Logotipo

El contenido generado por IA puede ser incorrecto.



**Formato para la documentación de la arquitectura de proyecto de analítica**

“Registre aquí el nombre del proyecto de analítica”

**Proceso**

**Gestión Estratégica de Tecnologías de la Información**

**Versión 1**

**3/04/2025**

**TABLA DE CONTENIDO**

[1. INTRODUCCIÓN 3](#_Toc193197736)

[2. DESCRIPCIÓN DEL PROBLEMA 3](#_Toc193197737)

[3. OBJETIVOS DEL PROYECTO 3](#_Toc193197738)

[3.1. Objetivo General 3](#_Toc193197739)

[3.2. Objetivos Específicos 4](#_Toc193197740)

[4. INVENTARIO DE FUENTES DISPONIBLES SELECCIONADOS PARA EL PROYECTO 4](#_Toc193197741)

[5. NIVELES DE INTERLOCUCIÓN PREVISTOS POR EL PROYECTO 5](#_Toc193197742)

[6. ESTADO DEL ARTE 7](#_Toc193197743)

[7. FASES PARA EL DESARROLLO DEL PROYECTO 7](#_Toc193197744)

[7.1. Definición de Necesidades Analíticas 8](#_Toc193197745)

[7.2. Creación de la Vista Minable 9](#_Toc193197746)

[7.3. Entrenamiento de la Solución Analítica 12](#_Toc193197747)

[7.4. Evaluación de la utilidad 15](#_Toc193197748)

[7.5. Incorporar de la solución analítica a los sistemas de Monitoreo Ambiental 16](#_Toc193197749)

[8. MATRIZ DE RIESGOS Y RESTRICCIONES 17](#_Toc193197750)

[9. REPOSITORIOS DE CÓDIGO Y MODELOS USADOS DENTRO DEL PROYECTO 18](#_Toc193197756)

[10. VISUALIZACIONES Y API’S SOBRE LOS COMPONENTES DESARROLLADOS. 18](#_Toc193197757)

[11. CRONOGRAMA 18](#_Toc193197758)

[12. TÉRMINOS Y/O CONCEPTOS 19](#_Toc193197759)

[13. BIBLIOGRAFÍA 19](#_Toc193197760)

[14. ANEXOS 20](#_Toc193197761)

[15. DATOS DE PRODUCCIÓN DEL PROYECTO 20](#_Toc193197762)

# INTRODUCCIÓN

En la introducción, se debe presentar una visión general del proyecto, destacando su propósito, alcance y relevancia dentro del contexto del Ministerio de Ambiente y Desarrollo Sostenible. Es importante incluir una descripción del problema que motiva el desarrollo de la solución analítica, así como los objetivos que se buscan alcanzar. Además, se debe mencionar la metodología utilizada para abordar el problema, resaltando la aplicación del enfoque CRISP-DM y cualquier otro marco de referencia empleado. Finalmente, se debe ofrecer una breve descripción de la estructura del documento, explicando cómo están organizadas las secciones para guiar al lector a través del contenido.

# DESCRIPCIÓN DEL PROBLEMA

En esta sección, se debe proporcionar una descripción clara y concisa del problema identificado dentro del Ministerio de Ambiente y Desarrollo Sostenible, especificando el área afectada y su impacto en los procesos administrativos, financieros, normativos o ambientales. Se recomienda respaldar el análisis con datos concretos, como indicadores de desempeño, costos asociados, tiempos de ejecución o niveles de cumplimiento normativo. Además, se debe detallar cómo este problema afecta la gestión institucional, incluyendo los riesgos financieros, operativos o regulatorios que podrían surgir si no se implementa una solución adecuada.

# OBJETIVOS DEL PROYECTO

## Objetivo General

Se debe definir de manera clara y precisa el propósito general del proyecto, destacando su meta principal y el impacto esperado. El objetivo general debe describir los resultados esperados y responder a preguntas clave como: ¿qué se busca lograr?, ¿para qué se realiza?, ¿cuándo se espera alcanzar? y ¿dónde se llevará a cabo?

**Ej.** Diseñar e implementar una solución avanzada para el análisis, descripción y predicción de información relacionada con [problema] en [ubicación]. La solución integrará diversos factores [factores] que influyen en la problemática, con el fin de proporcionar un respaldo sólido basado en datos para la toma de decisiones en [acción específica].

## Objetivos Específicos

En esta sección, se deben formular objetivos específicos que aborden las causas del problema identificado y orienten la generación de información para la toma de decisiones. Se recomienda seguir esta estructura, pero pueden agregarse o ajustarse objetivos según el alcance del proyecto.

Ej.

**1. Reducción de la Incertidumbre en la Información**

Mejorar la disponibilidad y precisión de los datos relevantes para el [problema], reduciendo brechas de información, inconsistencias o falta de representatividad en las fuentes utilizadas.

**2. Identificación de Factores Clave que Influyen en el Problema**

Determinar qué variables o condiciones tienen mayor impacto en la ocurrencia y evolución del [problema], permitiendo focalizar estrategias de intervención más efectivas.

**3. Optimización de la Interpretación y Uso de Datos**

Transformar datos en conocimiento accionable, facilitando su comprensión por parte de las partes interesadas y minimizando errores en la toma de decisiones debido a interpretaciones erróneas o falta de contexto.

**4. Anticipación de Escenarios de Riesgo o Oportunidad**

Reducir la reacción tardía ante eventos adversos o cambios en las condiciones del problema, proporcionando mecanismos analíticos que permitan prever tendencias y mejorar la capacidad de respuesta.

**5. Evaluación de Alternativas para la Mejora de Procesos**

Identificar y validar estrategias basadas en datos que optimicen el desempeño de los procesos asociados al [problema], permitiendo mejorar su eficiencia, impacto o sostenibilidad.

**6. Fortalecimiento de la Transparencia y Trazabilidad de la Información**

Garantizar que los resultados analíticos sean replicables, verificables y comprensibles, minimizando sesgos, ambigüedades o interpretaciones subjetivas que afecten la credibilidad de la información.

# INVENTARIO DE FUENTES DISPONIBLES SELECCIONADOS PARA EL PROYECTO

Esta sección debe documentar un inventario detallado de todas las fuentes de datos seleccionadas para el proyecto, proporcionando una visión clara de su disponibilidad, características y posibles restricciones.

Enumerar todas las fuentes de datos seleccionadas, incluyendo únicamente su nombre. La descripción detallada de cada fuente debe diligenciarse en los formatos correspondientes.

**Nota**

Diligenciar los formatos de caracterización de las fuentes de datos (diligenciar por cada fuente de datos:

|  |  |
| --- | --- |
| **F-E-GET-24** | [Roles y responsabilidades del gobierno del dato](https://somosig.minambiente.gov.co/portal/document_tab.php?id_doc=1348&version=2&opcion_regreso=1) |
| **F-E-GET-26** | [Catálogo de componentes de información](https://somosig.minambiente.gov.co/portal/document_tab.php?id_doc=1384&version=1&opcion_regreso=1) |
| **F-E-GET-20** | [Diccionario de Datos](https://somosig.minambiente.gov.co/portal/document_tab.php?id_doc=1344&version=2&opcion_regreso=1) |

* **Identificar restricciones o situaciones especiales** que puedan afectar la disponibilidad de la información, como:
  + Necesidad de permisos o acuerdos de acceso
  + Trámites con entidades externas
  + Procesos de anonimización o adecuación de datos
* **Garantizar la disponibilidad en el Datalake del Ministerio:** Si la fuente ya está disponible, especificar su ubicación en el Datalake. Si no está disponible, documentar las acciones requeridas para su incorporación, asegurando su persistencia para cada pregunta de analítica, especialmente si proviene de fuentes externas o abiertas.
* **Adjuntar los siguientes formatos validados y aprobados por el líder funcional para cada fuente de datos:**
* **F-E-GET-24** – Roles y responsabilidades del gobierno del dato
* **F-E-GET-26** – Catálogo de componentes de información
* **F-E-GET-20** – Diccionario de datos
* **Compromiso de disponibilidad**: Especificar formalmente la fecha a partir de la cual se realizará la ingesta de las fuentes de datos al Datalake, considerando los tiempos de aprobación del proyecto.

# 

# NIVELES DE INTERLOCUCIÓN PREVISTOS POR EL PROYECTO

En esta sección, se deben documentar los niveles de interlocución previstos en el proyecto, identificando los actores clave, sus responsabilidades y los roles específicos que desempeñan. Esto permitirá establecer una estructura clara de comunicación y garantizar la coordinación efectiva entre los diferentes niveles de gestión y ejecución.

* **Identificar los roles mínimos establecidos en el procedimiento oficializado por la entidad.** Estos son los roles base, pero si el proyecto requiere actores adicionales, pueden incluirse siempre que su función esté justificada.
* **Registrar las personas o áreas responsables en cada nivel de interlocución.** Se debe indicar quién o qué dependencia asume cada rol dentro del proyecto.
* **Especificar las responsabilidades de cada nivel de interlocución.** Hay que asegurar que cada rol asignado contribuya a la ejecución eficiente del proyecto y facilite la toma de decisiones basada en datos.
* **Verificar que los responsables de cada nivel cuenten con acceso a las herramientas necesarias para cumplir su función.** Esto incluye bases de datos, plataformas de documentación (wiki OTIC), repositorios de código (GitLab) y demás sistemas involucrados en la gestión del proyecto.
* **Mantener actualizada esta información.** Cualquier cambio en la estructura de interlocución debe ser documentado y comunicado a los involucrados.

A continuación, se presenta la estructura mínima requerida para la documentación de los niveles de interlocución:

Tabla *1* Niveles de interlocución previstos

| **Nivel de Interlocución** | **Responsabilidades** | **Roles Específicos** |
| --- | --- | --- |
| Líder Funcional | Definir los objetivos estratégicos del proyecto, priorizar actividades y asegurar la alineación con las necesidades del negocio. | Experto temático, Gerente de Proyecto |
| Enlace de TI | Ejecutar las tareas asignadas y coordinar con otros miembros del equipo para lograr los objetivos del proyecto. |  |
| Custodio de datos | Garantizar la disponibilidad, calidad e integridad de los datos; definir políticas de acceso y seguridad de la información. | Responsable de la gestión de datos en la entidad |
| Arquitecto de datos | Diseñar y mantener la estructura de datos del proyecto, asegurando su interoperabilidad y escalabilidad. | Diseñador de modelos de datos y lineamientos de almacenamiento |
| Profesional GIS  (si aplica) | Administrar y analizar datos espaciales, asegurando su correcta integración en el modelo analítico. | Especialista en sistemas de información geográfica |
| Ingeniero de datos | Desarrollar y optimizar los procesos ETL (extracción, transformación y carga de datos). | Desarrollador de pipelines de datos |
| Ingeniero DBA | Administrar bases de datos, garantizar su rendimiento y seguridad. | Administrador de bases de datos |
| Arquitecto de Software | Diseñar la infraestructura tecnológica y definir la arquitectura de las aplicaciones del proyecto. | Diseñador de soluciones tecnológicas |
| Analista de datos | Realizar análisis exploratorio y estadístico de los datos para generar información relevante. | Especialista en interpretación de datos |
| Administrador de la wiki | Gestionar la documentación del proyecto en la wiki OTIC. | Responsable de creación y acceso usuarios para la documentación técnica |
| Administrador Gitlab | Gestionar los repositorios de código y control de versiones del proyecto. | Administrador del control de versiones |
| Especialista Nivel 2 (responsable de componente de infraestructura de TI) | Asegurar la disponibilidad, operación y mantenimiento de la infraestructura tecnológica. | Responsable de la infraestructura TI |
| Científico de datos | Diseñar, entrenar e implementar modelos de analítica avanzada para extraer información valiosa. | Especialista en machine learning e inteligencia artificial |

# 

# ESTADO DEL ARTE

Se debe documentar el acercamiento previo al problema (estado del arte) con el objetivo de exponer el aprendizaje obtenido en el escenario pertinente. Esto permitirá enriquecer los modelos propuestos para el Ministerio.

* Identificar las señales y factores relevantes relacionados con el problema
* Evaluar los mecanismos y estrategias existentes para abordar esta problemática, destacando puntos relevantes que puedan contribuir a la solución.
* Plantear otras preguntas relevantes relacionadas con modelos, datos, información o recursos que puedan ser útiles para los modelos propuestos.

# 

# FASES PARA EL DESARROLLO DEL PROYECTO

En esta sección, se describe la estructura metodológica del proyecto a través de sus fases. Para cada fase, es necesario documentar los siguientes aspectos:

* **Propósito de la fase:** Definir el objetivo principal de la fase y su relevancia A continuación, se presentan las cinco fases definidas para el desarrollo de un proyecto de analítica (Ver gráfica 1):
* **Definición de necesidades analíticas:** Identificación de los problemas y objetivos del proyecto, selección de fuentes de datos y definición de criterios de análisis.
* **Creación de la vista minable:** Transformación y estructuración de los datos para su análisis, asegurando su calidad y coherencia.
* **Entrenamiento de los modelos:** Aplicación de técnicas analíticas y de machine learning para la construcción de modelos predictivos o explicativos.
* **Evaluación de la utilidad:** Validación y ajuste de los modelos mediante métricas de desempeño y pruebas con datos reales.
* **Incorporación a los sistemas de monitoreo:** Implementación de los modelos en plataformas de gestión y generación de reportes para la toma de decisiones.

Gráfica 1*.*  Fases del proyecto [nombre proyecto]

## Definición de Necesidades Analíticas

El propósito principal de esta fase es establecer las preguntas analíticas que respondan a las necesidades del Ministerio de Ambiente y Desarrollo Sostenible, en relación con la problemática de interés. Esta fase es un proceso iterativo en el que los temáticos expertos del ministerio y el equipo de BigData e Inteligencia Artificial (oficina TIC) colaborarán para definir de manera precisa las tecnologías, procesos, variables (datos) y las personas involucradas en el análisis del problema [problema]. La Oficina TIC asumirá el liderazgo en la organización de los talleres o reuniones, creando un ambiente propicio para la discusión y el intercambio de ideas. El resultado será un documento que detalla el método utilizado, que podrá ser utilizado por el Ministerio de Ambiente y Desarrollo Sostenible en futuros procesos similares relacionados con la analítica y el Big Data. Además, se especificará de manera detallada la pregunta analítica seleccionada para ser abordada con los datos disponibles. Esta actividad puede requerir de dos a tres iteraciones para llegar a un resultado satisfactorio.

En esta etapa se documentan las preguntas de analíticas que respondan a las necesidades del Ministerio de Ambiente y Desarrollo Sostenible, en relación con el [problema]. Identificar y definir las preguntas clave que guiarán el análisis de datos. Las preguntas deben ser específicas, relevantes y orientadas a la acción para abordar eficazmente la problemática identificada. Se registra detalladamente el método utilizado para definir las necesidades analíticas.

Se documentan las tecnologías, procesos, variables (datos) y las personas involucradas en el análisis, proporcionando una guía para futuros procesos similares en el Ministerio.

Se especifica con detalle la pregunta analítica seleccionada para ser abordada con los datos disponibles.

## Creación de la Vista Minable

El objetivo de esta fase es obtener un conjunto de datos unificados que puedan ser utilizados para entrenar los modelos. De forma iterativa se realizarán las siguientes actividades relacionadas con la ingeniería y entendimiento de los datos:

Gráfica 2*.*  Proceso Creación Vista Minable

### Construcción de ETL

En esta sección, se debe documentar el proceso de Extracción, Transformación y Carga (ETL) aplicado a los datos relevantes del proyecto. Se deben identificar las fuentes de datos utilizadas, como el DataLake del Ministerio de Ambiente y Desarrollo Sostenible, el IDEAM, las Corporaciones Autónomas Regionales u otras bases de datos con información ambiental pertinente. Además, es necesario describir los métodos de extracción empleados, especificando si se utilizan consultas SQL, APIs, descargas manuales u otros mecanismos.

Asimismo, se debe detallar el proceso de transformación, incluyendo la limpieza, normalización y estandarización de los datos para garantizar su calidad. Se debe indicar cómo y dónde se almacenarán los datos transformados, ya sea en un DataLake, una base de datos estructurada u otro repositorio. También es importante mencionar las herramientas y tecnologías utilizadas, como Python, SQL, o Power BI, y describir los mecanismos de validación implementados para asegurar la integridad y coherencia de los datos procesados.

### Limpieza y transformación

En esta sección, se debe describir el proceso de limpieza y transformación de los datos, asegurando su calidad y adecuación para las siguientes fases del proyecto. Se deben identificar y documentar las técnicas utilizadas para la detección y manejo de valores atípicos, datos faltantes, duplicados e inconsistencias en los conjuntos de datos. Además, es importante especificar las reglas de estandarización aplicadas, como la conversión de unidades, la homogenización de formatos y la validación de codificaciones.

También se debe indicar cómo se garantiza la idoneidad de los datos en función de la pregunta analítica definida, asegurando que la información transformada sea representativa y relevante para el análisis. Es necesario describir los criterios utilizados para la selección y filtrado de datos, así como cualquier ajuste o cálculo adicional realizado. Finalmente, se recomienda documentar las herramientas y tecnologías empleadas en este proceso, como Python, R, SQL o plataformas de procesamiento de datos en la nube.

### Exploración univariada

En esta sección, se debe documentar el análisis individual de cada variable con el objetivo de comprender su comportamiento y características estadísticas. Se recomienda calcular y reportar medidas de tendencia central (media, mediana, moda), dispersión (varianza, desviación estándar, rango intercuartílico) y distribución (asimetría, curtosis). Además, es importante incluir visualizaciones como histogramas, boxplots y diagramas de densidad para facilitar la interpretación de la variabilidad y detectar posibles valores atípicos.

También se deben establecer criterios para la selección de variables más representativas según su relevancia para la pregunta analítica. Es importante justificar por qué ciertas variables se consideran clave en el análisis y cuáles pueden ser descartadas o transformadas para mejorar la calidad de los datos. Finalmente, se recomienda documentar las herramientas utilizadas en esta exploración, como Python (Pandas, Seaborn, Matplotlib), R o software estadístico especializado.

### Exploración bivariada

En esta sección, se debe documentar el análisis de relaciones entre pares de variables para identificar patrones, dependencias y posibles asociaciones significativas. Se recomienda utilizar medidas estadísticas como coeficientes de correlación (Pearson, Spearman, Kendall) para evaluar la fuerza y dirección de las relaciones lineales o no lineales. Asimismo, es necesario generar visualizaciones como diagramas de dispersión, heatmaps de correlación y diagramas de cajas para facilitar la interpretación de los resultados.

Además, se debe considerar la detección de relaciones espurias o influenciadas por valores atípicos, documentando cualquier ajuste necesario para mejorar la calidad del análisis. Es importante destacar aquellas combinaciones de variables que aportan mayor valor a la pregunta analítica definida, justificando su inclusión en etapas posteriores del modelo. Finalmente, se deben registrar las herramientas utilizadas en este proceso, como Python (Seaborn, Matplotlib, Pandas), R o software especializado en estadística y análisis de datos.

### Análisis multivariado

Documentar el análisis de relaciones entre múltiples variables aplicando métodos estadísticos y de machine learning según la naturaleza del problema. Se debe especificar qué técnicas se utilizaron, como regresión múltiple, análisis de componentes principales (PCA) o clustering, y justificar su elección.

Registrar las variables de control consideradas, los supuestos estadísticos evaluados y los criterios utilizados para seleccionar las variables más relevantes. Incluir visualizaciones que faciliten la interpretación de los resultados, como matrices de correlación, gráficos tridimensionales o dendrogramas, y asegurar que los hallazgos obtenidos sean trazables dentro del análisis general del proyecto.

### Validación de supuestos

Evaluar el cumplimiento de las condiciones necesarias para la correcta aplicación de las técnicas de análisis utilizadas en el proyecto. Esto implica verificar si los datos y modelos cumplen con los criterios esperados, lo que permitirá obtener resultados confiables y coherentes.

Documentar el proceso de validación y las estrategias implementadas en caso de que los supuestos no se cumplan. Si es necesario, ajustar los datos o modificar el enfoque analítico para garantizar la solidez de los resultados y su correcta interpretación.

### Estimación y prueba de hipótesis

Cuantificar la magnitud de los efectos y evaluar la variabilidad en las relaciones identificadas entre las variables. Este análisis permite obtener una comprensión más profunda de los datos y establecer métricas clave que servirán como insumo para la fase de entrenamiento de los modelos.

Documentar los resultados de las pruebas realizadas, asegurando que las conclusiones derivadas sean consistentes y respaldadas por evidencia estadística. En caso de ser necesario, ajustar los enfoques de análisis para mejorar la precisión y utilidad de los hallazgos.

### Visualización

Presentar los resultados del análisis de datos a través de representaciones gráficas que permitan identificar y comunicar de manera clara las relaciones, correlaciones, efectos y desviaciones encontradas. Las visualizaciones deben facilitar la interpretación de los datos y servir como apoyo para la toma de decisiones.

Como resultado de esta fase, se generará la vista minable en formato parquet, la cual contendrá todas las variables clave identificadas para medir el problema según las bases de datos procesadas. Asimismo, incluirá registros representativos del comportamiento del fenómeno en Colombia, proporcionando un conjunto de datos históricos que facilitará los análisis posteriores.

## Entrenamiento de la Solución Analítica

En esta fase se lleva a cabo el ajuste y calibración de la solución analítica, aplicando técnicas estadísticas o matemáticas que permitan responder a la pregunta analítica planteada. El proceso de entrenamiento se divide en tres grandes actividades: representación, evaluación y optimización de la solución analítica (Gráfica 3).

Para documentar esta fase, se debe describir detalladamente cada una de las actividades realizadas en cada subfase, incluyendo los métodos utilizados, los resultados obtenidos y cualquier ajuste implementado. Además, es importante registrar los criterios empleados para evaluar el desempeño de la solución y las mejoras aplicadas para optimizarla.

Gráfica 3*.*  Proceso de entrenamiento de la solución analítica

### Representación de la solución analítica

En esta subfase, se debe documentar el proceso de selección de los modelos de machine learning que mejor representen los datos. Se deben incluir los criterios utilizados para la selección, las categorías de modelos evaluadas (clásicos, aprendizaje profundo o ensambles) y la justificación de los modelos elegidos para las pruebas iniciales. Además, es importante detallar las pruebas realizadas para evaluar el ajuste de cada modelo a los datos, describiendo los parámetros considerados y cualquier ajuste preliminar efectuado. Se debe registrar cómo se determinó la mejor representación de la solución analítica y qué modelos avanzarán a la siguiente fase de evaluación. Ver gráfica 4.



Gráfica 4 Representación y modelos de machine Learning

Para documentar esta fase, es importante detallar el proceso de búsqueda de modelos de machine learning así:

* Documenta las publicaciones académicas, libros y artículos relevantes en el campo del aprendizaje automático que aborden problemas similares.
* Documenta las bases de datos académicas, como PubMed, Google Scholar o IEEE Xplore, para buscar investigaciones relacionadas.
* Documenta la consulta de blogs, foros y comunidades en línea donde los profesionales discuten y comparten información sobre técnicas y modelos de machine learning.
* Documenta como se analiza la literatura y los recursos especializados para identificar modelos de machine learning que se hayan utilizado con éxito en problemas similares.

Documentar los modelos, algoritmos, metodologías o estrategias particulares que pueden ser utilizadas para abordar alguna de las tareas identificadas en los objetivos generales y específicos de la problemática.

Tabla 1 Ejemplo Inventario de modelos, algoritmos y metodologías preliminares

|  |  |
| --- | --- |
| **Modelo/Algoritmo/Metodología** | **Tarea con potencial de ser abordada** |
| Regresión Lineal | Predicción de variables continuas |
| Árboles de Decisión | Clasificación y análisis de factores |
| Redes Neuronales | Detección de patrones complejos |
| Clustering K-Means | Segmentación de datos |

### Evaluación de la solución analítica

En esta fase, se comparan los modelos seleccionados en la etapa previa utilizando diversas métricas de desempeño, como precisión, exactitud, error cuadrático, probabilidad, ganancia de información y costo/utilidad. El objetivo es diferenciar entre modelos con un buen rendimiento y aquellos menos adecuados para la solución analítica. Cabe resaltar que esta evaluación es interna y distinta de las métricas externas que se emplearán en la fase posterior.

Se debe documentar detalladamente los resultados obtenidos, especificando las métricas utilizadas y los valores alcanzados por cada modelo. Además, se debe incluir una comparación estructurada entre los modelos, resaltando sus fortalezas y debilidades. Finalmente, se debe justificar la selección del modelo óptimo con base en los resultados de la evaluación, argumentando por qué este es el más adecuado para responder a la pregunta analítica planteada.

### Optimización de la solución analítica

La optimización de la solución analítica consiste en la aplicación de un método que refine y mejore el modelo seleccionado, garantizando su eficiencia en la ejecución o despliegue y su capacidad de generalización ante nuevos datos. Este proceso busca reducir errores, mejorar el desempeño del modelo y asegurar su estabilidad en diferentes escenarios.

El resultado de esta fase es un código ejecutable en Python, acompañado de las visualizaciones necesarias para la validación de los resultados entre las partes interesadas. Las visualizaciones podrán generarse en Python o en la plataforma definida por la Oficina TIC del Ministerio de Ambiente y Desarrollo Sostenible.

En esta fase, se debe documentar lo siguiente:

* **Método de optimización:** Descripción detallada del enfoque utilizado para mejorar el modelo, incluyendo su justificación y aplicación.
* **Ajustes de parámetros:** Registro de las configuraciones y modificaciones realizadas en el modelo durante el proceso de optimización.
* **Código fuente:** Implementación en Python del modelo optimizado, con comentarios explicativos que faciliten su comprensión y mantenimiento.
* **Visualizaciones:** Descripción de los gráficos o reportes utilizados para la validación de los resultados entre las partes interesadas.
* **Evaluación del rendimiento:** Análisis de métricas relevantes para medir la mejora del modelo después de la optimización.
* **Comparación de resultados:** Evidencia de mejoras en el modelo optimizado frente al modelo original, con base en métricas clave.
* **Validación del modelo:** Información sobre los datos adicionales o conjuntos de validación utilizados para comprobar la eficacia del modelo optimizado.
* **Plataforma de visualización:** Especificación de la herramienta utilizada para la generación de reportes y gráficos.

## Evaluación de la utilidad

Esta fase se centra en analizar la utilidad y aplicabilidad de los modelos desarrollados desde una perspectiva empresarial, asegurando que estos puedan integrarse en los procesos de toma de decisiones relacionados con el objetivo definido en la fase de necesidades analíticas. La evaluación permitirá determinar si los modelos aportan valor a la problemática abordada y cómo pueden optimizar el proceso de toma de decisiones dentro del Ministerio de Ambiente y Desarrollo Sostenible.

Como parte de esta fase, se llevará a cabo la **documentación del proceso de ejecución basado en la metodología CRISP-DM**, asegurando que cada etapa sea descrita en función de las preguntas de investigación planteadas. Esto garantizará que el proceso seguido para la generación de los modelos sea transparente, replicable y alineado con estándares reconocidos en la analítica de datos.

Como parte de esta fase, se sugiere realizar talleres de transferencia de conocimientos, con el objetivo de capacitar a los participantes en el uso de los modelos como herramientas de apoyo en la toma de decisiones. Es importante que en estos talleres participen tanto técnicos especializados en gestión de datos como expertos del Ministerio con experiencia en la problemática de interés.

Para documentar esta fase, se debe incluir lo siguiente:

* **Proceso de ejecución CRISP-DM:** Descripción detallada de cada una de sus etapas y su aplicación en la generación de los modelos, incluyendo las preguntas de investigación abordadas.
* **Contribución de los modelos a la toma de decisiones:** Documentación sobre cómo los resultados obtenidos pueden ser utilizados en el contexto empresarial para mejorar los procesos de decisión.
* **Evaluación de la relevancia y utilidad:** Análisis de la pertinencia de los modelos para resolver la problemática identificada y su alineación con los objetivos empresariales.
* **Planificación y organización de los talleres:** Cronograma detallado, duración, modalidad (virtual o presencial) y audiencia objetivo.
* **Participación de expertos y técnicos:** Descripción del rol y la contribución de los participantes en la validación y aplicación de los modelos en el contexto empresarial.

Las fases del proyecto siguen un enfoque ágil y colaborativo, con reuniones semanales lideradas por la Oficina TIC, garantizando la participación de los expertos del negocio. La alineación con la metodología **CRISP-DM** proporciona un marco estructurado para el desarrollo y evaluación de los modelos, asegurando la trazabilidad y calidad del proceso. (Ver Gráfica 5).



Gráfica 5 Metodología CRISP-DM.

Fuente: <https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining>

## Incorporar de la solución analítica a los sistemas de Monitoreo Ambiental

En esta fase, se integrará la solución analítica desarrollada con los sistemas de Monitoreo Ambiental existentes, permitiendo su uso en tiempo real para la gestión del problema de interés. El proceso de integración debe garantizar la interoperabilidad con las plataformas actuales y su operatividad dentro del flujo de trabajo del monitoreo ambiental.

Para documentar esta fase, se deben incluir los siguientes elementos:

* **Arquitectura de datos y componentes desarrollados:** Descripción detallada de la estructura de datos, modelos implementados y su interacción con los sistemas de Monitoreo Ambiental.
* **Informe de pruebas y validación:** Registro de los resultados obtenidos durante las pruebas de integración, incluyendo métricas de desempeño, posibles ajustes y validación de la funcionalidad esperada.
* **Procedimientos y guías para mantenimiento:** Documentación de los pasos a seguir para la gestión continua de la integración en producción, abordando aspectos como monitoreo de rendimiento, actualización de modelos y resolución de incidentes.

Es necesario que la documentación generada en esta fase sea clara y estructurada, facilitando la replicabilidad del proceso y el mantenimiento de la solución analítica en el tiempo.

# MATRIZ DE RIESGOS Y RESTRICCIONES

Se debe documentar información sobre cualquier posible riesgo que se deba tener en cuenta para el proyecto en relación con los datos, formatos, confidencialidad, tiempos, recursos, entre otros aspectos relevantes, así como la identificación de restricciones asociadas.

Tabla 2 Ejemplo Matriz de Riesgos y restricciones

|  |  |
| --- | --- |
| **Riesgos** | **Descripción** |
| Disponibilidad de Datos | Posibles restricciones relacionadas con la disponibilidad, accesibilidad y calidad de los datos necesarios para el proyecto. |
| Formatos de Datos | Restricciones asociadas con los formatos de datos aceptables para el proyecto. |
| Confidencialidad | Riesgos relacionados con la confidencialidad de los datos y la necesidad de cumplir con regulaciones de privacidad. |
| Tiempo | Posibles restricciones de tiempo que podrían afectar la ejecución del proyecto, como plazos ajustados o fechas límite. |
| Recursos | Limitaciones en términos de personal, presupuesto, tecnología u otros recursos necesarios para el proyecto. |
| Cambios en los Requerimientos | Riesgo de que los requerimientos del proyecto cambien durante su desarrollo, lo que podría afectar la planificación y ejecución. |
| Tecnología | Restricciones relacionadas con la disponibilidad o compatibilidad de tecnologías necesarias para el proyecto. |
| Experiencia del Equipo | Posibles limitaciones en la experiencia y habilidades del equipo, lo que podría impactar en la ejecución del proyecto. |
| Regulaciones y Normativas | Riesgos asociados con el cumplimiento de regulaciones y normativas gubernamentales o industriales. |
| Stakeholders y Expectativas | Restricciones en relación con las expectativas y necesidades de los stakeholders, que pueden afectar la dirección del proyecto. |
| Comunicación | Posibles problemas de comunicación que podrían surgir entre los miembros del equipo o con los stakeholders. |
| Impacto de Eventos Externos | Riesgo de eventos externos imprevistos que podrían afectar el desarrollo del proyecto. |
| Sostenibilidad del Proyecto | Posibles limitaciones en términos de sostenibilidad a largo plazo del proyecto, incluyendo mantenimiento y soporte. |



# REPOSITORIOS DE CÓDIGO Y MODELOS USADOS DENTRO DEL PROYECTO

Documente la entrega técnica del código de entrenamiento, refinamiento, pruebas, despliegue y puesta en producción de los modelos de analítica avanzada, visualizaciones y demás información requerida en los tableros de control, esta información, repositorios, y archivos de visualización y cualquier artefacto producto de la consultoría deberá inventariarse y describirse.

# VISUALIZACIONES Y API’S SOBRE LOS COMPONENTES DESARROLLADOS.

En esta fase, se debe documentar de manera estructurada el manual técnico y el manual de usuario de las visualizaciones, APIs y componentes desarrollados, con el propósito de facilitar su uso y mantenimiento por parte del Ministerio de Ambiente y Desarrollo Sostenible.

El manual técnico debe proporcionar información detallada sobre la implementación y configuración de cada componente, incluyendo la arquitectura del sistema, requisitos de instalación, ajustes de parámetros y cualquier otro detalle relevante para su administración y mantenimiento. Este documento servirá de guía para los equipos técnicos encargados de la integración y gestión de las soluciones en los entornos operativos.

Por su parte, el manual de usuario está dirigido a los profesionales del Ministerio que emplearán estas herramientas en su trabajo diario. Debe incluir instrucciones claras sobre cómo acceder, navegar y utilizar las visualizaciones y APIs, con ejemplos prácticos, capturas de pantalla y explicaciones detalladas para garantizar una adopción efectiva.

Ambos manuales deben abarcar todas las visualizaciones, APIs y componentes generados en el proyecto de analítica, asegurando que la documentación sea comprensible, accesible y alineada con las necesidades del usuario final.

# CRONOGRAMA

Describa el cronograma que refleje la duración estimada de cada una de las fases del proyecto, como se muestra en la tabla a continuación:

Tabla 3 Cronograma del proyecto de analítica.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Fases** | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes ..** |
| Definición de necesidades analíticas |  |  |  |  |  |  |
| Creación de la vista minable |  |  |  |  |  |  |
| Entrenamiento modelos |  |  |  |  |  |  |
| Evaluación de la utilidad |  |  |  |  |  |  |

Detalle de los productos esperados:

Tabla 4 Detalle de los Productos esperados

| **No.** | | **Entregable /Productos** | | **Tiempo de entrega después de firmado el contrato.** | **Tiempo estimado para revisión y aceptación** | | **Revisión y aceptación a cargo** | **Avance porcentual** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | | Documento de Inventario de fuentes disponibles seleccionados para el proyecto. |  | |  |  |  |
| **2** | | Documento de acercamientos previos al problema |  | |  |  |  |
| **3** | | Matriz de riesgo y Restricciones |  | |  |  |  |
| **4** | | Inventario de modelos, algoritmos y metodologías preliminares |  | |  |  |  |
| **5** | | Repositorios de código y modelos usados dentro del proyecto. |  | |  |  |  |
| **6** | | Visualizaciones y API’s sobre los componentes desarrollados. |  | |  |  |  |

# TÉRMINOS Y/O CONCEPTOS

En esta sección, se deben definir de manera clara y precisa los términos y conceptos clave utilizados en el documento, asegurando una comprensión uniforme para todos los lectores. Se recomienda incluir definiciones técnicas, metodológicas y operativas relevantes para la solución analítica y su implementación en el contexto del Ministerio. Cada término o concepto debe presentarse con una explicación concisa, indicando su significado, aplicación y, si es necesario, la fuente de referencia. Esta sección facilita la interpretación del documento y minimiza posibles ambigüedades en su lectura.

# BIBLIOGRAFÍA

En esta sección se debe listar el material bibliográfico consultado durante el desarrollo del proyecto de analítica.

La bibliografía debe organizarse y citarse conforme a las normas APA, asegurando la correcta referenciación de libros, artículos, informes, normativas y otras fuentes utilizadas. A continuación, se presenta un ejemplo de formato:

Ejemplo de cita en APA 7:

* Autor, A. A. (Año). *Título del libro o documento en cursiva*. Editorial o fuente.
* Autor, B. B. (Año). Título del artículo. *Nombre de la revista o fuente en cursiva, volumen(número)*, páginas. https://doi.org/xxxxx
* Entidad o Institución. (Año). *Título del informe o norma en cursiva*. URL

# ANEXOS

En esta sección se debe incluir información adicional que complemente, sustente o amplíe los contenidos presentados en el documento. Los anexos pueden contener tablas, gráficos, formatos, cuestionarios, cuadros, esquemas funcionales y otros elementos que permitan una mejor comprensión y respaldo del proyecto.

# DATOS DE PRODUCCIÓN DEL PROYECTO

En esta sección se debe registrar la información sobre las personas responsables del proceso de producción del proyecto de analítica. Los roles aquí descritos están orientados al contenido y la metodología de la arquitectura del proyecto, más que a la validación de cumplimiento de lineamientos de documentación.

ELABORÓ: Persona(s) técnicas o profesionales responsable(s) del diseño y ejecución de la solución analítica.

REVISÓ: Líder(es) o Coordinador(es) encargado(s) de evaluar la metodológica, consistencia de la información y pertinencia de los resultados antes de su aprobación.

APROBÓ: Jefe de la Oficina responsable de validar y avalar el documento del proyecto de analítica para su publicación y uso oficial.

|  |  |
| --- | --- |
| ELABORÓ |  |
| REVISÓ |  |
| APROBÓ |  |