



Facultad de Ciencias Económicas y Empresariales
ICADE

Desempleo juvenil en Europa y su relación con el PIB: predicción por género y nivel educativo con modelos ARIMAX

Clave: 202011376

Doble Grado en ADE y Business Analytics (E-2+Analytics)

Autor: Maria del Carmen Gea Berruezo

Director: Maria José Martín Rodrigo

RESUMEN EJECUTIVO

El presente Trabajo Fin de Grado analiza el desempleo juvenil en Europa, centrándose en el impacto del género, el nivel educativo y el crecimiento económico. A través de un enfoque cuantitativo basado en modelos ARIMAX, se estudia la evolución de las tasas de desempleo juvenil entre 2006 y 2023, y se realizan predicciones hasta el año 2026. La metodología utilizada permite incorporar como variable exógena el Producto Interior Bruto (PIB) europeo.

El análisis revela diferencias significativas en las tasas de desempleo entre hombres y mujeres, así como entre distintos niveles educativos. Las mujeres y aquellos con menor nivel de formación (ED0-2) presentan mayores dificultades de inserción laboral, afectados tanto por barreras estructurales como por vulnerabilidades económicas. Los modelos predictivos muestran un incremento generalizado del desempleo juvenil en los próximos años, especialmente en los grupos más desfavorecidos, y confirman que el crecimiento del PIB tiene un efecto reductor sobre estas tasas, aunque con distinta intensidad según el subgrupo analizado.

Palabras clave: Desempleo juvenil, Europa, género, nivel educativo, PIB, modelos ARIMAX, desigualdad laboral.

EXECUTIVE SUMMARY

This project analyzes youth unemployment in Europe, focusing on the impact of gender, educational level, and economic growth. Using a quantitative approach based on ARIMAX models, the study examines the evolution of youth unemployment rates between 2006 and 2023 and provides forecasts up to 2026. The methodology allows for the inclusion of the European Gross Domestic Product (GDP) as an exogenous variable.

The analysis reveals significant differences in unemployment rates between men and women, as well as across educational levels. Women and those with lower levels of education (ED0-2) face greater challenges in accessing the labor market, affected by both structural barriers and economic vulnerabilities. The predictive models indicate a general increase in youth unemployment in the coming years, particularly among the most disadvantaged groups, and confirm that GDP growth has a reducing effect on these rates, although the magnitude varies by subgroup.

Keywords: Youth unemployment, Europe, gender, educational level, GDP, ARIMAX models, labor inequality.

Índice

Índice de figuras	5
Índice de tablas	5
1. Introducción.....	6
1.1. Contexto del desempleo juvenil en Europa.	6
1.2. Objetivo del estudio.....	7
2. Revisión de la literatura.....	8
2.1. Dinámicas de género en el desempleo juvenil.....	8
2.1.1. Desempleo juvenil masculino: sensibilidad a las crisis económicas.	8
2.1.2. Desempleo juvenil femenino: barreras estructurales persistentes.	9
2.2. Definición e influencia del nivel educativo.....	10
2.2.1. Nivel ED0-2: Menos que primaria, primaria y educación secundaria inferior.	10
2.2.2. Nivel ED3-4: Educación secundaria superior y post-secundaria no terciaria.	11
2.2.3. Nivel ED5-8: Educación terciaria.....	12
2.2.4. Consecuencias y problemas actuales: subempleo	13
2.3. Otros factores sociales que condicionan la inserción laboral juvenil.	14
2.4. Factores macroeconómicos y regionales.	16
2.4.1. Producto Interior Bruto (PIB).....	16
2.4.2. Automatización.	16
2.4.2. Otros factores.....	17
3. Explicación de los modelos.	18
3.1. Nociones básicas de los modelos: ARIMA y ARIMAX.....	19
3.1.1. Introducción a los modelos ARIMA.....	19
3.1.2. Extensión a ARIMAX.	20
3.2. Justificación de las variables analizadas.....	21

3.2.1. Género.	21
3.2.2. Nivel educativo.....	21
3.2.3. Crecimiento del Producto Interior Bruto (PIB).	21
4. Metodología del modelo.....	21
4.1. Objetivo.	21
4.2. Variables utilizadas.	22
4.3. Rango de años.....	23
4.4. Preprocesamiento de datos.	23
4.4.1. Manejo de valores ausentes.	24
4.4.2. Conversión a variables numéricas.	24
4.5. Ajuste de modelos.....	24
4.5.1. Selección de parámetros del modelo ARIMAX.	26
4.6. Técnicas de evaluación del modelo.	26
4.6.1. Métricas de error.	26
4.6.2. Validación de las predicciones para 2023.....	27
4.6.3. Análisis de residuos: test de Ljung-Box.	27
5. Resultados del modelo.....	28
5.1. Análisis descriptivo.	28
5.1.1. Tasas promedio de desempleo juvenil para hombres y mujeres.....	28
5.1.2. Tasas de desempleo juvenil para los diferentes niveles educativos.....	30
5.2. Modelos predictivos.	32
5.2.1. Modelos predictivos por género.	32
5.2.2. Modelos predictivos divididos por nivel educativo.....	36
6. Conclusiones.....	42
7. Anexos.	45
7.1. Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado	45

7.2. Código de R.....	46
7.2.1. Código género.....	46
7.2.2. Código nivel educativo.....	50
8. Bibliografía.....	54

Índice de figuras

Figura 1: Tasa de paro juvenil media de hombres, mujeres y diferencia a lo largo de los años; según nuestro dataset.....	29
Figura 2: Tasa de paro juvenil media de ED0-2, ED3-4, ED5-8 a lo largo de los años; según nuestro dataset.....	30
Figura 3: Tasas de paro juvenil históricas y predichas entre 2006 y 2026	34
Figura 4: Tasas de paro juvenil históricas y predichas entre 2006 y 2026.....	39

Índice de tablas

Tabla 1: Tabla comparativa de resultados del modelo según género (2023-2026).....	32
Tabla 2: Resumen de las métricas de error para modelos según género.....	35
Tabla 3: Tabla comparativa de resultados del modelo según nivel educativo (2023-2026).....	36
Tabla 4: Resumen de las métricas de error para modelos según nivel educativo.....	40

1. Introducción.

1.1. Contexto del desempleo juvenil en Europa.

El desempleo juvenil se define, de manera general, como la situación en la que personas de entre 15 y 24 años no logran acceder al mercado laboral, a pesar de estar disponibles y buscar empleo de forma activa (OCDE, s. f.). Históricamente, la tasa de desempleo juvenil ha sido considerablemente superior a la de la población adulta, lo que constituye una de las principales preocupaciones económicas y sociales a nivel mundial. De hecho, un aumento en la tasa de desempleo adulto se traduce en un incremento de más del doble en el desempleo juvenil (Bal-Domańska, 2022).

De acuerdo con la Organización Internacional del Trabajo (OIT, 2024), la tasa global de desempleo juvenil alcanzó un 13% el año pasado, lo que representa el nivel más bajo registrado desde el inicio del milenio. Sin embargo, las proyecciones indican que esta tasa podría aumentar en regiones como América del Norte y en las zonas norte, sur y oeste de Europa. En Europa, este fenómeno ha sido objeto de atención en la agenda de la Comisión Europea, dado su impacto no solo para el desarrollo económico y social actual, sino también para la estabilidad laboral de las futuras generaciones (Görlich, Stepanok y Al-Hussami, 2013).

En este contexto de incertidumbre, aparecen nuevos factores que aumentan la inquietud respecto al futuro profesional de los jóvenes. En los últimos años, la automatización y la digitalización han comenzado a transformar el mercado laboral, afectando de manera particular a este grupo poblacional. Diversos estudios han demostrado que los trabajadores jóvenes se concentran con mayor frecuencia en empleos con alto riesgo de automatización, debido a una menor cualificación y a la falta de experiencia profesional, lo que los convierte en un colectivo vulnerable ante la posible sustitución tecnológica (Pouliakas, 2018). Sin embargo, otros análisis subrayan que la digitalización, medida a través del índice DESI, también puede ser una herramienta útil para reducir el desempleo juvenil en los países de la Unión Europea, siempre que se acompañe de políticas de formación en competencias digitales (Başol, Sevgi & Yalçın, 2023). En este sentido, la automatización representa tanto una amenaza como una

oportunidad para el empleo juvenil, en función de la capacidad de los sistemas educativos y laborales para preparar a los jóvenes ante los desafíos del nuevo entorno digital.

1.2. Objetivo del estudio.

El presente Trabajo de Fin de Grado (TFG) tiene como objetivo principal analizar la influencia de diversas variables, entre ellas el género y el nivel educativo, en la tasa de desempleo juvenil en Europa y su evolución a lo largo del tiempo. Asimismo, se pretende explorar la interacción de estas desigualdades con factores macroeconómicos, destacando el Producto Interior Bruto (PIB).

Para lograr este propósito, se han definido los siguientes objetivos específicos, los cuales estructuran y guían el desarrollo de esta investigación:

1. Realizar una revisión de los estudios académicos existentes sobre el desempleo juvenil en Europa, con especial atención a las desigualdades tanto a nivel individual (género, nivel educativo, situación familiar) como macroeconómico (PIB, cultura regional).
2. Examinar las tasas de desempleo juvenil desde una perspectiva diferenciada por género y nivel educativo, identificando patrones históricos y tendencias recientes que permitan contextualizar las desigualdades observadas.
3. Desarrollar el modelo avanzado ARIMAX, para prever tendencias futuras del desempleo juvenil segmentadas por género y nivel educativo. Dichos modelos integrarán como variable exógena el crecimiento anual del PIB europeo, proporcionando una herramienta analítica para comprender la relación entre los factores individuales y macroeconómicos en el desempleo juvenil.

La elección del modelo ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) se basa en su capacidad para incorporar variables externas que influyen en la serie temporal analizada, lo que lo convierte en una herramienta útil para el objeto de este estudio. A diferencia de modelos ARIMA tradicionales, el enfoque

ARIMAX permite integrar indicadores económicos relevantes, como el PIB, mejorando la precisión de las predicciones al considerar tanto la evolución histórica de la tasa de desempleo juvenil como los factores macroeconómicos que la condicionan (Ben & Ben Salah, 2013; Woloszko, 2019). De este modo, se obtiene una comprensión más completa de las dinámicas del mercado laboral juvenil y se facilita la formulación de políticas públicas más eficaces.

2. Revisión de la literatura.

2.1. Desigualdades de género en el desempleo juvenil.

El género constituye un papel determinante en la comprensión del desempleo juvenil en Europa. A pesar de que tanto hombres como mujeres se enfrentan a desafíos específicos, las diferencias en las tasas de desempleo reflejan desigualdades estructurales que afectan de manera diferente a cada género. Mientras que los hombres muestran una mayor sensibilidad a las fluctuaciones económicas, las mujeres enfrentan barreras estructurales y sociales que limitan su acceso al mercado laboral. Ambos fenómenos están estrechamente relacionados con las características del mercado laboral y los patrones históricos de desempleo juvenil.

2.1.1. Desempleo juvenil masculino: sensibilidad a las crisis económicas.

El desempleo juvenil masculino se caracteriza por una alta sensibilidad a las fluctuaciones económicas, especialmente durante las crisis. En las fases iniciales de recesión, los hombres suelen experimentar mayores incrementos en sus tasas de desempleo, debido a su concentración en sectores económicos más vulnerables, como la construcción y la industria (European Training Foundation, 2024; Bal-Domańska, 2022). Este patrón está respaldado por la ley de Okun, que describe una relación directa entre el crecimiento económico y el desempleo, siendo esta relación más pronunciada en los hombres que en las mujeres (Bod'a y Považanová, 2020).

Sin embargo, esta mayor exposición a las crisis no se explica únicamente por el comportamiento macroeconómico, sino también por factores estructurales del mercado laboral. En economías avanzadas como la del Reino Unido, los hombres jóvenes enfrentan una mayor probabilidad de permanecer desempleados a largo plazo. Esto puede

deberse a que sectores tradicionalmente masculinos, como el de la construcción, tienden a ser más cíclicos, reforzando así la vulnerabilidad del empleo juvenil masculino en tiempos de crisis (Baussola y Mussida, 2017).

En contextos de recuperación económica, como ocurrió tras la crisis financiera de 2008, los hombres suelen beneficiarse más rápidamente del crecimiento económico, ya que los sectores donde están sobrerrepresentados tienden a reactivarse antes (Butkus et al., 2020). Este fenómeno resalta pues la necesidad de políticas laborales que diversifiquen las oportunidades de empleo para reducir su vulnerabilidad en épocas de recesión.

2.1.2. Desempleo juvenil femenino: barreras estructurales persistentes.

A pesar de los avances educativos alcanzados por las mujeres en Europa, que en muchos casos han superado a los hombres en términos de finalización de estudios superiores, las tasas de desempleo juvenil femenino siguen siendo consistentemente más altas. Esta disparidad pone de manifiesto que las barreras estructurales y sociales continúan limitando su acceso pleno al mercado laboral. Entre estas barreras destacan la segregación sectorial, la escasa representación en áreas estratégicas como STEM y la sobrecarga de responsabilidades familiares no remuneradas (European Training Foundation, 2024).

En países como la República Checa y Polonia, los estudios confirman que, aunque las mujeres con educación superior tienen mayores probabilidades de empleo que las de menor nivel educativo, siguen enfrentando desventajas frente a los hombres con cualificaciones equivalentes (Ryckowski y Zinecker, 2020). Esta desigualdad se ve agravada por normas sociales restrictivas y estereotipos de género que perpetúan la discriminación en el lugar de trabajo (European Training Foundation, 2024).

Otra barrera crítica es la discriminación laboral explícita, reflejada en prácticas como la preferencia de los empleadores por contratar a hombres debido a los costes asociados a la maternidad. Este fenómeno penaliza a las mujeres jóvenes, incluso cuando poseen cualificaciones iguales o superiores a las de sus homólogos masculinos (Stefanova-Lauerova y Terrell, 2002).

Asimismo, en contextos de crisis, las mujeres suelen verse afectadas de forma desproporcionada, especialmente en sectores como el turismo y los servicios, donde su presencia es mayoritaria. La pandemia de la COVID-19 constituye un claro ejemplo de este fenómeno, ya que aumentó las desigualdades de género en el empleo y puso de manifiesto la urgencia de implementar políticas que reduzcan estas brechas estructurales.

En definitiva, las desigualdades de género en el desempleo juvenil no solo responden a barreras históricas en el acceso al mercado laboral, sino también a una distribución desigual de las oportunidades en sectores clave. Estas diferencias justifican la necesidad de analizar por separado las tasas de desempleo entre hombres y mujeres, enfoque que se recoge en los modelos predictivos desarrollados en este estudio.

2.2. Definición e influencia del nivel educativo.

El nivel educativo es otro factor determinante en la inserción laboral de los jóvenes y, por ende, en las tasas de desempleo juvenil. A lo largo de la literatura académica, se ha destacado que el nivel de cualificación de los individuos tiene un impacto significativo en sus oportunidades de empleo, configurando patrones diferenciados de acceso al mercado laboral (Bal-Domańska, 2022). En el contexto europeo, la Clasificación Internacional Normalizada de la Educación (CINE) 2011, desarrollada por la UNESCO, proporciona una estructura estandarizada que permite estudiar cómo los niveles educativos influyen en la empleabilidad de los jóvenes (UNESCO, 2012).

2.2.1. Nivel ED0-2: Menos que primaria, primaria y educación secundaria inferior.

El nivel educativo ED0-2 comprende a aquellos jóvenes que han completado únicamente la educación básica o secundaria inferior, adquiriendo habilidades elementales como la lectura, la escritura y las matemáticas, pero sin finalizar la educación secundaria superior. Esta limitación educativa coloca a este grupo en una posición de desventaja significativa en el mercado laboral actual, caracterizado por una alta demanda de competencias técnicas y digitales avanzadas (Uslu y Tatli, 2024).

La falta de cualificaciones específicas hace que estos jóvenes enfrenten serias dificultades para acceder al empleo, especialmente en épocas de recesión económica, donde sus tasas de desempleo tienden a aumentar de manera desproporcionada en comparación con otros grupos educativos más cualificados (Butkus et al., 2020). Esta

vulnerabilidad no solo impacta en sus oportunidades laborales inmediatas, sino que también incrementa el riesgo de pobreza y exclusión social, perpetuando desigualdades intergeneracionales que afectan tanto a ellos como a sus futuras generaciones (Russell y O'Connell, 2001; Džunić et al., 2024).

Conscientes de esta problemática, las instituciones de la Unión Europea han implementado políticas para reducir el abandono escolar temprano, priorizando la retención de los jóvenes en el sistema educativo como una estrategia clave para mejorar su empleabilidad. Estas iniciativas han logrado avances significativos en la última década, evidenciando una tendencia positiva que, aunque alentadora, aún requiere esfuerzos sostenidos para abordar las desigualdades estructurales que persisten en este grupo (Bal-Domańska, 2022).

2.2.2. Nivel ED3-4: Educación secundaria superior y post-secundaria no terciaria.

El nivel ED3-4 comprende la educación secundaria superior y programas postsecundarios no terciarios, que proporcionan a los estudiantes una preparación técnica o vocacional orientada a las necesidades del mercado laboral. Esta formación específica dota a los jóvenes de habilidades prácticas que mejoran sus oportunidades de empleo, aunque no alcanzan el nivel de cualificación asociado a la educación universitaria (UNESCO, 2012).

En países como Alemania, el sistema educativo de orientación vocacional ha demostrado ser especialmente eficaz. Los jóvenes que completan este nivel suelen obtener diplomas de formación profesional que les permiten acceder a puestos cualificados muy valorados por los empleadores (Wolbers, 2007). Este fenómeno se refleja en que, a finales de la primera década de los 2000, las tasas de desempleo juvenil para este grupo alcanzaban niveles similares a los del nivel ED5-8 en algunos países, a pesar de las diferencias en la orientación educativa (Dietrich, 2012).

Sin embargo, esta efectividad no es uniforme en toda Europa. En países como Polonia, por ejemplo, el modelo vocacional no ha logrado los mismos niveles de éxito, lo que evidencia la influencia de factores estructurales y contextuales en los resultados de empleabilidad (Kogan et al., 2006).

La pandemia de COVID-19 exacerbó las dificultades para los jóvenes de este nivel educativo. Dado que los programas vocacionales dependen en gran medida de actividades prácticas presenciales, las restricciones sanitarias y la transición a la educación a distancia afectaron significativamente su continuidad y calidad. Además, el declive en los ingresos familiares durante este período dificultó que muchos estudiantes completaran su formación técnica (Bălan, 2021).

Por otro lado, un desafío persistente para este grupo es la falta de alineación entre algunos programas educativos de formación profesional y las demandas reales del mercado laboral. Este desajuste reduce las oportunidades de empleo y, en muchos casos, incrementa las tasas de desempleo juvenil incluso entre aquellos que han completado este nivel formativo (Uslu y Tatli, 2024).

2.2.3. Nivel ED5-8: Educación terciaria.

El nivel educativo ED5-8, que incluye la educación terciaria de ciclo corto, grados asociados, programas de doctorado y formación avanzada, presenta un perfil distintivo en relación con el mercado laboral. Los jóvenes que completan este nivel tienden a experimentar un tiempo de inserción laboral más corto que otros grupos. Según datos de la Organización Internacional del Trabajo, mientras que los graduados de educación primaria tardan, de media, 22.2 meses en encontrar empleo, los graduados de educación terciaria reducen este tiempo a 8.5 meses, lo que demuestra una ventaja inicial significativa (ILO, 2017).

Sin embargo, esta ventaja no asegura la estabilidad laboral, ya que el grupo se enfrenta a un desajuste constante entre las competencias adquiridas en la educación superior y las demandas específicas del mercado laboral. Este fenómeno, conocido como desajuste estructural, incrementa las tasas de desempleo juvenil incluso entre los más cualificados, debido a la falta de experiencia práctica que los empleadores demandan (Apaydın, 2018).

En respuesta a este problema, se implementó el Plan Bolonia, una iniciativa europea que busca armonizar los sistemas educativos del continente y fomentar la colaboración entre universidades y empresas. Este plan tiene como objetivo reducir el

desajuste entre la formación académica y las necesidades laborales, promoviendo una mayor exposición de los estudiantes a entornos empresariales desde etapas tempranas (Dietrich, 2012). Sin embargo, los resultados han sido desiguales. En países como Italia, las diferencias regionales limitan la efectividad de estas políticas. En el sur del país, por ejemplo, las altas tasas de cualificación no garantizan mejores oportunidades laborales debido a la falta de dinamismo económico en la región (Baussola y Mussida, 2017).

Aunque los jóvenes con educación terciaria son menos vulnerables a las fluctuaciones económicas, gracias a las competencias avanzadas que poseen (Butkus et al., 2020), el problema de la sobrecualificación se ha convertido en un obstáculo creciente. Muchos trabajadores con este nivel educativo se encuentran desempeñando empleos que no requieren las habilidades y formación adquiridas, lo que refleja un mal aprovechamiento del capital humano (O'Reilly et al., 2015).

2.2.4. Consecuencias y problemas actuales: subempleo

A pesar de que los niveles educativos más altos tienden a reducir el riesgo de desempleo, como se ha expuesto en los apartados anteriores, esto no garantiza el acceso a empleos acordes con la formación recibida. En este sentido, uno de los problemas emergentes más relevantes en el mercado laboral juvenil europeo es el subempleo, una situación en la que los jóvenes se ven empleados en condiciones inferiores a su capacidad o necesidades (OIT, 2018).

Este fenómeno se manifiesta principalmente de dos formas: por un lado, cuando los jóvenes trabajan menos horas de las deseadas (subempleo por insuficiencia de tiempo); y por otro, cuando desempeñan empleos que no requieren el nivel de cualificación que poseen (subempleo por infrautilización de competencias). Ambas situaciones conllevan una infrautilización del capital humano y limitan las posibilidades de crecimiento económico y personal (OIT, 2018).

Numerosos estudios han demostrado que el subempleo afecta tanto a jóvenes con baja cualificación como a aquellos con estudios superiores. En el primer caso, la falta de competencias demandadas por el mercado laboral les relega a empleos inestables. En el segundo, el desajuste entre la formación académica y las necesidades del mercado

laboral hace que muchos jóvenes sobrecualificados terminen aceptando trabajos por debajo de su nivel educativo, fenómeno cada vez más documentado en los países europeos (Cedefop, 2014; Butkus et al., 2020).

Las consecuencias del subempleo juvenil son muchas: reducción de ingresos, dificultad para independizarse, insatisfacción laboral, retraso en el desarrollo profesional, y en muchos casos, abandono del mercado laboral. En este sentido, mejorar la calidad del empleo juvenil exige ir más allá de la simple reducción de las tasas de desempleo. Es necesario implementar programas de orientación laboral, formación continua y adecuación entre la oferta educativa y las demandas del mercado, con el fin de garantizar una inserción laboral plena, estable y acorde al potencial formativo de los jóvenes europeos (OIT, 2018).

2.3. Otros factores sociales que condicionan la inserción laboral juvenil.

Además del género y de los niveles educativos, la inserción laboral juvenil está influida por una serie de factores sociales que, en conjunto, amplían o disminuyen las oportunidades de empleo para los jóvenes. Estos factores no solo actúan de manera individual, sino que interactúan entre sí, aumentando las desigualdades en diferentes contextos regionales y socioeconómicos.

- **Experiencia laboral**: La falta de experiencia laboral es uno de los obstáculos más significativos que enfrentan los jóvenes al buscar empleo. Los sistemas educativos que integran experiencias prácticas, como programas de formación dual o prácticas profesionales, contribuyen a reducir el desempleo juvenil al facilitar la adquisición de competencias prácticas relevantes. Por ejemplo, el Marco de Calidad para Prácticas Profesionales, adoptado por el Consejo de la Unión Europea en 2014, busca garantizar que los aprendices reciban experiencias laborales de calidad en condiciones seguras y justas, incrementando así su empleabilidad (Council of the European Union, 2014). Sin embargo, la implementación desigual de estos programas en los países miembros limita su impacto, especialmente en regiones con mercados laborales menos dinámicos.

- Lugar de residencia: El contexto geográfico también condiciona significativamente las oportunidades laborales de los jóvenes. Las áreas rurales suelen presentar mayores tasas de desempleo debido a la limitada disponibilidad de instituciones educativas de calidad y la escasez de empleos cualificados (European Training Foundation, 2024; Uslu & Tatli, 2024). Por otro lado, aunque las áreas urbanas ofrecen un mayor acceso a oportunidades laborales, en algunos países, durante la crisis de 2008, se observó que los jóvenes en zonas urbanas también enfrentaban altas tasas de desempleo debido a la saturación del mercado (Karamessini et al., 2019). Estas diferencias destacan la importancia de políticas que fomenten el desarrollo equilibrado entre áreas rurales y urbanas, minimizando las brechas regionales.
- Legado familiar: El entorno socioeconómico de las familias también desempeña un papel importante en las trayectorias laborales de los jóvenes. Los recursos económicos, el nivel educativo de los padres y el capital cultural transmitido dentro del hogar tienen un impacto directo en las oportunidades educativas y laborales. Por ejemplo, los jóvenes cuyos padres poseen niveles educativos más altos tienen menos probabilidades de estar desempleados, ya que acceden a una mejor orientación y apoyo en su transición al mercado laboral (Warmuth et al., 2014; Boudon, 1974). Además, los efectos intergeneracionales del desempleo son notables: el desempleo parental puede afectar negativamente las decisiones educativas de los hijos, perpetuando las desigualdades socioeconómicas (Lindemann y Gangl, 2019).
- Estado civil: El estado civil también influye en la vinculación de los jóvenes con el mercado laboral, especialmente en el caso de las mujeres. Las jóvenes casadas tienden a tener una menor participación laboral en comparación con las solteras, una tendencia acentuada en ciertos contextos culturales y económicos. Este fenómeno refleja desigualdades de género persistentes, que limitan el acceso de las mujeres a oportunidades laborales estables y de calidad (Booth, 2009; Booth, Francesconi, & Frank, 2003).

2.4. Factores macroeconómicos y regionales.

El desempleo juvenil en Europa está profundamente influido por factores macroeconómicos y regionales que, en muchos casos, condicionan de manera estructural las oportunidades laborales de los jóvenes. Estas variables no solo explican las disparidades entre países, sino también las diferencias dentro de una misma región. El análisis de estos factores resulta necesario para entender la dinámica del desempleo juvenil y su relación con el modelo predictivo que se aplicará posteriormente.

2.4.1. Producto Interior Bruto (PIB).

El crecimiento económico, medido a través del Producto Interior Bruto (PIB), es uno de los principales indicadores que afectan al desempleo juvenil. Según Dunsch (2017), las tasas de desempleo juvenil tienden a aumentar significativamente durante períodos de recesión económica, ya que los jóvenes son los primeros en ser afectados por las contracciones del mercado laboral. Este fenómeno es particularmente visible en sectores como el comercio, la hostelería y la construcción, donde predominan los contratos temporales y de corta duración.

Aunque el crecimiento del PIB contribuye a reducir el desempleo juvenil, estudios como los de Göçer y Erdal (2015) señalan que, por sí solo, no es suficiente para garantizar una inserción laboral equitativa. La relación entre el PIB y el desempleo juvenil es, además, desigual entre países, dependiendo de la estructura económica, el dinamismo del mercado laboral y la capacidad de las políticas públicas para mitigar los efectos de las crisis económicas. En el contexto del modelo ARIMAX que se realizará, el PIB actuará como variable exógena para comprender estas diferencias.

2.4.2. Automatización.

La automatización se refiere a la incorporación de tecnologías avanzadas, como robots e inteligencia artificial (IA), en los procesos productivos, con el fin de realizar tareas anteriormente hechas por humanos (Sostero et al., 2023). Este creciente fenómeno tiene un fuerte impacto en el mercado laboral juvenil europeo, tanto desde un punto de vista positivo como negativo.

Por un lado, la automatización puede sustituir a jóvenes trabajadores que ocupan puestos caracterizados por tareas rutinarias, pertenecientes principalmente a ED0-2 y ED3-4. De acuerdo con algunos estudios, aproximadamente el 14% de los trabajadores en la Unión Europea enfrentan un alto riesgo de que sus empleos sean automatizados, afectando especialmente a aquellos con menores niveles de cualificación (Pouliakas, 2018).

Por otro lado, la automatización también puede crear nuevas oportunidades laborales en sectores emergentes relacionados con la tecnología y la digitalización. Sin embargo, no todos los jóvenes pueden beneficiarse de estas oportunidades, sino únicamente aquellos con altas habilidades técnicas en competencias digitales (Başol, Sevgi & Yalçın, 2023). Por ello, es esencial que se considere, desde los organismos públicos, estrategias de adaptación de los sistemas educativos a las nuevas necesidades del mercado, para reducir los efectos de la automatización entre los jóvenes ubicados en categorías de educación ED0-2 y ED-3-4.

2.4.2. Otros factores.

- Pandemias y crisis globales: La pandemia de COVID-19 puso de manifiesto cómo los choques externos pueden agravar las desigualdades laborales entre los jóvenes. Durante este período, las tasas de desempleo juvenil aumentaron drásticamente debido a la paralización de sectores económicos esenciales y a la priorización de la retención de empleados con mayor antigüedad por parte de las empresas (Bălan, 2021). Estudios recientes, como el de Privara (2022), también señalan que indicadores como la tasa de mortalidad y la duración de las pandemias tienen una correlación directa con el desempleo juvenil. Estas dinámicas serán relevantes para el análisis predictivo del modelo, ya que permiten considerar escenarios de incertidumbre económica a futuro.
- Desigualdades regionales: Las desigualdades entre regiones europeas también juegan un papel crucial en las oportunidades laborales de los jóvenes. En regiones del norte y centro de Europa, caracterizadas por economías más diversificadas y mercados laborales dinámicos, las tasas de desempleo juvenil son considerablemente más bajas en comparación con el sur y el este del continente, donde las economías tienden a depender

de sectores más volátiles y menos tecnificados (Görlich et al., 2013). Además, factores como la corrupción, las infraestructuras insuficientes y la baja inversión en innovación tecnológica contribuyen a mantener estas desigualdades regionales. En el sur de Italia, por ejemplo, la falta de dinamismo económico limita las oportunidades laborales incluso para los jóvenes altamente cualificados (Tomić, 2018).

- Costes laborales y políticas públicas: El coste de contratación también afecta directamente al desempleo juvenil. En muchas regiones, los altos costes asociados a la contratación de nuevos empleados desincentivan la creación de empleo juvenil, especialmente en contextos de crisis. Esto se ve agravado por la rigidez de algunos mercados laborales, que dificultan la entrada de los jóvenes al mercado laboral formal (Bal-Domańska, 2022).

3. Explicación de los modelos.

A través de la revisión de la literatura se ha demostrado cómo factores individuales, sociales, macroeconómicos y regionales influyen en el desempleo juvenil europeo. Las desigualdades de género, el nivel educativo, y las condiciones económicas pueden amplificar o mitigar las barreras laborales que enfrentan los jóvenes.

En este contexto, la previsión de las tendencias de desempleo juvenil se posiciona como una herramienta para estudiar este problema. Si bien los modelos basados en series temporales permiten captar la dinámica histórica del fenómeno, la incorporación de variables externas como el PIB, a través de metodologías como ARIMAX, contribuye a mejorar la precisión del análisis y ofrecer una perspectiva más integral del problema (Dumičić, Čeh Časni y Žmuk, 2015).

Este enfoque ha sido respaldado en la literatura en el ámbito del análisis económico y social, especialmente cuando se busca modelizar fenómenos que dependen tanto de su comportamiento histórico como de factores contextuales que actúan desde el entorno. En consecuencia, ARIMAX ofrece un equilibrio entre robustez estadística y facilidad de interpretación, aspectos clave en el presente análisis.

Frente a otras alternativas posibles, ARIMAX presenta ventajas claras. Por ejemplo, los modelos VAR (Vector Autoregressive Models) permiten modelar múltiples

series endógenas de forma simultánea, pero requieren un elevado número de observaciones y suposiciones de estacionariedad conjunta que no siempre se cumplen en contextos sociales y económicos, como el del desempleo juvenil (Amoako, Kwofie, & Osei, 2023). Además, los modelos VAR están más orientados al análisis de relaciones dinámicas entre variables interdependientes, mientras que el objetivo de este estudio se centra en la predicción de una sola variable dependiente (la tasa de desempleo juvenil) a partir de una exógena.

Por otro lado, modelos como Prophet, desarrollados por Facebook, destacan por su facilidad de uso y robustez frente a cambios estacionales o festivos. Sin embargo, presentan limitaciones al incorporar variables exógenas de forma directa, lo que puede restringir su utilidad para analizar el impacto de factores macroeconómicos (Taylor & Letham, 2018). Los modelos de regresión lineal con retardos también fueron considerados, pero no capturan adecuadamente la estructura temporal y la autocorrelación que caracteriza a las series económicas (Keele & Kelly, 2006).

En este punto del TFG se desarrollará la metodología aplicada en este estudio, profundizando en los modelos y variables seleccionados para analizar y prever las dinámicas del desempleo juvenil en Europa.

3.1. Nociones básicas de los modelos: ARIMA y ARIMAX.

3.1.1. Introducción a los modelos ARIMA.

El modelo ARIMA (Autoregressive Integrated Moving Average) es una herramienta utilizada en el análisis y la predicción de series temporales, gracias a su capacidad para identificar patrones históricos y prever su evolución futura. Este enfoque permite modelar series temporales no estacionarias transformándolas en estacionarias mediante procesos de diferenciación, lo que facilita la identificación de relaciones autoregresivas y de medias móviles entre los datos (Ho y Xie, 1998).

El modelo ARIMA se descompone en tres componentes fundamentales:

- **AR (Autoregresivo):** Representa la relación entre el valor actual de la serie y los valores anteriores, permitiendo que los datos pasados influyan en las predicciones futuras.

- I (Integrado): Define el número de veces que una serie debe diferenciarse para alcanzar la estacionariedad, eliminando tendencias y fluctuaciones estacionales.
- MA (Media Móvil): Capta la relación entre el valor actual de la serie y los errores pasados en las predicciones, ajustando los rezagos mediante un promedio móvil.

La flexibilidad del modelo ARIMA lo convierte en una herramienta útil para el análisis de datos históricos cuando no se dispone de información sobre factores externos que puedan influir en la variable estudiada. Sin embargo, esta limitación se ve superada mediante extensiones como el modelo ARIMAX, que permite incorporar variables exógenas en el análisis, mejorando su adaptabilidad a contextos más complejos.

3.1.2. Extensión a ARIMAX.

El modelo ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) representa una extensión del modelo ARIMA que permite incorporar variables exógenas en el análisis de series temporales. Esta característica es especialmente útil para estudios en los que el comportamiento de la serie temporal no puede explicarse únicamente por sus valores históricos, sino que está influido por factores externos que afectan su evolución (Haydier, Ali y Salih, 2023).

El modelo ARIMAX combina los componentes autoregresivos, integrados y de media móvil del ARIMA con variables exógenas, que actúan como predictores adicionales. Su ecuación general se define como:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_r x_{rt} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q} + e_t$$

Donde:

- y_t es la variable dependiente en el tiempo t .
- $x_{1,t}, x_{2,t}, \dots, x_{r,t}$ son las variables exógenas en el tiempo t , con $\beta_1, \beta_2, \dots, \beta_r$ como sus coeficientes respectivos.
- $\phi_1, \phi_2, \dots, \phi_p$ son los coeficientes del componente autoregresivo (AR).
- $\theta_1, \theta_2, \dots, \theta_q$ corresponden a los coeficientes del componente de media móvil (MA).
- e_t es el término de error aleatorio en el tiempo t .

En este estudio, el modelo ARIMAX será utilizado para analizar las tasas de desempleo juvenil diferenciadas por género y nivel educativo, incorporando como variable exógena el Producto Interior Bruto (PIB) de la Unión Europea. Este enfoque permitirá explorar cómo el crecimiento económico afecta de manera diferenciada a los subgrupos analizados.

3.2. Justificación de las variables analizadas.

3.2.1. Género.

Como hemos comprobado previamente en la revisión de la literatura, las tasas de desempleo juvenil entre hombres y mujeres funcionan de manera diferente entre sí. En este estudio, el análisis por género permitirá identificar patrones diferenciales y evaluar cómo variables exógenas, como el PIB, afectan a hombres y mujeres.

3.2.2. Nivel educativo.

Del mismo modo, los estudios previos han demostrado que las tasas de paro varían significativamente entre los diferentes niveles educativos. Con este modelo, se busca analizar cómo evoluciona cada nivel educativo a lo largo del tiempo y cómo se relaciona con la variable exógena, el PIB.

3.2.3. Crecimiento del Producto Interior Bruto (PIB).

El crecimiento del Producto Interior Bruto (PIB) es una de las variables exógenas más relevantes en el análisis económico, al ser un indicador clave del desarrollo y la salud económica de un país. Tal y como reflejan estudios previos (Dunsch, 2017; Bal-Domańska, 2022), las variaciones en el PIB tienen un impacto significativo en diversos aspectos de la economía, incluida la tasa de desempleo juvenil, lo que demuestra la importancia de su inclusión en este modelo de análisis.

4. Metodología del modelo.

4.1. Objetivo.

Este modelo se centra en dos objetivos principales. En primer lugar, busca analizar la evolución futura de las tasas de desempleo juvenil, diferenciadas por género y nivel

educativo, con el fin de identificar tendencias y desigualdades en los próximos años. En segundo lugar, pretende estudiar la relación entre estas tasas y la variable exógena, el Producto Interior Bruto (PIB), para detectar patrones que expliquen cómo las fluctuaciones económicas afectan de manera desigual a los distintos subgrupos analizados.

4.2. Variables utilizadas.

El presente Trabajo de Fin de Grado se fundamenta en datos proporcionados por Eurostat, la oficina estadística oficial de la Unión Europea. Esta institución ofrece información estadística de alta calidad así como homogénea para todos los países, lo que asegura la comparabilidad de la información, convirtiéndola en una fuente fiable para realizar un análisis de las dinámicas laborales en Europa (Eurostat, n.d.).

En este estudio, se han empleado tres conjuntos de datos clave para analizar el desempleo juvenil desde diferentes perspectivas:

1. **Desempleo juvenil por género:** Este dataset, titulado *Youth unemployment rate by sex, age and NUTS 2 region (código: yth_empl_110)*, proporciona información detallada sobre la evolución de las tasas de desempleo juvenil desglosadas por género.

Los principales parámetros considerados son:

- Género: masculino y femenino.
- Grupo de edad: jóvenes de entre 15 y 24 años.
- Nivel geográfico: regiones NUTS 1 y NUTS 2 de Europa.
- Periodo: 2006-2023, abarcando el periodo más completo y consistente disponible para este análisis.

2. **Desempleo juvenil por nivel educativo:** El segundo dataset utilizado, titulado *Unemployment rates by sex, age, educational attainment level and NUTS 2 region (código: lfst_r_lfu3rt)*, aporta información crucial sobre las tasas de desempleo juvenil en función del nivel educativo. Los principales parámetros seleccionados son:

- Nivel educativo:
 - ED0-2: Educación baja (nivel primario o inferior).
 - ED3-4: Educación media (secundaria superior o postsecundaria no terciaria).

- ED5-8: Educación alta (terciaria).
- Grupo de edad: jóvenes de entre 15 y 24 años.
- Nivel geográfico: regiones NUTS 1 y NUTS 2 de Europa.
- Periodo: 2006-2023, consistente con el dataset anterior.

3. Producto Interior Bruto (PIB):

- Datos históricos (2016-2023): Se utilizó el dataset proporcionado por Eurostat titulado *GDP and main components (output, expenditure and income)*, identificado con el código de datos en línea *nama_10_gdp*. Este conjunto de datos contiene información detallada sobre el Producto Interno Bruto (PIB) de la Unión Europea y sus principales componentes.
- Predicciones: Se incluyen estimaciones del crecimiento del PIB para el periodo 2024-2026, proporcionadas por la Comisión Europea (2024). Estas proyecciones son esenciales para modelar la variable exógena en el análisis y evaluar su impacto en el desempleo juvenil.

4.3. Rango de años.

El análisis de este trabajo se basa en dos períodos diferenciados: un período histórico (2006-2022) y un período de predicción (2023-2026). El rango histórico comprende los datos utilizados para entrenar el modelo, permitiendo analizar tendencias clave en el desempleo juvenil por género y nivel educativo, así como el impacto de eventos como la crisis financiera de 2008 y la pandemia de COVID-19. Por otro lado, el período de predicción se centra en proyectar las tasas de desempleo juvenil mediante el modelo ARIMAX, incorporando como variable exógena las proyecciones de crecimiento del PIB de la Unión Europea. En este contexto, el año 2023 se utiliza como punto de validación del modelo, al comparar las predicciones con datos reales, lo que permite evaluar su precisión y capacidad para capturar las dinámicas del mercado laboral juvenil.

4.4. Preprocesamiento de datos.

Para garantizar un análisis coherente, los datos fueron organizados y estructurados en hojas de Excel, separando las tasas de desempleo por género y nivel educativo para cada año y región.

4.4.1. Manejo de valores ausentes.

En los datasets empleados, encontramos valores ausentes debido a la falta de información en determinados años o regiones. Para abordar esta problemática, se optó por el método de imputación utilizando la media ponderada del año correspondiente. Esta decisión se justifica porque:

1. Permite mantener la estructura temporal de los datos, considerando la media de ese año en particular, lo cual resulta crucial para analizar tendencias longitudinales.
2. Dado que el objetivo incluye el análisis de tendencias y la predicción de desempleo juvenil, la imputación temporalmente contextualizada asegura que los modelos no se vean afectados por lagunas de datos.

4.4.2. Conversión a variables numéricas.

Dado que los datos originales de las tasas de desempleo eran caracteres no numéricos, fue necesario convertir las columnas a formato numérico para permitir los cálculos y análisis posteriores. Este paso garantizó que los modelos predictivos y análisis estadísticos pudieran ejecutarse sin problemas.

4.5. Ajuste de modelos.

El desarrollo del modelo se llevará a cabo utilizando R Studio, un entorno de desarrollo integrado (IDE) diseñado para trabajar con el lenguaje de programación R. Este entorno es utilizado en el análisis estadístico, la ciencia de datos y el modelado predictivo, ya que proporciona herramientas para la escritura de código y visualización de datos. R Studio facilita la organización y ejecución de proyectos de investigación gracias a su interfaz intuitiva y su compatibilidad con una amplia gama de paquetes de R que permiten realizar análisis avanzados de forma eficiente y estructurada.

El primer paso en el desarrollo del modelo es descargar y cargar las librerías necesarias para ejecutar el análisis y procesar los datos en R. Las librerías utilizadas incluyen:

- **readxl**: Permite la lectura y manipulación de archivos Excel. Es fundamental para importar datos estructurados en hojas de cálculo, como los datasets que se emplearán en el análisis.

- **ggplot2:** Especializada en la creación de gráficos, esta librería facilita la visualización de patrones y tendencias a través de gráficos como líneas.
- **forecast:** Diseñada para el análisis y modelado de series temporales, esta librería es clave para ajustar modelos ARIMA y ARIMAX, y realizar predicciones basadas en patrones históricos y variables exógenas.

Una vez cargadas las librerías, el siguiente paso es importar los datos desde archivos Excel. Los datos principales, que contienen las tasas de desempleo desagregadas por género y nivel educativo, se cargan utilizando la función `read_excel` de la librería *readxl*. Estos datos se almacenan en un objeto denominado `df`. Adicionalmente, se importa un archivo con la información sobre el crecimiento anual del PIB de la Unión Europea, que se usará como variable exógena en los modelos. Posteriormente, se calcula la tasa de crecimiento anual del PIB a partir de los valores absolutos disponibles en el dataset, añadiendo proyecciones para los años 2024-2026.

Cabe señalar que, dado que el modelo ARIMAX empleado en este estudio incorpora una única variable exógena —el crecimiento anual del Producto Interior Bruto (PIB) de la Unión Europea—, no se considera necesaria la realización de una matriz de correlaciones entre variables explicativas para descartar problemas de multicolinealidad. En este caso, al limitarse a una única variable exógena, se evita este riesgo y se garantiza la robustez del modelo.

A continuación, se definen las series temporales correspondientes a las tasas promedio de desempleo para hombres, mujeres y los diferentes niveles educativos (ED0-2, ED3-4, ED5-8). Estas series se construyen calculando los promedios anuales de desempleo y convirtiéndolos en objetos de tipo `ts` (time series) en R.

Posteriormente, se ajustan modelos ARIMAX para cada grupo, utilizando el PIB como variable exógena. Los modelos se seleccionan automáticamente utilizando la función `auto.arima`, como se explica en el apartado siguiente.

Una vez ajustados los modelos, se realizan predicciones para el período 2023-2026, integrando las proyecciones de crecimiento del PIB. Las predicciones obtenidas se

combinan con los datos históricos para generar series completas que permitan analizar la evolución del desempleo y realizar comparaciones detalladas.

4.5.1. Selección de parámetros del modelo ARIMAX.

Una vez definidas y construidas las series temporales y ajustados los modelos ARIMAX para cada grupo de análisis, es importante detallar el proceso de selección de sus parámetros, dado el papel que desempeñan en la calidad de las predicciones.

En particular, se optó por utilizar la función `auto.arima()` del paquete `forecast` en R. Esta herramienta automatiza la identificación de los mejores valores para los componentes del modelo (p, d, q), que corresponden a los términos autorregresivos, de diferenciación e incorporación de medias móviles, respectivamente. La selección se realiza en función de criterios de penalización estadística como el Akaike Information Criterion (AIC) y el Bayesian Information Criterion (BIC), buscando siempre el modelo más parsimonioso (mejor explicación del dato con la menor cantidad de parámetros posible).

El uso de `auto.arima()` resulta especialmente útil en investigaciones como esta, donde se trabaja con múltiples series segmentadas por género y nivel educativo. Esta función permite asegurar que cada modelo esté correctamente adaptado a las características específicas de su serie temporal correspondiente, integrando el PIB como variable exógena (`xreg`) de forma coherente.

4.6. Técnicas de evaluación del modelo.

Con el fin de determinar la validez y la precisión de cada uno de los modelos creados se analizarán diferentes determinantes.

4.6.1. Métricas de error.

Entre las diferentes métricas de error a analizar encontramos el MAE, MSE y RMSE.

El Mean Absolute Error (MAE) es una métrica de evaluación utilizada en modelos estadísticos predictivos que mide el promedio de los errores absolutos entre los valores

reales y los valores predichos. El MAE indica, en promedio, cuánto se desvía el modelo de los valores reales, siendo una métrica intuitiva y fácil de interpretar. Su rango va de 0 (predicción perfecta) a valores positivos más altos, donde un MAE más bajo implica mejor desempeño del modelo.

El Mean Squared Error (MSE) es una métrica de evaluación utilizada en modelos estadísticos predictivos que mide el promedio de los errores al cuadrado entre los valores reales y los valores predichos. El MSE penaliza los errores grandes más que los pequeños debido al término al cuadrado, lo que lo hace útil para modelos en los que los errores grandes son menos deseables. Un MSE más bajo indica un mejor desempeño del modelo.

El Root Mean Squared Error (RMSE) es una métrica de evaluación utilizada en modelos estadísticos predictivos que mide la raíz cuadrada del promedio de los errores al cuadrado entre los valores reales y los valores predichos. El RMSE tiene la ventaja de estar en la misma escala que los valores reales, lo que facilita su interpretación. Al igual que el MSE, penaliza más los errores grandes que los pequeños. Un RMSE más bajo indica un mejor desempeño del modelo.

4.6.2. Validación de las predicciones para 2023.

Dado que los datos de desempleo juvenil para el periodo 2023 no han sido utilizados en el entrenamiento del modelo, estos serán utilizados para la validación del mismo. Este proceso permite medir la eficacia del modelo para realizar predicciones sobre nuevos datos, asegurando que no solo se ajusta bien a los datos de entrenamiento, sino que también ofrece resultados precisos y confiables en escenarios reales.

4.6.3. Análisis de residuos: test de Ljung-Box.

El test de Ljung-Box es una prueba estadística utilizada para evaluar si los residuos de un modelo estadístico, como una regresión o un modelo de series temporales, son independientes entre sí. Es decir, determina si existe autocorrelación significativa en los residuos en diferentes retardos (lags). El objetivo principal es verificar si los residuos de un modelo son ruido blanco, es decir, si no presentan patrones sistemáticos o correlaciones entre los valores a lo largo del tiempo.

Se plantea como hipótesis nula que no existe autocorrelación en los residuos hasta un determinado número de retardos, mientras que la hipótesis alternativa establece la presencia de dicha autocorrelación.

Para evaluar esta relación, se analizará el p-valor resultante del test de autocorrelación. Si el p-valor es mayor a 0.05, no se rechazará la hipótesis nula, lo que indicará independencia en los residuos. Por el contrario, si el p-valor es menor a 0.05, se concluirá que existen patrones significativos en los residuos, evidenciando la presencia de autocorrelación.

5. Resultados del modelo.

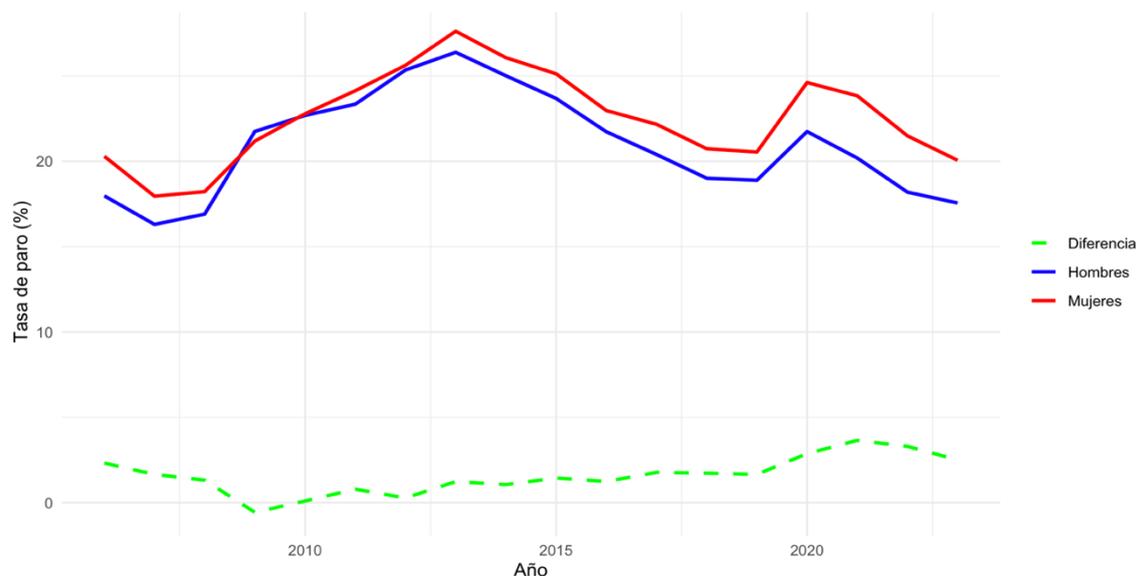
5.1. Análisis descriptivo.

Antes de proceder con la construcción del modelo de predicción temporal, es fundamental realizar un análisis descriptivo de los datos disponibles. Este paso inicial proporciona una visión general de las tasas de desempleo juvenil, permitiendo identificar tendencias, diferencias significativas y patrones relevantes en función del género y el nivel educativo. Este análisis preliminar no solo facilita la comprensión del comportamiento de los datos, sino que también sienta las bases para garantizar la consistencia y calidad del modelo predictivo.

5.1.1. Tasas promedio de desempleo juvenil para hombres y mujeres.

El análisis de nuestro dataset muestra que la tasa de desempleo juvenil en Europa es significativamente distinta entre hombres y mujeres. En promedio, la tasa de desempleo para los hombres jóvenes se sitúa en un 20.95%, mientras que para las mujeres jóvenes asciende a un 22.52%. Esto representa una diferencia promedio de 1.57 puntos porcentuales a favor de los hombres.

Figura 1: Tasa de paro juvenil media de hombres, mujeres y diferencia a lo largo de los años; según nuestro dataset.



Fuente: *Elaboración propia a partir de datos de Eurostat (2024).*

El análisis del gráfico revela varias tendencias históricas y eventos significativos en la evolución del desempleo juvenil por género. Antes de la crisis financiera de 2008, se evidenciaba una desigualdad marcada entre las tasas de desempleo juvenil de hombres y mujeres, con las mujeres consistentemente enfrentando tasas más altas. Esta disparidad subraya la persistencia de barreras estructurales que limitan el acceso equitativo al mercado laboral para las mujeres jóvenes.

Durante la crisis de 2008, ambas tasas de desempleo juvenil experimentaron incrementos significativos. Sin embargo, el impacto inicial fue más pronunciado en los hombres, lo que concuerda con teorías económicas previas que destacan la vulnerabilidad de los jóvenes varones en sectores cíclicos como la construcción, uno de los más afectados durante esta recesión (Butkus et al., 2020; Bal-Domańska, 2022).

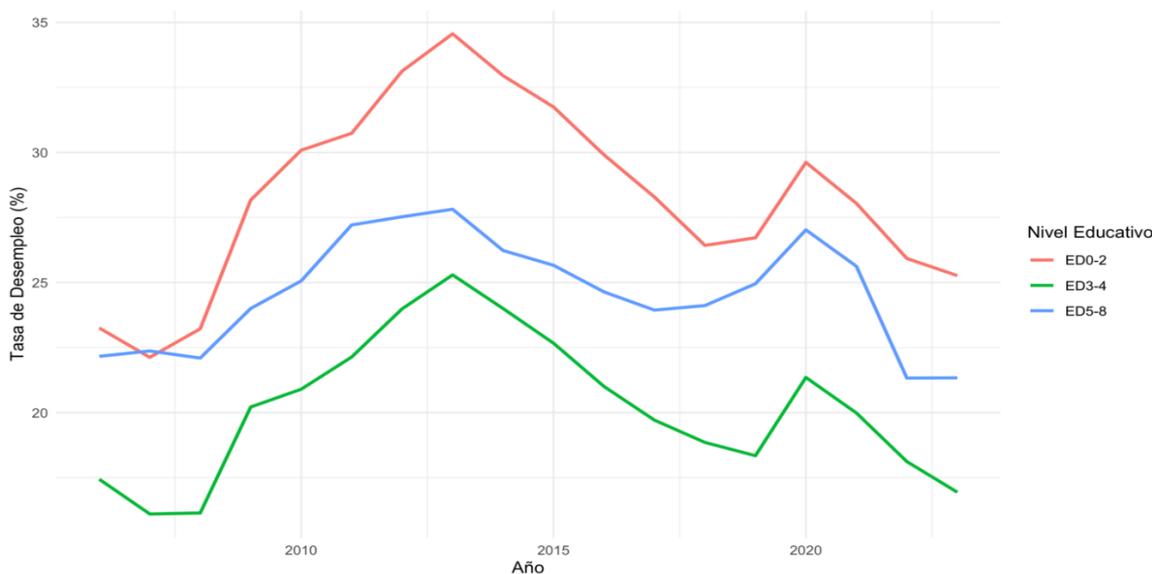
Tras la crisis, se observó una recuperación gradual en las tasas de desempleo juvenil para ambos géneros. Sin embargo, las mujeres continuaron enfrentando niveles de desempleo más elevados, lo que indica que la recuperación económica no benefició por igual a todos los grupos. Esto refuerza la idea de que las desigualdades de género son una constante en el mercado laboral, incluso en períodos de crecimiento económico, donde los hombres suelen reintegrarse más rápidamente al empleo formal.

Finalmente, la pandemia de COVID-19 supuso un nuevo punto de inflexión con un marcado aumento en las tasas de desempleo juvenil. En este caso, no solo se mantuvieron las diferencias entre hombres y mujeres, sino que estas se ampliaron. Este fenómeno puede estar vinculado a la sobrerrepresentación de las mujeres en sectores como el turismo, el comercio minorista y los servicios, que fueron particularmente vulnerables a las restricciones impuestas durante la pandemia.

5.1.2. Tasas de desempleo juvenil para los diferentes niveles educativos.

En el análisis del desempleo juvenil basado en nuestro dataset, se ha calculado la tasa promedio de desempleo para cada nivel educativo, con el objetivo de identificar patrones que reflejen la relación entre el nivel de formación y las oportunidades laborales. Los niveles educativos considerados en este estudio son: ED0-2 (educación básica o inferior), ED3-4 (educación secundaria o postsecundaria no terciaria) y ED5-8 (educación terciaria o superior). Los resultados obtenidos son los siguientes:

Figura 2: Tasa de paro juvenil media de ED0-2, ED3-4, ED5-8 a lo largo de los años; según nuestro dataset.



Fuente: Elaboración propia a partir de datos de Eurostat (2024).

Los resultados obtenidos destacan el gran impacto del nivel educativo en las tasas de desempleo juvenil, aunque también revelan ciertas particularidades que complementan la revisión de la literatura. Si bien la tendencia general indica que una mayor educación reduce el desempleo, el análisis muestra que los jóvenes con niveles educativos intermedios (ED3-4) presentan las tasas más bajas de desempleo, superando tanto a aquellos con educación básica (ED0-2) como a los de educación terciaria (ED5-8).

Este comportamiento puede explicarse, en parte, por la distribución regional de los datos en el dataset. Es importante destacar que no todas las regiones analizadas cuentan con datos completos para el nivel educativo ED5-8, lo que podría influir en los resultados generales del análisis. Entre las regiones con datos disponibles, sobresale el caso de Turquía, donde las tasas de desempleo juvenil para el nivel ED5-8 son significativamente más altas en comparación con los niveles educativos inferiores (ED0-2 y ED3-4). A pesar de ello, se observa que los jóvenes con educación terciaria presentan una mayor estabilidad en sus tasas de desempleo a lo largo del tiempo, lo que se alinea con las teorías revisadas en la literatura.

Por otro lado, los jóvenes con menor nivel educativo (ED0-2) enfrentan barreras estructurales importantes para acceder al mercado laboral, siendo este el grupo más afectado durante periodos de crisis económicas. Este comportamiento refuerza las teorías que señalan que la falta de cualificación limita significativamente las oportunidades laborales y amplifica la vulnerabilidad en momentos de recesión.

En contraste, los jóvenes con educación secundaria (ED3-4) parecen tener un perfil más alineado con las demandas del mercado laboral. Este nivel formativo ofrece competencias técnicas y profesionales que son altamente valoradas en sectores diversos, lo que explica sus mejores resultados en términos de empleabilidad. Sin embargo, este grupo también experimenta variaciones significativas dependiendo de las condiciones económicas generales, con picos de desempleo notables en periodos de crisis.

En conclusión, el análisis refuerza la fuerte correlación entre el nivel educativo y la vulnerabilidad al desempleo juvenil. Aunque los jóvenes con educación terciaria (ED5-8) demuestran mayor resiliencia, aquellos con educación básica (ED0-2) y secundaria (ED3-4) enfrentan desafíos particulares que se intensifican en contextos de crisis.

5.2. Modelos predictivos.

5.2.1. Modelos predictivos por género.

Tabla 1: Tabla comparativa de resultados del modelo según género (2023-2026).

Género	Predicción.2023	Predicción.2026	Incremento.Total	Intervalo.2023..IC.95..	Coef..PIB	Coef..AR.1.
Hombres	17.9%	20.6%	+2.7 pp	[16.2% - 19.6%]	-0.196	1.344
Mujeres	20.6%	22.4%	+1.8 pp	[18.7% - 22.5%]	-0.145	1.205

Fuente: Elaboración propia según modelo ARIMAX.

5.2.1.1. Predicciones de las tasas de desempleo para hombres y mujeres (2023-2026)

El modelo ARIMAX ajustado para las tasas de desempleo masculino predice una evolución relativamente estable, aunque con un incremento gradual a lo largo del período de predicción. Según las estimaciones, la tasa de desempleo masculino pasará de un 17.9% en 2023 a un 20.6% en 2026, lo que representa un crecimiento acumulado de 2.7 puntos porcentuales. Este aumento refleja la persistencia de desafíos estructurales en la integración laboral de los hombres jóvenes, especialmente en sectores cíclicos.

En el caso de las mujeres, las predicciones del modelo ARIMAX muestran también un incremento en las tasas de desempleo juvenil, aunque a un ritmo más moderado. Las tasas proyectadas comienzan en un 20.6% en 2023 y alcanzan un 22.4% en 2026, con un crecimiento acumulado de 1.8 puntos porcentuales. Este aumento, aunque inferior al de los hombres, refuerza la tendencia histórica de tasas más elevadas de desempleo juvenil femenino, probablemente influenciada por barreras estructurales y sectoriales que limitan su acceso a empleos de calidad.

Un aspecto relevante del modelo es el análisis de los intervalos de confianza, que permite evaluar la fiabilidad de las proyecciones. Por ejemplo, para 2023, las tasas de desempleo proyectadas para los hombres se sitúan dentro de un rango de entre 16.2% y 19.6%, mientras que para las mujeres oscilan entre 18.7% y 22.5%. Estos intervalos reflejan un nivel razonable de certidumbre en las estimaciones, lo que respalda la robustez del modelo ARIMAX para predecir tendencias diferenciadas por género en el desempleo juvenil.

5.2.1.2. Análisis de coeficientes: contribución del PIB como variable exógena.

El objetivo principal de este apartado es interpretar el impacto de cada componente del modelo ARIMAX, con especial atención a la contribución del PIB como variable exógena en las predicciones realizadas para el período 2023-2026. Los coeficientes estimados reflejan cómo las dependencias temporales (autorregresión y media móvil) y el crecimiento económico influyen en las tasas de desempleo juvenil por género.

Desempleo masculino:

- **Coefficiente autorregresivo (AR1):** El valor estimado de 1.344 indica una fuerte dependencia de los valores históricos de la serie temporal. Esto sugiere que la evolución previa del desempleo masculino tiene un peso considerable en su proyección futura.
- **Coefficiente del PIB:** El valor de -0.196 muestra que un crecimiento económico positivo reduce las tasas de desempleo masculino. Este efecto más pronunciado podría explicarse por la mayor participación de los hombres en sectores cíclicos como la construcción y la industria, que suelen beneficiarse más rápidamente de los períodos de expansión económica.

Desempleo femenino:

- **Coefficiente autorregresivo (AR1):** El coeficiente de 1.205 también indica una dependencia significativa de los valores históricos, aunque algo menor que en el caso masculino.
- **Coefficiente del PIB:** Con un valor de -0.145, el impacto del crecimiento económico en el desempleo femenino es menos marcado que en el masculino. Esto podría estar relacionado con factores como la segregación sectorial, las barreras culturales y las responsabilidades familiares no remuneradas, que limitan la capacidad de las mujeres para beneficiarse plenamente de las mejoras macroeconómicas.

Los coeficientes negativos asociados al PIB demuestran que un crecimiento económico positivo tiene un efecto reductor en las tasas de desempleo de ambos géneros, aunque con diferencias en su magnitud:

- Para los hombres, el impacto del PIB es más significativo, probablemente debido a su mayor presencia en sectores sensibles a los ciclos económicos.
- Para las mujeres, aunque el crecimiento económico también contribuye a la reducción del desempleo, su efecto es más limitado. Esto puede atribuirse a barreras estructurales que dificultan su plena integración laboral incluso en contextos de expansión económica.

Estos hallazgos subrayan la importancia de diseñar políticas públicas que no solo promuevan el crecimiento económico, sino que también aborden las desigualdades estructurales de género en el mercado laboral.

5.2.1.3. Visualización de predicciones y tendencias.

Figura 3: Tasas de paro juvenil históricas y predichas entre 2006 y 2026



Fuente: Elaboración propia según modelo ARIMAX.

Los resultados del modelo predicen que, en los próximos años, las desigualdades entre hombres y mujeres en las tasas de desempleo juvenil se mantendrán, reflejando barreras estructurales persistentes en el mercado laboral. Asimismo, se proyecta un incremento generalizado en las tasas de paro juvenil, alineándose con las tendencias identificadas en estudios previos.

5.2.1.4. Evaluación del modelo.

La evaluación del modelo ARIMAX se llevó a cabo utilizando diversas métricas de error y pruebas estadísticas que permiten validar su precisión y adecuación. Estas incluyen las métricas de error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y el test de Ljung-Box para evaluar la independencia de los residuos.

1. Métricas de error: Para los hombres, el modelo presentó un MAE de 0.96, un MSE de 1.43 y un RMSE de 1.20. Esto indica un nivel razonable de precisión en las predicciones, con una baja variabilidad en los errores respecto a los valores observados. En el caso de las mujeres, el modelo mostró un MAE de 1.11, un MSE de 1.82 y un RMSE de 1.35. Aunque los errores son ligeramente mayores que para los hombres, los valores se encuentran dentro de un rango aceptable, destacando la capacidad del modelo para capturar las tendencias generales en ambas categorías.

Tabla 2: Resumen de las métricas de error para modelos según género.

<i>MÉTRICA</i>	<i>RESULTADO PARA EL MODELO DE HOMBRES</i>	<i>RESULTADO PARA EL MODELO DE MUJERES</i>
MAE	0.96	1.11
MSE	1.43	1.82
RMSE	1.20	1.35

Fuente: Elaboración propia a partir del modelo ARIMAX.

2. Validación de predicciones para 2023: Las predicciones para el año 2023 fueron contrastadas con los valores reales de la tasa de desempleo. Para los hombres, el modelo predijo una tasa de 17.92%, cercana al valor real de 17.55%, mostrando una alta precisión. Para las mujeres, la predicción fue de 20.58%, frente a un valor real de 20.05%. Estas diferencias mínimas reflejan que el modelo puede prever con precisión las tasas de desempleo para ambos géneros.

3. Análisis de residuos: El test de Ljung-Box fue aplicado a los residuos del modelo para verificar su independencia. Los resultados (p-valores de 0.58 para hombres y 0.89 para mujeres) indican que no se rechaza la hipótesis nula de independencia de los residuos. Esto sugiere que el modelo no presenta patrones residuales significativos y que los errores están distribuidos aleatoriamente, confirmando la validez estadística del modelo.

5.2.2. Modelos predictivos divididos por nivel educativo.

Tabla 3: Tabla comparativa de resultados del modelo según nivel educativo (2023-2026).

Nivel.Educativo	Predicción.2023	Predicción.2026	Incremento.Total	Intervalo.2023..IC.95..	Coef..PIB	Coef..AR.1
ED0-2	25.3%	27.7%	+2.4 pp	[23.4% - 27.2%]	-0.169	1.404
ED3-4	17.8%	20.2%	+2.4 pp	[16.2% - 19.3%]	-0.198	1.337
ED5-8	20.1%	26.7%	+6.6 pp	[18.5% - 21.7%]	-0.080	1.182

Fuente: Elaboración propia según modelo ARIMAX.

5.2.2.1. Predicciones de las tasas de desempleo para los distintos niveles educativos (2023-2026)

El modelo ARIMAX ajustado para el nivel educativo ED0-2 sugiere un incremento gradual en las tasas de desempleo durante el período de predicción. En 2023, la tasa proyectada es del 25.3%, aumentando progresivamente hasta alcanzar el 27.7% en 2026. Esto representa un crecimiento total de 2.4 puntos porcentuales en este grupo, destacando su alta vulnerabilidad en el mercado laboral debido a la falta de cualificaciones avanzadas.

Para el nivel ED3-4, el modelo ARIMAX muestra una trayectoria moderada al alza. Las tasas de desempleo comienzan en un 17.8% en 2023 y alcanzan un 20.2% en

2026, con un incremento total de 2.4 puntos porcentuales. Este comportamiento refleja cierta estabilidad relativa, posiblemente asociada a la formación técnica o vocacional adquirida en este nivel.

En el caso del nivel ED5-8, las predicciones muestran un aumento más marcado en comparación con otros niveles. La tasa de desempleo proyectada en 2023 es del 20.1%, subiendo hasta un 26.7% en 2026, lo que equivale a un incremento de 6.6 puntos porcentuales. Este aumento subraya los desafíos específicos de este grupo, como la sobre cualificación y el desajuste entre habilidades y demandas del mercado laboral.

Como previamente hemos indicado, las predicciones incluyen intervalos de confianza que aportan una medida de fiabilidad en los resultados. En este caso, por ejemplo, para el nivel ED0-2 en 2023, las tasas proyectadas oscilan entre un 23.4% y un 27.2%, mientras que para ED3-4 se sitúan entre el 16.2% y el 19.3%, y para ED5-8 varían entre el 18.5% y el 21.7%. Estos rangos muestran un nivel razonable de certidumbre en las estimaciones realizadas.

5.2.2.2. Análisis de coeficientes: contribución del PIB como variable exógena.

El modelo ARIMAX para el nivel educativo ED0-2 indica que el coeficiente de la variable exógena (crecimiento del PIB de la UE) es de -0.169, lo que sugiere que un incremento en el PIB se asocia con una disminución moderada en la tasa de desempleo de este nivel educativo. Además, el coeficiente autorregresivo (AR1) es de 1.404, lo que implica que los valores pasados de la serie tienen una fuerte influencia en las tasas actuales de desempleo. Estos resultados coinciden con la evidencia que muestra que los jóvenes con menor nivel educativo son menos resilientes a cambios en el mercado laboral.

Para el nivel educativo ED3-4, el coeficiente asociado al crecimiento del PIB es de -0.197, indicando un impacto más significativo en comparación con ED0-2, probablemente debido a la mayor capacidad de este grupo para beneficiarse de una economía en expansión. El coeficiente autorregresivo (AR1) es de 1.337, mostrando también una alta dependencia de los valores pasados, aunque algo menor que en ED0-2.

Este nivel educativo parece responder, junto al ED0-2, significativamente más a las fluctuaciones económicas.

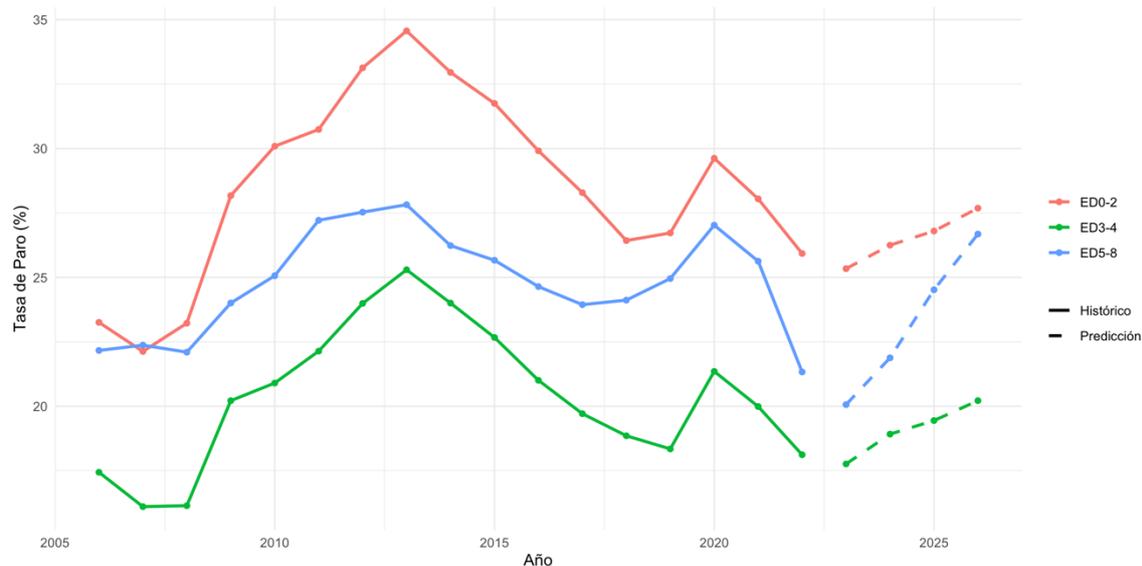
En el caso del nivel educativo ED5-8, el coeficiente del PIB es de -0.079, reflejando un impacto más limitado en comparación con los niveles inferiores. El coeficiente autorregresivo (AR1) es de 1.182, mostrando una menor influencia de los valores pasados en comparación con ED0-2 y ED3-4.

Los coeficientes autorregresivos (AR1) presentan valores elevados en todos los niveles educativos (1.404 en ED0-2, 1.337 en ED3-4 y 1.182 en ED5-8), lo que indica el gran peso de las dinámicas pasadas en la predicción del desempleo juvenil. Estos resultados indican que el comportamiento del desempleo en años anteriores influye notablemente en las estimaciones futuras, reforzando la importancia de los modelos que incorporan la dependencia temporal, como los ARIMAX utilizados en este estudio.

Por otro lado, los coeficientes asociados al PIB muestran diferencias en la sensibilidad al ciclo económico según el nivel educativo. Los niveles ED0-2 y ED3-4 presentan los coeficientes más negativos (-0.169 y -0.198, respectivamente), lo que indica una mayor reacción del desempleo juvenil en estos grupos ante las variaciones del crecimiento económico. En contraste, el grupo ED5-8 presenta un coeficiente menos pronunciado (-0.080), lo que sugiere que su desempleo es menos dependiente del contexto macroeconómico. Estos resultados respaldan las teorías expuestas en el marco conceptual, que apuntaban a una mayor vulnerabilidad de los jóvenes con menor nivel educativo ante los ciclos económicos.

5.2.2.3. Visualización de predicciones y tendencias.

Figura 4: Tasas de paro juvenil históricas y predichas entre 2006 y 2026.



Fuente: Elaboración propia según modelo ARIMAX.

Las tasas correspondientes a los niveles educativos ED0-2 y ED3-4 muestran un patrón muy similar, tanto en su evolución histórica como en las proyecciones futuras, lo que sugiere una respuesta homogénea a las dinámicas del mercado laboral en estos grupos. En contraste, el nivel educativo ED5-8 presenta un comportamiento distinto, posiblemente influenciado por las características de los datos en nuestro dataset, que reflejan economías menos estables y con menor capacidad predictiva.

A pesar de estas particularidades, las tendencias generales están alineadas con los hallazgos de modelos previos: se prevé un aumento en las tasas de desempleo juvenil en Europa durante los próximos años, lo que refuerza la necesidad de abordar las desigualdades estructurales a nivel educativo y regional.

5.2.2.4. Evaluación del modelo.

Como indicamos previamente, se han utilizados diferentes métricas para evaluar la validez de los distintos modelos.

1. Métricas de error:

- Para el nivel ED0-2, el modelo presentó un MAE de 1.15, un MSE de 1.78 y un RMSE de 1.33. Esto refleja una precisión aceptable, aunque ligeramente menor que los otros niveles educativos.
- En el caso de ED3-4, el modelo mostró un MAE de 0.84, un MSE de 1.14 y un RMSE de 1.07, lo que indica una excelente capacidad para capturar las variaciones en esta categoría.
- Para ED5-8, el modelo presentó los menores errores, con un MAE de 0.76, un MSE de 1.25 y un RMSE de 1.12, demostrando un ajuste sólido y fiable.

Tabla 4: Resumen de las métricas de error para modelos según nivel educativo.

<i>MÉTRICA</i>	<i>RESULTADO PARA ED0-2</i>	<i>RESULTADO PARA ED3-4</i>	<i>RESULTADO PARA ED5-8</i>
MAE	1.15	0.84	0.76
MSE	1.78	1.14	1.25
RMSE	1.33	1.07	1.12

Fuente: Elaboración propia a partir del modelo ARIMAX.

2. Validación de predicciones para 2023: En el nivel ED0-2, el modelo predijo una tasa de desempleo del 25.34%, muy próxima al valor real de 25.27%, con una diferencia mínima de 0.07 puntos porcentuales. Para el nivel ED3-4, la predicción fue de 17.76%, frente a un valor real de 16.94%, con una diferencia de 0.82 puntos porcentuales. En ED5-8, el modelo proyectó una tasa de 20.06%, mientras que el valor real fue de 21.34%, lo que resultó en una diferencia de -1.28 puntos porcentuales. Estas cifras reflejan una alta precisión general del modelo, especialmente en los niveles ED0-2 y ED3-4.
3. Análisis de residuos: El test de Ljung-Box se utilizó para analizar la independencia de los residuos generados por el modelo. Para ED0-2, el p-valor obtenido fue de 0.5281. En el caso del nivel ED3-4 el p-valor fue de 0.5265. Para ED5-8 el p-valor

fue de 0.9604. En todos los casos, se asegura que los errores están distribuidos aleatoriamente y que el modelo es robusto para las predicciones realizadas.

En conclusión, el modelo ARIMAX demuestra ser una herramienta eficaz para predecir las tasas de desempleo juvenil por niveles educativos. Las métricas de error bajas, combinadas con la independencia de los residuos, validan la capacidad del modelo para capturar las peculiaridades del desempleo juvenil.

6. Conclusiones.

El presente Trabajo Fin de Grado ha tenido como finalidad analizar la influencia de diversas variables —en concreto, el género y el nivel educativo— sobre la tasa de desempleo juvenil en Europa, así como estudiar su evolución en el tiempo y su interacción con factores macroeconómicos, destacando el Producto Interior Bruto (PIB). A partir de este propósito general, se definieron tres objetivos específicos, los cuales han sido abordados a lo largo del estudio.

En primer lugar, se ha cumplido el objetivo de realizar un análisis exhaustivo de la literatura académica sobre el desempleo juvenil en el contexto europeo. La revisión bibliográfica ha permitido identificar los factores que explican las disparidades existentes entre jóvenes, tanto desde una perspectiva individual —considerando variables como el género, el nivel educativo o la situación familiar— como desde una dimensión estructural, incluyendo indicadores macroeconómicos como el PIB y elementos regionales como la cultura laboral.

En segundo lugar, se ha llevado a cabo una revisión de las tasas de desempleo juvenil desagregadas por género y nivel educativo, utilizando datos estadísticos oficiales que abarcan el periodo 2006–2023. Este análisis ha permitido identificar patrones de desigualdad: por un lado, los jóvenes con menor nivel educativo (ED0-2) presentan sistemáticamente tasas de desempleo más elevadas que aquellos con educación terciaria (ED5-8); por otro, persisten diferencias entre hombres y mujeres, si bien con signos de convergencia en los últimos años.

En tercer lugar, se ha cumplido el objetivo metodológico de desarrollar modelos ARIMAX con el fin de prever la evolución del desempleo juvenil hasta el año 2026. Estos modelos, contruidos de forma independiente para cada grupo poblacional, han incorporado el crecimiento del PIB como variable exógena. La elección del modelo ARIMAX ha permitido captar tanto la dinámica interna de las series temporales como el efecto de condiciones macroeconómicas. La validación de las predicciones con los datos reales de 2023 ha mostrado un alto grado de precisión, lo que refuerza la utilidad del modelo como herramienta de análisis de tendencias.

Más allá de las predicciones, los resultados del modelo permiten extraer recomendaciones útiles para el diseño de políticas públicas. Se observa que el crecimiento económico ayuda a reducir el desempleo juvenil, pero no lo hace con la misma fuerza en todos los grupos. Por ejemplo, los hombres jóvenes y quienes tienen un nivel educativo medio (formación profesional o secundaria superior) se benefician más de la mejora del PIB. En cambio, las mujeres jóvenes y los jóvenes con niveles educativos bajos apenas ven mejoras en sus tasas de empleo, incluso en contextos de crecimiento económico.

Esto sugiere que las políticas generales de crecimiento no bastan para mejorar la situación de todos. Por ello, se recomienda complementar estas medidas con otras más específicas: programas de formación dirigidos a jóvenes con menor cualificación, ayudas a la contratación de mujeres jóvenes, y orientación laboral temprana que ayude a evitar largos periodos de desempleo. Además, el hecho de que el desempleo juvenil esté muy influido por su evolución pasada indica que intervenir pronto es clave para evitar que estos jóvenes queden fuera del mercado laboral durante años.

En resumen, el modelo no solo permite anticipar tendencias, sino que también ofrece pistas sobre dónde enfocar las políticas públicas para lograr un impacto real y reducir las desigualdades que afectan al empleo juvenil en Europa.

Sin embargo, este Trabajo de Fin de Grado, como cualquier investigación empírica, presenta ciertas limitaciones que deben ser tenidas en cuenta al interpretar sus resultados.

En primer lugar, la disponibilidad de datos ha sido una de las principales restricciones. Todos los conjuntos de datos utilizados —tanto para el análisis descriptivo como para la modelización predictiva— disponen de información únicamente hasta el año 2023. Esto ha limitado la posibilidad de contrastar las predicciones a medio plazo con datos reales posteriores y de capturar con mayor precisión las tendencias más recientes del desempleo juvenil en Europa.

En segundo lugar, aunque la automatización constituye un factor cada vez más relevante en la transformación del mercado laboral, no se ha podido incorporar como variable exógena en los modelos ARIMAX. Esto se debe a la falta de datos temporales

suficientemente armonizados a nivel europeo que permitan medir de forma cuantitativa su impacto sobre el empleo juvenil.

A partir de los resultados de este estudio, se abren diversas líneas de investigación que podrían ser desarrolladas para profundizar en el análisis del desempleo juvenil desde una perspectiva aplicada.

Una de las direcciones consiste en trasladar el conocimiento generado hacia el diseño y evaluación de políticas públicas concretas, especialmente aquellas dirigidas a reducir las desigualdades en el acceso al empleo entre jóvenes.

Asimismo, futuros trabajos podrían ampliar el marco actual incorporando nuevas variables exógenas, como indicadores de automatización, o transición verde, siempre que se disponga de datos suficientemente robustos. La inclusión de estas dimensiones permitiría un análisis más completo del impacto de los cambios estructurales en el empleo juvenil, y contribuiría al desarrollo de políticas más adaptadas a los retos del mercado laboral del futuro.

Por último, también resultaría relevante aplicar esta metodología a nivel nacional o regional, comparando distintas áreas geográficas dentro de Europa, con el objetivo de identificar buenas prácticas y adaptar las estrategias de intervención a contextos socioeconómicos específicos.

7. Anexos.

7.1. Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Maria del Carmen Gea Berruezo, estudiante de E2 + Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "**Desempleo juvenil en Europa y su relación con el PIB: predicción por género y nivel educativo con modelos ARIMAX**", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Interpretador de código:** Para realizar análisis de datos preliminares.
4. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
5. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 5 de abril de 2025

Firma: Maria del Carmen Gea Berruezo

7.2. Código de R.

El siguiente anexo presenta el código desarrollado en R para el análisis, a través de modelos ARIMAX, del desempleo juvenil en Europa. El conocimiento aplicado se adquirió en la asignatura *Machine Learning II: Forecasting*.

El código se estructura en dos bloques principales: el primero se centra en el análisis del desempleo juvenil según el género, y el segundo aborda el estudio del desempleo en función del nivel educativo.

Como se explicó previamente, se han tomado una serie de decisiones metodológicas basadas en hipótesis realistas, con el fin de garantizar un análisis alineado con los objetivos del estudio. En primer lugar, para proyectar el crecimiento del PIB, se utilizaron valores fijos procedentes de estimaciones de organismos europeos, priorizando la estabilidad del modelo frente a la complejidad de generar una proyección ARIMA propia. En segundo lugar, los valores ausentes fueron imputados mediante la media de cada columna, dada la baja proporción de datos faltantes y con el objetivo de evitar distorsiones. Aunque existen métodos más avanzados, como la interpolación o imputación múltiple, estos implican supuestos adicionales poco justificables en series anuales agregadas. Finalmente, no se aplicó normalización al PIB, ya que se trabaja con tasas de variación porcentuales, que ya están en una escala homogénea.

```
#####PASOS GENERALES#####
```

```
#Instalamos paquetes necesarios y cargamos librerías
```

```
install.packages("readxl")
```

```
install.packages("ggplot2")
```

```
install.packages("forecast")
```

```
library(forecast)
```

```
library(readxl)
```

```
library(ggplot2)
```

7.2.1. Código género: Modelo según género.

1º Carga y limpieza de datos.

```
#Vemos el directorio y cargamos los datos
```

```
getwd()
```

```
file_path <- "hombresymujeresinpaises.xlsx"
```

```
df <- read_excel(file_path)
```

```
#Eliminamos la primera fila, que contiene hombre/mujer y renombramos ajustando a nuestro database  
df_cleaned <- df[-1, ]
```

```
colnames(df_cleaned) <- c("Region", "2006_hombres", "2006_mujeres", "2007_hombres",
```

```
"2007_mujeres", "2008_hombres", "2008_mujeres",
```

```
"2009_hombres", "2009_mujeres", "2010_hombres", "2010_mujeres", "2011_hombres",
```

```
"2011_mujeres", "2012_hombres", "2012_mujeres",
```

```
"2013_hombres", "2013_mujeres", "2014_hombres", "2014_mujeres", "2015_hombres",
```

```
"2015_mujeres", "2016_hombres", "2016_mujeres",
```

```
"2017_hombres", "2017_mujeres", "2018_hombres", "2018_mujeres", "2019_hombres",
```

```
"2019_mujeres", "2020_hombres", "2020_mujeres",
```

```
"2021_hombres", "2021_mujeres", "2022_hombres", "2022_mujeres", "2023_hombres",
```

```
"2023_mujeres")
```

```

#Para los años en que falten datos, rellenamos con la media de la columna
# 1° Convertimos todas las columnas a numéricos, excepto la columna de "Region"
df_cleaned[, 2:ncol(df_cleaned)] <- lapply(df_cleaned[, 2:ncol(df_cleaned)], as.numeric) #Empezamos
en columna 2
# 2° Rellenamos valores NA con la media de cada columna #si el es NA, se sustituye por la media de la
columna
df_cleaned[, 2:ncol(df_cleaned)] <- lapply(df_cleaned[, 2:ncol(df_cleaned)], function(x) {
  ifelse(is.na(x), mean(x, na.rm = TRUE), x) })

```

2° Análisis descriptivo.

```

#Realizamos un análisis descriptivo para ver la validez del dato
# Calculamos la tasa media de desempleo por género y vemos los resultados
average_hombres <- mean(as.numeric(unlist(df_cleaned[, grep("hombres", colnames(df_cleaned))])),
na.rm = TRUE)
average_mujeres <- mean(as.numeric(unlist(df_cleaned[, grep("mujeres", colnames(df_cleaned))])),
na.rm = TRUE)
diferencia <- average_mujeres - average_hombres ##Vemos una diferencia de 1.53%
# Sacamos resultados
cat("Tasa media de desempleo para hombres:", round(average_hombres, 2), "%\n")
cat("Tasa media de desempleo para mujeres:", round(average_mujeres, 2), "%\n")
cat("Diferencia:", round(diferencia, 2), "%\n")

#Vemos la evolución de la diferencia a lo largo del tiempo
years <- c(2006:2023)
differences <- sapply(years, function(year) {
  hombres_col <- paste0(year, "_hombres")
  mujeres_col <- paste0(year, "_mujeres")
  df_cleaned[[hombres_col]] <- gsub("[^0-9.-]", "", df_cleaned[[hombres_col]]) # Eliminamos espacios
en blanco o caracteres especiales
  df_cleaned[[mujeres_col]] <- gsub("[^0-9.-]", "", df_cleaned[[mujeres_col]])
  hombres_values <- as.numeric(df_cleaned[[hombres_col]]) # Convertimos a numérico después de
limpiar
  mujeres_values <- as.numeric(df_cleaned[[mujeres_col]])
  mean(mujeres_values, na.rm = TRUE) - mean(hombres_values, na.rm = TRUE) }) # Calcular la
diferencia promedio ignorando NA

# Creamos un gráfico con todo: tasa promedio y diferencias a lo largo de los años
average_hombres_per_year <- sapply(years, function(year) {
  hombres_col <- paste0(year, "_hombres")
  mean(df_cleaned[[hombres_col]], na.rm = TRUE)})
average_mujeres_per_year <- sapply(years, function(year) {
  mujeres_col <- paste0(year, "_mujeres")
  mean(df_cleaned[[mujeres_col]], na.rm = TRUE)})

# Creamos un data frame con las tasas promedio y la diferencia por año
df_rates <- data.frame(
  Year = years,
  Hombres = average_hombres_per_year,
  Mujeres = average_mujeres_per_year,
  Diferencia = average_mujeres_per_year - average_hombres_per_year)

```

```

df_rates$Hombres <- as.numeric(df_rates$Hombres) # Convertimos a numérico y manejar valores NA si
existen
df_rates$Mujeres <- as.numeric(df_rates$Mujeres)
df_rates$Diferencia <- as.numeric(df_rates$Diferencia)

# Pintamos las tasas de desempleo de hombres, mujeres y la diferencia
ggplot(df_rates, aes(x = Year)) +
  geom_line(aes(y = Hombres, color = "Hombres"), size = 1) +
  geom_line(aes(y = Mujeres, color = "Mujeres"), size = 1) +
  geom_line(aes(y = Diferencia, color = "Diferencia"), size = 1, linetype = "dashed") +
  scale_color_manual(values = c("Hombres" = "blue", "Mujeres" = "red", "Diferencia" = "green")) +
  labs(title = "Tasa de paro juvenil media de hombres, mujeres y su diferencia a lo largo de los años",
        x = "Año", y = "Tasa de paro (%)") +
  theme_minimal() +
  theme(legend.title = element_blank())

```

3° Preparación del dato para las series temporales: carga de la variable exógena y creación del objeto serie temporal (ts)

```

#Realizamos el modelo predictivo con series temporales y PIB como variable exógena

#1° Tratamiento de la variable exógena
gdp_eu_data <- read_excel("gdpEU.xlsx") # Cargamos la variable exógena
gdp_eu_values <- as.numeric(gdp_eu_data[2, 2:19]) # Extraemos los datos de PIB de la UE desde 2006
hasta 2023
# Calculamos la tasa de crecimiento anual en porcentaje
gdp_eu_growth <- c(NA, diff(gdp_eu_values) / head(gdp_eu_values, -1)) * 100 # Se elimina e 1° valor
como NA porque no hay cambio con respecto al año anterior
gdp_eu_future <- c(0.9, 1.5, 1.8) # Añadimos royecciones en porcentaje para 2024, 2025 y 2026
gdp_eu_growth <- c(gdp_eu_growth, gdp_eu_future)

# 2°: Calculamos la tasa media de desempleo por género en todas las regiones para cada año
avg_unemployment_hombres <- sapply(years, function(year) {
  mean(df_cleaned[[paste0(year, "_hombres")]], na.rm = TRUE)}) # Creamos vectores para almacenar
las medias anuales
avg_unemployment_mujeres <- sapply(years, function(year) {
  mean(df_cleaned[[paste0(year, "_mujeres")]], na.rm = TRUE)})
# Creamos la serie temporal de desempleo medio (hombres y mujeres) desde 2006 a 2022
ts_hombres <- ts(avg_unemployment_hombres[1:17], start = 2006, frequency = 1) # Solo hasta 2022
(17 años)
ts_mujeres <- ts(avg_unemployment_mujeres[1:17], start = 2006, frequency = 1)

```

4° Ajuste de modelos ARIMAX

```

# Ajustamos modelos ARIMA para hombres y mujeres usando el PIB (hasta 2022) como variable exógena
arimax_hombres <- auto.arima(ts_hombres, xreg = gdp_eu_growth[1:17])
arimax_mujeres <- auto.arima(ts_mujeres, xreg = gdp_eu_growth[1:17])

# Realizamos la predicción para 2023-2026 usando proyecciones de crecimiento del PIB UE
forecast_hombres <- forecast(arimax_hombres, xreg = gdp_eu_growth[18:21], h = 4)
forecast_mujeres <- forecast(arimax_mujeres, xreg = gdp_eu_growth[18:21], h = 4)

```

```

# Creamos un data frame con los años y las tasas de paro promedio de hombres y mujeres
years <- 2006:2022
tasa_paro_promedio_hombres <- avg_unemployment_hombres[1:17]
tasa_paro_promedio_mujeres <- avg_unemployment_mujeres[1:17]

# Añadimos las predicciones para 2023-2026 al final de las series
predicted_hombres <- c(tasa_paro_promedio_hombres, forecast_hombres$mean)
predicted_mujeres <- c(tasa_paro_promedio_mujeres, forecast_mujeres$mean)
years_full <- c(years, 2023:2026)

# Obtenemos la predicción para 2023
prediccion_2023_hombres <- forecast_hombres$mean[1] # 1º predicción (2023)
prediccion_2023_mujeres <- forecast_mujeres$mean[1]

# Calculamos la media real de desempleo en 2023 a partir de df_cleaned
media_real_2023_hombres <- mean(df_cleaned$`2023_hombres`, na.rm = TRUE)
media_real_2023_mujeres <- mean(df_cleaned$`2023_mujeres`, na.rm = TRUE)

# Extraemos los resultados
cat("Predicción para 2023 (Hombres):", prediccion_2023_hombres, "\n")
cat("Valor Real para 2023 (Hombres):", media_real_2023_hombres, "\n")
cat("Predicción para 2023 (Mujeres):", prediccion_2023_mujeres, "\n")
cat("Valor Real para 2023 (Mujeres):", media_real_2023_mujeres, "\n")

# Creamos un data frame para el gráfico
df_evolucion <- data.frame(
  Year = rep(years_full, 2),
  Tasa_Paro = c(predicted_hombres, predicted_mujeres),
  Genero = rep(c("Hombres", "Mujeres"), each = length(years_full)),
  Tipo = c(rep("Histórico", length(years)), rep("Predicción", 4)))
df_evolucion$Year <- as.numeric(df_evolucion$Year)
df_evolucion$Tasa_Paro <- as.numeric(df_evolucion$Tasa_Paro)
df_evolucion$Genero <- as.factor(df_evolucion$Genero)
df_evolucion$Tipo <- as.factor(df_evolucion$Tipo)

# Pintamos la evolución del desempleo en hombres y mujeres
library(ggplot2)
ggplot(df_evolucion, aes(x = Year, y = Tasa_Paro, color = Genero, linetype = Tipo)) +
  geom_line(size = 1) +
  geom_point() +
  scale_linetype_manual(values = c("Histórico" = "solid", "Predicción" = "dashed")) +
  labs(title = "Evolución del Desempleo Juvenil en Hombres y Mujeres (2006-2026)",
    x = "Año", y = "Tasa de Paro (%)") +
  theme_minimal() +
  theme(legend.title = element_blank())

```

5º Evaluación de la validez del modelo

```

#Evaluación de la validez del modelo
# 1º Errores de predicción: calculamos el MAE, MSE y RMSE para hombres y mujeres
mae_hombres <- mean(abs(forecast_hombres$residuals), na.rm = TRUE)
mse_hombres <- mean((forecast_hombres$residuals)^2, na.rm = TRUE)
rmse_hombres <- sqrt(mse_hombres)

```

```

mae_mujeres <- mean(abs(forecast_mujeres$residuals), na.rm = TRUE)
mse_mujeres <- mean((forecast_mujeres$residuals)^2, na.rm = TRUE)
rmse_mujeres <- sqrt(mse_mujeres)
# Extraemos los resultados
cat("MAE para Hombres:", mae_hombres, "\n")
cat("MSE para Hombres:", mse_hombres, "\n")
cat("RMSE para Hombres:", rmse_hombres, "\n\n")
cat("MAE para Mujeres:", mae_mujeres, "\n")
cat("MSE para Mujeres:", mse_mujeres, "\n")
cat("RMSE para Mujeres:", rmse_mujeres, "\n")

#2° Prueba de autocorrelación de residuos: Ljung-Box de hombres y mujeres (sin valores NA)
Box.test(na.omit(forecast_hombres$residuals), lag = 10, type = "Ljung-Box")
Box.test(na.omit(forecast_mujeres$residuals), lag = 10, type = "Ljung-Box")
# Pintamos residuos para hombres y mujeres
par(mfrow = c(2, 1))
plot(forecast_hombres$residuals, main = "Residuos del Modelo para Hombres", ylab = "Residuos", xlab = "Tiempo")
abline(h = 0, col = "red")
plot(forecast_mujeres$residuals, main = "Residuos del Modelo para Mujeres", ylab = "Residuos", xlab = "Tiempo")
abline(h = 0, col = "red")
par(mfrow = c(1, 1))

# 3° Comparamos con valores reales en 2023
tasa_paro_real_hombres_2023 <- mean(df_cleaned$`2023_hombres`, na.rm = TRUE)
tasa_paro_real_mujeres_2023 <- mean(df_cleaned$`2023_mujeres`, na.rm = TRUE)
cat("Predicción 2023 para Hombres:", forecast_hombres$mean[1], "\n")
cat("Valor Real 2023 para Hombres:", tasa_paro_real_hombres_2023, "\n\n")
cat("Predicción 2023 para Mujeres:", forecast_mujeres$mean[1], "\n")
cat("Valor Real 2023 para Mujeres:", tasa_paro_real_mujeres_2023, "\n")

```

7.2.2. Código nivel educativo.

1° Carga y limpieza de datos.

```

#Cargamos el database
#Vemos el directorio donde se almacena el fichero y cargamos los datos en el dataframe
getwd()
file_path <- "educ_com.xlsx"
df <- read_excel(file_path)
#Realizamos la limpieza y preparación de los datos
#Eliminamos la primera fila
df_cleaned <- df[-1, ]
# Creamos una lista de nombres de columnas preparadas para nuestro estudio
years <- 2006:2023
levels <- c("ED0-2", "ED3-4", "ED5-8")
column_names <- c("Region", unlist(lapply(years, function(year) paste0(year, "_", levels))))
# Asignamos los nombres de las columnas al dataframe
colnames(df_cleaned) <- column_names
# Convertimos todas las columnas, excepto la columna "Region", a valores numéricos
cols_to_convert <- colnames(df_cleaned)[-1] # Transformamos todas las columnas menos "Region"
df_cleaned[cols_to_convert] <- lapply(df_cleaned[cols_to_convert], function(x) {
  as.numeric(gsub(":", NA, x)) # Reemplazamos ":" por NA y convertimos a numérico
})

```

```
# Rellenamos valores NA con la media de la columna->con el objetivo de disminuir el impacto de valores ausentes
df_cleaned[cols_to_convert] <- lapply(df_cleaned[cols_to_convert], function(x) {
  ifelse(is.na(x), mean(x, na.rm = TRUE), x)})
```

2º Análisis descriptivo.

```
#Realizamos un análisis descriptivo de los datos
# Calculamos las tasas medias de desempleo para cada nivel educativo
mean_ED0_2 <- colMeans(df_cleaned[, grep("ED0-2", colnames(df_cleaned))], na.rm = TRUE)
mean_ED3_4 <- colMeans(df_cleaned[, grep("ED3-4", colnames(df_cleaned))], na.rm = TRUE)
mean_ED5_8 <- colMeans(df_cleaned[, grep("ED5-8", colnames(df_cleaned))], na.rm = TRUE)

# Calculamos las tasas medias de desempleo generales para cada nivel educativo
average_ED0_2 <- round(mean(mean_ED0_2, na.rm = TRUE), 2)
average_ED3_4 <- round(mean(mean_ED3_4, na.rm = TRUE), 2)
average_ED5_8 <- round(mean(mean_ED5_8, na.rm = TRUE), 2)
# Mostramos resultados
cat("Tasa promedio de desempleo para nivel ED0-2:", average_ED0_2, "%\n")
cat("Tasa promedio de desempleo para nivel ED3-4:", average_ED3_4, "%\n")
cat("Tasa promedio de desempleo para nivel ED5-8:", average_ED5_8, "%\n")

####Gráfico análisis descriptivo de los datos
# Creamos un dataframe para almacenar las tasas medias
years <- as.numeric(gsub("_.*", "", names(mean_ED0_2))) # Extraemos los años de los nombres de columnas
df_means <- data.frame(
  Year = years,
  ED0_2 = mean_ED0_2,
  ED3_4 = mean_ED3_4,
  ED5_8 = mean_ED5_8)
# Pintamos el gráfico utilizando ggplot2
library(ggplot2)
ggplot(df_means, aes(x = Year)) +
  geom_line(aes(y = ED0_2, color = "ED0-2"), size = 1) +
  geom_line(aes(y = ED3_4, color = "ED3-4"), size = 1) +
  geom_line(aes(y = ED5_8, color = "ED5-8"), size = 1) +
  labs(title = "Tasa Media de Desempleo Juvenil por Nivel Educativo (2006-2023)",
    x = "Año",
    y = "Tasa de Desempleo (%)", color = "Nivel Educativo" ) + theme_minimal()
names(mean_ED0_2)
#Una vez realizado el análisis descriptivo, y viendo la idoneidad del dato, pasamos a realizar las series temporales
```

3º Preparación del dato para las series temporales: carga de la variable exógena y creación del objeto serie temporal (ts)

```
#En primer lugar, cargamos la variable exógena utilizada: crecimiento del PIB de la UE
gdp_eu_data <- read_excel("gdpEU.xlsx")
# Extraemos los datos de PIB de la UE desde 2006 hasta 2023
gdp_eu_values <- as.numeric(gdp_eu_data[2, 2:19])
# Calculamos la tasa de crecimiento anual en porcentaje
gdp_eu_growth <- c(NA, diff(gdp_eu_values) / head(gdp_eu_values, -1)) * 100 # Al igual que en el modelo previo, el primer valor es NA porque no hay cambio con respecto al año anterior
```

```

# Añadimos las proyecciones para 2024, 2025 y 2026
gdp_eu_future <- c(0.9, 1.5, 1.8) # Proyecciones de crecimiento para 2024, 2025 y 2026
gdp_eu_growth <- c(gdp_eu_growth, gdp_eu_future)

# Definimos años y niveles educativos
years <- 2006:2022
levels <- c("ED0-2", "ED3-4", "ED5-8")
# Calculamos, de manera agrupada, la tasa promedio de desempleo para cada nivel educativo por año
avg_unemployment_levels <- lapply(levels, function(level) {
  sapply(years, function(year) {
    mean(df_cleaned[[paste0(year, "_", level)]], na.rm = TRUE)}})
# Creamos series temporales (ts) para cada nivel educativo
ts_levels <- lapply(avg_unemployment_levels, function(avg) {
  ts(avg, start = 2006, frequency = 1)})

```

4° Ajuste de modelos ARIMAX

```

# Ajustamos cada modelo ARIMA para cada nivel educativo usando el PIB como variable exógena
arimax_levels <- lapply(ts_levels, function(ts_level) {
  auto.arima(ts_level, xreg = gdp_eu_growth[1:17])})
# Realizamos predicciones para 2023-2026 usando las proyecciones de crecimiento del PIB de la UE
forecasts_levels <- lapply(arimax_levels, function(arima_model) {
  forecast(arima_model, xreg = gdp_eu_growth[18:21], h = 4)})

```

```

# Creamos un data frame para almacenar los datos históricos y predicciones
years_full <- c(years, 2023:2026)
predicted_levels <- lapply(1:length(levels), function(i) {
  c(avg_unemployment_levels[[i]], forecasts_levels[[i]]$mean)})
df_evolucion <- data.frame(
  Year = rep(years_full, length(levels)),
  Tasa_Paro = unlist(predicted_levels),
  Nivel_Educativo = rep(levels, each = length(years_full)),
  Tipo = c(rep("Histórico", length(years)), rep("Predicción", 4)))
df_evolucion$Year <- as.numeric(df_evolucion$Year)
df_evolucion$Tasa_Paro <- as.numeric(df_evolucion$Tasa_Paro)
df_evolucion$Nivel_Educativo <- as.factor(df_evolucion$Nivel_Educativo)
df_evolucion$Tipo <- as.factor(df_evolucion$Tipo)

```

```

# Pintamos la evolución del desempleo por nivel educativo
ggplot(df_evolucion, aes(x = Year, y = Tasa_Paro, color = Nivel_Educativo, linetype = Tipo)) +
  geom_line(size = 1) +
  geom_point() +
  scale_linetype_manual(values = c("Histórico" = "solid", "Predicción" = "dashed")) +
  labs(title = "Evolución del desempleo por nivel educativo (2006-2026)",
  x = "Año", y = "Tasa de paro (%)") +
  theme_minimal() +
  theme(legend.title = element_blank())

```

5° Evaluación de la validez del modelo

```

#Análisis de la calidad del modelo
#Calculamos el MAE, MSE y RMSE para cada nivel educativo
evaluation_metrics <- lapply(1:length(levels), function(i) {
  residuals <- forecasts_levels[[i]]$residuals

```

```

mae <- mean(abs(residuals), na.rm = TRUE)
mse <- mean((residuals)^2, na.rm = TRUE)
rmse <- sqrt(mse)
list(MAE = mae, MSE = mse, RMSE = rmse)})

# Imprimimos los resultados
for (i in 1:length(levels)) {
  cat("Nivel Educativo:", levels[i], "\n")
  cat("MAE:", round(evaluation_metrics[[i]]$MAE, 2), "\n")
  cat("MSE:", round(evaluation_metrics[[i]]$MSE, 2), "\n")
  cat("RMSE:", round(evaluation_metrics[[i]]$RMSE, 2), "\n\n")}

# Prueba de autocorrelación de residuos: Ljung-Box de cada nivel educativo
ljung_box_results <- lapply(1:length(levels), function(i) {
  residuals <- na.omit(forecasts_levels[[i]]$residuals)
  Box.test(residuals, lag = 10, type = "Ljung-Box")})
# Extraemos los resultados de Ljung-Box
for (i in 1:length(levels)) {
  cat("Nivel Educativo:", levels[i], "\n")
  print(ljung_box_results[[i]])
  cat("\n")}

# Pintamos los residuos para verlo de manera más visual
par(mfrow = c(length(levels), 1)) # Creamos un layout con una fila por nivel educativo
for (i in 1:length(levels)) {
  residuals <- forecasts_levels[[i]]$residuals
  plot(residuals, main = paste("Residuos del Modelo para", levels[i]),
       ylab = "Residuos", xlab = "Tiempo")
  abline(h = 0, col = "red")}
par(mfrow = c(1, 1)) # Restauramos el layout original

# Calculamos la predicción y valores reales para 2023
prediccion_2023 <- sapply(1:length(levels), function(i) {
  forecasts_levels[[i]]$mean[1]})
media_real_2023 <- sapply(levels, function(level) {
  mean(df_cleaned[[paste0("2023_", level)]], na.rm = TRUE)})

# Imprimir los resultados
for (i in 1:length(levels)) {
  cat("Nivel Educativo:", levels[i], "\n")
  cat("Predicción para 2023:", round(prediccion_2023[i], 2), "\n")
  cat("Valor Real para 2023:", round(media_real_2023[i], 2), "\n")
  cat("Diferencia:", round(prediccion_2023[i] - media_real_2023[i], 2), "\n\n")}

```

8. Bibliografia.

Amoako, S., Kwofie, S., & Osei, P. (2023). VAR, ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends. *Journal of Electrical Systems and Information Technology*, 10(1), 1–11. <https://doi.org/10.1186/s43067-023-00078-1>

Apaydın, F. (2018). A comparative analysis of the characteristics of unemployment in Turkey. *Journal of Süleyman Demirel University Institute of Social Sciences*, 30(1), 159–200.

Bălan, M. (2021). *Youth labor market and COVID-19*. Hyperion University. https://openurl.ebsco.com/EPDB%3Agcd%3A16%3A20152704/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A170961934&crl=c&link_origin=scholar.google.es

Bal-Domańska, B. (2022). The impact of macroeconomic and structural factors on the unemployment of young women and men. *Economic Change and Restructuring*, 55(4), 1141–1172. <https://doi.org/10.1007/s10644-021-09341-9>

Başol, O., Sevgi, H., & Yalçın, E. C. (2023). *The effect of digitalization on youth unemployment for EU countries: Treat or threat? Sustainability*, 15(14), 11080. <https://doi.org/10.3390/su151411080>

Baussola, M., & Mussida, C. (2017). Regional and gender differentials in the persistence of unemployment in Europe. *International Review of Applied Economics*, 31(2), 173–190. <https://doi.org/10.1080/02692171.2016.1231801>

Bektemyssova, G., Ahmad, A. R., Mirzakulova, S., & Ibraeva, Z. (2022). *Time series forecasting by the ARIMA method*. Scientific Journal of Astana IT University. <https://doi.org/10.37943/HFCH4395>

Ben, F., & Ben Salah, H. (2013). *Forecasting application rates using ARIMA and ARIMAX models*. Society of Actuaries. <https://dr.soa.org/49384d/globalassets/assets/files/research/projects/research-2013-arima-arimax-ben-appl-rates.pdf>

Bod'a, M., & Považanová, M. (2020). Output–unemployment asymmetry in Okun coefficients for OECD countries. *Economic Analysis and Policy*. <https://doi.org/10.1016/j.eap.2020.12.004>

Booth, A. L. (2009). Gender and competition. *Labour Economics*, 16(6), 599–606. <https://doi.org/10.1016/j.labeco.2009.08.002>

Booth, A. L., Francesconi, M., & Frank, J. (2003). A sticky floors model of promotion, pay, and gender. *European Economic Review*, 47(2), 295–322. [https://doi.org/10.1016/S0014-2921\(01\)00197-0](https://doi.org/10.1016/S0014-2921(01)00197-0)

Boudon, R. (1974). *Education, opportunity and social inequality: Changing prospects in Western society*. Wiley. <https://archive.org/details/educationopportu0000boud/page/n5/mode/2up>

Butkus, M., Matuzeviciute, K., Rupliene, D., & Seputiene, J. (2020). Does unemployment responsiveness to output change depend on age, gender, education, and the phase of the business cycle? *Economies*, 8(4), 98. <https://doi.org/10.3390/economies8040098>

Butkus, M., Seputienė, J., & Rimas, T. (2020). Youth unemployment and employment: The case of the European Union. *Economics and Sociology*, 13(3), 248–263. <https://doi.org/10.14254/2071-789X.2020/13-3/16>

Cedefop. (2014). *Skill mismatch: More than meets the eye*. Publications Office of the European Union. https://www.cedefop.europa.eu/files/9087_en.pdf

Comisión Europea. (2024). *Autumn 2024 economic forecast: Gradual rebound in an adverse environment*. Recuperado el 20 de noviembre de 2024 de https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/economic-forecasts/autumn-2024-economic-forecast-gradual-rebound-adverse-environment_en

Council of the European Union. (2014). *Council recommendation on a quality framework for traineeships (Employment, Social Policy, Health and Consumer Affairs Council Meeting)*. Recuperado el 20 de noviembre de 2024 de http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/lsa/141424.pdf

Dietrich, H. (2012). Youth unemployment in Europe: Theoretical considerations and empirical findings. Friedrich-Ebert-Stiftung. <https://library.fes.de/pdf-files/id/ipa/09227.pdf>

Dumičić, K., Čeh Časni, A., & Žmuk, B. (2015). Forecasting unemployment rate in selected European countries using smoothing methods. *World Academy of Science, Engineering and Technology, International Journal of Social, Education, Economics and Management Engineering*, 9(4), 1041-1046.

Dunsch, S. (2017). Age- and gender-specific unemployment and Okun's law in CEE countries. *Empirica*, 44(4), 583–605. <https://doi.org/10.1080/00128775.2017.1338962>

Džunić, M., Lakić Golubović, N., & Janković-Milić, V. (2024). Educational attainment as a predictor of poverty and social exclusion: Empirical analysis of Serbian case. *E+M Ekonomie a Management*, 27(1), 40–52. <https://doi.org/10.15240/tul/001/2024-1-003>

European Training Foundation (ETF). (2024). *Gender dimension of labour market transitions: Implications for activation and skills development policies of the EU neighbouring countries*. Recuperado el 19 de noviembre de 2024 de <https://www.etf.europa.eu/en/publications-and-resources/publications/gender-dimension-labour-market-transitions>

Eurostat. (2024). *Gross domestic product (GDP) at current market prices by NUTS 2 region* [Dataset]. https://doi.org/10.2908/nama_10r_2gdp

Eurostat. (2024). *Unemployment rates by sex, age, educational attainment level and NUTS 2 region (%)* [Dataset]. https://doi.org/10.2908/lfst_r_lfu3rt

Eurostat. (2024). *Youth unemployment rate by sex, age and NUTS 2 region* [Dataset]. https://doi.org/10.2908/yth_empl_110

Eurostat. (n.d.). *Who we are*. Recuperado el 16 de noviembre de 2024 de <https://ec.europa.eu/eurostat/web/main/about-us/who-we-are>

Göçer, İ., & Erdal, L. (2015). The Relationship between youth unemployment and economic growth in central and eastern european countries: An empirical analysis. *Çankırı Karatekin Üniversitesi İİBF Dergisi*, 5(1), 173-188

Görlich, D., Stepanok, I., & Al-Hussami, F. (2013). Youth unemployment in Europe and the world: Causes, consequences and solutions (Kiel Policy Brief, No. 59). Kiel Institute for the World Economy (IfW). <https://hdl.handle.net/10419/70102>

Ho, S. L., & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers & Industrial Engineering*, 35(1–2), 213–216. [https://doi.org/10.1016/S0360-8352\(98\)00066-7](https://doi.org/10.1016/S0360-8352(98)00066-7)

International Labour Organization (ILO). (2020). *Global employment trends for youth 2020: Technology and the future of jobs*. Recuperado el 19 de noviembre de 2024 de: https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40dgreports/%40dcomm/%40publ/documents/publication/wcms_737648.pdf

Karamessini, M. (2019). Factors explaining youth unemployment and early job insecurity in Europe. En *Youth unemployment and job insecurity in Europe* (pp. 45–69). Edward Elgar Publishing. <https://doi.org/10.4337/9781788118897.00009>

Keele, L., & Kelly, N. J. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, 14(2), 186–205. <https://doi.org/10.1093/pan/mpj006>

Kogan, I., Unt, M., & Saar, E. (2006). *Youth unemployment in the enlarged Europe*. University of Mannheim. Recuperado el 27 de noviembre de 2024 de https://www.academia.edu/39702536/Youth_Unemployment_in_the_Enlarged_Europe

Lindemann, K., & Gangl, M. (2019). *The intergenerational effects of unemployment: How parental unemployment affects educational transitions in Germany* (SOEPpapers on Multidisciplinary Panel Data Research, No. 971). German Institute for Economic Research (DIW Berlin). <https://hdl.handle.net/10419/180393>

OCDE. (s. f.). *Youth unemployment rate*. Recuperado el 18 de noviembre de 2024 de <https://www.oecd.org/en/data/indicators/youth-unemployment-rate.html>

O'Reilly, J., Eichhorst, W., Gábos, A., Hadjivassiliou, K., Lain, D., Leschke, J., McGuinness, S., Kureková, L. M., Nazio, T., Ortlieb, R., Russell, H., & Villa, P. (2015). Five characteristics of youth unemployment in Europe: Flexibility, education, migration, family legacies, and EU policy. *SAGE Open*, 5(1), 1–19. <https://doi.org/10.1177/2158244015574962>

Organización Internacional del Trabajo (ILO). (2024). *Global employment trends for youth 2024: Decent work, brighter futures* [Executive Summary]. Recuperado el 14 de noviembre de 2024 de https://www.ilo.org/sites/default/files/2024-08/GET_2024_ExecSum_EN_0.pdf

Organización Internacional del Trabajo. (2018). *Perspectivas sociales y del empleo en el mundo: Tendencias del empleo juvenil 2018*. Recuperado el 20 de noviembre de 2024 de https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_646146.pdf

Privara, A. (2022). Economic growth and labour market in the European Union: Lessons from COVID-19. *Oeconomia Copernicana*, 13(2), 355–377. <https://doi.org/10.24136/oc.2022.011>

Pouliakas, K. (2018). *Automation risk in the EU labour market: A skill-needs approach*. European Centre for the Development of Vocational Training (Cedefop). https://www.cedefop.europa.eu/files/automation_risk_in_the_eu_labour_market.pdf

Russell, H., & O'Connell, P. J. (2001). Getting a job in Europe: The transition from unemployment to work among young people in nine European countries. *Work, Employment and Society*, 15(1), 1-24. <https://doi.org/10.1017/S0950017001000010>

Ryczkowski M, Zinecker M (2020) Gender unemployment in the Czech and Polish labour market. *Argumenta Oeconomica* 2:213–229. <https://doi.org/10.15611/aoe.2020.2.09>

Sostero, M., et al. (2023). Automation technologies and their impact on employment: A review of the current state of research. *Technological Forecasting and Social Change*, 186, 122064. <https://doi.org/10.1016/j.techfore.2022.122064>

Stefanova-Lauerová, J., & Terrell, K. (2002). *Explaining gender differences in unemployment with micro data on flows in post-communist economies* (IZA Discussion Paper No. 600). Institute for the Study of Labor (IZA). <https://hdl.handle.net/10419/21286>

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>

Tomić, I. (2018). What drives youth unemployment in Europe?: Economic vs non-economic determinants. *International Labour Review*, 157(3), 379–408. <https://doi.org/10.1111/ilr.12113>

UNESCO. (2012). *International Standard Classification of Education (ISCED) 2011*. Recuperado el 13 de noviembre de 2024 de: <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>

Uslu, G., & Tatli, H. (2024). Factors affecting regional youth unemployment: Micro-evidence on individuals. *International Journal of Management Economics and Business*, 20(1), 16-42. <https://doi.org/10.17130/ijmeb.1374411>

Warmuth, J., Kittel, B., Steiber, N., & Mühlböck, M. (2014). *CUPESSE theoretical overview D1.2*. https://cupesse.eu/fileadmin/cupesse/downloads/working-papers/CUPESSE_Working-Paper_1.pdf

Wolbers, M. H. J. (2007). Patterns of labour market entry: A comparative perspective on school-to-work transitions in 11 European countries. *Acta Sociologica*, 50(3), 249–269. <https://doi.org/10.1177/0001699307080924>

Woloszko, N. (2019). Tracking GDP in real time using machine learning and big data: A new perspective from the OECD Weekly Tracker. *Journal for Labour Market*

Research, 53(1), 1–13. Recuperado el 12 de noviembre de 2024 de <https://labourmarketresearch.springeropen.com/articles/10.1186/s12651-019-0253-4>