Maestría en Analítica de Datos
Automatización e Integración de Datos

UNIVERSIDAD CENTRAL

Automatización y desarrollo de modelos de Machine Learning para la clasificación de clientes potenciales en el mercado de automóviles

Proyecto de Investigación

2024

Juan Esteban Pulido Lancheros
Jorge Hernando Chávez Camargo

Universidad Central
Bogotá D.C.

Tabla de contenido

1.	Introducción	3
2.	Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA	4
2.1.	Título del proyecto de investigación	5
2.2.	Objetivo general	5
2.3.	Objetivos específicos	5
2.4.	Pregunta de investigación	6
2.5.	Hipótesis	6
3.	Origen de los datos	6
3.1.	Acerca del conjunto de datos	7
3.2.	Consideraciones legales o éticas del uso de la información	8
3.3.	Retos de la información y los datos para Integración y Automatización de Datos para IA	8
3.4.	Expectativas de la utilización de Integración y Automatización de Datos para IA en el proyecto	8
4.	Diagrama de integración y Automatización de Datos	9
5.	Integración de Datos	9
6.	Automatización de Datos	10
7.	Aplicación de IA	11

1. Introducción

La capacidad de una empresa para comprender y segmentar a sus clientes es fundamental para el éxito en un mercado altamente competitivo.

En el contexto de una compañía de venta de automóviles y que está incursionando en un nuevo mercado, la clasificación efectiva de los clientes es fundamental para determinar cuál es la estrategia más acertada para abordarlos a través de sus campañas de marketing.

Este proyecto tiene como objetivo diseñar una arquitectura de datos que tenga la capacidad de cargar datos de una fuente de información, permitir su análisis, generar un modelo de machine learning y finalmente presentar los resultados en una herramienta de visualización de datos.

Todo este proceso busca que, a través de la clasificación de los clientes en grupos homogéneos, permita a la empresa desarrollar estrategias específicas para cada segmento, lo que maximizará el impacto de las campañas de marketing en busca de llegar al mayor nivel de ventas.

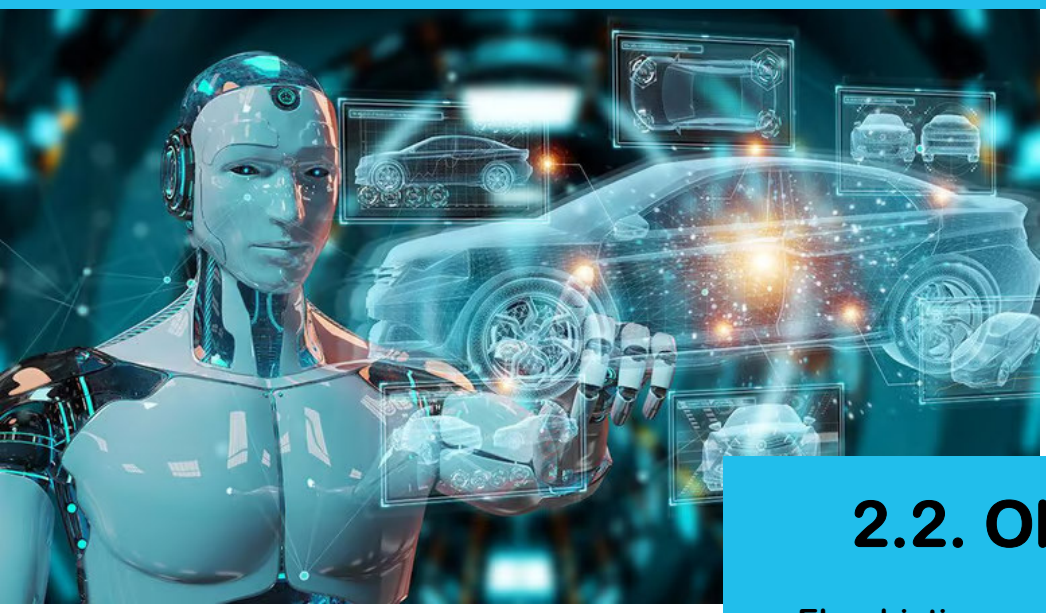
A su vez se pretende que, con el flujo de datos construido y automatizado, se realice un análisis de datos avanzado que sea lo más eficiente en términos de tiempo de procesamiento de información y generación eficiente de resultados relevantes para la compañía.

2. Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA

El proyecto de investigación se enfoca en la Integración y Automatización de Datos para potenciar la Inteligencia Artificial (IA) en la clasificación de clientes de una empresa dedicada a la venta de automóviles. Su objetivo principal es desarrollar un sistema avanzado que aproveche metodologías de Machine Learning para clasificar de manera eficiente y precisa la base de datos de clientes potenciales de la compañía. Una de las características centrales del proyecto es su enfoque integral, que abarca desde la recopilación y limpieza de datos hasta la implementación del algoritmo de aprendizaje automático que nos permitirá clasificar de manera correcta a los clientes. Para lograr esto, se utilizarán técnicas de integración de datos que permitan extraer la información proveniente de su fuente. La automatización desempeñará un papel crucial al simplificar y agilizar procesos como la extracción y transformación de datos, permitiendo así que cuando se actualice la base de datos una actualización continua de la base de datos y asegurando que la información utilizada por los modelos de IA sea siempre relevante y actualizada. Además, el proyecto se caracteriza por su enfoque en la escalabilidad y la adaptabilidad, buscando desarrollar una arquitectura de datos pueda generar la nueva información estratégica de manera ágil y eficiente de acuerdo con las necesidades de la empresa. Se establecerán métricas claras para evaluar el desempeño del sistema, como la precisión de la clasificación de clientes y la eficiencia en el procesamiento de datos. En última instancia, se espera que el proyecto mejore significativamente la capacidad de la empresa para comprender y atender las necesidades individuales de sus clientes, impulsando así su competitividad y rentabilidad en el mercado.

2.1. Título del Proyecto

“Automatización y desarrollo de un modelo de Machine Learning para la clasificación de clientes potenciales en el mercado de automóviles”



2.2. Objetivo General

El objetivo general de este proyecto de investigación es desarrollar un sistema de Integración y Automatización de Datos para potenciar la Inteligencia Artificial (IA) en la clasificación de clientes de una empresa dedicada a la venta de automóviles.

2.3. Objetivos Específicos

- Generar un mecanismo que integre los datos para la clasificación de clientes.
- Realizar la limpieza al conjunto de datos, garantizando la coherencia y calidad de la información.
- Desarrollar algoritmos de aprendizaje automático y técnicas de Inteligencia Artificial adecuadas para la clasificación de clientes, teniendo en cuenta la precisión, la escalabilidad y la interpretabilidad de los resultados.
- Implementar una interfaz de usuario intuitiva que permita a los usuarios de la empresa interactuar con el sistema, visualizar resultados y realizar ajustes según sea necesario.

2.4 Pregunta de investigación

¿Cómo puede la integración y automatización de datos potenciar la inteligencia artificial en la clasificación de clientes para una empresa dedicada a la venta de automóviles?

2.5 Hipótesis

La implementación de un sistema de integración y automatización de datos permitirá a la empresa de venta de automóviles mejorar la precisión y eficiencia de la clasificación de clientes mediante algoritmos de inteligencia artificial, lo que resultará en un aumento de la efectividad en las estrategias de marketing y ventas.

3. Origen de los datos

kaggle

https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation

3.1. Acerca del conjunto de datos

Contexto

Una empresa de automóviles tiene planes de ingresar a nuevos mercados con sus productos existentes (P1, P2, P3, P4 y P5). Después de una intensa investigación de mercado, dedujeron que el comportamiento del nuevo mercado es similar al del mercado existente.

En su mercado actual, el equipo de ventas ha clasificado a todos los clientes en 4 segmentos (A, B, C, D). Luego, realizaron actividades de divulgación y comunicación segmentadas para un segmento diferente de clientes. Esta estrategia les ha funcionado bien. Planean utilizar la misma estrategia para los nuevos mercados y han identificado 2627 nuevos clientes potenciales.

Contenido

Variable	Definición
ID	Identificación única
Gender	Género del cliente
Ever_Married	Estado civil del cliente
Age	Edad del cliente
Graduated	¿El cliente es un graduado?
Profession	Profesión del cliente
Work_Experience	Experiencia laboral en años.
Spending_Score	Puntuación de gasto del cliente.
Family_Size	Número de familiares del cliente (incluido el cliente)
Var_1	Categoría anónima para el cliente
Segmentation	(objetivo) Cliente Segmento del cliente

Licencia

3.2. Consideraciones legales o éticas del uso de la información

El conjunto de datos se encuentra en la página web de Kaggle que es una comunidad que pone a disposición de estudiantes e investigadores conjuntos de datos sobre diversos temas. Sin embargo, al ser de dominio público, puede ser utilizada con total libertad para fines educativos. Además, no contiene datos sensibles ni personales de ninguno de los clientes que hay dentro de la misma. Los datos están debidamente anonimizados y no es posible perfilar directamente a ningún individuo. Los datos serán destinados netamente al propósito académico que se desarrollará en el presente proyecto y no causará daño a las personas o grupos involucrados.

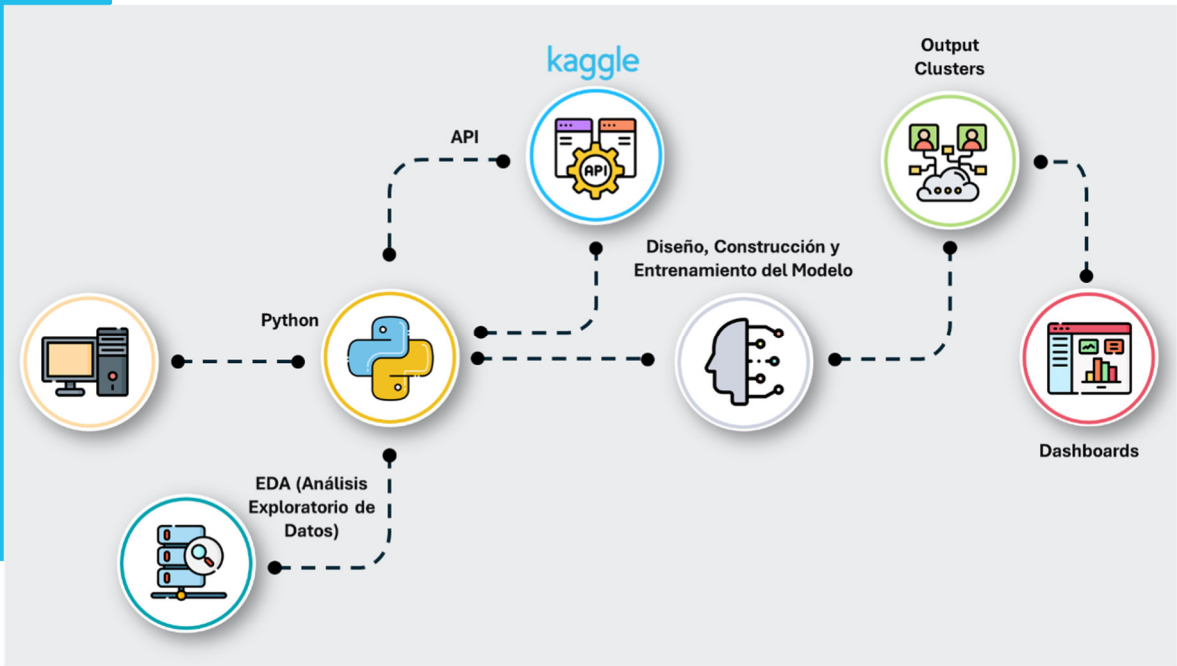
3.3. Retos de la información y los datos para Integración y Automatización de Datos para IA

Al integrar y automatizar datos para aplicaciones de IA, es fundamental garantizar la calidad y la coherencia de los datos. Esto puede implicar abordar varios retos en la integración, limpieza y carga de datos hacia una herramienta de visualización de datos.

3.4. Expectativas de la utilización de Integración y Automatización de Datos para IA en el proyecto

Se espera que la integración y automatización de datos simplifiquen y agilicen el proceso de preparación de datos, lo que permitirá hacer los análisis pertinentes para identificar el modelo apropiado para la clasificación de los clientes. Una vez desarrolladas las soluciones a los retos que hay en materia de la integración, modelación y carga de datos se logrará de manera satisfactoria el desarrollo efectivo de automatización de datos en proyectos de IA.

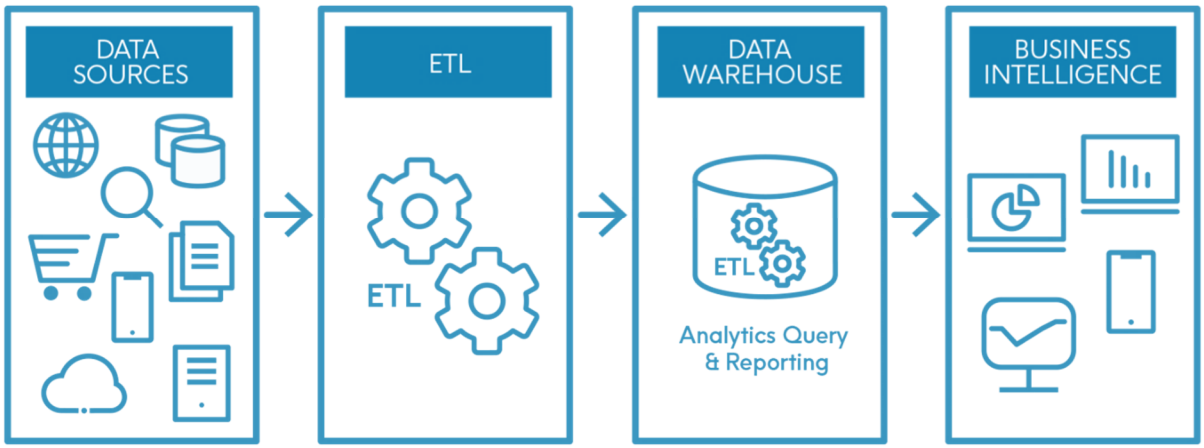
4. Diagrama de integración y Automatización de Datos



5. Integración de Datos

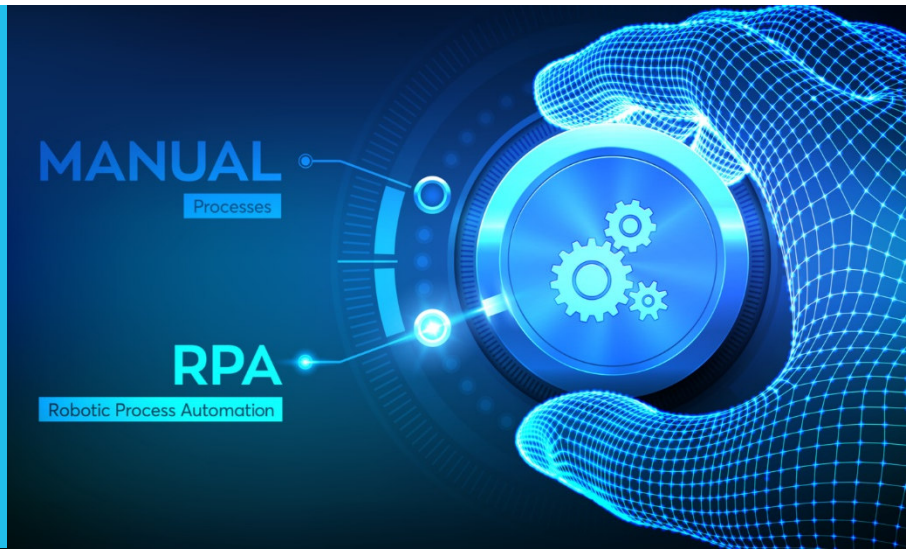
Para integrar los datos se utilizarála plataforma de computación en la nube Google Cloud Plataform (GCP) a través del servicio Google App Engine y mediante el uso de Python que es compatible con este servicio. Una vez obtenida la información se realizará el análisis descriptivo a los datos y su respectiva normalización si es necesario para asegurar que todas las variables estén en la misma escala y tengan un impacto equitativo en el análisis.

Para almacenar los datos utilizaremos Google Cloud Storage (GCS) que permite alojar un gran volumen de datos de forma segura y es altamente escalable.



6. Automatización de Datos

La automatización de datos es clave para garantizar que las necesidades de información de la empresa sean atendidas de manera eficiente y regular sin intervención manual. A continuación, se describe de manera detallada la forma en que se va a implementar la automatización de datos en el contexto del proyecto:



1. **Fuente de datos:** El primer paso es definir la fuente de datos relevantes que serán procesados y analizados para dar respuesta a las necesidades de información de la compañía.
2. **Conexión a la fuente de datos:** Una vez identificada la fuente de datos, se establecerá la conexión vía API que se ejecutará a través de los scripts

desarrollos y realizar la extracción de la información necesaria de forma periódica.

3. **Procesamiento de datos:** Una vez que se hayan extraído los datos de las diferentes fuentes, se realizará el procesamiento de datos para limpiar y estructurar la información de manera coherente junto con la respectiva normalización de datos y la estandarización de formatos para de esta forma contar con la información lista para el modelado y posterior clasificación.
4. **Modelo de clasificación:** Para realizar la clasificación de clientes se aplicará un modelo basado en árbol de decisión. Este tipo de modelos son flexibles y permiten manejar tanto datos numéricos como categóricos; son altamente interpretables y fáciles de visualizar, además de su rendimiento óptimo para conjuntos de datos de diferentes volúmenes.

Para la clasificación de los clientes a través del árbol de decisión. Este algoritmo de aprendizaje automático supervisado es especialmente útil para problemas de clasificación. Los árboles de decisión dividen el conjunto de datos en subconjuntos más pequeños basados en características específicas, con el objetivo de clasificar correctamente las instancias de datos en categorías predefinidas.

Durante el proceso de entrenamiento, los árboles aprenderán a clasificar a los clientes en diferentes segmentos en función variables como de una variedad de características, como género, edad, nivel de formación

académica, hábitos de gasto, estado civil, profesión y cantidad de personas a cargo.

Una vez que se hayan entrenado los árboles de decisión, se evaluará su rendimiento utilizando métricas como la precisión, la exhaustividad y el puntaje F1 en un conjunto de datos de prueba y se ajustarán los hiperparámetros de los árboles de decisión para optimizar su rendimiento y evitar el sobreajuste. Esto garantizará que el modelo sea capaz de analizar bien los datos nuevos.

El uso de árboles de decisión en la aplicación de inteligencia artificial proporcionará una metodología efectiva y transparente para la clasificación de los clientes en la empresa de automóviles, permitiendo una caracterización de los clientes y facilitando la personalización de las estrategias de marketing y ventas.