Contenido

Máster en Big Data

URL: https://www.imf-formacion.com/masters-profesionales/master-big-data-business-intelligence

MÓDULO 1 - Fundamentos de tratamiento de datos para Data Science

1. Uso de máquinas virtuales y *shell* de comandos
    - Concepto de máquina virtual box
    - Creación y configuración de una máquina virtual
    - Carga de una máquina virtual
    - La shell de comandos de linucreación de scripts

Una máquina virtual (VM) es un entorno de software que simula el hardware de una computadora física, lo que permite ejecutar sistemas operativos y aplicaciones como si se tratara de una máquina real. VirtualBox es una herramienta popular que permite crear y gestionar máquinas virtuales, facilitando la ejecución de varios sistemas operativos en un solo equipo. La creación y configuración de una máquina virtual implica definir parámetros como la cantidad de memoria, el tipo de procesador y la capacidad del disco. Una vez configurada, la máquina se puede cargar para realizar tareas como si fuera una computadora independiente. La shell de comandos de Linux es un potente entorno para ejecutar comandos y crear scripts que permiten automatizar procesos, realizar configuraciones y gestionar el sistema de manera eficiente.

2. Fundamentos de programación en Python
    - El lenguaje Python y el entorno Jupyter notebook
    - Elementos básicos de Python
    - Estructuras de control
    - Estructuras de datos
    - Funciones
    - Excepciones
    - Importación de módulos
    - Gestión de ficheros

Python es un lenguaje de programación sencillo y versátil, ampliamente utilizado tanto en desarrollo de software como en ciencia de datos. El entorno Jupyter Notebook proporciona un espacio interactivo que facilita la programación en Python, permitiendo a los usuarios combinar código con texto y visualizaciones. Los elementos básicos de

Python incluyen variables, tipos de datos y operaciones matemáticas. Las estructuras de control, como bucles y condicionales, permiten dirigir el flujo del programa. Python también cuenta con estructuras de datos como listas, diccionarios y tuplas que son fundamentales para organizar la información. Además, las funciones permiten dividir el código en bloques reutilizables, mientras que las excepciones ayudan a gestionar errores. Python facilita la importación de módulos que amplían su funcionalidad y ofrece herramientas para la gestión de ficheros, lo cual es crucial para la manipulación de datos.

3. Fundamentos de bases de datos relacionales
    - El modelo relacional
    - SQLite Studio
    - El lenguaje SQL

El modelo relacional es la base de la mayoría de las bases de datos modernas, permitiendo organizar la información en tablas interconectadas por relaciones. Este enfoque facilita la consulta y manipulación de datos de manera estructurada. SQLite Studio es una herramienta que permite trabajar con bases de datos SQLite de forma sencilla e intuitiva. El lenguaje SQL (Structured Query Language) es el principal medio para interactuar con bases de datos relacionales, proporcionando comandos para realizar tareas como insertar, actualizar, eliminar y consultar datos. A través de SQL, se pueden gestionar grandes volúmenes de datos de forma eficiente y segura.

4. Fundamentos de tecnologías de internet
    - Formatos de almacenamiento de datos en internet
    - Manipulación de documentos CSV
    - Manipulación de documentos JSON
    - Manipulación de documentos XML

En internet, los datos pueden ser almacenados y compartidos en varios formatos, como CSV, JSON y XML. Los archivos CSV (Comma Separated Values) se utilizan para almacenar datos tabulares de forma sencilla y fácil de manipular. Los documentos JSON (JavaScript Object Notation) son comunes para el intercambio de información entre aplicaciones debido a su estructura clara y ligera. XML (eXtensible Markup Language), aunque menos utilizado que JSON hoy en día, sigue siendo relevante para ciertos tipos de aplicaciones y servicios web. La manipulación de estos formatos permite a los desarrolladores procesar y aprovechar los datos provenientes de diversas fuentes en la web.

5. Compartir datos, código y recursos en repositorios

- Repositorios digitales para compartir
- La tecnología GITHUB
- Uso de Google Drive como repositorio digital

Los repositorios digitales permiten compartir código, datos y otros recursos de manera eficiente entre desarrolladores y colaboradores. GitHub es una plataforma líder para almacenar y gestionar proyectos de software, facilitando la colaboración a través del control de versiones y la gestión de cambios. Los usuarios pueden trabajar juntos en proyectos, realizar revisiones y compartir sus avances de manera pública o privada. Google Drive también puede ser utilizado como un repositorio digital, especialmente útil para compartir documentos y archivos de forma sencilla y accesible para cualquier colaborador.

6. Fundamentos de tratamiento de datos con el *stack* científico de Python
    - Gestión de matrices y cálculo estadístico con NUMPY
    - Representación gráfica con MATPLOTLIB
    - Manipulación y análisis de datos con PANDAS

Python cuenta con un potente ecosistema de herramientas para el tratamiento y análisis de datos, conocido como el stack científico. NumPy es una librería fundamental para la gestión de matrices y la realización de cálculos estadísticos, ofreciendo funciones de gran rendimiento para la manipulación de datos numéricos. Matplotlib es la librería más popular para la representación gráfica en Python, permitiendo crear visualizaciones que facilitan el entendimiento de los datos. Pandas, por otro lado, es esencial para la manipulación y análisis de datos estructurados, facilitando la limpieza y transformación de grandes volúmenes de información. Estas herramientas en conjunto permiten a los usuarios realizar desde análisis simples hasta tareas complejas de ciencia de datos.

MÓDULO 2 - Business intelligence

1. Introducción a la inteligencia de negocio
    - Qué es la inteligencia de negocio
    - Importancia de los sistemas de inteligencia de negocio
    - Componentes de los sistemas de BI: arquitectura de inteligencia de negocio
    - Tipos de análisis que se pueden realizar
    - Inteligencia de negocio y analítica de negocio: BI y BA

- Inteligencia de negocio para Big Data

La inteligencia de negocio (BI, por sus siglas en inglés) se refiere al uso de tecnologías y estrategias para analizar datos y convertirlos en información accionable que apoye la toma de decisiones empresariales. Los sistemas de BI son fundamentales para mejorar la eficiencia operativa y la competitividad de las organizaciones. Los componentes de un sistema de BI incluyen una arquitectura que abarca fuentes de datos, almacenamiento, herramientas de análisis y visualización. Los análisis que se pueden realizar incluyen análisis descriptivo, predictivo y prescriptivo. Además, existe una diferencia importante entre inteligencia de negocio (BI) y analítica de negocio (BA), siendo BI más descriptivo y BA más predictivo. BI también desempeña un papel crucial en el manejo de Big Data, ayudando a procesar grandes volúmenes de datos y proporcionando insights estratégicos.

2. Almacenes de datos y bases de datos analíticas
    - Almacenes de datos
    - Herramientas de análisis de un almacén de datos: OLAP
    - Multidimensionalidad y el modelo multidimensional
    - Desnormalización
    - Lenguajes de consulta analísticos: MDX

Los almacenes de datos (Data Warehouses) son sistemas diseñados para almacenar grandes volúmenes de información de manera organizada y facilitar el análisis de los datos históricos. Las herramientas OLAP (Online Analytical Processing) permiten explorar los datos almacenados desde múltiples perspectivas y realizar consultas complejas para identificar patrones y tendencias. El modelo multidimensional es clave en el análisis OLAP, ya que permite organizar la información en dimensiones que representan diferentes aspectos del negocio. La desnormalización es una técnica utilizada en almacenes de datos para optimizar la velocidad de consulta. El lenguaje MDX (Multidimensional Expressions) es el lenguaje de consulta empleado para trabajar con datos multidimensionales.

3. Herramientas de extracción y carga
    - Qué es el proceso de extracción, transformación y carga (ETL)
    - Proceso ETL en un proyecto de inteligencia de negocio
    - Tipos de cargas
    - Gobierno del dato y orquestación

- Buenas prácticas
- Herramientas ETL: Pentaho Data Integration

El proceso de Extracción, Transformación y Carga (ETL) es esencial en los proyectos de inteligencia de negocio, ya que permite extraer datos de diferentes fuentes, transformarlos para adecuarlos a los estándares y cargarlos en un almacén de datos. Las cargas pueden ser completas o incrementales, dependiendo de las necesidades del proyecto. El gobierno del dato y la orquestación son aspectos clave para asegurar la calidad y disponibilidad de los datos durante todo el proceso ETL. Pentaho Data Integration es una de las herramientas ETL más conocidas, permitiendo llevar a cabo procesos de integración de datos de manera eficiente. Además, se recomienda seguir buenas prácticas para asegurar la integridad y calidad de los datos en cada etapa del proceso.

4. Aplicaciones de inteligencia de negocio
    - Aplicaciones de inteligencia de negocio
    - Herramientas de inteligencia de negocio
    - Herramienta de inteligencia de negocio: Pentaho Business Analytics

Las aplicaciones de inteligencia de negocio permiten transformar grandes volúmenes de datos en información comprensible para apoyar la toma de decisiones. Existen diversas herramientas de BI que permiten realizar análisis detallados y visualizaciones claras. Pentaho Business Analytics es una herramienta poderosa que ofrece un conjunto completo de funciones para análisis de datos, desde la integración hasta la visualización, facilitando la interpretación de los resultados y el soporte a las decisiones empresariales.

5. Análisis de datos masivos aplicados al negocio
    - Datos externos
    - DEMO

El análisis de datos masivos o Big Data permite a las empresas tomar decisiones informadas al evaluar tanto datos internos como externos. Los datos externos, como redes sociales, información del mercado y fuentes públicas, complementan los datos internos para proporcionar una visión más completa. Este análisis se puede demostrar mediante aplicaciones específicas que muestran cómo los datos pueden impactar directamente en las estrategias empresariales.

6. Inteligencia de cliente (CRM)
    - CRM

- Inteligencia de cliente
- Ingesta de datos CRM

Los sistemas CRM (Customer Relationship Management) son esenciales para gestionar las relaciones con los clientes y analizar sus comportamientos y preferencias. La inteligencia de cliente se refiere al análisis profundo de estos datos para entender mejor a los clientes y ofrecerles experiencias personalizadas. La ingesta de datos CRM implica la recopilación de información de múltiples fuentes para alimentar el sistema y generar estrategias de marketing más efectivas y focalizadas.

MÓDULO 3 - Aprendizaje Automático Aplicado (Machine Learning)

1. Introducción al aprendizaje automático
    - El proceso de la minería de datos
    - Tipos de aprendizaje automático
    - Introducción a SCIKIT-LEARN y THEANO
    - Uso básico de un modelo

El aprendizaje automático es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos que permiten a las computadoras aprender de los datos y hacer predicciones o tomar decisiones sin ser programadas explícitamente para cada tarea. El proceso de la minería de datos consiste en extraer patrones útiles y conocimiento de grandes volúmenes de datos, utilizando técnicas como el preprocesamiento, la selección de características y la modelización. Existen diferentes tipos de aprendizaje automático, entre ellos el aprendizaje supervisado, no supervisado y por refuerzo. Herramientas como SCIKIT-LEARN y THEANO son fundamentales para implementar algoritmos de aprendizaje automático. SCIKIT-LEARN es una librería de Python que facilita el uso de una amplia variedad de modelos, mientras que THEANO es una biblioteca para la manipulación eficiente de tensores, muy útil en el desarrollo de redes neuronales. El uso básico de un modelo implica entrenar el algoritmo con datos etiquetados, validar su rendimiento y utilizarlo para hacer predicciones sobre datos nuevos.

2. Modelos supervisados
    - Predicción de valores continuos con regresión lineal
    - Clasificación mediante regresión logística
    - Árboles de decisión
    - Otros modelos supervisados

Los modelos supervisados son aquellos que se entrenan utilizando un conjunto de datos etiquetados, es decir, donde se conoce el resultado deseado. La regresión lineal se utiliza para predecir valores continuos, como por ejemplo el precio de una vivienda en función de sus características. Por otro lado, la regresión logística es una técnica utilizada para problemas de clasificación binaria, como predecir si un correo electrónico es spam o no. Los árboles de decisión son modelos que permiten clasificar datos dividiéndolos en subconjuntos según sus características. Además de estos, existen otros modelos supervisados como los vecinos más cercanos (KNN) y las máquinas de vectores de soporte (SVM), que también son herramientas poderosas para problemas de clasificación y regresión

3. Modelos no supervisados
    - Análisis de componentes principales
    - Identificación de objetos similares con k-means
    - Organización de clústeres como árbol jerárquico
    - Localización de regiones a través de DBSCAN

Los modelos no supervisados se utilizan cuando no se dispone de datos etiquetados, y el objetivo es encontrar patrones ocultos o relaciones entre los datos. El análisis de componentes principales (PCA) es una técnica que permite reducir la dimensionalidad de un conjunto de datos, facilitando la visualización y el análisis. El algoritmo k-means se utiliza para agrupar objetos similares en diferentes clústeres, mientras que la organización jerárquica permite crear una estructura de clústeres en forma de árbol, mostrando relaciones jerárquicas entre ellos. DBSCAN es otro algoritmo que se utiliza para encontrar regiones densas en los datos y agruparlas, siendo particularmente útil para detectar formas arbitrarias y eliminar ruido.

4. Ingeniería de características y selección de modelos
    - Diferentes tipos de características y transformación
    - Selección de características
    - Selección de modelos

La ingeniería de características es el proceso de transformar los datos brutos en características que puedan ser utilizadas por los modelos de aprendizaje. Esto puede incluir técnicas como la normalización, la codificación categórica y la extracción de características relevantes. La selección de características implica elegir aquellas variables que sean más relevantes para mejorar la precisión y reducir la complejidad del modelo. Además, la selección de modelos es un proceso

importante para determinar qué modelo se ajusta mejor a los datos, comparando diferentes algoritmos y configuraciones para maximizar el rendimiento.

5. Modelos conexionistas
    - Perceptrones
    - Redes neuronales
    - Clasificación de dígitos escritos a mano

Los modelos conexionistas, también conocidos como redes neuronales, están inspirados en el funcionamiento del cerebro humano. El perceptrón es la unidad básica de las redes neuronales y funciona como un clasificador lineal. Las redes neuronales consisten en múltiples capas de perceptrones que permiten a los modelos aprender patrones complejos y no lineales. Estos modelos han sido ampliamente utilizados en aplicaciones como la clasificación de dígitos escritos a mano, donde una red neuronal es capaz de reconocer números con gran precisión. Las redes neuronales profundas o deep learning son una extensión de este concepto y se utilizan en una amplia variedad de tareas complejas, desde el reconocimiento de imágenes hasta el procesamiento del lenguaje natural.

6. Reglas de asociación y *market basket analysis*
    - Soporte, confianza y lift
    - Algoritmo apriori
    - Otros algoritmos asociativos

Las reglas de asociación son una técnica de minería de datos utilizada para descubrir relaciones interesantes entre elementos dentro de grandes conjuntos de datos. En el contexto de market basket analysis, se busca encontrar patrones de compra que indiquen qué productos suelen ser comprados juntos. Los conceptos de soporte, confianza y lift son fundamentales para evaluar la calidad de las reglas descubiertas. El algoritmo apriori es uno de los métodos más conocidos para generar estas reglas, pero existen otros algoritmos asociativos que también pueden ser utilizados dependiendo de la naturaleza de los datos y los objetivos del análisis.

MÓDULO 4 - Minería de Texto y Procesamiento del Lenguaje Natural (PLN)

1. Introducción histórica y tecnológica
    - Contexto histórico
    - Cadenas de procesamiento clásicas

- Preprocesamiento de textos
- Tokenización
- Segmentación de frases
- Análisis léxico o morfológico
- Análisis sintáctico
- Análisis semántico

El procesamiento del lenguaje natural (PLN) tiene una historia larga y fascinante. Todo comenzó con los primeros intentos de hacer que las computadoras comprendieran el lenguaje humano. Las cadenas de procesamiento clásicas consisten en una serie de pasos que ayudan a convertir el texto en una forma comprensible para las máquinas. El preprocesamiento de textos incluye tareas como la limpieza del texto y la normalización. La tokenización divide el texto en unidades más pequeñas, como palabras o frases, facilitando el análisis. La segmentación de frases permite dividir textos largos en oraciones más cortas. El análisis léxico y morfológico ayuda a entender la estructura y significado de las palabras. El análisis sintáctico se centra en la estructura gramatical, mientras que el análisis semántico se ocupa del significado del texto.

2. Herramientas PLN I: NLTK
    - Conceptos básicos de NLTK
    - Instalación
    - Funcionalidades de NLTK
    - Usos y operaciones de NLTK en español
    - Taggers: análisis y extracción de la información

NLTK (Natural Language Toolkit) es una de las bibliotecas más populares para el procesamiento del lenguaje natural en Python. Para empezar a usar NLTK, primero es necesario instalarlo, lo cual se puede hacer fácilmente mediante comandos de Python. NLTK ofrece una gran variedad de funcionalidades, como tokenización, análisis sintáctico y extracción de información. En español, NLTK se puede usar para realizar diferentes operaciones, como etiquetar palabras y realizar análisis morfológico. Los taggers son componentes clave para el análisis del texto, permitiendo etiquetar palabras con sus respectivas categorías gramaticales, lo que facilita la extracción de información relevante.

3. Herramientas PLN II: Inception y Gate
    - Inception

- Gate

Inception es una herramienta que permite la anotación de textos para tareas de PLN. Facilita la colaboración entre investigadores para mejorar la precisión del análisis. Gate, por otro lado, es una plataforma de código abierto que ofrece un conjunto de herramientas para la extracción de información y el procesamiento de textos. Ambas herramientas son esenciales para tareas complejas de PLN y se utilizan ampliamente en el ámbito académico y empresarial.

4. Text mining
    - Introducción al clustering de textos
    - Clustering de textos con k-means
    - Reconocimiento y síntesis de voz
    - Sentiment analysis
    - Topic modeling

El text mining o minería de textos es el proceso de extraer información útil de grandes volúmenes de texto. El clustering de textos es una técnica que agrupa documentos similares, y uno de los métodos más comunes es k-means. El reconocimiento y síntesis de voz permiten a las máquinas comprender y generar lenguaje hablado. El sentiment analysis se utiliza para determinar la polaridad de los textos, es decir, si son positivos, negativos o neutrales. El topic modeling es una técnica que permite descubrir temas ocultos dentro de un conjunto de documentos, lo cual es útil para organizar grandes volúmenes de información.

5. Otras aplicaciones y técnicas PLN
    - Otras aplicaciones de PLN
    - Ejemplo práctico: asistentes conversacionales

El procesamiento del lenguaje natural tiene muchas aplicaciones, como la traducción automática, la generación de texto y los sistemas de recomendación. Un ejemplo práctico de PLN son los asistentes conversacionales, que utilizan técnicas avanzadas para comprender las consultas del usuario y proporcionar respuestas relevantes. Estas tecnologías están cada vez más presentes en nuestras vidas diarias, mejorando la interacción entre humanos y máquinas.

MÓDULO 5 - Inteligencia de Negocio y Visualización

1. Introducción al *business intelligence*
    - Business intelligence

- Estructura de un sistema de business intelligence
- Datos, información y conocimiento
- Alfabetización de datos

El business intelligence (BI) se refiere al conjunto de procesos, tecnologías y herramientas que convierten los datos en información útil para apoyar la toma de decisiones empresariales. Un sistema de BI se estructura a partir de varias capas que incluyen la recopilación, integración, análisis y presentación de los datos. Los datos se transforman en información y posteriormente en conocimiento para guiar decisiones estratégicas. La alfabetización de datos es crucial en este contexto, ya que permite que los usuarios comprendan y trabajen con los datos de forma efectiva.

2. BI vs. *reporting* tradicional
    - Paradigma actual
    - Reporting tradicional vs. BI
    - Modern BI vs. Traditional BI
    - Encaje de BI con big data

El paradigma actual de BI ha evolucionado considerablemente respecto al reporting tradicional. Mientras que el reporting tradicional se centra en la presentación de información histórica, BI proporciona una visión más amplia, incluyendo análisis en tiempo real y capacidad predictiva. El BI moderno se diferencia del tradicional en su capacidad de interactuar con grandes volúmenes de datos y proporcionar insights más profundos. BI también encaja perfectamente con big data, ayudando a extraer valor de fuentes masivas de información.

3. Fundamentos tecnológicos para el tratamiento y análisis de datos
    - Componentes de entornos BI
    - Tipos de almacenamiento y sistemas de comunicación
    - Bases de datos relacionales
    - Principales herramientas
    - Introducción a Google Big Query

Los entornos de BI cuentan con diversos componentes tecnológicos, como sistemas de almacenamiento, bases de datos y herramientas de análisis. Existen diferentes tipos de almacenamiento, desde bases de datos relacionales hasta soluciones distribuidas para big data. Las bases de datos relacionales siguen

siendo fundamentales en muchos contextos para la gestión de datos estructurados. Herramientas como Google Big Query permiten analizar grandes volúmenes de datos de manera rápida y eficiente, facilitando la toma de decisiones en tiempo real.

4. Fundamentos de visualización de datos
    - Teoría de la visualización
    - Optimización
    - Propiedades de la visualización
    - Tipos de gráficos y su uso
    - Optimización de las técnicas de visualización

La visualización de datos es un aspecto esencial del BI, ya que permite transformar los datos en información visualmente accesible. La teoría de la visualización se basa en representar datos de manera que los usuarios puedan comprenderlos de forma rápida y efectiva. La optimización de las visualizaciones implica elegir el tipo de gráfico adecuado según el tipo de datos y los objetivos del análisis. Existen muchos tipos de gráficos, como barras, líneas y diagramas de dispersión, cada uno con su uso específico según el contexto de los datos. Optimizar estas técnicas de visualización mejora la comunicación y el entendimiento de la información.

5. Visualización avanzada de datos
    - Técnicas de visualización
    - Informes y cuadros de mando
    - Tipos de análisis
    - Informes por departamento
    - Qué hacer y qué no en una visualización

Las técnicas avanzadas de visualización ayudan a extraer información significativa de grandes volúmenes de datos. Los informes y cuadros de mando son herramientas clave en BI que permiten a los usuarios visualizar métricas y KPI de manera clara. Existen varios tipos de análisis, como descriptivo, predictivo y prescriptivo, y cada departamento dentro de una organización puede beneficiarse de informes específicos que se ajusten a sus necesidades. Para una visualización efectiva, es importante seguir buenas prácticas y evitar sobrecargar los gráficos con información innecesaria.

6. Herramientas de visualización

- Calificación de herramientas
- Tipos de herramientas de visualización
- Tableau: visión general
- Tableau: interfaz
- Tableau: ejercicio guiado
- Power BI: visión general
- Power BI: interfaz
- Power BI: ejercicio guiado

Existen diversas herramientas para la visualización de datos, cada una con características únicas que se ajustan a diferentes necesidades. Tableau es una herramienta muy popular que permite crear visualizaciones interactivas de forma intuitiva; su interfaz facilita el proceso de diseño de gráficos y cuadros de mando. En un ejercicio guiado, los usuarios pueden aprender a crear dashboards personalizados. Power BI es otra herramienta potente que ofrece capacidades similares, con integración nativa a otros servicios de Microsoft. La interfaz de Power BI permite construir visualizaciones dinámicas, y sus ejercicios guiados ayudan a los usuarios a sacar el máximo provecho de la plataforma.

MÓDULO 6 - Infraestructura Big Data

1. Procesamiento de datos con Hadoop
    - HADOOP. Conceptos básicos
    - ¿Qué es APACHE HADOOP?
    - Ecosistema HADOOP
    - ¿Quién usa HADOOP y por qué se usa?
    - Módulos de HADOOP
    - Instalación de HADOOP
    - Ejemplo de comprobación de la instalación
    - HDFS
    - MAPREDUCE

Hadoop es un marco de software de código abierto que permite el procesamiento distribuido de grandes conjuntos de datos. Apache Hadoop es la implementación más conocida, utilizada ampliamente para almacenar y procesar datos masivos. El

ecosistema Hadoop incluye varias herramientas, como HDFS (sistema de archivos distribuido) y MapReduce (modelo de programación para el procesamiento de datos). Muchas empresas usan Hadoop por su capacidad de manejar grandes volúmenes de datos de manera económica. Los módulos de Hadoop incluyen HDFS, MapReduce, YARN y Hadoop Common. La instalación de Hadoop requiere configurar estos componentes, y se puede verificar mediante ejemplos de comprobación como el uso de HDFS y la ejecución de trabajos MapReduce.

2. Herramientas del ecosistema Hadoop
    - PIG
    - HIVE
    - SQOOP
    - HBASE

El ecosistema Hadoop incluye diversas herramientas que complementan sus capacidades. Pig es un lenguaje de alto nivel para el procesamiento de datos, mientras que Hive proporciona una interfaz SQL para trabajar con datos almacenados en Hadoop. Sqoop se utiliza para transferir datos entre bases de datos relacionales y Hadoop. HBase, por otro lado, es una base de datos NoSQL que permite el acceso aleatorio a grandes volúmenes de datos, siendo ideal para necesidades de consulta rápida.

3. Procesamiento de datos con Spark
    - SPARK: historia y evolución
    - Componentes
    - SPARK SHELL
    - Descarga y configuración
    - Conceptos básicos
    - Ejemplos

Apache Spark es una herramienta poderosa para el procesamiento de datos a gran escala, y surgió como una evolución de Hadoop para proporcionar una alternativa más rápida y flexible. Spark tiene varios componentes importantes, como Spark Core, Spark SQL y MLlib para el aprendizaje automático. Spark Shell es una interfaz interactiva que permite experimentar con las funciones de Spark en tiempo real. La descarga y configuración de Spark son sencillas, y se pueden encontrar ejemplos prácticos para ayudar a los usuarios a aprender conceptos básicos rápidamente.

4. Arquitecturas de *streaming*

- Conceptos y casos de uso
- Patrones de arquitectura
- Componentes tecnológicos
- Ejemplos de arquitecturas de Streaming
- Algoritmos para procesamiento de Streams

Las arquitecturas de streaming permiten el procesamiento de datos en tiempo real, siendo cruciales en aplicaciones donde la inmediatez es fundamental. Los casos de uso de estas arquitecturas incluyen la monitorización de eventos y el análisis de redes sociales. Los patrones de arquitectura de streaming suelen incluir componentes como productores de datos, colas de mensajes y consumidores que procesan la información. Algunos algoritmos para el procesamiento de streams se centran en el análisis continuo y en la actualización de modelos predictivos conforme se reciben nuevos datos.

5. Componentes de arquitecturas en *streaming*
    - Conceptos de procesamiento de datos con SPARK Streaming
    - Ejemplo de procesamiento de un Stream
    - Arquitectura interna
    - Transformaciones y operaciones de salida
    - Integración, garantías y rendimiento

Spark Streaming es un componente de Apache Spark diseñado para el procesamiento de datos en tiempo real. Un ejemplo típico de procesamiento de un stream podría ser la ingesta y análisis de datos provenientes de sensores. La arquitectura interna de Spark Streaming se basa en el modelo de micro-batches, que permite realizar transformaciones y operaciones de salida de manera eficiente. Integrar Spark Streaming con otros sistemas asegura una alta disponibilidad y mejora el rendimiento, lo cual es clave para aplicaciones críticas

6. Plataformas y API en la nube
    - Servicios cloud
    - Microsoft Azure
    - Amazon Web Services (AWS)
    - Google Cloud Platform

Las plataformas en la nube ofrecen servicios que permiten a las empresas gestionar y analizar grandes volúmenes de datos sin necesidad de infraestructura física. Microsoft Azure, Amazon Web Services (AWS) y Google Cloud Platform (GCP) son tres de los proveedores de servicios en la nube más conocidos. Estos servicios cloud ofrecen herramientas para almacenamiento, procesamiento y análisis de datos, así como APIs que permiten la integración y automatización de tareas de análisis de datos de manera sencilla y escalable.

MÓDULO 7 - Almacenamiento e Integración de Datos

1. Bases de datos no convencionales
    - Los datos y el Big Data
    - Las bases de datos relacionales
    - Limitaciones de las bases de datos relacionales
    - Bases de datos NOSQL
    - Bases de datos orientadas hacia agregados
    - Modelos de distribución en las bases de datos NOSQL
    - El teorema CAP
    - Cuándo usar NOSQL o SQL

Los datos masivos o Big Data requieren soluciones de almacenamiento y procesamiento más flexibles que las bases de datos tradicionales. Las bases de datos relacionales son útiles para datos estructurados, pero presentan limitaciones cuando se manejan grandes volúmenes o datos no estructurados. En estos casos, las bases de datos NoSQL son una opción eficaz. Estas bases de datos están orientadas hacia agregados, lo que permite almacenar datos complejos de forma eficiente. Además, los modelos de distribución en NoSQL permiten escalar fácilmente. El teorema CAP explica las características fundamentales de los sistemas distribuidos: consistencia, disponibilidad y tolerancia a particiones. Elegir entre NoSQL y SQL depende del tipo de datos y los requisitos de escalabilidad.

2. Modelos de base de datos basados en documentos
    - Bases de datos orientadas a documentos
    - Instalación de MONGODB
    - Instalación de ROBO 3T
    - Conceptos básicos de MONGODB
    - La shell de comandos de MONGODB

- Operaciones CRUD
- Consultas

Las bases de datos orientadas a documentos permiten almacenar datos en formatos como JSON, lo cual resulta útil para aplicaciones con datos no estructurados. MongoDB es uno de los sistemas más populares en este ámbito. La instalación de MongoDB y herramientas como Robo 3T facilitan la gestión de datos. MongoDB tiene una shell de comandos específica que permite realizar operaciones CRUD (crear, leer, actualizar, eliminar) y ejecutar consultas complejas de forma sencilla, ofreciendo gran flexibilidad en el manejo de la información.

3. Modelos de base de datos orientados a columnas
    - Instalación de CASSANDRA
    - Características generales
    - Bases de datos en CASSANDRA
    - El lenguaje CQL

Las bases de datos orientadas a columnas están diseñadas para manejar grandes cantidades de datos de forma eficiente, especialmente en análisis y consultas de lectura intensiva. Cassandra es una de las bases de datos orientadas a columnas más conocidas, y su instalación es relativamente simple. Cassandra utiliza el lenguaje CQL (Cassandra Query Language) para gestionar las bases de datos, lo cual permite realizar consultas de forma similar a SQL, pero optimizado para las características únicas de este sistema distribuido.

4. Modelos de base de datos orientados a grafos
    - Bases de datos orientadas a grafos
    - Características principales
    - Introducción a NEO4J y al lenguaje CYPHER

Las bases de datos orientadas a grafos son ideales para almacenar y consultar datos que presentan relaciones complejas, como redes sociales o recomendaciones. Entre sus características principales se encuentra la capacidad de modelar entidades y sus relaciones de manera directa. Neo4j es una base de datos orientada a grafos muy popular y utiliza el lenguaje Cypher, que facilita la creación y consulta de relaciones entre nodos. Estas bases de datos son ideales para casos donde las relaciones entre datos son fundamentales.

5. Modelos de bases de datos clave-valor
    - Bases de datos clave-valor

- REDIS
- Instalación de REDIS
- Elementos básicos
- Estructuras de datos
- Replicación

Las bases de datos clave-valor son una de las formas más simples de bases de datos NoSQL, donde cada dato se almacena como un par clave-valor. Redis es un ejemplo popular de base de datos clave-valor, conocido por su rapidez y eficiencia en la gestión de datos en memoria. La instalación de Redis es sencilla y ofrece varios elementos básicos y estructuras de datos, como listas y conjuntos. Redis también soporta replicación, lo cual permite tener copias de seguridad y asegurar la disponibilidad del sistema.

6. Adquisición de datos
    - APACHE KAFKA
    - APACHE FLUME

Flume son herramientas clave para la adquisición de datos en sistemas distribuidos. Kafka se utiliza para la transmisión de datos en tiempo real, permitiendo manejar flujos masivos de datos de manera eficiente y fiable. Flume, por otro lado, está diseñado para recopilar y transportar grandes cantidades de datos de múltiples fuentes hacia un almacén central, como Hadoop. Estas herramientas son fundamentales para garantizar que los datos se recojan y distribuyan de manera confiable en sistemas que manejan grandes volúmenes de información.

MÓDULO 8 - Valor y Contexto de la Analítica Big Data

1. El *business case* de big data
    - Contexto económico
    - El valor del dato
    - Tipos de datos desde la perspectiva de negocio
    - Tipos de analítica
    - Puntos clave en la transformación
    - La organización analítica
    - El Big Data y la empresa

El contexto económico actual destaca la importancia del Big Data para el crecimiento empresarial. El valor del dato radica en su capacidad de generar información relevante y estratégica. Existen distintos tipos de datos, desde los estructurados hasta los no estructurados, que cada empresa puede utilizar desde una perspectiva de negocio. Los tipos de analítica se dividen en descriptiva, predictiva y prescriptiva, cada una aportando distintos niveles de conocimiento. La transformación hacia una organización analítica requiere puntos clave como la cultura de datos y una infraestructura adecuada. El Big Data ayuda a las empresas a ser más competitivas y a responder mejor a las demandas del mercado.

2. Proyectos de *big data*
    - Factores clave
    - Perfiles
    - Áreas de aplicación
    - Fases de un proyecto Big Data
    - Entornos en un proyecto Big Data
    - Proyectos de datos

Un proyecto de Big Data requiere considerar varios factores clave, como la infraestructura tecnológica y el talento especializado. Los perfiles necesarios incluyen científicos de datos, ingenieros de datos y analistas de negocio. Las áreas de aplicación del Big Data son diversas, abarcando desde marketing hasta producción. Un proyecto de Big Data pasa por diferentes fases, como la definición del problema, la recopilación de datos, el análisis y la implementación de soluciones. Los entornos en un proyecto de Big Data varían según las necesidades, e incluyen desde sistemas locales hasta soluciones en la nube para garantizar la escalabilidad y disponibilidad de los datos.

3. Aplicaciones analíticas por sectores
    - Google y el buscador
    - Amazon
    - Walmart
    - Recursos humanos
    - Financiero
    - Netflix

Las aplicaciones analíticas se extienden por varios sectores. Google utiliza técnicas avanzadas de Big Data en su buscador para optimizar los resultados de búsqueda. Amazon aplica análisis predictivos para personalizar la experiencia del usuario y gestionar su inventario. Walmart emplea análisis en tiempo real para mejorar la cadena de suministro. En recursos humanos, el análisis de datos se usa para identificar candidatos adecuados y mejorar la retención de empleados. El sector financiero utiliza Big Data para evaluar riesgos y detectar fraudes. Netflix emplea modelos analíticos para personalizar recomendaciones y mejorar la experiencia del usuario.

4. Tecnologías emergentes en analítica
    - Consideraciones del ecosistema Big Data
    - Aspectos clave en la definición de una iniciativa Big Data
    - Consideraciones técnicas en el planteamiento de un proyecto Big Data
    - Casos de éxito
    - Casos de fracaso

El ecosistema Big Data está en constante evolución, y es importante considerar los aspectos clave al definir una iniciativa de Big Data, como la infraestructura, los objetivos de negocio y la seguridad. Las consideraciones técnicas incluyen la selección de herramientas y tecnologías adecuadas según el tipo de proyecto. Existen casos de éxito, como empresas que han logrado mejorar sus procesos y obtener ventajas competitivas, así como casos de fracaso donde una mala planificación llevó a resultados negativos. Aprender de ambos tipos de experiencias es fundamental para implementar proyectos de Big Data con éxito.

5. Gestión de equipos y métodos ágiles
    - El agilismo. ¿Qué son las metodologías ágiles?
    - La metodología agile/SCRUM. Introducción a SCRUM
    - Roles y responsabilidades
    - Artefactos de SCRUM
    - Fases de un proyecto SCRUM
    - Métricas y medidas SCRUM
    - Estimación de proyectos agile

oyectos de forma flexible y rápida. Entre ellas destaca SCRUM, una metodología ágil que estructura los proyectos en ciclos iterativos llamados sprints. Los roles principales en SCRUM incluyen el Product Owner, el Scrum Master y el equipo de desarrollo. Los artefactos de SCRUM, como el Product Backlog y el Sprint Backlog, ayudan a organizar y priorizar las tareas. Un proyecto SCRUM pasa por varias fases, desde la planificación hasta la revisión del sprint. Las métricas SCRUM permiten evaluar el progreso y la eficiencia del equipo, mientras que la estimación ayuda a definir el alcance del proyecto de forma realista.

6. Aspectos regulatorios del tratamiento de datos
    - Contexto jurídico en materia de protección de datos
    - Derecho a la protección de datos
    - Principios
    - Figuras específicas
    - Transferencias internacionales de datos
    - Prácticas específicas
    - Evaluación de impacto

El tratamiento de datos está regulado por un marco jurídico que garantiza la protección de la información personal. El derecho a la protección de datos es fundamental y está basado en principios como la transparencia, la finalidad y la minimización de datos. Existen figuras específicas como el Responsable del Tratamiento y el Delegado de Protección de Datos que tienen roles claros dentro de la organización. Las transferencias internacionales de datos deben cumplir con normativas estrictas para asegurar la privacidad. Las prácticas específicas y la evaluación de impacto son herramientas que permiten identificar riesgos y garantizar el cumplimiento de la normativa vigente.

MÓDULO 9 - Aplicaciones Analíticas. Casos prácticos

1. Caso de Health Tech
    - Fundamentos del análisis en Health Tech
    - Contexto
    - Presentación del caso
    - Desarrollo del caso
2. Caso de estudio de analítica en redes sociales
    - Fundamentos de análisis de redes sociales

- Contexto
- Presentación del caso
- Desarrollo del caso

3. Caso de estudio en *internet of things* (IoT)
    - Concepto de Internet of Things
    - Sistemas integrados y su relación con los dispositivos IoT
    - Hardware y software
    - Redes y protocolos
4. Caso de estudio en analítica financiera (el *rating* de empresas) - *Fintech*
    - Presentación del caso
    - Fundamentos del rating de empresas
    - Contexto
    - Desarrollo del caso
5. Caso de estudio en el sector *retail*
    - Fundamentos del análisis en el sector *retail*
    - Contexto
    - Presentación del caso
    - Desarrollo del caso
6. Caso de estudio de *Business Analytics*
    - Fundamentos de *Business Analytics*
    - Contexto
    - Presentación del caso
    - Desarrollo del caso

MÓDULO 10 - Trabajo Fin de Máster en Big Data

Máster en Inteligencia Artificial y Deep Learning

URL: https://www.imf-formacion.com/masters-profesionales/master-deep-learning-inteligencia-artificial

MÓDULO 1 - Las herramientas del científico de datos

1. Fundamentos de Python
    - El lenguaje de programación Python
    - Entorno de programación: Jupyter Notebook
    - Sintaxis básica de Python
    - Herramientas para control de flujo
    - Estructuras de datos
    - Funciones

Python es un lenguaje de programación versátil y fácil de aprender, utilizado ampliamente en ciencia de datos, desarrollo web y automatización. Jupyter Notebook es un entorno interactivo muy popular para programar en Python, que permite combinar código, visualizaciones y texto explicativo. La sintaxis básica de Python es sencilla y clara, lo que facilita su aprendizaje. Las herramientas para el control de flujo, como condicionales y bucles, permiten dirigir el comportamiento del programa. Python también cuenta con diversas estructuras de datos, como listas, tuplas y diccionarios, que permiten almacenar y organizar información de manera eficiente. Las funciones en Python son fundamentales para modularizar el código y hacerlo más reutilizable.

2. Librerías para ciencia de datos: Numpy, Pandas, etc.
    - Procesamiento de archivos
    - Numpy
    - Pandas

Numpy, Pandas, etc.

Python cuenta con una gran cantidad de librerías útiles para la ciencia de datos. La manipulación de archivos, como CSV y Excel, es sencilla gracias a las funcionalidades nativas y a librerías adicionales. Numpy es esencial para trabajar con arrays y realizar operaciones matemáticas de manera eficiente. Pandas permite la manipulación y análisis de datos a través de estructuras como DataFrames, facilitando la limpieza y transformación de grandes volúmenes de información.

3. Procesamiento de datos y visualización con Python
    - Matplotilib
    - GGPLOT
    - Seaborn
    - BOKEH

La visualización de datos es clave para interpretar los resultados del análisis. Matplotlib es una librería básica para crear gráficos simples y personalizables en Python. GGPLOT, inspirado en la gramática de gráficos de R, y Seaborn, que se basa en Matplotlib, facilitan la creación de visualizaciones más sofisticadas. Bokeh permite la creación de gráficos interactivos, lo cual es útil para análisis exploratorios donde la interacción del usuario es fundamental.

4. Fundamentos de R
    - El lenguaje de programación R
    - Entorno de programación: Rstudio
    - Sintaxis básica de R
    - Estructuras de datos
    - Herramientas de control de flujo
    - Funciones

R es un lenguaje de programación diseñado específicamente para el análisis estadístico y la visualización de datos. RStudio es el entorno de desarrollo más utilizado para programar en R, proporcionando herramientas avanzadas para trabajar cómodamente. La sintaxis de R es diferente a la de otros lenguajes, pero está optimizada para trabajar con datos. Las estructuras de datos principales incluyen vectores, matrices, data frames y listas. R también ofrece herramientas de control de flujo, como bucles y condicionales, y permite definir funciones para mejorar la modularidad del código.

5. Paquetes de R
    - Instalación y carga de paquetes
    - Procesamiento de datos con Dplyr
    - Procesamiento de datos con Tidyr

R cuenta con numerosos paquetes que extienden su funcionalidad. La instalación y carga de paquetes es sencilla y permite acceder a herramientas especializadas

para diversas tareas. Dplyr es un paquete fundamental para la manipulación de datos, permitiendo realizar operaciones como filtrado, agrupamiento y transformación de manera intuitiva. Tidyr es otra librería importante que se usa para organizar y limpiar los datos, asegurando que estén en el formato correcto para el análisis.

6. Procesamiento de datos y visualización con R
    - Gráficos básicos en R
    - Gráficos en capas Ggplot2
    - Gráficos dinámicos Plotly

R es muy poderoso en la creación de gráficos y visualizaciones de datos. Los gráficos básicos en R permiten una comprensión rápida de los datos, mientras que Ggplot2 ofrece una forma flexible de crear gráficos en capas, personalizando cada aspecto de la visualización. Plotly es una herramienta que permite la creación de gráficos dinámicos e interactivos, que son particularmente útiles para presentaciones y exploración visual de datos.

MÓDULO 2 - Impacto y valor del big data

1. Introducción al mundo big data
    - Hadoop y la aparición del fenómeno big data
    - ¿Por qué ahora?
    - Definición de big data
    - Analítica descriptiva, predictiva y prescriptiva
    - Relaciones entre big data, ciencia de datos y aprendizaje automático

Hadoop marcó el inicio del fenómeno del big data, ofreciendo una solución escalable y distribuida para manejar grandes volúmenes de datos. Este fenómeno ha ganado relevancia en los últimos años debido al aumento de la capacidad de almacenamiento y procesamiento de datos. Big data se define por las características de volumen, variedad y velocidad. La analítica descriptiva, predictiva y prescriptiva son las tres categorías principales de análisis aplicadas a big data para comprender el pasado, prever el futuro y recomendar acciones. Big data, ciencia de datos y aprendizaje automático están estrechamente relacionados, siendo el aprendizaje automático una herramienta clave para extraer valor de los grandes volúmenes de datos.

2. Inteligencia de negocio vs. big data
    - Inteligencia de negocio y big data

- El almacén de datos
- El lago de datos

La inteligencia de negocio (BI) y big data comparten el objetivo de ayudar a las organizaciones a tomar decisiones informadas, pero se diferencian en sus enfoques y herramientas. BI utiliza principalmente el almacén de datos para realizar análisis históricos, mientras que big data emplea el lago de datos, que permite almacenar datos en su forma bruta, proporcionando mayor flexibilidad. Ambos enfoques son complementarios y pueden trabajar juntos para proporcionar una visión integral del negocio.

3. Tecnologías big data
    - Las tres uves como marco de análisis
    - Volumen. Hadoop y Spark
    - Velocidad. Spark Streaming
    - Variedad. NoSQL
    - Arquitectura de referencia
    - Arquitecturas batch y streaming
    - Arquitecturas lambda y kappa

El análisis de big data se basa en las "tres uves": volumen, velocidad y variedad. Hadoop y Spark son tecnologías clave para manejar el volumen de datos, mientras que Spark Streaming se usa para procesar datos a gran velocidad en tiempo real. Las bases de datos NoSQL permiten gestionar la variedad de datos que no se ajustan al modelo relacional tradicional. La arquitectura de referencia para big data combina arquitecturas batch y streaming para cubrir diferentes necesidades de procesamiento. Las arquitecturas Lambda y Kappa se utilizan para equilibrar el procesamiento en tiempo real y en lotes, optimizando la eficiencia y flexibilidad.

4. Impacto sobre la organización
    - El papel de la tecnología big data en la transformación digital
    - Modelos de madurez
    - La estrategia big data
    - La empresa data-driven
    - Perfiles profesionales
    - Gestión del cambio

La tecnología big data juega un papel crucial en la transformación digital de las organizaciones, permitiéndoles ser más eficientes y estar mejor preparadas para el cambio. Los modelos de madurez ayudan a evaluar el estado actual de una empresa en términos de adopción de big data y definir una estrategia clara. La empresa "data-driven" toma decisiones basadas en datos y se apoya en perfiles profesionales como científicos de datos, ingenieros de datos y analistas de negocio. La gestión del cambio es esencial para asegurar una transición exitosa hacia un enfoque centrado en datos.

5. Valor del dato y aplicaciones por sectores
    - El valor del dato
    - Aplicaciones por sectores

El valor del dato se traduce en una mejor toma de decisiones y eficiencia operativa. Las aplicaciones del big data varían según el sector: en el comercio minorista, se utiliza para optimizar el inventario y personalizar la experiencia del cliente; en el sector financiero, para evaluar riesgos y detectar fraudes; y en la atención médica, para mejorar el diagnóstico y el tratamiento. Cada sector puede aprovechar el valor del big data para obtener ventajas competitivas significativas.

MÓDULO 3 - Inteligencia artificial para la empresa

1. Introducción a la inteligencia artificial
    - Introducción a la inteligencia artificial
    - Tipos de inteligencia artificial
    - Historia
    - Ramas de inteligencia artificial
    - Aplicaciones prácticas y tendencias

La inteligencia artificial (IA) es un campo de la informática que se enfoca en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana. Los tipos de IA incluyen la inteligencia artificial estrecha (ANI), que realiza tareas específicas, la inteligencia general (AGI), que aspira a ser tan capaz como un ser humano, y la superinteligencia (ASI), que superaría las capacidades humanas. La historia de la IA comienza en los años 50 con la invención de las primeras computadoras y ha evolucionado rápidamente con la aparición del aprendizaje automático y el deep learning. Las ramas principales de la IA incluyen el aprendizaje automático, el procesamiento del lenguaje natural (NLP), la visión por computadora y la robótica. Las aplicaciones prácticas abarcan desde

asistentes virtuales hasta vehículos autónomos, y las tendencias actuales apuntan hacia una mayor personalización y automatización en diversos sectores.

2. Técnicas y aplicaciones para la toma de decisiones
    - Introducción a las técnicas y aplicaciones para la toma de decisiones
    - Sistemas expertos
    - Aprendizaje supervisado

Las técnicas y aplicaciones de IA para la toma de decisiones se utilizan para ayudar a las organizaciones a seleccionar la mejor opción entre múltiples alternativas. Los sistemas expertos son un tipo de sistema de IA que utiliza reglas y conocimiento experto para proporcionar recomendaciones en áreas específicas, como el diagnóstico médico. El aprendizaje supervisado también se usa para la toma de decisiones, ya que permite a los modelos aprender de datos etiquetados y predecir resultados futuros que ayudan a guiar la toma de decisiones empresariales.

3. Aprendizaje por refuerzo y aplicaciones
    - ¿Qué es el aprendizaje por refuerzo? Ciclo de vida
    - Explorar vs. explotar
    - Ecuación de Bellman: programación dinámica
    - Q-Function: state-action value function
    - Algoritmos
    - Ejemplo: optimización de tareas en un almacén de comercio electrónico con Q-Learning
    - Bandido multibrazo
    - Ejemplo: maximización de visitas en campañas de marketing online con el bandido multibrazo

El aprendizaje por refuerzo es una técnica de IA en la que un agente aprende a tomar decisiones mediante prueba y error, recibiendo recompensas o penalizaciones según sus acciones. El ciclo de vida del aprendizaje por refuerzo comienza con la definición del entorno y la recompensa. El dilema de explorar vs. explotar se refiere a la elección entre explorar nuevas acciones o explotar el conocimiento adquirido. La ecuación de Bellman y la función Q (Q-Function) son conceptos clave que ayudan al agente a calcular el valor de las acciones. Ejemplos de aplicaciones incluyen la optimización de tareas en un almacén de comercio electrónico mediante Q-Learning, y la maximización de visitas en campañas de marketing

online utilizando el bandido multibrazo, un enfoque que equilibra la exploración de nuevas estrategias con la explotación de las que ya han demostrado ser efectivas.

4. Técnicas y aplicaciones del procesamiento del lenguaje natural (NLP)
    - Introducción al procesamiento del lenguaje natural
    - Preprocesamiento de textos
    - Análisis de sentimientos
    - Topic modeling
    - Chatbots

El procesamiento del lenguaje natural (NLP) es una rama de la IA que permite a las máquinas entender y generar lenguaje humano. El preprocesamiento de textos incluye tareas como la tokenización, eliminación de ruido y normalización, que preparan el texto para el análisis. El análisis de sentimientos es una técnica común de NLP que determina la actitud expresada en un texto, como si es positivo, negativo o neutral. Topic modeling se usa para identificar temas dentro de grandes volúmenes de textos, mientras que los chatbots son aplicaciones de NLP que permiten a las empresas proporcionar atención al cliente automatizada y eficaz.

5. Sistemas de recomendación y aplicaciones
    - Introducción a los sistemas de recomendación de aplicaciones
    - ¿Por qué los sistemas de recomendación?
    - Tipos de sistemas de recomendación
    - Filtrado colaborativo
    - Filtrado basado en contenido
    - Filtrado demográfico
    - Ejemplos de sistemas de recomendación comerciales

Los sistemas de recomendación son herramientas de IA que sugieren productos o contenidos a los usuarios en función de sus preferencias. Estos sistemas mejoran la experiencia del usuario y aumentan el compromiso con el contenido. Los sistemas de recomendación se clasifican en filtrado colaborativo, que se basa en los gustos de usuarios similares; filtrado basado en contenido, que analiza las características de los ítems que ha consumido el usuario; y filtrado demográfico, que utiliza información del perfil del usuario. Ejemplos de sistemas de recomendación comerciales incluyen las sugerencias de productos en Amazon, las

recomendaciones de películas en Netflix y las listas de reproducción personalizadas en Spotify.

MÓDULO 4 - Tecnologías y herramientas big data

1. Hadoop y su ecosistema
    - Componentes Hadoop. HDFS y MapReduce
    - Clúster Hadoop
    - Herramientas del ecosistema Hadoop
    - Cloudera

Hadoop es una plataforma de software de código abierto que facilita el almacenamiento y procesamiento distribuido de grandes volúmenes de datos. Los componentes principales de Hadoop son HDFS (Hadoop Distributed File System), que maneja el almacenamiento de datos, y MapReduce, que se encarga del procesamiento paralelo de grandes cantidades de datos. Un clúster Hadoop es un conjunto de nodos que trabajan juntos para almacenar y procesar datos de manera distribuida. El ecosistema de Hadoop incluye herramientas como Hive, Pig y Sqoop, que amplían sus capacidades para la manipulación y el análisis de datos. Cloudera es una distribución comercial de Hadoop que proporciona herramientas adicionales para la gestión, monitoreo y escalabilidad de los clústeres Hadoop en entornos empresariales.

2. Spark. Fundamentos y aplicaciones
    - Características de Apache Spark
    - Arquitectura
    - Estructuras de datos
    - Componentes de SPARK
    - GraphX
    - SparkR
    - Databricks

Apache Spark es una plataforma de procesamiento de datos en memoria que ofrece una alternativa más rápida a Hadoop MapReduce. Spark se destaca por su capacidad para procesar datos tanto en tiempo real como en modo batch. Su arquitectura permite el procesamiento distribuido y escalable, utilizando estructuras de datos como los RDDs (Resilient Distributed Datasets). Los componentes principales de Spark incluyen Spark SQL para el procesamiento

estructurado, Spark Streaming para el análisis en tiempo real, GraphX para el análisis de grafos y SparkR para el análisis estadístico con R. Databricks es una plataforma que proporciona un entorno gestionado para ejecutar Spark en la nube, facilitando el análisis colaborativo y la escalabilidad.

3. Bases de datos NoSQL
    - Hbase
    - MongoDB
    - Cassandra
    - Modelos de BBDD orientadas a grafos

Las bases de datos NoSQL están diseñadas para manejar grandes volúmenes de datos no estructurados o semi-estructurados, y proporcionan una mayor flexibilidad en comparación con las bases de datos relacionales tradicionales. HBase es una base de datos NoSQL que funciona sobre HDFS y está optimizada para accesos aleatorios a grandes volúmenes de datos. MongoDB es una base de datos orientada a documentos que utiliza el formato JSON, ideal para aplicaciones con datos no estructurados y cambiantes. Cassandra es una base de datos distribuida orientada a columnas, conocida por su alta disponibilidad y capacidad de escalabilidad horizontal. Además, existen bases de datos orientadas a grafos, como Neo4j, que permiten modelar y explorar relaciones complejas entre datos de manera eficiente.

4. Plataforma Cloud
    - Servicios cloud
    - Microsoft Azure
    - Amazon Web Services (AWS)
    - Google Cloud Platform

Las plataformas en la nube proporcionan servicios que permiten a las organizaciones almacenar, procesar y analizar datos sin necesidad de infraestructura física propia. Los servicios cloud se dividen en tres categorías principales: IaaS (Infraestructura como Servicio), PaaS (Plataforma como Servicio) y SaaS (Software como Servicio). Microsoft Azure, Amazon Web Services (AWS) y Google Cloud Platform (GCP) son los principales proveedores de servicios en la nube. Microsoft Azure ofrece soluciones como Azure Machine Learning y Azure Databricks para el análisis de datos. AWS proporciona servicios como Amazon S3 para almacenamiento y EMR para el procesamiento de datos con Hadoop y Spark.

Google Cloud Platform cuenta con herramientas como BigQuery para el análisis rápido y eficiente de grandes volúmenes de datos.

MÓDULO 5 - El Big Data en la empresa

1. Estándares de gestión de proyectos
    - Fundamentos de la gestión de proyectos
    - El estándar PMBOK
    - Otros estándares

La gestión de proyectos se basa en un conjunto de principios y prácticas que permiten planificar, ejecutar y controlar proyectos de manera eficiente. El estándar PMBOK (Project Management Body of Knowledge) es uno de los más reconocidos a nivel mundial y proporciona una guía para la gestión de proyectos a través de procesos estandarizados. Además de PMBOK, existen otros estándares como PRINCE2 y Agile PM, que ofrecen diferentes enfoques para la gestión de proyectos dependiendo de los objetivos y la naturaleza del trabajo.

2. Gestión ágil de proyectos
    - Principios ágiles
    - SCRUM
    - Otros marcos ágiles

La gestión ágil de proyectos se basa en los principios ágiles, que promueven la adaptabilidad, la colaboración y la entrega incremental de valor. SCRUM es uno de los marcos ágiles más utilizados, y divide los proyectos en ciclos cortos llamados sprints para mejorar la flexibilidad y la capacidad de respuesta ante cambios. Existen otros marcos ágiles, como Kanban y Lean, que se centran en la optimización del flujo de trabajo y la mejora continua, siendo ideales para proyectos que requieren alta adaptabilidad.

3. Aspectos regulatorios y éticos
    - El marco regulatorio actual en el ámbito de la UE
    - El Reglamento General de Protección de Datos
    - Aspectos éticos

En el ámbito de la Unión Europea, el marco regulatorio actual establece directrices claras para la protección de datos y la privacidad. El Reglamento General de Protección de Datos (GDPR) es una normativa que garantiza los derechos de los ciudadanos sobre sus datos personales y regula cómo las empresas deben

manejarlos. Además de los aspectos regulatorios, es importante tener en cuenta los aspectos éticos, que incluyen la transparencia, la equidad y la responsabilidad en el uso de los datos y las tecnologías emergentes.

4. Gobierno del dato
    - La importancia del gobierno del dato
    - Componentes y herramientas de gobierno del dato
    - Marcos de gestión de datos
    - Pasos hacia el gobierno del dato

El gobierno del dato es fundamental para asegurar la calidad, seguridad y valor de la información en una organización. Los componentes del gobierno del dato incluyen políticas, estándares y roles definidos que aseguran una correcta gestión de los datos. Las herramientas de gobierno del dato permiten la implementación efectiva de estas políticas. Los marcos de gestión de datos, como DAMA-DMBOK, ofrecen pautas para organizar y administrar los datos de manera eficiente. Implementar un gobierno del dato exitoso implica seguir varios pasos, como establecer una estrategia clara, definir roles y responsabilidades, y utilizar tecnología adecuada para el monitoreo y control de los datos.

MÓDULO 6 - Aplicaciones por sectores. Masterclasses, estudio de casos y talleres prácticos

1. E-commerce y marketing
    - Personalización de la experiencia del usuario
    - Seguimiento del comportamiento del consumidor
    - Predecir la demanda y evitar roturas de stock
2. Banca y finanzas
    - Retos del sector bancario
    - Era del big data: competir en un mundo data-driven
    - Implantación de una estrategia big data en banca
    - Buscando nuevas fuentes de crecimiento
    - Apificación del sistema bancario
    - Generación de un ecosistema
    - Colaboración con competidores no tradicionales

3. People analytics
    - People analytics
    - Casos de estudio
    - Voz del empleado
4. Telecomunicaciones
    - El sector de las telecomunicaciones
    - Aplicaciones de ciencia de datos en telecomunicaciones
5. Ciencia y salud
    - Medicina basada en la evidencia
    - Otras dimensiones del uso de los datos
    - Ejemplos
6. Industria 4.0, internet de las cosas (IoT) y Smart cities
    - Internet de las cosas
    - Aplicaciones sectoriales

MÓDULO 7 - Cloud, MLops, productivización de modelos. Introducción a process mining

1. BI
    - Modelo de datos y ETL
    - Herramientas de BI
    - Dashboards

El Business Intelligence (BI) se centra en la recolección, transformación y análisis de datos para apoyar la toma de decisiones en una organización. El modelo de datos y los procesos ETL (Extract, Transform, Load) son fundamentales para organizar los datos de manera estructurada, permitiendo su posterior análisis. Las herramientas de BI, como Power BI, Tableau y QlikView, son ampliamente utilizadas para visualizar y explorar la información. Los dashboards, o paneles de control, son elementos clave que permiten a los usuarios visualizar indicadores clave de rendimiento (KPIs) y tomar decisiones informadas de forma rápida y precisa.

2. Process mining
    - La captura de datos de los procesos

- Celonis

El process mining es una disciplina que se enfoca en la captura y análisis de datos de procesos empresariales para mejorar su eficiencia. La captura de datos se realiza mediante el registro de eventos que muestran cómo se ejecutan los procesos en realidad. Celonis es una de las herramientas líderes en process mining y permite identificar cuellos de botella, desviaciones y áreas de mejora en los procesos empresariales mediante el análisis detallado de los datos recopilados.

3. Cloud
    - Conceptos básicos sobre la nube
    - Servicios básicos
    - Ejemplos de los servicios básicos

La nube, o cloud computing, es un modelo que permite acceder a recursos informáticos y servicios a través de internet. Los conceptos básicos sobre la nube incluyen la computación a demanda, escalabilidad y flexibilidad. Los servicios básicos en la nube se dividen en Infraestructura como Servicio (IaaS), Plataforma como Servicio (PaaS) y Software como Servicio (SaaS). Ejemplos de estos servicios incluyen Amazon EC2 para IaaS, Google App Engine para PaaS y Microsoft Office 365 para SaaS, que proporcionan soluciones adecuadas para diferentes necesidades empresariales.

4. Inteligencia artificial en cloud
    - API cognitivas
    - Herramientas de inteligencia artificial
    - Herramientas de inteligencia artificial escalables

La nube también facilita el acceso a tecnologías avanzadas como la inteligencia artificial (IA). Las API cognitivas son servicios que permiten integrar capacidades de IA, como el reconocimiento de voz, la traducción automática y el análisis de imágenes, en aplicaciones sin necesidad de desarrollar modelos complejos. Herramientas de inteligencia artificial como TensorFlow y PyTorch están disponibles en la nube, permitiendo un acceso fácil y escalable a los recursos necesarios para entrenar y ejecutar modelos de IA. Las herramientas escalables de IA en la nube permiten a las empresas implementar soluciones de machine learning y deep learning sin tener que preocuparse por la infraestructura subyacente.

5. Productividad de modelos
    - La explotación de un modelo

- Explotación de la ingeniería de características
- Modelos en producción

La explotación de un modelo implica su uso para resolver problemas prácticos y generar valor para la organización. La explotación de la ingeniería de características consiste en mejorar la calidad de los datos utilizados por un modelo para maximizar su rendimiento. Los modelos en producción deben estar bien monitorizados para garantizar su efectividad y adaptarse a cambios en los datos o en el entorno de negocio, asegurando que continúan proporcionando resultados precisos y útiles.

6. MLOps
    - Requerimientos de MLOps
    - MLOps en Azure
    - MLOps en AWS
    - MLOps en Google

MLOps es la práctica que combina el desarrollo de modelos de machine learning con las operaciones de TI para facilitar la implementación, monitoreo y mantenimiento de modelos en producción. Los requerimientos de MLOps incluyen la automatización del proceso de desarrollo y la integración continua. MLOps en Azure se implementa a través de herramientas como Azure Machine Learning, que facilitan la colaboración y el monitoreo de modelos. AWS ofrece SageMaker, una solución integral para gestionar el ciclo de vida de los modelos de machine learning. En Google, MLOps se gestiona mediante Vertex AI, que permite a los equipos de datos crear y desplegar modelos de forma eficiente y escalable.

MÓDULO 8 - Series temporales y modelos prescriptivos. Optimización. Modelos de grafos

1. Optimización
    - Descripción de la optimización matemática
    - Programación lineal
    - Programación entera
    - Programación no lineal
    - Heurísticas y metaheurísticas
    - Optimización bajo incertidumbre
    - Optimización y machine learning

La optimización matemática se refiere al proceso de encontrar la mejor solución posible a un problema bajo ciertas restricciones. La programación lineal busca optimizar una función objetivo sujeta a restricciones lineales, mientras que la programación entera es una extensión que limita algunas o todas las variables a valores enteros. La programación no lineal se ocupa de problemas donde las funciones objetivo o las restricciones no son lineales. Las heurísticas y metaheurísticas son enfoques para resolver problemas de optimización complejos donde las soluciones exactas son difíciles de encontrar. La optimización bajo incertidumbre considera la variabilidad y el riesgo en los parámetros del problema. Además, la optimización es un componente crucial del machine learning, ya que se utiliza para entrenar modelos ajustando parámetros para minimizar errores.

2. Teoría de grafos
    - Introducción a los grafos
    - Análisis básico de grafos
    - Análisis de importancia y centralidad
    - Detección de comunidades y clustering
    - Simulación de sistemas

La teoría de grafos es un campo de las matemáticas que estudia las relaciones entre objetos representados por nodos y aristas. Los grafos se utilizan para modelar redes y sistemas complejos. El análisis básico de grafos permite calcular propiedades como la conectividad y los caminos mínimos entre nodos. Los análisis de importancia y centralidad identifican los nodos más influyentes dentro de una red. La detección de comunidades y el clustering se emplean para identificar grupos de nodos que están más conectados entre sí. La simulación de sistemas basada en grafos permite modelar y analizar dinámicas complejas en redes, como la propagación de información o enfermedades.

3. Series temporales
    - Definición de series temporales
    - Modelos ARMA y ARIMA
    - Modelos SARMA y SARIMA
    - Metodología Box-Jenkins
    - Función de transferencia
    - Análisis de outliers e intervenciones
    - Series temporales en la era del Big Data

Las series temporales son secuencias de datos recogidos en intervalos de tiempo regulares, utilizadas para analizar cómo evolucionan los valores a lo largo del tiempo. Los modelos ARMA (AutoRegressive Moving Average) y ARIMA (AutoRegressive Integrated Moving Average) se utilizan para modelar datos estacionarios y predecir valores futuros. Los modelos SARMA y SARIMA son extensiones de ARIMA que incorporan estacionalidad en los datos. La metodología Box-Jenkins es un enfoque sistemático para identificar, estimar y verificar modelos de series temporales. La función de transferencia se utiliza para modelar el efecto de una o más series sobre otra. El análisis de outliers e intervenciones permite detectar eventos atípicos que podrían afectar el comportamiento de la serie. En la era del Big Data, las series temporales se analizan a gran escala para aplicaciones como la predicción de demanda y la monitorización de sistemas.

MÓDULO 9 - Deep learning aplicada: NLP y visión artificial

1. Machine Learning aplicado al procesamiento del lenguaje natural (NLP)
    - Limpieza de texto
    - Expresiones regulares
    - Vectorización (TF-IDF)
    - Concepto de distancia: distancias de textos
    - Algoritmos no supervisados
    - Algoritmos supervisados

El machine learning aplicado al procesamiento del lenguaje natural (NLP) permite automatizar y mejorar el análisis de textos. La limpieza de texto es el primer paso fundamental y consiste en eliminar elementos no deseados como signos de puntuación y caracteres especiales. Las expresiones regulares son una herramienta potente para realizar este tipo de limpieza de manera eficiente. La vectorización, como TF-IDF (Term Frequency-Inverse Document Frequency), convierte el texto en representaciones numéricas para ser procesadas por los algoritmos de aprendizaje. El concepto de distancia se usa para medir la similitud entre textos, utilizando métricas como la distancia coseno o la distancia Euclidiana. Se aplican tanto algoritmos no supervisados, como el clustering para agrupar documentos similares, como algoritmos supervisados para tareas de clasificación de texto.

2. Modelos avanzados de NLP. Enfoques con Deep Learning
    - Vectrorización basada en deep learning: Word2Vect
    - Modelos de clasificación (RRNN, GRU y convolucionales 1D)

- Attention mechanism

Los enfoques avanzados de NLP utilizan técnicas de deep learning para capturar mejor los matices del lenguaje. Word2Vec es una técnica de vectorización basada en deep learning que genera representaciones vectoriales de palabras manteniendo sus relaciones semánticas. Para la clasificación de textos, se utilizan modelos de redes neuronales recurrentes (RNN), GRU y convolucionales 1D, que capturan patrones secuenciales y espaciales en los datos de texto. El mecanismo de atención (Attention mechanism) permite a los modelos enfocarse en las partes más relevantes de la entrada, mejorando la precisión y eficiencia de las tareas de NL

3. Introducción a la visión artificial
    - Introducción a la teoría bajo las redes neuronales convolucionales
    - Numpy y Open CV: cómo trabajar con imágenes
    - YOLO: detección automática de objetos en imágenes

La visión artificial es un campo de la inteligencia artificial que se enfoca en enseñar a las computadoras a interpretar y entender imágenes. Las redes neuronales convolucionales (CNN) son la base de la teoría aplicada a la visión artificial, permitiendo la extracción automática de características visuales. Herramientas como Numpy y OpenCV son fundamentales para trabajar con imágenes y realizar operaciones básicas como la manipulación y filtrado de imágenes. YOLO (You Only Look Once) es un modelo popular para la detección automática de objetos, capaz de identificar múltiples objetos en una imagen de forma rápida y precisa.

4. Aplicación del Deep learning a la visión artificial
    - Arquitecturas de redes convolucionales
    - Optimización de la inicialización de pesos y regularización de modelos
    - Transfer learning: qué es y cómo funciona
    - Qué son y cómo funcionan los Autoencoders
    - Principales frameworks de deep learning: Keras-TensorFlow y Pytorch

Las arquitecturas de redes convolucionales (CNN) son ampliamente utilizadas en la visión artificial para tareas como clasificación de imágenes y detección de objetos. La optimización de la inicialización de pesos y la regularización de modelos son prácticas importantes para evitar problemas como el sobreajuste. Transfer learning es una técnica que reutiliza modelos preentrenados para nuevas tareas, reduciendo el tiempo de entrenamiento y mejorando los resultados. Los

autoencoders son redes neuronales utilizadas para la reducción de dimensionalidad y la detección de anomalías. Los principales frameworks de deep learning, como Keras-TensorFlow y PyTorch, proporcionan herramientas para implementar y entrenar estos modelos de manera eficiente.

MÓDULO 10 - Trabajo Fin de Máster en IA

Máster en Data Science

URL: https://www.imf-formacion.com/masters-profesionales/master-data-science

MÓDULO 1 - Las herramientas del científico de datos

1. Fundamentos de Python.
    - El lenguaje de programación Python
    - Entorno de programación: Jupyter Notebook
    - Sintaxis básica de Python
    - Herramientas para control de flujo
    - Estructuras de datos
    - Funciones

Python es un lenguaje de programación popular por su simplicidad y versatilidad, adecuado para principiantes y expertos. Jupyter Notebook es un entorno interactivo que permite escribir y ejecutar código Python, ideal para proyectos de ciencia de datos y análisis interactivo. La sintaxis de Python es clara y concisa, lo que facilita su aprendizaje. Herramientas para el control de flujo, como condicionales y bucles, permiten gestionar la lógica de los programas. Las estructuras de datos, como listas, diccionarios y tuplas, son esenciales para organizar información. Además, las funciones ayudan a modularizar el código, mejorando su reutilización y legibilidad.

2. Librerías para ciencia de datos: Numpy, Pandas, etc.
    - Procesamiento de archivos
    - Numpy
    - Pandas

Python cuenta con varias librerías esenciales para la ciencia de datos. El procesamiento de archivos permite leer y escribir datos de diferentes formatos, como CSV o Excel. Numpy es una librería fundamental para el manejo de arrays y operaciones matemáticas complejas. Pandas facilita la manipulación y análisis de datos estructurados, proporcionando estructuras como DataFrames para trabajar con grandes conjuntos de datos de forma eficiente.

3. Procesamiento de datos y visualización con Python.
    - Matplotilib

- GGPLOT
- Seaborn
- BOKEH

La visualización de datos es clave para entender y comunicar los resultados del análisis. Matplotlib es una de las librerías más utilizadas para crear gráficos en Python, ofreciendo flexibilidad para generar diferentes tipos de visualizaciones. Seaborn y GGPLOT están construidos sobre Matplotlib y facilitan la creación de gráficos estadísticos. Bokeh, por otro lado, es útil para crear visualizaciones interactivas que permiten una exploración más dinámica de los datos.

4. Fundamentos de R.
    - El lenguaje de programación R
    - Entorno de programación: Rstudio
    - Sintaxis básica de R
    - Estructuras de datos
    - Herramientas de control de flujo
    - Funciones

R es un lenguaje de programación enfocado en el análisis estadístico y la visualización de datos. RStudio es el entorno de programación más utilizado para R, proporcionando una interfaz amigable y herramientas útiles para el desarrollo de proyectos. La sintaxis de R permite realizar operaciones matemáticas y manipular datos de manera eficiente. Las estructuras de datos, como vectores, listas y data frames, son fundamentales en R, y las herramientas de control de flujo permiten gestionar la lógica del análisis. Además, las funciones ayudan a automatizar tareas repetitivas y organizar el código.

5. Paquetes de R.
    - Instalación y carga de paquetes
    - Procesamiento de datos con Dplyr
    - Procesamiento de datos con Tidyr

R tiene una amplia gama de paquetes que extienden sus capacidades. La instalación y carga de paquetes es simple, y permite acceder a funciones específicas según las necesidades del análisis. Dplyr es un paquete muy popular para la manipulación de datos, que facilita operaciones como filtrado,

agrupamiento y resumen de datos. Tidyr, por su parte, se utiliza para limpiar y transformar datos, permitiendo prepararlos para el análisis de forma eficiente.

6. Procesamiento de datos y visualización con R.
    - Gráficos básicos en R
    - Gráficos en capas Ggplot2
    - Gráficos dinámicos Plotly

R es muy potente para la visualización de datos, ofreciendo diferentes herramientas para crear gráficos. Los gráficos básicos en R permiten visualizar datos de manera rápida y sencilla. Ggplot2 es un paquete popular que permite crear gráficos en capas, proporcionando flexibilidad para combinar diferentes elementos visuales. Plotly se usa para crear gráficos dinámicos e interactivos, lo cual es muy útil para presentaciones y análisis exploratorios en los que se busca una mayor interacción con los datos.

MÓDULO 2 - La ciencia de datos. Técnicas de análisis, minería y visualización

1. El ciclo de vida del dato
    - Definición de ciencia de datos
    - El ciclo de vida de los datos
    - Definición de objetivos en un proyecto de datos
    - Identificación de los datos necesarios
    - Preparación y preproceso
    - Análisis y modelado
    - Validación y prueba

La ciencia de datos se refiere al uso de técnicas y herramientas para extraer información y conocimiento de los datos. El ciclo de vida del dato comienza con la definición de los objetivos de un proyecto de datos, lo que implica entender el problema a resolver y los resultados esperados. Luego, se identifican los datos necesarios para cumplir esos objetivos. La preparación y preproceso de datos incluyen la limpieza, transformación y estructuración de los datos para su análisis. En la fase de análisis y modelado, se aplican técnicas estadísticas y de aprendizaje automático para descubrir patrones y generar modelos predictivos. La validación y prueba aseguran que los modelos son precisos y útiles. Finalmente, la fase de

explotación se enfoca en implementar los resultados para generar valor para la organización.Explotación

2. Calidad del dato
    - Fundamentos de calidad de datos
    - Las dimensiones de calidad de datos
    - Los procesos de incremento de la calidad de datos

La calidad de los datos es fundamental para asegurar que los resultados del análisis sean precisos y útiles. Los fundamentos de la calidad de los datos incluyen la precisión, completitud y coherencia de los mismos. Las dimensiones de calidad de datos abarcan aspectos como la validez, la accesibilidad y la actualidad. Los procesos de incremento de la calidad de los datos implican tareas de limpieza, deduplicación y enriquecimiento de los datos, lo cual asegura que la información sea confiable y adecuada para el análisis.

3. Tecnologías big data.
    - La recolección de datos
    - Limpieza y preproceso de datos
    - Transformación de datos

Las tecnologías big data son cruciales para gestionar y analizar grandes volúmenes de datos. La recolección de datos implica el uso de herramientas para recopilar información de diversas fuentes, como sensores, redes sociales y sistemas empresariales. La limpieza y preproceso de datos aseguran que los datos sean útiles y estén en el formato adecuado para el análisis. La transformación de datos permite convertir la información en una estructura que pueda ser analizada de manera efectiva, facilitando el descubrimiento de patrones y tendencias.

4. Modelos analíticos
    - El uso de modelos en ciencia de datos
    - Modelos usados en análisis descriptivo
    - Modelos usados en análisis predictivo
    - Modelos usados en análisis prescriptivo
    - Validación y prueba de los modelos

Los modelos analíticos son fundamentales en la ciencia de datos para extraer información útil. Los modelos usados en análisis descriptivo permiten resumir los datos y entender qué ocurrió en el pasado. Los modelos predictivos, por otro lado, se centran en predecir eventos futuros basándose en datos históricos. Los modelos prescriptivos sugieren acciones para alcanzar objetivos específicos. La validación y prueba de los modelos son pasos críticos para asegurar que los resultados sean precisos y se puedan confiar en las predicciones y recomendaciones.

5. Herramientas y técnicas de visualización
    - El papel de la visualización en la ciencia de datos
    - Elementos de comunicación visual
    - La gramática de los gráficos
    - Ejemplo
    - Tipología de gráficos elementales
    - Otros tipos de gráficos de datos
    - Herramientas de visualización

La visualización juega un papel esencial en la ciencia de datos, permitiendo que la información compleja sea comprensible para los usuarios finales. Los elementos de comunicación visual incluyen gráficos, tablas y diagramas que representan los datos de forma clara. La gramática de los gráficos proporciona reglas sobre cómo construir visualizaciones efectivas. Existen diferentes tipos de gráficos elementales, como los gráficos de barras y de líneas, así como gráficos más avanzados que permiten explorar datos de forma interactiva. Herramientas como Tableau, Power BI y Matplotlib son ampliamente usadas para crear visualizaciones impactantes que facilitan la toma de decisiones.

MÓDULO 3 - Estadística para científicos de datos

1. Lenguaje y tratamiento de datos
    - Análisis estadístico con R y RStudio
    - Comando esenciales en R y tipos de datos
    - Operaciones útiles sobre tablas, carga y descarga de datos
    - Funciones y bucles

El análisis estadístico con R y RStudio es una de las principales aplicaciones del lenguaje R, que permite trabajar de manera efectiva con datos. Los comandos esenciales en R, junto con los diferentes tipos de datos, facilitan el manejo de

información y su manipulación. Además, operaciones útiles sobre tablas, así como la carga y descarga de datos, permiten a los usuarios trabajar con grandes volúmenes de información de forma eficiente. Las funciones y los bucles son herramientas fundamentales para automatizar tareas repetitivas y realizar análisis complejos.

2. Análisis exploratorio de datos
    - Tipos de gráficos y sus usos
    - Librería GGPlot
    - Casos de análisis exploratorio de datos

El análisis exploratorio de datos es crucial para entender las características y patrones presentes en los datos. Existen diferentes tipos de gráficos que se utilizan según la naturaleza de los datos y los insights que se buscan. La librería GGPlot es una herramienta poderosa en R para la creación de visualizaciones complejas y en capas, facilitando la exploración y presentación de datos. Los casos de análisis exploratorio ayudan a identificar correlaciones y tendencias antes de aplicar modelos más avanzados.

3. Probabilidad e inferencia estadística
    - Probabilidad
    - Distribuciones de probabilidad
    - Inferencia estadística

La probabilidad es un concepto esencial en la estadística que permite cuantificar la incertidumbre. Las distribuciones de probabilidad describen cómo se distribuyen los valores de una variable aleatoria, y son fundamentales para realizar inferencias. La inferencia estadística se utiliza para hacer generalizaciones sobre una población basándose en una muestra, permitiendo extraer conclusiones y tomar decisiones informadas.

4. Modelos lineales y aprendizaje estadístico
    - Probabilidad
    - Análisis mutivariable: covarianza y correlación
    - Regresión lineal múltiple
    - Selección de modelos: equilibrio entre sesgo y varianza
    - Criterios estadísticos de selección de modelos

Los modelos lineales son una base importante en el aprendizaje estadístico. El análisis multivariable, que incluye conceptos como la covarianza y la correlación, permite entender las relaciones entre múltiples variables. La regresión lineal múltiple se utiliza para modelar la relación entre una variable dependiente y varias independientes. La selección de modelos es crucial para encontrar el equilibrio entre sesgo y varianza, utilizando criterios estadísticos que aseguren un buen ajuste sin sobreajustar los datos.

5. Regresión logística, modelos restringidos de Ridge y Lasso y gradiente
    - Regresión logística
    - Interpretación de coeficientes en la regresión logística
    - Métricas en problemas de clasificación binaria
    - Modelos restringidos de Ridge y Lasso y elastic net
    - GLM y series temporales
    - Algoritmo de gradiente descendente

La regresión logística es una técnica utilizada para problemas de clasificación, como predecir categorías binarias. La interpretación de los coeficientes en la regresión logística ayuda a entender cómo afectan las variables independientes al resultado. Las métricas en problemas de clasificación binaria, como la precisión y la sensibilidad, son fundamentales para evaluar el rendimiento del modelo. Los modelos restringidos de Ridge y Lasso se usan para prevenir el sobreajuste, aplicando penalizaciones a los coeficientes. GLM y las series temporales permiten modelar datos con dependencia temporal, mientras que el algoritmo de gradiente descendente se usa para optimizar los parámetros del modelo y minimizar errores.

MÓDULO 4 - Aprendizaje automático

1. Herramientas para machine learning
    - Visión general de machine learning
    - Estructurar un proyecto de machine learning

El machine learning es una rama de la inteligencia artificial que permite a las computadoras aprender a partir de los datos. La visión general del machine learning incluye conceptos como el aprendizaje supervisado, no supervisado y por refuerzo. Para estructurar un proyecto de machine learning, es fundamental definir los objetivos, recopilar y preparar los datos, seleccionar los algoritmos adecuados y evaluar los resultados para mejorar continuamente el modelo.

2. Técnicas y aplicaciones del aprendizaje supervisado

- Problemas de regresión
- Problemas de clasificación

El aprendizaje supervisado se utiliza cuando se tiene un conjunto de datos etiquetados. En los problemas de regresión, el objetivo es predecir un valor continuo, como el precio de una casa. En problemas de clasificación, el modelo busca asignar una etiqueta o categoría, por ejemplo, clasificar correos electrónicos como spam o no spam. Estas técnicas son muy útiles para aplicaciones empresariales y científicas.

3. Técnicas y aplicaciones del aprendizaje no supervisado
    - Algoritmos de clustering
    - Algoritmos de reducción dimensional
    - Algoritmos de detección de anomalías

El aprendizaje no supervisado se aplica cuando los datos no están etiquetados, y el objetivo es descubrir patrones ocultos. Los algoritmos de clustering se utilizan para agrupar datos similares, como segmentar clientes en un análisis de mercado. Los algoritmos de reducción dimensional, como PCA, ayudan a simplificar conjuntos de datos grandes y facilitar su análisis. La detección de anomalías permite identificar comportamientos atípicos en los datos, como posibles fraudes en transacciones.

4. Modalidades y técnicas de deep learning
    - Redes neuronales y deep learning
    - Deep learning en la práctica

El deep learning se basa en redes neuronales profundas, que imitan la forma en que el cerebro humano procesa la información. Las redes neuronales y el deep learning se utilizan en aplicaciones como el reconocimiento de imágenes y el procesamiento del lenguaje natural. En la práctica, el deep learning requiere grandes volúmenes de datos y potencia computacional, pero ofrece resultados excepcionales en problemas complejos.

5. Soluciones en la nube para machine learning
    - Herramientas de AutoML
    - Infraestructura de aprendizaje automático como servicio (ML IaaS)
    - Otros servicios en la nube

Las soluciones en la nube permiten a las empresas implementar modelos de machine learning sin necesidad de invertir en infraestructura propia. Las herramientas de AutoML facilitan el proceso de entrenamiento y optimización de modelos automáticamente. La infraestructura de aprendizaje automático como servicio (ML IaaS) proporciona recursos escalables para entrenar modelos. Además, otros servicios en la nube permiten gestionar el ciclo de vida del machine learning, desde la recopilación de datos hasta la implementación de modelos.

MÓDULO 5 - Inteligencia artificial para la empresa

1. Introducción a la inteligencia artificial
    - Introducción a la inteligencia artificial
    - Tipos de inteligencia artificial
    - Historia
    - Ramas de inteligencia artificial
    - Aplicaciones prácticas y tendencias

La inteligencia artificial (IA) se refiere a la capacidad de las máquinas para realizar tareas que normalmente requieren inteligencia humana. Los tipos de IA se dividen en inteligencia artificial estrecha (ANI), inteligencia general (AGI) y superinteligencia (ASI). La historia de la IA comienza en los años 50 con el desarrollo de las primeras computadoras programables, evolucionando hasta la actualidad con avances como el aprendizaje automático y el deep learning. Existen varias ramas de la IA, incluyendo el machine learning, el procesamiento del lenguaje natural y la robótica. Las aplicaciones prácticas de la IA incluyen asistentes virtuales, vehículos autónomos y sistemas de recomendación, mientras que las tendencias actuales apuntan hacia una mayor automatización y personalización.

2. Técnicas y aplicaciones para la toma de decisiones
    - Introducción a las técnicas y aplicaciones para la toma de decisiones
    - Sistemas expertos
    - Aprendizaje supervisado

Las técnicas y aplicaciones para la toma de decisiones se basan en el uso de algoritmos y modelos para seleccionar la mejor opción entre varias alternativas. Los sistemas expertos son un tipo de IA que imita el conocimiento de un especialista en un dominio específico, ofreciendo recomendaciones. El aprendizaje supervisado también se aplica en la toma de decisiones, ya que permite a los modelos predecir resultados futuros basados en datos históricos y ayudar a seleccionar acciones óptimas.

3. Aprendizaje por refuerzo y aplicaciones
    - ¿Qué es el aprendizaje por refuerzo? Ciclo de vida
    - Explorar vs. explotar
    - Ecuación de Bellman: programación dinámica
    - Q-Function: state-action value function
    - Algoritmos
    - Ejemplo: optimización de tareas en un almacén de comercio electrónico con Q-Learning
    - Bandido multibrazo
    - Ejemplo: maximización de visitas en campañas de marketing online con el bandido multibrazo

El aprendizaje por refuerzo es un tipo de aprendizaje automático en el que un agente aprende mediante la interacción con su entorno y la obtención de recompensas. El ciclo de vida del aprendizaje por refuerzo implica definir el entorno, las recompensas y la política que guiará las acciones del agente. El dilema de explorar vs. explotar se refiere a la necesidad de equilibrar la búsqueda de nuevas estrategias (explorar) con el uso de las estrategias ya conocidas (explotar). La ecuación de Bellman y la función Q (Q-Function) son fundamentales para calcular el valor de un estado-acción. Ejemplos de aplicaciones incluyen la optimización de tareas en un almacén de comercio electrónico con Q-Learning y la maximización de visitas en campañas de marketing online usando el bandido multibrazo.

4. Técnicas y aplicaciones del procesamiento del lenguaje natural (NLP)
    - Introducción al procesamiento del lenguaje natural
    - Preprocesamiento de textos
    - Análisis de sentimientos
    - Topic modeling
    - Chatbots

El procesamiento del lenguaje natural (NLP) es una rama de la IA que permite a las máquinas entender y generar lenguaje humano. El preprocesamiento de textos incluye tareas como la limpieza, tokenización y normalización del texto para prepararlo para el análisis. El análisis de sentimientos es una técnica que determina la polaridad de un texto, mientras que el topic modeling ayuda a identificar temas ocultos en grandes conjuntos de documentos. Los chatbots son aplicaciones

prácticas del NLP, y se utilizan ampliamente en atención al cliente para ofrecer respuestas automatizadas y eficientes.

5. Sistemas de recomendación y aplicaciones
    - Introducción a los sistemas de recomendación de aplicaciones
    - ¿Por qué los sistemas de recomendación?
    - Tipos de sistemas de recomendación
    - Filtrado colaborativo
    - Filtrado basado en contenido
    - Filtrado demográfico
    - Ejemplos de sistemas de recomendación comerciales

Los sistemas de recomendación son herramientas que sugieren productos o contenidos a los usuarios basándose en sus preferencias. Estos sistemas son esenciales para mejorar la experiencia del usuario y aumentar el compromiso. Existen varios tipos de sistemas de recomendación: el filtrado colaborativo se basa en las preferencias de usuarios similares, el filtrado basado en contenido analiza las características de los ítems, y el filtrado demográfico utiliza información del perfil del usuario. Ejemplos comerciales incluyen las recomendaciones de películas en Netflix, productos en Amazon y música en Spotify.

MÓDULO 6 - Tecnologías y herramientas big data

1. Hadoop y su ecosistema
    - Componentes Hadoop. HDFS y MapReduce
    - Clúster Hadoop
    - Herramientas del ecosistema Hadoop
    - Cloudera

Hadoop es una plataforma de software de código abierto que permite el almacenamiento y procesamiento distribuido de grandes volúmenes de datos. Sus componentes principales son HDFS (Hadoop Distributed File System), que gestiona el almacenamiento de los datos, y MapReduce, que se encarga del procesamiento paralelo. Un clúster Hadoop está formado por varios nodos que colaboran para almacenar y procesar los datos de manera eficiente. Además, el ecosistema Hadoop incluye herramientas como Hive, Pig y Sqoop, que amplían sus capacidades. Cloudera es una distribución comercial de Hadoop que facilita la implementación y gestión de clústeres Hadoop en entornos empresariales.

2. Spark. Fundamentos y aplicaciones
    - Características de Apache Spark
    - Arquitectura
    - Estructuras de datos
    - Componentes de SPARK
    - GraphX
    - SparkR
    - Databricks

Apache Spark es un marco de procesamiento de datos en memoria que es mucho más rápido que MapReduce de Hadoop. Spark se destaca por su arquitectura resiliente y su capacidad para procesar datos en tiempo real. Sus estructuras de datos, como RDDs (Resilient Distributed Datasets), permiten la manipulación de datos a gran escala. Spark tiene varios componentes importantes, como Spark SQL para consultas estructuradas, GraphX para análisis de grafos, y SparkR para análisis estadístico en R. Databricks es una plataforma que permite ejecutar Spark en la nube de forma eficiente, proporcionando un entorno colaborativo para el análisis de datos.

3. Bases de datos NoSQL
    - Hbase
    - MongoDB
    - Cassandra
    - Modelos de BBDD orientadas a grafos

Las bases de datos NoSQL se utilizan para almacenar y gestionar datos no estructurados o semi-estructurados. HBase es una base de datos NoSQL que se ejecuta sobre HDFS y es adecuada para acceso aleatorio a grandes volúmenes de datos. MongoDB es una base de datos orientada a documentos, ideal para manejar datos flexibles y sin una estructura definida. Cassandra es una base de datos distribuida y orientada a columnas que se destaca por su alta disponibilidad y escalabilidad. También existen bases de datos orientadas a grafos, como Neo4j, que permiten modelar y consultar relaciones complejas entre datos de manera eficiente.

4. Plataformas Cloud
    - Servicios cloud

- Microsoft Azure
- Amazon Web Services (AWS)
- Google Cloud Platform

Las plataformas en la nube proporcionan una infraestructura flexible para la gestión y análisis de datos. Los servicios cloud incluyen almacenamiento, procesamiento y análisis, entre otros. Microsoft Azure ofrece soluciones como Azure Machine Learning y Azure Databricks para ejecutar proyectos de Big Data y Machine Learning. Amazon Web Services (AWS) cuenta con servicios como Amazon S3 para almacenamiento y EMR para el procesamiento de datos a gran escala. Google Cloud Platform (GCP) proporciona herramientas como BigQuery para el análisis rápido de grandes volúmenes de datos. Estas plataformas permiten escalar recursos según las necesidades del proyecto, proporcionando flexibilidad y eficiencia.

MÓDULO 7 - El trabajo del científico de datos: pasos y técnicas en el análisis. Storytelling

1. Introducción: conceptos de data science
    - El método científico
    - Conociendo el dato
    - Técnicas de análisis de datos
    - Herramientas
2. Pasos en el análisis de datos
    - Metodología de la ciencia de datos: CRISP-DM
    - Entendimiento del problema
    - Entendimiento de los datos
    - Preparación de los datos
    - Desarrollo del modelo
    - Evaluación del modelo
    - Despliegue del modelo
3. Storytelling: poner en valor y transmitir los resultados del análisis
    - Narrativa de datos como medio de creación de conocimiento
    - Mejores prácticas en visualización de datos

- Técnicas y herramientas de storytelling

MÓDULO 8 - El proceso de aprendizaje automático: qué es y qué no es. Dónde aplicar la inteligencia artificial

1. Concepto de aprendizaje automático
    - Mapa general de la analítica de datos
    - Introducción al aprendizaje automático
    - Aprendizaje automático supervisado I: algoritmos de regresión
    - Aprendizaje automático supervisado II: algoritmos de clasificación
    - Aprendizaje automático supervisado III: series temporales
    - Aprendizaje automático no supervisado I: clustering
    - Aprendizaje automático no supervisado II: modelos de asociación

El data science o ciencia de datos es un campo interdisciplinario que utiliza el método científico para extraer conocimiento e insights de los datos. El método científico aplicado al análisis de datos implica formular hipótesis, realizar experimentos y evaluar resultados para tomar decisiones informadas. Conocer el dato implica entender su origen, calidad y relevancia para el problema a resolver. Las técnicas de análisis de datos incluyen métodos estadísticos, aprendizaje automático y visualización. Existen diversas herramientas que ayudan a realizar análisis de datos, como Python, R, y librerías especializadas como Pandas, NumPy y Matplotlib.

2. Cómo conseguir que siga aprendiendo
    - Aprendizaje por refuerzo (reinforcement learning)
    - Deep learning

El proceso de análisis de datos sigue la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que proporciona una guía para desarrollar proyectos de ciencia de datos de manera estructurada. El primer paso es el entendimiento del problema, que implica definir claramente los objetivos del proyecto. Luego viene el entendimiento de los datos, donde se analiza la calidad y la relevancia de los datos disponibles. La preparación de los datos incluye tareas de limpieza, transformación y selección de variables relevantes. Durante el desarrollo del modelo se seleccionan y entrenan algoritmos para extraer patrones útiles. La evaluación del modelo permite medir su rendimiento y verificar si cumple con los objetivos. Finalmente, el despliegue del modelo se refiere a su

implementación en un entorno productivo donde pueda ser utilizado por los usuarios.

3. Casos de uso típicos
    - Modelos de fuga
    - Modelos de propensión
    - Clasificación de clientes
    - Detección de anomalías
    - Predicción de ventas
    - Mantenimiento predictivo
    - Sistemas de recomendación

El storytelling con datos es una técnica para comunicar los resultados del análisis de forma clara y efectiva. La narrativa de datos ayuda a transformar la información compleja en conocimiento accesible para la audiencia. Las mejores prácticas en visualización de datos incluyen elegir gráficos adecuados, simplificar la presentación visual y resaltar los puntos clave. Utilizar técnicas y herramientas de storytelling, como gráficos interactivos y narrativas visuales, facilita la comunicación de los resultados y permite que las personas no técnicas comprendan el valor del análisis de datos.

MÓDULO 9 - Nuevas tendencias: process mining, MLOps, cloud

1. Process mining
    - Concepto, potencial y cómo posicionarlo en los clientes
    - Principales soluciones del mercado
    - Principales procesos

Process mining es una técnica que permite analizar los procesos empresariales a partir de los datos generados por los sistemas informáticos. Su potencial radica en la capacidad de descubrir, monitorear y mejorar los procesos a través del análisis de datos reales, lo cual permite una mejor eficiencia operativa. Para posicionarlo en los clientes, es importante destacar sus beneficios en términos de optimización de procesos y reducción de costos. Existen diversas soluciones en el mercado, como Celonis, Disco y UiPath Process Mining, que se utilizan para analizar procesos como la gestión de pedidos, la gestión de reclamaciones y los procesos de producción.

2. Cloud

- Concepto y componentes
- Principales soluciones del mercado

El concepto de cloud, o computación en la nube, se refiere a la entrega de servicios informáticos, como almacenamiento, procesamiento y redes, a través de internet. Los componentes principales de la nube incluyen infraestructura como servicio (IaaS), plataforma como servicio (PaaS) y software como servicio (SaaS). Entre las principales soluciones del mercado se encuentran Microsoft Azure, Amazon Web Services (AWS) y Google Cloud Platform (GCP), que ofrecen servicios escalables y flexibles para satisfacer diferentes necesidades empresariales.

3. MLOps
    - Concepto
    - Herramientas MLOps

MLOps es una práctica que integra el desarrollo de modelos de machine learning con las operaciones de TI para facilitar la implementación y mantenimiento de dichos modelos en entornos productivos. Las herramientas de MLOps, como MLflow, Kubeflow y TFX, ayudan a automatizar el ciclo de vida de los modelos, desde su desarrollo hasta el monitoreo continuo, garantizando la calidad y eficiencia de los modelos en producción.

MÓDULO 10 - Trabajo de Fin de Master en Data Science