**ÍNDICE**

[**1. Introducción al Big Data 1**](#_appxs8otwo35)

[**2. Modelos de Negocio Basados en Datos 1**](#_e3iqg4ifekjn)

[**3. Procesamiento y Análisis de Datos 1**](#_zauiurjfy7tc)

[**4. Tipos de Datos y su Estructuración 2**](#_9fxuws3oalmu)

[**5. Machine Learning y Técnicas de Análisis 2**](#_j6x554ba943h)

[**6. Técnicas y procesos de extracción de la información (IMPORTANTE) 3**](#_67zehuyrlsrs)

[**7. Visualización de Datos 3**](#_s51yrsc25jxr)

[**8. Implicaciones Éticas y Legales 3**](#_53uelbkpp3ig)

[**9. Desafíos y Oportunidades 3**](#_8hytc7588t15)

[**10. Competencias del Científico de Datos 4**](#_zh0cskeu9tyh)

[**Preguntas Tipo Test con Respuestas 4**](#_l7t1q89c7hjq)

[**Respuestas a las Preguntas de Desarrollo 9**](#_d1czisgf9t1)

**Guía de Estudio: Big Data e Inteligencia Artificial**

### **1. Introducción al Big Data**

* **Definición**: Se refiere a grandes volúmenes de datos que requieren nuevas formas de procesamiento.
* **Las 5Vs del Big Data**:
  + **Volumen**: Cantidad masiva de datos generados.
  + **Velocidad**: Rapidez de creación y procesamiento.
  + **Variedad**: Diferentes tipos de datos (estructurados y no estructurados).
  + **Veracidad**: Fiabilidad de los datos.
  + **Valor**: Potencial utilidad de los datos.

### **2. Modelos de Negocio Basados en Datos**

* **Empresas orientadas a datos**: Utilizan Big Data para decisiones precisas.
* **Proceso de transformación**:
  1. Exploración.
  2. Implementación.
  3. Optimización.
  4. Innovación continua.
* **Impacto en decisiones**: Walmart, por ejemplo, usa Big Data para optimizar inventarios.

### **3. Procesamiento y Análisis de Datos**

* **Herramientas**: Apache Hadoop, Apache Spark, Pentaho.
* **Fases**:
  + Recolección, limpieza, transformación, modelado y análisis.
  + Técnicas: minería de datos, machine learning, business intelligence.

### **4. Tipos de Datos y su Estructuración**

**Clasificación de Datos**:

1. **Datos Estructurados**:
   * **Definición**: Organizados en un formato rígido (tablas).
   * **Ejemplos**: Bases de datos SQL, hojas de cálculo.
   * **Aplicaciones**: Informes financieros, CRM.
2. **Datos Semi-Estructurados**:
   * **Definición**: Tienen una estructura flexible con etiquetas.
   * **Ejemplos**: XML, JSON.
   * **Aplicaciones**: Bases de datos NoSQL, correos electrónicos.
3. **Datos No Estructurados**:
   * **Definición**: Sin formato específico, difícil de organizar en tablas.
   * **Ejemplos**: Videos, audios, publicaciones en redes sociales.
   * **Desafíos**: Requieren técnicas avanzadas como NLP (procesamiento de lenguaje natural).
   * **Aplicaciones**: Análisis de opiniones, reconocimiento facial.

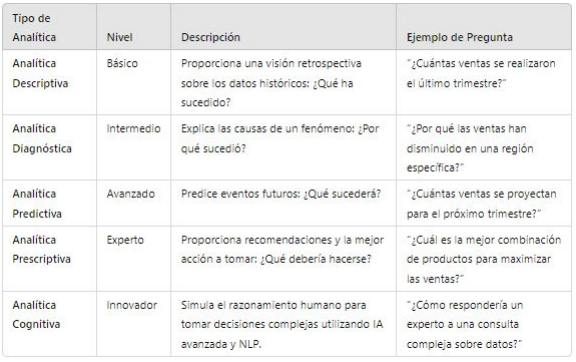
**Procesamiento y Almacenamiento**:

* **Datos estructurados**: Procesados con herramientas convencionales.
* **Semi-estructurados y no estructurados**: Requieren Hadoop, Spark y data lakes.

### **5. Machine Learning y Técnicas de Análisis**

* **Tipos de aprendizaje**:
  + **Supervisado**: Con etiquetas.
  + **No supervisado**: Sin etiquetas.
* **Problemas**:
  + **Clustering**: Agrupación.
  + **Regresión**: Predicción de valores numéricos.
  + **Clasificación**: Asignación de categorías.
* **Aplicaciones**: Análisis de texto, imágenes, audio y video.

### **6. Técnicas y procesos de extracción de la información (IMPORTANTE)**



### **7. Visualización de Datos**

* **Propósito**: Interpretar resultados para la toma de decisiones.
* **Características**: Resumir grandes volúmenes y descubrir patrones.
* **Herramientas**: Tableau, Power BI.

### **8. Implicaciones Éticas y Legales**

* **Protección de datos**: Cumplimiento de regulaciones como GDPR.
* **Desafíos éticos**:
  + Discriminación algorítmica.
  + Transparencia y consentimiento informado.
* **Ejemplos de incumplimiento**: Cambridge Analytica, Equifax.

### **9. Desafíos y Oportunidades**

* **Técnicos**:
  + Escalabilidad, seguridad y velocidad de procesamiento.
* **Oportunidades**:
  + Personalización de servicios, análisis predictivo y aplicaciones en smart cities.

### **10. Competencias del Científico de Datos**

* **Habilidades**:
  + IA, machine learning, estadísticas.
  + Programación y comunicación.
* **Capacidades**:
  + Evaluar resultados, trabajar en equipos multidisciplinares.

**—------------------------------------------------------------------------------------------------------------------------**

### **Preguntas Tipo Test con Respuestas**

1. **¿Qué característica de Big Data se refiere a la cantidad de datos generados?**
   * a) Velocidad
   * **b) Volumen**
   * c) Variedad
   * d) Veracidad
2. **¿Cuál de las siguientes opciones es un ejemplo de datos estructurados?**
   * a) Videos
   * **b) Hojas de cálculo**
   * c) Publicaciones en redes sociales
   * d) Correos electrónicos
3. **¿Qué herramienta se utiliza comúnmente para procesar grandes volúmenes de datos?**
   * a) Excel
   * **b) Apache Hadoop**
   * c) Word
   * d) Google Docs
4. **¿Cuál de las siguientes no es una de las 5Vs de Big Data?**
   * a) Volumen
   * b) Variedad
   * **c) Visibilidad**
   * d) Veracidad
5. **¿Qué tipo de datos se caracterizan por tener etiquetas pero no un formato rígido?**
   * a) Estructurados
   * b) No estructurados
   * **c) Semi-estructurados**
   * d) Agregados
6. **¿Qué técnica de aprendizaje automático se usa para predecir un valor numérico?**
   * a) Clustering
   * b) Clasificación
   * **c) Regresión**
   * d) Análisis de sentimientos
7. **Un ejemplo de aplicación de análisis de texto es:**
   * a) Clasificación de imágenes
   * b) Reconocimiento facial
   * **c) Análisis de sentimientos**
   * d) Creación de resúmenes automáticos de video
8. **¿Cuál de las siguientes es una herramienta de visualización de datos?**
   * a) Spark
   * b) Hadoop
   * **c) Tableau**
   * d) TensorFlow
9. **Los datos que carecen de un formato predefinido se denominan:**
   * a) Datos estructurados
   * b) Datos semi-estructurados
   * **c) Datos no estructurados**
   * d) Datos agregados
10. **El uso de Big Data en salud incluye:**
    * a) Gestión de inventarios
    * b) Personalización de anuncios
    * **c) Detección temprana de enfermedades**
    * d) Análisis de opiniones en redes sociales
11. **¿Qué tipo de análisis proporciona recomendaciones de acción?**
    * a) Descriptivo
    * b) Predictivo
    * **c) Prescriptivo**
    * d) Diagnóstico
12. **Una base de datos que almacena datos con etiquetas flexibles, como JSON, es:**
    * a) Relacional
    * **b) Semi-estructurada**
    * c) No estructurada
    * d) Estructurada
13. **El análisis de datos en tiempo real es más común en:**
    * a) Revisión de historiales pasados
    * **b) Monitoreo industrial**
    * c) Archivos históricos
    * d) Libros de contabilidad
14. **¿Qué aspecto del Big Data aborda la precisión y confiabilidad de los datos?**
    * a) Volumen
    * **b) Veracidad**
    * c) Variedad
    * d) Velocidad
15. **¿Cuál es un desafío ético en el uso de Big Data?**
    * a) Personalización de servicios
    * **b) Discriminación algorítmica**
    * c) Análisis de ventas
    * d) Detección de fraudes
16. **El análisis de datos usando técnicas estadísticas, minería de datos y machine learning es parte de:**
    * a) Visualización
    * b) Limpieza de datos
    * **c) Modelado de datos**
    * d) Análisis de datos
17. **La etapa inicial en el procesamiento de datos es:**
    * a) Análisis
    * b) Limpieza
    * **c) Recolección**
    * d) Interpretación
18. **La detección de outliers es una técnica que se aplica en:**
    * **a) Limpieza de datos**
    * b) Visualización
    * c) Análisis descriptivo
    * d) Codificación de datos
19. **El aprendizaje supervisado se caracteriza por:**
    * a) No tener etiquetas
    * **b) Tener etiquetas para entrenamiento**
    * c) Usar redes neuronales únicamente
    * d) Ser siempre en tiempo real
20. **Un ejemplo de datos generados por máquinas es:**
    * a) Respuestas de encuestas
    * b) Comentarios en redes sociales
    * **c) Sensores IoT**
    * d) Entradas de blogs
21. **Una técnica que agrupa datos similares sin etiquetas es:**
    * a) Clasificación
    * **b) Clustering**
    * c) Regresión
    * d) Prescripción
22. **La visualización de datos ayuda principalmente a:**
    * a) Mejorar la velocidad de procesamiento
    * **b) Facilitar la interpretación y toma de decisiones**
    * c) Almacenar datos
    * d) Codificar datos
23. **Los cuadros de mando se utilizan para:**
    * a) Procesar datos
    * **b) Monitorear KPIs y obtener una visión en tiempo real**
    * c) Extraer datos de la web
    * d) Limpiar datos
24. **El modelo conceptual de datos se encarga de:**
    * a) Implementar la base de datos
    * **b) Definir qué contiene el sistema y cómo se relacionan las entidades**
    * c) Almacenar los datos
    * d) Definir tipos de datos de atributos
25. **Los datos semi-estructurados pueden ser representados en:**
    * a) Tablas SQL
    * **b) JSON o XML**
    * c) Archivos de video
    * d) Hojas de cálculo
26. **El GDPR es una regulación que:**
    * a) Facilita la recolección de datos
    * **b) Regula la privacidad de los datos personales**
    * c) Obliga a usar técnicas de análisis de datos
    * d) Controla la creación de datos estructurados
27. **Un ejemplo de aprendizaje no supervisado es:**
    * a) Clasificación de correos como spam/no spam
    * b) Predicción del precio de una acción
    * **c) Clustering de clientes según preferencias**
    * d) Detección de fraude
28. **El uso de Big Data para análisis de comportamiento de clientes es común en:**
    * a) Finanzas
    * **b) Publicidad**
    * c) Investigación médica
    * d) Energía
29. **Una desventaja de los datos no estructurados es:**
    * a) Alta facilidad de procesamiento
    * **b) Análisis complejo y costoso**
    * c) Alta organización
    * d) Fácil uso en bases de datos relacionales
30. **El procesamiento de datos no estructurados a menudo requiere:**
    * a) Excel
    * b) Procesos manuales simples
    * **c) Técnicas avanzadas como NLP y machine learning**
    * d) Solo bases de datos relacionales

**—------------------------------------------------------------------------------------------------------------------------**

### **Respuestas a las Preguntas de Desarrollo**

1. **Explica las 5Vs de Big Data y cómo influyen en el manejo de grandes volúmenes de información.**
   * **Volumen**: Se refiere a la cantidad masiva de datos generados constantemente. Manejar grandes volúmenes implica la necesidad de tecnologías avanzadas de almacenamiento y procesamiento, como clústeres distribuidos y bases de datos escalables.
   * **Velocidad**: Indica la rapidez con la que los datos se generan y procesan. En aplicaciones como transacciones financieras o monitoreo en tiempo real, la capacidad de procesar datos rápidamente es esencial para la toma de decisiones.
   * **Variedad**: Representa la diversidad de tipos de datos (estructurados, semi-estructurados y no estructurados). Esto requiere herramientas versátiles capaces de integrar y analizar diferentes tipos de datos, como bases de datos NoSQL y sistemas de procesamiento distribuido.
   * **Veracidad**: Se relaciona con la calidad y fiabilidad de los datos. Datos inexactos o ruidosos pueden sesgar los análisis, por lo que es vital implementar técnicas de limpieza y validación de datos.
   * **Valor**: El potencial de los datos para proporcionar información útil. No solo es importante tener grandes volúmenes de datos, sino también saber extraer información relevante para obtener insights valiosos.
2. **Describe las diferencias entre datos estructurados, semi-estructurados y no estructurados, y da ejemplos de cada uno.**
   * **Datos estructurados**: Siguen un formato predefinido, como tablas con filas y columnas. Ejemplos: bases de datos SQL, hojas de cálculo. Son fáciles de almacenar y procesar con herramientas tradicionales.
   * **Datos semi-estructurados**: Tienen una estructura flexible con etiquetas, pero no un formato fijo. Ejemplos: documentos JSON, XML. Permiten más flexibilidad que los datos estructurados y se utilizan en aplicaciones web y bases de datos NoSQL.
   * **Datos no estructurados**: Carecen de un formato específico y no se pueden organizar fácilmente en tablas. Ejemplos: videos, imágenes, audios. Su análisis requiere técnicas avanzadas como NLP y machine learning.
3. **¿Cuáles son las etapas principales en el procesamiento de datos y qué técnicas se utilizan en cada una?**
   * **Recolección**: Obtención de datos desde diferentes fuentes. Puede incluir sensores, registros, encuestas y bases de datos externas.
   * **Limpieza**: Corrección de errores, eliminación de datos duplicados, manejo de valores faltantes. Se utilizan herramientas como Trifacta y Pentaho.
   * **Transformación**: Convertir datos a un formato adecuado para el análisis. Técnicas de normalización y codificación de categorías se aplican en esta etapa.
   * **Modelado y Análisis**: Aplicación de algoritmos de machine learning, minería de datos y análisis estadístico.
   * **Interpretación**: Traducción de los resultados analíticos en insights prácticos y su presentación en reportes o cuadros de mando.
4. **Discute los desafíos éticos y legales en el uso de Big Data, mencionando ejemplos relevantes.**
   * **Privacidad y protección de datos**: El uso masivo de datos requiere cumplir con normativas como GDPR para proteger la información personal de los usuarios. Un ejemplo de incumplimiento fue el caso de Cambridge Analytica, donde se usaron datos de usuarios sin consentimiento.
   * **Discriminación algorítmica**: Algoritmos pueden perpetuar sesgos si se entrenan con datos históricos sesgados. Esto ocurrió en sistemas de contratación automatizada que penalizaban a ciertos grupos por sesgos en los datos.
   * **Transparencia**: Las empresas deben ser claras sobre cómo usan los datos y obtener el consentimiento de los usuarios.
5. **Explica cómo se aplican las técnicas de machine learning en el análisis de datos y menciona al menos tres ejemplos de aplicaciones.**
   * **Aplicación de técnicas**: El machine learning permite a las máquinas aprender de los datos para hacer predicciones o identificar patrones sin ser programadas explícitamente. Se utilizan algoritmos supervisados (como regresión y clasificación) y no supervisados (como clustering).
   * **Ejemplos de aplicaciones**:
     + Detección de fraudes en transacciones financieras.
     + Análisis de sentimientos en redes sociales para evaluar la opinión pública sobre productos o servicios.
     + Reconocimiento de imágenes en sistemas médicos para detectar enfermedades en imágenes de rayos X o resonancias.