Instituto Tecnológico de Costa Rica

Escuela de Computación

Bases de Datos II

Grupo 20

Investigación: Data Science

Profesor (a):
Alberto Shum Chan

Estudiante (s):

José Adrián Amador Ávila - 2016101574
Pablo Jesús Mora Barrantes - 2019205110
Jose Andrés Vargas Serrano - 2019211290

Alajuela, I Semestre 2023

Descripción General

¿Qué es Data Science?

Data science o Ciencia de Datos, como lo describe Amazon, es el estudio de datos con el fin de extraer información significativa, además, es un campo multidisciplinario que combina áreas como la Matemática, Estadística, la Inteligencia Artificial y la Ingeniería en Computación para analizar grandes cantidades de datos. Este análisis permite que los científicos de datos planteen y respondan preguntas como: "¿Qué pasó?", "¿Por qué pasó?", "¿Qué pasará?" y "¿Qué se puede hacer con los resultados?". IBM también agrega a la lista de áreas el Machine Learning.

Importancia del Data Science

Es importante porque combina herramientas, métodos y tecnología para generar significado a partir de los datos.

- Las organizaciones modernas están inundadas en datos (proliferación de dispositivos que generan y recolectan datos)
- Sistemas en línea y portales de pago capturan más datos en los campos del comercio electrónico, medicina, finanzas y cualquier otro aspecto de la vida humana.
- Disposición de grandes cantidades de datos no estructurados (texto, audio, vídeo e imágenes)

Descripción de la funcionalidad que abarque los principales usos que se le puede dar

Ciclo de vida del Data Science

- **Ingesta de datos (data ingestion):** se comienza con la recopilación de datos, tanto estructurados como no estructurados, obtenidos de todas las fuentes relevantes con diferentes métodos (de forma manual, web scraping *(bots)*, transmisión de datos en tiempo real).
- **Almacenamiento y procesamiento de los datos:** las empresas deben considerar diferentes tipos de almacenamiento dado la variedad de formatos en los datos obtenidos. Los equipos de administración de datos ayudan a establecer estándares en torno al almacenamiento y la estructura de datos, facilitando así el flujo de trabajo en torno a modelos de análisis, machine learning y deep learning (aprendizaje profundo). En esta etapa se incluye:
    - Limpieza de los datos
    - La deduplicación

    - La transformación
    - Y combinación de datos mediante procesos de ETL (extracción, transformación y carga)
- **Análisis de datos:** los científicos de datos realizan un análisis exploratorio de datos para examinar sesgos, patrones, rangos y distribuciones de valores dentro de los datos. Esto impulsa la generación de hipótesis para las pruebas a/b. Permite determinar la relevancia de los datos para su uso dentro del análisis predictivo, machine learning y/o deep learning. Si el modelo es preciso, las organizaciones pueden depender de estos conocimientos para la toma de decisiones comerciales, permitiendo así una mayor escalabilidad.
- **Comunicar:** el análisis y sus resultados se presentan como informes y otras visualizaciones (gráficos) de datos que facilitan la comprensión para los analistas comerciales y otros tomadores de decisiones. Lenguajes como R y Python incluyen componentes para generar visualizaciones, también existen herramientas dedicadas para este fin.

Algunas responsabilidades del Científico de Datos

IBM enlista algunas responsabilidades que el Científico de Datos posee:

- Conocer lo suficiente acerca del negocio para realizar preguntas pertinentes e identificar puntos débiles del negocio
- Aplicar estadística y ciencias de la computación (y otros conocimientos pertinentes) al análisis de datos
- Usar una gran variedad de herramientas y técnicas para preparar y extraer datos (estructurados como no estructurados)
- Extraer conocimientos del Big Data utilizando análisis predictivo e inteligencia artificial, incluido modelos de machine learning, procesamiento de lenguaje natural y deep learning
- Escribir programas que automatizan el procesamiento de datos y cálculos.
- Expresar el significado de los resultados del análisis de datos a los tomadores de decisiones y otras entidades de la manera más fácil para ellos entender
- Explicar cómo los resultados pueden ser usados para resolver un problema en particular del negocio
- Colaborar con otros miembros del equipo de ciencia de datos: analistas de datos y negocios, arquitectos de TI, ingenieros de datos, desarrolladores de aplicaciones.

Principales usos

Algunos de sus principales usos, según IBM, se enlistan a continuación :

- Un banco internacional ofrece servicios de préstamos más rápidos con una aplicación móvil que utiliza modelos de riesgo crediticio basados en aprendizaje automático y una arquitectura híbrida de computación en la nube que es poderosa y segura.
- Una empresa de tecnología de medios digitales creó una plataforma de análisis de audiencia que permite a sus clientes ver qué atrae a las audiencias de televisión a medida que se les ofrece una gama cada vez mayor de canales digitales. La solución emplea análisis profundos y machine learning para recopilar información en tiempo real sobre el comportamiento del espectador.
- Un departamento de policía urbana creó herramientas de análisis de incidentes estadísticos para ayudar a los oficiales a comprender cuándo y dónde desplegar recursos para prevenir el crimen. La solución basada en datos crea informes y paneles para aumentar la conciencia situacional de los oficiales de campo.

Ventajas y desventajas

Ventajas

Algunas ventajas brindadas por Amazon son las siguientes:

- **Descubrir patrones desconocidos de transformación:** Permite a las empresas descubrir nuevos patrones y relaciones con el potencial de transformar la organización. Puede revelar cambios de bajo coste en la administración de recursos para obtener el máximo impacto en los márgenes de beneficio.
- **Innovar con nuevos productos y soluciones:** Puede revelar lagunas y problemas que de otro modo pasarían desapercibidos. Mejor información sobre las decisiones de compra, los comentarios de los clientes y los procesos empresariales puede impulsar la innovación en las operaciones internas y las soluciones externas.
- **Optimización en tiempo real:** Para las empresas, en especial las grandes, es un gran reto responder en tiempo real a las condiciones cambiantes. Esto puede causar importantes pérdidas o interrupciones en la actividad empresarial. La ciencia de datos puede ayudar a las empresas a predecir los cambios y reaccionar de forma óptima ante las distintas circunstancias.

Desventajas

- Se necesita un amplio conocimiento de muchas áreas (nivel de expertis muy alto), por lo tanto poderlo dominar es muy poco probable
- Los datos arbitrarios pueden producir resultados inesperados
- El problema de la privacidad de los datos (datos de los clientes por ejemplo)
- Es un proceso costoso: herramientas complejas, necesidad de entrenamiento previo para el equipo de trabajo, necesidad de los roles para realizar todo el proceso.

Referencias:

https://www.ibm.com/topics/data-science

https://aws.amazon.com/es/what-is/data-science/

https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

https://www.oracle.com/what-is-data-science/

Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64–73. https://doi.org/10.1145/2500499

https://www.imperva.com/learn/application-security/web-scraping-attack/

https://www.devopsschool.com/blog/what-is-data-science-advantages-and-disadvantages-of-data-science/

https://data-flair.training/blogs/pros-and-cons-of-data-science/