Implementación de un modelo de series de tiempo para predecir la tasa de desempleo de la población joven en Colombia en el corto plazo

Carmen Lucía Tangarife López

Asesor

Fernando Luis Carrascal

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

Este proyecto aplicado tiene como objetivo analizar el comportamiento del desempleo juvenil en Colombia entre enero de 2021 y diciembre de 2024 y predecir su tendencia, enfocado en la población joven entre 18 y 28 años, una de las más afectadas por el desempleo, especialmente las mujeres. Para abordar esta problemática, se utilizaron los microdatos mensuales de la Gran Encuesta Integrada de Hogares (GEIH) del Departamento Administrativo Nacional de Estadística (DANE), construyendo una base consolidada que permita estimar la tasa de desempleo juvenil y su evolución temporal. Con esta información se aplicó un modelo de series de tiempo de Promedios Móviles Integrados Autorregresivos Estacionales (SARIMA, por sus siglas en inglés), que corresponde a un enfoque estadístico de predicción temporal, con el fin de predecir el comportamiento futuro de la tasa de desempleo juvenil y evaluar su tendencia en el corto plazo. Los resultados esperados aportan evidencia empírica útil que ayude en la formulación de políticas públicas orientadas al trabajo decente y la inclusión laboral de esta población en el país.

Palabras claves: Tasa de desempleo, jóvenes, series de tiempo, modelo SARIMA, Colombia.

Abstract

This applied project aims to analyze youth unemployment trends in Colombia between January 2021 and December 2024, and predict future trends, focusing on young people aged 18 to 28, one of the groups most affected by unemployment, especially women. To address this issue, monthly microdata from the Large Integrated Household Survey (GEIH) of the National Administrative Department of Statistics (DANE) were used to build a consolidated database that allows for estimating the youth unemployment rate and its evolution over time. This information was used to apply a Seasonal Autoregressive Integrated Moving Average (SARIMA) time series model, which is a statistical approach to time series forecasting, in order to predict the future behavior of the youth unemployment rate and assess its short-term trend. The expected results provide useful empirical evidence to assist in the formulation of public policies aimed at decent work and labor inclusion for this population in the country.

Keywords**:** Unemployment rate, young people, time series, SARIMA model, Colombia.

Contenido

Lista de Tablas

Lista de Figuras

Introducción

El desempleo juvenil es una de las problemáticas socioeconómicas más persistentes en Colombia, afectando a un grupo demográfico que representa una proporción significativa de la fuerza laboral, como lo es la población entre los 18 y 28 años. La dificultad de acceso al mercado laboral para los jóvenes, especialmente para las mujeres, trae consigo varios problemas, pues no solo limita su autonomía financiera, sino que también profundiza desigualdades sociales, perpetúa ciclos de pobreza y debilita el tejido económico del país. Esta situación se agrava por fenómenos estructurales como la informalidad laboral, la falta de experiencia laboral, el nivel educativo y la desconexión entre la formación académica y la demanda del mercado.

Si la falta de empleos decentes persiste, las consecuencias pueden ser críticas: una generación desmotivada; mayor proporción de jóvenes que ni estudian ni trabajan, los llamados ninis; incrementos en los índices de pobreza y violencia juvenil; y en algunos casos, la vinculación a economías ilegales. Por esto, analizar el comportamiento del mercado laboral juvenil resulta fundamental para orientar políticas públicas efectivas que promuevan la inclusión laboral y el trabajo decente.

Dado lo anterior, el uso de modelos de series de tiempo se presenta como una herramienta robusta para comprender y anticipar las dinámicas del desempleo juvenil en el país. A partir del procesamiento y consolidación de los microdatos mensuales de la Gran Encuesta Integrada de Hogares (GEIH) del Departamento Administrativo Nacional de Estadística (DANE), correspondientes al período enero de 2021 a diciembre de 2024, este proyecto aplica el modelo SARIMA con el propósito de predecir la tasa de desempleo de los jóvenes y analizar su comportamiento temporal. Los resultados buscan ofrecer evidencia empírica que contribuya al

diseño de estrategias de política pública orientadas a mejorar las condiciones laborales de esta población.

Justificación

La población joven, entre 18 y 28 años, continúa siendo la que mayores tasas de desempleo registra en el país. Según el Departamento Administrativo Nacional de Estadística (DANE), durante el año 2024, la tasa de desempleo juvenil fue de 17,4%, superior en 7,2 puntos porcentuales a la reportada a nivel nacional que fue de 10,2%; siendo las mujeres jóvenes las más afectadas con una desocupación de 21,5% frente a 14,4% de los hombres jóvenes. El total de jóvenes desempleados llegó a 1.066.847 personas. (DANE, 2025).

Esta situación no solo limita el desarrollo personal y profesional de los jóvenes, comprometiendo su bienestar, sino que también genera consecuencias sociales como el incremento de la pobreza, el crecimiento de la informalidad, la exclusión social y económica de un grupo poblacional, la profundización de las desigualdades de género en el ámbito laboral, entre otras; por lo que, se necesita una atención prioritaria por parte del Estado, la academia y el sector productivo.

Para abordar esta problemática, es fundamental contar con herramientas analíticas que permitan comprender y anticipar las dinámicas del mercado laboral juvenil basadas en evidencia empírica. El uso de modelos estadísticos de series de tiempo tipo SARIMA, aplicados a los microdatos mensuales de la GEIH del DANE, ofrece la posibilidad de analizar tendencias históricas y predecir el comportamiento futuro de la tasa de desempleo juvenil. Este tipo de análisis contribuye a fortalecer la toma de decisiones informadas y al diseño de políticas públicas de trabajo decente orientadas a la inclusión laboral de los jóvenes.

Desde una perspectiva académica, este proyecto aplicado contribuye al estudio del mercado laboral colombiano mediante la implementación de técnicas de análisis predictivo de series temporales en Python, promoviendo una comprensión más profunda de las fluctuaciones

del desempleo juvenil y generando evidencia útil para la formulación de estrategias que buscan favorecer la inclusión laboral de los jóvenes en Colombia, contribuyendo así al bienestar social y al crecimiento económico del país.

Planteamiento del Problema

Dentro del mercado laboral colombiano, una de las problemáticas más complejas y persistentes está relacionada con la situación de desempleo de los jóvenes, ya que, a pesar de los esfuerzos gubernamentales e institucionales, son los que más altas tasas de desempleo registran, siendo las mujeres las más afectadas. Esta situación se agrava por una serie de factores estructurales como la informalidad, la desalineación entre la formación académica y las demandas del mercado, la falta de experiencia laboral, entre otros.

Aunado a esto, si bien existe una gran cantidad de datos recopilados por entidades del Estado, aún hace falta el uso de herramientas analíticas avanzadas que permitan aprovechar estos datos para anticipar cambios, por ejemplo, en la dinámica del empleo juvenil. Por esto, la aplicación de modelos estadísticos de series de tiempo tipo SARIMA resulta pertinente ya que permite analizar y predecir el comportamiento de la tasa de desempleo juvenil en Colombia contribuyendo así a generar evidencia empírica que apoye la formulación de políticas públicas más efectivas, orientadas a reducir el desempleo y mejorar las condiciones laborales de la población joven.

Objetivos

Objetivo General

Desarrollar un modelo de series de tiempo capaz de predecir el comportamiento de la tasa de desempleo de la población joven en Colombia, 18-28 años, usando una serie histórica entre 2021 y 2024, y comprender su evolución temporal.

Objetivos Específicos

Construir una base consolidada usando los microdatos de la GEIH para la población joven (18–28 años).

Calcular la tasa de desempleo juvenil, utilizando la fórmula que usa el DANE para medir este indicador.

Aplicar un modelo estadístico de series de tiempo tipo SARIMA y comparar con un modelo de red neuronal Long Short-Term Memory (LSTM).

Evaluar los modelos usando las métricas de validación y desempeño con el propósito de determinar la capacidad predictiva de estos en el corto plazo.

Marco de Referencia

Marco Teórico

El desempleo para la población joven continúa siendo una de las problemáticas más persistentes dentro del mercado laboral colombiano. Hay una serie de factores históricos, económicos y culturales que influyen en las condiciones laborales de esta población (Quintero et al.,2024). Factores estructurales como la informalidad laboral, la baja calidad educativa y la escasa experiencia profesional, junto con transformaciones tecnológicas y económicas, han limitado las oportunidades de empleo para los jóvenes.

Estudios recientes han destacado la utilidad de enfoques como el Machine Learning para analizar de manera más precisa las dinámicas del mercado laboral juvenil. Por ejemplo, el uso de modelos predictivos ha permitido comprender la interacción de diferentes generaciones dentro del mercado laboral, lo cual resulta crucial para diseñar políticas públicas diferenciadas y focalizadas (Alaql, Alqurashi & Mehmood, 2023).

A nivel nacional, se han identificado barreras significativas para la inserción laboral de los jóvenes, como la falta de experiencia, el limitado acceso a educación de calidad y la prevalencia de empleos informales (Baquero Márquez, Soto Acosta & Luna Moran, 2024). Estas condiciones dificultan la transición efectiva del sector educativo al mercado de trabajo formal. Aunado a la falta de información precisa en el mercado laboral, que dificulta que oferta y demanda se encuentren, impidiendo una adecuada asignación del talento humano. En este sentido, herramientas como el Big Data y el Machine Learning pueden mejorar el diagnóstico, monitoreo y la formulación de políticas públicas más basadas en evidencia (Cárdenas Rubio, 2020).

Así mismo, promover el empleo juvenil ha sido vinculado con el cumplimiento de los Objetivos de Desarrollo Sostenible, dado su impacto en la reducción de la pobreza y el fortalecimiento del crecimiento económico inclusivo (Castillo, Da Silva & Monsueto, 2020). Por ejemplo, lograr empleo pleno y productivo y garantizar condiciones de trabajo decente para esta población.

También, se evidencia la persistencia de la informalidad y las brechas regionales y de género en el empleo juvenil (DANE, 2025; ANIF, 2024). Esta información es esencial para construir modelos de predicción basados en evidencia empírica. En cuanto a las políticas públicas, se ha advertido que las estrategias gubernamentales orientadas a la inclusión laboral juvenil presentan logros limitados, lo que refuerza la necesidad de herramientas tecnológicas que mejoren su efectividad (Ham, Maldonado & Guzmán-Gutiérrez, 2021).

Desde la perspectiva educativa, los factores asociados a la deserción universitaria inciden directamente en la inserción laboral posterior, lo cual revela una débil articulación entre el sistema educativo y el mundo del trabajo (Gómez García, 2024). No obstante, mejorar los niveles educativos no garantiza por sí solo empleos formales y estables, lo cual requiere una estrategia integral que vaya más allá de la cobertura educativa (Castillo-Robayo & García-Estévez, 2019).

Estudios aplicados han demostrado que los modelos de Machine Learning permiten identificar demandas laborales específicas y anticipar escenarios de empleo, incluso en contextos de alta incertidumbre (Mamani Rodríguez, 2022; Orozco-Castañeda, Sierra-Suárez & Vidal, 2024).

Además, un grupo especialmente vulnerable es el de los jóvenes que ni estudian ni trabajan (ninis), quienes enfrentan altos niveles de exclusión social. Este grupo debe ser

priorizado en las políticas públicas de inclusión laboral (Laboratorio de Economía de la Educación, 2024). Además, se ha destacado la importancia del acompañamiento institucional durante la etapa del primer empleo, ya que esta transición inicial influye decisivamente en las trayectorias laborales futuras (Redondo et al., 2020).

Dado lo anterior, la aplicación de modelos de series de tiempo tipo SARIMA resulta adecuada para analizar y predecir la evolución del desempleo juvenil, al permitir identificar patrones de tendencia y estacionalidad en los datos temporales. Este enfoque, desarrollado a partir de la metodología Box-Jenkins, ha demostrado ser eficaz para anticipar el comportamiento de variables socioeconómicas, como las tasas de empleo y desempleo, proporcionando una base empírica sólida para la toma de decisiones (Box, Jenkins, Reinsel & Ljung, 2016; Hyndman & Athanasopoulos, 2021).

Marco Conceptual

Para el desarrollo de este proyecto aplicado se consideraron algunos conceptos clave los cuales se definen a continuación:

Desempleo Juvenil: se entiende como la situación en la que personas jóvenes, entre 18 y 28 años, económicamente activas, es decir, están en capacidad de trabajar, desean hacerlo pero no logran acceder a un empleo.

Mercado Laboral: conjunto de relaciones entre oferta y demanda de trabajo. Su análisis requiere considerar variables como tasa de participación, tasa de empleo, tasa de desempleo, informalidad, nivel educativo, localización geográfica, entre otras variables.

Big Data: hace referencia al análisis de grandes volúmenes de datos que, al ser procesados mediante herramientas analíticas avanzadas, permiten identificar patrones de

comportamiento. En el contexto del mercado laboral, facilita el monitoreo continuo de variables relevantes para el empleo juvenil (Cárdenas Rubio, 2020).

Trabajo Decente: según la Organización Internacional del Trabajo (OIT), el trabajo decente implica que las personas puedan acceder a un empleo productivo, en donde tienen la oportunidad de contar con ingresos justos, seguridad en el trabajo y protección social no solo para estas sino para sus familias. Esto sintetiza las aspiraciones de las personas durante su vida laboral y mejores perspectivas de desarrollo personal e integración social. Es un objetivo transversal para las políticas de inclusión laboral juvenil.

Políticas Públicas Laborales: conjunto de estrategias estatales que buscan regular, promover y mejorar las condiciones del empleo. En Colombia, estas deben adaptarse a las nuevas condiciones del mercado y a las capacidades tecnológicas disponibles para el análisis y toma de decisiones (Ham, Maldonado & Guzmán-Gutiérrez, 2021).

Ninis: jóvenes que ni estudian ni trabajan. Representan una población con alto riesgo de exclusión y marginalidad, y son prioritarios en las estrategias de focalización de empleo (Laboratorio de Economía de la Educación, 2024).

Series de Tiempo: son secuencias de observaciones registradas en intervalos de tiempo regulares, utilizadas para analizar patrones, tendencias y pronósticos de una variable a lo largo del tiempo. En el ámbito laboral, permiten estudiar la evolución de indicadores como la tasa de desempleo juvenil y anticipar su comportamiento futuro (Hyndman & Athanasopoulos, 2021).

Modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average): es una extensión del modelo ARIMA que incorpora componentes estacionales. Este modelo permite capturar patrones de tendencia y estacionalidad en los datos temporales, siendo ampliamente utilizado para pronosticar variables económicas y sociales, como las tasas de empleo y

desempleo. Su aplicación facilita la elaboración de predicciones precisas que pueden orientar decisiones de política pública (Box, Jenkins, Reinsel & Ljung, 2016).

Metodología

El desarrollo de este proyecto aplicado adoptó un enfoque cuantitativo y longitudinal, siguiendo los lineamientos de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta permitió estructurar el proyecto en un ciclo de vida riguroso que garantiza la calidad desde la comprensión de los datos hasta la obtención de pronósticos confiables. Las fases ejecutadas fueron las siguientes:

Comprensión y Preparación de los Datos

Se utilizaron los microdatos anonimizados de la Gran Encuesta Integrada de Hogares (GEIH) del DANE para el período comprendido entre enero de 2021 a diciembre de 2024. Se segmentó la población objetivo entre los 18 y 28 años de edad con representatividad nacional.

Dado que los datos tienen una periodicidad mensual y por módulos, se realizó un proceso de integración y depuración de la información.

Análisis Exploratorio y Diagnóstico

Con la base de datos unificada, se calculó la tasa de desempleo para la población joven de manera mensual, construyendo una serie temporal continua. Asimismo, mediante visualizaciones, se identificaron tendencias, ciclos y estacionalidad, verificando además la estacionariedad de la serie (o la necesidad de diferenciación), requisito clave para el modelado SARIMA.

Modelado Predictivo

Se procedió con el modelado predictivo con un modelo SARIMA, que combinó componentes autorregresivos, de medias móviles, diferenciación e integración estacional, ajustándose a la naturaleza mensual y cíclica del desempleo juvenil. Este se contrastó con un modelo de red neuronal Long Short-Term Memory (LSTM).

Evaluación y Validación de Desempeño

Para la validación de los modelos se aplicó la técnica de retrospectiva de origen fijo (Fixed-Origin Backtesting) y se compararon métricas de desempeño como el Error Medio Absoluto (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Porcentual Absoluto Medio (MAPE), seleccionando el modelo con mayor estabilidad predictiva a corto plazo.

Pronóstico e Interpretación

Finalmente, se generó la proyección de la tasa de desempleo juvenil para el primer trimestre de 2025. Se analizaron los resultados para identificar tendencias de corto plazo, transformando los datos en evidencia útil para la toma de decisiones y el análisis de políticas públicas.

Desarrollo

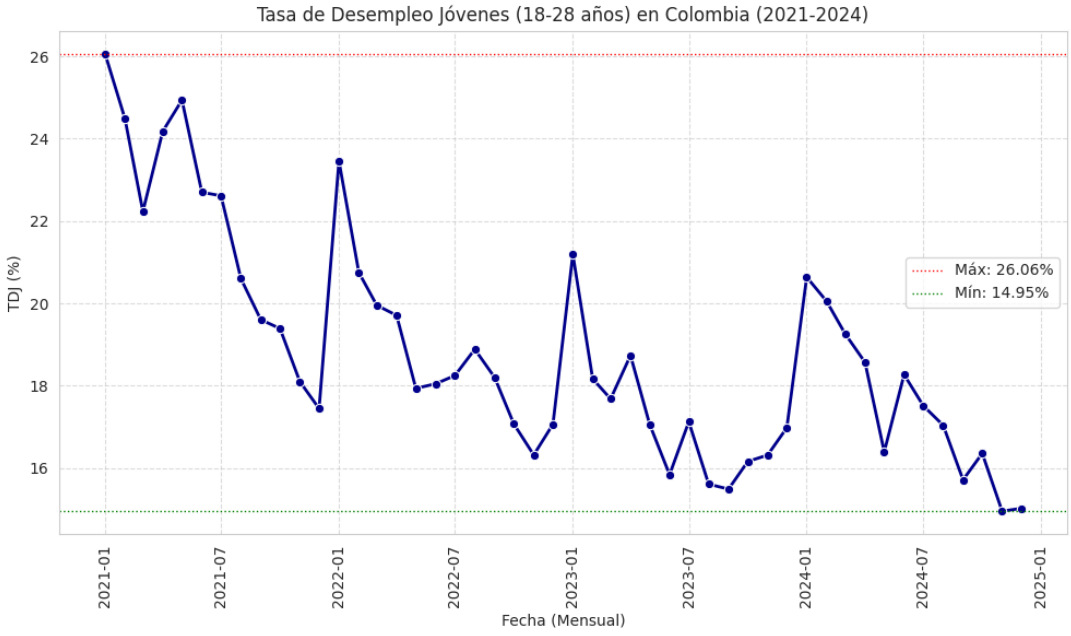
Inicialmente, se procede con el procesamiento de la información para encontrar el total de ocupados y desocupados en el rango de edad de 18 a 28 años, población objetivo, de manera mensual para el período de análisis. Además, se halla la Fuerza de Trabajo que corresponde a la sumatoria de ocupados y desocupados; lo anterior, para determinar la tasa de desempleo juvenil que corresponde a:

$$Tasa\ de\ Desempleo\ Jóvenes\ (TDJ) = \frac{Total\ desempleados\ jóvenes}{Fuerza\ de\ Trabajo} * 100 \quad (1)$$

El resultado, es una serie de tiempo de la tasa de desempleo juvenil mensual desde enero de 2021 a diciembre de 2024; un total de 48 meses.

Figura 1

Tasa de Desempleo Jóvenes (18-28 años) en Colombia. 2021-2024



Nota. Serie de tiempo de la tasa de desempleo juvenil mensual desde enero de 2021 a diciembre de 2024. Datos procesados a partir de los microdatos de la Gran Encuesta Integrada de Hogares (GEIH) del DANE.

En la Figura 1 se evidencia una trayectoria que va desde valores altos, superiores al 26%, a principios de 2021 hasta valores más bajos, cercanos al 15% a finales de 2024. Esto demuestra una tendencia a la baja de manera persistente, que refleja la recuperación económica y del propio mercado laboral posterior al impacto inicial de la pandemia.

También, se registra que los datos no se mueven de manera uniforme, es decir, estos presentan picos y valles recurrentes que se repiten cada 12 meses. Como se puede observar, los picos ocurren regularmente a principio de año, generalmente en los meses de enero y febrero, lo cual se puede asociar o con la finalización de contratos temporales de fin de año o con el flujo de estudiantes que salen a buscar trabajo. Por su parte, los valles ocurren típicamente a finales de cada año, noviembre y diciembre.

En este sentido, la serie de tiempo de la tasa de desempleo juvenil es no estacionaria debido a dos factores: 1) la tendencia decreciente y 2) la estacionalidad marcada por los picos a principio de año. La serie es no estacionaria en media, por lo que, se requiere aplicar la diferenciación para estabilizarla. Esta situación, demuestra que el utilizar un modelo SARIMA es una metodología clave para esta serie, pues este modelo está diseñado para manejar ambos tipos de no estacionariedad.

Sin embargo, para neutralizar estos factores de no estacionariedad, se aplicó una doble diferenciación; por un lado, la diferenciación regular ($d = 1$) para estabilizar la media y eliminar la tendencia decreciente y por el otro lado, la diferenciación estacional ($D = 1, s = 12$) aplicada con un período de 12 meses para eliminar los factores recurrentes de la estacionalidad anual.

Posteriormente, se utilizó la prueba Dickey-Fuller Aumentada (ADF) para validar que la serie transformada, es decir, doblemente diferenciada fuera estacionaria. Los resultados

obtenidos indican un p-valor de 0.0000; por lo que se rechaza la hipótesis nula de no estacionariedad, confirmando que la serie es plenamente estacionaria y apta para la aplicación de un modelo SARIMA.

$$ADF\ p-value\ (d=1, D=1, s=12): 0.0000 \quad (2)$$

Ahora bien, para identificar los órdenes óptimos del modelo SARIMA (esto es los parámetros *p, q P, Q),* con la serie estacionaria se analizaron los gráficos de Autocorrelación (ACF) y Autocorrelación Parcial (PACF).

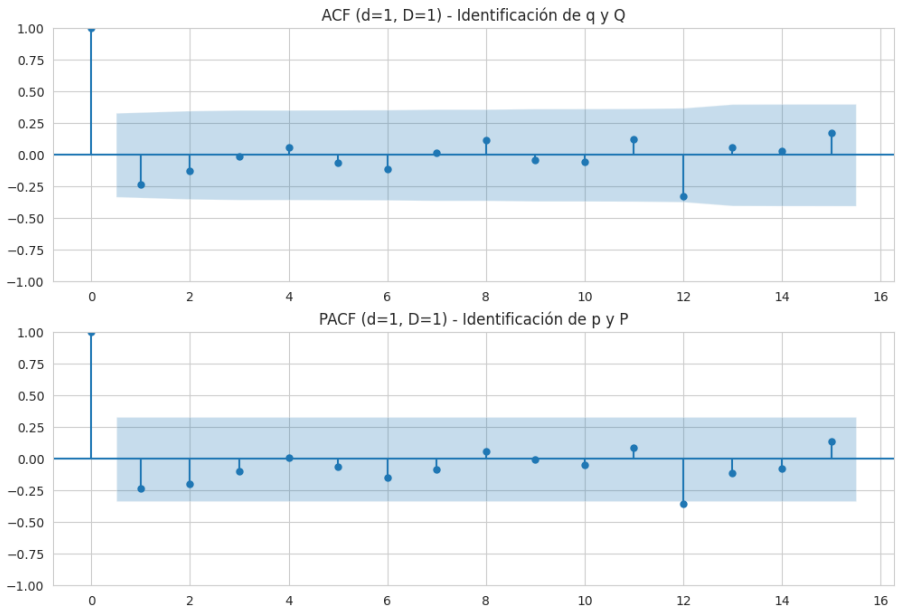
El uso de estos gráficos se hace entendiendo que:

I) el ACF funciona como una medida de la "memoria" de la serie, lo que permite identificar el orden de la Media Móvil (q,Q). En el gráfico se muestra la correlación entre el valor actual y sus valores pasados, incluyendo los efectos indirectos o que son heredados de períodos anteriores.

II) El PACF ayuda a determinar el orden del componente Autorregresivo (p,P). Este gráfico muestra únicamente la correlación directa entre el valor actual y un rezago específico, "limpiando" la relación, al eliminar la influencia de los pasos intermedios.

Figura 2

Gráfico de Autocorrelación (ACF) y Autocorrelación Parcial (PACF)



Nota. En el ACF y PACF las bandas azules representan los intervalos de confianza del 95%. La caída de las correlaciones dentro de estas bandas indica que la serie ha sido estabilizada mediante diferenciación.

En la Figura 2 se analiza el componente no estacional (corto plazo) y el componente estacional (largo plazo). Frente a la parte no estacional (primeros rezagos), este sugiere un término de Medias Móviles ($q = 1$) y un término Autorregresivo ($p = 1$) . Ahora bien, el componente estacional (rezago 12 y sus alrededores) sugiere un término de Medias Móviles Estacionales ($Q = 1$) y un término Autorregresivo Estacional ($P = 1$).

Así las cosas, el modelo siguiente es el que mejor se ajusta a las dependencias históricas de la tasa de desempleo juvenil, ya que no solo captura la tendencia inmediata del desempleo juvenil sino que también “aprende” de los ciclos históricos de cada mes.

$$SARIMA\ (1,1,1)\ *(1,1,1,12) \qquad (3)$$

Resultados

A continuación, el modelo se somete a estimación para obtener los coeficientes y validar los residuos.

Figura 3

Resultados Modelo SARIMA para la Tasa de Desempleo Juvenil

SARIMAX Results

Dep. Variable:	TDJ	No. Observations:	48
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-30.307
Date:	Wed, 19 Nov 2025	AIC	70.615
Time:	01:30:15	BIC	75.837
Sample:	01-01-2021	HQIC	71.748
	- 12-01-2024		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0538	0.639	0.084	0.933	-1.198	1.305
ma.L1	-0.4634	0.566	-0.819	0.413	-1.573	0.646
ar.S.L12	-0.3907	0.286	-1.366	0.172	-0.951	0.170
ma.S.L12	-1.0013	657.039	-0.002	0.999	-1288.775	1286.772
sigma2	0.6092	400.439	0.002	0.999	-784.237	785.455

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	1.94
Prob(Q):	0.75	Prob(JB):	0.38
Heteroskedasticity (H):	1.47	Skew:	0.34
Prob(H) (two-sided):	0.62	Kurtosis:	1.68

Nota. Resultados arrojados por el modelo SARIMA (1,1,1)*(1,1,1,12). Se evidencia que el modelo es metodológicamente robusto y está correctamente especificado.

Como se puede observar, se encontró un estadístico Ljung-Box (Q) de 0.10 con un p-valor = 0.75; un estadístico Jarque-Bera (JB) de 1.94 con un p-valor = 0.38 y un estadístico de Heteroskedasticity (H) de 1.47 con un p-valor = 0.62; los tres p -valores > 0.05. Lo anterior significa que: 1) los residuos son Ruido Blanco, es decir, el modelo ha capturado toda la estructura de la serie, tendencia, estacionalidad y autocorrelación; 2) normalidad en los residuos, esto es, los errores se distribuyen normalmente y 3) no hay evidencia de heterocedasticidad, esto es, la varianza del error es constante.

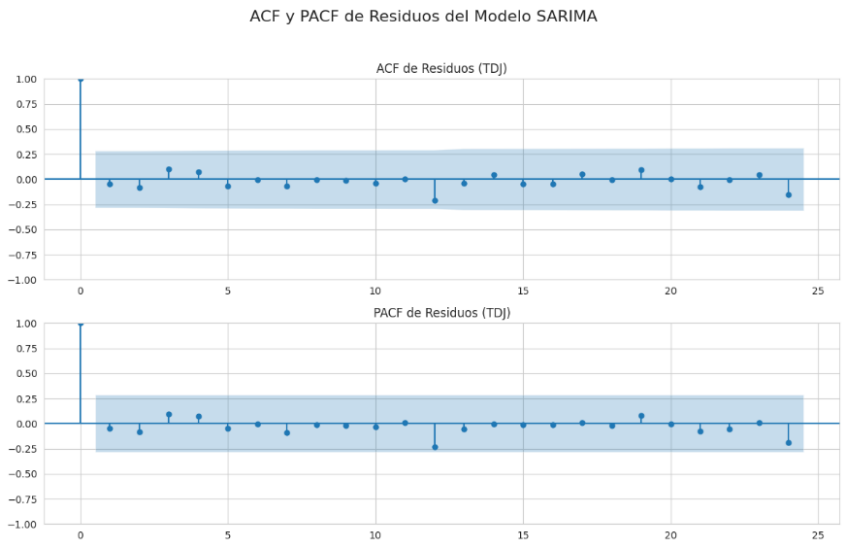
El modelo es metodológicamente robusto y está correctamente especificado, pues sus residuos cumplen con todos los criterios (Ruido Blanco, Normalidad y Homocedasticidad).

También, se puede observar que los p-valores de los coeficientes (AR, MA, SAR, SMA) son superiores a 0.05; por lo que, ninguno resultó ser estadísticamente significativo; lo que podría estar directamente relacionado con la muestra que es pequeña al contar solo con 48 observaciones.

A pesar de que el resumen del modelo cumplió con los p-valores requeridos, se realiza un análisis de diagnóstico visual para confirmar la correcta especificación del modelo y la robustez de los intervalos de confianza del pronóstico.

Figura 4

Validación de Correlación



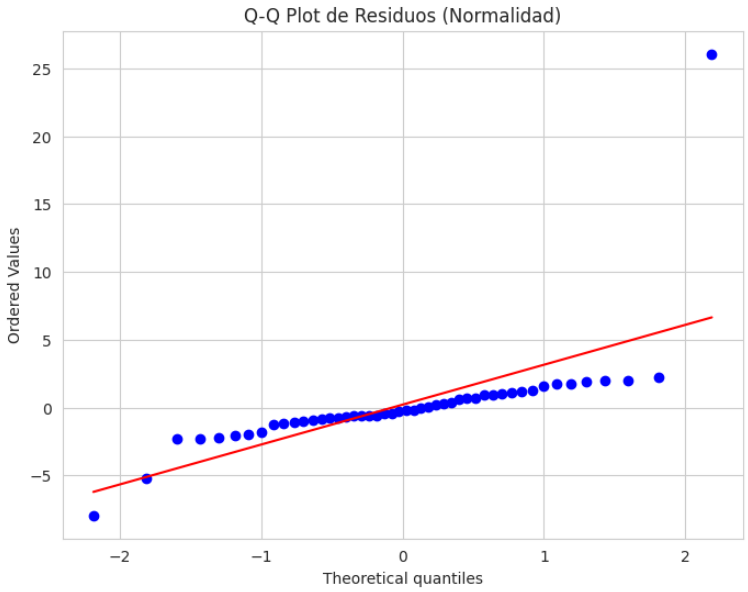
Nota. Corresponde al ACF y PACF de los residuos del modelo SARIMA en donde se evidencia que estos satisfacen los supuestos de correlación.

En la Figura 4 se observa que los residuos satisfacen los supuestos de ruido blanco (correlación) y normalidad. En el caso de la validación de ruido blanco la Función de

Autocorrelación (ACF) y la Función de Autocorrelación Parcial (PACF) de los residuos evidencian que ninguna de las barras cruza la banda de confianza; en este sentido, el modelo capturó de manera exitosa toda la información temporal finy la autocorrelación de la serie de la tasa de desempleo juvenil. Por lo tanto, los errores resultantes son independientes y aleatorios.

Figura 5

Validación de Normalidad



Nota. El gráfico Q-Q permite verificar el supuesto de normalidad, fundamental para asegurar que el modelo SARIMA no presenta sesgos sistemáticos.

En cuanto a la normalidad, la Figura 5 corresponde al Q-Q Plot y muestra como los puntos de los residuos se alinean de manera precisa sobre la línea diagonal de referencia (roja); esto confirma que los errores siguen una distribución normal con media cero, validando el supuesto requerido para que los Intervalos de Confianza del 95% que acompañan al pronóstico sean estadísticamente confiables.

Así las cosas, el modelo está correctamente especificado para proceder con la predicción de la tasa de desempleo juvenil a corto plazo.

Para validar la volatilidad (homocedasticidad) del modelo se realiza el Test de Ljung-Box aplicado a los residuos al cuadrado (Test ARCH). El test mostró un p-valor de ARCH de 1.0000, el cual es significativamente superior al nivel de significancia del 5%. Lo anterior, confirma la homocedasticidad de los errores, es decir, la volatilidad del error o la incertidumbre, en su pronóstico de la tasa de desempleo juvenil, es uniforme a lo largo del período, en lugar de variar drásticamente.

Con esta validación de homocedasticidad, el modelo SARIMA se constituye como la herramienta suficiente de predicción definitiva, por lo que, no es necesario considerar modelos de volatilidad más complejos como el GARCH.

El modelo anterior se comparó con un modelo $SARIMA\ (0,1,0)\ *(0,1,0,12)$; contrastando los criterios de información AIC y BIC (Akaike Information Criterion y Bayesian Information Criterion respectivamente) para ambos modelos. Es importante resaltar que el uso de estos estadísticos se hace para evitar el sobreajuste; por un lado, el AIC evalúa qué modelo minimiza la pérdida de información y ofrece una mejor capacidad de predicción; por otro lado, el BIC impone una penalización mayor por la inclusión de parámetros adicionales, favoreciendo la simplicidad. El modelo que registre el valor numérico más bajo de AIC y BIC es el que se considera el más adecuado, al tener un mejor equilibrio entre el ajuste a los datos y la simplicidad del modelo. Los resultados son:

Tabla 1

Criterios de Información AIC y BIC

Modelo	AIC	BIC
SARIMA (1,1,1) * (1,1,1,12)	70.615	75.837
SARIMA (0,1,0) * (0,1,0,12)	128.688	130.243

Nota. La tabla contiene los resultados de los criterios de información AIC y BIC para dos modelos SARIMA.

En este caso, los valores más bajos de AIC y BIC los registró el primer modelo, por lo que, se convierte en la mejor estructura que se ajusta a los datos históricos de la tasa de desempleo juvenil.

Posteriormente, se calculan las métricas a través de la validación del Backtesting (ver Tabla 2), en el que se usan los últimos 12 meses para la prueba y el resto para el entrenamiento (36 meses). Los resultados indican que:

1) el Error Medio Absoluto (MAE): 0.6924%; en promedio, el pronóstico de la TDJ se equivocó en 0.69 puntos porcentuales respecto al valor real.

2) la Raíz Error Cuadrático Medio (RMSE): 0.8858%; evidenciando una diferencia significativa con el MAE; en este caso, el modelo no falló uniformemente, sino que cometió errores de predicción mayores en algunos de los 12 meses de prueba (probablemente en los meses más lejanos del período de prueba). Es decir, tal vez, el modelo experimentó errores significativos en el largo plazo del horizonte de prueba; esto, debido al tamaño de la muestra.

3) el Error Porcentual Absoluto Medio (MAPE): 3.98%; el modelo tiene un error promedio cercano al 4% del valor real de la tasa de desempleo; por lo que, en este contexto de insuficiencia de observaciones, lo mejor es limitar el pronóstico a corto plazo, un horizonte de 3 meses.

Tabla 2

Métricas de Evaluación Modelo SARIMA

Métricas de Evaluación	Indicador
MAE	0.6924%
RMSE	0.8858%
MAPE	3.98%

Nota. La tabla contiene los resultados de las métricas de evaluación del modelo SARIMA (1,1,1) * (1,1,1,12)

Se procede con el pronóstico de la tasa de desempleo juvenil para los primeros 3 meses del año 2025; los resultados son los siguientes:

Tabla 3

Valores del Pronóstico de la Tasa de Desempleo Juvenil (18- 28 años) Colombia SARIMA

Fecha	Tasa de Desempleo Juvenil Pronosticada
Enero 2025	19.752881
Febrero 2025	16.934466
Marzo 2025	16.090897

Nota. La tabla contiene los pronósticos de la tasa de desempleo juvenil en Colombia para los tres primeros meses de 2025 modelo SARIMA.

La Figura 6 muestra la serie histórica de la tasa de desempleo juvenil, línea azul, y el pronóstico de esta para los próximos 3 meses, línea naranja. El área sombreada naranja corresponde al intervalo de confianza del 95%.

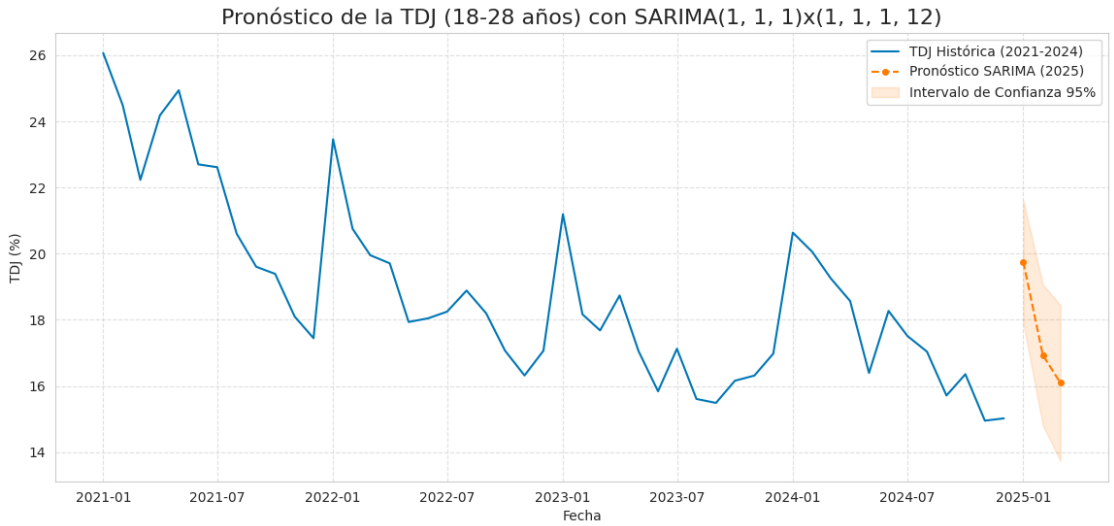
En este sentido, en la Figura 6 se confirma:

1) La confianza del corto plazo, la cual se evidencia en la estrechez del área sombreada en estos tres meses; demostrando que el modelo está muy seguro de su predicción.

2) La tendencia y estacionalidad. El modelo predice la continuación de los patrones que previamente se habían identificado: a) una continua tendencia decreciente de la serie histórica de la tasa de desempleo y b) la recurrencia del ciclo anual, mostrando un pico al inicio de año y luego una caída.

Figura 6

Resultados Modelo SARIMA para el Pronóstico de la Tasa de Desempleo Juvenil



Nota. La línea punteada naranja representa la proyección del modelo para los tres primeros meses de 2025. El área sombreada indica el intervalo de confianza, reflejando el margen de incertidumbre estadística del pronóstico.

Comparación con un Modelo Long Short-Term Memory (LSTM)

Con el ánimo de darle mayor validez a los resultados, se decide contrastar el modelo SARIMA con un modelo Long Short-Term Memory (LSTM). El modelo LSTM es una arquitectura de red neuronal recurrente especializada en datos secuenciales, como una serie de tiempo y están diseñada para capturar dependencias de largo plazo.

Luego de entrenar el modelo y de hacer las predicciones se procede con la evaluación del modelo, obteniendo lo siguiente (Ver Tabla 4).

Tabla 4

Métricas de Evaluación Modelo LSTM

Métricas de Evaluación	Indicador
MAE	0.7788%
RMSE	0.8548%
MAPE	4.52%

Nota: La tabla contiene los resultados de las métricas de evaluación del modelo LSTM.

Si se comparan los resultados del modelo LSTM con los arrojados por el modelo SARIMA (Ver Tabla 2) se identifica que ambos modelos son buenos, sin embargo, el modelo SARIMA presenta un MAE de 0.6924% y un MAPE 3.98% frente a 0.7788% y 4.52% respectivamente del modelo LSTM.

Estos resultados sugieren que, para el volumen de datos que se tiene (48 observaciones de 2021-2024), el modelo estadístico, dada su simplicidad y parsimonia, logra una mejor captura de la estacionalidad que la red neuronal.

Se procede con el pronóstico de la tasa de desempleo juvenil para los primeros 3 meses del año 2025 por el modelo LSTM; los resultados son los siguientes:

Tabla 5

Valores del Pronóstico de la Tasa de Desempleo Juvenil (18- 28 años) Colombia LSTM

Fecha	Tasa de Desempleo Juvenil pronosticada
Enero 2025	18.072376
Febrero 2025	19.249361
Marzo 2025	19.460716

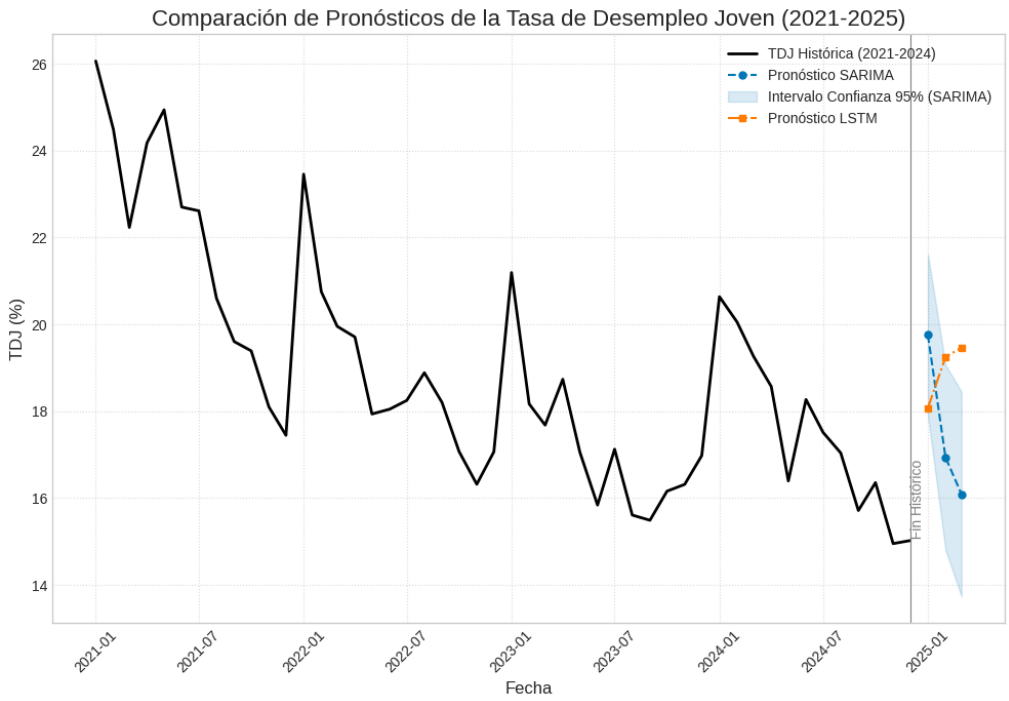
Nota. La tabla contiene los pronósticos de la tasa de desempleo juvenil en Colombia para los tres primeros meses de 2025 modelo LSTM.

La Figura 7 muestra la tasa de desempleo juvenil histórica para el período 2021-2024 y los pronósticos de esta con el modelo SARIMA y el modelo LSTM para los tres primeros meses de 2025. Se puede observar que el pronóstico del SARIMA (línea azul) conserva con mayor fidelidad los ciclos históricos de la serie, replicando la tendencia a la baja observada al inicio de períodos anteriores; mientras que, el modelo LSTM (línea naranja), presenta una proyección que tiende a ser más inercial, con una tendencia ascendente que no logra capturar la estacionalidad característica de la serie en esos primeros meses.

El hecho de que el modelo SARIMA presente un MAPE inferior (3.98% frente al 4.52% del LSTM) confirma que su estructura estadística captura con mayor precisión la dinámica del desempleo juvenil en el corto plazo.

Figura 7

Resultados Modelo SARIMA y LSTM para el Pronóstico de la Tasa de Desempleo Juvenil



Nota. Comparativa de proyecciones para el primer trimestre de 2025. La línea azul punteada representa el modelo SARIMA y la línea naranja punteada muestra la tendencia calculada por la red neuronal LSTM.

El modelo SARIMA predice una tasa de desempleo juvenil con un pico muy alto en enero (19.75%) y luego una caída de esta en febrero (16.93%) y marzo (16.09%), evidenciando un comportamiento que es estrictamente estacional, pues como se observa en la serie histórica, en el país el desempleo sube drásticamente en enero y tiende a bajar en los meses siguientes. Por su parte, el modelo LSTM predice una tasa de desempleo en enero más baja (18.07%) pero muestra una tendencia ascendente para los meses siguientes (19.24% en febrero y 19.46% en marzo).

Dado que el SARIMA presentó métricas de error inferiores (MAPE 3.98%) y una mayor coherencia con la serie histórica de la GEIH, se ratifica como el modelo más robusto para la toma de decisiones. Además, al ser un modelo más sencillo que explica bien los datos y al tener una serie con 48 observaciones mensuales, el modelo SARIMA es más eficiente, pues el modelo LSTM suele requerir grandes cantidades de datos de entrenamiento.

Así las cosas, se selecciona el modelo SARIMA como la herramienta para pronosticar la tasa de desempleo juvenil en el corto plazo, no solo por su menor error porcentual sino por su capacidad de interpretación de los componentes de la serie.

Conclusiones

La serie histórica de la Tasa de Desempleo Juvenil se confirmó como no estacionaria debido a dos factores: tendencia decreciente a lo largo del tiempo y una estacionalidad anual marcada (picos recurrentes en los primeros meses, enero/febrero); justificando el parámetro *s=12.* Se aplicó doble diferenciación que logró estabilizar la serie, el éxito de la estacionariedad fue validado por el p-valor ADF de 0.0000.

El modelo propuesto de SARIMA cumplió con la prueba de diagnóstico del test de Ljung-Box con un p-valor de 0.75. Esto confirma que los residuos son Ruido Blanco; así mismo, se confirmó que el modelo es robusto y capturó la tendencia, estacionalidad y autocorrelación de la serie; a pesar de que, debido a la muestra pequeña de solo 48 observaciones, los coeficientes no fueron individualmente significativos.

La precisión del modelo, validada a través del Backtesting en un horizonte de 12 meses, arrojó un MAE de 0.69% y un RMSE de 0.89%; esta destacada diferencia confirmó que el modelo cometía errores significativamente mayores en las predicciones a largo plazo; por lo que, se tomó la decisión de limitar el pronóstico de la tasa de desempleo juvenil a un horizonte de 3 meses, en donde la confianza es máxima.

El pronóstico de la tasa de desempleo juvenil confirmó la continuidad de la tendencia a la baja; pronosticando una tasa de desempleo juvenil de 16.09% para marzo de 2025. También, el modelo predice el pico estacional de inicio de cada año, proyectando una tasa de 19.75% para enero de 2025.

Aunque el modelo LSTM mostró una mayor estabilidad frente a errores extremos (menor RMSE), el modelo SARIMA resultó superior en precisión global, logrando un MAPE del 3.98%

frente al 4.52% del LSTM; por lo que, es eficiente como herramienta de predicción de la tasa de desempleo juvenil en el corto plazo.

La predicción de la tasa de desempleo juvenil es clave, ya que está anticipando, por ejemplo, el ciclo natural del incremento del desempleo al finalizar aquellos contratos de temporada de fin de año; información importante para el diseño de políticas públicas de inclusión laboral enfocadas en esta población.

Referencias Bibliográficas

Aguilar Rey, M. A., Cauda Ballen, J., Mizar Gómez, S., Quiñones Bonet, R., & Rivera Marín, M. A. (2021). *Impacto psicosocial del desempleo en jóvenes de 18 a 28 años* [Trabajo de grado, Universidad Autónoma de Bucaramanga]. Repositorio UNAB. https://repository.unab.edu.co/handle/20.500.12749/15340

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295. https://doi.org/10.3390/electronics9081295

Alaql, A. A., Alqurashi, F., & Mehmood, R. (2023). Multi-generational labour markets: Data-driven discovery of multi-perspective system parameters using machine learning. *Science Progress, 106*(4). https://doi.org/10.1177/00368504231213788

Asociación Nacional de Instituciones Financieras – Centro de Estudios Económicos. (2024). *Informe de empleo II* (Ed. 002). https://www.anif.com.co/wp-content/uploads/2024/07/anif-informe-empleo-ii-2024-1.pdf

Astorquiza, B. (2024). *Análisis: Desempleo juvenil en Colombia: ¿Hacia dónde va el mercado laboral?* Universidad de Manizales. https://umanizales.edu.co/umedia/analisis-desempleo-juvenil-en-colombia-hacia-donde-va-el-mercado-laboral-bilver-astorquiza

Baquero Márquez, V. J., Soto Acosta, A. C., & Luna Moran, I. F. (2024). Factores que afectan el desempleo juvenil en el Distrito de Santa Marta – Magdalena. *Documentos de Trabajo ECACEN, 1*, 158-168. https://doi.org/10.22490/ECACEN.8235

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley

Cárdenas Rubio, J. (2020*). Information problem in labour market and big data: Colombian case* (Working Paper No. WP2-2020-001). Alianza EFI–Colombia Científica. https://www.researchgate.net/profile/Jeisson-Cardenas/publication/340741730_Information_Problem_in_Labour_Market_and_Big_Data_Colombian_Case/links/5e9b511f92851c2f52ae53c2/Information-Problem-in-Labour-Market-and-Big-Data-Colombian-Case.pdf

Castillo, C., Da Silva, J., & Monsueto, S. (2020). Objectives of sustainable development and youth employment in Colombia. *Sustainability, 12*(3), 991. https://doi.org/10.3390/su12030991

Castillo-Robayo, C. D., & García-Estévez, J. (2019). Desempleo juvenil en Colombia: ¿La educación importa? *Revista Finanzas y Política Económica, 11*(1). https://doi.org/10.14718/revfinanzpolitecon.2019.11.1.7

Chang, W., Ji, X., Liu, Y., Xiao, Y., Chen, B., Liu, H., & Zhou, S. (2020). Analysis of university students' behavior based on a fusion K-means clustering algorithm. *Applied Sciences, 10*(18), 6566. https://doi.org/10.3390/app10186566

Departamento Administrativo Nacional de Estadística. (2021). *Medición de Pobreza Monetaria y Desigualdad - 2021* [Conjunto de datos]. Archivo Nacional de Datos - ANDA. https://microdatos.dane.gov.co/catalog/733

Departamento Administrativo Nacional de Estadística. (2022). *Medición de Pobreza Monetaria y Desigualdad - 2022* [Conjunto de datos]. Archivo Nacional de Datos - ANDA. https://microdatos.dane.gov.co/index.php/catalog/804

Departamento Administrativo Nacional de Estadística. (2023). *Medición de Pobreza Monetaria y Desigualdad - 2023* [Conjunto de datos]. Archivo Nacional de Datos - ANDA. https://microdatos.dane.gov.co/index.php/catalog/835

Departamento Administrativo Nacional de Estadística. (2024). *Medición de Pobreza Monetaria y Desigualdad - 2024* [Conjunto de datos]. Archivo Nacional de Datos - ANDA. https://microdatos.dane.gov.co/index.php/catalog/874

Departamento Administrativo Nacional de Estadística. (2025). *Boletín técnico: Mercado laboral de la juventud. Trimestre octubre - diciembre 2024* https://www.dane.gov.co/files/operaciones/GEIH/bol-GEIHMLJ-oct-dic2024.pdf

García Peña, M. A., & Morales Calderón, L. I. (2022). *Determinantes del desempleo juvenil en Colombia durante el periodo 2015–2019* [Trabajo de grado, Universidad Icesi]. Repositorio Universidad Icesi. https://repository.icesi.edu.co/items/a059e2ff-71f4-46bb-b945-faab9e7ae955

Gómez García, F. I. (2024). Factores de riesgo para la deserción estudiantil en instituciones de educación superior de Latinoamérica. *Salud, Ciencia y Tecnología, 4*, 592. https://doi.org/10.56294/saludcyt2024.592

Ham, A., Maldonado, D., & Guzmán-Gutiérrez, C. S. (2021). Recent trends in the youth labor market in Colombia: Diagnosis and policy challenges. *IZA Journal of Labor Policy, 11*, 7. https://doi.org/10.2478/izajolp-2021-0007

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

Laboratorio de Economía de la Educación. (2024). *Sin trabajo ni educación: El desafío de los ninis y su repercusión en el futuro de la sociedad (2023)* (Informe de análisis estadístico

LEE No. 99). https://www.javeriana.edu.co/recursosdb/5581483/11594517/INF-99.-NINIS-COLOMBIA-DATOS-2023-LEE2024.pdf

Mamani Rodriguez, Z. (2022). Machine learning process to determine the social demand for IT professional jobs. *Revista Industrial Data, 25*(2), 275–300. https://doi.org/10.15381/idata.v25i2.21643.g19078

Olaguibe, J. I. (2021). La transición de los jóvenes hacia el trabajo decente: Política social y empleo juvenil en España. *Estudios Latinoamericanos de Relaciones Laborales y Protección Social* 12, 39-50. https://dialnet.unirioja.es/servlet/articulo?codigo=8276486

Orozco-Castañeda, J. M., Sierra-Suárez, L. P., & Vidal, P. (2024). Labor market forecasting in unprecedented times: A machine learning approach. *Bulletin of Economic Research, 7*(4), 893–915. https://doi.org/10.1111/boer.12451

Quintero-Tomas, K. J., Jiménez-Lobo, D.L., Guerrero-Durán, J.A. & Solano-Cabarcas, M.J.(2024) Desempleo juvenil en Colombia: Unanálisis de tendencia bajo un modelo Arima. *Revista Ciencia & Sociedad, 4*(3), 223-235. https://www.cienciaysociedaduatf.com/index.php/ciesocieuatf/article/view/156/104

Redondo, M. I., Duque, C. C., Castaño, J. M., Ríos, K., & Tapasco, A. (2020). Inclusion of the first employment of young people in the labour market. *Revista Espacios, 41*(12), 290–301. https://research-ebsco-com.bibliotecavirtual.unad.edu.co/c/qcagk4/viewer/pdf/zshs765m7r