Universidad Externado de Colombia

Facultad de Economía

Curso de Inteligencia Artificial con Aplicaciones en Economía I

Educación, Juventud e Informalidad en Bogotá: Un análisis basado en la GEIH

David Leonardo Martínez Pinzón

11 de noviembre de 2025

Resumen:

Bogotá concentra buena parte del capital educativo del país, miles de estudiantes, decenas de campus y una narrativa dominante que repite que estudiar es la ruta "natural" hacia la formalidad laboral. La promesa que nos venden a quienes con perspectiva estudiamos es que si inviertes millones en tu formación, el mercado te devolverá un empleo estable, con seguridad social y proyección. Este trabajo contrasta esa promesa con datos. A partir de la GEIH 2024 se analiza la informalidad laboral de los jóvenes ocupados entre 18 y 28 años en Bogotá, combinando un análisis exploratorio detallado con modelos de inteligencia artificial supervisados.

Los resultados son incómodos pero claros, la educación reduce el riesgo de informalidad, pero no lo elimina; incluso los jóvenes con educación universitaria mantienen tasas de informalidad relevantes. Las diferencias por sexo y por identidad de género, así como las jornadas laborales extendidas, muestran que el problema excede la idea de "falta de educación" y se ancla en la estructura misma del mercado laboral juvenil. Regresión logística y Random Forest permiten cuantificar la probabilidad de ser informal y detectar qué variables pesan más en esa predicción. La hipótesis que guía el trabajo se mantiene: si aún en la ciudad con mayor concentración universitaria la formalidad no está garantizada para la juventud, el problema no es educativo. El problema es estructural.

Palabras clave: Informalidad laboral, Juventud, Mercado laboral, Educación superior, Bogotá, IA aplicada a Economía.

1. Introducción

Bogotá funciona como laboratorio de la promesa educativa, universidades en cada esquina, campañas públicas que venden la idea de que “el estudio es la llave de la movilidad social” y programas de becas que se justifican en que más educación implica más formalidad. Sin embargo, la experiencia cotidiana de muchos jóvenes va en otra dirección. Aun con títulos técnicos, tecnológicos o universitarios, una parte importante termina en trabajos sin afiliación a salud ni pensión, con horarios extendidos y alta inestabilidad.

Este trabajo parte de una hipótesis sencilla pero incómoda, si en Bogotá, donde se concentra buena parte del capital educativo del país, la informalidad juvenil sigue siendo alta, entonces no estamos frente a un problema de “falta de estudio”. Estamos frente a un problema estructural del mercado laboral.

Para ponerlo a prueba, se utiliza la GEIH 2024 y se trabaja exclusivamente con jóvenes ocupados entre 18 y 28 años residentes en la ciudad. El análisis combina: i) un EDA que muestra cómo se distribuye la informalidad por sexo, nivel educativo y horas trabajadas, y ii) modelos de clasificación (regresión logística y Random Forest) que estiman la probabilidad de ser informal y permiten comparar el peso de la educación frente a otros factores. La idea es pasar del discurso a la evidencia.

2. Pregunta de investigación.

¿Qué factores explican que un joven ocupado en Bogotá, aun con educación media o superior, caiga en la informalidad laboral?

3. Objetivos

Estimar, con la GEIH 2024, la probabilidad de informalidad laboral juvenil en Bogotá y cuantificar el peso real de la educación frente a otros determinantes como sexo, edad y horas trabajadas.

Caracterizar, mediante análisis exploratorio, los perfiles de jóvenes más expuestos a la informalidad según sexo, nivel educativo e intensidad de trabajo.

Comparar, mediante regresión logística y Random Forest, qué variables resultan más influyentes en la predicción de informalidad y qué tan lejos está el mercado laboral de cumplir la promesa de la profesionalización.

3. Datos y alcance

El estudio trabaja con una sola fuente pública la Gran Encuesta Integrada de Hogares (GEIH) del DANE, corte diciembre de 2024. Se emplean los módulos de:

- Características generales, salud y educación.
- Fuerza de trabajo.
- Hogar y vivienda.
- Ocupados.

Se construye una base unificada a nivel de individuo utilizando las llaves DIRECTORIO, SECUENCIA_P y ORDEN, incorporando HOGAR cuando es necesario.

Población de estudio:

- Jóvenes entre 18 y 28 años (edad entre 18 y 28).
- Residentes en Bogotá (DPTO == 11).
- En condición de ocupados (P6240 == 1).

Sobre esta submuestra se definen la variable objetivo de informalidad y los predictores como nivel educativo, sexo, edad, horas trabajadas, estrato del hogar y sector económico. El análisis no incluye por ahora encuestas empresariales ni otras fuentes; se centra en el lado de la oferta laboral medida por la GEIH.

En la GEIH, la variable objetivo es la informalidad. Se define como un joven que no cotiza ni a salud ni a pensión. Pretendese computar las variables de cotización P5090 correspondiente a Salud y P5100 que responde a Pensión. Esa es la medida más clara y reconocida para identificar informaliad en los microdatos del DANE. Los predictores son directos. Edad, sexo, nivel educativo alcanzado, horas trabajadas, sector económico y estrato del hogar. Opcionalmente también se puede incorporar ingreso y posición ocupacional si se cuenta con el módulo de ocupados, pero dadas las restricciones operacionales y temporales se opta por mantenerse en esa ambiciosa linea.

En la EDL, las variables claveon sector económico, tamaño de empresa, existencia de vacantes, nivel educativo que exigen y la modalidad de contratación. A partir de estas variables se construyen dos targets derivados de empresas que formalizan frente a las que no, y empresas que exigen educación universitaria frente las que no lo hacen. Esto con el objetivo de demostrar que la paradoja de la informalidad en Bogotá no solo es de oferta, sino de demanda también.

En la GEIH la variable objetivo es la informalidad laboral juvenil. Se define, siguiendo criterios del DANE y la OIT, como la situación de un joven ocupado que no cotiza ni a salud ni a pensión:

- No cotiza salud: P6920 == 2.
- No cotiza pensión: P6940 == 2 o valor faltante.

Se construye la variable binaria informal que toma valor 1 si se cumplen ambas condiciones y 0 en caso contrario, y su versión categórica informal_cat (Informal / Formal).

Los predictores principales son:

- edad: años cumplidos (P6040).
- sexo_cat: categorización de P4000 en Hombre, Mujer, Otro/NS.

- edu_cat: transformación de P6070 en niveles educativos agregados:
- Primaria o menos
- Secundaria
- Media
- Técnica / tecnológica
- Universitaria
- P6800: horas trabajadas a la semana.
- Variables adicionales de contexto (estrato, sector económico) que pueden entrar en extensiones del modelo.

5. Metodología

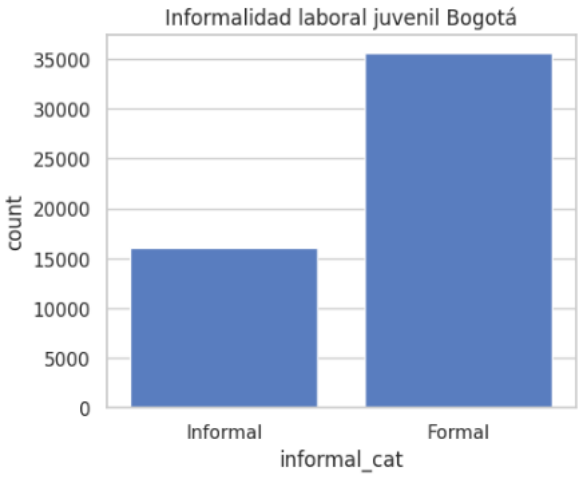
6. Resultados y Discusión

Una vez consolidada la base, se pasa al EDA. Ahí se produce una primera radiografía de la informalidad juvenil. En el código se generan:

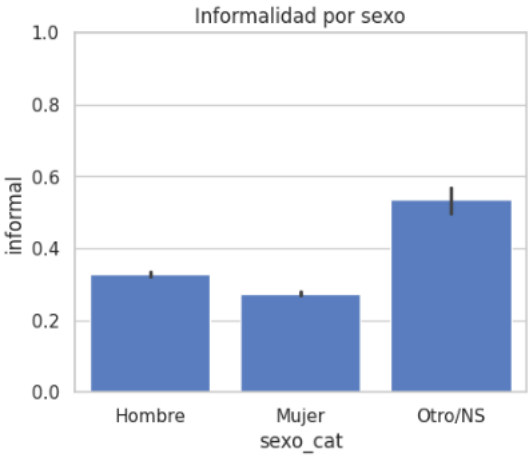
- Distribuciones de informalidad total (informal_cat).
- Gráficos de informalidad por sexo y por nivel educativo.

- Un mapa de calor que cruza sexo × educación y muestra el porcentaje promedio de informalidad por celda.
- Boxplots de horas trabajadas (P6800) según si la persona es formal o informal.
- Tablas de NA y revisión de posibles outliers.

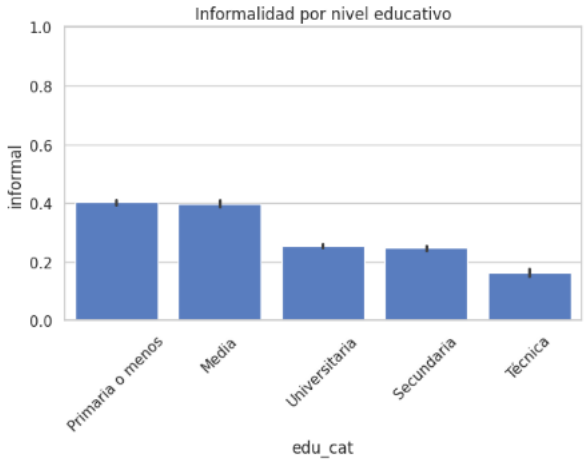
[Figura 1. Distribución general de informalidad]



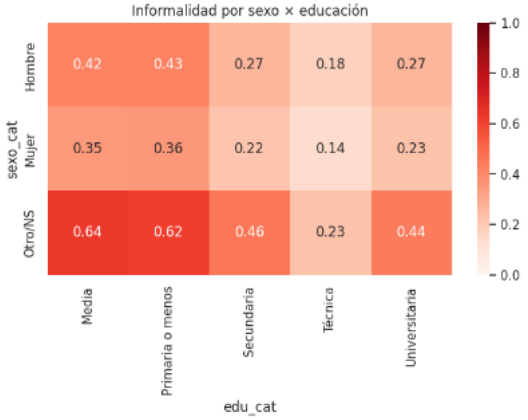
[Figura 2. Informalidad por sexo]



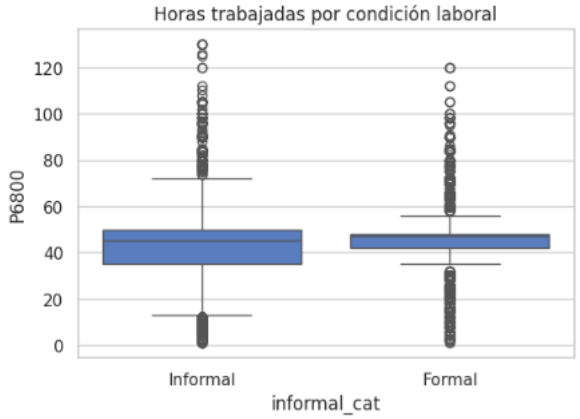
[Figura 3. Informalidad por nivel educativo]



[Figura 4. Heatmap sexo × educación]



[Figura 5. Boxplot de horas trabajadas por condición de informalidad]



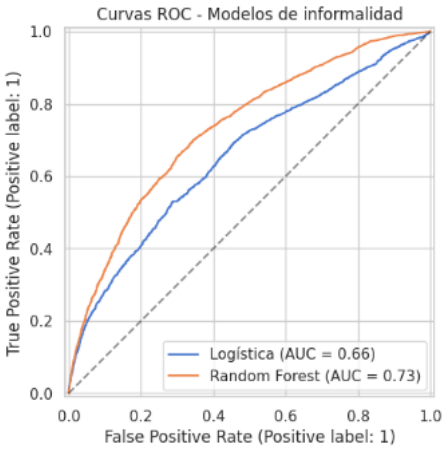
En la segunda etapa, se entra de lleno a los modelos de IA. Se arma una matriz de características donde conviven variables numéricas (edad, P6800) y categóricas (sexo_cat,

edu_cat). En el notebook de modelos esto se hace a través de un ColumnTransformer que aplica OneHotEncoder a las categóricas y StandardScaler a las numéricas.

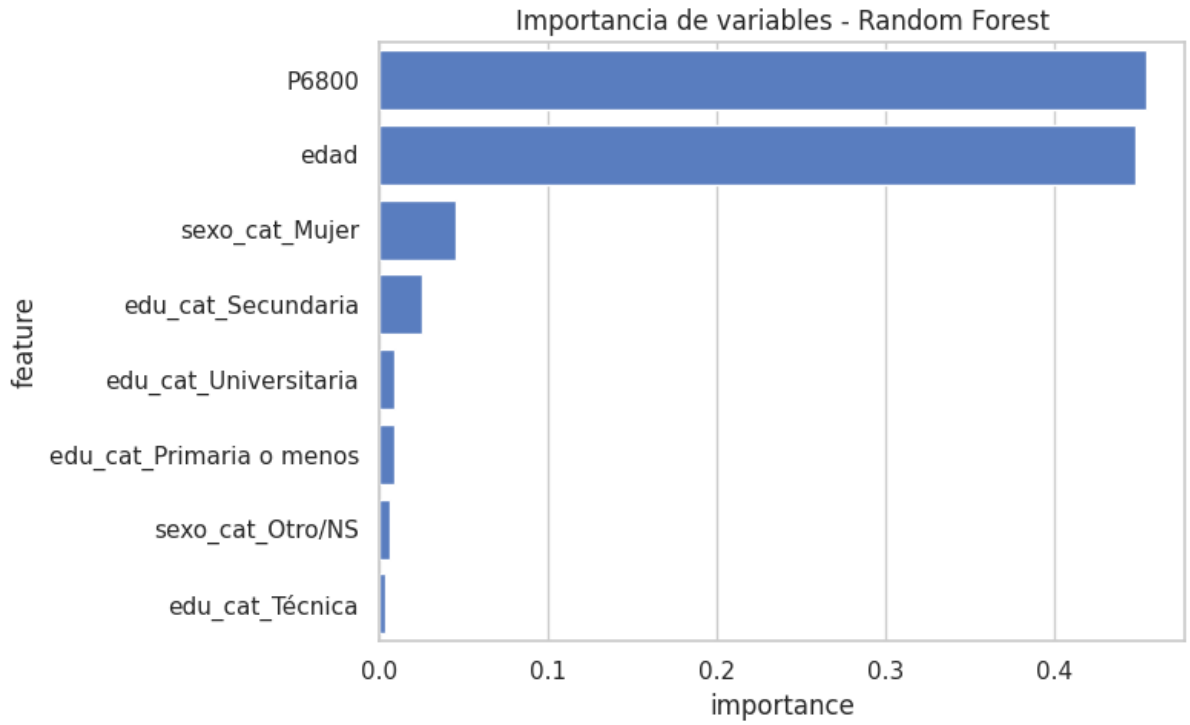
Luego se divide la muestra en conjuntos de entrenamiento y prueba (70% / 30%), usando estratificación para preservar la proporción de formales e informales (train_test_split(..., stratify=y)).

Sobre esta base se estiman dos modelos supervisados este Una Regresión Logística, que funciona como modelo base. Es útil para entender cómo cambian las probabilidades de informalidad cuando se pasa de un nivel educativo a otro, o de un sexo a otro, manteniendo el resto constante. Un Random Forest, que no asume linealidad y captura interacciones entre variables. El modelo se entrena dentro del mismo pipeline, y luego se evalúa con métricas estándar como accuracy, matriz de confusión y, sobre todo, área bajo la curva ROC (AUC).

[Figura 6. Curva ROC: Logit vs Random Forest]



[Figura 7. Importancia de variables en el Random Forest]



Finalmente, se extraen las importancias de variables del Random Forest para ver qué pesa más en la clasificación. Esto se refleja en un gráfico donde aparecen, ordenadas, las contribuciones de horas trabajadas, edad, sexo y niveles educativos a la predicción de informalidad, tal como se generó en el código.

En primer lugar, la informalidad total ronda el orden de tres de cada diez jóvenes ocupados. La barra correspondiente a “Informal” en la Figura 1 no es pequeña ni marginal; es suficientemente grande como para hablar de un rasgo estructural, no de un comportamiento residual. La informalidad acompaña de forma constante la inserción laboral de la juventud bogotana.

Cuando se mira la informalidad por sexo (Figura 2), aparece el primer matiz clave ara nuestro ejercuci los hombres muestran una tasa de informalidad mayor que las mujeres, y el grupo identificado como Otro/NS concentra una proporción aún más alta. En términos políticos, esto indica que hay identidades que cargan con una vulnerabilidad adicional, que va más allá del título educativo.

El análisis por nivel educativo (Figura 3) confirma la intuición básica, pero con matices que valen oro. La proporción de informales es más alta entre quienes tienen primaria o menos y media, y se reduce en secundaria, técnica y universitaria. La educación sirve, sí, pero el detalle fino es que incluso entre quienes alcanzan nivel universitario la informalidad se mantiene alrededor de una cuarta parte. Además, los niveles técnicos y tecnológicos presentan tasas relativamente más bajas, lo que sugiere que ciertos programas de inserción rápida al mercado laboral son, en la práctica, más protectores que un título universitario que el mercado no está absorbiendo bien.

El cruce sexo × educación, sintetizado en el mapa de calor (Figura 4), muestra que la informalidad no se reparte homogéneamente. Los hombres con baja educación concentran niveles altos de informalidad; las mujeres, aun con mayor educación, logran reducir el riesgo, pero no lo eliminan; el grupo Otro/NS sufre tasas altas casi en cualquier nivel educativo. La idea de que "la educación por sí sola corrige las brechas" queda en entredicho.

Finalmente, los boxplots de horas trabajadas (Figura 5) son probablemente la imagen más cruda de la precariedad. Las y los informales tienden a trabajar más horas por semana que los formales, y la distribución muestra una cola larga de casos con 60, 70 horas o más. La combinación de jornadas extensas y ausencia de cotización en salud y pensión configura un escenario duro: una juventud que sostiene al mercado con mucho trabajo y poca protección.

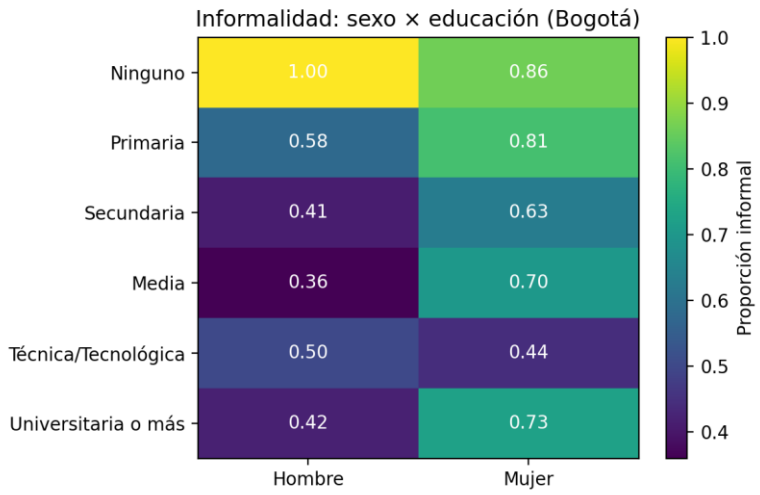
En conjunto, el EDA deja una imagen nítida y espectacular la educación ayuda, pero no alcanza; y la intensidad del trabajo y la posición en el mercado pesan tanto o más que el título.

La regresión logística alcanza un AUC cercano a 0.66 (Figura 6), lo que significa que el modelo clasifica mejor que el azar, pero no de manera perfecta. Los coeficientes –que se

recuperan en el notebook de modelos– confirman que mayores niveles educativos reducen la probabilidad de informalidad, mientras que ciertas combinaciones de sexo y educación la aumentan o la moderan. Sin embargo, el nivel de ajuste también deja claro que la realidad laboral juvenil es más compleja que una relación lineal del logit. Hay cosas que la GEIH no captura y que el modelo tampoco puede adivinar.

El Random Forest mejora el desempeño, con un AUC que se mueve alrededor de 0.73 (Figura 6). No es un oráculo, pero sí un modelo que captura mucho mejor las interacciones y no fuerza la realidad a un plano estrictamente lineal. Lo interesante no es solo el número, sino lo que sale cuando se miran las importancias de variables (Figura 7), Horas trabajadas (P6800) y edad aparecen como los predictores más influyentes en la clasificación entre formal e informal. Las variables asociadas a educación (edu_cat) y sexo (sexo_cat) también pesan, pero con un rol secundario frente a la combinación juventud + intensidad de trabajo.

En otras palabras, los modelos confirman lo que el EDA ya insinuaba: la informalidad no se explica únicamente porque alguien tenga “poca educación”. Se dispara cuando se combinan trayectoria educativa, sexo, edad y, sobre todo, condiciones concretas de trabajo como las horas semanales. La comparación entre Logit y Random Forest permite afirmar dos cosas a la vez, hays patrones regulares que la IA es capaz de detectar, interesante que seab invisibles a nuestros ojos, no estamos ante un caos completo y Hay componentes estructurales que no aparecen explícitamente en la encuesta (tipo de contrato, redes, poder de negociación, prácticas de contratación) y que dejan su huella en ese margen de error que el modelo no logra cerrar.



Heatmap sexo × educación

El mapa de calor resume lo anterior con claridad visual. A simple vista, los cuadros más oscuros están en las mujeres, y los más claros en los hombres, salvo pequeñas excepciones. Los extremos son reveladores, los hombres sin educación están en informalidad total (100%), mientras que las mujeres universitarias siguen con un 73%. El mensaje que deja este gráfico es casi doloroso, es la base de la política de Genero!, ni el máximo nivel educativo es capaz de blindar a las mujeres de la informalidad en Bogotá.

El resultado refuerza la idea de que la hipótesis central se cumplem, no es la educación lo que explica el fenómeno, sino la estructura del mercado laboral.

La hipótesis de este proyecto se confirma en Bogotá, incluso en los contextos de mayor concentración universitaria, la informalidad persiste. No es un problema de falta de educación. Es un problema estructural del mercado laboral colombiano, que combina discriminación de género, desajuste entre formación y demanda, y una economía que sigue precarizando a los jóvenes.

Al final, los datos y los modelos terminan diciendo lo mismo que intuíamos, pero ahora sin escapatoria estadística.

La informalidad juvenil en Bogotá no es un accidente ni una desviación menor, afecta a en torno a un tercio de los jóvenes ocupados, incluso en la ciudad que concentra buena parte del capital educativo del país. La narrativa de que "si estudias, el mercado te recibirá con un empleo digno" no queda respaldada en las cifras.

La educación sí reduce el riesgo de caer en la informalidad, pero no cumple la promesa de blindar al joven frente a la precariedad. Las tasas de informalidad entre quienes tienen educación universitaria lo evidencian. Al mismo tiempo, ciertos programas técnicos parecen ofrecer una protección relativa mayor, lo que habla de un sistema educativo desalineado con lo que el mercado realmente está dispuesto a contratar en condiciones formales.

Las diferencias por sexo e identidad muestran que la desigualdad de género y las violencias simbólicas también se expresan en el acceso a empleos con seguridad social. El hecho de que el grupo Otro/NS concentre más informalidad es una señal clara de exclusión, no de "malas decisiones individuales".

Las horas trabajadas emergen, tanto en los gráficos como en el Random Forest, como una señal fuerte de precarización quienes están por fuera de la protección social tienden a trabajar más, no menos. Es una juventud que sostiene trabajos largos, inestables y mal protegidos.

La hipótesis central del trabajo se confirma, el problema no es únicamente educativo. No basta con decirle a la juventud que estudie; la estructura del mercado laboral, las prácticas de contratación y la forma en que se diseñan los incentivos a la formalización están produciendo informalidad incluso entre quienes cumplen a rajatabla la narrativa del mérito.

Si la política pública quiere ser honesta, tiene que mover el foco. La cobertura educativa es necesaria, pero no suficiente. Hace falta entrarle al terreno incómodo de la regulación laboral juvenil, la inspección real, los incentivos a la formalización en sectores que se alimentan de mano de obra joven y, sobre todo, la garantía de protección social en las primeras experiencias de trabajo. Mientras eso no cambie, los modelos de IA seguirán diciendo lo mismo de que la informalidad no es una falla individual; es un diseño estructural.

La promesa de la profesionalización, esa idea de que "si estudias tendrás trabajo formal", no se cumple en los datos. Los resultados son una alertaorquwe ne necesitamos políticas que ataquen directamente las prácticas de contratación y las condiciones de empleo, no solamente la oferta educativa.

Bibliografías:

Aldana, E. A. (2025). Sistema simplificado de análisis del mercado laboral con GEIH 2024 (Trabajo de grado). Universidad Nacional Abierta y a Distancia.

Banco de la República. (2025). Nueva evidencia sobre la informalidad laboral y empresarial en Colombia (Ensayos sobre Política Económica, No. 108). Banco de la República.

Banco Mundial. (2013). Incluir a la juventud en el crecimiento: Retos del empleo juvenil en América Latina. Banco Mundial.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Chollet, F. (2021). Deep Learning with Python (2.ª ed.). Manning Publications.

Consej de Bogotá. (2025). Educación superior y mercado laboral: Brechas entre formación y empleo. Concejo de Bogotá.

Corficolombiana. (2025). Realidades del mercado laboral III: Jóvenes y empleo. Corficolombiana.

DANE. (2022). Nueva medición de la informalidad laboral en Colombia. Departamento Administrativo Nacional de Estadística.

DANE. (2024). Gran Encuesta Integrada de Hogares (GEIH) – Microdatos y documentación técnica (2024). Departamento Administrativo Nacional de Estadística.

DANE. (2024). Manual de recolección y conceptos básicos – GEIH 2024. Departamento Administrativo Nacional de Estadística.

DANE. (2024). Metodología de empleo informal y seguridad social: GEIH. Departamento Administrativo Nacional de Estadística.

DANE. (2024). Metodología general de la Gran Encuesta Integrada de Hogares (GEIH). Departamento Administrativo Nacional de Estadística

DANE. (2025). Boletín técnico: Mercado laboral de la juventud (Trimestre julio–septiembre 2025). Departamento Administrativo Nacional de Estadística.

Fundación Corona. (2024). 1 de cada 2 jóvenes se encuentran en informalidad o desempleo. Fundación Corona.

Fundación Corona & Aliados. (2025). Jóvenes, informalidad y brechas de acceso al empleo formal en Colombia. Fundación Corona.

Galvis, L. A. (2012). Informalidad laboral y calidad del empleo en las principales ciudades de Colombia. Revista de Economía del Rosario, 15(1), 1–45.

García, G. A., Guataquí, J., & Rodríguez, C. (2010). Educación, informalidad y retornos en el mercado laboral colombiano. Borradores de Economía, Banco de la República.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2.ª ed.). O'Reilly Media.

Google Research. (s.f.). Welcome to Google Colaboratory. https://colab.research.google.com

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2.ª ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R and Python (2.ª ed.). Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS 2017).

McKinney, W. (2022). Python for Data Analysis (3.ª ed.). O'Reilly Media.

Molnar, C. (2022). Interpretable Machine Learning.

OIT. (2023). Tendencias del empleo juvenil en América Latina y el Caribe. Organización Internacional del Trabajo.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Periódico UNAL. (2025). El título profesional no garantiza empleo ni formalidad. Periódico de la Universidad Nacional de Colombia.

ProBogotá Región. (2021). Bogotano 2051: Análisis del mercado laboral juvenil de Bogotá. Fundación ProBogotá Región.

Python Software Foundation. (2023). Python Language Reference (3.x). https://docs.python.org

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning (3.ª ed.). Packt Publishing.

Rincón-Báez, W. U., & Solear-Hurtado, A. J. (2015). Perspectiva socioeconómica de los vendedores informales de Chapinero, en Bogotá, Colombia. Cooperativismo y Desarrollo, 23(107), 93–110.

Robayo, C. D. C. (2019). Desempleo juvenil en Colombia: ¿La educación importa? Finanzas y Política Económica, 11(2), 247–268.

scikit-learn developers. (2023). Scikit-learn User Guide. https://scikit-learn.org

Secretaría Distrital de Desarrollo Económico. (2024). Panorama laboral de la población joven en Bogotá (junio–agosto 2024). Observatorio de Desarrollo Económico.

Secretaría Distrital de Planeación. (s.f.). Caracterización de la informalidad laboral en Bogotá. Alcaldía Mayor de Bogotá.

Tabares, L. C. (2025). Análisis de los factores que inciden en el desempleo juvenil en Bogotá durante 2024 (Trabajo de grado). Corporación Universitaria Minuto de Dios.

VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3–28.

Wooldridge, J. M. (2010). Introducción a la econometría: Un enfoque moderno (4.ª ed.). Cengage Learning.

Rincón-Báez, W. U. y Solear-Hurtado, A. J. (2015). Perspectiva socioeconómica de los vendedores informales de Chapinero, en Bogotá, Colombia. Cooperativismo y Desarrollo, 23(107), xx-xx. doi: http://dx.doi.org/10.16925/co.v23i107.1255