ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Análisis predictivo del potencial comercial de industrias en el Ecuador para una empresa gestora ambiental

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencia de Datos

Presentado por:
Paulina Michell Garzón Arreaga
Marco Andrés Garófalo Cervantes

GUAYAQUIL - ECUADOR
Año: 2026

DEDICATORIA

Dedico este trabajo a mis padres, como reconocimiento a su esfuerzo, amor y ejemplo, que han sido la base de mi crecimiento personal y profesional.

De igual manera, dedico esta tesis a Carlos Garzón y Tanya Arreaga, por su apoyo y acompañamiento a lo largo de este camino.

Finalmente, dedico este logro a mi esposo, Luis Torres, por caminar a mi lado, creer en mí y ser un pilar fundamental en cada etapa de este proceso.

Paulina Garzón

Dedicado a Dios, por la vida y la sabiduría. A mis padres, por su ejemplo de perseverancia y por creer en mí incluso cuando yo no lo hacía

Marco Garófalo

AGRADECIMIENTOS

En primer lugar, agradezco a Dios por la vida, la fortaleza y la constancia que me ha permitido culminar esta etapa académica.

Expreso mi sincero agradecimiento a mis docentes y a la institución académica por los conocimientos impartidos, el acompañamiento y la guía brindada a lo largo del desarrollo de esta tesis, los cuales fueron fundamentales para mi formación profesional.

Paulina Garzón

Doy gracias a Dios por la sabiduría otorgada durante este proceso. A mis padres, les agradezco de todo corazón por su guía, sus palabras de aliento y por ser el pilar fundamental de mi vida y mi carrera.

Marco Garófalo

DECLARACIÓN EXPRESA

Nosotros Paulina Garzón y Marco Garófalo acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al/los autor/es que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 23 de enero del 2026.

Firmado electrónicamente por:
PAULINA MICHELL GARZON ARREAGA
Validar únicamente con FirmaEC

Paulina Garzón

Marco Andres Garofalo Cervantes
Time Stamping Security Data

Marco Garófalo

COMITÉ EVALUADOR

Christian Eduardo Galarza Morales
Time Stamping Security Data

Dr. Christian Galarza Morales

PROFESOR TUTOR

Maria Isabel Mera Collantes
Time Stamping Security Data

Dra. María Isabel Mera

PROFESOR EVALUADOR

RESUMEN

La siguiente investigación se desarrolla en el contexto de la gestión de residuos industriales y la transición hacia modelos de economía circular en el Ecuador, donde las empresas gestoras ambientales enfrentan el desafío de identificar y priorizar clientes industriales con alto potencial comercial. Por lo tanto el objetivo del estudio se centra en desarrollar un sistema de inteligencia comercial basado en análisis predictivo que permita estimar la probabilidad de conversión de empresas industriales en clientes efectivos, justificándose en la necesidad de reducir la dependencia de criterios empíricos y optimizar la toma de decisiones comerciales mediante el uso de datos.

Durante el desarrollo del proyecto se utilizó un enfoque cuantitativo, que integró la información proveniente de fuentes internas de una empresa gestora ambiental con bases de datos públicas provenientes de instituciones oficiales. Para la unión de esta información se utilizaron procesos de limpieza, normalización e integración de datos, incorporando variables firmográficas, financieras, contextuales y ambientales. Posteriormente, se entrenaron y validaron modelos de aprendizaje automático con el fin de evaluar su desempeño predictivo.

Los resultados obtenidos evidenciaron que el modelo desarrollado permitió asignar un puntaje por empresa, mejorando la priorización de prospectos y reduciendo la carga operativa asociada a la prospección manual. Asimismo, se observó una mayor alineación entre los criterios de marketing y ventas.

En conclusión, el sistema propuesto contribuye a fortalecer la eficiencia operativa, la inteligencia comercial y la toma de decisiones basada en datos en el sector industrial ecuatoriano.

Palabras clave: gestión de residuos, inteligencia comercial, aprendizaje automático, análisis predictivo, prospección de clientes

ABSTRACT

The present research is developed in the context of industrial waste management and the transition toward circular economy models in Ecuador, where environmental management companies face the challenge of identifying and prioritizing industrial clients with high commercial potential. Therefore, the objective of this study is to develop a commercial intelligence system based on predictive analytics that estimates the probability of industrial companies converting into effective clients. This approach is justified by the need to reduce reliance on empirical criteria and to optimize commercial decision-making using structured data.

During the development of the project, a quantitative approach was applied, integrating information from internal sources of an environmental management company with public databases obtained from official institutions. Data cleaning, normalization, and integration processes were performed, incorporating firmographic, financial, contextual, and environmental variables. Subsequently, supervised machine learning models were trained and validated to evaluate their predictive performance.

The results showed that the developed model made it possible to assign a conversion score to each company, improving prospect prioritization and reducing the operational workload associated with manual prospecting. Additionally, greater alignment between marketing and sales criteria was observed.

In conclusion, the proposed system contributes to strengthening operational efficiency, commercial intelligence, and data-driven decision-making in the Ecuadorian industrial sector.

Keywords: *Industrial waste management, Commercial intelligence, Predictive analytics, Machine learning, Client prospecting.*

ÍNDICE GENERAL

ABREVIATURAS

CIIU	Clasificación Industrial Internacional Uniforme
CRM	Sistema de gestión de relaciones con el cliente
ESG	Estándar de informes de sostenibilidad
ESPOL	Escuela Superior Politécnica del Litoral
GBM	*Gradient Boosting*, Modelo de ensamble de clasificación
INEC	Instituto Nacional de Estadística y Censos
LTV	Valor de Vida (Lifetime Value)
MAATE	Ministerio del Ambiente, Agua y Transición Ecológica
MIEAE	Módulo de Información Económica Ambiental en Empresas (del INEC)
ML	Aprendizaje Automático
n.d.	*No date* (Sin fecha)
ODS	Objetivos de Desarrollo Sostenible
PECCS	Programa Ecuador Carbono Cero
p. ej.	Por ejemplo
Publ.	Publicación
Rev.	Revisión
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
ROI	Retorno de la Inversión
SCVS	Superintendencia de Compañías, Valores y Seguros
SHAP	Técnica de explicabilidad para analizar la influencia de cada variable en el resultado del modelo.
SVM	*Support Vector Machine*, Tipo de modelo supervisado de clasificación.
VAN	Valor Actual Neto
ENESEM	Encuesta Estructural Empresarial

SIMBOLOGÍA

t	Tonelada(s)
USD	Dólar estadounidense

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La transición global hacia modelos de economía circular se ha consolidado como un pilar fundamental para el desarrollo sostenible, promoviendo la eficiencia en el uso de recursos y la reducción sustancial de residuos a través de la prevención, el reciclaje y la reutilización. Este enfoque está alineado con los Objetivos de Desarrollo Sostenible (ODS), que instan a transformar la gestión de desechos industriales. En este marco, las empresas gestoras ambientales actúan como agentes estratégicos para el cumplimiento normativo y la promoción de prácticas de sostenibilidad. Específicamente en Ecuador, la regulación exige a las empresas generadoras de residuos peligrosos y especiales contratar operadores autorizados para garantizar la disposición final adecuada de los desechos.

Pese al mandato regulatorio y el alto valor estratégico que posee el mercado B2B de gestión de residuos, las empresas gestoras ambientales en el país enfrentan un desafío crítico: la ineficiencia en la identificación y priorización de clientes industriales con alto potencial de generación de residuos y, por ende, de contratación de servicios especializados.

Actualmente, la prospección de clientes industriales se apoya en métodos empíricos y no sistematizados, que dependen del conocimiento tácito de los vendedores o de la interpretación subjetiva de indicadores. Esta clasificación manual de *leads* consume recursos operacionales significativos.

Ante la limitación de los métodos heurísticos y la necesidad de maximizar la rentabilidad en un mercado industrial de alto valor por transacción, este estudio se justifica en el desarrollo de una herramienta basada en datos que fortalezca la inteligencia comercial.

En respuesta a la ineficiencia operativa y la brecha analítica, se propone el diseño e implementación de un sistema de inteligencia comercial basado en aprendizaje automático supervisado. Este modelo predictivo asignará un *score* de conversión objetivo a cada empresa prospecto, clasificándolas en categorías de prioridad a partir de una base de datos enriquecida con información interna y fuentes públicas (financieras, ambientales y estructurales).

CAPÍTULO 1

1. PLANTEAMIENTO DE LA PROBLEMÁTICA

1.1 Descripción del problema

En el contexto de la transición hacia modelos de economía circular, las empresas gestoras ambientales enfrentan el reto de identificar y priorizar a los clientes industriales con mayor potencial de generación de residuos y contratación de servicios especializados. En Ecuador, esta transición hacia una producción más sostenible exige soluciones basadas en datos que optimicen la gestión de recursos y aumenten la eficiencia en los procesos comerciales (Espinoza, 2023).

A nivel internacional, la economía circular constituye un pilar del desarrollo sostenible. En particular, los Objetivos de Desarrollo Sostenible (ODS) promueven “reducir sustancialmente la generación de residuos mediante actividades de prevención, reducción, reciclaje y reutilización” al 2030 (Espinoza, 2023). Alcanzar esta meta requiere transformaciones no solo en los procesos productivos, sino también en la manera en que las empresas gestionan sus residuos industriales y articulan sus decisiones de compra de servicios ambientales.

En Ecuador, la normativa vigente —como el Acuerdo Ministerial No. 061 emitido por el MAATE (2022)— obliga a las empresas generadoras de residuos peligrosos y especiales a contratar operadores autorizados que aseguren la trazabilidad, el tratamiento y la disposición final de los desechos (MAATE, 2015). No obstante, datos del Instituto Nacional de Estadística y Censos evidencian que solo el 30,89 % de las empresas conoce la cantidad de residuos no peligrosos que genera, y apenas el 29,22 % y 32,16 % dispone de información sobre los residuos especiales y peligrosos, respectivamente (INEC, 2025). Esta brecha de información limita la capacidad de las gestoras ambientales para anticipar la demanda de servicios y diseñar estrategias comerciales basadas en evidencia.

Actualmente, la identificación de industrias generadoras de residuos con alto potencial contractual requiere el análisis manual de información dispersa en fuentes públicas y privadas, lo que incrementa la carga operativa y reduce la efectividad comercial. En el caso de la empresa gestora ambiental analizada, una organización multinacional con presencia activa en Ecuador y 63 millones de toneladas métricas de residuos recuperados a nivel global (Veolia), esta limitación se traduce en una clasificación manual de leads poco precisa, con tiempos elevados de procesamiento y resultados heterogéneos entre los equipos de marketing y comercial.

La prospección de clientes industriales para la gestora sujeto de este estudio se apoya en métodos empíricos y no sistematizados que dependen del conocimiento tácito de los vendedores, las relaciones previas o la interpretación subjetiva de indicadores financieros. Aunque este enfoque heurístico puede resultar útil en contextos acotados, carece de escalabilidad y reproducibilidad, afectando la productividad y la asignación de recursos. En la práctica, el ingeniero de marketing digital y el asignador comercial destinan cerca del 30 % de su jornada a priorizar e identificar leads, mientras que los asesores de ventas invierten hasta el 35 % de su tiempo en seguimientos que no se concretan en contratos.

A ello se suma la falta de una definición estandarizada entre marketing y comercial sobre los criterios que determinan cuándo un contacto debe considerarse un lead con alta probabilidad de conversión. Esta desalineación genera discrepancias en la clasificación, dificulta la trazabilidad de los procesos y limita la capacidad de medir el retorno de la inversión publicitaria. Actualmente, el área de marketing invierte alrededor de USD 3.000 mensuales en pauta digital para la generación de leads, sin contar con una herramienta analítica que permita perfilar de manera precisa los públicos objetivos u optimizar las audiencias según su potencial de conversión.

En conjunto, esta situación refleja la necesidad de desarrollar una herramienta basada en datos que estandarice los criterios de priorización, automatice la

clasificación de prospectos y reduzca la dependencia de la intuición o el trabajo manual.

1.2 Justificación del problema

La gestión adecuada de residuos industriales es esencial para el cumplimiento de los compromisos ambientales y los ODS asumidos por el Ecuador. Las empresas generadoras de residuos peligrosos, especiales y no peligrosos deben contratar operadores certificados que garanticen la trazabilidad y disposición final adecuada (MAATE, 2015). En este escenario, las empresas gestoras ambientales certificadas actúan como agentes estratégicos para el cumplimiento normativo y la promoción de la economía circular.

Sin embargo, la falta de bases de datos integradas que combinen información técnica (tipo de residuos), económica (ingresos y tamaño empresarial) y ambiental (certificaciones y cumplimiento) obstaculiza el desarrollo de estrategias comerciales basadas en evidencia. Esto genera ineficiencias en la asignación de recursos, pérdida de oportunidades de negocio y una menor tasa de conversión de prospectos en contratos efectivos.

En respuesta a esta brecha, la empresa ha conformado una base de más de 134.000 empresas industriales activas, priorizando aquellas clasificadas como grandes por la Superintendencia de Compañías, dado que su escala operativa está asociada a mayores volúmenes de residuos. Este enfoque permite focalizar esfuerzos en clientes con alto potencial económico y operativo, mejorando la rentabilidad y sostenibilidad del servicio.

El desarrollo de un modelo de aprendizaje automático orientado a la priorización de clientes industriales representa una herramienta estratégica para fortalecer la inteligencia comercial. La ciencia de datos permite analizar múltiples variables simultáneamente, identificar patrones no evidentes y estimar probabilidades de conversión de leads con base en evidencia objetiva. Investigaciones como las de Chollet y Gouveia & Costa confirman la eficacia de los modelos supervisados como Random Forest, XGBoost y SVM para

predecir comportamientos comerciales en entornos industriales (Chollet, 2018) (Cunha & Coelho, 2022).

La implementación de esta herramienta beneficiará directamente a los equipos de Marketing Digital, Asignación Comercial y Ventas Industriales, al reducir el tiempo invertido en tareas operativas y mejorar la calidad de los leads gestionados. Se estima un ahorro operativo anual de USD 21.900 por optimización del tiempo de trabajo y una mejora del 10 % en la tasa de conversión de leads. Además, la optimización del presupuesto publicitario permitirá incrementar el retorno de inversión (ROI) en un 73 % y alcanzar un Valor Actual Neto (VAN) positivo durante los primeros cinco años.

Este estudio se enmarca en la tendencia global de transformación digital en sectores regulados, donde la toma de decisiones basada en evidencia constituye un factor clave de eficiencia, cumplimiento y sostenibilidad. La propuesta no solo ofrece una solución técnica, sino que aporta un modelo replicable de inteligencia comercial aplicable a la gestión ambiental en Ecuador y América Latina.

1.3 Solución propuesta

Frente a las limitaciones descritas, se propone el diseño e implementación de un sistema de inteligencia comercial basado en aprendizaje automático supervisado, orientado a estimar la probabilidad de conversión de cada empresa prospecto en cliente efectivo.

El modelo predictivo asignará un score de conversión a cada empresa industrial dentro de la base de prospectos, clasificándolos en categorías de alta, media y baja prioridad. Este puntaje se calculará a partir de una base enriquecida que integra fuentes internas y externas públicas, como la Superintendencia de Compañías, el MAATE, y el Ranking EKOS 2024.

El sistema aprovechará algoritmos de clasificación supervisada y se evaluará mediante métricas de desempeño, garantizando su validez técnica. Adicionalmente, incluirá un dashboard o app que facilitarán la visualización de patrones de conversión y la priorización de leads, promoviendo la optimización de los flujos de trabajo actuales del CRM.

La implementación de este modelo permitirá sustituir la clasificación empírica por un proceso basado en evidencia y analítica predictiva, incrementando la eficiencia, precisión y trazabilidad en la gestión de leads. Asimismo, estandariza el lenguaje y los criterios utilizados por marketing y comercial para definir qué contactos presentan alta probabilidad de conversión, contribuyendo a una mayor alineación estratégica y optimización de la inversión publicitaria.

En síntesis, la solución propuesta constituye una herramienta de transformación digital que fortalece la competitividad de la empresa gestora ambiental, mejora la eficiencia operativa y promueve una gestión comercial sostenible y sustentada en datos.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un sistema de inteligencia comercial basado en análisis predictivo que evalúe el potencial comercial de empresas industriales en Ecuador para una multinacional gestora ambiental, mediante la integración de fuentes internas y externas y la aplicación de técnicas de aprendizaje automático supervisado.

1.4.2 Objetivos Específicos

- Evaluar la calidad, estructura y consistencia de la base de datos de prospectos industriales, identificando y documentando las variables clave necesarias para su enriquecimiento y posterior uso en análisis predictivos.

- Integrar y enriquecer la base de datos con información proveniente de fuentes públicas (Superintendencia de Compañías, MAATE, ranking EKOS 2024), incorporando indicadores financieros, ambientales y estructurales que permitan caracterizar el perfil de las empresas.
- Entrenar y validar un modelo predictivo supervisado que estime el porcentaje de probabilidad de conversión de cada empresa prospecto en cliente, evaluando su desempeño mediante métricas de clasificación adecuadas (accuracy, precision, recall y ROC-AUC).
- Desarrollar un producto digital que consolide la influencia de las variables del modelo y facilite la visualización de porcentaje de conversión y toma de decisiones estratégicas basadas en datos.
- Estandarizar los criterios de calificación de leads entre los equipos de marketing y comercial, utilizando el puntaje predictivo como herramienta objetiva para la priorización de contactos y la optimización del presupuesto publicitario.

1.5 Metodología

El presente estudio adopta un enfoque cuantitativo y un nivel de investigación explicativo, orientado a identificar los factores que determinan el potencial comercial de empresas industriales en Ecuador y a modelar su probabilidad de conversión en clientes efectivos de una empresa gestora ambiental. La metodología se sustenta en técnicas de ciencia de datos y aprendizaje automático supervisado, con el propósito de transformar la información disponible en conocimiento aplicable a la toma de decisiones comerciales.

El diseño metodológico se estructura en cuatro etapas secuenciales y complementarias:

1. Preprocesamiento y validación de datos
2. Enriquecimiento e integración de fuentes externas
3. Modelado predictivo supervisado

4. Visualización aplicada mediante dashboards de inteligencia comercial.

En la primera etapa, se analiza la calidad y estructura de la base de datos proporcionada por la empresa, compuesta por registros de ventas correspondiente al periodo de 2024. Esta base constituye uno de los principales insumos del modelo y será auditada bajo criterios de completitud, unicidad, consistencia, validez y actualidad, siguiendo lineamientos metodológicos propuestos por Brei y Singh (Brei, 2020) (Singh, 2024) (Singh, 2024).

Posteriormente, se ejecutará un proceso de enriquecimiento que integrará información proveniente de fuentes públicas como la Superintendencia de Compañías, el Ministerio del Ambiente, Agua y Transición Ecológica (MAATE) y el "ranking EKOS 2024" lista empresas que participan voluntariamente en certificaciones como "Carbono Cero", demostrando su compromiso con la gestión ambiental y la sostenibilidad. Este paso permitirá incorporar variables financieras, ambientales y estructurales que amplíen la caracterización de las empresas y fortalezcan la capacidad predictiva del modelo.

A continuación, se desarrollará un modelo de aprendizaje automático supervisado orientado a estimar el porcentaje de probabilidad de conversión de cada empresa prospecto. El modelo se entrenará y validará utilizando técnicas de clasificación, seleccionadas por su eficacia comprobada en la predicción de comportamientos comerciales (Chollet, 2018). Las métricas de evaluación incluirán accuracy, precision, recall y ROC-AUC, con el fin de garantizar su validez técnica y su capacidad discriminante.

Finalmente, se desarrollará un producto interactivo, un aplicativo web, que permitirá visualizar la influencia de las variables sobre el modelo y las industrias priorizadas. Esta herramienta servirá como apoyo operativo para los equipos de marketing y comercial, estandarizando los criterios de calificación de leads y optimizando el uso de la inversión publicitaria.

En síntesis, la metodología combina el rigor del análisis estadístico con el potencial del aprendizaje automático para construir una herramienta predictiva de inteligencia comercial. Este enfoque permitirá reducir la dependencia de la intuición en la prospección de clientes, mejorar la eficiencia operativa y promover una toma de decisiones basada en evidencia dentro de la empresa gestora ambiental.

1.6 Resultado esperado

Se espera que el desarrollo e implementación del sistema de inteligencia comercial basado en ciencia de datos genera los siguientes resultados concretos:

1. Base de datos enriquecida y validada, que consolide información técnica, económica, estructural y normativa de empresas industriales en Ecuador, conformando un repositorio robusto y homogéneo de prospectos comerciales para la empresa gestora ambiental.
2. Modelo predictivo funcional y validado, capaz de estimar con precisión la probabilidad de conversión de cada prospecto en cliente. El modelo deberá alcanzar métricas de desempeño satisfactorias utilizando algoritmos supervisados como Random Forest, XGBoost o SVM.
3. Producto interactivo de inteligencia comercial que presenta los resultados del análisis predictivo de manera visual, accesible y operativa. Este producto incluirá visualizaciones que faciliten la interpretación de información que respalden la toma de decisiones estratégicas.
4. Propuesta metodológica replicable y escalable, que sienta las bases para la aplicación futura de modelos de inteligencia comercial basados en ciencia de datos en otras líneas de negocio o mercados del sector ambiental, contribuyendo al proceso de transformación digital y sostenibilidad corporativa de la organización.

Además, aportarán un marco técnico y metodológico innovador, alineado con las tendencias globales de transformación digital y con los Objetivos de Desarrollo Sostenible, promoviendo una gestión empresarial más inteligente, sostenible y basada en evidencia.

1.7 Dataset

El presente estudio se basa en el análisis de un conjunto de datos proporcionado por la empresa gestora ambiental del año 2024, el cual consolida información comercial, operativa y logística sobre la prestación de servicios a clientes industriales en Ecuador. El dataset está compuesto por 369.492 registros y 47 variables, constituyendo la fuente principal para el desarrollo del sistema de inteligencia comercial propuesto.

Estos datos se integran a bases de entidades públicas de Superintendencia compañía del año 2024 y datos del INEC para enriquecer la información.

Dado que el análisis se desarrolla en un contexto B2B industrial, donde no se dispone de datos individuales de comportamiento digital, la selección de variables se enfoca en dimensiones firmográficas, contextuales, regulatorias y operativas de las empresas. La unidad de análisis corresponde a la entidad jurídica (empresa), y no al usuario final.

1.7.1 Origen y proceso de recopilación

El dataset fue extraído de los sistemas SAP de la empresa, integrando información proveniente de los módulos de logística, facturación y gestión de clientes. A este conjunto se añadió un segundo bloque de datos que incluye información cualitativa sobre oportunidades comerciales, segmento industrial, tamaño de empresa, ingresos anuales estimados y tipo de residuo generado. Parte de estos atributos proviene de fuentes públicas, como la Superintendencia de Compañías, Valores y Seguros (SCVS).

Asimismo, se incorporaron variables derivadas de la revisión de sitios web corporativos e informes públicos, con el fin de construir indicadores de madurez ambiental y cumplimiento regulatorio.

Durante la revisión técnica, se identificaron registros del sistema CRM (Salesforce) con inconsistencias de completitud y duplicidad en campos como estado del lead o fase comercial. Debido a su baja confiabilidad histórica, estas variables fueron excluidas del modelo en esta versión, priorizando la integridad analítica y la estabilidad predictiva.

El conjunto de datos integra, en su mayoría, registros de entregas de residuos, servicios prestados, pesos, precios, fechas, clientes y asesores comerciales, así como atributos del contexto empresarial que permiten caracterizar el perfil de cada prospecto.

1.7.2 Variables seleccionadas para el análisis

Las variables relevantes fueron seleccionadas a partir de un proceso de evaluación técnica y se agrupan en cinco dimensiones analíticas principales:

1. Identificación y perfil empresarial: código de cliente, nombre, tipo de persona jurídica, ingresos anuales, tamaño de empresa.
2. Segmento industrial: sector productivo, subsegmento económico, ubicación geográfica (provincia, ciudad).
3. Información operativa: tipo de residuo, peso gestionado, cantidad entregada, tipo de servicio o tratamiento aplicado.
4. Indicadores comerciales: oportunidades detectadas, precios, montos facturados, estado de entrega o servicio.
5. Información administrativa: asesor comercial asignado, dirección de facturación, certificaciones y observaciones de cumplimiento.

Tabla 1.1.1 Datos a utilizar para el preprocesamiento de las bases

Nombre	Tipo	Descripción
Ruc	String	Número único de identificación tributaria de la empresa.
Compañía	String	Nombre de la empresa o cliente.
Provincia	String	Provincia en la que se ubica la empresa.
Ciudad	String	Ciudad en la que se encuentra la empresa.
Market segments	String	Segmento de mercado al que pertenece la empresa (e.g., Industrial, Comercial).
Tipo de lead	String	Clasificación del lead
Tamaño empresa	String	Tamaño de la empresa basado en ingresos o empleados (e.g., Pequeña, Mediana, Grande).
Línea de negocio	String	Sector o actividad económica principal de la empresa.
Es cliente	Integer	Indicador binario (1 = Cliente, 0 = No Cliente).

Estas variables serán utilizadas para el entrenamiento de modelos de clasificación supervisada, así como para la segmentación y priorización de prospectos comerciales en el desarrollo del sistema de inteligencia propuesta

CAPÍTULO 2

2. ESTADO DEL ARTE

2.1 Contextualización del Lead Scoring en entornos B2B industriales

"En los procesos de ventas entre empresas (B2B), los equipos de marketing y ventas se enfrentan al reto de identificar, cualificar y priorizar clientes potenciales. La cualificación de clientes potenciales es una tarea crítica porque impacta las tasas de conversión y maximiza la eficacia de los esfuerzos y estrategias de marketing y ventas" (González, Rubiano, & Sosa, 2025)

Dentro de los principales desafíos que enfrentan las empresas en el mercado, está la clasificación y priorización de potenciales clientes o leads. Con el objetivo de resolver esta problemática muchas han optado por el uso de plataformas de CRM (Customer Relationship Management) que facilitan de forma automatizada, la priorización de prospectos por un sistema de Lead Scoring. Esta automatización asigna una puntuación cuantitativa a los prospectos, reflejando su probabilidad de convertirse en clientes reales por medio de su comportamiento y de la información entregada en los campos de los formularios.

Esta priorización toma aún mayor relevancia en entornos B2B industriales, debido al ciclo de venta prolongado, el alto valor por transacción, y la alta inversión operativa que requiere la prestación de servicios complejos, como la disposición final de residuos peligrosos, lo que obliga a los equipos comerciales a enfocar sus recursos exclusivamente en las oportunidades de mayor valor estratégico y potencial de rentabilidad.

Esta dinámica separa al mercado B2B industrial del retail o servicios de consumo masivo, donde los leads pueden convertirse rápidamente y con mayor frecuencia, sin embargo el valor por transacción es bajo. Por

consiguiente, esto expone la importancia de cada contrato, en el que uno solo puede facturar miles o millones de dólares al año bajo la descripción de servicios complejos como, en este caso, disposición final de residuos peligrosos. Este tipo de servicios suelen implicar compromisos de largo plazo que incluyen transporte especializado, cumplimiento regulatorio, tratamiento en plantas y seguimiento ambiental, por lo que cada oportunidad comercial tiene un impacto estratégico elevado, es decir una inversión mal dirigida en prospectos poco cualificados no solo incrementa costos de ventas, sino que también representa un alto costo de oportunidad en mercados donde los clientes potenciales son limitados y la complejidad del servicio establece un umbral de inversión que solo ciertos segmentos empresariales pueden asumir (Tomar & Shriram, 2023).

Por esta razón la literatura académica (Wu, Andreev, & Morad, 2023) resalta la importancia de los métodos de “lead scoring” y los clasifica en tradicionales y predictivos. Los modelos tradicionales se basan en reglas manuales, donde se asignan puntos a variables como el tamaño de la empresa, la industria o el cargo del contacto. Aunque este tipo de modelos es simple de implementar, carecen de precisión y de la capacidad de ajustarse automáticamente a nuevas tendencias o cambios en el comportamiento de los clientes, lo que puede resultar en priorizaciones inexactas frente a mercados cambiantes (Wu, Andreev, & Morad, 2023). Por el contrario, los modelos predictivos utilizan algoritmos de machine learning (ML) para encontrar patrones históricos y actualizar el score, basándose en la entrada, comportamiento y evolución de la nueva data.

2.2 Evolución y comparación de algoritmos de clasificación en Lead Scoring B2B

En una primera etapa, la regresión logística y los árboles de decisión constituyeron las herramientas más comunes en el Lead Scoring (Wu, Andreev, & Morad, 2023). Con esta afirmación se destacan la sencillez y explicabilidad de estos modelos, sin embargo, en escenarios más complejos con múltiples variables, presentan limitaciones para captar interacciones no

lineales. Este tipo de comportamiento es muy habitual en mercados industriales, en el que la dinámica se ve afectada por múltiples variables internas y externas.

Posteriormente, la literatura señala el surgimiento de los métodos de ensamble, entre los cuales el Gradient Boosting (GBM) representó un hito al combinar múltiples clasificadores débiles para poder generar un modelo más robusto, reduciendo sesgo y mejorando la precisión (Friedman, 2001). Su aplicación en el lead scoring ha demostrado ser efectiva como en el caso de estudios en compañías de software industrial muestran que GBM superó consistentemente a la regresión logística y a los árboles de decisión en métricas como el AUC y la sensibilidad (Wu, Andreev, & Morad, 2023). El Gradient Boosting marcó un punto de inflexión en la disciplina, sentando las bases para algoritmos más eficientes que hoy dominan la práctica.

Una vez puestos estos cimientos, se integran métodos de boosting optimizados como XGBoost y LightGBM. XGBoost incorpora mecanismos de regularización y una gestión más eficiente de valores faltantes, lo que lo hace particularmente adecuado para datasets industriales con información incompleta o desbalanceada (Ha-Thuc & Sinha, 2016).

LightGBM, por su parte, optimiza aún más el proceso mediante técnicas como Gradient-based One-Side Sampling (GOSS) y Exclusive Feature Bundling (EFB), que reducen el tiempo de entrenamiento sin sacrificar precisión (Gabriel, 2024). Estas mejoras lo convierten en una alternativa de gran valor en proyectos que requieren escalabilidad y rapidez, como el procesamiento de bases de datos de prospectos industriales con miles de registros.

Por último, se encuentra el modelo de Random Forest, que aunque no pertenece a la familia de boosting, se mantiene vigente como un referente por excelencia debido a su estabilidad, robustez frente a sobreajuste y facilidad de interpretación. En escenarios B2B industriales, su uso se justifica cuando el volumen de datos no es lo suficientemente grande para justificar algoritmos

más complejos o cuando la prioridad es contar con explicaciones claras para equipos comerciales y de gestión (Verma & Dong, 2016)

Para el caso de Ecuador, la aplicación de algoritmos cobra aún mayor relevancia debido a la disponibilidad limitada de los datos. Los segmentos industriales estratégicos definidos por la CIIU Rev.4, como manufactura de químicos, minería, alimentos y bebidas, cemento y construcción, generación de energía y gestión de residuos son altamente regulados debido a los procesos de producción, lo que los convierte en prospectos valiosos para la gestión ambiental.

2.3 Aplicaciones reales de Lead Scoring con Machine Learning en entornos industriales y servicios B2B

La aplicación de algoritmos de machine learning en procesos de lead scoring B2B ha demostrado resultados tangibles en la mejora de la eficiencia comercial, reduciendo el tiempo de prospección y aumentando la tasa de conversión de oportunidades (Wu, Andreev, & Morad, 2023). La literatura científica y empresarial documenta varios casos relevantes que permiten entender el impacto de estas metodologías en distintos sectores industriales.

2.4 Modelo supervisado en manufactura B2B (Rubiano, 2025)

En el estudio desarrollado por Rubiano, se analizó la implementación de un modelo de clasificación supervisado en una compañía manufacturera B2B, lo que permitió automatizar la priorización de prospectos y mejorar la precisión en la identificación de oportunidades con verdadera intención de compra. El impacto más importante se vió reflejado en la optimización de los recursos comerciales, al concentrar sus esfuerzos en leads con mayor probabilidad de conversión, reduciendo los costos asociados a visitas comerciales y mejorando los indicadores de rentabilidad (González, Rubiano, & Sosa, 2025).

2.5 Estudio experimental de Nygård y Mezei (2020)

Nygård y Mezei, en su estudio *Automating Lead Scoring with Machine Learning: An Experimental Study*, evidenciaron cómo los modelos predictivos superaron a los modelos basados en reglas en un conjunto de datos con miles de prospectos. Los algoritmos de boosting lograron diferenciar con mayor precisión los leads de calidad, corrigiendo la toma de decisiones de los equipos comerciales, que, de forma empírica, sobrevaloras prospectos de gran tamaño, pero con baja intención real de compra. Estos resultados subrayan que la capacidad de priorización basada en datos permite aumentar la calidad del *pipeline* y disminuir pérdidas por perseguir clientes sin potencial inmediato (Mezei & Nygård , 2020).

2.6 Codificación bayesiana en WeWork (Slakey et al., 2019)

En un caso de aplicación corporativa, Slakey documentaron cómo WeWork empleó *Bayesian target encoding* para tratar variables categóricas de alta cardinalidad en su sistema de *lead scoring*. La mejora obtenida fue significativa: el AUC pasó de 0.87 a 0.97, optimizando la predicción de cierre de contratos a partir de bases de datos con miles de *leads* provenientes de distintas industrias. Este resultado es especialmente relevante para contextos B2B industriales como el ecuatoriano, donde la información categórica como actividad económica, tipo de empresa, ubicación regional, constituyen una de las principales fuentes de datos (Slakey, Salas, & Schamroth, 2019).

2.7 Técnicas de ranking y explicabilidad en LinkedIn

En plataformas empresariales como LinkedIn, la integración de modelos explicativos ha sido esencial para priorizar resultados y mejorar la transparencia de los sistemas predictivos. En su estudio Learning to Rank Personalized Search Results in Professional Networks, los investigadores de LinkedIn desarrollaron un modelo de ranking basado en aprendizaje supervisado, optimizado con explicabilidad y métricas de relevancia contextual (Ha-Thuc & Sinha, 2016). Este enfoque evidencia la aplicabilidad

de técnicas como SHAP que analizan la influencia de cada variable sobre el resultado, fortaleciendo la confianza de los equipos comerciales y de marketing en los resultados generados por los algoritmos.

2.8 Revisión de variables predictoras utilizadas en modelos de Lead Scoring B2B industriales

Es fundamental realizar una correcta selección de variables predictoras, ya que esto determina la efectividad del modelo de lead scoring B2B. En la literatura, estas variables se agrupan principalmente en tres categorías: firmográficas, contextuales y comportamentales (también denominadas conductuales) (Wu, Andreev, & Morad, 2023).

En los entornos industriales, los clientes son organizaciones y no individuos, lo que limita la disponibilidad de datos de comportamiento digital y la convierten en fragmentaria o indirecta; por ello, los modelos se apoyan fundamentalmente en variables firmográficas y contextuales (Mezei & Nygård , 2020) (Ha-Thuc & Sinha, 2016).

Aterrizando esto al contexto ecuatoriano, el *Módulo de Información Económica Ambiental en Empresas* (INEC) reporta para 2023 que el 96 % de empresas generan residuos no peligrosos (11,55 millones de toneladas), 76 % desechos especiales (109.087 t) y 92 % desechos peligrosos (709.082 t). De estas empresas solo el 30,9 %, 29,2 % y 32,2 %, respectivamente, declararon conocer la cantidad generada (INEC, 2025). Esta información confirma la transversalidad de la generación de residuos y la falta de datos cuantificables, lo que respalda la decisión de priorizar variables firmográficas y contextuales e incorporar ambientales verificables.

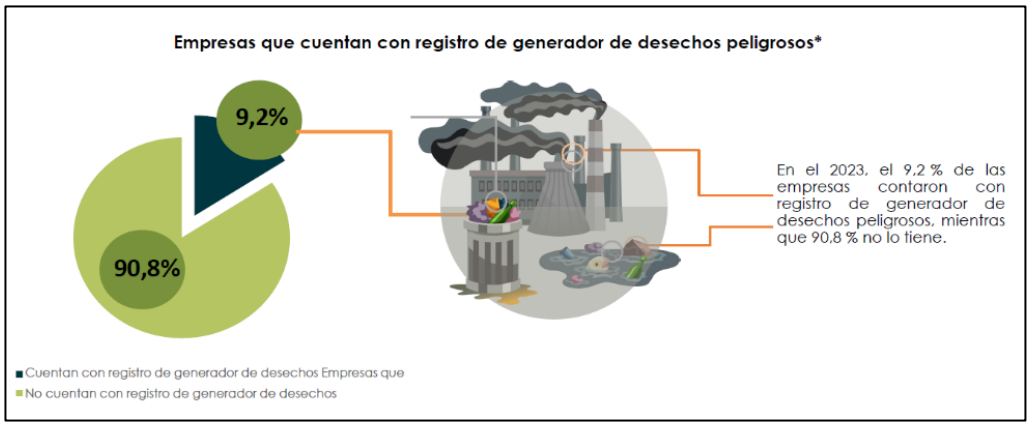

Figura 2.1. Distribución de empresas registradas como generadores de desechos

2.8.1 Variables firmográficas

Estas describen las características estructurales de las empresas como el sector económico, tamaño (segmento), ubicación geográfica y forma jurídica, por lo que son utilizadas en modelos B2B por su estabilidad, trazabilidad y capacidad para capturar diferencias en la organización y escala de las empresas (Wu, Andreev, & Morad, 2023).

En el caso de las características de actividad económica en el sector ecuatoriano, estas se encuentran estandarizadas bajo la Clasificación Industrial Internacional Uniforme (CIIU Rev. 4 A.C.) del INEC. Esta clasificación no mide comportamiento sino la naturaleza productiva de la empresa (p. ej., industrias manufactureras, químicos, construcción), lo que permite segmentar leads de manera homogénea entre industrias.

De acuerdo con el tablero Ranking 2024 – Resumen gráfico de la Superintendencia de Compañías, Valores y Seguros (SCVS), en 2024 presentaron balances 134.754 compañías y 3.778 pertenecen al segmento “Grande”, equivalente al 2,8 % del universo (SCVS, 2025). Este recorte es consistente con mercados B2B industriales caracterizados por alto valor por transacción y menor frecuencia de compra, y con una mayor trazabilidad de la información corporativa (Mezei & Nygård , 2020).

Aunque la generación de residuos es transversal en el mercado (como lo demuestran los datos del INEC), la estrategia de mercado para servicios de alto costo y alta complejidad operativa prioriza el enfoque en el 2.8% de las empresas 'grandes'. Este enfoque asegura que el modelo se centre en la máxima rentabilidad potencial y la optimización de los recursos comerciales en cuentas con el mayor valor de vida (LTV), que son las únicas capaces de absorber la estructura de costos de los servicios ofrecidos.

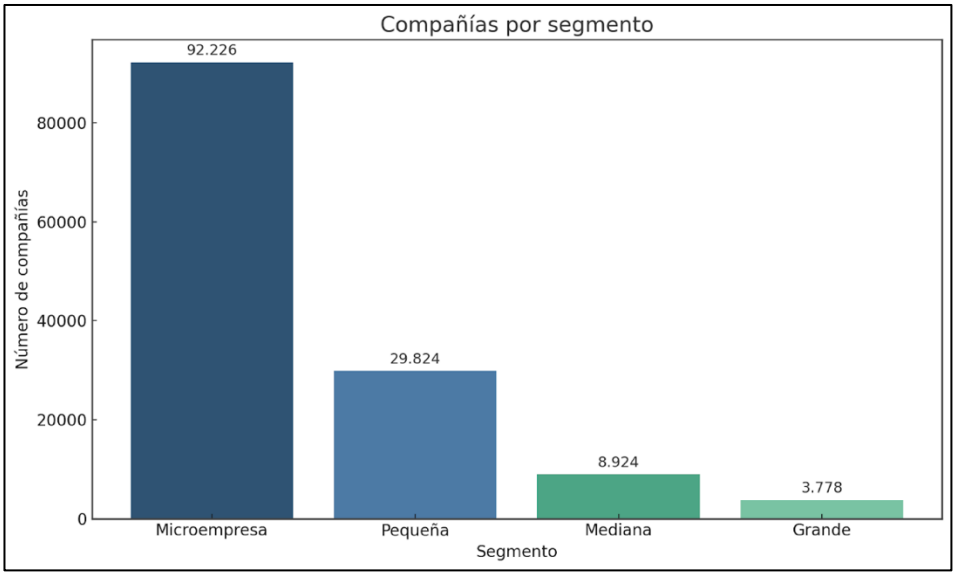

Figura 2.2 Distribución de empresas según SRI

En la Figura 2 se puede observar que, del total de 134 754 compañías que presentaron balances ante la Superintendencia de Compañías, Valores y Seguros (SCVS), 3 778 pertenecen al segmento “Grande”, equivalente al 2,8 % del universo empresarial. Este grupo concentra los principales actores industriales del país, caracterizados por un alto valor por transacción y mayor trazabilidad en su información financiera.

En el contexto nacional, el desempeño ambiental empresarial puede verse evidenciado mediante referentes del sector manufacturero químico, particularmente en la industria de pinturas y recubrimientos. Por ejemplo, la empresa Pinturas Unidas reporta indicadores internos de reuso y reproceso

mensual de materiales que reflejan prácticas alineadas con los principios de economía circular. En 2024, la compañía alcanzó una tasa anual de recuperación del 61,05 % sobre el total de producto ingresado, con una eficiencia de proceso del 100 % en conformidad de producto y una meta de recuperación mínima del 75%. Estos resultados evidencian la factibilidad de incorporar variables contextuales ambientales en la evaluación de desempeño industrial y apoyan el desarrollo de modelos predictivos que consideren

TABLA DE REPROCESO MENSUAL				
2024				
Mes	**Producto ingresado Pinturas y afines**	**Producto Recuperado**	**Producto de baja Pintura y afines**	**% Recuperado relacion ingreso**
Enero	5042,14	931,68	5761,9	18,48
Febrero	2711,39	692,9	1738,16	25,56
Marzo	1711,81	1402,71	771,12	81,94
Abril	3032,42	1725,25	22,13	56,89
Mayo	2888,42	742,5	956,96	25,71
Junio	1604,72	2934,65	0	182,88
Julio	1556,51	1575,17	1414,78	101,2
Agosto	1408,83	1468,22	0	104,22
Septiembre	3530,81	2980,75	722,63	84,42
Octubre	2218,24	1418,5	406,79	63,95
Noviembre	2440,02	1432,83	540,75	58,72
Diciembre	7158,04	4247,27	943,38	59,34
Total	35303,35	21552,43	13278,6	
% Total recuperado / ingresado x 100				61,05

Figura 2.3 Indicadores de gestión ambiental de una empresa de pinturas

2.9 Asignación determinista por actividad

Para enriquecer el modelo sin encadenar modelos adicionales, se construye un diccionario documental CIIU (nivel división, 2 dígitos) con sus familias de residuos a partir de INEC–MIEAE (INEC, 2025). Cada división CIIU se etiqueta con las familias de residuos habitualmente presentes en esa actividad en el contexto ecuatoriano. Con ese diccionario se cruza el dataset y se generan tres variables binarias:

- **prob_np** = 1 si la división CIIU de la empresa está etiquetada con residuos no peligrosos; en caso contrario 0.

- **prob_especial** = 1 si la división está etiquetada con desechos especiales; en caso contrario 0.
- **prob_peligroso** = 1 si la división está etiquetada con desechos peligrosos; en caso contrario 0.

Desechos especiales	Empresas que generan	Empresas que conocen	kilotoneladas (kT)	
	Recuento	Recuento	Absoluto 2023	Relativo 2023
Escorias de acería*	70	14	80,0	73,3%
Neumáticos usados	12.446	3.199	21,4	19,6%
Equipos eléctricos y electrónicos en desuso	1.690	731	4,0	3,7%
Aceites vegetales usados generados en procesos de fritura	634	118	2,7	2,4%
Envase de agroquímicos y otros químicos tóxicos	386	328	1,0	0,9%

Figura 2.4 Generación de desechos especiales de las industrias según ENESEM

Residuos No Peligrosos	Empresas que generan	Empresas que conocen	kilotoneladas (kT)	
	Recuento	Recuento	Absoluto 2023	Relativo 2023
Chatarra Liviana	1.471	1.282	8.674,4	75,1%
Escombros de construcción	513	254	1241,3	10,7%
Cartón	10.974	3.337	1007,8	8,7%
Orgánicos	3.102	1.099	304,5	2,6%
Otros Residuos No Peligrosos *	25.830	8.008	319,4	2,8%

Figura 2.5 Generación de residuos no peligrosos de la industrias según ENESEM

Desechos peligrosos	Empresas que generan	Empresas que conocen	kilotoneladas (kT)	
	Recuento	Recuento	Absoluto 2023	Relativo 2023
(E.38.02) - Lixiviados generados en vertederos, rellenos y celdas de seguridad	1	1	292,9	41,3%
(G.46.01) - Lodos de las plantas de tratamiento de aguas residuales industriales que contienen sustancias peligrosas	357	302	117,8	16,6%
(B.06.02) - Lodos, ripios y desechos de perforación en superficie que contienen, hidrocarburos, HAP's, Cadmio, Cromo (VI), Vanadio, Bario, Mercurio, Níquel	7	7	67,4	9,5%
(Q.86.08) - Fármacos caducados o fuera de especificaciones	656	514	43,9	6,2%
Otros desechos peligrosos*	2.590	904	187,2	26,4%

Figura 2.6 Gestión de desechos peligrosos en las industrias según ENESEM

Tomar en cuenta estas variables, no significa inferir cantidades ni probabilidades, sino identificar o etiquetar cada industria mediante su actividad, lo que es auditable y reproducible.

2.9.1 Variables contextuales

Las variables contextuales se refieren a aquellas que influyen en la generación y gestión de residuos dentro del marco regulatorio aplicable. En modelos de lead scoring B2B resultan útiles para la segmentación territorial, el dimensionamiento del mercado y priorización de cuentas cuando la información conductual es limitada (Wu, Andreev, & Morad, 2023) (Mezei & Nygård , 2020).

2.10 Ubicación geográfica por provincia

Esta variable se incorpora siguiendo la codificación oficial del INEC, lo que permite comparar empresas por territorio y sector económico (CIIU). Como referencia pública, el tablero Ranking 2024 – Resumen gráfico de la Superintendencia de Compañías, Valores y Seguros (SCVS) identifica a

Guayas como la provincia con mayor ingreso total de 68.123.778.591,88 en 2024, seguida por Pichincha con 58.059.390.998,24 lo que confirma la concentración de actividad empresarial en estos polos industriales (Superintendencia de compañías, 2024).

Este indicador sustenta la conveniencia de publicar scores y deciles del modelo por provincia o sector, con el fin de orientar al equipo comercial, priorizando áreas de alta densidad empresarial.

En provincias con alta concentración industrial y abarcados por la cobertura de gestores se observa un entorno competitivo más desarrollado, mientras que en zonas con menor concentración podría existir oportunidad de expansión para nuevos operadores, especialmente cuando hay concentración de industrias sin cobertura directa.

Sin embargo, de acuerdo al Listado de Gestores Ambientales sobre los principales gestores ambientales del Ecuador como ADS-PECS, Incinerox, Geocycle, Hazwat, Ecoreesa, Pecs Ambiente y Plusambiente, la mayoría declara cobertura nacional y presencia operativa en las tres regiones principales (Sierra, Costa y Oriente) (Gobierno del Ecuador, 2023). Esta cobertura generalizada reduce la necesidad de una para explicar diferencias territoriales en la generación o gestión de residuos. Por esta razón, el modelo prescinde de esta variable como predictora y la conserva únicamente con valor descriptivo, dentro del contexto competitivo nacional

2.11 Cobertura del registro y brecha de información

El Módulo de Información Económica Ambiental en Empresas (MIEAE) del INEC señala que, en 2023, solo el 9,2 % de las empresas contaban con el Registro de Generador de Desechos Peligrosos, mientras que el 90,8 % no lo tenía (INEC, 2025). Esta información evidencia la limitada formalización y

trazabilidad de los procesos de gestión de residuos en la industria ecuatoriana.

2.11.1 Variables e interés ambientales

Uno de los factores que pueden enriquecer este estudio, en conjunto con los anteriormente mencionados, son los de carácter regulatorio ambiental, que reflejan el nivel de compromiso de cada uno de los prospectos ante la sostenibilidad, lo que los obliga a adherirse a políticas activas o que participan en programas ambientales de reconocimiento público, Esto puede ser un indicador relevante del grado de madurez ambiental y potencial de colaboración con una gestora de residuos, por lo que en este sentido, se consideran tres fuentes principales:

- Empresas adherentes al Programa Ecuador Carbono Cero (PECCS), publicadas por la Revista Ekos en el año 2022.
- Compañías que reportan indicadores ambientales bajo el estándar del Informe de Sostenibilidad ESG (Environmental, Social and Governance), disponibles en los portales corporativos y reportes públicos.
- Empresas certificadas con la norma ISO 14001, enfocada en sistemas de gestión ambiental.

Cada una de estas variables se codifica como binaria (1 = participa o posee la certificación; 0 = no). Su incorporación al *dataset* permite al modelo identificar organizaciones con mayor compromiso ambiental, aprovechando la necesidad por alcanzar las metas de sostenibilidad y economía circular promovidas por el sector.

En resumen, la revisión del estado del arte evidencia que los modelos predictivos en el área del lead scoring B2B basados en variables firmográficas y contextuales permiten a los diferentes stakeholders, construir una visión más completa del mercado industrial ecuatoriano, lo que transforma la toma de

decisiones en una acción orientada a resultados basada en datos. A pesar de que los registros en fuentes oficiales son limitados, se pueden encontrar datos firmográficos y contextuales en fuentes como el INEC y SCVS que aportan estructura y trazabilidad a la construcción del dataset. Estos se refuerzan con variables categóricas provenientes de indicadores ambientales como presencia en PECCS, reportes ESG e ISO 14001, lo que enriquece las variables, refuerza el valor predictivo del modelo y le brinda un factor diferenciador frente a los demás modelos aplicados y revisados previamente en el sector industrial.

CAPÍTULO 3

3. DISEÑO

3.1 Introducción a la exploración de datos

La exploración y validación de datos constituye una fase crítica en cualquier proyecto de ciencia de datos, ya que define la calidad del insumo que alimentará al modelo predictivo y, en consecuencia, su capacidad para generar resultados confiables (Provost & Fawcett, 2013)

En el siguiente capítulo se realiza la descripción de la metodología utilizada para realizar el diseño, construcción y validación del modelo predictivo de lead scoring orientado a identificar empresas con alto potencial de convertirse en clientes del segmento ambiental. A partir del contexto plasmado en el capítulo 2, se ponen en práctica las variables firmográficas, contextuales y ambientales mediante un enfoque de aprendizaje supervisado (supervised machine learning).

Estos procesos se realizan con el objetivo de desarrollar un modelo que permita estimar la probabilidad de que una empresa en el mercado ecuatoriano requiera los servicios de gestión ambiental, haciendo uso de

fuentes públicas y corporativas verificables. Los datos integrados para llevar a cabo el proyecto provienen de fuentes oficiales como a Superintendencia de Compañías, Valores y Seguros (SCVS), del Instituto Nacional de Estadística y Censos (INEC), del Programa Ecuador Carbono Cero (PECC) y de los registros de certificaciones ISO 14001 y ESG reportadas en portales públicos.

3.2 Metodología general del proyecto

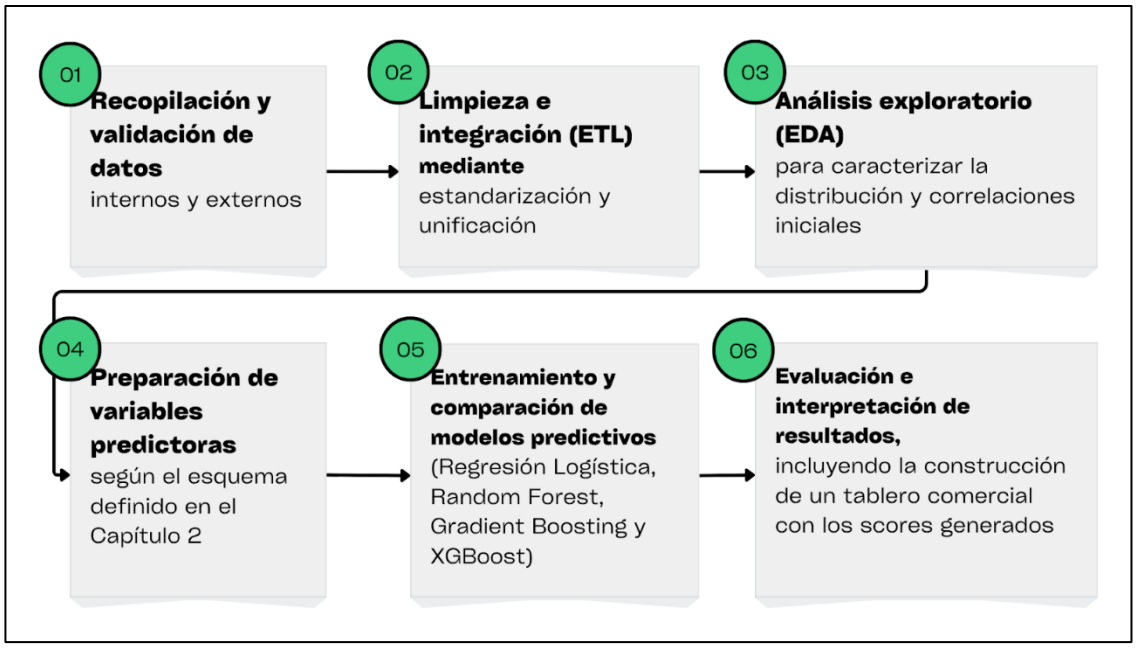

Figura 3.1 Esquema general de la metodología del proyecto

El proyecto adopta una metodología de enfoque cuantitativo, justificada por la necesidad de integrar datos estructurales y trazables. La selección de este enfoque obedece a la intención de superar las limitaciones observadas en sistemas de lead scoring previos, los cuales se han caracterizado por la insuficiencia de información cuantitativa verificable de las empresas. Por consiguiente, el estudio se centra en el desarrollo de un modelo predictivo basado en el análisis de registros e indicadores públicos, con la finalidad de identificar aquellas organizaciones con mayor probabilidad de requerir los servicios de gestión de residuos ofrecidos por la gestora ambiental.

La propuesta se basa en un diseño no experimental y transversal al analizar conjuntos de datos existentes, tanto internos como externos, correspondientes al año 2024, que no dependen de la manipulación para la generación de variables. Esto con el propósito de tener un análisis transparente que observe las relaciones existentes entre las características firmográficas, contextuales y ambientales mencionadas en el capítulo 2, junto con su condición durante el periodo 2024 para ser consideradas como prospectos y así contribuir a la construcción de un modelo predictivo explicativo.

Como señalan Batini & Scannapieco, un modelo de analítica avanzada no puede compensar deficiencias estructurales en los datos, por lo que la validación de su integridad, consistencia y actualidad es indispensable antes de proceder a fases de preprocesamiento o modelado (Stefani & Vassiliadis, 2021).

Durante el análisis se incluyen fuentes externas provenientes de organismos oficiales como el Instituto Nacional de Estadística y Censos (INEC, 2025), la Superintendencia de Compañías, Valores y Seguros (Superintendencia de compañías, 2024), el Programa Ecuador Carbono Cero (PECCS) y la revista Ekos (Ekos, 2022), estas se utilizan únicamente con el fin de enriquecer el contexto y generar variables ambientales complementarias binarias, por lo que el estudio no evalúa la evolución de cada una de las empresas sino su situación estructural y contextual dentro del año 2024.

Metodológicamente, el proyecto se apoya en un método analítico–correlacional, que busca identificar patrones estadísticos entre las variables predictoras y la probabilidad de conversión de los leads. Este enfoque permite combinar el rigor del análisis estadístico con la aplicabilidad de los modelos de aprendizaje automático, garantizando la reproducibilidad del proceso y la interpretabilidad de los resultados para los equipos comerciales.

3.3 Origen y recopilación de datos

El modelo se alimenta de un conjunto de datos heterogéneos, tanto estructurados como semiestructurados, que se clasifican en internos y externos, según su naturaleza. Esta integración de fuentes busca construir una base consolidada que permita identificar patrones relevantes entre las características de las empresas y su probabilidad de convertirse en clientes.

3.4 Fuentes internas

Estas corresponden a la información propia de la empresa, generada por sus sistemas comerciales y operativos durante el año 2024. La fuente principal corresponde a la base de datos histórica de ventas por gestión de residuos a diferentes clientes industriales entregada por la empresa gestora, con 74.269 registros y 48 variables provenientes de SAP (módulos de logística, facturación y gestión de clientes). Es importante destacar que este elevado número de registros se debe a que la información está organizada a nivel de transacción, lo que implica que un mismo cliente puede aparecer múltiples veces en función de las operaciones realizadas a lo largo del año.

En consecuencia, el número real de clientes únicos se reduce a 1513, segmentados por sector y actividad económica, tipo de persona, RUC, ubicación geográfica, monto y número de facturación, cantidad de los desechos/residuos recolectados por kg, frecuencia, descripción del servicio, asesor asignado y rutas.

Dado que para entrenar el modelo se requiere de un dataset completo que incluya datos tanto de clientes como de no clientes, vamos a considerar únicamente las siguientes 6 variables:

- Nombre de Cliente distingue a las organizaciones activas dentro de la cartera comercial
- Ciudad Sucursal, Provincia Sucursal, DirecciónSucursal indican la ubicación geográfica de la empresa a la que se le realiza el servicio

- Cód.Según Listado Nacional (CIIU) clasifica según la actividad económica de la empresa la tipología de la Superintendencia de Compañías, Valores y Seguros (SCVS, 2024)
- Descripción del servicio detalla la gestión ambiental ofrecida en el periodo 2024

La naturaleza de los datos anteriormente mencionados se considera estructurada ya que se almacenan en el SAP como formatos tabulares (Excel o CSV) y mantienen consistencia de campos y codificación. Gracias al carácter institucional de esta información se garantiza la trazabilidad y confiabilidad para desarrollar el modelo predictivo empresarial. Esta fuente permite además generar la variable dependiente binaria **potencial_cliente**, que identifica si la empresa pertenece o no a la cartera activa de la empresa gestora ambiental.

3.5 Fuentes externas

Estas corresponden a la información que proviene de organismos públicos y privados, que refuerzan el perfil estructural (tamaño, sector, ubicación), desempeño financiero y potencial generación de residuos (según su actividad) tanto de prospectos como de clientes actuales de la empresa gestora ambiental.

El formato es semiestructurado, ya que puede extraerse de informes, PDFs o portales web, por lo que se requiere de una preparación, extracción, limpieza y estandarización previos al uso dentro del dataset principal.

Además, estas fuentes nutren el dataset con datos de indicadores verificables de desempeño ambiental, cumplimiento regulatorio y compromiso institucional, que definen el interés ambiental o afinidad con políticas sostenibles de las empresas, a su vez que fortalecen la interpretación de la variable **interes_ambiental** definida en el Capítulo 2.

3.5.1 Instituto Nacional de Estadística y Censos (INEC, 2025)

Aporta indicadores contextuales sobre la generación y manejo de residuos peligrosos, especiales y no peligrosos. A partir de esta información se derivan las variables binarias prob_np, prob_especial y prob_peligroso, que permiten inferir el tipo de residuos predominantes por sector CIIU.

3.5.2 Superintendencia de Compañías, Valores y Seguros (SCVS, 2024)

Se obtienen variables financieras y estructurales, tales como ingresos por ventas, activos, patrimonio, utilidad neta, número de empleados, rentabilidad (ROE, ROA), márgenes brutos y operacionales, liquidez y apalancamiento, además de los campos firmográficos de provincia, código de segmento y clasificación CIIU. Estas variables constituyen la base cuantitativa principal del modelo, ya que reflejan la capacidad económica, el tamaño y la actividad productiva de cada empresa.

3.5.3 Listado de Empresas Adherentes al Programa Ecuador Carbono Cero publicado por Revista Ekos (2022).

El Programa Ecuador Carbono Cero (PECCS), administrado por el Ministerio del Ambiente, Agua y Transición Ecológica (MAATE), brinda la información de las empresas que buscan reducir las emisiones de CO2 dentro de su cadena de producción para aportar a la sostenibilidad financiera de las iniciativas de conservación a través de la primera Norma de Compensación avalada por Ecuador, enmarcada en el marco regulatorio nacional. Además, se incluyen registros públicos de certificaciones ISO 14001, información disponible en portales web corporativos de empresas que han divulgado informes de sostenibilidad o ESG que ayudan a la construcción de la variable permitiendo identificar aquellas empresas con gestión ambiental certificada. Esta combinación de fuentes permite identificar empresas con gestión ambiental formalmente certificada y con alto grado de compromiso institucional.

3.6 Diseño del dataset

Durante esta fase del proceso metodológico se define la estructura. tipología y formato de los datos que se utilizan en el dataset de entrenamiento, validación y evaluación del modelo predictivo. En esta fase se realiza la integración de las fuentes internas y externas mencionadas en los apartados anteriores, alcanzando así la compatibilidad, coherencia y representatividad del universo de empresas prospectos que se requiere analizar.

El propósito principal del diseño es consolidar en una base única, toda la información relevante, con las variables firmográficas, geográficas y ambientales anteriormente mencionadas, para identificar patrones claros que diferencien a los clientes actuales de la empresa gestora ambiental, con los prospectos de potenciales clientes. Esta estructura constituye la base analítica para el modelo de lead scoring orientado al sector industrial ecuatoriano.

3.6.1 Criterios de segmentación comercial

El dataset está diseñado en base a los criterios estratégicos de segmentación utilizados por la empresa gestora ambiental para identificar prospectos de interés dentro del mercado industrial, por lo que se pretende cumplir con las siguientes condiciones:

- **Tamaño empresarial:** 134.754 empresas registradas y clasificadas en 4 distintos niveles como “GRANDE”, "Medianas”, “Pequeñas” y “Microempresas” por la Superintendencia de Compañías, Valores y Seguros (SCVS), debido a su volumen de facturación y capacidad operativa. Según el interés de la empresa hay que considerar que el modelo priorice aquellas catalogadas como “GRANDE” empresas.
- **Actividad económica:** sectores con alto potencial de generación de residuos o uso intensivo de recursos (por ejemplo, manufactura, química, alimentos, metalmecánica, papelera y farmacéutica).

- **Ubicación geográfica:** presencia en provincias con concentración industrial (Guayas, Pichincha, Azuay, Tungurahua y Manabí).
- **Compromiso ambiental:** adhesión a programas de sostenibilidad o certificaciones ambientales verificables.

Estos criterios ayudan a delimitar las empresas o el universo objetivo con el que se requiere realizar el estudio, además de ser una guía para las variables que se incluyen en el dataset, garantizando que el dataset pueda ser reproducido y auditado en fases posteriores del proyecto, siguiendo las buenas prácticas de gestión de datos para proyectos de analítica avanzada recomendadas por Singh (Singh, 2024).

3.6.2 Integración y estructura general

Para el resultado del dataset final se necesitó de un proceso exhaustivo de depuración, normalización y correspondencia entre los distintos orígenes de la información, esto evidencia el nivel de dificultad de trabajo al integrar fuentes internas con externas al presentar diferentes grados de granularidad, formatos, criterios de codificación. Esto requirió de aplicar secuencias de transformaciones para garantizar la estructura unificada y consistente del dataset.

3.6.3 Integración de la base pública

Como fuente principal de los organismos públicos, se toma la información de la Superintendencia de Compañías, Valores y Seguros (SCVS). Estos datasets se encuentran disponibles al público desde la aplicación de la institución y permiten una visualización previa desde un dashboard de Power Bi por lo que se verificó que exista la información necesaria antes de descargar la metadata.

Los archivos se encuentran distribuidos en múltiples dominios contables (balances generales, estados de resultados, ratios financieros y datos firmográficos), por lo que, para lograr su consolidación, se realiza un proceso

de integración por etapas. En la unión inicial se logró mediante el campo “expediente” y posteriormente una segunda unión por el “RUC”, garantizando la información correspondiente para cada una de las entidades registradas y consolidar la información financiera y estructural en un solo dataset.

Como siguiente paso, se efectúa un proceso de preprocesamiento y control de calidad de datos basado en la detección y tratamiento de valores nulos, la eliminación de duplicados, la normalización de formatos (por ejemplo, conversión de variables textuales a categóricas o numéricas) y el control de consistencia entre variables interrelacionadas (como ingresos, activos y pasivos).

Durante esta etapa se conservan valores negativos en variables como patrimonio y utilidad antes de impuestos, al considerarse representativos de situaciones reales (como pérdidas o alto endeudamiento). Este preprocesamiento a detalle en los datos financieros permite detectar y depurar registros cuyo comportamiento no es coherente con la estructura contable esperada, verificando relaciones básicas como activos = pasivos + patrimonio, que las utilidades netas no superen los ingresos por ventas, y que los márgenes financieros y niveles de endeudamiento se mantengan dentro de rangos razonables para empresas operativas. Gracias a estas acciones se garantiza que las variables derivadas reflejen condiciones económicas reales y no errores de registro o digitación, muy comunes en cifras de portales públicos.

Como resultado final, se obtuvo una base pública consolidada que integra información identificatoria, financiera y estructural de 134.754 empresas, conformando el insumo base para la etapa siguiente de integración con fuentes ambientales.

3.6.4 Construcción de variables ambientales

Con el propósito de enriquecer y complementar la información financiera y estructural de las empresas se toman indicadores de sostenibilidad para

desarrollar variables ambientales derivadas de fuentes secundarias verificadas como lo son las revistas EKOS y Merco. Estas variables permiten identificar tanto el grado de compromiso público de las compañías con las políticas de sostenibilidad como su probabilidad de generar distintos tipos de residuos según su actividad económica.

El enriquecimiento empieza por la creación de una variable binaria llamada interes_ambiental, para la que se tomaron registros provenientes de dos fuentes, Listado de Socios Estratégicos Programa Ecuador Carbono Cero publicado por EKOS y los resultados del MERCO ESG Ecuador 2024.

La integración se realiza mediante un proceso de homologación de nombres empresariales entre ambas fuentes y la base de la SCVS, apoyado en métricas de similitud de texto y revisión manual de coincidencias. Este procedimiento permitió asegurar la correspondencia entre denominaciones legales y comerciales, depurando los falsos positivos y confirmando los casos reales de coincidencia.

A continuación, se procede en la construcción de tres variables adicionales, así mismo binarias **prob_np**, **prob_peligroso** y **prob_especial**, que reflejan la probabilidad de presencia o generación de residuos en cada una de las empresas según su actividad económica declarada (CIIU Rev. 4 A.C.).

Estas variables se elaboran mediante un mapeo experto por letra CIIU (nivel 1), empleando como referencia la información publicada en el Módulo de Información Económica Ambiental en Empresas (ENESEM 2023).

El mapeo permite clasificar las divisiones económicas de nivel 1 (A–U) en tres categorías de probabilidad:

- **Residuos no peligrosos (prob_np)**: actividades con generación regular de desechos ordinarios, orgánicos o reciclables (por ejemplo, manufactura de alimentos, textiles, madera y comercio minorista).

- **Residuos peligrosos (prob_peligroso)**: actividades industriales o extractivas con potencial de producir desechos químicos, metálicos o contaminantes (por ejemplo, fabricación de productos químicos, metalurgia, tratamiento de superficies o salud humana).
- **Desechos especiales (prob_especial)**: actividades con generación de desechos de gestión diferenciada, como neumáticos, aceites usados, equipos electrónicos o envases contaminados.

Como resultado del proceso se obtiene un diccionario que relaciona cada letra CIIU con su categoría de probabilidad de residuos, permitiendo la unión automática con el dataset principal de la SCVS a través de la variable **ciiu_n1**. Estas variables amplían la capacidad analítica del estudio, al examinar el comportamiento ambiental de los sectores industriales junto con sus características financieras y estructurales.

3.6.5 Preprocesamiento e integración de la base interna

La fuente interna se toma a partir de la base transaccional de la empresa gestora ambiental durante el año 2024, esta contiene todos los registros de ventas y servicios prestados a su cartera de clientes a nivel nacional. Esta base de información se obtiene del SAP en un formato tabular con un total de 74.269 registros y 48 variables representados por las transacciones individuales por cada cliente, tratamiento aplicado, peso y monto de facturación por los residuos gestionados.

Debido al formato original de este dataset es fundamental realizar un proceso de agrupación, depuración y normalización para consolidar la información y así construir una base única por empresa, compatible con la estructura de la base pública proveniente de la Superintendencia de Compañías (SCVS). Este proceso obligatorio permite la integración de ambos datasets.

3.6.6 Normalización y limpieza inicial

Como primer paso, se realiza la limpieza de los datos de identificación y ubicación, ya que estos son los que van a permitir identificar en el dataset final cuáles empresas son clientes de la empresa gestora ambiental y cuáles no. Luego se transformaron los nombres de las empresas a mayúscula y los de las columnas en minúscula, además de eliminar espacios y caracteres especiales que pueden causar ruido al momento de la integración. De igual manera se unificaron los formatos de provincias y ciudades.

El paso más importante dentro del procesamiento de las variables de identificación es la normalización del RUC al asegurar que cada valor contenga una longitud de 13 dígitos, para ello se diseña una función de corrección y así poder tratar las inconsistencias:

- RUC con 12 dígitos y terminado en “001”, se completa con unas cero iniciales.
- RUC entre 10 y 12 dígitos, se rellena con ceros a la izquierda hasta alcanzar los 13.
- Los valores no recuperables fueron marcados por "revisar" para su posterior inspección manual y en algunos casos realizar la corrección de este por medio de la búsqueda en fuentes de información pública.

Este procedimiento permitió mantener la trazabilidad de todos los registros, incluso de aquellos parcialmente incompletos, sin descartar información útil.

3.6.7 Construcción de la vista empresa

Una vez realizada la normalización y estandarización, los registros se agrupan por RUC, consolidando la información a nivel de cliente. Esta base inicial conserva columnas como **nombre_cliente_norm, provincia_norm, ciudad_norm, ciiu_n6_raw** y la letra CIIU de nivel 1 (**ciiu_n1_letra**), que

posteriormente serán estandarizadas con los nombres de las variables en el dataset público para la integración.

Este proceso generó un archivo unificado bajo el nombre de **df_clientes**, compuesta por 1509 clientes únicos, que representan el universo total de empresas que mantuvieron durante el 2024 una relación comercial con la empresa gestora ambiental.

En esta fase también se crea la variable **ciiu_n1_letra** en el dataset interno, obtenido a partir del primer carácter alfabético del código CIIU a nivel 6 que ya constaba desde el inicio. La creación de esta columna ayuda a simplificar la trazabilidad de la actividad económica sin ruido, reduciendo la granularidad para su posterior vinculación.

3.6.8 Integración de la base pública con la interna

Para lograr la integración de la información interna con la de la base pública consolidada de la SCVS, se usa la columna **RUC_norm** como clave común de enlace. Durante el proceso de unión se realiza un “left join” para mantener los registros de la base pública e incorporar las coincidencias correspondientes desde la base de la empresa gestora ambiental.

Esta unión nos permite obtener como resultado la variable objetivo del tipo binaria llamada **potencial_cliente** que adopta el valor de 1 cuando el RUC viene desde la base de clientes de y 0 cuando solo se encuentra en la base externa. La nueva columna permite distinguir de forma clara las empresas que ya son clientes y aquellas que constituyen potenciales prospectos dentro del universo de grandes compañías del país.

3.6.9 Resultado de la integración

Una vez unificada la información, se procede a aplicar los criterios comerciales mencionados en el apartado 3.4.1 con el objetivo de priorizar el análisis sobre las empresas clasificadas como “GRANDE” (categoría 4 según la tipología institucional de la SCVS). Sin embargo, no se limita el dataset final únicamente

a este segmento, sino que se conserva el universo completo de 134.754 registros, correspondiente a todas las compañías registradas en la SCVS. Esta decisión permite que el producto final del modelo pueda generar insights predictivos para cualquier empresa ingresada por RUC, asegurando la cobertura integral del mercado y manteniendo la trazabilidad de las empresas de todos los tamaños, lo que facilita a los equipos comerciales y de marketing identificar las diferentes empresas del mercado y su relevancia según los intereses establecidos. El segmento de grandes empresas se utiliza únicamente como subconjunto de referencia analítica dentro del estudio, dada su relevancia comercial y su peso dentro del mercado industrial ecuatoriano.

La base final consolidada denominada **df_merged_completo** integra las siguientes categorías de información:

- Identificación estructural: **ruc, nombre, tipo, provincia, ciiu_n1_norm, ciiu_n6_norm, ciiu_descripcion, segmento**.
- Variables financieras: **ingresos_ventas, activos, patrimonio, total_gastos, costos_ventas_prod, utilidad_neta, liquidez_corriente, roa, roe**, entre otras.
- Indicadores ambientales: **interes_ambiental, prob_np, prob_peligroso, prob_especial.**
- Variable comercial dependiente: **potencial_cliente**.

El resultado es un dataset homogéneo y estructurado, 134.756 registros con y 52 columnas provenientes de la SCVS, enriquecidos con los atributos ambientales y la información interna de la empresa gestora ambiental. Este dataset constituye el insumo final para el análisis exploratorio de datos (EDA) y la etapa posterior de modelado predictivo.

3.6.10 Consideraciones metodológicas

Vale la pena mencionar que para el proceso de integración se evitó incluir variables temporales (como fechas de entrega o contabilización) por no existir equivalentes en la base pública. Del mismo modo, las variables transaccionales de peso, cantidad o precio unitario fueron excluidas del dataset final, al tratarse de métricas operativas sin correspondencia estructural en el universo externo.

Como resultado la integración prioriza la coherencia estructural y la comparabilidad de las unidades de análisis, permitiendo disponer de un dataset consistente entre fuentes internas y externas, apto para análisis estadístico, exploratorio y predictivo.

3.7 Exploración y análisis de datos

El dataset a analizar llamado **df_merged_completo** contiene 134.754 empresas y 52 variables, correspondientes al universo de las 4 clases de industrias en el mercado ecuatoriano según tipología de la SCVS. Esta base enriquecida con variables ambientales y un objetivo binario llamada **potencial_cliente**.

Durante la limpieza se revisa si contiene valores nulos o duplicados por ruc para garantizar la integridad del conjunto de análisis.

3.7.1 Participación de los clientes actuales respecto al total de empresas en el mercado industrial ecuatoriano

El análisis de la variable **potencial_cliente** muestra que únicamente el 0,8 % de las empresas del dataset (1.123 registros) corresponde a clientes activos, mientras que el 99,2 % restante (133.631 registros) representa empresas no clientes.

El desequilibrio de clases también confirma la necesidad de emplear técnicas de balanceo o muestreo estratificado en etapas posteriores del modelado para evitar sesgos hacia la clase mayoritaria además de justificar la aplicación de

un modelo predictivo para priorizar prospectos con mayor probabilidad de conversión.

3.7.2 Análisis de la distribución empresarial por tamaño y participación del segmento "GRANDE"

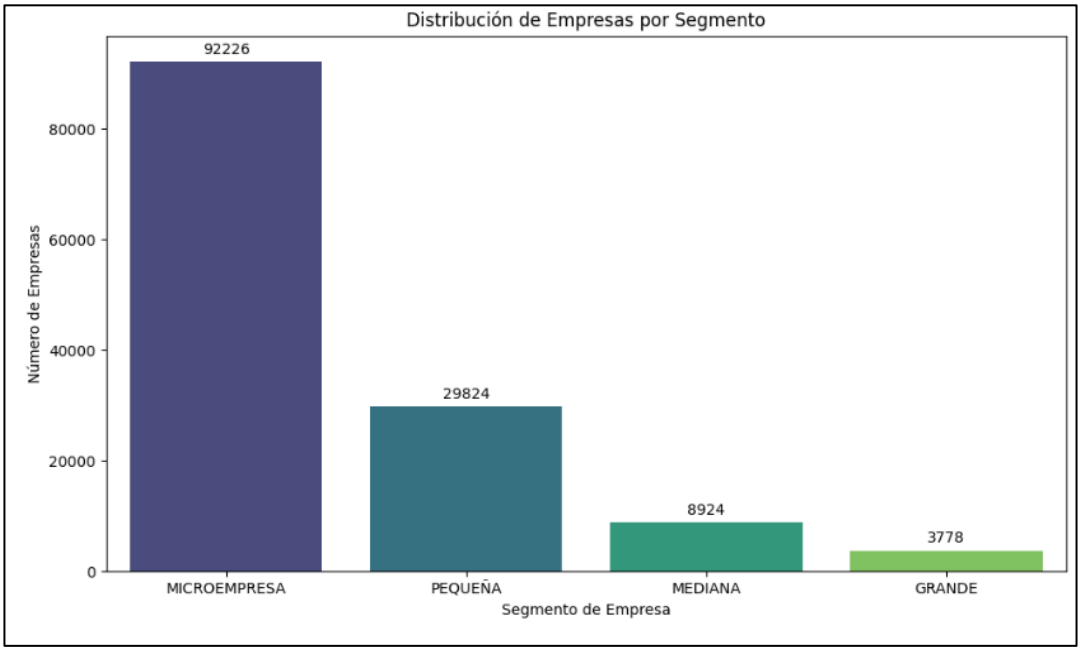

Figura 3.2 Distribución de empresas por segmento

La clasificación por segmentos empresariales evidencia una fuerte concentración de microempresas (68 % del total), seguidas por pequeñas (22 %) y medianas (7 %). El segmento de grandes empresas agrupa 3.778 compañías (3 %), las cuales constituyen el segmento prioritario de este estudio.

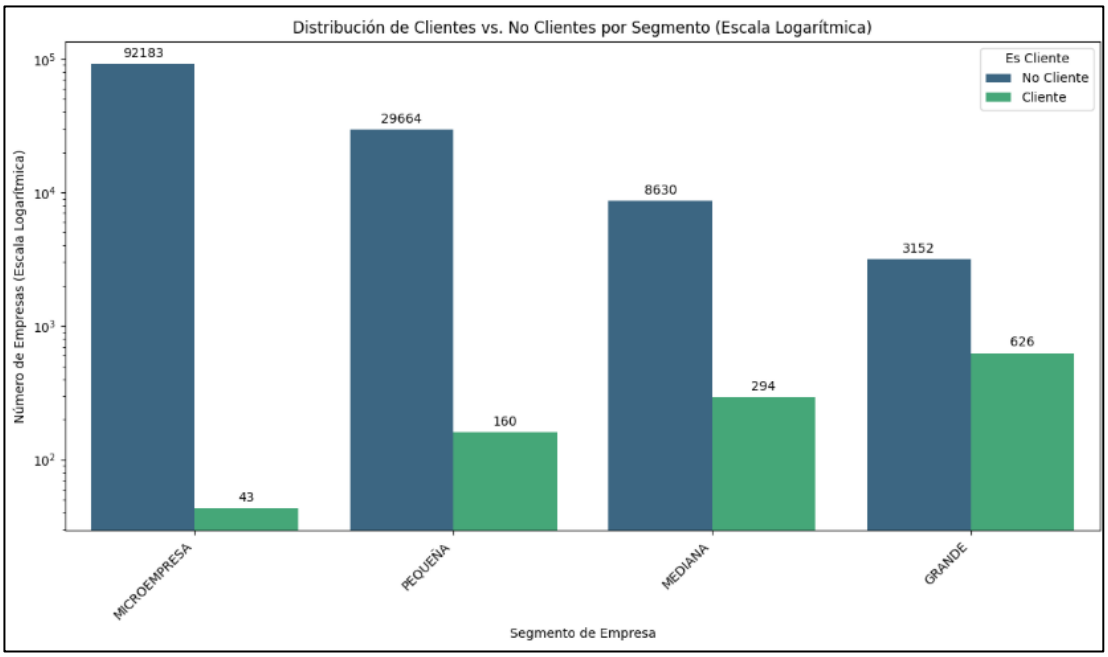

Figura 3.3 Distribución de clientes vs. no clientes por segmento

Dentro de este segmento, la gestora ambiental mantiene 626 clientes activos (19 % del total de grandes empresas), que a su vez representan aproximadamente el 55 % de toda la cartera corporativa actual. Estos resultados confirman que la base comercial de la empresa gestora ambiental está fuertemente concentrada en industrias de gran escala, coherente con su estrategia de negocio B2B.

3.7.3 Distribución geográfica de la actividad empresarial y nivel de cobertura de la cartera de clientes

La distribución territorial muestra una concentración empresarial muy marcada en Guayas (51.139 empresas), Pichincha (42.559) y Azuay (7.515). En conjunto, estas tres provincias reúnen más del 80 % de las empresas registradas en el país, lo que coincide con los principales polos industriales y logísticos nacionales descritos en el Capítulo 2.

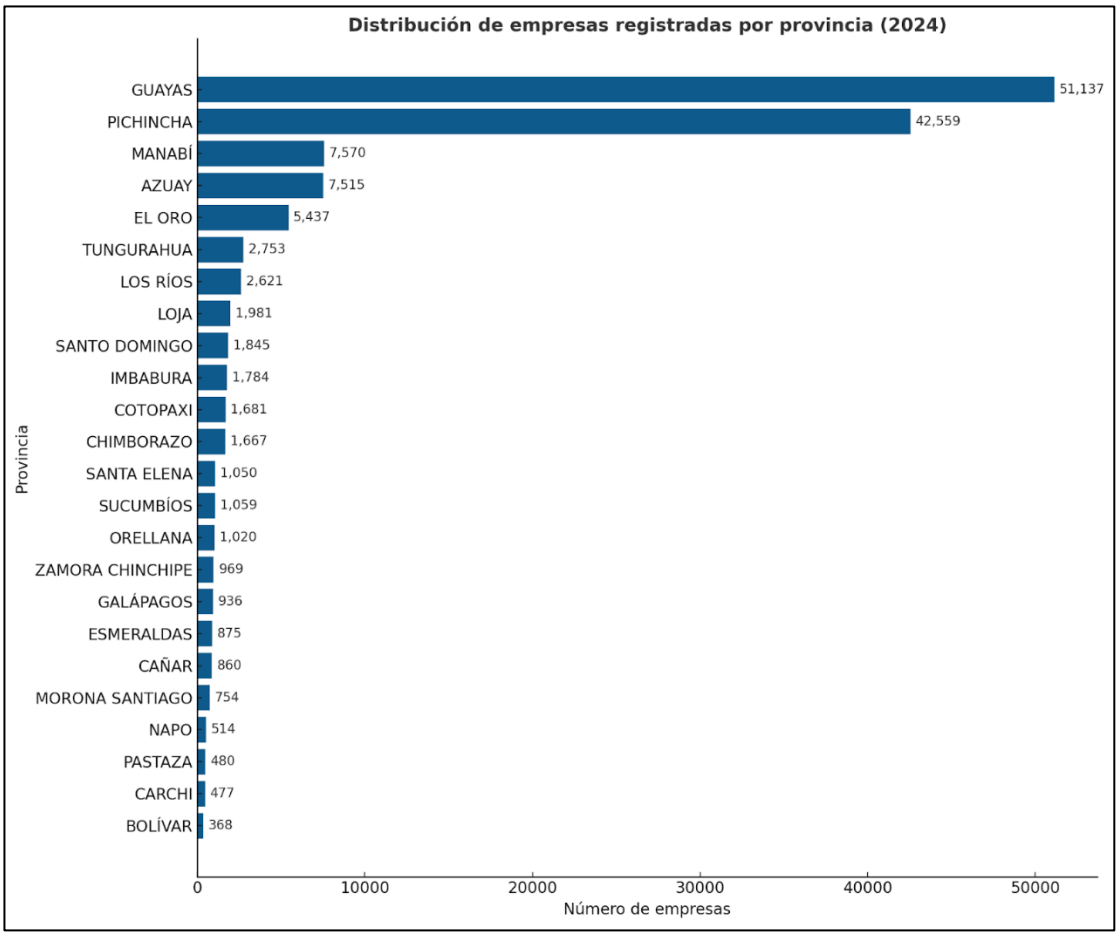

Figura 3.4 Distribución de empresas por provincia

Al analizar la proporción de clientes por provincia, se observa que Guayas lidera con un 1,24 % de penetración (632 clientes), seguida por Azuay (0,92 %) y El Oro (0,86 %). En conjunto, estas tres provincias concentran más del 50 % de la cartera activa, evidenciando una fuerte orientación comercial hacia la región Costa.

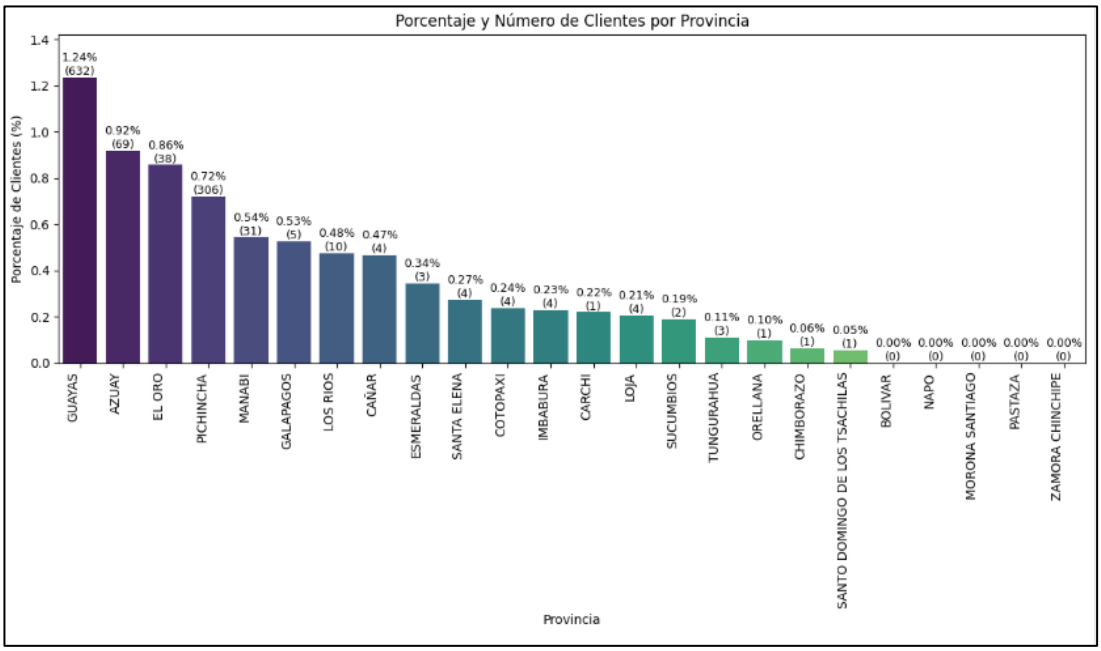

Figura 3.5 Porcentaje y número de clientes por provincia

Desde una perspectiva cuantitativa, Guayas lidera en número absoluto con 345 grandes clientes, seguida de Pichincha (197) y Azuay (31). Esta dualidad entre penetración relativa y volumen absoluto permite identificar provincias consolidadas (Guayas, Pichincha, Azuay) frente a zonas de oportunidad emergente (Manabí, Los Ríos, Cotopaxi) donde la base industrial es relevante pero la presencia comercial aún incipiente.

3.7.4 Concentración sectorial de las empresas grandes y cobertura actual

El análisis sectorial basado en el código CIIU Rev. 4 A.C. muestra una alta concentración en las divisiones G (Comercio al por mayor y menor) y C (Industrias manufactureras), especialmente en Guayas y Pichincha. En Guayas destacan 121 clientes en el sector G y 117 en C, mientras que en Pichincha se registran 77 y 58 respectivamente.

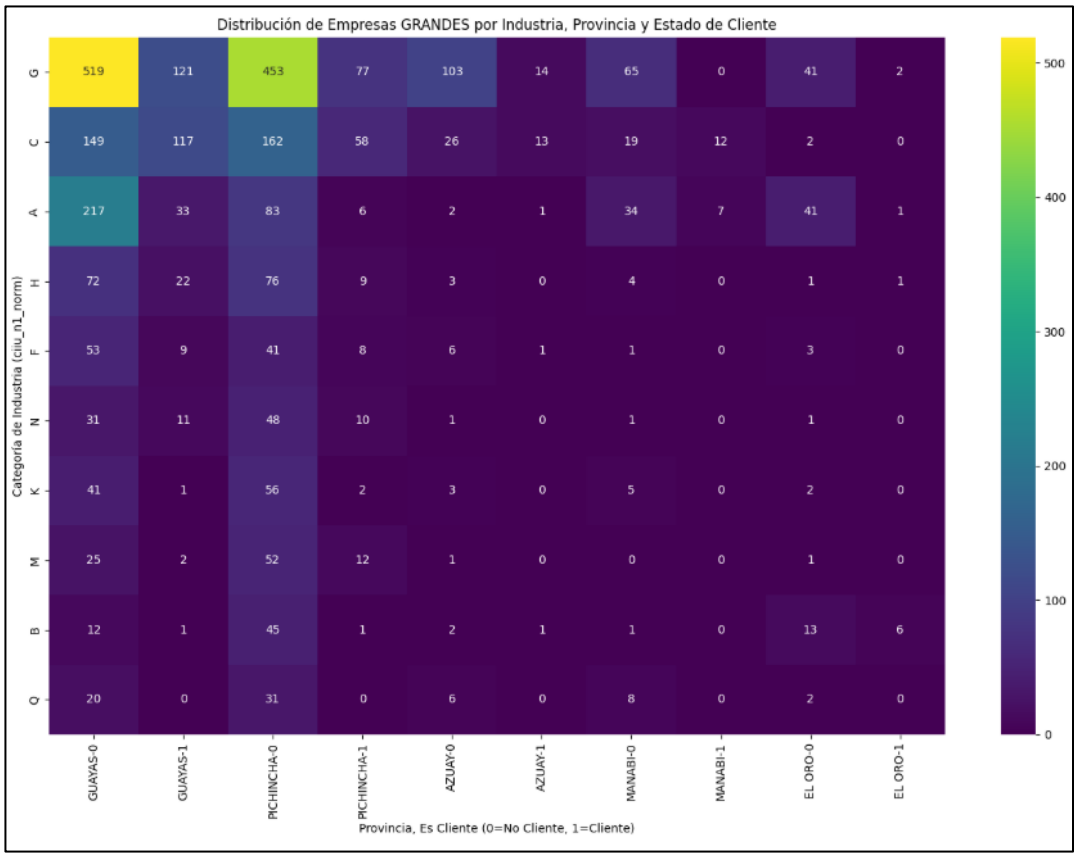

Figura 3.6 Distribución de empresas GRANDES por industria, provincia y estado de cliente

Estas cifras confirman que los principales polos industriales y comerciales del país constituyen el núcleo de la cartera activa.

Sin embargo, los mismos sectores presentan un alto número de empresas no clientes 519 en G y 149 en C solo en Guayas, lo que evidencia brechas significativas de captación. Este contraste indica que, aunque existe una base sólida en sectores industriales clave, el potencial de expansión sigue siendo elevado en segmentos manufactureros, de comercio y logística.

3.7.5 Análisis de la relación entre las características económicas y la condición de cliente/no cliente

Para evaluar las diferencias financieras entre empresas clientes y no clientes, se analizaron las variables **ingresos_ventas**, **activos_totales**, **patrimonio** y **utilidad_neta** mediante gráficos boxplot en escala logarítmica.

Esta transformación permitió reducir la dispersión causada por valores extremos y facilitar la comparación visual de distribuciones. Los resultados muestran que las empresas clientes exhiben, en promedio, valores significativamente superiores en todas las variables financieras analizadas.

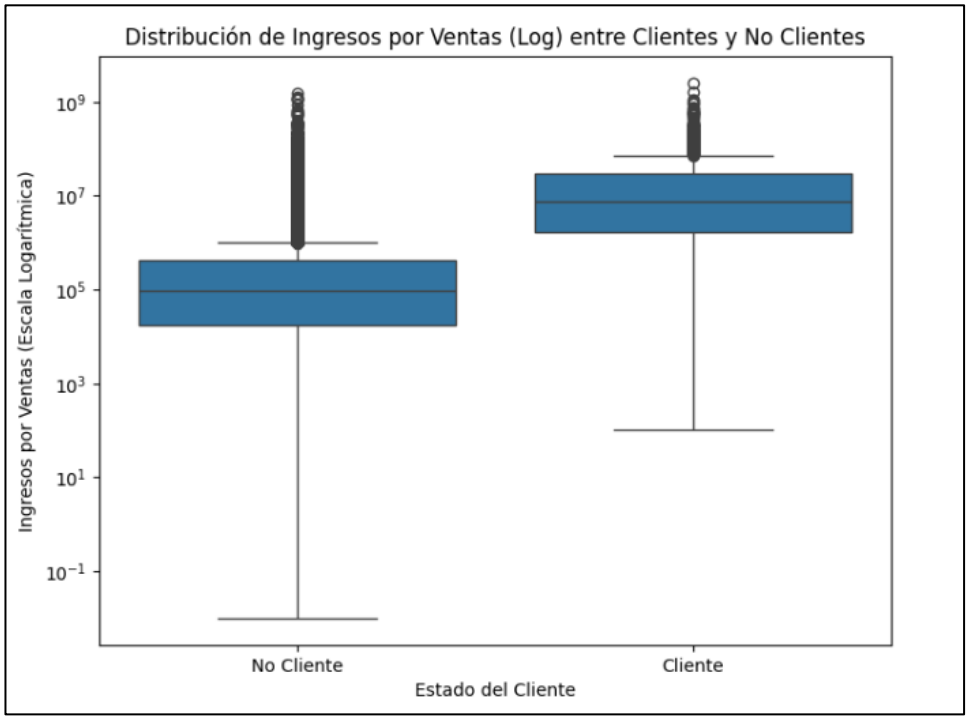

Figura 3.7 Distribución de ingresos por venta entre cliente y no cliente

En el caso de ingresos por ventas, la mediana y el rango intercuartílico se ubican a niveles más altos, lo que indica mayor capacidad de generación de ingresos y diversificación económica.

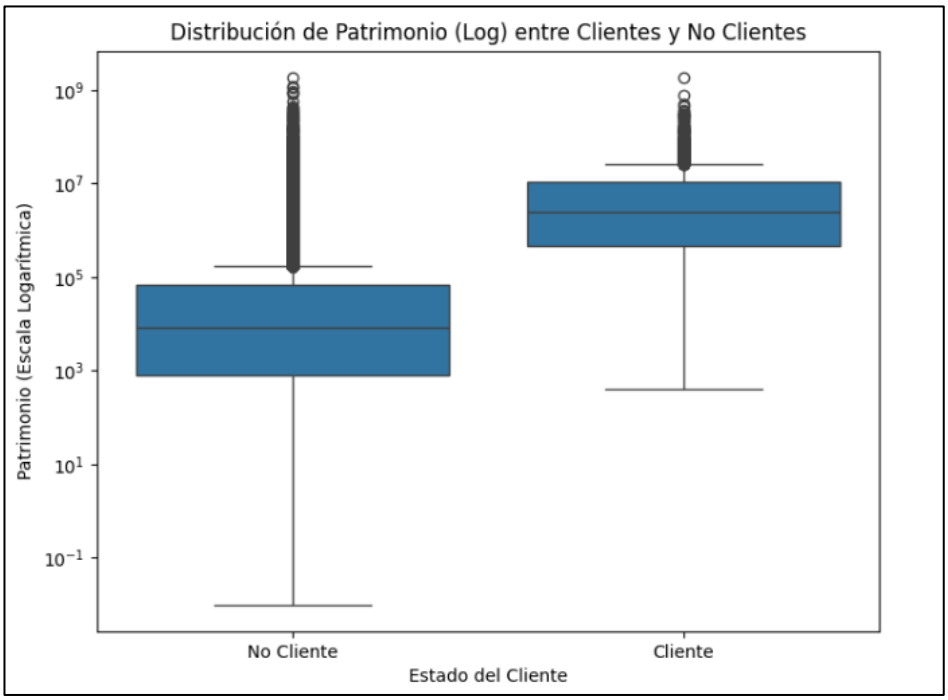

Figura 3.8 Distribución de patrimonio entre cliente y no cliente

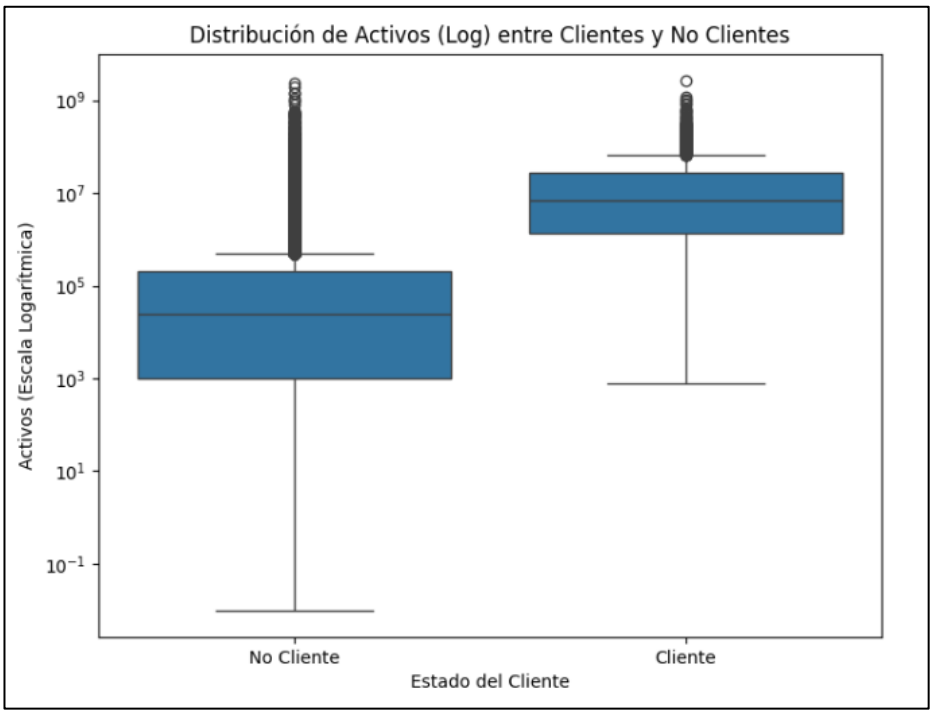

Figura 3.9 Distribución de activos entre cliente y no cliente

Patrimonio y activos totales reflejan igualmente una mayor solidez estructural entre los clientes, con distribuciones amplias que sugieren heterogeneidad

dentro del grupo atendido, pero siempre por encima del nivel medio del mercado.

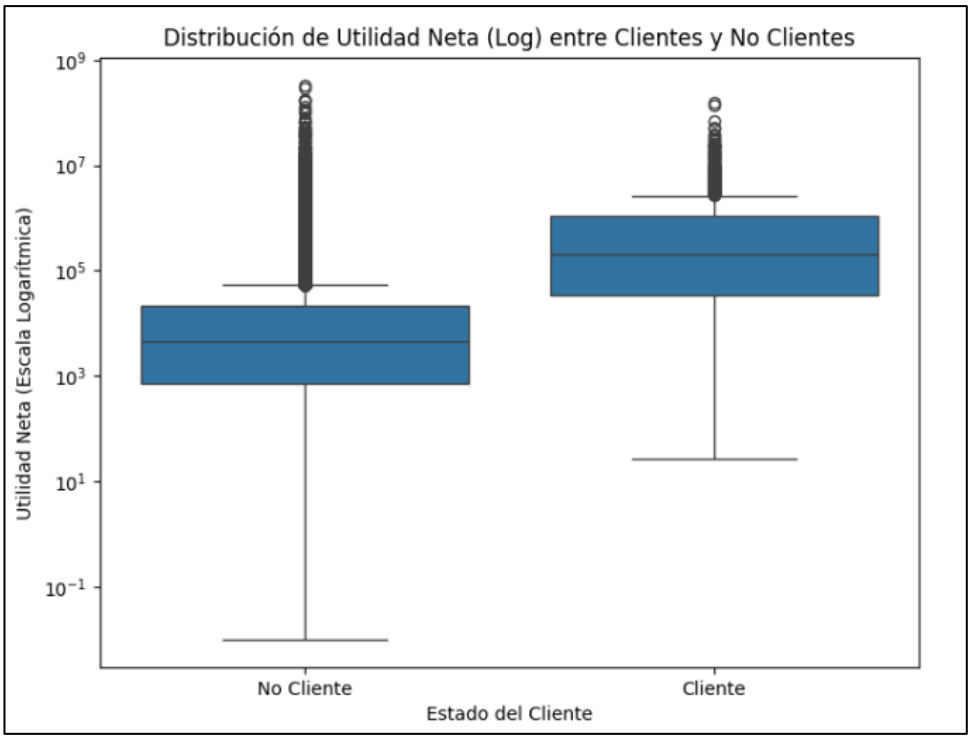

Figura 3.10 Distribución de utilidad neta entre cliente y no cliente

Las utilidades, aunque más variables, siguen el mismo patrón, mostrando mayor rentabilidad en las empresas que forman parte de la cartera de la gestora ambiental.

Estos resultados respaldan la hipótesis de que la gestora ambiental mantiene relaciones con compañías de alto rendimiento financiero, lo que alinea su posicionamiento con clientes de gran escala y mayor estabilidad económica.

3.7.6 Presencia de atributos ambientales en el universo empresarial analizado

Basado en el análisis de este dataset, se observa que dentro del segmento de empresas "GRANDES", solo una fracción limitada presenta un interés ambiental explícito (identificado por la variable **interes_ambiental** = 1). Específicamente, 120 de las 3778 empresas grandes en el dataset muestran

esta característica, lo que representa aproximadamente el 3.18% del total de empresas grandes. A pesar de ser un grupo pequeño, estas compañías grandes con interés ambiental explícito muestran una probabilidad significativamente mayor de ser clientes en comparación con aquellas empresas grandes que no presentan esta acreditación o reconocimiento ambiental. Un análisis comparativo revela que el 30.83% de las empresas grandes con interés ambiental son clientes, mientras que solo el 16.10% de las empresas grandes sin interés ambiental son clientes. Esta diferencia sugiere que el interés ambiental explícito, tal como se captura en este dataset, es un factor que está asociado con una mayor propensión a ser cliente dentro del segmento de grandes empresas.

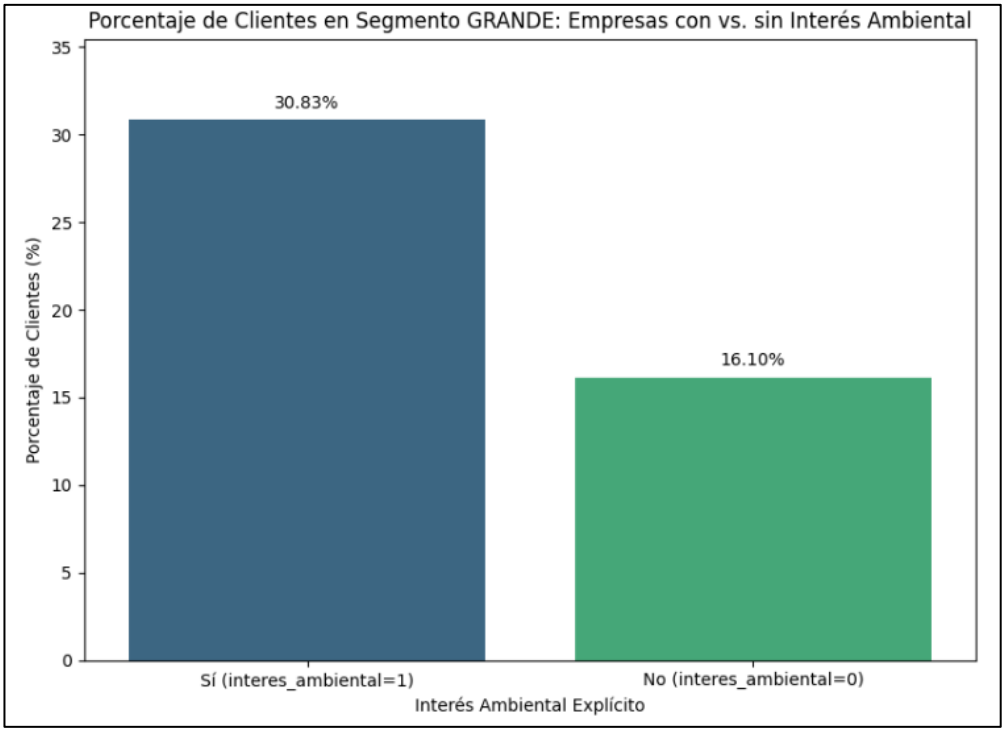

Figura 3.11 Porcentaje de grandes empresas con interés ambiental

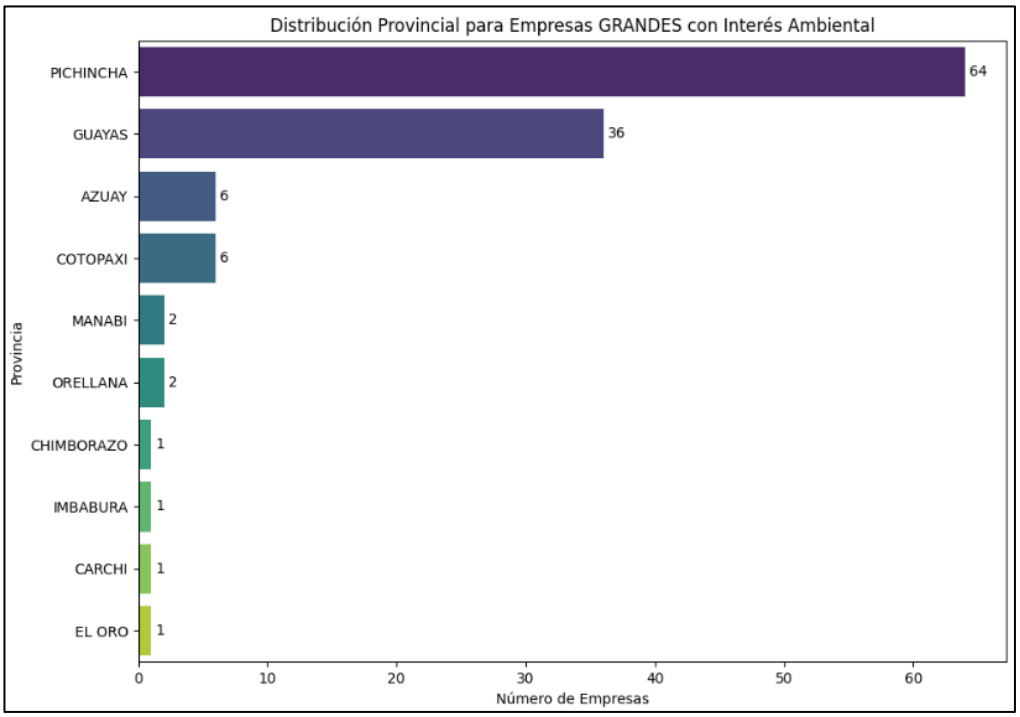

Figura 3.12 Distribución de grandes empresas por provincias con interés ambiental

Tabla 3.1 Clasificación de CIIU N1

Código	Descripción	Nivel
A	Agricultura, ganadería, silvicultura y pesca.	1
B	Explotación de minas y canteras.	1
C	Industrias Manufactureras.	1
D	Suministro De electricidad, gas, vapor y aire acondicionado.	1
E	Distribución de agua; alcantarillado, gestión de desechos y actividades de saneamiento.	1
F	Construcción.	1
G	Comercio Al por mayor y al por menor; reparación de vehículos automotores y Motocicletas.	1
H	Transporte y almacenamiento.	1
I	Actividades de alojamiento y de servicio de comidas.	1
J	Información y comunicación.	1
K	Actividades financieras y de seguros.	1
L	Actividades inmobiliarias.	1
M	Actividades profesionales, científicas y técnicas.	1
N	Actividades de servicios administrativos y de apoyo.	1
O	Administración pública y defensa; planes de seguridad social de afiliación obligatoria.	1
P	Enseñanza.	1

Q	Actividades de atención de la salud humana y de asistencia social.	1
R	Artes, Entretenimiento y recreación.	1
S	Otras Actividades de servicios.	1
T	Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio.	1
U	Actividades de organizaciones y órganos extraterritoriales.	1
S	Otras Actividades de servicios.	1
T	Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio.	1
U	Actividades de organizaciones y órganos extraterritoriales.	1

Las 120 empresas grandes con interés ambiental explícito tienden a concentrarse más en las industrias de Manufactura y Agricultura/Pesca (en relación con el segmento GRANDE total), se ubican principalmente en PICHINCHA y GUAYAS, y son financieramente muy robustas, incluyendo algunas de las empresas más grandes del dataset. Su perfil financiero, basado en estas estadísticas resumidas, parece ser incluso un poco más alto que el del cliente GRANDE promedio.

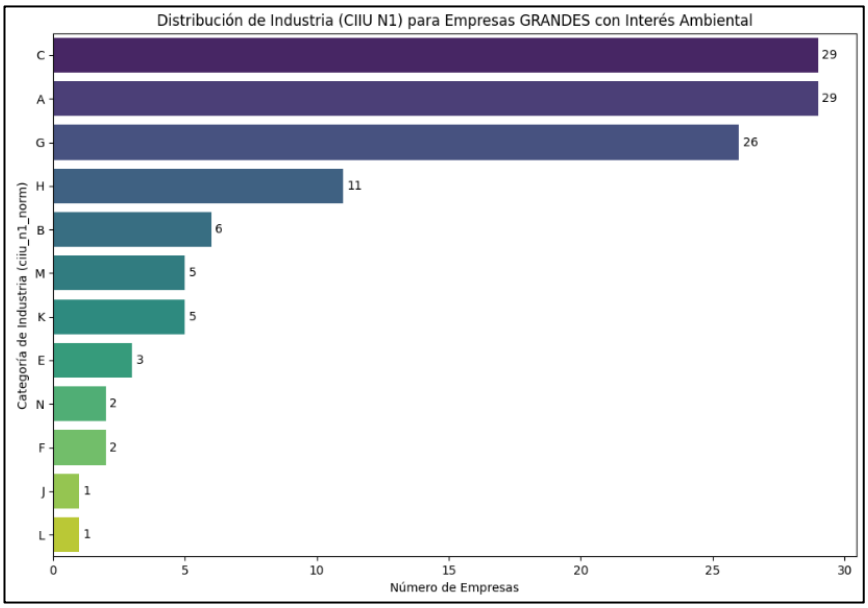

Figura 3.13 Distribución de grandes empresas por provincias con interés ambiental

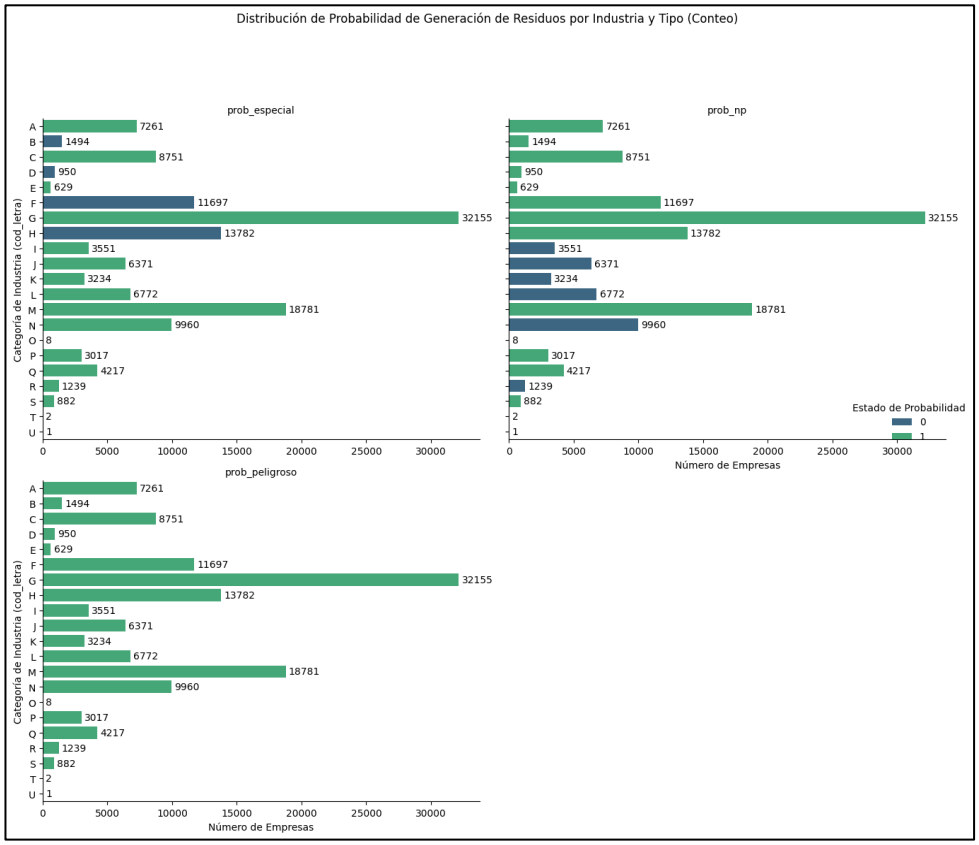

Figura 3.14 Distribución de probabilidad de generación de residuos por industria y tipo

Este gráfico de conteos confirma que la Industria C (Industrias Manufactureras) es, por mucho, el sector con el mayor volumen absoluto de empresas asociadas a la probabilidad de generar los tres tipos de residuos (**prob_peligroso, prob_np, prob_especial**) en este dataset. Las industrias A y G también son significativas en términos de volumen para residuos no peligrosos y especiales.

Esto te da una idea clara de dónde encontrar la mayor cantidad de empresas que probablemente generen estos tipos de residuos, basándose en la clasificación CIIU. Para entender la intensidad o proporción dentro de cada industria (es decir, qué tan común es tener probabilidad 1 en una industria específica, independientemente de cuántas empresas totales tenga esa industria), sería más útil ver los gráficos de proporciones que generamos anteriormente.

3.7.7 Síntesis del análisis exploratorio

- El análisis exploratorio revela un mercado altamente concentrado y subatendido, con una penetración comercial menor al 1 % en el total de empresas del país, pero con fuerte presencia en el segmento de grandes industrias.

- Territorialmente, la cobertura se concentra en Guayas, Azuay y El Oro, mientras que Pichincha y la Sierra Centro constituyen zonas de oportunidad.

- Sectorialmente, los sectores manufactureros y comerciales dominan la cartera, aunque mantienen amplias brechas de captación.

3.8 Plataformas y Prototipos de Visualización

La fase de plataforma y visualización corresponde a la materialización de los resultados del modelo de lead scoring mediante el desarrollo de un prototipo

funcional de análisis y predicción empresarial. Este prototipo integra un flujo completo de procesamiento, predicción y visualización de datos, permitiendo a los usuarios consultar la probabilidad de conversión de empresas industriales hacia proyectos de gestión ambiental.

3.8.1 Entrenamiento y Selección del Modelo

Para la fase de entrenamiento del modelo predictivo, se empleó una metodología que prioriza la velocidad de experimentación, la gestión automatizada del preprocesamiento y la evaluación simultánea de múltiples algoritmos de clasificación bajo una configuración estandarizada.

Se definió los parámetros fundamentales del proceso de entrenamiento, donde se estableció una división del 80% para entrenamiento y 20% para validación, aplicando una estrategia de validación cruzada estratificada con 5 pliegues (stratified k-fold) para garantizar representatividad de clases en cada subconjunto.
Asimismo, se implementaron técnicas automáticas para eliminar colinealidad entre variables (threshold = 0.95), realizar selección de características relevantes y balancear la variable objetivo en caso de desbalanceo.

El modelo seleccionado finalmente fue desplegado dentro del prototipo descrito a continuación.

3.8.2 Flujo de Datos y Arquitectura Técnica

El flujo de datos implementado sigue una arquitectura modular, diseñada para garantizar eficiencia, escalabilidad y respuesta en tiempo real:

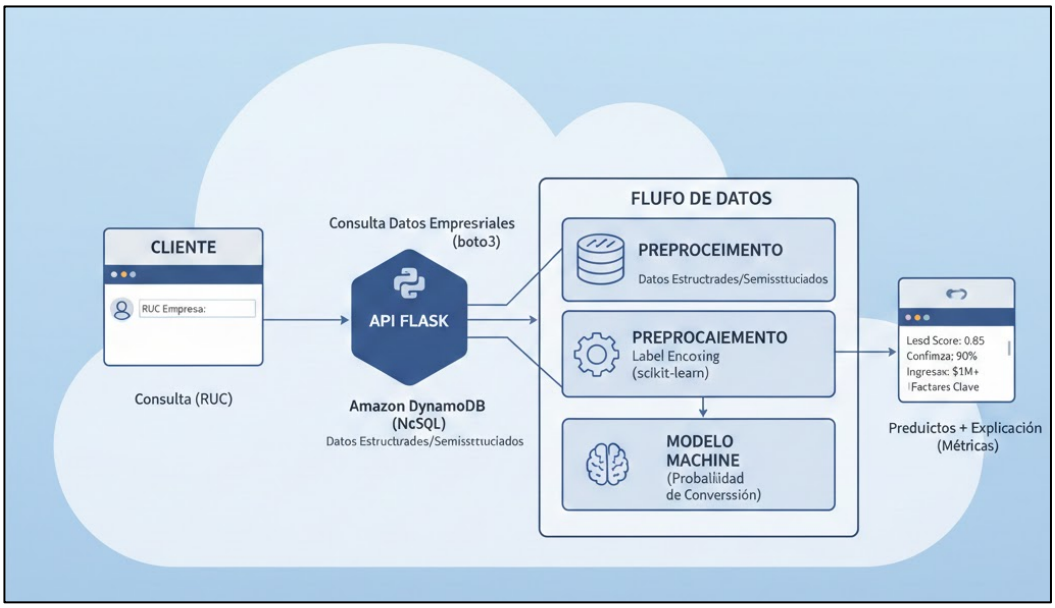

Figura 3.15 Arquitectura de la solución

- **Cliente→ API Flask** El usuario realiza una consulta desde la interfaz web ingresando el RUC de una empresa. Esta solicitud es enviada a la API desarrollada con Flask, la cual actúa como intermediario entre el cliente y el modelo de predicción.
- **Flask → DynamoDB (boto3)** La API consulta los datos empresariales almacenados en Amazon DynamoDB a través del SDK boto3. Esta base de datos NoSQL permite almacenar información estructurada y semiestructurada de manera eficiente, soportando altas tasas de lectura y escritura.
- **Preprocesamiento → Label Encoding** Los datos recuperados son transformados mediante técnicas de preprocesamiento, aplicando codificación categórica (Label Encoding) previamente definida en el entrenamiento con la biblioteca scikit-learn, garantizando la coherencia con las variables utilizadas y evitando un posible data drift.
- **Modelo de Machine Learning → Predicción** El modelo de lead scoring, previamente entrenado, recibe los datos procesados y devuelve una predicción de conversión (probabilidad de que la empresa contrate un

servicio ambiental). El modelo se ejecuta dentro del entorno Flask, asegurando baja latencia y escalabilidad en la nube.

- **Resultado → Cliente (JSON)** Finalmente, los resultados son enviados al cliente en formato JSON, donde son renderizados dinámicamente en la interfaz web, junto con los indicadores financieros, métricas de confianza y explicación del modelo.

3.8.3 Tecnologías de Desarrollo Frontend

La interfaz de usuario fue implementada utilizando React 18, con la herramienta de compilación y desarrollo Vite (v7.1.12). Esta combinación permite una experiencia fluida, con tiempos de carga mínimos y componentes reutilizables que aseguran una alta mantenibilidad del sistema. El diseño prioriza la claridad visual y la interpretación inmediata de resultados, empleando una paleta de colores suaves y jerarquía tipográfica que facilita la lectura de métricas y recomendaciones.

Las siguientes figuras muestran las principales vistas del prototipo:

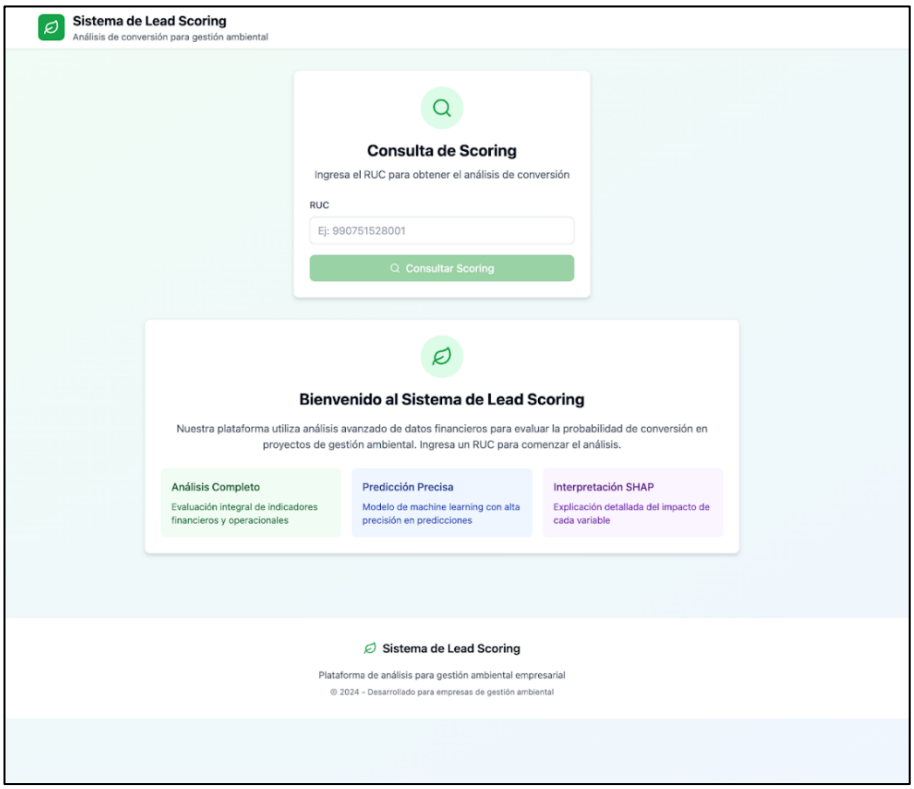

Figura 3.16 Página de consulta inicial del Sistema de Lead Scoring

La figura 3.16 muestra cómo se habilita al usuario a ingresar el RUC para obtener la evaluación de la probabilidad de conversión.

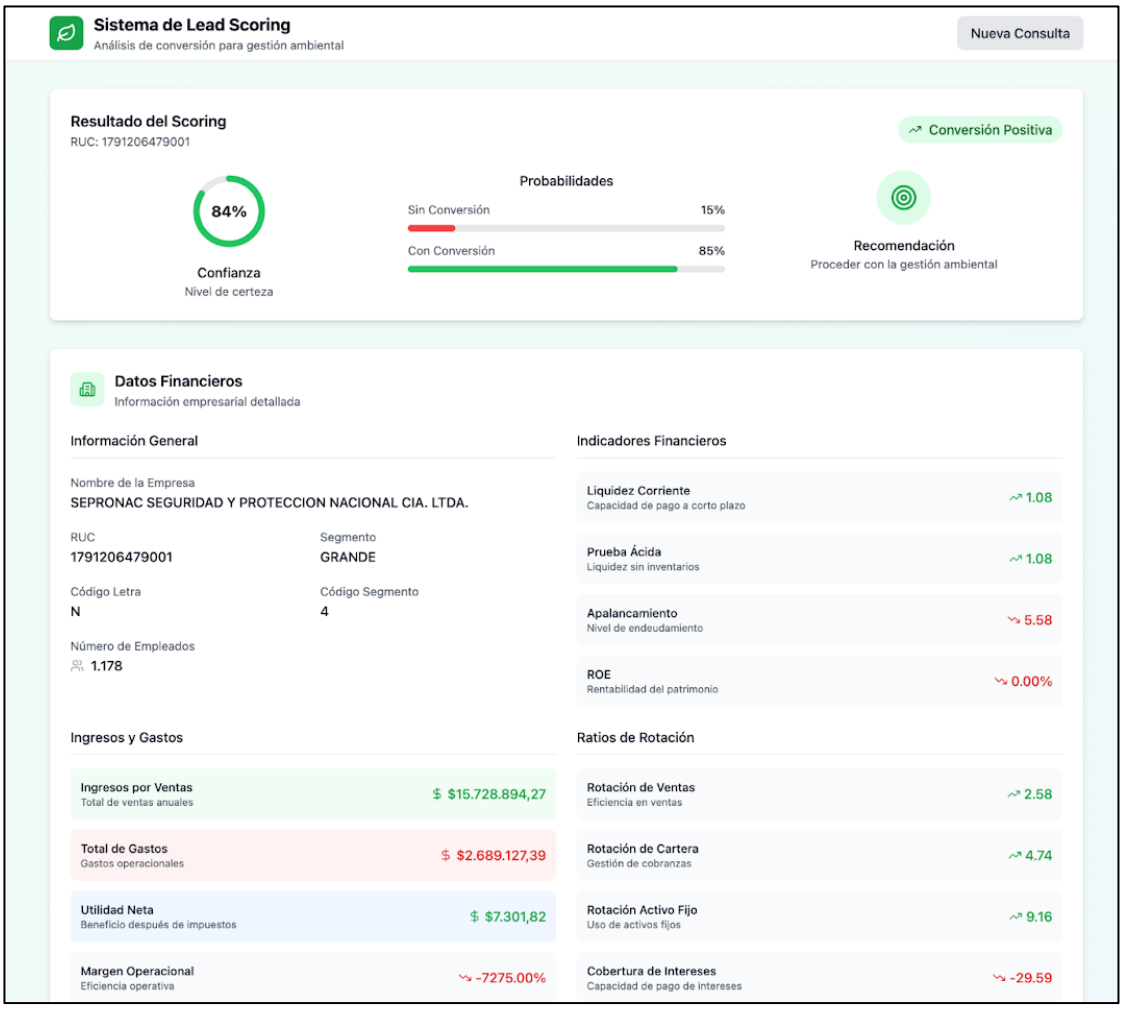

Figura 3.17 Panel principal con resultados de scoring financiero y recomendación de conversión

La Figura 3.17 presenta el nivel de confianza, los indicadores financieros principales y la recomendación automática para continuar con el proceso de gestión ambiental.

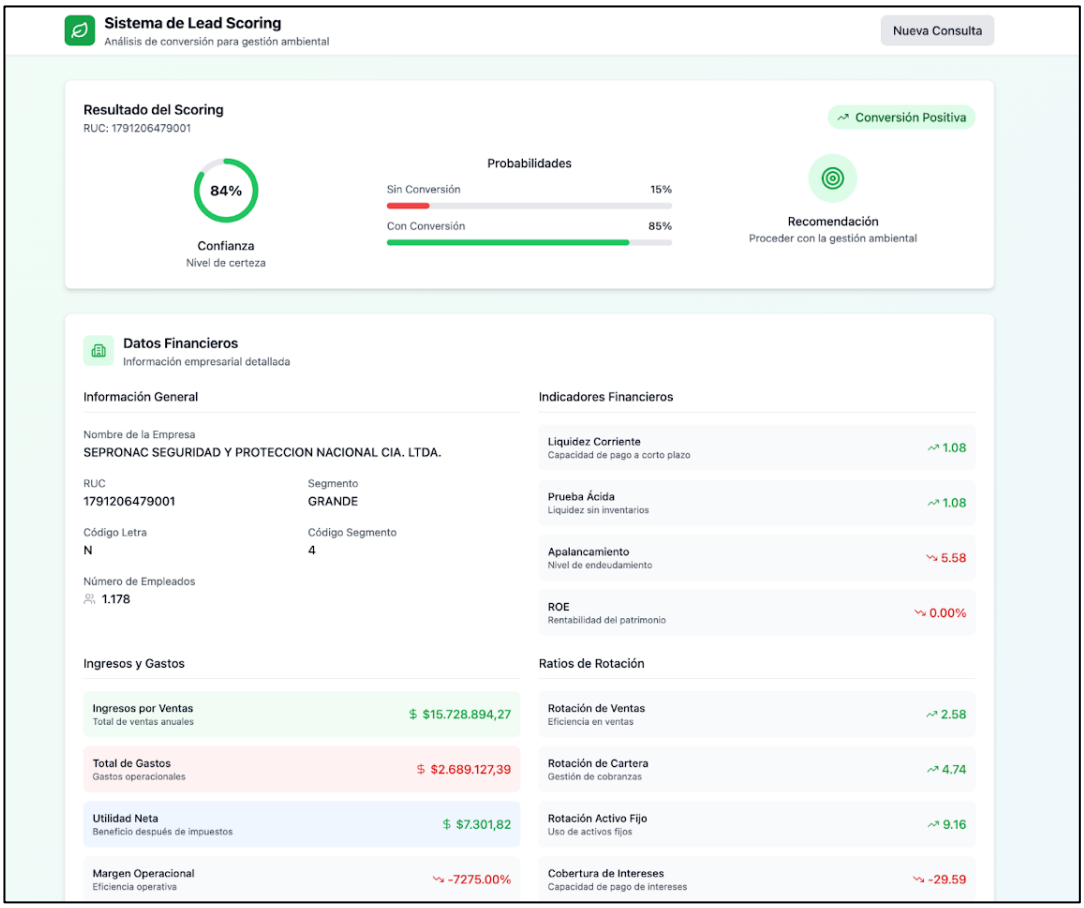

Figura 3.18 Panel principal con resultados de scoring financiero y recomendación de conversión

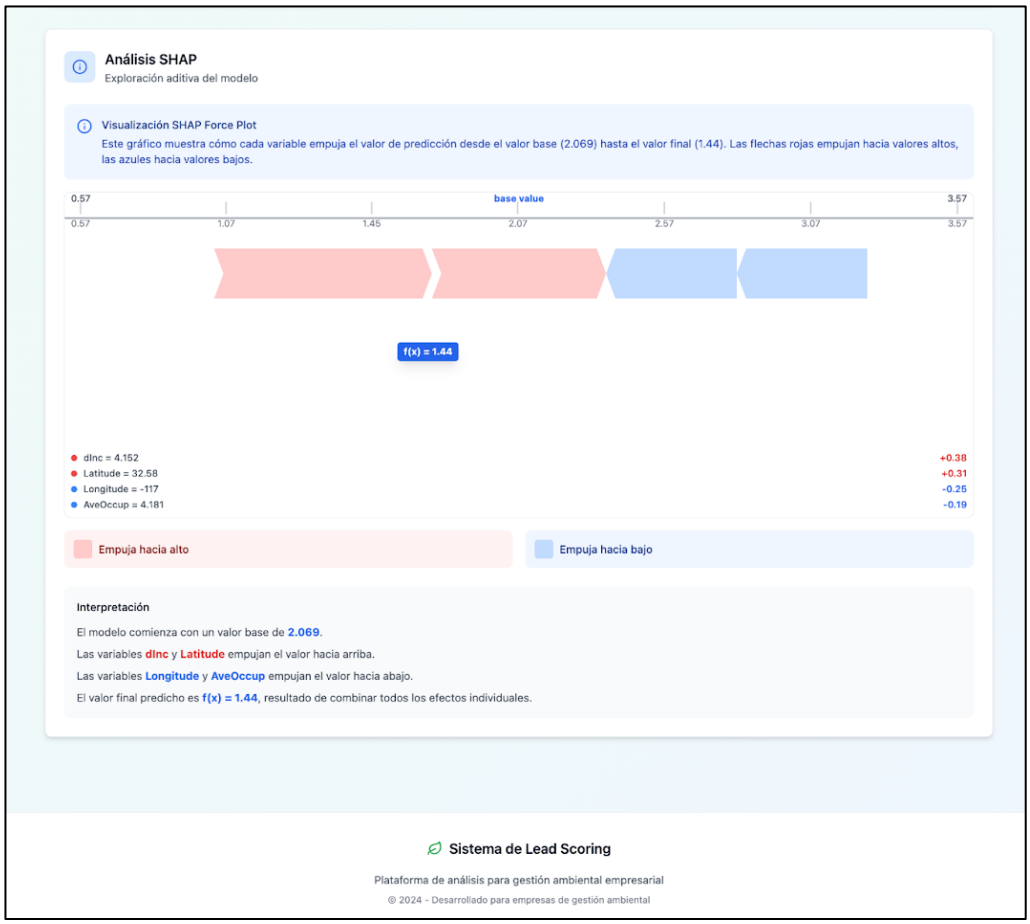

Figura 3.19 Visualización SHAP Force Plot — Interpretabilidad del modelo de machine learning

La Figura 19 presenta el impacto individual de las variables en la predicción final, evidenciando la transparencia del modelo y su capacidad explicativa.

3.8.4 Funcionalidades del Prototipo

El sistema integra las siguientes capacidades analíticas y operativas:

- **Predicción de Conversión:** cálculo automático de la probabilidad de conversión basada en indicadores financieros, geográficos y operativos de cada empresa.
- **Interpretabilidad del Modelo:** visualización de resultados mediante gráficos SHAP, explicando el efecto de cada variable sobre la predicción.

- **Análisis Financiero Integrado:** extracción y presentación de métricas como liquidez corriente, apalancamiento, rentabilidad (ROE) y rotaciones operativas.
- **Recomendación Automatizada:** generación de una recomendación final (Conversión Positiva / Negativa) para orientar decisiones de los equipos de marketing y ventas.

3.9 Métricas de negocio

Las métricas de negocio cuantifican el impacto estratégico del modelo predictivo sobre la eficiencia comercial y la rentabilidad de la empresa gestora ambiental. Su objetivo es vincular los resultados técnicos del modelo con indicadores financieros tangibles que evidencien el aporte del modelo al retorno sobre la inversión (ROI), la generación de ingresos y la eficiencia operativa.

3.9.1 Fundamentación de las métricas

Para la evaluación económica se emplearon tres indicadores financieros ampliamente utilizados en análisis de proyectos de inversión:

- **Retorno sobre la inversión (ROI)**: mide la rentabilidad generada por el proyecto respecto a la inversión inicial.
- **Valor actual neto (VAN)**: determina el valor presente de los beneficios futuros descontados a una tasa determinada.
- **Periodo de recuperación (Payback)**: estima el tiempo requerido para recuperar la inversión inicial.

Estos indicadores se basan en los flujos de caja generados por el incremento de ingresos y los ahorros operativos derivados de la implementación del modelo.

3.9.2 Ingresos proyectados

El modelo predictivo permite aumentar la eficiencia comercial mediante la priorización basada en datos de los leads con mayor probabilidad de conversión. Durante 2024, la empresa gestora ambiental registró 301 leads generados a través de campañas digitales, con una tasa de conversión del 51,2 %, lo que dio como resultado 154 contratos. Con un ticket promedio de USD 2.275, los ingresos actuales ascienden a USD 350.350.

Con la implementación del modelo se proyecta un incremento de 10 puntos porcentuales en la tasa de conversión (61,2 %), generando 30 contratos adicionales, sin aumentar la inversión publicitaria. Esto representa un incremento de USD 68.260 anuales, como se muestra en la Tabla 3.2.

Tabla 3.2 Ingresos actuales y proyectados del modelo

Concepto	Valor	Fuente/Calculo
Leads generados (base 2024)	301	Datos de campañas reales
Tasa de conversión 2024	51,2 %	Leads → contratos
Contratos actuales	301 × 51,2 % = 154	Valor de referencia
Ticket promedio	USD 2.275	Promedio por contrato
Ingresos actuales	154 × 2.275 = USD 350.350	Base actual
Tasa de conversión con modelo	61,2 %	+10 puntos porcentuales
Contratos proyectados	301 × 61,2 % = 184	30 adicionales
Ingresos proyectados	184 × 2.275 = USD 418.610	Con modelo
Incremento en ingresos	418.610 – 350.350 = USD 68.260	Beneficio directo

3.9.3 Costos del proyecto

El costo total anual del proyecto corresponde al salario del ingeniero en sistemas o datos responsable del mantenimiento y supervisión del modelo. Dado que la actualización se realiza con recursos internos, no se incluyen costos adicionales de operación o mantenimiento.

La inversión anual asciende a USD 15.600, como se presenta en la Tabla 3.3.

Tabla 3.3 Costo anual del proyecto

Concepto	Costo anual (USD)	Descripción
Sueldo del ingeniero en sistemas o datos	$ 15.600	Salario anual del responsable técnico del modelo
Inversión total	$ 15.600	Costo anual total del proyecto

3.9.4 Ahorro operativo anual

Previo a la implementación del modelo, el equipo comercial y de marketing dedicaba una fracción significativa de su tiempo a la clasificación y priorización manual de leads dentro del CRM. La automatización de esta tarea mediante analítica predictiva permite liberar horas laborales relevantes, reduciendo carga operativa y generando un ahorro equivalente al valor mensual del tiempo liberado.

El ahorro total anual es de USD 21.900, como se detalla en la Tabla 3.4.

Tabla 3.4 Ahorro operativo anual por perfil

Cargo	Sueldo mensual (USD)	% de tiempo ahorrado	Horas liberadas (8h/día)	Valor mensual del tiempo ahorrado	Ahorro anual
Ing. de Marketing Digital	$ 1.250	30%	480	1.250 × 0.30 = 375	375 × 12 = 4.500
Asignador Comercial	$ 1.100	30%	480	1.100 × 0.30 = 330	330 × 12 = 3.960

4 Asesores Industriales	800 × 4 = 3.200	35%	2.688	3.200 × 0.35 = 1.120	1.120 × 12 = 13.440
Total ahorro operativo anual					**$ 21.900**

3.10 Análisis financiero

La evaluación financiera integra los ingresos incrementales y los ahorros operativos, en contraste con la inversión total anual. El flujo neto anual corresponde a la suma de ambos beneficios:

Tabla 3.5 Indicadores financieros

Flujo neto anual = Ingresos adicionales + Ahorro operativo
Flujo neto anual = 68.260 + 21.900 = 90.160

Con base en este flujo neto anual, se calcularon los indicadores financieros.

3.10.1 VAN (5 años, 10 %)

La Figura 23 presenta la proyección del VAN considerando una inversión inicial de USD 15.600 y una tasa de descuento del 10 %. El valor presente acumulado en cinco años asciende a USD 326.283, lo que evidencia una generación significativa de valor.

Tabla 3.6 Proyección del VAN por año

Año	Flujo neto (USD)	Factor (10%)	Valor presente (USD)	VAN acumulado (USD)
0	-15.600	1,000	-15.600	-15.600

1	90.160	0,909	81.964	66.364
2	90.160	0,826	74.970	141.334
3	90.160	0,751	67.779	209.113
4	90.160	0,683	61.080	270.193
5	90.160	0,621	56.090	326.283

3.10.2 ROI acumulado

Tabla 3.7 Fórmula para ROI acumulado

ROI = Beneficio acumulado / Inversión inicial × 100

En el quinto año, los beneficios acumulados alcanzan USD 450.800, lo que representa un ROI acumulado de 2.886 %. Por lo tanto, en un horizonte de cinco años, el proyecto devuelve 28,88 veces la inversión inicial.

Tabla 3.8 ROI acumulado por año

Año	Flujo neto (USD)	Factor (10%)	Valor presente (USD)	VAN acumulado (USD)
0	-15.600	1,000	-15.600	-15.600
1	90.160	0,909	81.964	66.364
2	90.160	0,826	74.970	141.334
3	90.160	0,751	67.779	209.113
4	90.160	0,683	61.080	270.193
5	90.160	0,621	56.090	326.283

3.10.3 Interpretación de resultados

Los resultados evidencian que la implementación del modelo predictivo es altamente rentable y contribuye de manera significativa a la eficiencia operativa. El ROI anual de 73,3 % indica que el modelo genera USD 1,73 por

cada dólar invertido. El VAN positivo superior a USD 300.000 confirma que el proyecto genera valor aún bajo una tasa de descuento del 10 %.

El periodo de recuperación inferior a un año (aproximadamente 7 meses) reafirma su viabilidad financiera. En términos operativos, el ahorro anual de USD 21.900 refleja una mejora sustancial en la eficiencia del equipo comercial, permitiendo concentrar esfuerzos en oportunidades reales de venta.

A nivel estratégico, los resultados demuestran que la analítica predictiva optimiza la inversión existente en marketing, potenciando la generación de ingresos sin incrementar costos. Finalmente, la implementación del modelo fortalece las capacidades analíticas de la empresa gestora ambiental, posicionándola como una organización que incorpora inteligencia artificial aplicada a procesos comerciales y sostenibilidad empresarial.

3.11 Comunicación de resultados y periodicidad

Los resultados del modelo serán comunicados a tres grupos objetivo con diferente nivel de detalle:

Equipo técnico (IT / Inteligencia de la Información) Recibe métricas técnicas detalladas y diagnósticos del modelo.

- Objetivo: evaluar rendimiento, identificar mejoras y garantizar la reproducibilidad.
- Periodicidad: mensual, para reaccionar rápidamente a degradaciones de performance.

Equipos de comercial y marketing Recibe listados clasificados con score individual y categoría (alta, media, baja).

- Objetivo: priorizar esfuerzos en prospección y seguimiento.

- Periodicidad: semanal, asegurando que trabajen con datos actualizados y relevantes.

Gerencia y mandos medios (direcciones, jefaturas) Reciben informes de impacto en métricas de negocio: tasas de conversión, ingresos potenciales captados, ROI.

- Objetivo: evaluar si la estrategia comercial y la priorización están alineadas con metas anuales.

- Periodicidad: trimestral, para permitir análisis de tendencias y decisiones estratégicas.

La combinación de métricas técnicas y de negocio, junto con una estrategia de comunicación segmentada y periódica, garantiza que el modelo de priorización no solo sea técnicamente sólido, sino también un instrumento de gestión estratégica. Su aplicación permitirá maximizar el ROI, incrementar ventas y optimizar recursos comerciales, asegurando que cada decisión de contacto con un prospecto esté respaldada por datos y análisis.

CAPÍTULO 4

4. RESULTADOS Y ANÁLISIS

Este capítulo presenta la evidencia empírica y cuantitativa del proyecto, las pruebas de funcionalidad, y el análisis costo/beneficio (ROI), así como a la validación del modelo para predecir clientes. La segmentación de la base de empresas utilizada en este capítulo se alinea con el Estudio de Mercado Industrial 2025 otorgado por la empresa.

Este apartado detalla el proceso de validación del proyecto, abarcando la rigurosa selección y verificación de las variables predictivas, la validación del modelo para asegurar su capacidad predictiva y generalización, y la presentación del producto digital final. Se explica cómo se abordó la identificación y mitigación de la redundancia de información, la evaluación de la importancia de las variables, y la implementación de técnicas para garantizar la robustez del modelo. Finalmente, se describe el sistema de lead scoring desarrollado, el cual consolida los insights del modelo y facilita su aplicación estratégica en los procesos comerciales y de marketing.

4.1 Validación de variables

La validación se abordó en dos niveles complementarios: validación de variables, para garantizar informatividad, no redundancia y estabilidad; y validación de modelos, para estimar capacidad predictiva y generalización. Este proceso detalla la rigurosa selección y verificación de las variables predictivas, la identificación y mitigación de la redundancia de información, la evaluación de la importancia de las variables, y la implementación de técnicas para garantizar la robustez del modelo. Las variables iniciales consideradas para el modelo fueron:

- **ingresos_ventas**

- **activos**
- **patrimonio**
- **total_gastos**
- **costos_ventas_prod**
- **n_empleados**
- **liquidez_corriente**
- **deuda_total**
- **deuda_total_c_plazo**
- **apalancamiento**
- **apalancamiento_financiero**
- **end_corto_plazo**
- **end_largo_plazo**
- **fortaleza_patrimonial**
- **cobertura_interes**
- **gastos_financieros**
- **impac_carga_finan**
- **utilidad_an_imp**
- **utilidad_neta**
- **margen_bruto**
- **margen_operacional**

- **rent_neta_ventas**
- **roa**
- **roe**
- **rent_ope_activo**
- **rent_ope_patrimonio**
- **rent_neta_activo**
- **rot_ventas**
- **rot_cartera**
- **rot_activo_fijo**
- **per_med_cobranza**
- **per_med_pago**
- **gastos_admin_ventas**
- **depreciaciones**
- **amortizaciones**
- **impuesto_renta**

Si bien muchos de los algoritmos de aprendizaje automático tienen la propiedad de seleccionar variables (shrinkage), se utilizaron varias técnicas para la selección preliminar de variables que serían después utilizadas en los diferentes algoritmos.

4.2 Correlación y redundancia

Se calcularon matrices de correlación de Pearson sobre variables numéricas Se calcularon matrices de correlación de Pearson sobre variables numéricas,

las cuales permitieron determinar el grado y la dirección de la relación lineal entre pares de variables. Este análisis fue fundamental para identificar patrones y dependencias, revelando qué variables tienden a moverse juntas y en qué sentido.

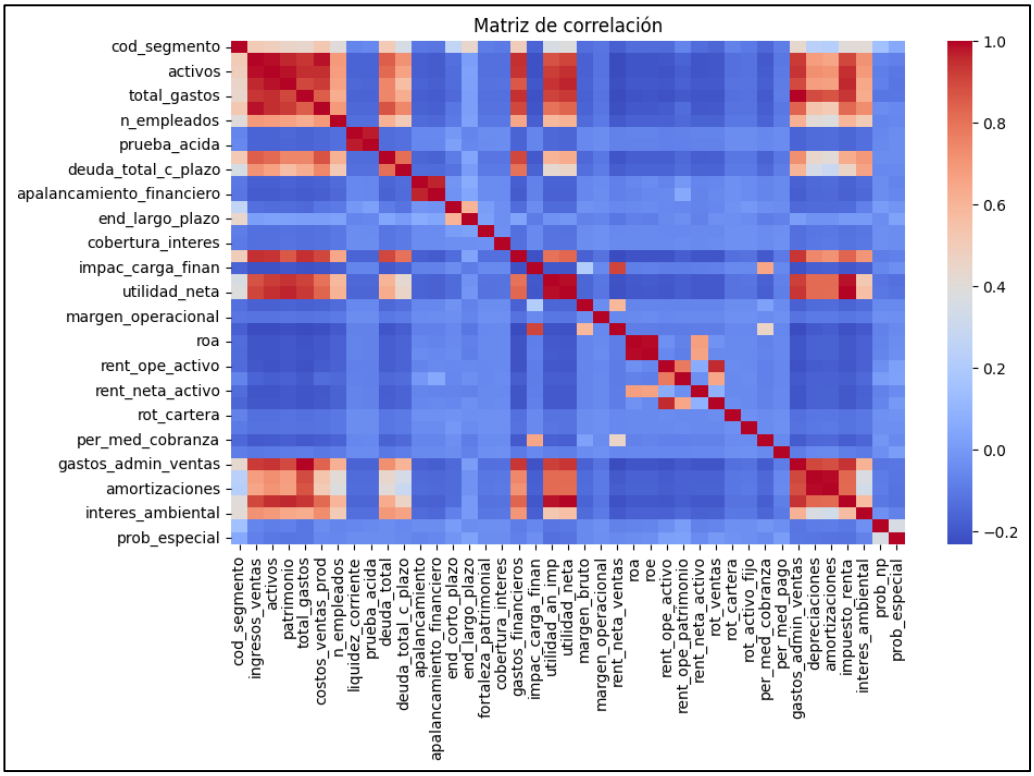

Figura 4.1 Matriz de Correlación (Pearson)

A partir de la Matriz de Correlación, se identificaron las variables con alta intercorrelación, lo que indica un problema de redundancia de información. La decisión de eliminar las variables 'patrimonio', '**costos_ventas_prod**', y '**gastos_admin_ventas**' se tomó debido a la fuerte correlación positiva (tonos rojos intensos en el heatmap) que presentaban con otras variables ya incluidas en el modelo (ej., 'activos' o '**total_gastos**'). Esta remoción inicial mitigó el riesgo de multicolinealidad y simplificó el modelo, siguiendo el principio de parsimonia.

Para refinar aún más el conjunto de features y optimizar la capacidad predictiva, se aplicaron dos métodos secuenciales de selección de variables:

4.3 Métodos Secuenciales de Selección (Random Forest y VIF)

Posteriormente a la reducción inicial por alta correlación, se aplicó un modelo de Random Forest para evaluar la importancia relativa de cada variable. Aquellas variables que demostraron una baja contribución al poder predictivo del modelo fueron eliminadas.

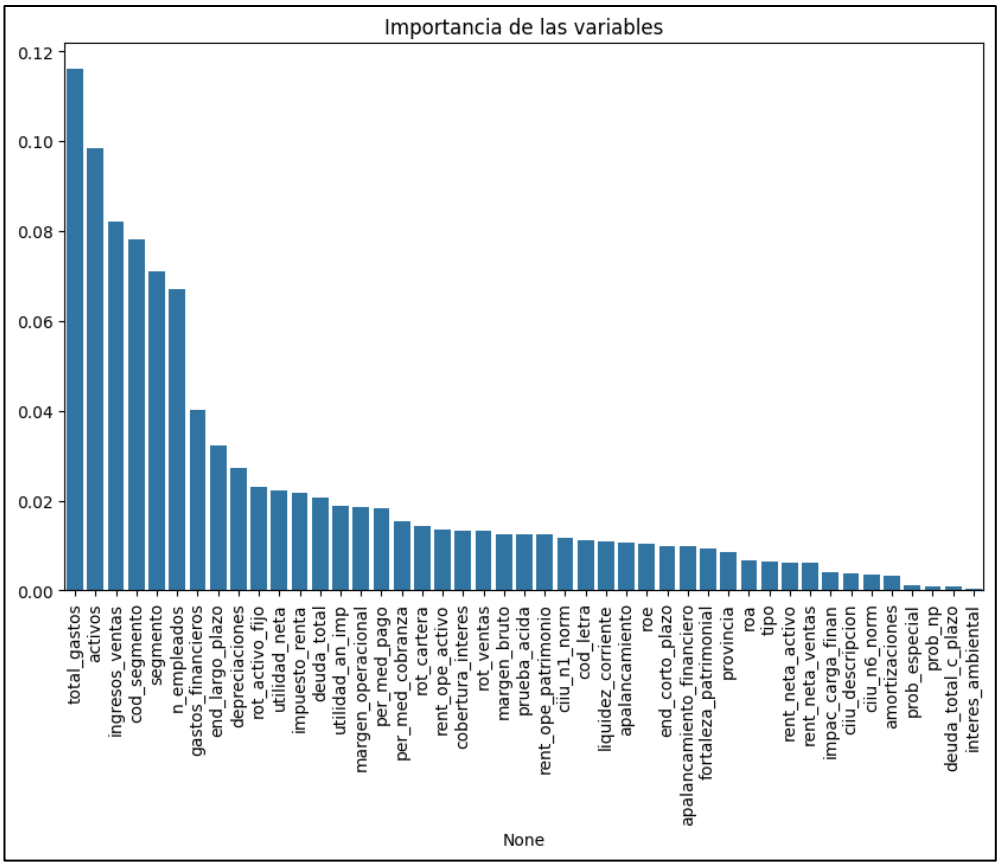

Figura 4.2 Importancia de las variables

Las variables eliminadas en esta fase por baja importancia fueron:

- **end_corto_plazo**
- **apalancamiento_financiero**

- **fortaleza_patrimonial**
- **provincia**
- **roa**
- **tipo**
- **rent_neta_activo**
- **rent_neta_ventas**
- **impac_carga_finan**
- **ciiu_descripcion**
- **ciiu_n6_norm**
- **amortizaciones**
- **prob_especial**
- **prob_np**
- **deuda_total_c_plazo**
- **interes_ambiental**

Finalmente, se empleó el Factor de Inflación de Varianza (VIF) para garantizar que no existiera multicolinealidad residual entre las variables restantes. El VIF cuantifica cuánto se infla la varianza de un coeficiente de regresión debido a la correlación con otras predictoras. Las variables con valores de VIF superiores a un umbral predefinido de diez (10), fueron consideradas redundantes y se eliminaron con el fin de estabilizar el modelo.

El conjunto final de variables seleccionadas, listo para el entrenamiento del modelo predictivo supervisado, fue:

- **cod_segmento**
- **segmento**
- **ingresos_ventas**
- **total_gastos**
- **n_empleados**
- **liquidez_corriente**
- **prueba_acida**

- **deuda_total**
- **apalancamiento**
- **end_largo_plazo**
- **cobertura_interes**
- **gastos_financieros**
- **utilidad_an_imp**
- **utilidad_neta**
- **margen_bruto**
- **margen_operacional**
- **roe**
- **rent_ope_activo**
- **rent_ope_patrimonio**
- **rot_ventas**
- **rot_cartera**
- **rot_activo_fijo**
- **per_med_cobranza**
- **per_med_pago**
- **depreciaciones**
- **potencial_cliente**

Este proceso riguroso garantizó un modelo predictivo robusto, basado en atributos de alta importancia y mínima redundancia.

4.4 Validación del Modelo: Desempeño y Optimización

La validación del modelo se realizó utilizando una muestra estratificada, la cual fue diseñada para preservar la distribución real de clientes en la población de estudio. Esto garantiza que el modelo se entrene y evalúe bajo proporciones realistas, donde los clientes representan una minoría. Adicionalmente, se realizó un backtesting temporal, entrenando el modelo con datos históricos de 2023-2024 y validándolo con datos de 2025, asegurando un enfoque prospectivo y la capacidad de generalización en el tiempo.

En esta sección se detalla la metodología utilizada para la selección y optimización del modelo predictivo, así como la cuantificación de su rendimiento en la tarea de estimar el porcentaje de probabilidad de conversión.

Para identificar el algoritmo de clasificación más adecuado para la predicción de conversión de leads, se compararon y evaluaron múltiples modelos de Machine Learning.

El modelo que demostró el mejor desempeño inicial fue el Gradient Boosting Classifier. Este modelo fue seleccionado y posteriormente optimizado (tuned) mediante la búsqueda de hiper parámetros para maximizar su rendimiento, obteniendo los siguientes resultados en el conjunto de datos de prueba:

4.5 Métricas de Clasificación y Matriz de Confusión

Los resultados del modelo Gradient Boosting Classifier se evaluaron mediante la matriz de confusión y métricas clave de clasificación (Accuracy, Precision, Recall y ROC-AUC), con el objetivo de cuantificar su capacidad para distinguir entre clientes convertidos y no convertidos.

La matriz de confusión muestra el número de predicciones correctas e incorrectas:

- **Verdaderos Negativos (TN):** 57.81% Empresas que no se convirtieron y fueron predichas correctamente como no convertidas.
- **Falsos Positivos (FP):** 8.41% Empresas que no se convirtieron, pero fueron predichas incorrectamente como convertidas.
- **Falsos Negativos (FN):** 4.95% Empresas que se convirtieron, pero fueron predichas incorrectamente como no convertidas.
- **Verdaderos Positivos (TP):** 28.83% Empresas que se convirtieron y fueron predichas correctamente como convertidas.

A partir de la Matriz de Confusión, se calculan las métricas clave:

Tabla 4.1 Métricas clave de la matriz de confusión

Métrica	Valor Obtenido	Interpretación
Accuracy	0.88	El modelo clasifica correctamente al 88% de los prospectos.
Precisión	0.77	Del total de prospectos que el modelo predijo como convertidos, el 77\% realmente lo fueron.
Recall	0.85	El modelo identificó correctamente al 85% de todos los prospectos que se convirtieron.

4.6 Curvas ROC y Precision-Recall

Las curvas visualizan el desempeño del clasificador a través de diferentes umbrales de decisión:

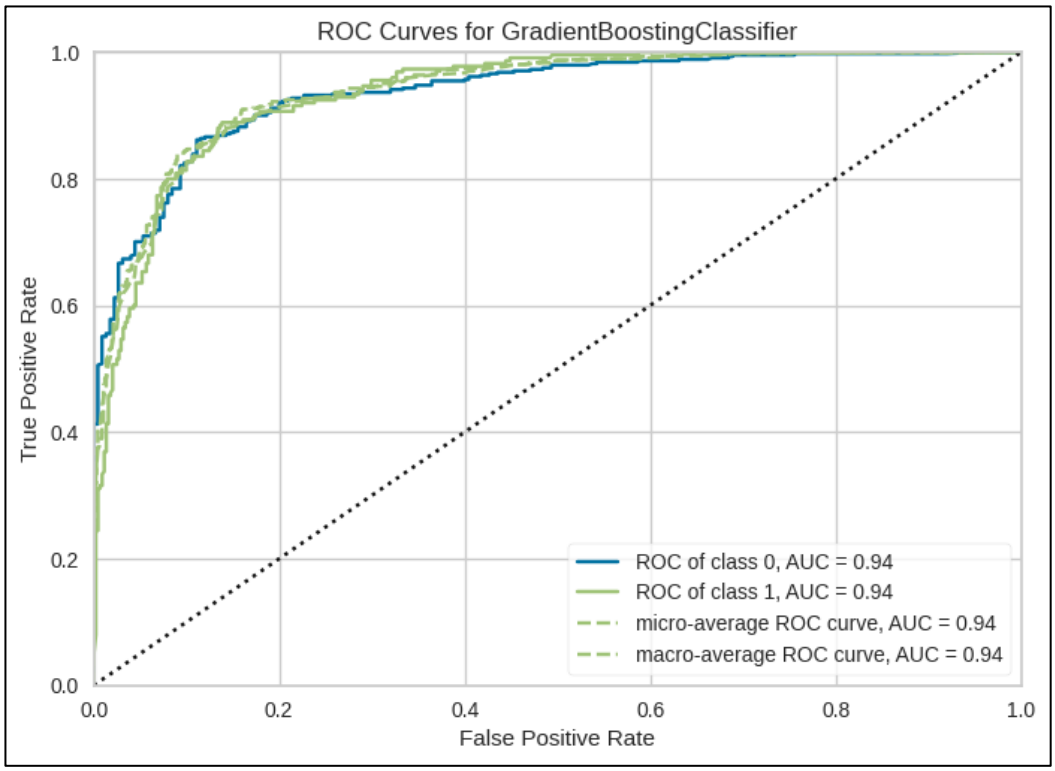

Figura 4.3 Resultado de curva ROC

Curva ROC (Receiver Operating Characteristic): La curva ROC muestra el compromiso entre la Tasa de Verdaderos Positivos (Recall) y la Tasa de Falsos Positivos. El Área Bajo la Curva (AUC) obtenida fue de 0.94 para ambas clases y los promedios micro y macro. Este valor, cercano a 1.0, indica una alta capacidad discriminatoria del modelo para diferenciar entre leads convertidos y no convertidos, validando su utilidad.

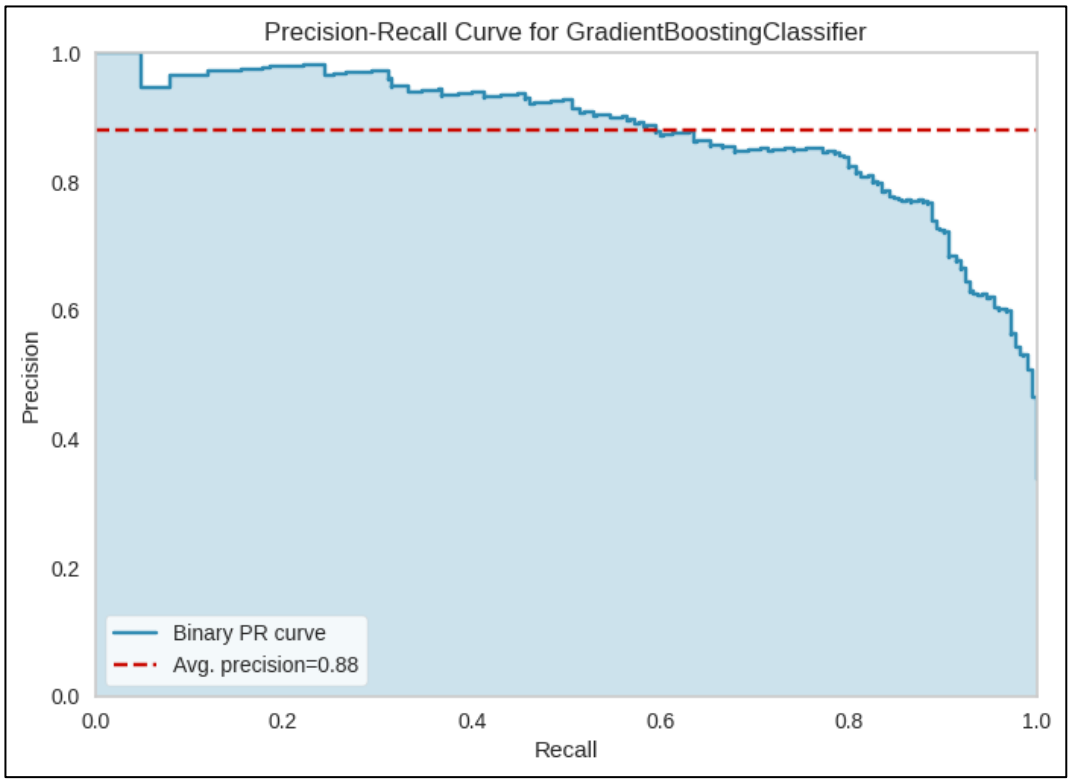

Figura 4.4 Resultados de Recall y Precisión

La curva Precision-Recall es crucial para problemas de clasificación donde una clase puede estar desbalanceada. El gráfico muestra un Average Precision de 0.88. Este valor confirma que el modelo mantiene una alta precisión incluso cuando se intenta capturar una alta proporción de los leads convertidos (Recall).

4.7 Variables con Mayor Influencia Predictiva

El gráfico de importancia de características muestra la contribución relativa de las variables más influyentes del conjunto final y enriquecido. Los indicadores financieros y estructurales dominan la predicción de conversión:

- **total_gastos:** Es, por un margen significativo, el factor más importante, destacando que el volumen operativo y de inversión de una empresa es el principal predictor de su potencial de conversión.

- **n_empleados:** La dimensión y estructura de la fuerza laboral es el segundo factor más influyente.
- **rot_cartera:** Un indicador financiero clave de eficiencia en la gestión de cuentas por cobrar.
- **apalancamiento:** Mide la relación deuda/patrimonio de la empresa.
- **ciiu_n1_norm:** La categorización sectorial (CIIU) de primer nivel normalizada aporta valor, aunque en menor medida.

4.8 Insights para el Acercamiento Comercial

La priorización de estas variables en el modelo proporciona al equipo comercial una base de insights sólida para el acercamiento social y la formulación de una mejor oferta. El modelo no solo da un puntaje, sino que también indica qué tipo de empresa (por su tamaño y gestión) es más probable que se convierta, permitiendo:

Personalización de la Oferta: Adaptar los servicios y la propuesta de valor a empresas con un alto '**total_gastos**' y '**n_empleados**', entendiendo su escala operativa.

Enfoque Estratégico: Priorizar a empresas con características de gestión financiera particulares (ej. '**rot_cartera**' y '**apalancamiento**'), lo que facilita un diálogo más informado sobre sus necesidades estructurales.

4.9 Estandarización y Priorización de Leads

La implementación del producto digital representó un avance significativo en la gestión comercial. El principal logro del producto digital fue estandarizar los criterios de calificación de leads entre los equipos de marketing y comercial.

El puntaje predictivo (probabilidad de conversión del Gradient Boosting Classifier) es utilizado como la herramienta objetiva para la priorización de contactos, eliminando la dependencia de criterios subjetivos.

Esto permite al equipo de Marketing optimizar el presupuesto publicitario, enfocando los recursos solo en los leads con el score más alto.

4.10 Beneficios Operacionales y Estratégicos

El producto digital transformó las herramientas disponibles para el equipo comercial. Antes, el equipo solo tenía acceso a SAP o HubSpot para la gestión de la relación con el cliente. El producto digital actúa como una capa de Inteligencia de Negocios que consolida el score predictivo con los indicadores financieros, ambientales y estructurales enriquecidos.

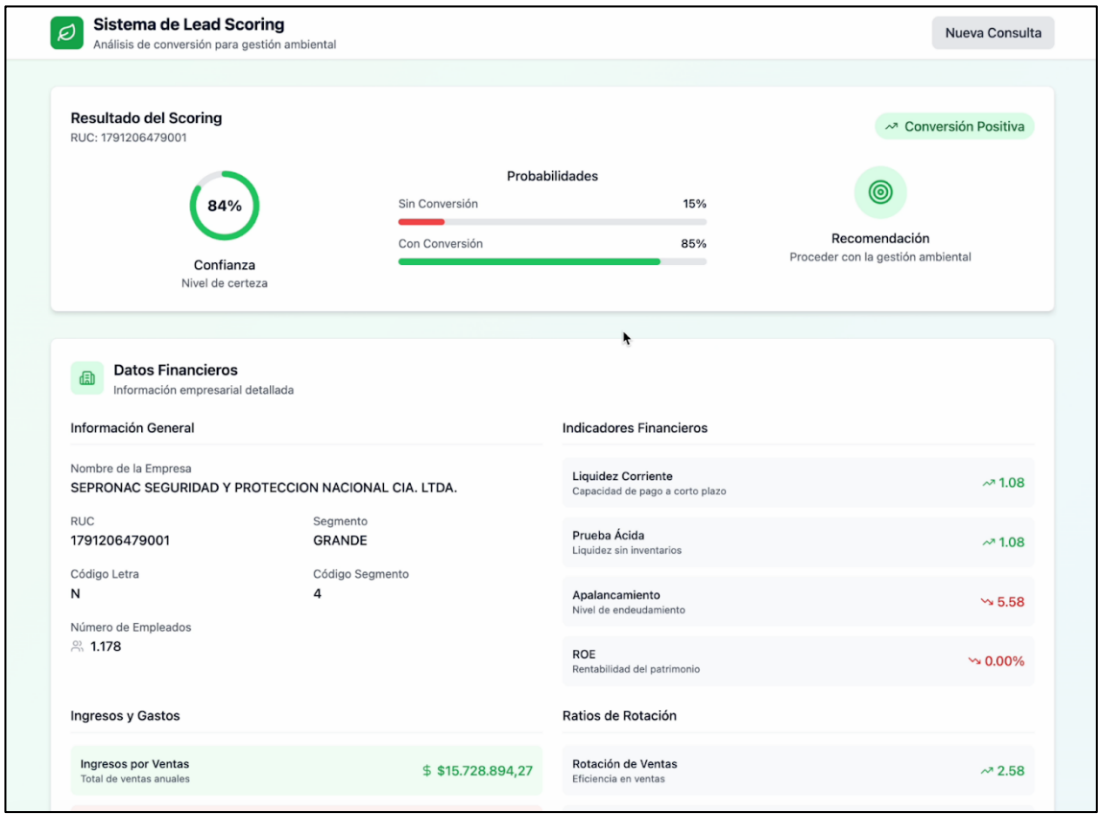

Figura 4.5 Dashboard de precisión

La herramienta proporciona a los comerciales más insights sobre el potencial y las características operacionales de cada prospecto. Esto les permite planificar un acercamiento más estratégico y con una mejor comprensión del

negocio del cliente, elevando la calidad de la interacción y la oferta presentada.

- **Página de Inicio y Objetivo:** La interfaz da la bienvenida al usuario, explicando que la plataforma realiza análisis avanzado de datos financieros y operativos para predecir la probabilidad de conversión en proyectos de gestión ambiental.
- **Consulta de Conversión:** El usuario interactúa con un formulario que solicita el RUC (Registro Único de Contribuyentes) de una empresa prospecto.
- **Ejecución del Modelo (Scoring):** Tras ingresar un RUC, el sistema ejecuta el modelo predictivo (Gradient Boosting Classifier, según el contexto anterior) para generar el score de conversión.
- **Presentación de Resultados (Dashboard):** Se presenta un dashboard completo dividido en varias secciones críticas:
- **Resultado del Scoring:** Muestra las probabilidades: de no conversión y de conversión, con un indicador de "Confianza" y una Recomendación ("Proceder con la gestión ambiental").
- **Datos Financieros:** Presenta información detallada y enriquecida sobre la empresa consultada, incluyendo:
- **Información General:** RUC, Segmento ("GRANDE"), Número de Empleados.
- **Indicadores Financieros y de Rotación:** Métricas como Liquidez Corriente, Prueba Ácida, Apalancamiento, ROE, y Ratios de Rotación (Ventas, Cartera, Activo Fijo, Cobertura de Intereses), muchos de ellos con flags de color verde/rojo que indican tendencia o nivel.
- **Análisis SHAP (SHapley Additive exPlanations):** Se presenta un gráfico explicativo donde los valores positivos favorecen la conversión y los negativos la reducen. Esto permite al usuario ver la contribución individual de cada variable a la probabilidad.

El producto desarrollado e implementado como un producto digital consolida la influencia de las variables del modelo y facilita la visualización

de patrones de conversión. La plataforma opera como una herramienta estratégica de Inteligencia Empresarial para los equipos comerciales y de marketing, permitiendo la consulta instantánea de una empresa a través de su RUC. Al integrarse el puntaje predictivo (probabilidad de conversión) con datos financieros, operativos, y un análisis SHAP individualizado, el sistema provee una base objetiva para la priorización de leads y la optimización del presupuesto publicitario. Este dashboard interactivo ofrece la visión 360° necesaria para que el equipo comercial obtenga mejores insights para el acercamiento social y la formulación de una oferta de valor superior, trascendiendo las limitaciones de herramientas transaccionales previas como SAP.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Se desarrolló e implementó con éxito el sistema de inteligencia comercial para el sector industrial ecuatoriano. La integración de variables financieras de la Superintendencia de Compañías, indicadores de cumplimiento ambiental del MAATE y el posicionamiento en el ranking EKOS permitió una caracterización multidimensional de los prospectos.

- El modelo de aprendizaje automático supervisado demostró una alta capacidad predictiva, alcanzando métricas de ROC-AUC, Accuracy y Recall >0.80 que garantizan una clasificación confiable de la probabilidad de conversión.
- El producto digital muestra la influencia de las variables clave, permitiendo que la toma de decisiones deje de ser intuitiva y pase a ser basada en datos, cumpliendo con la necesidad de priorizar esfuerzos en un mercado industrial competitivo.
- La estandarización de criterios entre Marketing y Ventas se materializó a través del puntaje predictivo (Lead Scoring). Este mecanismo elimina la subjetividad en la calificación de prospectos.
- El modelo predice "quién comprará", pero no "qué servicio específico" o "en qué volumen", debido a la heterogeneidad de la data histórica interna.

5.2 Recomendaciones

- Es prioritario desarrollar una API o proceso ETL automatizado que conecte el modelo predictivo con el CRM (HubSpot) y el ERP (SAP). La carga manual de datos degrada la agilidad comercial y aumenta el riesgo de errores humanos.
- Dado que el sector industrial ecuatoriano es sensible a cambios regulatorios y macroeconómicos, se debe establecer un protocolo de

monitoreo trimestral para detectar desviaciones en los datos de entrada (Data Drift) que puedan invalidar las predicciones actuales.

- Para superar la limitación de la "oferta específica", se sugiere que la siguiente fase del proyecto utilice algoritmos de clasificación multiclase o sistemas de recomendación que sugieran el servicio ambiental óptimo basado en el perfil de la empresa.
- Realizar talleres conjuntos entre los equipos de Marketing y Comercial para asegurar la interpretación correcta del puntaje predictivo, evitando que el sistema sea percibido como una "caja negra" y fomentando la confianza en la herramienta.

BIBLIOGRAFÍA

Espinoza, A. (2023). Economía circular: una aproximación a su origen, evolución e importancia como modelo de desarrollo sostenible. *Revista de economía institucional*, 109-134.

MAATE. (4 de 5 de 2015). *Portal Único de Trámites Ciudadanos.* Obtenido de www.gob.ec

INEC. (1 de 3 de 2025). *https://www.ecuadorencifras.gob.ec.* Obtenido de Instituto Nacional de Estadísticas y Censos: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Encuestas_Ambientales/EMPRESAS/2023/MIEAE%20_03_2025.pdf

Veolia. (s.f.).

Chollet, F. (2018). *Deep Learning with Python .* MANEJO.

Cunha, C., & Coelho, M. (2022). Industry 4.0: Predicting Lead Conversion Opportunities with Machine Learning in Small and Medium Sized Enterprises. *authenticus*, 54-64.

Brei, V. A. (2020). Machine Learning in Marketing: Overview, Learning Strategies, Applications, and Future Developments. *Research Gate*, 173-236.

Singh, Y. (2024). Predicting Marketing Campaign Effectiveness through Consumer Personality Analysis using Machine Learning . *Research Gate*.

González, L., Rubiano, J., & Sosa, G. (2025). The relevance of lead prioritization: a B2B lead scoring model based on machine learning. *frontiers*.

Tomar, V., & Shriram, N. (2023). Lead Score Analysis. *International Journal for Research Trends and Innovation*.

Wu, M., Andreev, P., & Morad, B. (2023). The state of lead scoring models and their impact on sales performance. *ResearchGate*, 69-98.

Ha-Thuc, V., & Sinha, S. (2016). Learning to Rank Personalized Search Results in Professional Networks. *arxiv*.

Gabriel, J. (2024). A Machine Learning-Based Web Application for Heart Disease Prediction. *Intelligent Control and Automation*.

Verma, A., & Dong, X. (2016). Detección de fibrilación ventricular mediante clasificador de bosque aleatorio. *Journal of Biomedical Science and Engineering*.

Mezei, J., & Nygård , R. (2020). Automating Lead Scoring with Machine Learning: An Experimental Study. *University of Hawai'i at Manoa* .

Slakey, A., Salas, D., & Schamroth, Y. (2019). Encoding Categorical Variables with Conjugate Bayesian Models for WeWork Lead Scoring Engine. *arXiv*.

Superintendencia de compañías, v. y. (2024). Obtenido de Superintencia de Compañías: https://appscvsmovil.supercias.gob.ec/ranking/reporte.html?utm_source=chatgpt.com

Gobierno del Ecuador. (16 de 03 de 2023). *Scribd.* Obtenido de https://www.scribd.com/document/881077516/Listado-de-Gestores-Ambientales-2023-03-16

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Mary Ann Liebert*.

Ekos. (2022). *Ekos*. Obtenido de https://ekosnegocios.com/articulo/empresas-adherentes-programa-ecuador-carbono-cero

Stefani, A., & Vassiliadis, B. (2021). A Quality Assurance Reference Framework for Assessing Educational Data. *Journal of Data Analysis and Information Processing*.