UNIVERSIDAD CATÓLICA DE CUENCA

UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE INFORMÁTICA, CIENCIAS DE LA COMPUTACIÓN E INNOVACIÓN TECNOLÓGICA

CARRERA DE INGENIERÍA EN SISTEMAS DE INFORMACIÓN

PRONÓSTICO DE MIGRACIÓN HUMANA DEL ECUADOR, UTILIZANDO MODELOS DE SERIES TEMPORALES

TRABAJO DE TITULACIÓN PREVIO

A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS DE INFORMACIÓN

AUTOR: CRISTIAN ANDRÉS ANDRADE GUALLPA.

DIRECTOR: ING. CRISTINA MARIUXI FLORES URGILÉS, MGS.

CAÑAR - ECUADOR

2023

DIOS, PATRIA, CULTURA Y DESARROLLO

UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE INFORMÁTICA, CIENCIAS DE LA COMPUTACIÓN E INNOVACIÓN TECNOLÓGICA

CARRERA DE INGENIERIA EN SISTEMAS DE INFORMACIÓN

"PRONÓSTICO DE MIGRACIÓN HUMANA DEL ECUADOR, UTILIZANDO MODELOS DE SERIES TEMPORALES"

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE

INGENIERO DE SISTEMAS DE INFORMACIÓN

AUTOR: CRISTIAN ANDRÉS ANDRADE GUALLPA.

DIRECTOR: ING. CRISTINA MARIUXI FLORES URGILÉS, MGS.

CAÑAR - ECUADOR

2023

DIOS, PATRIA, CULTURA Y DESARROLLO.

DECLARACIÓN

Yo, Cristian Andrés Andrade Guallpa, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

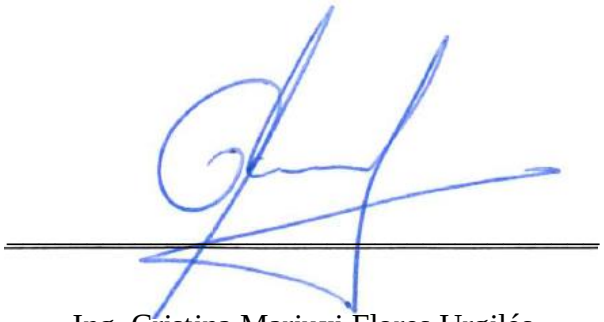
La Universidad Católica de Cuenca extensión Cañar puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y la Normativa actual de la institución.

Andrade Guallpa Cristian Andrés

C.I: **0302792007**

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por el Est. Cristian Andrés Andrade Guallpa, bajo mi supervisión.



Ing. Cristina Mariuxi Flores Urgilés

DIRECTOR DEL TRABAJO INVESTIGATIVO

UNIVERSIDAD CATÓLICA DE CUENCA

RESUMEN

El artículo presenta un estudio para generar pronósticos sobre la migración humana en Ecuador, utilizando técnicas de series temporales. Los objetivos planteados fueron a) Pronosticar la migración humana del Ecuador, utilizando modelos de series temporales, b) Realizar un estudio de estado del arte y revisión bibliográfica sobre minería de datos y metodologías de maching learning, c) Evaluar y comparar diferentes modelos de series temporales, utilizando medidas de precisión y desempeño, con el fin de seleccionar el modelo que ofrezca las mejoras predicciones para la migración humana del Ecuador a futuro d) Realizar el pronóstico de la migración en el Ecuador utilizando modelos de series temporales.

Inicialmente se presenta el contexto de la migración en Ecuador, la cual se ha visto afectada en los últimos años por varios factores como terremotos, cambios de gobierno y la pandemia de COVID-19. Luego se detallan conceptos teóricos relacionados con las series de tiempo, componentes de una serie temporal (tendencia, estacionalidad, ciclos), métodos de pronóstico como ARIMA y modelos de regresión.

El estudio aplica la metodología CRISP-DM utilizando datos de migración del Instituto Nacional de Estadística y Censos de Ecuador entre 2019-2022. Se aplicaron dos modelos: regresión lineal y ARIMA estacional. Los resultados muestran que el modelo de regresión lineal tuvo un error de predicción menor, por lo que se seleccionó como modelo definitivo. El artículo concluye que los modelos de series de tiempo son una herramienta valiosa para realizar pronósticos en temas sociales como la migración.

***Palabras Clave*:** migración, pronóstico, CRISP-DM, ARIMA.

ABSTRACT

The article presents a study to generate forecasts on human migration in Ecuador using time series techniques. The objectives set out were a) To forecast human migration in Ecuador using time series models, b) To conduct a state-of-the-art study and literature review on data mining and machine learning methodologies, c) To evaluate and compare different time series models using precision and performance measures, aiming to select the model that offers the best predictions for future human migration in Ecuador d) To forecast migration in Ecuador using time series models. Initially, the context of migration in Ecuador is presented, which has been affected in recent years by several factors such as earthquakes, changes in government, and the COVID-19 pandemic. The article then delves into theoretical concepts related to time series, components of a time series (trend, seasonality, cycles), and forecasting methods like ARIMA and regression models. The study applies the CRISP-DM methodology using migration data from the National Institute of Statistics and Censuses of Ecuador between 2019-2022. Two models were applied: linear regression and seasonal ARIMA. The results show that the linear regression model had a lower prediction error, so it was chosen as the definitive model. The article concludes that time series models are a valuable tool for making forecasts on social topics such as migration**.**

Key words: migration, prognosis, CRISP-DM, ARIMA.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Pronóstico de migración humana del Ecuador, utilizando modelos de series temporales

"Forecast of human migration from Ecuador, using time series models."

Cristian Andrés Andrade Guallpa [1]

Estudiante, Universidad Católica de Cuenca, Ecuador,
cristian.andrade.07@est.ucacue.edu.ec

Cristina Mariuxi Flores Urgilés [2]

Docente Universidad Católica de Cuenca, Ecuador

cmfloresu@ucacue.edu.ec

Luis Fernando Pinos Castillo [3]

Docente, Universidad Católica de Cuenca, Ecuador

lfpinosc@ucacue.edu.ec

Danny Patricio Andrade Cárdenas [4]

Docente, Universidad Católica de Cuenca

dpandradec@ucacue.edu.ec

ORCID

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

Introducción

En el vasto dominio del análisis de series temporales, la predicción precisa y oportuna de eventos futuros basada en observaciones pasadas y presentes ha sido un desafío constante en diversas disciplinas. Estas series, definidas como secuencias de puntos de datos observados en intervalos de tiempo sucesivos, ofrecen una variedad de patrones, tendencias y estacionalidades que, si se interpretan correctamente, pueden ofrecer percepciones proféticas sobre eventos futuros. Una de las aplicaciones menos exploradas pero críticamente relevante de los modelos de series temporales es en la predicción de flujos migratorios. La migración, impulsada por una mezcla de factores socioeconómicos, políticos y ambientales, posee inherentemente dinámicas temporales que la hacen susceptible a un análisis predictivo.

Prever con precisión estos flujos migratorios no es solo una cuestión académica, sino una necesidad imperante para la planificación estratégica de recursos, la diplomacia y la seguridad nacional en muchos países. Por lo tanto, la integración efectiva de modelos de series temporales en el análisis migratorio promete una gestión más informada y proactiva de los patrones migratorios.

Bases teóricas

Series temporales

Peris (2022) comenta que un modelo de serie temporal hace referencia a métodos estadísticos que se encargan de analizar datos secuenciales; tienen como objetivo entender la estructura subyacente de la serie temporal para prever o pronosticar valores futuros.

Una serie temporal o cronológica es un conjunto de observaciones de una variable realizadas a lo largo del tiempo de forma organizada según los valores que ha tomado la variable, tiempo, haciendo que los valores que ha medido la variable se muestren en orden cronológico.

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

Cualquier serie temporal muestra cómo ha cambiado una variable a lo largo del tiempo (Fernandez, 2021, p. 03).

Descomposición de una serie temporal	
Tendencia(T)	Da a conocer el crecimiento o descenso de un evento.
Estacional (E	Representa las irresoluciones estacionales, las mismas que se encuentran en datos trimestrales, mensuales y semanales.
Cíclica(C)	Se refiere a los ciclos que duran más de un año.
Irracional o accidentales (A)	Representas las vacilaciones impredecibles aleatorias.

Fuente: (Cardenas, 2020, p. 32).

Métodos predictivos

Los métodos predictivos hacen mención a las técnicas que se utilizan para prever o estimar futuros resultados o tendencias basándose en datos históricos y actuales, se basan en estadísticas, aprendizaje automático y minería de datos para modelar los datos y realizar predicciones (Fierro, 2020).

Métodos predictivos estadísticos

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

Los modelos predictivos son un conjunto de métodos que utilizan el aprendizaje automático, la extracción de datos históricos, los macrodatos y el reconocimiento de patrones para anticipar resultados futuros con el fin de mejorar la toma de decisiones mediante métodos de análisis de datos (Fernandez, 2021).

- *Modelo Aditivo.* - tiene como objetivo representar o expresar la variable de respuesta X, o un componente de esta, combinando las funciones suaves que se han aplicado a las distintas variables, que permiten establecer una relación no lineal entre la variable respuesta y cada una de las explicativas (Barreiro, 2019, p. 29).
- *Modelo Multiplicativo.* – Este modelo es de gran utilidad cuanto las fluctuaciones estacionales aumentan o disminuyen a medida que le aserie temporal avanza, es decir, la amplitud de la estacionalidad no es constante sino que varía en función del nivel de la serie.
- *Modelo Mixto.* – Se pueden aplicar cuando se dispone de grandes conjuntos de datos con estructuras jerárquicas o agrupaciones, y se quiere tener en cuenta la variabilidad tanto a nivel de grupo como a nivel individual. Esto permite un análisis más preciso y flexibilidad al modelar diferentes tipos de variabilidad en los datos. (al. A. L., 2016, p. 4).

Modelos para el pronóstico de una serie temporal

- *Modelo Autorregresivo Integrado de Media Móvil (ARIMA): "*Este modelo utiliza datos no estacionarios, es univariado y se encuentra entre los métodos estadísticos más simples y populares que se aplican en series de tiempo" (Gualoto, 2021, pág. 4).
- *Modelo Autorregresivo Integrado de Media Móvil Estacional (SARIMA):* Este modelo añade el componente estacional al proceso regular alterando la

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

estacionariedad de la serie a lo largo del tiempo e influyendo sustancialmente en el proceso estocástico, parte del procedimiento habitual. El modelo se compone de términos de orden estacional (P, d, q) de media móvil, integración y autorregresivos, así como de términos de orden no estacionario (p (p, d, q)) y términos de orden estacional autorregresiva, de integración y de media móvil (P, D, Q) s (Bravo C. C., 2021, p. 28).

- *Función de autocorrelación (PACF) ACR:* "La autocorrelación parcial calcula el grado de correlación entre dos variables que se dividen en n periodos, cuando se neutraliza la dependencia línea que se crea por los retrasos intermedios entre ambas" (Gualoto, 2021, p. 4).
- *Modelos Autorregresivo (AR):* Consiste en reunir información sobre el pasado de la variable, observar su pasado, observar su trayectoria a lo largo del tiempo trayectoria a lo largo del tiempo y explorar el patrón de regularidad que muestran los datos. Dado que la modelización univariante es sencilla y se utiliza para definir el valor que se designa en un momento t, una variable económica que depende del tiempo, la dependencia temporal, un método de método consiste en recopilar información sobre el pasado de la variable (Sarmiento & Alayón, 2013, p. 3).
- *Modelo de Media Móvil (MA):* El modelo de media móvil es una agrupación lineal de términos de errores presentes y rezagados (Gualoto, 2021, p. 14).

Algoritmos de regresión

Regresión Lineal. – El método de los mínimos cuadrados se basa en determinar la suma de los cuadrados de las diferencias entre los puntos reales y los que se definen por la recta trazada a partir de las variables presentes en el modelo. En este caso, la mejor estimación es aquella que minimiza estas diferencias. Para discernir cuál modelo se ajusta mejor a los datos disponibles en

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

el modelo de regresión lineal, se comparan los valores de F obtenidos en cada uno de los modelos de regresión desarrollados. Si se aplican las técnicas de selección de variables mencionadas anteriormente, este coeficiente se calculará cada vez que se añada o retire una variable. Esto se debe a que, al realizar este procedimiento, en realidad se están generando nuevos modelos de regresión (Peláez, 2016).

Regresión Múltiple. - es la extensión del modelo de regresión simple a k variables explicativas (Carrasquilla-Batista1 & Alfonso, 2016, p. 9).

Metodologías de Desarrollos de proyectos Big Data

CRISP-DM

CRISP-DM, es una metodología iterativa, flexible y adaptable. Proporciona un enfoque estructurado y detallado para planificar y llevar a cabo un proyecto de minería de datos. Cuenta con seis fases tales como la comprensión del negocio; comprensión de los datos: preparación de los datos; modelado; evaluación; despliegue (Organiche, Alfaro, & Barrera, 2020).

SEMMA

Es un enfoque estructurado para el análisis de datos que fue desarrollado por SAS Institute. SEMMA es el acrónimo de las cinco etapas principales del proceso: *Sample, Explore, Modify, Model y Assess*. Proporciona un enfoque lógico y secuencial que guía a los analistas a través de las fases mencionadas anteriormente (Campaña Chanta & Chambi Vargas, 2022).

KDD

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

KDD, combina el descubrimiento y el análisis de los datos, el proceso permite buscar patrones de interés y conocimientos en grandes conjuntos de datos. El proceso KDD es iterativo, es decir es posible regresar a la etapa anterior en caso de que sea necesario.

DATLAS

La metodología DATLAS tiene como objetivo manejar la uso de la Ciencia de Datos, consta de cinco fases (definir el problema; análisis de los datos; preparación de los datos; modelado; interpretación de los resultados). Enfatiza la importancia de trabajar con los usuarios finales para comprender sus necesidades y de comunicar los resultados de los proyectos a los interesados de manera efectiva (Cardozo, 2021).

Análisis comparativo de las metodologías de desarrollo de proyectos Big Data

La siguiente matriz presenta un análisis comparativo cualitativo de las principales metodologías empleadas en el desarrollo de proyectos de Big Data. Se han considerado indicadores relacionados con el enfoque, la flexibilidad y la integración al negocio que caracterizan a cada metodología.

Tabla 1. Cuadro comparativo metodologías para proyectos Big Data. Fuente: Autoría Propia.

	CRISP-DM	**SEMMA**	**KDD**	**DATLAS**
Enfoque	Proceso general de minería de datos	Proceso orientado a la solución SAS	Generalización del proceso de descubrimiento de conocimientos	Modelo de analítica de datos y aprendizaje automático

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Flexibilidad	Diseñado para ser genérico y aplicable en diferentes sectores y tecnologías	Específico para herramientas SAS, aunque las fases son aplicables en general.	Muy general y amplio, requiere adaptaciones específicas para implementaciones concretas.	Diseñado para ser adaptable y con un enfoque práctico en proyectos de análisis de datos.
Integración de negocios	Fuerte énfasis en la comprensión y adaptación al negocio.	Más centrado en el análisis técnico de los datos.	Comienza con un enfoque más técnico, aunque considera la interpretación y evaluación en el contexto de negocio.	Enfoque equilibrado entre el negocio y la técnica, con énfasis en la implementación y seguimiento del proyecto.

El cuadro compara 4 metodologías para proyectos de Big Data: CRISP-DM, SEMMA, KDD y DATLAS. Cada una provee un marco de trabajo estructurado con diferentes fortalezas y enfoques. Sin embargo, tras analizar las variables comparadas, se concluye que CRISP-DM es la opción más recomendable por su flexibilidad y equilibrio entre perspectivas de negocio y técnicas.

CRISP-DM se destaca por ser un proceso genérico diseñado para adaptarse a diversos sectores y tecnologías. Cuenta con una fuerte orientación al entendimiento del negocio y a la integración entre requerimientos y soluciones analíticas. Sus características la convierten en una metodología integral y balanceada para encarar con éxito proyectos de minería de datos y responder a problemáticas de distinta índole.

METODOLOGÍA

Enfoque de la investigación

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

El enfoque de la investigación es cuantitativo, ya que se basa en la recopilación y análisis de datos numéricos. El objetivo es realizar un pronóstico de la migración humana del Ecuador, utilizando modelos de series temporales.

La presente investigación utiliza la metodología CRISP-DM para el desarrollo del modelo predictivo, por ser un proceso estructurado e iterativo diseñado específicamente para proyectos de minería de datos.

Nivel de la investigación

La investigación es de tipo descriptivo, ya que se pretende describir los patrones de migración humana del Ecuador. Además, es predictivo, ya que se pretende generar pronósticos de la migración humana del Ecuador.

Técnicas e instrumentos de la investigación

Los datos para la investigación se recopilarán de fuentes secundarias, como la Oficina Nacional de Estadísticas y Censos (INEC) y la Organización Internacional para las Migraciones (OIM). Los datos se recopilarán en forma de series de tiempo, que representan la cantidad de migrantes que ingresan o salen del Ecuador en un determinado periodo de tiempo.

RESULTADOS

Para el desarrollo del modelo de pronóstico de migración humana se ha utilizado la metodología CRISP DM, así como un conjunto de datos de las entradas y salidas del país de ciudadanos ecuatorianos desde el año 2019.

Comprensión del negocio.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

En los últimos años, Ecuador ha experimentado una serie de eventos inesperados que han tenido un impacto significativo en su migración. Estos eventos incluyen un terremoto en 2016, un cambio de gobierno hacia la derecha en 2017, y la llegada de la pandemia de COVID-19 en 2020. Estos factores llevaron a un deterioro en las condiciones de vida y a un aumento en la migración de ecuatorianos en busca de mejores oportunidades en el extranjero. En 2022, se agregó un problema adicional a la situación, con un aumento en la violencia y las muertes en el país. La combinación de factores negativos, como las políticas neoliberales, la pobreza, el desempleo, el mal manejo de la pandemia y la inseguridad, provocó un aumento en la migración. El año 2022 cerró con una gran cantidad de ecuatorianos emigrando, de acuerdo con el Instituto Nacional de Estadística y Censos (INEC), en 2022, el Ecuador experimentó un éxodo notable. Con 1.127.891 emigrantes, esto representó el 6,44% de la población total, dirigiéndose principalmente a España, Estados Unidos e Italia.

Un modelo para el pronóstico de migración humana aporta a los siguientes puntos:

Planificación Gubernamental:

- **Políticas Públicas:** Un pronóstico preciso permitirá al gobierno ecuatoriano diseñar políticas públicas más efectivas en áreas como empleo, educación, salud y vivienda, adaptándose a las necesidades cambiantes de la población.
- **Seguridad y Control:** La anticipación de flujos migratorios puede ayudar a las autoridades a prepararse para posibles desafíos de seguridad y control fronterizo.

Desarrollo Económico:

- **Mercado Laboral:** Conocer las tendencias migratorias permitirá anticipar la demanda de empleo y las necesidades de capacitación, así como identificar posibles faltantes o excedentes en ciertas profesiones o sectores.

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

- **Remesas:** Las remesas son una fuente crucial de ingresos para muchas familias ecuatorianas. Prever las tendencias migratorias puede ayudar a anticipar las fluctuaciones en las remesas y su impacto en la economía.

Desarrollo Sostenible:

- **Recursos:** La anticipación de flujos migratorios puede ayudar en la gestión adecuada de recursos como agua, energía y alimentos.

Comprensión de los datos.

Para el análisis de los datos se ha tomado los conjuntos de datos de las entradas y salidas internacionales de los años 2019, 2020, 2021, 2022; cada conjunto de datos cuenta con 24 columnas que se encuentran detalladas en la siguiente tabla:

Tabla 2. Conjuntos de datos de entradas y salidas internacionales. Fuente: Autoría Propia

Nombre del campo	**Descripción del campo**
tip_movi	Tipo de movimiento
tip_naci	Tipo de nacionalidad
anio_movi	Año de movimiento
mes_movi	Mes de movimiento
dia_movi	Día de movimiento
sex_migr	Sexo
nac_migr	Nacionalidad
subcont_nac	Subcontinente de Nacionalidad
cont_nac	Continente de Nacionalidad
via_tran	Medio de transporte
mot_viam	Motivo de viaje
pais_prod	País de procedencia destino
subcont_prod	Subcontinente de Procedencia destino
cont_prod	Continente de Procedencia Destino
lug_prod	Lugar de procedencia destino

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

pais_res	País de residencia
subcont_res	Subcontinente de Residencia
cont_res	Continente de Residencia
jef_migr	Jefatura de Migración
pro_jefm	Provincia Jefatura de Migración
can_jefm	Cantón Jefatura de Migración
cla_migr	Clase de migración
ocu_migr	Ocupación
Edad	Edad

Estos datos fueron extraídos del INEC (Instituto Nacional de Estadísticas y Censos), el registro estadístico de entradas y salidas cuantifica los movimientos internacionales en Ecuador, considerando diversas nacionalidades y medios de transporte. Esta información es recolectada por las jefaturas de control migratorio bajo la Subsecretaría de Migración del Ministerio del Interior.

Preparación de los datos.

Para la generación del modelo se ha realizado un filtro de los registros pertenecientes a las salidas de ciudadanos que residen en el Ecuador. Además, se ha tomado los campos *mes_movi y dia_movi*. A estos registros se los ha agrupado por *mes_movi y dia_movi* para obtener la serie temporal con el *recuento* de los registros y se ha agregado el campo *índice*, obteniendo los siguientes resultados.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

	indice	anio_movi	mes_movi	Recuento
0	1	2019	Enero	65645
1	2	2019	Febrero	98212
2	3	2019	Marzo	127380
3	4	2019	Abril	114383
4	5	2019	Mayo	114935
5	6	2019	Junio	86652
6	7	2019	Julio	109185
7	8	2019	Agosto	152782
8	9	2019	Septiembre	118464
9	10	2019	Octubre	111900
10	11	2019	Noviembre	118164
11	12	2019	Diciembre	111106
12	13	2020	Enero	83959
13	14	2020	Febrero	125151
14	15	2021	Enero	30809

Ilustración 1. Serie Temporal de salidas del país, por mes y año de referencia.

Se ha generado la siguiente gráfica para conocer el comportamiento de los datos pertenecientes a la serie temporal. En donde se puede considerar que la distribución de los datos manifiesta una estacionalidad en las primeras observaciones, en el año 2020 existe un comportamiento atípico, correspondiente a la emergencia sanitaria COVID 2019, en donde no existió flujo migratorio. Por otro lado, se ha podido constatar a partir del año 2021 existe una tendencia creciente del flujo migratorio.

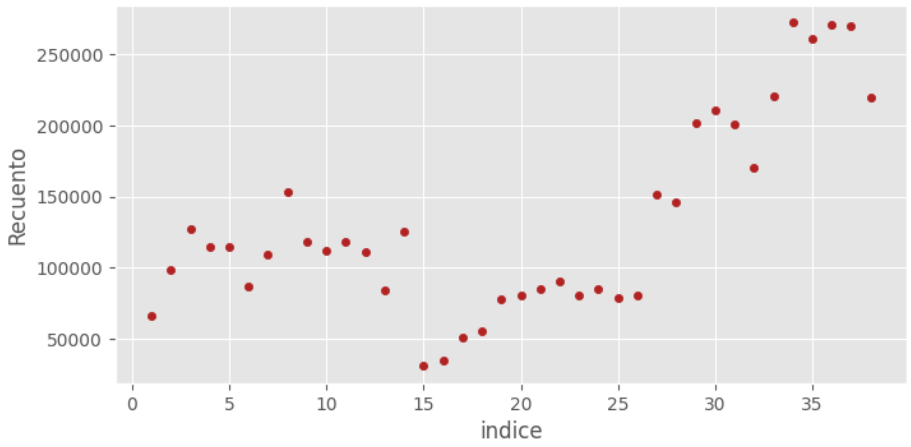


Ilustración 2. Serie Temporal de salidas del país, por mes y año de referencia.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

En la siguiente grafica se puede observar la serie temporal completa, en donde se puede observar que durante el año 2020 no se obtiene datos considerando que la salida del país estaba restringida por las medidas gubernamentales frente a la crisis sanitaria COVID 2019.

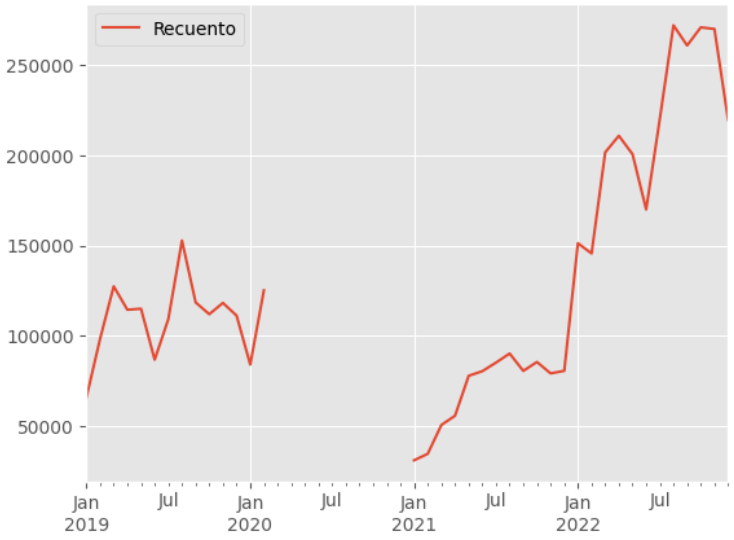


Ilustración 3. Serie Temporal completa.

Considerando que el conjunto de datos tiene valores nulos, atípicos y un notable cambio de tendencia posterior al año 2020, se toma como medida para el tratamiento de los datos omitir los registros anteriores al año 2021. Obteniendo la distribución presentada en la siguiente ilustración:

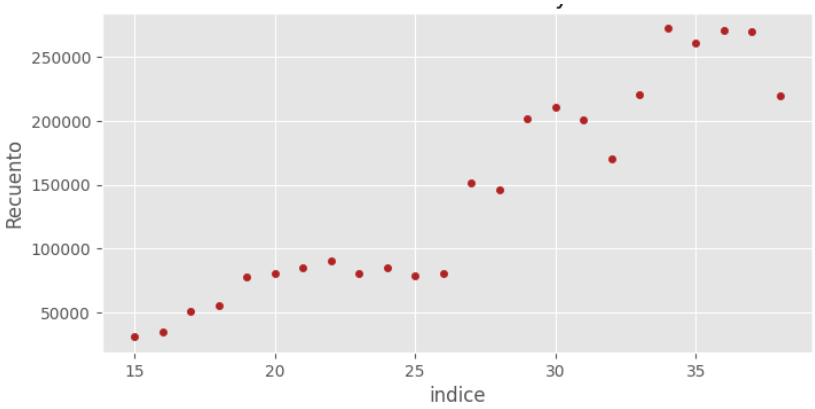


Ilustración 4. Serie Temporal distribución de datos

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Modelado

Para la generación del modelo se ha utilizado regresión lineal y el método estadístico ARIMA, se han realizado una serie de pruebas para determinar cuál se ajusta mejor a los datos y presenta un pronóstico con menor error.

Regresión lineal

Para el modelo de regresión lineal se ha utilizado como variable predictora a *índice*, y como valor resultado a *recuento*, quedando de la siguiente manera:

$$Y = \beta 0 + \beta 1\ X1 + \varepsilon$$

Donde, β0 y β1 son los coeficientes de regresión, X será la variable predictor, en este caso el *Índice*, y ε será el error aleatorio.

Previo a la generación del modelo se realiza el análisis de correlación entre las dos variables, obteniendo los resultados mostrados en la siguiente ilustración:

```
Coeficiente de correlación de Pearson:  0.9444148500199893
P-value:  4.165710610829909e-12
```

Ilustración 5. Resultados del análisis de correlación entre la variable predictora y la variable resultado

El coeficiente de correlación de Pearson mide la relación lineal entre dos conjuntos de datos. Los valores del coeficiente varían entre -1 y 1, donde: 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta y 0 indica que no hay correlación. En el caso del presente estudio, un coeficiente de correlación de Pearson de 0.9444 sugiere una correlación positiva muy fuerte entre las variables "índice" y "recuento". Esto significa que, en general, cuando una variable aumenta, la otra también tiende a aumentar en una proporción cercana

El valor P (P-value) es una medida que nos ayuda a determinar la significancia estadística de los resultados. Un valor P pequeño (típicamente ≤ 0.05) indica que podemos rechazar la hipótesis nula. En este contexto, la hipótesis nula suele ser que no hay correlación entre las dos variables. Dado que tu P-value es mucho menor que 0.05, podemos rechazar la hipótesis nula con confianza y concluir que hay una correlación estadísticamente significativa entre las variables "indice" y "recuento".

En resumen, los resultados sugieren que hay una correlación positiva muy fuerte y estadísticamente significativa entre las variables "índice" y "recuento". Es uno de los valores más

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

altos posibles para el coeficiente de Pearson, lo que indica una relación lineal casi perfecta entre las dos variables.

División de los datos en train y test

Para la generación del modelo predictivo se divide el dataset en, dataset de entrenamiento y el dataset de prueba, con el objeto de entrenar el modelo para obtener una predicción de la variable *recuento* de salidas del país, en el dataset de prueba. Para ello se toma el 80% de datos para entrenamiento y el 20% para datos prueba.

Creación del modelo

Una vez realizada la división del conjunto de datos, se procede a entrenar al modelo, obteniendo los siguientes resultados

```
Intercept: [-161028.0242609]
Coeficiente: [('indice', 11378.047858942062)]
Coeficiente de determinación R^2: 0.8903997127916282
```

El intercepto representa el valor esperado de la variable respuesta cuando la variable predictora ("índice") es 0. En otras palabras, para la primera observación del modelo la respuesta será aproximadamente -161,028.0243.

El coeficiente es 11,378.0479. Esto significa cada mes u observación de la serie temporal, se espera que la respuesta aumente en 11,378.0479 unidades, manteniendo todo lo demás constante.

El coeficiente de determinación, conocido como R2, mide la proporción de la variabilidad total de la variable respuesta que es explicada por el modelo. En este caso indica que aproximadamente el 89.04% de la variabilidad en la respuesta es explicada por el modelo de regresión lineal que has ajustado con la variable "índice". Es un valor muy alto, lo que sugiere que el modelo se ajusta bien a los datos y explica una gran proporción de la variabilidad observada.

Utilizando el método de mínimos cuadrados ordinarios (OLS), se obtiene los siguientes resultados:

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

OLS Regression Results

Dep. Variable:	y	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.853
Method:	Least Squares	F-statistic:	105.7
Date:	Wed, 06 Sep 2023	Prob (F-statistic):	1.03e-08
Time:	21:07:17	Log-Likelihood:	-222.51
No. Observations:	19	AIC:	449.0
Df Residuals:	17	BIC:	450.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.61e+05	3.18e+04	-5.063	0.000	-2.28e+05	-9.39e+04
x1	1.138e+04	1106.825	10.280	0.000	9042.852	1.37e+04

Omnibus:	2.054	Durbin-Watson:	1.486
Prob(Omnibus):	0.358	Jarque-Bera (JB):	1.472
Skew:	-0.482	Prob(JB):	0.479
Kurtosis:	2.035	Cond. No.	128.

Ilustración 6. Resultados de regresión

El modelo parece ser estadísticamente significativo y la variable predictora x1 tiene un efecto positivo significativo en la variable respuesta y. El modelo explica aproximadamente el 86.1% de la variabilidad.

En la siguiente gráfica se puede observar las observaciones originales frente a los valores que son predichos por el modelo de regresión lineal, frente a un intervalo de confianza del 95%. En donde se puede observar un alto porcentaje de observaciones se ajustan al modelo.

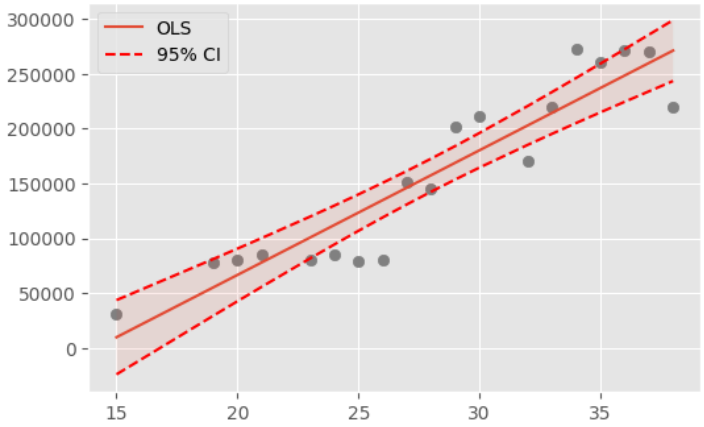


Ilustración 7. Resultados modelos de regresión lineal.

ARIMA

ARIMA, "Modelo autorregresivo integrado de media móvil" (*AutoRegressive Integrated Moving Average*, en inglés), es un modelo ampliamente utilizado para predecir series temporales. ARIMA

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

combina tres componentes principales: modelos autorregresivos (AR), modelos de media móvil (MA) e integración (I). Antes de implementar un modelo ARIMA, es esencial realizar un análisis exploratorio y preparar la serie para asegurar que el modelo ARIMA sea adecuado y efectivo. Para ello es necesario verificar si la serie es estacionaria, observar si hay alguna tendencia o patrón estacional en la serie, verificar la Autocorrelación y Autocorrelación Parcial.

Para el primer punto se ejecuta prueba Dickey-Fuller, para verificar si la serie es estacionaria, obteniendo para *p-value* un valor 0.998477 y dado que el valor p es mucho mayor que 0.05, no se puede rechazar la hipótesis nula. Esto sugiere que la serie no es estacionaria. Por lo cual es esencial transformarla en una serie estacionaria antes de aplicar ARIMA.

Para comprender el comportamiento de la serie y la existencia de una tendencia o patrón estacional procedemos a descomponer la serie obteniendo los siguientes resultados:

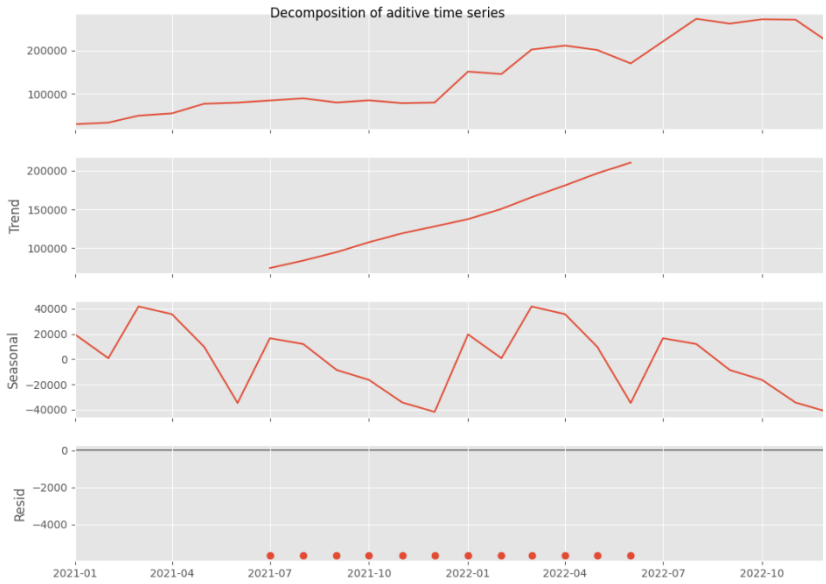


Ilustración 8. Resultados modelo ARIMA.

De acuerdo a los valores presentados en la tendencia, se puede observar que desde septiembre de 2021 existe un fuerte aumento en la variable de interés durante este período. Se puede asociar este comportamiento a factores económicos, políticos y de seguridad que han afectado al país. Además, se presenta estacionalidad en el modelo pues se visualiza patrones repetidos de manera cíclica, después de eliminar la tendencia y la estacionalidad, la variabilidad restante en la serie es relativamente baja durante este período.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Dado que la serie muestra una tendencia y estacionalidad claras, se considera usar modelos como ARIMA estacional (SARIMA), utilizando el método *AutoArima* se obtiene el modelo que se ajusta de manera eficiente a los datos (0,1,0) (0,1,0) [12], es un modelo simple que solo utiliza la diferenciación (tanto a nivel regular como estacional) para hacer predicciones. No utiliza observaciones pasadas ni errores pasados para predecir valores futuros.

En la siguiente gráfica se presenta un resumen del modelo generado, en el cuál entre los aspectos importantes se puede destacar los siguientes aspectos: El coeficiente "sigma2" es significativo a un nivel de significancia del 5% (ya que el p-valor es menor que 0.05), hay evidencia suficiente para rechazar la hipótesis nula de que "sigma2" es igual a cero, la varianza del error o residuo en el modelo es significativamente diferente de cero.

El intervalo de confianza del 95% nos da una idea de la incertidumbre asociada con la estimación de "sigma2". Es probable que el verdadero valor de "sigma2" se encuentre dentro de este intervalo.

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	7.014e+08	2.91e+08	2.408	0.016	1.3e+08	1.27e+09

Una vez entrenado el modelo ARIMA con los datos de entrenamiento, obtenemos los resultados mostrados en la siguiente gráfica, en donde se puede observar que el pronóstico se ajusta adecuadamente a los datos.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

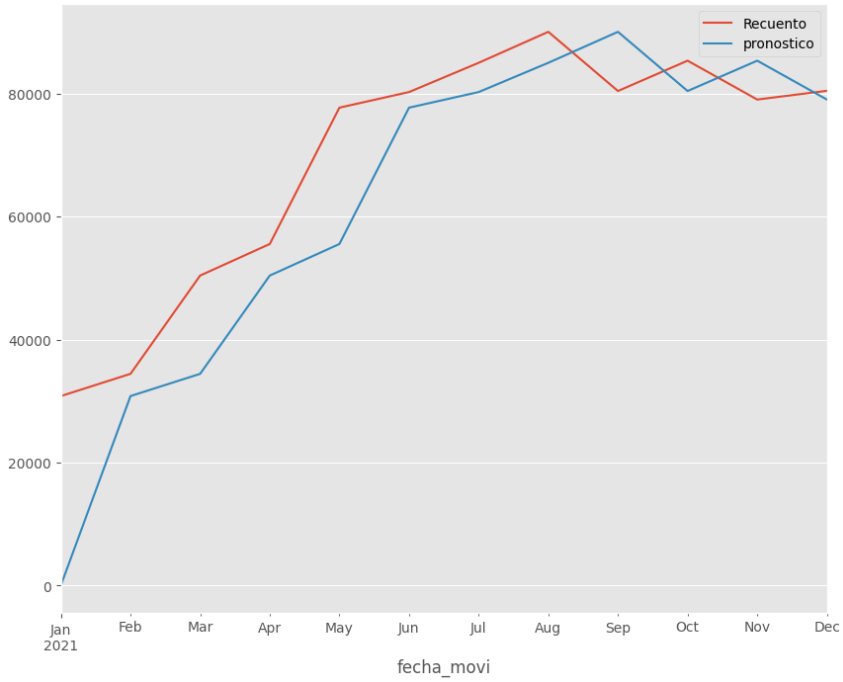


Ilustración 9. Resultados modelo ARIMA

Evaluación.

Para la evaluación de los modelos se ha tomado como referencia MAE (Error Absoluto Medio): que representa la diferencia promedio entre los valores observados y los valores predichos y RMSE (Raíz del Error Cuadrático Medio): que es la raíz cuadrada de la media de los errores al cuadrado y penaliza más los errores grandes, por lo que es más sensible a outliers que el MAE.

En la siguiente tabla se puede observar los valores obtenidos de cada uno de los modelos:

Tabla 3. Valores finales de los modelos

MODELO	MAE	RMSE
REGRESIÓN LINEAL	12043.10	12043.10
RMSE	25980.58	33220.56

El modelo de regresión lineal tiene un MAE mucho más bajo (12043.10) en comparación con el modelo ARIMA (25980.58), esto sugiere que, en promedio, el modelo de regresión lineal predice más cerca de los valores reales que el modelo ARIMA. Por otro lado, el RMSE del modelo de

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

regresión lineal es 12043.10, mientras que el RMSE del modelo ARIMA es 33220.56, lo que indica que el modelo ARIMA tiene errores más grandes en comparación con el modelo de regresión lineal.

En base a los errores MAE y RMSE, el modelo de regresión lineal parece ser más preciso en sus predicciones en comparación con el modelo ARIMA para tu serie temporal. Es importante tener en cuenta que, aunque el modelo de regresión lineal tiene errores más bajos en este caso, no siempre será el mejor modelo para todas las series temporales. La elección del modelo debe basarse en la naturaleza de los datos, la presencia de tendencias y estacionalidades, y otros factores específicos del problema.

Visualmente se puede determinar los siguientes aspectos:

La regresión lineal ha capturado la tendencia general de los datos, lo cual genera ciertas imprecisiones debido a la naturaleza de los mismos. Por otro lado, ARIMA ha capturado patrones más complejos y estacionales, pero se visibiliza un desfase que puede generar problemas y requiere una revisión adicional de los parámetros o una consideración de otros factores externos para mejorar la predicción.

Despliegue

Para implementar un modelo de pronóstico es necesario establecer el modelo que mejor se adapte a las necesidades para ellos se toma en referencia las conclusiones de la evaluación del modelo y de determina los siguientes aspectos:

El modelo de regresión lineal tiene un MAE y RMSE mucho más bajos que el modelo ARIMA. Estas métricas indican que, en promedio, las predicciones del modelo de regresión lineal están más cerca de los valores reales que las del modelo ARIMA. Aunque los datos no son netamente lineales y algunos se ajustan a la línea de regresión, el error es significativamente menor en comparación con ARIMA.

Por otro lado, aunque el modelo ARIMA captura un comportamiento similar, está desfasado, lo que significa que no está prediciendo correctamente en el tiempo correcto. Este desfase puede ser problemático, especialmente si las predicciones temporales precisas son cruciales.

En base a las métricas de error y al comportamiento grafico de la predicción, el modelo de regresión lineal es más exacto en sus predicciones en comparación con el modelo ARIMA, por lo que para este estudio se toma como referencia el modelo de regresión lineal.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

DISCUSIÓN

Los resultados obtenidos de la regresión lineal y ARIMA reflejan una diferencia fundamental entre los modelos lineales y no lineales. Mientras que la regresión lineal asume una relación lineal entre las variables y puede no capturar patrones más complejos, ARIMA, siendo un modelo no lineal, tiene la capacidad de capturar tendencias y estacionalidades en los datos. Sin embargo, la elección de parámetros en ARIMA es crucial, y un mal ajuste puede llevar a predicciones desfasadas o inexactas, como observaste.

Aunque la regresión lineal proporcionó un error más bajo en los resultados, es esencial considerar la naturaleza de los datos. Si los datos tienen patrones no lineales o estacionalidades, un modelo lineal puede no ser adecuado a largo plazo, incluso si se proporciona buenos resultados a corto plazo. Por otro lado, mientras que ARIMA puede capturar patrones más complejos, la elección correcta de parámetros y la consideración de factores externos son cruciales para su precisión.

Dado que ambos modelos presentan ventajas y limitaciones, podría ser útil considerar enfoques híbridos o explorar otros modelos avanzados. Por ejemplo, las redes neuronales, especialmente las redes LSTM (Long Short-Term Memory), que han demostrado ser efectivas para el pronóstico de series temporales, ya que pueden capturar relaciones a largo plazo y patrones no lineales en los datos. Sin embargo, requieren una gran cantidad de datos y pueden ser computacionalmente intensivas.

En temas tan críticos como la migración, la capacidad de prever tendencias futuras permite a los gobiernos y organizaciones tomar decisiones informadas. Ya sea en la asignación de recursos, la preparación de infraestructura o la formulación de políticas, un pronóstico preciso puede ser la diferencia entre una respuesta efectiva y una crisis. Los modelos de series temporales ofrecen herramientas valiosas para el pronóstico en temas sociales como la migración, es esencial aplicarlos con cuidado, comprensión y una fuerte orientación ética. La combinación de análisis cuantitativo con consideraciones humanitarias y éticas puede conducir a soluciones más efectivas y compasivas.

CONCLUSIONES

La aplicación de la metodología CRISP-DM permitió estructurar el desarrollo del modelo predictivo siguiendo un enfoque iterativo y adaptable a las necesidades del problema. Las fases

e-ISSN: 2588-1000
Pro Sciences
Revista de Producción, Ciencias e Investigación

de comprensión del negocio, análisis de datos, preparación de datos, modelado y evaluación facilitaron la construcción de un modelo de pronóstico de migración humana sólido y confiable.

Se exploraron dos modelos de pronóstico: regresión lineal y ARIMA estacional. En donde se determina que la regresión lineal obtuvo menor error de predicción, explicando cerca del 90% de la variabilidad de los datos. El modelo ARIMA capturó patrones estacionales y complejos de los datos, pero tuvo un desfase en las predicciones que afectó su precisión.

Cabe destacar la importancia de seleccionar un modelo de pronóstico que se alinee con la naturaleza de los datos, reconociendo la existencia de tendencias y patrones no lineales. En este estudio, la regresión lineal demostró ser el enfoque más preciso. Al analizar los datos migratorios de Ecuador entre 2019 y 2022, se evidenció un crecimiento en la tendencia migratoria, matizada por fluctuaciones estacionales y una notable perturbación debido a la pandemia de COVID-19 en 2020.

De esta manera, se ultima que los modelos de series de tiempo son valiosos para realizar pronósticos en temas sociales como migración, pero se debe combinar con consideraciones éticas y humanitarias. Asimismo, el pronóstico preciso de la migración permite una mejor planificación de recursos, infraestructura y políticas públicas por parte de los gobiernos.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Referencias

al., A. L. (2016). Ronda clínica y epidemiológica. Series de tiempo. *Scielo*, 9.

Arana, C. (2021). Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales. *ECONSTOR*, 1-26.

Ballester, D. G. (24 de 07 de 2018). *riunet.upv.es.* Obtenido de riunet.upv.es: https://riunet.upv.es/bitstream/handle/10251/152398/Garc%c3%ada%20-%20Predicci%c3%b3n%20del%20precio%20de%20billetes%20de%20avi%c3%b3n%20a%20partir%20de%20una%20red%20neuronal%20caracterizada%20po....pdf?sequence=1&isAllowed=y

Barragan, F. D., Cartagena, N. G., Arroyo, G. F., & Mina, J. R. (2022). Inobervancia a los derechos humanos: migración irregular de grupos vulnerables a Ecuador. *Revista Universidad y Sociedad*, 108-117.

Barreiro, P. A. (07 de 2019). *idus.us.es*. Obtenido de idus.us.es: https://idus.us.es/handle/11441/89999

Basso, M. (2021). Análisis de las migraciones internas en Argentina en el período 2005-2010. *Scielo*, 112-143.

Bianchi, F. D. (2001). *catedra.ing.unlp.edu.ar*. Obtenido de https://catedra.ing.unlp.edu.ar: https://catedra.ing.unlp.edu.ar/electrotecnia/senysis/files/apuntes/ResumenMatlab2.pdf

Bravo, C. C. (01 de 01 de 2021). *https://repositorio.unal.edu.co.* Obtenido de https://repositorio.unal.edu.co: https://repositorio.unal.edu.co/bitstream/handle/unal/81124/1018467917.2021.pdf?sequence=3&isAllowed=y

Bravo, C. C. (2022). Estudio del fenómeno migratorio Ecuador-Estados Unidos:implicaciones desde la política internacional en el periodo 2018-2020. *REVISTA UNIVERSIDAD DE GUAYAQUIL* , 47-66.

Campaña Chanta, T., & Chambi Vargas, G. L. (14 de 06 de 2022). *repositorioacademico.upc.edu.pe.* Obtenido de repositorioacademico.upc.edu.pe: https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/667677/Campa%c3%b1a_CT.pdf?sequence=3&isAllowed=y

Cardenas, A. P. (2020). *repositorio.unsaac.edu.pe.* Obtenido de repositorio.unsaac.edu.pe: http://repositorio.unsaac.edu.pe/bitstream/handle/20.500.12918/5669/253T20201031_TC.pdf?sequence=1&isAllowed=y

Cardozo, L. (2021). Metodología Datlas.

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Carrasquilla-Batista1, A., & A. C.-R. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Scielo*, 13.

Cruz, B. (2007). Redes neuronales recurrentes para el análisis de secuencias. *Revista Cubana de Ciencias Informaticas*, 57.

Fernandez, S. d. (2021). *estadistica.net*. Obtenido de estadistica.net: https://www.estadistica.net/Algoritmos2/series-temporales.pdf

Fierro, A. A. (01 de 06 de 2020). *sedici.unlp.edu.ar.* Obtenido de sedici.unlp.edu.ar: http://sedici.unlp.edu.ar/bitstream/handle/10915/114857/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Gualoto, S. X. (01 de 05 de 2021). *dspace.ups.edu.ec.* Obtenido de dspace.ups.edu.ec: https://dspace.ups.edu.ec/bitstream/123456789/20237/1/UPS%20-%20TTS356.pdf

Guerrero, J. P. (06 de 2020). *idus.us.es.* Obtenido de idus.us.es: https://idus.us.es/bitstream/handle/11441/115230/TFG%20DGMyE%20P%C3%A9rez%20Guerrero%2C%20Jes%C3%BAs.pdf?sequence=1&isAllowed=y

Instituto nacional de estadística y censos. (22 de 01 de 2022). *www.ecuadorencifras.gob.ec*. Obtenido de www.ecuadorencifras.gob.ec: https://www.ecuadorencifras.gob.ec/entradas-y-salidas-internacionales/#:~:text=En%202022%2C%20el%20flujo%20migratorio,es%20el%20encargado%20del%20procesamiento.

Méndez, M., & Gómez, J. (2022). La migración internacional como agente de desarrollo local para las naciones. *Revista Latinoamericana de Difusión Científica*, 257-269.

Organiche, E. C., Alfaro, A. J., & Barrera, G. C. (2020). Principales Metodologías en el Desarrollo de Proyectos de Minería de Datos. *TECNOCULTURA*.

Ortiz Martos, A. J., Martín Valdivia, M. T., & Ureña López, L. A. (s.f.). Detección automática de Spam utilizando Regresión Logística. *Procesamiento del Lenguaje Natural*, 8.

Pacheco, J. L., Suárez, A. I., & Argüello, M. V. (01 de 01 de 2020). *core.ac.uk.* Obtenido de https://core.ac.uk/: https://core.ac.uk/download/pdf/288306071.pdf

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. *Revista Seden*, 20.

Sarmiento, D. A., & Alayón, C. A. (2013). Modelado de pérdidas en una transmisión de video por medio de series de tiempo ARIMA y SARIMA. *Scielo*, 11.

Silva, J. M., Borré, J. R., Montero, S. R., & Mendoza, X. F. (202). Migración: Contexto, impacto y desafío. *Redalyc*, 299-311.

Tena, F. P. (15 de 10 de 2022). *repositori.uji.es.* Obtenido de repositori.uji.es: https://repositori.uji.es/xmlui/bitstream/handle/10234/201358/TFG_2022_Peris_Tena_Francisco.pdf?sequence=1

e-ISSN: 2588-1000

Pro Sciences

Revista de Producción, Ciencias e Investigación

Valls, J., & Badiella, L. (s.f.). *sct.uab.cat.* Obtenido de sct.uab.cat: https://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/ManualSAS.PDF

Vargas, L. E., & Fuquen, E. M. (2021). *Introducción al análisis de datos con RStudio.* Bogotá: Cenipalma.

Walker, J. S. (2018). *Python: La Guía Definitiva para Principiantes para Dominar Python.* Babelcube.

Cuenca, 13 de julio 2023

Asunto: Embargo Temporal del Trabajo de Titulación

Señor,
Ing. Leopoldo Pauta Ayabaca
DECANO DE LA UNIDAD ACADÉMICA DE INFORMÁTICA, CIENCIAS DE LA COMPUTACIÓN E INNOVACIÓN TECNOLÓGICA
De mi consideración:

Yo, CRISTIAN ANDRÉS ANDRADE GUALLPA , como autor del Trabajo de Titulación "PRONÓSTICO DE MIGRACIÓN HUMANA DEL ECUADOR, UTILIZANDO MODELOS DE SERIES TEMPORALES " y CRISTINA MARIUXI FLORES URGILÉS , MGS, como director de la misma, solicitamos a usted y por su digno intermedio a Biblioteca y al responsable del repositorio institucional, el EMBARGO TEMPORAL del mismo, por un lapso de 6 meses, con la finalidad de evaluar su contenido con fines de: evaluación de artículo científico para publicación en revista indexada. Entiendo que luego de vencido este período automáticamente la obra será puesta a disposición del público bajo las normas de gestión de la Universidad.

Por la atención que sepa dar al presente, nos suscribimos de usted muy agradecidos.

Atentamente,

_____
Cristian Andrés Andrade Guallpa
CI: 0302792007
Autor

C.C.: Biblioteca