Computación Científica

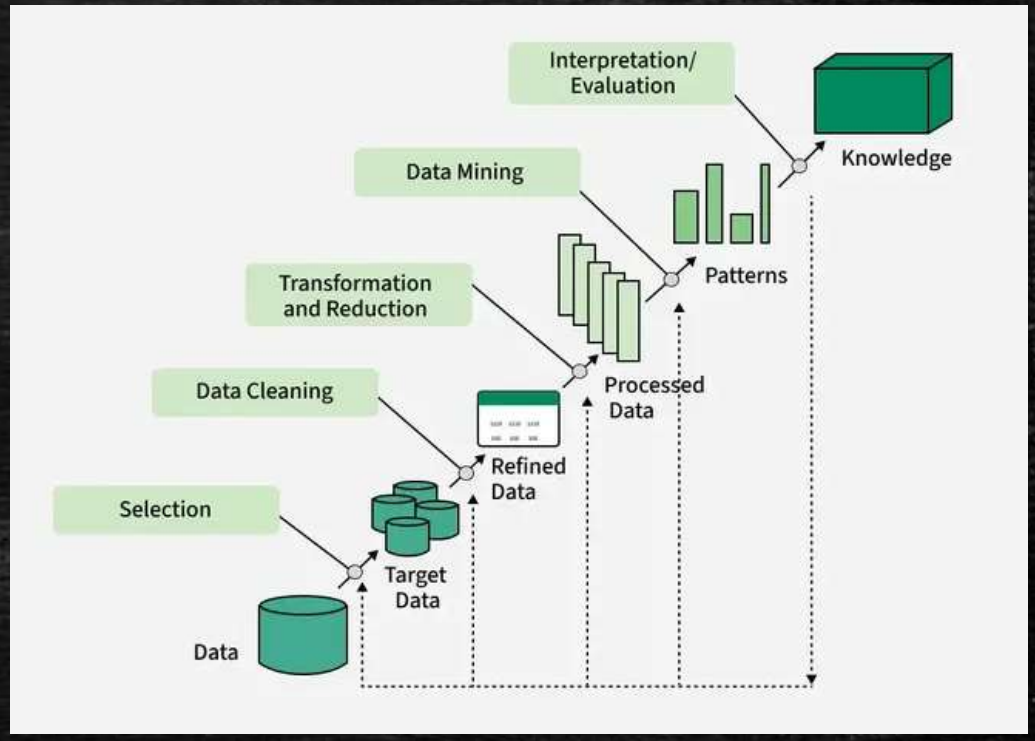
Juan Sebastián Valencia Villa
juan.valencia72@eia.edu.co

Metodologías

KDD	CRISP - DM	TDSP	ASUM - DM
• Knowledge Discovery in Databases	• Cross Industry Standard Process for Data Mining • Definida por IBM	• Team Data Science Process • Definida por Microsoft	• Analytic Solution Unified Method for Data Mining • Evolución de CRISP – DM

Metodología 1: KDD

- **Selección:** Se identifican las fuentes de datos relevantes.
- **Preprocesamiento:** Tratamiento de valores faltantes, ruido e inconsistencias.
- **Transformación:** Conversión de datos a formatos adecuados.
- **Data Mining:** Aplicación de algoritmos de estadística avanzada y machine learning para identificar patrones.
- **Interpretación/Evaluación:** Análisis de resultados y generación de conocimiento.



Fase 1: Selección

Actividades Clave

- Identificación de fuentes de datos (Bases de datos, logs, archivos, sensores).
- Selección de variables relevantes.
- Definición de población y periodo de análisis.

Artefactos

- Documento de fuentes de datos.
- Diccionario preliminar de variables.
- Conjunto de datos bruto.
- Justificación de inclusión o exclusión de variables.

Fase 2: Preprocesamiento

Actividades Clave

- Tratamiento de valores faltantes.
- Eliminación de duplicados.
- Corrección de errores lógicos.
- Detección de atípicos.

Artefactos

- Conjunto de datos limpio.
- Reporte de calidad de datos.
- Reglas de limpieza aplicadas.
- Métricas de calidad.

Fase 3: Transformación

Actividades Clave	Artefactos
• Normalización o estandarización. • Discretización. • Agregaciones. • Reducción de dimensionalidad. • Codificación de variables categóricas.	• Conjunto de datos transformado. • Código con transformaciones. • Nuevas variables. • Documentación de transformaciones.

Fase 4: Data Mining

Actividades Clave

- Selección de tarea.
- Entrenamiento de modelos.
- Ajuste de parámetros.

Artefactos

- Modelos ajustados.
- Reglas descubiertas.
- Clústeres.
- Métricas de desempeño.

Fase 5: Interpretación y Evaluación

Actividades Clave

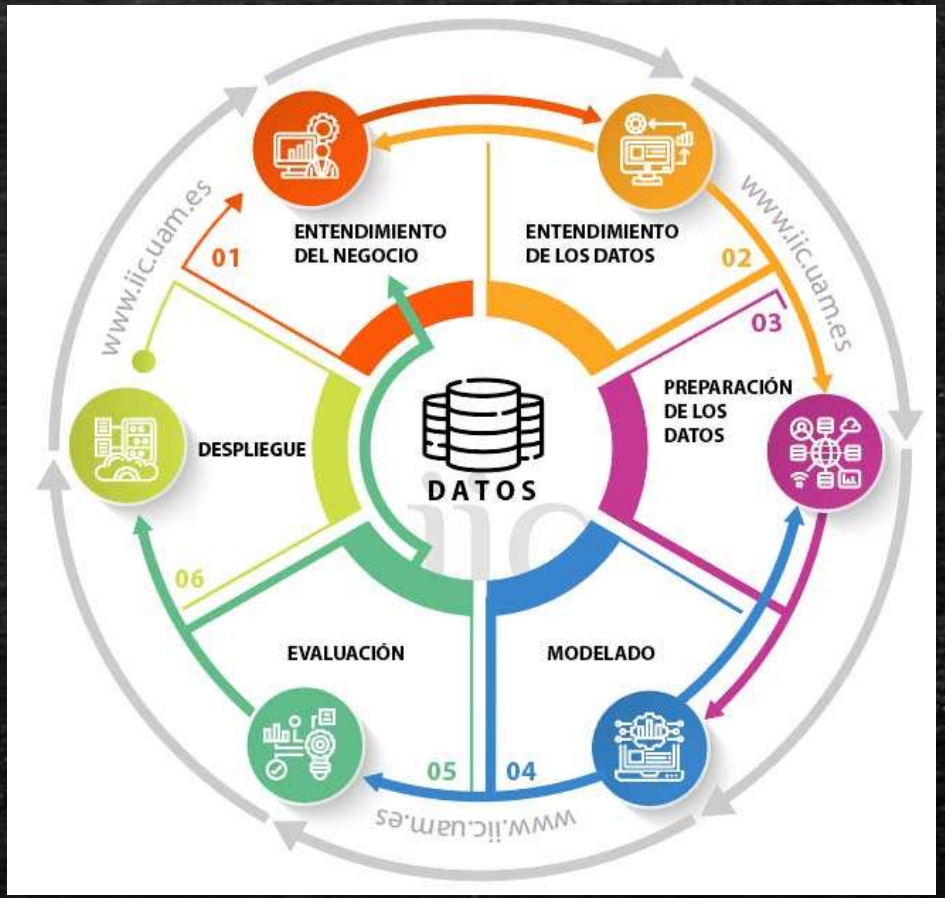
- Validación de resultados.
- Interpretación de patrones.
- Evaluación de utilidad.
- Identificación de conocimiento accionable.

Artefactos

- Informe de conocimiento descubierto.
- Visualización de explicativas.
- Reglas interpretadas.
- Conclusiones y limitaciones.

Metodología 2: CRISP - DM

- **Entendimiento del Negocio:** Determinar los objetivos del negocio, determinar las metas analíticas, definir la situación actual y construir el plan de desarrollo del proyecto.
- **Entendimiento de los datos:** Recolección inicial de los datos, descripción de los datos, exploración de los datos y verificación de la calidad del dato.
- **Preparación de los datos:** Seleccionar datos, limpieza de datos, construcción de atributos, integrar datos, formatear datos.
- **Modelado:** Seleccionar técnicas de modelado de datos, diseñar el esquema de pruebas, construir el modelo, evaluar el modelo.
- **Evaluación:** Evaluar resultados, revisar el proceso, determinar siguientes pasos.
- **Despliegue:** Plan de despliegue, plan de monitoreo y mantenimiento, producción del reporte final y revisar el proyecto.



Fase 1: Entendimiento del negocio

- Obtener la máxima información posible de los objetivos "comerciales".
  - Recopilar información sobre la situación actual del negocio.
  - Registrar los objetivos organizacionales específicos.
  - Definir los criterios de determinación del rendimiento del proceso analítico.
- Generación del plan de proyecto.

Determinar los objetivos del negocio

- Definir el contexto
- Determinar los objetivos del negocio
- Definir los criterios de éxito

Detectar fraude

Asegurar el éxito de una campaña

Evaluación de la situación

- Realizar el inventario de recursos
- Definir requisitos supuestos y restricciones
- Identificar las contingencias
- Generar la terminología base
- Establecer los costos y beneficios

¿Qué información se tiene disponible?

¿Hay suficientes datos?

Determinar los objetivos de desarrollo analítico

- Definir los objetivos analíticos
- Definir los criterios de éxito del modelo analítico

Determinar el perfil de cliente

Realizar el plan del proyecto

- Estructurar el plan del proyecto
- Ejecutar esquemas de evaluación inicial

Definición del paso a paso del proyecto

Fase 2: Entendimiento de los datos

- Acceder a los datos y explorarlos con la ayuda de estadística descriptiva.
- Recopilación de los datos:
  - Datos existentes.
  - Datos adquiridos.
  - Datos adicionales.
- Necesidades:
  - Identificación de los atributos.
  - Relevancia de los atributos.
  - Disponibilidad de los datos.
  - Calidad y completitud.

Recolectar los datos iniciales

- Identificar los datos adquiridos
- Identificar la ubicación de los datos
- Determinar los mecanismos de recolección y adquisición

Descripción de los datos

- Establecer el volumen de los datos
- Generar un compendio de términos del negocio

Exploración de los datos

- Establecer métricas iniciales de los datos
- Generar un informe exploratorio de los datos

Verificar la calidad de los datos

- Identificar la consistencia de los valores individuales
- Determinar correspondencia, completitud y esquemas de corrección

Fase 3: Preparación de los datos

- Es uno de los aspectos más importantes y que más tiempo exige.
  - Fusión de conjuntos y/o registros de datos.
  - Selección de una muestra de datos.
  - Agregación de registros.
  - Derivación de nuevos atributos.
  - Clasificación de datos de modelado.
  - Eliminación o sustitución de valores en blanco o ausentes.
  - División en conjunto de prueba y entrenamiento.

Seleccionar datos

- Identificar un subconjunto de datos

Limpiar datos

- Enriquecer los datos
- Optimizar la calidad de los datos

Construir datos

- Definir uso de ingeniería de características
- Transformar, eliminar o integrar datos

Integrar datos

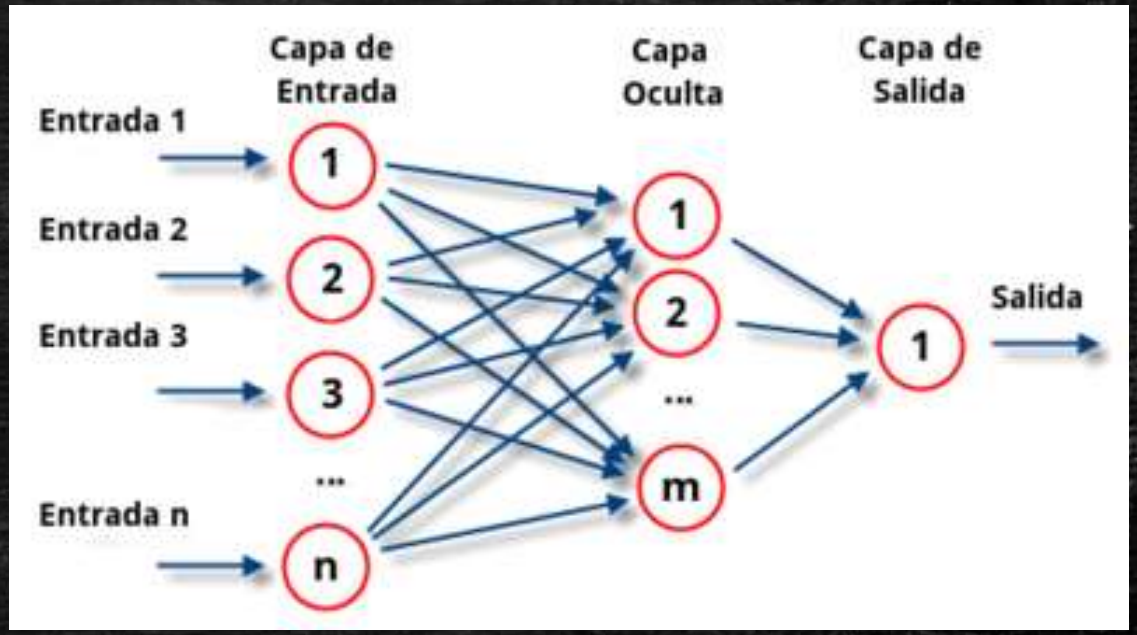
- Generar nuevos campos

Formatear datos

- Estandarizar datos
- Eliminar caracteres específicos

Fase 4: Modelado

- No es ciencia exacta, requiere iteraciones para llegar a una respuesta deseada.
  - Definir el modelo a través del tipo de datos, categórico o numérico.
  - Definir el modelo a través de los objetivos analíticos.
  - Definir el modelo por el tamaño de datos.
  - Definir el modelo según la explicabilidad.



Escoger la técnica de modelado

Elegir la técnica considerando el objetivo de negocio

Generar el plan de prueba

Determinar los conjuntos de validación

Definir las métricas de ajuste y desempeño

Construir el modelo

Identificar hiper parámetros

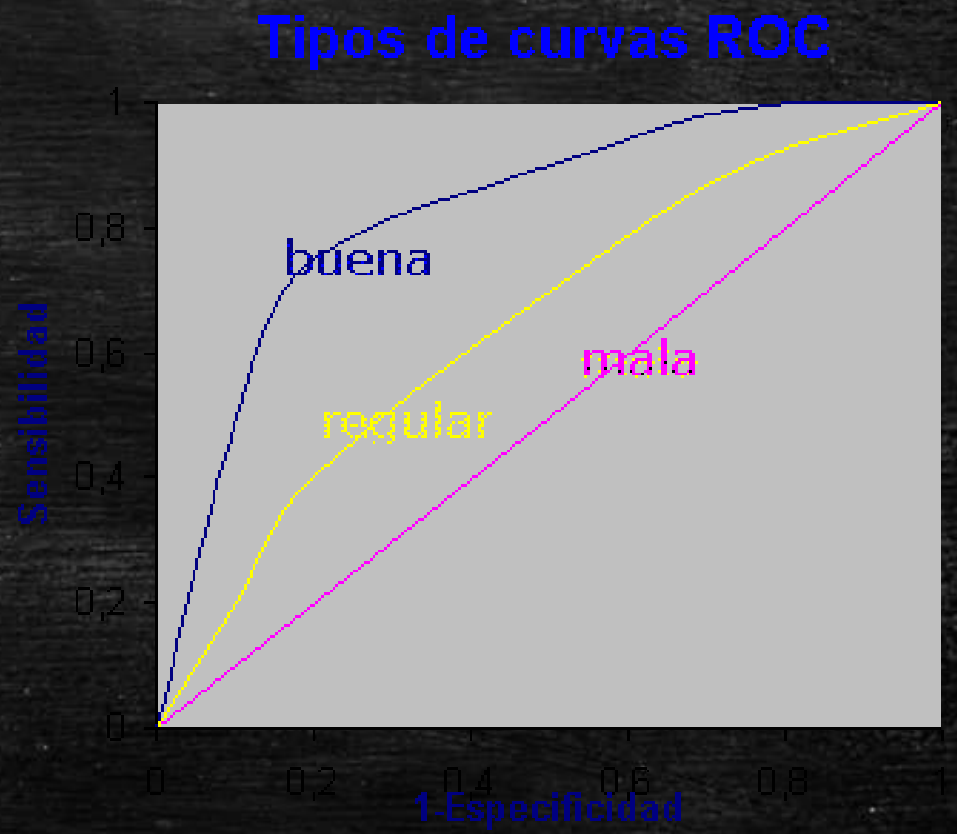
Generar procesos iterativos de identificación

Evaluar el modelo

Comparar según los criterios de éxito

Fase 5: Evaluación

- Se ejecuta desde dos miradas:
  - Cumplimiento de métricas:
    - Ajuste.
    - Error.
      - ECM
      - Curva ROC
      - Distancia
  - Cumplimiento de objetivo de proyecto:
    - Explicabilidad.
    - Predicción.
    - Definición.



Evaluar los resultados

Contrastar el modelo con los objetivos analíticos y de negocio

Revisar el proceso

Definir esquemas de mejora

Determinar próximos pasos

Finalizar el proceso de modelado

Iterar para encontrar otra respuesta

Fase 6: Despliegue

- Planificación y control de la distribución de resultados:
  - Resumir los resultados.
  - Integración con los sistemas.
  - Difusión de la información.
  - Visualización.
  - Medición de usabilidad.
  - Tiempo de vida del modelo.
  - Planes de contingencia.
- Finalización de tareas de presentación

Planear la implementación

- Implementar la respuesta en el negocio
- Generar documentación

Planear monitoreo y soporte

- Establecer estrategias de monitoreo
- Definir tiempos de redefinición o actualización

Producir el informe final

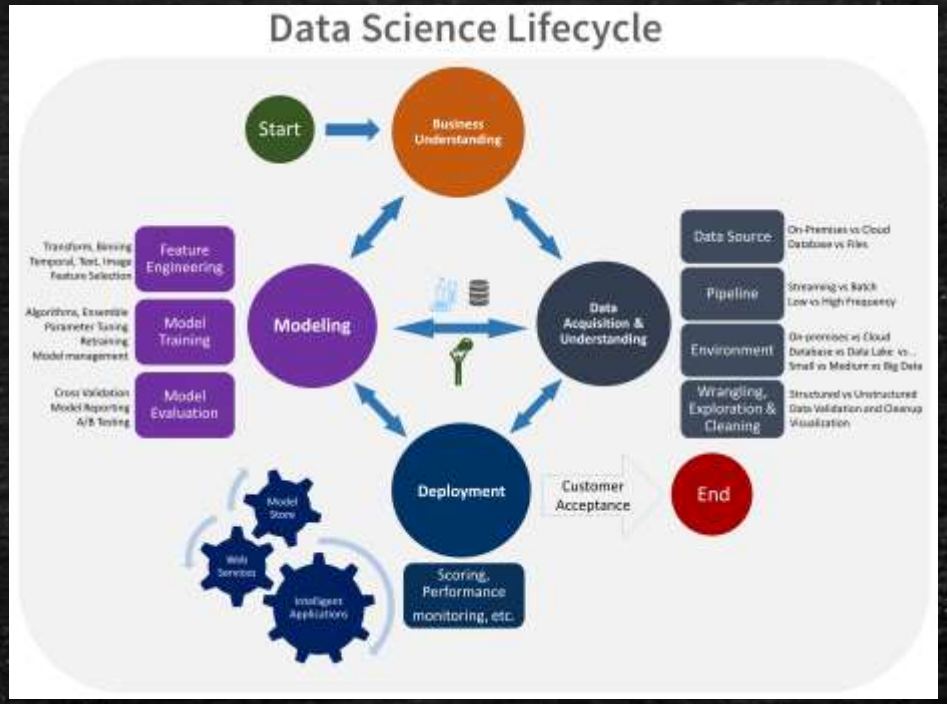
- Generar documento de cierre

Revisar el proyecto

- Documentar lecciones aprendidas

Metodología 3: TDSP

- **Entendimiento del Negocio:** Determinar los objetivos del negocio, determinar las metas analíticas, definir la situación actual y construir el plan de desarrollo del proyecto.
- **Adquisición y entendimiento de los datos:** Recolección inicial de los datos, descripción de los datos, exploración de los datos y verificación de la calidad del dato. Seleccionar, limpiar, integrar y formatear datos.
- **Modelado:** Seleccionar técnicas de modelado de datos, diseñar el esquema de pruebas, construir el modelo, evaluar el modelo.
- **Despliegue:** Plan de despliegue, plan de monitoreo y mantenimiento, producción del reporte final y revisar el proyecto.



Etapa 1: Entendimiento del negocio

Objetivo

- Especificar las variables objetivo del negocio y las métricas de éxito
- Identificar los orígenes de datos existentes o por adquirir

Forma de hacerlo

- **Definición de objetivos:** trabajar de la mano con el cliente y formular las preguntas que lleven a desarrollar los objetivos.
- **Identificar orígenes de datos:** buscar los datos pertinentes que lleven a responder las preguntas que definen los objetivos.

Artefactos

- **Documento marco:** Documento que evoluciona a lo largo del desarrollo donde se describen las necesidades y avances.
- **Orígenes de datos:** Especifica los orígenes de los datos involucrados.
- **Diccionario de datos:** Descripciones de los datos en la que se identifican tipologías y reglas de validación.

Etapa 2: Adquisición y entendimiento de los datos

Objetivo

- Generar un conjunto de datos limpio y de alta calidad.
- Desarrollar una arquitectura de solución.

Forma de hacerlo

- **Introducción de los datos:** Configurar los mecanismos de movilización de los datos hacia el ambiente de análisis.
- **Exploración de los datos:** Visualizar los datos para establecer métricas iniciales de calidad y completitud.
- **Configuración de una canalización de datos:** Establecer mecanismos para puntear datos nuevos y su actualización con regularidad.

Artefactos

- **Informe de calidad de datos:** Contiene el resumen de los datos, las relaciones entre cada atributo y la clasificación de las variables.
- **Arquitectura de la solución:** Diagrama o descripción de la canalización de los datos para llevar a cabo la tarea de puntuación de datos o predicciones.
- **Decisión de punto de control:** Evaluar el proyecto para establecer su pertinencia.

Etapa 3: Modelado

Objetivo

- Determinar las características óptimas de los datos para el modelo de aprendizaje automático.
- Crear un modelo de aprendizaje automático informativo que predice el objetivo con la máxima precisión.
- Crear un modelo de aprendizaje automático que es adecuado para entornos de producción.

Forma de hacerlo

- **Diseño de características:** Crear características a través de los datos sin procesar para facilitar su entrenamiento.
- **Entrenamiento del modelo:** Buscar el modelo que responda a la pregunta de negocio con la máxima precisión.
- **Validación del modelo:** Determinar si el modelo es adecuado para su uso en producción.

Artefactos

- **Conjuntos de características:** Contiene los esquemas de desarrollo de las nuevas características.
- **Informe del modelo:** Informe detallado de cada experimento definido.
- **Decisión de punto de control:** Evaluar el modelo para establecer su pertinencia.

Etapa 4: Implementación

Objetivo

- Implementar modelos con canalización de datos en un entorno de producción o similar para que el usuario final los acepte.

Forma de hacerlo

- **Uso del modelo:** Identificar qué tipos de aplicaciones deben o pueden consumir el modelo.

Artefactos

- Panel de estado que muestra el estado del sistema y métricas clave.
- Informe de modelado final con detalles de implementación.
- Documento de arquitectura final de solución.

Etapa 5: Aceptación del usuario final

Objetivo

- Confirmar que la canalización, el modelo y su implementación cumplen con los objetivos del cliente.

Forma de hacerlo

- **Validación del sistema:** confirmar que el modelo implementado cumplen con las necesidades del cliente.
- **Entrega del proyecto:** Entregar el proyecto a la entidad que va a ejecutar el sistema en producción.

Artefactos

- Informe de salida del proyecto.

Metodología 4: ASUM - DM

- **Entendimiento del Negocio:** Entender los objetivos y requerimientos desde la perspectiva del negocio, para luego convertirlos en un problema que se pueda abordar desde la analítica.
- **Enfoque analítico:** Traducir los objetivos del negocio en metas de analítica.
- **Requerimientos de datos:** Definir los datos que se usarán para cumplir los objetivos del proyecto.
- **Recolección de datos:** Adquirir los datos de diferentes fuentes, tanto internas como externas al negocio.
- **Entendimiento de datos:** Explorar profundamente los datos para generar estadísticas que caractericen los datos adquiridos.
- **Preparación de datos:** Ejecutar de transformaciones necesarias para dar uso de los datos en etapas de modelado.
- **Construcción del modelo:** Construir de modelos analíticos enfocados a la solución del negocio.
- **Evaluación del modelo:** Aplicar métricas de evaluación con miras a verificar que los modelos tenga un correcto desempeño y resuelvan los objetivos de negocio.
- **Despliegue de solución:** Definir cómo se presentarán los resultados del modelo.
- **Retroalimentación:** Retroalimentar la solución por parte del negocio.

Etapas de entendimiento

- **Comprensión del negocio:**
  - Busca definir el problema, los objetivos de cara al negocio y los requisitos de la solución a nivel empresarial.
  - Sienta las bases para que el problema empresarial sea resuelto.
- **Enfoque analítico:**
  - Busca expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático.
  - Determinar el tipo de problema y definir el método de construcción.

Etapas de preparación

- **Requisitos de datos:**
  - Dependiendo del enfoque analítico se determinan los contenidos de datos, formatos y representaciones.
  - Busca definir de manera clara el insumo de datos que requiere el problema.
- **Recopilación de datos:**
  - Se reúnen los recursos de datos disponibles que son relevantes para resolver el problema.
  - Busca entonces definir los mecanismos por los cuáles se capturan los insumos necesarios para la implementación de un modelo.
- **Comprensión de datos:**
  - Busca conocer el estado actual del insumo.
  - Implementa herramientas estadísticas descriptivas enfocadas a descubrir elementos iniciales de los datos.

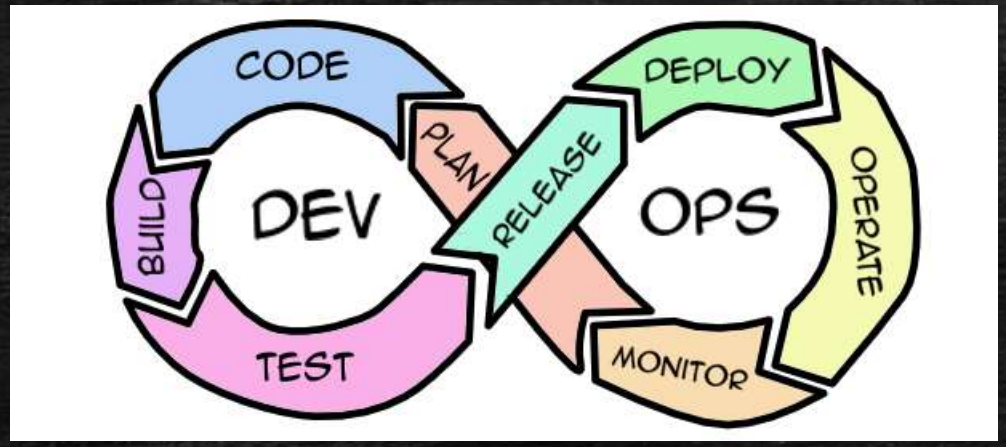
Etapas de ejecución

- **Preparación de datos:**
  - Busca construir el conjunto de datos que se utilizará para modelar.
  - Se implementan mecanismos de ingeniería de características enfocados a mejorar la capacidad predictora de las variables.
- **Modelado:**
  - Busca identificar la mejor técnica de modelado de acuerdo al enfoque analítico.
  - Se determina un esquema de desempeño que permita decidir el mejor método aplicado.
- **Evaluación:**
  - Busca interpretar la calidad del componente entregado bajo mecanismos gráficos y numéricos.

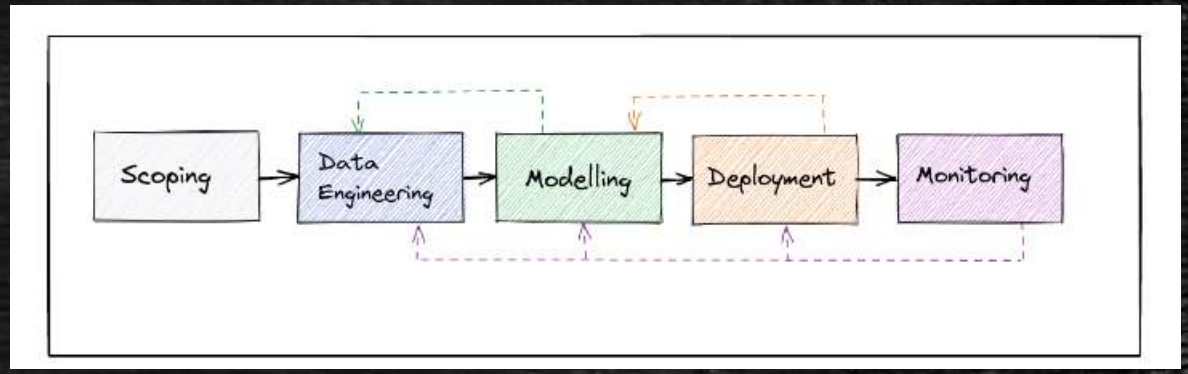
Etapas de cierre

- **Implementación:**
  - Busca implementar el modelo definido en ambientes de producción.
- **Retroalimentación:**
  - Busca determinar la efectividad futura del modelo implementado y sus posibles modificaciones en soporte.

DevOps VS MLOps

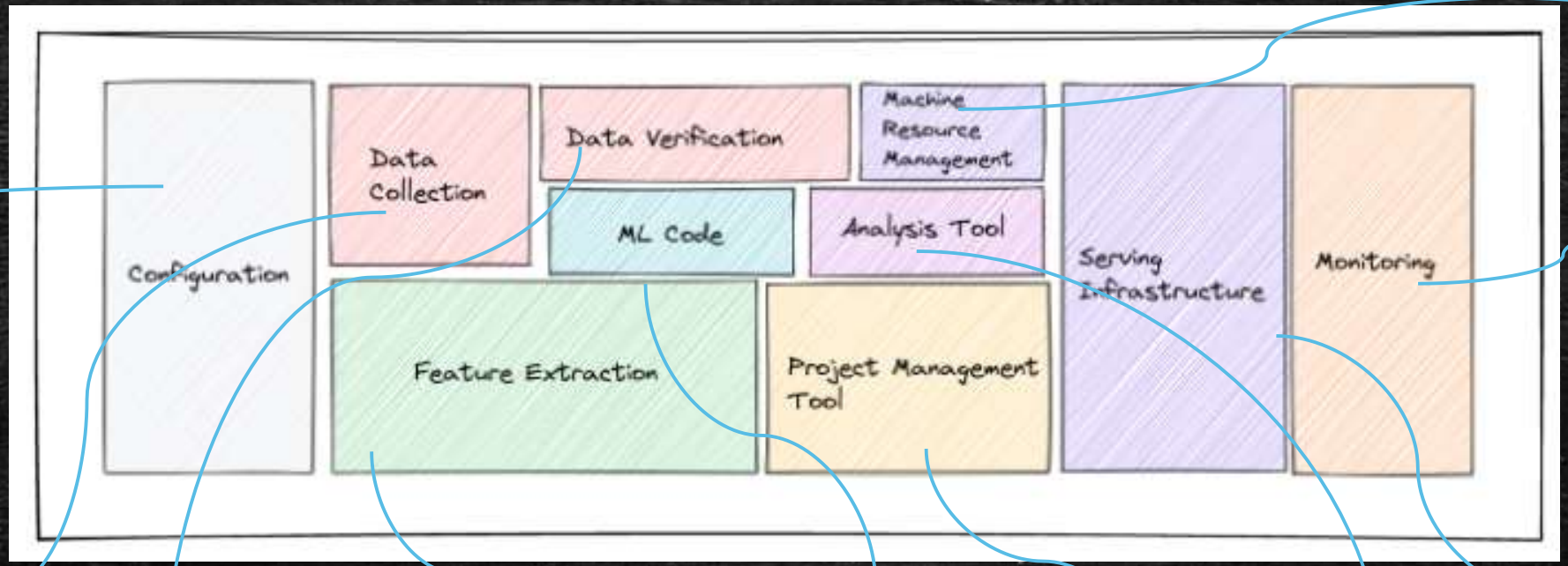


+ Se usa para el desarrollo de software. Se planean características, se escribe y construye el código, se prueba, crea un release y se despliega. Finalmente se opera y monitorea su comportamiento.



+ Se usa para el desarrollo de modelos de aprendizaje automático. Se valida si se puede desarrollar por modelos de aprendizaje, se hace ingeniería de características, se modela, despliega y monitorea el comportamiento del modelo en arquitectura.

Infraestructura de producción de ML



Configuración: Protocolos de comunicaciones, integraciones y tuberías de datos para desarrollo.

Recolección de datos: Recolección de datos desde diferentes fuentes. Consolidan datos de fuentes diversas.

Verificación de datos: Validación si los datos son correctos para el modelo y se encuentran estructurados.

Extracción de características: Se seleccionan las mejores variables para predecir.

Código de ML: Se desarrolla el código base del modelo que se utilizará para predecir.

Herramienta de administración de proyecto: Se hace seguimiento de las actividades necesarias.

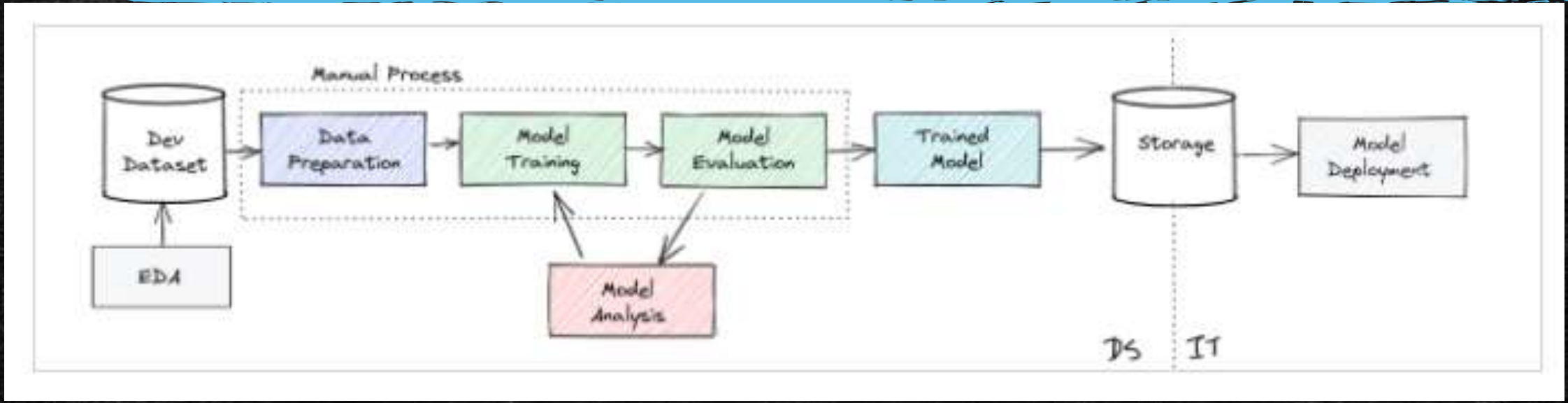
Herramienta de análisis: Calcula el desempeño del modelo en operación.

Infraestructura de servicio: Se implementa la arquitectura analítica que soporta el modelo de decisión.

Monitoreo: Se implementa un esquema de monitoreo para conocer uso, fallos y predicciones.

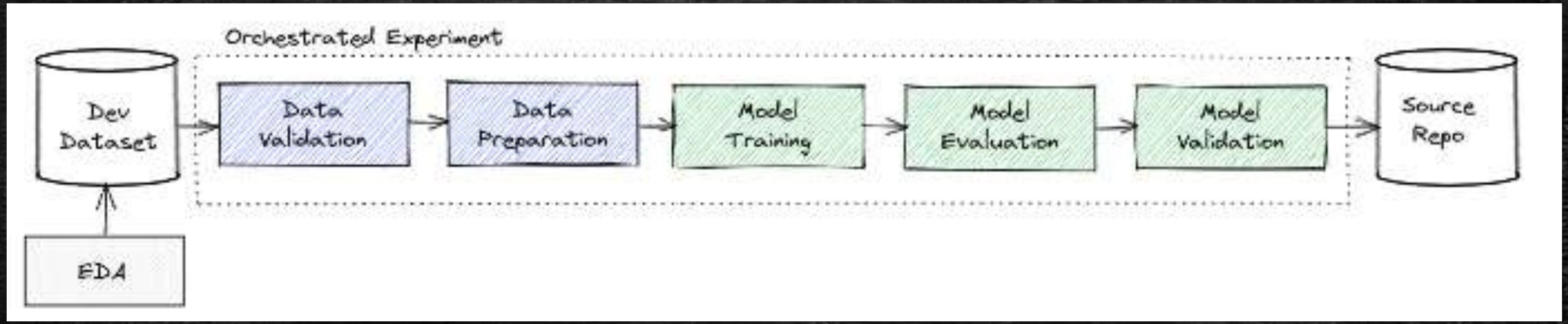
Administración de recursos: Se hacen disponibles los recursos necesarios para operar un modelo.

MLOps Nivel 0



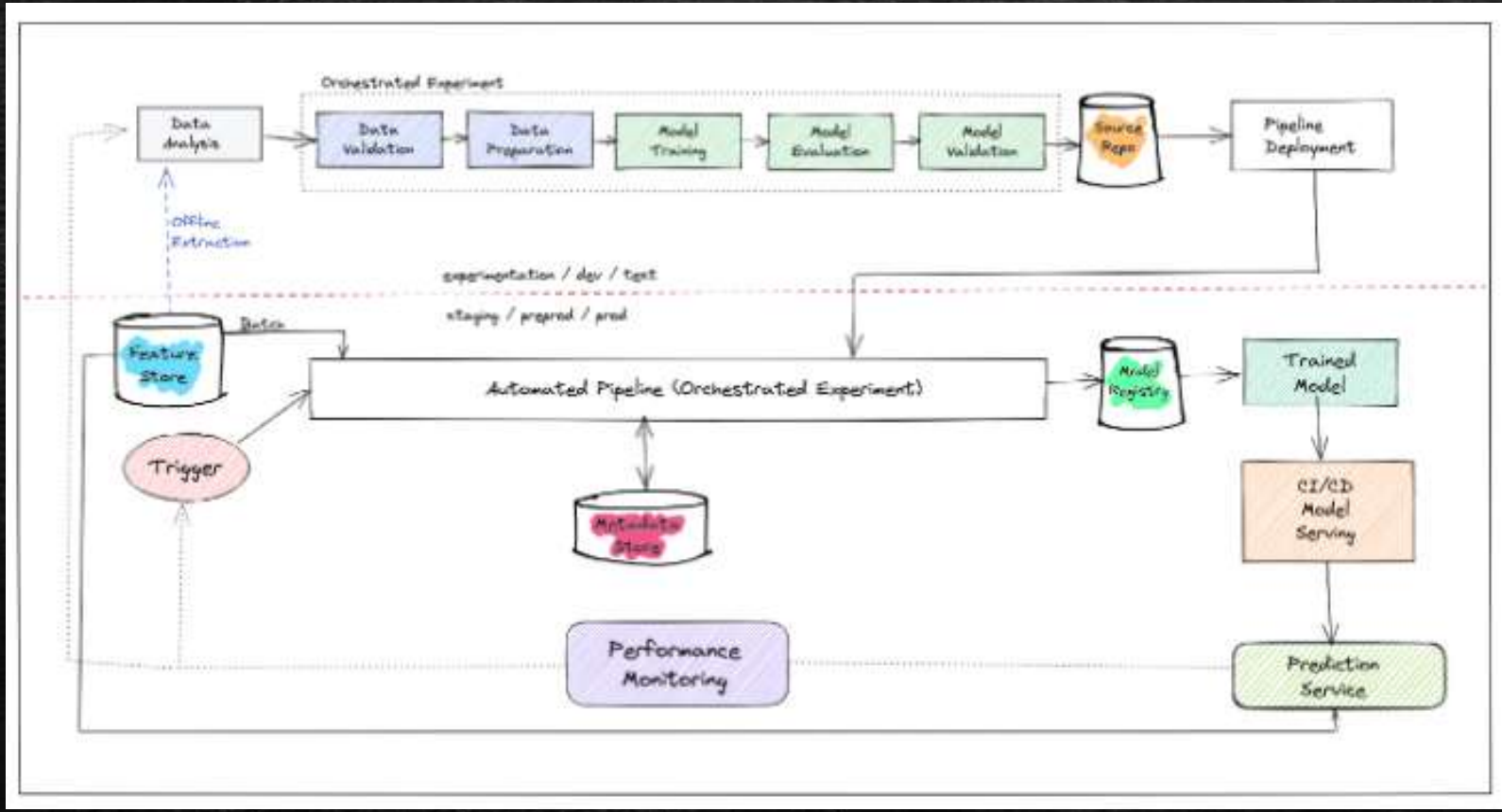
Cuando los procesos de aprendizaje se ejecutan de forma manual, experimentación para implementación, estamos hablando de un nivel 0.

MLOps Nivel 1

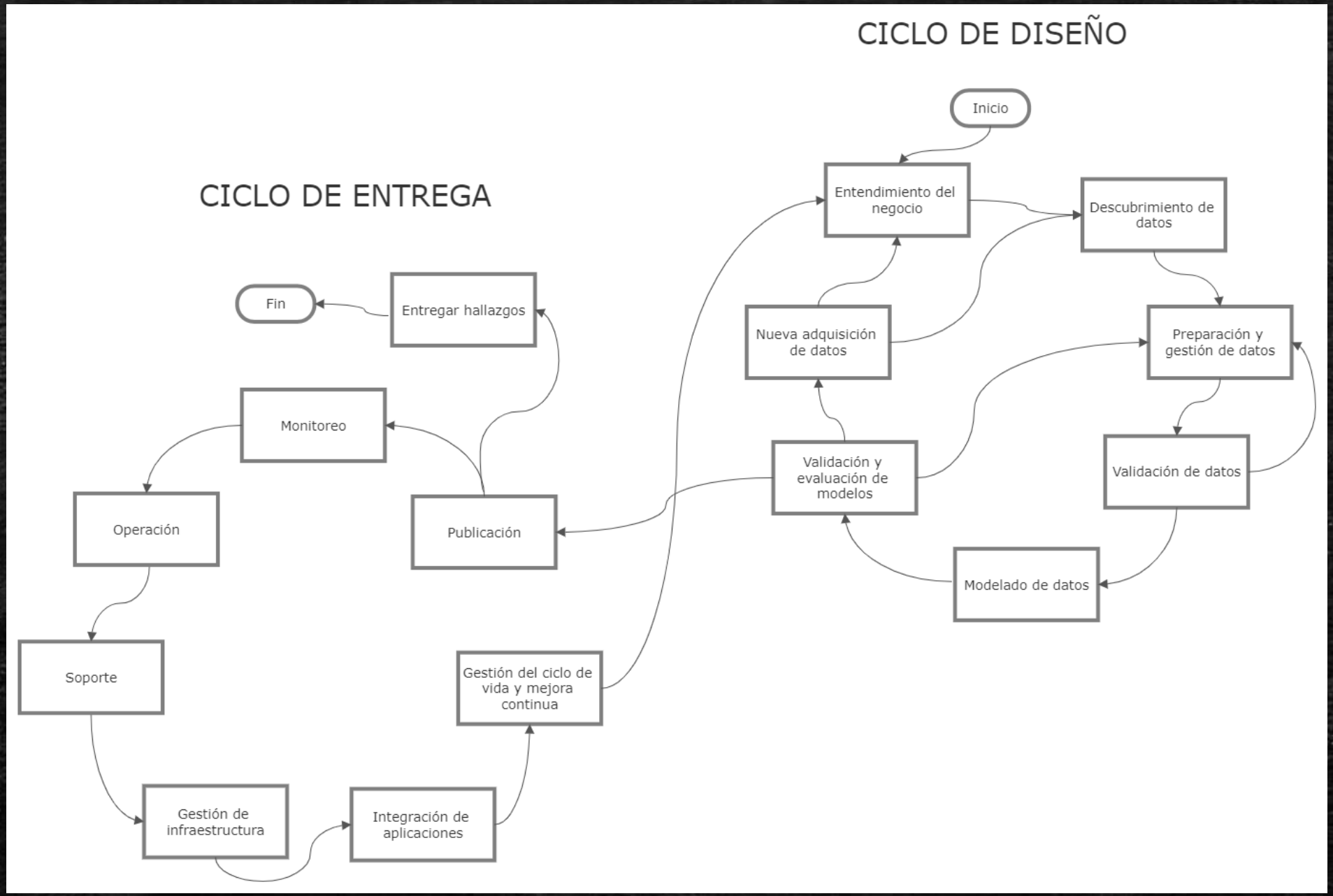


Cuando creamos un proceso automatizado que permite realizar validación, preparación y experimentación para implementación, estamos hablando de un nivel 1.

MLOps Nivel 2



Cuando automatizamos el modelo al nivel que se reentrene automáticamente bajo nuevos datos, estamos en nivel 2.


CICLO DE DISEÑO
Inicio
Entendimiento del negocio
Descubrimiento de datos
Preparación y gestión de datos
Validación de datos
Modelado de datos
Validación y evaluación de modelos
Nueva adquisición de datos
CICLO DE ENTREGA
Fin
Entregar hallazgos
Monitoreo
Operación
Publicación
Soporte
Gestión de infraestructura
Integración de aplicaciones
Gestión del ciclo de vida y mejora continua

Referencias

+ https://www.ibm.com/downloads/cas/WKK9DX51
+ https://view.genial.ly/60832dde2e6fa40d5be899fb/presentation-infografiaasum-dm
+ https://towardsdatascience.com/a-gentle-introduction-to-mlops-7d64a3e890ff
+ https://docs.microsoft.com/es-es/azure/machine-learning/algorithm-cheat-sheet