Universidad del Desarrollo
Facultad de Ingeniería

Diversificación de Rubros en Licitaciones Públicas

Un Enfoque Basado en Cálculo de Density Relatedness y Machine Learning para Maximizar el Éxito de la Diversificación en Rubros Potenciales

POR: RICARDO ALBERTO CHACON ACOSTA

MATIAS JOSE MERCANDINO SEPULVEDA

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dr. CRISTIAN ESTEBAN CANDIA VALLEJOS

Enero 2025

SANTIAGO

Dedicamos este trabajo a nuestras familias, quienes con su apoyo incondicional, amor y motivación constante hicieron posible que llegáramos hasta aquí. También, a todos aquellos que nos inspiraron a superar los desafíos y a crecer durante este proceso.

AGRADECIMIENTO

Queremos expresar nuestro más profundo agradecimiento a todas las personas que contribuyeron a la realización de este trabajo. A nuestros asesores, Cristian Candia Vallejos y Adolfo Fuentes Jofre, por su guía y consejos que enriquecieron nuestro proyecto.

Agradecemos también a nuestros compañeros, Matias Bunster y Nicolas Lagos , quienes con su apoyo y colaboración hicieron de esta experiencia algo aún más significativo.

Finalmente, agradecemos a nuestras familias por ser nuestra mayor fortaleza y brindarnos siempre el respaldo necesario para alcanzar nuestras metas.

TABLA DE CONTENIDO

Resumen

En el contexto de las licitaciones públicas en Chile entre 2014 y 2023, las empresas buscan constantemente oportunidades para diversificar y aumentar su éxito en futuros procesos de licitación. Este estudio tiene como objetivo explorar cómo las empresas podrían utilizar modelos de *machine learning* para clasificar y evaluar automáticamente oportunidades de diversificación hacia rubros relacionados, basándose en el concepto de *density relatedness*. A través de un análisis planificado de datos históricos de licitaciones públicas, se desarrollaron metodologías para calcular estas métricas y se aplicó un modelo predictivo que permitirá identificar las mejores oportunidades de expansión. Se espera que los resultados demuestren que una alta *density relatedness* indica que el rubro potencial (r') está estrechamente relacionado con los rubros actuales del proveedor, lo que sugiere una mayor probabilidad de éxito en la diversificación hacia ese rubro.

1. Introducción

La diversificación estratégica es esencial para el crecimiento y la sostenibilidad de las empresas en mercados competitivos. Igor Ansoff, en su artículo "Strategies for Diversification" (1957), señala que para que las empresas mantengan su posición en mercados competitivos, es fundamental que crezcan y se adapten continuamente. Según Ansoff, la diversificación es una de las cuatro principales estrategias de crecimiento disponibles para las empresas, junto con el desarrollo de productos, penetración de mercados y la expansión de mercados. La estrategia de diversificación es clave para que las organizaciones no solo mantengan su competitividad, sino también para que se posicionen de manera sólida en industrias en constante evolución. En el ámbito de las licitaciones públicas en Chile, las empresas enfrentan el desafío de identificar oportunidades que no solo sean rentables, sino también alineadas con sus capacidades y experiencias existentes. El concepto de *relatedness*, propuesto por Neffke, Henning, y Boschma (2011), mide la similitud o conexión entre dos productos o servicios, basándose en la frecuencia con la que aparecen juntos en las ofertas de los proveedores. Esta co-ocurrencia refleja la proximidad entre actividades económicas, lo que facilita que las empresas se diversifiquen hacia nuevos productos o servicios relacionados, aprovechando sus capacidades.

Este estudio tiene como objetivo aplicar modelos de *machine learning* para cuantificar métricas clave, como el *density relatedness*, y desarrollar un sistema automatizado que ayude a las empresas a tomar decisiones estratégicas informadas. La pregunta de investigación central que guía este trabajo es: **¿Cómo pueden las empresas utilizar modelos de *machine learning* para clasificar y evaluar automáticamente oportunidades de diversificación hacia rubros relacionados, basándose en el concepto de *density relatedness*, aplicados a datos históricos de licitaciones públicas chilenas desde 2014 hasta 2023, con el objetivo de maximizar su éxito en futuros procesos de licitación?.**

La motivación de este trabajo radica en mejorar el proceso de toma de decisiones estratégicas en licitaciones públicas para las empresas en Chile que buscan diversificarse. A través de este proyecto, se busca aportar valor facilitando que las empresas identifiquen y aprovechen eficazmente oportunidades de diversificación, lo que aumentaría sus probabilidades de éxito en futuros procesos de licitación, ayudándoles a maximizar su competitividad y adaptabilidad en mercados en evolución.

2. Trabajo Relacionado

La teoría de la complejidad económica ha sido fundamental para explicar patrones de diversificación tanto en economías como en empresas. Hidalgo y Hausmann (2009) desarrollaron el concepto de *complejidad económica* para cuantificar la estructura económica de los países y cómo esto impulsa el crecimiento. A su vez, Neffke, Henning y Boschma (2011) abordaron el concepto de *relatedness* para explorar la diversificación regional, mostrando cómo la similitud entre sectores fomenta el éxito económico. Estos enfoques complementan la estrategia de diversificación de Ansoff (1957). En este contexto, LicitaLAB, en su página web https://www.licitalab.cl/data-science-compras-publicas/, señala que utiliza ciencia de datos para mejorar los procesos de licitaciones públicas mediante el uso de machine learning. Esta metodología busca optimizar la transparencia y eficiencia en las compras públicas, lo que resulta relevante para el análisis de licitaciones en el marco de este estudio. El enfoque de LicitaLAB subraya cómo los datos pueden ser clave para facilitar la toma de decisiones estratégicas en el ámbito empresarial.

3. Hipótesis y Objetivos

Hipótesis:

- **H1:** Los modelos de machine learning, al utilizar la métrica density relatedness, pueden clasificar oportunidades de diversificación que garanticen el éxito de las empresas al expandirse hacia nuevos rubros relacionados en licitaciones públicas.

Objetivo General:

Desarrollar un modelo basado en machine learning que utiliza la métrica *density relatedness* para clasificar oportunidades de diversificación en licitaciones públicas, permitiendo a las empresas identificar rubros relacionados que maximicen su probabilidad de éxito al expandirse hacia nuevos rubros.

Objetivos Específicos:

1. **Calcular la Métrica de Density Relatedness:** Analizar datos históricos de licitaciones públicas en Chile entre 2014 y 2023 para calcular la métrica de *density relatedness* entre distintos rubros, identificando la proximidad y conexión entre ellos.
2. **Implementar Modelos de Machine Learning:** Diseñar y entrenar modelos predictivos que incorporan la métrica de *density relatedness* como variable clave, con el objetivo de predecir las mejores oportunidades de diversificación para empresas específicas.
3. **Validar la Hipótesis y Evaluar el Modelo:** Evaluar el desempeño de los modelos predictivos mediante métricas de éxito en licitaciones públicas, confirmando si el uso de *density relatedness* efectivamente maximiza las oportunidades de diversificación y competitividad de las empresas.

4. Datos y Metodología

4.1. Datos

Fuente de datos: La base de datos empleada en esta investigación fue extraída íntegramente del portal de compras públicas de Chile, Mercado Público. Se recolectó información detallada de todas las licitaciones publicadas en el periodo comprendido entre 2014 y 2023. Esta fuente proporciona un conjunto de datos exhaustivo, incluyendo información sobre los proveedores participantes, los rubros a los que postulaban, los montos adjudicados y otras variables relevantes.

Variables: Para este estudio, se seleccionó y construyó un conjunto completo de variables con el objetivo de caracterizar tanto a los proveedores como a los diferentes rubros de productos y servicios licitados, así como de cuantificar la relación entre estos elementos.

La siguiente tabla presenta las principales variables utilizadas en este estudio, extraídas de la base de datos de Mercado Público:

Tabla 1: Diccionario de variables que provienen del conjunto de datos que entrega mercado público.

Variable	Descripción
Código	Identificador único asignado a cada licitación.
NombreOrganismo	Nombre de la entidad pública (organismo, servicio, municipio, etc.) que realiza la licitación y contrata los bienes o servicios.
FechaPublicación	Fecha exacta en la que se publica la licitación en el portal de Mercado Público, iniciando el proceso de postulación para los proveedores.
CodigoProductoONU	Código estándar utilizado para clasificar los productos o servicios licitados, siguiendo la nomenclatura internacional de productos (Clasificación Uniforme de Productos).
NombreProveedor	Razón social o nombre comercial del proveedor que presenta una oferta para participar en la licitación.
MontoLineaAdjudica	Valor monetario total adjudicado al proveedor ganador de la licitación, correspondiente al costo de los bienes o servicios contratados.
OfertaSeleccionada	Variable categórica que indica si la oferta presentada por un proveedor específico fue seleccionada como ganadora de la licitación (1 = seleccionada, 0 = no seleccionada).
Año	Año en el que se publicó la licitación, permitiendo realizar análisis temporales y comparaciones entre diferentes periodos.
NombreProducto	Descripción textual del producto o servicio objeto de la licitación, complementando la información proporcionada por el código del producto ONU.

Enriquecimiento de datos: Con el objetivo de profundizar en el análisis de la diversificación de proveedores y facilitar la interpretación de los resultados, se procedió a enriquecer la base de datos original. Se incorporó la variable 'RubroN2', obtenida a partir de los datos de Mercado Público, la cual representa una categorización más general de los productos y servicios licitados. Esta variable permitió reducir la dimensionalidad del análisis y agrupar los rubros en categorías más amplias, facilitando así el estudio de las relaciones entre proveedores y sectores.

Limpieza y transformación: En esta etapa del proceso de investigación se llevó a cabo una exhaustiva limpieza y transformación de los datos relacionados con los nombres de los proveedores. El objetivo principal fue garantizar la calidad y consistencia de esta variable, fundamental para los análisis posteriores.

- **Eliminación de registros inválidos:** Se eliminaron aquellos registros donde el campo "NombreProveedor" contenía valores vacíos, nulos o caracteres especiales que no corresponden a nombres de empresas.
- **Normalización y estandarización:** Se aplicó una serie de transformaciones a los nombres de proveedores, incluyendo:
    - Eliminación de acentos y caracteres especiales.
    - Conversión a minúsculas.
    - Eliminación de caracteres no alfanuméricos.
    - Eliminación de palabras comunes en nombres de empresas (e.g., "SPA", "LTDA").
    - Unificación de formatos (por ejemplo, eliminación de espacios en blanco adicionales).

Generación de Variables Derivadas: Después de la etapa de limpieza y transformación de datos, se procedió a la generación de variables adicionales que permiten obtener información más detallada sobre los proveedores y los rubros. Este proceso es fundamental para describir y analizar sus comportamientos en el contexto de las licitaciones públicas, facilitando la identificación de patrones y tendencias relevantes para el estudio.

Tabla 2: Diccionario de variables calculadas utilizadas.

Variable	**Descripción**
TotalAdjudicacionesProveedor	Cantidad total de adjudicaciones obtenidas por el proveedor.
TotalParticipacionesProveedor	Cantidad total de participaciones del proveedor en procesos de compra.
%AdjudicacionProveedor	Porcentaje de adjudicación del proveedor (relación entre adjudicaciones y participaciones).
TotalMontoAdjudicadoProveedor	Monto total adjudicado al proveedor a lo largo de todos sus contratos.
TotalMontoPromedioAdjudicadoProveedor	Promedio del monto adjudicado al proveedor por cada adjudicación obtenida.
NumeroRubrosConVentaja	Número de rubros o categorías en las que el proveedor presenta una ventaja comparativa.
TotalMontoAdjudicadoRubro	Monto total adjudicado en el rubro específico.
TotalMontoPromedioAdjudicadoRubro	Monto promedio adjudicado en el rubro específico.
NumeroProveedoresRubro	Cantidad de proveedores que participan en el rubro especificado.

Estas variables calculadas proporcionan indicadores clave sobre la actividad de los proveedores, su nivel de participación en diferentes rubros y su grado de especialización o diversificación en el mercado. Al enriquecer el conjunto de datos con estas métricas, se habilita un análisis más profundo y preciso de los factores que influyen en el éxito de los proveedores en las licitaciones.

Cálculo de la Variable de Éxito para la Diversificación: Se creó una variable binaria para evaluar el éxito en la diversificación de los proveedores, considerando que un proveedor es exitoso si alcanza una Ventaja Comparativa Revelada con valor VCR ≥ 1 en el rubro potencial entre 2019 y 2023.

4.2. Metodología

Este proyecto busca identificar oportunidades de diversificación hacia rubros relacionados para los proveedores en el contexto de licitaciones públicas chilenas, utilizando modelos de machine learning y el concepto de density relatedness. Para ello, se desarrolló la siguiente metodología:

4.2.1. Análisis Descriptivo

Como paso inicial, se realizó un análisis descriptivo de la base de datos para comprender su composición y características principales. Este análisis incluyó:

- **Cantidad total de registros:** Se contabilizó el número total de registros en la base de datos, proporcionando una visión general del tamaño del conjunto de datos.
- **Distribución de licitaciones por rubro:** Se identificaron los 10 rubros con mayor cantidad de licitaciones, mostrando su porcentaje relativo respecto al total. Este análisis permitió identificar los sectores más dinámicos en el contexto de las licitaciones públicas.
- **Participación de proveedores:** Se analizó la cantidad de licitaciones por proveedor, destacando los 10 proveedores con mayor participación y su porcentaje relativo. Esto facilitó la identificación de actores clave en el mercado.
- **Porcentaje de adjudicaciones por proveedor:** Se evaluó el porcentaje de éxito (adjudicaciones) de los proveedores, proporcionando una medida de su competitividad en el mercado de licitaciones.

Este análisis permitió comprender las características iniciales de los datos y sentó las bases para el desarrollo de las etapas posteriores en el estudio.

4.2.2. Cálculo Ventaja Comparativa Revelada (VCR)

Para identificar en qué rubros los proveedores poseen una ventaja comparativa, se calculó la Ventaja Comparativa Revelada (VCR) para cada proveedor *p* en cada rubro *r*. Mide el grado de especialización de un proveedor en un rubro específico en comparación con su desempeño general en todos los rubros.

La VCR se calcula utilizando una adaptación del Índice de Balassa (Balassa, 1965), como se muestra a continuación:

$$VCR_{p,r} = \left(\frac{X_{p,r}}{\sum_{r'} X_{p,r'}} \Big/ \frac{\sum_{p'} X_{p',r}}{\sum_{p',r'} X_{p',r'}} \right)$$

- $X_{p,r}$ Es el número de licitaciones en las que el proveedor p participó en el rubro r.
- $\Sigma_{r'} X_{p,r'}$ Es el total de licitaciones en las que participó el proveedor p en todos los rubros.
- $\Sigma_{p'} X_{p',r}$ Es el total de licitaciones de todos los proveedores en el rubro r.
- $\Sigma_{p',r'} X_{p',r'}$ Es el número total de licitaciones en todos los rubros y proveedores.

Una VCR mayor a 1 indica que la empresa tiene una ventaja comparativa en ese rubro, es decir, está más especializada en comparación con el promedio.

4.2.3. Cálculo de la matriz de especialización

En el cálculo del VCR, las licitaciones se normalizan con respecto al total del proveedor y del rubro. Para obtener una variable binaria, decimos que un proveedor p participa competitivamente en un rubro r si su VCR es ≥ 1; con este umbral, y siguiendo la metodología de Hidalgo (2021), definimos la matriz binaria M_{pr} de la siguiente manera:

$$M_{pr} = \begin{cases} 1 & \text{si } VCR_{pr} \geq 1 \\ 0 & \text{si } VCR_{pr} < 1 \end{cases}$$

Donde:

- M_{pr} es la matriz de especialización para el proveedor p en el rubro r.
- VCR_{pr} representa la ventaja comparativa revelada para el proveedor p en el rubro r.

4.2.4. Cálculo de Proximidad

Una vez se obtuvo la matriz de especialización, se mide la proximidad o similitud entre diferentes rubros para identificar oportunidades de diversificación hacia áreas relacionadas. Esta proximidad se calcula utilizando la probabilidad condicional de que un proveedor que tiene ventaja comparativa en un rubro también la tenga en otro. La fórmula, basada en Hidalgo (2021), es:

$$\phi_{r,r'} = \frac{\sum_p M_{pr} M_{pr'}}{\max(M_r, M_{r'})}$$

- $\phi_{r,r'}$ es la proximidad entre los rubros r y r'.
- $\Sigma_p M_{pr}\ M_{pr'}$ es la suma de los proveedores que tienen ventaja comparativa en ambos rubros.
- $max\ (M_r\ ,\ M_{r'})$ es el valor máximo entre el número total de proveedores que tienen ventaja comparativa en el rubro r y el rubro r'.

Esta medida permite cuantificar qué tan relacionados están dos rubros basándose en las actividades reales de los proveedores.

4.2.5. Cálculo de Density Relatedness

Una vez calculada la proximidad, el siguiente paso es calcular el "density relatedness", que evalúa cuán conectado está un rubro potencial con los rubros en los que el

proveedor ya tiene ventaja comparativa. La fórmula, basada y fundamentada en el trabajo de Hidalgo (2021), y es:

$$\omega_{p,r'} = \frac{\sum_r M_{pr} * \emptyset_{rr'}}{\sum_r \emptyset_{rr'}}$$

Donde:

- $\omega_{p,r'}$ es la densidad para el proveedor p en el rubro r´.
- M_{pr} es el conjunto de rubros en los que el proveedor p tiene ventaja comparativa.
- $\phi_{r,r'}$ es la proximidad entre los rubros r y r´.

Una densidad alta sugiere que el rubro r' está estrechamente relacionado con los rubros actuales del proveedor, lo que indica una mayor probabilidad de éxito en la diversificación hacia ese rubro.

4.2.6. Definición de la Variable Objetivo

Antes de desarrollar el modelo de machine learning, fue fundamental definir una etiqueta clara que represente el objetivo de predicción: identificar si un proveedor logró diversificarse exitosamente hacia rubros potenciales. Este proceso se basó en los datos históricos futuros de licitaciones públicas del período 2019-2023, lo que permitió evaluar el desempeño de los proveedores en relación con los rubros identificados previamente como potenciales a través de la métrica density relatedness.

Para construir esta etiqueta, se utilizó el mismo concepto de antes llamado ventaja comparativa revelada (VCR), una métrica que indica si un proveedor tiene una posición competitiva significativa en un rubro específico. El cálculo de la etiqueta se realizó bajo los siguientes criterios:

- Éxito: Si un proveedor alcanzó un valor de VCR ≥ 1 en alguno de los rubros potenciales, se consideró que logró diversificarse exitosamente hacia ese rubro.
- No Éxito: En caso de que un proveedor no lograra alcanzar un valor de VCR ≥ 1 en los rubros potenciales, se clasificó como no éxito.

La elección del umbral de VCR se fundamenta en la literatura económica, donde un valor igual o superior a 1 refleja una especialización efectiva en un rubro. Este enfoque permite capturar si un proveedor logró consolidar su presencia en un nuevo mercado, diferenciando entre los casos en los que la diversificación fue exitosa y aquellos en los que no lo fue.

4.2.7. Modelos de Machine Learning

Una vez calculada la density relatedness, esta se utilizará como una variable clave en los modelos de machine learning que clasificarán y evaluarán automáticamente las oportunidades de diversificación para los proveedores. El objetivo principal de estos modelos es predecir cuáles rubros relacionados ofrecen el mayor potencial de éxito, considerando el historial de participación en licitaciones del proveedor y su conexión con otros rubros.

Como señalan Watt, Borhani y Katsaggelos (2016), el uso de machine learning permite refinar modelos para lograr predicciones más robustas y precisas. Este enfoque es fundamental en el análisis de diversificación, ya que optimiza los modelos predictivos mediante el ajuste de parámetros, la prevención del sobreajuste (overfitting) y la selección de características clave, mejorando así el rendimiento en diversas aplicaciones.

En este proyecto, se utilizó XGBoost como el modelo principal de machine learning debido a sus características avanzadas y su capacidad para manejar de manera

eficiente los retos inherentes a los datos tabulares. Según Chen y Guestrin (2016), XGBoost destaca como un sistema de boosting altamente escalable y eficiente, diseñado para optimizar tanto la precisión como el tiempo de ejecución. Este enfoque lo hace ideal para analizar las licitaciones públicas, que involucran múltiples variables numéricas y categóricas con relaciones complejas.

Una de las principales razones para seleccionar XGBoost fue su capacidad para implementar técnicas de regularización avanzadas que ayudan a reducir el riesgo de sobreajuste y mejoran la generalización del modelo. Estas técnicas son esenciales en contextos donde los datos pueden contener ruido o redundancias, como se describe en Hastie, Tibshirani y Friedman (2009). Además, XGBoost incluye funcionalidades específicas para manejar desbalances extremos en las clases, como el ajuste de pesos por clase, lo que permite mejorar la precisión del modelo al mitigar el sesgo hacia la clase mayoritaria (Chen & Guestrin, 2016).

Su arquitectura eficiente, que incluye paralelización y optimización en el uso de memoria, permite entrenar modelos complejos en menor tiempo comparado con alternativas como Random Forest o Regresión Logística. Estas características refuerzan la elección de XGBoost como la herramienta más adecuada para abordar los objetivos de este proyecto, proporcionando un balance óptimo entre precisión, eficiencia y robustez en un escenario de alta complejidad.

Por otro lado, el documento "Revisiting economic diversification in Africa’s largest resource-rich nation: Empirical insights from unsupervised machine learning" utiliza técnicas de aprendizaje automático para identificar agrupaciones de sectores económicos, orientando así políticas de diversificación. Esto subraya la importancia de aplicar métricas cuantitativas, como la density relatedness, para guiar decisiones estratégicas.

Así, el uso de machine learning en el contexto de las licitaciones públicas no solo optimiza la predicción del éxito de las propuestas, sino que también establece un marco analítico que permite a los proveedores navegar con mayor confianza en un entorno competitivo y lleno de incertidumbres.

Finalmente el desarrollo del modelo de machine learning, se utilizó SHAP (*SHapley Additive exPlanations*) como técnica para interpretar la importancia de las variables incluidas en el modelo. SHAP es ampliamente reconocido por su capacidad para descomponer las predicciones de un modelo en contribuciones individuales de cada variable, basándose en valores de Shapley provenientes de la teoría de juegos (Lundberg & Lee, 2017).

- **Comprensión de la importancia de las variables:** SHAP permitió evaluar cómo cada característica influía en las predicciones del modelo, identificando aquellas variables que tenían un impacto significativo en la clasificación de oportunidades de diversificación.
- **Validación del modelo:** A través de los valores de SHAP, se garantiza que las predicciones estuvieran alineadas con las expectativas teóricas y prácticas del comportamiento de las variables, proporcionando confianza en la robustez y explicabilidad del modelo.
- **Facilidad para la interpretación:** El uso de SHAP facilitó la comunicación de los resultados del modelo, destacando las contribuciones específicas de las variables clave. Esto fue especialmente relevante en un contexto donde las decisiones basadas en el modelo impactan estrategias empresariales.

Como explican Lundberg y Lee (2017), SHAP no solo permite interpretar modelos complejos de manera consistente, sino que también proporciona un marco matemático sólido para entender las interacciones entre variables, lo cual es crucial en aplicaciones prácticas como la diversificación empresarial.

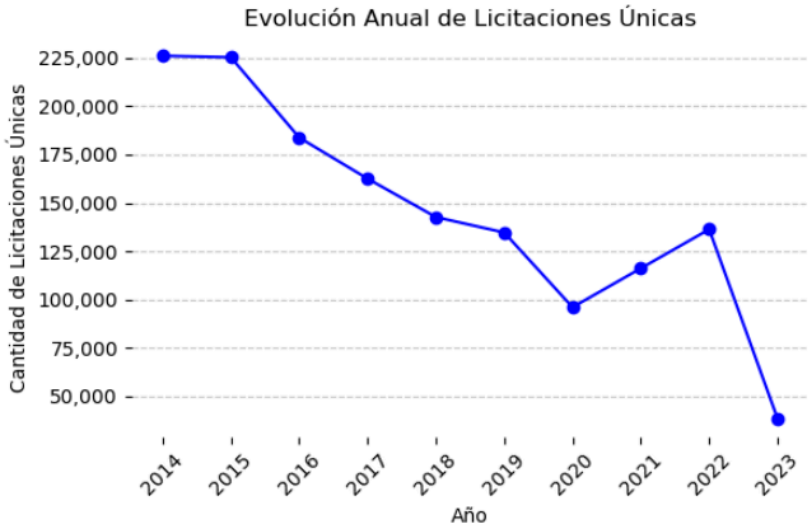
5. Resultados

Análisis Descriptivo

El análisis exploratorio tiene como objetivo proporcionar un panorama general de nuestro conjunto de datos, centrándonos en las variables más relevantes para nuestro modelo: rubros, proveedores y montos adjudicados. Estas variables son clave para entender la dinámica de las licitaciones públicas y sentar las bases para las siguientes etapas del proyecto.

Comenzamos analizando la evolución anual de licitaciones públicas, como se muestra en la Figura 1. Esta visualización nos permite identificar tendencias a lo largo del tiempo, evaluando si el volumen de licitaciones ha aumentado, disminuido o permanecido estable.

Figura 1: Evolución Anual de Licitaciones Únicas



En la Figura 2, se ilustra el análisis de los Top 10 rubros con mayor participación de proveedores. Este gráfico destaca los sectores donde existe mayor competitividad entre los actores del mercado, lo que refleja una diversificación en la oferta.

Por otro lado, en la Figura 3 se presenta el análisis de los Top 10 rubros con mayor cantidad de licitaciones públicas. Este enfoque permite identificar sectores con alta frecuencia de procesos licitatorios, lo cual puede estar vinculado a necesidades recurrentes o estratégicas en determinadas áreas.

Figura 2: Rubros con mayor cantidad de licitaciones únicas (Top 10)

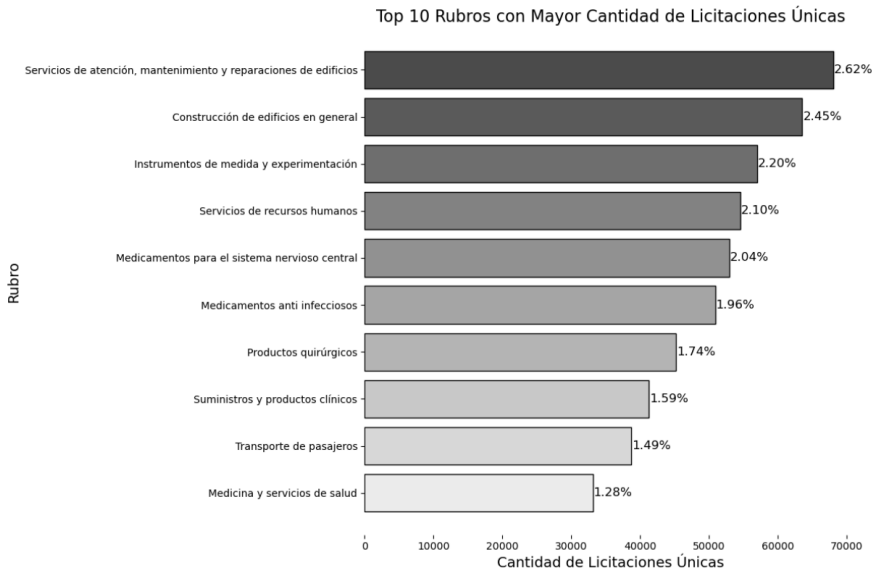
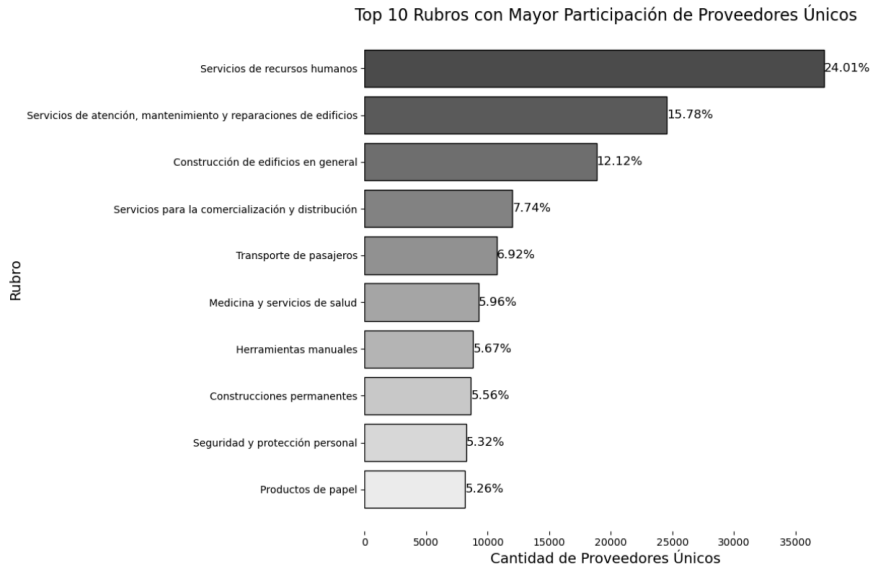
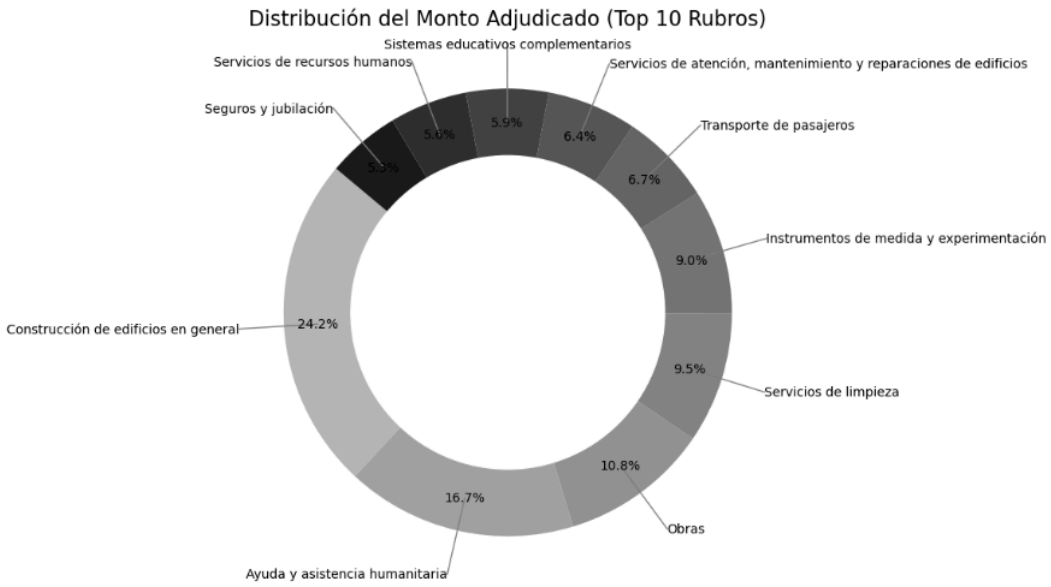


Figura 3: Rubros con mayor participación de proveedores únicos (Top 10)



La Figura 4 muestra la distribución de montos adjudicados por rubro, proporcionando información clave sobre la asignación de recursos. Este análisis evidencia cómo se distribuye el presupuesto en los diferentes sectores y permite identificar rubros con asignaciones atípicas o concentraciones significativas de presupuesto.

Figura 4: Distribución del Monto Adjudicado (Top 10 rubros)



RubroN2	
Construcción de edificios en general	7.447027e+12
Ayuda y asistencia humanitaria	5.130304e+12
Obras	3.314068e+12
Servicios de limpieza	2.914403e+12
Instrumentos de medida y experimentación	2.775870e+12
Transporte de pasajeros	2.053501e+12
Servicios de atención, mantenimiento y reparaciones de edificios	1.986048e+12
Sistemas educativos complementarios	1.831997e+12
Servicios de recursos humanos	1.721484e+12
Seguros y jubilación	1.630770e+12

En la Figura 5, se analizan los Top 10 proveedores con mayor participación en licitaciones únicas, lo que nos permite identificar a los actores más relevantes dentro de cada rubro. Por otro lado, la Figura 6 presenta los Top 10 proveedores con participación en el mayor número de rubros, revelando aquellos con una mayor diversificación en su actividad.

Figura 5: Proveedores con mayor participación en rubros (Top 10)

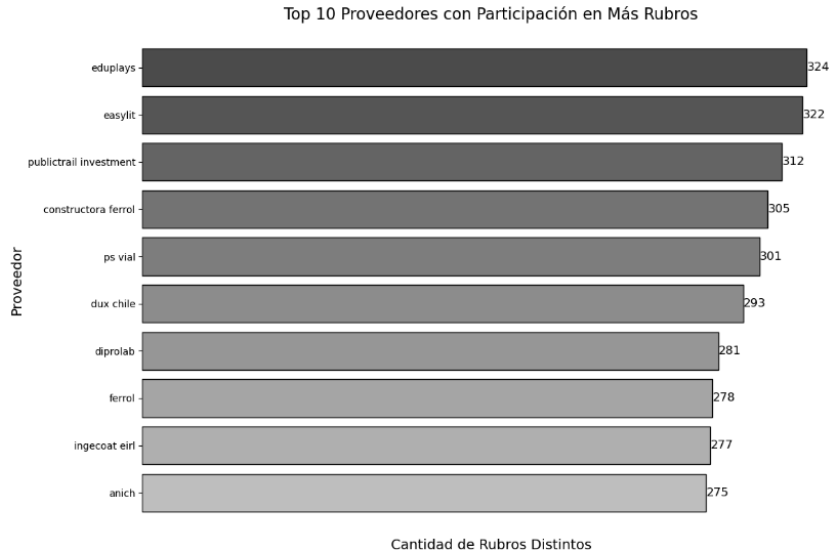
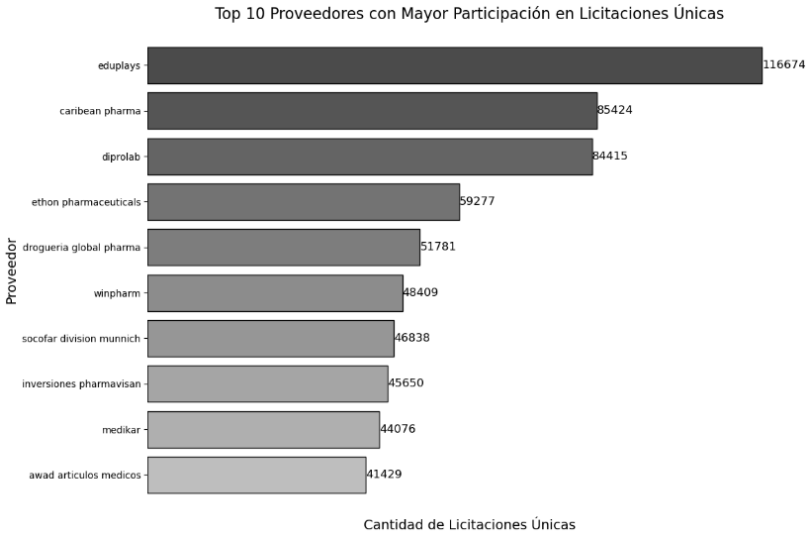
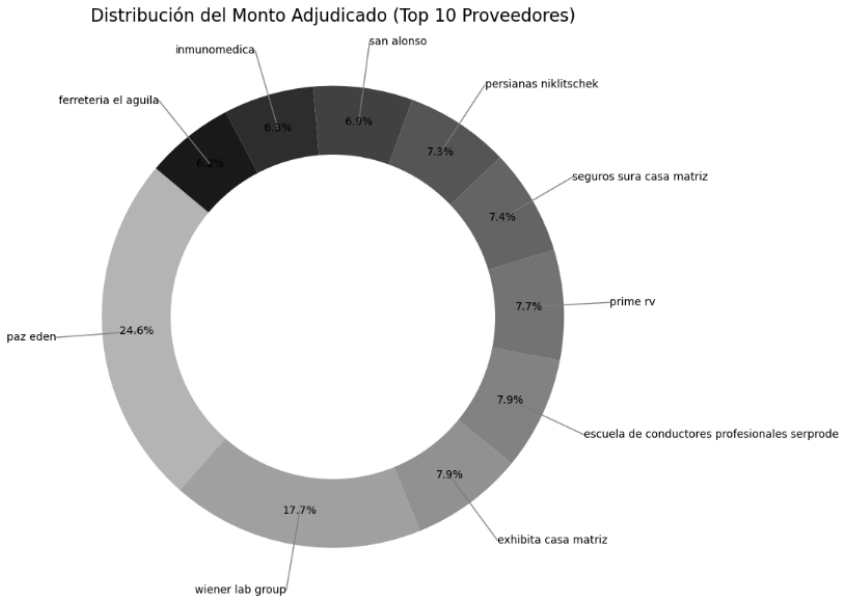


Figura 6: Proveedores con mayor participación en licitaciones únicas (Top 10)



Finalmente, en la Figura 7, se analiza la distribución del monto adjudicado por proveedor, proporcionando una perspectiva integral sobre cómo se concentran los recursos entre los principales actores del mercado.

Figura 7: Distribución del monto adjudicado (Top 10 proveedores)



paz eden	3.106859e+12
wiener lab group	2.232550e+12
exhibita casa matriz	1.000091e+12
escuela de conductores profesionales serprode	1.000012e+12
prime rv	9.743818e+11
seguros sura casa matriz	9.403692e+11
persianas niklitschek	9.216062e+11
san alonso	8.778729e+11
inmunomedica	7.963990e+11
ferreteria el aguila	7.862062e+11

Técnicas de desbalanceo

El problema del desbalance en los datos fue uno de los mayores desafíos del proyecto, dado que solo el 0.3% de los casos correspondían a diversificaciones exitosas, mientras que el 99.7% eran negativas. Este desbalance afectaba gravemente la capacidad del modelo para identificar correctamente los casos positivos, ya que la clase mayoritaria dominaba el proceso de aprendizaje, llevando a una baja precisión en general, pero a un recall decente para la clase minoritaria. Para mitigar este problema, se implementaron varias estrategias. Primero, se utilizó downsampling, reduciendo el número de ejemplos de la clase mayoritaria de forma aleatoria

para equilibrar la proporción entre ambas clases. El conjunto de datos resultante tenía una proporción final de 1:7 entre clases positiva y negativa.

También, se exploró el uso de SMOTE (Synthetic Minority Oversampling Technique) para aumentar artificialmente el tamaño de la clase minoritaria mediante la creación de ejemplos sintéticos. Aunque SMOTE mostró cierta mejora en el recall, también incrementó los falsos positivos y, en este caso, no superó el desempeño de las estrategias anteriores. En resumen, la técnica de downsampling ofreció un balance óptimo entre eficiencia computacional y rendimiento del modelo, mientras que SMOTE se dejó como una técnica exploratoria con potencial para futuros proyectos con características diferentes. Estas estrategias permitieron mitigar los efectos del desbalance extremo, mejorando la capacidad del modelo para identificar oportunidades de diversificación exitosas.

Descubriendo el Espacio de Conocimiento de Rubros (ECR)

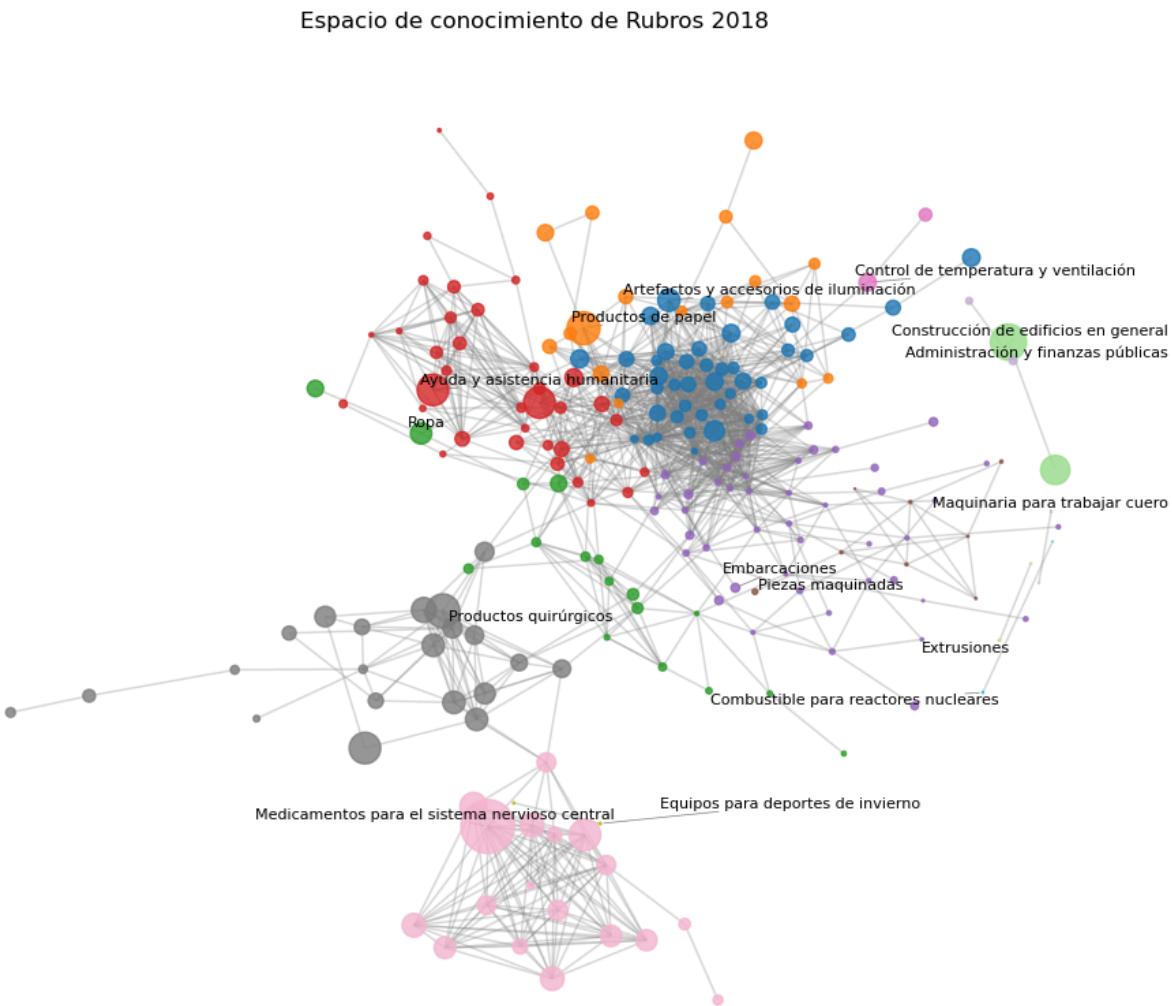
La **Figura 8** muestra la representación gráfica de la estructura de proximidad entre rubros en 2018, denominada Espacio de Conocimiento de Rubros (ECR). Esta red se construyó identificando los enlaces entre rubros y reteniendo únicamente aquellos con un nivel de proximidad superior a 0,2. En lugar de estar organizada alrededor de un grupo principal de actividades altamente conectadas, la estructura del ECR está más distribuida, con pequeños clústeres de rubros relacionados entre sí pero con menos conexiones hacia otros grupos. Esto implica que no existe un conjunto dominante de rubros que articule toda la red, lo que podría dificultar la diversificación hacia rubros más alejados en términos de proximidad.

Más allá de su estructura inmediata, el ECR refleja los caminos más comunes seguidos por proveedores activos en el mercado de licitaciones públicas. Por ejemplo, como se observa en la **Figura 1**, un proveedor que busca diversificarse en rubros como medicamentos para el sistema nervioso central y productos quirúrgicos enfrenta menos rutas directas de conexión, lo que

representa un desafío mayor. Por otro lado, rubros como “productos de papel” y “artefactos y accesorios de iluminación” tienen más conexiones cercanas, facilitando la diversificación hacia nuevos rubros dentro de sus clústeres.

En comparación con configuraciones más limitadas, el ECR en este análisis ofrece una amplia variedad de caminos para diversificación gracias a la gran cantidad de rubros y su interconexión. Esto permite a los proveedores considerar tanto oportunidades cercanas como la exploración de rubros más alejados, en función de sus capacidades y objetivos estratégicos. Sin embargo, la configuración del ECR no es estática, ya que está influenciada por cambios dinámicos en el mercado y las decisiones tomadas por las empresas, convirtiéndose en una herramienta valiosa para planificar estrategias de diversificación basadas en datos.

Figura 8: Espacio de conocimiento de rubros (ECR). Representación gráfica de la proximidad entre 345 rubros distintos. El gráfico incluye la red de rubros del año 2018 en donde los nodos son rubros y el enlace las proximidades superiores a 20%. Los nodos coloreados resaltan los 15 clusters principales identificados utilizando el algoritmo de Louvain Blondel et al. (2008), tiene una modularidad de 0,525 y el grado promedio de los nodos es 13,46.

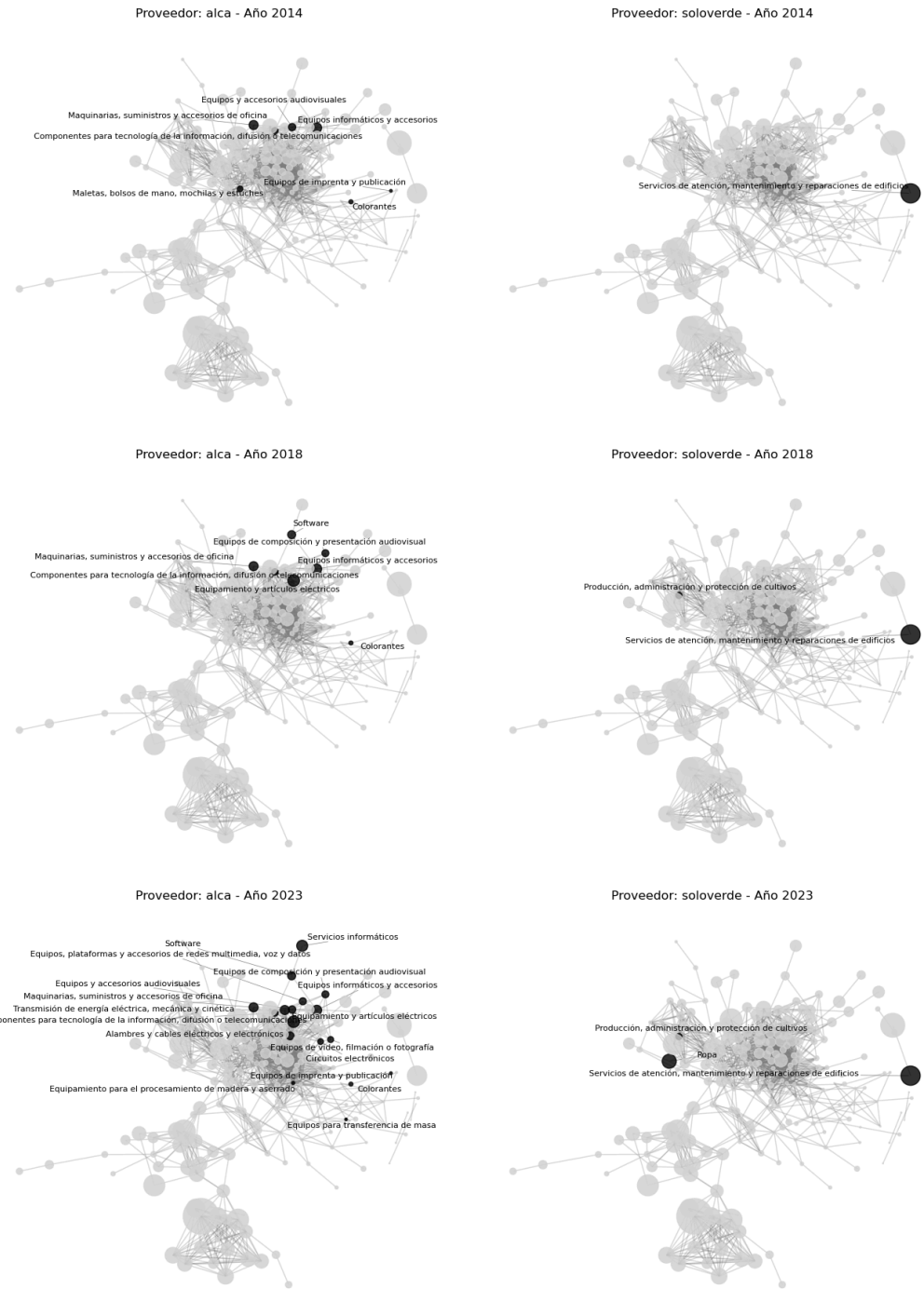


Density relatedness y la entrada a nuevos rubros

En la **Figura 9**, hacemos un seguimiento de la trayectoria de diversificación de los proveedores Soloverde y Alca a lo largo de los años 2014, 2018 y 2023. Para facilitar la comparación, optamos por utilizar la representación de la red de rubros de 2018 como base para las tres instantáneas.

En general, Alca muestra un comportamiento más diversificado, expandiéndose hacia nuevos rubros tanto dentro como fuera de su clúster. Por otro lado, Solverde registra una diversificación más limitada y alejada. En ambos casos, durante el período 2018-2023, ambos proveedores ganan ventaja comparativa revelada (VCR) en algunos rubros específicos, lo que podría indicar un reequilibrio de sus estrategias al abrirse hacia nuevos rubros.

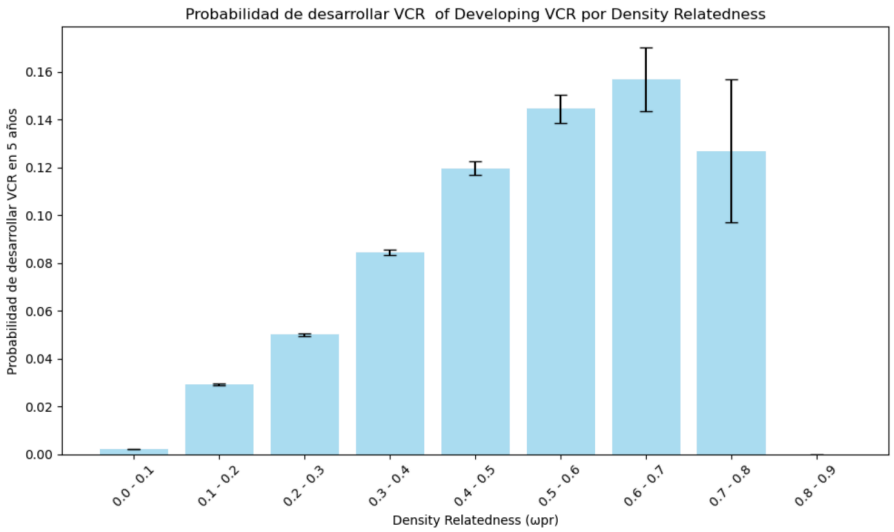
Figura 9: Diversificación de los proveedores alca y soloverde evaluados en 3 periodos de tiempos distintos utilizando el espacio de conocimiento de rubros del año 2018 como red base.



Luego, en el contexto de licitaciones públicas, utilizando un modelo de machine learning basado en XGBoost. Este modelo permite evaluar la probabilidad de que un proveedor desarrolle ventaja comparativa revelada (VCR) en un rubro, en función de métricas como la density relatedness, métricas de red, etc. A través de este análisis, buscamos validar si los proveedores tienden a diversificarse hacia rubros cercanos a su portafolio actual, tal como propone el principio de relación.

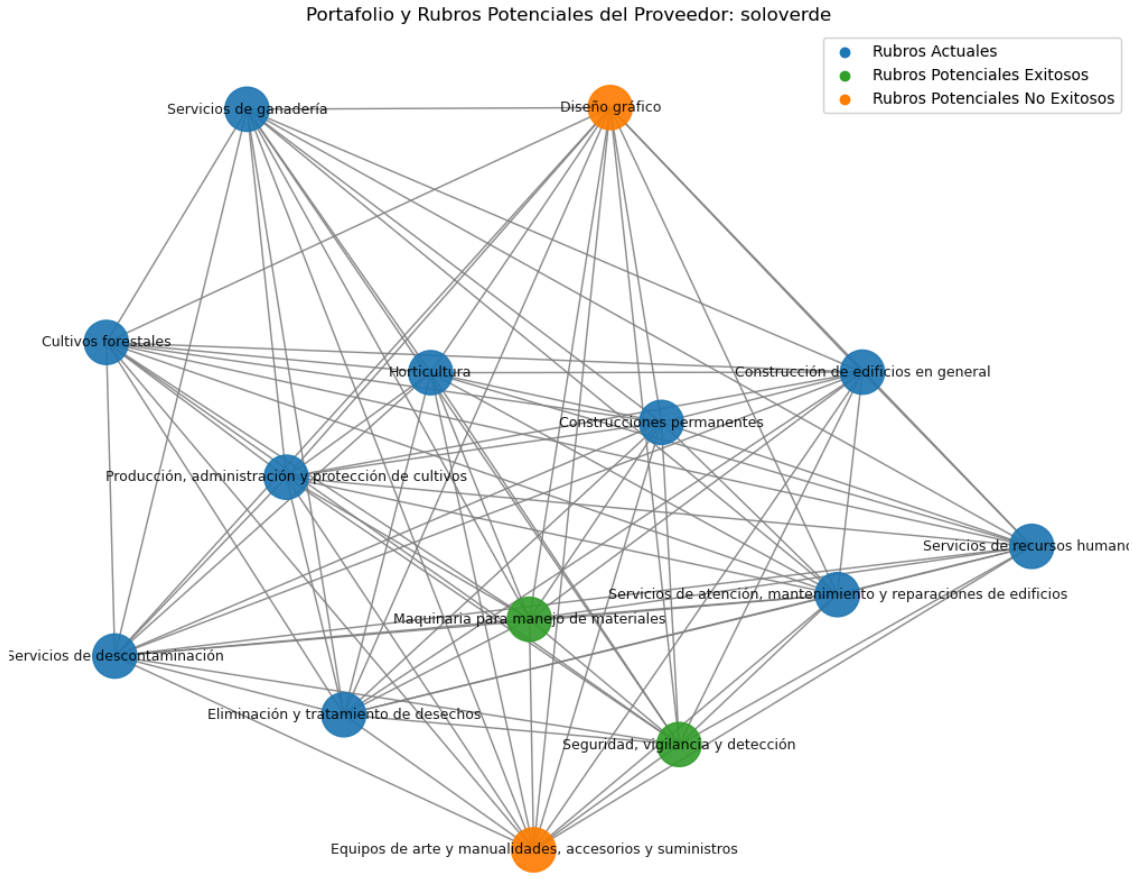
Figura 10: Probabilidad de desarrollar VCR en función de density relatedness.

Muestra que existe una relación cuadrática invertida entre la density relatedness y la probabilidad de que un proveedor desarrolle VCR en un nuevo rubro, tal como lo indica la importancia de esta variable en el modelo XGBoost. Esto implica que los rubros con valores intermedios de density relatedness tienen mayores probabilidades de éxito. En este contexto, los proveedores parecen diversificarse hacia rubros suficientemente conectados para aprovechar sus capacidades actuales, pero no tan cercanos como para ofrecer un crecimiento limitado.



Por último, en la **Figura 11**, se muestra la red del proveedor soloverde conformado por el portafolio de rubros en los cuales tiene ventaja comparativa y los rubros potenciales a los cuales podría diversificar en base a su valor de density relatedness. Se puede ver que este proveedor para este periodo de 5 años (2019-2023) logró tener éxito en 2 de los 4 rubros potenciales.

Figura 11: Red completa del proveedor soloverde que muestra los rubros actuales, los rubros donde tuvo éxito y rubros donde no.



Density relatedness y Machine Learning

Para evaluar la hipótesis planteada: *"Los modelos de machine learning, al utilizar la métrica density relatedness, pueden clasificar oportunidades de diversificación que maximicen el éxito de las empresas en licitaciones públicas"*, se compararon tres modelos de clasificación: Regresión Logística, Random Forest y XGBoost. Cada modelo fue entrenado con el conjunto de datos balanceado utilizando la técnica de downsampling (proporción 1:7) previamente descrita.

En la **Tabla 3** se presentan las métricas de desempeño de los tres modelos. Aunque todos lograron un desempeño aceptable en términos de precisión global (accuracy), XGBoost sobresalió significativamente en métricas clave como AUC-ROC (0.917) y recall (0.487), lo que lo convierte en el modelo más adecuado para este problema. Por su parte, Regresión Logística y Random Forest tuvieron valores más bajos de AUC-ROC y recall, lo que limitó su capacidad para identificar oportunidades de diversificación exitosas en el escenario altamente desbalanceado.

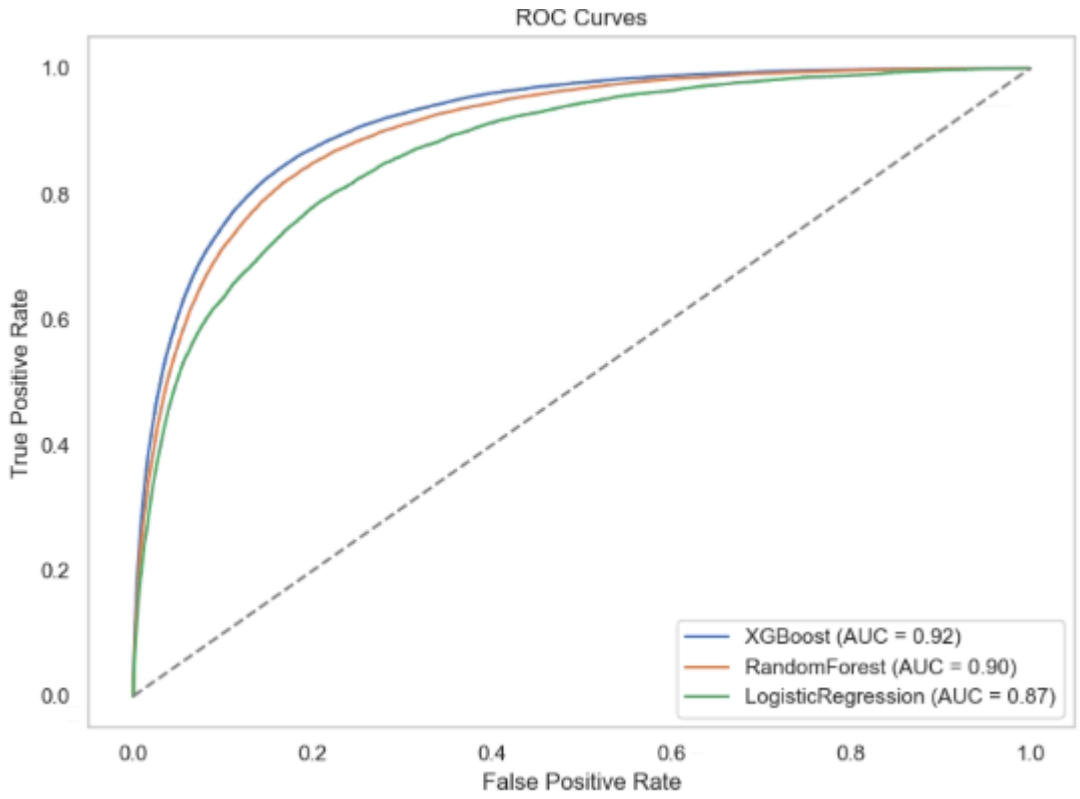
Tabla 3: Comparación de métricas entre los tres modelos probados con downsampling en proporción 1:7.

Tabla de comparación de métricas con tiempo de ejecución:

Modelo	Técnica	Parámetros	Accuracy	Precision	Recall	F1-Score	ROC AUC	Tiempo(s)
XGBoost	Downsampling 1:7	Default (Boosted Trees)	0.97	0.047	0.487	0.086	0.917	3.11
RandomForest	Downsampling 1:7	Árboles: 10, Profundidad: 10	0.974	0.046	0.407	0.083	0.903	16.13
LogisticRegression	Downsampling 1:7	Solver: lbfgs	0.979	0.045	0.305	0.079	0.872	1.44

En la **Figura 12** se comparan las curvas ROC, destacando el valor de AUC-ROC de 0.9210, que respalda su eficacia general para clasificar correctamente ambas clases. Este resultado es importante, ya que en escenarios desbalanceados, mantener una alta capacidad de discriminación es esencial.

Figura 12: Comparación Curvas AUC-ROC



También, en la **Tabla 4** se ve la matriz de confusión de estos modelo, donde se observa que el modelo XGBoost fue el modelo que más identificó casos positivos de diversificación exitosa (10,940) , sin embargo, (219,963) casos negativos que fueron clasificados incorrectamente como éxitos (falsos positivos), lo que influye en la precisión baja (0.047)

Tabla 4: Matriz que representa los casos exitosos y no exitosos del modelo XGBoost en base a las métricas obtenidas.

Tabla compacta de matrices de confusión:

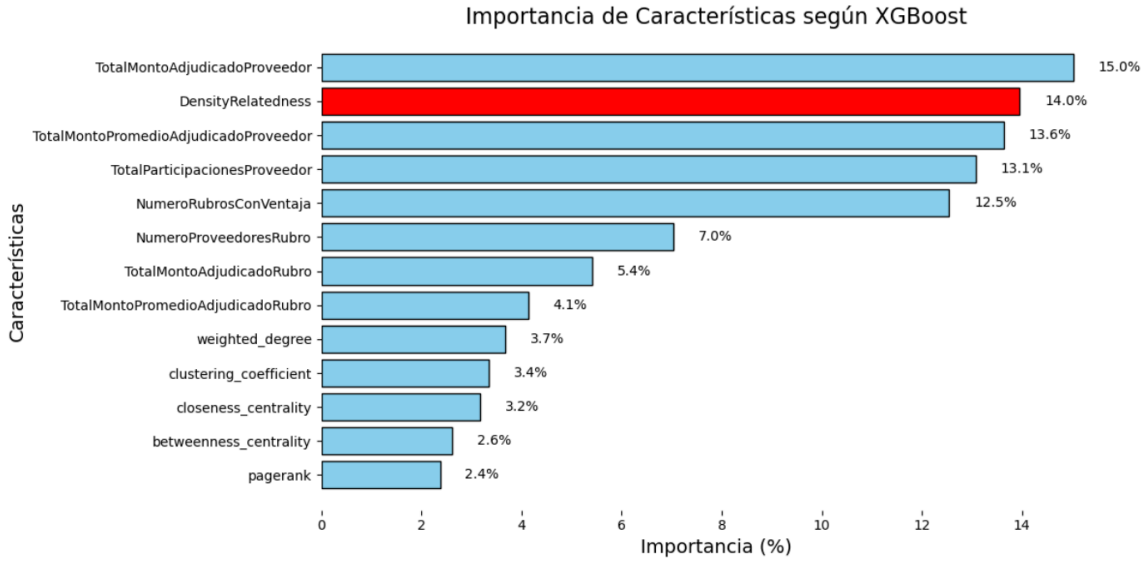
Modelo	Técnica	Parámetros	TN	FP	FN	TP
XGBoost	Downsampling 1:7	Default (Boosted Trees)	7519734	219963	11536	10940
RandomForest	Downsampling 1:7	Árboles: 10, Profundidad: 10	7551771	187926	13325	9151
LogisticRegression	Downsampling 1:7	Solver: lbfgs	7594797	144900	15620	6856

Dicho esto, el modelo **XGBoost** fue seleccionado como el modelo final debido a su desempeño destacado en varios criterios clave de evaluación: **AUC-ROC**, **F1-score**, la cantidad de **falsos positivos** y **falsos negativos**, y su eficiencia en términos de **rendimiento computacional**. Aunque los otros modelos (Regresión Logística y Random Forest) ofrecieron resultados aceptables en métricas como accuracy y recall, XGBoost sobresalió al proporcionar un equilibrio en los errores que entrega el modelo, junto con una capacidad superior para discriminar entre clases

Una vez seleccionado **XGBoost** como el modelo final, se procedió a calcular la importancia de las características tanto utilizando la importancia intrínseca del modelo como mediante el análisis de interpretabilidad con SHAP (**SHapley Additive Explanations**).

En la **Figura 13**, se muestra la importancia de las características según el modelo XGBoost. Las características **DensityRelatedness** y **NumeroProveedoresRubro** se destacan como las más relevantes, indicando que los proveedores con alta densidad de relación en los rubros y con experiencia previa en diversos rubros tienen mayores probabilidades de éxito en diversificación.

Figura 13: Importancia de características según el modelo XGBoost.

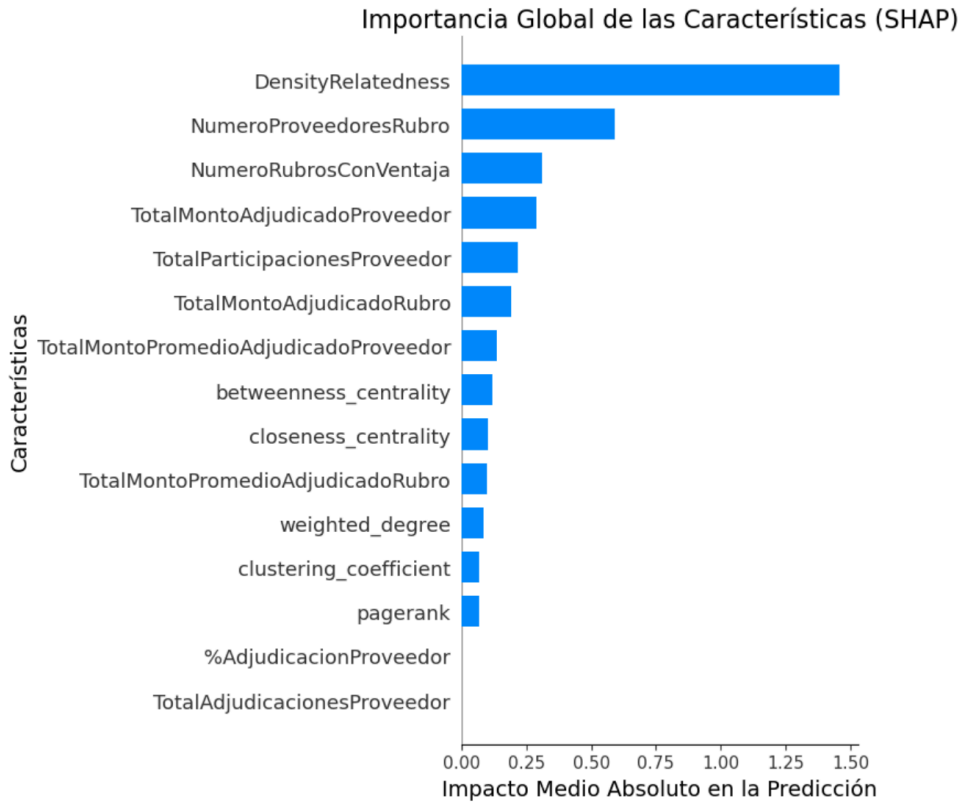


La métrica density relatedness jugó un rol importante en el modelo, como se observa en la **Figura 13**, donde se presenta el porcentaje de contribución de cada característica del modelo. Con un aporte del 14.0%, fue la segunda variable más importante, demostrando su capacidad para conectar el portafolio actual del proveedor con posibles rubros de diversificación. Este hallazgo valida la hipótesis, ya que empresas con altos valores de density relatedness muestran mayor probabilidad de éxito en nuevos rubros.

Otras variables clave incluyen el monto total adjudicado al proveedor (15.0%) y el monto promedio adjudicado (13.6%), reflejando la importancia de la experiencia y el desempeño económico en la predicción del éxito. Métricas derivadas de teoría de grafos, como clustering coefficient (2.6%) y pagerank (2.4%), aunque menos dominantes, complementaron el modelo al aportar una dimensión estructural útil para evaluar patrones en la red de rubros.

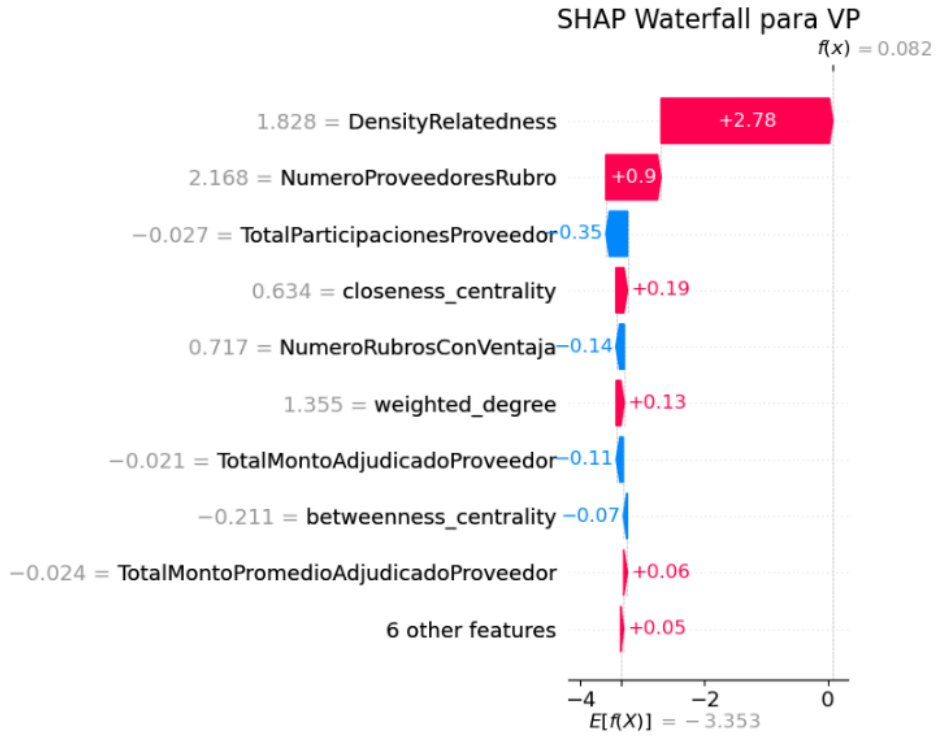
Además, se complementa este análisis con SHAP para desglosar la influencia de cada característica en las predicciones del modelo. En la **Figura 14**, se presenta un gráfico resumen de SHAP, que valida la importancia global de las características destacadas por el modelo. Por ejemplo, valores altos de Density Relatedness impactan positivamente en la probabilidad de éxito, mientras que características como %AdjudicacionProveedor y TotalAdjudicacionesProveedor tienen impactos negativos o no tienen impacto en algunos casos.

Figura 14: Importancia global de las características según SHAP.

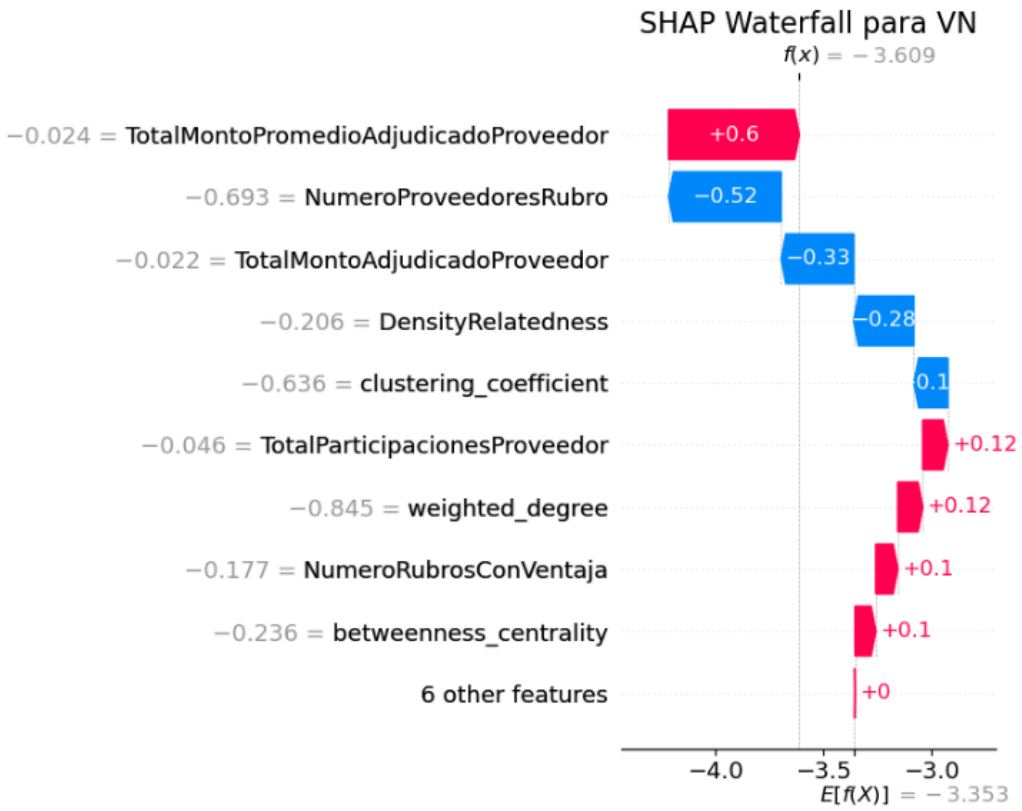


Para explicar casos específicos, se utilizó un análisis de predicción individual con SHAP. En la Figura 15, se muestra cómo las características del modelo influyen en las predicciones de un proveedor.

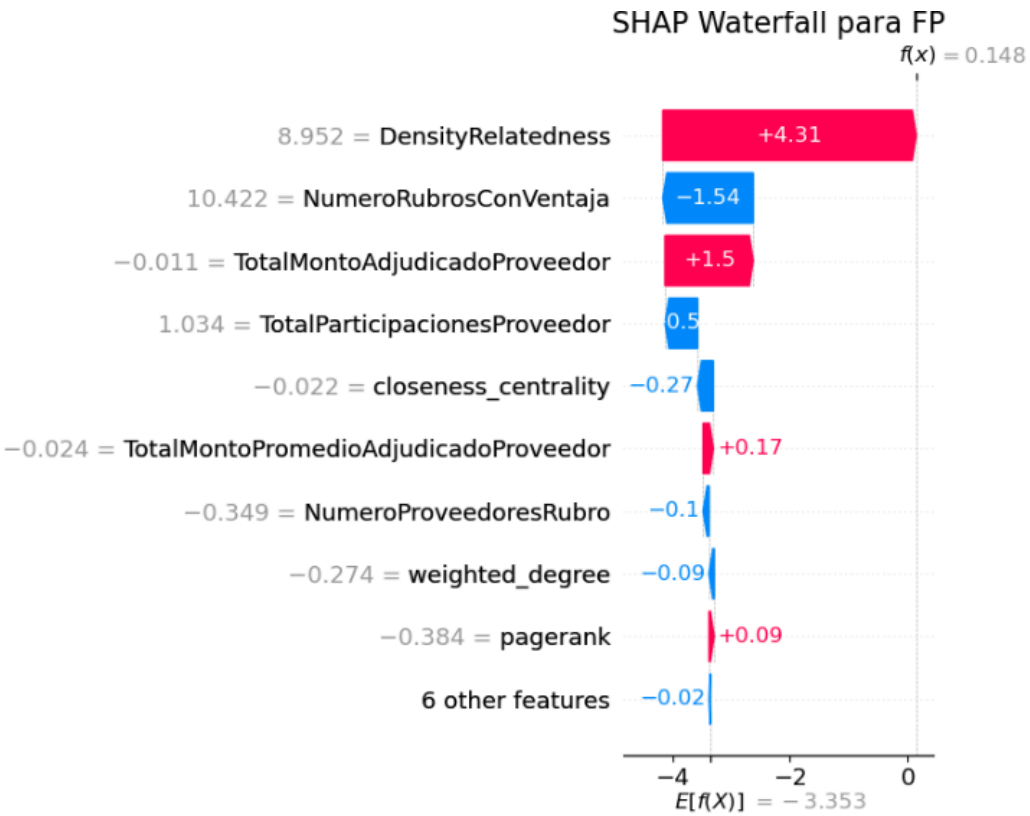
Figura 15: Desglose de una predicción individual para cada escenario (VP, VN, FN, FP) utilizando SHAP.



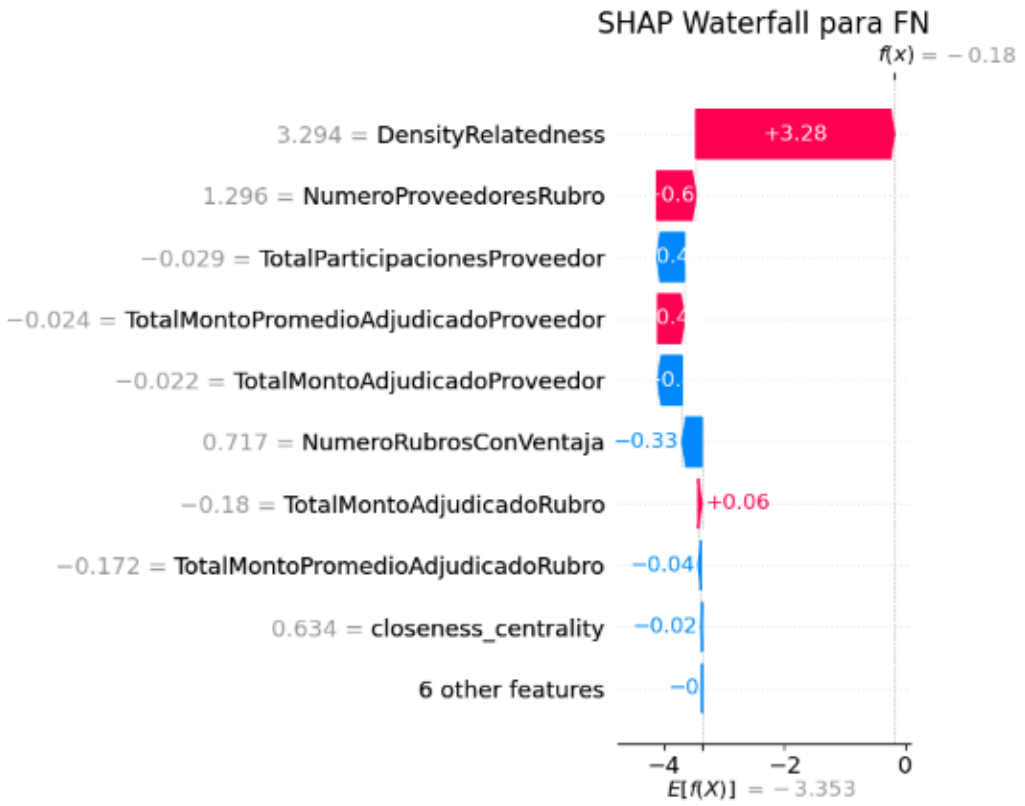
Este proveedor fue clasificado correctamente como exitoso en su diversificación. La característica más influyente fue DensityRelatedness, con un impacto significativo que refuerza la predicción de éxito. Adicionalmente, el número de proveedores en el rubro (NumeroProveedoresRubro) también contribuyó positivamente, indicando que este proveedor tiene alta capacidad de diversificación en rubros relacionados. Esto demuestra que el modelo puede identificar correctamente proveedores con características favorables para el éxito.



Clasificado correctamente como no exitoso, este proveedor presenta características desfavorables para diversificarse. La baja densidad de relatedness (DensityRelatedness) y el alto número de proveedores en el rubro (NumeroProveedoresRubro) redujeron significativamente la probabilidad de éxito. Este resultado refleja cómo el modelo identifica correctamente proveedores con menor potencial para diversificación en mercados altamente competitivos.



Este proveedor fue clasificado erróneamente como exitoso. Aunque DensityRelatedness tuvo un impacto muy positivo, otros factores como NumeroRubrosConVentaja afectaron negativamente. Esto indica que, aunque el modelo reconoce su capacidad de diversificación teórica, no tuvo en cuenta limitaciones prácticas que impidieron el éxito real. Este caso resalta una oportunidad para ajustar el peso de las características secundarias.



Clasificado erróneamente como no exitoso, este proveedor muestra cierto potencial de diversificación debido a una densidad de relatedness moderadamente alta (DensityRelatedness). Sin embargo, características como NumeroRubrosConVentaja disminuyeron la predicción, lo que llevó al modelo a subestimar su capacidad. Este caso destaca cómo el modelo podría mejorar en la identificación de proveedores con potencial moderado para diversificación.

6. Conclusiones

Implicancias del Modelo

El modelo XGBoost, entrenado con downsampling en proporción 1:7, se destacó como la mejor alternativa entre los modelos evaluados. Este enfoque permitió identificar casos positivos de diversificación exitosa (10,940 proveedores), demostrando la capacidad del modelo para destacar oportunidades relevantes en un escenario altamente desbalanceado.

Aunque el recall alcanzó un 48.67%, lo que indica una buena capacidad para identificar casos de éxito, se observó un alto número de falsos positivos (219,963), lo que refleja que el modelo tiende a sobreestimar algunas oportunidades de diversificación. Esto podría interpretarse como una invitación a explorar mercados con potencial no reconocido previamente, aunque con cierta cautela.

En comparación con SMOTE y otros modelos como Random Forest y Regresión Logística, el uso de downsampling produjo un equilibrio adecuado entre las métricas clave (AUC-ROC, recall y F1-score) y el tiempo de ejecución, consolidándose como la mejor estrategia para abordar el problema en este conjunto de datos. La importancia de características tanto global por SHAP como intrínseca del modelo son resultados que validan la hipótesis inicial y demuestra que la metodología basada en Density Relatedness y machine learning puede apoyar la toma de decisiones estratégicas para proveedores.

Limitaciones

Si bien el modelo XGBoost logró resultados satisfactorios, enfrenta varias limitaciones:

- **Desbalance de clases:** A pesar del downsampling, la representación de casos minoritarios sigue siendo un desafío, lo que influye en métricas como precisión (0.047).
- **Falsos positivos:** Un alto número de falsos positivos indica que el modelo clasifica como exitosos algunos casos que en la realidad no lo son, lo que podría generar expectativas equivocadas en los proveedores.
- **Dependencia de variables clave**: La métrica DensityRelatedness mostró ser crítica para las predicciones del modelo, pero su efectividad depende de la calidad y precisión de los datos originales.
- **Escalabilidad:** Evaluar un conjunto de datos de más de 8 millones de registros requirió importantes ajustes en los parámetros y el uso de técnicas de muestreo para manejar la carga computacional..

Propuestas de Mejora

- **Incorporación de variables adicionales:** Explorar la inclusión de nuevas características que capturen dinámicas no consideradas, como análisis de tendencias históricas o datos externos sobre sectores económicos.
- **Optimización del modelo:** Experimentar con técnicas avanzadas como ensembles híbridos (combinando XGBoost con otros modelos) o algoritmos más recientes como LightGBM, que podrían ofrecer mejoras en precisión y eficiencia computacional.
- **Evaluación contextual:** Ajustar el modelo según sectores específicos, permitiendo adaptar las predicciones a las dinámicas propias de cada rubro.

- **Reducción de falsos positivos**: Implementar un paso adicional de validación manual para casos clasificados como exitosos, mitigando riesgos asociados a falsas expectativas.
- **Mayor interpretabilidad:** Ampliar el uso de herramientas como SHAP para identificar patrones más detallados y comunicar los resultados a los proveedores de manera comprensible.

7. Bibliografía

- Ansoff, I. (1957). *Strategies for diversification*. Harvard Business Review, 35(5), 113-124. https://www.casrilanka.com/casl/images/stories/2017/2017_pdfs/sab_portal/course_material/strategies_for_diversification.pdf
- Neffke, F., Henning, M., & Boschma, R. (2011). How Do Regions Diversify Over Time? Industry Relatedness and the Development of New Growth Paths in Regions. Economic Geography, 87(3), 237-265. https://www.researchgate.net/publication/46454619_How_do_regions_diversify_over_time_Industry_relatedness_and_the_development_of_new_growth_paths_in_regions
- Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. Proceedings of the National Academy of Sciences, 106(26), 10570-10575. https://doi.org/10.1073/pnas.0900943106
- Balassa, B. (1965). Trade liberalisation and “revealed” comparative advantage1. The Manchester School, 33(2):99–123. https://www.scirp.org/reference/referencespapers?referenceid=1933496
- Hidalgo, C. A. (2021). Economic complexity theory and applications. Nature Reviews Physics, 3(2):92–113. https://oec.world/pdf/economic-complexity-theory-and-applications.pdf
- Machine Learning Refined: Foundations, Algorithms, and Applications, Watt, J., Borhani, R., Katsaggelos, A. Cambridge University Press, 2016) https://people.engr.tamu.edu/guni/csce421/files/Machine_Learning_Refined.pdf
- Louvain, Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://www.researchgate.net/publication/1913681_Fast_Unfolding_of_Communities_in_Large_Networks
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.