Análisis predictivo de series temporales de producción de gas natural en Colombia implementando metodologías estadísticas y de machine learning

Héctor Andrés Llanos Vargas

Asesor

Eduardo Sánchez Sandoval

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica.

2026

Nota de Aceptación

Eduardo Sánchez Sandoval

_____

Nombre Director de Trabajo de Grado

Lina Rocio Rivadeneira

_____ _____

Jurado Jurado

2026

Resumen

La industria energética de Colombia desempeña un papel importante en la estabilidad económica del país y la explotación de hidrocarburos y en particular de gas, es una de las principales fuentes de ingresos y generación de empleo en las regiones productoras. De esta manera, La predicción de gas en Colombia resulta ser un indicador económico de gran importancia ya que impacta en la cantidad de regalías que recibe el departamento y al precio de venta para el consumo domiciliario o industrial que pagan los usuarios. En este contexto, se analizan diversos algoritmos relacionados con el pronóstico de la producción de gas utilizando técnicas estadísticas y modelos de machine learning.

Los datos fueron obtenidos de la página de datos abiertos y corresponde con la producción fiscalizada de gas por campo medida en millones de pies cúbicos por mes - mpcpm. Los datos se encuentran desagregados por departamento, municipio, operadora, campo y contrato para cada uno de los meses del año desde 2013 hasta noviembre de 2025. Se destacan enfoques estadísticos clásicos de series temporales ARIMA y SARIMA. Modelos de machine learning como ridge regression, XGBoost, random forest regressor y support vector machine regressor fueron usados en el presente trabajo. Como principal resultado se encontró que el método ridge regressor es el que mejor desempeño tiene por encima de otros métodos, con un R^2 de 0.82 y un MAPE de solo 2.65%. En conjunto, estos referentes respaldan la propuesta de aplicar y comparar modelos de machine learning y estadísticos tradicionales para evaluar la predicción final con el objetivo de plantear escenarios futuros en la producción de gas.

Palabras clave: Producción de gas, series temporales, machine learning, random forest, ridge regression.

Abstract

Colombia's energy industry plays a significant role in the country's economic stability, and hydrocarbon exploitation, particularly gas, is a major source of income and employment in producing regions. Therefore, gas forecasting in Colombia is a crucial economic indicator, as it impacts the amount of royalties received by the department and the retail price paid by consumers for residential and industrial use. In this context, various algorithms related to gas production forecasting are analyzed using statistical techniques and machine learning models.

The data were obtained from open data sources and corresponds to the audited gas production per field, measured in millions of cubic feet per month. The data is disaggregated by department, municipality, operator, field, and contract for each month of the year from 2013 to November 2025. Classical statistical approaches to time series analysis, ARIMA and SARIMA, are highlighted. Machine learning models such as ridge regression, XGBoost, random forest regressor, and support vector machine regressor were used in this work. The main finding was that the ridge regressor method performed best compared to other methods, with an R^2 of 0.82 and a MAPE of only 2.65%. Taken together, these results support the proposal to apply and compare machine learning and traditional statistical models to evaluate the final prediction with the aim of developing future scenarios for gas production.

Keywords: Gas production, time series, machine learning, random forest, ridge regression.

Contenido

Lista de Tablas

Lista de Figuras

Introducción

El estado crítico del sector gasífero en Colombia, documentado por la ANH y Ecopetrol, y cuyos datos históricos de producción están disponibles en portales como Datos Abiertos Colombia (Producción Fiscalizada de Gas 2017 | Datos Abiertos Colombia, n.d.) exige el uso de herramientas de pronóstico avanzadas. La literatura respalda claramente la transición de métodos puramente tradicionales a enfoques basados en machine learning y modelos híbridos para el pronóstico en el sector energético (Safiyari et al., 2022).

La propuesta tiene un enfoque comparativo para el problema de aplicación. Primero, se utilizarán modelos estadísticos como ARIMA (AutoRegressive Integrated Moving Average) y SARIMA (Seasonal AutoRegressive Integrated Moving Average) para establecer una línea base de rendimiento. Segundo, se implementarán y evaluarán diversos algoritmos de machine learning como ridge regression, XGBoost, random forest regressor y support vector machine regressor que han demostrado ser efectivos tanto en la predicción de la producción de gas (Amaechi et al., 2019) y (AlShafeey & Csáki, 2021)como en el pronóstico de la demanda energética en general. La comparación sistemática de estos modelos permitirá identificar la metodología más robusta para el pronóstico de la producción de gas, proporcionando una herramienta de alto valor para la gestión de la seguridad energética del país.

Descripción del Problema

El gas natural se ha consolidado como un pilar fundamental en la matriz energética de Colombia, impulsando el desarrollo industrial, la generación eléctrica y mejorando la calidad de vida en millones de hogares. Sin embargo, el país se encuentra en una encrucijada crítica. La producción nacional ha entrado en una fase de declive, las reservas probadas disminuyen a un ritmo preocupante y la dependencia de las importaciones va en aumento, generando un escenario de incertidumbre sobre la futura seguridad y soberanía energética.(Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos, n.d.-a) En este contexto, la implementación de metodologías avanzadas de análisis predictivo, como las series temporales y el aprendizaje automático (machine learning), emerge como una herramienta estratégica para optimizar la gestión de este recurso vital, anticipar escenarios y fundamentar decisiones críticas a nivel gubernamental y empresarial.

El panorama actual del gas natural en Colombia es complejo y desafiante. Durante el primer semestre de 2024, la producción de gas a nivel nacional fue de 991 millones de pies cúbicos por día (MPCD), lo que representó una disminución del 6.1% en comparación con el mismo período del año anterior, según datos de la cámara colombiana de petróleo, gas y energía esta tendencia a la baja no es un hecho aislado. Informes del Ministerio de Minas y Energía y de la Agencia Nacional de Hidrocarburos (Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos, n.d.-b) revelan que las reservas de gas natural a finales de 2023 se estimaban en 2.37 terapíes cúbicos (TCF), con una relación reservas/producción (R/P) de apenas 6.1 años. Esta cifra representa una caída significativa frente a años anteriores y enciende las alarmas sobre la autosuficiencia del país a mediano y largo plazo.

La producción de gas en Colombia se concentra geográficamente en unos pocos departamentos. Históricamente, Casanare ha sido el principal productor, aportando más de la mitad del gas del país, seguido por Córdoba y La Guajira. Sin embargo, la producción en estos campos está declinando. Si bien han surgido nuevos campos productores en el departamento Bolívar, su contribución no logra compensar la caída de los grandes yacimientos. Esta concentración geográfica de la producción también implica vulnerabilidades logísticas y de abastecimiento que deben ser gestionadas con precisión.(Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos, n.d.-a).

Ante la caída de la oferta local, Colombia ha tenido que recurrir de manera creciente a la importación de Gas Natural Licuado (GNL). En 2024, las importaciones de gas experimentaron un aumento significativo para poder suplir la demanda interna, especialmente durante fenómenos climáticos como "El Niño", que reducen la generación hidroeléctrica y aumentan la demanda de gas para la generación térmica. Esta dependencia del mercado internacional no solo tiene implicaciones en la balanza comercial, sino que también el país se expone a la volatilidad de los precios internacionales y a riesgos geopolíticos, amenazando la competitividad de la industria nacional y el bolsillo de los consumidores.

Planteamiento del Problema

Frente a este escenario, la capacidad de anticipar con mayor precisión la producción futura de gas natural en Colombia se vuelve crucial. Tradicionalmente, las proyecciones de producción se han basado en modelos y análisis geológicos. Si bien estos métodos son fundamentales, pueden ser complementados con técnicas estadísticas y de machine learning. El análisis de series temporales, por ejemplo, permite modelar y predecir el comportamiento futuro de la producción basándose en sus datos históricos, identificando tendencias, ciclos y estacionalidades(Monroy, 2008). Metodologías como ARIMA (Promedio Móvil Integrado Autoregresivo) y SARIMA (Promedio Móvil Integrado Autoregresivo con Temporalidad) han sido aplicadas con éxito en el pronóstico de producción de hidrocarburos a nivel internacional y en estudios preliminares en Colombia.

A su vez, los algoritmos de machine learning como, ridge regression, gradient boosting regressor (como XGBoost, LightGBM, CatBoost), random forest regressor y support vector machine regressor, ofrecen la capacidad de analizar patrones complejos y no lineales en los datos de producción, incorporando múltiples variables que pueden influir en ella, como la inversión en exploración y producción, la actividad de perforación, e incluso factores macroeconómicos.

Aunque la aplicación de estas técnicas en el sector de hidrocarburos colombiano es aún incipiente, su potencial para mejorar la precisión de los pronósticos es innegable. La optimización de la producción, la planificación de la infraestructura de transporte y almacenamiento, y la definición de políticas públicas para incentivar la exploración y garantizar el abastecimiento a largo plazo, son solo algunas de las áreas que se beneficiarían directamente de un enfoque predictivo más sofisticado.

Por lo tanto, el problema central de este trabajo de aplicación radica en la necesidad de superar las limitaciones de los modelos de pronóstico convencionales mediante la aplicación de metodologías estadísticas y de machine learning que permitan mejorar significativamente la precisión de las predicciones de producción de gas natural a nivel departamental en Colombia.

Esto conduce a la siguiente pregunta principal: ¿Cuál es el desempeño de los algoritmos de machine learning y estadística clásica para el pronóstico de producción de gas en Colombia en el periodo 2013 – 2025?

Justificación

La seguridad energética de Colombia enfrenta una amenaza inminente y estructural debido a la creciente brecha entre la oferta interna de gas natural y la demanda nacional. El país ha entrado en una fase de declive productivo en sus principales campos gasíferos, una situación agravada por la insuficiente adición de nuevas reservas. La relación reservas/producción (R/P), que actualmente se sitúa en un preocupante umbral de 5.9 años (Informes de Gestión y Sostenibilidad, n.d.) (Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos, n.d.-b), evidencia un horizonte de autosuficiencia críticamente corto. Este escenario obliga a Colombia a depender cada vez más de la importación de Gas Natural Licuado (GNL), un mercado caracterizado por su alta volatilidad de precios y su sensibilidad a factores geopolíticos externos, lo que introduce un elemento de vulnerabilidad económica y estratégica para la nación.

La gestión de este riesgo se ve obstaculizada por una limitante fundamental: la incertidumbre en los pronósticos de producción a nivel regional, las inversiones en exploración y producción. La negociación de contratos de suministro a largo plazo depende de proyecciones confiables sobre la cantidad de gas que estará disponible en cada cuenca productora (Análisis de La Inflación y El Crecimiento Económico Un Enfoque de Machine Learning. México 1990-2021, n.d.). Sin embargo, los métodos de pronóstico tradicionales, aunque valiosos, a menudo se basan en modelos estadísticos que pueden no capturar adecuadamente la complejidad y el comportamiento de los yacimientos, ni la interacción de múltiples variables operativas, geológicas y económicas que influyen en la producción.

Esta falta de herramientas predictivas avanzadas y de alta precisión constituye el núcleo del problema. La incapacidad para anticipar con mayor certeza el declive en la producción en

departamentos clave como Casanare, La Guajira o Córdoba, o para estimar el potencial real de nuevas áreas en desarrollo, dificulta la planificación energética. Las decisiones se toman con un grado de incertidumbre mayor al deseable, lo que puede conducir a una asignación ineficiente de recursos, retrasos en proyectos de infraestructura o una subestimación del déficit de suministro futuro, con graves consecuencias para la industria, las generadoras térmicas y los consumidores residenciales.

Objetivos

Objetivo General

Aplicar modelos estadísticos y de machine learning para el pronóstico de la producción de gas natural en Colombia para el periodo 2013 – 2025.

Objetivos Específicos

Implementar algoritmos de predicción estadística tradicional como ARIMA y SARIMA para encontrar escenarios de producción de gas en Colombia.

Implementar algoritmos de machine learning para encontrar escenarios de producción de gas en Colombia.

Comparar entre los modelos de predicción estadística y de machine learning para saber cuál de ellos ofrece la mejor precisión en las métricas de evaluación.

Marco de Referencia

Marco Contextual

En este marco contextual se establece el fundamento metodológico sobre el cual se desarrolla el presente trabajo de aplicación. Se abordan las teorías y técnicas relacionadas con el análisis de series temporales, desde los modelos estadísticos tradicionales hasta los enfoques computacionales avanzados basados en machine learning, como máquinas de soporte vectorial, ridge regression, XGBoost y random forest. En particular estos dos últimos algoritmos de ensamble permiten manejar la complejidad de las interacciones no lineales inherentes a series temporales. (Géron, 2019). Al integrar diferentes métodos, el trabajo realizado no solo se limita a proyecciones o gráficos estadísticos, sino que, se puede llegar a posicionar como una herramienta de soporte en la toma de decisiones informadas para la planeación del sector del gas en Colombia. Este documento establece las bases técnicas para el uso de diversos modelos que pueden dar luces sobre el impacto económico en la producción y por ende en las tarifas que se cobran a la industria y al sector domiciliario. Dado que Colombia enfrenta un panorama de reservas crítico, el uso de herramientas avanzadas de modelamiento permite a tomadores de decisión anticiparse a momentos de déficit, optimizar recursos económicos y técnicos, planear infraestructura y estabilizar los precios que pueden llegar a afectar el mercado interno colombiano. Se exponen las herramientas que se emplearán para el pronóstico de la producción de gas natural en Colombia, basándose en la literatura científica que existe y se aplican las técnicas al caso colombiano.

Marco Teórico

Análisis de Series Temporales y Pronóstico de Producción

Una serie de tiempo es una secuencia de observaciones registradas en intervalos de tiempo sucesivos (Wang & Guo, 2020). En el sector energético, el análisis de estas series es fundamental para la planificación y la toma de decisiones. La producción de un recurso como el gas natural se ve influenciada, por factores geológicos, operativos y económicos. El objetivo principal de este análisis es modelar su comportamiento histórico para realizar pronósticos confiables sobre su evolución futura.

Modelos Estadísticos Tradicionales

Los modelos estadísticos han sido la herramienta principal para el pronóstico de producción en la industria del petróleo y el gas. Entre ellos, el modelo ARIMA (Promedio Móvil Integrado Autorregresivo) y SARIMA (Promedio Móvil Integrado Autorregresivo con Temporalidad) es uno de los más reconocidos. (Peña, 2010).

Modelo ARIMA y SARIMA

Un modelo ARIMA (AutoRegressive Integrated Moving Average) (p,d,q) intenta modelar la dependencia de una observación actual en sus propios valores anteriores, p términos AR, Auto Regresivos, con sus errores de pronósticos anteriores, q términos MA, Media Movil y se aplica una diferenciación d veces para que la serie sea estacionaria antes de aplicar los componentes AR y MA. El modelo ARIMA y sus variantes son eficaces para capturar las correlaciones temporales lineales en los datos. Se componen de un término Autorregresivo (AR) que refiere a que el valor actual de la serie temporal depende linealmente de sus propios valores anteriores, rezagados. Esencialmente, es una regresión del valor actual de la serie sobre sus valores pasados. Uno de Integración (I) para lograr la estacionariedad de la serie que refiere al

uso de la diferenciación para hacer que la serie temporal sea estacionaria. Una serie estacionaria es aquella cuyas propiedades estadísticas como la media y la varianza no cambian con el tiempo, lo cual es un requisito para los modelos AR y MA. La diferenciación implica restar el valor anterior al valor actual de la serie. Esto ayuda a eliminar tendencias y estacionalidad, y uno de promedio móvil (MA). Es una modelo muy eficaz para capturar patrones lineales y dependencias temporales en datos de series temporales, aunque puede requerir un análisis cuidadoso para determinar los parámetros óptimos (p, d, q). Para series temporales con estacionalidad clara, se usa una extensión llamada SARIMA (Seasonal AutoRegressive Integrated Moving Average) Su aplicación en la predicción de la producción de pozos de gas ha sido muy estudiada, (Duan et al., 2022) aplicaron un algoritmo de suavizado óptimo basado en ARIMA (ARIMA-RTS) para mejorar la precisión de la predicción de producción en pozos de gas, demostrando la aplicación que tiene estos modelos. (Peña, 2010; Wooldridge, 2010).

Enfoques de Machine Learning para Pronóstico Energético

Con el aumento de la capacidad computacional, los modelos de machine learning (ML) (Rahman Mahin et al., 2025) se han convertido en una alternativa poderosa, para determinar patrones complejos. Su aplicación en el sector energético es muy amplia y creciente. Uno de los métodos más usados en las técnicas de ML es Random Forest RF o arboles aleatorios cuyo poder se encuentra en la generación de variables predictoras y la definición de patrones no lineales entre las variables, de manera general se puede clasificar este método como de aprendizaje supervisado y a diferencia de los métodos estadísticos clásicos como ARIMA es que RF no modelan la autocorrelación directamente (Harrington, 2012). Es así como, se podrían incluir variables exógenas que se pueden considerar para los análisis:

- Factores climáticos: Temperatura, lluvias o eventos climáticos.
- Precios internacionales: Brent.
- Factores regulatorios: Fechas de concesiones y políticas internas.
- Demanda interna: Consumo eléctrico, industrial, etc.

Gradient Boosting Regressor

Algorimtmos como XGBoost, LightGBM y CatBoost son algoritmos de ensamble que construyen modelos secuencialmente, corrigiendo los errores de los modelos anteriores. Son muy potentes y versátiles para capturar relaciones no lineales y complejas en los datos(Theobald, 2017). Por ejemplo, gradient boosting regressor construye árboles de decisión de forma secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por los árboles anteriores. Es extremadamente potente y ha demostrado un gran éxito en una amplia variedad de problemas, incluyendo series temporales. (Géron, 2019).

Random Forest Regressor

Es un algoritmo de ensamble que construye múltiples árboles de decisión y promedia sus predicciones. Es robusto al sobreajuste y puede manejar características no lineales. Este algoritmo construye múltiples árboles de decisión de forma independiente y luego promedia sus predicciones lo que reduce el sobreajuste y mejora la estabilidad y precisión del modelo.(Géron, 2019).

Support Vector Regressor (SVR)

Una Máquina de Vectores de Soporte (SVM) es un modelo de aprendizaje automático muy potente y versátil, capaz de realizar clasificación lineal o no lineal, regresión e incluso detección de valores atípicos. Es uno de los modelos más populares en aprendizaje automático, y cualquier persona interesada en él debería tenerlo en su arsenal. Las SVM son especialmente

adecuadas para la clasificación de conjuntos de datos complejos, pero de tamaño pequeño o mediano. (Géron, 2019)

Ridge Regression

La regresión de ridge también es un modelo lineal de regresión, por lo que la fórmula que utiliza para realizar predicciones es la misma que se utiliza para los mínimos cuadrados ordinarios. Sin embargo, en la regresión de ridge, los coeficientes (w) se eligen no solo para que predigan bien con los datos de entrenamiento, sino también para que se ajusten a una restricción adicional. También se busca que la magnitud de los coeficientes sea lo más pequeña posible; en otras palabras, todas las entradas de w deben ser cercanas a cero. Intuitivamente, esto significa que cada característica debe tener el menor efecto posible en el resultado (lo que se traduce en una pendiente pequeña), sin dejar de predecir bien. Esta restricción es un ejemplo de lo que se denomina regularización. Regularizar significa restringir explícitamente un modelo para evitar el sobreajuste. El tipo particular utilizado por la regresión de cresta se conoce como regularización L2. (Müller & Guido, 2017).

Métodos de Validación

Para ARIMA y SARIMA, se emplean principalmente métodos de validación basados en split temporal simple (train-test) y validación cruzada en bloque (time series cross-validation), respetando el orden temporal (Brockwell & Davis, 2016). Para Random Forest en series temporales, aunque se puede usar validación cruzada en bloques temporal, es crucial evitar la aleatorización que rompa la dependencia temporal y se priorizan métricas como MAPE (Mean Absolute Percentage Error - Error Porcentual Absoluto Medio), RMSE (Root Mean Square Error - Raíz del Error Cuadrático Medio) y MAE (Mean Absolute Error - Error Absoluto Medio) en el conjunto de prueba o test definido al inicio del modelado.(Géron, 2019)

Para validar algoritmos de Machine Learning como Gradient Boosting Regressor (XGBoost, LightGBM, CatBoost), Random Forest y SVM, especialmente cuando se aplican a series temporales, es crucial utilizar métodos de validación que respeten la estructura de los datos.(Harrington, 2012)

División Entrenamiento-Prueba (Train-Test Split)

Este es el método más básico y lo que hace es dividir el conjunto de datos en dos conjuntos, uno para entrenar el modelo y otro para evaluarlo. Para series temporales es fundamental que la división sea temporal, es decir, entrenar con datos hasta un cierto punto en el tiempo y probar con los datos futuros. No se deben seleccionar datos de prueba al azar de todo el historial, ya que esto introduciría data leakage (fuga de información) y el modelo estaría viendo el futuro.(Kubat, 2017)

Validación Cruzada K-Fold (K-Fold Cross-Validation)

En este método los datos se dividen en K folds o subconjuntos de tamaño aproximadamente igual. El modelo se entrena K veces en cada iteración, se usa un fold diferente como conjunto de validación y los K-1 folds restantes como conjunto de entrenamiento. La métrica final es el promedio de las K evaluaciones.(Kubat, 2017)

Validación Cruzada para Series Temporales (Time Series Cross-Validation / Walk-Forward Validation)

Esta es la técnica de validación cruzada correcta para series temporales y funciona de la siguiente manera:

- Se define una ventana inicial de entrenamiento.
- El modelo se entrena con esta ventana.
- Se pronostica el siguiente período (o varios períodos) de la serie.

- Luego, la ventana de entrenamiento se avanza, incluyendo los datos más recientes (reales) en el conjunto de entrenamiento, y el proceso se repite.

Algunas consideraciones clave para algoritmos como gradient boosting, random forest y svm, es que una vez se ha transformado la serie temporal en un problema de regresión supervisada mediante el feature engineering creación de lags, medias móviles, etc., estos algoritmos se comportan como cualquier otro modelo de regresión. (McKinney, 2022)

Una vez validados los modelos de regresión, el siguiente paso es evaluar su rendimiento a través de métricas comunes como el error medio absoluto (MAE - Mean Absolute Error), la raíz del error medio cuadrático (RMSE - Root Mean Squared Error), el error porcentual absoluto medio (MAPE - Mean Absolute Percentage Error y el coeficiente de determinación R^2 (Géron, 2019). De igual manera, además de estas métricas cuantitativas, la evaluación visual a través de gráficos de las predicciones contra valores reales es fundamental en series temporales, también se deben revisar los gráficos de residuos ya que permite revelar patrones no capturados por el modelo, como estacionalidad restante, heterocedasticidad o autocorrelación, lo que indica que el modelo puede ser mejorado. Los intervalos de confianza son una buena estrategia ya que junto a las predicciones permite evaluar la incertidumbre en los pronósticos.

Metodología

Recolección de Datos

La serie temporal de datos hace parte del repositorio institucional de la agencia nacional de hidrocarburos ANH y de la página de datos abiertos del gobierno de Colombia. Sobre estas dos fuentes se selecciona la información más actualizada que va desde el mes de enero de 2013 hasta noviembre de 2025. El conjunto de datos incorpora información histórica asociada a la producción fiscalizada por año, mes, campo, contrato, empresa, departamentos, municipios, producción fiscalizada, gas lift, gas reinyectado, gas quemado, consumo en campo, gas enviado a planta, gas transformado y entregado a gaseoductos.

Método

Para el procesamiento de la información se usó un entorno basado en la nube a través de Google Colab en su versión free, aprovechando las capacidades que tiene para manejo de grandes volúmenes de datos y su interoperabilidad con las principales librearas de estadística, gráficos y machine learning en un entorno de desarrollo Python. Se emplearon las siguientes librerías principales:

- Manipulación y análisis: Pandas y NumPy.
- Modelado Estadístico: Statsmodels (para la implementación de los modelos ARIMA y SARIMA).
- Machine Learning: Scikit-learn (para Ridge Regression, Random Forest y SVM) y la librería XGBoost para algoritmos de ensamble optimizados.
- Visualización: Matplotlib y Seaborn.

En cuanto al análisis y modelado se incluyeron análisis descriptivos y comparativos además de las predicciones finales. Como primer paso se hizo un análisis exploratorio de datos

con la generación de cuadros de estadísticas descriptivas, graficas de la serie temporal, funciones de autocorrelación, análisis de estacionalidad para entender la distribución de la producción fiscalizada y poder llevar una estrategia de modelamiento de la serie temporal de datos. Para la visualización de resultados se crearon gráficas de las series de tiempo junto con su respectiva línea de predicción e intervalos de confianza. Se construyeron tablas comparativas utilizando métricas de error como el MAE, MSE, RMSE, MAPE y R^2 para determinar qué algoritmo ofrece la mayor precisión en los escenarios futuros de producción.

Resultados

El conjunto de datos utilizado contiene información histórica asociada a la producción fiscalizada, organizada con una frecuencia temporal mensual. Cada registro corresponde a observaciones agregadas por mes y año, lo que permite el análisis de la evolución de la producción en el tiempo y la aplicación de técnicas de series temporales y modelos predictivos. Las variables principales del conjunto de datos son las siguientes:

Tabla 1

Listado y Tipo de Variables

Variable	Tipo
YEAR	Número
MES	Número
CAMPO	Texto
CONTRATO	Texto
EMPRESA	Texto
DEPARTAMENTO	Texto
MUNICIPIO	Texto
PROD_FISCALIZADA	Número
GAS_LIFT	Número
GAS_REINYECTADO	Número
GAS_QUEMADO	Número
CONSUMO_CAMPO	Número
ENVIADO_PLANTA	Número
GAS_TRANSFORMADO	Número
ENTREGADO_GASEODUCTOS	Número

Nota. Nombres de las variables y su tipo.

Tabla 2

Descripción de Variables

Variable	Tipo
YEAR	Año de la medición
MES	Año de la medición
CAMPO	Nombre del campo de exploración
CONTRATO	Nombre del contrato de exploración
EMPRESA	Nombre de la empresa
DEPARTAMENTO	Nombre departamento
MUNICIPIO	Nombre municipio
PROD_FISCALIZADA	Producción fiscalizada de gas en MPC
GAS_LIFT	Número
GAS_REINYECTADO	Gas reinyectado en campo
GAS_QUEMADO	Gas quemado en campo
CONSUMO_CAMPO	Consumo de gas en campo
ENVIADO_PLANTA	Gas enviado a la planta
GAS_TRANSFORMADO	Gas transformado
ENTREGADO_GASEODUCTOS	Gas entregado al gasoducto

Nota. Nombres de las variables y su descripción

La variable YEAR es de tipo numérico entero y representa el año calendario de la observación. La variable MES también es de tipo numérico entero y toma valores entre 1 y 12, indicando el mes del año. A partir de estas dos variables se construye una variable de fecha, generalmente denominada date, de tipo fecha, que identifica de forma única cada periodo mensual y facilita el ordenamiento temporal de la serie.

La variable PROD_FISCALIZADA es una variable numérica continua, medida en unidades de producción de millones de pies cúbicos (mpc), y constituye la variable objetivo del análisis. Esta variable cuantifica el volumen de producción fiscalizada observado en cada

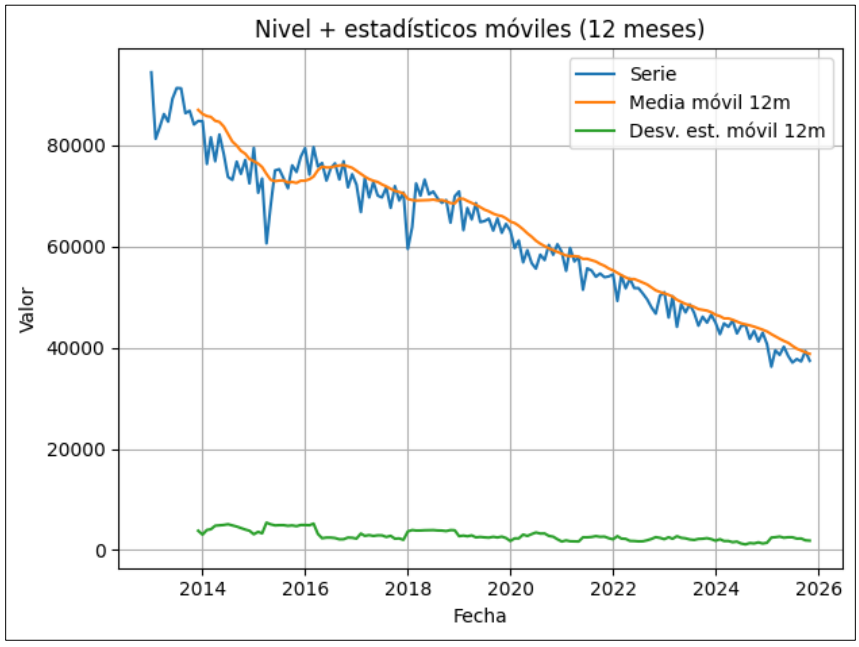
periodo. Para el análisis y modelación, los valores de PROD_FISCALIZADA se agregan a nivel mensual mediante la suma de los registros correspondientes del mismo año y mes teniendo millones de pies cúbicos por mes (mpcpm).

En conjunto, el dataset combina variables temporales discretas (año y mes) con una variable cuantitativa continua, lo que lo hace adecuado para análisis exploratorio de series temporales, con los métodos de modelación estadística ARIMA y SARIMA y también con enfoques de aprendizaje automático orientados al pronóstico como ridge regression, XGBoost, random forest regressor y support vector machine regressor.

Modelación SARIMA Y ARIMA

Figura 1

Serie Temporal Producción de Gas Fiscalizada



Este gráfico representa la serie temporal de la producción fiscalizada total mensual. Se puede observar varias características importantes:

- La línea azul representa la serie original PROD_FISCALIZADA.
- La línea naranja muestra la media móvil de 12 meses. Esta línea suaviza las fluctuaciones a corto plazo y ayuda a identificar la tendencia subyacente de la serie, eliminando el componente estacional. La media móvil claramente muestra una tendencia decreciente constante a lo largo del tiempo. Esto refuerza la idea de que la producción fiscalizada ha estado disminuyendo en promedio año tras año.
- La línea verde representa la desviación estándar móvil de 12 meses. Esta métrica indica cómo varía la dispersión o volatilidad de la serie a lo largo del tiempo. Se observa que la desviación estándar móvil también disminuye a medida que la media móvil y el nivel de la serie bajan. Esto es una fuerte indicación de heterocedasticidad, es decir, que la variabilidad de la serie no es constante, las fluctuaciones son mayores cuando los niveles de producción son altos y se reducen cuando los niveles son bajos.
- La serie muestra un patrón repetitivo y predecible que se repite anualmente y se manifiesta como picos y valles que ocurren aproximadamente en los mismos meses de cada año. Esta estacionalidad es bastante clara y es una característica muy particular de la serie.
- Varianza no constante (Heterocedasticidad): A medida que la serie disminuye su nivel, la magnitud de sus fluctuaciones, es decir, la variabilidad también se reduce. Esto indica que, cuando la producción es alta, hay mayores oscilaciones, y cuando es más baja, las oscilaciones son menores. Esto es una característica clara de heterocedasticidad, donde la varianza de la serie no es constante a lo largo del tiempo.
- Debido a la presencia de una tendencia clara y una varianza no constante, es evidente que la serie original no es estacionaria. Para modelos de series temporales, como ARIMA y SARIMA, la estacionariedad es un requisito fundamental, lo que implicaría la

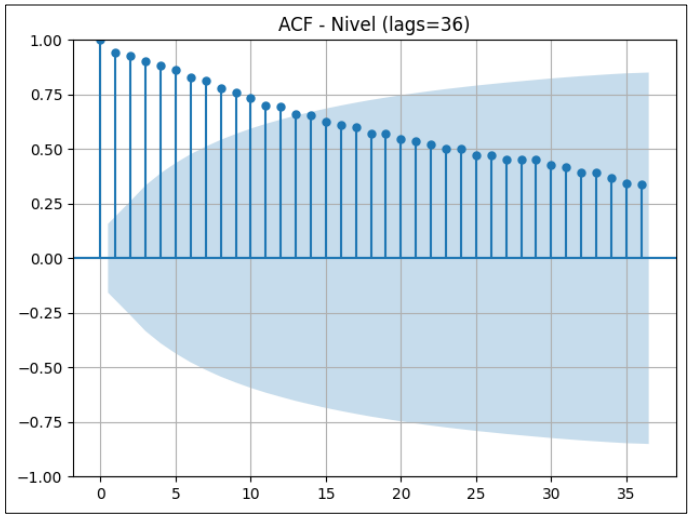
necesidad de aplicar transformaciones como diferenciación o transformación logarítmica para estabilizar la serie antes de modelarla.

Análisis Exploratorios de Datos

La Función de Autocorrelación (ACF) mide la correlación de una serie temporal con sus propios valores retardados (lags). Es una herramienta fundamental para identificar patrones, como la tendencia y la estacionalidad, además, para determinar el orden de los componentes de media móvil (MA) en modelos ARIMA y SARIMA.

Figura 2

Función de Autocorrelación (ACF)



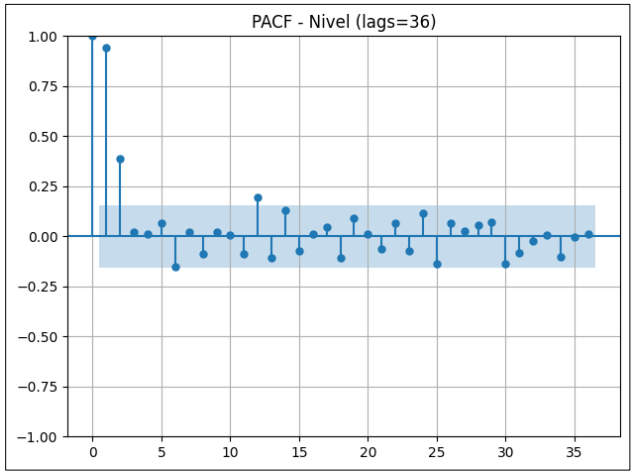
En el caso de la serie PROD_FISCALIZADA, se ha establecido que no es estacionaria y presenta una fuerte tendencia decreciente como estacionalidad. De igual manera, se observa un decaimiento lento y lineal, es decir con una tendencia clara. El patrón más prominente que se esperaría en la ACF de una serie con tendencia es un decaimiento muy lento y gradual de las autocorrelaciones. Esto significa que los valores de la serie actual están fuertemente correlacionados con valores pasados a lo largo de muchos lags, porque la tendencia hace que los

valores altos sigan a valores altos y los valores bajos sigan a valores bajos. El decaimiento no es rápido ni exponencial, sino que se mantiene significativo por un número considerable de retardos (lags).

La Función de Autocorrelación Parcial (PACF) mide la correlación entre una observación y una observación en un lapso anterior (lag), después de eliminar la influencia de las observaciones intermedias. Es fundamental para identificar el orden de los componentes autorregresivos (AR) en un modelo ARIMA y SARIMA. En el contexto de la serie PROD_FISCALIZADA, que ya se ha establecido que no es estacionaria y presenta una fuerte tendencia decreciente y estacionalidad se observa retardos (lags) significativos al inicio, Se observa en el gráfico de PACF un pico significativo en el lag 1, 2 y 3, indicando una fuerte correlación directa de una observación con la anterior, como es típico en series con tendencia.

Figura 3

Función de Autocorrelación Parcial (PACF)



Análisis de Estacionariedad

La prueba de Dickey-Fuller Aumentada (ADF - Augmented Dickey-Fuller) se utiliza para determinar si una serie temporal tiene una raíz unitaria, lo que indicaría que no es

estacionaria., En tal caso la hipótesis nula (H0) seria; la serie temporal tiene una raíz unitaria, es decir, no es estacionaria. La hipótesis alternativa (H1) es; la serie temporal no tiene una raíz unitaria, es decir, es estacionaria.

La prueba de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) se utiliza para probar la estacionariedad de una serie temporal en torno a una tendencia determinística. La hipótesis nula (H0) es que la serie temporal es estacionaria o estacionaria en torno a una tendencia determinística y la hipótesis alternativa (H1) seria cuando la serie temporal no es estacionaria, es decir, tiene una raíz unitaria. Para el caso de la serie temporal de producción fiscalizada de gas se tiene que:

- Ambas pruebas ADF y KPSS en la seria original indicaron claramente que la serie PROD_FISCALIZADA en su estado original no es estacionaria.
- La aplicación de la transformación logarítmica por sí sola no fue suficiente para lograr la estacionariedad en ninguna de las pruebas.
- Al aplicar la primera diferencia a la serie, ambas pruebas ADF y KPSS confirmaron que la serie resultante es estacionaria. Esto indica que eliminar la tendencia lineal fue efectivo.
- La aplicación de la diferencia estacional con un periodo de 12 meses también resultó en una serie estacionaria según ambas pruebas. Esto demuestra la presencia de una fuerte componente estacional que fue exitosamente eliminada.
- Aunque la prueba ADF indicó estacionariedad, la prueba KPSS mostró una ligera inconsistencia con un p-valor < 0.05, sugiriendo que podría quedar alguna forma de no-estacionariedad, posiblemente relacionada con la varianza o una tendencia sutil que la primera diferencia no eliminó por completo.

- Log + Diferencia Estacional (Log + Dif estacional (12)): Esta combinación también logró la estacionariedad según ambas pruebas, lo que la convierte en una opción robusta junto con la Dif(1) y Dif estacional (12) individuales.
- Para trabajar con la serie PROD_FISCALIZADA de manera estacionaria, las transformaciones más efectivas y consistentes fueron la primera diferencia, la diferencia estacional y la combinación de la transformación logarítmica con la diferencia estacional. Estas transformaciones lograron eliminar las tendencias y la estacionalidad, que son requisitos fundamentales para modelos de series temporales como ARIMA y SARIMA.

Tabla 3

Análisis de Estacionariedad

Test	Statistic	p-value	Critical Values (1%)	Critical Values (5%)	Critical Values (10%)	Stationary (p<0.05)	Stationary (p>0.05)
ADF - Nivel	-0.680	0.852	-3.477	-2.882	-2.578	No	N/A
KPSS - Nivel	1.780	0.010	0.739	0.463	0.347	N/A	No
ADF - Log(nivel)	0.808	0.992	-3.477	-2.882	-2.578	No	N/A
KPSS - Log(nivel)	1.756	0.010	0.739	0.463	0.347	N/A	No
ADF – Dif(1)	-4.975	0.000	-3.477	-2.882	-2.578	Yes	N/A
KPSS - Dif(1)	0.299	0.100	0.739	0.463	0.347	N/A	Yes
ADF - Log + Dif(1)	-5.128	0.000	-3.477	-2.882	-2.578	Yes	N/A
KPSS - Log + Dif(1)	0.500	0.042	0.739	0.463	0.347	N/A	No

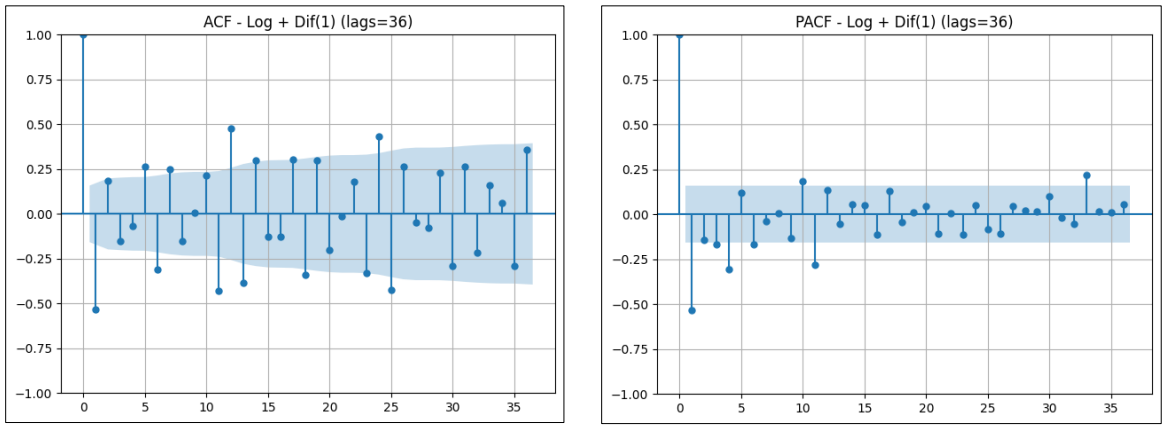
Test	Statistic	p-value	Critical Values (1%)	Critical Values (5%)	Critical Values (10%)	Stationary (p<0.05)	Stationary (p>0.05)
ADF - Dif estacional (12)	-3.957	0.002	-3.482	-2.884	-2.579	Yes	N/A
KPSS – Dif estacional (12)	0.087	0.100	0.739	0.463	0.347	N/A	Yes
ADF - Log + Dif estacional (12)	-3.019	0.033	-3.482	-2.884	-2.579	Yes	N/A
KPSS - Log + Dif estacional (12)	0.324	0.100	0.739	0.463	0.347	N/A	Yes

Nota. Diagnóstico ADF y KPSS

Un diagnóstico aún más preciso se evalúa con la correlación entre la media móvil y la desviación estándar móvil. Para este caso el resultado para la serie original es de ***0.680***. Este es un valor positivo y relativamente alto, lo que indica una fuerte presencia de heterocedasticidad. Después de aplicar la transformación logarítmica la correlación entre la media móvil y la desviación estándar móvil se redujo drásticamente a aproximadamente ***0.122***. Este valor mucho más bajo y cercano a cero, lo que indica que la transformación logarítmica fue muy efectiva en estabilizar la varianza de la serie, reduciendo significativamente la heterocedasticidad. Esto es importante porque los modelos de series temporales ARIMA y SARIMA asumen una varianza constante, es decir, homocedasticidad en los residuales para que sus estimaciones sean válidas. De esta manera y en base a la evidencia estadística anterior se selecciona el conjunto de datos con la trasformación logarítmica y con la primera diferencia. Las gráficas ACF y PACF con los datos trasformados muestran:

Figura 4

Graficas ACF y PACF con Datos Trasformados



En la gráfica ACF y PACF se observa un decaimiento rápido de las autocorrelaciones a cero y permanecer dentro de las bandas de confianza. Un decaimiento rápido es una fuerte señal de que la serie es ahora estacionaria, ya que la tendencia de largo plazo ha sido eliminada por la diferenciación. Basado en la interpretación de los gráficos ACF y PACF de la serie log-transformada y con diferenciación estacional, los órdenes para el modelo SARIMA son:

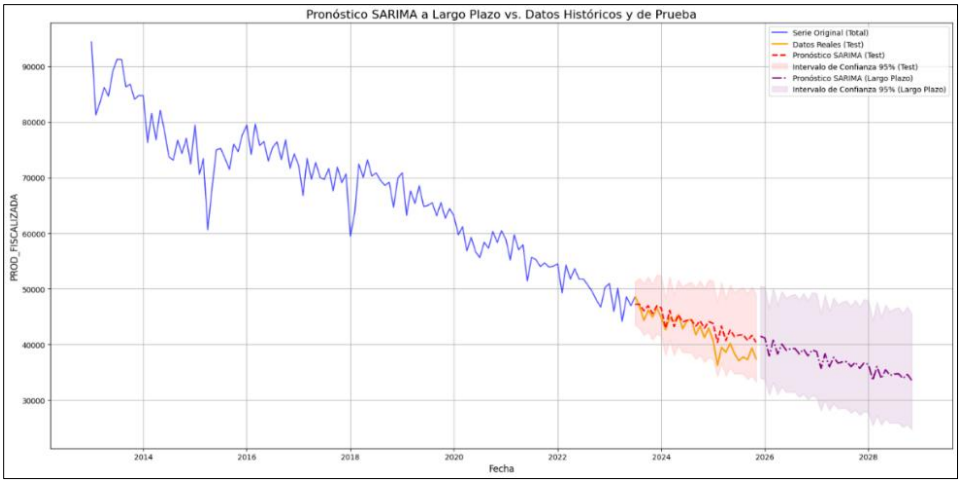
- Componentes no estacionales (p, d, q):
  - p = 1 Pico significativo en PACF lag 1, luego se corta
  - d = 1 La serie ya ha sido diferenciada estacionalmente, y parece estacionaria en media sin diferenciación adicional no estacional
  - q = 2 Picos significativos en ACF lag 1 y 2, luego decaimiento
- Componentes estacionales (P, D, Q, s):
  - P = 1 Hay picos significativos en el PACF en lags estacionales
  - D = 1 Ya se aplicó una diferencia estacional de orden 1
  - Q = 1 Hay picos significativos en el ACF en lags estacionales

- s = 12 Los picos claros en los múltiplos de 12 (12, 24) confirman que los datos tienen un ciclo anual, es decir, estacionalidad anual

- Por lo tanto, el modelo SARIMA propuesto es (1, 1, 2) (1, 1, 1, 12).

Figura 5

Diagnóstico del Modelo



Se observa que en el periodo final de datos de 2023 a 2025 la línea roja punteada que representa el pronóstico con el modelo SARIMA sigue muy cerca la línea naranja, es decir, los datos reales, esto indica que el modelo es capaz de capturar la dinámica de la serie temporal y hacer predicciones futuras confiables.

Los órdenes para el modelo ARIMA son:

- Componentes no estacionales (p, d, q):

  - p = 1 Pico significativo en PACF lag 1, luego se corta

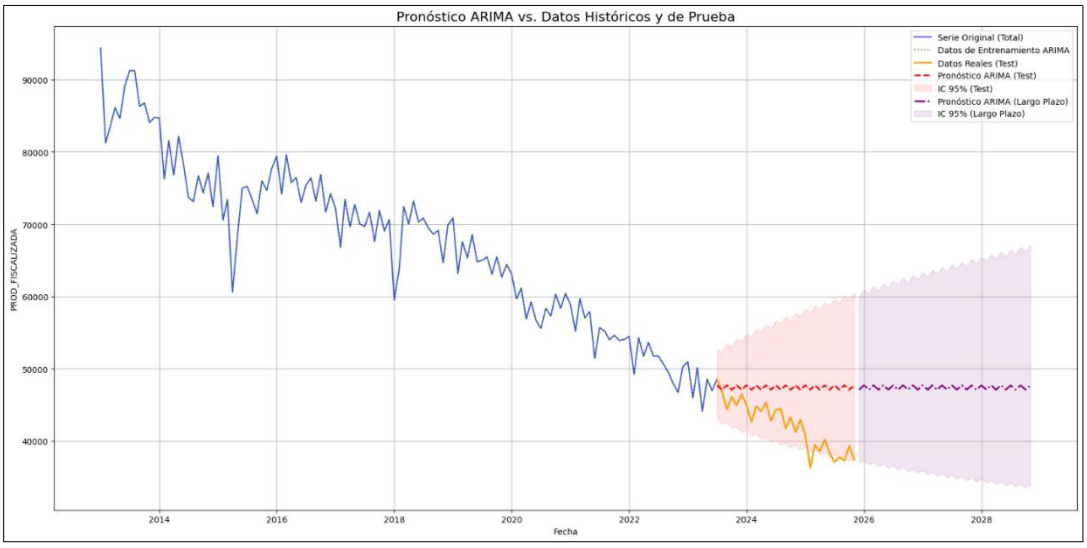
  - d = 1 La serie ya ha sido diferenciada estacionalmente, y parece estacionaria en media sin diferenciación adicional no estacional

  - q = 2 Picos significativos en ACF lag 1 y 2, luego decaimiento

- Por lo tanto, el modelo ARIMA propuesto es (1, 1, 2).

Figura 6

Pronóstico ARIMA



El modelo ARIMA no presenta un ajuste satisfactorio ya que la línea roja punteada se desvía completamente del conjunto de datos reales y esto sucede porque este modelo no logra captar los componentes estacionales o los ciclos de 12 meses, de esta manera, sus predicciones se desvían de la realidad de la serie temporal.

Tabla 4

Métricas ARIMA y SARIMA

Modelo	MAE	MSE	RMSE	MAPE	R2	Durbin - Watson
SARIMA	1788.86	4,965,915.71	2228.43	4.49%	0.56	1.22
ARIMA	5319.61	39,084,952.16	6251.80	13.33%	-2.45	1.02

Nota. Métricas de evaluación de los modelos ARIMA Y SARIMA

El modelo SARIMA es el más adecuado basándose en las métricas de evaluación iniciales, ya que presentó errores consistentemente más bajos en todas las métricas (MAE, MSE, RMSE, MAPE, R^2 Y Durbin - Watson) en comparación con el modelo ARIMA.

Los gráficos ACF de los residuales muestran que la mayoría de los picos caen dentro de las bandas de confianza, lo que es una buena señal de que la autocorrelación se ha reducido. Sin embargo, todavía hay algunos picos que sugieren que el modelo no ha capturado completamente algunos eventos o volatilidades extremas. Se observa un pico en el lag 24 en el ACF que podría indicar estacionalidad residual. El MAPE para SARIMA es de solo 4.49%, lo cual indica un excelente pronóstico de producción. En cambio, para el modelo ARIMA este indicador es de 13.33% lo que indica una perdida significativa en la precisión. Los estadísticos de Durbin – Watson mide la autocorrelación de los errores y teóricamente debe estar sobre 2. Para el caso modelo SARIMA 1.22 y ARIMA 1.02, se observa que ambos valores están por debajo de 1.5, lo que indica que todavía existe cierta autocorrelación positiva en los residuos, sin embargo, el valor de SARIMA es mejor, lo que indica que es más preciso.

Párrafo de terminación del modelo estadístico

Modelación con Ridge Regression

Para este caso se procede de manera similar al modelo estadístico SARIMA donde primero se construye la serie temporal agrupando por mes y como variable de trabajo es PROD_FISCALIZADA. Se hace la transformación logarítmica a la variable para estabilizar la varianza, normalizar la distribución y reducir el impacto de posibles outliers, es decir, se logra la homocedasticidad, ya que es un requisito fundamental en la modelación con *ridge regression.* También se trabaja con la primera diferencia como en el caso de SARIMA para tener una serie estacionaria y lograr consistencia en los análisis. Luego de seguir estos primeros pasos se hace el

feature engineering, el cual se compone por los rezagos o lags del mes 1 a 12 para capturar patrones estacionales. Luego se calcularon las medias móviles en los meses 3, 6, 12 para lograr capturar la tendencia local. Se calcularon las diferencias del mes 1 al 12 en la escala logarítmica para capturar el crecimiento porcentual de la variable, lo cual es muy útil en series con tendencia. Para las características cíclicas se calcularon el seno y el coseno del mes 1 al 12 con el objetivo de capturar la estacionalidad anual. Por último, se hace el escalamiento de la variable para que tenga una media de 0 y una desviación estándar de 1 lo que asegura que la penalización sea igual para todas las características, permitiendo que el modelo aprenda cuáles son realmente importantes para predecir la producción de gas. Se procede a separar los datos en 20% para test y 80% para train de la misma manera que con el modelo SARIMA Y ARIMA. En cuanto al resultado de la predicción se encontró el siguiente gráfico y métricas de desempeño del modelo.

Figura 7

Pronóstico Ridge Regression

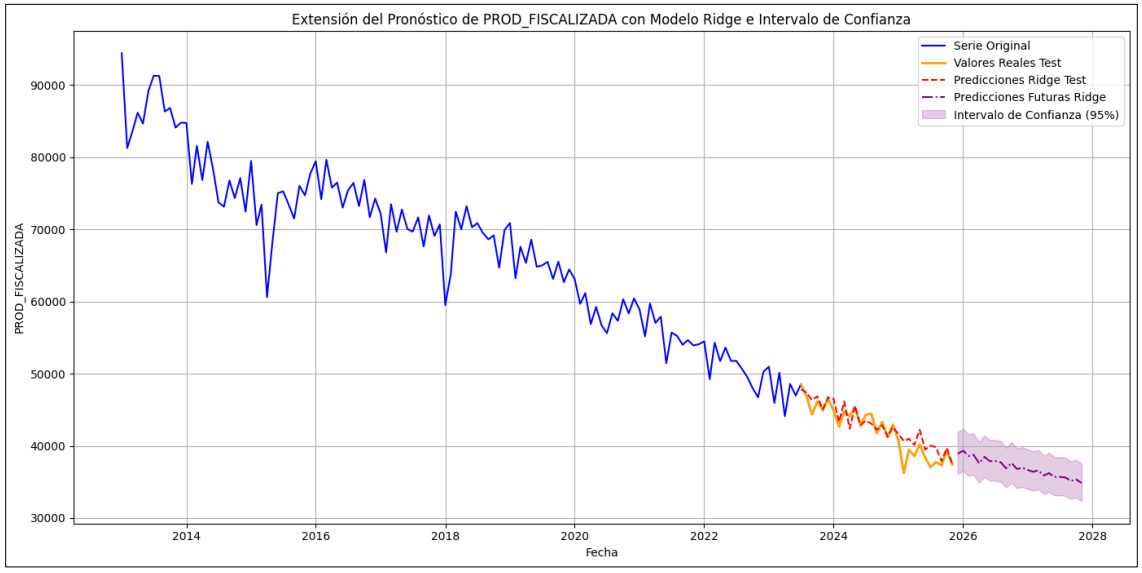


Tabla 5

Métricas Modelo Ridge

Modelo	MAE	MSE	RMSE	MAPE	R2
Ridge Regression	1073.22	2083757.81	1443.52	2.65%	0.82

Nota. Métricas de evaluación del modelo Ridge

La línea punteada roja muestra cómo se comportó el modelo sobre el conjunto de datos de prueba. El modelo rigde logra capturar la tendencia general y la magnitud de la producción, así como, hacer una proyección de la producción de gas para los próximos 24 meses y se observa una estabilidad con una tendencia leventemente decreciente y coherente con el comportamiento histórico de la serie temporal

Support Vector Machines SVM

Se hace el mismo análisis inicial descrito en el titulo anterior para lograr una serie estacionaria. Los pasos fueron:

1. Indexación Temporal: Creación de la columna fecha para asegurar el orden temporal.
2. Agregación de la producción nacional total por mes.
3. Log-Transform: Uso de trasformación logarítmica para reducir el impacto de la heterocedasticidad.
4. Diferenciación: Aplicación para eliminar la tendencia.
5. Variables de Retraso (Lags): Generación de 3 características autorregresivas.
6. Escalamiento: Ajuste de las características a una escala común mediante.
7. Se hace una partición de 80% para train y 20% para test.

8. Se hace optimización con el algoritmo Grid Search Cross-Validation para buscar la mejor combinación de hiperparámetros del modelo de machine learning.

Tabla 6

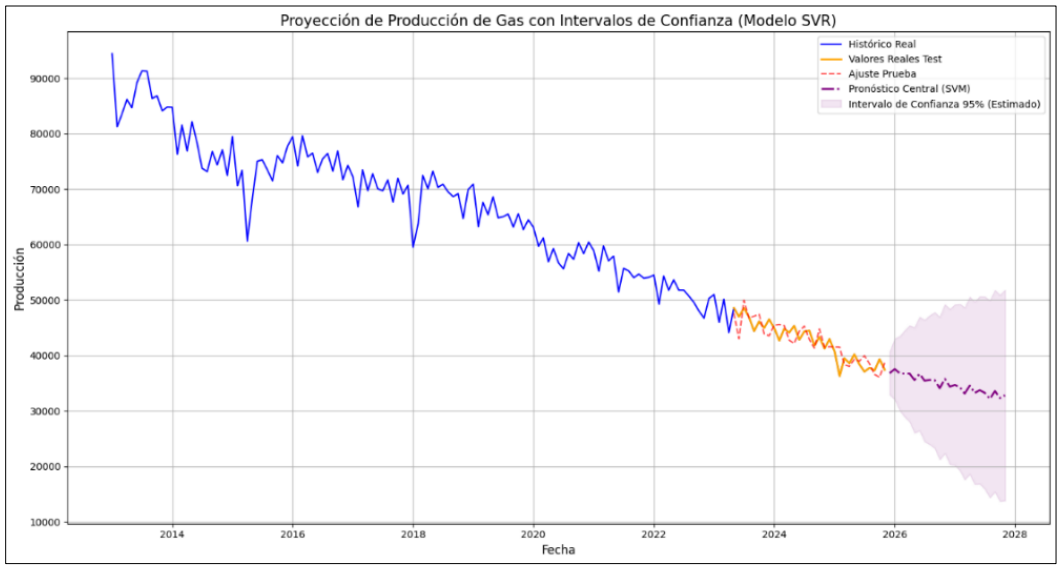
Métricas de Modelo SVM

Modelo	MAE	MSE	RMSE	MAPE	R2
Support Vector Machines	1574.25	3913569.1	1978.27	3.75%	0.69

Nota. Métricas de evaluación del modelo SVM

Figura 8

Pronóstico SVM



El modelo SVR alcanza un R^2 de 0.69, lo que significa que el modelo logra explicar aproximadamente el 69% de la variabilidad de la producción de gas original. El valor de MAE significa que, en promedio, el modelo se equivoca en 1574 unidades de producción. El RMSE es un poco más grande que el MAE indica que hay valores atípicos o outliers que el modelo puede

estar ignorando y que son susceptibles de ser mejorados. La métrica más importante es el MAPE y significa que en promedio el error en predicciones solo es de 3.75% con respecto al valor real, lo cual es muy bueno.

Random Forest

Se sigue la misma lógica inicial que en los dos modelos anteriores para asegurar una serie estacionaria y se hace una partición de 80% para train y 20% para test

Tabla 7

Métricas de Modelo Random Forest

Modelo	MAE	MSE	RMSE	MAPE	R2
Random Forest	1352.1	2907574.28	1705.16	3.24%	0.77

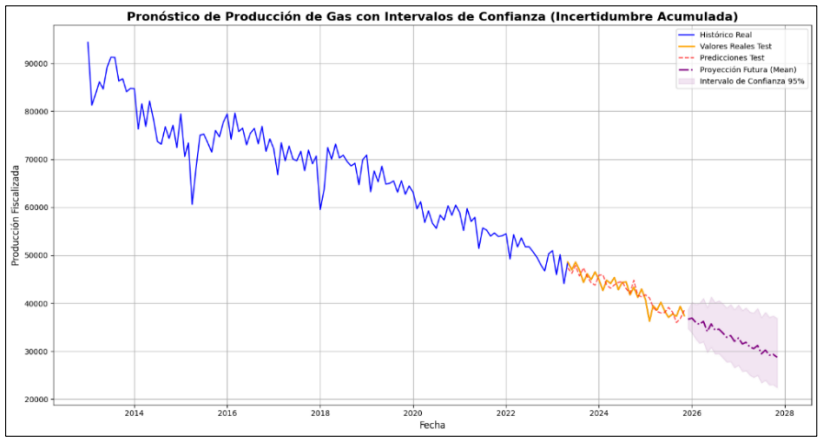
Nota. Métricas de evaluación del modelo random forest

Para este caso el R^2 de 0.77, indica que logra explicar el 77% de la variabilidad de la producción de gas, lo cual es un resultado robusto para la serie temporal. El MAPE es bajo, lo que indica que las predicciones se podrían desviar en un 3.2% del valor real. El RMSE es de aproximadamente 1700 unidades.

Para el grafico de la serie temporal se observa que el modelo proyecta una continuidad en la dinámica actual de la serie, capturando la estabilidad o ligeras correcciones en la producción fiscalizada de gas.

Figura 9

Pronóstico Random Forest

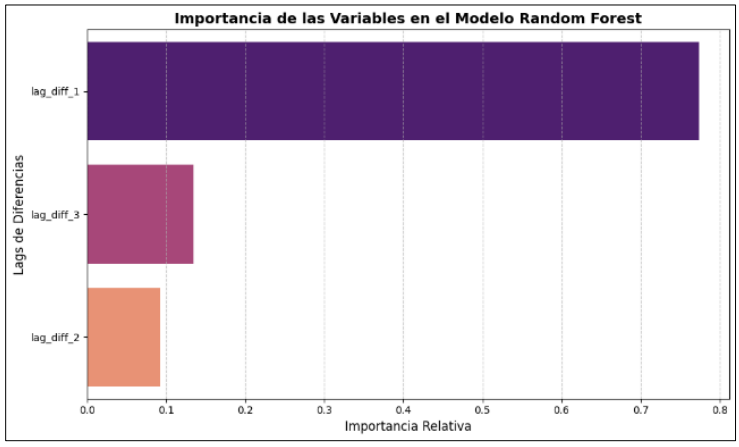


La Importancia de variables refleja:

El primer rezago, lag_diff_1, aporta el 77.4% de la importancia. Esto confirma que la producción de gas tiene una fuerte memoria en el valor del último mes y es el mejor predictor para la estimación del mes siguiente. En cuanto a los rezagos 2 y 3, se observa un impacto significativamente menor, lo que indica que la producción no se ve afectada por valores o tendencias antiguas.

Figura 10

Importancia de Variables

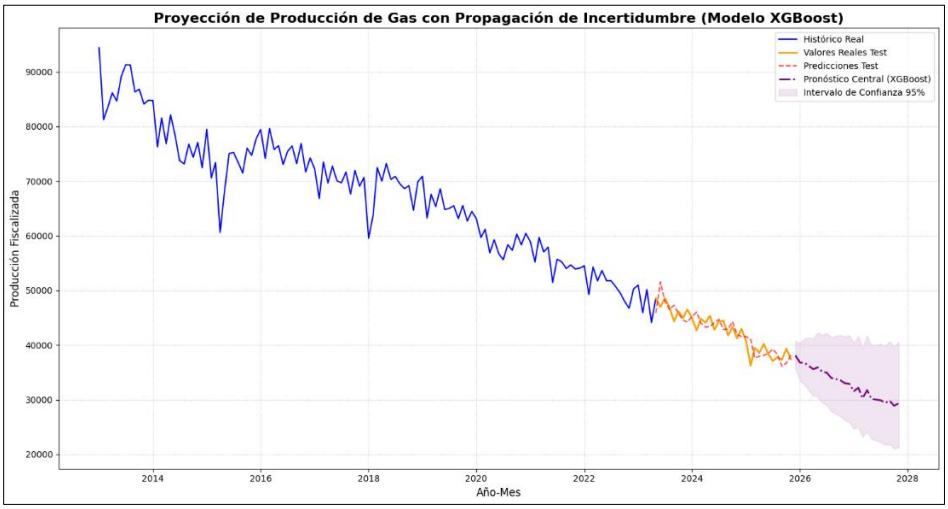


Para el grafico de la serie temporal se observa que el modelo proyecta una continuidad en la dinámica actual de la serie, capturando la estabilidad o ligeras correcciones en la producción fiscalizada de gas.

XGBoost

Figura 11

Pronóstico XGBoost



MAPE indica que en promedio las predicciones del modelo fallan en un 3.56% respecto del valor real. Para este caso el R^2 de 0.71, indica que logra explicar el 77% de la variabilidad de la producción de gas, lo cual es un resultado robusto para la serie temporal.

Tabla 8

Métricas de Modelo XGBoost

Modelo	MAE	MSE	RMSE	MAPE	R2
XGBoost	1494.55	3686025.14	1919.9	3.56%	0.71

Nota. Métricas de evaluación del modelo XGBoost

El análisis comparativo de los modelos de regresión y series temporales presentados en este capitulo permite observar las diferencias de rendimiento entre los métodos de machine learning y los puramente estadísticos, mientras que el modelo Ridge Regression se consolidó con un valor R^2 de 0.82, el más alto, y un MAPE de apenas 2.65%, el modelo ARIMA mostró una incapacidad estructural para modelar la serie, resultando en un coeficiente de determinación negativo -2.45 y un error porcentual del 13.33%. El modelo SARIMA, aunque logró una corrección significativa al integrar el componente estacional reduciendo el MAPE a 4.49% y elevando el R^2 a 0.56, no alcanzó el mismo nivel de confiabilidad de la regresión Ridge, lo que sugiere que la clave del pronóstico está en la inclusión de variables externas y no solo en la dependencia del pasado histórico de la serie temporal. Es de resaltar que, el análisis de los estadísticos de Durbin-Watson en los modelos autorregresivos, con valores de 1.02 y 1.22, indica la persistencia de una autocorrelación residual. En contraste, la estabilidad del RMSE 1443.52 en el modelo Ridge confirma su robustez ante valores atípicos, consolidándolo como el modelo más confiable. Los resultados expuestos validan la transición hacia enfoques de aprendizaje automático, demostrando la precisión que se gana con el modelo Ridge Regression proporciona una base sólida para las conclusiones generales y las futuras implementaciones para tareas de pronóstico en la producción de gas en colombia.

Análisis Comparativo de Modelos de Pronóstico

El desempeño de los modelos evaluados muestra una diferencia entre los enfoques estadísticos tradicionales y los algoritmos de aprendizaje supervisado de machine learning. A continuación, se detallan los hallazgos principales basados en las métricas de error y precisión.

Tabla 9

Análisis Comparativo de Métricas

Modelo	MAE	MSE	RMSE	MAPE	R^2	Durbin - Watson
SARIMA	1788.86	4,965,915.71	2228.43	4.49%	0.56	1.22
ARIMA	5319.61	39,084,952.16	6251.80	13.33%	-2.45	1.02
Ridge Regression	1073.22	2083757.81	1443.52	2.65%	0.82	
Support Vector machines	1574.25	3913569.1	1978.27	3.75%	0.69	
Random Forest	1352.1	2907574.28	1705.16	3.24%	0.77	
XGBoost	1494.55	3686025.14	1919.9	3.56%	0.71	

Nota. Análisis comparativo de las métricas de evaluación

- El modelo de Ridge Regression resultó ser el más robusto para este conjunto de datos. Presenta los errores más bajos.
- MAE 1073.22 y MAPE 2.65% en el modelo ridge indican una desviación mínima respecto a los valores reales. Un MAPE por debajo del 5% se considera como un pronóstico excelente.
- En el modelo ridge el R2 de 0.82, es el coeficiente de determinación más alto, lo que significa que el modelo explica el 82% de la variabilidad de los datos, superando a los modelos de Random Forest y XGBoost.
- El valor de 2.65% en el MAPE para el modelo Ridge es bajo y esto indica que, en promedio, las predicciones del modelo solo se desvían un 2.65% del valor real.
- Todos estos modelos se encuentran en un rango muy bueno, sin embargo, la diferencia entre Ridge Regression y SARIMA representa una ganancia significativa en contextos donde resulta importante minimizar el error de predicción frente a datos reales. Un ejemplo de ellas es cuando se usan los resultados en contextos financieros o para optimizar recursos

económicos, técnicos, planear infraestructura y estabilizar los precios de mercados. En este tipo de escenarios es muy importante trabajar con el mínimo error.

- Aunque los modelos de ensamble como Random Forest con un R2 0.77 y XGBoost R2 0.71 muestran resultados competitivos, no logran superar a la regresión Ridge.
- Es notable el desempeño deficiente de ARIMA, con un R2 de -2.45. Un coeficiente de determinación negativo indica que el modelo es incluso peor que predecir simplemente usando el promedio de los datos históricos. Esto sugiere que la serie temporal presenta una estacionalidad fuerte que el modelo ARIMA simple no pudo capturar.
- En el caso del modelo SARIMA con un R2 de 0.56, se observa una mejora sustancial, lo que confirma que el componente estacional es muy importante para tener en cuenta en este fenómeno.
- Para los modelos ARIMA y SARIMA, los valores de 1.02 y 1.22 en el estadístico de Durbin - Watson sugieren la presencia de autocorrelación positiva en los residuos ya que están lejos del valor ideal de 2.
- Tras el análisis comparativo, se determina que el modelo de Ridge Regression es la técnica óptima para el fenómeno estudiado.
- Al evaluar la precisión con el MAPE, se observa que el modelo Ridge Regression ofrece el desempeño más alto, con un error de apenas 2.65%. Este resultado es significativamente superior al obtenido por el modelo tradicional ARIMA de 13.33%, lo que representa una reducción del error en más de un 80%.
- Un MAPE inferior al 5% se considera como un pronóstico de alta precisión, los resultados validan la idoneidad del modelo seleccionado para los pronósticos de la producción de gas.

Conclusiones

El conjunto de datos descargado de la página de datos abiertos y de la agencia nacional de hidrocarburos ANH, presentaba un buen estado en cuanto a los tipos de datos y estructura general, es decir, que no fue necesario pasar por un proceso exhaustivo de limpieza o formateo de la información.

La serie temporal en su estado original presenta una tendencia clara, el grafico de media móvil confirma una tendencia decreciente significativa en la producción a largo del periodo de estudio.

Para trabajar con la serie PROD_FISCALIZADA de manera estacionaria, las transformaciones más efectivas fueron la primera diferencia, la diferencia estacional y la combinación de la transformación logarítmica con la diferencia estacional. Estas transformaciones lograron eliminar las tendencias y la estacionalidad, lo cual es un requisito fundamental para modelar series temporales como ARIMA y SARIMA.

Analizando los diagnósticos, el modelo SARIMA (1, 1, 2) (1, 1, 1, 12) se logró eliminar la estructura de dependencia y la tendencia. Sin embargo, las desviaciones de la normalidad en el histograma, Q-Q plot y posibles patrones en los residuales estandarizados y en el correlograma pueden indicar áreas de mejora.

Se evidencia una clara insuficiencia en los modelos de series temporales clásicos, en particular ARIMA, cuyo R^2 negativo de -2.45 y elevado MAPE 13.33% lo descartan como herramienta confiable el modelamiento de escenarios futuros o de proyección con un componente estacional marcado.

Si bien la inclusión del componente estacional en el modelo SARIMA logró corregir gran parte del sesgo, reduciendo el error a un 4.49%, no alcanzó la precisión de los algoritmos de aprendizaje supervisado.

La producción de gas en el conjunto de datos se ve fuertemente ligada a ciclos anuales. No considerar al menos 12 meses de historia (lags) invalida cualquier intento de pronóstico.

Aunque el pronóstico central es preciso a corto plazo, los intervalos de confianza muestran que hacia los 24 meses la incertidumbre crece. Esto sugiere que los modelos deben ser reentrenados mensualmente con nuevos datos reales para mantener la precisión.

Los resultados demuestran que el modelo de Ridge Regression tiene la mayor capacidad predictiva, alcanzando un coeficiente de determinación R^2 de 0.82 y el error porcentual más bajo MAPE de 2.65%. Este desempeño está por encima incluso de modelos de ensamble complejos como Random Forest y XGBoost.

Los modelos de machine learning y aun SARIMA son altamente confiables para la planificación operativa de corto y mediano plazo, es decir, de 1 a 6 meses, y sirve como una buena base de referencia para decisiones estratégicas a largo plazo, siempre considerando los límites superiores e inferiores de confianza calculados.

El modelo Ridge resulto ser un modelo confiable y robusto ya que, ofrece una visión realista de la producción futura, con márgenes de error inferiores al 3%, lo cual es sobresaliente para hacer predicciones.

Para el modelo SVMR el error de las predicciones es de solo un 3.75% respecto al valor real. Generalmente, un MAPE por debajo del 5% se considera un pronóstico de excelente precisión.

El modelo Random Forest proyecta continuidad en la serie, capturando la estabilidad o ligeras correcciones en la producción fiscalizada. Los intervalos de confianza al 95% muestran cómo el riesgo aumenta a medida que se aleja del presente. El sombreado en la gráfica permite a los tomadores de decisiones visualizar no solo el caso base, sino también los escenarios optimistas y pesimistas basados en la variabilidad estadística del modelo.

Recomendaciones

Para mejorar los resultados en la modelación de la producción de gas fiscalizado se sugiere integrar variables exógenas ya sean de nivel económico, climático, geológico y social que podrían mejorar de manera significativa el pronóstico final

Aunque el modelo SARIMA presenta un buen ajuste en sus parámetros, se recomienda simplificar aún más el modelo reduciendo los órdenes p, q, P y Q y luego añadir otros términos de forma iterativa hasta alcanzar un mejor resultado.

Es necesario fomentar la distribución de información gubernamental para poder desde espacios académicos dar soluciones técnicas y eficientes a problemas de la industria nacional. Trabajos de aplicación como este podría dar aportes al problema energético al que se enfrenta el país por la reducción de las reservas de gas e hidrocarburos que son esenciales para el buen estado de la economía nacional.

En Colombia, aunque no tenemos estaciones climáticas marcadas, el consumo de gas tiene ciclos claros ligados a la actividad industrial y la generación eléctrica, es decir, existe la posibilidad de que cuando las hidroeléctricas bajan por el fenómeno del niño, el gas sube, por lo tanto, se recomienda el uso de variables exógenas en los modelos de regresión.

A diferencia de ARIMA, que tiene una memoria corta, la IA robusta con las LSTM (Long Short-Term Memory) son ideales para el modelar series temporales. Estas redes neuronales pueden ser una buena solución para detectar patrones no lineales que los modelos estadísticos ignoran, como los picos repentinos de demanda de gas cuando hay sequías prolongadas.

Referencias Bibliográficas

AlShafeey, M., & Csáki, C. (2021). Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods. *Energy Reports*, *7*, 7601–7614. https://doi.org/10.1016/J.EGYR.2021.10.125

Amaechi, U. C., Ikpeka, P. M., Xianlin, M., & Ugwu, J. O. (2019). Application of machine learning models in predicting initial gas production rate from tight gas reservoirs. *Rudarsko Geolosko Naftni Zbornik*, *34*(3), 29–40. https://doi.org/10.17794/rgn.2019.3.4

Análisis de la inflación y el crecimiento económico un enfoque de Machine Learning. México 1990-2021. (n.d.). Retrieved May 28, 2025, from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202024000300174&lang=es

Brockwell, P., & Davis, R. (2016). *Introduction to Time Series and Forecasting* (3rd Edition). Springer International Publishing.

Duan, Y., Wang, H., Wei, M., Tan, L., & Yue, T. (2022). Application of ARIMA-RTS optimal smoothing algorithm in gas well production prediction. *Petroleum*, *8*(2), 270–277. https://doi.org/10.1016/J.PETLM.2021.09.001

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems* (N. Tache, Ed.; 2nd Edition.). O'Reilly Media, Inc.

Harrington, Peter. (2012). *Machine Learning in Action* (1st, Edition ed.). Manning Publications Co.

Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos. (n.d.-a). Retrieved January 23, 2026, from https://www.anh.gov.co/es/noticias/informe-anual-de-reservas-y-recursos-irr-2023/

Informe Anual de Reservas y Recursos IRR 2023 - Agencia Nacional de Hidrocarburos. (n.d.-b). Retrieved July 6, 2025, from https://www.anh.gov.co/es/noticias/informe-anual-de-reservas-y-recursos-irr-2023/

Informes de Gestión y Sostenibilidad. (n.d.). Retrieved July 6, 2025, from https://www.ecopetrol.com.co/wps/portal/Home/es/ResponsabilidadEtiqueta/InformesGestionSostenibilidad/Informesdegestion

Kubat, M. (2017). *An Introduction to Machine Learning* (2nd Edition.). Springer International Publishing.

McKinney, W. (2022). *Python for Data Analysis* (3rd Edition). O'Reilly Media, Inc.

Monroy, Salvador. (2008). *Estadística Descriptiva* (1ra Edición). Instituto Politécnico Nacional.

Müller, A. C., & Guido, S. G. (2017). *Introduction to Machine Learning with Python* (Dawn. Schanafelt, Ed.; 1st Edition). O'Reilly Media, Inc.

Peña, D. (2010). *Análisis de series temporales*. Alianza Editorial.

Producción Fiscalizada de Gas 2017 | Datos Abiertos Colombia. (n.d.). Retrieved June 2, 2025, from https://www.datos.gov.co/Minas-y-Energ-a/Producci-n-Fiscalizada-de-Gas-2017/bgru-y57z/about_data

Rahman Mahin, M. P., Shahriar, M., Das, R. R., Roy, A., & Reza, A. W. (2025). Enhancing Sustainable Supply Chain Forecasting Using Machine Learning for Sales Prediction. *Procedia Computer Science*, *252*, 470–479. https://doi.org/10.1016/J.PROCS.2025.01.006

Safiyari, M. H., Shavvalpour, S., & Tarighi, S. (2022). From traditional to modern methods: Comparing and introducing the most powerful model for forecasting the residential natural gas demand. *Energy Reports*, *8*, 14699–14715. https://doi.org/10.1016/J.EGYR.2022.10.397

Theobald, Oliver. (2017). *Machine Learning For Absolute Beginners* (2da Edición.).

Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*, *17*(3), 205–221. https://doi.org/10.23919/JCC.2020.03.017

Wooldridge, J. M. (2010). *Introducción a la econometría Un enfoque moderno* (4a. edición). Cengage Learning Edi to res, S.A. de C.V.