



Tecnológico de Monterrey

Procesamiento de lenguaje natural (Equipo 41)

10.3 Actividad: Actividad: modelos LLM y IA en tu lugar de trabajo.

Integrantes:

Daniela Sánchez	A01749492
Esteban Sánchez	A01740631
Francisco Rupit	A01795106
Bruno Sánchez García	A01378960
Jose Antonio Saenz	A01793884

Profesores:

Luis Eduardo Falcón Morales

Yetnalezi Quintas Ruiz

25 de Junio de 2024

Proyecto 1 Detección temprana de problemas de calidad mediante el uso de códigos de diagnóstico (DTC)

Modelo: ChatGPT 4

Empresa: Ford Motor Company SA de CV

Giro: Automotriz

Area: Connected Vehicles Software/Early warning response team

Descripción de Actividad: Encontrar la correlación que hay entre los problemas que están teniendo los vehículos a través de la data enviada por los mismos con las garantías que están apareciendo en el campo. Entender de todos los DTCs que tiene un vehículo, cuales son los que fueron trabajados en cada una de las garantías y corroborar que las acciones técnicas que se les dan al equipo de servicio están teniendo el efecto deseado.

Interacción con: áreas de desarrollo de producto encargados del desarrollo de partes y software, IT encargados de mantener actualizadas las tablas de donde proviene la información, manufactura que se encargan de la implementación de las acciones correctivas y Project Management

Ingeniería de Instrucciones

Persona:

Actúa como un científico de datos.

GPT4:

¡Claro! Como científico de datos, puedo ayudarte con diversas tareas relacionadas con el análisis y la interpretación de datos. Esto puede incluir la limpieza de datos, el análisis estadístico, la visualización de datos, la construcción de modelos predictivos, y la generación de informes.

¿Cómo puedo asistirte hoy? ¿Tienes algún conjunto de datos específico con el que necesitas ayuda, o hay alguna pregunta o tarea en particular que te gustaría que abordara?

Persona:

Hazme las preguntas necesarias para que puedas darme ideas de como implementar algún modelo que me ayude a encontrar la relación que hay entre los DTCs que está generando un vehículo y las garantías que han aparecido para ese vehículo.

GPT4:

Para ayudarte a implementar un modelo que encuentre la relación entre los DTCs (Códigos de Diagnóstico de Problemas) generados por un vehículo y las garantías que han aparecido para ese vehículo, necesitaré recopilar cierta información clave. Aquí tienes algunas preguntas que me ayudarán a entender mejor tu situación y necesidades:

1. Datos Disponibles:

- ¿Tienes acceso a un conjunto de datos que contenga los DTCs generados por los vehículos? Si es así, ¿qué variables incluye este conjunto de datos (por ejemplo, ID del vehículo, fecha, DTC, etc.)?
- ¿Tienes acceso a un conjunto de datos que contenga información sobre las garantías? Si es así, ¿qué variables incluye este conjunto de datos (por ejemplo, ID del vehículo, fecha de la garantía, tipo de garantía, etc.)?

- ¿Existen otros conjuntos de datos relacionados que puedan ser útiles, como historial de mantenimiento, datos de sensores del vehículo, etc.?

2. Detalles de los Datos:

- ¿Cuántos registros hay en cada conjunto de datos (aproximadamente)?
- ¿Los conjuntos de datos tienen identificadores comunes para vincular los DTCs y las garantías (por ejemplo, ID del vehículo)?
- ¿Hay alguna información temporal disponible, como fechas de los DTCs y fechas de las garantías?

3. Objetivos del Modelo:

- ¿Cuál es el objetivo principal del modelo? ¿Es predecir la aparición de garantías basándose en los DTCs generados, identificar patrones comunes entre DTCs y garantías, o algo más?
- ¿Hay algún tipo de garantía específica en la que estés más interesado?

4. Características del Vehículo:

- ¿Qué tipo de vehículos están siendo analizados (por ejemplo, coches, camiones, etc.)?
- ¿Existen características adicionales de los vehículos que puedan ser relevantes (por ejemplo, marca, modelo, año de fabricación, etc.)?

5. Metodología y Resultados Esperados:

- ¿Tienes alguna preferencia o restricción en cuanto a la metodología de modelado (por ejemplo, modelos de aprendizaje supervisado, no supervisado, etc.)?
- ¿Qué métricas de evaluación son importantes para ti (por ejemplo, precisión, recall, F1-score, etc.)?
- ¿Cómo planeas utilizar los resultados del modelo?

6. Recursos y Limitaciones:

- ¿Qué herramientas y tecnologías tienes disponibles para el análisis (por ejemplo, Python, R, SQL, herramientas de visualización, etc.)?
- ¿Existen limitaciones en cuanto a tiempo, recursos computacionales, o acceso a datos que debamos considerar?

Responder a estas preguntas me permitirá ofrecerte ideas más concretas y detalladas sobre cómo implementar un modelo que pueda ayudarte a encontrar la relación entre los DTCs y las garantías en los vehículos.

Persona:

- 1.- Si hay un conjunto de datos que contienen los DTCs. Este conjunto contiene el ID del vehículo, DTC, fecha cuando sucedió el DTC, Odómetro, planta donde se construyó, fecha cuando se construyó. También existe un conjunto de datos con las garantías donde se encuentra el ID del vehículo, ID de la garantía, fecha de reparación, comentarios del usuario, comentarios del técnico, pasos que siguieron para reparar el vehículo. También hay un historial del mantenimiento hecho al vehículo.
- 2.- Hay millones de registros en cada conjunto de datos. El ID del vehículo es un identificador común. Si existen las fechas de los DTCs y las garantías
- 3.-El objetivo principal del modelo es entender cuáles de los DTCs que tiene el vehículo son reparados.
- 4.-Están siendo analizados vehículos de uso particular. Si existen diferentes modelos y años de fabricación.
- 5.-No hay restricciones. Precisión. Los resultados se utilizarán para ver que DTCs son los que más se han reparado
- 6.-Python, GCP. No existen limitaciones

GPT4:

Gracias por la información detallada. Con base en tus respuestas, aquí tienes un plan para implementar un modelo que te ayude a encontrar la relación entre los DTCs generados por los vehículos y las garantías que han aparecido para esos vehículos:

1. Preparación de los Datos

a. Integración de Datos:

- Combinar los conjuntos de datos de DTCs y garantías utilizando el ID del vehículo como clave primaria.
- Incluir datos relevantes de historial de mantenimiento para enriquecer el análisis.

b. Limpieza de Datos:

- Eliminar o imputar registros con valores faltantes.
- Filtrar registros duplicados.
- Asegurar la consistencia de los formatos de fecha y otros campos.

2. Exploración de Datos

a. Análisis Descriptivo:

- Realizar análisis descriptivos para entender la distribución de DTCs y garantías.
- Generar visualizaciones de la frecuencia de DTCs y tipos de garantías.
- Analizar la distribución de DTCs y garantías a lo largo del tiempo.

b. Análisis de Correlación:

- Calcular las correlaciones entre las distintas variables (como odómetro, planta de construcción, modelo del vehículo) y la aparición de DTCs y garantías.

3. Ingeniería de Características

a. Creación de Variables:

- Crear variables que indiquen la frecuencia de DTCs por vehículo.
- Crear variables que indiquen la frecuencia de garantías por vehículo.
- Crear variables temporales (por ejemplo, tiempo desde la última reparación).

b. **Codificación de Variables Categóricas:

- Codificar variables categóricas como el modelo del vehículo y la planta de construcción.

4. Modelado

a. Selección del Modelo:

- Considerar modelos de clasificación como Random Forest, Gradient Boosting, o redes neuronales para predecir si un DTC resultará en una garantía.
- Considerar el uso de modelos de series temporales, si la secuencia temporal de los DTCs y las garantías es relevante.

b. Entrenamiento y Validación:

- Dividir los datos en conjuntos de entrenamiento y prueba.
- Entrenar el modelo utilizando el conjunto de entrenamiento.
- Validar el modelo utilizando métricas como precisión, recall, y F1-score en el conjunto de prueba.

5. Evaluación del Modelo

a. Métricas de Evaluación:

- Evaluar el modelo usando precisión para identificar qué DTCs son más propensos a resultar en garantías.
- Generar una matriz de confusión para entender el rendimiento del modelo.

b. Interpretación de Resultados:

- Interpretar las características más importantes del modelo para entender los factores que más contribuyen a la predicción de garantías.

6. Implementación y Monitoreo

a. Despliegue del Modelo:

- Implementar el modelo en un entorno de producción utilizando GCP.
- Desarrollar una API para que otras aplicaciones puedan consultar el modelo.

b. Monitoreo y Mantenimiento:

- Establecer un sistema de monitoreo para verificar el rendimiento continuo del modelo.
- Actualizar el modelo periódicamente con nuevos datos para mantener su precisión.

7. Visualización y Reportes

a. Dashboards:

- Crear dashboards interactivos para visualizar los resultados del modelo y las tendencias de DTCs y garantías.
- Incluir gráficos que muestren la frecuencia de los DTCs reparados y no reparados.

b. Reportes:

- Generar reportes periódicos que resuman los hallazgos clave y las recomendaciones basadas en el análisis de datos.

Herramientas y Tecnologías

- **Lenguaje de Programación:** Python
- **Bibliotecas de Análisis y Modelado:** Pandas, NumPy, Scikit-learn, TensorFlow/Keras (si se usan redes neuronales)
- **Visualización:** Matplotlib, Seaborn, Plotly
- **Plataforma de Despliegue:** Google Cloud Platform (GCP)
- **Herramientas de Monitoreo:** Google Cloud Monitoring

Este enfoque te permitirá encontrar patrones y relaciones entre los DTCs y las garantías, y entender mejor cuáles DTCs son más propensos a resultar en reparaciones bajo garantía. Si necesitas ayuda con algún paso específico o quieres que genere algún código de ejemplo, no dudes en decírmelo.

Persona:

¿Puedes darme otras opciones de posibles análisis que le pudiera hacer a los datos para lograr el objetivo final?

GPT4:

¡Por supuesto! Además del análisis descriptivo inicial, hay varias otras técnicas y enfoques analíticos que puedes considerar para lograr el objetivo de entender la relación entre los DTCs y las garantías en los vehículos. Aquí tienes algunas opciones adicionales de análisis que podrías realizar:

1. Análisis de Series Temporales

a. Análisis de Tendencias Temporales:

- Analizar la tendencia de aparición de DTCs y garantías a lo largo del tiempo.
- Identificar patrones estacionales o cíclicos.

b. Descomposición de Series Temporales:

- Descomponer las series temporales de DTCs y garantías en componentes de tendencia, estacionalidad y ruido.

2. Análisis de Cohortes

a. Definición de Cohortes:

- Agrupar los vehículos en cohortes basadas en el año de fabricación, modelo, o planta de construcción.

b. Comparación de Cohortes:

- Comparar la frecuencia y tipo de DTCs y garantías entre diferentes cohortes para identificar diferencias significativas.

3. Análisis de Supervivencia

a. Curvas de Supervivencia:

- Utilizar análisis de supervivencia para estudiar el tiempo hasta la aparición de un DTC o una garantía.

b. Modelos de Riesgo Proporcional (Cox):

- Aplicar modelos de riesgo proporcional para identificar factores que afectan el tiempo hasta la aparición de DTCs y garantías.

4. Análisis de Clúster

a. Segmentación de Vehículos:

- Aplicar algoritmos de clustering (como K-means o DBSCAN) para segmentar los vehículos en grupos basados en sus características y patrones de DTCs y garantías.

b. Análisis de Clústeres:

- Analizar las características de los distintos clústeres para identificar patrones comunes y diferencias.

5. Análisis de Texto

a. Procesamiento de Comentarios:

- Utilizar técnicas de procesamiento de lenguaje natural (NLP) para analizar los comentarios de los usuarios y técnicos en los registros de garantías.

b. Análisis de Sentimientos:

- Realizar análisis de sentimientos en los comentarios para entender mejor las experiencias de los usuarios y técnicos.

6. Análisis Predictivo

a. Modelos de Clasificación:

- Entrenar modelos de clasificación (como Random Forest, Gradient Boosting, SVM) para predecir si un DTC resultará en una garantía.

b. Modelos de Regresión:

- Utilizar modelos de regresión para predecir la severidad de un DTC en términos de costo o tiempo de reparación.

7. Análisis de Asociación

a. Reglas de Asociación:

- Aplicar algoritmos de minería de datos (como Apriori) para descubrir reglas de asociación entre diferentes DTCs y garantías.

8. Análisis de Componentes Principales (PCA)

a. Reducción de Dimensionalidad:

- Aplicar PCA para reducir la dimensionalidad de los datos y visualizar relaciones entre las variables.

9. Análisis de Anomalías

a. Detección de Anomalías:

- Utilizar técnicas de detección de anomalías para identificar patrones inusuales en los DTCs y garantías que podrían indicar problemas subyacentes.

10. Análisis de Causalidad

a. Modelos Causales:

- Aplicar modelos causales (como el modelo de variables instrumentales) para identificar relaciones causales entre los DTCs y las garantías.

Estas técnicas adicionales te permitirán explorar tus datos desde diferentes ángulos y obtener una comprensión más profunda de las relaciones entre los DTCs y las garantías en los vehículos. Si necesitas ejemplos más detallados o ayuda con algún análisis específico, no dudes en pedírmelo.

Proyecto 2: Detección de Fraude en Transacciones Financieras

Empresa: SAS Institute

Giro: Consultoría y Gestión de Datos

Area: Fraude e Inteligencia de Seguridad/ Apoyo en Aseguramiento de Calidad (QA)

SAS ofrece productos especializados para el área de fraude y riesgo, como SAS Fraud Management y SAS Risk Management.

El proyecto implica el análisis de transacciones financieras para identificar patrones sospechosos y prevenir el fraude utilizando SAS Fraud Management. Utilizamos este software para desarrollar modelos de detección de fraude y realizar análisis estadísticos detallados.

Interacción con Otras Áreas:

Regularmente se interactúa con el departamento de IT para la integración de sistemas y con el departamento de Finanzas para coordinar la implementación de medidas antifraude.

Aplicación de la Inteligencia Artificial:

Para obtener ideas sobre cómo mejorar la detección de fraude, utilizamos técnicas de ingeniería de instrucciones con ChatGPT. A continuación, se presenta una posible implementación:

1. Identificación del Problema:

"¿Cuáles son las mejores técnicas de inteligencia artificial para la detección de fraude en transacciones financieras y cómo se pueden implementar utilizando SAS Fraud Management?"

2. Soluciones Propuestas:

CHAGPT4:

“

- Implementar algoritmos de machine learning, como el aprendizaje supervisado y no supervisado, para identificar patrones anómalos.
- Utilizar redes neuronales profundas y algoritmos de clustering para detectar comportamientos sospechosos en tiempo real.

3. Pasos a Seguir:

- Recopilar y preprocesar datos de transacciones financieras.
- Entrenar modelos de detección de fraude utilizando SAS Fraud Management.

- *Integrar los modelos en el sistema existente y monitorear su rendimiento para ajustar los parámetros según sea necesario.*

“

Proyecto 3: Optimización de la Calidad del Software

Empresa: SAS Institute

Giro: Consultoría y Gestión de Datos

Area: Fraude e Inteligencia de Seguridad/ Apoyo en Aseguramiento de Calidad (QA)

SAS ofrece productos especializados para el área de fraude y riesgo, como SAS Fraud Management y SAS Risk Management.

El proyecto implica la revisión y mejora continua de los procesos de desarrollo de software para garantizar la calidad y fiabilidad de los productos de SAS. Esto incluye pruebas automáticas, revisiones de código y análisis de rendimiento.

Interacción con Otras Áreas:

Regularmente se interactúa con los departamentos de Desarrollo de Software y Soporte Técnico para implementar y evaluar las mejoras.

Aplicación de la Inteligencia Artificial:

Para obtener ideas sobre cómo aplicar la IA en la optimización de la calidad del software, utilizamos técnicas de ingeniería de instrucciones con ChatGPT. A continuación, se presenta una posible implementación:

1. Identificación del Problema:

"¿Cuáles son las mejores técnicas de inteligencia artificial para la optimización de la calidad del software y cómo se pueden implementar en el contexto de QA en SAS?"

2. Soluciones Propuestas:

ChatGPT4:

“

- *Implementar algoritmos de machine learning para predecir errores y fallos en el software antes de su lanzamiento.*

- *Utilizar técnicas de procesamiento de lenguaje natural (NLP) para automatizar la revisión de código y la generación de pruebas.*

3. Pasos a Seguir:

- *Recopilar datos históricos de errores y fallos de software.*

- *Entrenar modelos predictivos utilizando las herramientas de SAS.*

- *Desplegar los modelos y crear un sistema de alerta temprana para identificar y corregir problemas antes de que afecten a los usuarios.*

“

Conclusiones

Esta actividad demostró cómo la utilización de un modelo LLM, como ChatGPT, puede ser una herramienta poderosa para identificar y diseñar aplicaciones de IA en diversas tareas diarias del trabajo. Al implementar técnicas avanzadas de detección de fraude y optimización de la calidad del software utilizando productos de SAS como SAS Fraud Management, las empresas pueden mejorar significativamente su eficiencia y precisión. Además, es crucial tener en cuenta que los verbos y otros componentes lingüísticos pueden influir en la agrupación de tópicos, por lo que es esencial aplicar técnicas de preprocesamiento adecuadas para mejorar la calidad del análisis. La colaboración entre departamentos y el uso de herramientas de SAS son fundamentales para el éxito de estos proyectos.

Proyecto 4: Generación de Audiencias para Optimización de Campañas Publicitarias

Modelo: ChatGPT 4

Empresa: Stellantis México

Giro: Automotriz

Área: Marketing / Data Governance

Actividad: Explotación de data de prospectos y clientes para eficientar gastos de campañas digitales

Descripción de Actividad: Analizar los datos de clientes y prospectos segregados en diferentes plataformas dentro de la empresa, con el objetivo de encontrar patrones que ayuden a generar audiencias que sirvan de retroalimentación a las diferentes plataformas de Marketing Digital.

Interacción con: Ventas, Post-venta, Calidad, Publicidad, Financiera

Ingeniería de Instrucciones

Persona:

Actúa como un científico de datos.

GPT4:

¡Claro! ¿En qué puedo ayudarte hoy? ¿Tienes algún proyecto o análisis específico en mente?

Persona:

¿Qué información requieres saber para ayudarme a implementar algún modelo que encuentre la relación entre los diferentes tipos de datos que se generan para prospectos y clientes en una empresa grande? Contextualiza tus preguntas bajo el objetivo de poder generar grupos de clientes o prospectos que compartan las mismas características de aquellos que concluyeron con la compra del producto que comercializa la misma empresa.

GPT4:

Entendido. Para ayudarte a implementar un modelo que encuentre la relación entre los diferentes tipos de datos de prospectos y clientes, y que permita generar grupos de clientes o prospectos con características similares a aquellos que concluyeron con la compra del producto, necesitare la siguiente información:

1. Definición del Problema y Objetivos

- **Objetivo principal:** Confirmar que el objetivo es identificar características comunes de los clientes que compraron el producto para aplicarlas a prospectos y segmentar clientes potenciales.
- **Objetivos secundarios:** Por ejemplo, mejorar las campañas de marketing, optimizar los esfuerzos de ventas, etc.

2. Descripción del Producto

- **Tipo de producto:** Entender el producto comercializado para identificar posibles características relevantes.
- **Ciclo de ventas:** Tiempo promedio desde el primer contacto hasta la compra.

3. Datos Disponibles

- **Datos de Prospectos y Clientes:**
- **Datos demográficos:** edad, género, ubicación, etc.
- **Datos psicográficos:** intereses, hábitos de compra, etc.
- **Datos de comportamiento:** visitas a la web, interacciones con el marketing, etc.
- **Datos transaccionales:** historial de compras, montos gastados, frecuencia de compra, etc.
- **Datos de engagement:** respuestas a campañas de marketing, participación en eventos, etc.
- **Etiquetas de compra:** Identificación de quiénes compraron y quiénes no.

4. Calidad y Disponibilidad de Datos

- **Consistencia:** Verificar si los datos están completos y son precisos.
- **Historial de datos:** ¿Cuánto tiempo de historial de datos está disponible?
- **Formato de los datos:** Bases de datos relacionales, hojas de cálculo, CRM, etc.

5. Preprocesamiento de Datos

- **Limpieza de datos:** Cómo se manejarán los datos faltantes o inconsistentes.
- **Transformaciones:** Normalización, codificación de variables categóricas, etc.
- **Feature Engineering:** Creación de nuevas variables a partir de las existentes.

6. Modelos y Algoritmos

- **Algoritmos:** Algoritmos de clasificación y clustering como árboles de decisión, regresión logística, K-means, etc.
- **Evaluación del modelo:** Métodos para evaluar la precisión del modelo (ROC-AUC, precisión, recall, etc.).

7. Segmentación y Clustering

- **Métodos de clustering:** K-means, DBSCAN, jerárquico, etc.
- **Criterios de segmentación:** Variables principales para segmentar a los clientes.

8. Herramientas y Tecnologías

- **Herramientas de análisis:** Python, R, SQL, etc.
- **Plataformas:** Jupyter Notebooks, Google Colab, herramientas de visualización como Tableau o Power BI.

9. Resultados Esperados y Presentación

- **Formatos de entrega:** Dashboards, informes, visualizaciones interactivas, etc.
- **Frecuencia de actualización:** ¿Con qué frecuencia se debe actualizar el modelo y el análisis?

- Con esta información, podré guiarte de manera más efectiva en la implementación del modelo. ¿Podrías proporcionarme algunos detalles específicos sobre los puntos anteriores?

Persona:

1. Definición del Problema y Objetivos

- Objetivo principal: El objetivo es identificar características comunes de los clientes que compraron el producto para aplicarlas a prospectos y segmentar clientes potenciales.
- Objetivos secundarios: Mejorar el desempeño de las campañas de marketing, mejorar la tasa de conversión a venta.

2. Descripción del Producto

- Tipo de producto: Automóviles.
- Ciclo de ventas: El tiempo promedio desde el primer contacto hasta la compra es de 90 días.

3. Datos Disponibles

- Datos de Prospectos y Clientes: almacenados en bases de datos SQL, DB2, archivos Excel.
- Datos demográficos: Se cuenta con información de edad, género, ubicación.
- Datos psicográficos: Se cuenta con datos sobre intereses, hábitos de compra, etc.
- Datos de comportamiento: Se cuenta con datos sobre visitas a la web, interacciones con anuncios digitales, niveles de búsqueda de productos en Google.
- Datos transaccionales: Se cuenta con datos sobre historial de compras, montos gastados, frecuencia de compra, créditos otorgados.
- Datos de engagement: Se cuenta con datos sobre respuestas a campañas de marketing, participación en eventos, etc.
- Etiquetas de compra: Hay forma de identificar quiénes compraron y quiénes no.

4. Calidad y Disponibilidad de Datos

- Consistencia: Los datos pueden no estar completos y no ser tan precisos. El nivel de consistencia de los datos es superior al 90%.
- Historial de datos: Se cuenta con un historial de datos mayor a 5 años.
- Formato de los datos: Bases de datos relacionales, hojas de cálculo, CRM, etc.

5. Preprocesamiento de Datos

- Limpieza de datos: Se pide omitir registros que contengan datos faltantes y hacer correcciones sobre datos mal escritos y correos y teléfonos con estructura incorrecta.
- Transformaciones: Se pide realizar normalización de los datos y codificación de variables categóricas.
- Feature Engineering: Creación de nuevas variables a partir de las existentes.

6. Modelos y Algoritmos

- Algoritmos: Se puede emplear alguno de los siguientes algoritmos de clasificación y clustering: árboles de decisión, regresión logística, K-means.
- Evaluación del modelo: Proponer algunos Métodos para evaluar la precisión del modelo (ROC-AUC, precisión, recall, etc.).

7. Segmentación y Clustering

- Métodos de clustering: Utilizar K-means.
- Criterios de segmentación: Interacciones con anuncios, historial de compras, edad, nivel socioeconómico, ubicación geográfica, intención de compra.

8. Herramientas y Tecnologías

- Herramientas de análisis: Python, R, SQL, etc.

- Plataformas: Jupyter Notebooks, Google Colab, herramientas de visualización como Tableau o Power BI.

9. Resultados Esperados y Presentación

- Formatos de entrega: Se espera obtener grupos de prospectos que compartan características similares a aquellos clientes que compraron un vehículo en los últimos 24 meses. Además, un Dashboard de medición de tasas de conversión para evaluar el desempeño del modelo y la optimización de las campañas publicitarias digitales.
- Frecuencia de actualización: El modelo debe actualizarse de forma diaria, tomando los nuevos datos de cada fuente de datos y haciendo el análisis correspondiente.

GPT4:

Implementación del Modelo

Con la información proporcionada, podemos proceder a diseñar e implementar un modelo de segmentación y predicción. A continuación, se detalla un plan paso a paso:

Paso 1: Recolección y Exploración de Datos

1. **Acceso a las Bases de Datos:**
 - Conectar a las bases de datos SQL y DB2.
 - Importar los datos de los archivos Excel.
2. **Exploración Inicial:**
 - Realizar un análisis exploratorio de datos (EDA) para comprender la distribución y relaciones entre las variables.
 - Visualizar las principales características demográficas, psicográficas, de comportamiento, transaccionales y de engagement.

Paso 2: Preprocesamiento de Datos

1. **Limpieza de Datos:**
 - Eliminar registros con datos faltantes.
 - Corregir datos mal escritos y estandarizar formatos de correos y teléfonos.
2. **Transformaciones:**
 - Normalizar los datos numéricos para que tengan una media de 0 y una desviación estándar de 1.
 - Codificar variables categóricas utilizando técnicas como One-Hot Encoding.
3. **Feature Engineering:**
 - Crear nuevas variables a partir de las existentes (por ejemplo, interacción total con anuncios, promedio de gasto por compra, etc.).

Paso 3: División de los Datos

1. **Separar en Conjuntos de Entrenamiento y Prueba:**
 - Dividir los datos en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%).

Paso 4: Entrenamiento del Modelo

1. **Clasificación:**
 - Entrenar modelos de árboles de decisión y regresión logística para predecir la probabilidad de compra.

- Evaluar los modelos usando métricas como ROC-AUC, precisión y recall.
- 2. **Clustering:**
 - Aplicar K-means para segmentar a los clientes y prospectos en grupos con características similares.
 - Determinar el número óptimo de clusters utilizando el método del codo o el análisis de la silueta.

Paso 5: Evaluación y Validación del Modelo

1. **Evaluación de Clasificación:**
 - Evaluar el desempeño de los modelos de clasificación en el conjunto de prueba.
 - Seleccionar el modelo con mejor desempeño basado en las métricas de evaluación.
2. **Evaluación de Clustering:**
 - Validar la coherencia de los clusters y la homogeneidad interna.

Paso 6: Implementación y Actualización del Modelo

1. **Automatización:**
 - Configurar procesos automatizados para la actualización diaria de los datos y la re-ejecución del modelo.
 - Utilizar pipelines en Python con herramientas como Airflow o cron jobs para automatizar la actualización.
2. **Integración:**
 - Integrar el modelo en las plataformas de análisis y visualización (Tableau, Power BI) para monitorear el desempeño y la conversión.
 - Crear un Dashboard interactivo que muestre las tasas de conversión y la efectividad de las campañas de marketing.

Ejemplo de Código para el Preprocesamiento y Clustering

```
import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.cluster import KMeans
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer

# Cargar datos
data = pd.read_csv('clientes_prospectos.csv')

# Limpieza de datos
data = data.dropna()
data['telefono'] = data['telefono'].str.replace(r'\D', '')

# Transformaciones
numeric_features = ['edad', 'monto_gastado', 'frecuencia_compra']
categorical_features = ['genero', 'ubicacion', 'intereses']

numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())])

categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

# Feature Engineering
data['interacciones_totales'] = data['interacciones_anuncios'] + data['participacion_eventos']

# Aplicar preprocesamiento
X = preprocessor.fit_transform(data)
```

```
# K-means Clustering
kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(X)

# Añadir clusters al dataframe original
data['cluster'] = clusters

# Guardar resultados
data.to_csv('clientes_segmentados.csv', index=False)
```

Presentación y Seguimiento

- **Dashboard en Tableau o Power BI:**
 - Crear visualizaciones para mostrar la distribución de clientes por clusters, tasas de conversión, y el rendimiento de las campañas de marketing.
- **Monitoreo y Actualización:**
 - Configurar alertas y reportes automáticos para el equipo de marketing y ventas.

Este plan proporciona una guía estructurada para implementar el modelo de segmentación y predicción de prospectos y clientes. ¿Te gustaría profundizar en algún paso específico o necesitas más detalles sobre algún aspecto?

Conclusiones Finales

Con esta actividad se comprueba la gran ventaja que los LLM brindan a los científicos de datos para el modelado de soluciones que ayuden en la ejecución de proyectos con manejo de grandes volúmenes de información al interior de las empresas. Las IAs se convierten más y más en aliados estratégicos para la implementación de soluciones en el corto y mediano plazo con tasas de efectividades cada vez más altas. También, se pudo corroborar las diferentes respuestas que generan los modelos comerciales (como Llama o Chat-GPT) ante los mismos *prompts* y entender la personalidad que adopta cada uno de éstos, según el rol que se le asigne.