UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGIAS DE LA INFORMACIÓN

EXTRACCION DE CONOCIMIENTOS EN BASES DE DATOS

DIAGNOSTICO

IDGS91N

PRESENTA:

SEBASTIÁN ACOSTA ORTIZ

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 04 de septiembre de 2025

Realizar documento digital donde explique los siguientes conceptos:

Inteligencia artificial y sus aplicaciones

La inteligencia artificial es una tecnología que permite a las computadoras y máquinas simular el aprendizaje humano, la comprensión, la resolución de problemas, la toma de decisiones, la creatividad y la autonomía.

- Ver e identificar objetos
- Entender y responder al lenguaje humano
- Aprender de nueva información y experiencia
- Hacer recomendaciones detalladas a usuarios y expertos
- Actuar de manera independiente, reemplazando la necesidad de inteligencia o intervención humana

Minería de datos

La minería de datos es el uso del machine learning y el análisis estadístico para descubrir patrones y otra información valiosa de grandes conjuntos de datos.

Big data

Big data se refiere a conjuntos de datos masivos y complejos que los sistemas tradicionales de gestión de datos no pueden manejar. Cuando se recopilan, gestionan y analizan adecuadamente, los big data pueden ayudar a las organizaciones a descubrir nuevos insights y tomar mejores decisiones empresariales.

Metodologías para el análisis de datos

El análisis de datos es el proceso de recopilar información con el propósito de estudiarla para generar conocimientos. El análisis de alto nivel es realizado principalmente por científicos de datos, pero las plataformas más recientes de análisis de datos cuentan con herramientas, como consultas basadas en procesamiento de lenguaje natural e información automatizada, que permiten a los usuarios de negocio explorar conjuntos de datos.

1. Análisis de datos predictivo (Predictive Data Analytics)

Es probablemente el tipo más utilizado de análisis de datos.

- Se emplea para identificar tendencias, correlaciones y causalidades.
- Se divide en modelado predictivo y modelado estadístico, que trabajan de la mano.

Ejemplo: una campaña de playeras en Facebook podría usar análisis predictivo para ver cómo se relacionan las tasas de conversión con el área geográfica, nivel de ingresos e intereses de la audiencia. Luego, con el modelado predictivo, se compararían diferentes segmentos de público para estimar ingresos posibles de cada demográfico.

2. Análisis de datos prescriptivo (Prescriptive Data Analytics)

Es donde la **inteligencia artificial y el big data** se combinan para **predecir resultados y recomendar acciones a tomar**.

- Se divide en **optimización** y **pruebas aleatorias (random testing)**.
- Gracias al **machine learning**, permite responder preguntas como:
    - "¿Qué pasa si probamos este eslogan?"
    - "¿Cuál es el mejor color de camiseta para un público mayor?"

3. Análisis de datos diagnóstico (Diagnostic Data Analytics)

Analiza datos del pasado para **entender causas y efectos**.

- Utiliza técnicas como: *drill down* (análisis a detalle), *data discovery* (descubrimiento de datos), minería de datos y correlaciones.
- Responde a la pregunta: **¿por qué ocurrió algo?**
- Se divide en:
    - **Descubrimiento y alertas** → notifica posibles problemas antes de que sucedan (ejemplo: menos horas trabajadas podrían anticipar menos ventas).
    - **Consultas y drill down** → permiten profundizar en los reportes (ejemplo: investigar por qué un vendedor cerró menos tratos un mes, revelando que estuvo de vacaciones dos semanas).

4. Análisis de datos descriptivo (Descriptive Data Analytics)

Es la base de los informes y dashboards de BI.

- Responde preguntas como: **¿cuántos, cuándo, dónde y qué?**
- Se divide en:
    - **Reportes predefinidos (canned reports)** → diseñados previamente con métricas específicas (ejemplo: reporte mensual de agencia de publicidad sobre campañas de playeras).
    - **Reportes ad hoc** → creados en el momento para responder preguntas concretas (ejemplo: analizar audiencia de redes sociales en cierta ciudad y hora del día para conocer mejor al público).

Conteste el siguiente cuestionario en el mismo documento (si no sabe una respuesta no la busque en la red):

Examen Diagnóstico General

Objetivo: Evaluar conocimientos previos sobre fundamentos de análisis de datos, estadística básica y uso de herramientas.

Formato:

- 20 preguntas: 10 de opción múltiple, 5 de verdadero/falso, 5 de respuesta corta.

Sección 1: Opción Múltiple (10 pts; 1 pto c/u)

1. ¿Cuál de estos dominios se enfoca en extraer patrones de datos históricos para descubrir relaciones ocultas?
   A) Inteligencia Artificial
   B) Data Mining
   C) Big Data
   D) DevOps
2. ¿Qué librería de Python se usa habitualmente para manipulación de tablas (DataFrames)?
   A) matplotlib
   B) pandas
   C) numpy
   D) seaborn
3. En un proceso ETL, la "T" corresponde a:
   A) Testing
   B) Transfer
   C) Transform
   D) Transmission
4. ¿Cuál de estas no es una fase de CRISP-DM?
   A) Comprensión del negocio
   B) Preparación de datos
   C) Deploy en producción
   D) Evaluación del modelo
5. El método batch processing se caracteriza por:
   A) Procesar datos a medida que llegan
   B) Procesar grandes volúmenes en intervalos programados
   C) Procesar datos en la nube exclusivamente
   D) No requiere transformación de datos
6. En un árbol de decisión, ¿qué indica un "nodo hoja"?
   A) Un punto de decisión intermedio

B) La variable de entrada
C) Una predicción final
D) Una rama cancelada

7. ¿Qué métrica combina precision y recall en un solo valor?
   A) Accuracy
   B) F1-score
   C) AUC-ROC
   D) MSE
8. Un vectorizador TF-IDF convierte texto en:
   A) Secuencias de audio
   B) Vectores numéricos
   C) Imágenes
   D) Tablas SQL
9. En SQL, para eliminar duplicados se usa típicamente:
   A) DELETE DUPLICATE
   B) DISTINCT
   C) UNIQUE
   D) REMOVE
10. Un diagrama de Gantt se usa para:
    A) Visualizar clusters en 2D
    B) Mostrar la arquitectura de un DW
    C) Planificar actividades en el tiempo
    D) Evaluar la precisión de un modelo

Sección 2: Verdadero / Falso (5 pts; 1 pto c/u)

11. V/F: Big Data siempre implica inteligencia artificial. F
12. V/F: Un modelo LSTM es útil para series de tiempo. V
13. V/F: En clustering jerárquico, los clusters nunca se fusionan. F
14. V/F: La matriz de confusión solo aplica en problemas de clasificación. V
15. V/F: Un dashboard interactivo permite filtrar datos al vuelo. V

Sección 3: Respuesta Corta (5 pts; hasta 2 líneas cada una)

16. Define brevemente qué es un Data Warehouse.

    No c

17. Menciona dos diferencias clave entre batch y streaming.

    No c

18. ¿Para qué usamos PCA antes de hacer clustering?

    No sepo

19. Nombra un caso de uso real de Machine Learning en industria.

    Sabra

20. ¿Qué representa el “error cuadrático medio” (MSE)?

    No se