Instituto Tecnológico Beltrán

Carrera: Tecnicatura Superior en Ciencia de Datos e IA

Asignatura: Modelizado de Minería de Datos

GUÍA DE ESTUDIO: Modelizado de Minería de Datos

Objetivo: Desarrollar competencias analíticas, técnicas y visuales aplicadas al modelado de datos, orientadas a la resolución de problemas de negocio y la toma de decisiones basada en datos.

1 La Ciencia de Datos: Definición y Conceptos Clave

Definición

La **Ciencia de Datos** es un enfoque interdisciplinario que combina conocimientos de estadísticas, informática, matemáticas y comprensión del negocio para **extraer valor a partir de grandes volúmenes de datos**, transformándolos en **conocimiento útil para la toma de decisiones**.

Componentes clave:

- **Datos:** estructurados, semi-estructurados y no estructurados.
- **Análisis exploratorio:** estadísticas descriptivas, visualización.
- **Machine Learning:**
    - modelos predictivos: Es una herramienta estadística y matemática que analiza datos históricos para identificar patrones y tendencias, con el fin de predecir resultados futuros. Estos modelos son fundamentales para la toma de decisiones basadas en datos, ya que permiten anticipar comportamientos, riesgos y oportunidades en diversas áreas como finanzas, marketing y salud, ayudando a las organizaciones a planificar estratégicamente.
    - modelos prescriptivos: Las soluciones de análisis **prescriptivos** implican la creación de **modelos** matemáticos y algoritmos de optimización para recomendar decisiones empresariales que conduzcan a los mejores resultados empresariales posibles.
- **Infraestructura:** bases de datos, big data, cloudcomputing.
- **Narrativa con datos:** storytelling visual. Es el arte de contar historias y comunicar mensajes utilizando principalmente elementos visuales como imágenes, videos, infografías y animaciones. Su objetivo es captar la atención, transmitir información compleja de forma clara y memorable, evocar emociones y crear una conexión más profunda con la audiencia, logrando un impacto duradero que va más allá de lo que el texto por sí solo puede lograr.

Herramientas comunes:

- **Lenguajes:** Python, R, SQL.
- **Plataformas:** Jupyter, Google Colab, Power BI, Tableau.
- **Librerías:** Pandas, Scikit-learn, Matplotlib, Seaborn.

Rol del Científico de Datos:

- Recopilar, limpiar y analizar datos.
- Formular preguntas de negocio.
- Comunicar resultados y tomar decisiones informadas.

2 Problemáticas Específicas Vinculadas a Negocios

Ejemplos Reales:

- **Bancos:** detección de fraudes.
- **Retail (Venta al por menor o comercio minorista):** predicción de demanda.
- **Telecomunicaciones:** fuga de clientes (churn).
- **Marketing:** segmentación de clientes.
- **Logística:** optimización de rutas.

Enfoque desde la minería de datos:

Planteo del problema → Recolección de datos → Limpieza → Modelado → Evaluación → Comunicación de resultados.

En minería de datos, el foco no es solo "ver qué pasó", sino predecir o clasificar eventos futuros**.**

3 Inteligencia de Negocios vs. Análisis Predictivo

Aspecto	**Inteligencia de Negocios (BI)**	**Análisis Predictivo**
Enfoque	Descriptivo (qué pasó)	Predictivo (qué podría pasar)
Tipo de análisis	Reportes, dashboards	Modelos estadísticos y ML
Herramientas	Power BI, Tableau, Excel	Python, R, Scikit-learn
Datos usados	Históricos agregados	Datos históricos + actuales
Objetivo	Comprender el pasado y optimizar procesos	Anticipar el futuro y tomar decisiones preventivas

Conexión con minería de datos:

El análisis predictivo es una **parte central** del modelizado en minería de datos. Se busca construir modelos que generalicen patrones y los apliquen a nuevos datos.

4 Capacidad Analítica para el Manejo de Información en Negocios

¿Qué implica ser analítico?

- Formular preguntas clave.
- Interpretar patrones en los datos.
- Tomar decisiones basadas en evidencia, no en intuición.
- Evaluar riesgos y oportunidades.

Aplicaciones prácticas:

- Score de riesgo crediticio.
- Recomendaciones personalizadas (ej. Netflix, Spotify).
- Optimización de precios dinámicos.

5 Visualización y Transformación de Información para Decisiones

Transformación de datos:

- **Limpieza:** eliminar nulos, duplicados.
- **Formateo:** normalización, categorización.
- **Enriquecimiento:** agregar nuevas variables derivadas.

"Garbage in, garbageout": la calidad del modelo depende de la calidad de los datos.

Visualización como base innovadora:

- **Ayuda a detectar** patrones ocultos**.**
- **Facilita la comunicación entre técnicos y no técnicos.**
- **Permite una rápida** toma de decisiones gerenciales**.**

6 Visualización como Ahorro de Tiempo en las Organizaciones

Ventajas clave:

- Automatización de reportes**.**
- Acceso en tiempo real a indicadores clave (KPIs).
- Menor tiempo de análisis manual.
- Detección temprana de problemas.

Ejemplo:

Un dashboard de ventas permite al gerente visualizar en segundos qué producto está bajando su rendimiento y actuar de inmediato, sin revisar hojas de cálculo extensas.

COMPETENCIAS QUE DEBÉS DESARROLLAR

Competencia	Nivel Esperado
Comprender el ciclo de vida de un proyecto de ciencia de datos	Intermedio
Modelar datos usando algoritmos supervisados y no supervisados	Intermedio
Aplicar técnicas de preprocesamiento de datos	Avanzado
Evaluar modelos con métricas apropiadas (accuracy, precision, recall, etc.)	Intermedio
Visualizar datos con herramientas BI y Python	Avanzado
Comunicar hallazgos de forma clara y visual	Avanzado

RECURSOS SUGERIDOS

Libros:

- *Data Sciencefor Business* – Provost&Fawcett
- *Hands-On Machine LearningwithScikit-Learn, Keras&TensorFlow* – AurélienGéron
- *Storytellingwith Data* – Cole NussbaumerKnaflic

Cursos gratuitos:

- Google Data Analytics (Coursera)
- Machine Learning con Python (freeCodeCamp)
- Minería de datos (Universidad Politécnica de Valencia – edX)

EJERCICIO FINAL DE REPASO

Caso: Una cadena de gimnasios quiere predecir qué clientes tienen mayor riesgo de cancelar su suscripción.

Pasos sugeridos:

- Importar y explorar datos de clientes (edad, frecuencia de asistencia, pagos, etc.).
- Transformar y limpiar datos.
- Crear modelo de clasificación (ej. árbol de decisión).
- Evaluar modelo.
- Visualizar resultados y entregar recomendaciones para reducir la pérdida de clientes.

Ejercicio práctico guiado para aplicar los contenidos clave de **Modelizado de Minería de Datos**, incluyendo:

- Preprocesamiento
- Visualización
- Modelado predictivo (clasificación)
- Toma de decisiones
- Representación visual

Ejercicio Práctico: Análisis de Recompra en una Campaña de Marketing

Objetivo:

Usar técnicas de modelizado y visualización de datos para **predecir si un cliente realizará una recompra** luego de haber recibido una promoción.

Dataset Sugerido:

Usar el archivo Mini_Proyecto_Clientes_Promociones.xlsx, con estos campos o crearlo a mano:

Cliente_ID	Género	Edad	Recibió_Promo	Monto_Promo	Recompra	Total_Compras	Ingreso_Mensual
1	F	23	Sí	500	Sí	2	30000
…	…	…	…	…	…	…	…

PASOS PARA REALIZAR EL EJERCICIO

1. Comprensión del problema

Contexto: Un área de marketing quiere saber si los clientes que recibieron promociones **volverán a comprar** (recompra), y qué variables influyen más en esa decisión.

Preguntas guía:

- ¿Recibir una promoción realmente influye en la recompra?
- ¿Importa el monto de la promoción?
- ¿Influye la edad o el ingreso?

2. Carga y exploración del dataset

- Cargar los datos en Python (o Excel).
- Verificar datos faltantes, valores atípicos, etc.
- Analizar la distribución de las variables.

Python:

```
import pandas as pd

df = pd.read_excel("Mini_Proyecto_Clientes_Promociones.xlsx")

df.info()

df.describe()
```

3. Transformación y codificación

- Convertir variables categóricas a numéricas.
- Crear nuevas variables si es necesario.

Python:

```
df['Genero'] = df['Genero'].map({'F': 0, 'M': 1})

df['Recibio_Promo'] = df['Recibio_Promo'].map({'Sí': 1, 'No': 0})

df['Recompra'] = df['Recompra'].map({'Sí': 1, 'No': 0})
```

4. Visualización de relaciones clave

- Recompra vs. Monto de promoción
- Recompra vs. Ingreso mensual
- Distribución por género y edad

Python:

```
importseaborn as sns

importmatplotlib.pyplot as plt



sns.boxplot(x="Recompra", y="Monto_Promocion", data=df)

plt.title("Recompra según el Monto Promocional")

plt.show()
```

5. Modelado Predictivo: Clasificación

- Elegir un modelo (por ejemplo, árbol de decisión).
- Entrenar y evaluar el modelo.

Python:

fromsklearn.model_selectionimporttrain_test_split

fromsklearn.treeimportDecisionTreeClassifier

fromsklearn.metricsimportclassification_report, confusion_matrix

X = df.drop(['Cliente_ID', 'Recompra'], axis=1)

y = df['Recompra']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo = DecisionTreeClassifier()

modelo.fit(X_train, y_train)

y_pred = modelo.predict(X_test)

print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))

6. Toma de decisiones basada en visualización

- Crear un **dashboard simple** (puede ser en Excel, Power BI o Seaborn) que ayude al equipo de marketing a:
    - Ver qué perfiles de cliente más probablemente recompran.
    - Detectar si invertir en promociones es rentable.

7. Preguntas para discusión o informe

- ¿Qué variables son más importantes para predecir la recompra?
- ¿Se justificaría invertir más en promociones para ciertos grupos?
- ¿Cómo podrías mejorar el modelo?
- ¿Qué tipo de visualización fue más útil para comunicar tus hallazgos?

Entregables (opcional)

- Código en Jupyter Notebook o Python Script
- Visualizaciones relevantes
- Informe (1-2 páginas) con:
    - Introducción al problema
    - Descripción del análisis y modelo
    - Resultados principales
    - Conclusiones y recomendaciones

ANEXOS:

Anexo 1:

Código en Python para generar el Excel:

```
import pandas as pd


# Crear dataset simulado
data = {
    "Cliente_ID": range(1, 21),
    "Genero": ["F", "M"] * 10,
    "Edad": [23, 34, 45, 29, 31, 38, 27, 50, 40, 36, 25, 33, 46, 28, 39, 42, 30, 48, 35, 37],
    "Recibio_Promo": ["Sí", "No", "Sí", "Sí", "No", "Sí", "No", "Sí", "No", "Sí", "No", "Sí", "Sí", "No", "No", "Sí", "No", "Sí", "No", "Sí"],
    "Monto_Promocion": [500, 0, 700, 300, 0, 600, 0, 800, 0, 450, 0, 620, 710, 0, 0, 480, 0, 750, 0, 520],
    "Recompra": ["Sí", "No", "Sí", "No", "No", "Sí", "No", "Sí", "No", "Sí", "No", "No", "Sí", "No", "No", "Sí", "No", "Sí", "No", "Sí"],
    "Total_Compras": [2, 1, 3, 1, 1, 4, 1, 5, 1, 3, 1, 2, 4, 1, 1, 3, 1, 5, 1, 3],
    "Ingreso_Mensual": [30000, 45000, 40000, 28000, 32000, 50000, 31000, 60000, 29000, 37000, 31000, 34000, 47000, 30000, 29000, 43000, 33000, 55000, 30000, 41000]
}
df = pd.DataFrame(data)


# Guardar en archivo Excel
df.to_excel("Mini_Proyecto_Clientes_Promociones.xlsx", index=False)
print("Archivo Excel generado correctamente.")
```

¿Qué hacer con este Excel?

NOTA: Usarlo como base para el siguiente **mini proyecto guiado**

Anexo 2:

Generación en Python de un archivo Word con el ejercicio práctico:

```
fromdocximportDocument

fromdocx.enum.textimport WD_ALIGN_PARAGRAPH


# Crear documento Word

doc = Document()


# Título

title = doc.add_heading("Ejercicio Práctico – Modelizado de Minería de Datos", 0)

title.alignment = WD_ALIGN_PARAGRAPH.CENTER


# Introducción

doc.add_paragraph("Este ejercicio está diseñado para aplicar los contenidos clave de la asignatura 'Modelizado de Minería de Datos', incluyendo preprocesamiento, visualización, modelado predictivo, y toma de decisiones basada en datos.")


# Secciones del ejercicio

sections = {

    "Objetivo": [

        "Predecir si un cliente realizará una recompra después de recibir una promoción, utilizando técnicas de modelado y visualización de datos."

    ],

    "Dataset Sugerido": [

        "Usar el archivo 'Mini_Proyecto_Clientes_Promociones.xlsx' con los siguientes campos:",

        "- Cliente_ID, Género, Edad, Recibió_Promo, Monto_Promocion, Recompra, Total_Compras, Ingreso_Mensual"
```

```
],

"Pasos para realizar el ejercicio": [],

"1. Comprensión del Problema": [

    "Contexto: Un área de marketing quiere saber si los clientes que recibieron promociones volverán a comprar.",

    "Preguntas guía:",

    "- ¿Recibir una promoción influye en la recompra?",

    "- ¿Importa el monto?",

    "- ¿Influyen edad o ingreso?"

],

"2. Carga y Exploración del Dataset": [

    "- Cargar los datos en Python o Excel.",

    "- Verificar datos faltantes, valores extremos, etc.",

    "- Analizar la distribución de las variables."

],

"3. Transformación y Codificación": [

    "- Convertir variables categóricas a numéricas.",

    "- Crear nuevas variables si es necesario."

],

"4. Visualización de Relaciones Clave": [

    "- Recompra vs. Monto de promoción.",

    "- Recompra vs. Ingreso mensual.",

    "- Distribución por género y edad."

],

"5. Modelado Predictivo – Clasificación": [

    "- Modelo sugerido: Árbol de Decisión.",
```

```
        "- Entrenar, predecir y evaluar el modelo."
    ],
    "6. Toma de Decisiones Basada en Visualización": [
        "- Crear un dashboard simple con insights útiles para el equipo de marketing.",
        "- Identificar perfiles con mayor probabilidad de recompra."
    ],
    "7. Preguntas para Discusión o Informe": [
        "- ¿Qué variables son más importantes para predecir la recompra?",
        "- ¿Qué tipo de cliente conviene incentivar?",
        "- ¿Cómo comunicarías tus hallazgos a alguien sin conocimientos técnicos?"
    ],
    "Entregables (opcional)": [
        "- Código en Jupyter Notebook o Python Script.",
        "- Visualizaciones relevantes.",
        "- Informe con introducción, análisis, resultados, y recomendaciones."
    ],
    "Rúbrica de Evaluación (opcional)": [
        "- Análisis Exploratorio: 20%",
        "- Preprocesamiento de datos: 15%",
        "- Modelado predictivo: 25%",
        "- Visualización: 15%",
        "- Interpretación y comunicación de resultados: 25%"
    ]
}
```

```
# Agregar secciones al documento
forheading, content in sections.items():
doc.add_heading(heading, level=1)
forparagraph in content:
doc.add_paragraph(paragraph)

# Guardar el archivo
doc.save("Ejercicio_Practico_Modelizado_Mineria_Datos.docx")
print("Archivo generado correctamente.")
```