



**Modelo de analítica de datos para la optimización de flujos de trabajo  
en consultoría ambiental para INGISOT SAS**

Oscar Armando Gavidia Albarracín

Juan Camilo Rubiano Usaquén

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

10/noviembre/2025

**Modelo de analítica de datos para la optimización de flujos de trabajo  
en consultoría ambiental para INGISOT SAS**

**Oscar Armando Gavidia Albarracín**

**Juan Camilo Rubiano Usaquén**

Trabajo de grado presentado como requisito para optar al título de:

**Magister en Ciencia de Datos**

Director (a):

José Alexander Fuentes Montoya

Modalidad:

**Trabajo Dirigido**

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

10/noviembre/2025

Nota de aceptación:

---

---

---

---

---

---

Firma del jurado

---

Firma del jurado

---

Firma del director del trabajo de grado

10/noviembre/2025

## Resumen

La presente tesis propone un modelo de analítica de datos orientado a la optimización de los flujos de trabajo en la consultoría ambiental desarrollada por INGISOT SAS. El estudio se fundamenta en el mapeo de procesos internos de las principales líneas de servicio consultoría ambiental, catastro, recursos hídricos, ordenamiento territorial y sistemas de información geográfica con el fin de identificar redundancias, cuellos de botella y oportunidades de mejora.

El modelo integra metodologías de ciencia de datos como CRISP-DM, SEMMA, TDSP y OSEMN, junto con enfoques ágiles de gestión (Scrum y Kanban), estructurando un ciclo iterativo de mejora continua que articula tres niveles analíticos: descriptivo, predictivo y prescriptivo. Asimismo, plantea el uso potencial de herramientas de código abierto (Python, R, Power BI, PostgreSQL/PostGIS, MLflow y Docker) como soporte para la toma de decisiones y la automatización de procesos.

El resultado es una propuesta integral que fortalece la trazabilidad de la información, la interoperabilidad entre áreas y la capacidad institucional para gestionar datos de manera estructurada. Se formulan tres fases de implementación preparación, integración-validación y optimización-escalamiento que orientan la adopción tecnológica y la sostenibilidad del modelo en la organización.

**Palabras clave:** analítica de datos, optimización de procesos, consultoría ambiental, ciencia de datos, gestión organizacional.

## **Abstract**

This thesis proposes a data analytics model aimed at optimizing workflow processes within the environmental consulting activities carried out by INGISOT SAS. The study is based on the mapping of internal processes across the main lines of service —environmental consulting, cadastre, water resources, territorial planning, geology, and geographic information systems— in order to identify redundancies, bottlenecks, and opportunities for improvement.

The model integrates well-established data science methodologies such as CRISP-DM, SEMMA, TDSP, and OSEMN, together with agile management approaches (Scrum and Kanban), structuring an iterative cycle of continuous improvement that encompasses three analytical levels: descriptive, predictive, and prescriptive. It also proposes the potential use of open-source tools (Python, R, Power BI, PostgreSQL/PostGIS, MLflow, and Docker) to support decision-making and the progressive automation of processes.

The result is a comprehensive proposal that strengthens information traceability, enhances interoperability between organizational areas, and consolidates institutional capacity for structured data management. Three implementation phases —preparation, integration and validation, and optimization and scaling— are outlined to guide technological adoption and ensure the model's sustainability within the organization.

**Keywords:** data analytics, process optimization, environmental consulting, data science, organizational management.

---

## Contenido

<b>1. Introducción .....</b>	<b>10</b>
<b>2. Objetivos.....</b>	<b>12</b>
2.1.    Objetivo General .....	12
2.2.    Objetivos específicos.....	12
<b>3. Justificación .....</b>	<b>13</b>
<b>4. Marco institucional.....</b>	<b>15</b>
4.1.    Estructura organizacional .....	16
4.1.1.    Gerencia general .....	16
4.1.2.    Dirección de proyectos .....	17
4.1.3.    Líder técnico. ....	17
4.1.4.    Personal técnico especializado. ....	18
4.2.    Productos o servicios ofertados por INGISOT SAS .....	18
4.3.    Análisis de sector. ....	18
<b>5. Marco de referencia .....</b>	<b>22</b>
5.1.    Fundamentos en ciencia de datos .....	22
5.1.1.    ¿Qué es la ciencia de datos? .....	22
5.1.2.    Ciencia de datos espacial .....	23
5.1.3.    Flujos de trabajo en ciencia de datos .....	23
5.2.    Modelos en la ciencia de datos .....	24
5.2.1.    Modelos descriptivos.....	25
5.2.2.    Modelos predictivos .....	25
5.2.3.    Modelos prescriptivos.....	26
5.3.    Metodologías para minería de datos y ciencia de datos .....	26
5.3.1.    CRISP-DM.....	26
5.4.    Herramientas para la ciencia de datos .....	30
5.5.    Despliegue y sostenibilidad del modelo.....	35
<b>6. Diseño metodológico.....</b>	<b>40</b>
6.1.    Tipo de investigación.....	40
6.2.    Análisis externo - PESTEL .....	42

---

6.2.1.	Cinco fuerzas de Poder .....	44
6.2.2.	Matriz DOFA .....	46
6.3.	Análisis interno .....	47
6.3.1.	Tema .....	47
6.3.2.	Objetivo de la encuesta. ....	47
6.3.3.	Variables para medir .....	47
6.3.4.	Escala de respuestas sugerida (tipo Likert): .....	49
6.4.	Población, muestra y ficha técnica. ....	49
6.4.1.	Tipo de muestreo y composición de la muestra .....	50
6.5.	Identificación de las variables .....	50
6.6.	Instrumento de análisis interno .....	51
6.7.	Validación del instrumento de medición .....	52
6.7.1.	Relación entre diagnóstico y diseño del modelo .....	53
<b>7.</b>	<b>Diagnóstico organizacional.....</b>	<b>54</b>
7.1.	Metodología .....	54
7.1.1.	Mapeo de procesos .....	54
7.1.2.	Matriz de oportunidades .....	56
7.2.	Líneas estratégicas INGISOT SAS.....	57
7.2.1.	Consultoría ambiental .....	57
7.2.2.	Geografía y catastro .....	60
7.2.3.	Ordenamiento territorial .....	63
7.2.4.	Recurso hídrico y modelación .....	67
7.2.5.	Sistemas de información geográfica .....	71
7.2.6.	Análisis de resultados.....	75
<b>8.</b>	<b>Modelo de datos para INGISOT SAS .....</b>	<b>77</b>
8.1.	Propuesta del modelo.....	77
8.2.	Elementos del modelo .....	77
8.2.1.	Actividades o subactividades .....	80
8.2.2.	Tipología de modelos.....	80
8.2.3.	Ciclo de vida .....	81

---

8.2.4.	Metodologías ágiles .....	82
8.2.5.	Resultados esperados .....	82
8.2.6.	Despliegue.....	83
<b>9.</b>	<b>Plan de intervención – propuesta de implementación.....</b>	<b>85</b>
9.1.	Fases de la implementación. ....	85
9.1.1.	Fase I. Preparación organizacional y técnica.....	86
9.1.2.	Fase II, Integración y validación del modelo.....	88
9.1.3.	Fase III, Optimización y escalamiento .....	89
9.2.	Componente técnico.....	90
9.3.	Componente organizacional .....	90
9.4.	Recursos requeridos .....	91
9.5.	Seguimiento y evaluación.....	92
9.6.	Plan de inversión.....	92
	<b>Conclusiones y recomendaciones .....</b>	<b>96</b>
	<b>Referencias .....</b>	<b>99</b>

### Lista de figuras

Figura 1 Estructura organizacional de INGISOT SAS .....	17
Figura 2 Top 10 de los departamentos con mayor número de títulos concesionados	19
Figura 3 Distribución del top 10 de departamentos con más títulos mineros concesionados. ....	20
Figura 4 Distribución de los 5 departamentos a nivel nacional con la mayor cantidad de pozos de petróleo registrados. ....	21
Figura 5 El proceso de minería de datos CRISP-DM .....	28
Figura 6 El proceso de la minería de datos en la metodología SEMMA.....	29
Figura 7 Reproducción del diagrama de flujo del proceso TDPS .....	30
Figura 8 Diferentes componentes de un sistema de aprendizaje profundo.....	33
Figura 9 Ciclo de vida del agua proporcionando una analogía para los cinco pasos del ciclo de vida de los datos. ....	35
Figura 10 Ciclo de vida del dato simplificado .....	36
Figura 11 Objetos de un repositorio GIT.....	37
Figura 12 Marco de migración.....	38

---

Figura 13 Mapeo de procesos .....	55
Figura 14 Mapeo de procesos línea de consultoría ambiental .....	58
Figura 15 Mapeo de procesos línea de geografía y catastro .....	61
Figura 16 Mapeo de procesos línea de ordenamiento territorial .....	65
Figura 17 Mapeo de procesos línea de recurso hídrico. ....	68
Figura 18 Mapeo de procesos línea SIG .....	72
Figura 19 Modelo de datos a implementar .....	78
Figura 20 Fases para la implementación propuestas .....	86
Figura 21 Fases para la preparación organizacional .....	87
Figura 22 sub-fases para la integración y validación del modelo.....	88
Figura 23 Roles dentro del componente organización y su relación con el modelo. ..	90

### Lista de tablas

<b>Tabla 1 Etapas de la metodología CRISP - DM.....</b>	<b>27</b>
<b>Tabla 2 Encuesta en escala Likert (formato: Totalmente en desacuerdo a totalmente de acuerdo) .....</b>	<b>48</b>
<b>Tabla 3 Tamaño de la muestra, ficha técnica. ....</b>	<b>49</b>
<b>Tabla 4 Definición de variables en la encuesta likert .....</b>	<b>50</b>
<b>Tabla 5 Expertos temáticos consultados.....</b>	<b>52</b>
<b>Tabla 6 Resultados de la ponderación V- Aiken.....</b>	<b>52</b>
<b>Tabla 7 Síntesis de hallazgos y su implicación en el modelo.....</b>	<b>53</b>
<b>Tabla 8 Elementos del mapa del proceso. ....</b>	<b>54</b>
<b>Tabla 9 Matriz de oportunidad en consultoría ambiental .....</b>	<b>59</b>
<b>Tabla 10 Matriz de oportunidad línea de geografía y catastro .....</b>	<b>62</b>
<b>Tabla 11 Matriz de oportunidad en el ordenamiento territorial .....</b>	<b>66</b>
<b>Tabla 12 Matriz de oportunidad recurso hídrico y modelación .....</b>	<b>70</b>
<b>Tabla 13 Matriz de oportunidad en SIG .....</b>	<b>75</b>
<b>Tabla 14 Fase I. Preparación e infraestructura inicial para la analítica de datos ..</b>	<b>87</b>
<b>Tabla 15 Propuesta de inversión progresiva .....</b>	<b>93</b>

---

## 1. Introducción

La consultoría ambiental opera en un entorno caracterizado por alta complejidad técnica, múltiples fuentes de información y cambios frecuentes en los requerimientos de los proyectos. En este contexto, la eficiencia operativa y la capacidad de adaptación de los equipos de trabajo se convierten en factores determinantes para el cumplimiento de plazos, la calidad de los productos y la sostenibilidad organizacional.

INGISOT SAS enfrenta limitaciones en la organización, seguimiento y control de sus flujos de trabajo de consultoría ambiental. Estas limitaciones se originan en la dispersión de la información técnica, la ejecución reiterada de procesos manuales y la ausencia de herramientas analíticas integradas que faciliten el ajuste oportuno de actividades ante cambios en los proyectos. Como consecuencia, se presentan incrementos en el tiempo operativo, reprocesos técnicos, retrasos en la entrega de documentos y dificultades para mantener un estándar homogéneo de calidad en los productos entregados.

El problema identificado se manifiesta principalmente a través de tres aspectos operativos. Primero, el tiempo dedicado a tareas repetitivas y ajustes manuales dentro de los procesos internos. Segundo, el tiempo total de entrega de los documentos de consultoría ambiental. Tercero, la calidad técnica de los productos finales, entendida como la suficiencia metodológica, el soporte analítico y la trazabilidad entre datos, análisis y conclusiones. Estos aspectos constituyen indicadores base para evaluar el impacto de una intervención orientada a la optimización de los flujos de trabajo.

La aplicación de la ciencia de datos en contextos organizacionales ha sido documentada principalmente a través de enfoques orientados a la automatización de procesos, el análisis descriptivo del desempeño operativo y el soporte a la toma de decisiones en entornos de proyectos. Experiencias reportadas en literatura técnica, estudios de caso empresariales y repositorios abiertos de analítica aplicada muestran que la incorporación progresiva de herramientas basadas en datos, desde scripts automatizados hasta tableros de seguimiento, contribuye a mejorar la eficiencia operativa, la trazabilidad de la información y la calidad de los productos generados, especialmente en organizaciones con alta carga técnica y variabilidad en sus procesos.

En este contexto, el presente trabajo propone una ruta de implementación de la ciencia de datos aplicada a los flujos de trabajo internos de INGISOT SAS, concebida como un esquema flexible y progresivo. La propuesta parte del análisis de las prácticas actuales de la organización y plantea la incorporación gradual de componentes analíticos, desde la automatización de procesos mediante código hasta el uso de herramientas de análisis y visualización, en función de necesidades operativas concretas. Esta aproximación permite responder a los cambios propios de los proyectos de consultoría ambiental, reducir los tiempos de ejecución y entrega, y fortalecer la calidad técnica de los productos, alineando los objetivos del modelo analítico con resultados verificables y con la capacidad real de adopción de la organización.

---

## **2. Objetivos**

### **2.1. Objetivo General**

Diseñar un modelo de analítica de datos orientado a la optimización de flujos de trabajo en la consultoría ambiental desarrollada por INGISOT SAS.

### **2.2. Objetivos específicos**

- Identificar los referentes teóricos y metodológicos sobre modelos de analítica de datos aplicados a la optimización de flujos de trabajo en consultoría ambiental.
- Diagnosticar los flujos de trabajo en consultoría ambiental de INGISOT SAS, focalizando procedimientos, herramientas y metodologías para el análisis de datos.
- Estructurar un modelo de analítica de datos orientado a la optimización de flujos de trabajo integrando componentes conceptuales y técnicos relacionados con la adquisición, procesamiento, análisis y visualización de datos.
- Formular una propuesta de implementación del modelo diseñado, considerando aspectos técnicos, organizacionales y de recursos para adopción progresiva en INGISOT SAS.

---

### 3. Justificación

La incorporación de herramientas de código abierto para la analítica de datos representa una estrategia clave en la transformación de los flujos de trabajo empresariales. En el caso de INGISOT SAS, cuya operación actual en consultoría ambiental se apoya en herramientas tradicionales como hojas de cálculo, la falta de sistematización y replicabilidad en el análisis limita la eficiencia operativa y la escalabilidad. En este contexto, identificar modelos de analítica de datos orientados a la optimización se representa como una fase preliminar clave dentro del ciclo de madurez digital organizacional.

Investigar este tema resulta relevante no solo por su aplicabilidad inmediata, sino también por su alineación con las tendencias actuales en ciencia de datos, donde la demanda de soluciones analíticas reproducibles y basadas en datos está en crecimiento exponencial. Además, este tipo de modelo puede servir de referente para otras pequeñas y medianas empresas del sector ambiental que enfrentan desafíos similares de madurez digital y transición tecnológica.

Desde el ámbito académico y aplicado, la pertinencia de este tipo de iniciativas se sustenta en enfoques ampliamente documentados en la ciencia de datos aplicada y la analítica de procesos, donde se reconoce el valor de la automatización, la reproducibilidad del análisis y el uso de modelos descriptivos para mejorar la eficiencia operativa y el control interno. Experiencias reportadas en contextos de consultoría, gestión de proyectos e ingeniería evidencian que la incorporación progresiva de herramientas analíticas basadas en datos contribuye a reducir tiempos operativos y a fortalecer la calidad técnica de los productos generados.

En el contexto actual de transformación digital, las organizaciones enfrentan crecientes presiones económicas que exigen mayor eficiencia, reducción de costos y decisiones basadas en evidencia. Aunque no todos los sectores avanzan al mismo ritmo, la consultoría ambiental, en particular las pequeñas empresas, enfrenta una creciente presión para revisar sus procesos internos ante los desafíos de una era digital.

En este sentido, la ciencia de datos se consolida como una respuesta tecnológica clave, no solo por la disponibilidad de herramientas de código abierto y su capacidad para

automatizar tareas repetitivas, sino también por su potencial para optimizar el análisis, la toma de decisiones y la generación de valor.

En coherencia con lo anterior, el presente trabajo se sustenta en la decisión institucional de incorporar herramientas de ciencia de datos de código abierto como estrategia para optimizar los procesos internos y fortalecer la calidad del análisis y la presentación de resultados. Esta decisión prioriza los procesos más recurrentes que implican análisis sistemático de datos y que presentan una lógica operativa repetitiva, en los cuales la mecánica del análisis se mantiene constante aun cuando cambian las fuentes o los contextos de aplicación. La selección de estos procesos responde a su alta frecuencia de ejecución y a la posibilidad de automatizar tareas previamente identificadas, orientando el trabajo hacia la replicabilidad, la eficiencia operativa y la coherencia entre el diagnóstico realizado, la priorización institucional y el plan de intervención propuesto.

---

#### 4. Marco institucional

INGISOT SAS, que por sus siglas corresponde a Ingeniería, sistemas de información geográfica y ordenamiento territorial, es una empresa colombiana constituida legalmente como Sociedad de Acciones Simplificadas SAS, con sede principal en el municipio de Sogamoso, departamento de Boyacá. Desde su creación a finales del año 2020, ha incursionado y participado en diversos estudios de consultoría, destacándose en consultoría medioambiental, análisis espacial y proyectos de ordenamiento territorial a nivel nacional.

La empresa tiene como misión ofrecer servicios especializados y soluciones integrales en el ámbito del análisis geoespacial, la gestión ambiental y el ordenamiento territorial, aportando valor agregado mediante la aplicación de herramientas tecnológicas avanzadas y un enfoque interdisciplinario.

Su visión es posicionarse como una empresa líder a nivel nacional en la implementación de sistemas de información geográfica automatizados aplicados a la consultoría ambiental y al manejo del territorio, fortaleciendo la toma de decisiones sostenibles y estratégicas en los sectores público y privado.

INGISOT SAS, atiende diversos renglones económicos entre estos a citar le corresponde el sector minero – energético, gubernamental y de infraestructura, dentro de sus clientes se incluyen entidades gubernamentales, corporaciones autónomas regionales y empresas privadas.

La empresa cuenta con un equipo multidisciplinario de profesionales altamente calificados con experiencia asociada a geología, medio ambiente, ciencias naturales y sociales, así como de profesionales especializado en sistemas de información geográfica e hidráulica. Este equipo usa herramientas tecnológicas desde el manejo de software especializado como ArcGIS, QGIS, Autocadmap, R y Python, así como el uso de plataformas digitales que le permite una labor integral dentro de los requerimientos técnicos que se atienden desde sus diversas líneas de acción.

A lo largo de sus 4 años de trayectoria INGISOT SAS ha apoyado y participado en diversos estudios medio ambientales en sectores de minería, infraestructura e hidrocarburos, destacándose su capacidad de adaptación ante los diversos requerimientos técnicos “términos de referencia” y requerimientos normativos a citar; decreto 1232 de 2020 del que trata el ordenamiento territorial.

De otra parte, ha venido incorporando estrategias de innovación tecnológica y mejora continua con énfasis en la transformación progresiva de sus procesos internos en adquisición de equipos especializados de fotogrametría y topografía en línea con misión.

En la actualidad la empresa cuenta con registro ante la cámara de comercio vigente e inscripción en el RUP lo que le permite participar en diversas licitaciones de carácter gubernamental asociados a su línea de trabajo. Internamente se adelanta el sistema de gestión con miras a la presentación dentro del RUC “registro único de contratistas”.

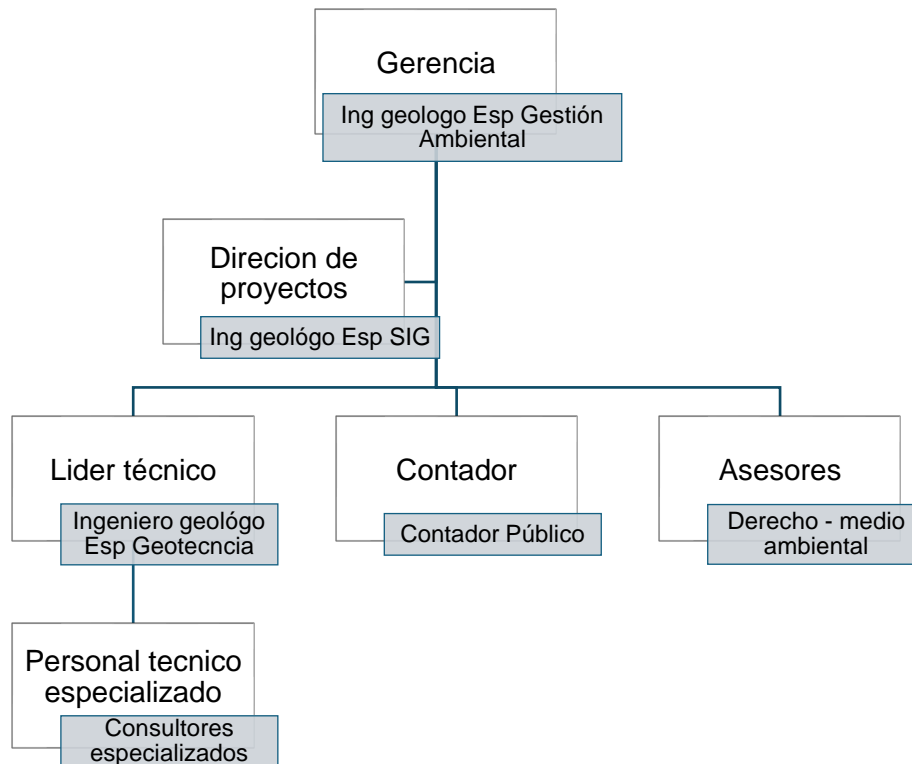
#### **4.1. Estructura organizacional**

Su estructura organización es simple y se basa en un modelo funcional técnico diseñado para garantizar una eficiencia operativa, la especificidad técnica y el seguimiento a los proyectos de ejecución redistribuidos en 4 niveles jerárquicos.

##### **4.1.1. Gerencia general**

Es la instancia máxima de representación, define y articula la visión estratégica liderando los procesos administrativos y comerciales garantizando la ejecución contractual, términos y plazos dentro de las diversas ordenes de trabajo o contratos de consultoría de INGISOT SAS.

**Figura 1 Estructura organizacional de INGISOT SAS**



Fuente: INGISOT SAS.

#### **4.1.2. Dirección de proyectos**

Coordina la ejecución de los proyectos contratados, desde la asignación, guía seguimiento y acompañamiento al coordinador operativo y el personal técnico especializado actuando como puente entre la gerencia y las necesidades técnicas para la consecución cumplimiento y estandarización de resultados en atención específica de los términos de referencia o particularidades de cada contrato u ordenes de trabajo.

#### **4.1.3. Líder técnico.**

Es el responsable de la ejecución metodológica y supervisión técnica en validación de los productos presentados por el personal técnico especializado en cada una de las líneas estratégicas, cada líder técnico cuenta con autonomía para la construcción de su equipo de trabajo aplicando criterios técnicos entorno a la calidad del producto.

Transversalmente se localiza en este nivel la administración contable y equipo de asesores según necesidades particulares de cada proyecto.

#### **4.1.4. Personal técnico especializado.**

Corresponde al equipo o grupo técnico conformado por profesionales de diversas líneas estratégicas conformado para la atención de cada ejecución contractual en áreas como la ingeniería civil, ambiental, social, biótica, apoyos SIG y la topografía. Ejecutan tareas específicas bajo la directriz técnica del líder y director de proyectos.

#### **4.2. Productos o servicios ofertados por INGISOT SAS**

Dentro de sus principales líneas de servicio se destacan:

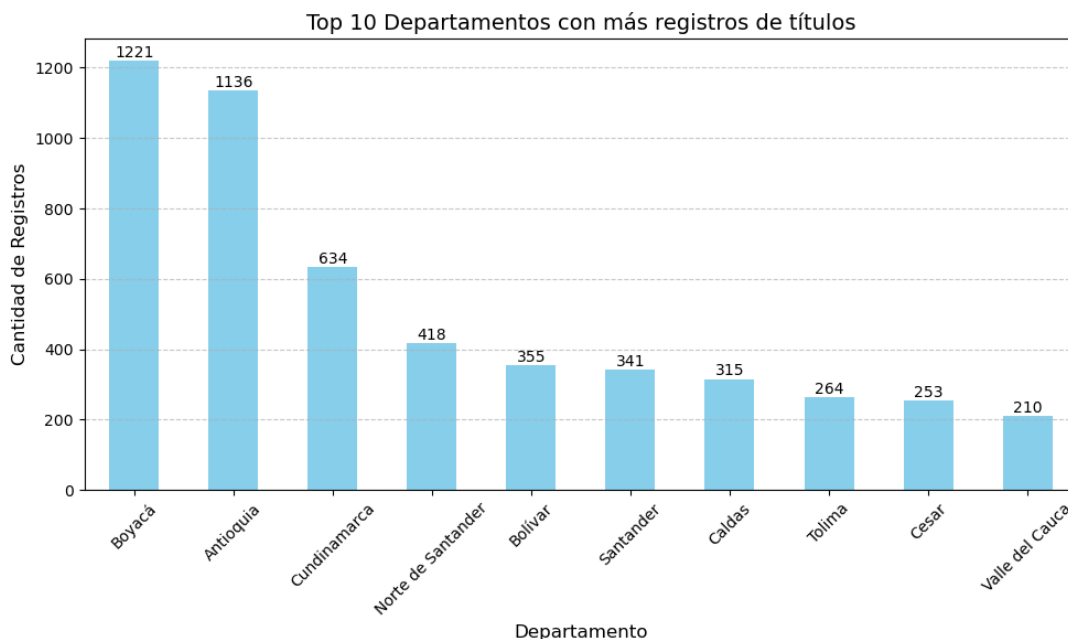
- Consultoría ambiental en atención de diversos requerimientos técnicos emitidos por las autoridades ambientales (CAR'S, ANLA & ANM)
- Análisis espaciales a través del uso de sistemas de información geográfica en sinergia con la adquisición, tratamiento, análisis y determinación de variables fisiográficas.
- Modelación hidrológica y climatología para la generación de modelos hidrogeológicos conceptuales y numéricos a través del uso de software especializado.
- Análisis, prospección y cuantificación de recursos naturales en integración con normatividad en cuantificación de recursos y reservas de la CRIRSCO.
- Levantamiento, adquisición y tratamiento de imágenes obtenidas con naves aéreas no tripuladas UAV/RPAS
- Adquisición, procesamiento, análisis e interpretación de datos geofísicos en utilización de métodos resistivos.
- Estudios de suelos y rocas para el diseño de obras, análisis de estabilidad, diseños hidráulicos y zonificaciones geológica geotécnicas en caracterización de unidades geológicas para ingeniería.
- Estudios de ordenamiento territorial e inclusión de la gestión del riesgo y la solución a conflictos territoriales en apoyo a procesos de delimitación territorial político administrativa.

#### **4.3. Análisis de sector.**

El renglón económico al que INGISOT SAS dirige sus servicios es amplio a escala nacional y local, concentrándose en tres grandes líneas de trabajo: el sector minero, el sector energético y las administraciones municipales, abarcando tanto el sector público como el privado. En el contexto local, tomando como referencia el departamento de Boyacá,

éste ocupa el primer lugar a nivel nacional en cantidad de títulos mineros concesionados por la Agencia Nacional de Minería (ANM), como se observa en la **¡Error! No se encuentra el origen de la referencia..**

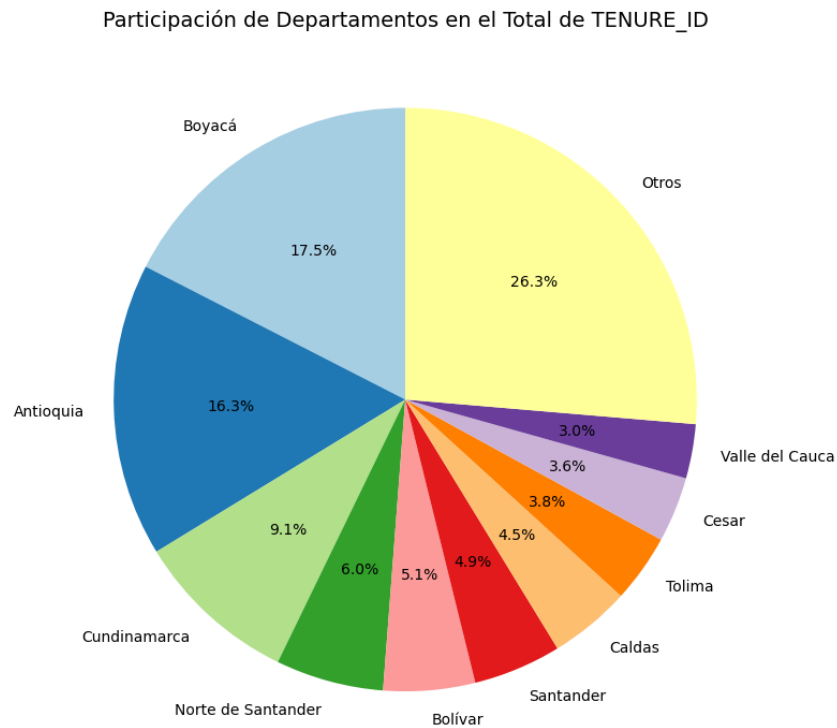
**Figura 2 Top 10 de los departamentos con mayor número de títulos concesionados**



Fuente: descarga y análisis de datos del 06 de abril de 2025.

Esta posición estratégica representa una oportunidad significativa para INGISOT SAS, ya que cada título minero requiere cumplir con los requerimientos de la Agencia Nacional de Minería, así como tramitar y obtener la licencia ambiental de explotación ante la autoridad ambiental competente. Esto duplica las posibilidades de oferta de servicios para la empresa. Los datos muestran que el departamento de Boyacá representa una participación del 17.5% del total nacional de títulos mineros, y sumando el 9.1% de títulos ubicados en el cercano departamento de Cundinamarca, se alcanza una porción superior al 25% del total nacional.

**Figura 3 Distribución del top 10 de departamentos con más títulos mineros concesionados.**

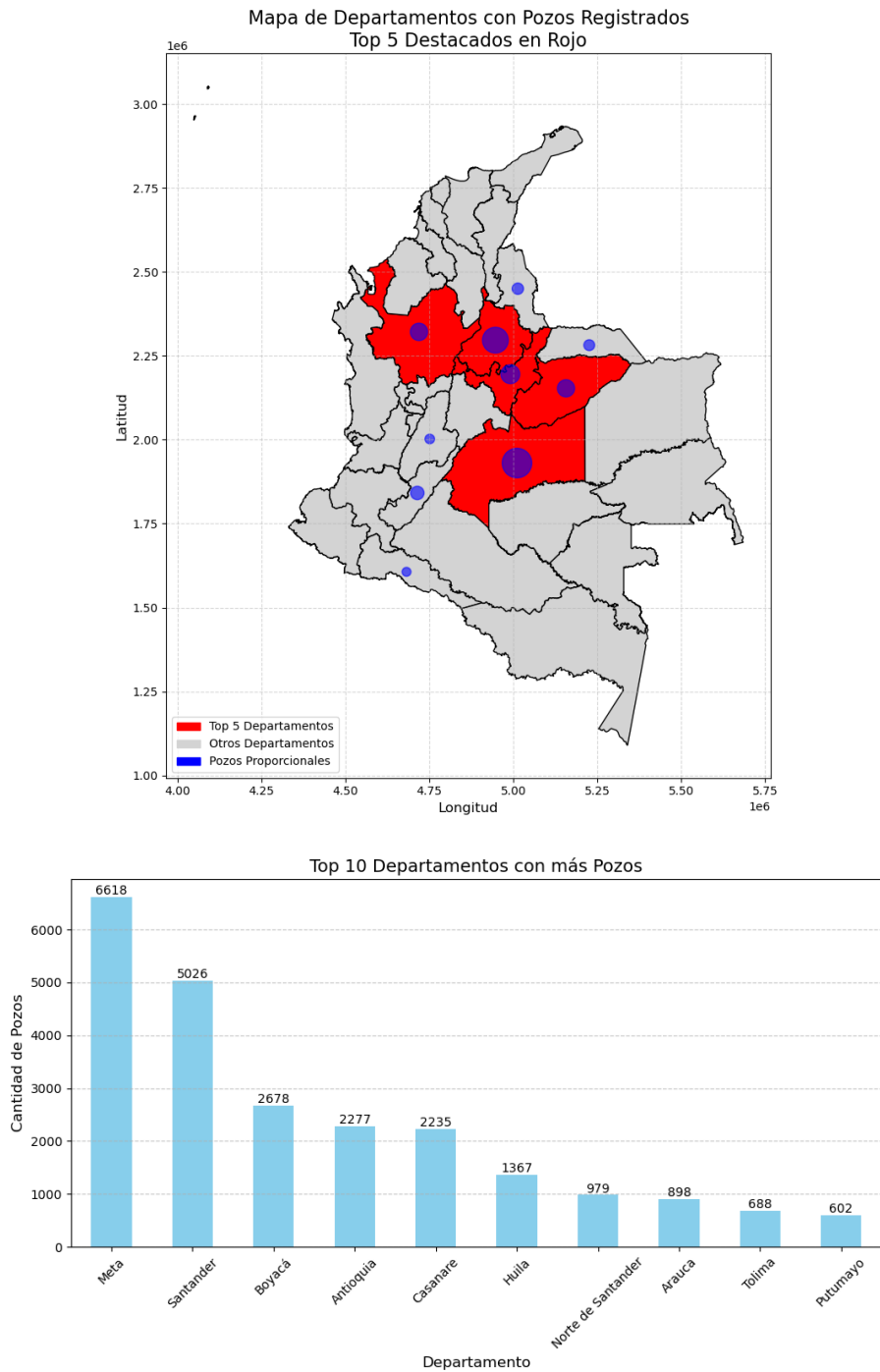


Fuente: descarga y análisis de datos del 06 de abril de 2025.

En cuanto al sector energético, la empresa participa en servicios relacionados con licencias, seguimiento e informes de cumplimiento ambiental para concesiones de hidrocarburos. Gracias a su ubicación estratégica, ha desarrollado proyectos técnicos en los departamentos de Casanare y Boyacá.

Dentro de una mirada paralela a este departamento le corresponde el quinto lugar dentro del top 10, y en cercanía el departamento del meta en el cual se localiza la mayor cantidad de pozos registrados por la Agencia Nacional de hidrocarburos. Por su cercanía y posición estratégica INGISOT SAS cuenta con la posibilidad de incursionar en estos lugares en la actualidad participa en procesos de consultoría ambiental en el Casanare en un tiempo de desplazamiento no superior a 3 horas de su capital Yopal.

**Figura 4 Distribución de los 5 departamentos a nivel nacional con la mayor cantidad de pozos de petróleo registrados.**



Fuente: ANH descarga y análisis de datos del 06 de abril de 2025.

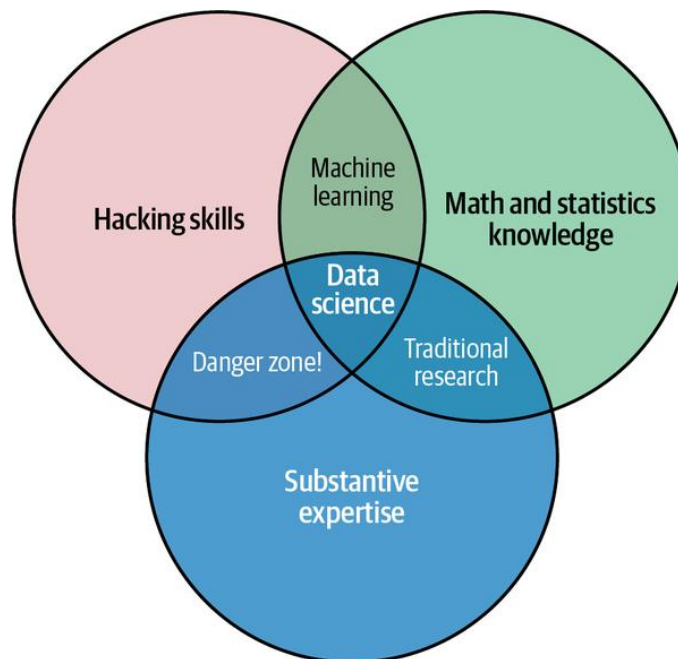
## 5. Marco de referencia

### 5.1. Fundamentos en ciencia de datos

#### 5.1.1. ¿Qué es la ciencia de datos?

Para (Conway, 2010), la ciencia de datos corresponde a convergencia entre las habilidades computacionales, el conocimiento matemático, estadístico y la experticia temática. (Grus, 2019) señala así la posibilidad de encontrarse con científicos de datos bien sea como acérrimos estadístico o en su defecto ingenieros de software y generaliza expresando que un científico de datos es aquel que extrae información de datos desordenados.

**Figura 5 Skills, y herramientas para la ciencia de datos**



Fuente: Tomado de (Conway, 2010)

(Nield T. , 2022), argumenta que tanto la matemática como la estadística captan el interés general entorno a la alta disponibilidad de datos y la necesidad de comprender cual es la lógica detrás de las herramientas de aprendizaje automático desde una analogía del uso sin una comprensión real de lo que se esconde o aloja detrás de estos algoritmos.

En términos de (VanderPlas, 2022) este considera a la ciencia de datos como una disciplina interdisciplinaria en la que concurren las habilidades (skills) de un estadístico un

---

informático para el diseño de algoritmos y la experiencia en el dominio (substantive expertise), que en otras palabras correspondería a la formación base de un científico de datos capaz de integrar su experticia temática. Siendo esta la raíz de investigación en la que se han identificado interdisciplinaridades temáticas ofertadas por INGISOT SAS en su ejercicio de consultoría entre ellas la ambiental, sistemas de información geográfica ordenamiento del territorio entre otras.

En un recuento histórico de (Nathn, 2021) en su libro "Practical data science with python" señala que con el auge de ofertas universitarias en maestrías y licenciaturas en ciencia de datos, la llegada de Kaggle<sup>1</sup> en la década de 2010 y el crecimiento amplio de librerías asociadas a la ciencia de datos y de lenguaje profundo (Deep learning) se impone un ritmo acelerado de popularidad que continua en la década de 2020, y el arribo de servicios en la nube para crear, entrenar e implementar modelos aprendizaje profundo como Amazon SageMaker Studio, Auto ML & Azure así como el crecimiento de ofertas laborales asociadas a la ciencia de datos.

#### **5.1.2. Ciencia de datos espacial**

De forma paralela quizás aislada y que han convergido gracias al código y las librerías usadas en software como QGIS y ArcGIS y Open source, los datos geoespaciales adquieren necesidades de tratamiento desde los modelos vector y las imágenes obtenidas por teledetección y su uso se extiende a razón de la automatización que otorga el código o los servicios en la nube. (Joel, 2023) identificándose así una vertiente denominada por algunos autores como spatial data science o ciencia de datos espaciales.

#### **5.1.3. Flujos de trabajo en ciencia de datos**

En el contexto de la ciencia de datos aplicada a entornos profesionales, un flujo de trabajo puede entenderse como un conjunto de actividades organizadas, interdependientes y orientadas a extraer conocimiento útil a partir de datos. Estas actividades, que incluyen etapas como formulación de problemas, adquisición de datos, procesamiento, modelado, evaluación e interpretación, que configuran un proceso iterativo y no lineal. Este tipo de estructuras metodológicas similares a los modelos CRISP-DM o TDSP han sido ampliamente utilizadas en dominios como el turismo (Egger & Yu, 2022), ciencias sociales (Hazzan & Mike, 2023), medicina (Shanmuga & Paulchamy, 2023) entre otras donde la

---

<sup>1</sup> Comunidad web en ciencia de datos.

---

especificidad temática exige adaptar las herramientas computacionales a las necesidades particulares de cada campo.

La semi automatización de los flujos de trabajo corresponde a desarrollos parciales en etapas previas a la producción dentro de scripts locales de autoservicio y ad hoc o alojados en repositorios web o mejor llamados controles de versiones como GitHub o el tratamiento de datos en servicios web de respaldo de máquinas virtuales o servidores de Azure o Amazon.

Song y Zhu (2016) argumentan que la capacidad de estructurar flujos de trabajo reproducibles, automatizables y escalables es un rasgo distintivo de la ciencia de datos moderna, En este sentido, la evolución desde procedimientos manuales hacia flujos programables ya sean visuales o codificados no solo incrementa la eficiencia operativa, sino que posiciona a las organizaciones en un nivel superior de madurez tecnológica. Para empresas como INGISOT SAS, la implementación de estos modelos no constituye únicamente una mejora técnica, sino una **estrategia de diferenciación competitiva** en el ámbito de la consultoría ambiental.

## 5.2. Modelos en la ciencia de datos

En datos estructurados “bases de datos” el termino genérico modelo se asocia a lo propuesto por Peter Chen en 1976 como el modelo entidad-relación, en el que se incluyen el modelo conceptual, lógico y físico (Oppel, 2009). Entorno a la forma de uso o lenguaje de consulta el más popular en bases relacionales corresponde a SQL de tipo declarativo en el que se especifica el patrón de datos deseado, la tabla o tablas de consulta, condicionales y transformaciones. (Nield T. , 2016)

No obstante, los modelos en ciencia de datos se abordan desde el proceso de la **minería de datos** diferenciando la extracción de datos para construcción de modelos o el uso de los resultados para predecir nuevos datos utilizando algoritmos de clasificación o regresión (Provost & Fawcett, 2013). En el rigor espacial este proceso se ha denominado minería de datos geoespaciales buscando cuantificar o descubrir patrones desconocidos a partir de datos geoespaciales tales como (puntos críticos, co - ubicaciones, valores atípicos espaciales, predicción y cambio espaciales) (Li, Xie, & Sheaker, 2023). En un término más simple como lo indica (Witten, Frank, & et.al, 2016, pág. 5) “la minería de datos consiste en resolver problemas mediante el análisis de datos ya presentes en las bases de datos”.

---

### 5.2.1. Modelos descriptivos

Los modelos descriptivos se centran en comprender y **explicar** fenómenos existentes a través del análisis de datos históricos. Su objetivo principal es identificar patrones, tendencias y relaciones que permitan entender "qué ha ocurrido" y "por qué ha ocurrido". (Provost & Fawcett, 2013).

Dentro de lo métodos se hayan la **segmentación** o dicha de otro modo como la clasificación de registros a través de la partición de registros utilizando algoritmos de clusterización como el K-means, y a su vez aplicando técnicas de minería de datos directas e indirectas que permiten descubrir las relaciones entre los datos y su explicación recurriendo a técnicas de **asociación** apoyadas mediante medidas de similitud o distancia numérica. (Collica, 2011). Lo anterior corresponden a alternativas útiles a ser verificadas de uso a través de datos sociales asociados a la población en términos de edad, altura, sexo, educación, etc del que tratan los requerimientos de consultoría en la atención de los estudios de impacto ambiental medio socioeconómico.

### 5.2.2. Modelos predictivos

Los modelos predictivos van más allá de la descripción, ya que permiten anticipar valores o categorías futuras basándose en datos históricos. Su enfoque se centra en pronosticar "qué ocurrirá" bajo determinadas condiciones (Larose & Larose, 2019). A esto se integran los métodos de ML y el análisis de series temporales.

Agrupados en tres categorías los **métodos** de machine learnig ML se utilizan para extraer patrones o predicción de eventos desde los datos, en la primera categoría del **tipo supervisado** esta se subdivide en los subgrupos de clasificación y regresión los datos son del tipo categóricos de clasificaciones binarias, multiclase o multi etiqueta, en regresiones se predicen valores numéricos continuos, mientras en que el **aprendizaje no supervisado** se obtienen patrones, mediante técnicas de agrupaciones o reducciones y el **aprendizaje de refuerzo** (George, 2021)

Por otra parte un campo amplio de los métodos predictivos corresponde al análisis de series temporales en pronóstico de datos simples es decir que se acompaña la columna de tiempo "fecha" y el valor en dichos momentos para los cuales a este tipo de análisis le corresponden los modelos de media móvil (q) o el modelo autorregresivo (p) como entrada al análisis de modelos de tendencias y patrones como ARMA (p,q), ARIMA (p,d,q)), SARIMA (p,d,q) (P,D,Q), & SARIMAX en el que se incluyen variables externas de pronóstico o el método VAR(p) para predecir múltiples series temporales. (Peixeiro, 2022). Este tema

---

corresponde en rigor a lo requerido en análisis climatológicos, hidrología y atmosfera en las que se analiza en valor precipitación, temperatura, humedad, etc dentro de una serie o línea de tiempo los cuales son frecuentemente requeridos en consultoría ambiental de la cual atiende INGISOT SAS.

### **5.2.3. Modelos prescriptivos**

Los modelos prescriptivos van un paso más allá de los predictivos al recomendar acciones específicas basadas en predicciones. Estos modelos responden a la pregunta "¿qué deberíamos hacer?", teniendo en cuenta múltiples escenarios y restricciones (Nathn, 2021). allí se incluyen los resultados de la modelación predictiva a través de la inteligencia de decisiones DI, sus aplicaciones son múltiples tales como: negocios, medicina, gobierno, finanzas, marketing, fabricación, transporte, agricultura (determinación del momento óptimo de siembra), energía (determinación de las fuentes de energía más rentable). (Kulkarni, Shivananda, & et al., 2023) un ejemplo homologo aplicado a consultoría correspondería a las zonificaciones ambientales que en base a los resultados se recomiendan acciones de intervención.

### **5.3. Metodologías para minería de datos y ciencia de datos**

Diversos autores en el ámbito de la ciencia de datos describen metodologías estructuradas como **CRISP-DM, SEMMA y TDSP** para guiar la construcción de modelos de datos del tipo predictivo y descriptivo donde se emplean técnicas procesos estadísticos y minería de datos , (Vikas, Sanjeet, & Monica, 2004), (Provost & Fawcett, 2013) Estas metodologías permiten establecer un flujo lógico y reproducible desde la comprensión del problema hasta la evaluación del modelo. En línea con esta evolución, (Nathn, 2021) destaca cómo CRISP-DM ha sido extendida en entornos modernos mediante la incorporación roles, entornos y servicios web, lo que favorece el despliegue y el mantenimiento colaborativo de modelos en producción. Es decir, se destaca la inclusión tecnológica y la identificación del personal necesario u óptimo para el desarrollo de los flujos de trabajo a ser optimizados actividad necesaria a ser realizada en INGISOT SAS.

#### **5.3.1. CRISP-DM**

La metodología más usada para abordar el proceso corresponde al proceso estándar intersectorial para la minería de datos (CRISP-DM), propuesto por (Shearer, 2000), La metodología CRISP – DM Consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación dentro de un

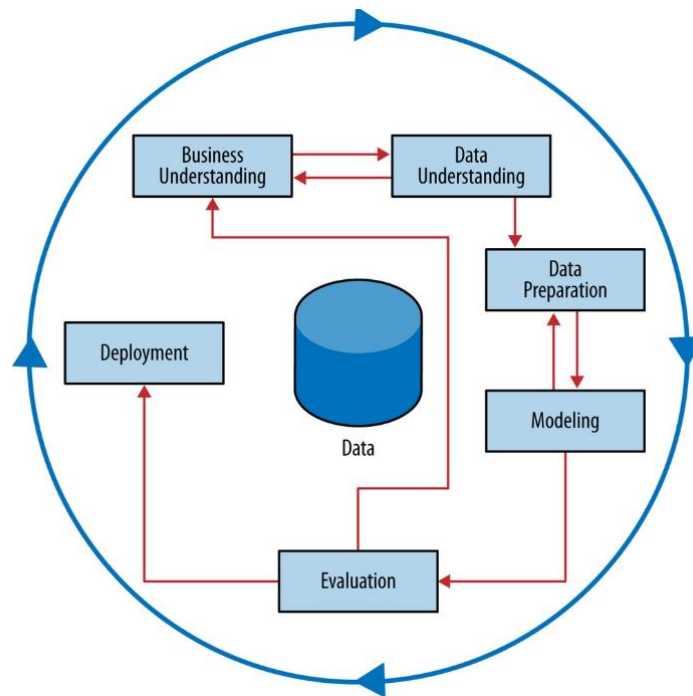
proceso iterativo dentro del cual cada fase es revisada lo que plantea un escenario flexible y adaptable a los objetivos y alcances de cada problema industrial o en nuestro caso de estudio la consultoría ambiental.

**Tabla 1 Etapas de la metodología CRISP - DM**

<b>Fase</b>	<b>Característica</b>
Comprensión del negocio	Comprensión del problema y definición de objetivos trabajando con el interesado, identificando las fuentes de datos, análisis de procesos, entrevistas y recopilación de la documentación requerida.
Comprensión de los datos	Se identifican problemas, anomalías, tipos, calidad e integridad de los datos, se exploran variables y posibles relaciones entre estas guiando así la técnica de modelado más adecuada a fin de responder a los objetivos planteados en la comprensión del negocio.
Preparación de datos	Durante esta etapa los datos son explorados, transformados y depurados, llevando a cabo el tratamiento de datos ausentes, atípicos en aplicación de técnicas de escalado, normalización y la codificación de variables categóricas. La relevancia del proceso radica en que la precisión y eficacia del modelo dependerán de la calidad de los datos.
Modelado	Se alude al uso de técnicas y algoritmos a fin de resolver el problema planteado en uso de los datos anteriormente preparados a fin de desarrollar un modelo de tipo predictivo o descriptivo.
Evaluación	Dentro de esta fase como su nombre lo indica se prueba el modelo con la inclusión de nuevos datos a fin de evaluar su presión y eficacia, identificando la necesidad de realizar ajustes o refinado del modelo.
Implementación	Aquí el modelo es implementado en un entorno de producción en constante monitoreo de resultados. Apoyado de la documentación necesaria de respaldo en la que se incluyan guías y la integración al proceso de negocio. Desde allí se continua con la retroalimentación y refinado de modelo.

Fuente: Adaptado de (Vikas, Sanjeet, & Monica, 2004)

**Figura 5 El proceso de minería de datos CRISP-DM**

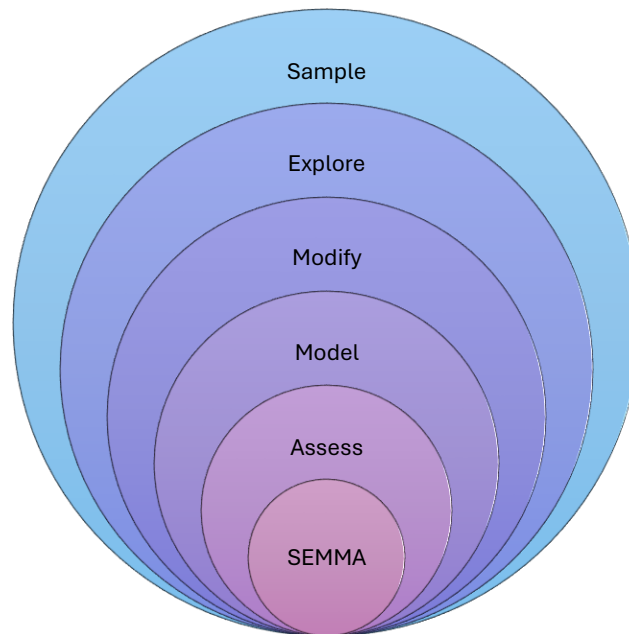


Fuente: (Provost & Fawcett, 2013)

### 5.3.2. SEMMA (Muestrear, explorar, modificar, modelar y evaluar)

SEMMA; Metodología para el análisis y modelado de tipo predictivo desarrollada por SAS Instituto Inc., en la cual se desarrollan 5 fases, partiendo de la sección de una muestra (SAMPLE) representativa de datos en uso de diversas técnicas de muestreo de tipo aleatorio o estratificado. Explorando (EXPLORE) mediante visualización la presencia de patrones estadísticos, modificando, transformando, limpiando o suprimiendo datos atípicos previo a la aplicación de una técnica de modelado (regresión lineal, logística, arboles de decisión y redes neuronales u otras) en contraste con métricas de rendimiento del modelo y terminando en un proceso de evaluación (ASSES) en las que se incluyen validaciones del tipo cruzada, reserva o Bootstrap. (Vikas, Sanjeet, & Monica, 2004). Estos pasos requieren la identificación dentro de los flujos de trabajo en términos de los propuesto en la metodología SEMMA para INGISOT SAS.

**Figura 6 El proceso de la minería de datos en la metodología SEMMA**

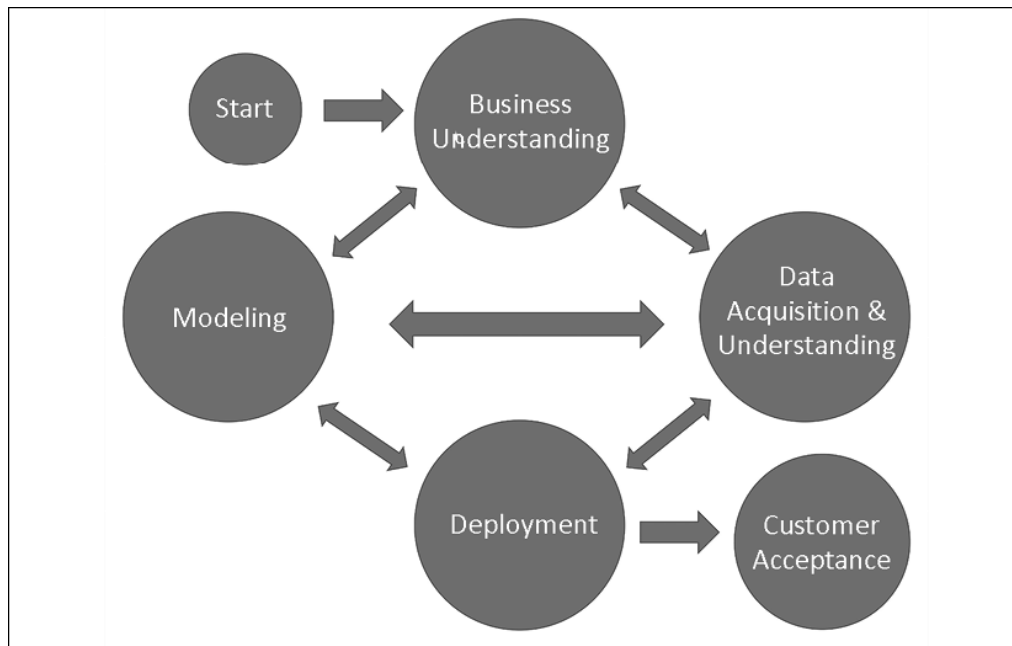


Fuente: (Vikas, Sanjeet, & Monica, 2004)

### 5.3.3. TPSP (Proceso de ciencia de datos en equipo)

Desarrollado y lanzado por Microsoft en el 2016, erige como una propuesta mejorada del CRISP-DM, dentro del cual se integran roles, servicios y accesos web como el controlador de versiones de GitHub dentro de un enfoque moderado a través del uso de software e IA (Nathn, 2021). (Microsoft, 2025) Integra alternativas para evaluar el nivel de la madurez en IA, habilidades, acceso, casos de uso, conceptos e implementación de una IA responsable.

**Figura 7 Reproducción del diagrama de flujo del proceso TDPS**



Fuente: (Nathn, 2021)

## 5.4. Herramientas para la ciencia de datos

### 5.4.1. Lenguajes para la ciencia de datos Python, R & SQL

En lo referente a las herramientas utilizadas en ciencia de datos, existe un consenso generalizado en torno al uso de lenguajes de programación como Python y R. Diversos estudios, tanto a nivel global como aquellos centrados en el mercado estadounidense, les identifican como las principales herramientas de código abierto para el análisis de datos (Larose & Larose, 2019).

**Python** se caracteriza por una estructura modular y una sintaxis clara, que facilita su adopción en el ámbito de la ciencia de datos. Entre sus principales componentes destacan las estructuras de datos como conjuntos, tuplas y diccionarios, así como los flujos de control (for, while) y el soporte para la programación orientada a objetos. Además, ofrece funcionalidades avanzadas para la definición de funciones y visualización de datos. Estas características convierten a Python en una herramienta fundamental para quienes se inician en el análisis de datos, como lo describe (Grus, 2019) en su obra *Data Science from Scratch: First Principles with Python*.

---

En el caso de R, se utilizó RStudio para el análisis estadístico y visualización de datos. Para Python, se empleó Jupyter Notebook como entorno principal que permite la ejecución interactiva de código y facilita la documentación del proceso analítico, sin necesidad de instalación local.

**SQL** (lenguaje estructurado de consulta), permite la gestión de bases de datos bien sean de tipo ligero (Microsoft Access, SQLite) o en su defecto bases de datos centralizadas (MySQL, PostgreSQL, Oracle) en configuraciones cliente servidor, en las que se incluye el uso de diversas herramientas de consulta (select, where, group by, order by) así como funciones de agregación (count, sum, max, min, avg), uniones (join) en las que son consultadas bases de datos normalizadas por las relaciones entre campos de bases de datos relacionales, a través de diversas alternativas de consulta (inner join, left join, right join, outer join). (Nield T. , 2016).

En la implementación del modelo para INGISOT SAS, se empleó SQL para la extracción y consulta de datos desde bases de datos relacionales (PostgreSQL), facilitando la selección y limpieza de registros específicos. Python se utilizó posteriormente para la transformación, análisis exploratorio y modelado de datos, aprovechando sus librerías especializadas (pandas, geopandas) para mayor flexibilidad en el procesamiento. Esta combinación permite aprovechar la eficiencia de SQL en consultas complejas con la versatilidad de Python para análisis estadísticos y visualización avanzada.

Ahora bien la relación expresa entre SQL y la ciencia de datos se basa así en la extracción y separación de datos a fin de responder preguntas directas en las consultas (informes ad hoc) o en su defecto de ingresar los datos a modelos predictivos, creación de consultas para informes, análisis exploratorio (EDA) y la creación de consultas avanzadas que recurre a funciones de anidación o unión exportables a formatos de consulta CSV e importables a software de inteligencia empresarial (BI) como tableau, y mediante script en Python para realizar así la transferencia de datos procesados desde un dataframe de pandas a la base de datos alojada en un servicio web. (Teate, 2021). Características que describen en detalle la forma o data que se procesa en INGISOT SAS en el cual se realiza descarga en servidores web o repositorios de información geoespacial a fin de dar inicio al

---

tratamiento y exploración de datos y un posterior ingreso a modelos bien sea del tipo descriptivo o predictivo y por qué no prescriptivo.

#### 5.4.2. librerías o paquetes de programación

El ecosistema de Python para ciencia de datos se complementa con paquetes o librerías especializadas que proporcionan funciones y estructuras optimizadas para el análisis, modelado y visualización de datos. En un rigor técnico (Nelli, 2018) señala que estas conforman el grupo sciPy siendo estas **Numpy** (Python numérico) en el que se incluyen arreglos matriciales, vectores ndarray. **Pandas** que proporciona estructuras de datos complejas como el Dataframe el cual corresponde a datos tabulares bidimensionales con etiquetas de fila y columna y **Matplotlib** que en síntesis aporta la parte gráfica en formato 2D. & **seaborn** como una derivación para la representación de elementos estadísticos. y **Seaborn**, una extensión basada en Matplotlib orientada a la visualización estadística avanzada.

No obstante, y sin dejar atrás el enfoque espacial objeto del interés por la experticia temática de INGISOT a este grupo le correspondería ingresar a **Geopandas**. útil para el manejo de datos vectoriales basada en las potencialidades y bondades de pandas en la que se incluye el objeto de datos *GeoDataframe* en adición de una columna de geometría al *dataframe* de pandas. Y **rasterio** para el manejo de datos raster que analógicamente corresponde a un arreglo matricial que entiende Numpy. (Toms, Van Rees, & Crickard, 2019).

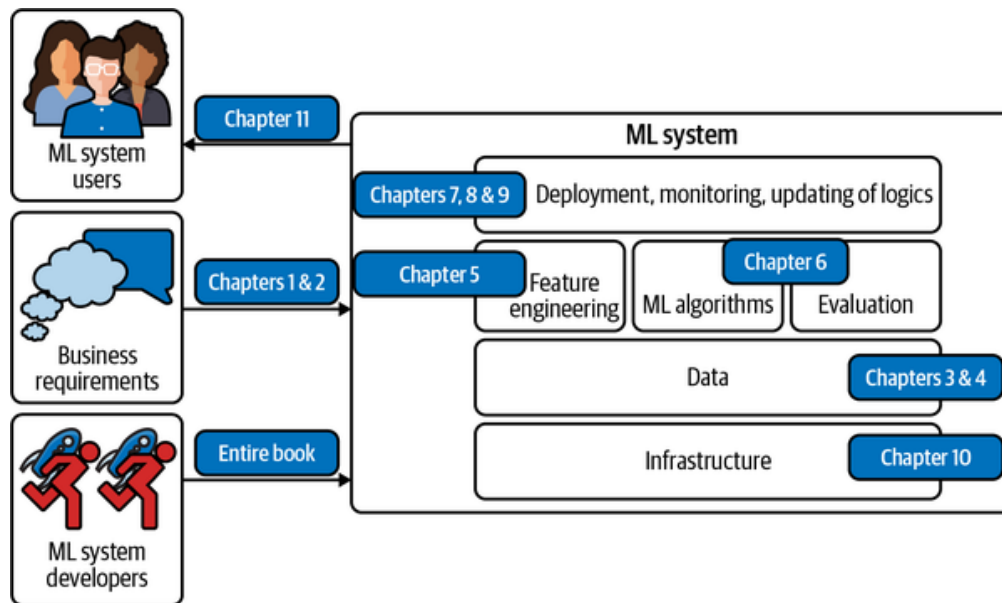
Entorno al software comercial es indispensable adentrarse en las bondades del módulo **Arcpy** el cual consta de funciones, módulos (arcpy.as – análisis espacial, arcpy.da – acceso de datos, arcpy.mp – mapeo, arcpy.geocoding – geocodificación y arcpy.na – analista de redes) y clases adicionales a fin de contribuir con la automatización de elementos geoespaciales. (Toms & Parker, 2022)

#### 5.4.3. Frameworks como herramienta estructural de ML

Cuando nos adentramos en modelos de predicción es imposible dejar de lado el ecosistema del machine learning (ML) e identificar que este “consta de numerosos componentes (algoritmos de ML, datos, lógicas de negocio, métricas de evaluación, e infraestructura subyacente). (Huyen, 2022)

En aprendizaje automático, los frameworks se entienden como conjuntos de herramientas modulares que facilitan el desarrollo, entrenamiento y despliegue de modelos. Estas estructuras suelen incorporar componentes para la construcción de grafos computacionales, la diferenciación automática y la ejecución de rutinas de optimización (Géron, 2022).

**Figura 8** Diferentes componentes de un sistema de aprendizaje profundo.



Fuente (Huyen, 2022)

Scikit-learn es una biblioteca para la implementación de algoritmos de aprendizaje automático, desarrollada sobre NumPy, SciPy y Matplotlib. Está enfocada principalmente en el modelado de datos, en lugar de en su carga, manipulación o visualización (Paper, 2019). Entre sus funcionalidades se destacan la transformación de datos, la automatización de pipelines y un diseño modular, lo que permite construir flujos de trabajo replicables (Quantum Technologies LLC, 2025).

TensorFlow framework creado por Google, en la que se incluyen algoritmos de redes neuronales profundas como el transformer, para cálculos matemáticos complejos de aprendizaje profundo con diversas funcionalidades dentro de las que se encuentran: el desarrollo de modelos, monitoreo, depuración y modelo de servicio, acompañado de un API para el acceso a datos (tensorflow-datasets), y el generador de datos o submódulo keras capaz de realizar la carga de datos específicos (imágenes, series temporales ) de diversas

---

fuentes de información. (Ganegedara, 2022). En tal sentido comprender la estructura, bondades y herramientas será necesario a fin de evaluar la inclusión dentro de los flujos de trabajo en consultoría teniendo siempre presente la experticia temática necesaria para su optimización.

#### **5.4.4. Pipelines como técnica estructural**

En Scikit-Learn, un pipeline (canalizador o flujo de procesamiento) corresponde a una herramienta que permite ejecutar de manera secuencial una serie de transformaciones sobre los datos, tanto de entrenamiento como de prueba. Estas transformaciones pueden incluir tareas como el escalado de variables, la codificación de variables categóricas, la imputación de valores faltantes, la reducción de la dimensionalidad y la ingeniería de características. El uso de pipelines contribuye a mejorar la eficiencia del flujo de trabajo, prevenir la fuga de datos (data leakage) y favorecer la legibilidad y mantenibilidad del código (Quantum Technologies LLC, 2025).

Con la inclusión de Action en Github las capacidades de integración/ entrega continua (CI/CD) en desarrollos o Develop se integran alternativas de canalización (pipelines) que gracias a las estructuras para ejecutar flujos de automatización definen un flujo de trabajo (workflow) en iniciación de trabajos a través de etapas de un proceso entre los que se tienen: despliegue (contenedores), seguridad, integración continua, automatización y paginado. (Laster, 2023)

#### **5.4.5. Soluciones de business Intelligence BI para la ciencia de datos**

Dentro del ámbito empresarial, la síntesis de resultados a través de indicadores es preponderantemente necesaria en presentación de analítica avanzada mediante reportes y es allí donde se entra POWER BI, como una herramienta de visualización interactiva que se soporta gracias modelos de datos, trabajados o estructurados con Power Query y DAX, así como la representación clásica de grafos (columnas, barras, círculos añillos, líneas de tiempo, tablas e inclusión de tarjetas o indicadores) con similitudes funcionales a las tablas dinámicas de Excel desde el panel de datos, a fin de presentar o mostrar aspectos categóricos, dinámicos de estructura o tablas y matrices de estructura combinada. (Kolokolov & Zelensky, 2024)

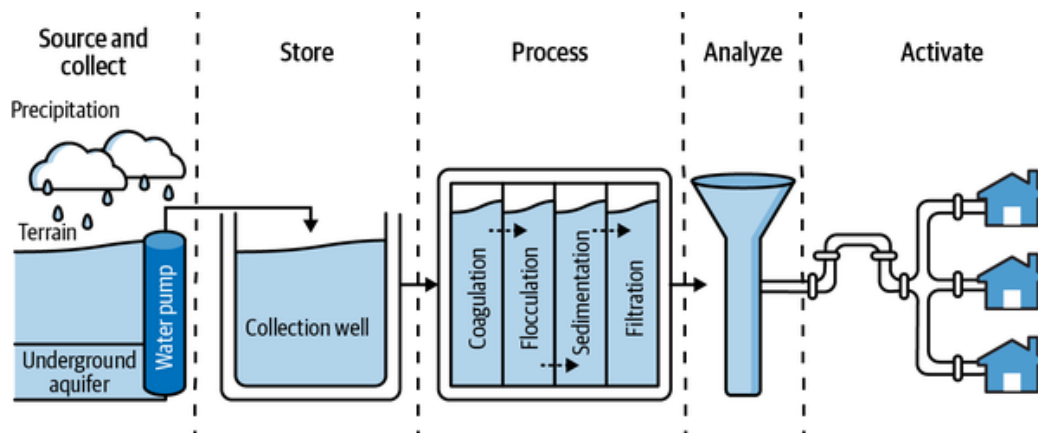
Para el tratamiento de datos en Power BI, se ara necesario contar o estructurar un modelo de datos en el cual se realice la identificación de entidades que se acompañan de

atributos “la data de consulta” estructuradas en tablas interrelacionadas producto de la cardinalidad creada entre las primary key o llaves principales y el foran key o llave foránea, consultadas mediante operaciones de conjunto como; unión, intersección, excepción, en alusión a datos estructurados relacionados que cuentan con diversos niveles de normalización o el uso de modelos dimensionales optimizados para el análisis mediante un proceso de tipo ETL (extracción, transformación y carga) realizado desde Power Query tomados del almacén de datos (base relacional o un data lake o delta lake). (Ehrenmueller, 2022)

### 5.5. Despliegue y sostenibilidad del modelo

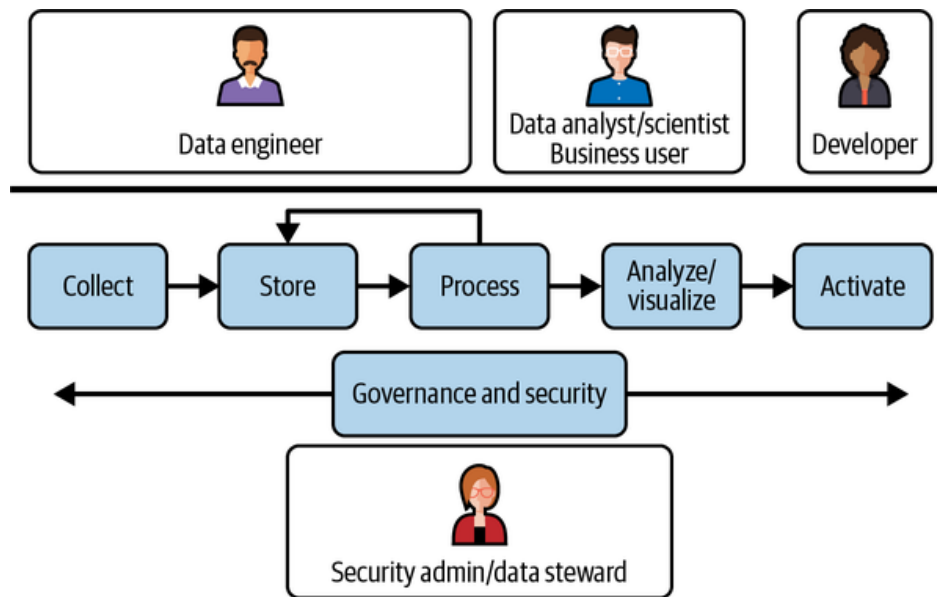
El despliegue dentro del ciclo de vida de los datos corresponde así a fase de activación luego de recopilar, almacenar, procesar y visualizar, que en analogía correspondería a la entrega de un sistema de agua, en el que se requiere de la precipitación “datos” y almacenado o ingesta en términos del volumen, velocidad y variedad, un sistema de recolección “infraestructura”, el procesamiento realizado por el plomero o “científico de datos”, en análisis del agua y por último la entrega mediante un sistema o red de acueducto – herramientas web (AWS, Google Cloud, Azure) que incluyen incluso la capacidad de manejar grandes modelos de lenguaje (LLM). (Tranquilin, Lakshmanan, & Tekiner, 2023)

**Figura 9** Ciclo de vida del agua proporcionando una analogía para los cinco pasos del ciclo de vida de los datos.



Fuente: Tranquilin, Lakshmanan, & Tekiner, 2023. *Water lifecycle, providing an analogy for the five steps in the data lifecycle*

Figura 10 Ciclo de vida del dato simplificado



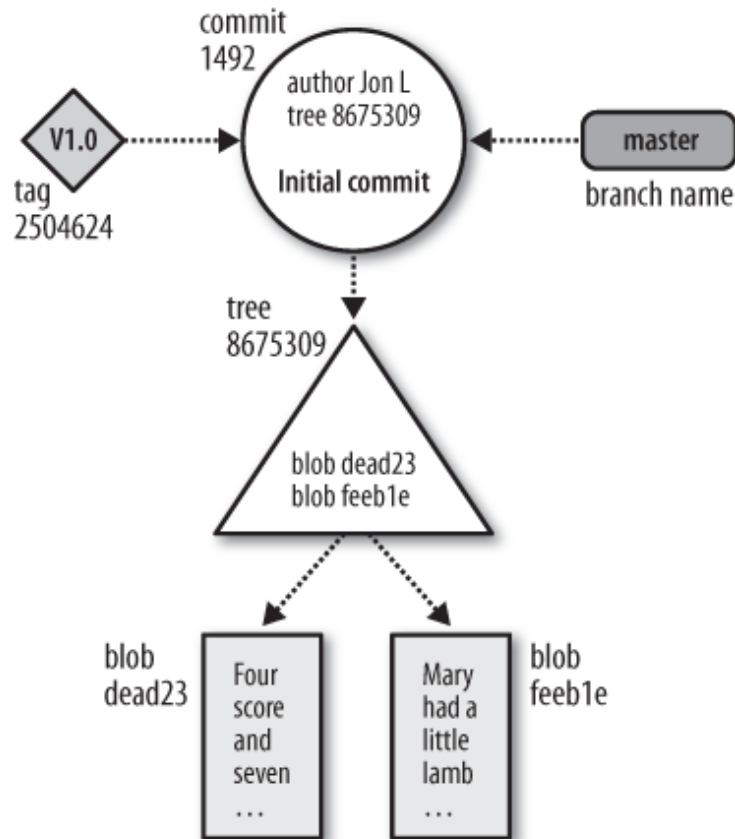
Fuente: Tranquilin, Lakshmanan, & Tekiner, 2023. *Simplified data lifecycle*

### 5.5.1. Herramientas de MLOps para el control, empaquetado y gestión de modelos en producción.

El sistema de control de versión (VCS) en el marco de la generación de modelos de datos para ciencia de datos como GIT, parte de la creación de un repositorio o la clonación de un espacio de trabajo a través de la iniciación git init ya sea que este se encuentre vacío o con diversos archivos (scripts, notebook, py. etc). la inclusión de nuevos archivos es simple desde un git add y se confirma mediante el commit, su estructura contiene *blobs* para referenciar variables o archivos, *tree* o árboles para representar la identificaciones del blob, el *commit* o confirmación de cambios guarda la metadata de cada cambio en el que se incluyen autor, fecha y registro y las *tags* (etiquetas) designada para objetos específicos dentro del repositorio y por último el índice que capta una versión del proyecto en un momento dado hasta que se realice la confirmación de los cambios del contenido de los archivos mas no del nombre del proyecto en base al identificador único o global hash

indiferente de su localización o directorio así este se localice en diversos equipos. (Loeliger, 2022)

**Figura 11 Objetos de un repositorio GIT**



Fuente: Loeliger, 2022. Objetos de un repositorio Git

El *repo* o versión controlada de un proyecto ofrece alternativas de repositorios locales o remotos alojado en un servicio de hosting como (github, gitlab y gitbucket) a través de transferencia local hacia los servicios de alojamiento usando la carga o el *push* o de manera inversa creando el repositorio remoto y posteriormente la copia en el *repo* local, el repositorio remoto permite la contribución compartida, así como facilidades de acceso y respaldo en la nube de acceso público o privado. (Skoulikari, 2023).

Lo anterior se presenta, así como una herramienta versátil y útil a fin de contribuir con las etapas de desarrollo y pruebas del modelo en sus fases de notebooks y modulación de los flujos de trabajo a ser optimizados al interior de INGISOT SAS, a lo que (Gandhi, 2022) recalca que “todo proyecto de software comienza con una idea implementada en código

---

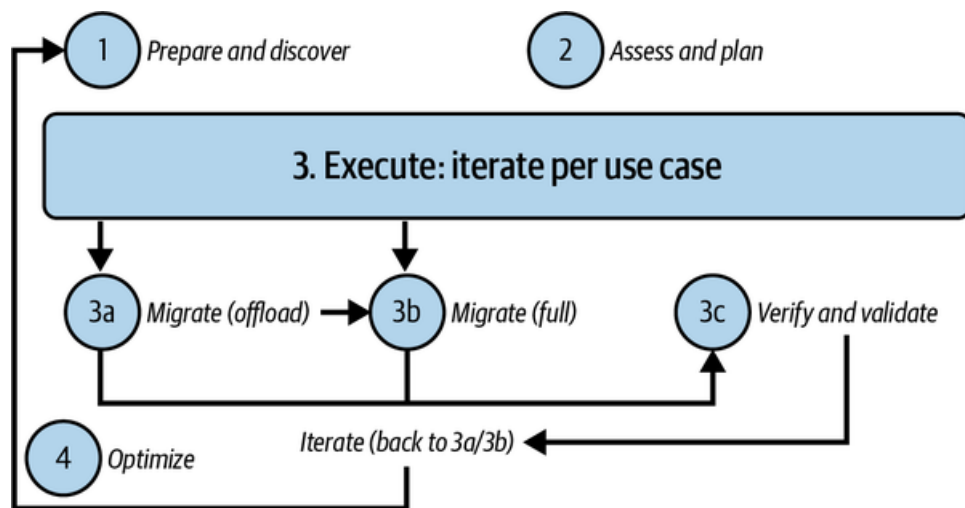
fuente”. Antes de evaluar las necesidades de despliegue en Docker o Mlops desde interfaz web como Amazon, Azure o google cloud.

En términos de (Schenker, 2018) el Docker o contenedor aloja y empaqueta las aplicaciones y dependencias (frameworks y bibliotecas) para su despliegue que en sus inicios se realizaba mediante máquinas virtuales en subutilización y un gran despliegue de capacidades que en analogía presentada este corresponde a un gigantesco buque o barco para transportar un único contenedor, lo que representa una gran pérdida de recursos. La solución analógica correspondió así en crear un Docker o paquete capas de “construir, enviar y ejecutar en cualquier lugar” (Schenker, 2018) los recursos de aplicación escrita es decir el código junto con todas las dependencias necesarias para la ejecución siendo ejemplo de ellos kubernetes, Apache Mesos, Rancher y Docker entre muchos otros.

### 5.5.2. Factores organizacionales

Cuando de modernizar los flujos de trabajo se trata es necesario realizar una mirada holística a los componentes o ingredientes necesarios para el despliegue o producción de modelos en los cuales se evalúan los resultados a fin de identificar pro y contras dentro del negocio, el personal o partes interesadas capaces de construir los modelos en búsqueda de adquirir suficiencia técnica, así como de evaluar la implementación en arquitecturas de datos mediante tecnología en que se recalca la construcción o miradas a los flujos de trabajo, a fin de que este se realice de formas más sencillas y por ende económicas apoyándose en marcos de migración, productos mínimos, gobernanza y estructura de los datos (políticas, catálogos, clases de datos entre otros (Tranquilin, Lakshmanan, & Tekiner, 2023), aspectos A ser evaluados e identificados dentro de los flujos de trabajo a ser optimizados en el modelo de datos propuesto para la implementación en INGISOT SAS.

### **Figura 12 Marco de migración**



Fuente: (Tranquilin, Lakshmanan, & Tekiner, 2023)

### 5.5.3. Gobernanza de datos

La gobernanza de datos del modelo desarrollado para INGISOT SAS asigna al personal técnico la responsabilidad sobre la gestión y calidad de los datos, los cuales se almacenan en una base de datos PostgreSQL con extensión PostGIS ubicada en un servidor local con un respaldo automático. Se implementan políticas de seguridad basadas en roles, cifrado de datos en tránsito (HTTPS/TLS) y en reposo (AES-256) conforme a la normativa colombiana, y se mantiene un catálogo centralizado que documenta metadatos, linaje, calidad y restricciones de uso para cada dataset.

En síntesis, el marco de referencia consolidó los fundamentos conceptuales y metodológicos que sustentan el diseño del modelo, garantizando que este se apoye en estándares internacionales y buenas prácticas de la ciencia de datos aplicadas al sector ambiental.

---

## 6. Diseño metodológico

### 6.1. Tipo de investigación.

La presente investigación corresponde a un enfoque de investigación aplicada, ya que se fundamenta en teorías, metodologías y herramientas consolidadas en el ámbito de la ciencia de datos, como CRISP-DM, Python, R y arquitecturas de almacenamiento tipo data lake con el propósito de desarrollar un modelo de datos funcional y contextualizado.

En este sentido, se implementó CRISP-DM como marco metodológico estructurando el proyecto en seis fases: (1) comprensión del negocio mediante el diagnóstico de flujos en INGISOT, (2) comprensión de datos a través de la caracterización de bases de datos existentes, (3) preparación de datos con limpieza y normalización utilizando Python y librerías especializadas como pandas para manipulación de dataframes, geopandas para procesamiento de datos espaciales, scikit-learn para modelado predictivo y matplotlib para visualización, (4) modelado mediante algoritmos de clustering y predictivos, (5) evaluación con validación de métricas de desempeño, y (6) despliegue con recomendaciones de implementación.

Complementariamente, se empleó R para análisis estadísticos descriptivos y se propusieron arquitecturas de almacenamiento tipo data lake como infraestructura para centralizar los datos de fuentes específicas de INGISOT, conformando así un producto analítico evaluable y adoptable en escenarios reales.

A diferencia de la investigación básica, la investigación aplicada tiene como objetivo resolver problemas concretos en contextos específicos, utilizando el conocimiento disponible para intervenir sobre la realidad (Hernández, Fernández, & Baptista, 2014) En este sentido, el estudio busca la implementación de un producto analítico que pueda ser evaluado y adoptado en escenarios reales.

Igualmente puede clasificarse como descriptiva, ya que se orienta a identificar, organizar y caracterizar los elementos teóricos y técnicos que conforman el proceso de modelado en ciencia de datos. Según (Hernández S. , 2022) este tipo de estudios busca detallar cómo es y cómo se manifiesta un fenómeno determinado, sin manipular variables,

---

sino describiendo sus componentes, relaciones y dinámicas internas. En el caso del presente trabajo, el énfasis recae en el análisis sistemático de metodologías como CRISP-DM y TDSP, lenguajes de programación como Python y R, y arquitecturas como los data lakes, lo cual permite establecer una visión estructurada del proceso analítico sin intervenir directamente en él. De esta forma, se contribuye al conocimiento contextual a fin de diseñar o desarrollara aplicaciones futuras.

El enfoque metodológico de la presente investigación es mixto, ya que articula elementos tanto cualitativos como cuantitativos. Por un lado, se realiza una descripción sistemática de enfoques, herramientas y marcos metodológicos propios de la ciencia de datos (componente cualitativo); y por otro, se contempla la implementación de modelos y estructuras cuantificables mediante programación y análisis estadístico (componente cuantitativo). Según Hernández Sampieri (2022), el enfoque mixto permite una comprensión más amplia de los fenómenos al integrar métodos de ambos paradigmas, lo cual resulta pertinente en investigaciones aplicadas que requieren tanto interpretación conceptual como validación empírica.

El presente estudio adopta un razonamiento inductivo, en tanto parte del análisis de teorías, metodologías y herramientas generales del campo de la ciencia de datos como los marcos metodológicos CRISP-DM y TDSP, lenguajes de programación y arquitecturas Mlops y controles de versión como GIT para luego proyectar su aplicación en el diseño de un modelo específico. Según Sampieri (2022), el enfoque inductivo se caracteriza por inferir conclusiones particulares a partir de observaciones o principios generales, permitiendo aplicar conocimientos previos a contextos delimitados. Esta lógica se alinea con el propósito de la investigación, que busca estructurar una propuesta técnica concreta a partir del estudio de referentes ampliamente aceptados.

En cuanto a la temporalidad, esta investigación es de corte transversal y duración limitada, ya que se desarrolla en un periodo estimado de seis meses y se concentra en la etapa de diagnóstico, fundamentación teórica y diseño del modelo de datos. No contempla la implementación operativa ni la validación longitudinal del modelo propuesto, sino que establece una ruta de trabajo técnica y metodológica que podrá ser aplicada en futuras fases o investigaciones. Esta característica es propia de estudios que buscan estructurar

---

propuestas desde el análisis de referentes existentes y el contexto actual del problema (Hernández Sampieri, 2022).

## **6.2. Análisis externo - PESTEL**

Se considera que el entorno operativo actual presenta condiciones potencialmente favorables para empresas de consultoría técnica especializadas en ciencia de datos e inteligencia artificial en Colombia. Potencialmente, el marco político-regulatorio colombiano podría desarrollarse en torno a principios de inteligencia artificial ética, con iniciativas como la posible presentación de proyectos de ley sobre IA por parte del Ministerio TIC, aunque esto aún se encuentra en fase de formulación, estableciendo un ecosistema normativo que, aunque demandará adaptación, también legitimará a los actores del sector que cumplan con estos estándares. Simultáneamente, el gobierno colombiano impulsa activamente la transformación digital mediante políticas públicas e iniciativas como la convocatoria 950 "Colombia Inteligente", que no solo demuestra compromiso estatal con estas tecnologías, sino que también crea oportunidades tangibles de contratación y alianzas estratégicas.

En el ámbito económico, las organizaciones colombianas atraviesan un periodo en que la optimización de recursos y eficiencia operativa resultan imperativas, convirtiendo las soluciones basadas en análisis de datos en inversiones estratégicas con alto retorno potencial. Los incentivos económicos específicos para proyectos de IA aplicada representan fuentes adicionales de financiación que pueden catalizar el crecimiento empresarial en este sector. Se observa una creciente disposición del mercado a invertir en consultoría técnica especializada, reconociendo el valor diferencial que aporta la toma de decisiones basada en evidencia analítica frente a aproximaciones tradicionales.

La dimensión social revela una transformación cultural significativa hacia la aceptación y valoración de la inteligencia artificial como herramienta decisoria, facilitando la adopción de servicios consultivos tecnológicos. El mercado laboral colombiano refleja esta tendencia mediante la alta demanda de profesionales con competencias en ciencia de datos, creando tanto oportunidades para formar equipos de élite como desafíos de retención de talento. Paralelamente, crece la expectativa de que estas tecnologías aborden efectivamente problemas sociales relevantes, abriendo nichos de mercado en consultoría de impacto. La

---

persistencia de una brecha digital entre organizaciones con diferente madurez tecnológica genera oportunidades adicionales en segmentos de consultoría básica e intermedia.

Tecnológicamente, el ecosistema colombiano presenta una considerable madurez con herramientas de código abierto y plataformas en la nube que reducen dramáticamente las barreras de entrada para implementar soluciones avanzadas. La proliferación de programas educativos especializados en universidades prestigiosas como Los Andes, Javeriana y EAN garantiza un flujo constante de talento calificado y posibilidades de colaboración academia-empresa. La convergencia entre inteligencia artificial y otras tecnologías emergentes amplía continuamente el espectro de servicios potenciales, mientras que la democratización de tecnologías como IA generativa y autoML redefine constantemente el panorama competitivo, exigiendo innovación continua en las propuestas de valor.

En materia ambiental, se vislumbran oportunidades significativas para fortalecer los servicios de Estudios de Impacto Ambiental mediante modelos predictivos que mejoren la precisión en caracterizaciones físicas, bióticas y sociales. El mercado para consultoría en prevención y adaptación al cambio climático mediante análisis avanzado de datos muestra un crecimiento sostenido, al igual que la demanda por soluciones que optimicen el uso de recursos naturales y sistemas de trazabilidad ambiental basados en datos para cumplimiento normativo. El marco legal complementa este panorama imponiendo requisitos específicos sobre protección de datos personales, propiedad intelectual de soluciones algorítmicas y responsabilidad en sistemas automatizados de decisión, elementos que configuran tanto desafíos como oportunidades de diferenciación para consultoras con conocimiento especializado.

Este análisis integral del entorno macro revela que, lejos de representar un riesgo, el contexto actual constituye una extraordinaria ventana de oportunidad para que empresas de consultoría técnica potenciadas por ciencia de datos consoliden ventajas competitivas sostenibles. La convergencia de políticas proactivas, demanda creciente, madurez tecnológica y disponibilidad de talento especializado crea condiciones ideales para el posicionamiento estratégico en este sector emergente. No obstante, será fundamental mantener capacidad adaptativa frente al rápido desarrollo regulatorio y las expectativas crecientes sobre ética, transparencia y responsabilidad en el uso de tecnologías analíticas

---

avanzadas, factores que determinarán el éxito a largo plazo en este prometedor ecosistema colombiano

### **6.2.1. Cinco fuerzas de Poder**

El análisis competitivo de INGISOT se estructura conforme al modelo de las Cinco Fuerzas de Porter, que identifica los cinco factores competitivos que moldean la rentabilidad y atractivo de un mercado. El poder de negociación de los clientes de INGISOT SAS se posiciona en un nivel moderado-alto dentro del mercado de consultoría ambiental en Colombia. Los principales demandantes de estos servicios empresas energéticas, mineras, constructoras y entidades gubernamentales cuentan con considerable tamaño y capacidad financiera que les otorga significativa influencia en las negociaciones. Sin embargo, este poder se ve parcialmente contrarrestado por los requisitos regulatorios ambientales obligatorios, ya que las empresas deben cumplir con normativas gubernamentales independientemente de sus preferencias de costos. Adicionalmente, los clientes muestran alta sensibilidad a la calidad y precisión de los estudios ambientales debido a los riesgos legales y reputacionales asociados con incumplimientos normativos. Si bien existe relativa facilidad para cambiar entre consultoras ambientales tradicionales, la implementación de soluciones avanzadas de analítica de datos crea dependencias técnicas que reducen gradualmente este poder en relaciones de largo plazo. Los grandes proyectos de infraestructura o extractivos generan relaciones contractuales significativas que tienden a equilibrar el poder negociador entre INGISOT y sus clientes estratégicos.

En cuanto al poder de negociación de los proveedores, INGISOT enfrenta una situación de nivel moderado caracterizada por dinámicas diversificadas según el tipo de recurso. Los proveedores de talento especializado en ciencia de datos con conocimiento sectorial ambiental son particularmente escasos en Colombia, lo que les otorga poder significativo en la cadena de valor. Por su parte, los proveedores de infraestructura tecnológica como servicios cloud y herramientas analíticas mantienen posiciones importantes, aunque su influencia se ve mitigada por la creciente competencia entre plataformas alternativas. Los fabricantes de equipos especializados para monitoreo ambiental ejercen poder moderado debido a la especificidad técnica de sus productos y la precisión requerida en mediciones ambientales. El acceso a datos históricos ambientales y climatológicos constituye otro insumo crítico que depende frecuentemente de instituciones específicas con políticas de

disponibilidad y costos variables. Como contrapeso favorable, la creciente disponibilidad de herramientas de código abierto para análisis de datos reduce la dependencia de proveedores de software propietario, permitiendo a INGISOT mayor flexibilidad en la selección de tecnologías base para su modelo analítico.

La amenaza de productos sustitutos para el servicio especializado de INGISOT se evalúa como baja-moderada, reflejando oportunidades de diferenciación significativas. Las metodologías tradicionales de consultoría ambiental sin componente analítico avanzado representan el sustituto más directo, pero ofrecen menor precisión, eficiencia y capacidad predictiva, limitando su atractivo para proyectos complejos. Las soluciones genéricas de gestión ambiental disponibles en el mercado, especialmente aquellas no adaptadas al contexto regulatorio colombiano, tienen limitada efectividad como sustitutos viables. El desarrollo interno de capacidades analíticas por parte de grandes clientes constituye un sustituto parcial a considerar, aunque requiere inversiones significativas en infraestructura, talento especializado y curva de aprendizaje organizacional. Los servicios de monitoreo automatizado sin interpretación experta y contextualización normativa no sustituyen completamente el valor agregado de la consultoría ambiental integral que combina ambas dimensiones. Las plataformas emergentes de autoservicio para reportes ambientales básicos pueden sustituir componentes elementales del servicio, pero carecen de la profundidad analítica necesaria para abordar análisis complejos de impacto ambiental en proyectos de alta sensibilidad o complejidad técnica.

La amenaza de nuevos competidores en el espacio que busca ocupar INGISOT presenta un nivel moderado, con barreras de entrada significativas, pero no infranqueables. Las barreras técnicas son considerables, requiriendo conocimiento especializado tanto en normativa ambiental colombiana como en ciencia de datos e inteligencia artificial aplicada a problemas ambientales. La credibilidad y trayectoria resultan cruciales en consultoría ambiental, favoreciendo a empresas establecidas como INGISOT frente a nuevos entrantes sin historial verificable. Si bien la inversión inicial en infraestructura tecnológica es moderada, la curva de aprendizaje organizacional para integrar efectivamente capacidades analíticas en procesos de consultoría ambiental es pronunciada y requiere tiempo significativo. Las relaciones establecidas con entidades reguladoras ambientales y clientes corporativos constituyen una barrera relacional relevante, particularmente en un sector

donde la confianza desempeña un papel determinante. La necesidad de referencias verificables en proyectos complejos dificulta la entrada efectiva de nuevos competidores sin historial demostrable. No obstante, existe un riesgo creciente de entrada lateral por parte de consultoras internacionales o firmas tecnológicas que podrían incorporar capacidades analíticas avanzadas a servicios ambientales existentes, aprovechando su experiencia en otros mercados o sectores.

La rivalidad entre competidores existentes en el sector donde opera INGISOT se caracteriza por una intensidad moderada-alta con múltiples dimensiones competitivas en evolución. El sector de consultoría ambiental en Colombia presenta fragmentación moderada con algunos actores dominantes de escala nacional y múltiples firmas especializadas en nichos geográficos o técnicos específicos. La diferenciación entre competidores tradicionales es relativamente limitada en términos de metodologías y entregables, intensificando la competencia basada en precio y relaciones preestablecidas. En este contexto, la incorporación de analítica avanzada representa un vector emergente de diferenciación en un mercado tradicionalmente conservador, permitiendo a INGISOT establecer una posición distintiva. Los costos de cambio para clientes son moderados una vez establecidas relaciones de trabajo y familiaridad con las particularidades del cliente, lo que atenúa parcialmente la rivalidad directa. La creciente importancia de la sostenibilidad y cumplimiento ambiental en agendas corporativas expande el mercado total disponible, reduciendo parcialmente la presión competitiva por participación de mercado. Sin embargo, existe competencia pronunciada por talento especializado que combina conocimiento ambiental con capacidades analíticas, recurso escaso que todas las firmas del sector buscan incorporar para desarrollar ventajas competitivas sostenibles.

### 6.2.2. Matriz DOFA

Fortalezas	Debilidades
<ul style="list-style-type: none"><li>• Conocimiento especializado en normativa ambiental colombiana y procesos regulatorios.</li><li>• Trayectoria y credibilidad establecida en el sector de consultoría ambiental.</li></ul>	<ul style="list-style-type: none"><li>• Curva de aprendizaje organizacional en implementación de soluciones analíticas avanzadas.</li><li>• Limitada disponibilidad interna de especialistas en ciencia de datos con conocimiento sectorial.</li></ul>

<ul style="list-style-type: none"> <li>• Base de datos histórica de proyectos ambientales para entrenamiento de modelos predictivos.</li> <li>• Conocimiento específico de ecosistemas locales y particularidades territoriales.</li> <li>• Relaciones consolidadas con autoridades ambientales y clientes estratégicos.</li> </ul>	<ul style="list-style-type: none"> <li>• Infraestructura tecnológica en proceso de adaptación para procesamiento analítico intensivo.</li> <li>• Dependencia de metodologías tradicionales en algunos procesos clave.</li> <li>• Resistencia interna al cambio en flujos de trabajo establecidos.</li> </ul>
Oportunidades	Amenazas
<ul style="list-style-type: none"> <li>• Creciente exigencia regulatoria ambiental que demanda análisis más precisos y eficientes.</li> <li>• Disponibilidad de datos satelitales y sensores remotos que enriquecen análisis ambiental.</li> <li>• Mayor conciencia corporativa sobre gestión de riesgos ambientales genera demanda.</li> <li>• Potencial de desarrollo de productos analíticos específicos para monitoreo continuo.</li> <li>• Expansión hacia mercados regionales con regulaciones ambientales similares.</li> </ul>	<ul style="list-style-type: none"> <li>• Entrada de competidores tecnológicos con capacidades analíticas nativas.</li> <li>• Evolución acelerada de tecnologías analíticas que requiere actualización constante.</li> <li>• Incertidumbre regulatoria sobre uso de modelos predictivos en trámites ambientales.</li> <li>• Riesgos reputacionales por predicciones incorrectas en modelos ambientales.</li> <li>• Cuestionamientos sobre transparencia y trazabilidad de decisiones algorítmicas</li> </ul>

Fuente: Autores.

### 6.3. Análisis interno

#### 6.3.1. Tema

"Análisis interno de la capacidad organizacional y tecnológica para la implementación de un modelo de analítica de datos en INGISOT SAS"

#### 6.3.2. Objetivo de la encuesta.

Identificar las percepciones del personal técnico y directivo de INGISOT SAS sobre los factores organizacionales, técnicos y operativos que inciden en la adopción de un modelo de analítica de datos para la optimización de los flujos de trabajo.

#### 6.3.3. Variables para medir

- Cultura organizacional hacia la innovación
- Niveles de sistematización actual
- Competencias digitales del personal
- Infraestructura y recursos tecnológicos

- percepción del impacto del modelo
- 

**Tabla 2 Encuesta en escala Likert (formato: Totalmente en desacuerdo a totalmente de acuerdo)**

ID	Pregunta propuesta	Variable asociada
1	En INGISOT SAS se promueve activamente la adopción de nuevas tecnologías.	Cultura organizacional
2	La empresa tiene una visión clara sobre el uso de ciencia de datos en sus procesos.	Cultura organizacional
3	Existe interés por parte de los líderes para implementar soluciones basadas en modelos analíticos.	Cultura organizacional
4	La empresa invierte recursos financieros suficientes en proyectos de innovación tecnológica.	Cultura organizacional
5	Se fomenta el aprendizaje continuo y la actualización tecnológica en el personal.	Cultura organizacional
6	La mayoría de los procesos técnicos en la empresa están sistematizados.	Nivel de sistematización
7	Se utilizan herramientas digitales para documentar y seguir los flujos de trabajo.	Nivel de sistematización
8	Los procedimientos actuales permiten replicar tareas sin mayor intervención manual.	Nivel de sistematización
9	Los flujos de trabajo actuales son susceptibles de mejora y automatización con IA	Nivel de sistematización
10	Se utilizan bases de datos centralizadas para almacenar información de proyectos.	Nivel de sistematización
11	El personal domina lenguajes como Python o R para análisis de datos.	Competencias digitales
12	El equipo tiene experiencia usando librerías como Pandas, Scikit-learn o Geopandas.	Competencias digitales
13	Los colaboradores conocen las fases de metodologías como CRISP-DM o SEMMA.	Competencias digitales
14	Hay interés y disposición para capacitarse en ciencia de datos dentro del equipo.	Competencias digitales
15	El equipo maneja conceptos de estadística avanzada y modelado predictivo.	Competencias digitales
16	La empresa cuenta con equipos y software adecuados para análisis de datos.	Infraestructura y recursos
17	Se dispone de conectividad estable y suficiente para ejecutar herramientas en la nube.	Infraestructura y recursos
18	La infraestructura tecnológica actual puede escalarse fácilmente si se requiere.	Infraestructura y recursos
19	La empresa cuenta con servidores o máquinas virtuales en la nube (cloud computing) que permiten el procesamiento de grandes volúmenes de datos.	Infraestructura y recursos
20	La empresa tiene políticas de seguridad de datos claramente definidas.	Infraestructura y recursos
21	La implementación de un modelo de datos mejoraría la calidad técnica de los entregables.	Percepción del impacto del modelo
22	Automatizar procesos reduciría significativamente los tiempos de entrega.	Percepción del impacto del modelo
23	El uso de ciencia de datos facilitaría el cumplimiento de los términos de referencia.	Percepción del impacto del modelo

24	La adopción del modelo contribuiría a una mayor competitividad de la empresa.	Percepción del impacto del modelo
25	Considero viable integrar modelos analíticos como parte de los procesos permanentes de consultoría.	Percepción del impacto del modelo
1	En INGISOT SAS se promueve activamente la adopción de nuevas tecnologías.	Cultura organizacional

#### 6.3.4. Escala de respuestas sugerida (tipo Likert):

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

#### 6.4. Población, muestra y ficha técnica.

Se construyó un instrumento de medición tipo Likert en forma de encuesta digital, a través de una plataforma como Microsoft Office Forms, con el objetivo de recopilar de manera estructurada y eficiente las percepciones del personal frente a diversas dimensiones del estudio. Este formato permite estandarizar las respuestas, facilitar el análisis cuantitativo de datos subjetivos y garantizar una recolección ágil y accesible, especialmente en contextos organizacionales donde el tiempo y la disponibilidad de los participantes pueden ser limitados.

<https://forms.gle/Coy2QKzjC2NsSqWP7>

**Tabla 3 Tamaño de la muestra, ficha técnica.**

Característica	Descripción
Periodo de recolección de datos	Del 01 de julio al 30 de agosto 2025
Ciudad de aplicación	Bogotá Colombia
Cargo de las personas entrevistadas	Empleados de diferentes áreas con mínimo 6 meses de antigüedad
Población	60 empleados
Muestra	53 empleados
Nivel de confianza	95%
Grado de precisión	5%
Medio de recolección	Encuesta en línea (Google Forms)

Fuente: Autores.

#### 6.4.1. Tipo de muestreo y composición de la muestra

La población objetivo estuvo conformada por 60 colaboradores de INGISOT SAS. La muestra utilizada para el análisis estuvo compuesta por 53 registros, de los cuales 9 correspondieron a respuestas reales obtenidas mediante la aplicación del formulario propuesto y 44 a datos que a través de la simulación de Montecarlo se obtuvieron debido a que por causas del personal y sus diferentes labores dentro de la compañía era difícil coordinar y conseguir nuestra muestra de 53 datos.

Las respuestas reales se obtuvieron mediante un muestreo por conveniencia, en función de la disponibilidad del personal. La simulación de datos no tuvo como propósito realizar inferencias estadísticas, más que todo fue un recurso para complementar el total de datos sin sesgar la muestra basándose en los 9 datos obtenidos realmente.

#### 6.5. Identificación de las variables

En esta investigación se busca medir condiciones clave para la adopción de un modelo de ciencia de datos en consultoría ambiental. Para ello, se definieron variables como la cultura organizacional hacia la innovación, el nivel de sistematización actual, las competencias digitales del personal, la infraestructura tecnológica disponible y la percepción del impacto del modelo. Estas variables permiten identificar la disposición, capacidad técnica y recursos existentes en la organización para implementar soluciones analíticas.

**Tabla 4** Definición de variables en la encuesta likert

Variable	Definición conceptual	Lo que se desea medir
Cultura organizacional hacia la innovación	Grado en que la organización promueve, valora y facilita la adopción de nuevas tecnologías como parte de sus procesos internos.	Disposición del personal y la empresa a implementar tecnologías emergentes.

Nivel de sistematización actual	Nivel de formalización y automatización de los procesos actuales dentro de los flujos de trabajo en consultoría ambiental.	Grado de estructura, repetibilidad y documentación de los procesos internos.
Competencias digitales del personal	Conocimientos, habilidades y disposición del personal en el uso de herramientas, lenguajes y librerías de ciencia de datos.	Capacidades técnicas del equipo humano para adoptar e implementar modelos analíticos.
Infraestructura y recursos tecnológicos	Disponibilidad de recursos físicos, virtuales y de conectividad necesarios para ejecutar un modelo de ciencia de datos.	Capacidad instalada de hardware, software y conectividad en la empresa.
Percepción del impacto del modelo	Opinión del personal sobre el valor agregado que un modelo de analítica de datos puede aportar a la calidad, eficiencia y escalabilidad de la consultoría.	Nivel de aceptación y expectativa sobre los beneficios de implementar el modelo propuesto.

Fuente: Autores

### 6.6. Instrumento de análisis interno

En el marco de esta tesis de maestría en Ciencia de Datos aplicada, se utilizó el coeficiente V de Aiken como método para validar el contenido de los ítems asociados a las variables clave del estudio: cultura organizacional hacia la innovación, nivel de sistematización actual, competencias digitales del personal, infraestructura y recursos tecnológicos, y percepción del impacto del modelo. La fórmula utilizada es:  $V = \frac{\sum S}{[n(c-1)]}$  donde  $\sum S$  es la suma de valores asignados por los expertos,  $n$  es el número de expertos (5) y  $c$  número de categorías de evaluación.

Estas variables fueron operacionalizadas mediante una escala tipo Likert, lo cual requiere asegurar que cada ítem sea claro, pertinente y verdaderamente representativo del constructo que busca medir. Dado que estas dimensiones abarcan tanto aspectos técnicos como percepciones organizacionales, el juicio experto es fundamental para determinar la relevancia conceptual de cada afirmación incluida en el instrumento.

El uso del V de Aiken permitió cuantificar el nivel de acuerdo entre expertos respecto a la validez de cada ítem, asegurando así que el instrumento de medición sea sólido desde el punto de vista metodológico y adecuado para generar datos confiables en el análisis posterior.

**Tabla 5 Expertos temáticos consultados.**

Cargo	Nombre	Perfil profesional
Director área social de INGISOT SAS	Jose Armando Santiago Garnica	Geógrafo, Msc Geografía y Phd en Geografía
Profesional en SIG	German Torres	Ingeniero geólogo, especialista en gestión ambiental y estudiante de especialización en sistemas de información geográfica.
Geotecnista vial	Damian Perez	Ingeniero geólogo, especialista en geotecnia vial
Geólogo Senior	Sebastián Galvis	Ingeniero geólogo, especialista en gestión ambiental, especialista en SIG.
Coordinadora técnico-administrativa	Nohora Plazas	Ingeniero geólogo, especialista en gestión ambiental, estudiante de especialización en sistemas de información geográfica.

Fuente: Autores

## 6.7. Validación del instrumento de medición

### MODELO DE ANALÍTICA DE DATOS PARA LA OPTIMIZACIÓN DE FLUJOS DE TRABAJO EN CONSULTORÍA AMBIENTAL PARA INGISOT SAS

Para cada una de las siguientes preguntas evaluaremos la claridad de la redacción, el enfoque conceptual y relación de la pregunta con la variable que se quiere medir. Por favor califique con 1 si considera que es adecuada la pregunta en relación al aspecto evaluado, o con 0 si considera que no lo es.

Nombre del evaluador: Jose Armando Santiago Garnica

Cargo: Director del area social de INGISOT SAS

NUM	VARIABLE	PREGUNTA	CLARIDAD DE LA REDACCIÓN	ENFOQUE CONCEPTUAL	RELACIÓN PREGUNTA VARIABLE	PROMEDIO	COMENTARIOS
1	Cultura organizacional	En INGISOT SAS se promueve activamente la adopción de nuevas tecnologías.	1	1	1	1,00	
2	Cultura organizacional	La empresa tiene una visión clara sobre el uso de ciencia de datos en sus procesos.	1	1	1	1,00	
3	Cultura organizacional	Existe apertura por parte de los líderes para implementar soluciones basadas en modelos analíticos.	1	1	1	1,00	
4	Cultura organizacional	La empresa invierte recursos financieros suficientes en proyectos de innovación tecnológica.	1	1	1	1,00	

**Tabla 6 Resultados de la ponderación V-Aiken.**

Valores Aiken	Cantidad de preguntas	Porcentaje según Aiken
0,87	5	11,08%
0,93	2	6,15%
1,00	18	82,77%
Total, general	25	100,00%

Fuente: Elaboración propia (2025)

### 6.7.1. Relación entre diagnóstico y diseño del modelo

Los resultados del instrumento, validados mediante el coeficiente **V de Aiken**, permitieron definir esas variables y dimensiones que nos ayudan a estructurar el modelo propuesto. Gracias a la incorporación de datos simulados como complemento al instrumento propuesto se facilitó la evaluación del modelo en diferentes escenarios, fortaleciendo su consistencia técnica.

A continuación, se presenta una síntesis de los principales hallazgos del diagnóstico y su implicación directa en el diseño del modelo, con el fin de hacer explícita la trazabilidad entre ambas.

**Tabla 7 Síntesis de hallazgos y su implicación en el modelo**

Dimensión	Hallazgo	Implicación
Cultura de Innovación	Apertura hacia innovación	Gestión del cambio favorable
Impacto Esperado	Alta percepción de beneficios	Priorizar quick wins
Infraestructura	Infraestructura básica adecuada	Arquitectura híbrida local-cloud
Madurez de Procesos	Procesos parcialmente sistematizados	Enfoque iterativo CRISP-DM
Capacidades Técnicas	Conocimiento limitado Python/R	Capacitación progresiva

Fuente: Elaboración propia (2025)

## 7. Diagnóstico organizacional






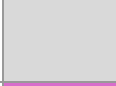


### 7.1. Metodología

#### 7.1.1. Mapeo de procesos

Para el diagnóstico de los flujos de trabajo en los servicios de INGISOT SAS se empleó la técnica de diagramación de procesos (mapeo de procesos), utilizando símbolos estandarizados de diagramas de flujo. Esta técnica permite representar gráficamente las secuencias de actividades, decisiones, herramientas y productos, facilitando la identificación de puntos críticos en la gestión de datos y la comunicación clara de procedimientos complejos.

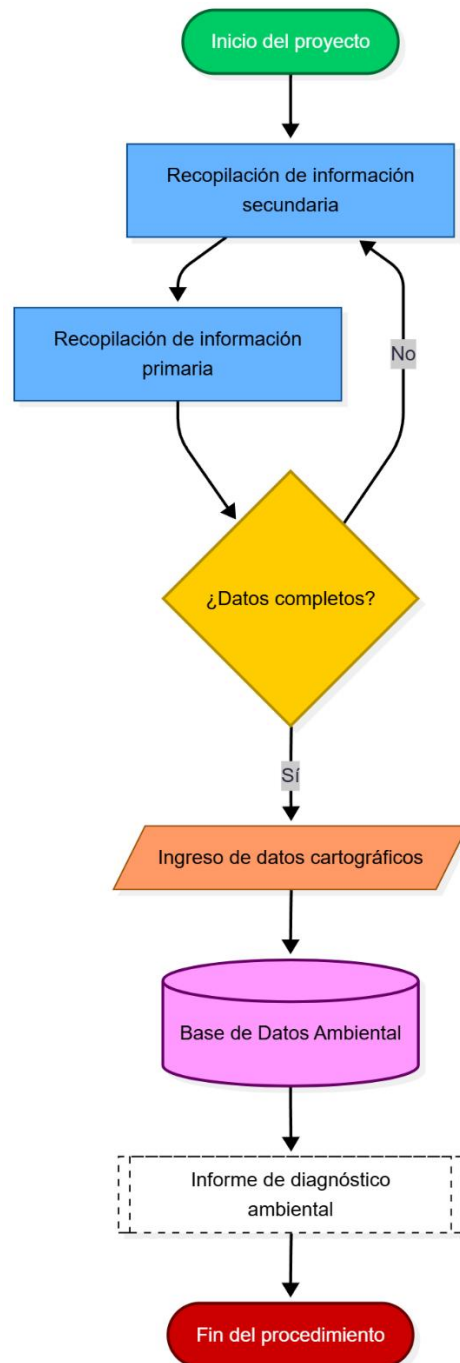
La elección de esta técnica responde a tres razones principales: (1) facilita la comprensión visual de procesos con múltiples actores y etapas, (2) permite evidenciar problemas y oportunidades de mejora relacionados con los datos, y (3) constituye una base metodológica para el diseño del modelo de analítica de datos orientado a la optimización de flujos de trabajo.

**Tabla 8 Elementos del mapa del proceso.**

Símbolo	Color	Nombre	Uso en los servicios	Ejemplo aplicado
Rectángulo		Actividad / Proceso	Representa una acción concreta dentro del flujo de trabajo.	"Recopilación de información secundaria"
Rombo		Decisión	Indica un punto de bifurcación o condición, donde se debe elegir un camino.	"¿Los datos son completos?"
Ovalo		Inicio / Fin	Señala el arranque o la finalización del procedimiento.	"Inicio del proyecto de EOT" / "Entrega del DTS"
Paralelogramo		Entrada / Salida de datos	Usado cuando se captura o se produce información en el flujo.	"Ingreso de datos cartográficos"
Rectángulo		Producto documental	Identifica la generación de reportes, informes o entregables oficiales.	"Informe de diagnóstico ambiental"
Fecha		Conexión	Une partes del diagrama cuando es complejo o se continúa en otra sección.	"Conexión al subproceso de depuración de datos"
Cilindro		Bases de datos	Alfanuméricas o geográficas	Base de datos catastral en formato GPKG
Nube		Herramientas	Herramientas usadas en los flujos de trabajo identificadas	Python, Excel, ArcGIS

Fuente: Elaboración propia (2025)

Figura 13 Mapeo de procesos



Fuente: Elaboración propia (2025)

---

### 7.1.2. Matriz de oportunidades

Además de la diagramación de procesos, se elaboraron tablas analíticas para cada línea de servicio de INGISOT SAS. Estas tablas constituyen un instrumento metodológico que permite sistematizar la información levantada en el diagnóstico, estableciendo un puente entre los flujos de trabajo actuales y las oportunidades de optimización mediante ciencia de datos.

La construcción de las tablas respondió a los siguientes criterios metodológicos:

**Descomposición en subactividades:** Cada servicio fue desagregado en tareas o fases específicas del flujo de trabajo, identificando aquellas con mayor dependencia de datos alfanuméricos y espaciales.

**Identificación de insumos y productos de datos:** Para cada subactividad se consignaron los tipos de datos generados o utilizados (series numéricas, registros alfanuméricos, cartografía, imágenes satelitales, etc.), con el fin de visibilizar la diversidad y volumen de información gestionada.

**Problemas de datos:** Se registraron los principales desafíos asociados a la gestión de información (faltantes, dispersión, desactualización, duplicidad, baja interoperabilidad), los cuales representan cuellos de botella en los flujos actuales.

**Potencial analítico:** Se evaluó el alcance de la ciencia de datos en cada subactividad, diferenciando entre análisis descriptivo (estadísticas e indicadores), predictivo (modelos de proyección y aprendizaje automático) y prescriptivo (escenarios y recomendaciones de acción).

**Técnicas y herramientas aplicables:** Se señalaron las metodologías de ciencia de datos pertinentes en cada caso (análisis estadístico, machine learning, minería de datos espaciales, integración en entornos GIS), priorizando aquellas que permiten replicabilidad y escalabilidad.

**Priorización estratégica:** Finalmente, cada subactividad se clasificó según su relevancia (alta, media o baja), tomando como referencia el impacto en la calidad del servicio, la criticidad de los datos y la posibilidad de automatización futura.

En conjunto, estas tablas constituyen una matriz de análisis metodológico que no solo organiza la información del diagnóstico, sino que también anticipa los puntos de inserción de la ciencia de datos y la inteligencia artificial dentro de los flujos de trabajo de INGISOT SAS, preparando el camino para el diseño del modelo planteado en el capítulo siguiente.

## **7.2. Líneas estratégicas INGISOT SAS**

### **7.2.1. Consultoría ambiental**

INGISOT SAS desarrolla consultoría ambiental en múltiples temáticas, siempre enmarcadas en los términos de referencia definidos por la autoridad competente según la naturaleza, magnitud e impacto del proyecto a licenciar. Estos términos de referencia determinan no solo los alcances técnicos del estudio, sino también el rol de la autoridad ambiental en la evaluación y otorgamiento de permisos y trámites, así como la conformación de equipos interdisciplinarios responsables de la caracterización ambiental.

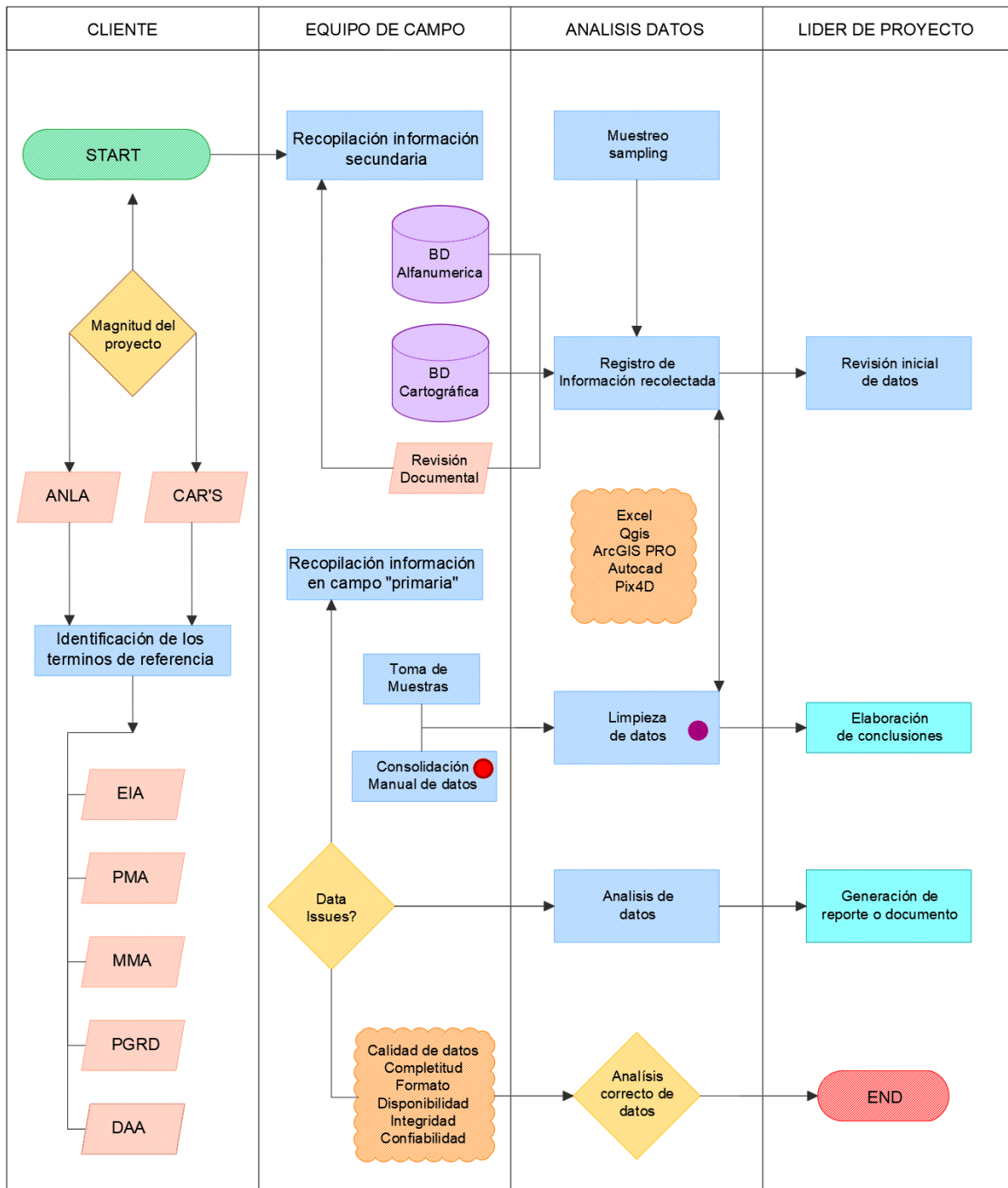
En términos generales, la ejecución de estos estudios comprende dos etapas fundamentales:

- Recopilación de información secundaria, a partir de bases de datos institucionales, literatura técnica y cartografía oficial.
- Recolección de información primaria en campo, que abarca levantamientos de datos físicos, bióticos y sociales en el área de influencia del proyecto.

Ambas etapas generan un volumen significativo de información —numérica, alfanumérica y cartográfica— que requiere procesos de tratamiento manual y semiautomatizado. Estos procesos buscan garantizar la calidad, completitud, homogeneidad, disponibilidad, integridad y confiabilidad de los datos. Para ello, se emplean herramientas como Excel, QGIS y ArcGIS, siendo estas últimas fundamentales para la gestión de información geográfica.

Posteriormente, la información depurada se organiza y analiza de acuerdo con las subtemáticas establecidas en cada término de referencia. Los resultados se presentan en documentos técnicos acompañados de anexos gráficos, tablas estadísticas y cartografía temática.

**Figura 14 Mapeo de procesos línea de consultoría ambiental**



Fuente: Elaboración propia (2025)

De manera estructural, los estudios ambientales se componen de las siguientes dimensiones:

Física y abiótica (clima, geología, suelos, hidrología, calidad del aire).

Biótica (ecosistemas, flora y fauna).

Socioeconómica y cultural.

Zonificaciones ambientales.

Evaluación de impactos.

Plan de manejo ambiental.

De forma transversal, los estudios también incluyen la **descripción del proyecto**, la **delimitación de áreas de influencia**, y la formulación de **planes de cierre y abandono**. Estos componentes integrados garantizan un análisis coherente entre la caracterización ambiental, la evaluación de impactos y la gestión de medidas de manejo.

En suma, la consultoría ambiental de INGISOT SAS combina la rigurosidad técnica en la recolección y análisis de datos con la necesidad de generar productos claros y aplicables para la toma de decisiones en materia de licenciamiento ambiental.

**Tabla 9 Matriz de oportunidad en consultoría ambiental**

<b>Servicio / Subactividad</b>	<b>Datos generados o usados</b>	<b>Problemas de datos</b>	<b>Potencial analítico (descriptivo, predictivo, prescriptivo)</b>
<b>Morfometría de cuencas y geomorfología</b>	Modelos digitales de elevación (DEM), imágenes satelitales, derivados topográficos (pendiente, curvatura, rugosidad, dirección de flujo, red de drenaje)	Diferentes resoluciones espaciales, ruido en datos satelitales, limitaciones de cobertura, necesidad de correcciones geométricas	Descriptivo (extracción de parámetros geomorfológicos), Predictivo (modelos de erosión, escorrentía, susceptibilidad a remoción en masa), Prescriptivo (identificación de áreas críticas para planificación territorial)
<b>Suelos (caracterización fisicoquímica y clasificación agrológica)</b>	Resultados de laboratorio (pH, textura, materia orgánica, nutrientes, contaminantes), mapas de suelos, unidades cartográficas, capacidad de uso	Cobertura espacial limitada, falta de actualización en cartografía, variabilidad espacial alta, datos dispersos entre diferentes fuentes	Descriptivo (estadísticas de parámetros fisicoquímicos, mapas de capacidad de uso), Predictivo (modelos de fertilidad, riesgo de contaminación, interpolación espacial de propiedades), Prescriptivo (clasificación agrológica, zonificación de aptitud agrícola)
<b>Zonificación ambiental y de manejo</b>	Capas temáticas: uso del suelo, cobertura vegetal, hidrografía, suelos, áreas protegidas, riesgo ambiental, infraestructura	Diferentes escalas cartográficas, inconsistencias entre fuentes, normatividad cambiante, necesidad de homogenización espacial	Descriptivo (mapas de restricción y factibilidad), Predictivo (modelos de susceptibilidad ambiental y espacial), Prescriptivo (escenarios de zonificación óptima mediante metodologías multicriterio)
<b>Atmósfera (determinación del clima según balance hídrico)</b>	Series temporales hidrometeorológicas: temperatura, precipitación, evapotranspiración	Series incompletas, estaciones con periodos cortos, variabilidad interanual, problemas de homogeneidad	Descriptivo (estadísticas climáticas y balances hídricos históricos), Predictivo (estimación de datos faltantes y proyección de tendencias), Prescriptivo (clasificación climática y escenarios de disponibilidad hídrica)

<b>Índices de fragmentación y análisis de especies vulnerables</b>	Imágenes multispectrales (Sentinel, Landsat), ortofotos de alta resolución, cartografía de cobertura y uso del suelo, registros de biodiversidad	Variabilidad en resolución y calidad de imágenes, limitaciones de series temporales, ausencia de datos actualizados sobre fauna/flora, dificultad de integración espacial	Descriptivo (cálculo de índices de fragmentación, métricas de paisaje), Predictivo (detección automática de áreas críticas y pérdida de hábitat), Prescriptivo (priorización de corredores biológicos y áreas de conservación para especies vulnerables)
<b>Vulnerabilidad de especies (UICN y Libros Rojos de Colombia)</b>	Listados de especies identificadas en campo (fauna y flora), bases de datos de la UICN, Libros Rojos de Colombia, registros de biodiversidad (SiB Colombia, GBIF)	Dificultades de homologación taxonómica, diferencias en nomenclatura, desactualización en categorías de amenaza, vacíos de información local	Descriptivo (estado de conservación y distribución de especies), Predictivo (proyección de riesgo de extinción en áreas específicas), Prescriptivo (recomendaciones de conservación y manejo en planes ambientales y territoriales)
<b>Categorización taxonómica de especies (campo vs GBIF)</b>	Listas de especies levantadas en campo, bases de datos globales (GBIF), metadatos taxonómicos (familia, género, orden)	Nombres científicos inconsistentes, sinónimos taxonómicos, duplicidad en registros, vacíos en metadatos taxonómicos locales	Descriptivo (clasificación taxonómica estandarizada de especies), Predictivo (validación automática de nuevas especies reportadas en campo), Prescriptivo (alineación de la taxonomía para reportes oficiales y planes de manejo ambiental)
<b>Transferencia de datos biológicos a formatos Darwin Core</b>	Resultados de caracterización biótica: género, familia, especie, lugar de colecta, número de individuos, colecciones biológicas; estándares Darwin Core	Errores en la estandarización de campos, ausencia de metadatos completos, duplicación de registros, incompatibilidad con sistemas receptores de información	Descriptivo (estructuración y validación de datos biológicos en estándares abiertos), Predictivo (detección automática de inconsistencias en registros biológicos), Prescriptivo (alineación de datos para interoperabilidad con SiB Colombia, GBIF y colecciones biológicas oficiales)

### 7.2.2. Geografía y catastro

La línea de geografía y catastro en INGISOT SAS se estructura en dos grandes subgrupos de trabajo:

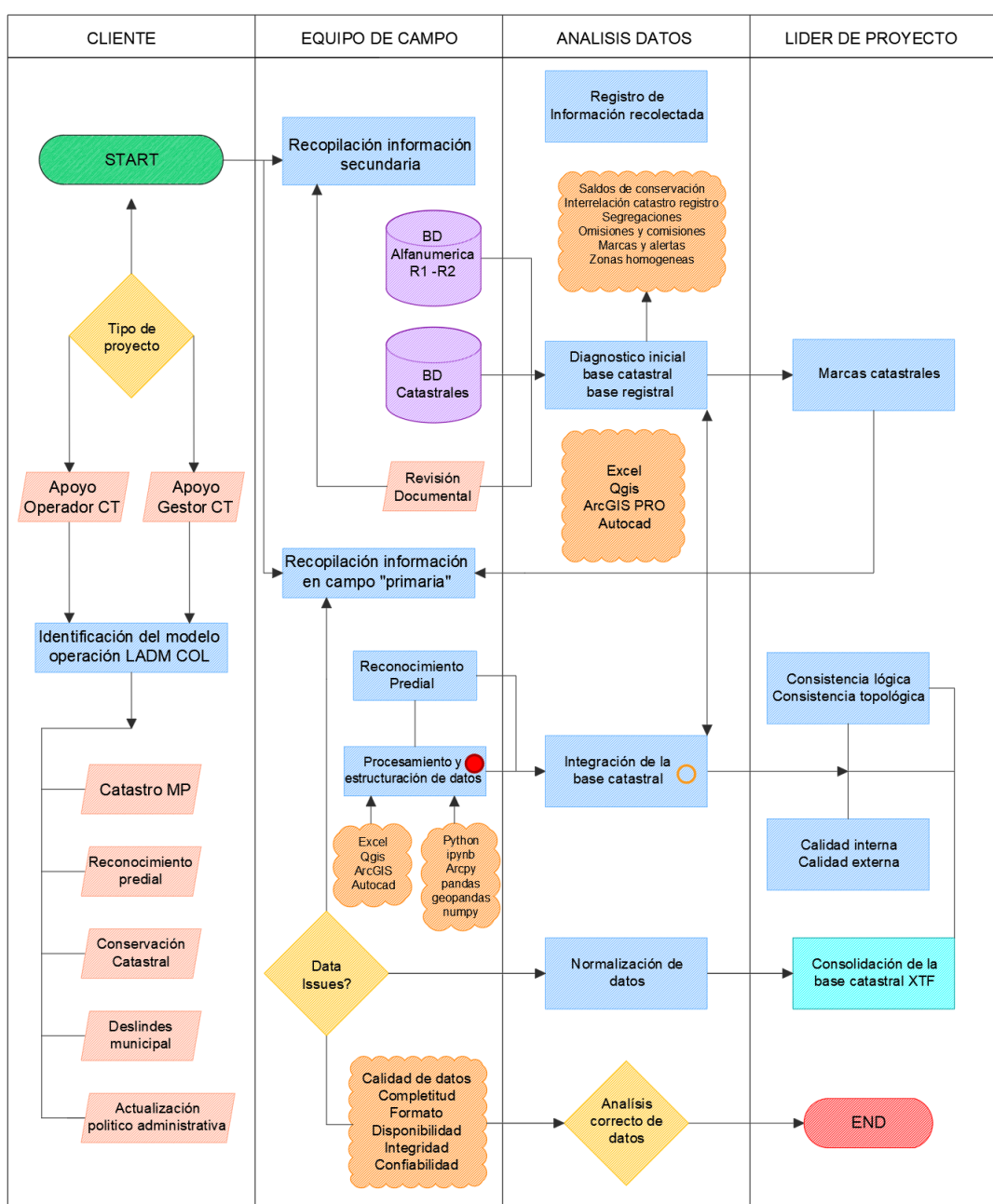
Apoyos transversales al ordenamiento territorial, que incluyen procesos como deslindes, actualizaciones político-administrativas y caracterizaciones técnicas, topográficas y sociales necesarias para la delimitación y regulación de los territorios.

Procesos de actualización catastral, que abarcan desde la identificación predial y el reconocimiento de la base física, hasta la construcción de bases geográficas en formatos de intercambio (XTF), exigidos por el modelo LADM-COL. Estas actividades requieren experticia técnica en sistemas de información geográfica y conocimiento específico en catastro, involucrando subactividades como apoyo social, reconocimiento predial, control de calidad (interna y externa), zonificaciones económicas y avalúos catastrales.

En ambas líneas confluyen datos alfanuméricos y geográficos. Los primeros provienen de bases catastrales usualmente gestionadas en formatos planos como CSV o Excel,

mientras que los segundos se apoyan en formatos SIG estructurados. El tratamiento de estos datos requiere diversas operaciones: control de calidad, verificación de completitud, estandarización de formatos, disponibilidad, integridad y confiabilidad. Estos procesos se apoyan en registros catastrales previos (R1 y R2) que constituyen la base para actualizaciones y permiten responder a requerimientos específicos según los contratos de los gestores catastrales.

**Figura 15 Mapeo de procesos línea de geografía y catastro**



Fuente: Elaboración propia (2025)

Hasta ahora, gran parte de las actividades se han desarrollado mediante métodos **ortodoxos**, principalmente con Excel, o recurriendo a **herramientas gubernamentales** desarrolladas por el Instituto Geográfico Agustín Codazzi (IGAC). Si bien estas soluciones han introducido la lógica de la ciencia de datos en catastro, su replicabilidad en la consultoría privada es limitada. Este contexto abre para INGISOT SAS la oportunidad de **replicar, adaptar y escalar herramientas propias**, orientadas a la gestión integral de información catastral.

La principal oportunidad identificada en esta línea consiste en avanzar hacia la **ciencia de datos geoespacial**, entendida como la integración de información alfanumérica y geográfica en entornos de análisis reproducibles. Herramientas de código abierto como **GeoPandas** (para datos vectoriales) y **Rasterio** (para datos raster) en Python, o sus equivalentes en R, permiten estructurar un flujo de trabajo moderno y escalable para el manejo de catastro. Con ello, INGISOT SAS puede migrar de procesos manuales y fragmentados hacia un modelo basado en automatización, integración y análisis espacial avanzado, fortaleciendo así su competitividad en esta línea estratégica.

El diagnóstico anterior evidencia que la gestión catastral de INGISOT SAS enfrenta retos asociados a la dispersión de formatos, dependencia de procesos manuales y baja replicabilidad de soluciones desarrolladas por entidades gubernamentales. Sin embargo, también identifica un alto potencial para integrar la ciencia de datos geoespacial en sus flujos de trabajo. La Tabla X sintetiza los datos utilizados, los problemas recurrentes, las oportunidades analíticas y las técnicas de ciencia de datos aplicables en esta línea de servicio

**Tabla 10 Matriz de oportunidad línea de geografía y catastro**

Servicio / Subactividad	Datos generados o usados	Problemas de datos	Potencial analítico (descriptivo, predictivo, prescriptivo)
<b>Catastro multipropósito (calidad y consistencia de datos)</b>	Bases catastrales multipropósito estructuradas en el modelo LADM-COL: predios, terrenos, construcciones, interesados, derechos y restricciones	Inconsistencias lógicas y topológicas, duplicidad de registros, datos incompletos, heterogeneidad en captura y actualización	Descriptivo (auditoría de calidad de datos catastrales, análisis de consistencia topológica), Predictivo (detección automática de anomalías en registros), Prescriptivo (propuestas de normalización y reglas de calidad aplicables al modelo LADM-COL)

<b>Catastro multipropósito (calidad externa)</b>	Versiones de base catastral de conservación y actualización (XTF R1 y R2), información registral (SICRE), tipologías constructivas y clasificación estandarizada	Inconsistencias entre versiones, cambios no documentados, diferencias en tipologías, ausencia de interoperabilidad plena entre catastro y registro	Descriptivo (comparación y análisis de cambios entre versiones catastrales), Predictivo (detección de patrones de cambio en construcciones y predios), Prescriptivo (propuesta de homologación y estandarización en tipologías y atributos)
<b>Reconocimiento predial (marcas catastrales y conservación)</b>	Información predial catastral (geometría de terrenos y construcciones), información registral (folios, actos jurídicos), límites físicos en campo, coincidencias topográficas	Omisiones y comisiones en registros, duplicidad de predios, inconsistencias entre catastro y registro, diferencias entre límites jurídicos y físicos	Descriptivo (identificación de omisiones/comisiones y saldos prediales), Predictivo (detección automática de inconsistencias topográficas), Prescriptivo (propuestas de conciliación Catastro-Registro y ajustes de conservación)
<b>Catastro multipropósito (contraste entre trámites y saldos de conservación vs base actualizada XTF)</b>	Trámites y saldos de conservación catastral, versiones de base catastral actualizada en formato XTF, metadatos de actualización	Diferencias entre registros de conservación y base actualizada, errores de carga en XTF, falta de trazabilidad de cambios, ausencia de reglas homogéneas para saldos	Descriptivo (comparación de saldos y cambios entre conservación y actualización), Predictivo (detección de inconsistencias recurrentes en los procesos de conservación), Prescriptivo (diseño de reglas de control y reconciliación de saldos en la actualización catastral)

Fuente: Elaboración propia (2025)

### 7.2.3. Ordenamiento territorial

El ordenamiento territorial, dada su extensión e importancia como instrumento de planificación, constituye una de las líneas estratégicas más amplias de INGISOT SAS. Su desarrollo implica un tratamiento intensivo de datos alfanuméricos y geográficos, que abarca desde las etapas iniciales del ciclo de vida de los datos —consolidación, adquisición, limpieza, depuración, organización y reducción— hasta fases más complejas como la presentación y visualización de resultados. Estos resultados pueden expresarse en múltiples formatos: gráficos estadísticos, tableros de control, indicadores, cartografía temática o despliegues web mediante geoservicios y geovisores.

Lo anterior configura una ruta de trabajo lineal pero bifurcada, en la que coexisten actividades de nivel básico (consolidación y organización de datos) con procesos de mayor complejidad técnica (desarrollo de geovisores, generación de indicadores avanzados o despliegue de servicios interoperables).

Dentro de esta línea, dos grandes bloques merecen especial atención:

- El diagnóstico territorial y ambiental, que en cualquiera de sus dimensiones (ambiental, social, institucional, económica y funcional) genera un gran volumen

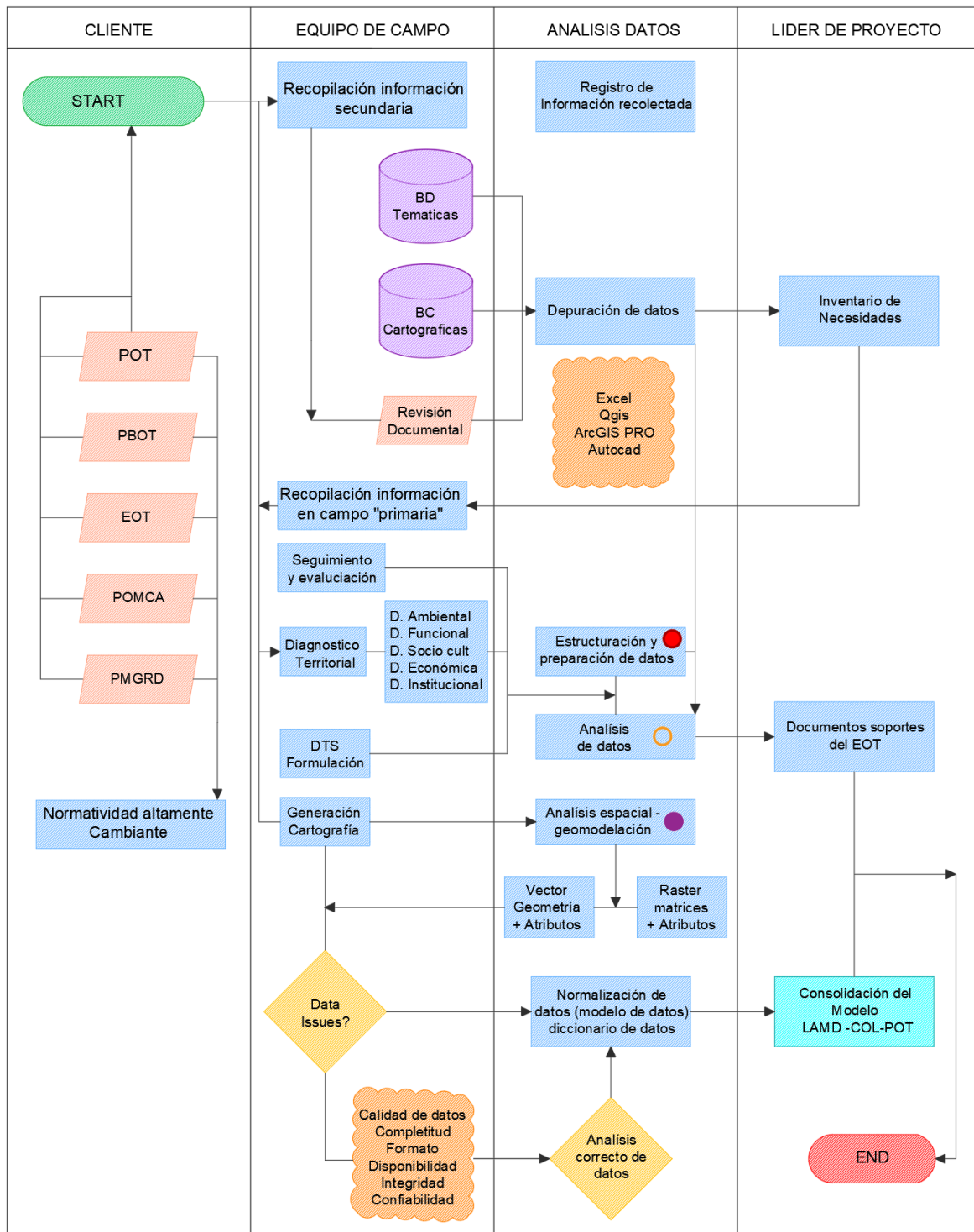
---

de datos con potencial de gobernanza, estandarización y alineación a infraestructuras de datos.

- La gestión del riesgo y el cambio climático, que constituyen temas de alta experticia técnica y son esenciales para la concertación de asuntos ambientales con las Corporaciones Autónomas Regionales (CAR) y otras autoridades ambientales.

El ordenamiento territorial, además de ser un proceso técnico, es también un instrumento normativo y político, soportado en la Ley 388 de 1997 y en el Decreto 1232 de 2020, los cuales orientan la formulación de los POT y EOT en Colombia. Esta normativa exige integrar información proveniente de diversas instituciones (IDEAM, IGAC, DANE, ANM, CAR, SIRH, entre otras), lo que incrementa los retos en la gestión y articulación de bases de datos. En este contexto, la naturaleza interdisciplinaria del ordenamiento —que involucra componentes ambientales, urbanos, sociales, económicos e institucionales— lo convierte en un espacio idóneo para la aplicación de ciencia de datos, no solo para la caracterización descriptiva del territorio, sino también para la construcción de escenarios prospectivos y prescriptivos que apoyen la toma de decisiones en planificación territorial.

Figura 16 Mapeo de procesos línea de ordenamiento territorial



Fuente: Elaboración propia (2025)

La síntesis de resultados en esta línea se refleja en la construcción de los componentes general, rural y urbano que conforman los documentos de formulación del ordenamiento. Dichos componentes se apoyan en estadísticas, cifras e indicadores derivados del diagnóstico territorial. A partir de este escenario, la incorporación de ciencia de datos se proyecta como una herramienta estratégica para estandarizar procesos, reducir tiempos de análisis y generar productos reproducibles y comparables entre municipios y regiones. Más allá de esto, el potencial de la inteligencia artificial (IA) abre nuevas posibilidades: desde algoritmos de aprendizaje automático capaces de anticipar escenarios de crecimiento urbano y cambios de cobertura, hasta modelos prescriptivos que recomienden alternativas de planificación óptima bajo criterios de sostenibilidad y riesgo. Aunque en la actualidad la IA se vincula principalmente al uso de código en la ciencia de datos, su integración futura permitirá escalar hacia sistemas inteligentes de apoyo a la decisión territorial, transformando la manera en que se conciben, formulan y gestionan los instrumentos de ordenamiento territorial.

A partir de lo anterior, resulta pertinente sistematizar las principales actividades y subprocesos asociados al ordenamiento territorial, identificando para cada uno de ellos los datos generados o utilizados, los problemas recurrentes de información, el potencial analítico y las técnicas de ciencia de datos aplicables. Esta estructuración permite no solo organizar de manera clara los componentes estratégicos de la línea, sino también evidenciar los puntos críticos donde la analítica descriptiva, predictiva y prescriptiva puede generar mayor valor agregado. La siguiente tabla resume dichos aspectos, priorizando aquellos con mayor impacto en la eficiencia y calidad de los procesos de ordenamiento territorial.

**Tabla 11 Matriz de oportunidad en el ordenamiento territorial**

<b>Servicio / Subactividad</b>	<b>Datos generados o usados</b>	<b>Problemas de datos</b>	<b>Potencial analítico (descriptivo, predictivo, prescriptivo)</b>
--------------------------------	---------------------------------	---------------------------	--

<b>Cambio climático (escenarios y tendencias)</b>	Escenarios climáticos del IDEAM (precipitación, temperatura), captura de carbono, series históricas, proyecciones (Cuarta Comunicación Nacional)	Incertidumbre en proyecciones, diferentes escalas espaciales/temporales, carencia de series largas y consistentes	Descriptivo (análisis de tendencias históricas), Predictivo (modelos de escenarios climáticos a futuro), Prescriptivo (integración de escenarios en planificación territorial, medidas de adaptación en EOT)
<b>Análisis socioeconómico (EIA)</b>	Datos demográficos: población, edad, educación, empleo, ingresos (fuentes: DANE, censos, encuestas locales)	Desactualización de censos, falta de datos desagregados a nivel local, heterogeneidad en formatos	Descriptivo (caracterización poblacional y socioeconómica), Predictivo (proyecciones de crecimiento poblacional, tendencias educativas/laborales), Prescriptivo (modelación de escenarios de impacto social según proyectos)
<b>Ordenamiento territorial (análisis de la dimensión económica)</b>	Datos agrícolas (áreas cultivadas, producción), pecuarios (inventarios de ganado, producción lechera/cárnica), minería (volumen extraído, regalías), información económica municipal (DANE, ANM, MADR)	Datos dispersos entre instituciones, series desactualizadas, ausencia de homogeneización, escalas municipales poco detalladas	Descriptivo (estadísticas económicas locales, indicadores de productividad), Predictivo (tendencias en producción agrícola, pecuaria y minera), Prescriptivo (escenarios de desarrollo económico y priorización de actividades productivas para el ordenamiento)
<b>Cambio climático (proyección de extremos climáticos)</b>	Escenarios de la Cuarta Comunicación de Cambio Climático del IDEAM: precipitaciones máximas, temperaturas extremas, emisiones y concentración de CO <sub>2</sub>	Incertidumbre en escenarios globales/regionales, baja resolución espacial, series históricas incompletas para comparación, dificultades en downscaling	Descriptivo (análisis de máximos históricos y tendencias de extremos), Predictivo (proyección de precipitaciones extremas, olas de calor/frío, niveles de CO <sub>2</sub> ), Prescriptivo (incorporación de escenarios de riesgo climático en la planificación territorial y EOT)

Fuente: Elaboración propia (2025)

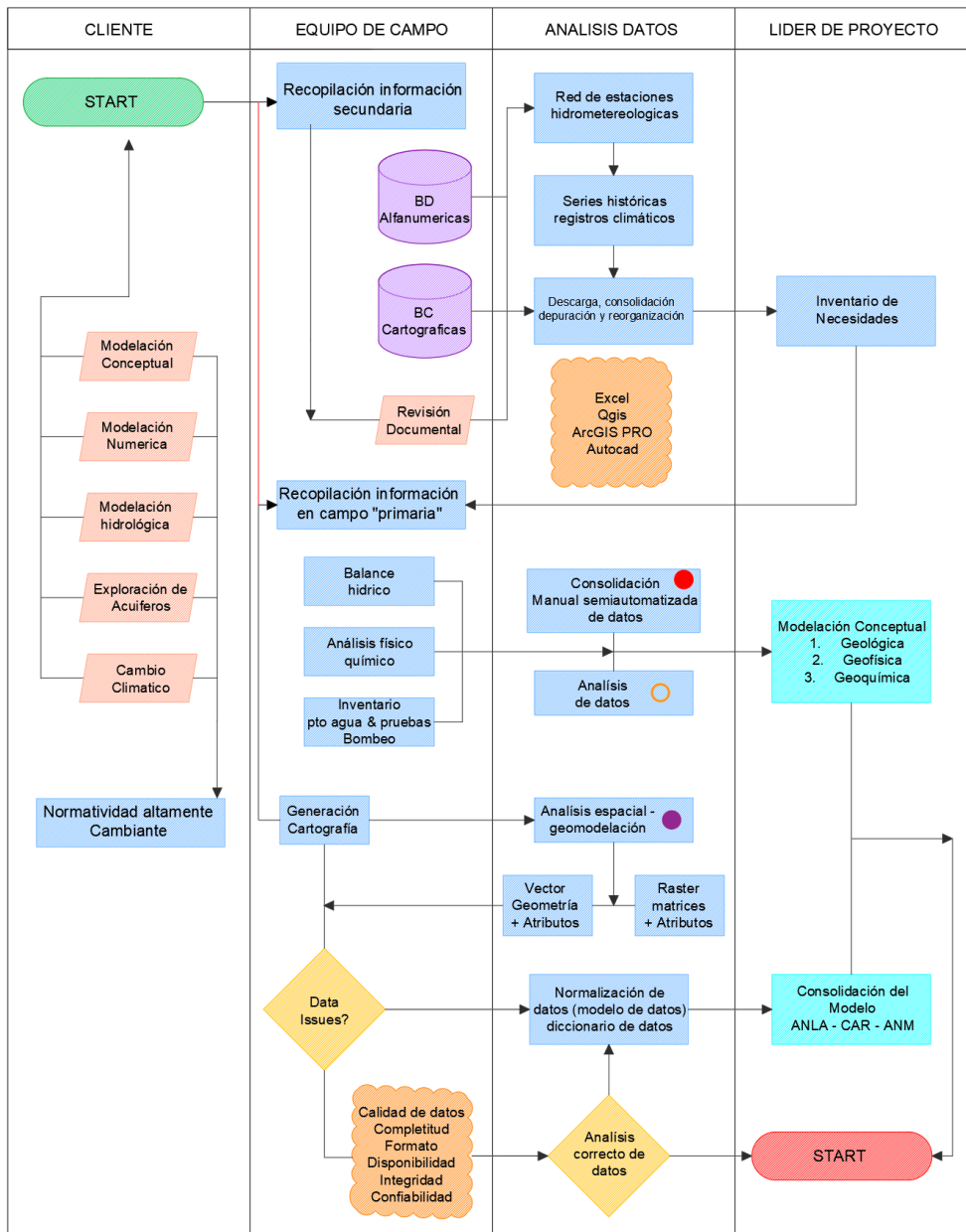
#### 7.2.4. Recurso hídrico y modelación

La modelación del recurso hídrico superficial y subterráneo constituye una de las líneas de acción específicas de INGISOT SAS. Su desarrollo parte de la caracterización del área de estudio y de la consolidación de información secundaria, principalmente las bases de

datos alfanuméricas correspondientes a las estaciones hidrometeorológicas del IDEAM, que constituyen el punto de partida para los análisis hidrológicos y climáticos.

El flujo de trabajo actual se estructura en varias etapas: descarga, consolidación, depuración y reorganización de datos, tareas que se realizan principalmente mediante herramientas como Excel, QGIS, ArcGIS y AutoCAD. A partir de este proceso se derivan parámetros, series, gráficos, estadísticas y zonificaciones que permiten caracterizar aspectos fundamentales como calidad de datos, detección de outliers, identificación de vacíos o valores nulos, análisis del régimen de lluvias (monomodal o bimodal) y comportamiento climatológico de la zona de estudio. Entre los principales productos destacan los cálculos de balance hídrico, caudales, oferta y escasez hídrica, caudal ambiental, así como indicadores máximos, mínimos y medios para variables como temperatura, precipitación, nubosidad y evapotranspiración.

**Figura 17 Mapeo de procesos línea de recurso hídrico.**



Fuente: Elaboración propia (2025)

El tratamiento de datos faltantes constituye un aspecto central de este proceso. Actualmente se recurre a métodos estadísticos convencionales como media, mediana o moda; sin embargo, esta aproximación deja de lado el potencial que ofrecen las librerías de ciencia de datos y machine learning para series temporales, que permiten un manejo más robusto y validado de la completitud de datos.

El flujo actual demanda una elevada inversión de tiempo y depende en gran medida de la experticia del profesional asignado. Esto genera riesgos operativos asociados a la rotación de personal o al manejo simultáneo de varios proyectos. De allí surge la necesidad de contar con flujos internos validados y estandarizados que reduzcan la carga mecánica de procesamiento y permitan que el equipo concentre sus esfuerzos en el análisis e interpretación de los resultados, incrementando así la calidad y pertinencia de los productos entregados.

Adicionalmente, la modelación hídrica no es ajena a la convergencia entre datos alfanuméricos y espaciales. La georreferenciación de estaciones hidrometeorológicas y la integración con información de campo refuerzan la importancia del componente espacial, que puede potenciarse mediante el uso de librerías específicas como Arcpy y PyQGIS, y su integración en entornos de programación (Spyder, Visual Studio Code, Jupyter Notebooks) y plataformas colaborativas como GitHub. Esto abre un abanico de posibilidades para transformar los flujos actuales en procesos semi-automatizados, reproducibles y escalables hacia infraestructuras de datos espaciales (IDE).

**Tabla 12 Matriz de oportunidad recurso hídrico y modelación**

<b>Servicio / Subactividad</b>	<b>Datos generados o usados</b>	<b>Problemas de datos</b>	<b>Potencial analítico (descriptivo, predictivo, prescriptivo)</b>
--------------------------------	---------------------------------	---------------------------	--

<b>Hidrología (balance hídrico, régimen de lluvias y caudales)</b>	Series históricas de precipitación, caudales, evapotranspiración (descargados del IDEAM)	Datos faltantes en series de tiempo, inconsistencias en estaciones, escalas temporales distintas	Descriptivo (estadísticas hidrológicas, tendencias), Predictivo (interpolación y estimación de datos faltantes, proyecciones de caudales), Prescriptivo (escenarios de disponibilidad hídrica)
<b>Atmósfera (clima y zonas de vida)</b>	Series del IDEAM: temperatura, evapotranspiración, humedad relativa, vientos; altitud (DEM)	Series incompletas, baja densidad de estaciones, heterogeneidad espacial	Descriptivo (estadísticas climáticas, zonificación), Predictivo (estimación de variables climáticas en zonas sin estación), Prescriptivo (escenarios de cambio climático o clasificación de zonas de vida)
<b>Hidrogeología (inventario de puntos de agua y susceptibilidad de acuíferos)</b>	Parámetros fisicoquímicos de agua (pH, conductividad, nitratos, metales), inventario de pozos, niveles freáticos, litología, vulnerabilidad (índices DRASTIC, GOD)	Escasez de puntos de monitoreo, datos heterogéneos en espacio y tiempo, ausencia de series largas, falta de normalización de parámetros	Descriptivo (caracterización de calidad y disponibilidad de agua subterránea), Predictivo (estimación de parámetros en sitios sin muestreo, modelos de susceptibilidad de acuíferos), Prescriptivo (zonificación de vulnerabilidad, soporte a decisiones de gestión del recurso hídrico)

Fuente: Elaboración propia (2025)

### 7.2.5. Sistemas de información geográfica

Los sistemas de información geográfica (SIG) han sido, desde la creación de INGISOT SAS, la herramienta transversal que dio origen a su nombre y una de sus mayores fortalezas dentro de la consultoría ambiental. El dominio tanto de software libre como comercial ha permitido a la empresa integrarse a múltiples líneas de trabajo, siempre con el componente espacial como eje central. En cada proyecto, los datos adquieren un adjetivo particular según su naturaleza y propósito, pero todos convergen en la infraestructura SIG.

## 1. Ciclo de vida de la información espacial

A lo largo del tiempo, y con la masificación del uso de estas herramientas, INGISOT ha consolidado un flujo de trabajo dinámico que puede representarse como un ciclo de vida de los datos espaciales. Este comprende:

Creación de información: generación de insumos cartográficos y temáticos.

Conversión de formatos: migración de información antigua (p. ej. CAD a GIS).

Adquisición y procesamiento digital de imágenes: aprovechando drones con sensores RGB, multiespectrales y LIDAR.

Actualización y conversión de datos geográficos y alfanuméricos: desde geodatabases (GDB) hasta formatos más versátiles como Geopackage (GPKG) o XML Feature Types (XTF). Generación de cartografía temática multipropósito: aplicada a estudios de riesgo, cambio climático, ordenamiento territorial, POMCAS y licenciamiento ambiental (EIA, PMA, DAA, MMA). Visualización y divulgación de resultados: mediante aplicaciones, geoportales o servicios interoperables (WFS, WMS).

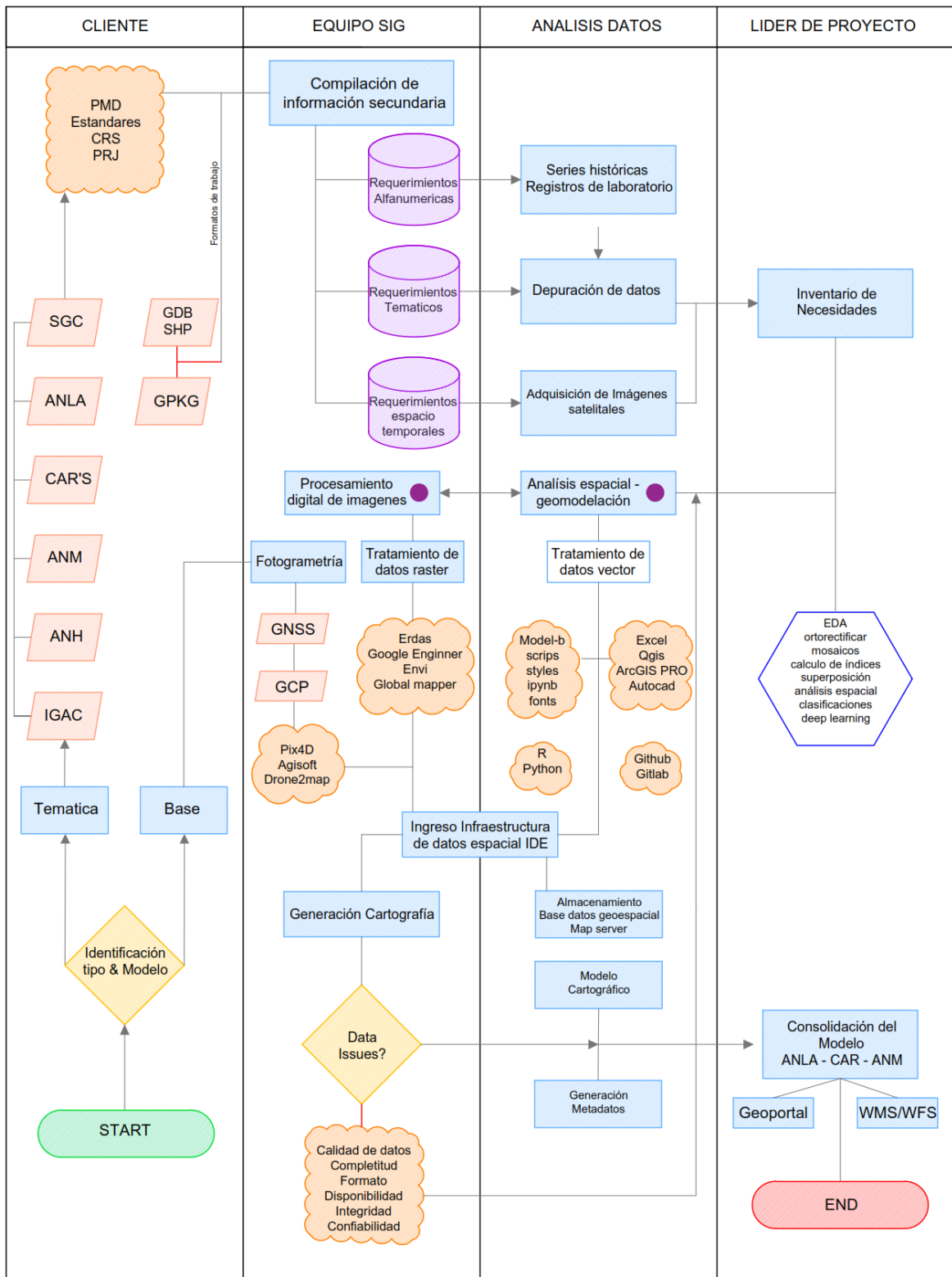
## **2. Volumen de información y retos de automatización**

El aporte de los SIG en los proyectos varía según el volumen de información gestionada. Puede abarcar desde pequeñas tablas de atributos hasta bases con miles o cientos de miles de registros. Entre los más significativos destacan los procesos de actualización catastral, que requieren modelos relacionales robustos y propensos a la automatización gracias a su estructura de datos. Esto ha permitido la inserción progresiva de resultados dentro de infraestructuras de datos espaciales (IDE), consolidando la interoperabilidad institucional.

## **3. Pilotos locales y programación aplicada**

Diversos pilotos locales han demostrado los beneficios de incorporar código en los flujos SIG. Desde simples interconexiones y conversiones hasta análisis complejos, la programación ha permitido reconstruir modelos relacionales, identificar capas principales, establecer relaciones y mapeos de dominios, tanto para un municipio en particular como para varios municipios integrados en estructuras de datos jerárquicas. El uso de librerías como pandas ha sido clave para organizar diccionarios y manejar grandes volúmenes de datos alfanuméricos con atributos espaciales, logrando reducción de tiempos, mayor integridad de la información y velocidad de respuesta.

### **Figura 18 Mapeo de procesos línea SIG**



Fuente: Elaboración propia (2025)

---

#### **4. Tratamiento de datos raster y analítica avanzada**

El manejo de datos raster, grillas matriciales alojadas en píxeles también ha sido potenciado por la programación. A través de notebooks estructurados en módulos, se ejecutan procesos de superposición, reclasificación, limpieza, generalización y conversión de formatos. Este flujo ha evolucionado hacia la integración con técnicas de machine learning, aplicadas a modelos predictivos que identifican áreas de riesgo, escenarios de cambio climático o zonificaciones ambientales, generando así productos cartográficos y temáticos de alto valor para la toma de decisiones.

#### **5. Enfoque estratégico y ciencia de datos**

Ante un marco normativo cambiante y cada vez más exigente, INGISOT reconoce el riesgo de una involución tecnológica que limite su participación en procesos contractuales o licitatorios. Por ello, identifica un nicho de mercado en liderar la transición de un modelo ortodoxo, basado en procesamiento manual, hacia uno soportado en ciencia de datos. En este, los modelos descriptivos, predictivos y prescriptivos se encajan en cada línea de negocio, con la generación de código capaz de semi-automatizar las etapas iniciales del ciclo de vida de los datos: adquisición, limpieza, depuración y normalización, o bien integrarlos directamente a modelos de datos e infraestructuras espaciales (IDE).

#### **6. Otros componentes técnicos: GNSS y cartografía base**

Un campo complementario, aunque no exclusivo de la ciencia de datos, es el tratamiento de datos obtenidos mediante equipos GNSS y software especializado para el procesamiento y modelado de puntos de control. Estos insumos potencian la generación de cartografía base de alta precisión, reforzando la capacidad técnica de INGISOT en el ámbito SIG.

**Tabla 13 Matriz de oportunidad en SIG**

Servicio / Subactividad	Datos generados o usados	Problemas de datos	Potencial analítico (descriptivo, predictivo, prescriptivo)	Técnicas de ciencia de datos aplicables	Priorización
<b>Descarga y consolidación de datos hidrometeorológicos</b>	Series históricas de precipitación, temperatura, caudales, evapotranspiración (IDEAM, estaciones locales)	Datos faltantes, registros discontinuos, diferentes formatos, escalas temporales heterogéneas	Descriptivo: estadísticas básicas, identificación de tendencias y distribución espacial. Predictivo: proyección de series de precipitación y temperatura. Prescriptivo: escenarios de disponibilidad hídrica.	Procesamiento de series temporales, análisis convencionales y matrices de datos Excel.	Alta

Fuente: Elaboración propia (2025)

### 7.2.6. Análisis de resultados

El diagnóstico realizado permitió descomponer y representar gráficamente los flujos de trabajo de INGISOT SAS en cada una de sus cinco líneas estratégicas. A través de los diagramas y tablas analíticas, se evidenció con claridad la manera en que la organización gestiona la información alfanumérica, geográfica y estadística, así como los principales problemas recurrentes: dispersión de datos, ausencia de estandarización, dependencia de procesos manuales y limitada interoperabilidad entre fuentes.

Más allá de las particularidades de cada línea, el análisis revela un patrón transversal: la fortaleza técnica de INGISOT SAS descansa en la experticia de su equipo humano y en el uso de herramientas convencionales, pero esta fortaleza se convierte también en una limitación cuando los volúmenes de información aumentan, los tiempos de respuesta se reducen y la complejidad de los proyectos exige escalabilidad y reproducibilidad.

Es relevante señalar que INGISOT SAS ya ha explorado el uso de elementos básicos de ciencia de datos, aunque de manera inicial, fragmentada y sin un marco estandarizado. Este antecedente, lejos de ser una debilidad, constituye un punto de apoyo para la transición: permite reconocer el potencial de la analítica en la optimización de flujos de trabajo y plantea la necesidad de organizar, sistematizar y escalar su aplicación. En este sentido, el presente diagnóstico no solo constituye un retrato fiel del estado actual de los procesos, sino también un punto de partida para su transformación.

La identificación de subactividades críticas y de oportunidades de mejora mediante ciencia de datos abre la posibilidad de evolucionar hacia procesos semi-automatizados, estandarizados y escalables. Este será el eje central del capítulo siguiente, en el cual se diseña un modelo de analítica de datos para INGISOT SAS, orientado a optimizar los flujos

de trabajo, reducir la dependencia de operaciones manuales y fortalecer la capacidad institucional de responder con mayor eficiencia, rigor y competitividad en el mercado de la consultoría ambiental.

Este diagnóstico permitió visibilizar los cuellos de botella, la dispersión de formatos y la necesidad de estandarización, estableciendo la base empírica para la construcción del modelo analítico propuesto.

---

## 8. Modelo de datos para INGISOT SAS

### 8.1. Propuesta del modelo

La metodología propuesta establece un marco de trabajo integral para la construcción de mapas generales y específicos que guíen la ejecución de proyectos de analítica y ciencia de datos. Este modelo se fundamenta en la necesidad de estandarizar el proceso de toma de decisiones técnicas y metodológicas, permitiendo que cada subactividad sea abordada con criterios claros, replicables y alineados con las mejores prácticas de la industria.

El modelo surge como respuesta a la complejidad inherente a los proyectos de datos, donde la diversidad de objetivos, tipos de información y contextos de aplicación requiere un enfoque sistemático que permita seleccionar, de manera fundamentada, las técnicas, herramientas y metodologías más apropiadas para cada caso particular. A diferencia de enfoques prescriptivos rígidos, esta propuesta adopta un carácter adaptativo, reconociendo que no existe una solución única aplicable a todos los escenarios.

El valor diferencial de este modelo radica en su capacidad para integrar múltiples dimensiones del desarrollo de proyectos analíticos: desde la identificación de la naturaleza del problema hasta la definición de estrategias de despliegue, pasando por la selección de ciclos de vida metodológicos y marcos de gestión ágil. Esta integración holística asegura coherencia entre los distintos componentes del proyecto y facilita la comunicación entre equipos técnicos y stakeholders.

Adicionalmente, el modelo contempla la construcción de dos niveles de abstracción complementarios: el **mapa general**, que funciona como marco conceptual y guía de referencia universal, y los **mapas específicos**, que materializan las decisiones concretas para cada línea de trabajo o subactividad particular. Esta dualidad permite mantener la consistencia metodológica mientras se respeta la especificidad de cada contexto.

### 8.2. Elementos del modelo

El modelo de analítica de datos propuesto para INGISOT SAS se compone de un conjunto articulado de elementos metodológicos (azul), tecnológicos (verde) y organizacionales (magenta) que permiten integrar la ciencia de datos en los procesos

---

operativos de la empresa. En el nivel metodológico se estructura el ciclo de vida analítico mediante marcos como CRISP-DM, SEMMA, TDSP y MLOps, los cuales guían las fases de comprensión del negocio, preparación, modelado, evaluación y despliegue. Estos se combinan con metodologías ágiles como Scrum y Kanban, asegurando una gestión iterativa y colaborativa que facilita la adaptación del modelo a diversas líneas de servicio ambiental, catastral, geológica y territorial.

En el nivel tecnológico, el modelo se sustenta en un ecosistema integrado de lenguajes, frameworks y herramientas (Python, R, Scikit-learn, TensorFlow, Power BI, PostgreSQL/PostGIS, MLflow, Docker) que permiten desarrollar, versionar y desplegar modelos descriptivos, predictivos y prescriptivos. Este entorno se conecta con infraestructuras locales o en la nube, sistemas de bases de datos espaciales y plataformas de inteligencia de negocio, asegurando la trazabilidad de los datos y la automatización de los flujos de trabajo. La estandarización de scripts, datos, modelos y visualizaciones garantiza la reproducibilidad de los resultados y la integración entre áreas técnicas.

Finalmente, el componente organizacional articula los roles humanos, los recursos financieros y los mecanismos de seguimiento necesarios para consolidar la adopción tecnológica. La interacción entre estos elementos metodología, tecnología y organización permite que el modelo sea funcional, escalable y aplicable, transformando los procesos de INGISOT en una operación basada en evidencia, automatizada y orientada a la mejora continua.

### **Figura 19 Modelo de datos a implementar**

MAPEO PROCESOS	CLIENTE	EQUIPO SIG	ANALISIS DATOS	LIDER DE PROYECTO				
COMPARACION & SELECCION	MODELOS DESCRIPTIVO	MODELOS PREDICTIVO	MOD PRESCRIPTIVO					
	Análisis exploratorio de datos (EDA) (Análisis estadístico descriptivo)  Análisis de correlación y covarianza (Coeficiente de Pearson, Kendall, matrices)  Análisis de series temporales (Medias móviles y autocorrelación)  Análisis espaciales (identify, union, overlay)	Modelos clasificación (Knn, Svm, Arboles de decision, XGBoost, RNN)  Modelos de regresión (Lineal, polinómica, Random forest, SVR)  Modelos de clustering (K-Means, PCA Analisis de componentes, Autoencoders)  Modelos de series temporales (ARIMA, SARIMA, SARIMAX, Transformes)	Modelos de optimización (Programación lineal, mixta, cuadrática, heurística)  Modelos de reducción dimensional (Análisis de componentes independientes ICA)  Modelos de procesamiento NPL (Modelos de sentimientos, chatbots - transformes o LLMs)  Modelos de simulación y escenarios (Monte carlo, agentes & dinámica de sistemas)					
Lenguajes, Frameworks, librerías, ETL, Almacenamiento, Business Intelligence (BI), MLOps (Python & R)(Scikit-learn, TensorFlow, PyTorch) (numpy, pandas, matplotlib, geopandas, ArcPy, seaborn) (PostgreSQL/PostGIS, MongoDB)(Power BI, Tableau)(MLflow, DVC, Docker)								
CICLO DE VIDA	CRISP-DM	SEMMA	KDD	TDSP	OSEMN	MLOps		
	BU Business Understanding DP Data understanding DT Data preparation DM Modeling IE Evaluation IE Deployment	S E M M A  Sample Explore Modify Model Assess	DS Data selection DP Data preprocessing DT Data transformation DM Data mining IE Interpretation/Eval	BU Business Understanding DAU Data acquisition M Modeling D Deployment CA Customer Acceptance	O Obtain - data S Scrub - data E Explore - data M Model - data N Interpret - results	Design Pipeline / Workflow  		
METODOLOGIAS AGILES	Requirements analysis (RA), System design (SD), Implementation (IMP), Verification & testing (VT), Maintenance  RA → SD → IMP → VT → MNT  Lineal & predictivo →		Project vision (PV), Iteration planning (IP), Development execution (DEV), Testing review (T&R), Retrospective adaptation (RA)  PV → IP → DEV → T&R → RA  Iterativo e incremental		Lista de tareas (PB), planificación sprint (SP) ciclos de desarrollo (S), Reunión diaria (DS) Revisión del sprint (SR), Retrospección (R)  PB → SP → S → DS → SR  Ciclico e iterativo RETRO		Visualize workflow (vw), limite word in progress (WIP), Manage flow (MF), Make process (MPE), Implement feedback loops (FBL), Improve collaboratively (IC)  VW → WIP → MF → MPE → FBL + IC  Flujo continuo →	
	WATERFALL		AGILE		SCRUM - (Job teams)		KANBAN	
RESULTADOS	CODIGO	DATOS	MODELOS	VISUALIZACION	DOCUMENTOS	METADATOS		
	scripts (.py, .py, .js) Integración con APIs o GEE	Datos raw Datos limpios (.csv, .parquet) Bases de datos (.gdb, .gpkg) Imágenes (.geotiff, .grid, .tiff)	Modelos (.pkl, .h5, .joblib) Métricas (.csv, .pdf, .png) Curvas evaluación (.png, .pdf) Matriz de confusión (.png, .csv)	Tableros (.pbix, .twd) Gráficos (.png, .pdf, .jpg) Mapas temáticos (.geojson, .shp)	Memorias (.pdf, .docx) Archivo README (.md) Archivo de entorno (.yml, .txt) Informes técnicos (.pdf, .docx)	Metadatos (.xlsx, .json, .xml) Logs o bitacorras (.txt, .log)		
DESPLIEGUE	Entornos (Local vs producción)	Docker	AWS (SageMaker, Lambda)	CI/CD (GitHub Actions, Gitlab)	Servidor on-premise	API REST (FastAPI, Django)		
	Versionamiento de modelos	Compose (modelo + API + BD)	Azure (Machine learning studio)	Versionamiento (MLflow, DVC)	Maquina virtual	Microservicios (Docker + Kuber)		
Pipelines	Virtualenv / conda	Google Cloud (Vertex AI, Cloud Compute Engine)	Monitoreo	Servicios de bases de datos	Dashboards o apps			
APIs		IA (Gemini, gpt, claude)	Automatiza el ciclo de vida	API REST				
				Sistema (Linux / Windows server)				
	FUNDAMENTOS	CONTENEDORES	NUBE & SERV	MLOPs	ARQUITECTURA	EXPOSICIÓN		
ADOPCIÓN	I. Preparación Organizacional 		Gerencia (\$ presupuesto) Dirección de proyectos (Alineación de objetivos) Lider tecnico Personal técnico esp.		Humanos Tecnológicos Financieros	Reducción porcentual tiempos Grado de automatización Nivel de adopción tecnológica Seguimiento al plan de inversión		
	FASES				COMPONENTE ORGANIZACIONAL		RECURSOS Y SEGUIMIENTO	

Fuente: Elaboración propia (2025)

---

La Figura 19 se concibe como un diagrama integrador y flexible que sintetiza los principales componentes conceptuales, metodológicos y tecnológicos involucrados en la aplicación de la ciencia de datos a los flujos de trabajo de la consultoría ambiental. Su propósito no es representar una secuencia rígida de ejecución paso a paso, sino servir como un marco de referencia que permita identificar, a partir del análisis de la necesidad, los grupos de herramientas, metodologías y enfoques más adecuados según el contexto del proyecto, el tipo de datos y el nivel de madurez organizacional. En este sentido, la figura orienta la selección progresiva de un flujo de trabajo ideal para materializar automatizaciones, desarrollar código reproducible y generar productos funcionales, los cuales son validados mediante pruebas piloto y procesos iterativos de desarrollo, en coherencia con el enfoque aplicado del modelo propuesto.

### **8.2.1. Actividades o subactividades**

El modelo propuesto para INGISOT SAS parte de la identificación precisa de las actividades y subactividades dentro de cada servicio técnico o consultoría. Estas representan la unidad mínima de análisis y constituyen el punto de partida para la estructuración del flujo analítico. Cada subactividad como depuración de datos, análisis espacial, fotogrametría o generación de metadatos se define en función de los requerimientos temáticos, espaciales y alfanuméricos establecidos en el proyecto.

El objetivo de esta etapa es reconocer la naturaleza analítica de cada proceso (descriptivo, predictivo o prescriptivo), pues esta clasificación orienta el tipo de datos, los métodos y las herramientas que se emplearán. Así, un proceso descriptivo podrá centrarse en análisis exploratorios y visualización; uno predictivo, en estimación de tendencias o riesgos; y uno prescriptivo, en la optimización de recursos o generación de escenarios de decisión. Esta primera definición es clave para estructurar las siguientes fases del modelo.

### **8.2.2. Tipología de modelos**

La selección del tipo de modelo constituye una decisión técnica crucial que depende directamente de la naturaleza de la subactividad identificada, la estructura de los datos disponibles y los objetivos específicos del análisis. Siguiendo el marco conceptual de (Bishop, 2006) y las taxonomías establecidas en la literatura de aprendizaje automático.

El modelo mantiene una arquitectura flexible que permite seleccionar técnicas de análisis según la naturaleza de los datos y el propósito de cada subactividad. En el ámbito ambiental y geográfico, se emplearán modelos de regresión y series temporales (ARIMA, Prophet, Random Forest) para proyectar variables como calidad del agua, deforestación o precipitación. En los procesos catastrales, se aplicarán modelos de clasificación y clustering (árboles de decisión, k-means, PCA) para identificar inconsistencias, segmentar predios o caracterizar tipologías constructivas.

Los modelos prescriptivos y de optimización se integrarán cuando sea necesario evaluar escenarios o asignar recursos (por ejemplo, priorización de visitas técnicas o simulación de cobertura de muestreo). Cada modelo incorporará métricas de desempeño específicas como precisión, RMSE, F1-score o error relativo y un umbral de aceptación definido por proyecto, asegurando comparabilidad y trazabilidad entre fases.

### **8.2.3. Ciclo de vida**

El ciclo de vida metodológico del modelo no se restringe a una sola estructura como CRISP-DM o TDSP, sino que se adapta a la naturaleza del servicio y a las fases del mapeo de procesos.

Las etapas generales son:

1. Compilación y depuración de datos (verificación de calidad, formato, completitud e integridad);
2. Inventario de necesidades y consolidación temática, donde se definen variables, fuentes y criterios de integración;
3. Tratamiento analítico y modelado, que incluye análisis espacial, procesamiento raster y vectorial, cálculo de índices y generación de modelos cartográficos;
4. Validación y consolidación de resultados, donde se verifican métricas técnicas y consistencia temática; y
5. Publicación e integración en infraestructura de datos, incluyendo estandarización, generación de metadatos y almacenamiento en bases geospaciales (GPKG, GDB o MapServer).

Cada fase puede incorporar técnicas propias de diferentes marcos (SEMMA, OSEMN, TDSP, MLOps) según el alcance del proyecto. Lo esencial es mantener una secuencia

---

lógica, trazable y replicable que permita pasar de datos dispersos a conocimiento estructurado y utilizable por la organización.

#### **8.2.4. Metodologías ágiles**

La gestión del modelo se apoya en principios ágiles y de mejora continua, adaptando el enfoque según el tipo de tarea. Para proyectos con objetivos definidos y entregas progresivas (por ejemplo, entrenamiento de modelos predictivos o Integración de bases ambientales), se aplicará una estructura Scrum, con sprints cortos, planificación iterativa y revisiones conjuntas con la dirección de proyectos. Los procesos de mantenimiento, control de calidad y actualización de información se gestionarán mediante Kanban, asegurando flujo continuo y priorización visual de tareas.

De forma transversal, se incorporan prácticas de MLOps y DevOps, que garantizan el versionamiento de modelos, la automatización de pipelines, la integración continua y el monitoreo operativo. Esta flexibilidad metodológica asegura que el modelo evolucione junto con las necesidades técnicas y organizacionales de INGISOT, permitiendo escalar desde prototipos hasta sistemas analíticos productivos.

#### **8.2.5. Resultados esperados**

Los resultados esperados del modelo se traducen en la mejora verificable de los procesos técnicos y en la consolidación de una cultura basada en datos dentro de la empresa.

Entre los productos más relevantes se incluyen:

- Bases limpias y estructuradas, listas para análisis espacial o estadístico.
- Modelos reproducibles y validados, con documentación técnica, scripts y métricas estandarizadas.
- Dashboards e informes ejecutivos, que permitan interpretar resultados en tiempo real.
- Sistemas de apoyo a la decisión, aplicados a la planeación, priorización y seguimiento de proyectos.
- Repositorios geospaciales consolidados, con metadatos interoperables y trazabilidad completa.

---

El éxito de cada proyecto se evaluará mediante indicadores técnicos (precisión, error, tiempo de procesamiento) y organizacionales (reducción de tiempos, automatización, grado de adopción). El umbral de aceptación se definirá según la naturaleza del análisis, garantizando equilibrio entre rigor técnico y aplicabilidad operativa.

#### **8.2.6. Despliegue**

El despliegue constituye la transición del modelo analítico hacia la operación real dentro de los flujos de trabajo de INGISOT SAS. Se implementará bajo un enfoque modular y progresivo, permitiendo que cada componente modelos, scripts o bases— sea incorporado gradualmente según su nivel de madurez.

Los entornos de ejecución se gestionarán mediante contenedores Docker y pipelines automatizados (MLflow, DVC), asegurando versionamiento y reproducibilidad. Las soluciones se integrarán con las plataformas SIG y los tableros de Power BI a través de APIs RESTful, garantizando interoperabilidad con los sistemas de gestión interna.

El monitoreo de desempeño y la detección de desviaciones (drift) permitirán ajustar modelos de forma continua, consolidando un ciclo de mejora sostenida. El criterio de aceptación final será la estabilidad, utilidad y eficiencia del modelo en la operación diaria: un sistema analítico se considerará exitoso cuando aporte valor verificable a la gestión técnica y reduzca los tiempos de respuesta frente a los procesos tradicionales.

**Ejemplo 1 – Proceso ambiental:** Durante la evaluación de calidad del agua, el modelo inicia con la integración de registros históricos, depuración de datos de laboratorio y cálculo de índices fisicoquímicos. Se aplican modelos de regresión y análisis temporal para identificar tendencias y umbrales críticos. El despliegue se realiza en Power BI mediante dashboards conectados a una base PostgreSQL/PostGIS actualizada en tiempo real.

**Ejemplo 2 – Proceso catastral:** Para detectar inconsistencias entre información física y jurídica, se ejecuta una rutina automatizada que combina geoprocursos (QGIS/ArcPy) con un modelo de clasificación Random Forest entrenado sobre atributos catastrales. Las

alertas de posibles errores se visualizan en un tablero de control, y los resultados se almacenan en la base geoespacial institucional (GPKG), permitiendo trazabilidad y validación continua.

El modelo resultante integra componentes conceptuales, tecnológicos y organizacionales, constituyendo un marco adaptable que guía la transición de INGISOT SAS hacia una operación basada en evidencia y analítica reproducible.

---

## **9. Plan de intervención – propuesta de implementación.**

La propuesta de implementación del modelo de analítica de datos para INGISOT SAS se estructura como una estrategia progresiva orientada a integrar la analítica descriptiva, predictiva y prescriptiva dentro de los flujos de trabajo operativos de la empresa. Este proceso busca traducir el modelo diseñado en una herramienta efectiva de gestión, capaz de optimizar la planeación, ejecución y seguimiento de los proyectos de consultoría ambiental.

La adopción del modelo se plantea bajo principios de escalabilidad técnica, sostenibilidad organizacional y alineación estratégica, garantizando que los avances en analítica de datos sean coherentes con las capacidades institucionales y las metas de transformación digital de INGISOT SAS. En este sentido, se promueve una implementación gradual que permita la apropiación tecnológica y metodológica del modelo por parte de los equipos técnicos y administrativos.

El enfoque propuesto reconoce que la madurez analítica de la empresa se encuentra en una fase inicial, caracterizada por la dependencia de herramientas tradicionales (hojas de cálculo y archivos locales), lo que limita la integración de datos, la trazabilidad y la automatización. Por tanto, la estrategia de implementación se concibe como un proceso de mejora continua, en el cual la formación del talento humano, la infraestructura tecnológica y la gestión del conocimiento son elementos centrales para el éxito del modelo.

### **9.1. Fases de la implementación.**

La implementación del modelo se desarrollará en tres fases secuenciales e iterativas, que responden al ciclo de vida de los datos y al grado de madurez digital de la organización.

**Figura 20 Fases para la implementación propuestas**



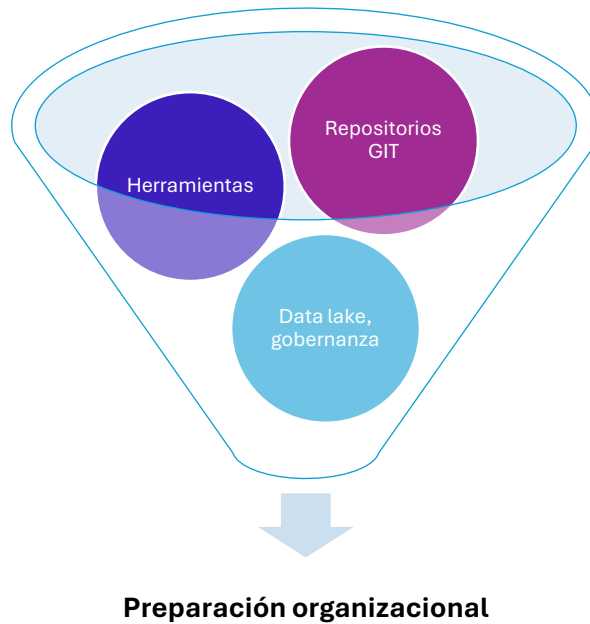
Fuente: Elaboración propia (2025)

### **9.1.1. Fase I. Preparación organizacional y técnica**

Esta fase contempla las acciones iniciales para preparar la empresa en términos tecnológicos y humanos. Se recomienda:

- Desarrollar un plan de capacitación interna en fundamentos de analítica de datos, manejo de Python, Power BI y control de versiones con Git.
- Configurar un entorno de trabajo colaborativo que incluya un repositorio institucional en GitHub, un almacenamiento centralizado (Data Lake) y protocolos de gobernanza de datos.
- Definir una política de calidad de datos, orientada a estandarizar formatos, metadatos y nomenclaturas, con el fin de asegurar la interoperabilidad entre proyectos y líneas de negocio.

**Figura 21 Fases para la preparación organizacional**



Fuente: Elaboración propia (2025)

**Tabla 14 Fase I. Preparación e infraestructura inicial para la analítica de datos**

<b>Acción</b>	<b>Responsable</b>	<b>Duración estimada</b>	<b>Recursos mínimos</b>
Capacitación básica en ciencia de datos aplicada a los procesos de la empresa	Director técnico, personal técnico	Mediano plazo	Tiempo del equipo, herramientas de capacitación internas, documentación técnica
Definición y puesta en marcha de un repositorio de código y documentación	Personal técnico	Corto plazo	Python, Git, equipos de cómputo existentes
Organización inicial de fuentes de datos y estructura de almacenamiento	Personal técnico, director técnico	Mediano plazo	Almacenamiento en nube básica, lineamientos de organización
Implementación de una estructura básica de	Director técnico, personal técnico	Mediano plazo	Nube básica Microsoft, equipos existentes

Acción	Responsable	Duración estimada	Recursos mínimos
almacenamiento centralizado de datos			
Definición de lineamientos iniciales de gobernanza de datos	Director técnico	Corto plazo	Tiempo de análisis, criterios técnicos internos
Formulación de una política básica de calidad de datos	Director técnico, personal técnico	Mediano plazo	Lineamientos técnicos, revisión de prácticas actuales

Fuente: Elaboración propia (2025)

### 9.1.2. Fase II, Integración y validación del modelo

En esta etapa se propone la incorporación del modelo analítico en los procesos operativos prioritarios de INGISOT SAS. Las acciones incluyen:

- Implementar pipelines de procesamiento de datos que automaticen la limpieza, transformación y análisis mediante librerías como *pandas*, *geopandas* y *rasterio*. Partiendo del mapeo de procesos y realizando el traslado hacia el modelo de datos en el que se aborden las fases propuestas.
- Desarrollar tableros de control en Power BI para el monitoreo de proyectos ambientales, indicadores de avance y alertas tempranas de desviaciones.
- Realizar una prueba piloto en una línea estratégica (por ejemplo, consultoría ambiental o SIG “series de tiempo”), donde se evalúe la eficacia del modelo en la reducción de tiempos de entrega y mejora de la trazabilidad.

**Figura 22 sub-fases para la integración y validación del modelo.**



Fuente: Elaboración propia (2025)

### **Medición del piloto y criterios de éxito**

La evaluación del piloto propuesto en la Fase II se realizará mediante la comparación de los tiempos de ejecución del análisis antes y después de la incorporación del modelo de analítica de datos, considerando rangos de tiempo habituales para un mismo tipo de proceso. La reducción de tiempos se evidenciará a partir de la automatización de tareas repetitivas previamente ejecutadas de forma manual, tomando como referencia la capacidad del modelo para replicar el análisis sin redefinir la lógica de procesamiento.

La mejora en la trazabilidad se abordará como un componente de la calidad técnica del análisis, evaluada a partir de la relación clara entre los datos utilizados, el código desarrollado y los resultados obtenidos. En este sentido, la trazabilidad se verificará mediante la capacidad de repetir el análisis de forma consistente, manteniendo la estructura del proceso independientemente del volumen o la fuente de los datos.

El piloto se considerará exitoso si el modelo demuestra escalabilidad operativa, entendida como la posibilidad de aplicar el mismo esquema analítico a uno o varios conjuntos de datos con características similares, reduciendo el esfuerzo manual y asegurando la reproducibilidad del análisis. El énfasis de esta fase será demostrativo y replicable, en coherencia con el alcance del trabajo y con los objetivos planteados

#### **9.1.3. Fase III, Optimización y escalamiento**

Una vez validado el modelo, se procederá a su consolidación institucional. Esta fase busca incorporar módulos de analítica avanzada y fortalecer la capacidad interna de interpretación y decisión basada en datos. Se recomienda:

- Implementar algoritmos de **machine learning** para predicción de variables ambientales o de desempeño operativo.
- Integrar el modelo dentro de un esquema de **MLOps** que garantice la trazabilidad, control de versiones y actualización periódica.
- Establecer un sistema de indicadores de desempeño analítico (**KPI's**) para monitorear la eficiencia, la productividad y el impacto del modelo.

---

## 9.2. Componente técnico

La implementación técnica se sustentará en una arquitectura híbrida que combine infraestructura local con servicios en la nube.

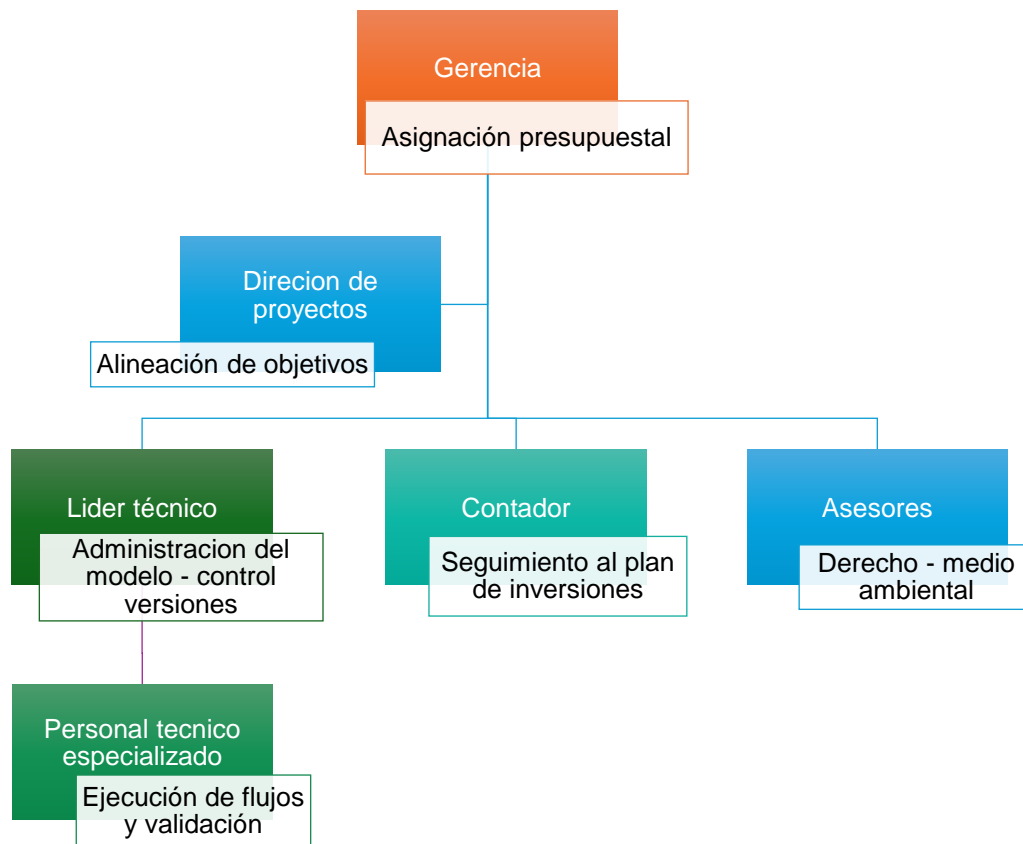
- Infraestructura de datos: almacenamiento estructurado en bases PostgreSQL con extensión PostGIS, y respaldo en nube mediante servicios como Google Cloud o Azure.
- Lenguajes y librerías: Python (pandas, geopandas, scikit-learn), R y SQL, permitiendo el desarrollo de flujos reproducibles y escalables.
- Automatización de procesos: uso de pipelines integrados y notebooks colaborativos (Jupyter o Colab) para la ejecución de scripts de análisis.
- Visualización y comunicación de resultados: implementación de tableros dinámicos en Power BI con conexión directa a las fuentes de datos procesadas.
- Control de versiones: adopción de Git y GitHub Actions para la gestión del código, documentación y seguimiento de cambios en tiempo real.

## 9.3. Componente organizacional

La adopción del modelo exige una estructura organizacional adaptativa que combine liderazgo estratégico con competencias técnicas. Para ello se proponen los siguientes roles:

- **Gerencia General:** liderazgo estratégico y asignación presupuestal, asegurando la integración del modelo con la planeación corporativa.
- **Dirección de Proyectos:** enlace entre las áreas técnicas y la gerencia, velando por la alineación entre los objetivos analíticos y los requerimientos contractuales.
- **Líder técnico en analítica de datos:** responsable de la administración del modelo, control de versiones, documentación y supervisión de procesos analíticos.
- **Equipo operativo y técnico:** encargados de la ejecución de los flujos de trabajo, interpretación de resultados y validación de productos intermedios.

**Figura 23 Roles dentro del componente organización y su relación con el modelo.**



Fuente: Elaboración propia (2025)

La gestión del cambio será un eje transversal de la implementación. Se sugiere conformar un comité interno de transformación digital, responsable de monitorear los avances, resolver obstáculos técnicos y promover la cultura de innovación analítica dentro de la organización.

#### 9.4. Recursos requeridos

La ejecución del modelo requiere tres tipos de recursos:

- **Humanos:** fortalecimiento del equipo actual con un analista de datos, un científico de datos junior y un ingeniero SIG especializado en automatización.
- **Tecnológicos:** estaciones de trabajo de alto rendimiento, servidores locales, licencias de Power BI Pro y servicios de nube para almacenamiento y procesamiento.
- **Financieros:** inversión inicial estimada en un 8-10 % del presupuesto anual destinado a innovación, cubriendo capacitación, adquisición de hardware y contratación de servicios tecnológicos.

---

A la presente INGISOT SAS, cuenta con servicios de almacenamiento en la nube, y un repositorio GIT. así como 3 estaciones de trabajo y licencias de servicios tecnológicos básicos como Microsoft y ArcGIS PRO.

#### **9.5. Seguimiento y evaluación.**

Para garantizar la sostenibilidad del modelo se propone un esquema de seguimiento basado en indicadores de desempeño (KPI), tales como:

- Reducción porcentual en tiempos de procesamiento de datos.
- Grado de automatización de procesos.
- Nivel de adopción tecnológica por parte de los equipos.
- Número de proyectos que integran componentes analíticos.

El modelo deberá ser revisado semestralmente, incorporando mecanismos de retroalimentación continua a través del ciclo de vida, lo cual permitirá ajustes metodológicos y técnicos de acuerdo con la evolución de la empresa y los requerimientos del sector ambiental y líneas estratégicas en general.

La implementación del modelo de analítica de datos en INGISOT SAS constituye un paso decisivo hacia la consolidación de una cultura organizacional basada en evidencia. La aplicación progresiva de los componentes técnicos y organizacionales permitirá fortalecer la eficiencia operativa, reducir la incertidumbre en la toma de decisiones y aumentar la competitividad de la empresa dentro del sector de la consultoría ambiental.

El modelo, además de mejorar la trazabilidad y el control de los procesos, ofrece una oportunidad para posicionar a INGISOT SAS como referente en la adopción de soluciones analíticas dentro del ámbito ambiental, contribuyendo a la transformación digital del sector y al desarrollo sostenible del territorio.

#### **9.6. Plan de inversión.**

El plan de inversión constituye un componente esencial dentro de la propuesta de implementación del modelo de analítica de datos, dado que permite traducir la estrategia técnica y organizacional en un marco financiero realista y sostenible. En el contexto de INGISOT SAS, su relevancia radica en que garantiza la disponibilidad progresiva de recursos humanos, tecnológicos y formativos necesarios para adoptar el modelo diseñado,

asegurando que la transición hacia una cultura de gestión basada en datos se realice de manera ordenada y con respaldo institucional.

En términos estructurales, el plan de inversión representa la hoja de ruta financiera del proceso de transformación analítica de la empresa. Cada etapa de implementación capacitación, adquisición de infraestructura, automatización y consolidación requiere de recursos que, de no planificarse anticipadamente, podrían limitar el alcance del modelo o interrumpir su desarrollo. Por tanto, el plan no solo cuantifica los costos asociados, sino que también sincroniza la ejecución técnica con la capacidad presupuestal y operativa de INGISOT SAS.

Con base en la información de los estados financieros de INGISOT SAS para el año 2024, se observa que la empresa cuenta con una capacidad operativa y financiera sólida, caracterizada por:

**Ingresos operacionales estables** provenientes de la ejecución de contratos de consultoría ambiental.

**Liquidez positiva**, con márgenes que permiten cubrir los gastos operativos y disponer de excedentes moderados.

**Ausencia significativa de deuda corriente**, lo cual otorga margen para destinar recursos a inversión sin comprometer el flujo de caja.

A partir de estos elementos, es posible proyectar la **capacidad de provisión e inversión progresiva** para la implementación del modelo analítico durante tres años, de acuerdo con los porcentajes razonables de reinversión recomendados para empresas de servicios técnicos (entre el 4 % y el 6 % de los ingresos anuales).

**Tabla 15 Propuesta de inversión progresiva**

Etapa / Año	Componente	Descripción	Inversión estimada (COP)	Resultado esperado
<b>Año 1 (2025)</b>	Capacitación y fortalecimiento técnico	Realización de <b>2 cursos estratégicos Anuales</b> : • <i>Curso 1</i> : “Herramientas de análisis espacial con Python aplicado” 30 horas UNIANDES. • <i>Curso 2</i> : “Análisis geoespacial avanzado con software SIG y deep learning”.	\$3,350,000	Personal técnico con competencias en analítica, SIG y automatización de flujos de trabajo.

<b>Año 1 (2025)</b>	Infraestructura tecnológica inicial	Adquisición de un servidor local de respaldo + almacenamiento externo (4 TB), instalación de entornos de desarrollo (Anaconda, PostgreSQL/PostGIS, Power BI Pro).	\$10,910,000	Configuración de entorno de datos y repositorio interno.
<b>Año 1 (2025)</b>	Software y licencias	Licencias ArcGIS PRO (2 usuarios x \$75.000/mes x 12 meses) + servicios en la nube (Google Cloud / Azure – cuota mensual básica).	\$6.000.000	Disponibilidad de entornos colaborativos y paneles interactivos.
<b>Año 2 (2026)</b>	Automatización y despliegue del modelo	Desarrollo de scripts en Python para limpieza, normalización y visualización de datos; integración con dashboards Power BI.	\$5,000,000	Primer flujo automatizado y tablero operativo funcional.
<b>Año 2 (2026)</b>	Formación complementaria	Curso corto o diplomado: “ <i>Machine Learning aplicado a gestión ambiental y predicción de indicadores de desempeño</i> ” (mínimo 40 horas).	\$3,000,000	Integración de analítica predictiva en procesos de consultoría.
<b>Año 2 (2026)</b>	Software y licencias	Licencias para el procesamiento de datos geofísicos	\$16,000,000	Procesamiento de datos de resistividad eléctrica, para interpretación y posprocesamiento
<b>Año 2 (2026)</b>	Capacitación y fortalecimiento técnico	Realización de <b>2 cursos estratégicos Anuales</b> : • <i>Curso 1</i> : “Inteligencia de negocios con Power BI” 15 horas UNIANDES. • <i>Curso 2</i> : “Análisis geoespacial avanzado con software SIG y deep learning”.	\$4,000,000	Personal técnico con competencias en analítica, SIG y automatización de flujos de trabajo.
<b>Año 2 (2026)</b>	Hardware especializado	Estaciones de trabajo de alto rendimiento (2 equipos, GPU intermedia, 32 GB RAM).	\$10,000,000	Capacidad de procesamiento local y entrenamiento de modelos
<b>Año 3 (2027)</b>	Consolidación y mantenimiento del sistema analítico	Monitoreo MLOps, integración con GEE o plataformas web, actualización de librerías y políticas de seguridad de datos.	\$4,000,000	Modelo operativo estable, automatizado y documentado.
<b>Año 3 (2027)</b>	Asesoría o acompañamiento externo	Consultoría de corto plazo con experto en implementación de arquitecturas de datos y gobernanza analítica	\$5,000,000	Transferencia de conocimiento y auditoría técnica
			<b>\$61,260,000</b>	

Fuente: Elaboración propia (2025)

La inversión propuesta representa un esfuerzo financiero controlado y sostenible, equivalente a un promedio del 2,1 % de los ingresos anuales durante el trienio. Este nivel de compromiso financiero se encuentra por debajo del umbral de riesgo operativo (5 %), lo cual garantiza que INGISOT SAS puede asumir el proceso de adopción tecnológica sin afectar la liquidez ni comprometer sus obligaciones contractuales.

---

El modelo de provisión progresiva también permite que cada fase de inversión genere retorno operativo medible, ya sea en reducción de tiempos, incremento de productividad o disminución de costos indirectos. A partir del segundo año, se prevé que los beneficios derivados de la automatización y la analítica avanzada compensen parcialmente los desembolsos iniciales, fortaleciendo la sostenibilidad financiera del modelo.

Adicionalmente, se recomienda que la empresa registre estas inversiones como activos intangibles y tecnológicos, siguiendo las disposiciones de la NIC 38, lo que permitiría capitalizar los desarrollos analíticos como parte del valor organizacional de INGISOT SAS, y reflejar su aporte en la productividad y competitividad de la compañía.

---

## Conclusiones y recomendaciones

El diagnóstico realizado en INGISOT SAS demuestra que el modelo de analítica de datos propuesto es técnicamente viable y alineado con la capacidad operativa de la empresa, dado que cuenta con un equipo multidisciplinario calificado, acceso a herramientas de código abierto de bajo costo (Python, R, PostgreSQL con extensión PostGIS) y una infraestructura híbrida que permite escalabilidad. Además, su diseño se encuentra en concordancia con estándares internacionales de gestión analítica, como CRISP-DM y MLOps, lo que garantiza compatibilidad con prácticas globales. No obstante, la principal barrera para su implementación exitosa no es de carácter tecnológico, sino cultural y organizacional. La transición hacia un enfoque *data-driven* exige la consolidación de una cultura analítica, la capacitación continua del personal, la adopción de protocolos claros de gobernanza de datos y la promoción de espacios de experimentación e innovación. Por tanto, el éxito del modelo dependerá tanto de su calidad técnica como del compromiso institucional con su adopción y sostenibilidad.

El análisis transversal de los flujos de trabajo reveló que el potencial de optimización mediante analítica de datos varía según la naturaleza y la madurez digital de cada línea de negocio de INGISOT SAS. Las áreas de hidrología y catastro presentan las mayores oportunidades de mejora inmediata, al permitir automatizar la depuración y normalización de datos, con reducciones de tiempo operativo estimadas entre un 35 % y un 40 %. En cambio, el componente de ordenamiento territorial, aunque requiere una inversión inicial más alta, ofrece el mayor impacto estratégico a mediano plazo mediante el desarrollo de modelos predictivos aplicados a escenarios de cambio climático y planificación territorial. En consecuencia, se recomienda iniciar la implementación del modelo mediante un programa piloto en la línea de hidrología, con el fin de validar la arquitectura técnica y generar evidencia de retorno antes de su escalamiento progresivo hacia las demás líneas. Esta estrategia garantiza una adopción gradual, medible y sostenible del enfoque analítico dentro de la organización.

El análisis bibliográfico permitió consolidar un marco teórico sólido y actualizado, en el cual se integran los fundamentos de la ciencia de datos, la minería de datos y la inteligencia artificial con las dinámicas de la gestión ambiental. Se abordaron modelos descriptivos,

---

predictivos y prescriptivos, así como metodologías estructuradas —CRISP-DM, SEMMA y TDSP— que aportan rigor al proceso de diseño. Esta revisión evidenció la convergencia entre los enfoques cuantitativos y cualitativos y su pertinencia para entornos técnicos interdisciplinarios.

A partir de los referentes identificados, se configuró una base conceptual que guía el modelo propuesto, garantizando la coherencia entre las fases analíticas, los lenguajes y las herramientas. El estudio demostró que la ciencia de datos aplicada al campo ambiental no solo optimiza procesos operativos, sino que fortalece la capacidad de análisis estratégico, el control de la calidad de los datos y la sostenibilidad técnica de las organizaciones.

El diagnóstico permitió caracterizar de manera precisa los procesos internos de INGISOT SAS, evidenciando tanto sus fortalezas técnicas como los vacíos estructurales en materia de sistematización, trazabilidad y automatización. A través del mapeo de procesos y las matrices analíticas, se identificaron cuellos de botella asociados a la dispersión de datos, la dependencia de herramientas manuales y la baja interoperabilidad entre líneas de trabajo. Este ejercicio proporcionó una visión integral de la organización, revelando su potencial para evolucionar hacia un esquema de gestión de datos más ágil y centralizado.

El resultado del diagnóstico fue fundamental para orientar el diseño del modelo, ya que permitió definir las áreas prioritarias de intervención, las oportunidades analíticas por la línea de negocio y las técnicas de ciencia de datos más apropiadas. En consecuencia, el diagnóstico no solo cumplió su propósito descriptivo, sino que se constituyó en el punto de partida empírico que justificó y direccionó la propuesta técnica presentada en los capítulos posteriores.

El modelo diseñado responde a un enfoque modular, integrando elementos metodológicos, tecnológicos y organizacionales que actúan de forma articulada. Se definieron tipologías de modelos analíticos (descriptivos, predictivos y prescriptivos) y ciclos de vida flexibles basados en CRISP-DM y MLOps, asegurando la reproducibilidad y la escalabilidad del proceso. Además, se incorporaron herramientas como Python, R, PostgreSQL/PostGIS y Power BI, que facilitan la implementación práctica del modelo en entornos de consultoría ambiental.

---

La propuesta plantea una arquitectura adaptable al nivel de madurez digital de INGISOT SAS, priorizando la estandarización de datos, la automatización de tareas repetitivas y la integración con infraestructuras geoespaciales. Este diseño constituye una innovación organizacional y metodológica, ya que traduce los principios de la ciencia de datos en un marco aplicable al contexto real de la empresa, con impacto directo sobre la calidad, la eficiencia y la competitividad técnica de sus operaciones.

La propuesta de implementación se estructuró en tres fases progresivas: preparación; integración y validación; y optimización y escalamiento, que garantizan una transición ordenada hacia una cultura basada en datos. Cada fase incorpora acciones técnicas, organizativas y financieras que permiten consolidar la adopción del modelo sin afectar la operatividad de la empresa. El plan de inversión asociado demuestra viabilidad económica y sostenibilidad, con recursos distribuidos de manera racional entre formación, infraestructura y automatización.

La estrategia propuesta fortalece la gobernanza de datos, define roles claros (gerencia, dirección de proyectos, liderazgo técnico y personal operativo) y establece mecanismos de seguimiento mediante indicadores de desempeño. Su ejecución progresiva permitirá que INGISOT SAS consolide capacidades analíticas internas y posicione su modelo como referente de transformación digital en la consultoría ambiental colombiana.

---

## Referencias

- Beck, K. B. (2001). *Manifiesto for Agile Software Development*. Obtenido de Agile Manifiesto: <https://agilemanifesto.org>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Chapman, P. C. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.
- Collica, R. (2011). *Customer segmentation and clustering using SAS enterprise miner. Second edition*. SAS Institute.
- Conway, D. (2010). *The Data Science Venn Diagram*. Drew Conway.
- Quantum Technologies LLC. (2025). *Feature Engineering for Modern Machine Learning with Scikit-Learn*. Packt Publishing.
- Davenport, T. H. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press.
- Egger, R., & Yu, J. (2022). Data science and interdisciplinarity. En R. Egger, *Applied data science in tourism* (págs. 31- 46). doi:[https://doi.org/10.1007/978-3-030-88389-8\\_3](https://doi.org/10.1007/978-3-030-88389-8_3)
- Ehrenmueller, M. (2022). *Data modeling with microsoft Power BI*. O'Reilly Media, Inc.
- Gandhi, R. (2022). *Head first Git, A learner's guide to understanding git from the inside out*. O'Reilly Media, Inc.
- Ganegedara, T. (2022). *TensorFlow in action*. Maning publicaciones.
- Gartner. (2019). *Agile in Data & Analytics Projects*. Gartner Research.
- George, N. (2021). *Practical Data Science with Python*. Packt Publishing.
- Grus, J. (2019). *Data science from scratch*. O'Reilly Media, Inc.
- Hazzan, O., & Mike, K. (2023). Data Science for Social Science and Digital Humanities Research. En E. O. (Ed.), *Guide to Teaching Data Science*. Springer. doi:[https://doi-org.bdbiblioteca.universidadean.edu.co/10.1007/978-3-031-24758-3\\_19](https://doi-org.bdbiblioteca.universidadean.edu.co/10.1007/978-3-031-24758-3_19)
- Hernández, R., Fernández, C., & Baptista, P. (2014). *Metología de la Investigación*. McGraw-Hill Education.

- 
- Hernández, S. (2022). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. (7a, Ed.) McGraw-Hill Education.
- Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly Media.
- Kolokolov, A., & Zelensky, M. (2024). *Data visualization with Microsoft Power BI*. O'Reilly Media, Inc.
- Kulkarni, A., Shivananda, A., & et al. (2023). *Introduction to prespective AI: A primer for decision Intelligence Solutioning with Pytoon*. Apress.
- Larose, C., & Larose, D. (2019). *Data science using Python and R*. John Wiley & Sons Inc.
- Laster, B. (2023). *Learning GitHub Actions*. O'Reilly Media, Inc.
- Li, Y., Xie, Y., & Sheaker, S. (2023). Spatial data science. En L. Rokach, O. Maimon, & E. Shmueli, *Machine learning for data science handbook* (págs. 401-422). Springer. doi:[https://doi.org/10.1007/978-3-031-24628-9\\_18](https://doi.org/10.1007/978-3-031-24628-9_18)
- Loeliger, J. (2022). *Version control with Git*. O'Reilly Media, Inc.
- Microsoft. (19 de 04 de 2025). *The future is yours*. Obtenido de <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/plan>
- Nathn, G. (2021). *Practical data science with Python: Learn tools and techniques from hands-on examples to extract insights from data*. Packt publishing. Obtenido de <https://www.packtpub.com/en-us/product/practical-data-science-with-python-9781801071970>
- Nelli, F. (2018). *Python data analytics: With pandas, NumPy, and Matplotlib* (Vol. 2nd ed). Apress.
- Nield, T. (2016). *Getting Started with SQL*. O'Reilly Media, Inc.
- Nield, T. (2022). *Essential math for data science*. O'Reilly Media, Inc.
- Oppel, A. (2009). *Data Modeling, A Beginner's Guide*. McGraw-Hill. Obtenido de <https://learning-oreilly-com.bdbiblioteca.universidadean.edu.co/library/view/data-modeling-a/9780071623988/>

- 
- Paper, D. (2019). *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*. Apress.
- Peixeiro, M. (2022). *Time series Forecasting in Python*. Manning Publications.
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media. Obtenido de <https://www.oreilly.com/library/view/data-science-for/9781449374273/>
- Royce, W. W. (1970). Managing the Development of Large Software Systems. *Proceedings of IEEE WESCON*, (págs. 1-9).
- Schenker, G. (2018). *Learn Docker - Fundamentals, of Docker 18.x*. Packt publishing.
- Shanmuga, R., & Paulchamy, B. (2023). The future of medicine in cognitive technologies and automatic detection using computational techniques. En K. Ram, & S. Karthic, *Translating healthcare through intelligent computational methods* (págs. 372 - 393). Springer. doi:[https://doi.org/10.1007/978-3-031-27700-9\\_23](https://doi.org/10.1007/978-3-031-27700-9_23)
- Shearer, C. (2000). The CRISP-DM model : : The new blueprint for data mining. *Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5 (4), 13-22.*
- Skoulikari, A. (2023). *Learning Git*. O'Reilly Media, Inc.
- Tanimura, C. (2021). *SQL for Data Analysis*. O'Reilly Media, Inc.
- Teate, R. (2021). *SQL for Data Scientists: A beginner's guide for building datasets for analysis*. Wiley.
- Toms, S., & Parker, B. (2022). *Python for ArcGIS Pro*. Pack Publishing.
- Toms, S., Van Rees, E., & Crickard, P. (2019). *Mastering geospatial analysis with Python* (Vol. 2nd ed). Packt Publishing.
- Tranquilin, M., Lakshmanan, V., & Tekiner, F. (2023). *Architecting Data and Machine Learning Platforms*. O'Reilly Media, Inc.
- VanderPlas, J. (2022). *Python Data Science Handbook* (Vol. 2nd Edition). O'Reilly Media, Inc.
- Vikas, K., Sanjeet, K., & Monica, B. (2004). *Cognitive Science, Computational Intelligence, and Data Analytics: Methods and Applications with Python*. (M. K.

---

(Elsevier), Ed.) Obtenido de <https://shop.elsevier.com/books/cognitive-science-computational-intelligence-and-data-analytics/khare/978-0-443-16078-3>

Witten, L., Frank, E., & et.al. (2016). *Data Mining* (Vol. 4th). Morgan Kaufman.