Diplomatura en Ciencia de Datos y Análisis Avanzado

Trabajo final integrador

“*Predicción del Costo Monómico Medio del Mercado Eléctrico Mayorista en Argentina: Un Enfoque de Series Temporales y Machine Learning*”

GRUPO J:

Angel Ayuso

Nicolas Ezequiel Divecchi

Gonzalo Garcia

Ignacio Lacabanne

Milena Mirabito

Martin Rodriguez

1. Resumen Ejecutivo

El presente trabajo tiene como objetivo principal proyectar el Costo Monómico Medio del Mercado Eléctrico Mayorista (MEM) de Argentina hacia el año 2026. El costo monómico constituye un indicador estratégico para el sector energético, dado que refleja el precio promedio ponderado de generación eléctrica y sirve como referencia para la planificación tarifaria, la gestión de compras y la cobertura financiera de las distribuidoras.
Se trabajó con una serie temporal mensual de enero de 2018 a julio de 2025**,** obtenida de fuentes públicas de CAMMESA. De la base completa se seleccionaron 14 predictores energéticos, económicos y climáticos, siguiendo las pautas de especialistas del sector que solicitaron este informe.
La evaluación de modelos incluyó enfoques clásicos de series de tiempo (ARIMA, SARIMA) y modelos de aprendizaje automático con variables rezagadas (Ridge, Support Vector Regressor, XGBoost y LightGBM). La elección de múltiples algoritmos respondió al interés de: (i) capturar la estacionalidad anual del costo monómico, (ii) interpretar la relevancia de predictores y (iii) alcanzar la máxima precisión posible en la proyección**.**
Los lags seleccionados fueron 1, 2, 10, 11 y 12, definidos a partir de los gráficos de autocorrelación y autocorrelación parcial. La validación se realizó con un train-test temporal (train hasta 2024, test en 2025), complementada con esquemas de *Expanding Window* y *Sliding Window.* En el caso de ARIMA y SARIMA, se trabajó directamente sobre la serie original y se compararon bajo los mismos criterios de éxito.
El KPI (*key performance indicator*) se definió de manera conjunta con el cliente como un conjunto de: (i) RMSE inferior al 10% respecto al valor promedio real, métrica seleccionada por ser directamente comprensible en las mismas unidades que la variable objetivo (USD/MWh) y (ii) Mejor desempeño frente a dos benchmarks de referencia: el modelo Naive y el modelo Naive estacional.
En cuanto a la interpretabilidad de los modelos, se adoptó un doble enfoque: (i) un análisis exploratorio clásico de correlaciones entre variables predictoras y la variable objetivo, reconociendo que este procedimiento rompe la naturaleza temporal de la serie, ya que evalúa relaciones sin considerar la secuencia cronológica, y (ii) la aplicación de feature importance y SHAP values, que permitieron interpretar de manera más rigurosa el aporte relativo de cada predictor en los modelos entrenados.

En conclusión, el proyecto permitió desarrollar un sistema de proyección confiable y validado bajo condiciones de negocio reales**,** con errores por debajo del umbral del 10% y

mejoras claras frente a los benchmarks, habilitando su aplicación en la planificación financiera y tarifaria de las distribuidoras**.**

2.Definición del Problema y Relevancia

El desafío central del proyecto consiste en estimar el Costo Monómico Medio del Mercado Eléctrico Mayorista (MEM) de Argentina hacia el año 2026, utilizando información histórica y un conjunto de variables energéticas, económicas y climáticas seleccionadas en función de la experiencia y lineamientos provistos por el cliente. El costo monómico medio consiste en la suma de los costos de generación energética (y asociados) y la demanda abastecida en el MEM. El objetivo concreto es contar con un modelo predictivo capaz de entregar proyecciones confiables del costo monómico medio, cuya utilidad práctica radica en reducir la incertidumbre en la planificación tarifaria, anticipar escenarios de costos frente a la volatilidad de los combustibles y la demanda, y aportar insumos para la gestión de riesgos financieros y regulatorios.

La relevancia estratégica se fundamenta en que el costo monómico medio constituye un insumo crítico para las decisiones de inversión y cobertura de las distribuidoras, así como para la sostenibilidad del sistema eléctrico en su conjunto. En este marco, el desarrollo de un modelo robusto y validado representa un valor tangible para el negocio, al mejorar la capacidad de proyección y planificación en un sector altamente sensible a variaciones macroeconómicas y regulatorias.

1. Datos y Calidad

La base de datos proviene de CAMMESA (descargada de la página https://cammesaweb.cammesa.com/variables-relevantes-mem/), con una serie temporal mensual entre enero de 2018 y julio de 2025 (91 observaciones), de la cual se seleccionaron 14 variables energéticas, económicas y climáticas siguiendo las pautas de especialistas del sector, finalizando en la base “MEM final.xlsx”.

Operacionalización de las variables y sus unidades:

Fecha: Fecha de la observación mensual (AAAA-MM-DD)

BarrilPetroleo: Precio internacional del barril de petróleo (USD/barril)

TempMedia: Temperatura media mensual (°C)

DemandaLocal+Exp: Demanda local de energía más las exportaciones de energía (GWh)

OfertaTotal: Oferta total de energía generada (GWh)

GasNatural: Precio del gas natural (USD/MBtu)

FuelOil: Precio del Fuel Oil (USD/ton). Abreviado como FO.

GasOil: Precio del Gas Oil (USD/m^3). Abreviado como GO.

CarbonMineral: Precio del carbón mineral (USD/ton). Abreviado como CM.

CombustTotal: Costo total de combustibles (millones USD)

CostoMonomico: Costo Monómico Medio. Precio medio en función del costo total de la generación de energía (USD/MWh). Variable *target*.

CombustAlt: Costo combinado de combustibles alternativos [Fuel Oil + Gas Oil + carbón mineral] (millones USD).

La revisión inicial incluyó control de valores faltantes, los cuales no se registraron, y detección de *outliers* mediante el método clásico de rango intercuartílico (Q1–1.5·IQR; Q3+1.5·IQR). Los *outliers* detectados fueron interpretados como variaciones estacionales esperables, debidas a la variación de otras variables correlacionadas (Ej: Costo monómico total y costo del barril del petróleo) en los meses comprendidos entre marzo y agosto, principalmente en invierno (Figura 1). Por esta razón no consideramos correspondiente eliminarlos ni reemplazarlos. No se requirió ingeniería de variables.

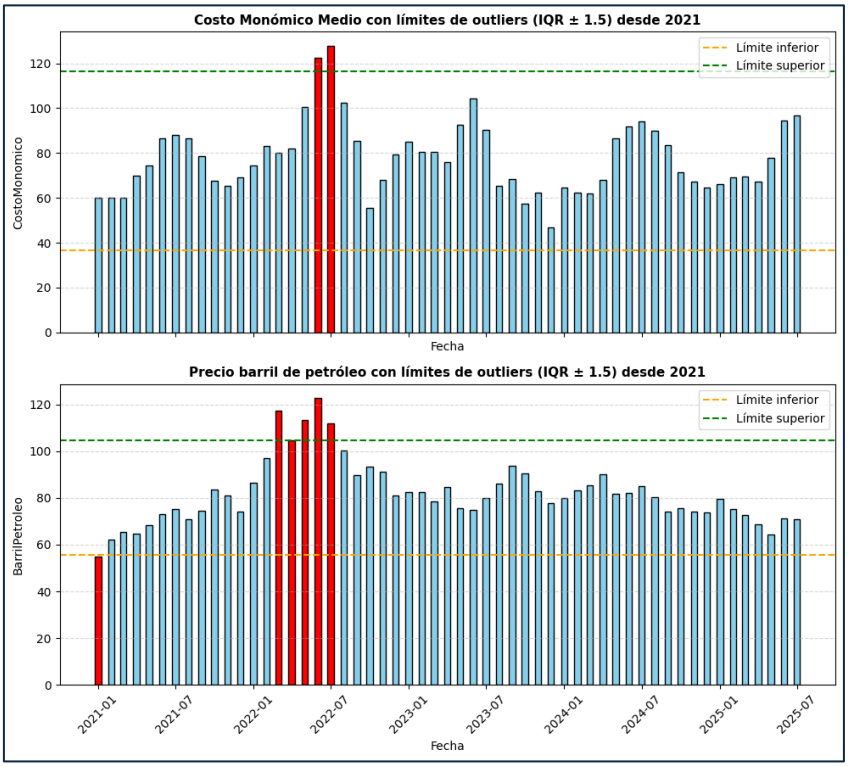


Figura 1. Valores mensuales del costo monómico medio (*arriba*) y el precio del barril de petróleo (*abajo*) desde enero 2021 a julio 2025. Las barras rojas representan aquellos meses en donde hubo *outliers* (IQR±1.5).

Posteriormente se avanzó con un EDA convencional de análisis univariado, aun reconociendo que rompe la secuencia temporal, con el único fin de obtener una aproximación preliminar explicativa; este análisis mostró fuertes correlaciones positivas entre el costo monómico medio y variables como combustible total (r=0.90), gas natural (r=0.87), combustibles alternativos FO+GO+CM (r=0.70) y barril de petróleo (r=0.70). Por otro lado, el costo monómico total local presentó una correlación negativa con la temperatura media con un coeficiente de correlación (r) igual a -0.49 (Figura 2).

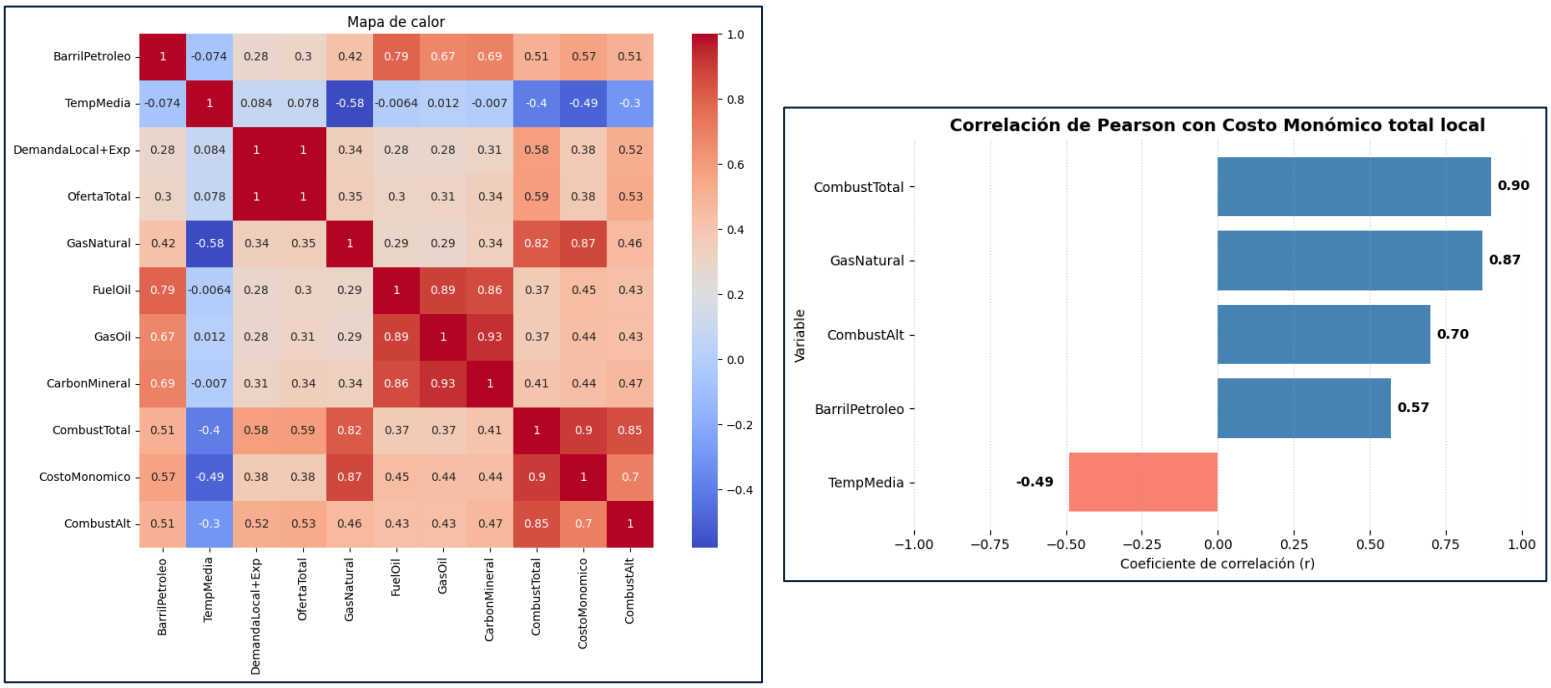


Figura 2. Correlación de Pearson entre todas las variables estudiadas en un mapa de calor (*izquierda*) y la correlación de Pearson entre el Costo Monómico total local y CombustTotal, GasNatural y otras tres variables más (*derecha*).

Posteriormente la serie se trató bajo un esquema temporal clásico, con descomposición en tendencia, estacionalidad y residuo (Figura 3), evidenciando una ausencia de tendencia lineal al alza o a la baja, confirmado a su vez por test estadísticos propios (Mark-Kendall clásico y Dick-Fuller Aumentado con constante y tendencia y KPSS con constante y tendencia), marcada estacionalidad anual con picos invernales, corroborando la conclusión previa con los *outliers* observados (Figura 4), confirmada por ACF y test Ljung–Box ($p<0.001$).

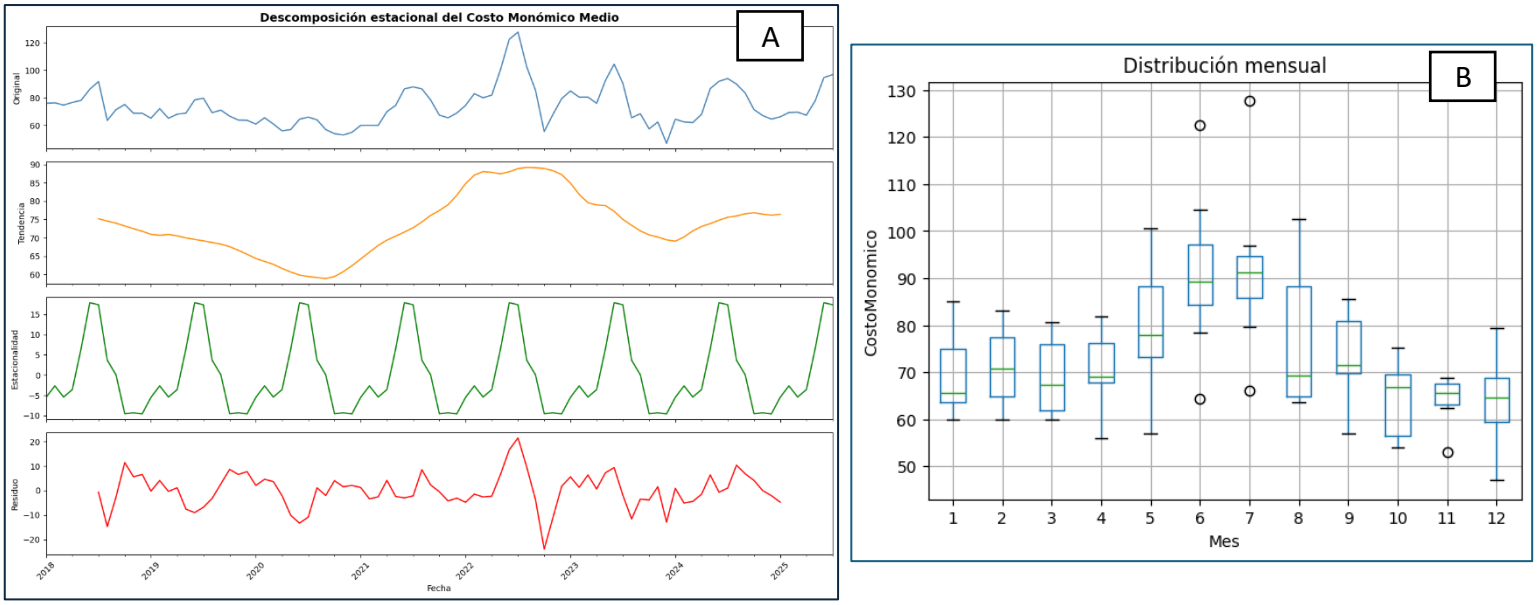


Figura 3. A) Gráfico de descomposición estacional de costo monómico medio. B) Boxplot del costo monómico medio por mes del año, siendo enero el mes 1.

El análisis cualitativo de los gráficos de autocorrelación (ACF) y el análisis de autocorrelación parcial (PACF) permitió seleccionar los lags 1, 2, 10, 11 y 12 para incorporar como variables rezagadas en la base destinada a los modelos de machine learning tradicionales (Ver ANEXO 1).

Con esta información quedó conformada la base de trabajo para la etapa de modelado. Por un lado, la serie original fue utilizada en modelos de ARIMA y SARIMA, aplicando la forma estándar de parametrización:

ARIMA: Los gráficos de ACF y PACF mostraron correlaciones significativas en los rezagos 1 y 2, mientras que a partir del rezago 3 los valores se ubicaron dentro de los intervalos de confianza, indicando ausencia de correlación estadísticamente significativa. Esto sugiere la presencia de componentes AR y/o MA de bajo orden (máximo 2), lo que justificó la evaluación de configuraciones como ARIMA(2,0,0)**,** ARIMA(0,0,2) y combinaciones como ARIMA(2,0,2)**.**

SARIMA: Se modeló como un SARIMAX con seasonal_order=(P,D,Q,12)**,** incorporando la estacionalidad anual de 12 meses identificada en el análisis exploratorio.

Por otro lado, la base enriquecida con variables rezagadas (lags 1, 2, 10, 11 y 12) quedó disponible para el entrenamiento de los modelos de machine learning tradicionales (Ridge, Support Vector Regressor, XGBoost y LightGBM).

4. Modelado y Evaluación

4.1 Modelos de Machine Learning tradicionales

Se evaluaron cuatro algoritmos de regresión: Ridge Regression, Support Vector Regressor (SVR), XGBoost y LightGBM. La selección respondió al objetivo de contar con un espectro de modelos que fueran desde opciones más simples y explicativas (Ridge), hasta algoritmos de mayor complejidad y capacidad predictiva (XGBoost y LightGBM), con SVR como un término intermedio.

Los modelos de boosting (XGBoost y LightGBM) se adaptaron para series de tiempo, removiendo el bootstrapping tradicional con el fin de no romper la secuencia temporal de los datos. Todos los modelos se entrenaron y evaluaron respetando la estructura temporal**,** utilizando como referencia dos benchmarks: Naive y Naive estacional**.**

Para la validación se aplicó un train-test split temporal, utilizando como conjunto de entrenamiento las 72 observaciones mensuales comprendidas entre enero de 2018 y diciembre de 2024, y como conjunto de prueba las 7 observaciones correspondientes al año 2025. La base final de modelado incluyó 52 variables predictoras resultantes de los predictores originales, las variables rezagadas (lags 1, 2, 10, 11 y 12) y las transformaciones estacionales (mes_sin, mes_cos).

El ajuste de hiperparámetros se realizó mediante GridSearchCV**,** siguiendo buenas prácticas de validación, y se utilizó como métrica principal el *Root Mean Squared Error* (RMSE)**,** por su fácil interpretación en las mismas unidades que la variable objetivo (USD/MWh). Adicionalmente, se reportaron métricas secundarias como *Mean Absolute Error* (MAE) y *Mean Absolute Percentage Error (*MAPE), *Mean Absolute Scaled Error* (MASE) y R^2 para complementar la comparación.

Todos los modelos probados, excepto XGBoost, alcanzaron el objetivo de negocio definido junto con el cliente de obtener un RMSE inferior al 10% respecto al valor promedio real del costo monómico (Tabla 1). Sin embargo, al compararlos con el benchmark más exigente, el Naive estacional (lag12), solo el SVR logró superarlo, alcanzando un RMSE de 5.22 frente a 5.40 del Naive estacional (Tabla 1). Este resultado confirma que el SVR no solo cumple con el umbral de error aceptable, si no que además aporta una mejora real frente a los modelos de referencia, justificando su selección como modelo final para la proyección del costo monómico.

Modelo	MAE	RMSE	MAPE	MASE	R^2
Naive (lag1)	5.24	7.71	6.30%	0.73	0.60
Naive estacional (lag12)	4.45	5.40	5.92%	0.32	0.80
Ridge	4.67	6.30	6.04%	0.65	0.73
SVR	4.06	5.22	4.92%	0.57	0.81
LightGBM	4.76	6.73	5.90%	0.67	0.69
XGBoost	9.53	12.32	11.66%	1.33	-0.03

Tabla 1. Comparación de métricas de los cuatro modelos de regresión entrenados y los brenchmark utilizados.

Adicionalmente, se evaluó el desempeño del modelo bajo otros esquemas de train-test split (Expanding Window y Sliding Window), con el objetivo de verificar si el SVR podía mejorar su performance utilizando distintas formas de validación temporal. Los resultados mostraron que esto no ocurrió: en Expanding Window arrojó un RMSE de 11.43, mientras que en Sliding Window métricas similares (RMSE de 9.71). En contraste, el corte fijo sobre 2025 confirmó el mejor rendimiento del SVR (RMSE 5.22), validando este esquema como el más apropiado para el objetivo del proyecto (Figura 4). En consecuencia, si el objetivo es la proyección a futuro inmediato (2025–2026), respalda la elección del SVR como modelo final entrenado en un train test temporal tradicional.

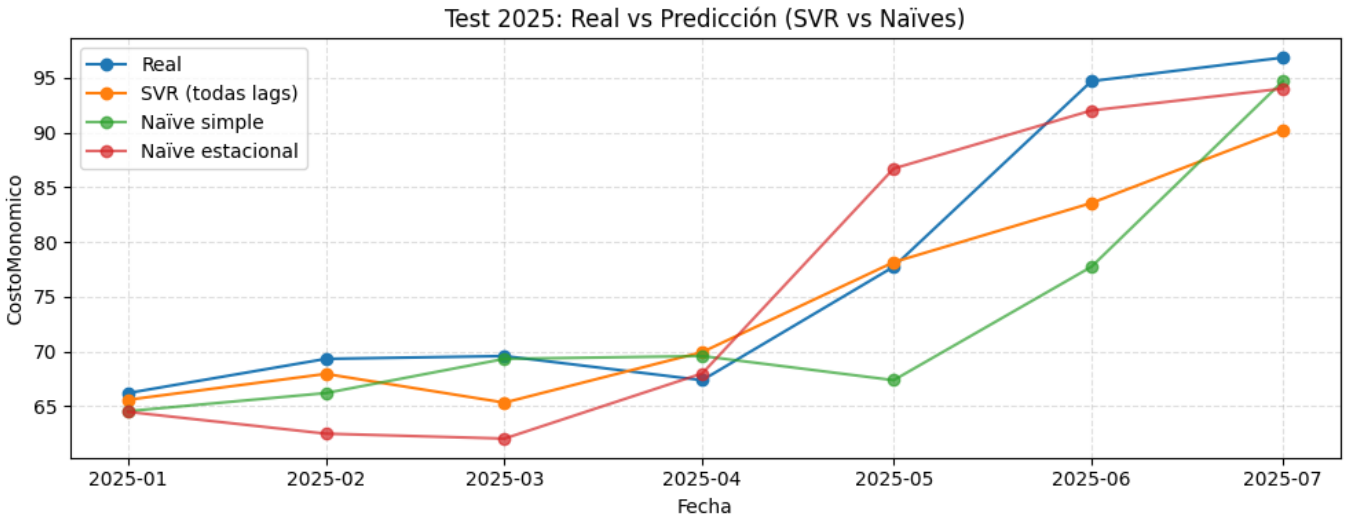


Figura 4. Gráfico comparativo entre las predicciones respecto al costo monómico medio en 2025 del modelo SVR entrenado, los brenchmarks (Naive simple y Naive estacional) y las observaciones reales del costo monómico en los primeros 7 meses del año.

4.2 Variables Explicativas

El uso de SHAP values constituye la forma adecuada de interpretar un modelo kernelizado como el SVR, en el cual las medidas tradicionales de *feature importance* no son aplicables. El análisis reveló que el modelo captura principalmente dos patrones: por un lado, la estacionalidad climática anual, reflejada en la marcada influencia de la temperatura media rezagada 11–12 meses, y por otro, el peso de los insumos energéticos inmediatos y de ciclo anual**,** en particular el consumo de combustibles alternativos y gas natural, tanto en rezagos cortos (1–2 meses) como en rezagos largos de 11–12 meses (Figura 5A). En conjunto, estos hallazgos confirman que el costo monómico medio está determinado por una combinación de factores estacionales (clima y demanda invernal) y factores de mercado energético (disponibilidad y costo de combustibles), lo que alinea la interpretación del modelo con la lógica de negocio del sector eléctrico. El grafico representando la variación del SHAP según el Feature Value confirma la fuerte estacionalidad climática anual y la dependencia de los costos de insumos energéticos en el costo monómico (Figura 5B). En conjunto, el modelo SVR captura con fidelidad que los inviernos fríos y el aumento en el uso de combustibles elevan el costo monómico, mientras que los períodos cálidos y de menor consumo lo reducen, lo cual se alinea con la lógica operativa del mercado eléctrico argentino.

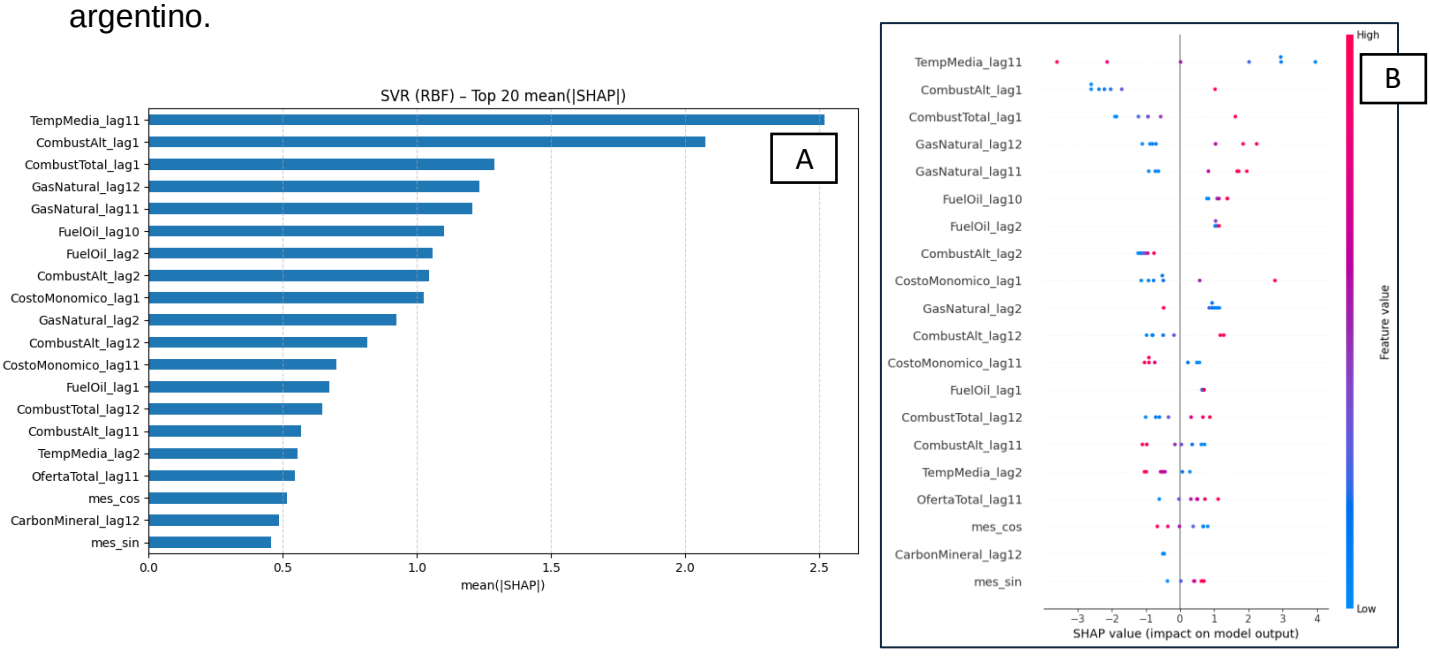


Figura 5: A) Principales variables que influyen en el modelo SVR (RBF) según la influencia media de SHAP. B) Variación en el SHAPE value según el feature value del modelo SVR entrenado.

El análisis de importancia de variables mediante coeficientes absolutos en el modelo *Ridge* mostró resultados consistentes con los obtenidos a través de SHAP en el modelo SVR. En

ambos casos se destacan como predictores principales los combustibles totales y la temperatura media rezagada (particularmente a 10–12 meses), junto con variables derivadas de estacionalidad como *mes_cos*. Esto confirma que, independientemente de la técnica utilizada, los determinantes del costo monómico medio son robustos y se alinean con la lógica de negocio del sector energético.

4.3 Modelos ARIMA Y SARIMA

En base al análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF), se observó evidencia de correlaciones significativas en los rezagos 1 y 2, lo que sugirió la posible presencia de componentes AR y/o MA de bajo orden. Por este motivo se evaluaron configuraciones como ARIMA(2,0,0), ARIMA(0,0,2) y ARIMA(2,0,2), mientras que para capturar la estacionalidad anual se estimaron modelos SARIMA con periodicidad de 12 meses, en particular SARIMA(2,0,2) con distintas combinaciones estacionales (P,D,Q,12).

Los resultados mostraron que el mejor ARIMA fue el (2,0,2), con un RMSE de 10.81 y R^2 de 0.20, muy por debajo de Naive estacional y de los modelos de *machine learning*. En contraste, el SARIMA con *seasonal_order*=(1,0,0,12) presentó un desempeño más razonable (RMSE de 6.97, R^2 de 0.67), aunque aún inferior al benchmark Naive estacional (RMSE 5.37, R^2 0.80) y al modelo final SVR (RMSE 5.22, R^2 0.81) (Ver **Anexo 2**).

Estos resultados sugieren que, si bien el SARIMA logra capturar la estacionalidad anual y mejora respecto a ARIMA, ninguno de los enfoques clásicos superó al Naive estacional, lo cual es un hallazgo relevante: la fuerte estacionalidad del costo monómico ya queda bien representada por este benchmark simple, y los modelos tradicionales de series de tiempo no añadieron valor predictivo adicional (Figura 6).

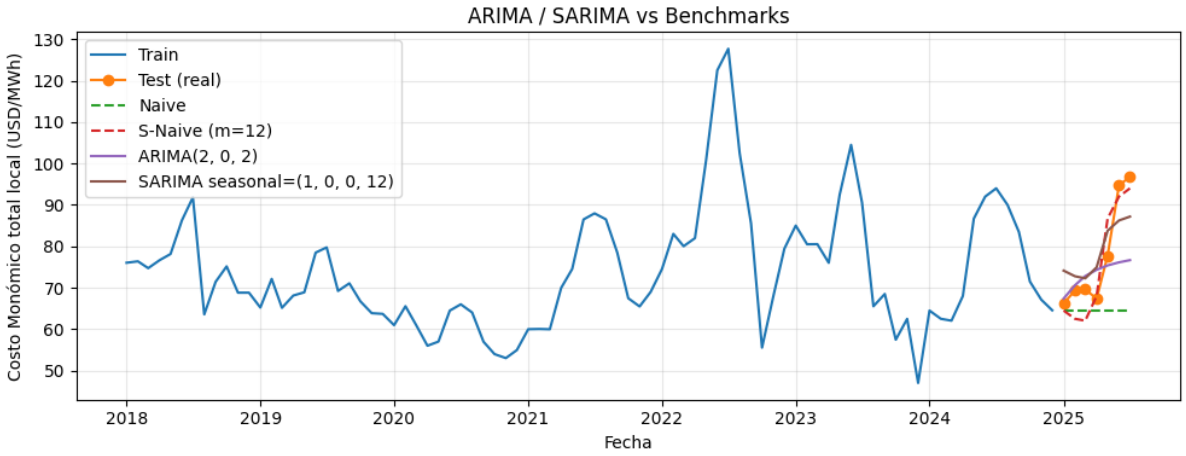


Figura 6. Gráfico comparativo entre las predicciones respecto al costo monómico en 2025 de los modelos ARIMA y SARIMA entrenados, los brenchmarks Naive simple y Naive estacional (S-Naive), y las observaciones reales del costo monómico en los primeros 7 meses del año.

5. Resultados

Los resultados de proyección obtenidos con el modelo SVR permiten anticipar la evolución del costo monómico para el período agosto 2025–diciembre 2026. El modelo respeta la estructura temporal histórica de la serie, reproduciendo la marcada estacionalidad anual: valores más bajos durante los meses de verano–otoño (≈63–70 USD/MWh) y un ascenso progresivo hacia los picos invernales (≈105–107 USD/MWh en junio–julio de 2026) (Tabla 2). Estas proyecciones cumplen con el objetivo de error definido (<10% de RMSE respecto a los valores reales promedio) y constituyen un insumo estratégico para la planificación tarifaria y financiera de las distribuidoras, al ofrecer una estimación confiable de los costos de energía en el corto plazo (Figura 7).

Proyecciones SVR 2025-2026 (Tabla Compacta)

	Fecha 1	SVR Forecast 1	Fecha 2	SVR Forecast 2
1	2025-08-01	97.241142	2026-05-01	72.996282
2	2025-09-01	88.209732	2026-06-01	90.398028
3	2025-10-01	75.300876	2026-07-01	106.907452
4	2025-11-01	69.674303	2026-08-01	105.730215
5	2025-12-01	69.923483	2026-09-01	92.52836
6	2026-01-01	70.637607	2026-10-01	77.795302
7	2026-02-01	69.403146	2026-11-01	71.864099
8	2026-03-01	66.47717	2026-12-01	71.841032
9	2026-04-01	63.683245		

Tabla 2. Predicciones del modelo SVR entrenado respecto al costo monómico durante el período agosto 2025-diciembre 2026.

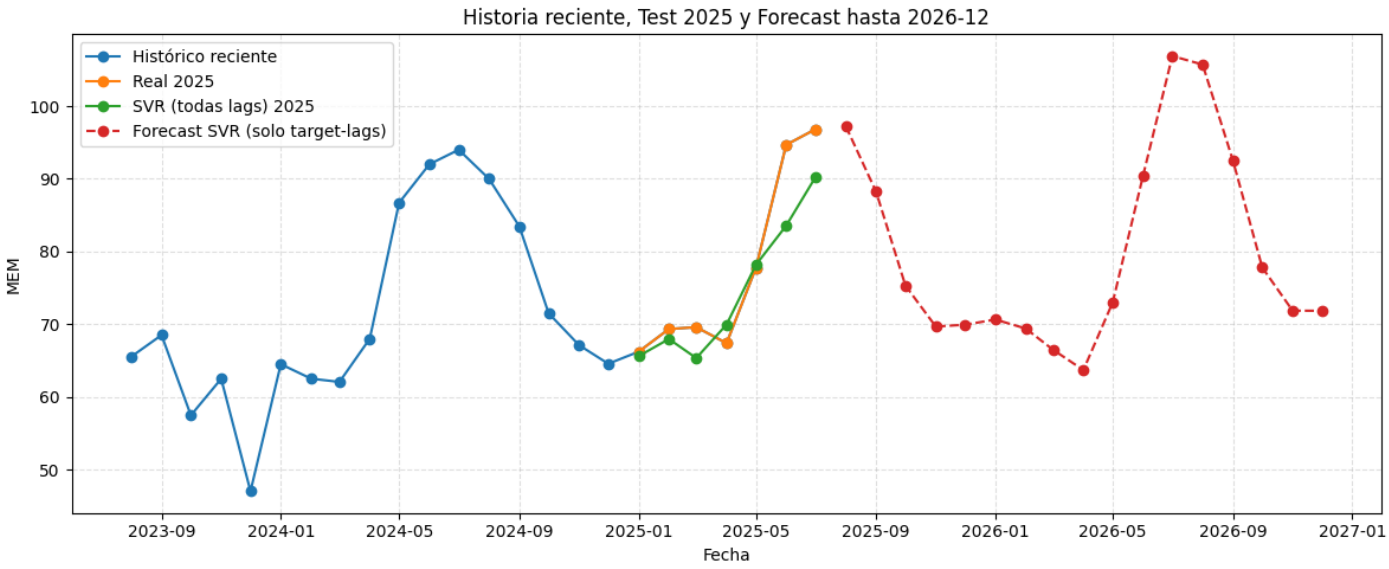


Figura 7. Gráfico comparativo de series temporales representando los valores históricos del costo monómico, las observaciones reales de la primer parte del 2025, las predicciones del modelo SVR entrenado y las predicciones basadas en los valores de los diferentes lags de costo monómico (variable *target*).

6. Conclusión

Siguiendo el marco metodológico CRISP-DM, se abordó el problema de proyectar el costo monómico del Mercado Eléctrico Mayorista argentino hacia 2026, partiendo de una base de datos de CAMMESA con 91 observaciones mensuales entre 2018 y 2025. El proceso incluyó la definición del objetivo de negocio, la exploración y depuración de los datos, la creación de variables rezagadas, el modelado con diversos enfoques y la evaluación comparativa de los resultados frente a benchmarks.

Un primer hallazgo relevante es que el benchmark Naive estacional (lag=12) ofrece un desempeño muy competitivo, con un RMSE de 5.37 y R^2=0.80, reflejando que gran parte de la dinámica del costo monómico está dominada por una marcada estacionalidad anual. Esto explica porqué los modelos clásicos como ARIMA y SARIMA, a pesar de capturar parcialmente la autocorrelación y la estacionalidad de la serie, no lograron superar al benchmark: el mejor SARIMA alcanzó un RMSE de 6.97 y R^2=0.67, mientras que el mejor ARIMA quedó muy por detrás (RMSE 10.81, R^2=0.20).

En contraste, los modelos de machine learning mostraron mayor capacidad para integrar la influencia de variables exógenas. Entre ellos, el SVR fue el de mejor desempeño, superando al Naive estacional y alcanzando un RMSE de 5.22 y R^2=0.81**.** Otros modelos

como *Ridge* y LightGBM también lograron métricas aceptables, pero con menor precisión. El XGBoost, en cambio, presentó dificultades y quedó descartado por errores muy superiores al umbral definido.

El análisis de interpretabilidad mediante SHAP values confirmó que el SVR no solo predice mejor, si no que además captura patrones alineados con la lógica de negocio: 1) Factores estacionales climáticos → temperatura media rezagada 11–12 meses, que refleja la mayor demanda y presión sobre los costos en los meses invernales y 2) Factores energéticos inmediatos → combustibles alternativos (fuel oil, gas oil, carbón mineral) y gas natural, tanto en rezagos cortos (1–2 meses) como largos (11–12 meses), que determinan el costo marginal de generación eléctrica.

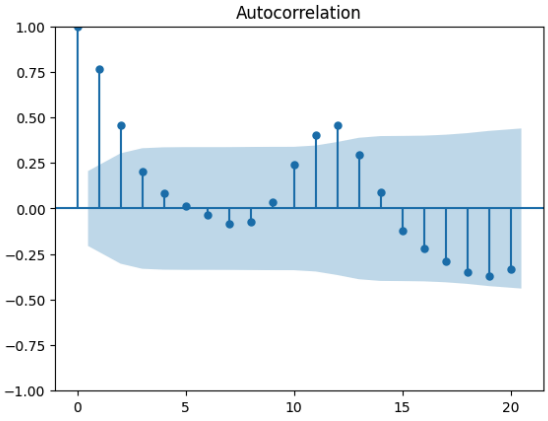
Esto implica que, para anticipar el costo monómico, los clientes deben prestar especial atención a la evolución de estas variables, que combinan dinámicas de corto plazo (mercado de combustibles) con **ciclos** estacionales recurrentes (clima y demanda invernal)**.**

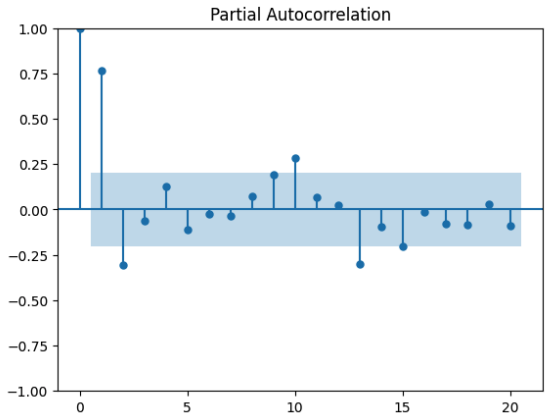
Las proyecciones para 2026 muestran valores en torno a 63–70 USD/MWh durante los meses estivales y picos cercanos a 107 USD/MWh en los meses de invierno, lo que respeta la estructura histórica del costo monómico. Esta información representa un insumo estratégico para las distribuidoras, ya que permite anticipar escenarios de costos, diseñar coberturas financieras y definir estrategias tarifarias más precisas.

Como próximos pasos, sería recomendable: 1. Ampliar la base de datos y considerar nuevas variables macroeconómicas (inflación, tipo de cambio, precios internacionales de *commodities*) y regulatorias (ajustes tarifarios o subsidios), que pueden aportar señales adicionales. 2. Implementar un sistema de monitoreo continuo, que permita reentrenar periódicamente los modelos con datos más recientes, ajustando automáticamente las proyecciones a los cambios en el entorno energético.

En conclusión, el trabajo demuestra que, si bien la estacionalidad explica gran parte del costo monómico y modelos simples como el Naive estacional ya ofrecen resultados sólidos, la incorporación de métodos avanzados de *machine learning* con variables exógenas permite mejorar la precisión y obtener interpretaciones valiosas. El SVR se destaca como el modelo final más adecuado para la proyección a corto plazo, ofreciendo a los clientes un instrumento confiable para planificación estratégica, gestión de riesgos y toma de decisiones en un sector crítico para la economía argentina.

Anexo 1: Gráficos de autocorrelación y autocorrelación parcial de la variable costo monómico en la serie temporal.





Anexo 2: Tabla comparativa de métricas de ARIMA y SARIMA vs Benchmarks tradicionales

Resultados comparativos ARIMA / SARIMA vs Benchmarks

Modelo	MAE	RMSE	MAPE	MASE	R^2
Naïve estacional (lag=12)	4.453	5.377	5.93%	0.345	0.803
SARIMA (2,0,2), seasonal=(1,0,0,12)	6.536	6.970	8.38%	0.506	0.669
ARIMA (2,0,2)	7.674	10.814	8.86%	1.104	0.204
Naïve simple (lag=1)	12.843	17.660	14.71%	1.848	-1.122