1

Conferencia 1

Introducción a la Ingeniería de Datos. Metodologías para la Ciencia de Datos

¿Cómo definirían ustedes la ciencia de datos?

Pausa para que los estudiantes compartan sus ideas.

Diferentes autores han intentado definir el campo de la ciencia de datos; algunas de las prespectivas son:

- campo multidisciplinario que combina áreas como la informática, las matemáticas y la estadística. Su objetivo principal es utilizar métodos y técnicas científicas para extraer conocimiento y valor de grandes volúmenes de datos, estructurados o no estructurados. (2019)
- campo interdisciplinario que integra disciplinas como las ciencias de la computación, el aprendizaje automático, las matemáticas y las estadísticas. (2021)
- tema multidisciplinario cuyo propósito es descubrir conocimiento para apoyar la toma de decisiones en diversos contextos empresariales.(2020)

En general, hay un consenso en que la ciencia de datos es un campo multidisciplinario o interdisciplinario que se nutre de áreas como la informática, la estadística y las ciencias de la computación. Su enfoque principal es el estudio de los datos, con el propósito de extraer conocimiento y valor a partir de ellos.

Ahora, ¿qué entienden ustedes por una metodología?

Pausa para que los estudiantes compartan sus ideas.

Una metodología puede entenderse como una **estrategia, guía o conjunto de pautas** que nos ayudan a desarrollar un proceso o actividad de manera estructurada. A diferencia de las herramientas o tecnologías específicas, las metodologías no están ligadas a un software o hardware en particular. En cambio, proporcionan un **marco de trabajo** que nos indica cómo proceder de manera sistemática para alcanzar nuestros objetivos.

¿Por qué es importante esto?
En el contexto de la ciencia de datos, las metodologías ayudan a abordar problemas complejos de manera organizada, intentando que cada paso del proceso esté bien definido y alineado con los objetivos del proyecto.

1.1 Metodologías en Ciencia de Datos

A continuación se presentan varias metodologías de la Ciencia de Datos, analizando críticamente su estructura, enfoque y aplicabilidad en diferentes contextos. El objetivo es comprender cómo cada metodología se adapta -o no- a distintos escenarios organizacionales, técnicos y de complejidad. A través de este análisis, descubriremos por qué ninguna metodología puede aplicarse universalmente a todas las circunstancias, y cómo su implementación rígida y sin adaptación puede llevar a resultados subóptimos.

1.1.1 KDD (Knowledge Discovery in Databases)

Definido como un proceso interactivo e iterativo de 5 fases para descubrir conocimiento por sus creadores en 1996:

- **Selección**: Los datos cambian de acuerdo con los objetivos del proceso. Se establece un grupo de datos con el que proceder.
- **Procesamiento/Limpieza**: En la fase de data cleaning se examina la calidad de los datos y se manejan situaciones como datos con parámetros faltantes/nulos, datos duplicados, etc.
- **Transformación/Reducción**: Convertir datos pre-procesados en utilizables, identificando características importantes según los objetivos del proceso.

- **Minería de Datos**: Búsqueda de patrones de interés mediante técnicas como clasificación, regresión, clustering, correlaciones, etc.

- **Interpretación/Evaluación**: Fase final, de consolidación del conocimiento encontrado en los datos. Se preparan los resultados para documentación y toma de decisiones. Los datos se han transformado en visualizaciones para facilitar la evaluación del resultado depurado.

1.1.2 CRISP-DM (Cross-Industry Standard Process for Data Mining)

Publicada en 1999 con el objetivo de estandarizar los procesos de minería de datos en diversos sectores, se ha consolidado como la metodología más utilizada en proyectos de minería de datos, análisis y ciencia de datos:

- **Comprensión Empresarial**: se centra en entender los objetivos del proyecto, para luego ser evaluados y descubrir si los datos son aptos para cumplir con los objetivos y producir un plan de proyecto.

- **Comprensión de Datos**: Recopilación, descripción y exploración de los datos iniciales.

- **Preparación de Datos**: 5 tareas: selección de datos, limpieza de datos, construcción de datos, integración de datos y ajuste del formato.

- **Modelado**: Construcción y evaluación de varios modelos con diferentes técnicas algorítmicas. Se determina que algoritmos probar (por ejemplo, regresión, red neuronal); se realiza un diseño de experimentos; construir el modelo y evaluar el modelo.

- **Evaluación**: Se evalúa y revisa la creación de modelos respecto a los objetivos comerciales. Para ello se evaluan los resultados, se revisan los procesos y se determinan los próximos pasos.

- **Implementación**: Despliegue, seguimiento, mantenimiento y revisión final de los resultados obtenidos.

CRISP-DM presenta un proceso iterativo estructurado, definido y documentado. Es una metodología empleada como referencia por otras metodologías.

1.1.3 SEMMA (Sample, Explore, Modify, Model, Assess)

Propuesta para manejo de grandes volúmenes de datos:

- **Muestreo**: Selección de una muestra representativa de datos del problema que está investigando. La forma correcta de obtener una muestra es la selección aleatoria.

- **Exploración**: Exploración de información útil, con la finalidad de sintetizar el problema y mejorar la eficiencia del modelo.

- **Modificación**: Manipulación de los datos con base en la investigación realizada para que los datos ingresados al modelo estén definidos y en un formato adecuado.

- **Modelado**: Modelado de datos, con el propósito de establecer una relación entre las variables explicativas y el objeto de estudio.

- **Evaluación**: Validación comparativa de los resultados, a través del análisis de los modelos, comparado con otros modelos estadísticos o una nueva muestra poblacional.

1.1.4 RAMSYS (Rapid collaborative data Mining System)

Desarrollada por Steve Moyle en 2002. Es una metodología que apoya proyectos de minería de datos, por ello amplía el método CRISP-DM.
Define su metodología en tres roles:

- **Modeladores**: encargados de probar la viabilidad de las hipótesis y generar nuevos conocimientos.

- **Data Master**: responsable de mantener la versión actual de la base de datos, las transformaciones y la información sobre los datos, como metadatos e información sobre la calidad de los datos.

- **Comité de Dirección**: responsable de establecer los desafíos del proyecto, definir criterios, recibir y seleccionar las presentaciones.

1.1.5 TDSP (Team Data Science Process)

Metodología de Microsoft de 2017. Es de cierta forma una combinación de Scrum y CRISP-DM. El ciclo de vida de TDSP se compone de cinco etapas principales:

- **Comprensión empresarial**: Se definen los objetivos y se identifican las fuentes de datos.
- **Adquisición y comprensión de datos**: Se incorporan los datos y se determina si se puede responder a la pregunta planteada (combina efectivamente la Comprensión de los Datos y la Limpieza de los Datos de CRISP-DM).
- **Modelado**: Ingeniería de características (feature engineering) y entrenamiento de modelos (model training). Combina Modelado y Evaluación de CRISP-DM).
- **Implementación**: Implementar en un entorno de producción.
- **Aceptación del cliente**: Validación por parte del cliente de si el sistema satisface las necesidades del negocio (una fase no cubierta explícitamente por CRISP-DM).

TDSP aborda la debilidad de CRISP-DM en cuanto a la falta de definición del equipo, definiendo seis roles:

- Arquitecto de soluciones
- Project Manager
- Ingeniero de datos
- Científico de datos
- Desarrollador de aplicaciones
- Líder de proyecto (Project lead)

1.1.6 Conclusiones del tema

A lo largo de esta conferencia, hemos explorado diversas metodologías utilizadas en la Ciencia de Datos. Aunque cada una tiene sus particularidades, todas comparten elementos comunes:

- **Comprensión del Negocio**: Definir el problema empresarial y se identifican los objetivos del análisis. El equipo de ciencia de datos debe trabajar en estrecha colaboración con los clientes para entender el problema y definir los objetivos.

- **Comprensión de los Datos**: Identificar y recopilar los datos requerido para el análisis. Exploración de los datos para entender su estructura, calidad y completitud.

- **Preparación de los Datos**: Limpiar, transformar y preparar los datos para garantizar que estén en el formato y calidad adecuados para el análisis.

- **Modelado de Datos**: Seleccionar técnicas de modelado adecuadas para analizar los datos e implementar modelos predictivos. Esta etapa también involucra la selección de algoritmos, ajuste de parámetros y validación del modelo.

- **Evaluación**: Evaluar el rendimiento del modelo y su capacidad para resolver el problema empresarial. Utilizando métricas de evaluación adecuadas y realizando mejoras al modelo si es necesario.

- **Despliegue**: Desplegar el modelo en un entorno de producción, integrándolo en los procesos de la negocio y asegurando su correcto funcionamiento.

- **Monitoreo y Mantenimiento**: Supervisar el rendimiento del modelo en producción y realizar ajustes para mantener su efectividad.

En resumen, una metodología de Ciencia de Datos es un enfoque estructurado que combina comprensión del negocio, manejo de datos, modelado y evaluación para transformar datos en soluciones efectivas. Sin embargo, es crucial recordar que:

- **No existe una metodología universal**: Cada proyecto tiene características únicas que pueden requerir adaptaciones o combinaciones de enfoques.

- **La flexibilidad es clave**: Las metodologías no deben aplicarse de manera rígida, sino como guías que permitan ajustarse a las necesidades específicas del problema.

- **La colaboración es esencial**: La comunicación entre científicos de datos, ingenieros y stakeholders es fundamental para alinear objetivos y garantizar resultados útiles.

- **El ciclo nunca termina**: La Ciencia de Datos es un proceso iterativo, donde el monitoreo y la mejora continua son parte integral del éxito.

Para concluir vamos a definir en que consiste el trabajo de un ingeniero de datos.

En el libro *Fundamentals of Data Engineering* de *Joe Reis and Matt Housley* se define el término como:

La Ingeniería de Datos consiste en el desarrollo, interpretación y mantenimiento de sistemas y procesos que toman datos crudos y producen información consistente de alta calidad que soporta downstream casos de uso como analíticas y machine learning.

Un ingeniero de datos maneja el ciclo de vida de la ingeniería de datos que abarca el proceso que comienza en obtener los datos de las fuentes y termina sirviéndolos, después de procesados para los casos de uso.

En este curso estaremos viendo la ingeniería de datos como proceso integral que abarca la recolección, almacenamiento, procesamiento, y disponibilidad de datos.

1.2 Breve historia de la evolución de la ingeniería de datos

Los comienzos: 1980 a 2000 de Data Warehouse a la Web.
El nacimiento de la ingeniería de datos tiene sus raíces en data warehousing[1].
Que data desde la década de 1970 y toma forma en los 80s cuando se crea el término data warehouse. Luego surge el Structured Query Language (SQL). Mientras crecían los nacientes sistemas de datos, los negocios necesitaban dedicar herramientas a reportar y a business intelligence (BI). Surgiendo roles como BI ingeniero[2], desarrolladores ETL[3] e ingenieros de data warehouse. Precursores de lo que se considera actualmente los ingenieros de datos. También se popularizó el internet a mediados de los 90s, surgiendo una ola de compañías centradas en la web.

[1]Centralized repository that aggregates data from various sources to support data analysis, data mining and artificial intelligence

[2]A business intelligence engineer designs, implements, and maintains systems used to collect and analyze business intelligence data.

[3]Responsible for designing, building, managing, and maintaining ETL (Extract, Transform, Load) processes.

Los inicios de los 2000: El nacimiento de la ingeniería de datos contemporánea.
Las grandes compañías sobrevivientes como Yahoo, Google y Amazon crecerían hasta convertirse en gigantes tecnológicos. La necesidad de sistemas escalables, con alta disponibilidad, confiables y cost-effective llevó a innovaciones en sistemas distribuidos y almacenamiento, marcando el inicio de la era del 'big data'. La combinación del surgimiento de nuevos algoritmos y metodologías como: 'Google File System', 'MapReduce', Apache Hadoop, Amazon Elastic Compute Cloud y su apertura como servicio al público a través de Amazon Web Services (AWS) creó una nueva era en el manejo y tratamiento de datos. Nacía la era de ingeniero de big data.
Mientras AWS se volvió muy rentable otras compañías lanzaros sus propios ecosistemas en la nube: Google Cloud, Microsoft Azure, DigitalOcean. La nube es discutiblemente una de las innovaciones más significativas del siglo 21; iniciando una revolución en la forma en que el software y las aplicaciones de datos se desarrollan y despliegan

Finales de los 2000 y la década de 2010: The Big Data Engineering Era
Surgen herramientas open source que democratizan el acceso a tecnologías de big data; que ya no estarían limitadas solo a las grandes compañías.
También comienza la transformación de procesamiento en batch a streaming a partir de eventos.
Ocurre una explosión de herramientas de manejo de datos. Y los ingenieros de bog data debían ser proficientes en desarrollo de software y configuración de infraestructuras de bajo nivel.
Big data engineers se centraban en manejar sistemas de datos de gran escala, pero la complejidad y el costo de mantener estas nuevas herramientas impulsaron a optar por simplificaciones.
El término "big data" perdió su brillo mientras se volvían más accesibles las herramientas para procesarlos; los ingenieros de big data engineers pasan a ser simplemente ingenieros de datos.

2020s: Ingeniería por el ciclo de vida de los datos.
El rol de los ingenieros de datos se encuentra en rápida evolución. La tendencia está siendo centrarse en herramientas descentralizadas, modulares y abstractas. Las tendencias populares a principios de la década de 2020 incluyen el modern data stack (MDS), que representa una colección de productos "listos para usar", tanto de código abierto como de terceros, ensamblados para facilitar el trabajo de los analistas. Al mismo tiempo, las fuentes de datos y los formatos de datos están creciendo tanto en variedad como en tamaño. La ingeniería de datos es, cada vez más, una disciplina de interconexión, que conecta varias tecnologías como si fueran piezas de LEGO, para servir a los objetivos empresariales finales.

1.3 Relación entre la Ingeniería de Datos y la Ciencia de Datos

La ingeniería de datos se sitúa upstream de la ciencia de datos, lo que significa que los ingenieros de datos proporcionan las entradas utilizadas por los científicos de datos.
Para muchos lo más interesante de la ciencia de datos es construir y optimizar modelos de Machine Learning; la realidad es que se estima que entre el 70% y el 80% del tiempo se dedica a recopilar, limpiar y procesar datos.
Además; usualmente los científicos de datos no están entrenados para diseñar sistemas de datos de grado de producción, y terminan haciendo este trabajo improvisadamente porque carecen del soporte y los recursos de un ingeniero de datos.
En un mundo ideal, los científicos de datos deberían dedicar más del 90% de su tiempo al análisis, la experimentación y el ML. Esto se logra cuando los ingenieros de datos se centran en construir una base sólida para que los científicos de datos tengan éxito.

1.4 Habilidades del Ingeniero de Datos

El conjunto de habilidades de un ingeniero de datos debe abarcar las ideas subyacentes de la ingeniería de datos: seguridad, gestión de datos, DataOps, arquitectura de datos e ingeniería de software. Se requiere un entendimeinto de como evaluar herramientas de datos y como estas encajan en el ciclo de vida de la ingeniería de datos.

Un ingeniero de datos maneja una gran cantidad de piezas móviles complejas y debe optimizar constantemente a lo largo de los ejes de costo, agilidad, escalabilidad, simplicidad, reutilización e interoperabilidad. También se espera que el ingeniero de datos cree arquitecturas de datos ágiles que evolucionen a medida que surjan nuevas tendencias.

Por definición, un ingeniero de datos debe comprender tanto los datos como la tecnología. Con respecto a los datos, esto implica conocer varias de las mejores prácticas en torno a la gestión de datos. En el extremo tecnológico, un ingeniero de datos debe estar al tanto de varias opciones de herramientas, su interrelación y sus trade-offs.