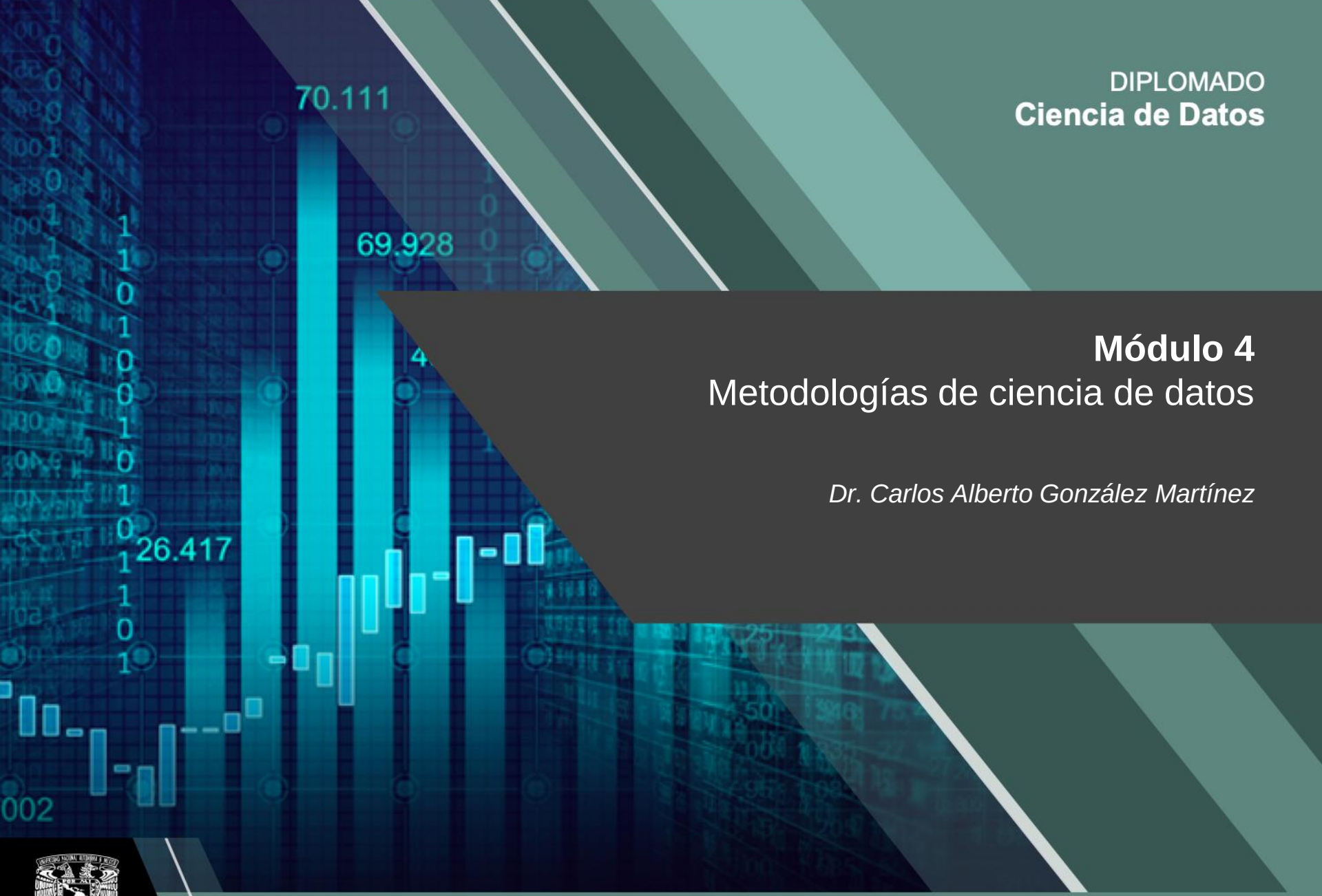


DIPLOMADO
Ciencia de Datos

Módulo 4
Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

El ciclo de la ciencia de datos

Objetivo

El participante identificará las etapas del ciclo de la ciencia de datos.

El ciclo de la ciencia de datos

Contenido

1. **Conocimiento del negocio**
2. Adquisición y comprensión de los datos
3. Modelado
4. Implementación
5. Aceptación del cliente

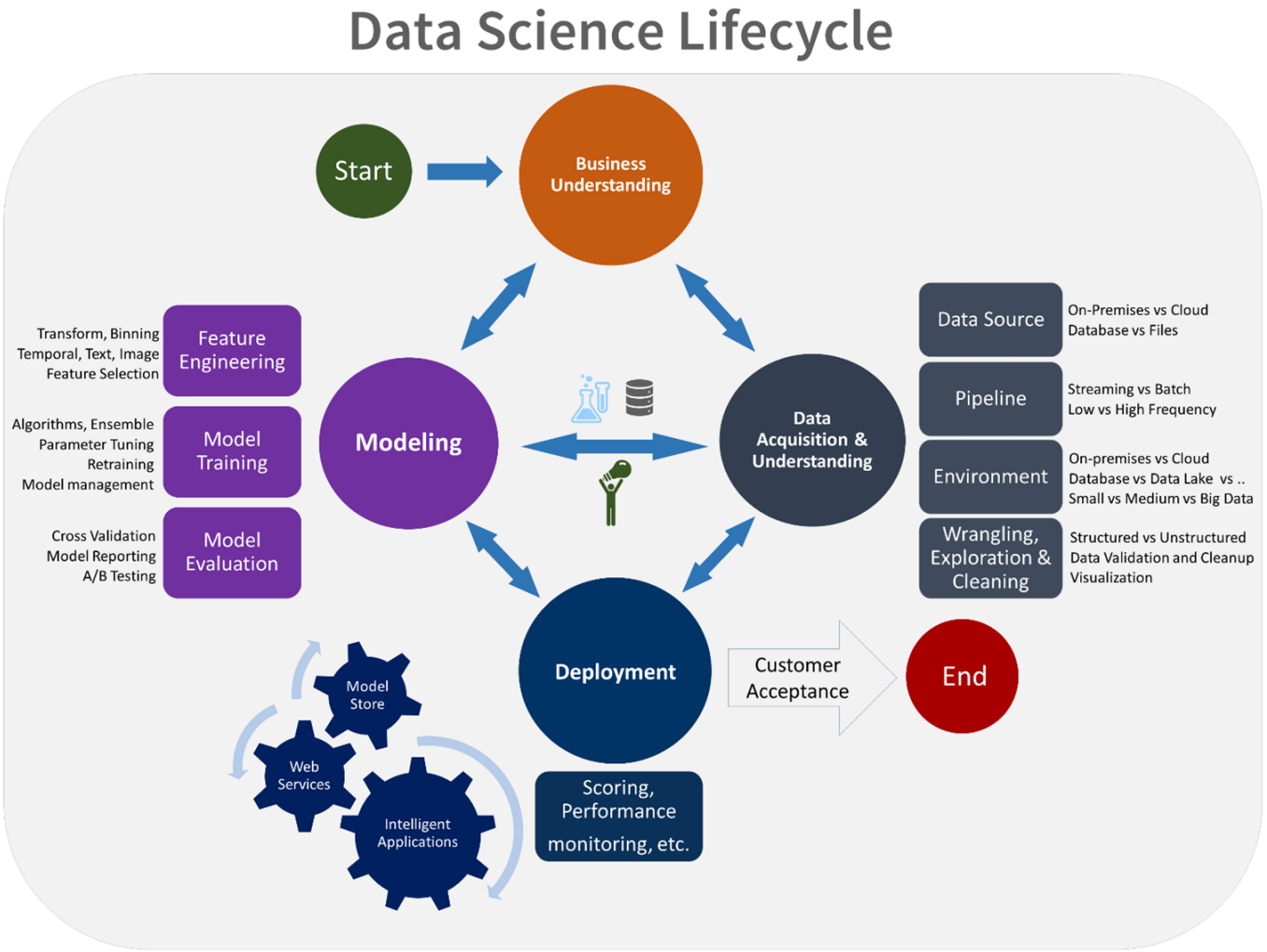
El ciclo de la ciencia de datos

El ciclo de la ciencia de datos representa las diferentes etapas recomendables, por las que debe de pasar un proyecto de ciencia de datos.

El Proceso de ciencia de datos en equipo (TDSP) ofrece un ciclo de vida recomendado, que se puede usar para estructurar los proyectos.

El ciclo de vida del TDSP se modela como una secuencia de pasos repetidos, que le orientan con respecto a las tareas necesarias para usar modelos predictivos.

El ciclo de la ciencia de datos



gmc

El ciclo de la ciencia de datos

En cada fase es importante contar con:

- **Objetivos**: objetivos específicos.
- **Cómo hacerlo**: un esquema de las tareas específicas y orientación sobre cómo realizarlas.
- **Artefactos**: las entregas y la asistencia para producirlas.

El ciclo de la ciencia de datos

1. **Conocimiento del negocio**

En esta fase se abordan dos tareas principales:

Definición de objetivos: trabaje con el cliente y con otras partes interesadas para comprender e identificar los problemas de la empresa. Formule preguntas que definan los objetivos empresariales y a las que puedan aplicarse las técnicas de ciencia de datos.

Para definir los objetivos del proyecto, plantee y ajuste preguntas "certeras“, que sean pertinentes, específicas y sin ambigüedad alguna.

El ciclo de la ciencia de datos

1. Conocimiento del negocio

La ciencia de datos o el aprendizaje automático suelen utilizarse para responder a cinco tipos de preguntas, las cuales se asocian al tipo de solución:

¿Cuánto? o ¿cuántos? (regresión)
¿Qué categoría? (clasificación)
¿Qué grupo? (agrupación en clústeres)
¿Es extraño? (detección de anomalías)
¿Qué opción se debe elegir? (recomendación)

El ciclo de la ciencia de datos

1. **Conocimiento del negocio**

Los objetivos del proyecto deben quedar perfectamente definidos, al igual que su criterio de éxito, el cual se sugiere que sea de carácter cuantitativo, así como determinar si se va a implementar en el ambiente de producción o solo se entregan los resultados del modelo.

Defina las métricas del éxito. Las métricas deben cumplir los requisitos **SMART**:

S pecific (específicas)
M easurable (medibles)
A chievable (alcanzables)
R elevant (pertinentes)
T ime-bound (con un límite de tiempo)

El ciclo de la ciencia de datos

1. Conocimiento del negocio

En esta fase se abordan dos tareas principales:

Identifique los orígenes de datos: busque los datos pertinentes que lo ayuden a responder las preguntas que definen los objetivos del proyecto.

Identificar las diferentes fuentes de datos, las cuales pueden ser:

- Bases de datos
- Archivos xls
- Archivos de texto plano, etcétera

Es necesario contar con el acceso a los datos

El ciclo de la ciencia de datos

1. **Conocimiento del negocio**

Se recomienda contar con un artefacto para documentar el origen de los datos.

IBM propone el siguiente artefacto para el origen de los datos:

Dataset Name	Original Location	Destination Location	Data Movement Tools / Scripts	Link to Report
Dataset 1	Brief description of its orignal location	Brief description of its destination location	script1.py	Dataset 1 Report
Dataset 2	Brief description of its orignal location	Brief description of its destination location	script2.R	Dataset 2 Report

gmc

El ciclo de la ciencia de datos

Contenido

1. Conocimiento del negocio
2. **Adquisición y comprensión de los datos**
3. Modelado
4. Implementación
5. Aceptación del cliente

El ciclo de la ciencia de datos

2. Adquisición y comprensión de los datos

En esta fase se abordan tres tareas principales:

- **Introducción de los datos** en el entorno de análisis de destino.
- **Exploración de los datos** para determinar si su calidad es suficiente para responder a la pregunta planteada.
- **Configuración de una canalización de datos** para puntuar los datos nuevos o que se actualizan con regularidad.

El ciclo de la ciencia de datos

Introducción de los datos

Consiste en configurar el proceso para mover los datos desde las ubicaciones de origen a las ubicaciones de destino, donde se ejecutan las operaciones de análisis, como el entrenamiento y las predicciones.

Exploración de los datos

Antes de entrenar los modelos, debe desarrollar una comprensión sólida de los datos. A menudo, los conjuntos de datos reales contienen ruido, les faltan datos o presentan un sinfín de discrepancias de otros tipos.

El ciclo de la ciencia de datos

Configuración de una canalización de datos

Además de la introducción y la limpieza iniciales de los datos, suele ser preciso configurar un proceso para puntuar los datos nuevos o actualizarlos con regularidad durante el proceso de aprendizaje continuo. La puntuación puede completarse con una canalización de datos o un flujo de trabajo.

Al finalizar esta etapa, es recomendable generar un reporte de la calidad de los datos, el cual cuente con:

- Descripción de los datos
- Calidad de los datos
- Información de las variables necesarias para el proyecto
- Análisis exploratorio de los datos

El ciclo de la ciencia de datos

Contenido

1. Conocimiento del negocio
2. Adquisición y comprensión de los datos
3. **Modelado**
4. Implementación
5. Aceptación del cliente

El ciclo de la ciencia de datos

3. Modelado

En esta fase se abordan tres tareas principales:

- **Diseño de características**: cree características de datos a partir de los datos sin procesar, para facilitar el entrenamiento del modelo.
- **Entrenamiento del modelo**: busque el modelo que responda a la pregunta con la máxima precisión, comparando sus métricas de éxito.
- Determine si el modelo es **adecuado para su uso en producción**.

El ciclo de la ciencia de datos

El diseño de características consiste en incluir, agregar y transformar variables sin procesar, para crear las características que se utilizan en el análisis.

En otras metodologías, esta etapa se conoce como la derivación de variables. Si desea conocer con detalle los factores en que se basa un modelo, debe entender cómo se relacionan entre sí las características y cómo deben utilizarlas los algoritmos de aprendizaje automático.

Se recomienda obtener las correlaciones entre las variables consideradas como importantes.

El ciclo de la ciencia de datos

Entrenamiento del modelo

Según el tipo de pregunta que intenta responder, existen muchos algoritmos de modelado disponibles.

El proceso de entrenamiento del modelo incluye los pasos siguientes:

Dividir los datos de entrada aleatoriamente, para el modelado en un conjunto de datos de aprendizaje y un conjunto de datos de prueba.

Compilar los modelos mediante el conjunto de datos de aprendizaje.

El ciclo de la ciencia de datos

Entrenamiento del modelo

Evaluar el entrenamiento y el conjunto de datos de prueba. Use una serie de algoritmos de aprendizaje automático paralelos, junto con los diversos parámetros de ajuste asociados (lo que se denomina *barrido de parámetros*), que están orientados a responder la pregunta de interés con los datos actuales.

Determinar la "mejor" solución para responder a la pregunta, mediante la comparación de las métricas de éxito entre los métodos alternativos.

Es recomendable generar un reporte del modelo utilizado, las variables utilizadas, la puesta a punto del modelo y los resultados obtenidos.

El ciclo de la ciencia de datos

Contenido

1. Conocimiento del negocio
2. Adquisición y comprensión de los datos
3. Modelado
4. **Implementación**
5. Aceptación del cliente

El ciclo de la ciencia de datos

4. Implementación

Esta etapa se encarga de la implementación del modelo en el ambiente de producción.

En los objetivos del proyecto se estableció si se va a implementar en el ambiente de producción y solo se entregarán los resultados y el modelo desarrollado.

Usualmente, las empresas solo están interesadas en el desarrollo del modelo y en conocer sus resultados. En este caso el proyecto termina con los resultados del mismo. La empresa se encarga, a través de su área de sistemas, de la implementación de la solución en su ambiente de producción.

El ciclo de la ciencia de datos

4. Implementación

Otra solución comúnmente usada por las empresas, consiste en adquirir el conocimiento, no solo el modelo. En este caso se estila integrar a uno o varios elementos de la empresa en el desarrollo del modelo. De esta forma el know how queda en casa.

La solución menos común consiste en la implementación en el ambiente de producción. Para este caso, es necesario involucrarse en conocer el sistema que utiliza la empresa, o bien, contratar a un experto en el tema. De esta forma la implementación en el ambiente productivo no causa tanto problema.

Es recomendable generar un reporte de la implementación, según sea el caso.

El ciclo de la ciencia de datos

Contenido

1. Conocimiento del negocio
2. Adquisición y comprensión de los datos
3. Modelado
4. Implementación
5. **Aceptación del cliente**

gmc

El ciclo de la ciencia de datos

5. Aceptación del cliente

En esta fase se abordan dos tareas principales:

Validación del sistema: confirme que el modelo implementado y la canalización cumplen las necesidades del cliente.

Entrega del proyecto: entregue el proyecto a la entidad que va a ejecutar el sistema en producción.

El cliente debe validar que el sistema satisface sus necesidades empresariales y responde a las preguntas con una precisión aceptable, para implementarlo en el entorno de producción y usarlo con la aplicación cliente. Se finaliza y revisa toda la documentación. El proyecto se entrega a la entidad responsable de las operaciones.

El ciclo de la ciencia de datos

5. Aceptación del cliente

Es recomendable generar un reporte final, que incluya:

- Contexto del proyecto
- Resumen de la empresa
- Exposición del problema a resolver
- Seguimiento de los datos (fuentes, tratamiento, calidad de datos, etcétera)
- Modelo utilizado
- Validación del modelo
- Implementación
- Guía de implementación
- Diccionario de datos
- Carta de aceptación por parte del cliente

Referencias bibliográficas

Microsoft, (2020). ¿Qué es el Proceso de ciencia de datos en equipo (TDSP)?. Recuperado de: https://www.ibm.com/downloads/cas/6RZMKDN8

IBM, (2020). Metodología fundamental para la ciencia de datos. Recuperado de: https://www.ibm.com/downloads/cas/6RZMKDN8

Contacto

Carlos Alberto González Martínez

Jefe de departamento de correlaciones, cruces y alertas (C5i)

gmcmmiv@hotmail.com