UNIVERSIDAD ADOLFO IBAÑEZ

TESIS DE MAGÍSTER

Análisis Diferenciado del Impacto de las Industrias Extractivistas en Chile: Un Enfoque de Data Science

---

Autor:
Matías Valenzuela

Profesores guía:
Moreno Bevilacqua
Christian Caamaño
Martín Arias

Comité de defensa:
Moreno Bevilacqua
Martín Arias
Tania Roa

Tesis realizada acorde a los requerimientos para el grado de
Master of Science in Data Science

de la

Facultad de Ingeniería y Ciencias

11 de enero de 2024

UAI | FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBAÑEZ

Resumen

Facultad de Ingeniería y Ciencias

Master of Science in Data Science

Análisis Diferenciado del Impacto de las Industrias Extractivistas en Chile: Un Enfoque de Data Science

por Matías Valenzuela

Esta tesis examina los impactos de las industrias extractivistas en Chile, con un enfoque específico en los sectores minero, acuícola y forestal. A pesar de la creencia generalizada en la literatura, como se menciona en (Maillet et al., 2021), de que estos impactos son homogéneos, este estudio adopta un enfoque innovador para identificar efectos únicos y significativos que cada industria tiene en las comunas donde operan. Este análisis detallado desafía la noción de un impacto uniforme y revela la complejidad y diversidad de las consecuencias de las actividades extractivistas en distintas áreas geográficas.

El proceso metodológico comienza con la detección y winzorización de outliers utilizando el rango intercuartílico (IQR). Luego, se empleó una matriz de criterios para seleccionar variables candidatas, basada en correlaciones, importancia en un modelo de Random Forest y resultados de pruebas estadísticas, todo normalizado y ponderado para otorgar puntuaciones a cada variable.

Dado el desequilibrio de clases en los datos y que las observaciones con la que se entrenaron los modelos no son numerosas, se utilizó LOOCV (Leave-One-Out Cross-Validation) en modelos de regresión logística, iterando las seis variables de mayor puntuación en la matriz de criterios de cada industria. Las iteraciones fueron evaluadas con métricas como F1-score, valor kappa y AUC-ROC, elegidas por su aptitud para contextos de desequilibrio de clases.

Los conjuntos de variables con mejores resultados se compararon usando Máquinas de Vectores de Soporte (SVM) con LOOCV, optimización de pesos y kernel gaussiano. Se incluyeron métricas adicionales como precisión, recall y MCC, relevantes en escenarios de desequilibrio de clases.

Finalmente, se desarrolló un Indicador de Impacto a partir de una suma ponderada de variables específicas de cada industria, determinadas por su importancia en un modelo de Random Forest. Este Indicador mide el grado de influencia de las características de una industria en una comuna, con valores más altos indicando un mayor impacto. Este enfoque cuantitativo refleja el impacto diferenciado de las industrias en diversas comunas.

Este estudio revela una notable heterogeneidad en los impactos de las industrias extractivistas en Chile, especialmente en los sectores minero, acuícola y forestal. Sus hallazgos, que ofrecen nuevas perspectivas para el estudio del extractivismo tanto a nivel nacional como internacional, podrían guiar futuras investigaciones y políticas de desarrollo sostenible.

Palabras clave: Aprendizaje Supervisado, Extractivismo, Modelos de Clasificación, Regresión Logística, Máquinas de Vectores de Soporte, Random Forest.

UNIVERSIDAD ADOLFO IBAÑEZ

Abstract

Faculty of Engineering and Science

Master of Science in Data Science

Differentiated Analysis of the Impact of Extractive Industries in Chile: A Data Science Approach

by Matías Valenzuela

This thesis examines the impacts of extractivist industries in Chile, focusing specifically on the mining, aquaculture, and forestry sectors. Despite the widespread belief in the literature, as mentioned in (Maillet et al., 2021), that these impacts are homogeneous, this study adopts an innovative approach to identify unique and significant effects that each industry has on the communes where they operate. This detailed analysis challenges the notion of a uniform impact and reveals the complexity and diversity of the consequences of extractivist activities in different geographical areas.

The methodological process begins with the detection and winsorization of outliers using the interquartile range (IQR). Then, a criteria matrix was used to select candidate variables, based on correlations, importance in a Random Forest model, and statistical test results, all normalized and weighted to assign scores to each variable.

Given the class imbalance in the data, LOOCV (Leave-One-Out Cross-Validation) was used in logistic regression models, iterating the six best variables of each industry. The iterations were evaluated with metrics such as F1-score, kappa value, and AUC-ROC, chosen for their suitability in contexts of class imbalance.

The variable sets with the best results were compared using Support Vector Machines (SVM) with LOOCV, weight optimization, and Gaussian kernel. Additional metrics such as precision, recall, and MCC, relevant in scenarios of class imbalance, were included.

Finally, an Impact Indicator was developed from a weighted sum of specific variables of each industry, determined by their importance in a Random Forest model. This Indicator measures the degree of influence of an industry's characteristics on a commune, with higher values indicating a greater impact. This quantitative approach reflects the differentiated impact of the industries in various communes.

This study reveals notable heterogeneity in the impacts of extractive industries in Chile, particularly in the mining, aquaculture, and forestry sectors. Its findings, offering new perspectives for the study of extractivism both nationally and internationally, could guide future research and sustainable development policies.

Keywords: Supervised Learning, Extractivism, Classification Models, Logistic Regression, Support Vector Machines, Random Forest.

Índice general

Capítulo 1

Introducción

Chile, reconocido por su desarrollo económico impulsado por la industria extractivista (Hatzold, 2013), se destaca en sectores como la minería (U.S. Geological Survey (USGS), 2020), la acuicultura, especialmente la salmonicultura (Bustos-Gallardo, 2017), y la silvicultura. La distribución de estas industrias, con la minería predominante en el norte del pais, principalmente en el Desierto de Atacama, las plantaciones forestales en las áreas costeras del sur, y la acuicultura en las zonas maritimas cercanas a la Isla de Chiloé, refleja la estratégica adaptación geográfica del país. A pesar de ser considerado un modelo de superación de la 'Maldición de los Recursos' (Heller, 2014; *OECD Territorial Reviews: Chile* 2009; World Bank e International Finance Corporation, 2002), basado en indicadores económicos como crecimiento, ingreso per cápita y flujos de inversión extranjera, la literatura académica ha generalizado los impactos de estas industrias (Maillet et al., 2021). Este estudio desafía tal homogeneización, revelando que los impactos son significativamente distintos entre los diferentes sectores extractivistas, con consecuencias únicas en lo económico, social y ambiental en las comunas afectadas.

Esta investigación ofrece una visión detallada y precisa de cómo cada sector extractivo influye en su entorno local en Chile. Desafiando las generalizaciones comunes, la hipótesis central sostiene que los impactos de las industrias extractivistas varían según las características y operaciones específicas de cada sector. Esto subraya la necesidad de políticas de desarrollo local adaptadas a cada sector. Utilizando técnicas de machine learning, el estudio ha identificado y cuantificado impactos únicos en diferentes sectores, proporcionando insights fundamentales para el fomento de un desarrollo sostenible y equitativo en las comunidades afectadas.

Capítulo 2

Definición de la oportunidad

Este proyecto desafía la concepción generalizada en la literatura de un impacto homogéneo de las industrias extractivistas, promoviendo el concepto de *extractivismos* frente a un solo extractivismo, para reconocer la diversidad en los tipos de extracción y sus impactos específicos. Enfrentando retos en la cuantificación directa en ciencias sociales, este análisis detallado es esencial (Arias-Loyola et al., 2022), especialmente dada la influencia significativa del extractivismo en las dinámicas económicas y espaciales del país (Arboleda, 2020; Bebbington, 2009). Investigaciones previas a menudo han ignorado la variedad de efectos según el tipo de industria extractivista, llevando a políticas públicas que no consideran adecuadamente el contexto espacial, resultando en soluciones incompletas o ineficaces.

El estudio aborda esta falta de comprensión sobre los impactos específicos de las industrias extractivistas, analizando datos de CONAF, SERNAGEOMIN, SERNAPESCA, INDH, BCN y CIT UAI. Se enfoca en cómo las industrias minera, forestal y acuícola impactan de manera única en sus respectivas zonas.

Los hallazgos revelan diferencias significativas en los impactos de las industrias extractivistas en Chile, con variaciones notables entre la minería, silvicultura y acuicultura. Esta variedad de impactos resalta la necesidad de innovar en el estudio del extractivismo (Arias-Loyola et al., 2022), y enfatiza la importancia de diseñar políticas y estrategias de desarrollo adaptadas a estas particularidades regionales, asegurando intervenciones más precisas y efectivas.

Capítulo 3

Estado del arte

En la investigación sobre extractivismo, hay una notable falta de uso de técnicas de machine learning. A pesar de su aplicación en otros campos, el extractivismo aún no ha integrado estas tecnologías de manera significativa. Esta tesis explora esta brecha, aplicando innovaciones en machine learning para ofrecer una nueva perspectiva en el estudio del extractivismo y destacando oportunidades de innovación en un campo aún emergente en cuanto a tecnología.

La industria extractivista, incluyendo minería, acuicultura y silvicultura, ha sido crucial en la evolución económica de Chile (U.S. Geological Survey (USGS), 2020; World Bank e International Finance Corporation, 2002; *OECD Territorial Reviews: Chile* 2009). Cada una de estas industrias se concentra en zonas geográficas específicas del país, lo que ofrece una oportunidad única para analizar sus impactos en el desarrollo local. Sin embargo, la comprensión de estos efectos sigue siendo un campo en desarrollo. Fuentes como (Maillet et al., 2021; Arias-Loyola et al., 2022) destacan la necesidad de metodologías de investigación más innovadoras y critican la tendencia a generalizar los impactos extractivistas. Sugieren, en cambio, un enfoque más diversificado y detallado, que refleje adecuadamente la complejidad y la especificidad geográfica de estos impactos.

(Arias-Loyola et al., 2022) destaca el crecimiento económico de Chile bajo una política neoliberal extractivista, sugiriendo que la actual dinámica económica del país tiene sus raíces en la explotación intensiva de recursos naturales. Este enfoque podría haber ocasionado costos sociales y ambientales significativos, no completamente reconocidos. En línea con esto, (Romero-Toledo, 2019) enfatiza la conflictividad asociada al extractivismo, particularmente en su interacción con territorios y culturas indígenas en el norte de Chile. La tensión entre el desarrollo económico y la preservación de tradiciones y el medio ambiente ilustra las complejas dinámicas que emergen en el contexto del extractivismo.

El estudio de (Pino y Carrasco, 2018) ofrece una perspectiva crítica sobre el impacto del extractivismo forestal en Arauco, utilizando métodos etnográficos para explorar su influencia en la cultura y la vida social local. Esta actividad ha generado resistencia y ha redefinido la relación de la comunidad con su entorno, destacando el extractivismo como un fenómeno sociocultural complejo con efectos significativos en la identidad territorial. Complementariamente, (Grosfoguel, 2015) examina el extractivismo desde una perspectiva decolonial, identificando sus raíces en el capitalismo y el colonialismo occidental. Este enfoque resalta la explotación de recursos y el "pillaje cognitivo", junto con la violencia estructural que

exacerba la devastación medioambiental y social, subrayando las resistencias indígenas y cuestionando la epistemología dominante.

El trabajo de (Bolados, 2016) introduce el concepto de "zonas de sacrificio", refiriéndose a áreas desproporcionadamente afectadas por la industria extractivista, y (Bolados y Babidge, 2017) profundiza en la depredación de los recursos hídricos como una consecuencia directa de la minería. Estas observaciones resaltan las externalidades negativas del desarrollo extractivista. Complementando esto, (Svampa y Antonelli, 2009) señala que la minería en Chile no solo es un sector económico clave, sino que se ha integrado en la política de estado. Esto indica que el extractivismo ha trascendido lo económico, influenciando estructuras políticas y la formulación de políticas a nivel nacional.

Resumiendo, este estudio amplía la comprensión del extractivismo en Chile, examinando sus impactos económicos, socioculturales y medioambientales desde una perspectiva innovadora. Mediante la aplicación de técnicas de machine learning, se ofrece un análisis más preciso y diferenciado de los efectos de cada sector extractivista, contribuyendo así a una visión más matizada del extractivismo.

En este estudio, la detección de outliers es fundamental para asegurar la validez de los análisis. Por ello, se empleó el rango intercuartílico (IQR), una técnica probada en diversos campos como la minería de datos y el smart farming, según (Kumar et al., 2023) y (Durai y Shamili, 2022), para identificar y manejar outliers de manera efectiva. Este paso es crucial para revelar patrones auténticos y garantizar resultados precisos y confiables.

La winsorización, utilizada para manejar outliers, es preferida por su habilidad para preservar la estructura de los datos. Esta técnica, eficaz en mantener datos válidos, es destacada por (Kumar et al., 2023) y confirmada en su utilidad en contextos financieros por (Nyitrai y Miklós, 2018). Además, (Sullivan, Wallace y Warkentin, 2021) la reconoce como una herramienta valiosa para manejar outliers, y estudios lingüísticos por (Nicklin y Plonsky, 2020) y (Baayen, 2010) demuestran su superioridad frente a métodos como el trimming.

En (Gregorutti, Michel y Saint-Pierre, 2017) resalta Random Forest para la selección de variables, enfatizando su habilidad para evaluar la importancia e interacciones. El Coeficiente de Información Máxima (MIC), abordado por (Kinney y Atwal, 2014) y (Reshef et al., 2013), es usado para detectar relaciones complejas. Por otro lado, (Bonett, 2020) resalta la importancia de la correlación biserial puntual en el análisis de diseños de dos grupos, y, (Chicco y Jurman, 2020) enfatiza la relevancia del Coeficiente de Correlación de Matthews (MCC) en contextos con desequilibrio de clases

La regresión logística es un modelo estadístico ampliamente utilizado para predecir resultados binarios en diversas áreas. (Zaidi y Ofori-Abebrese, 2016) y (Syed et al., 2018) lo han empleado en análisis financieros, mientras que (Lizares Castillo, 2017) y (Heredia, Rodriguez y Vilalta, 2012) lo aplicaron en el sector educativo. Su eficacia también se extiende a las industrias extractivas, adaptándose a relaciones complejas y no lineales entre las variables analizadas.

Las Máquinas de Vectores de Soporte (SVM) son destacadas en inteligencia artificial para clasificación y predicción. (Herrera et al., 2015) y (Reyes et al., 2019) han aplicado SVM en educación y señales electromiográficas, respectivamente, demostrando su capacidad para manejar datos complejos. Paralelamente, (Zamudio et al., 2021) muestra su uso con un kernel gaussiano en salud, enfocado en la clasificación del personal médico para prever el burnout.

Este uso resalta la habilidad de SVM, especialmente con kernels gaussianos, para abordar conjuntos de datos con fronteras de decisión no lineales.

Capítulo 4

Hipótesis y objetivos

4.0.1. Hipótesis

El impacto de las industrias extractivistas en las comunas en las que están presentes varía según el tipo de industria. Esta variabilidad se refleja en aspectos socioeconómicos, ambientales y de accesibilidad a servicios, y es posible identificarla y cuantificarla mediante técnicas avanzadas de machine learning.

4.0.2. Objetivo General

Esta tesis tiene como objetivo el diseño de un indicador comprensivo que examine el impacto de la industria extractiva sobre las comunas en Chile, focalizándose en las variables más significativamente impactadas. Este indicador será formulado para detectar y calibrar el impacto de manera diferenciada, poniendo de relieve las comunas donde la actividad extractiva es prominente en contraposición a las que son menos afectadas. Del mismo modo, el indicador proporcionará una categorización de las distintas industrias extractivas, evidenciando su impacto único en el entorno comunal.

4.0.3. Objetivos Específicos

Para alcanzar el objetivo general, se establecen los siguientes objetivos específicos:

1. **Preparación de Bases de Datos:** Estandarizar y organizar las bases de datos existentes relacionadas con las comunas para asegurar su coherencia y compatibilidad, facilitando su análisis posterior.
2. **Delimitación de Macrozonas Geográficas:** Establecer macrozonas geográficas que representen las áreas de influencia de las industrias extractivistas, basándose en criterios como la ubicación y la extensión de las operaciones industriales.
3. **Sistema de Clasificación Comunal:** Desarrollar un método de clasificación para las comunas según la presencia y el nivel de actividad de las industrias extractivistas, permitiendo una evaluación detallada del impacto industrial.
4. **Gestión de Outliers y Estandarización:** Identificar y manejar adecuadamente los outliers, junto a escalar los conjuntos de datos para mejorar la precisión y fiabilidad de los análisis realizados.
5. **Identificación de Variables Características:** Determinar un conjunto óptimo de variables características para cada industria extractivista. Esto incluye el desarrollo de una

matriz de criterios de selección, la experimentación con diferentes combinaciones de variables y la validación final de estos conjuntos para asegurar que reflejen con precisión los impactos específicos de cada industria.

6. **Construcción de un Indicador Integral:** Crear un indicador comprensivo que mida el impacto de las industrias extractivistas en las comunas, utilizando los sets de variables características validadas.

4.0.4. Resultados

El objetivo de **Preparación de Bases de Datos** se ha cumplido satisfactoriamente, como se muestra en la tabla 5.1 y en la sección de metodología 5.1.2. Este trabajo incluyó recopilar y organizar una amplia variedad de datos para un análisis eficiente a nivel comunal, con énfasis en la estandarización y adaptación de los niveles de desagregación de datos. Se ha realizado también la geolocalización de los datos de conflictos socioambientales. Los detalles y la importancia de cada base de datos se explican en la sección de metodología y en la tabla 5.1.

El objetivo de **Delimitación de Macrozonas Geográficas** se ha cumplido eficazmente, estableciendo zonas que reflejan las áreas de influencia de las industrias extractivistas a través de un análisis geoespacial y la integración de múltiples fuentes de datos. Los detalles de la metodología y los criterios utilizados se explican en la sección 5.2.

El objetivo de **Desarrollar un Sistema de Clasificación Comunal** se ha logrado, estableciendo un método de clasificación basado en la actividad de las industrias extractivistas. Este sistema, que asigna comunas a categorías según percentiles de actividad industrial, permite diferenciar su influencia extractivista y permite analizar el impacto de estas industrias. Los detalles de la metodología y categorización se presentan en la sección de *Preprocesamiento de Datos*.

El objetivo de **Gestión de Outliers y Estandarización** se ha cumplido, aplicando el método del Rango Intercuartílico (IQR) y la winsorización para manejar outliers, como se detalla en la sección de *Preprocesamiento de Datos*. Esta aproximación preservó la integridad de los datos y mejoró la eficacia de los análisis.

Se ha logrado con éxito el objetivo de **Identificar Variables Características** para cada industria extractivista, siguiendo un proceso meticuloso y detallado. Inicialmente, se aseguró la robustez de los datos mediante la detección y winsorización de outliers. Posteriormente, se implemento una matriz de criterios para la selección preliminar de variables, basándonos en factores como correlaciones, importancia en un modelo de Random Forest y resultados de pruebas estadísticas. Así, a cada variable se le asigno una puntuación, los criterios están especificados en 5.4 y la estructura de la matriz en (ver tabla A.1).

Luego, se experimento con combinaciones de las seis variables con mayor puntuación de la matriz de criterios para cada industria en un modelo de regresión logística. Las iteraciones se evaluaron con métricas de evaluación para determinar los conjuntos de variables más efectivos, el proceso se detalla en 5.5. Estos conjuntos se compararon ulteriormente mediante el uso de Máquinas de Vectores de Soporte (SVM). El proceso se detalla en 5.6. La metodología se ilustra en el diagrama A.1 y las variables en A.2.

El objetivo de **Construir un Indicador Integral** ha sido exitosamente cumplido. Mediante el uso de RandomForest, se seleccionó y ponderó las variables caracteristicas para evaluar el impacto de las industrias extractivistas en las comunas. Este proceso y su justificación se detallan en la sección *Construcción del Indicador* del estudio.

Los resultados se concentran en tres mapas clave para cada industria extractivista: uno que muestra la presencia de la industria en las comunas, otro que muestra clasificación de las comunas utilizada en el estudio, y un tercero que presenta el valor del indicador integral, los mapas pueden ser consultados en A.0.4.

Capítulo 5

Metodología

5.1. Preparación y Unificación de las Bases de Datos

5.1.1. Bases de Datos Utilizadas

Nombre de la Base	Descripción	Fuente
Catastro Vegetacional de CONAF	Esta base de datos proporciona información detallada sobre la vegetación en Chile, incluyendo la extensión de las plantaciones forestales por comuna y otros datos relacionados con la cobertura vegetal y el uso de la tierra.	SIG de CONAF
Faenas Chilenas (de minería) de SERNAGEOMIN	Contiene información sobre la ubicación y características de las faenas mineras en Chile, incluyendo el tipo de faena, su categoría según SERNAGEOMIN y otros datos relevantes sobre cada faena.	https://datos.gob.cl/dataset/faenas-en-chile
Listado de concesiones vigentes (marzo 2023) de SUBPESCA	Proporciona un registro actualizado de las concesiones de acuicultura en Chile, incluyendo la ubicación de las faenas y otros detalles relevantes sobre las faenas acuícolas.	SUBPESCA
Listado de conflictos sociambientales INDH	Esta base de datos recopila información sobre conflictos socioambientales ocurridos en Chile, ofreciendo descripciones breves de cada conflicto y datos adicionales que permiten entender su contexto y magnitud.	INDH
Polígonos de las comunas del país	Contiene el mapa del país, sin capas, que muestra los limites de las comunas	BCN
Matriz de Bienestar Humano Territorial	Esta base de datos incluye una amplia gama de indicadores por manzana, abarcando aspectos de accesibilidad a servicios, seguridad, variables socioeconómicas y ambientales, ofreciendo una visión integral del bienestar humano a nivel local.	CIT UAI
Código Unico Territorial	Esta base de datos contiene el código único territorial de cada comuna, región y provincia.	Superintendencia de Seguridad Social

CUADRO 5.1: Resumen de Bases de Datos

5.1.2. Preparación de las Bases de Datos

Para analizar coherente y conjuntamente diversas bases de datos, cada una con características únicas, fue esencial adaptarlas a un formato desagregado por comuna. Este capítulo detalla cómo se reestructuraron estas bases para correlacionar y analizar los datos a nivel comunal.

Código Único Territorial

Esta base de datos, con los nombres e identificaciones únicas de cada comuna, es fundamental para estandarizar y unificar los identificadores en todas las bases de datos, asegurando una correspondencia precisa entre ellas.

Catastro Vegetacional de CONAF

Esta base de datos contiene información desagregada por comuna, detallando las hectáreas plantadas de especies forestales como pino, eucalipto, entre otras. Para facilitar la comparativa entre comunas, se asignó una columna adicional que acumula el valor total de las hectáreas plantadas en cada comuna.

Faenas Chilenas de SERNAGEOMIN

La segunda base de datos consta de un compendio de faenas mineras, cada una con múltiples atributos que reflejan su operación y categorización. La puntuación de cada faena se determinó siguiendo una progresión geométrica basada en la clasificación de SERNAGEOMIN, que incluye las siguientes categorías:

- Categoría A: Grandes empresas con más de un millón de horas hombre al año.
- Categoría B: Empresas con entre 200 mil y un millón de horas-hombre al año.
- Categoría C: Empresas con entre 27 mil y 200 mil horas-hombre al año.
- Categoría D: Pequeñas operaciones con hasta 27 mil horas-hombre al año.

La progresión geométrica, asignando 64 puntos a la Categoría A, 16 a la B, 4 a la C, y 1 a la D, permite ponderar adecuadamente la influencia de cada faena en función de su tamaño y actividad económica. Luego, a cada comuna se le calculo una puntuación total, que es la suma de las puntuaciones de las faenas mineras que están presentes en cada comuna. Se filtraron las faenas para incluir solo aquellas categorizadas como mina subterránea y mina a rajo abierto.

Listado de Concesiones Vigentes de SUBPESCA

Esta base de datos se enfocó en la representación cuantitativa de la actividad acuicola, asignando una columna que refleja el número total de faenas presentes en cada comuna.

Polígonos

Esta base de datos, consistente en polígonos geoespaciales, fue instrumental para la visualización de los datos.

Listado de Conflictos Socioambientales INDH

Esta base de datos presento un desafío debido a la falta de una columna con ubicaciones geográficas claras. Para resolverlo, se extrajo la información de la localización de las descripciones de cada registro de conflicto socioambiental. Utilizando la librería googleway de R Cooley, 2022 y la API de Google Maps utilizando la función *google_geocode()* para determinar la comuna correspondiente de cada conflicto.

Matriz de Bienestar Humano Territorial

Esta base de datos necesitaba adaptarse al análisis comunal. La información, originalmente a nivel de manzana, fue ajustada ponderando los indicadores según la proporción de la población de cada manzana respecto a la comuna entera. Este enfoque de suma ponderada permite que los indicadores reflejen de forma precisa la realidad comunal, incorporando las variaciones a menor escala.

Con la unificación de todas estas bases de datos, se ha establecido una plataforma robusta para el análisis integrado y comparativo, permitiendo una interpretación detallada y precisa del impacto de las industrias extractivistas en las comunas de Chile.

5.1.3. Variables de cada Industria

Para la industria forestal, se examinó la proporción del área de cada comuna ocupada por plantaciones forestales. En la industria de acuicultura, se utilizó el número de faenas en cada comuna. En cuanto a la industria minera, se aplicó un sistema de puntuación ponderada a nivel comunal. Esto implicó sumar las puntuaciones de cada faena minera según su categoría SERNAGEOMIN para evaluar la actividad minera en cada comuna.

5.2. Definición de Macrozonas

Las macrozonas se crearon al agrupar regiones según la concentración de industrias forestales, acuícolas y mineras, reduciendo el ruido estadístico. Esto permitió un análisis más enfocado y mejoró la calidad de los datos. La selección de macrozonas optimizó recursos computacionales, facilitó la validación de modelos y proporcionó información valiosa sobre el impacto de las operaciones extractivas en las comunas.

Industria	**Regiones que Conforman la Macrozona**
Minería	Tarapacá, Antofagasta, Atacama, Coquimbo, Valparaíso, Región Metropolitana de Santiago
Acuicultura	Los Lagos, Aysén
Forestal	BioBio, O'Higgins, Maule, Ñuble, Araucanía

CUADRO 5.2: Macrozonas por Industria

5.3. Preprocesamiento de Datos

5.3.1. Proceso de Clasificación de las Industrias

Para analizar el impacto de la industria extractivista en las comunas, se clasificaron en dos categorías: aquellas con presencia significativa de faenas extractivistas y las que tienen poca o ninguna. Esto permitió evaluar cómo diferentes conjuntos de variables afectan la clasificación. Si los modelos muestran un alto rendimiento, se pueden identificar las variables más determinantes que reflejan el impacto de la industria extractivista en las comunas del país.

Industria Minera

Para clasificar las comunas en la macrozona minera, se calculó primero el percentil 25 de la variable de industria, pero solo para aquellas comunas con puntuaciones mayores a cero. Posteriormente, este umbral se aplicó a todas las comunas, incluyendo aquellas con valor cero. De esta manera, se diferenciaron dos grupos: las comunas con actividad minera significativa, con puntuaciones iguales o superiores al percentil 25 (contemplando solo a las comunas que si tienen presencia de industria), y aquellas con poca o ninguna actividad minera, con puntuaciones menores a dicho percentil.

Acuicultura

Para el sector acuícola, se adoptó un enfoque similar al del análisis minero en la clasificación de las comunas. El cálculo del umbral, basado en el percentil 25, se realizó solo para aquellas comunas con operaciones acuícolas activas (valor distinto de cero). Luego, este umbral se aplicó a la totalidad de las comunas, permitiendo clasificarlas en dos categorías: las que presentan una alta concentración de operaciones acuícolas (igual o superior al percentil 25) y aquellas con menor o ninguna actividad acuícola (inferior al percentil 25).

Industria Forestal

Considerando que las plantaciones forestales son comunes en la mayoría de las comunas, se adoptó un criterio diferente para la clasificación, utilizando el percentil 75 como referencia. Las comunas con una proporción de plantaciones forestales igual o superior al percentil 75 se clasificaron como zonas de alta densidad forestal. Por otro lado, las que quedaron por debajo de este umbral se consideraron de menor densidad forestal.

Industria	Mayor o Igual al Umbral de Clasificación	Menor al Umbral de Clasificación
Minera	33	78
Acuicultura	17	23
Forestal	37	112

CUADRO 5.3: Resultados de la Clasificación

5.3.2. Detección de Outliers con el Método del Rango Intercuartílico (IQR)

El Rango Intercuartílico (IQR) es una técnica estadística para identificar outliers, definiendo la dispersión estadística como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1).

Los cuartiles dividen los datos en cuatro partes, con el IQR enfocándose en la mitad central y excluyendo extremos.

$$\text{IQR} = Q3 - Q1$$

donde $Q1$ es el valor medio entre el menor valor y la mediana, y $Q3$ es el medio entre la mediana y el valor más alto. Los outliers se identifican estableciendo límites: valores menores que $Q1 - 1{,}5 \times \text{IQR}$ o mayores que $Q3 + 1{,}5 \times \text{IQR}$ son considerados atípicos.

5.3.3. Winsorización: Un Enfoque para Manejar Outliers

La winsorización es una técnica de transformación de datos que reduce el impacto de outliers al reemplazar valores extremos por otros menos extremos, preservando la estructura general del conjunto de datos.

Consiste en ajustar los valores extremos de un conjunto de datos para disminuir el impacto de outliers. Se reemplazan valores por debajo y por encima de ciertos percentiles con valores en estos percentiles. Por ejemplo, en winsorización al 5 %, los valores debajo del percentil 5 se reemplazan por el valor en dicho percentil, y los valores sobre el percentil 95 por el valor en ese percentil.

5.3.4. Aplicación de la Winsorización y el Método IQR en el Estudio

Identificación de Outliers con IQR

Se utilizó el método IQR para identificar outliers, determinando límites superior e inferior de la distribución de los datos.

Aplicación de la Winsorización con Límites del IQR

Como se hizo en (Kumar et al., 2023), tras identificar outliers con IQR, se aplicó winsorización usando límites IQR como valores de corte. Outliers sobre el límite superior (Q3 + 1.5 * IQR) se reemplazaron por este valor, y los debajo del límite inferior (Q1 - 1.5 * IQR) por el valor mínimo del IQR.

Esta combinación de IQR y winsorización mitiga el efecto de valores extremos sin descartarlos. Ajustando estos puntos a límites extremos pero representativos, se conserva la integridad de los datos, manteniendo la robustez del análisis estadístico y modelado predictivo. Se decidio winsorizar ya que mejora la capacidad predictiva de los modelos, como se puede ver en 6.0.1.

5.3.5. Estandarización de Variables en Análisis Estadístico y Modelado Predictivo

La estandarización de datos, crucial en análisis estadístico y aprendizaje supervisado, implica reescalar variables para que tengan media cero y desviación estándar de uno, conforme a la fórmula $Z = \frac{(X-\mu)}{\sigma}$. Este proceso uniformiza las variables y mejora el rendimiento de modelos de clasificación.

5.4. Selección de Variables Mediante Matriz de Criterios

5.4.1. Introducción de la Matriz de Criterios

La matriz de criterios combina criterios binarios basados en pruebas estadísticas y criterios continuos como la importancia en modelos de Random Forest o correlaciones, evaluando las variables en una escala de 0 a 1. Se utilizó para preseleccionar variables antes de probarlas en modelos de clasificación, lo que redujo la carga computacional y permitió identificar de manera eficiente y objetiva las variables más relevantes.

5.4.2. Estructura de la Matriz de Criterios

La matriz de criterios en este estudio incluyó diversos criterios para evaluar y ponderar la importancia de las variables:

1. **Importancia en Modelos de Random Forest:** Se calculó la importancia de cada variable basándose en su contribución a la precisión de modelos de Random Forest, utilizando tanto la validación cruzada Leave-One-Out (LOOCV) como la de K-Folds. Las importancias se sumaron y normalizaron, proporcionando una puntuación ponderada por variable. Una Importancia mayor, significa que esa variable aporta mas a la capacidad de clasificación del modelo en comparación a variables con una importancia menor.

2. **Correlación MIC:** Se aplicó la correlación MIC para evaluar la asociación entre cada variable independiente y la variable de cada industria (5.1.3). Estos valores, normalizados entre 0 y 1, proporcionaron una medida cuantitativa de la fuerza de las relaciones. Un valor mas cercano a 1 representa una relación mas fuerte, MIC muestra también relaciones mas complejas que la lineal.

3. **Correlación Biserial:** Se utilizó la correlación biserial para examinar la relación entre variables numéricas y la variable objetivo binaria, identificando asociaciones significativas. Estos valores, también fueron normalizados entre 0 y 1.

4. **Test de Kruskal-Wallis:** Este test no paramétrico se empleó para comparar las medianas de las variables entre categorías, traduciendo los resultados en valores binarios basados en la significancia estadística. Se utiliza este criterio ya que si una variable muestra una diferencia significativa en su mediana según a que categoría pertenezca, se entiende que es mas probable que ayude a los modelos de clasificación a tener un mejor rendimiento.

   Hipótesis Nula (H_0): Todas las poblaciones tienen la misma mediana.
   Hipótesis Alternativa (H_a): Al menos una población tiene una mediana diferente.

5. **Test de Fligner-Killeen:** Se utilizó para evaluar la homogeneidad de las varianzas entre grupos, incorporando los resultados en la matriz como valores binarios según su significancia. Se utilizó este criterio ya que si una variable muestra tener distinta varianza según su categoría, es mas probable que esta variable ayude a los modelos de clasificación a tener un mejor rendimiento.

   Hipótesis Nula (H_0): Las varianzas en todos los grupos son iguales.
   Hipótesis Alternativa (H_a): Al menos un grupo tiene una varianza diferente.

6. **Test de Kolmogorov-Smirnov:** Se aplicó para comparar la distribución de las variables entre clases, convirtiendo los resultados en valores binarios basados en la significancia estadística. Este criterio fue utilizado ya que si una variable tiene distintas distribuciones según su categoria, es mas probable que esta variable ayude a los modelos de clasificación a tener un mejor rendimiento.

   Para dos muestras comparadas entre sí:

   Hipótesis Nula (H_0): Las dos muestras provienen de la misma distribución.
   Hipótesis Alternativa (H_a): Las dos muestras provienen de distribuciones diferentes.

Se puede ver la estructura de la matriz de criterios en A.1.

Selección de Estadísticas No Paramétricas y Validación Cruzada

La selección de estadísticas no paramétricas se basó en un análisis de la distribución de las variables, confirmándose que la mayoría no seguía una distribución normal (según el test de Shapiro-Wilk). Dado el desequilibrio de clases y el número limitado de observaciones, se optó por utilizar validación cruzada Leave-One-Out (LOOCV) en los modelos para maximizar el uso de las observaciones.

Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado que genera múltiples árboles de decisión durante el entrenamiento y realiza predicciones basándose en la agregación de sus resultados. Para clasificación, predice la clase más votada entre todos los árboles; para regresión, calcula la media de las predicciones. Introduce aleatoriedad mediante muestras de bootstrap y selección aleatoria de características en los nodos, lo que ayuda a prevenir el sobreajuste y proporciona una indicación de la importancia de las características. (Hsieh, 2023)

Se aplicó el modelo Random Forest con LOOCV para evaluar la importancia de las variables en la clasificación. Se utilizó la función *train_control()* de la librería caret de R (Kuhn, 2022) y la función *train()* de la librería randomForest de R (Liaw y Wiener, 2002). Esto permitió construir múltiples árboles de decisión y combinarlos para mejorar la precisión y generalización del modelo.

Correlación MIC

El Coeficiente de Información Máxima (MIC) es una medida estadística que evalúa la intensidad y la fuerza de la relación entre dos variables, ya sea lineal o no lineal. Se destaca por su habilidad para identificar y cuantificar relaciones fuertes en conjuntos de datos, analizando diversas particiones de datos para maximizar la información mutua. Esta característica lo hace especialmente valioso para detectar asociaciones no lineales en análisis de datos complejos, donde las relaciones no siempre son obvias. En este estudio, MIC se utilizó para cuantificar la fuerza de las relaciones entre variables, proporcionando una comprensión más profunda de su importancia en el contexto del análisis (Kinney y Atwal, 2014) y (Reshef et al., 2013).

Correlación Biserial

La correlación biserial es una técnica estadística diseñada para medir la relación entre una variable binaria (por ejemplo, éxito/fracaso) y una variable continua. Es particularmente útil en situaciones donde se necesita entender la fuerza y dirección de la asociación entre una característica cualitativa dicotómica y una cuantitativa. Este tipo de correlación evalúa cómo la variabilidad en la variable continua se relaciona con la diferencia entre las dos categorías de la variable binaria. La correlación biserial proporciona una estimación de la correlación lineal y es valiosa en contextos donde las variables no cumplen con las suposiciones de las pruebas de correlación más tradicionales (Kornbrot, 2005).

Suma de Puntuaciones

A partir de las matrices de criterios (una por cada macrozona), se procedió a calcular una puntuación total para cada variable. Esto se logró sumando los valores asignados a cada criterio por variable. Esta puntuación representa la relevancia comprensiva y el cumplimiento de cada variable en relación con todos los criterios considerados, como los criterios continuos están normalizados entre 0 y 1, se asegura un influencia equitativa en el calculo de la puntuación.

Selección de Variables Principales

Para cada macrozona, se seleccionaron las seis variables con las puntuaciones más altas. Esta selección se basó en el principio de que las variables con mayores puntuaciones son aquellas que probablemente reflejan la diferencia de las clases.

Macrozona	Set de variable
Minera	Participación juvenil en empleo y estudio; Conflictos socioambientales; Empleo; Delitos leves contra la propiedad; Equipamientos de educación; Equipamiento deportivo
Forestal	Equipamientos culturales; Empleo; Equipamiento de servicios publicos; Suficiencia de vivienda; Participacion juvenil en empleo y estudio; Delitos leves contra la propiedad
Acuicultura	Empleo; Equipamientos deportivos; Suficiencia de vivienda; Conflictos socioambientales; Equipamiento de educación; Resilencia de hogares

CUADRO 5.4: Variables de mayor puntuación de la Matriz de Criterios

5.5. Análisis de Subconjuntos de Variables

Ahora se pretende encontrar el subconjunto de variables que mejor clasifiquen a las comunas. Si bien, se podía intentar probar la combinación de todas las variables de la base de datos, computacionalmente este proceso era muy costoso, es por esto que se creó la matriz de criterios.

Se itero las 6 variables con mayor puntuación en la matriz de criterios, se probó cada combinación en un modelo de regresión logística con validación cruzada LOOCV. De todas estas combinaciones, se seleccionó el set que tuvo un mejor valor para la métrica AUC-ROC, otro para la métrica F1 score y otro para el valor Kappa. En este paso se utilizo la libreria caret

(Kuhn, 2022) de R para entrenar los modelos y la libreria pROC de R (Robin et al., 2021) para la métrica AUC-ROC y la libreria irr de R (Gamer, Lemon y Puspendra, 2019).

5.5.1. Regresión Logística

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de una variable binaria. La relación entre la variable dependiente y una o más variables independientes se modela utilizando la función logística.

La función logística, que es el núcleo de este modelo, se define como:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

donde $P(Y = 1)$ es la probabilidad de que la variable dependiente Y sea 1, β_0 es la intersección, $\beta_1, \ldots, \beta_k$ son los coeficientes de las variables independientes $X_1, \ldots, X_k$, y e es la base del logaritmo natural.

Este modelo se ajusta utilizando el método de máxima verosimilitud, buscando estimar los coeficientes β que maximizan la probabilidad de observar los datos dados. (Jr., Lemeshow y Sturdivant, 2013)

5.5.2. Importancia de las Métricas en Datos con Desequilibrio de Clases

En contextos donde los datos presentan un desequilibrio de clases, es crucial seleccionar métricas de evaluación adecuadas:

- **AUC (Área Bajo la Curva ROC):** Esta métrica, que representa el área bajo la curva del gráfico de Características Operativas del Receptor (ROC), evalúa la capacidad de un modelo de clasificación para distinguir entre clases. Una de las razones por las que el AUC es particularmente útil en conjuntos de datos desequilibrados es que proporciona una medida agregada del rendimiento del modelo a lo largo de todos los umbrales de clasificación posibles. Esto significa que no se ve directamente afectada por la prevalencia de una clase sobre otra, a diferencia de métricas como la precisión o la tasa de verdaderos positivos, que pueden dar una visión sesgada del rendimiento del modelo en presencia de un desequilibrio de clases. Al integrar el rendimiento del modelo en todo el espectro de umbrales, el AUC ofrece una visión más holística y equilibrada.
- **F1 Score:** El F1 Score es una métrica que combina la precisión (la proporción de identificaciones positivas que fueron correctas) y la sensibilidad (también conocida como tasa de verdaderos positivos o recall, que mide la proporción de positivos reales que fueron identificados correctamente) en un solo número. Se calcula como el promedio armónico de la precisión y la sensibilidad, ofreciendo un balance entre ambas. Matemáticamente, se expresa como:

  $$F1 = 2 \cdot \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

  Esta métrica es especialmente útil en situaciones donde las falsas positivas y las falsas negativas son igualmente costosas, y es más informativa que usar la precisión o la

sensibilidad por separado. El F1 Score se vuelve particularmente importante en escenarios con desequilibrio de clases, ya que en estos casos, las métricas como la precisión pueden dar una idea equivocada del rendimiento del modelo. Un F1 Score alto indica un equilibrio adecuado entre precisión y sensibilidad.

- **Valor Kappa:** El coeficiente Kappa es una métrica que mide la concordancia entre las predicciones de un modelo y las observaciones reales, ajustándose por la posibilidad de coincidencia aleatoria. Esta métrica es particularmente útil en la evaluación de modelos sobre datos con desequilibrio de clases, ya que toma en cuenta tanto los aciertos como los errores en todas las categorías de la clasificación. El valor Kappa no solo compara la concordancia con lo que se esperaría por azar, sino que también ajusta este acuerdo en consideración al desequilibrio potencial en la distribución de clases (Cohen, 1960).

  El valor Kappa se calcula de la siguiente manera:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

  donde:

  - P_o es la proporción de acuerdo observado entre las clasificaciones.
  - P_e es la proporción de acuerdo que se esperaría por azar.

  El valor de κ puede variar entre -1 y 1. Un valor de 0 indica que el acuerdo entre las clasificaciones es el que se esperaría por azar, mientras que un valor de 1 indica un acuerdo perfecto. Valores negativos sugieren un acuerdo peor que el azar, lo que indica una discordancia significativa entre las predicciones y las observaciones reales.

Estas métricas proporcionan una visión mas robusta frente al desequilibrio de clases presente en los datos. Durante la selección de conjuntos de variables, se notó que en las tres macrozonas, los mejores conjuntos para F1 Score y valor Kappa eran iguales, por lo que, en vez de comparar tres sets de variables (uno por cada métrica) solo se compararon dos.

Macrozona	Mejor AUC	Mejor F1-Score y Valor Kappa
Minera	Participación juvenil en empleo y estudio; Conflictos socioambientales; Equipamiento de educación	Participación juvenil en empleo y estudio; Conflictos socioambientales; Equipamiento de educación; Delitos leves contra la propiedad; Equipamientos deportivos
Forestal	Equipamientos culturales; Equipamiento de servicios públicos; Participación juvenil en empleo y estudio; Delitos leves contra la propiedad	Equipamientos culturales; Equipamiento de servicios públicos; Empleo; Delitos leves contra la propiedad
Acuicultura	Equipamientos deportivos; Suficiencia de vivienda; Equipamientos de educación; Resiliencia de hogares	Equipamientos deportivos; Equipamientos de educación

CUADRO 5.5: Sets de Variables con Mejores Metricas por Macrozona

El rendimiento de los modelos se puede ver en 6.0.2.

5.6. Selección de Variables Utilizando Modelos SVM

En el proceso de iteraciones previo, se identificaron los dos subsets de variables que mostraron un mejor rendimiento en las métricas. Ahora, el objetivo es determinar, para cada macrozona, cuál de los dos subsets identificados es más eficaz en la predicción de las clasificaciones de las observaciones. Para lograr esto, se emplearon modelos de Máquinas de Soporte Vectorial (SVM)

Este proceso se dividió en dos pasos. En el primero, se buscó una distribución de pesos adecuada para mitigar el efecto del desbalance de clases en los modelos SVM. Se llevaron a cabo iteraciones con diferentes distribuciones de pesos en un modelo SVM para cada conjunto de variables. La distribución de pesos que resultó en el mejor valor de métrica F1-Score en los modelos SVM entrenados para cada conjunto de variables se seleccionó para la siguiente etapa.

El siguiente paso consistió en comparar y determinar cuál de estos subsets era superior. Para llevar a cabo esta comparación, se evaluaron modelos SVM con la distribución de pesos obtenida, con LOOCV y kernel gaussiano (o raidal) para cada subset. Para entrenar los modelos SVM se utilizo la libreria e1071 de R (Meyer et al., 2021).

Fundamentos de las Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM) son una técnica de aprendizaje automático supervisado ampliamente utilizada tanto para tareas de clasificación como de regresión. Lo que distingue a las SVM es su enfoque en encontrar un hiperplano óptimo que separa las clases con el mayor margen posible. Este hiperplano actúa como un límite de decisión entre diferentes clases en el espacio de características. La eficacia de las SVM reside en su capacidad para gestionar datos complejos, incluso aquellos que no son linealmente separables en su espacio original.

Kernel El concepto de kernel es fundamental en las SVM, especialmente para tratar con datos no lineales. El uso de una función kernel permite a las SVM transformar el espacio de características original en un espacio de mayor dimensión donde los datos pueden ser linealmente separables. Este proceso se conoce como ”truco del kernel” y es clave para la flexibilidad y potencia de las SVM. En lugar de realizar cálculos en un espacio de alta dimensión, lo que sería computacionalmente costoso, las funciones kernel permiten que estos cálculos se realicen en el espacio original de manera eficiente.

5.7. Evaluación de los Modelos

La matriz de confusión es crucial para evaluar modelos predictivos en aprendizaje automático. A partir de esta matriz, se calculan diversas métricas de evaluación que permiten medir la calidad del modelo.

5.7.1. Matriz de Confusión

Una matriz de confusión es una tabla que permite la visualización del rendimiento de un modelo de clasificación. Cada fila de la matriz representa las instancias en una clase real,

mientras que cada columna representa las instancias en una clase predicha. Los elementos de la matriz son:

- Verdaderos Positivos (TP): Casos correctamente identificados como positivos.
- Falsos Positivos (FP): Casos incorrectamente identificados como positivos.
- Verdaderos Negativos (TN): Casos correctamente identificados como negativos.
- Falsos Negativos (FN): Casos incorrectamente identificados como negativos.

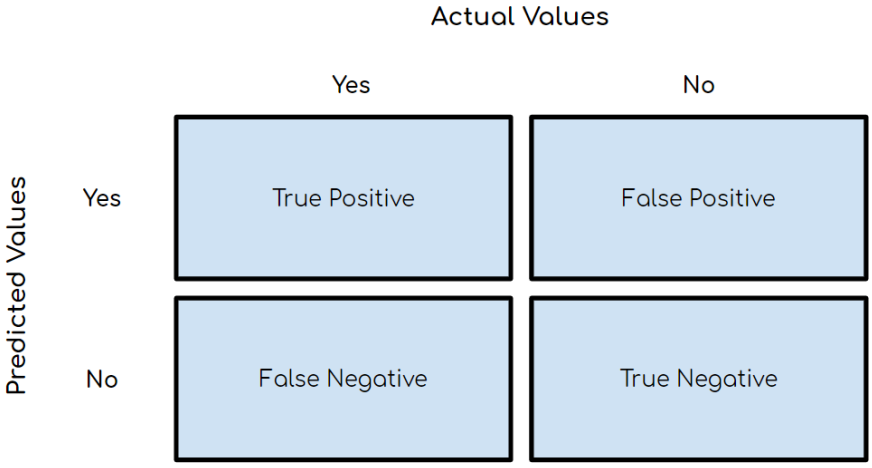


FIGURA 5.1: Matriz de Confusión.

Las siguientes métricas se derivan de estos valores:

Precision

La precisión se calcula como:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Esta métrica es importante para evaluar la capacidad del modelo para no clasificar como positiva una muestra negativa.

Recall

El recall, o sensibilidad, se calcula como:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Es fundamental para medir la capacidad del modelo de identificar todas las muestras positivas.

MCC

El MCC (Coeficiente de Correlación de Matthews) es una métrica robusta para evaluar modelos de clasificación, considerando verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Es útil en casos de datos desbalanceados.

El MCC se calcula con la fórmula:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Otras métricas utilizadas

F1-Score y AUC estan detalladas en 5.5.2

Resultados SVM

El rendimiento de los modelos esta en 6.0.3.

5.8. Matriz de Variables caracteristicas por Macrozona

Macrozonas	Variables caracteristicas
Minería	Participación juvenil en empleo y estudio; Conflictos socioambientales; Equipamiento de educación; Delitos leves contra la propiedad; Equipamiento deportivo
Acuicultura	Equipamiento de educación; Equipamiento de deportes
Forestal	Equipamientos culturales; Empleo; Equipamiento de servicios públicos; Delitos leves contra la propiedad

CUADRO 5.6: Variables caracteristicas de cada Macrozona

5.9. Construcción del Indicador

5.9.1. Selección y Evaluación de Variables

Utilizando RandomForest, se seleccionaron variables de alto rendimiento en distintas macrozonas para determinar su importancia relativa en la clasificación.

5.9.2. Cálculo de la Importancia Relativa

Para normalizar las importancias obtenidas de las variables en un modelo RandomForest, se ajustaron los valores para que la suma total sea igual a 1. Esto se hace para garantizar que cada variable contribuya de manera proporcional en el modelo. Matemáticamente, si denotamos la importancia de la variable i-ésima como I_i, la importancia relativa normalizada I_i^{norm} se calcula como:

$$I_i^{\text{norm}} = \frac{I_i}{\sum_{j=1}^{n} I_j}$$

donde n es el número total de variables. Cada I_i^{norm} representa la contribución proporcional de la variable i-ésima en el modelo.

5.9.3. Elaboración y Normalización del Indicador

Para construir un indicador compuesto, se utilizó una suma ponderada de variables previamente normalizadas, asignando pesos de acuerdo con la importancia relativa obtenida del modelo RandomForest. La normalización de las variables se llevó a cabo para que sus valores estuvieran en una escala común de 0 a 1, lo cual es crucial para mantener la coherencia y comparabilidad del indicador final. La fórmula para calcular el indicador normalizado es la siguiente:

$$I = \sum_{i=1}^{n} w_i \cdot x_i^{\text{norm}}$$

donde I es el valor del indicador, n es el número total de variables, w_i es el peso o importancia relativa de la variable i-ésima derivado de RandomForest, y x_i^{norm} es el valor normalizado de la variable i-ésima. La normalización de cada variable x_i se realiza mediante:

$$x_i^{\text{norm}} = \frac{x_i - \text{mín}(x_i)}{\text{máx}(x_i) - \text{mín}(x_i)}$$

donde $\text{mín}(x_i)$ y $\text{máx}(x_i)$ son, respectivamente, el mínimo y el máximo de la variable i-ésima en el conjunto de datos.

5.9.4. Asignación de Signos y Ajuste Final del Indicador

Se determinaron los signos de las variables basándose en su distribución y efecto en la clasificación. Finalmente, se normalizó el indicador para que sus valores estén entre 0 y 1, facilitando así su interpretación y comparación.

Las métricas de los modelos Random Forest y los signos de las variables estan en 6.1.5, se destaca que el signo de la variable en el indicador indica en que sentido afecta la presencia de industria.

Capítulo 6

Resultados y análisis

6.0.1. Elección de Winsorización

Para evaluar el potencial impacto positivo de la winsorización de los outliers detectados mediante el método IQR en la investigación, se llevó a cabo una comparación de diferentes conjuntos de variables para cada macrozona. Este paso se realizó antes de implementar la matriz de criterio, por lo que se armaron los set a evaluar en base el test de Kruskal-Wallis, el test de Fligner-Killeen y el test de Kolmogorov-Smirnov.

Cada test genero un conjunto de variables que se evaluó tanto con datos winsorizados como sin winsorizar, y las métricas de evaluación utilizadas incluyeron Recall, F1-Score y AUC-ROC.

CUADRO 6.1: Set Random Forest

Set	Recall	F1 Score	AUC-ROC
Sin winsorizar	0.18	0.3	0.86
Winsorizado	0.6	0.68	0.88

Las tablas de las otras metricas estan en A.3 para Kruskan-Wallis, A.4 Para Flinger-Killen, A.5 para Kolmogorov-Smirnov.

6.0.2. Iteración de variables candidatas con Reg. Log.

En la etapa actual del análisis, se enfoco en identificar un subconjunto de variables óptimo para la predicción de clases. Dada la alta complejidad computacional de examinar todas las combinaciones posibles de variables, se seleccionaron seis variables clave para cada macrozona con la puntuación mas alta de una matriz de criterios que garanticen variables potencialmente caracteristicas.

Se implemento un modelo de regresión logística, evaluado a través de validación cruzada LOOCV (Leave-One-Out Cross-Validation), escogida por su efectividad en conjuntos de datos con desequilibrio de clases, así como sacar el maximo provecho de las observaciones con el que el modelo se entrena, ya que en el contexto de la investigación las macrozonas no contienen a tantas comunas.

De las diversas combinaciones probadas, se eligieron las mejores basandose en tres métricas: AUC-ROC, F1 Score y valor Kappa. Estas métricas fueron seleccionadas por su relevancia

en contextos con desequilibrio de clases, ofreciendo una visión integral de la capacidad del modelo para clasificar correctamente las observaciones.

CUADRO 6.2: Industria Forestal

Métrica	Variables	Valor
AUC-ROC	Eq. Cultura; Empleo; Eq. Serv. Públicos; Delitos Leves a la Propiedad	0.73
F1 Score	Eq. Cultura; Empleo	0.44
Kappa	Eq. Cultura; Empleo	0.38

CUADRO 6.3: Industria Acuicultura

Métrica	Variables	Valor
AUC-ROC	Eq. Deporte; Suficiencia de Vivienda; Eq. Educación; Resilencia de Hogares	0.93
F1 Score	Eq. Deporte; Eq. Educación	0.83
Kappa	Eq. Deporte; Eq. Educación	0.71

CUADRO 6.4: Industria Minera

Métrica	Variables	Valor
AUC-ROC	Part. Juv. empleo y estudio; Conflictos Socioambientales; Eq. Educación	0.92
F1 Score	Part. Juv. empleo y estudio; Conflictos Socioambientales; Delitos Leves a la Propiedad; Eq. Educación; Eq. Deporte	0.9
Kappa	Part. Juv. empleo y estudio; Conflictos Socioambientales; Delitos Leves a la Propiedad; Eq. Educación; Eq. Deporte	0.86

6.0.3. Selección con SVM

Para abordar el desequilibrio de clases en el modelo SVM se siguieron las siguientes estrategias:

1. **Optimización de pesos:**Al ajustar los pesos, se otorga mayor importancia a las clases menos representadas, equilibrando así su influencia en el modelo.
2. **Kernel gaussiano (radial):** Se eligio este kernel por su capacidad para manejar la no linealidad en los datos, lo que es especialmente útil en conjuntos de datos complejos donde las relaciones entre clases no son claramente definidas o separables linealmente, como lo son los datos con los que se trabaja el estudio.
3. **Validación Cruzada LOOCV:** Esta validación cruzada maximiza la utilización de los datos disponibles. Al dejar una observación fuera para validar el modelo en cada iteración, se aprovecha al máximo el conjunto de datos limitado, lo que ayuda a obtener estimaciones más precisas del rendimiento del modelo, incluso en clases minoritarias.

Las métricas seleccionadas — Precisión, Recall, F1-Score, AUC-ROC y MCC — fueron elegidas específicamente por su relevancia en escenarios con desequilibrio de clases, se habla de ellas en 5.7

CUADRO 6.5: Resultados de clasificación para la Industria Forestal

Set	Precision	Recall	F1 Score	AUC-ROC	MCC
Set AUC-ROC	0.68	0.45	0.56	0.71	0.46
Set F1-Score, Kappa	0.64	0.18	0.29	0.57	0.25

CUADRO 6.6: Resultados de clasificación para la Industria Acuícola

Set	Precision	Recall	F1 Score	AUC-ROC	MCC
Set AUC-ROC	0.68	1.00	0.76	0.79	0.63
Set F1-Score, Kappa	0.90	0.62	0.73	0.79	0.64

CUADRO 6.7: Resultados de clasificación para la Industria Minera

Set	Precision	Recall	F1 Score	AUC-ROC	MCC
Set AUC-ROC	0.59	0.75	0.66	0.76	0.50
Set F1-Score, Kappa	0.73	0.79	0.76	0.84	0.69

6.1. Construcción del Indicador

6.1.1. Selección y Evaluación de Variables

En esta fase, se utilizo el método RandomForest para identificar la importancia relativa de las variables caracteristicas para predecir la clase de cada comuna.

6.1.2. Normalización de la Importancia Relativa

A continuación, se ajusto la importancia relativa de las variables caracteristicas para que la suma total sea igual a 1.

6.1.3. Elaboración y Normalización del Indicador

Basándose en la importancia relativa se elaboro un indicador mediante su suma ponderada. Las variables fueron previamente normalizadas para asegurar su alineación en una escala de 0 a 1.

6.1.4. Asignación de Signos y Ajuste Final del Indicador

Finalmente, se determino los signos de las variables y se ajusto el indicador para normalizar sus valores entre 0 y 1. Las tablas finales presentan este ajuste, es importante destacar que el signo que acompaña a cada variable caracteristica refleja el sentido en que la presencia de faenas extractivistas afecta dicha variable Los boxplot en que se baso la desición del signo que acompaña a cada variable caracteristica esta en A.0.6

6.1.5. Tablas de Rendimiento e Importancia relativa

Industria Forestal

CUADRO 6.8: Métricas de Evaluación

Métrica	Valor
F1 Score	0.47
AUC-ROC	0.77

CUADRO 6.9: Importancias del modelo

Variable	Importancia relativa	Importancia normalizada
Eq. Cultura	10.14	0.25
Empleo	12.83	0.29
Eq. Serv. Públicos	8.53	0.21
Delitos Leves a la Propiedad	10.64	0.23

Industria Acuicultura

CUADRO 6.10: Métricas de Evaluación

Métrica	Valor
F1 Score	0.61
AUC-ROC	0.79

CUADRO 6.11: Importancias del modelo

Variable	Importancia relativa	Importancia normalizada
Eq. Deporte	7.14	0.49
Eq. Educación	7.35	0.5

Industria Minera

CUADRO 6.12: Métricas de Evaluación

Métrica	Valor
F1 Score	0.33
AUC-ROC	0.71

CUADRO 6.13: Importancias del modelo

Variable	Importancia relativa	Importancia normalizada
Delitos Leves a la Propiedad	8.57	0.27
Part. Juv. empleo y estudio	8.4	0.26
Eq. Educación	6.6	0.2
Eq. Deporte	5.77	0.18
Conflictos Socioambientales	2.12	0.06

6.1.6. Tablas de signos

CUADRO 6.14: Signos por variable, Forestal

Variable	Signo en la fórmula del indicador
Eq. Cultura	positivo
Empleo	negativo
Eq. Educación	negativo
Delitos Leves a la Propiedad	positivo

CUADRO 6.15: Signos por variable, Acuicultura

Variable	Signo en la fórmula del indicador
Eq. Deporte	positivo
Eq. Educación	positivo

CUADRO 6.16: Signos por variable, Minera

Variable	Signo en la fórmula del indicador
Part. Juv. empleo y estudio	negativo
Conflictos Socioambientales	positivo
Eq. Educación	positivo
Delitos Leves a la Propiedad	negativo
Eq. Deporte	negativo

6.1.7. Interpretación de los Mapas de Resultado

los mapas estan en A.0.4.

Respuesta a las diferencias entre proporción y categorías en los mapas de rendimiento

En los mapas presentados en A.0.4, existen diferencias entre proporción y categorías con el indicador del rendimiento. Deberíamos considerar varias explicaciones para este fenómeno:

1. **Clasificación Discreta:** El proceso de clasificación se realizó de manera discreta, lo que implica que los datos se agruparon en categorías o intervalos específicos. Esta discretización puede resultar en la pérdida de matices y detalles en la representación de los

datos, especialmente en áreas donde los valores están cerca de los límites de las categorías.

2. **Factores Complejos:** La presencia de industria extractivista no es el único factor que influye en el comportamiento de las variables caracteristicas. Las variables que componen cada indicador pueden estar relacionadas de manera más compleja con sus causas, y esta investigación puede no captar todas las relaciones y factores relevantes que afectan el comportamiento de cada área.

3. **Elección de Signos de Variables:** La elección de los signos de las variables que conforman los indicadores puede influir en la interpretación de los resultados. En algunas áreas, los efectos locales pueden no estar completamente representados debido a estas elecciones.

4. **Delimitación de Zonas:** Los límites de las zonas se basaron en regiones, que son delimitaciones administrativas. Esto puede no reflejar adecuadamente las áreas donde la industria extractivista tiene un impacto significativo.

5. **No se busca Causalidad:** Es importante recordar que estos mapas no intentan establecer causalidad, sino proporcionar una medida que refleje en qué medida las distintas variables se desvían en presencia de la industria extractivista. Por lo tanto, se enfocan en patrones y asociaciones, pero no en relaciones causales directas.

Capítulo 7

Conclusiones

La presente investigación ha validado la hipótesis de que los impactos de las industrias extractivistas en las comunas chilenas varían significativamente según el sector industrial. Los objetivos planteados se han cumplido satisfactoriamente, evidenciando la variabilidad de los efectos de cada industria y demostrando la viabilidad de su cuantificación mediante técnicas de machine learning.

El objetivo general de desarrollar un indicador que clasifique las comunas según el grado de impacto de la industria dominante se alcanzó a través de un meticuloso proceso de selección de variables y modelado estadístico. Los objetivos específicos, incluyendo la estandarización de las bases de datos, la delimitación de macrozonas, la gestión de outliers y la construcción de un indicador integral, se cumplieron integralmente, proporcionando un modelo robusto y confiable para la evaluación del impacto industrial.

La hipótesis se vio confirmada por el análisis empírico y las técnicas de modelización utilizadas, subrayando la importancia de adaptar las políticas de desarrollo local a las características únicas de cada industria extractiva. El Indicador desarrollado representa un avance significativo en la metodología de evaluación del impacto industrial y puede servir como una herramienta esencial para la toma de decisiones informadas en el contexto ambiental y de desarrollo.

Al encontrar los impactos diferenciados, se abre el debate en torno a la discusión sobre el extractivismo, destacando la necesidad de políticas de desarrollo local más específicas y adaptadas a las características de cada industria. Como líneas futuras de investigación, se sugiere la aplicación del indicador en estudios longitudinales para evaluar la dinámica del impacto industrial a lo largo del tiempo. Además, sería provechoso explorar la extensión del modelo a otras industrias y regiones, así como la integración de datos de percepción pública y salud ambiental para enriquecer la comprensión de los efectos intangibles de la actividad extractivista.

Apéndice A

Material Complementario de la Investigación

A.0.1. Estructura de Matriz de Criterios

	Método de Evaluación						
	Random Forest	MIC	Cor. Biseral	Kruskal-Wallis	Flinger-Killen	Kolmogorov-Smirnov	Puntuación
Variable 1							
Variable 2							
Variable 3							
Variable 4							

CUADRO A.1: Estructura de Matriz de Criterios

A.0.2. Proceso de Selección de Variables Caracteristicas

Industria	**Matriz de Criterios**	**Iteración Reg. Log.**	**Selección SVM**
Forestal	Eq. Cultural; Empleo; Eq. Serv. Públicos; Suficiencia de Vivienda; Part. Juv. empleo y estudio; Delitos Leves a la Propiedad	Eq. Cultura; Empleo; Eq. Serv. Públicos; Delitos Leves a la Propiedad	Eq. Cultura; Empleo; Eq. Serv. Públicos; Delitos Leves a la Propiedad
		Eq. Cultura; Empleo	
Acuicultura	Empleo; Eq. Deportes; Suficiencia de Vivienda; Conflictos Socioambientales; Eq. Educación; Resilencia de Hogares	Eq. Deporte; Suficiencia de Vivienda; Eq. Educación; Resilencia de Hogares	Eq. Deporte; Eq. Educación
		Eq. Deporte; Eq. Educación	
Minera	Part. Juv. empleo y estudio; Conflictos Socioambientales; Empleo; Delitos Leves a la Propiedad; Eq. Educación; Eq. Deportes	Part. Juv. empleo y estudio; Conflictos Socioambientales; Eq. Educación	Part. Juv. empleo y estudio; Conflictos Socioambientales; Eq. Educación; Delitos Leves a la Propiedad; Eq. Deporte
		Part. Juv. empleo y estudio; Eq. Educación; Delitos Leves a la Propiedad; Eq. Deporte	

CUADRO A.2: Selección de Variables

A.0.3. Diagrama de Metodologia

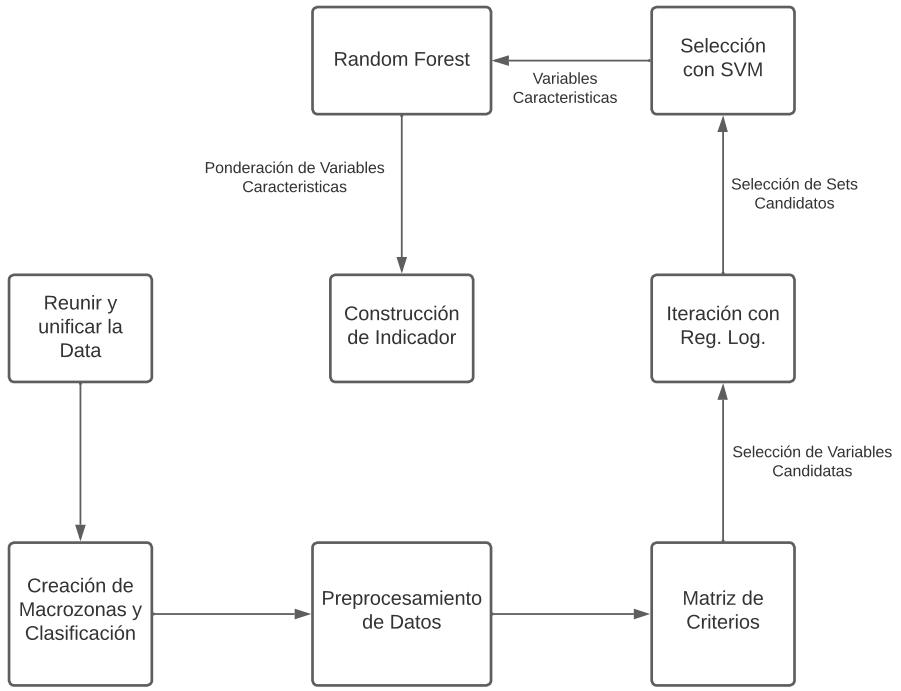
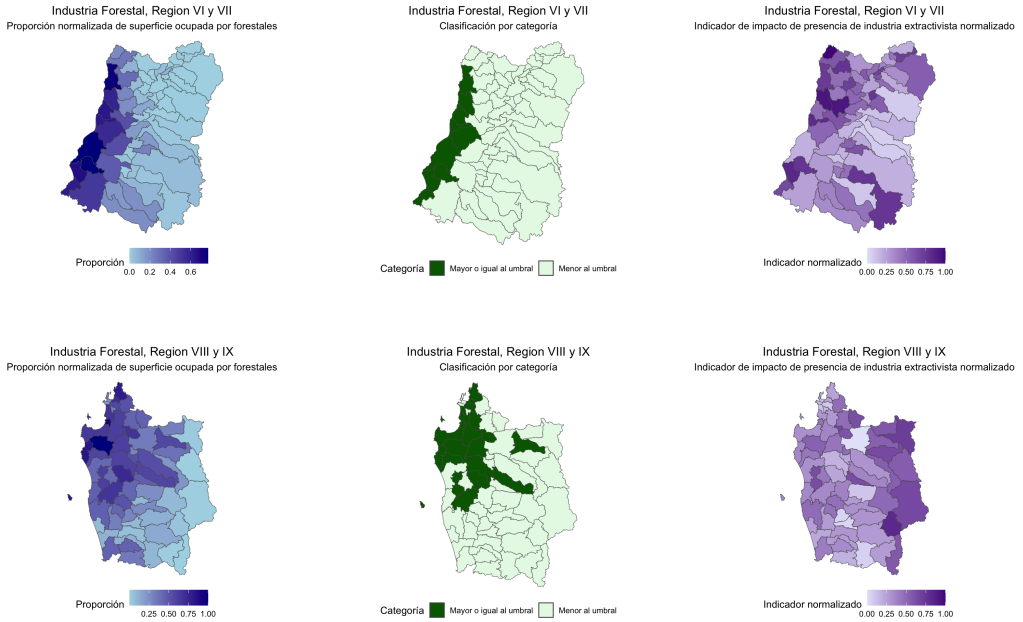
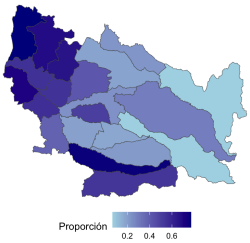


FIGURA A.1: Diagrama de Metodología.

A.0.4. Mapas de resultados

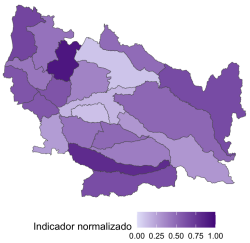


Industria Forestal, Region XVI
Proporción normalizada de superficie ocupada por forestales

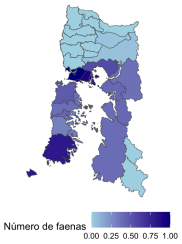


Industria Forestal, Region XVI
Clasificación por categoría

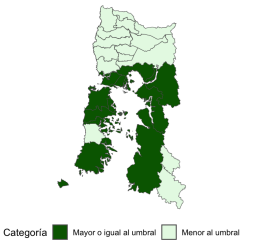
Industria Forestal, Region XVI
Indicador de impacto de presencia de industria extractivista normalizado



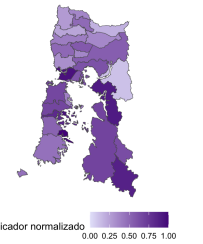
Industria Acuicultura, Región X
Número de faenas normalizado



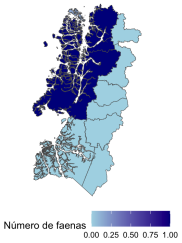
Industria Acuicultura, Region X
Clasificación por categoría



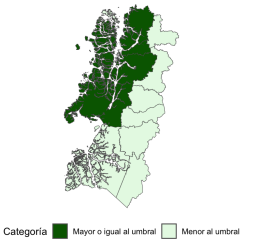
Industria Acuicultura, Region X
Indicador de impacto de presencia de industria extractivista normalizado



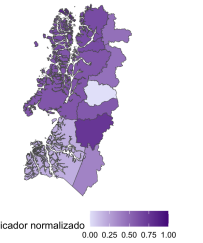
Industria Acuicultura, Región XI
Número de faenas normalizado



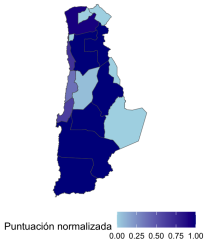
Industria Acuicultura, Region XI
Clasificación por categoría



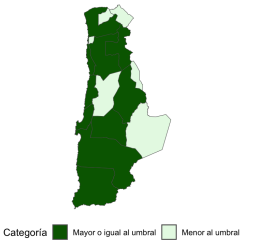
Industria Acuicultura, Region XI
Indicador de impacto de presencia de industria extractivista normalizado



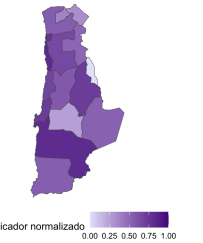
Industria Minera, Región I y II
Puntuación normalizada



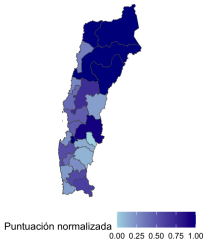
Industria Minera, Region I y II
Clasificación por categoría



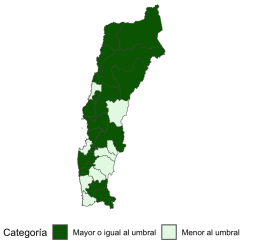
Industria Minera, Region I y II
Indicador de impacto de presencia de industria extractivista normalizado



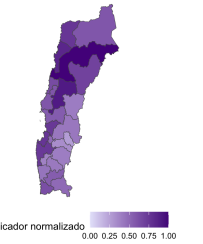
Industria Minera, Región III y IV
Puntuación normalizada

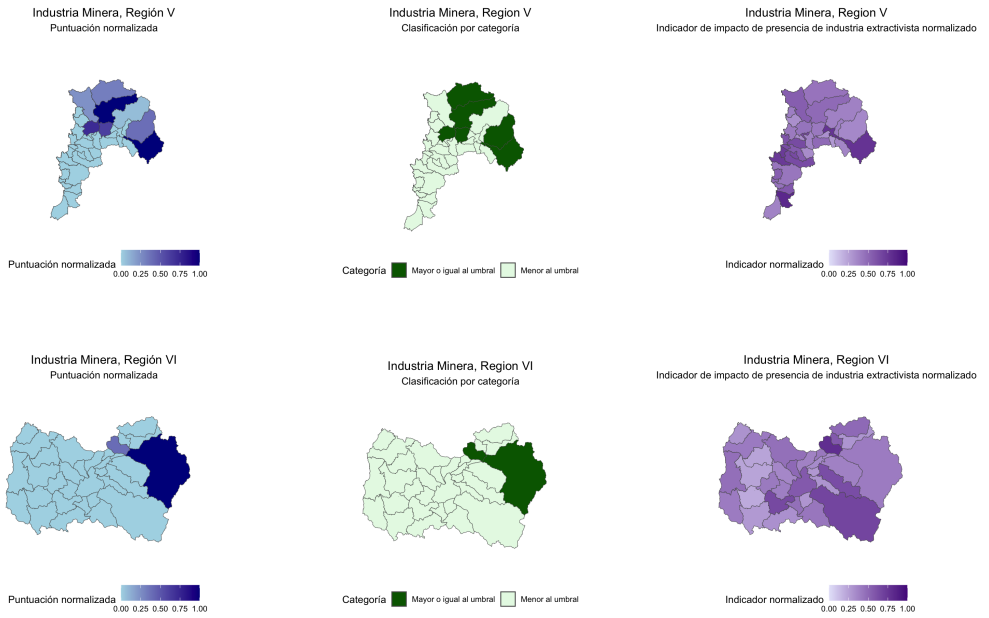


Industria Minera, Region III y IV
Clasificación por categoría



Industria Minera, Region III y IV
Indicador de impacto de presencia de industria extractivista normalizado





A.0.5. Tablas complementarias

CUADRO A.3: Set Kruskan-Wallis

Set	Recall	F1 Score	AUC-ROC
Sin winsorizar	0.57	0.66	0.85
Winsorizado	0.56	0.62	0.86

CUADRO A.4: Set Flinger-Killen

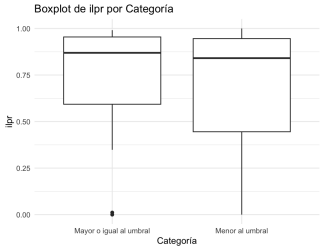
Set	Recall	F1 Score	AUC-ROC
Sin winsorizar	0.45	0.56	0.79
Winsorizado	0.5	0.62	0.83

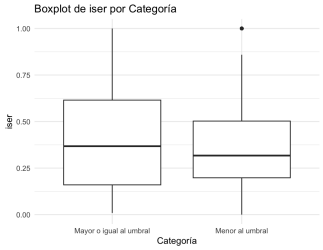
CUADRO A.5: Set Kolmogorov-Smirnov

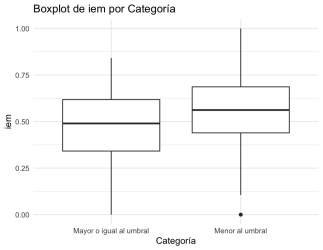
Set	Recall	F1 Score	AUC-ROC
Sin winsorizar	0.52	0.6	0.84
Winsorizado	0.52	0.62	0.84

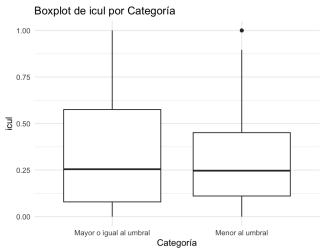
A.0.6. Boxplots para la decisión del signo del indicador

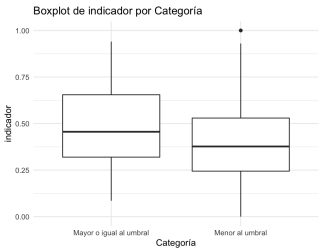
Forestal



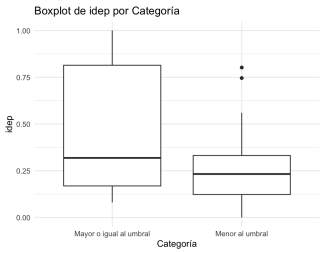


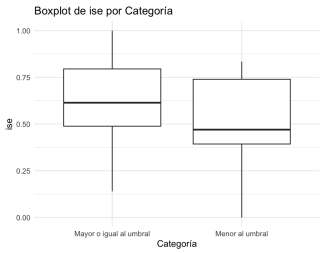


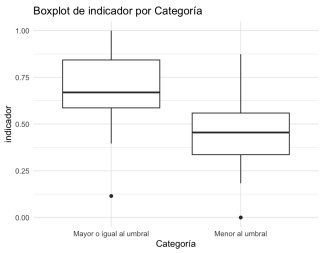




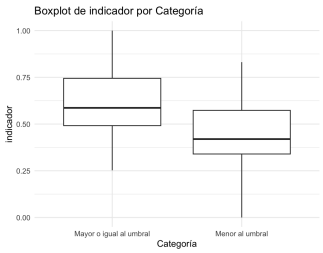
Acuicultura

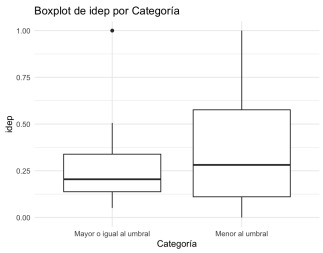


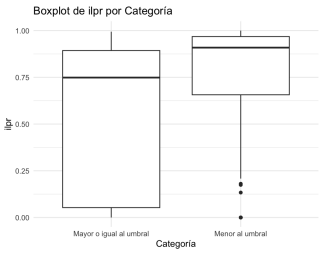


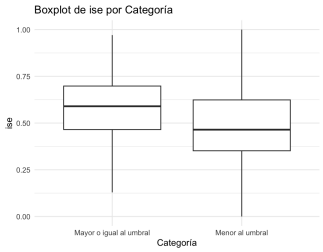


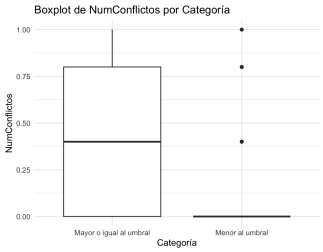
Minera

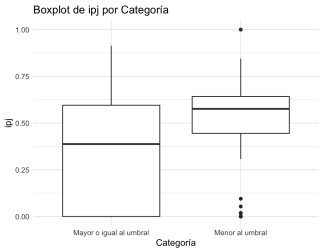












Bibliografía

Arboleda, M. (2020). *Planetary Mine: Territories of Extraction Under Late Capitalism*. 1.ª ed. New York: Verso.

Arias-Loyola, Martín et al. (2022). «Beyond the resource curse: The redistributive challenge of sustainable resource-led development in Australia, Chile and Zambia». En: *The Extractive Industries and Society* 11, pág. 101084. DOI: https://doi.org/10.1016/j.exis.2022.101084.

Baayen, R. H. (2010). «Analyzing linguistic data: A practical introduction to statistics using R». En: *Journal of Child Language* 37, págs. 465-470. DOI: 810.1017/S0305000909990080.

Bebbington, A. (2009). «Latin America: contesting extraction, producing geographies». En: *Singapore Journal of Tropical Geography* 30.1, págs. 7-12.

Bolados, Paola (2016). «Conflictos socio-ambientales/territoriales y el surgimiento de identidades post neoliberales (Valparaíso-Chile)». En: *Izquierdas* 31, págs. 102-129.

Bolados, Paola y Sally Babidge (2017). «Ritualidad y extractivismo. La limpia de canales y las disputas por el agua en el salar de atacama-norte de Chile». En: *Estudios Atacameños* 54, págs. 201-216.

Bonett, Douglas G (2020). «Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination». En: *PubMed*.

Bustos-Gallardo, B. (2017). «The post 2008 Chilean Salmon Industry: an example of an enclave economy». En: *The Geographical Journal* 183.2, págs. 152-163.

Chicco, Davide y Giuseppe Jurman (2020). «The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation». En: *BMC Genomics* 21, pág. 6. DOI: 10.1186/s12864-019-6413-7.

Cohen, Jacob (1960). «A coefficient of agreement for nominal scales». En: *Educational and psychological measurement* 20.1, págs. 37-46.

Cooley, David (2022). *googleway: Access Google Maps API*. R package version 4.0.1. URL: https://cran.r-project.org/web/packages/googleway/index.html.

Durai, Senthil y Mary Shamili (2022). «Smart farming using Machine Learning and Deep Learning techniques». En: *Decision Analytics Journal* 3, pág. 100041. DOI: https://doi.org/10.1016/j.dajour.2022.100041.

Gamer, Matthias, Jim Lemon y Ian Fellows Puspendra (2019). *Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. URL: https://CRAN.R-project.org/package=irr.

Gregorutti, Baptiste, Bertrand Michel y Philippe Saint-Pierre (2017). «Correlation and variable importance in random forests». En: *Statistics and Computing* 27.3, págs. 659-678. DOI: https://doi.org/10.1007/s11222-016-9646-1.

Grosfoguel, Ramón (2015). «Del extractivismo económico al extractivismo epistémico y ontológico». En: *Revista Internacional de Comunicación y Desarrollo* 4, págs. 33-45.

Hatzold, Max-Emanuel (2013). «Social Conflict, Economic Development and Extractive Industry: Evidence from South America». En: *Community Development Journal* 48.3. URL: https://academic.oup.com/cdj/article-abstract/48/3/401/299547.

Heller, H. (2014). *Teoria del Estado*. 2.ª ed. Mexico: Fondo de Cultura Economica.

Heredia, Jobany, Aida Rodriguez y Jose Vilalta (2012). «Empleo de la regresión logística ordinal para la predicción del rendimiento academico». En: *Revista Investigación Operacional* 33, págs. 252-267.

Herrera, E. G. Rejón et al. (2015). «Clasificación de Indicadores de Interacción del uso de la plataforma Moodle para cursos de modalidad B-learning». En: *Tecnología Educativa* 2.1, Artículo 8. DOI: 10.32671/terc.v2i1.170. URL: https://terc.mx/index.php/terc/article/view/170.

Hsieh, William W. (2023). *Decision Trees, Random Forests and Boosting*. Cambridge University Press, págs. 473-493. DOI: 10.1017/9781107588493.015. URL: https://doi.org/10.1017/9781107588493.015.

Jr., David W. Hosmer, Stanley Lemeshow y Rodney X. Sturdivant (2013). *Applied Logistic Regression*. John Wiley y Sons, Inc. ISBN: 9780470582473. DOI: 10.1002/9781118548387. URL: https://doi.org/10.1002/9781118548387.

Kinney, Justin y Gurinder Atwal (2014). «Equitability, mutual information, and the maximal information coefficient». En: *Proceedings of the National Academy of Sciences* 111.9, págs. 3354-3359. DOI: 10.1073/pnas.1309933111.

Kornbrot, Diana (2005). «Point Biserial Correlation». En: *Encyclopedia of Statistics in Behavioral Science*. John Wiley y Sons, Ltd. ISBN: 9780470013199. DOI: 10.1002/0470013192.bsa485. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa485.

Kuhn, Max (2022). *caret: Classification and Regression Training*. R package version 6.0-91. URL: https://CRAN.R-project.org/package=caret.

Kumar, Sanjeev et al. (2023). «An outliers detection and elimination framework in classification task of data mining». En: *Decision Analytics Journal* 6, pág. 100164. DOI: https://doi.org/10.1016/j.dajour.2023.100164.

Liaw, Andy y Matthew Wiener (2002). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-14. URL: https://CRAN.R-project.org/package=randomForest.

Lizares Castillo, Mónica (2017). «Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico». Tesina. Lima, Perú: Universidad Nacional Mayor de San Marcos.

Maillet, Antoine et al. (2021). «Conflicto, territorio y extractivismo en Chile. Aportes y límites de la producción académica reciente». En: *Revista de Geografía Norte Grande* 80, págs. 59-80.

Meyer, David et al. (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-9. URL: https://CRAN.R-project.org/package=e1071.

Nicklin, Christopher y Luke Plonsky (2020). «Outliers in L2 Research in Applied Linguistics: A Synthesis and Data Re-Analysis». En: *Annual Review of Applied Linguistics* 40, págs. 26-55. DOI: https://doi.org/10.1017/S0267190520000057.

Nyitrai, Tamás y Virág Miklós (2018). «The effects of handling outliers on the performance of bankruptcy prediction models». En: *Socio-Economic Planning Sciences* 67, págs. 34-42. DOI: https://doi.org/10.1016/j.seps.2018.08.004.

OECD Territorial Reviews: Chile (2009). Inf. téc. Retrieved from. OECD.

Pino, Anyela y Noelia Carrasco (2018). «Extractivismo forestal en la comuna de Arauco (Chile): internalización y formas de resistencia». En: *Rev. Colomb. Soc.* 42, págs. 207-226.

Reshef, David et al. (2013). «Equitability Analysis of the Maximal Information Coefficient, with Comparisons». En: *ArXiv* abs/1301.6314. URL: https://api.semanticscholar.org/CorpusID:9609553.

Reyes, Luis et al. (2019). «Sistema de clasificación SVM de señales electromiográficas extraídas en un sistema embebido». En: *Research in Computing Science* 148, págs. 135-141.

Robin, Xavier et al. (2021). *pROC: Display and Analyze ROC Curves*. R package version 1.18.0. URL: https://CRAN.R-project.org/package=pROC.

Romero-Toledo, Hugo (2019). «Extractivismo en Chile: la producción del territorio minero y las luchas del pueblo aimara en el Norte Grande». En: *Colombia Internacional* 98, págs. 03-30. DOI: https://doi.org/10.7440/colombiaint98.2019.01.

Sullivan, Joe, Linda Wallace y Merril Warkentin (2021). «So Many Ways to Assess Outliers: What Really Works and Does it Matters?» En: *Journal of Business Research* 132, págs. 530-543. DOI: https://doi.org/10.1016/j.jbusres.2021.03.066.

Svampa, Maristella y Mirta Antonelli, eds. (2009). *Minería transnacional, narrativas del desarrollo y resistencias sociales*. Buenos Aires: Biblos. ISBN: 978-950-786-709-5.

Syed, Shahan et al. (2018). «Prediction of Stock Performance by Using Logistic Regression Model: Evidence from Pakistan Stock Exchange (PSX)». En: *Asian Journal of Empirical Research* 8, págs. 247-258. DOI: https://doi.org/10.18488/journal.1007/2018.8.7/1007.7.247.258.

U.S. Geological Survey (USGS) (2020). *Minerals Commodities Summaries 2020*. Inf. téc. Washington, DC: U.S. Geological Survey.

World Bank e International Finance Corporation (2002). *Global Mining Treasure or Trouble? Mining in Developing Countries*. Inf. téc. Retrieved from. Washington, D.C.

Zaidi, Makram y Amina Amirat Grace Ofori-Abebrese (2016). «Forecasting Stock Market Trends by Logistic Regression and Neural Networks: Evidence from KSA Stock Market». En: URL: https://api.semanticscholar.org/CorpusID:37561825.

Zamudio, Alejandro et al. (2021). «Burnout en profesionales de la salud en contexto de pandemia: Una propuesta metodológica para la detección de patrones basada en inteligencia artificial». En: *Revista Digital Internacional de Psicología y Ciencia Social* 7, e-ISSN 2448-8119.