Medición de la actividad económica desde un enfoque geoespacial

Camilo Valladares

1 de septiembre de 2025

Resumen

Utilizando modelos de Deep Learning y Machine Learning (basados en árboles) abordamos la detección de plantas y la estimación de su actividad económica a partir de datos geoespaciales. Para cumplir con este objetivo, nos centramos en la industria siderúrgica española, un sector desafiante por sus bajas emisiones atmosféricas debido al uso de hornos de arco eléctrico. Se combinan datos de producción con variables geoespaciales de teledetección, incluyendo emisiones al aire (CO, SO_2, $NO2$ y $PM_{2,5}$), luz nocturna (NTL) y temperatura de la tierra (LST), para entrenar modelos y predecir la ubicación de plantas y producción. En la evaluación con datos no vistos, el modelo de clasificación basado en redes neuronales para identificar plantas logró una precisión de 98 %, mientras que los modelos basados en árboles de decisión, para estimar la producción, obtuvieron un R^2 por sobre 0.9. Estos resultado muestran un avance significativo respecto a estudios previos (Yang et al., 2025) al demostrar que la información geoespacial es eficaz incluso en escenarios de baja emisión. Este enfoque muestra una gran utilidad para la toma de decisiones públicas e investigadores, ofreciendo un método robusto para medir la actividad económica casi en tiempo real.

1. Introducción

La medición económica es crucial para obtener una visión objetiva del desempeño económico de un país. Indicadores como el PIB en sus distintas frecuencias son herramientas clave para medir la salud de una economía, evaluar su desempeño y anticipar tendencias, y sobre todo, son la base para la toma de decisiones de gobiernos, empresas e inversores. A pesar de su importancia, los métodos tradicionales de medición de la actividad económica presentan tres limitaciones principales: rezago en la publicación de los datos, baja granularidad y un alto costo. Por ejemplo, el Indicador de Producción Industrial (IPI) de España se publica con un rezago aproximado de cinco semanas respecto al mes de referencia y se ofrece como un agregado para toda la producción industrial, ocultando que está conformado por actividades económicas heterogéneas. Esta metodología, además, representa un alto costo por ser intensivos en el uso de encuestas, lo que ha provocado que países en desarrollo tengan estadísticas económicas de baja frecuencia e incompletos, afectando negativamente la robustez de dichas estimaciones y su planificación.

Los datos geoespaciales han emergido como una nueva fuente de información que permite mitigar estas limitaciones. Investigadores han demostrado que es posible estimar el PIB utilizando datos satelitales de luces nocturna (Henderson et al., 2012), aunque estos estudios se enfocan en mediciones agregadas y pueden carecer de precisión temporal.

En este estudio desarrollamos una nueva metodología para detectar y medir la actividad económica con alta precisión temporal y espacial. Nuestro enfoque combina la producción industrial y su localización con señales de ambientales casi en tiempo real, lo que nos permite ofrecer una herramienta de monitoreo económico más precisa y escalable. La idea principal es que cada actividad económica tiene una huella ambiental distintiva, lo que nos permite identificar su ubicación e intensidad, para una actividad predefinida, a través de indicadores geoespaciales. Para demostrarlo, nos centramos en la industria siderúrgica española, caracterizada por bajos niveles de emisión atmosférica al utilizar predominantemente hornos de arco eléctrico.

La metodología de nuestro trabajo comprende dos etapas: primero, utilizamos modelos de Deep Learning para predecir la ubicación de plantas siderúrgicas a partir de datos geoespaciales y las coor-

denadas de plantas conocidas. Luego, con los datos de las plantas identificadas, aplicamos modelos de Machine Learning para predecir la producción de acero.

La principal contribución de este trabajo es doble. En primer lugar, demostramos que es posible predecir con alta precisión la ubicación de una planta productiva no incluida en el conjunto de entrenamiento. A partir de un modelo de red neuronal, logramos predecir con gran precisión la presencia de una planta siderúrgica. Las variables que mostraron una mayor importancia en esta etapa fueron la luz nocturna (NTL) y la temperatura de la superficie terrestre (LST).

En segundo lugar, probamos que es posible obtener una alta precisión al predecir la producción industrial utilizando una variedad de indicadores ambientales. En esta etapa, los modelos basados en árboles que entrenamos mostraron que (NTL) y el dióxido de nitrógeno ($NO2$) son predictores claves de la producción. Para la validación del modelo se utilizaron distintas ventanas temporales y lo complementamos con k-fold cross validation. Finalmente, encontramos que nuestras predicciones usando datos geoespaciales tuvieron un R^2 sobre un 0.9.

2. Industria siderúrgica y factores ambientales geoespaciales

A diferencia de las plantas siderúrgicas tradicionales, que se basan en la fundición de mineral de hierro en altos hornos, la producción de acero en España se caracteriza por el uso predominante de hornos de arco eléctrico (EAFs). Este proceso, que representa la mayor parte del acero bruto producido en el país, utiliza chatarra de acero y otros materiales reciclados como materia prima. Si bien, es común que las plantas de acero tengan integrada la producción hasta variados productos de acero, en este estudio nos enfocamos en la producción de acero bruto, siguiendo la convención de indicadores económicos buscando la homogeneidad en la canasta de productos.

El uso de hornos de arco eléctrico tiene implicaciones directas en la huella ambiental que deja la industria, lo que hace posible el monitoreo con datos geoespaciales. Durante el proceso de producción, la intensa actividad genera una serie de indicadores ambientales distintivos. Por ejemplo, aunque este tipo de producción tiene menores emisiones de SO_2 y NO_x, en comparación con los altos hornos, aún libera ciertos contaminantes, como las partículas finas ($PM_{2,5}$) y CO, lo que puede ser detectado por satélites. Además, la operación continua de estos hornos y el manejo de material caliente tienen un efecto significativo en la temperatura de la superficie terrestre y en la luz nocturna de las áreas circundantes, incluso en comparación con las operaciones de día. Esto convierte a estas variables en proxies excelentes para monitorear la actividad de la producción. Más adelante detallaremos las fuentes de información y la construcción del dataset.

2.1. Indicadores ambientales geoespaciales utilizados

Para este estudio, empleamos una variedad de indicadores geoespaciales derivados de datos satelitales que reflejan la huella ambiental de la producción industrial. La combinación de estas variables nos permite capturar diferentes aspectos de la actividad de las plantas siderúrgicas, desde la contaminación del aire hasta la intensidad energética. El uso de gases como NO_2, SO_2 y CO como indicadores de actividad industrial ha sido ampliamente documentado en la literatura. Por ejemplo, estudios como Wei et al. (2023) han demostrado la posibilidad de generar mapas diarios continuos de contaminantes, validando su utilidad para análisis espaciales de alta resolución.

Utilizamos la columna total de tres contaminantes atmosféricos clave: el dióxido de azufre (SO_2), el dióxido de nitrógeno (NO_2), el monóxido de carbono (CO) y el ozono (O_3). El SO_2 y el NO_2 son subproductos comunes de la combustión a alta temperatura en procesos industriales como los hornos de arco eléctrico, mientras que el CO se genera por la combustión incompleta, lo que lo convierte en un indicador relevante de la actividad. El O_3, si bien se forma a partir de precursores en la atmósfera, puede ser un indicador de las condiciones atmosféricas locales que influyen en las emisiones. Para el período de 2018 en adelante, las mediciones de estos gases se obtuvieron del instrumento TROPOMI a bordo del satélite Sentinel-5P de la misión Copernicus, gracias a su alta resolución espacial y temporal. Para cubrir los vacíos de información, especialmente en los meses de invierno, las series de SO_2 y CO se complementaron con datos de misiones anteriores de la NASA, como el Ozone Monitoring Instrument (OMI) a bordo del satélite Aura (Levelt et al., 2006; Veefkind et al., 2012) y el instrumento Measurements of Pollution in the Troposphere (MOPITT) a bordo del satélite Terra (Deeter et al., 2019), respectivamente. Adicionalmente, se incluyó el material particulado con un diámetro

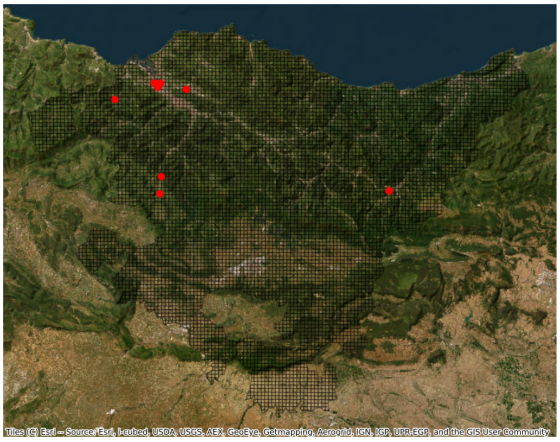
aerodinámico menor a 2.5 micrómetros ($PM_{2,5}$), un contaminante relevante para la salud pública y la huella industrial.

Adicionalmente a las emisiones, incorporamos variables que reflejan la actividad energética y térmica de las plantas. La operación de las plantas siderúrgicas 24 horas al día genera una luz nocturna (NTL) detectable por satélites como el VIIRS (Chen and Nordhaus, 2019). Tal como señalan Levin et al. (2020), la evolución en la teledetección de NTL ofrece nuevas oportunidades para monitorear procesos económicos con mayor resolución temporal y espacial. Del mismo modo, la operación de hornos a alta temperatura genera un calor residual que eleva la temperatura de la superficie terrestre (LST) en las zonas circundantes, lo que puede ser medido por satélites como MODIS. Estas variables complementan la información sobre emisiones y brindan una visión más completa de la intensidad de la actividad industrial.

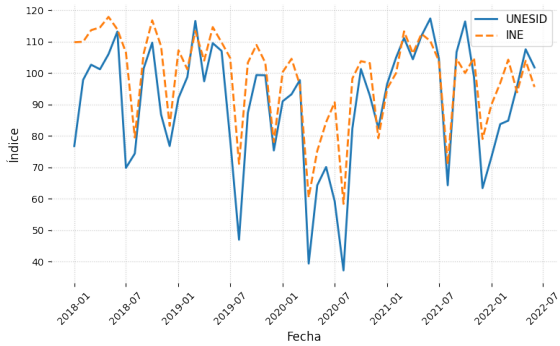
2.2. Datos de localización de plantas y producción

Para este estudio, la información sobre la producción y localización de las plantas siderúrgicas en España fue proporcionada por la Unión de Empresas Siderúrgicas (UNESID), la asociación que agrupa a las empresas productoras de acero en el país. UNESID nos facilitó la serie de producción a nivel de planta, específicamente para las ocho fábricas de acero que operan con hornos de arco eléctrico en el País Vasco, cubriendo el período desde enero de 2018 hasta junio de 2022. La naturaleza de esta industria, con pocas plantas concentradas en una región específica, facilitó la verificación de sus ubicaciones. Al no ser un número elevado de plantas, sus coordenadas geográficas fueron fácilmente confirmadas mediante el uso de imágenes satelitales, permitiendo asociar cada planta con una celda única de 1x1 km en la cuadrícula de nuestro estudio. Es importante destacar que nuestro modelo de regresión para estimar la producción fue entrenado exclusivamente con los datos de las plantas ubicadas en el País Vasco.

La precisión de los datos de producción de UNESID fue validada comparándolos con el Índice de Producción Industrial (IPI) publicado por el Instituto Nacional de Estadística (INE) de España. Para esta comparación, los niveles de producción de UNESID se transformaron en un índice con base en 2021, la misma base que utiliza el IPI del INE (ver panel b de la Figura 1). A nivel agregado, ambos indicadores mostraron una dinámica temporal muy similar, lo que confirma la fiabilidad de los datos de UNESID. Las diferencias observadas entre ambos índices se deben principalmente a que el IPI del INE incluye la producción de una importante planta de altos hornos ubicada en Asturias, mientras que nuestro conjunto de datos se centra exclusivamente en las plantas de horno de arco eléctrico del País Vasco. Esta coherencia en la dinámica de producción nos permitió confiar en el conjunto de datos de UNESID como una base sólida para nuestro análisis.



(a) Grilla del País Vasco con plantas



(b) Producción en UNESID vs INE

Figura 1: Localización de plantas y producción.

Nota: El índice del INE corresponde al Índice de Producción Industrial en base 2021 para el grupo específico *Fabricación de productos básicos de hierro, acero y ferroaleaciones* a nivel nacional. Por su parte, los niveles de producción en UNESID fueron llevados a números índice con base 2021 para ser comparables con el índice publicado por el INE para la producción de acero. En UNESID solo contamos con la producción para el País Vasco.

2.3. Construcción de grilla y descarga de información

El manejo de datos comenzó por la creación de una grilla que divide el País Vasco en 7,718 celdas de 1x1 km. Esta cuadrícula sirvió como la base geográfica para nuestro análisis. Las 8 plantas siderúrgicas proporcionadas por UNESID se asociaron a la celda en la que se encontraba el centroide de sus coordenadas, permitiendo que la producción de cada fábrica se vinculara de forma única a un grid de 1x1 km. Posteriormente, para cada una de estas celdas, se extrajo la serie temporal mensual de los indicadores ambientales geoespaciales, permitiendo crear un conjunto de datos robusto y homogéneo para el entrenamiento de los modelos.

Para obtener los indicadores ambientales, se descargaron los datos satelitales directamente de las colecciones de Google Earth Engine (GEE). Los datos de SO_2, NO_2, CO Y O_3 provienen del instrumento TROPOMI a bordo del satélite Sentinel-5P de la misión Copernicus. Para el $PM_{2,5}$ se descargó la información del proyecto Copernicus Atmosphere Monitoring Service (CAMS) Global Near-Real-Time. NTL se obtuvo del instrumento VIIRS y LST de los datos de MODIS.

En su origen, todos estos conjuntos de datos poseen una frecuencia diaria y una resolución espacial nativa variable. Para nuestro estudio, los datos se reproyectaron y agregaron a nuestra cuadrícula de celdas de 1x1 km, y se calculó el valor promedio mensual para cada indicador. Esto nos permitió crear una serie temporal unificada y homogénea para el análisis.

Dada la fecha de inicio de operación de Sentinel-5P, la serie de datos para las emisiones al aire se complementaron con las mediciones del instrumento MOPITT del satélite Terra de la NASA, que ha proporcionado datos de forma continua desde el año 2000, asegurando una cobertura temporal completa. Misma fuente fue utilizada para mejorar los meses con alta nubosidad, donde, en particular para el SO_2, tuvimos menos días de medición.

En el Apéndice se puede encontrar un mayor detalle de las variables incorporadas en el conjunto de datos, con su descripción y estadística descriptiva.

3. Metodología

Desarrollamos una metodología de dos etapas para abordar el problema de medir la actividad industrial con datos geoespaciales. Primero, se utilizó un enfoque de clasificación para detectar la ubicación de las plantas siderúrgicas en el País Vasco. Posteriormente, se aplicó un modelo de regresión basado en árboles para estimar la producción de acero en las ubicaciones identificadas. La implementación de

esta metodología se basó en la construcción de un conjunto de datos robusto y homogéneo, la selección rigurosa de variables clave y el entrenamiento de modelos de Machine Learning y Deep Learning.

3.1. Modelos

Nuestra primera etapa metodológica se centró en un problema de clasificación: predecir la probabilidad de que una celda de nuestra cuadrícula contenga una planta siderúrgica. Para abordar este desafío, utilizamos un modelo de Deep Learning basado en redes neuronales, esto ya que fue el modelo con mejor desempeño contra otros modelos de Machine Learning (Ver Apéndice B). Reconocimos que este problema presentaba un desafío significativo debido al desequilibrio de clases, ya que la cantidad de celdas con una planta es considerablemente menor que el número de celdas sin una. Un desbalance tan marcado puede sesgar los modelos, llevándolos a predecir de forma predominante la clase mayoritaria. Para mitigar este problema y permitir que nuestro modelo aprendiera de manera efectiva de ambas clases, aplicamos la técnica de SMOTE (Synthetic Minority Over-sampling Technique), la cual genera ejemplos sintéticos de la clase minoritaria para equilibrar el conjunto de datos de entrenamiento.

Para la clasificación, diseñamos una red neuronal secuencial con tres capas densas ocultas como la de la Figura 2. Cada una de estas capas utiliza la función de activación ReLU y un regularizador L2, lo que ayuda a prevenir el sobreajuste del modelo. Se incluyeron capas de Dropout después de cada capa densa para mejorar la generalización del modelo. La capa de salida es una capa densa con una sola neurona y una función de activación sigmoide para producir la probabilidad de que una celda contenga una planta. El modelo se compiló utilizando el optimizador Adam y la función de pérdida Binary Cross-entropy, y se evaluó con métricas como la precisión, el recall y la exactitud. Es importante destacar que, para manejar el desequilibrio de clases, los pesos de clase fueron ajustados para dar mayor importancia a la clase minoritaria (las celdas con plantas), lo cual complementó la técnica de SMOTE aplicada previamente para balancear el conjunto de datos de entrenamiento.

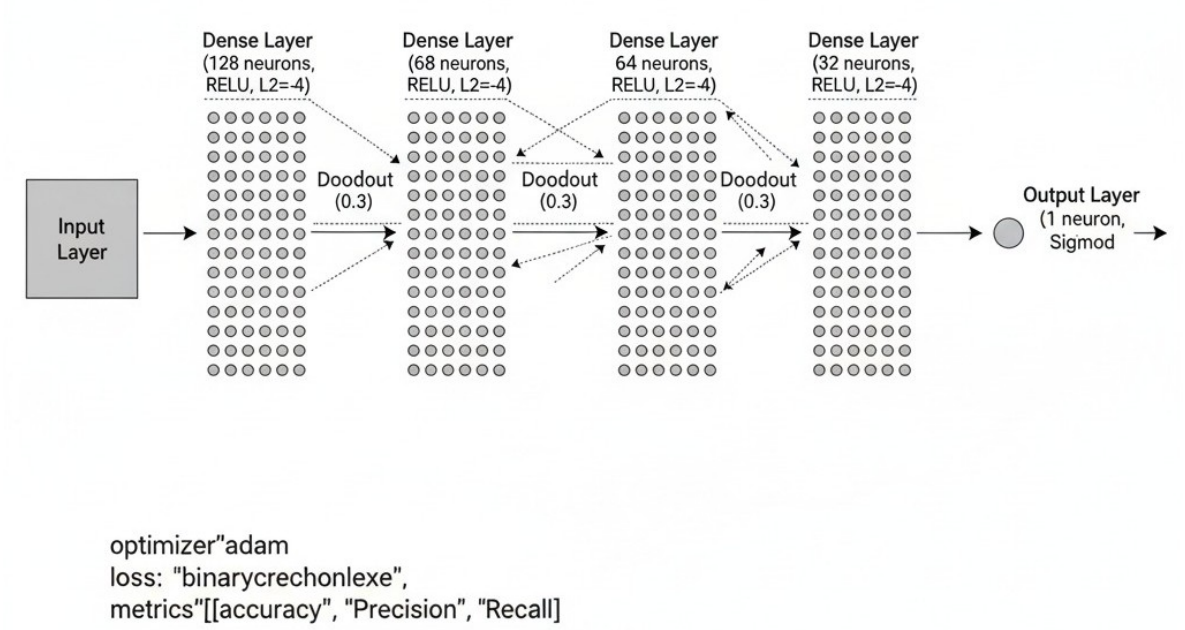


Figura 2: Modelo de red neuronal

Para la segunda etapa de nuestro análisis, nos centramos en la predicción de la producción industrial, trabajando exclusivamente con las celdas de la cuadrícula que contenían plantas. Dado que se trata de un problema de regresión, el modelo que mostró un mejor resultado fue el XGBoost, un algoritmo de aprendizaje supervisado basado en árboles que ha demostrado un excelente desempeño en la generalización con datos complejos, y, que además, sigue permaneciendo como recomendado para conjuntos de datos medianos. Los hiperparámetros del modelo, incluyendo el número de estimadores, la profundidad máxima y el número de características a considerar en cada división, fueron optimizados a través de una búsqueda de grilla (GridSearch) para maximizar su precisión.

Modelamos el valor esperado de Y basado en un conjunto de p predictores, X, utilizando la siguiente especificación:

$$Y = f(X_1, X_2, ..., X_p) + \epsilon$$

, donde Y representa la producción de acero crudo a nivel de grid, X_1, X_2,...,X_p representan las variables predictoras y ϵ el error de predicción, con $\mathbb{E}[Y|X_1, X_2, ..., X_p] = f(X_1, X_2, ...X_p)$ y esperanza condicional cero.

Para evaluar y comparar la performance del modelo utilizado, previamente estimamos diferentes modelos, incluyendo regresión lineal, Lasso, Kernel ridge regression, ElasticNet, random forest, LightGBM y XGBoost. Siendo este último el de mejor desempeño y generalización. Mostramos un mayor detalle de estos resultados en el Apéndice C.

3.2. Variables a utilizar

Para el desarrollo de los modelos, se seleccionaron y prepararon variables geoespaciales clave, las cuales se agruparon en tres tipos principales para el análisis. En primer lugar, se incluyeron las emisiones de contaminantes atmosféricos (SO_2, NO_2, CO y O_3), que sirven como indicadores directos de la actividad industrial. Para capturar las dinámicas temporales y la inercia del sistema, se incorporaron sus valores rezagados (lags) de uno, dos y tres meses. En segundo lugar, se incluyeron factores ambientales adicionales, como la Luz Nocturna (NTL) y la Temperatura de la Superficie Terrestre (LST), también con rezagos de uno, dos y tres meses, para capturar la intensidad energética y la huella térmica de las operaciones de las plantas. Adicionalmente, se incluyó el mes como una variable categórica para capturar cualquier patrón de estacionalidad no reflejado en las variables de emisiones. Finalmente, y de forma crucial para la estimación de la producción, se incorporaron la latitud y la longitud del centroide de cada celda, lo que permitió que el modelo considerara la ubicación geográfica como un factor predictivo relevante para capturar la heterogeneidad espacial inherente a la industria.

3.3. Entrenamiento y validación

En el desarrollo de nuestro trabajo, seguimos la práctica estándar en Machine Learning de dividir el conjunto de datos completo en un subconjunto de entrenamiento y uno de prueba, utilizando una proporción de 80/20. Esta división nos permitió entrenar nuestros modelos con una parte de los datos y, posteriormente, validar su rendimiento con un conjunto de datos completamente no visto, lo que asegura una evaluación objetiva. Es importante señalar que, para ambos modelos, tanto el de clasificación para la detección de plantas como el de regresión para la estimación de la producción, nos enfocamos en los grids ubicados en la región del País Vasco y restringimos el análisis a la serie temporal entre enero de 2018 y junio de 2022.

4. Principales resultados

Para evaluar el rendimiento de nuestros modelos, utilizamos un conjunto de métricas estándar en el campo del aprendizaje automático, adaptadas a la naturaleza de cada tarea. Para el modelo de clasificación basado en redes neuronales, que predice la probabilidad de que una celda contenga una planta, evaluamos su desempeño con métricas como la exactitud (accuracy), la precisión (precision), el recall y el área bajo la curva ROC (AUC). Para el modelo de regresión que estima la producción de acero, nos centramos en el coeficiente de determinación (R^2). Además de estas métricas, pusimos en la interpretación de las variables para entender qué indicadores geoespaciales influyen más en las predicciones (en la medida que los modelos nos lo permitan).

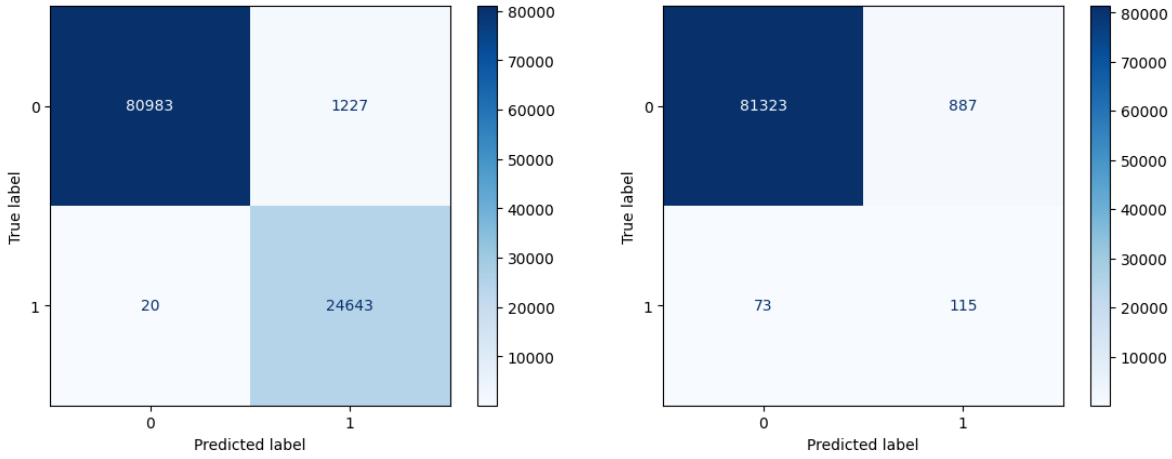
4.1. Predicción de ubicación

Nuestra principal meta fue entrenar un modelo de clasificación que utiliza como input datos y atributos geoespaciales para predecir si un grid contiene o no una planta siderúrgica. Para el entrenamiento y prueba de este modelo, se incluyó la información mensual para los 7,718 grids de 1x1 km de la región del País Vasco, de los cuales solo ocho contienen las plantas de acero compartidas por UNESID. La evaluación de este modelo se realizó en dos etapas. Una primera etapa, se siguió la práctica estándar del machine learning al utilizar la división train-test-split de la biblioteca scikit-learn en una

proporción de 80/20. En la segunda etapa, validamos la capacidad de generalización del modelo con un conjunto de datos externos: tomamos plantas sudamericanas que no fueron incluidas en los conjuntos de entrenamiento y prueba, las geolocalizamos en grids y descargamos la información geoespacial para que nuestro modelo predijera si contenían o no una planta.

Para evaluar el desempeño del modelo de clasificación en la predicción de la ubicación de las plantas, utilizamos las matrices de confusión mostradas en la Figura 3. El modelo fue entrenado con datos balanceados utilizando SMOTE, pero los resultados de prueba se analizaron con los datos originales, sin y con casos sintéticos. En el panel a de la Figura 3, con un umbral estándar de 0.5, se observa un alto número de verdaderos negativos (grids correctamente identificados sin planta). Los falsos negativos, donde el modelo predice erróneamente la ausencia de una planta, son muy bajos, con solo 20 casos. Aunque hubo 1,227 falsos positivos, la precisión y el recall del modelo fueron excepcionalmente altos, alcanzando un 98.8 % en ambos casos. Este análisis inicial se basa en la predicción a nivel de cada observación mensual.

En la matrix de confusión sin casos sintéticos (panel b de la Figura 3), vemos que el desempeño del modelo parece empeorar. Sin embargo, para refinar el rendimiento y obtener una clasificación a nivel de grid completo, exploramos un enfoque alternativo en este caso. Cuando redefinimos la clasificación de una celda de la grilla como çon planta"solo si un porcentaje mínimo de sus observaciones mensuales (por ejemplo, el 20 %) se clasificaban positivamente, el rendimiento del modelo mejoró significativamente. A través de este proceso, logramos un recall del 91.6 %, lo que demuestra la capacidad del modelo para identificar de manera confiable la presencia de una planta siderúrgica. Al llevar los resultado a nivel de planta (agrupando la serie mensual), vemos que los errores de predicción ocurren en áreas cercanas a grids con planta. Esto sería solucionable ampliando el tamaño del grid (ver panel a de la Figura 4) .



(a) Resultados con SMOTE

(b) Resultados sin SMOTE

Figura 3: Matriz de confusión en conjunto de test

Nota: En estos resultados, SMOTE efectivamente aborda el problema de desbalanceo tanto en entrenamiento como en test. Sin embargo, vemos que los resultados difieren claramente si evaluamos el modelo en el conjunto de test con o sin SMOTE. La cantidad de casos sintéticos generados corresponden a un 0.3 de os casos reales. Se utilizó un umbral de clasificación estándar de 0.5.

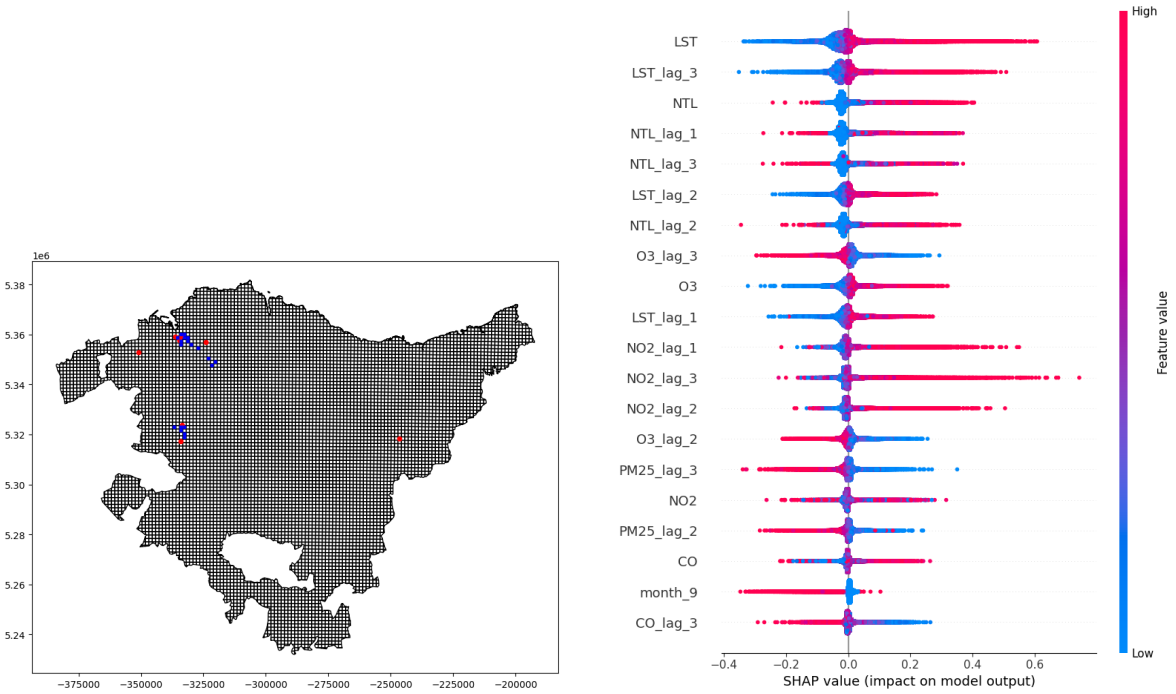
Para interpretar los resultados de nuestro modelo y entender la relación entre las variables explicativas y la predicción de la existencia de una planta, utilizamos la técnica de SHAP (SHapley Additive exPlanations). Esta herramienta aborda la crítica común de la falta de interpretabilidad de los modelos de çaja negra", como las redes neuronales y los modelos basados en árboles. Basada en la teoría de juegos, SHAP distribuye equitativamente la contribución de cada variable a la predicción del modelo. Esto nos permite analizar tanto la interpretabilidad global (determinando la importancia general de las características para el modelo) como la interpretabilidad local (explicando por qué se tomó una decisión concreta para una predicción individual). Una forma sencilla e intuitiva de expresar

la descomposición de SHAP es a través de un modelo aditivo:

$$f(X) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

, donde $f(X)$ es la predicción del modelo para una instancia específica X, ϕ_0 es el valor base del modelo, que es la predicción esperada si no tuviéramos información de ninguna variable (generalmente, el promedio de las predicciones del conjunto de entrenamiento) y ϕ_i es el valor SHAP de la variable i, que representa la contribución de esa variable para mover la predicción desde el valor base hasta la predicción final.

El panel b de la Figura 4 muestra la importancia de las variables para la predicción del modelo de clasificación, analizada a través de los valores SHAP. Los resultados indican que la predicción de la ubicación de una planta siderúrgica depende en gran medida de la temperatura de la superficie terrestre (LST) y de la luz nocturna (NTL). En menor medida, las mediciones de ozono (O_3) también resultaron ser un predictor relevante. Estos hallazgos son particularmente interesantes, ya que se distancian de los resultados de estudios previos como el de (Yang et al., 2025). Mientras que en ese trabajo se identificó a las emisiones de CO y NO_2 como los principales predictores para una industria altamente contaminante (funcionaban con altos hornos principalmente), nuestro estudio, al enfocarse en un sector siderúrgico más limpio, muestra que la huella ambiental se relaciona más con factores como la huella térmica y la intensidad de la actividad energética nocturna que con las emisiones directas al aire.



(a) Grilla con plantas y errores de predicción

(b) Valor SHAP

Figura 4: Errores de predicción y Valor SHAP en el modelo de clasificación

Nota: El panel de la izquierda (a) muestra el área del País Vasco con las plantas reales en rojo y en azul los errores de predicción (celdas que se predice que contienen plantas erróneamente). El panel de la derecha (b) muestra el valor SHAP de cada variable de entrada en nuestro modelo de clasificación. Este valor muestra la contribución de cada variable en la predicción del modelo. Cada punto representa una muestra en el modelo y la densidad refleja la distribución de los valores de las características.

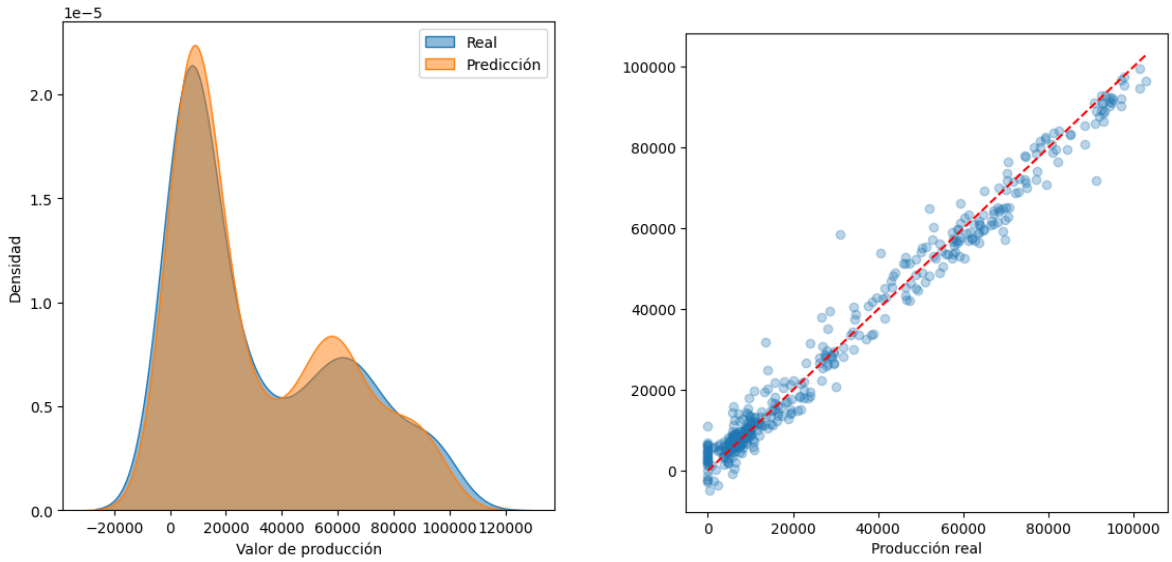
Para evaluar la capacidad de generalización de nuestro modelo de red neuronal para clasificar si una celda contiene o no una planta fuera de la región de entrenamiento, construimos un conjunto de datos con plantas siderúrgicas en América del Sur. Para cada planta, se descargó la información geoespacial para el grid de 1x1 km en el que se encontraba, cubriendo un período de tres años. Al alimentar estos datos al modelo, este predijo que aproximadamente el 70 % de las observaciones mensuales correspondían a una planta. Un análisis más profundo reveló que en todos los grids evaluados,

al menos el 50 % de las observaciones mensuales fueron clasificadas positivamente, lo que nos permite concluir que el modelo predijo correctamente la existencia de una planta en todas las ubicaciones de la prueba.

4.2. Predicción de producción

Una vez que se predijo la ubicación de las plantas, continuamos con la estimación de la producción, un problema de regresión que abordamos a nivel de cada celda de la cuadrícula donde ya se había identificado la existencia de una planta. Para identificar el modelo más adecuado, entrenamos y comparamos varios algoritmos basados en árboles utilizando nuestros datos geoespaciales. De los modelos evaluados, el XGBoost demostró el mejor desempeño, logrando un R^2 de 0.92. Un análisis más detallado de los modelos entrenados y su comparación se presenta en el Apéndice B.

La Figura 5 compara la producción de acero crudo real y la predicha a nivel de grid. El gráfico de densidad (panel a) muestra una similitud notable entre las distribuciones real y predicha, lo que indica que el modelo reproduce de manera efectiva la tendencia central, la dispersión y la multimodalidad de los datos. Sin embargo, se observa una ligera sobrestimación en el segundo pico de la distribución (aproximadamente en el rango de 60k a 70k toneladas), donde la densidad de las predicciones es ligeramente superior a la real. Además, aunque el pico principal (10k–20k) está bien ajustado, las predicciones parecen un poco más concentradas que los datos reales. La alta correlación entre los valores predichos y los reales se confirma en el panel a con el gráfico de dispersión, donde la nube de puntos se alinea estrechamente con la diagonal, lo que indica un desempeño sólido del modelo. Si bien se observa una ligera dispersión en los valores de producción bajos y medios, no se evidencia un sesgo sistemático fuerte, aunque la ligera sobrestimación en valores altos observada en la comparación de densidades también se refleja en este gráfico. En conclusión, el modelo demuestra un desempeño muy bueno al capturar la forma global de la distribución de la producción y la relación lineal entre los valores, con la principal área de mejora enfocada en corregir la ligera sobrestimación en el rango de producción más alto.



(a) Densidad de producción real y predicha

(b) Dispersión de valores reales y predichos

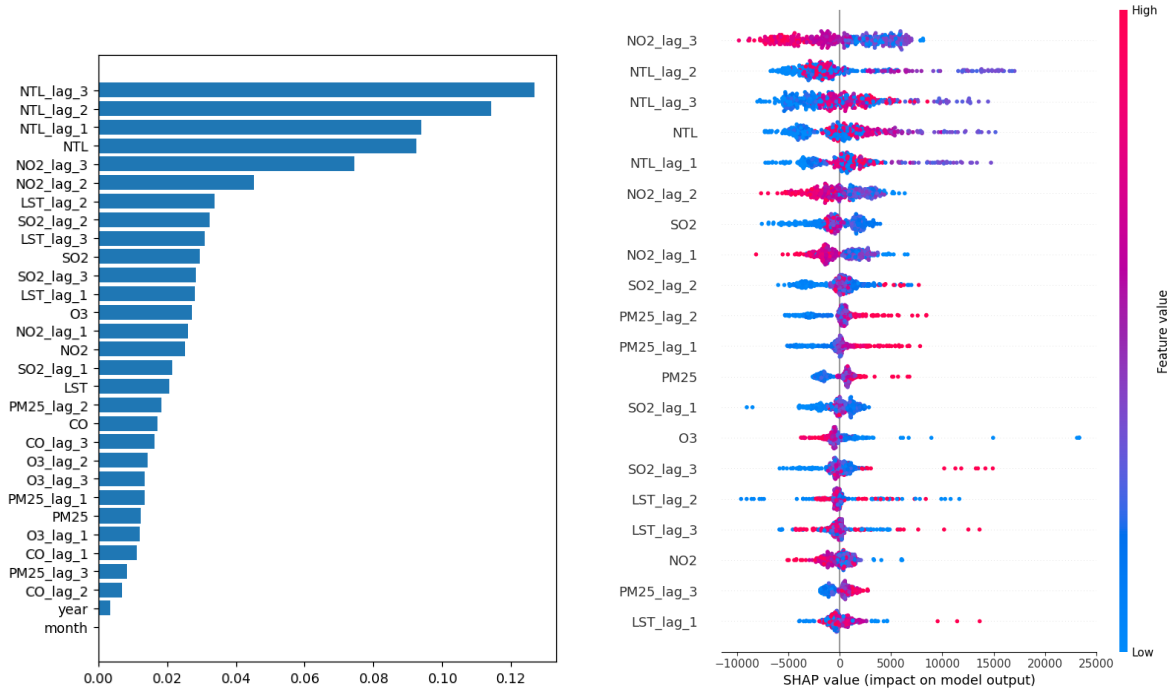
Figura 5: Valores reales y predichos

Nota: Esta figura muestra como se ajusta las predicciones del modelo a los valores reales. El panel (a) muestra el ajuste en las densidades, mientras que el panel (b) muestra el ajuste entre la cantidad de producción real y la cantidad de producción predicho.

Para interpretar los resultados del modelo de regresión XGBoost, nos basamos en un análisis detallado de la importancia de las variables, utilizando los gráficos en la Figura 6 (importancia de variables según ganancia promedio en la métrica de pérdida en el panel 6a y valores SHAP en el panel 6b). El ranking de importancia revela que los principales predictores de la producción de acero

crudo son el NO_2 con rezago de tres meses y la luz nocturna (NTL) en sus formas contemporáneas y rezagadas. Este hallazgo sugiere que la actividad económica a través de la luz nocturna y la dinámica de las emisiones de NO_2) son los factores más determinantes para la predicción de la producción.

El análisis de la dispersión y el color de los puntos SHAP proporciona una visión más profunda del impacto de cada variable. Como era de esperar, los valores altos de NTL (representados por el color rojo) se asocian consistentemente con predicciones de mayor producción, mientras que los valores bajos se asocian con predicciones de menor producción. Este patrón intuitivo refuerza la idea de que la NTL es un proxy robusto de la actividad industrial. Por otro lado, los valores altos de NO_2 (y en menor medida, SO_2 y $PM2{,}5$) con rezagos tienden a ejercer un efecto negativo en la predicción, lo que podría reflejar que altos niveles de contaminación anteceden a períodos de menor producción. En contraste, variables como la temperatura de la superficie terrestre (LST) y O_3 tienen una influencia menor en la predicción.



(a) Importancia de variables

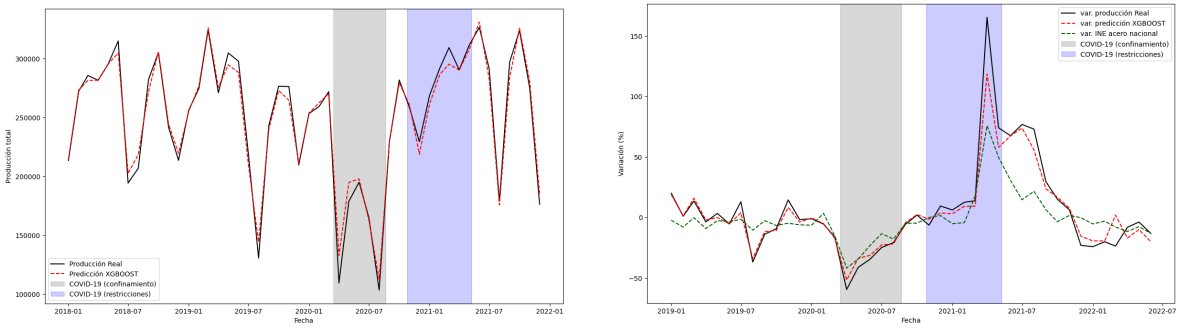
(b) Valor SHAP

Figura 6: Importancia de las variables en el modelo de regresión

Nota: En el panel (a), se presenta la importancia de las variables del modelo XGBoost, medida por la ganancia promedio en la métrica de pérdida. Esta métrica cuantifica cuán útil fue cada variable para reducir la función de pérdida del modelo durante el entrenamiento. En el panel (b), el diagrama de resumen de los valores SHAP nos permite una interpretación más profunda. Este gráfico muestra, para cada observación, la contribución de cada variable a la predicción final. Los valores SHAP nos dan una visión más robusta del impacto real de cada variable en las predicciones del modelo, a diferencia de la métrica de ganancia que se limita a los nodos de decisión.

Para evaluar el comportamiento del modelo frente a eventos clave, como la pandemia del COVID-19, y determinar si nuestra metodología es capaz de capturar fluctuaciones significativas en la actividad industrial, realizamos una validación visual de nuestras predicciones para la producción agregada de las plantas en el País Vasco. La Figura 7 ilustra el impacto de la pandemia en la producción, destacando dos hitos. El primero, el confinamiento nacional estricto entre marzo y junio de 2020, se muestra en el área gris sombreada y generó una fuerte caída en la producción real de acero, una tendencia que nuestro modelo predijo de manera precisa. Posteriormente, el periodo de restricciones continuadas entre octubre de 2020 y mayo de 2021, sombreado en azul, muestra una actividad más estable que también fue bien capturada por las predicciones. Además de estos eventos, la serie predicha también se ajusta a las caídas estacionales de producción que ocurren en agosto debido a periodos de vacaciones. En general, las predicciones de nuestro modelo se ajustan muy bien a las tendencias reales, con solo una ligera subestimación en los picos de producción. Para contextualizar, el panel (b) de la figura compara

la variación interanual de nuestra serie con la del Índice de Producción Industrial (IPI) del INE, que incluye la producción nacional de acero. A pesar de que el IPI abarca más plantas (incluyendo una importante planta de altos hornos), se observa una dinámica y tendencias muy similares entre ambas series, lo que refuerza la validez de nuestra metodología.



(a) Cantidad de producción de acero (toneladas)

(b) Variación interanual

Figura 7: Producción real vs predicha (2018-2022)

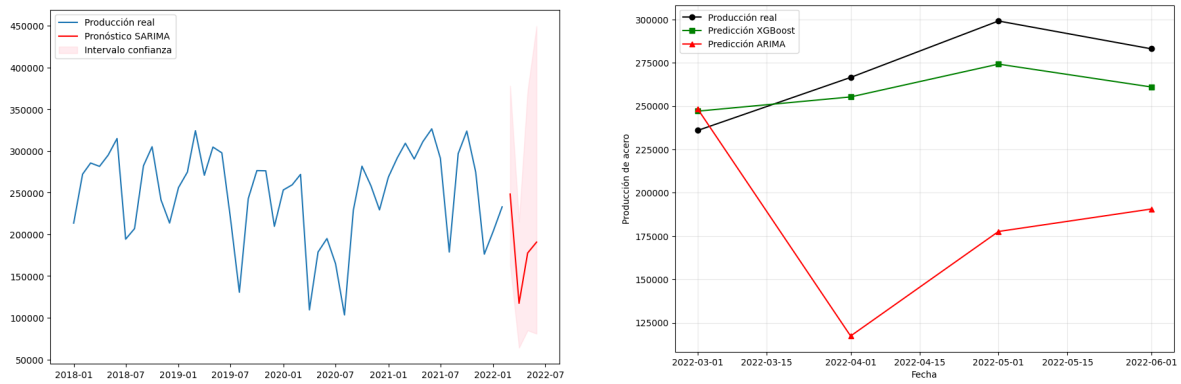
Nota: En el panel (a), se compara la serie de tiempo de la producción real de acero crudo con las predicciones generadas por nuestro modelo XGBoost. El panel (b) muestra la variación interanual de la producción predicha y real, junto con la variación del Índice de Producción Industrial (IPI) del INE, el cual sirve como referencia a nivel nacional en la producicón de acero crudo. Las áreas sombreadas en ambos paneles indican dos periodos clave de la pandemia de COVID-19 en España. El área gris corresponde al confinamiento estricto (marzo-junio 2020), que tuvo un fuerte impacto en la producción, mientras que el área azulada señala el periodo de restricciones continuadas (octubre 2020-mayo 2021), que, a diferencia del primero, no tuvo un impacto significativo en la actividad industrial del sector.

4.3. Capacidad de predicción a futuro

Para evaluar el poder predictivo de nuestro modelo XGBoost, lo contrastamos con un enfoque de series de tiempo más tradicional. Para ello, estimamos un modelo SARIMA, cuyos parámetros fueron seleccionados utilizando el criterio de Akaike (AIC), lo que garantiza la mejor combinación de estos para la serie de producción. Los modelos SARIMA son una extensión de los modelos ARIMA y son aún hoy ampliamente utilizados en finanzas y economía por bancos e inversores para predecir el comportamiento de series temporales. Su popularidad se debe a su capacidad para capturar patrones como tendencias y estacionalidad basándose exclusivamente en los datos históricos de la propia variable.

El Panel (a) de la Figura 8 ilustra el pronóstico del modelo SARIMA para el período de marzo a junio de 2022. Aunque estos modelos son robustos, su principal limitación es que su poder predictivo se desvanece a medida que el horizonte de pronóstico se extiende. Esto se refleja en los amplios intervalos de confianza que se generan, indicando una creciente incertidumbre en las predicciones. Esta incertidumbre se debe a que un modelo SARIMA no incorpora información externa (como variables geoespaciales), lo que lo hace vulnerable a fluctuaciones inesperadas en la serie de tiempo.

El Panel (b) de la misma figura contrasta el desempeño predictivo del modelo SARIMA con el de nuestro modelo XGBoost. Si bien el SARIMA mostró una ligera ventaja en el primer mes de pronóstico (marzo de 2022), el modelo XGBoost fue mucho más preciso en la estimación de la producción en los tres meses siguientes. El modelo XGBoost logró capturar de forma más robusta tanto la magnitud como la dinámica de la producción real. Estos resultados demuestran que una metodología que utiliza indicadores geoespaciales como predictores es más robusta que un enfoque univariado tradicional para la predicción de la actividad industrial.



(a) Pronóstico SARIMA

(b) Comparación de predicciones

Figura 8: Contraste de predicciones XGBoost vs SARIMA

Nota: Comparación de Predicciones. El panel (a) presenta las series de producción real de acero y las predicciones generadas por un modelo SARIMA (1,1,1)(1,1,0,12), cuyos parámetros fueron seleccionados en base al menor valor del Criterio de Información de Akaike (AIC). Al igual que el modelo XGBoost, el SARIMA fue entrenado con datos desde enero de 2018 hasta febrero de 2022. La proyección del SARIMA se realizó para el período de marzo a junio de 2022. El panel (b) compara las predicciones de este modelo SARIMA con las de nuestro modelo XGBoost para el mismo período.

5. Conclusiones

Este trabajo ha demostrado que el uso de datos geoespaciales combinados con técnicas avanzadas de Machine Learning y Deep Learning constituyen una alternativa robusta y escalable para medir la actividad económica con mayor granularidad temporal y espacial. Al centrarnos en la industria siderúrgica española —caracterizada por su bajo nivel de emisiones atmosféricas debido al uso de hornos de arco eléctrico— hemos podido validar que, incluso en contextos de señales ambientales débiles, es posible identificar patrones detectables mediante teledetección y traducirlos en información económica de valor.

Los resultados obtenidos refuerzan la idea de que cada actividad productiva deja una "huella ambiental" medible, donde variables como la luz nocturna (NTL), la temperatura de la superficie terrestre (LST) y las emisiones de NO_2 se consolidan como proxies fundamentales de la localización y el nivel de producción. Nuestros modelos alcanzaron una precisión del 98 % en la identificación de plantas y un coeficiente de determinación superior a 0.9 en la estimación de la producción, lo que evidencia un desempeño notable frente a enfoques tradicionales e incluso frente a otros estudios recientes.

Además, la comparación con métodos de series temporales clásicos, como SARIMA, muestra que los modelos que integran información exógena geoespacial son más precisos y estables en escenarios de incertidumbre, como el experimentado durante la pandemia del COVID-19. Esta capacidad predictiva, casi en tiempo real, abre nuevas posibilidades para complementar los indicadores oficiales —como el IPI— y mejorar la capacidad de respuesta de gobiernos, empresas e investigadores en contextos de crisis o cambios abruptos en la actividad económica.

En síntesis, este estudio no solo aporta evidencia empírica sobre la eficacia de los datos geoespaciales para medir la actividad industrial, sino que también plantea un marco metodológico replicable en otras industrias y regiones. Futuras líneas de investigación podrían explorar la integración de nuevas fuentes satelitales de mayor resolución, la extensión del enfoque a sectores con diferentes perfiles de emisiones, y la combinación con técnicas de Big Data para fortalecer aún más la capacidad de monitoreo económico en tiempo real.

Referencias

Chen, X. and Nordhaus, W. D. (2019). Viirs nighttime lights in the estimation of cross-sectional and time-series gdp. *Remote Sensing*, 11(9):1057.

Deeter, M. N., Edwards, D. P., Gille, J. C., Drummond, J. R., Worden, H. M., Francis, G. E., Francis, J. P., and Worden, H. M. (2019). The mopitt instrument. *Atmospheric Measurement Techniques*, 12:535–554.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028.

Levelt, P. F., van der A, R. J., Brinksma, E. J., van den Oord, G. H. J., R. Eskes, H., Veefkind, J. P., De Haan, J. F., Veefkind, J. P., and Spinhoven, J. (2006). The ozone monitoring instrument (omi) on aura: A new measurement for atmospheric composition. *Atmospheric Measurement Techniques*, 9:2463–2480.

Levin, N., Kyba, C. C. M., Zhang, Q., Sánchez de Miguel, A., Román, M. O., Li, X., Portnov, B. A., Molthan, A. L., Jechow, A., Miller, S. D., Wang, Z., Shrestha, R. M., and Elvidge, C. D. (2020). Remote sensing of night lights: A review and an outlook for the future. *Remote Sensing of Environment*, 237:111443.

Veefkind, J. P., de Haan, J. F., Vandaele, A. C., Levelt, P. F., and Vandaele, A. C. (2012). A multi-platform intercomparison of tropospheric no2 observations from satellites. *Atmospheric Chemistry and Physics*, 12:1495–1513.

Wei, J., Li, Z., Wang, J., Li, C., Gupta, P., and Cribb, M. (2023). Ground-level gaseous pollutants (no_2, so_2, and co) in china: daily seamless mapping and spatiotemporal variations. *Atmospheric Chemistry and Physics*, 23(2):1421–1440.

Yang, A., Ai, J., and Arkolakis, C. (2025). A geospatial approach to measuring economic activity. Technical Report Working Paper No. 33619, National Bureau of Economic Research.

Apéndices

A. Variables incorporadas

Las variables incorporadas en el estudio son:

- grid_id: Identificador único del cuadrante espacial (grid) donde se agrupan las observaciones. Cada grid_id corresponde a un polígono geográfico definido en el sistema de grillas de 1x1 km dentro del área del País Vasco. Tenemos 7.718 grids.
- year: Año de la observación o medición. Trabajaremos una serie desde el 2018 a 2022.
- month: Mes de la observación (formato numérico, 1 = enero, 12 = diciembre).
- NO_2: Concentración promedio mensual de dióxido de nitrógeno en el grid correspondiente. Unidad en mol/m^2.
- CO: Concentración promedio mensual de monóxido de carbono. Unidad en mol/m^2.
- SO_2: Concentración promedio mensual de dióxido de azufre. Unidad en mol/m^2.
- O_3: Concentración promedio mensual de ozono troposférico. Unidad en mol/m^2.
- NTL: Nighttime Lights: nivel promedio de luminancia nocturna (derivado de imágenes satelitales). Indica actividad humana e infraestructura.
- LST: Land Surface Temperature: temperatura de la superficie terrestre (en °K). Para interpretar se debe multiplicar x0.02. Y con eso se obtiene el °K. Luego, para ver en °C se debe restar -273.
- $PM_{2,5}$: Concentración estimada de material particulado fino. Unidad en mg/m^2.
- tiene_planta: Indicador binario: 1 si existe al menos una planta siderúrgica dentro del grid, 0 si no.
- produccion: Producción industrial total (suma) correspondiente al grid en ese mes y año. Calculada agregando los valores de todas las plantas ubicadas dentro del grid_id. Si es NaN, no hubo plantas reportadas en ese mes/grid.

Cuadro 1: Estadísticos descriptivos de variables ambientales según presencia de planta en el grid

Variable	Con planta	Count	Mean	Median	Std	Min	Max
NO2	0	433884	0.00007	0.00007	0.00001	0.00004	0.00011
	1	996	0.00009	0.00008	0.00002	0.00004	0.00024
CO	0	433884	0.02946	0.02956	0.00346	0.01821	0.04248
	1	996	0.03055	0.03061	0.00330	0.01940	0.04102
O3	0	433884	0.14273	0.14341	0.01268	0.10958	0.18184
	1	996	0.14191	0.14002	0.01142	0.11166	0.17979
SO2	0	433884	0.00091	0.00067	0.00081	0.00000	0.02450
	1	996	0.00093	0.00068	0.00075	0.00012	0.00730
PM25	0	433884	0.00616	0.00577	0.00211	0.00147	0.02650
	1	996	0.00683	0.00609	0.00308	0.00154	0.02640
NTL	0	433884	3.76626	0.85250	11.75724	0.00000	464.07248
	1	996	49.41456	49.15375	39.56972	1.73250	298.75000
LST	0	433884	14533.27735	14535.70833	365.26013	0.00000	15530.00000
	1	996	14773.51230	14761.75000	438.07444	13774.00000	15912.00000

B. Contraste de modelos de clasificación

En este apéndice, presentamos los principales resultados de los modelos que probamos para clasificar si un grid contiene o no una planta siderúrgica. Para esta tarea, evaluamos el desempeño de tres modelos basados en árboles y redes neuronales: Random Forest y XGBoost (cuyos hiperparámetros se optimizaron mediante una búsqueda de grilla o GridSearch), y un modelo de Redes Neuronales. En nuestra evaluación, nos enfocamos en el recall, ya que en un escenario de producción, la prioridad es que el modelo identifique correctamente todas las celdas que contienen una planta (minimizar los falsos negativos), lo que permite que un analista revise una lista completa de ubicaciones candidatas. En este sentido, el modelo que mostró el mejor desempeño fue el de Redes Neuronales, logrando un recall de 0.99.

Cuadro 2: Resultados de las métricas de clasificación, con casos sintéticos en conjunto Test

Modelo	**Clase**	**Precisión**	**Recall**	**F1-score**
Red neuronal	0	1.00	0.98	0.99
	1	0.95	1.00	0.97
	Accuracy			0.98
	Macro avg	0.97	0.99	0.98
	Weighted avg	0.98	0.99	0.99
Random Forest	0	0.98	0.85	0.91
	1	0.87	0.98	0.92
	Accuracy			0.92
	Macro avg	0.92	0.92	0.92
	Weighted avg	0.92	0.92	0.92
XGBoost	0	0.98	0.96	0.97
	1	0.88	0.92	0.90
	Accuracy			0.95
	Macro avg	0.93	0.94	0.93
	Weighted avg	0.95	0.95	0.95

Cuadro 3: Resultados de las métricas de clasificación, sin casos sintéticos en conjunto Test

Modelo	**Clase**	**Precisión**	**Recall**	**F1-score**
Red neuronal	0	0.99	0.99	0.99
	1	0.11	0.61	0.19
	Accuracy			0.99
	Macro avg	0.56	0.80	0.59
	Weighted avg	1.00	0.99	1.00
Random Forest	0	1.00	1.00	1.00
	1	0.35	0.26	0.30
	Accuracy			1.00
	Macro avg	0.68	0.63	0.65
	Weighted avg	1.00	1.00	1.00
XGBoost	0	1.00	0.99	0.99
	1	0.11	0.57	0.19
	Accuracy			0.99
	Macro avg	0.56	0.78	0.59
	Weighted avg	1.00	0.99	0.99

C. Contraste de modelos de regresión

Para la tarea de estimación de la producción, se evaluó el desempeño de varios modelos de regresión, con el fin de identificar el algoritmo que mejor se ajustara a las relaciones no lineales entre los datos geoespaciales y los niveles de producción de acero. De manera consistente con las expectativas para este tipo de problemas, los modelos lineales como ElasticNet, Ridge y Lasso mostraron un rendimiento significativamente inferior, con coeficientes de determinación (R^2) que no superaban el 0.41, lo que confirma que el problema no puede ser modelado con una relación lineal simple.

Por otro lado, los modelos basados en árboles ensamblados demostraron una capacidad superior para capturar la complejidad de los datos. En este grupo, el modelo XGBoost fue el de mejor desempeño, con un R^2 de 0.91 y un RMSE de 8.877, seguido de cerca por Gradient Boosting ($R^2 \approx 0{,}90$) y Random Forest ($R^2 \approx 0{,}89$). Estos resultados validan la elección de un enfoque no lineal para la predicción. Para maximizar la precisión de cada modelo, se realizó una búsqueda de hiperparámetros exhaustiva utilizando GridSearch para encontrar la combinación óptima de parámetros. Este análisis comparativo nos permitió seleccionar a XGBoost como el modelo principal de nuestro estudio, y la solidez de los resultados de otros modelos basados en árboles refuerza la fiabilidad de nuestro enfoque.

Cuadro 4: Resultados de las métricas de regresión en el conjunto Test

Modelo	**RMSE**	$\mathbf{R^2}$
XGBoost	8877.47	0.909
Gradient Boosting	9237.34	0.902
Random Forest	9832.09	0.889
LightGBM	11566.19	0.846
ElasticNet	22605.07	0.413
Ridge	23034.36	0.391
Regresión Lineal	23630.26	0.359
Lasso	23638.77	0.358
Kernel Ridge	44944.92	-1.320