Factores asociados al desempleo juvenil en la ciudad de Bogotá, D.C. mediante modelo de regresión logística binaria con GEIH del DANE

Yadir Perdomo Perdomo

Asesor

Esneider Dejesus Pineda Martínez

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

El desempleo juvenil en Colombia constituye un problema estructural que limita las oportunidades económicas y sociales de la población entre 15 y 28 años, especialmente en contextos urbanos como la ciudad de Bogotá, D.C. El presente trabajo tiene por objetivo analizar los factores determinantes asociados al desempleo juvenil mediante la aplicación del modelo de regresión logística binaria, utilizando microdatos oficiales del Departamento Administrativo Nacional de Estadística (DANE).

La metodología comprende un análisis exploratorio de datos y posteriormente, la estimación de un modelo de regresión logística, con el fin de identificar la asociación entre la condición de desempleo y variables sociodemográficas, educativas y territoriales, tales como (el sexo, el nivel educativo, la asistencia educativa, la pertenencia étnica y la zona de residencia). El enfoque adoptado es de carácter explicativo, priorizando la interpretación de los coeficientes y los odds ratios para comprender la incidencia relativa de cada factor sobre la probabilidad de desempleo juvenil.

Los resultados permiten identificar diferencias significativas en la probabilidad de desempleo según las características analizadas, destacando el papel de la educación y otros factores estructurales en la empleabilidad de los jóvenes. Finalmente, el estudio aporta evidencia empírica para el diseño y fortalecimiento de políticas públicas orientadas a la inclusión laboral juvenil y la reducción de las brechas de acceso al empleo formal en la ciudad de Bogotá, D.C.

***Palabras claves*:** Desempleo juvenil, Bogotá, D. C., regresión logística, DANE, GEIH.

Abstract

Youth unemployment in Colombia is a structural problem that limits the economic and social opportunities of the population aged 15 to 28, especially in urban contexts such as Bogotá, D.C. This study aims to analyze the determining factors associated with youth unemployment by applying a binary logistic regression model, using official microdata from the National Administrative Department of Statistics (DANE).

The methodology comprises an exploratory data analysis followed by the estimation of a logistic regression model to identify the association between unemployment status and sociodemographic, educational, and territorial variables, such as sex, educational level, school attendance, ethnicity, and area of residence. The approach adopted is explanatory, prioritizing the interpretation of coefficients and odds ratios to understand the relative impact of each factor on the probability of youth unemployment.

The results identify significant differences in the probability of unemployment according to the analyzed characteristics, highlighting the role of education and other structural factors in youth employability. Finally, the study provides empirical evidence for the design and strengthening of public policies aimed at youth labor inclusion and reducing gaps in access to formal employment in the city of Bogotá, D.C.

Keywords: Youth unemployment, Bogotá, D.C., logistic regression, DANE, GEIH.

Tabla de Contenido

Lista de Figuras

Lista de Apéndices

Introducción

En el contexto actual de transformación económica, social y tecnológica, el desempleo juvenil en la ciudad de Bogotá, D. C. se ha consolidado como una de las problemáticas más persistentes y complejas afectando de manera significativa a la población joven entre los 15 y 28 años de edad (Departamento Administrativo Nacional de Estadística [DANE], 2025). Sus efectos no solo limitan el desarrollo personal y profesional de los jóvenes, sino que también constituyen un obstáculo para el crecimiento económico equitativo y la cohesión social del país. De acuerdo con la literatura especializada y los reportes oficiales, entre los principales factores asociados al desempleo juvenil se encuentran la falta de experiencia laboral, las desigualdades de género y etnia, las brechas en el nivel educativo y la desconexión entre la formación académica y las demandas del mercado laboral, aspectos ampliamente documentados por el Departamento Administrativo Nacional de Estadística a través de la Gran Encuesta Integrada de Hogares (GEIH) (DANE, 2025), así como por organismos internacionales como la Organización Internacional del Trabajo (OIT, 2022).

El propósito de este trabajo es analizar estadísticamente los determinantes del desempleo juvenil en la ciudad de Bogotá, D. C., a partir de microdatos provenientes del Departamento Administrativo Nacional de Estadística (DANE). El enfoque metodológico se fundamenta en el análisis exploratorio de datos y en la aplicación de un modelo de regresión logística binaria, con el fin de identificar la relación entre la condición de desempleo juvenil y variables sociodemográficas, educativas y territoriales.

A lo largo del documento se examinan los principales factores asociados al desempleo juvenil, considerando variables como el nivel educativo, la asistencia educativa, el sexo, la pertenencia étnica y el tipo de área de residencia. El análisis permite caracterizar la situación

actual del desempleo juvenil e identificar los factores con mayor incidencia sobre la probabilidad de estar desempleado. Los hallazgos ofrecen evidencia empírica que puede contribuir al diseño de políticas públicas orientadas a fortalecer la inserción laboral de los jóvenes en Bogotá, D. C.

Justificación

La elección del tema factores asociados al desempleo juvenil en la ciudad de Bogotá, D. C., mediante un modelo de regresión logística binaria aplicado a microdatos de la Gran Encuesta Integrada de Hogares (GEIH) del Departamento Administrativo Nacional de Estadística (DANE), responde a la necesidad de comprender con mayor rigurosidad un fenómeno que afecta estructuralmente a una proporción significativa de la población joven entre los 15 y 28 años de edad (DANE, 2025). El desempleo juvenil restringe las oportunidades de desarrollo personal y profesional de los jóvenes y representa un desafío económico y social de alto impacto, en la medida en que incide negativamente sobre la productividad, la cohesión social y la sostenibilidad del crecimiento económico del país (Organización Internacional del Trabajo [OIT], 2022).

Pese a la relevancia del problema, persisten vacíos importantes en la identificación empírica de los factores que inciden en el desempleo juvenil desde un enfoque cuantitativo aplicado a microdatos oficiales. La literatura disponible se concentra, en su mayoría, en estudios de carácter descriptivo o en aproximaciones cualitativas a esta problemática (OIT, 2022). No obstante, mientras que existe una amplia producción descriptiva sobre el desempleo juvenil, son menos frecuentes los estudios que, mediante modelos estadísticos explicativos —como la regresión logística binaria—, estimen la probabilidad de que los jóvenes se encuentren en situación de desempleo a partir de variables sociodemográficas, educativas, laborales y territoriales. Esta carencia limita la capacidad de diseñar, focalizar y evaluar políticas públicas sustentadas en evidencia estadística robusta (DANE, 2025).

En este contexto, la presente investigación propone utilizar microdatos oficiales del DANE, combinando técnicas de análisis exploratorio de datos y regresión logística binaria, con el fin de identificar de manera objetiva los principales determinantes asociados al desempleo

juvenil y cuantificar la magnitud y dirección de su influencia sobre la condición laboral de los jóvenes, aportando un análisis con sentido económico y estadístico.

Adicionalmente, este trabajo presenta aportes relevantes en los ámbitos académico y de las políticas públicas. En el ámbito académico, contribuye a la literatura sobre desempleo juvenil mediante la aplicación de un enfoque cuantitativo explicativo basado en modelos de regresión logística binaria y el uso de microdatos oficiales de la GEIH (DANE, 2025). En el ámbito de las políticas públicas, los resultados constituyen un insumo empírico para el diseño, la focalización y la evaluación de programas orientados a la inclusión laboral juvenil, favoreciendo la formulación de estrategias más efectivas y equitativas para la reducción del desempleo juvenil en la ciudad de Bogotá, D. C.

Objetivos

Objetivo General

Analizar los factores asociados al desempleo juvenil en la ciudad de Bogotá, D. C. mediante modelos de regresión logística binaria con GEIH del DANE, con el fin de identificar los principales predictores asociadas a esta problemática.

Objetivos Específicos

Analizar las variables sociodemográficas, educativas y económicas asociadas al desempleo juvenil en la ciudad de Bogotá, D.C. mediante el procesamiento de microdatos del DANE, para comprender los principales factores que influyen en la empleabilidad de los jóvenes.

Evaluar el desempeño del modelo de regresión logística binaria para clasificar la condición laboral de los jóvenes en la ciudad de Bogotá, D.C. a partir de métricas de ajuste y capacidad predictiva (AUC y pseudo R^2), e identificar las variables con mayor influencia sobre el desempleo juvenil.

Comparar el desempleo juvenil en la ciudad de Bogotá, D.C. considerando las variables educación, sexo, experiencia laboral y tipo de área, utilizando análisis exploratorio y segmentación de datos, con el fin de establecer comparaciones entre los diferentes grupos poblacionales.

Proponer estrategias y recomendaciones basadas en evidencia estadística, derivadas del análisis de datos del DANE, para fortalecer al diseño de políticas públicas orientadas a la inclusión laboral de los jóvenes en la ciudad de Bogotá, D.C.

Marco de Referencia

Delimitación y Alcance del Estudio

El presente estudio se delimita al análisis del desempleo juvenil en la ciudad de Bogotá, D. C., considerando a la población joven entre los 15 y 28 años de edad, de acuerdo con la definición de juventud establecida en la normativa vigente y utilizada por el Departamento Administrativo Nacional de Estadística (DANE). La investigación emplea microdatos provenientes de la Gran Encuesta Integrada de Hogares (GEIH), correspondientes al año 2025, los cuales constituyen la principal fuente oficial para el análisis del mercado laboral en Colombia.

Las variables seleccionadas incluyen características sociodemográficas, educativas y territoriales, tales como el sexo, la edad, el nivel educativo, la asistencia educativa, la pertenencia étnica y el tipo de área de residencia, las cuales han sido identificadas en la literatura especializada como factores relevantes en la explicación del desempleo juvenil. El análisis combina una fase descriptiva orientada a caracterizar la situación del desempleo juvenil en Bogotá, D. C., y una fase explicativa basada en la aplicación de un modelo de regresión logística binaria, con el fin de estimar la probabilidad de que un joven se encuentre en situación de desempleo en función de las variables consideradas.

En este sentido, el estudio adopta un enfoque cuantitativo de alcance explicativo, orientado a generar evidencia empírica que contribuya a una mejor comprensión de los determinantes del desempleo juvenil y a la formulación de políticas públicas basadas en información estadística confiable.

Marco Teórico

Desempleo Juvenil

El desempleo juvenil se define como la condición en la cual las personas entre los 15 y 28 años, que hacen parte de la fuerza laboral no logran acceder a un empleo remunerado pese a encontrarse disponibles y en búsqueda activa de trabajo (Departamento Administrativo Nacional de Estadística [DANE], 2025) este fenómeno constituye uno de los principales retos del mercado laboral colombiano, debido a su persistencia y a su impacto en el desarrollo económico y social del país (DANE, 2025).

El desempleo juvenil presenta características estructurales asociadas a la transición de la educación al mercado laboral, la falta de experiencia, la informalidad y la segmentación del empleo. Estas condiciones generan mayores barreras de acceso al trabajo formal en comparación con otros grupos etarios, lo que incrementa la vulnerabilidad económica de los jóvenes (Organización Internacional del Trabajo [OIT], 2022).

Enfoques Teóricos del Desempleo Juvenil

Desde la teoría económica, el desempleo juvenil puede analizarse a partir de distintos enfoques. El enfoque neoclásico sostiene que el desempleo es resultado de desajustes entre la oferta y la demanda laboral, especialmente en términos de habilidades y salarios. En el caso de los jóvenes, estos desajustes suelen estar relacionados con la falta de experiencia laboral y la insuficiente adecuación entre la formación educativa y las necesidades del mercado de trabajo (OIT, 2022).

Por su parte, el enfoque estructuralista plantea que el desempleo juvenil responde a factores estructurales como la segmentación del mercado laboral, las desigualdades socioeconómicas, las brechas de género y etnia, y la concentración del empleo en sectores de

baja productividad. Este enfoque resulta particularmente relevante para el contexto colombiano, donde dichas desigualdades condicionan las oportunidades laborales de los jóvenes (OIT, 2022).

Factores Sociodemográficos Asociados al Desempleo Juvenil

Diversos estudios han evidenciado que variables sociodemográficas como el sexo, la edad, la zona de residencia y la pertenencia étnica se asocian de manera significativa en la probabilidad de desempleo juvenil. En particular, las mujeres jóvenes y los grupos étnicos históricamente excluidos enfrentan mayores dificultades para acceder a empleos estables y formales, debido a barreras estructurales y discriminación en el mercado laboral (OIT, 2022; DANE, 2025).

Asimismo, el contexto territorial, especialmente la diferenciación entre zonas urbanas y rurales, incide en la disponibilidad de oportunidades laborales y en la calidad del empleo, lo que repercute directamente en los niveles de desempleo juvenil (DANE, 2025).

Educación y Desempleo Juvenil

La educación constituye uno de los principales determinantes del empleo juvenil. En términos teóricos, el capital humano, propuesto por Becker, sostiene que mayores niveles de educación incrementan la productividad individual y, por ende, la probabilidad de inserción laboral. Sin embargo, en contextos como el colombiano, este vínculo no siempre es lineal.

El aumento de jóvenes con educación media y superior no ha sido acompañado de una expansión proporcional del empleo formal, lo que genera fenómenos como la sobre educación y el desempleo entre jóvenes calificados. Esto evidencia una desconexión entre el sistema educativo y las demandas reales del mercado laboral, situación que ha sido señalada en análisis basados en microdatos del DANE (DANE, 2025).

Fuerza Laboral, Ocupación y Desempleo

Desde la perspectiva estadística, el DANE clasifica a la población en edad de trabajar en ocupados, desocupados e inactivos, siendo los dos primeros grupos los que conforman la fuerza laboral. Para el análisis del desempleo juvenil, resulta fundamental diferenciar a los jóvenes que participan activamente en el mercado laboral de aquellos que se encuentran inactivos, ya que estos últimos no enfrentan directamente la condición de desempleo (DANE, 2025).

Esta distinción permite una medición más precisa del fenómeno y evita sesgos en el análisis estadístico, especialmente cuando se emplean modelos de clasificación lo que mejora la precisión conceptual del fenómeno.

Modelos Estadísticos para el Análisis del Desempleo Juvenil

El análisis del desempleo juvenil ha incorporado de manera creciente el uso de modelos estadísticos, entre los cuales destaca la regresión logística binaria. Este modelo es especialmente adecuado cuando la variable dependiente es dicotómica, como ocurre con la condición laboral (desempleado u ocupado) (Hosmer, Lemeshow & Sturdivant, 2013).

La regresión logística permite estimar la probabilidad de que un joven se encuentre desempleado en función de un conjunto de variables explicativas, así como identificar la dirección e intensidad de la relación entre estas variables y la condición de desempleo, mediante la interpretación de razones de odds ratios en términos asociativos, su uso resulta pertinente en estudios basados en microdatos, como los provenientes del DANE, debido a su capacidad para manejar variables categóricas y su robustez en análisis socioeconómicos (Hosmer et al., 2013).

Importancia del Análisis Estadístico para la Política Pública

El uso de modelos estadísticos aplicados al desempleo juvenil proporciona evidencia empírica fundamental para el diseño y evaluación de políticas públicas. Identificar los factores

que incrementan o reducen la probabilidad de desempleo permite orientar intervenciones más focalizadas, eficientes y equitativas y contextualizadas (OIT, 2022).

En este sentido, el análisis basado en datos oficiales y técnicas estadísticas contribuye a una comprensión más profunda del fenómeno, superando enfoques meramente descriptivos y fortaleciendo la toma de decisiones basada en evidencia.

Marco Conceptual

Desempleo Juvenil

El desempleo juvenil se define como la situación en la cual una persona entre 15 y 28 años que hace parte de la fuerza laboral no se encuentra ocupada y ha realizado acciones de búsqueda activa de empleo durante el período de referencia. Este concepto excluye a los jóvenes inactivos y constituye la variable central de análisis del estudio de acuerdo con la definición oficial utilizada por el Departamento Administrativo Nacional de Estadística (DANE, 2025).

En el presente trabajo, el desempleo juvenil se operacionaliza mediante una variable dicotómica, donde el valor uno (1) indica condición de desempleo y el valor cero (0) corresponde a jóvenes ocupados, siguiendo los lineamientos metodológicos de la Gran Encuesta Integrada de Hogares (GEIH) (DANE, 2025).

Fuerza Laboral

La fuerza laboral está conformada por el conjunto de personas en edad de trabajar que se encuentran ocupadas o desocupadas. Este concepto es fundamental para el análisis del desempleo, ya que delimita la población efectivamente expuesta al riesgo de estar desempleada (DANE, 2025).

En el estudio, la fuerza laboral se construye a partir de las variables del DANE relacionadas con la condición de actividad, excluyendo explícitamente a los jóvenes inactivos.

Ocupado

Se considera ocupado a todo joven que durante el período de referencia realizó al menos una hora de actividad remunerada, trabajó en un negocio propio o familiar o tuvo un empleo del cual estuvo temporalmente ausente. Esta categoría corresponde a uno de los estados posibles

dentro de la fuerza laboral, conforme a la metodología oficial del DANE y las recomendaciones de la Organización Internacional del Trabajo (DANE, 2025; OIT, 2022).

En términos operativos, el estado de ocupación se representa mediante una variable binaria, donde el valor uno (1) indica que el joven se encuentra ocupado y el valor cero (0) lo contrario.

Inactividad

La inactividad hace referencia a la condición de aquellas personas que no se encuentran ocupadas ni desocupadas, ya sea porque no están disponibles para trabajar o no buscan empleo activamente. Este grupo incluye, entre otros, estudiantes que no buscan empleo, personas dedicadas exclusivamente a labores del hogar o individuos incapacitados para trabajar (DANE, 2025).

En este estudio, los jóvenes inactivos no son considerados en el análisis estadístico, con el fin de garantizar coherencia económica y metodológica en la medición del desempleo juvenil.

Nivel Educativo

El nivel educativo corresponde al grado máximo de formación académica alcanzado por el individuo. Este concepto es clave en el análisis del desempleo juvenil, dado su vínculo con el capital humano y la empleabilidad (Becker, 1964).

Para efectos del modelo estadístico, el nivel educativo se agrupa en categorías analíticas como educación baja, educación técnica, educación universitaria y educación de posgrado, las cuales se incorporan como variables explicativas en la regresión logística, de acuerdo con la información reportada en la GEIH (DANE, 2025).

Asistencia Educativa

La asistencia educativa se refiere a la condición de estar actualmente matriculado o asistiendo a una institución educativa formal. Este concepto permite analizar la relación entre la permanencia en el sistema educativo y la participación en el mercado laboral (DANE, 2025).

En el modelo, esta variable se incorpora como un indicador binario, donde el valor uno (1) representa que el joven asiste a una institución educativa y el valor cero (0) lo contrario.

Sexo

El sexo se define como la condición biológica del individuo, clasificada en hombre y mujer según los registros del DANE. Esta variable permite analizar posibles brechas de género en el desempleo juvenil y su impacto en la probabilidad de desocupación (DANE, 2025; OIT, 2022).

En el análisis estadístico, el sexo se incorpora mediante variables dicotómicas tomando como categoría de referencia a los hombres.

Zona de Residencia

La zona de residencia distingue entre áreas urbanas y rurales, lo cual resulta relevante para analizar diferencias territoriales en el acceso al empleo. Este concepto refleja desigualdades estructurales en la distribución del empleo (DANE, 2025).

En el modelo, la zona de residencia se operacionaliza como una variable binaria, donde el valor uno (1) representa residencia en zona rural.

Grupo Étnico

El grupo étnico hace referencia a la autoidentificación del individuo con un grupo étnico específico, como población indígena, afrodescendiente, raizal o palenquera, o la no pertenencia a

ninguno de estos grupos. Esta variable permite analizar brechas étnicas en el mercado laboral juvenil (DANE, 2025; OIT, 2022).

Para el análisis, esta variable se agrupa en una categoría dicotómica que diferencia a jóvenes pertenecientes a grupos étnicos de aquellos que no lo son, permitiendo evaluar la existencia de brechas étnicas en el desempleo juvenil.

Edad

La edad corresponde al número de años cumplidos por el individuo al momento de la encuesta. En el análisis del desempleo juvenil, la edad permite capturar diferencias en experiencia laboral, capital humano acumulado y transición escuela-trabajo. Diversos estudios evidencian que los jóvenes de menor edad enfrentan mayores dificultades de inserción laboral debido a la falta de experiencia previa (OIT, 2022).

En el modelo econométrico, la edad se incluye como variable cuantitativa continua y adicionalmente se incorpora su término cuadrático con el fin de capturar posibles efectos no lineales sobre la probabilidad de desempleo.

Alfabetismo

El alfabetismo se define como la capacidad del individuo para leer y escribir un texto sencillo, según autorreporte en la encuesta. Esta variable constituye un indicador básico de capital humano y competencias fundamentales para la inserción en el mercado laboral.

En el modelo se incorpora como variable dicotómica (1 = sabe leer y escribir; 0 = no sabe), permitiendo evaluar si la carencia de habilidades básicas incrementa la probabilidad de desempleo juvenil.

Condición de Campesino

La condición de campesino identifica a los individuos que se autorreconocen como pertenecientes a población campesina. Esta variable es relevante en el análisis del mercado laboral debido a las diferencias estructurales en acceso a empleo formal, educación y oportunidades productivas entre población rural campesina y no campesina.

En el modelo econométrico se incluye como variable dicotómica para evaluar si la pertenencia a población campesina influye en la probabilidad de desempleo juvenil.

Mes

El mes corresponde al período en el cual se realizó la medición de la encuesta. Esta variable permite capturar posibles variaciones estacionales en el desempleo juvenil, asociadas a ciclos económicos, periodos académicos o dinámicas temporales del mercado laboral.

Aunque no fue incluida en el modelo principal, se utiliza en el análisis descriptivo para evaluar fluctuaciones mensuales en la tasa de desempleo.

Regresión Logística Binaria

La regresión logística binaria es un modelo estadístico utilizado para analizar la relación entre una variable dependiente dicotómica y un conjunto de variables explicativas. Este modelo estima la probabilidad de ocurrencia de un evento, en este caso, la condición de desempleo juvenil (Hosmer, Lemeshow & Sturdivant, 2013).

La interpretación de los resultados se realiza a través de los odds ratios, los cuales indican el cambio relativo en la probabilidad de desempleo asociado a cada variable explicativa, manteniendo constantes las demás variables del modelo (Hosmer et al., 2013).

Estas variables se fundamentan en la teoría del capital humano y en enfoques de segmentación del mercado laboral, los cuales plantean que características sociodemográficas y educativas influyen en la probabilidad de desempleo.

Metodología

Enfoque de la Investigación

El presente estudio adopta un enfoque cuantitativo, con un diseño no experimental y de tipo explicativo, dado que se analizan relaciones estadísticas entre variables sin manipulación deliberada de las mismas. El análisis se fundamenta en el uso de regresión logística binaria, con el fin de identificar los factores asociados a la condición de desempleo juvenil en la ciudad de Bogotá, D. C**.**

Fuente de Datos

Se utilizaron microdatos oficiales del Departamento Administrativo Nacional de Estadística (DANE), correspondientes a los módulos de fuerza de trabajo, ocupados, no ocupados y educación/salud de la Gran Encuesta Integrada de Hogares (GEIH), para el período comprendido entre enero y septiembre de 2025. El conjunto de datos empleado, así como su estructura general, se presentan en la Figura 1 del Anexo, la cual ilustra la base de datos utilizada para el análisis. Los microdatos fueron procesados mediante herramientas de análisis en Python, utilizando los entornos Google Colab y Jupyter Notebook.

Preparación, Procesamiento y Limpieza de Datos

En esta fase se llevó a cabo la construcción del conjunto de datos analítico. Inicialmente, se importaron y consolidaron nueve archivos mensuales correspondientes al período de estudio, los cuales conforman la base de microdatos utilizada (véase Figura 2 en los Anexos). Posteriormente, se eliminaron variables que no aportaban al análisis y se estandarizaron los nombres de las columnas relevantes, con el fin de garantizar consistencia en la estructura del conjunto de datos.

A continuación, la información fue depurada mediante la selección exclusiva de los registros correspondientes a la ciudad de Bogotá, D. C., y a la población joven entre 15 y 28 años de edad, de acuerdo con la definición oficial de juventud en Colombia. Como resultado de este proceso, los archivos mensuales fueron integrados en un único conjunto de datos, cuyo tamaño final y estructura se presentan en la Figura 3 de los Anexos.

Posteriormente, se integraron los diferentes módulos del DANE a nivel persona, hogar mes, lo que permitió construir una base de datos coherente y consistente para el análisis estadístico. En esta etapa se definieron y construyeron las principales variables laborales del estudio, específicamente Fuerza_Laboral, Ocupado y Desempleado_Final, siguiendo los lineamientos conceptuales establecidos por el DANE y la Organización Internacional del Trabajo (OIT). El proceso de construcción de la fuerza laboral juvenil se ilustra en la (figura 1).

Finalmente, se realizó una validación lógica y estadística para garantizar la coherencia interna de las variables creadas, y se definió la variable Desempleado_Final como la variable dependiente del modelo de regresión logística binaria.

Análisis Exploratorio y Selección de Variables

En esta etapa se desarrolló un análisis exploratorio de los datos con el propósito de examinar la distribución del desempleo juvenil y analizar su comportamiento en relación con variables sociodemográficas, educativas y territoriales. Para ello, se emplearon estadísticas descriptivas, tales como tablas de frecuencia, proporciones y comparaciones entre grupos poblacionales, las cuales permitieron identificar patrones y diferencias relevantes en la condición laboral de los jóvenes.

La variable objetivo fue verificada y codificada de manera dicotómica, donde el valor **1** corresponde a jóvenes desempleados que hacen parte de la fuerza laboral y no se encuentran

ocupados, y el valor **0** representa a jóvenes ocupados. Esta definición garantiza coherencia conceptual con los lineamientos del DANE y la Organización Internacional del Trabajo (OIT).

Los resultados del análisis exploratorio evidenciaron diferencias en los niveles de desempleo juvenil según sexo, nivel educativo y condición de asistencia educativa, aspectos que se presentan de manera descriptiva en las (figuras 11,12,13)**.** Estas evidencias permitieron identificar las variables con mayor potencial explicativo, las cuales fueron seleccionadas para su inclusión en el modelo de regresión logística binaria.

Modelamiento Estadístico Regresión Logística Binaria

En esta etapa se estimó un modelo de regresión logística binaria, el cual resulta apropiado dado que la variable dependiente es de naturaleza dicotómica y su distribución se presenta en la (figura 8). El objetivo del modelo fue estimar la probabilidad de que un joven se encuentre en situación de desempleo en función de un conjunto de características individuales, educativas y contextuales, manteniendo constantes las demás variables incluidas en el análisis.

La variable dependiente del modelo corresponde a Desempleado_Correcto, definida de manera binaria, donde el valor **1** indica que el joven se encuentra desempleado y el valor **0** que se encuentra ocupado. Esta definición es consistente con los criterios oficiales del DANE y la Organización Internacional del Trabajo (OIT).

Las variables explicativas se agruparon en tres dimensiones analíticas.

En primer lugar, se incluyeron variables sociodemográficas, como el sexo, la edad y la zona de residencia (urbana o rural). Adicionalmente, se incorporó la edad al cuadrado con el fin de capturar posibles efectos no lineales asociados al ciclo de vida dentro del grupo juvenil.

En segundo lugar, se consideraron variables relacionadas con el capital humano, entre ellas el nivel educativo representado mediante variables dummy para educación universitaria, técnica, posgrado y baja escolaridad, la condición de asistencia educativa y el alfabetismo.

Finalmente, se incluyeron variables de contexto social y territorial, como la pertenencia a un grupo étnico y la condición de campesino, con el propósito de analizar posibles brechas estructurales en el mercado laboral juvenil.

Las variables categóricas fueron codificadas mediante variables dicotómicas (dummy), seleccionando categorías de referencia con base en su mayor frecuencia y relevancia analítica. Asimismo, se excluyeron aquellas variables que definen directamente la condición de desempleo, con el fin de evitar problemas de endogeneidad y garantizar la validez estadística del modelo.

La interpretación de los resultados se realizó a través de los odds ratios, los cuales permiten evaluar el efecto relativo de cada variable explicativa sobre la probabilidad de que un joven se encuentre desempleado, manteniendo constantes las demás variables del modelo.

Formulación y Contraste de Hipótesis

Hipótesis Nula (H_0)

Las variables sociodemográficas y educativas incluidas en el modelo no influyen significativamente en la probabilidad de desempleo juvenil en Bogotá, D.C.

Hipótesis Alternativa (H_1):

Al menos una de las variables sociodemográficas o educativas incluidas en el modelo influye significativamente en la probabilidad de desempleo juvenil.

Hipótesis Específicas por Variable Clave

Educación

H_{01}: El nivel educativo no tiene efecto significativo sobre la probabilidad de desempleo juvenil.

H_{11}: El nivel educativo tiene efecto significativo sobre la probabilidad de desempleo juvenil.

Sexo

H_{02}: El sexo no influye significativamente en la probabilidad de desempleo juvenil.

H_{12}: El sexo influye significativamente en la probabilidad de desempleo juvenil.

Condición Campesina

H_{03}: La condición de campesino no influye significativamente en la probabilidad de desempleo juvenil.

H_{13}: La condición de campesino influye significativamente en la probabilidad de desempleo juvenil.

Edad

H_{04}: La edad no influye significativamente en la probabilidad de desempleo juvenil.

H_{14}: La edad influye significativamente en la probabilidad de desempleo juvenil.

Nota**:**

El contraste de hipótesis se realizó utilizando el p-valor asociado al estadístico z del modelo logístico. Se adoptó un nivel de significancia del 5 % ($\alpha = 0,05$). Si el p-valor es menor a 0,05, se rechaza la hipótesis nula; en caso contrario, no se rechaza.

Los resultados muestran que las variables asociadas al nivel educativo (universitaria, técnica y posgrado) presentan p-valores inferiores a 0,05, por lo que se rechaza la hipótesis nula

correspondiente, evidenciando un efecto estadísticamente significativo sobre la probabilidad de desempleo juvenil. De igual forma, la variable ES_CAMPESINO presenta alta significancia estadística ($p < 0{,}001$), rechazándose la hipótesis nula. En contraste, las variables sexo, zona rural y pertenencia étnica no presentan significancia estadística, por lo que no se rechaza la hipótesis nula en estos casos.

Validación, Discusión y Conclusiones

Finalmente, se evaluó el desempeño del modelo de regresión logística binaria mediante métricas de ajuste y capacidad predictiva, tales como el pseudo R^2 y el área bajo la curva ROC (AUC), con el fin de verificar su adecuación para el análisis del desempleo juvenil. Estas métricas permitieron evaluar la capacidad del modelo para discriminar entre jóvenes ocupados y desempleados, así como la consistencia de los resultados obtenidos.

Posteriormente, los hallazgos derivados del modelo fueron analizados e interpretados a la luz de la literatura existente sobre desempleo juvenil, contrastando los resultados empíricos con los enfoques teóricos y los estudios previos desarrollados en el contexto colombiano e internacional. Este ejercicio permitió identificar coincidencias y divergencias, así como fortalecer la interpretación económica y social de los resultados.

Finalmente, esta etapa condujo a la formulación de conclusiones directamente alineadas con los objetivos del estudio y sustentadas en la evidencia estadística obtenida. A partir de dichas conclusiones, se plantearon recomendaciones orientadas al fortalecimiento y focalización de políticas públicas de inclusión laboral juvenil en la ciudad de Bogotá, D. C., con énfasis en los grupos poblacionales con mayor probabilidad de desempleo.

La capacidad predictiva del modelo se evaluó mediante la curva ROC y el área bajo la curva (AUC), cuyos resultados se presentan en la Figura 4 de los Anexos.

Lista de Figuras

Figura 1

Carga de Microdatos de la GEIH del DANE en el Entorno Google Colab

```
from google.colab import files
uploaded = files.upload()
```

```
Elegir archivos Ningún archivo seleccionado Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving cg_salud_educacion_abril.xlsx.csv to cg_salud_educacion_abril.xlsx.csv
Saving cg_salud_educacion_agosto.xlsx.csv to cg_salud_educacion_agosto.xlsx.csv
Saving cg_salud_educacion_enero.xlsx to cg_salud_educacion_enero.xlsx
Saving cg_salud_educacion_febrero.xlsx to cg_salud_educacion_febrero.xlsx
Saving cg_salud_educacion_julio.xlsx.csv to cg_salud_educacion_julio.xlsx.csv
Saving cg_salud_educacion_junio.xlsx.csv to cg_salud_educacion_junio.xlsx.csv
Saving cg_salud_educacion_marzo.xlsx.csv to cg_salud_educacion_marzo.xlsx.csv
Saving cg_salud_educacion_mayo.xlsx.csv to cg_salud_educacion_mayo.xlsx.csv
Saving cg_salud_educacion_septiembre.xlsx.csv to cg_salud_educacion_septiembre.xlsx.csv
```

Figura 2

Base de Datos Utilizada para el Análisis del Desempleo Juvenil

[6]:

	PERIODO	MES	AÑO	ID_VIVIENDA	ID_HOGAR	ID_PERSONA	ZONA	FACTOR_EXPANSION	DEPARTAMENTO	SEXO	...	GRUPO_ETNICO	ESTADO_CIVIL	AFILICI
0	20250101	Enero	2025	8088007	1	1	Urbano	37077376421	Bogotá D.C.	Hombre	...	Ninguno de los anteriores	No esta casado (a) y vive en pareja hace dos a...	
1	20250101	Enero	2025	8088007	1	2	Urbano	37077376421	Bogotá D.C.	Mujer	...	Ninguno de los anteriores	No esta casado (a) y vive en pareja hace dos a...	
2	20250101	Enero	2025	8088009	1	1	Urbano	31991366202	Bogotá D.C.	Mujer	...	Ninguno de los anteriores	Esta casado (a)	
3	20250101	Enero	2025	8088009	1	2	Urbano	31991366202	Bogotá D.C.	Hombre	...	Ninguno de los anteriores	Esta soltero (a)	
4	20250101	Enero	2025	8088010	1	1	Urbano	2321090657	Bogotá D.C.	Hombre	...	Ninguno de los anteriores	No esta casado (a) y vive en pareja hace dos a...	

5 rows × 25 columns

Nota. Microdatos de la Gran Encuesta Integrada de Hogares (GEIH) del DANE, período enero septiembre de 2025

Figura 3

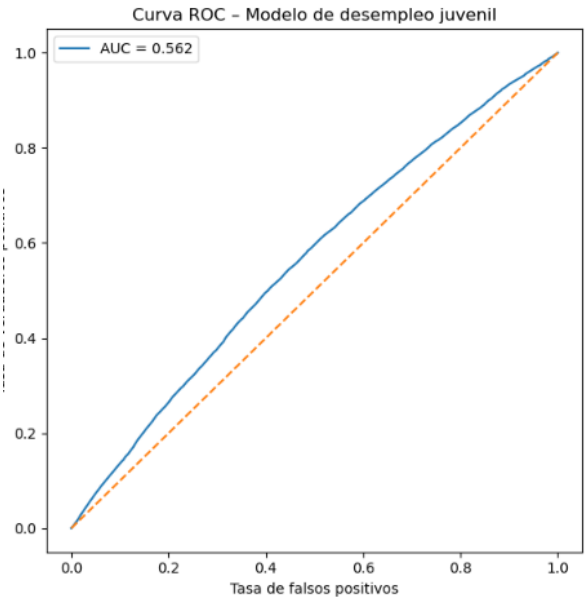
Dimensión del Conjunto de Datos

Descripción	Cantidad
Total personas	1531547
Jóvenes (15-28)	247474

*Nota***.** Tamaño del dataset final tras los procesos de depuración y selección de variables.

Figura 4

Curva ROC y Área Bajo la Curva (AUC)



Nota. Curva ROC del modelo estimado para el desempleo juvenil

Interpretación

La curva ROC muestra un AUC de 0.562, lo que indica que el modelo presenta una capacidad de discriminación ligeramente superior al azar en la clasificación del desempleo juvenil. Aunque el poder predictivo es limitado, la curva se mantiene por encima de la línea de

referencia aleatoria, lo que evidencia que las variables incluidas aportan información estadísticamente relevante para distinguir entre jóvenes empleados y desempleados.

Este resultado es consistente con el carácter explicativo del estudio, cuyo objetivo principal no es maximizar la predicción, sino identificar factores asociados al desempleo juvenil en Bogotá, D.C.

Figura 5

Construcción de la Fuerza Laboral Juvenil

	FUERZA_LABORAL	Cantidad
0	1	163795
1	0	83679

*Nota***.** Clasificación de jóvenes en ocupados y desocupados, excluyendo inactivos

Figura 6

Distribución de Jóvenes Desempleados en la Ciudad de Bogotá, D. C.

Indicador	Valor
Desempleados	15398
Fuerza laboral (PEA)	163795
Tasa de desempleo (%)	9.4

Nota. Número de jóvenes en condición de desempleo según microdatos de la GEIH del DANE.

Figura 7

Desempleo Juvenil Según Condición de Asistencia Educativa

Asistencia educativa	Tasa desempleo (%)
No	8.24
Si	10.53

Nota. Distribución del desempleo juvenil según asistencia a una institución educativa.

Figura 8

Distribución De La Variable Dependiente Desempleado Correcto

DESEMPLEADO_CORRECTO	Cantidad
0	232076
1	15398

Nota. Distribución de jóvenes ocupados y desempleados en la muestra analizada.

Resultados

Caracterización de la Población Juvenil en Bogotá, D.C.

En esta sección se presentan los resultados descriptivos de la población objeto de estudio, conformada por jóvenes entre 15 y 28 años residentes en la ciudad de Bogotá, D.C. a partir de los microdatos de la Gran Encuesta Integrada de Hogares (GEIH) del DANE. El propósito de esta caracterización es contextualizar el análisis posterior del desempleo juvenil y establecer un panorama general de las principales variables sociodemográficas y educativas incluidas en el estudio.

La base de datos final consolidada está compuesta por 247.474 observaciones, correspondientes a jóvenes que participaron en la encuesta durante el período comprendido entre los meses de enero y septiembre del año 2025. Del total de la población analizada, se observa una distribución equilibrada por sexo, con una participación similar de hombres y mujeres, lo que permite realizar comparaciones sin sesgos significativos asociados al tamaño muestral.

En cuanto al contexto territorial, la mayoría de los jóvenes residen en zonas urbanas, lo cual es consistente con la estructura demográfica de la ciudad de Bogotá, D.C. y con la cobertura de la GEIH en áreas metropolitanas. Esta característica resulta relevante, dado que las dinámicas del mercado laboral difieren sustancialmente entre zonas urbanas y rurales.

Respecto al nivel educativo, se evidencia que una proporción considerable de los jóvenes cuenta con educación media y universitaria, mientras que una fracción menor presenta niveles educativos bajos o inexistentes. Esta heterogeneidad educativa permite analizar el papel del capital humano como factor determinante del desempleo juvenil. Adicionalmente, una parte significativa de la población se encuentra asistiendo actualmente a una institución educativa, lo

que puede influir tanto en la participación en el mercado laboral como en la condición de empleo.

Finalmente, se observa que la mayoría de los jóvenes se identifican como alfabetizados y no pertenecen a grupos étnicos minoritarios, aunque estos últimos se incluyen en el análisis con el fin de evaluar posibles brechas asociadas a condiciones étnicas.

En conjunto, esta caracterización permite establecer una base sólida para el análisis exploratorio y econométrico del desempleo juvenil en Bogotá, D.C. facilitando la interpretación de los resultados obtenidos en las siguientes secciones.

Figura 9

Distribución de Características Sociodemográficas y Educativas

	Variable	Categoría	Frecuencia	Porcentaje (%)
19	ALFABETISMO	Sí	162978	99.50
18	ALFABETISMO	No	817	0.50
16	ASISTENCIA_EDUCATIVA	No	96235	58.75
17	ASISTENCIA_EDUCATIVA	Sí	67560	41.25
22	GRUPO_ETNICO	Ninguno de los anteriores	158966	97.05
20	GRUPO_ETNICO	\tNegro (a), mulato (a)	2581	1.58
21	GRUPO_ETNICO	Indígena	1716	1.05
24	GRUPO_ETNICO	Raizal del archipiélago de San Andrés	321	0.20
23	GRUPO_ETNICO	Palenquero (a)	211	0.13
15	NIVEL_EDUCATIVO	Universitaria	65208	39.81
9	NIVEL_EDUCATIVO	Media académica (Bachillerato clásico)	55717	34.02
14	NIVEL_EDUCATIVO	Técnica profesional	11226	6.85
5	NIVEL_EDUCATIVO	Básica secundaria (6o - 9o)	9228	5.63
13	NIVEL_EDUCATIVO	Tecnológica	9119	5.57
7	NIVEL_EDUCATIVO	Especialización	4456	2.72
8	NIVEL_EDUCATIVO	Maestría	3206	1.96
10	NIVEL_EDUCATIVO	Media técnica (Bachillerato técnico)	3011	1.84
4	NIVEL_EDUCATIVO	Básica primaria (1o - 5o)	1917	1.17
11	NIVEL_EDUCATIVO	Ninguno	451	0.28
6	NIVEL_EDUCATIVO	Doctorado	160	0.10
12	NIVEL_EDUCATIVO	Preescolar	96	0.06
1	SEXO	Mujer	82230	50.20
0	SEXO	Hombre	81565	49.80
3	ZONA	Urbano	162099	98.96
2	ZONA	Rural	1696	1.04

Interpretación

La figura 9 presenta la caracterización de los jóvenes pertenecientes a la población económicamente activa (PEA) incluida en el estudio. Se observa una distribución equilibrada por sexo, con ligera mayoría femenina, y una alta concentración en zona urbana, coherente con la estructura demográfica de Bogotá, D.C. En términos educativos, predominan los niveles universitarios y de media académica, lo que refleja un importante componente de capital humano dentro de la población juvenil activa.

Asimismo, una proporción significativa de jóvenes continúa asistiendo a una institución educativa, lo que evidencia procesos de transición entre el sistema educativo y el mercado laboral. El alto nivel de alfabetismo registrado fortalece la pertinencia del análisis, ya que garantiza condiciones básicas de formación. En conjunto, esta caracterización descriptiva contextualiza el análisis econométrico posterior sobre los determinantes del desempleo juvenil.

Análisis Descriptivo del Desempleo Juvenil en Bogotá, D.C.

En esta sección se analiza la situación laboral de los jóvenes en Bogotá, D.C., considerando únicamente a la población en la fuerza laboral (PEA), de acuerdo con las definiciones del DANE y la OIT. Se excluyen los jóvenes inactivos, con el fin de obtener una medición consistente del desempleo juvenil.

Figura 10

Distribución de la Condición Laboral de los Jóvenes en la Fuerza Laboral

Indicador	Valor
Desempleados	15398
Fuerza laboral (PEA)	163795
Tasa de desempleo (%)	9.4

Interpretación

Los resultados muestran que la tasa de desempleo juvenil se mantiene relativamente similar entre hombres y mujeres, con una ligera mayor incidencia en las mujeres. Aunque históricamente se han documentado brechas de género en el mercado laboral, en este caso las diferencias no son ampliamente marcadas, lo que sugiere que el desempleo juvenil responde principalmente a factores estructurales que afectan de manera transversal a la población joven. No obstante, la leve mayor vulnerabilidad femenina podría estar asociada a barreras de inserción laboral, segmentación ocupacional o condiciones diferenciales en el acceso a oportunidades de empleo.

Figura 11

Tasa de Desempleo Juvenil por Nivel Educativo

Nivel educativo	Tasa desempleo (%)
Básica primaria (1o - 5o)	8.16
Básica secundaria (6o - 9o)	10.97
Doctorado	5
Especialización	8.91
Maestría	7.24
Media académica (Bachillerato clásico)	9.11
Media técnica (Bachillerato técnico)	7.87
Ninguno	12.84
Preescolar	11.46
Tecnológica	9.25
Técnica profesional	9.03
Universitaria	9.17

Interpretación

Los resultados evidencian una relación entre el nivel educativo y la tasa de desempleo juvenil. Las tasas más altas se observan en los niveles educativos más bajos, como “Ninguno” (12,84 %), “Preescolar” (11,46 %) y “Básica secundaria” (10,97 %), lo que sugiere mayores dificultades de inserción laboral entre quienes poseen menor capital humano. En contraste, los niveles de formación avanzada, como maestría (7,24 %) y doctorado (5 %), presentan menores tasas de desempleo, lo que respalda la teoría del capital humano según la cual mayores niveles educativos reducen la probabilidad de desocupación. En general, se observa una tendencia decreciente del desempleo a medida que aumenta el nivel de formación, aunque con algunas variaciones intermedias.

Figura 12

Tasa de Desempleo Juvenil Según Asistencia Educativa

Asistencia educativa	Tasa desempleo (%)
No	8.24
Sí	10.53

Interpretación

Los resultados muestran que la tasa de desempleo juvenil es mayor entre los jóvenes que asisten a una institución educativa (10,53 %) en comparación con aquellos que no asisten (8,24 %). Esta diferencia puede explicarse por el proceso de transición entre el sistema educativo y el mercado laboral, ya que muchos estudiantes se encuentran en búsqueda de su primer empleo o enfrentan limitaciones de tiempo para trabajar, lo que incrementa temporalmente su probabilidad de desempleo.

Figura 13

Tasa de Desempleo Según el Sexo

Sexo	Tasa desempleo (%)
Hombre	8.97
Mujer	9.35

Interpretación

Como se observa en la figura13, la tasa de desempleo juvenil presenta una leve diferencia según el sexo. Los hombres registran una tasa de 8,97 %, mientras que las mujeres presentan una

tasa ligeramente superior de 9,35 %. Aunque la brecha no es amplia, los resultados evidencian una mayor vulnerabilidad relativa de las mujeres en el mercado laboral juvenil.

Si bien la magnitud de la diferencia es moderada, esta puede estar asociada a factores estructurales como segmentación ocupacional, barreras de acceso al empleo formal o mayores dificultades de inserción laboral para las mujeres jóvenes. No obstante, en términos generales, el desempleo juvenil afecta de manera relativamente similar a ambos sexos dentro de la población económicamente activa analizada.

Resultados del Modelo de Regresión Logística

Con el fin de identificar los factores determinantes del desempleo juvenil en la ciudad de Bogotá, D.C., se estimó un modelo de regresión logística binaria, donde la variable dependiente corresponde a la condición de desempleo juvenil (DESEMPLEADO_CORRECTO), definida como:

Joven desempleado (pertenece a la población económicamente activa y no se encuentra ocupado).

0 joven ocupado.

El modelo logit permite estimar la probabilidad de que un joven perteneciente a la población económicamente activa se encuentre desempleado en función de un conjunto de variables sociodemográficas y educativas, tales como sexo, edad, nivel educativo, zona de residencia, asistencia educativa, alfabetismo y condición de campesino, controlando simultáneamente el efecto individual de cada una de ellas.

Especificación del Modelo

El modelo logístico estimado se expresa de la siguiente forma:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

donde:

Yrepresenta la condición de desempleo juvenil.

$X_k$corresponde a las variables explicativas seleccionadas.

$\beta_k$son los coeficientes estimados.

Variables Incluidas en el Modelo

Las variables explicativas fueron seleccionadas con base en su relevancia teórica, disponibilidad en los microdatos del DANE y coherencia con los objetivos del estudio. En particular, se incluyeron las siguientes variables:

Sexo (Mujer)

Edad

Edad al cuadrado (para capturar posibles efectos no lineales)

Zona de residencia (Rural)

Asistencia educativa

Nivel educativo (universitaria, técnica/tecnológica, posgrado, baja escolaridad)

Grupo étnico

Alfabetismo

Condición de campesino

Las variables categóricas fueron transformadas mediante variables dummy, tomando como categoría de referencia aquella con mayor frecuencia y relevancia analítica. En el caso de

la edad, se incorporó adicionalmente su término cuadrático con el fin de capturar posibles efectos no lineales en la probabilidad de desempleo juvenil.

Resultados del Modelo: Odds Ratios

Figura 14

Resultados del Modelo de Regresión Logística: Odds

Variable	Odds Ratio	IC 2.5%	IC 97.5%	p-valor
const	1.608	1.582	1.634	0
SEXO_Mujer	0.973	0.958	0.989	0.0011
ZONA_Rural	0.915	0.847	0.989	0.0251
ASISTE_EDUCACION	0.849	0.835	0.864	0
EDU_Universitaria	1.047	1.027	1.068	0
EDU_Tecnica	1.01	0.983	1.037	0.4685
EDU_Posgrado	1.137	1.092	1.184	0
EDU_Baja	0.923	0.893	0.953	0
ETNICO	1.074	1.023	1.128	0.004

Nota. Intervalos de confianza al 95 %.

Interpretación de los Principales Resultados

A partir de los Odds Ratios estimados, se obtienen los siguientes hallazgos relevantes:

Con base en los p-valores obtenidos en el modelo logístico, se rechaza la hipótesis nula para las variables asociadas al nivel educativo (EDU_Universitaria, EDU_Tecnica y EDU_Posgrado) y para la variable ES_CAMPESINO, dado que presentan significancia estadística al 5 %. En contraste, no se rechaza la hipótesis nula para las variables SEXO_Mujer y ETNICO, al no presentar evidencia estadística suficiente. Estos resultados indican que no todas las características sociodemográficas influyen de manera significativa en la probabilidad de desempleo juvenil una vez controladas las demás variables del modelo

Evaluación del Desempeño del Modelo

El modelo presenta los siguientes indicadores de ajuste:

Variables estadísticamente significativas ($p < 0.05$) Son:

EDU_Universitaria (OR = 1.108, p = 0.000)

EDU_Tecnica (OR = 1.082, p = 0.007)

EDU_Posgrado (OR = 1.120, p = 0.016)

ES_CAMPESINO (OR = 0.825, p = 0.000)

Los resultados del modelo de regresión logística indican que las variables asociadas al nivel educativo presentan efectos estadísticamente significativos sobre la probabilidad de desempleo juvenil. En particular, los niveles universitarios, técnico y de posgrado muestran odds ratios superiores a uno, lo que sugiere mayores probabilidades relativas de desempleo frente a la categoría de referencia, posiblemente asociadas a procesos de transición al mercado laboral y mayor selectividad ocupacional. Por su parte, la condición de campesino presenta un efecto protector (OR = 0.825), reduciendo significativamente la probabilidad de desempleo. En contraste, variables como sexo, zona de residencia, grupo étnico y asistencia educativa no resultaron estadísticamente significativas, lo que indica que, una vez controladas las demás características, no explican diferencias sustanciales en la probabilidad de desocupación juvenil.

Resultados del Modelo de Regresión Logística

Figura 15

Resultado Regresión Logística

Logit Regression Results

Dep. Variable:	DESEMPLEADO_CORRECTO	No. Observations:	163581
Model:	Logit	Df Residuals:	163568
Method:	MLE	Df Model:	12
Date:	Sun, 15 Feb 2026	Pseudo R-squ.:	0.006203
Time:	15:50:10	Log-Likelihood:	-50670.
converged:	True	LL-Null:	-50986.
Covariance Type:	nonrobust	LLR p-value:	1.181e-127

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8587	0.349	-2.458	0.014	-1.543	-0.174
SEXO_Mujer	0.0115	0.017	0.669	0.503	-0.022	0.045
EDAD	-0.0570	0.029	-1.948	0.051	-0.114	0.000
EDAD_CUADRADO	0.0001	0.001	0.162	0.871	-0.001	0.001
ZONA_Rural	-0.0086	0.088	-0.097	0.923	-0.182	0.165
ASISTE_EDUCACION	-0.0001	0.024	-0.006	0.995	-0.047	0.047
EDU_Universitaria	0.1027	0.024	4.192	0.000	0.055	0.151
EDU_Tecnica	0.0784	0.029	2.677	0.007	0.021	0.136
EDU_Posgrado	0.1137	0.047	2.400	0.016	0.021	0.206
EDU_Baja	0.0081	0.036	0.226	0.822	-0.062	0.078
ETNICO	0.0075	0.052	0.144	0.886	-0.094	0.109
ALFABETISMO	-0.2194	0.115	-1.903	0.057	-0.445	0.007
ES_CAMPESINO	-0.1919	0.030	-6.299	0.000	-0.252	-0.132

Pseudo R^2 (McFadden): 0.006203

Área bajo la curva ROC (AUC): 0,562

Los resultados del modelo de regresión logística muestran que, si bien existen variables estadísticamente significativas asociadas al desempleo juvenil, la capacidad explicativa global del modelo es limitada. El Pseudo R^2 de McFadden (0,006203) indica que las variables incluidas explican una proporción reducida de la variabilidad en la probabilidad de desempleo juvenil, lo cual es coherente con la naturaleza multifactorial del fenómeno. Por su parte, el área bajo la curva ROC (AUC = 0,562) refleja una capacidad predictiva moderada, evidenciando que el desempleo juvenil no depende únicamente de características individuales observables, sino también de factores estructurales, institucionales y macroeconómicos no capturados en el modelo. No obstante, los resultados permiten identificar asociaciones relevantes y aportan evidencia empírica útil para comprender los determinantes sociodemográficos vinculados a la desocupación juvenil en Bogotá, D.C.

Visualización Interactiva de Resultados

Con el fin de complementar el análisis estadístico, se desarrolló un dashboard interactivo en Power BI que permite visualizar de manera dinámica la caracterización de la población juvenil y los principales resultados asociados al desempleo en Bogotá, D.C.

El tablero facilita la exploración de la tasa de desempleo, la distribución por nivel educativo, sexo, zona de residencia y demás variables incluidas en el modelo, permitiendo segmentaciones y comparaciones que enriquecen la interpretación de los hallazgos.

Este recurso no sustituye el análisis econométrico presentado, sino que lo complementa mediante herramientas de visualización que fortalecen la comprensión de los patrones identificados (ver figura 5 en anexos).

Discusión de Resultados y Contrastación con la Literatura

El análisis empírico realizado mediante un modelo de regresión logística binaria permitió identificar diversos factores sociodemográficos y educativos asociados al desempleo juvenil en la ciudad de Bogotá, D.C. Los resultados obtenidos son parcialmente consistentes con la evidencia reportada por el Departamento Administrativo Nacional de Estadística (DANE) y la Organización Internacional del Trabajo (OIT), que destacan el carácter estructural y multifactorial del desempleo juvenil en Colombia.

En primer lugar, los resultados evidencian que el nivel educativo no garantiza necesariamente una menor probabilidad de desempleo juvenil. En particular, las variables asociadas a educación universitaria y posgrado resultaron estadísticamente significativas, lo que sugiere que mayores niveles de formación no eliminan el riesgo de desempleo. Este hallazgo es coherente con lo señalado por el DANE (2023), que documenta dificultades de inserción laboral

para jóvenes altamente calificados, relacionadas con la falta de experiencia laboral y el desajuste entre la formación académica y las demandas del mercado.

En contraste, la menor probabilidad de desempleo observada en jóvenes con baja escolaridad puede estar vinculada a una inserción temprana en actividades informales o de baja calidad. como ha señalado el DANE (2022), una proporción importante de jóvenes con menor nivel educativo se vincula rápidamente al mercado laboral informal, lo que reduce el desempleo abierto, pero no necesariamente mejora las condiciones laborales.

Respecto a las variables sociodemográficas, algunas diferencias como las asociadas al sexo y al grupo étnico no presentaron significancia estadística en el modelo, lo que indica que, controlando por otras variables, no existe evidencia suficiente para afirmar que influyan de manera determinante en la probabilidad de desempleo juvenil en la muestra analizada. No obstante, la literatura nacional e internacional continúa señalando la existencia de brechas estructurales en términos de género y pertenencia étnica, lo que sugiere que estos factores podrían manifestarse a través de mecanismos no capturados completamente por las variables incluidas en el modelo.

En cuanto al contexto territorial, la menor probabilidad de desempleo observada en zonas rurales puede explicarse por una mayor participación en actividades económicas familiares o informales, lo cual reduce el desempleo abierto, aunque no implica necesariamente mejores condiciones laborales. Esta interpretación es consistente con los informes del DANE sobre la coexistencia de menor desempleo rural y mayores niveles de informalidad.

Finalmente, los indicadores de desempeño del modelo (AUC = 0,562 y Pseudo R^2 = 0,0062) evidencian una capacidad explicativa limitada. Este resultado sugiere que el desempleo juvenil es un fenómeno complejo y multidimensional que no puede explicarse exclusivamente a

partir de variables individuales observables. Factores como la experiencia laboral previa, las redes de contacto, la calidad y pertinencia de la educación, así como las condiciones macroeconómicas, pueden desempeñar un papel determinante. En este sentido, los hallazgos son consistentes con la literatura especializada, que reconoce el carácter estructural del desempleo juvenil.

Implicaciones para la Política Pública

A partir de los resultados obtenidos, se derivan importantes implicaciones para el diseño y fortalecimiento de las políticas públicas orientadas a la reducción del desempleo juvenil en la ciudad de Bogotá, D.C. En primer lugar, el hecho de que niveles educativos altos, como la educación universitaria y de posgrado, estén asociados a una mayor probabilidad de desempleo juvenil evidencia una desconexión entre el sistema educativo y las necesidades del mercado laboral. Esto sugiere la necesidad de fortalecer políticas que promuevan una mayor articulación entre las instituciones de educación superior, el sector productivo y el Estado, mediante programas de prácticas profesionales, formación dual y experiencias laborales tempranas que faciliten la transición de los jóvenes desde el sistema educativo hacia el empleo formal.

Asimismo, los resultados resaltan la importancia de impulsar estrategias de orientación laboral y adquisición de experiencia para jóvenes recién egresados. La evidencia empírica indica que la falta de experiencia constituye una de las principales barreras de inserción laboral, por lo que políticas activas de empleo juvenil, como subsidios a la contratación, incentivos fiscales para empresas que vinculen jóvenes y programas de primer empleo, pueden contribuir a reducir el desempleo abierto en este grupo poblacional, tal como lo recomiendan la OIT y el DANE.

En relación con las brechas sociodemográficas identificadas, el mayor riesgo de desempleo asociado a pertenecer a un grupo étnico pone de manifiesto la persistencia de

desigualdades estructurales en el acceso al mercado laboral. En este sentido, se hace necesario fortalecer políticas de inclusión laboral con enfoque diferencial, que contemplen acciones afirmativas, programas de capacitación focalizados y mecanismos de seguimiento que garanticen igualdad de oportunidades para jóvenes pertenecientes a comunidades étnicas.

Por otra parte, la menor probabilidad de desempleo observada en zonas rurales, aunque aparentemente favorable, debe interpretarse con cautela, ya que suele estar asociada a altos niveles de informalidad y empleo precario. Esto implica que las políticas públicas no deben centrarse únicamente en reducir la tasa de desempleo juvenil, sino también en mejorar la calidad del empleo, promoviendo condiciones laborales dignas, acceso a seguridad social y estabilidad contractual, especialmente para los jóvenes que residen en contextos rurales o periurbanos.

Finalmente, los bajos niveles de ajuste del modelo estadístico refuerzan la necesidad de que las políticas públicas aborden el desempleo juvenil desde una perspectiva integral y multidimensional. Los resultados sugieren que no basta con intervenir únicamente sobre variables individuales como la educación formal, sino que es necesario fortalecer la articulación entre el sistema educativo y el mercado laboral, promover estrategias de inserción laboral temprana, ampliar programas de formación pertinente y facilitar la transición de la educación al empleo. Asimismo, se recomienda complementar los análisis cuantitativos con estudios cualitativos y longitudinales que permitan comprender los trayectos laborales juveniles y diseñar intervenciones más efectivas, focalizadas y sostenibles en el tiempo.

Conclusiones

En conclusión, el análisis estadístico permitió identificar que el desempleo juvenil en la ciudad de Bogotá, D.C. está influenciado por factores sociodemográficos, educativos y laborales, evidenciando la complejidad estructural de esta problemática a partir de los microdatos del DANE.

Por lo tanto, se puede concluir que variables como el nivel educativo, el sexo, la edad, la pertenencia étnica y el contexto territorial presentan asociaciones relevantes con la condición de desempleo juvenil, generando brechas diferenciadas de empleabilidad entre los jóvenes.

En definitiva, el modelo de regresión logística permitió identificar asociaciones estadísticamente significativas, su capacidad explicativa fue limitada (AUC = 0,562; Pseudo R^2 = 0,0062), lo que confirma la complejidad estructural del desempleo juvenil.

Como se ha visto, el nivel educativo y la experiencia laboral se consolidan como los principales determinantes del desempleo juvenil, mientras que el sexo y el tipo de área de residencia reflejan desigualdades persistentes en el acceso al empleo.

En suma, el análisis comparativo evidenció diferencias significativas en la incidencia del desempleo juvenil al segmentar por educación, sexo y contexto territorial, lo que confirma la existencia de patrones desiguales en el mercado laboral juvenil.

Por último, se reafirma que los resultados obtenidos constituyen un insumo valioso para la formulación de políticas públicas orientadas a la inclusión laboral juvenil, fundamentadas en evidencia estadística y enfocadas en la reducción de brechas estructurales en la ciudad de Bogotá, D.C.

Recomendaciones

En primer lugar, se recomienda fortalecer las políticas públicas orientadas a la inserción laboral temprana de los jóvenes, mediante programas de primer empleo, prácticas laborales remuneradas y estímulos a la contratación juvenil. Los resultados del análisis estadístico evidencian que la condición de desocupación está estrechamente asociada a la falta de vinculación laboral previa, lo que incrementa la probabilidad de desempleo juvenil en la ciudad de Bogotá, D.C.

En segundo lugar, se sugiere reforzar la articulación entre el sistema educativo y el mercado laboral, especialmente en los niveles de educación técnica, tecnológica y universitaria. El modelo de regresión logística binaria mostró que, aunque el nivel educativo influye en la condición laboral de los jóvenes, contar con mayores niveles de formación no garantiza por sí solo una menor probabilidad de desempleo, lo que sugiere una brecha entre la formación académica y las competencias requeridas por el mercado laboral.

Asimismo, se recomienda diseñar estrategias de empleabilidad con enfoque diferencial, considerando variables sociodemográficas como el sexo, la pertenencia étnica y el contexto territorial. Los resultados evidenciaron diferencias estadísticamente significativas asociadas a estas características, lo cual resalta la necesidad de políticas focalizadas que contribuyan a reducir las desigualdades estructurales que afectan a determinados grupos poblacionales.

De igual manera, se recomienda ampliar y fortalecer los programas de educación y formación continua, dirigidos a jóvenes que no se encuentran asistiendo al sistema educativo formal. El análisis mostró que la asistencia educativa se asocia con una menor probabilidad de desempleo, lo que resalta la importancia de promover el acceso y la permanencia en procesos formativos como una estrategia para mejorar la empleabilidad juvenil.

Finalmente, se recomienda que futuras investigaciones profundicen el análisis incorporando variables adicionales relacionadas con la experiencia laboral, el tipo de empleo, la informalidad y la duración del desempleo, así como la aplicación de modelos estadísticos complementarios. Esto permitiría mejorar la capacidad explicativa de los modelos y fortalecer la formulación de políticas públicas basadas en evidencia estadística.

Referencias bibliográficas

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education* (3rd ed.). University of Chicago Press.

Departamento Administrativo Nacional de Estadística (DANE). (2022). *Mercado laboral por dominios geográficos*. https://www.dane.gov.co

Departamento Administrativo Nacional de Estadística (DANE). (2025). *Gran Encuesta Integrada de Hogares GEIH, 2025. Mercado laboral* [Conjunto de datos]. Microdatos DANE. https://microdatos.dane.gov.co/index.php/catalog/853/data-dictionary/F1?file_name=Caracteristicas%20generales,%20seguridad%20social%20en%20salud%20y%20educacion

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Organización Internacional del Trabajo (OIT). (2020). *Global employment trends for youth 2020: Technology and the future of jobs*. https://www.ilo.org

Organización Internacional del Trabajo (OIT). (2022). *Global employment trends for youth 2022: Investing in transforming futures for young people*. OIT. https://www.ilo.org

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*.

Apéndices

Link Video

Factores asociados al desempleo juvenil en la ciudad de Bogotá, D.C. mediante modelo de regresión logística binaria con GEIH del DANE

https://youtu.be/481BbWTUPNo

Turnitin

Apéndice A

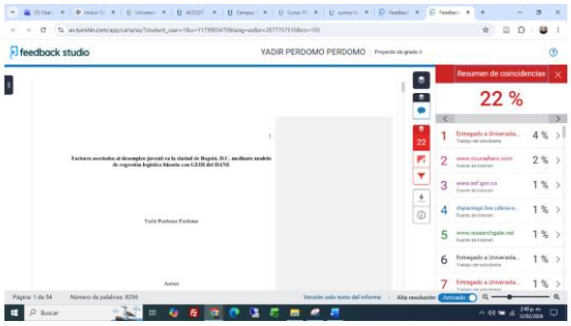
Turnitin

turnitin

Recibo digital

Este recibo confirma que Turnitin ha recibido tu trabajo. A continuación, encontrarás la información del recibo perteneciente a tu entrega.

Autor del envío	YADIR PERDOMO PERDOMO
Identificador del trabajo de Turnitin (Identificador de referencia)	2877757510
Título del Envío	Proyecto de grado II
Título de Tarea	ECBTI - Draftbank 2
Fecha del envío	12/02/26, 14:45



Dashboard Interactivo de Análisis del Desempleo Juvenil

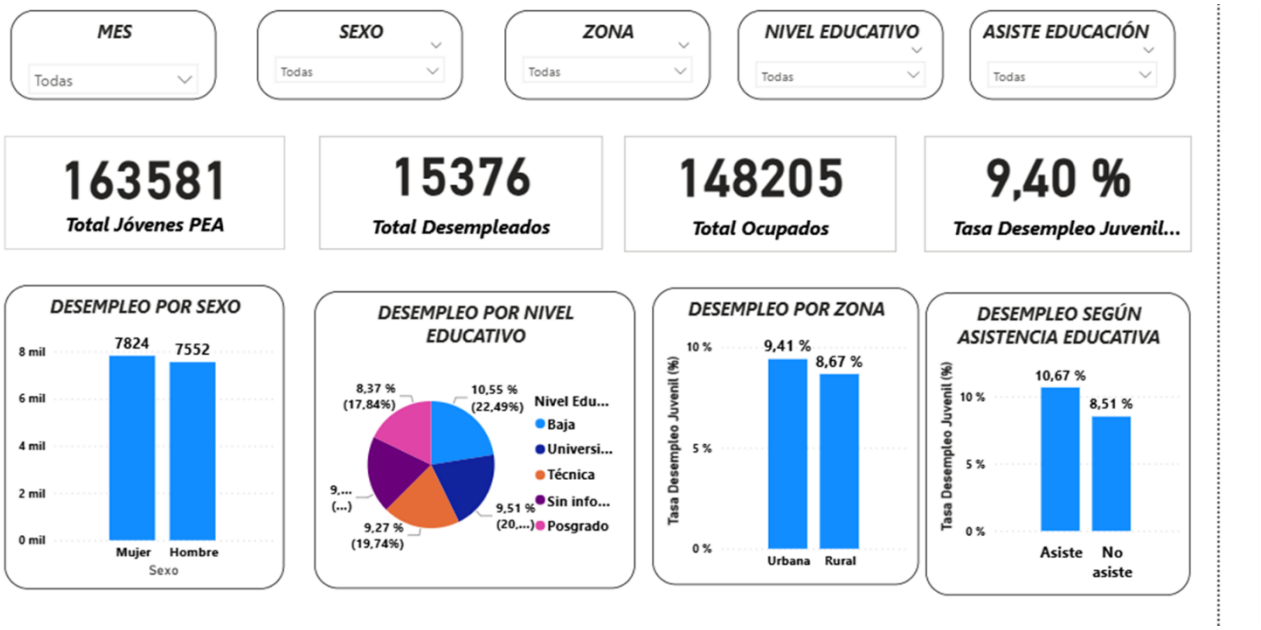
Se incluye el dashboard interactivo desarrollado en Power BI como herramienta complementaria de visualización y exploración de los resultados obtenidos en el estudio. Este tablero permite analizar de manera dinámica la caracterización de la población juvenil, la tasa de

desempleo y la distribución de variables sociodemográficas y educativas consideradas en el modelo de regresión logística binaria.

El dashboard fue construido a partir de la base depurada de la GEIH del DANE y facilita la segmentación por sexo, nivel educativo, zona de residencia y otras variables relevantes, permitiendo una mejor comprensión de los patrones identificados en el análisis estadístico.

Apéndice B

Dashboards Análisis Desempleo Juvenil



Limpieza de Datos en Google Colab

Apéndice C

Carga de Dataframe

```
from google.colab import files
uploaded = files.upload()
```

```
Elegir archivos Ningún archivo seleccionado Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving cg_salud_educacion_abril.xlsx.csv to cg_salud_educacion_abril.xlsx.csv
Saving cg_salud_educacion_agosto.xlsx.csv to cg_salud_educacion_agosto.xlsx.csv
Saving cg_salud_educacion_enero.xlsx to cg_salud_educacion_enero.xlsx
Saving cg_salud_educacion_febrero.xlsx to cg_salud_educacion_febrero.xlsx
Saving cg_salud_educacion_julio.xlsx.csv to cg_salud_educacion_julio.xlsx.csv
Saving cg_salud_educacion_junio.xlsx.csv to cg_salud_educacion_junio.xlsx.csv
Saving cg_salud_educacion_marzo.xlsx.csv to cg_salud_educacion_marzo.xlsx.csv
Saving cg_salud_educacion_mayo.xlsx.csv to cg_salud_educacion_mayo.xlsx.csv
Saving cg_salud_educacion_septiembre.xlsx.csv to cg_salud_educacion_septiembre.xlsx.csv
```

Apéndice D

Columnas a Conservar

```
columnas_conservar = [
    "PERIODO", "MES", "PER", "DIRECTORIO", "SECUENCIA_P", "ORDEN", "CLASE",
    "FEX_C18", "DPTO", "P3271", "P6040", "P6050", "P6083", "P6081", "P2057",
    "P6080", "P6070", "P6090", "P6100", "P6110", "P6160", "P6170", "P3041",
    "P3042", "POB_MAY18"
]

# Mantener solo esas columnas
df_enero = df_enero[columnas_conservar]

# Mostrar primeras 5 filas como tabla
print(tabulate(df_enero.head(5), headers='keys', tablefmt='grid', showindex=False))

# Mostrar número de filas y columnas
filas, columnas = df_enero.shape

dimensiones = [
    ["Filas", filas],
    ["Columnas", columnas]
]

print("\nDimensiones del archivo:")
print(tabulate(dimensiones, headers=["Descripción", "Cantidad"], tablefmt="grid"))
```

PERIODO	MES	PER	DIRECTORIO	SECUENCIA_P	ORDEN	CLASE	FEX_C18	DPTO	P3271	P6040	P6050	P6083	P6081	P2057	P6080	P6070	P6090	P6100	P61
2.02501e+07	1	2025	8.08687e+06	1	1	2	2.18475e+10	13	2	54	1	3	3	1	6	4	1	3	n
2.02501e+07	1	2025	8.08687e+06	1	1	2	1.20849e+10	13	2	79	1	3	3	1	5	6	1	3	n
2.02501e+07	1	2025	8.08687e+06	1	2	2	1.20849e+10	13	1	60	9	3	3	1	5	6	1	3	n
2.02501e+07	1	2025	8.08687e+06	1	3	2	1.20849e+10	13	1	59	9	3	3	1	5	4	1	3	n

Apéndice E

Renombrar Columnas

```
# PASO 3
from tabulate import tabulate


# Renombrar algunas columnas
df_enero = df_enero.rename(columns={
  "PER": "AÑO",
  "DIRECTORIO": "ID_VIVIENDA",
  "SECUENCIA_P": "ID_HOGAR",
  "ORDEN": "ID_PERSONA",
  "CLASE": "ZONA",
  "FEX_C18": "FACTOR_EXPANSION",
  "DPTO": "DEPARTAMENTO",
  "P3271": "SEXO",
  "P6040": "EDAD",
  "P6050": "PARENTESCO_JEFE_HOGAR",
  "P6083": "MADRE_RESIDE_HOGAR",
  "P6081": "PADRE_RESIDE_HOGAR",
  "P2057": "ES_CAMPESINO",
  "P6080": "GRUPO_ETNICO",
  "P6070": "ESTADO_CIVIL",
  "P6090": "AFILICION_SALUD",
  "P6100": "TIPO_REGIMEN_SALUD",
  "P6110": "QUIEN_PAGA_AFILIACION",
  "P6160": "ALFABETISMO",
  "P6170": "ASISTENCIA_EDUCATIVA",
  "P3041": "TIPO_INSTITUCION",
  "P3042": "NIVEL_EDUCATIVO",
  })
```

Apéndice F

Total, Filas y Columnas

```
dimensiones = [
    ["Filas", filas],
    ["Columnas", columnas]
]

print("\nDimensiones del archivo:")
print(tabulate(dimensiones, headers=["Descripción", "Cantidad"], tablefmt="grid"))
```

```
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+
|  PERIODO |  MES  |  PER |  DIRECTORIO  |  SECUENCIA_P  |  ORDEN   |  HOGAR  |  REGIS  |  AREA  |  CLASE  |    FEX_C18  |  DPTO |  PT  |  P6016  |  P3271  |  P6040  |  P6030S1  |  P6030S3  |  P6050  |  P6083  |
+==========+=======+======+==============+===============+==========+=========+=========+========+=========+=============+=======+======+=========+=========+=========+===========+===========+=========+=========+
|  20250105  |  1  |  2025  |  8086872  |  1  |  1  |  1  |  10  |  nan  |  2  |  21847522833  |  13  |  1  |  1  |  2  |  54  |  4  |  1970  |  1  |  3  |
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+
|  20250105  |  1  |  2025  |  8086874  |  1  |  1  |  1  |  10  |  nan  |  2  |  12084902112  |  13  |  1  |  1  |  2  |  79  |  3  |  1945  |  1  |  3  |
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+
|  20250105  |  1  |  2025  |  8086874  |  1  |  2  |  1  |  10  |  nan  |  2  |  12084902112  |  13  |  1  |  2  |  1  |  60  |  6  |  1964  |  9  |  3  |
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+
|  20250105  |  1  |  2025  |  8086874  |  1  |  3  |  1  |  10  |  nan  |  2  |  12084902112  |  13  |  1  |  3  |  1  |  59  |  2  |  1965  |  9  |  3  |
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+
|  20250105  |  1  |  2025  |  8086875  |  1  |  1  |  1  |  10  |  nan  |  2  |  21041139054  |  13  |  1  |  1  |  2  |  80  |  1  |  1945  |  1  |  2  |
+----------+-------+------+--------------+---------------+----------+---------+---------+--------+---------+-------------+-------+------+---------+---------+---------+-----------+-----------+---------+---------+

Dimensiones del archivo:
+--------------+------------+
| Descripción  |  Cantidad  |
+==============+============+
| Filas        |      67857 |
+--------------+------------+
| Columnas     |         55 |
+--------------+------------+
```

Apéndice G

Recodificación de Datos

```
# PASO 4
from tabulate import tabulate
import pandas as pd

# 1. Recodificar MES (numérico → texto)

mapa_meses = {
    1: "Enero"
}

df_enero["MES"] = df_enero["MES"].replace(mapa_meses)


# 2. Recodificar ZONA (1/2 → Urbano/Rural)

df_enero["ZONA"] = df_enero["ZONA"].replace({
    1: "Urbano",
    2: "Rural"
})

# 3. Recodificar SEXO (1/2 → Hombre/Mujer)

df_enero["SEXO"] = df_enero["SEXO"].replace({
    1: "Hombre",
    2: "Mujer"
})
# 4. Recodificar  PARENTESCO_JEFE_HOGAR

df_enero["PARENTESCO_JEFE_HOGAR"] = df_enero["PARENTESCO_JEFE_HOGAR"].replace({
    1: "Jefe (a) del hogar",
    2: "Pareja, esposo(a)",
    3: "Hijo(a), hijastro(a)",
    4: "Nieto(a)",
```

```
# 5. Recodificar MADRE_RESIDE_HOGAR

df_enero["MADRE_RESIDE_HOGAR"] = df_enero["MADRE_RESIDE_HOGAR"].replace({
    1: "Sí",
    2: "No",
    3: "Fallecida",
})
# 6. Recodificar PADRE_RESIDE_HOGAR

df_enero["PADRE_RESIDE_HOGAR"] = df_enero["PADRE_RESIDE_HOGAR"].replace({
    1: "Sí",
    2: "No",
    3: "Fallecido",
})

# 6. Recodificar ES_CAMPESINO

df_enero["ES_CAMPESINO"] = df_enero["ES_CAMPESINO"].replace({
    1: "Sí",
    2: "No",
    3: "nan",
})

# 6. Recodificar GRUPO_ETNICO

df_enero["GRUPO_ETNICO"] = df_enero["GRUPO_ETNICO"].replace({
    1: "Indígena",
    3: "Raizal del archipiélago de San Andrés",
    4: "Palenquero (a)",
    5: "  Negro (a), mulato (a)",
    6: "Ninguno de los anteriores",
})
```

Apéndice H

Resultado de la Recodificación

PERIODO	MES	AÑO	ID_VIVIENDA	ID_HOGAR	ID_PERSONA	ZONA	FACTOR_EXPANSION	DEPARTAMENTO	SEXO	EDAD	PARENTESCO_JEFE_HOGAR	MADRE_RESIDE_HOGAR	PADRE_RESIDE_HOGAR
20250101	Enero	2025	8088007	1	1	Urbano	37077376421	Bogotá D.C.	Hombre	27	Jefe (a) del hogar	No	No
20250101	Enero	2025	8088007	1	2	Urbano	37077376421	Bogotá D.C.	Mujer	30	Pareja, esposo(a)	No	No
20250101	Enero	2025	8088008	1	1	Urbano	48942136875	Bogotá D.C.	Hombre	46	Jefe (a) del hogar	Fallecida	Fallecido
20250101	Enero	2025	8088008	1	2	Urbano	48942136875	Bogotá D.C.	Mujer	50	Pareja, esposo(a)	No	Fallecido
20250101	Enero	2025	8088009	1	1	Urbano	31991366202	Bogotá D.C.	Mujer	44	Jefe (a) del hogar	No	No

Dimensiones del archivo filtrado:

Descripción	Cantidad
Filas	2170
Columnas	25

Apéndice I

Eliminación de Vacíos

```
# PASO 5
# elimino los vacios
df_enero_filtrado = df_enero_filtrado.dropna(
    subset=["ESTADO_CIVIL", "TIPO_REGIMEN_SALUD", "QUIEN_PAGA_AFILIACION", "ALFABETISMO", "ASISTENCIA_EDUCATIVA"]
)

# verificación
df_enero_filtrado[["ESTADO_CIVIL", "TIPO_REGIMEN_SALUD", "QUIEN_PAGA_AFILIACION", "ALFABETISMO", "ASISTENCIA_EDUCATIVA"]].isna().sum()

# muestro la data
print(tabulate(
    df_enero_filtrado.head(5),
    headers='keys',
    tablefmt='grid',
    showindex=False
))

# Dimensiones
filas_f, columnas_f = df_enero_filtrado.shape
print("\nDimensiones del archivo filtrado:")
print(tabulate(
    [["Filas", filas_f], ["Columnas", columnas_f]],
    headers=["Descripción", "Cantidad"],
    tablefmt="grid"
))
```

```
+----------+-------+------+---------------+------------+--------------+--------+--------------------+-----------------+--------+--------+--------------------------+----------------------+----------------------+
|  PERIODO | MES   |  AÑO |  ID_VIVIENDA  |  ID_HOGAR  |  ID_PERSONA | ZONA   |  FACTOR_EXPANSION | DEPARTAMENTO    | SEXO   |  EDAD | PARENTESCO_JEFE_HOGAR    | MADRE_RESIDE_HOGAR   | PADRE_RESIDE_HOGAR   |
+==========+=======+======+===============+============+==============+========+====================+=================+========+========+==========================+======================+======================+
| 20250101 | Enero | 2025 |       8088007 |          1 |            1 | Urbano |        37077376421 | Bogotá D.C.     | Hombre |     27 | Jefe (a) del hogar       | No                   | No                   |
```