



Guía Arquitectura de Proyectos de Analítica

Proceso
Gestión Estratégica de Tecnologías
de la Información
Versión 1
25/06/2024

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Contenido

Introducción	3
1 Objetivo	3
2 Alcance	3
3 Definiciones.....	3
4 Plantilla de documentación	7
4.1 Descripción del problema.....	7
4.2 Objetivos del proyecto.....	8
4.2.1 Objetivo general.....	8
4.2.2 Objetivos específicos	8
4.3 Inventario de fuentes disponibles seleccionados para el proyecto.....	9
4.4 Niveles de interlocución previstos por el proyecto	10
4.5 Acercamiento previo al problema.....	12
4.6 Actividades y responsabilidades - Fases para el desarrollo del proyecto	12
4.6.1 Definición de necesidades analíticas	13
4.6.2 Creación de la vista minable	13
4.6.3 Entrenar la solución analítica	16
4.6.4 Evaluación de la utilidad	19
4.6.5 Incorporar de la solución analítica a los sistemas de Monitoreo Ambiental	20
4.7 Matriz de riesgo y Restricciones.....	20
4.8 Repositorios de código y modelos usados dentro del proyecto	21
4.9 Visualizaciones y API's sobre los componentes desarrollados.....	21
4.10 Cronograma	22



SC-2000142



SA-2000143

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Introducción

La presente guía tiene como objetivo proporcionar un marco para la documentación de proyectos de analítica en el Ministerio de Ambiente y Desarrollo Sostenible, ya que es importante que las iniciativas de análisis de datos estén respaldadas por una documentación clara y precisa que garantice la reproducibilidad, la transparencia y la comprensión de los procesos involucrados.

Este documento servirá como referencia para los equipos encargados de desarrollar proyectos de analítica dentro del Ministerio, el cual proporciona una estructura coherente y unificada para la documentación de cada fase del proceso analítico, desde la descripción del problema inicial hasta la incorporación de soluciones analíticas en los sistemas de monitoreo ambiental.

Además, esta guía promueve las mejores prácticas en la documentación de proyectos de analítica, lo que garantiza que el conocimiento generado durante el desarrollo del proyecto se preserve y esté disponible para futuras referencias y mejoras.

A lo largo de esta guía, se proporcionarán instrucciones detalladas, ejemplos prácticos y plantillas para facilitar la tarea de documentar cada fase del proyecto de analítica.

1 Objetivo

El Objetivo de la presente guía es establecer los lineamientos para la identificación, elaboración y documentación de cada una de las etapas que conforman un proyecto de analítica del Ministerio de Ambiente y Desarrollo Sostenible, basados en la metodología CRISP-DM.

2 Alcance

La presente guía aplica para los proyectos de analítica, que se desarrollen o se planeen desarrollar en las diferentes dependencias del Ministerio de Ambiente y Desarrollo Sostenible y que estén relacionadas con la misionalidad de la entidad.

3 Definiciones

Algoritmo: Se trata de un conjunto de secuencias ordenadas de instrucciones matemáticas lógicas, que permiten resolver un problema o realizar una tarea específica. Se emplean para procesar datos, aprender patrones y tomar decisiones informadas. Tomado de: <https://openwebinars.net/blog/glosario-basico-de-ia-aclarando-conceptos/>

Análisis de datos (Data analytics): Proceso en el que se utiliza tecnología, técnicas estadísticas y grandes conjuntos de datos, para identificar patrones y correlaciones y así extraer conclusiones usando la

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

información proporcionada. Adaptado de Berkeley School of Information. (8 de Abril de 2020). What is Data Science? Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2632/ficha/>

Análítica diagnóstica: Tipo de análisis de datos que responde a la pregunta ¿por qué sucedió? Se enfoca en el pasado y se basa en la probabilidad. Datahack BiG Data Family. (24 de 05 de 2019). Tipos de Analítica Big Data. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2633/ficha/>

Análisis exploratorio de datos (EDA): Enfoque para el análisis de datos que emplea un número de diferentes técnicas para observar los datos y ver lo que parecen decir; descubrir las estructuras subyacentes; aislar variables importantes; detectar anomalías y sugerir modelos adecuados para estadísticas convencionales. Encyclopedia of Database Systems, 2018. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2630/ficha/>

Análítica predictiva: Tipo de análisis de datos que responde a la pregunta ¿qué podría suceder si una condición específica ocurre? Se enfoca en el futuro basándose en la probabilidad. Datahack BiG Data Family. (24 de 05 de 2019). Tipos de Analítica Big Data. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2634/ficha/>

Análítica prescriptiva: Tipo de análisis de datos que responde a la pregunta ¿cuáles son las mejores acciones basados en el resultado deseado? Se enfoca en el futuro y está sometida por un conjunto de reglas. Datahack BiG Data Family. (24 de 05 de 2019). Tipos de Analítica Big Data. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2635/ficha/>

API: significa “interfaz de programación de aplicaciones”. Las API son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software del instituto de meteorología contiene datos meteorológicos diarios. La aplicación meteorológica de su teléfono “habla” con este sistema a través de las API y le muestra las actualizaciones meteorológicas diarias en su teléfono. Amazon (17 de marzo de 2024). Tomado de: <https://aws.amazon.com/es/what-is/api/>

Aprendizaje automático (Machine Learning): El aprendizaje automático se refiere a la capacidad de las máquinas para aprender y mejorar automáticamente a partir de los datos sin ser programadas explícitamente. Se basa en el uso de algoritmos y modelos que permiten a las máquinas reconocer patrones, realizar predicciones y tomar decisiones basadas en la información adquirida. Tomado de: <https://openwebinars.net/blog/glosario-basico-de-ia-aclarando-conceptos/>

Aprendizaje de Ensamble (Machine Learning): Los métodos de ensamble de modelos o métodos combinados intentan ayudar a mejorar el rendimiento de los modelos de Machine Learning al mejorar su precisión. Este es un proceso mediante el cual se construyen estratégicamente varios modelos de Machine Learning para resolver un problema particular. Tomado de: <https://aprendeia.com/metodos-de-ensamble-de-modelos-machine-learning-ensemble-methods-en-espanol/>

Aprendizaje no supervisado (Machine Learning): Sistema de aprendizaje de Machine Learning que estudia los datos por sí mismos para encontrar patrones y relaciones presentes en un conjunto de datos. Extrae metodologías que resumen y agrupan los datos para ayudar a obtener información significativa, utilizando algoritmos de aprendizaje, que interpretan los datos sin necesitar de un operador y la máquina hace el análisis de los datos disponibles, a través de correlaciones y relaciones entre ellos mismos. Adoptado de SINGH, N. S. (1 de 03 de 2019). XENONSTACK. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2654/ficha/>



SC-2000142



SA-2000143

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Aprendizaje por refuerzo (Machine Learning): Sistema de aprendizaje de Machine Learning que encierra acciones, parámetros, valores finales y las observaciones recopiladas de la interacción con un entorno determinado para tomar decisiones que maximicen la recompensa o minimicen el riesgo. El sistema enseña la máquina a través del proceso de ensayo y error, dando la opción de aprender de experiencias pasadas, que permiten adaptar su enfoque según las situaciones y elegir la mejor opción. Adoptado de SINGH, N. S. (1 de 03 de 2019). XENONSTACK. Obtenido de <https://conceptos.dane.gov.co/conceptos/conceptos/2655/ficha/>

Aprendizaje supervisado (Machine Learning): Sistema de aprendizaje de Machine Learning cuya enseñanza se da a través de datos históricos y resuelve problemas de regresión y clasificación. El operador proporciona datos conocidos con entradas y salidas al algoritmo, quien debe encontrar un método de entrada o salida para dar una subcategoría. Adoptado de SINGH, N. S. (1 de 03 de 2019). XENONSTACK. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2656/ficha/>

Aprendizaje profundo (Deep Learning): es una rama del aprendizaje automático que se centra en el uso de redes neuronales artificiales profundas para extraer y aprender características jerárquicas de los datos. Estas redes están compuestas por múltiples capas de nodos interconectados y son capaces de realizar tareas complejas, como el reconocimiento de objetos en imágenes o el procesamiento del lenguaje natural (NLP). Tomado de: <https://openwebinars.net/blog/glosario-basico-de-ia-aclarando-conceptos/>

Bases de datos (BD): Conjunto o colección de datos interrelacionados entre sí, que se utilizan para la obtención de información de acuerdo con el contexto de los mismos y que son almacenados sistemáticamente para su posterior uso. Departamento Administrativo Nacional de Estadística (DANE). Resolución 1503 de 2011. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2744/ficha/>

Big Data: Proceso de análisis de datos masivos donde estos se trabajan a gran velocidad para extraer información en la toma de decisiones de las empresas. Adaptado de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) y Oficina de Estadística de las Comunidades Europeas (Eurostat), "Oslo Manual 2018: Guidelines for Collecting, Reporting and Use Data on Innovation", Cuarta edición, París, 2018. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/2760/ficha/>

Datos: son la representación primaria de variables cualitativas y cuantitativas que son almacenables, transferibles, pueden ser visualizadas, controladas y entendidas (Etymonline, 2016). Tomado de: <https://herramientas.datos.gov.co/sites/default/files/2020-11/3920.pdf>

Datos crudos (Raw): datos ingeridos que no se han transformado y que están en su formato de archivo original. Adaptado de Amazon: <https://docs.aws.amazon.com/prescriptive-guidance/latest/defining-bucket-names-data-lakes/raw-data-layer-naming-structure.html>

Datos procesados (Stage): se refieren a la etapa en la que los datos crudos o sin procesar han pasado por un proceso de transformación y limpieza para convertirse en una forma más estructurada y útil para su análisis.

Datos de entrenamiento: conjunto de datos generados por humanos y especialmente seleccionados con los cuales se entrena a la inteligencia artificial. Tomado de: <https://blogs.uoc.edu/elearning-innovation-center/es/glosario-sobre-ia-conceptos-basicos/>

Estándar de referencia (Gold estándar): El rendimiento de todo test diagnóstico se basa en su comparación con un gold standard (estándar de oro, patrón de oro, patrón de referencia). El GS es la técnica

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

diagnóstica que define la presencia de la condición con la máxima certeza conocida. Salech et al. (2008). Tomado de: <http://dx.doi.org/10.4067/S0034-98872008000900018>

Extracción, Transformación y Carga (ETL): significa extraer, transformar y cargar, es un proceso de integración de datos que combina datos de múltiples fuentes de datos en un único almacén de datos consistente que se carga en un almacén de datos u otro sistema de destino. IBM Tomado de: <https://www.ibm.com/mx-es/topics/etl#:~:text=el%20siguiente%20paso-,%C2%BFQu%C3%A9%20es%20ETL%3F,u%20otro%20sistema%20de%20destino.>

Lago de datos (DataLake): Son soluciones de gestión de datos híbridos de última generación que pueden hacer frente a los retos de Big data y que impulsan nuevos niveles de analítica en tiempo real. Su entorno altamente escalable soporta volúmenes de datos extremadamente grandes, y acepta datos en su formato nativo a partir de varios orígenes de datos. Como complementos para su almacenamiento de datos, proporcionan la plataforma para machine learning y analítica avanzada en tiempo real en un entorno colaborativo. IBM Data lake 2019. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/3782/ficha/>

CRISP-DM: son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos. Tomado de: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>

Modelo: Los modelos son representaciones de sistemas o procesos, los cuales mediante el uso de algoritmos se utilizan para realizar predicciones, clasificaciones o tomar decisiones basadas en la información aprendida. Estos modelos de IA se construyen a partir de los conjuntos de datos, para ser tratados en consecuencia. Tomado de: <https://openwebinars.net/blog/glosario-basico-de-ia-aclarando-conceptos/>

Parquet: es un formato de almacenamiento en columnas disponible para cualquier proyecto en el ecosistema Hadoop, independientemente de la elección del marco de procesamiento de datos, el modelo de datos o el lenguaje de programación. Tomado de: <https://parquet.apache.org/>

Python: es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes. El software Python se puede descargar gratis, se integra bien a todos los tipos de sistemas y aumenta la velocidad del desarrollo. Tomado de: <https://aws.amazon.com/es/what-is/python/#:~:text=Python%20es%20un%20lenguaje%20de,ejecutar%20en%20muchas%20plataformas%20diferentes.>

Variable: Característica de las unidades estadísticas que puede asumir un conjunto de valores los cuales pueden ser cuantitativos o cualitativos. Departamento Administrativo Nacional de Estadística (DANE). Norma técnica de la calidad del proceso estadístico.2020. Tomado de: <https://conceptos.dane.gov.co/conceptos/conceptos/4819/ficha/>

Vista minable: es la consolidación en una única tabla de todas las observaciones y los atributos sobre los que se aplicarán los algoritmos de minería de datos. Quinteros et al. Tomado de: <https://core.ac.uk/download/pdf/296389834.pdf>



SC-2000142



SA-2000143

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

4 Plantilla de documentación

Esta guía está diseñada para ayudar a documentar la arquitectura de proyectos de analítica en la wiki de manera clara y estructurada. A continuación, se presenta la estructura especificada:

<p>Documento Arquitectura de Proyecto de Analítica</p> <p>Contenido</p> <ul style="list-style-type: none"> Descripción del problema Objetivos del proyecto Objetivo general Objetivos específicos Inventario de fuentes disponibles seleccionados para el proyecto Niveles de interlocución previstos por el proyecto Acercamientos previos al problema Actividades y responsabilidades - Fases para el desarrollo del proyecto Definición de necesidades analíticas Creación de la vista minable Construcción de ETL Limpieza y transformación Exploración univariada Exploración bivariada Análisis multivariado Validación de supuestos Estimación y prueba de hipótesis Visualización Entrenar la solución analítica Representación de la Solución Analítica Evaluación de la solución analítica Optimización de la solución analítica Evaluación de la utilidad Incorporar de la solución analítica a los sistemas de Monitoreo Ambiental Matriz de riesgo y Restricciones Repositorios de código y modelos usados dentro del proyecto. Visualizaciones y API's sobre los componentes desarrollados. Cronograma
--

A continuación, se detalla la información necesaria para cada capítulo al documentar en la plantilla:

4.1 Descripción del problema

En esta sección se requiere que se documente un resumen claro y conciso del problema a resolver. Aquí se describe el problema en detalle, identificando sus principales características y desafíos. Se analiza y documenta el impacto que el problema tiene en el negocio u organización, incluyendo posibles consecuencias negativas

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

de no abordarlo. Se enumeran las variables clave que están involucradas en el problema, destacando aquellas que se consideran críticas para su comprensión y resolución.

4.2 Objetivos del proyecto

En esta sección se establecen los objetivos del proyecto, tanto generales como específicos:

4.2.1 Objetivo general

Se define el propósito general del proyecto, destacando la meta principal que se busca alcanzar. Define el objetivo general que dará cuenta de los resultados esperados respondiendo a preguntas como: ¿qué? ¿para qué? ¿cuándo? ¿dónde?

Ejemplo: Diseñar e implementar una solución avanzada que se enfoque en el análisis, descripción y predicción de información asociada a la problemática de [problema] en [ubicación]. Esta solución considerará los diversos factores [factores] que contribuyen a este problema ambiental. Su propósito final es proporcionar un apoyo sólido y basado en datos para la toma de decisiones [acción específica].

4.2.2 Objetivos específicos

En esta sección, se detallan los objetivos específicos que contribuyen a la consecución del objetivo general. Cada objetivo específico debe ser:

1. Llevar a cabo un proceso de descubrimiento y recolección de datos enfocados en el [problema], identificando y seleccionando fuentes de información pertinentes para un análisis mediante técnicas de analítica avanzada.
2. Ejecutar un análisis exploratorio de datos utilizando las fuentes previamente seleccionadas para determinar los tipos de datos disponibles, tanto estructurados como no estructurados, que puedan ser aplicados en el análisis de la problemática de [problema] en [ubicación].
3. Consolidar la información necesaria, en sus formatos crudos (Raw) y procesados (Stage), provenientes del lago de datos, con el objetivo de prepararla adecuadamente para el posterior procesamiento analítico.
4. Desarrollar e implementar una estrategia de analítica avanzada para producir información relevante que pueda respaldar eficazmente la toma de decisiones de las diversas partes interesadas.
5. Implementar un proceso de validación para garantizar la eficacia de las técnicas y enfoques analíticos utilizados. Esto se logrará mediante la configuración de puntos de referencia y el uso de conjuntos de datos de entrenamiento y referencias (gold estándar).



SC-2000142



SA-2000143

- Diseñar y desarrollar las visualizaciones y API's necesarios para presentar y compartir de manera efectiva los resultados y la información generada a partir del análisis realizado.

4.3 Inventario de fuentes disponibles seleccionados para el proyecto.

En esta fase, se recopila y documenta un inventario detallado de todas las fuentes de datos seleccionadas para el proyecto. Este documento proporciona una visión general de las fuentes disponibles, sus características y su disponibilidad, además de abordar cualquier situación especial que pueda afectar la disponibilidad de la información.

Se debe enumerar todas las fuentes de datos seleccionadas para el proyecto, incluyendo su nombre, descripción y detalles relevantes.

Descripción detallada de las características de cada fuente, como su origen, formato, tipo de información, frecuencia de generación y soporte.

Especificación de cualquier situación especial que pueda demorar o dificultar la disponibilidad de la información, como permisos necesarios, trámites ante terceras entidades o procesos de anonimización de datos.

Anexar de archivos con muestras de datos o estructuras, con el objetivo de proporcionar ejemplos concretos del contenido de cada fuente.

Hay que asegurar de que las muestras de datos estén disponibles en el Datalake del Ministerio. En caso contrario, se deberá garantizar su persistencia en el Datalake por cada pregunta de analítica, especialmente si provienen de fuentes abiertas o de otras entidades.

Tabla 1 Inventario de fuentes disponibles. Ejemplo.

Fuente	Descripción	Área Responsable	Productor (fuente oficial)	Clasificación	Tipo de Información	Frecuencia de Generación	Soporte	Formato	Datos Abiertos	Tipo de datos	URL de publicación	Cobertura Geográfica
Matriz Unificada de los Proyectos de Inversión - Código F-E-GIP-51	Es un registro de los proyectos de inversión que se lideran las corporaciones y municipios y en él se almacena la información requerida para hacer seguimiento al proyecto.	Oficina Asesora de Planeación	Ministerio de Ambiente y Desarrollo Sostenible	Clasificada	Semiestructurado	Mensual	Digital - Electrónico	Hoja de Cálculo	No	NA		

NA: No aplica

Compromiso de disponibilidad a partir de la aprobación del proyecto:

Se debe documentar el compromiso de disponibilidad a partir de la aprobación del proyecto, especificando de manera formal la fecha a partir de la cual el Equipo de BigData e Inteligencia Artificial tendrá acceso efectivo a los datos de cada fuente.

4.4 Niveles de interlocución previstos por el proyecto

En esta fase, se establecerán los niveles de interlocución previstos para el proyecto, detallando las personas involucradas en la comunicación y su respectivo nivel de responsabilidad. Esto garantizará una comunicación efectiva y fluida a lo largo de todo el proyecto.

Tabla 2 Ejemplo Niveles de interlocución previstos.

Nivel de Interlocución	Responsabilidades	Roles Específicos
Equipo de Proyecto	Ejecutar las tareas asignadas y coordinar con otros miembros del equipo para lograr los objetivos del proyecto.	Gerente de Proyecto
	Ejecutar las tareas asignadas para la formulación, seguimiento y evaluación del proyecto	Líder del Proyecto de analítica Profesional en Ingeniería de Sistemas e Informática, Industria, Electrónica o afines, con Maestría en Ingeniería de Sistemas, Industrial, Ciencia de Datos, o afines ó Doctorado en Ingeniería, Ciencia de Datos o afines. Experiencia profesional mínima de 10 años realizar procesos de analítica, en temas como: <ul style="list-style-type: none"> • Técnicas de filtrado colaborativo • Minería de datos / analítica de Datos • Ciencia de Datos, • Procesos en Batch o Streaming. Trabajo con dato estructurado y no estructurado.
	Ejecutar las tareas asignadas para el diseño y desarrollo de soluciones avanzadas de análisis mediante machine learning	Científico de datos Senior Profesional en Ingeniería sistemas, industrial, electrónica, o en áreas afines a las ciencias de la computación. Maestría en Ingeniería sistemas, o en áreas afines a las ciencias de la computación, Industrial, Ciencia de Datos, o afines. Doctorado en Ingeniería sistemas, o en áreas afines a las ciencias de la computación (Deseable). Experiencia profesional de 5 años.

Nivel de Interlocución	Responsabilidades	Roles Específicos
		Experiencia en al menos dos proyectos de analítica de datos donde sea el consultor principal de la solución implementada.
	Ejecutar las tareas asignadas para acopio, limpieza y exploración inicial de los datos, así como en la implementación de modelos y algoritmos de análisis bajo la supervisión de científicos de datos senior.	Científicos de Datos Junior Profesional en ingenierías o áreas afines, experiencia de 1 año trabajando en proyectos de Analítica y Bigdata y título profesional en áreas de conocimiento relacionadas con el objeto del trabajo, capaz de trabajar en entornos altamente colaborativos y de ritmo rápido. Núcleo Analítica. Profesional en matemáticas, o en áreas afines a las ciencias de la estadística y el análisis de datos. Experiencia de 1 año trabajando en proyectos de Analítica y Bigdata y título profesional en áreas de conocimiento relacionadas con el objeto del trabajo, capaz de trabajar en entornos altamente colaborativos y de ritmo rápido.
	Ejecutar las tareas asignadas para diseñar, construir y mantener la infraestructura de datos necesaria para la recopilación, almacenamiento, procesamiento y acceso eficiente a grandes volúmenes de datos.	Ingeniero de Datos Profesional en ingenierías o áreas afines, experiencia de 1 año trabajando en proyectos de Analítica y Bigdata y título profesional en áreas de conocimiento relacionadas con el objeto del trabajo, capaz de trabajar en entornos altamente colaborativos y de ritmo rápido. Núcleo Ingeniería de Información, profesional en Ingeniería sistemas, o en áreas afines a las ciencias de la computación. Experiencia de 1 año trabajando en proyectos de Analítica y Bigdata y título profesional en áreas de conocimiento relacionadas con el objeto del trabajo, capaz de trabajar en entornos altamente colaborativos y de ritmo rápido.
	Ejecutar las tareas asignadas para dirigir y guiar al equipo en la aplicación de su experiencia especializada en un área particular del análisis de datos.	Líder en temáticas específicas Profesional en Ingeniería ambiental, o en áreas afines. Maestría en temáticas relacionadas a temas ambientales. Experiencia profesional de 5 años. Experiencia en al menos dos proyectos de que vinculen temas ambientales.

Nivel de Interlocución	Responsabilidades	Roles Específicos
Stakeholders Clave	Proporcionar información relevante sobre los requisitos y expectativas del negocio. Participar en la validación y aprobación de entregables claves.	Viceministros, Directores, Jefes de Oficina y Asesores de Despacho,

4.5 Acercamiento previo al problema

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.

4.6 Actividades y responsabilidades - Fases para el desarrollo del proyecto

Para lograr los resultados del proyecto expresados en el punto anterior, se proponen cinco fases (ver gráfica 1):

- Definición de necesidades analíticas.
- Creación de la vista minable.
- Entrenamiento de los modelos.
- Evaluación de la utilidad.
- Incorporación a los sistemas de monitoreo de gestión ambiental.



Gráfica 1 Fases del proyecto

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

4.6.1 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 la problemática de [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.

4.6.2 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

4.6.2.1 Construcción de ETL¹

Realizar un proceso de Extracción, Transformación y Carga (ETL) sobre bases de datos (BD) o conjuntos de datos relevantes que puedan existir en el DataLake del Ministerio de Ambiente y Desarrollo Sostenible y otras fuentes de datos disponibles: BD del Ministerio de Ambiente y Desarrollo Sostenible, Ideam, Corporaciones Autónomas Regionales u otro cualquier conjunto de datos disponible que contengan información ambiental relevante.

4.6.2.2 Limpieza y transformación

Verificar los datos a través de los procesos de limpieza y transformación, considerando la idoneidad de los datos para la siguiente fase de acuerdo con la pregunta analítica definida.

4.6.2.3 Exploración univariada

Realizar un análisis individual de cada variable, examinando sus distribuciones estadísticas y eligiendo aquellas que sean más representativas para comprender los datos.

4.6.2.4 Exploración bivariada

Analizar las relaciones entre pares de variables, revisión de tablas y gráficos de correlación, cálculo de correlaciones y detección de relaciones lineales.

4.6.2.5 Análisis multivariado

¹ Extract, Transform and Load

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Explorar las relaciones entre múltiples variables, aplicando regresiones multivariadas y considerando variables de control para identificar relaciones complejas que puedan ser interpretadas.

4.6.2.6 Validación de supuestos

Verificar los supuestos y condiciones necesarios para aplicar las técnicas de análisis utilizadas.

4.6.2.7 Estimación y prueba de hipótesis

Determinar el tamaño del efecto y la variación existente entre las relaciones identificadas entre las variables, lo que representa una comprensión de los datos y métricas insumo para la fase entrenamiento de los modelos.

4.6.2.8 Visualización

Comunicar de manera efectiva los resultados, mediante visualizaciones gráficas de los efectos, desviaciones, correlaciones y relaciones identificadas entre las variables y datos de análisis.

Como resultado de esta fase se obtendrá la vista minable (archivo en formato parquet) que contienen, por un lado, todas las variables ideales para medir el problema, de acuerdo con las bases de datos procesadas, y por otro, los registros que representan un conjunto de datos históricos capaz de representar el comportamiento del fenómeno [fenómeno] en Colombia.

En la fase de Creación de la vista minable, es crucial documentar todas las actividades realizadas y los resultados obtenidos. Aquí hay una lista de los elementos que se deben documentar:

Construcción de ETL:

- Detalles del proceso de Extracción, Transformación y Carga (ETL) sobre las bases de datos relevantes.
- Fuentes de datos utilizadas
- Descripción de los pasos específicos de extracción, transformación y carga de datos.

Limpieza y transformación:

- Descripción de los métodos utilizados para limpiar y transformar los datos.
- Criterios para la limpieza de datos, como eliminación de valores atípicos, imputación de valores faltantes, etc.
- Transformaciones realizadas para preparar los datos para análisis.

Exploración univariada:

- Resultados de análisis individual de cada variable, incluyendo distribuciones estadísticas y características relevantes.
- Variables seleccionadas como más representativas para comprender los datos.



SC-2000142



SA-2000143

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	SOMOSIG Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Exploración bivariada:

- Análisis de relaciones entre pares de variables, incluyendo tablas y gráficos de correlación.
- Resultados de cálculos de correlaciones y detección de relaciones lineales.

Análisis multivariado:

- Resultados de exploración de relaciones entre múltiples variables, incluyendo aplicaciones de regresiones multivariadas.
- Consideraciones de variables de control para identificar relaciones complejas.

Validación de supuestos:

- Verificación de supuestos y condiciones necesarios para aplicar las técnicas de análisis utilizadas.
- Documentación de cualquier supuesto validado o no validado.

Estimación y prueba de hipótesis:

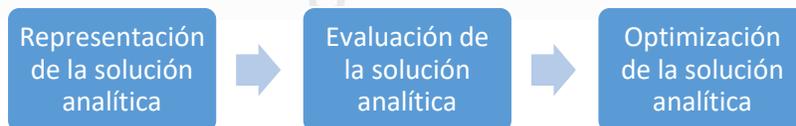
- Determinación del tamaño del efecto y variación existente entre las relaciones identificadas.
- Métricas obtenidas que informan la comprensión de los datos y métricas insumo para la fase de entrenamiento de modelos.

Visualización:

- Descripción de visualizaciones utilizadas para comunicar los resultados.

4.6.3 Entrenar la solución analítica

En esta fase se desarrollará una serie de actividades que consiste en ajustar la solución analítica, utilizando técnicas estadísticas o matemáticas que permita obtener una respuesta a la pregunta analítica. Esta fase tiene tres grandes actividades: representación, evaluación y optimización de la solución analítica.



Gráfica 3 Proceso de entrenamiento de la solución analítica

Para documentar la fase de "Entrenar la solución analítica" es importante detallar cada una de las actividades realizadas, los resultados obtenidos y cualquier otra información relevante para el desarrollo de la solución analítica por cada subfase.

4.6.3.1 Representación de la Solución Analítica

Consiste en realizar una búsqueda de cuáles modelos de machine learning pueden representar mejor los datos. Estos modelos pueden tener una representación en tres formas: clásica, aprendizaje profundo o ensamble, y dentro de éstas puede existir uno o varios modelos con los cuales se deben realizar pruebas para determinar su ajuste a los datos. 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:

- Documentar las publicaciones académicas, libros y artículos relevantes en el campo del aprendizaje automático que aborden problemas similares.
- Documentar las bases de datos académicas, como PubMed, Google Scholar o IEEE Xplore, para buscar investigaciones relacionadas.
- Documentar la consulta de blogs, foros y comunidades en línea donde los profesionales discuten y comparten información sobre técnicas y modelos de machine learning.
- Documentar 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 3 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

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	 Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

4.6.3.2 Evaluación de la solución analítica

Consiste en comparar los modelos seleccionados en la actividad anterior, de acuerdo con métricas de precisión, exactitud, error cuadrado, probabilidad, ganancia de información, costo/utilidad, entre otras, que permitan distinguir entre buenos y malos modelos. La evaluación en este punto es diferente a las métricas de valoración externas que se usarán en la fase siguiente.

En esta fase se debe documentar los resultados de la evaluación, incluyendo las métricas utilizadas y los valores obtenidos para cada modelo.

También se debe documentar los resultados de comparación de modelos destacando las fortalezas y debilidades de cada uno.

Justificación de la selección del modelo óptimo-basada en los resultados de la evaluación.

4.6.3.3 Optimización de la solución analítica

Entendida como la aplicación de un método que cualifique el modelo seleccionado, esto permitirá eficiencia en la ejecución o despliegue y determinará si el resultado es generalizable con nuevos datos.

El resultado de esta fase es código ejecutable (en Python) y las visualizaciones que se requieran para la validación de resultados entre las partes.

Las visualizaciones se realizarán en Python o en la plataforma de preferencia del Ministerio de Ambiente y Desarrollo Sostenible, la que será definida por la Oficina TIC.

En esta fase se debe documentar lo siguiente:

- Descripción detallada del método seleccionado para optimizar el modelo, incluyendo su justificación y cómo se aplicará.
- Detalles sobre los ajustes de parámetros y configuraciones realizados durante la optimización del modelo.
- El código en Python que implementa el modelo optimizado, con comentarios explicativos para facilitar su comprensión y mantenimiento.
- Descripción de las visualizaciones requeridas para la validación de resultados entre las partes interesadas.
- Evaluación del rendimiento del modelo optimizado utilizando métricas relevantes.
- Comparación de los resultados antes y después de la optimización para demostrar mejoras.
- Información sobre cómo se validaron los resultados del modelo optimizado, incluyendo datos adicionales utilizados o conjuntos de validación.
- Nombre de la plataforma específica para las visualizaciones.

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	SOMOSIG Sistema Integrado de Gestión
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

4.6.4 Evaluación de la utilidad

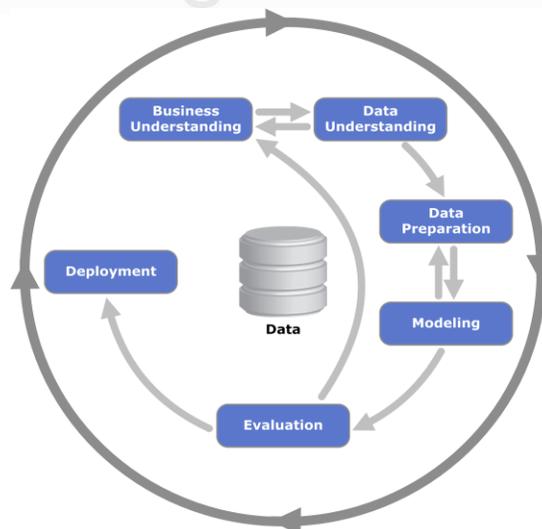
Esta fase implica evaluar los resultados de los modelos generados desde una perspectiva empresarial, con el objetivo de aplicarlos en una o varias actividades de toma de decisiones relacionadas con el objetivo empresarial definido en la fase de necesidades analíticas.

El resultado de esta etapa son dos talleres de transferencia de conocimientos, diseñados con el fin de capacitar a los participantes en el uso de los modelos desarrollados y su aplicación como herramientas de apoyo, en la toma de decisiones asociadas con el objetivo empresarial, en este caso, la problemática de interés. Se espera que cada taller tenga una duración de aproximadamente dos horas y debido a las circunstancias actuales, es probable que se realicen de forma virtual. Esta modalidad permitirá grabar las sesiones para su posterior revisión y replicación.

Es crucial que en los talleres o reuniones finales participen tanto técnicos especializados en la gestión de datos, como expertos del Ministerio de Ambiente y Desarrollo Sostenible con experiencia en la problemática de interés.

Las fases propuestas para este proyecto están alineadas con la Metodología CRISP-DM, la cual es ampliamente utilizada en el desarrollo de proyectos de análisis de datos y cuenta con un sólido soporte teórico y práctico. Ver Gráfica 5.

Cabe destacar que la ejecución del proyecto se orienta hacia un enfoque ágil y colaborativo, asegurando la participación continua de los expertos del negocio. La oficina TIC establecerá un cronograma de reuniones semanales, donde estos expertos tendrán la oportunidad de contribuir de manera activa en diversas actividades del proyecto.



Gráfica 5 Metodología CRISP-DM.

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

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Documentar el proceso de ejecución CRISP-DM llevado a cabo para la generación de los modelos a entregar, este debe contener todas las etapas de la metodología para describir el proceso que se llevó a cabo con cada una de las preguntas de investigación a resolver.

Para documentar la fase de "Evaluación de la utilidad", se debe incluir los siguientes elementos:

- Documentación de cómo los resultados de los modelos desarrollados pueden contribuir a la toma de decisiones en el contexto empresarial.
- Descripción de la evaluación de la relevancia y utilidad de los modelos para resolver el problema y alcanzar los objetivos empresariales.
- Descripción del cronograma de planificación y organización de los talleres diseñados para capacitar a los participantes en el uso de los modelos desarrollados que considere la duración de los talleres, su modalidad (virtual o presencial), y la audiencia objetivo.
- Documentación del involucramiento de técnicos especializados en la gestión de datos y expertos del Ministerio de Ambiente y Desarrollo Sostenible con experiencia en la problemática de interés en los talleres.

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

En esta fase, se llevará a cabo la integración de la solución analítica desarrollada en el proyecto con los sistemas de Monitoreo Ambiental existentes. El objetivo principal es aprovechar la solución analítica para el problema de interés en tiempo real.

En esta fase se realiza la documentación detallada de la arquitectura de datos y los componentes desarrollados.

Informe de pruebas y validación que documente los resultados obtenidos durante las pruebas de integración.

Procedimientos y guías para el mantenimiento y gestión continua de la integración en producción.

4.7 Matriz de riesgo 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 4 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.



SC-2000142



SA-2000143

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

Riesgos	Descripción
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.

4.8 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.

4.9 Visualizaciones y API's sobre los componentes desarrollados

Documente el manual técnico y de usuario de las visualizaciones, api's y componentes desarrollados para uso del ministerio como cliente final la solución de analítica.

El manual técnico proporciona información detallada sobre la implementación, configuración y funcionamiento interno de las visualizaciones, APIs y componentes desarrollados. Incluye detalles técnicos relevantes, como requisitos de instalación, arquitectura del sistema, configuración de parámetros, y cualquier otro aspecto técnico que pueda ser útil para su comprensión y mantenimiento.

Por otro lado, el manual de usuario está dirigido a los usuarios finales del Ministerio que utilizarán estas herramientas en su trabajo diario. Proporciona instrucciones claras y concisas sobre cómo acceder, navegar y

MINISTERIO DE AMBIENTE Y DESARROLLO SOSTENIBLE	GUIA ARQUITECTURA DE PROYECTOS DE ANALÍTICA	
	Proceso: Gestión Estratégica de Tecnologías de la Información	
Versión: 1	Vigencia: 25/06/2024	Código: G-E-GET-46

utilizar las visualizaciones y APIs de manera efectiva. Además, incluye ejemplos de uso, capturas de pantalla y cualquier otra información que pueda facilitar su comprensión y aprovechamiento.

Los dos manuales mencionados anteriormente deben incorporar las visualizaciones como APIs y componentes que fueron generados como parte del proyecto de analítica.

4.10 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 5 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						
Entrenar modelos						
Evaluar la utilidad						

Detalle de los productos esperados:

Tabla 6 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.				



SC-2000142



SA-2000143