Material Imprimible

Curso de Data Science

Módulo 1 - Introducción

Contenidos:

- Introducción a Data Science.
- Análisis de datos exploratorio.
- Machine Learning.
- Tipos de aprendizaje en Machine Learning.
- Etapas de la metodología Data Science.

¿Qué es Data Science?

La ciencia de datos o data science es una disciplina interdisciplinaria que combina técnicas estadísticas, de programación y de análisis de datos para extraer conocimientos y comprensión de conjuntos de datos complejos y grandes. Se utiliza para resolver problemas empresariales, científicos y sociales en diversos campos, desde la biología y la medicina hasta las finanzas y el marketing.

Consiste en emplear métodos analíticos basados en la estadística y la ciencia computacional para extraer conclusiones de un conjunto de datos

Es un campo multidisciplinar que tiene por objetivo principal identificar tendencias, motivos, conexiones y correlaciones en las grandes series de datos.

Engloba tópicos como el Machine Learning, Deep Learning, Chatbots, y muchas tecnologías disruptivas que han llegado al mercado en los últimos años.

Data Science y Data Analytics: Diferencias de enfoque

La ciencia de datos y el análisis de datos (o data analytics) son dos términos que a menudo se usan indistintamente, pero tienen algunas diferencias importantes.

Data analytics es una disciplina que se enfoca en el análisis de datos históricos para extraer información útil y tomar decisiones basadas en datos. El análisis de datos se utiliza comúnmente en áreas como el marketing, la publicidad y las finanzas para identificar patrones, tendencias y oportunidades en los datos existentes.

Data Science se enfoca principalmente en la exploración avanzada, estadística y predictiva, buscando estimar valores o resultados futuros con información pasada y presente.

¿Qué herramientas se usan típicamente?

Python / R -> Código abierto y gratuitas

Rapidminer

SPSS

SAS

SQL

¿Dónde se aplica Data Science?

Seguros, para predecir el riesgo de un asociado

Marketing, para predecir el comportamiento futuro del cliente y poder clusterizar los mismos

Finanzas, para estimar cuán riesgoso es otorgarle una línea de crédito a mis clientes

Recursos Humanos, para optimizar la performance de nuestros empleados y predecir cuándo presentarán la renuncia

Análisis de datos exploratorios

El análisis exploratorio de datos (EDA, por sus siglas en inglés de Exploratory Data Analysis) es una técnica utilizada en la ciencia de datos para examinar y comprender los datos en bruto de manera sistemática y sin prejuicios. Es el primer paso en el proceso de análisis de datos y se realiza antes de realizar cualquier análisis estadístico formal.

El objetivo del EDA es descubrir patrones, tendencias, relaciones y características clave en los datos, lo que ayuda a los científicos de datos a entender mejor la naturaleza de los datos y a identificar posibles problemas o errores en los mismos.

El EDA implica la visualización de datos utilizando técnicas gráficas, como histogramas, diagramas de dispersión y gráficos de caja, así como el cálculo de medidas estadísticas descriptivas, como la media, la mediana, la varianza y el rango intercuartílico.

El análisis exploratorio de datos se realiza de manera iterativa, lo que significa que los resultados del análisis se utilizan para ajustar y refinar la exploración de los datos hasta que se entienden los datos lo suficientemente bien como para avanzar en la modelización y el análisis estadístico más formal.

Machine Learning

El Machine Learning es una técnica fundamental utilizada en la ciencia de datos para modelar y analizar grandes conjuntos de datos, lo que permite a los científicos de datos descubrir patrones y relaciones en los datos y tomar decisiones informadas basadas en los conocimientos obtenidos.

Esta tecnología se enfoca en desarrollar algoritmos y modelos matemáticos que permiten a los sistemas aprender y mejorar su rendimiento en tareas específicas a partir de la experiencia adquirida a través de los datos.

Es una técnica de análisis de datos avanzada que aprende de manera autónoma sin ser programada explícitamente para una tarea en particular. El machine learning se basa en algoritmos que analizan grandes cantidades de datos y encuentran patrones y relaciones que permiten al sistema mejorar su desempeño y capacidad para tomar decisiones precisas, inclusive, llegando a predecir eventos futuros en el tiempo.

Ventajas:

- Permite automatizar modelos predictivos estadísticos que el ser humano no podría realizar
- Procesa grandes volúmenes de datos, evitando utilizar muestras pequeñas
- Permite realizar distintos tipos de aprendizaje
- Evoluciona constantemente con las nuevas tecnologías

Tipos de aprendizaje en Machine Learning

Hay tres tipos principales de aprendizaje automático: supervisado, no supervisado y por refuerzo.

En el aprendizaje supervisado, los modelos se entrenan utilizando conjuntos de datos etiquetados (es decir, datos que ya tienen una respuesta o etiqueta conocida), mientras que en el aprendizaje no supervisado, los modelos buscan patrones en los datos sin etiquetar. En el aprendizaje por refuerzo, los modelos aprenden a través de la retroalimentación continua de su entorno mientras intentan lograr una tarea específica.

Etapas de la metodología data science

- Definición del problema y entendimiento del negocio: En esta etapa, se identifica y define el problema o pregunta que se quiere responder utilizando los datos. Es importante definir claramente el problema para poder enfocar el análisis en la dirección correcta.

- Recopilación y entendimiento de los datos: En esta etapa, se recopilan los datos necesarios para abordar el problema definido. Los datos pueden provenir de diversas fuentes, como bases de datos internas, encuestas, registros públicos o fuentes en línea.

- Preparación de datos: En esta etapa, los datos recopilados se limpian, procesan y transforman en un formato adecuado para su análisis. Esto puede incluir la

eliminación de datos duplicados o incompletos, la normalización de los datos y la creación de variables adicionales. Se puede utilizar un análisis exploratorio para encontrar patrones, errores y desafíos en los datos preparados.

- Entrenamiento de modelo: En esta etapa, se construyen modelos estadísticos y algoritmos de aprendizaje automático para analizar los datos y responder a la pregunta o problema definido en la etapa 1.

- Validación de modelo: En esta etapa, se prueba y valida el modelo construido para asegurarse de que se ajuste bien a los datos y que tenga un buen rendimiento en la predicción o clasificación.

- Implementación y seguimiento: En esta etapa, se implementa el modelo y se realiza un seguimiento del desempeño para asegurarse de que se esté resolviendo el problema de manera efectiva. Si es necesario, se pueden realizar ajustes o mejoras al modelo en esta etapa.