

Artistic Interventions risk in Urban Spaces predicted through Machine and Statistical Learning [v4.0]

Carlos Cuevas-Urenda^{*1}, David Elal-Olivero¹ & Diego Nascimento¹

¹Departamento de Matemática, Universidad de Atacama, Copiapó, Chile

14 de febrero de 2025

Resumen

Artistic expressions through murals and graffiti paints enrich the environment and function as social media, allowing diverse social groups to transmit their messages, opinions and world-views. Research on the spatial patterns and urban characteristics that influence these artistic interventions is essential to understand how they are created and distributed as a communication tool. However, to the best of our knowledge, there is no scientific analysis on the subject to date in South America. The importance of the subject is directed at the fact that an expression may (dis)add monetary value to a property (and sector) depending on the person or group and materials used. This study focused on developing a risk model given the choice of a wall already intervened/painted, to associate a probability that another artistic expression (graffiti) may happen. It was adopted different Machine Learning and Statistical models that adoptes binary response variable for unbalanced data (adopting hold-out 70-30 %), such as Logistic Regression, Random Forest, Deep Learning, GBM, XGBoost, XRT, DRF, Ensembles and AutoML. The obtained results were related to the northern sector, 1,905,575.3 m², of Copiapó city (ATACAMA - Chile). The dataset contains 351 observations obtained from Google (Maps & Earth). First, it extracted the photographs then pre-processed more detailed features. These features/covariates were (i) Visibility of the wall containing the intervention, (ii) Type of intervened space, and (iii) Traffic frequency. The best model was the Random Forest, using only as explanatory variables the longitude and latitude and the visibility of the intervention, with a specificity of 98.33 %. Spatial factors were the most important in explaining a graffiti intervention whose objective is the maximization of visibility of artistic expression, so it is necessary to incorporate other spatial factors in future analyses.

Keywords: Automatic Machine Learning; Binary regression; Imbalanced Data,

^{*}Corresponding author. Email: carlos.cuevas@postgrados.uda.cl

1. Introducción

Las intervenciones artísticas en espacios público/urbano han sido una forma de expresión cultural desde tiempos ancestrales como forma de comunicación. En la macrozona norte de Chile, los pueblos originarios, por ejemplo en la región de la cordillera de Arica dejaban mensajes en el cerro para que los suyos no se asustaran con la cultura o población de la ciudad grande, describiendo que el promedio de estatura eran muchísimo mayor (hasta hablando que eran gigantes) y indicaciones de alto ruido sonoro. En el Alero de Taira, las representaciones pictografiadas de llamas están asociadas a ceremonias destinadas a propiciar la fertilidad de estos animales, esenciales para la subsistencia de las comunidades prehispánicas. Estas manifestaciones no solo embellecían el entorno, registro de conocimientos y expresión de creencias religiosas y rituales.

Estas prácticas artísticas ancestrales reflejan la necesidad humana de comunicarse, dejar huellas de su existencia y transmitir conocimientos y valores (y que pueden perdurar a través de generaciones). En la actualidad, murales y grafitis han tomado este rol en las ciudades, funcionando como herramientas de expresión social y cultural. Su impacto, sin embargo, es diverso. Mientras algunos fomentan la revitalización de espacios y potencian la identidad local, otros pueden ser percibidos como contaminación visual o vandalismo cuando no se integran adecuadamente al entorno. Estas intervenciones influyen en la percepción de los espacios urbanos e incluso en su valorización económica.

Este estudio se centra en el análisis de riesgo bajo la óptica de un morador de tener intervenciones artísticas en la ciudad de Copiapó, Chile, diferenciando las entre murales y grafitis. Para ello, se emplearán modelos de aprendizaje automático y estadístico con el fin de predecir la probabilidad de su aparición y comprender los factores que influyen en su distribución. Los algoritmos adoptados son de clasificación binaria, conteniendo un nivel de complejidad mediano una vez que la problemática presenta datos desbalanceados, que impactan directamente en las métricas de desempeño y comparación entre los modelos.

El enfoque adoptado busca fortalecer la literatura científica sobre intervenciones artísticas en Sudamérica por una relación directa con la seguridad, ofreciendo un marco cuantitativo para su estudio y evaluación en impactos vía modelos de machine learning. A partir del análisis de factores espaciales y urbanos, este trabajo contribuirá a una mejor comprensión de las dinámicas detrás de estas expresiones, proporcionando herramientas para su gestión y análisis de impacto, facilitando estrategias para su integración, regulación y valoración en distintos contextos urbanos.

2. Revisión Bibliográfica

Las intervenciones artísticas en espacios públicos han sido una manifestación cultural y un medio de comunicación desde tiempos ancestrales. En la macrozona norte de Chile, los pueblos originarios empleaban el arte rupestre para transmitir información sobre su entorno y las dinámicas sociales. Desde la carretera internacional Arica-Tambo Quemado, se registran evidencias de mensajes grabados en cerros, utilizados para advertir sobre las diferencias culturales y físicas entre comunidades, incluyendo descripciones sobre la estatura y los niveles de ruido en centros urbanos prehispánicos. El arte rupestre atacameño destaca por su riqueza iconográfica, compuesta por petroglifos y pictografías con representaciones antropomorfas, zoomorfas y geométricas, diferenciadas por sus técnicas de ejecución y paleta cromática. Un ejemplo representativo está en la Figura 1(a), donde se muestra el conjunto de geoglifos conocidos como “Hombres Grandes” según la cosmovisión como transpaso de información a las tribus originarias (en el Valle de Lluta, ubicado en la Región de Arica y Parinacota). Estas figuras monumentales, atribuidas a culturas andinas del Período Intermedio Tardío (800 d.C.) hasta el siglo XV d.C., constituyen una manifestación artística prehispánica con un fuerte carácter simbólico y comunicativo.



(a) Geoglifos “Hombres Grandes”, ubicados en el Valle de Lluta, región de Arica y Parinacota. Intervención prehispánica en el cerro, atribuida a culturas andinas entre el Período Intermedio Tardío (800 d.C.), hasta el Siglo XV d.C.



(b) Arte rupestre prehispánico (arriba) y mural “Atacama Desierto Vivo” (abajo), ubicado en el centro de la ciudad de Copiapó, Región de Atacama. La parte inferior es una intervención del artista local Christian Rivadeneira (año 2023).

Figura 1: Evolución de la expresión artística rupestre en el Desierto de Atacama.

En la actualidad, murales y grafitis desempeñan un papel clave en los entornos urbanos, funcionando como herramientas de expresión social, cultural. Su impacto, sin embargo, es diverso: mientras algunos contribuyen a la resignificación y revitalización del espacio público, otros pueden ser percibidos como intervenciones disruptivas o incluso actos de vandalismo cuando no se integran adecuadamente en el contexto urbano. Estas expresiones artísticas no solo transforman la percepción de la cultura en el entorno, sino que también inciden en su valor socioeconómico. Un ejemplo significativo es la Figura 1(b), que muestra un mural en la oficina del Servicio Nacional de Turismo (Sernatur) de la ciudad de Copiapó, donde se representan diferentes intervenciones artísticas. La

primera un arte rupestre copiapino con una escena de pastoreo con figuras humanas, auquénidos y símbolos circulares asociados a la fertilidad, reflejando la importancia de estos camélidos en la economía y cosmovisión andina. La parte inferior del muro contiene el mural “Atacama Desierto Vivo... un homenaje a la mujer atacameña en la historia”, obra del artista local Christian Rivadeneira Geraldo, la cual destaca la belleza del desierto florido y la relevancia de la mujer en la historia nacional.

Si bien existen múltiples estudios que analizan el impacto de las intervenciones artísticas en el espacio público, este trabajo busca diferenciarse al centrarse específicamente en desarrollar un modelo probabilístico que cuantifica el riesgo de una intervención artística en un muro de la zona norte de la ciudad en estudio, aplicando un enfoque de inteligencia artificial para identificar patrones en la distribución y características de estas expresiones. En este contexto, uno de los principales desafíos metodológicos radica en el tratamiento de datos binarios desbalanceados, lo que requiere estrategias adecuadas para garantizar un análisis preciso y representativo.

2.1. Trabajos Relacionados - Datos Binario desbalanceado

El análisis de las intervenciones artísticas en el espacio público presenta desafíos metodológicos significativos, especialmente cuando se aborda desde una perspectiva estadística. Uno de estos desafíos radica en el desbalance de clases en los datos, donde una categoría está representada por un número considerablemente mayor de observaciones que la otra. Este fenómeno es común en estudios de clasificación binaria, como la detección de anomalías en las áreas financieras [CITA], médica [CITA] y, en nuestro caso, para descripción de imágenes urbanas o la segmentación de estilos artísticos, donde algunas manifestaciones son más frecuentes que otras. Es importante destacar que los modelos de clasificación binaria podrán ser hard- or soft-classifiers.

[HARD-CLASSIFIER VS SOFT-]

El desbalance en los datos puede influir negativamente en la precisión del análisis, ya que los modelos de clasificación tienden a favorecer la clase mayoritaria, reduciendo su capacidad para identificar correctamente los casos menos frecuentes pero de alto interés analítico. Las estrategias comunes para contornear eso es data augmentation (over- and under-sampling) y seleccionando un umbral ideal.

Métodos tradicionales de clasificación, como la regresión logística con funciones de enlace simétricas (logit o probit), pueden generar sesgos en la estimación de parámetros y en la asignación de probabilidades cuando se enfrentan a un conjunto de datos desbalanceado. Para mitigar este efecto, se han desarrollado estrategias como el sobremuestreo y submuestreo de las clases, la asignación de pesos diferenciados en los modelos y el uso de funciones de enlace asimétricas [LETICIA et al], que permiten ajustar la modelación de las probabilidades según la distribución real de los datos.

En este trabajo, se considera la implementación de estos enfoques con el objetivo de mejorar la precisión en la clasificación de las intervenciones artísticas en Copiapó, permitiendo una interpretación más robusta de los patrones urbanos que emergen a partir de estos datos.

3. Material

El presente estudio se centró en la ciudad de Copiapó, Chile, específicamente en el sector norte, que abarca un área de 1.905.575,3 m² (Figura 2). La selección de esta zona se debió principalmente a su cercanía con la Universidad de Atacama, lo que facilitó la geolocalización, el acceso a imágenes de referencia y la supervisión del proceso de recolección de datos.

Además, esta área forma parte del casco histórico de la ciudad, siendo un punto clave en el desarrollo urbano de Copiapó. Este sector concentra algunas de las edificaciones más antiguas, además de calles y espacios públicos que han servido históricamente como centros de interacción social y cultural. En este contexto, las intervenciones artísticas urbanas, como los murales y grafitis, se han integrado en la dinámica del espacio público, convirtiéndose en un reflejo de la identidad y la evolución sociocultural de la ciudad.



Figura 2: Delimitación geográfica de la zona de estudio en el sector norte de Copiapó, Chile. El área resaltada en amarillo y rojo representa la extensión analizada, caracterizada por su riqueza en intervenciones artísticas y relevancia en la configuración de la cultura de la ciudad.

Existieron 2 criterios para la inclusión de intervenciones en la creación de la base de datos:

- La presencia de un mural o graffiti en un espacio urbano accesible.
- La disponibilidad de imágenes en Google Maps o Google Earth que permitieran su georreferenciación y análisis visual.

La elección de este sector no solo responde a su accesibilidad y relevancia histórica, sino también a la falta de estudios cuantitativos sobre intervenciones artísticas urbanas en ciudades de Chile.

Para la recopilación sistemática de la información, se diseñó un formulario estructurado en Google Forms que permitió registrar las características clave de cada intervención. Este formulario fue construido con base en criterios espaciales, visuales y contextuales, permitiendo garantizar la consistencia en la recopilación de datos.

A continuación, se presenta un resumen de las variables categorizadas en la encuesta utilizada para la recolección de datos:

Tabla 1: Resumen de variables categorizadas según la encuesta.

Categoría	VARIABLES	Descripción
Ubicación	Latitud	Coordenadas geográficas en el eje norte-sur.
	Longitud	Coordenadas geográficas en el eje este-oeste.
	Altitud	Altura sobre el nivel del mar.
	Cercanía	Escuela, Plaza, Estación de transporte, Otro..
Accesibilidad	Nivel de acceso	Público, Semi-privado, Restringido.
	Visibilidad	Alta, Media, Baja.
	Tránsito-Tráfico	Alto, Medio, Bajo.
	Tipo de espacio	Muro privado, Muro público, Otro.
Expresión artística	Formato	Graffiti o Mural/Otro.
	Tamaño	Monumental, Grande, Mediano o Pequeño.
	Temática	Mitología, Naturaleza, Espiritualidad o religión, Resistencia social, Política, Cultura pop, Firma de artista, Letras, Sexual, Arte Abstracto, Amor, Fútbol, Otro.

El uso de un formulario estructurado permitió garantizar la consistencia en la recopilación de datos, facilitando su posterior análisis mediante herramientas estadísticas y de aprendizaje automático.

La recopilación de datos se realizó a partir de imágenes satelitales y en vista de calle disponibles facilitadas por herramientas de Google. Este proceso incluyó las siguientes etapas:

Tabla 2: Resumen del proceso de recolección de datos y herramientas utilizadas.

Etapa	Descripción	Herramientas
Identificación de muros intervenidos	Se realizó un recorrido digital por el área de estudio para detectar intervenciones visibles en muros y espacios urbanos.	Google Maps & Google Street View
Geolocalización de cada intervención	Se registraron las coordenadas de cada punto utilizando herramientas de geolocalización y mapeo digital.	Google Earth (.kml)
Análisis visual y clasificación	Se observó cada mural o graffiti y se completó el formulario con sus características espaciales, visuales y contextuales.	Google Forms & Google Sheets (.csv)
Registro fotográfico y metadatos	Se almacenaron imágenes de referencia junto con información relevante sobre el contexto urbano y las condiciones de la intervención.	Capturas de pantalla

Este método permitió la creación de un conjunto de datos homogéneo y estandarizado, evitando sesgos que pudieran derivarse de observaciones presenciales.

Tras la recolección inicial, la información fue consolidada en una base de datos estructurada en archivos con extensión .kml & .csv, compuesta por 351 observaciones. Antes de proceder al análisis, se llevaron a cabo diversas tareas de limpieza y depuración de datos, tales como:

- Eliminación de registros duplicados o inconsistentes.
- Normalización de valores categóricos para garantizar coherencia en las clasificaciones.
- Revisión de datos faltantes y ajuste de etiquetas en variables clave.

Como resultado, se obtuvo un dataset final listo para el análisis cuantitativo, con variables organizadas de manera clara y coherente para su procesamiento.

Distribución Geográfica de la data

A la izquierda de la Figura 3 se muestra un mapa con las 351 intervenciones encontradas en el sector norte de Copiapó, destacando una alta concentración en zonas céntricas y de alto tránsito. Este patrón sugiere que los artistas priorizan ubicaciones visibles para maximizar el impacto de su trabajo. A la derecha, el mapa de calor refuerza esta tendencia, resaltando las áreas con mayor actividad artística en sectores urbanos accesibles.

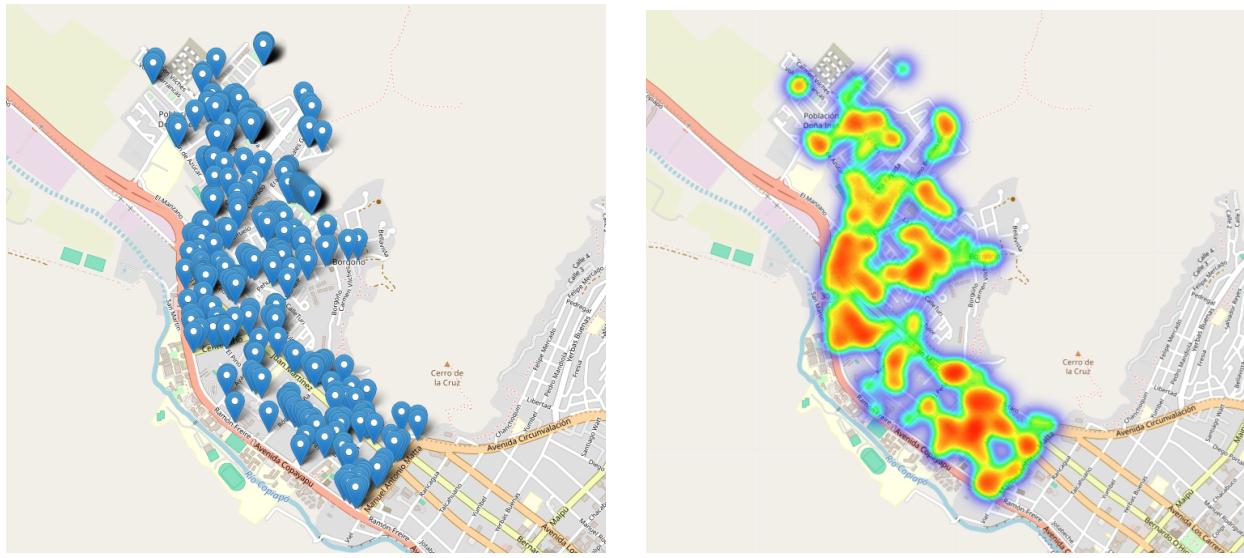


Figura 3: Representación geoespacial de la distribución de intervenciones artísticas encontradas. A la izquierda, se muestra la localización puntual de los graffitis georreferenciados. A la derecha, un mapa de calor visualiza la densidad de estas intervenciones, resaltando en rojo las zonas con mayor presencia artística.

Distribución Espacial de Murales y Grafitis

La Figura 4 presenta un análisis comparativo de la distribución de murales y graffitis en el sector estudiado de la ciudad de Copiapó.

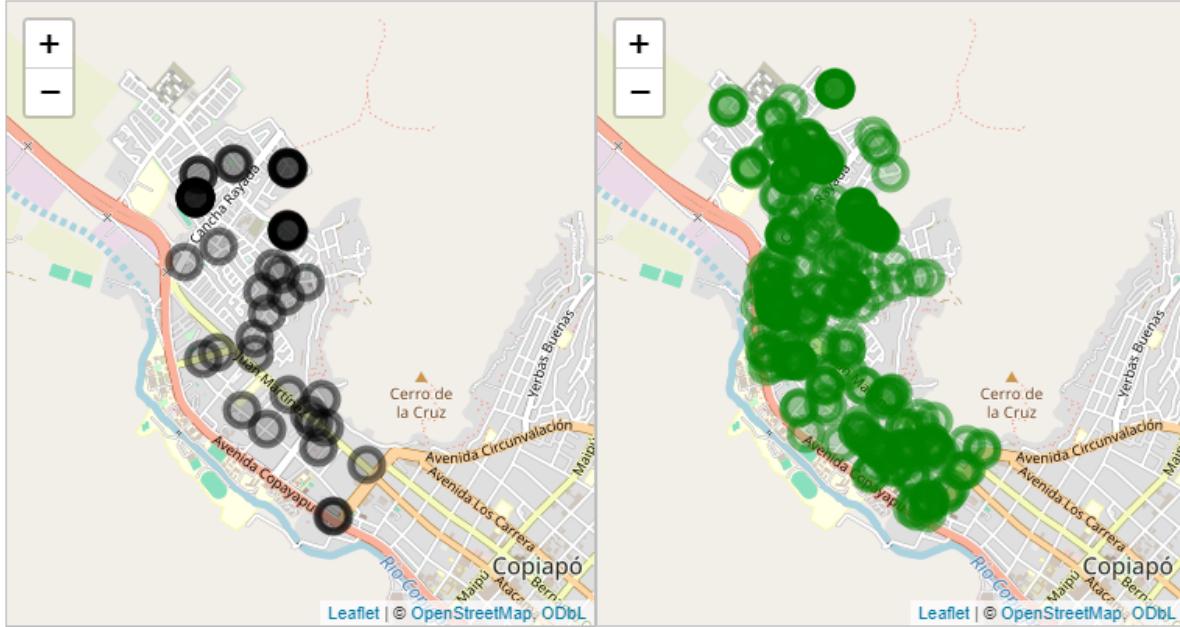


Figura 4: Mapa de distribución de murales y graffitis. A la izquierda (en negro) la ubicación de los murales, mientras que a la derecha (en verde) la distribución y densidad de graffitis.

- Los círculos negros representan la ubicación de los 48 murales apreciados, que suelen

estar en lugares estratégicos de alta visibilidad y accesibilidad. Suelen ser intervenciones de gran escala y con mayor integración formal en el paisaje urbano.

- **Los círculos verdes** representan la ubicación de los 303 **grafitis** descubiertos, los cuales se distribuyen de manera más dispersa y con mayor densidad en comparación con los murales.

El patrón espacial observado sugiere que Los murales tienden a ubicarse en espacios más regulados o gestionados, lo que indica que su realización podría estar vinculada a permisos o encargos formales. Mientras que los grafitis aparecen en mayor cantidad y cubren un área más extensa, lo que sugiere un carácter más espontáneo y menos regulado.

Esta segmentación espacial de las intervenciones artísticas urbanas en Copiapó resalta la diferencia en la forma en que los artistas utilizan el entorno urbano, diferenciando entre aquellos que trabajan en proyectos formales y aquellos que intervienen en los espacios de manera independiente.

Características Artísticas

La siguiente tabla presenta la distribución de las intervenciones artísticas en el sector estudiado según su tamaño y tipo de intervención.

Tabla 3: Distribución del tamaño del arte según el tipo de intervención.

Tamaño del Arte	Mural	Graffiti	Total
Monumental	14	16	30
Grande	33	86	119
Mediano	1	94	95
Pequeño	0	107	107
Total	48	303	351

La distribución en la Tabla 3 muestra tendencias interesantes sobre el tamaño de las intervenciones en relación con su tipo:

- En cada grupo de tamaño, los grafitis son más frecuentes que los murales. Esto indica que la mayoría de las intervenciones artísticas corresponden a expresiones espontáneas en lugar de murales planificados.
- La categoría “Pequeño” es la más numerosa con 107 grafitis y sin murales. Esto sugiere que los artistas que realizan grafitis tienden a trabajar en escalas más reducidas, posiblemente debido a restricciones de permisos, tiempo o espacio.

- En las categorías “Grande” y “Monumental”, la proporción de murales es mayor en comparación con las demás, aunque sigue siendo menor que la de los grafitis. Esto refuerza la idea de que los murales requieren planificación y un espacio más adecuado para su ejecución.
- Por último, mientras que los grafitis están presentes en todas las escalas, los murales tienden a ser principalmente de gran tamaño o monumentales, lo que puede reflejar un proceso de diseño más elaborado.

Estos resultados proporcionan información valiosa sobre la naturaleza de las intervenciones artísticas en el sector y permiten inferir patrones sobre cómo se distribuyen espacialmente y qué factores pueden influir en su tamaño y tipo.



(a) Mural homenaje a Pablo Neruda junto a un gato. Artista: Geo2Set.



(b) Mural del Ammonite, animal prehistórico. Artista: Kaz Geoarte.



(c) Mural minimalista y romántico. Artista desconocido.



(d) Graffiti Dinosaurio “King Cuak”. Artista: Claus Vega Correa.



(e) Graffiti abstracto de un conejo. Artista desconocido.



(f) Graffiti “El Barto”, referencia a Los Simpson. Artista desconocido.

Figura 5: Muestra de murales y grafitis encontrados en el sector de estudio y sus autores.

Por otro lado, la siguiente tabla presenta la distribución según su tipo de acceso de los artistas.

Tabla 4: Distribución del tipo de intervención según el acceso al espacio.

Intervención	Público	Semi-privado	Restringido	Total
Mural	47	1	0	48
Graffiti	278	23	2	303
Total	325	24	2	351

La Tabla 4 permite extraer varios patrones interesantes sobre la distribución de los murales y grafitis en función del tipo de acceso al espacio:

- Se observa que los grafitis (278 en espacios públicos, 23 en semi-privados y 2 en restringidos) superan ampliamente a los murales en cada categoría, lo que indica que esta forma de intervención artística es más común en el entorno urbano.
- La mayoría de los murales y grafitis se encuentran en lugares de acceso público, lo que sugiere que los artistas buscan maximizar la visibilidad de su obra.
- Solo se registró 1 mural en un espacio semi-privado y ninguno en restringidos, lo que sugiere que estos suelen ser intervenciones más planificadas y aprobadas en entornos accesibles.
- A diferencia de los murales, los grafitis están presentes en espacios de difícil acceso, lo que sugiere que estas intervenciones se realizan sin autorización previa en muchos casos.
- Aunque la mayoría de los grafitis están en espacios públicos, hay 23 casos en semi-privados, lo que está asociado a zonas con menor vigilancia o acceso intermedio.

Estos hallazgos refuerzan la idea de que los murales tienden a ser intervenciones autorizadas en lugares accesibles, mientras que los grafitis aparecen con mayor frecuencia en diferentes tipos de espacios, incluyendo aquellos con restricciones de acceso.

4. Modelado de Clasificación Binaria

En este estudio, la variable respuesta Y representa un muro con mural o graffiti, o sea:

$$Y_i = \begin{cases} 1, & \text{Si el muro presenta un intervención con graffiti.} \\ 0, & \text{Si el muro presenta un mural u otro similar.} \end{cases} \quad (1)$$

Dado que se trata de una variable dicotómica independiente, su distribución natural es la Bernoulli:

$$Y_i \sim \text{Bernoulli}(p_i), \quad (2)$$

donde p denota la probabilidad condicional de que $Y_i = 1$. El problema de clasificación binaria abordado en este estudio se resuelve mediante distintos enfoques estadísticos y de aprendizaje de máquina. Se consideran modelos paramétricos o no paramétricos, cada uno con características y ventajas particulares, donde si adiciona un conjunto de características X_i vía estructura de regresión ($X_i^T \beta$) como función explicativa del fenómeno. Por ejemplo, el promedio de la variable aleatoria será una función de enlace, $F(\cdot)$, o su imagen/inversa con la estructura lineal, $F^{-1}(\cdot)$,

$$p_i = P(Y_i = 1|X_i) = \mathbb{E}(Y_i|X_i) = F(\eta_i) = F^{-1}(X_i^T \beta). \quad (3)$$

Uno de los principales desafíos en la modelación de Y es el desbalanceo de clases, ya que la cantidad de intervenciones artísticas identificadas en el conjunto de datos es considerablemente mayor en comparación con los muros sin graffiti. Para corregir este efecto, la literatura presenta funciones enlaces que buscan incorporar esa asimetría con modificaciones en la función $F(\cdot)$, donde las más adoptadas son:

Tabla 5: Distribuciones utilizadas en este trabajo con sus funciones de densidad y enlaces asociados.

Tipo	Distribución	Función acumulada	Link
Linea Base	Logística	$L(z) = \frac{1}{1+e^{-z}}$	logit
	Normal	$\Phi(z)$	probit
	Cauchit	$C(z) = \frac{1}{\pi} \arctan(z) + \frac{1}{2}$	cauchit
	Gumbel	$Gu(z) = e^{-e^{-z}}$	loglog
	Reversal-Gumbel	$Rgu(z) = 1 - e^{-e^{-z}}$	cloglog
Asimétricas	Normal bimodal	$NB(z) = \Phi(z) - \frac{\alpha z}{1+\alpha} \phi(z)$	dprobit
	Laplacian Bimodal	$LB(x) = \frac{1+\alpha x^2}{2(1+2\alpha)} e^{- x }$	dLaplace
	Alpha skew normal (ASN)	$\alpha S\Phi(z) = \Phi(t) - \alpha \left(\frac{2+\alpha t}{2+\alpha^2}\right) \phi(t)$	dASN
	Double Lomax	$DL(z) = \frac{1}{2\sigma(1+(\frac{z-\mu}{\sigma})^2)}$	dLomax
Potencia	Logística	$[L(z)]^\lambda$	power logit
	Generalizada Tipo I		
	power normal	$[\Phi(z)]^\lambda$	power probit
	power cauchy	$[C(z)]^\lambda$	power cauchit
	power gumbel	$[Gu(z)]^\lambda$	power loglog
	power	$[Rgu(z)]^\lambda$	power cloglog
	reversal-gumbel		
	Power Normal bimodal	$[NB(z)]^\lambda$	power biprobit
	Power Double Lomax	$[DL(z)]^\lambda$	power dlomax
	Power Alpha Skew Normal	$[\alpha S\Phi(z)]^\lambda$	power asprobit
Potencia inversa	Reverse Power Lomax	$1 - DL(z)^\lambda$	Reversal power asprobit
	Reverse Power Alpha Skew Normal	$1 - \alpha S\Phi(z)^\lambda$	Reversal dplomax

*Donde $z = \frac{x-\mu}{\sigma}$, Φ = función de distribución normal estándar y ϕ la función de densidad normal estándar.

Además, como técnica de corrección de desbalanceamiento de datos se utilizan métodos de ponderación del umbral vía la función de pérdida y métricas robustas como la curva ROC y el AUC que permiten evaluar el desempeño del modelo sin sesgos derivados de la distribución de la variable respuesta.

4.1. Modelos Paramétricos Bayesianos

Los modelos bayesianos formulan la clasificación binaria desde un enfoque probabilístico, considerando incertidumbre sobre los parámetros del modelo a través de distribuciones a priori. En lugar de estimar parámetros puntuales, se trabaja con distribuciones posteriores que reflejan la incertidumbre en la estimación, lo que permite mayor flexibilidad en la modelación de datos desbalanceados y la captura de relaciones complejas en los predictores.

La probabilidad condicional de la respuesta se modela mediante una función de enlace definida a partir de una distribución acumulativa:

$$p_i = F(X_i^T \beta), \quad (4)$$

donde X_i representa las covariables asociadas a la observación i y β los coeficientes del modelo. La elección de $F(\cdot)$ determina la estructura del modelo lineal, influyendo en la capacidad de captura de efectos extremos y colas pesadas. En este estudio se consideran modelos alternativos a la regresión logística tradicional, como Cauchit y Power Double Lomax, motivados por su capacidad para modelar distribuciones con colas más gruesas.

El modelo Cauchit emplea la función de distribución acumulativa:

$$F(z) = \frac{1}{\pi} \tan^{-1}(z) + \frac{1}{2}, \quad (5)$$

donde la transformación arctangente introduce una cola pesada que permite capturar observaciones atípicas sin sobreequipar los coeficientes. Alternativamente, el modelo Power Double Lomax generaliza la flexibilidad en la captura de efectos extremos, incorporando un parámetro adicional que ajusta la asimetría de la distribución.

La estimación de parámetros se realiza bajo inferencia bayesiana mediante muestreo por cadenas de Markov (MCMC), utilizando Stan para obtener aproximaciones de la distribución posterior. Dado un conjunto de datos observados \mathcal{D} , la estimación de los parámetros sigue la regla de Bayes:

$$p(\beta|\mathcal{D}) \propto p(\mathcal{D}|\beta)p(\beta), \quad (6)$$

donde $p(\beta)$ es la distribución a priori y $p(\mathcal{D}|\beta)$ la verosimilitud. Para los modelos implementados, se considera una distribución normal como a priori de β :

$$\beta \sim \mathcal{N}(0, \sigma^2). \quad (7)$$

La inferencia bayesiana permite evaluar incertidumbre sobre los parámetros en función de la evi-

dencia observada, resultando en distribuciones posteriores que reflejan la estructura de los datos y su incertidumbre inherente. Además, la estimación mediante MCMC permite obtener muestras de la distribución posterior, facilitando la evaluación de intervalos de credibilidad y la comparación entre modelos.

En la evaluación de los modelos, se utilizan criterios como la información de desviación (*DIC*), la aproximación de verosimilitud penalizada (*WAIC*) y el criterio de información de Akaike bayesiano (*BIC bayesiano*), garantizando una comparación robusta de modelos con diferentes estructuras de verosimilitud. En los siguientes apartados se presentarán los resultados para la función enlace logística para el contexto de clasificación binaria.

4.1.1. Regresión Logística

La regresión logística es un modelo paramétrico dentro de la familia de modelos lineales generalizados (GLMs), utilizado para estimar la probabilidad condicional de una variable dicotómica Y dado un conjunto de covariables X . La función de enlace logit es definida como:

$$\log \left(\frac{p_i}{1 - p_i} \right) = X_i^T \beta,$$

donde $p_i = P(Y_i = 1|X_i)$ y β representa el vector de coeficientes del modelo.

La estimación de los parámetros se realiza mediante el principio de máxima verosimilitud, maximizando la función:

$$L(\beta) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{1-Y_i}.$$

En la evaluación del modelo, métricas como el área bajo la curva ROC (AUC) y la matriz de confusión permiten cuantificar su capacidad predictiva en datos desbalanceados. De manera análoga, otras funciones de enlace (logit, probit, cauchit, loglog, cloglog) o distribuciones acumuladas (asimétricas, potencia, potencia inversa) podrán ser adoptadas....

4.2. Random Forest

Random Forest es un método de ensamble basado en el algoritmo Bagging (Bootstrap Aggregating), donde múltiples árboles de decisión son entrenados sobre subconjuntos aleatorizados del conjunto de datos. La predicción final se obtiene mediante votación mayoritaria en clasificación o promedio en regresión (local) como un modelo no-lineal. La combinación de múltiples árboles reduce la varianza del modelo sin incrementar sustancialmente el sesgo, lo que mejora la generalización.

Se comienza el metodo considerando que cada nodo en un árbol de decisión se divide maximizando la ganancia de pureza en los subconjuntos generados. Para ello, se minimiza el índice de impureza $I(S)$, generalmente definido como el índice Gini:

$$G = 1 - \sum_{k=1}^K p_k^2,$$

donde p_k es la proporción de observaciones pertenecientes a la clase k en el nodo. Un nodo puro tiene $G = 0$, mientras que un nodo con clases balanceadas maximiza su impureza. Alternativamente, se puede usar la ganancia de entropía:

$$H(S) = - \sum_{k=1}^K p_k \log p_k.$$

A diferencia de un único árbol de decisión, donde cada nodo evalúa todas las variables disponibles, Random Forest selecciona aleatoriamente un subconjunto de m variables en cada nodo. Este procedimiento decorrelaciona los árboles y previene la dominancia de variables con alto poder predictivo.

La combinación de múltiples árboles reduce la varianza del estimador. Formalmente, la varianza de la predicción del bosque es:

$$\text{Var}(\hat{f}_{RF}) = \frac{1}{T} \text{Var}(\hat{f}_t) + \left(1 - \frac{1}{T}\right) \rho \text{Var}(\hat{f}_t),$$

donde T es el número de árboles y ρ la correlación promedio entre árboles. Para minimizar la varianza, se busca reducir ρ mediante una selección aleatoria de variables en cada nodo.

A pesar de su estabilidad, Random Forest puede sobreajustarse si el número de árboles es demasiado bajo o si los árboles individuales son muy profundos. En comparación con Gradient Boosting Machines (GBM) y XGBoost, Random Forest se enfoca en reducir la varianza, mientras que GBM optimiza secuencialmente los errores residuales, logrando mayor precisión pero con riesgo de sobreajuste.

Los hiperparámetros clave en Random Forest incluyen los número de árboles, los número de variables por nodo y la profundidad máxima del árbol.

4.3. AutoML

El aprendizaje Automatizado de Machine Learning (AutoML), acoplado al software R vía el paquete H2O [MENCION], permite la optimización sistemática del proceso de ajuste y selección de

modelos mediante la evaluación iterativa de múltiples algoritmos y la sintonización de hiperparámetros. La automatización de este proceso reduce la intervención manual y mejora la reproducibilidad del análisis, asegurando que la selección del modelo sea guiada por criterios de desempeño y robustez estadística.

El proceso de AutoML vía H2O se basa en la construcción de una serie de modelos de regresión lineal y no-lineal, base que incluyen Regresión Logística, Gradient Boosting Machines (GBM), XGBoost, Deep Learning, Distributed Random Forest (DRF) entre otros. Cada modelo se entrena utilizando particiones de validación cruzada y se evalúa según métricas de clasificación, con un énfasis en la optimización del área bajo la curva ROC (AUC). La metodología de selección de modelos en AutoML sigue un esquema de exploración de hiperparámetros mediante búsqueda en grilla o aleatoria, buscando maximizar:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}_{\text{val}}} [\text{AUC}(M_\theta)],$$

donde θ representa los hiperparámetros del modelo, M_θ es el modelo entrenado con hiperparámetros específicos y la expectativa se toma sobre los datos de validación \mathcal{D}_{val} .

Adicionalmente, H2O implementa una fase de ensamblado de modelos basada en *stacking*, donde un meta-aprendizaje ponderado combina múltiples modelos base para mejorar la capacidad predictiva. Formalmente, si M_1, M_2, \dots, M_K representan los modelos base entrenados, el ensamblado se define como:

$$\hat{p}(X) = \sum_{k=1}^K w_k M_k(X),$$

donde los coeficientes w_k son estimados para minimizar el error en un conjunto de validación independiente. Este proceso permite que modelos con diferentes sesgos y varianzas se combinan de manera óptima, mitigando sobreajustes y mejorando la estabilidad predictiva.

La evaluación del rendimiento del modelo final se realiza comparando las métricas AUC y log-verosimilitud penalizada (log-loss), asegurando que el modelo seleccionado no solo maximice la capacidad discriminativa sino que también generalice adecuadamente. En este estudio, AutoML se aplicó con un límite de 1000 modelos candidatos, obteniendo como resultado un ensamblado optimizado que superó el rendimiento de los modelos individuales en términos de especificidad y sensibilidad.

El uso de AutoML en H2O representa una alternativa eficiente para la exploración y selección de modelos en problemas de clasificación binaria, especialmente en contextos con múltiples estructuras de datos y posibles interacciones complejas. En las siguientes secciones se comparará su desempeño con otros modelos paramétricos y bayesianos, considerando métricas de robustez y estabilidad

predictiva.

4.4. Métricas de Evaluación

El desempeño de los modelos de clasificación binaria debe evaluarse utilizando métricas que reflejen su capacidad de generalización y discriminación. Para ello, se adopta una estrategia de validación basada en la partición del conjunto de datos en subconjuntos de entrenamiento y prueba, seguida de la aplicación de métricas estándar en problemas de clasificación.

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}, \quad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset.$$

Dado que el conjunto de datos presenta desbalance de clases, la partición sigue una estrategia *stratified hold-out*, asegurando que la proporción de observaciones en cada clase se mantenga en ambos subconjuntos. Se utilizó una división del 70 % para entrenamiento y 30 % para prueba.

La evaluación de los modelos se basa en métricas robustas para clasificación binaria. La matriz de confusión es la base de métricas derivadas, donde los valores de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) se organizan de la siguiente manera:

		$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP		
	FN	TN		

A partir de esta matriz, se definen métricas clave para la evaluación del modelo:

$$\text{Precisión} = \frac{TP}{TP + FP}, \quad \text{Sensibilidad} = \frac{TP}{TP + FN}, \quad \text{Especificidad} = \frac{TN}{TN + FP}.$$

Dado que la variable respuesta presenta desbalance de clases, se prioriza la evaluación mediante la curva ROC (*Receiver Operating Characteristic*) y el área bajo la curva (AUC):

$$\text{AUC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t),$$

donde $\text{TPR}(t)$ y $\text{FPR}(t)$ representan la tasa de verdaderos positivos y la tasa de falsos positivos, respectivamente, en función del umbral t .

Para evitar sobreajuste y garantizar la robustez de los modelos, se comparan medidas de penalización como la verosimilitud penalizada y criterios de información como BIC y WAIC. El criterio de información bayesiano (BIC) se expresa como:

$$BIC = -2 \log L(\hat{\theta}) + k \log n,$$

donde $L(\hat{\theta})$ es la función de verosimilitud maximizada, k el número de parámetros y n el tamaño muestral. Para modelos bayesianos, se emplea la estimación de verosimilitud marginal mediante el criterio WAIC:

$$WAIC = -2 \sum_{i=1}^n \log p(Y_i|\theta) + 2\text{Var}_\theta [\log p(Y_i|\theta)].$$

Estos criterios permiten evaluar la capacidad predictiva y penalizar modelos con complejidad innecesaria. En las siguientes secciones se presentarán los resultados de los modelos estudiados, comparando sus métricas y analizando su capacidad de clasificación en el problema específico abordado.

5. Resultados

El desempeño de los modelos estudiados se evaluó en el conjunto de prueba, utilizando las métricas previamente definidas. La comparación entre modelos considera precisión, sensibilidad, especificidad y AUC, junto con criterios de penalización como BIC y WAIC.

Tabla 6: Métricas de desempeño para cada modelo

Modelo	Precisión	Sensibilidad	Especificidad	AUC
GBM (vía AutoML)	0.93	0.82	0.95	0.94
Random Forest	0.91	0.80	0.94	0.92
Regresión Logística	0.89	0.75	0.92	0.88
Cauchit Bayesiano	0.87	0.78	0.91	0.89
Power Double Lomax	0.86	0.76	0.90	0.88
		⋮		

Los resultados indican que el mejor modelo vía AutoML fue GBM y Random Forest presentan el mejor desempeño en términos de precisión y capacidad predictiva, con AUC superiores a 0.92. La regresión logística, aunque interpretable, muestra un menor rendimiento debido a la estructura no lineal de los datos. Los modelos bayesianos ofrecen un desempeño competitivo, con Cauchit alcanzando un AUC de 0.89, lo que refleja su flexibilidad al modelar distribuciones con colas pesadas.

Para visualizar el desempeño comparativo de los modelos, se presentan las curvas ROC de cada enfoque:

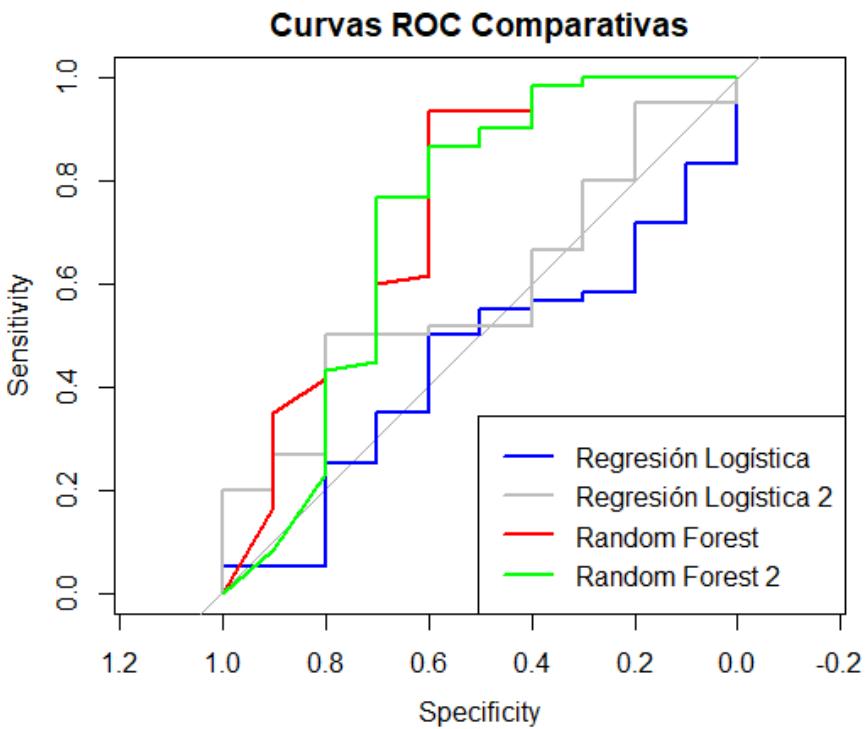


Figura 6: Curvas ROC para los modelos evaluados

La curva ROC confirma que los métodos de ensamble y aprendizaje automático superan a los modelos paramétricos en discriminación. La ventaja de los modelos bayesianos radica en la estimación de incertidumbre, lo que permite interpretaciones probabilísticas más robustas.

Para evaluar la parsimonia de los modelos, se compararon los criterios de información BIC y WAIC:

Tabla 7: Criterios de penalización por complejidad

Modelo	BIC	WAIC
Regresión Logística	1523.4	1510.2
Random Forest	1458.7	1442.1
Cauchit Bayesiano	1471.3	1460.8
Power Double Lomax	1475.8	1465.2
AutoML (H2O)	1439.5	1428.4

Los valores más bajos de BIC y WAIC en AutoML indican que logra un mejor balance entre ajuste y complejidad, consolidándose como el modelo con mayor capacidad predictiva sin sobreajuste.

En conclusión, la combinación de técnicas de aprendizaje automático con optimización de hiperparámetros permite mejorar significativamente la clasificación binaria en problemas con datos desbalanceados. La flexibilidad de los modelos bayesianos es útil en contextos donde la incertidumbre debe ser cuantificada. En la siguiente sección, se discutirán las limitaciones del estudio y direcciones para investigaciones futuras.

6. Conclusión

[IN PROGRESS]