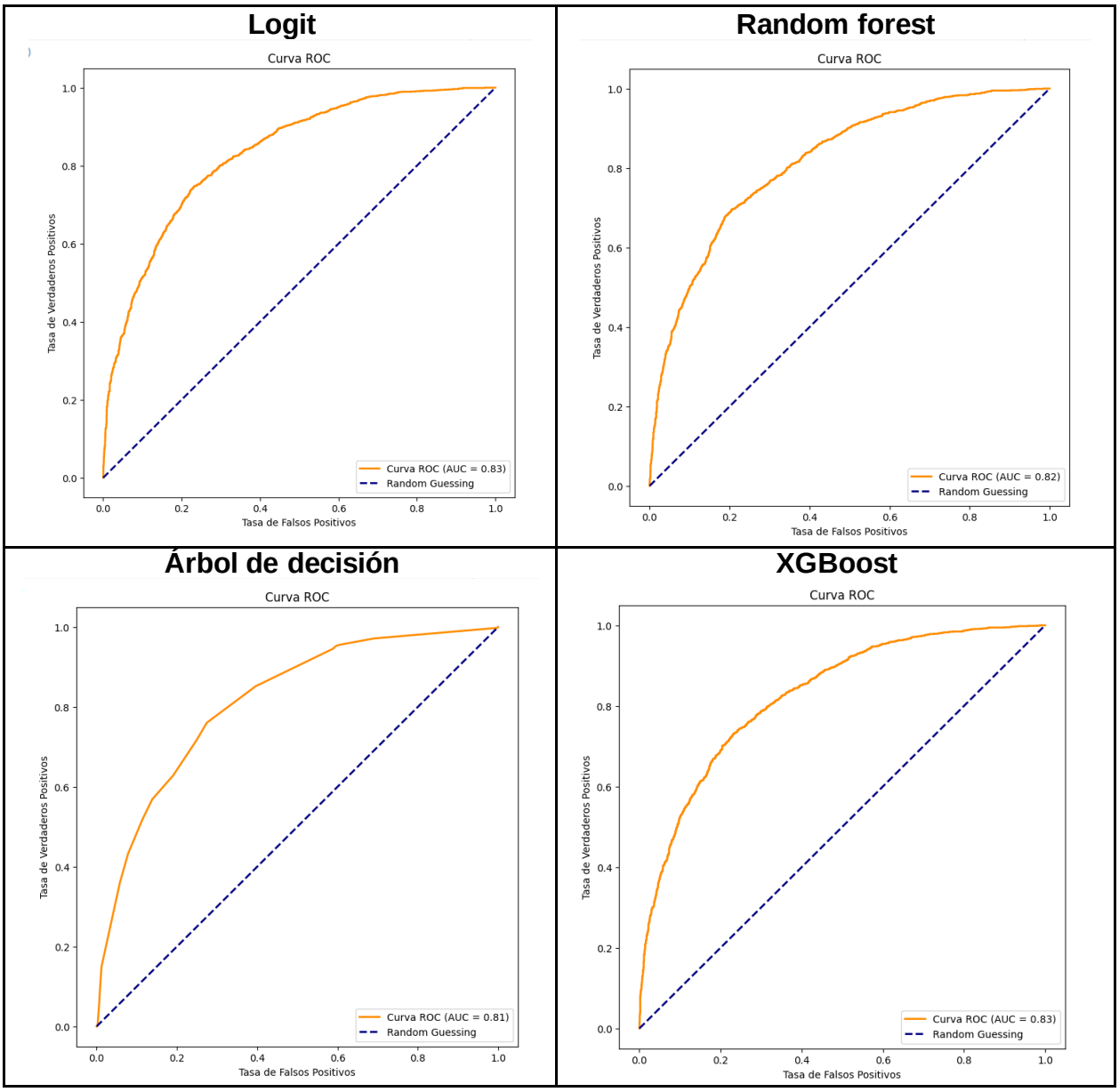
80

Modelo predice Informal	Sin ajuste hiperparámetros	Con ajuste hiperparámetros	Diferencia al F1 Score
Logit	**83**	82	+3
Decision Tree	79	82	**+2**
Random Forest	82	82	+2
XGboost	**83**	82	+3





Adicionalmente, para todos los modelos se graficaron las curvas ROC.



Cabe destacar que el modelo logit se presenta como una opción viable, considerando el equilibrio entre la métrica F1 y el tiempo de cómputo. Aunque tradicionalmente ha sido utilizado en contextos más simples o en situaciones donde se presupone una relación lineal entre las variables predictoras y la variable objetivo, el modelo logit ha demostrado ser eficaz en diversos escenarios. Su simplicidad computacional y eficiencia en términos de tiempo lo convierten en una alternativa atractiva, especialmente cuando se busca un rendimiento aceptable en métricas como F1 sin incurrir en una mayor carga computacional como lo fue en el caso de XGBoost.

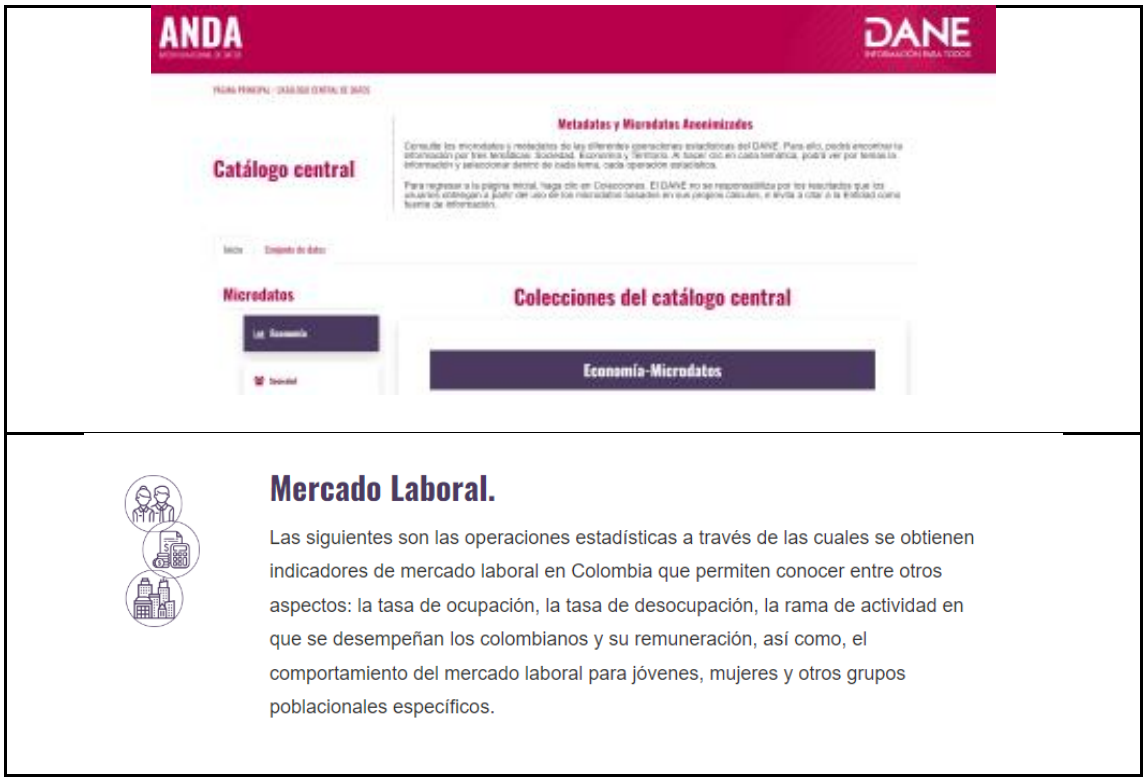
Tecnología

Desarrollo del proyecto

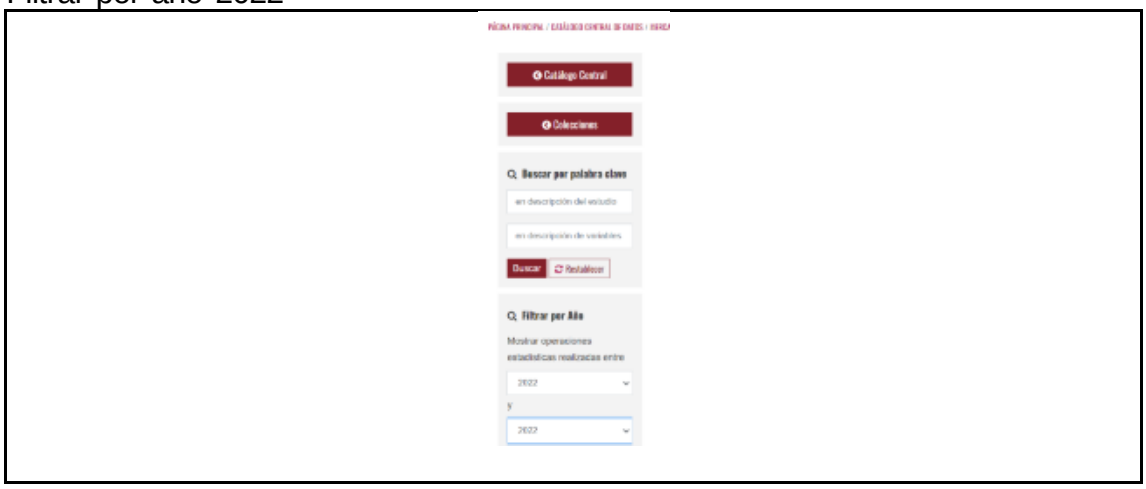
1. Adquisición de los datos

El DANE dispone los datos de la GEIH en el portal ANDA (Archivo Nacional de Datos) https://microdatos.dane.gov.co/ a través de un catálogo central en el cual se encuentran metadatos y microdatos anonimizados. Para obtenerlos se deben seleccionar las opciones:

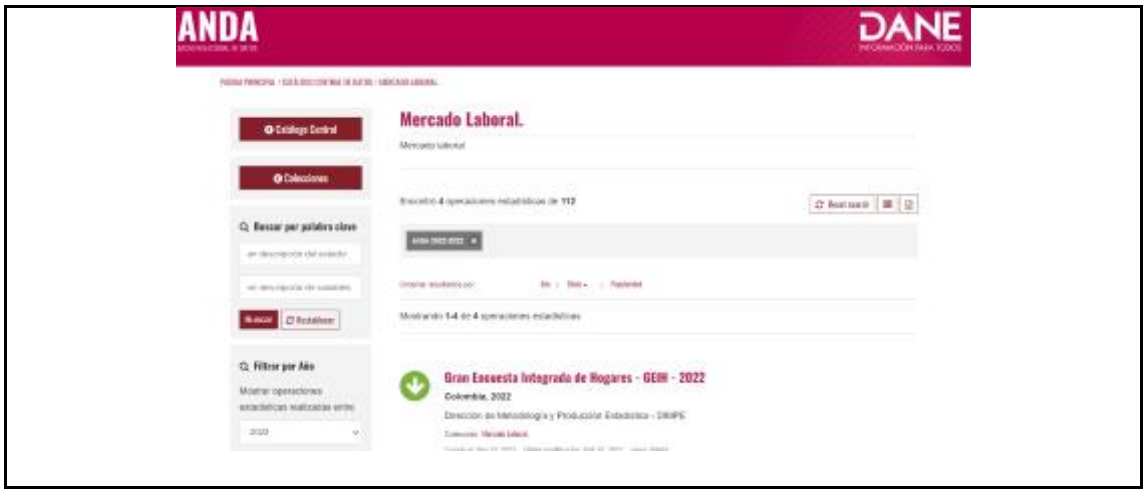
- Microdatos>Economía>Mercado Laboral



- Filtrar por año 2022

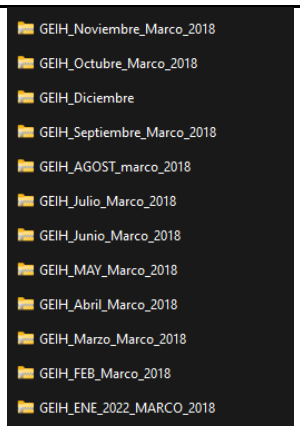


- Seleccionar Gran Encuesta Integrada de Hogares – GEIH – 2022 https://microdatos.dane.gov.co/index.php/catalog/771

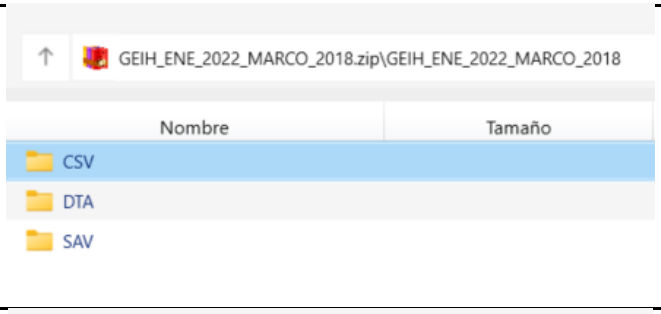


- Seleccionar Obtener Microdatos

- Seleccionar cada uno de los meses, descargar ZIP pasando reCaptcha



- Descomprimir cada zip seleccionado solo los archivos en formato CSV



- Como se indicó en el análisis exploratorio de datos, los archivos que se tomarán para el trabajo serán:

Características generales, seguridad social en salud y educación.csv

Ej. enero 2022

Nombre	Tipo	Tamaño
Características generales, seguridad social en salud y educación	Archivo de valores...	12.650 KB

Ocupados.csv

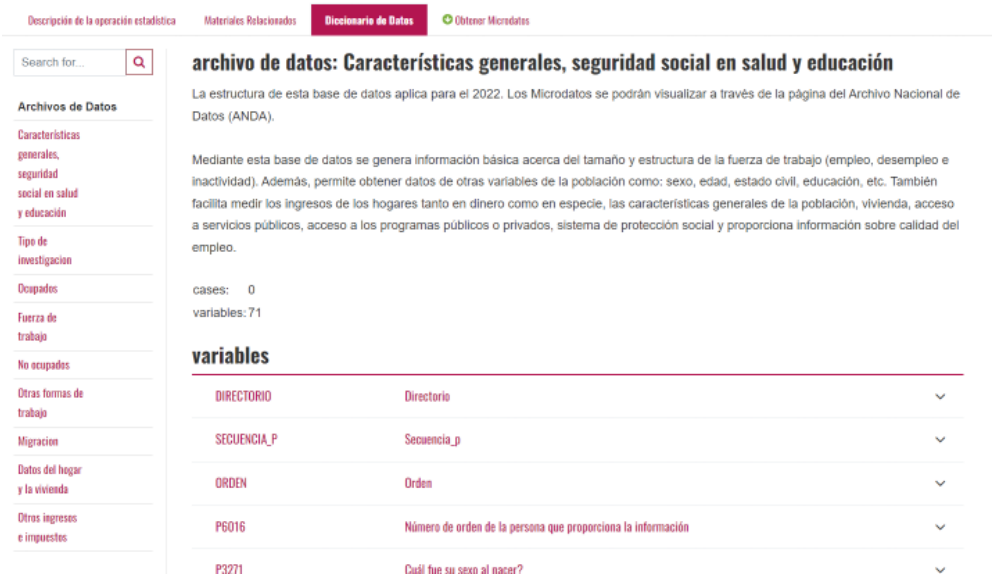
Ej. enero 2022

Nombre	Tipo	Tamaño
Ocupados	Archivo de valores...	11.545 KB

- La estructura establecida para cada archivo se encuentra definida en el diccionario de datos de la GEIH - 2022

Características generales, seguridad social en salud y educación

https://microdatos.dane.gov.co/index.php/catalog/771/data-dictionary/F54?file_name=Caracter%C3%ADsticas%20generales,%20seguridad%20social%20en%20salud%20y%20educaci%C3%B3n



Ocupados

https://microdatos.dane.gov.co/index.php/catalog/771/data-dictionary/F56?file_name=Ocupados

Por su parte, para descargar los datos de Bloomberg, se utilizó el laboratorio financiero de la Universidad EAFIT.

- Ingreso

- Identificación de variables

- Descarga

Mes	inflación	Desempleo	Salarios	Tasa de Interes a Empresas
ene-22	6,9	10,82	4,3	8,77
feb-22	8	9,11	3,8	9,88
mar-22	8,5	9,1	3,6	10,53
abr-22	9,2	10,45	3	11,4
may-22	9,1	10,82	5,8	12,15
jun-22	9,7	11,27	4,9	12,71
jul-22	10,2	11,73	3,6	14,25
ago-22	10,8	11,02	1	15,42
sep-22	11,4	11,05	-0,1	16,36
oct-22	13,1	12,64	-0,8	17,66
nov-22	12,5	12,72	-1,5	18,49
dic-22	13,1	14,83	-1,8	19,59

2. Herramientas utilizadas

En el desarrollo de este proyecto, se emplearon las plataformas colaborativas **Google Colab** y **GitHub** como piedras angulares para garantizar un entorno de trabajo eficiente y colaborativo. Google Colab ofreció una infraestructura en la nube que permitió ejecutar código de manera remota, facilitando así la compartición y colaboración en tiempo real entre los miembros del equipo. La colaboración en Colab minimizó los problemas de compatibilidad, ya que proporciona un entorno de ejecución uniforme sin requerir instalaciones locales específicas.

Por otro lado, GitHub se utilizó como repositorio central para gestionar el código fuente y demás entregables, asegurando un control de versiones definitivas. La combinación de estas herramientas no sólo simplificó la colaboración en el desarrollo del proyecto, sino que también posibilitó un seguimiento preciso de los cambios realizados, garantizando la integridad del código y promoviendo una gestión eficiente del flujo de trabajo. Esta elección estratégica de herramientas contribuyó significativamente a la agilidad y eficacia del equipo en el desarrollo y mantenimiento del proyecto.

Conclusiones generales del proyecto

A partir del estudio realizado, se pueden obtener dos grandes conclusiones:

1. En general, las variables seleccionadas son significativas, esto quiere decir que tienen influencia en la modalidad de ocupación de los hijos. Se destacan las variables relacionadas a la educación del individuo. Por otro lado, en este ejercicio identificamos que aunque las variables de los padres y el hogar tienen relevancia dadas las relaciones y herencias que se identifican en la literatura, son las variables de los individuos las que mayor importancia tienen. Estas variables se mantienen a lo largo de los cuatro modelos propuestos, por lo que adicionalmente, podemos concluir que existe movilidad intergeneracional ya que la modalidad de la ocupación no depende de las características o condiciones de sus antecesores.

2. Estas variables permitieron generar un modelo XGBoost el cual tuvo un F1 score cercano al 83%, esta métrica fue seleccionada ya que encontramos una proporción desbalanceada en la etiqueta. No obstante, la regresión logit también presenta resultados similares con un menor tiempo de cómputo.

Estos resultados fueron sometidos a diversos experimentos y métricas.

Finalmente, es importante destacar que la GEIH recopila variables relevantes para la creación, diseño de estrategias y políticas públicas que podrían reducir los niveles de informalidad existentes en Colombia. Si se logran diseñar estrategias efectivas para implementar en el país, no solo veremos un impacto positivo en la economía colombiana, sino, en la vida de los individuos y sus familias. La formalización del empleo garantiza

mejores condiciones laborales, mayor seguridad social, contribuyendo a una mayor productividad y crecimiento económico sostenible.

Los resultados anteriormente obtenidos, destacan la necesidad de implementar políticas y acciones específicas que aborden las variables sociales, económicas y geográficas relacionadas con la informalidad laboral y la movilidad intergeneracional en Colombia. Al enfocarnos en la promoción de empleos formales, la mejora en la educación y en las condiciones de los hogares, podremos avanzar hacia una sociedad con mayores posibilidades de movilidad ascendente para las generaciones futuras.

Referencias bibliográficas

Angulo, R., Azevedo, J. P., Gaviria, A., & Paez, G. N. (2012). Movilidad Social en Colombia. Recuperado de https://economia.uniandes.edu.co/sites/default/files/imagenes/eventos/Roberto-Angulo-Movilidad-Social-en-Colombia.pdf

Departamento Administrativo Nacional de Estadística. (11 de mayo de 2023). *Empleo Informal y Seguridad Social*. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social

Departamento Nacional de Planeación. (2017). La formalización del empleo en Colombia: diagnóstico, determinantes y recomendaciones de política.

Kropko, J. (2008). *Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data*. [Tesis for Master Degree, University of North Carolina]. https://doi.org/10.17615/wz24-qq92

Ministerio del Trabajo de Colombia. (2023). Formalización Laboral. Recuperado de https://www.mintrabajo.gov.co/empleo-y-pensiones/empleo/subdireccion-de-formalizacion-y-proteccion-del-empleo/formalizacion-laboral

Organización Internacional del Trabajo (2002). *El trabajo decente y la economía informal* [Informe IV]. Conferencia Internacional del Trabajo, Ginebra, Suiza. https://www.ilo.org/public/spanish/standards/relm/ilc/ilc90/pdf/rep-vi.pdf

Pérez, R. Q., Contreras, M. Y. y Hernández, K. C. (2014). *Determinantes de la informalidad laboral: un análisis para Colombia*. Investigación & desarrollo, *22*(1), 126-145. http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0121-32612014000100001

La República. (19 de agosto de 2022). *Colombia tiene una tasa de empleo informal de 53%, es de las más altas en el mundo*. https://www.larepublica.co/globoeconomia/colombia-tiene-una-tasa-de-empleo-informal-de-53-1-es-de-las-mas-altas-en-el-mundo-3428202

Lora, E. (2018). Informalidad laboral en Colombia: características y determinantes.

Semana. (22 de noviembre de 2016). *Las devastadoras consecuencias de la informalidad en Colombia*. https://www.semana.com/economia/articulo/los-problemas-y-las-consecuencias-de-la-informalidad-en-colombia/239203/

Shanahan, J. M., Elder, G. H. Jr., Miech, R. A., & Russo, F. M. (2007). The Influence of Parental Occupation and Educational Expectations on Vocational Aspirations in Adolescence.