Análisis descriptivo y Modelo Predictivo para las ventas de combustible la Estación de Servicio San Diego

Fernando Roa[1], Cristian Rivera[2]

[1−2]Facultad de Ingeniería y Ciencias Básicas,
Universidad Central
Maestría Analítica de Datos
Automatización e Integración de Datos para la IA
Bogotá, Colombia
{[1]froam1, [2]criveraa}@correo1.com

November 25, 2023

Contents

1 Introducción

En la era actual de la información, la capacidad de integrar y automatizar datos de múltiples fuentes es crucial para el desarrollo y la eficiencia de sistemas basados en Inteligencia Artificial (IA). El presente trabajo se enfoca en la implementación de estas capacidades creando un ETL (Extract, Transform, Load), una metodología consolidada que facilita la extracción de datos desde distintos orígenes, su transformación en un formato adecuado y, finalmente, su carga en sistemas de almacenamiento o análisis, además de la ejecución periódica de predicciones de ventas usando modelos de machine learning utilizando recursos de la nube de Oracle.

El núcleo de este proyecto radica en la concepción de un modelo predictivo especialmente diseñado para analizar información de ventas de combustible de una estación de servicio, la cual a través de un sistema POS registra todas la ventas que realiza. Dicho modelo busca no sólo comprender las tendencias y patrones históricos, sino también anticipar futuras variaciones y comportamientos en el mercado. La implementación de la IA en este contexto se vuelve una herramienta invaluable para las empresas que buscan mantenerse a la vanguardia y tomar decisiones basadas en datos precisos y proyecciones confiables.

Esta investigación abordará las técnicas y herramientas utilizadas para la integración y automatización de datos, así como la estructura y funcionamiento del modelo predictivo propuesto, ofreciendo una perspectiva detallada de su aplicabilidad y potencial impacto en el ámbito comercial.

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

El proyecto presenta una serie de características distintivas que lo posicionan como una herramienta esencial para el manejo de grandes volúmenes de datos comerciales. En primer lugar, cuenta con una robusta infraestructura de integración que facilita la recolección y consolidación de datos desde páginas web, asegurando la completitud de la información. La automatización incorporada garantiza flujos de trabajo eficientes y reduce el margen de error humano, maximizando la precisión en cada etapa del proceso ETL (Extracción, Transformación y Carga). En el aspecto de transformación, se han implementado algoritmos avanzados que adaptan y preparan los datos para análisis, garantizando su relevancia y utilidad. El corazón del proyecto es su modelo predictivo basado en técnicas de Inteligencia Artificial, diseñado para detectar tendencias y patrones en la información de venta. Este modelo se reentrena constantemente con nuevos datos, mejorando su precisión y adaptabilidad frente a cambiantes escenarios de mercado. Además, presenta interfaces amigables para los usuarios y sistemas de visualización de datos, permitiendo una interpretación intuitiva de los resultados. Asimismo, se ha puesto un especial énfasis en la escalabilidad y flexibilidad del sistema, permitiendo su adaptación a diferentes tamaños y tipos de organizaciones. Por último, pero no menos importante, el proyecto se rige por estrictas normas de seguridad y confidencialidad, en materia de privacidad y gestión de datos. El compromiso con estas normativas asegura que, además de proporcionar insights valiosos, el proyecto mantiene la integridad y confianza de sus usuarios. La combinación de estas características convierte a este proyecto en una solución integral, capaz de atender las crecientes demandas del análisis de datos en el mundo comercial y ofrecer previsiones acertadas que pueden guiar estrategias de negocio hacia el éxito. Esta innovadora propuesta representa un paso adelante en la confluencia de la Inteligencia Artificial, la gestión de datos y el análisis predictivo, consolidándose como un referente en el ámbito de la tecnología y la toma de decisiones basadas en

datos.

2.1 Integración y Automatización de Datos para Inteligencia Artificial en el Proyecto: ETL y Modelo Predictivo para Información de Venta

La estación de Servicio TuyaSolar cuya actividad económica es la venta de combustibles como gasolina corriente y diesel cuenta con un sistema Pos mediante el cual se almacena la información de todas las ventas realizadas. Dicha información puede ser consultada y descargada en archivos de formato .xls desde un sitio web suministrado por el proveedor del sistema.

Actualmente dicha información no es aprovechada en ningún sentido, por lo que el presente proyecto es vital para potenciar la toma de decisiones basadas en datos por parte de la estación.

2.2 Objetivo general

Desarrollar e implementar un sistema de Integración y Automatización de Datos mediante la creación de un ETL y un Modelo Predictivo para la proyección de ventas, constituyéndose en una herramienta para la toma de decisiones basadas en datos.

2.2.1 Objetivos especificos

- Crear un ETL para la recolección y consolidación de datos, garantizando su homogeneidad y consistencia, además de garantizar la adaptabilidad y escalabilidad del sistema a diferentes contextos organizacionales.
- Crear un modelo de machine learning para predecir las ventas de combustible de la estación de servicio entrenado a partir de datos históricos.
- Automatizar el ETL y adaptar el modelo predictivo para que éste se retroalimente y se adapte a nuevos escenarios de mercado, manteniéndolo ajustado a la realidad.
- Crear una interfaz de visualización intuitiva que facilite la comprensión y explotación de los insights derivados del modelo, al tiempo que se asegura el cumplimiento de las normativas de privacidad y gestión de datos.

2.3 Alcance

Nuestro alcance incluye el diseño de una infraestructura de integración que permita la armonización y consolidación de datos provenientes de páginas web, garantizando la calidad y coherencia de la información. Se dará especial énfasis a la implementación de algoritmos avanzados para la transformación de datos y la adaptabilidad constante del modelo predictivo a las dinámicas cambiantes del mercado. Además, se desarrollará una interfaz de usuario con un sistema de visualización que brindarán interpretaciones claras y accesibles de los resultados generados. Por último, el proyecto se compromete a respetar las normativas vigentes en materia de privacidad y gestión de datos, asegurando la protección y confiabilidad de la información manejada. Si bien el objetivo principal es el análisis de información de venta, la flexibilidad y adaptabilidad del sistema posibilitan futuras expansiones y adaptaciones a otros ámbitos comerciales y de análisis.

2.4 Pregunta de investigación

¿Cómo la implementación de un sistema de Integración y Automatización de Datos impacta en la toma de decisiones de Pymes en el desarrollo de su actividad comercial?

2.5 Hipotesis

La implementación de un sistema de Integración y Automatización de Datos se convertirá en una herramienta robusta para la toma de decisiones como: negociaciones de contratos con proveedores y clientes, estimación de fechas para la compra del combustible e identificación de temporadas de mayor y menor venta. El uso de la herramienta propuesta conducirá a decisiones estratégicas respaldadas en datos, optimizando las operaciones, la identificación de tendencias de mercado y la respuesta a cambios dinámicos, resultando en una ventaja competitiva sustancial para las organizaciones que adoptan dicho sistema.

3 Reflexiones sobre el origen de datos e información

La información utilizada para el desarrollo del presente proyecto es de tipo estructurado y consiste en una tabla en donde se alojan los registros de todas las ventas realizadas por la estación de servicio. Dentro de los campos más relevantes destacan: Fecha y hora de la venta, vendedor, producto, volumen de venta, total venta y nombre del cliente en caso de contratos.

Se cuenta con la autorización para el uso de la información por parte del dueño de la estación y de mutuo acuerdo se establece que el producto final será aprovecha solamente por personal autorizado de tal empresa.

3.1 ¿Cual es el origen de los datos e información ?

Los datos de ventas de la estación son tomados de la página web con url https://www.imedoil.online/nexus.php suministrada por la empresa Imedoil S.A.S. proveedora del sistema POS, con un formato de descarga con extensión .xls.

3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información?

Teniendo en cuenta que los datos utilizados para el proyecto no contienen datos personales, no existe la obligación a cumplir con todas las disposiciones de la Ley de Protección de Datos Personales 1581 de 2012.

A pesar de lo anterior, se tendrán en cuenta las siguientes consideraciones éticas:

- Privacidad: Recopilar y utilizar solo la información no personal que sea necesaria para el propósito previsto.
- Transparencia: Informar a los usuarios sobre cómo se recopila y utiliza su información no personal.
- Seguridad: Tomar medidas de seguridad para proteger la información no personal de los usuarios.

3.3 ¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA?

Es necesario realizar un análisis de valores faltantes, duplicados y posibles inconsistencias en la data descargada del sitio web, teniendo en cuenta que hasta el momento no ha sido utilizada y se desconoce la presencia o ausencia de posibles inconsistencias.

3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto?

La integración y automatización de datos en este proyecto es el corazón del mismo, ya que se espera una automatización del 100% de los procesos tales como: extración, transformación y carga de datos en la nube, predicción y visualización de ventas en tableros cuya disponibilidad será únicamente para usuarios autorizados por la estación de Servicio Tuya Solar, evitando al máximo cualquier tipo de mantenimiento o intervención manual.

4 Diseño de integración y Automatización de Datos para IA (Diagrama)

El diseño del proyecto desde la extracción de los datos hasta el tablero de visualización se presenta en el siguiente diagrama:

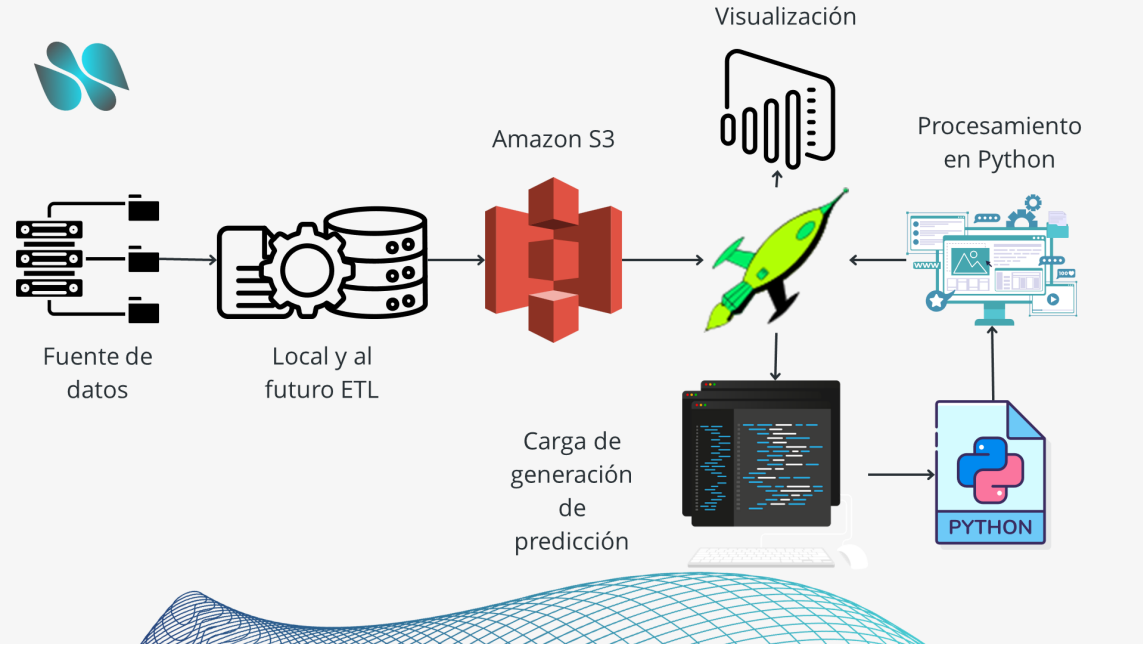

Figure 1: Diagrama del proyecto - End to End

5 Integración de Datos

En el escenario descrito, se inicia con la recolección de datos de una fuente específica, una URL dada, mediante un sistema ETL (Extracción, Transformación y Carga) que puede ser implementado tanto localmente como en la nube. El propósito del ETL es extraer los datos de la fuente, transformarlos de manera adecuada, a través de procesos como la limpieza y la normalización, y finalmente cargarlos para su posterior almacenamiento o análisis.

Una vez procesados los datos mediante el ETL, estos se almacenan en AWS S3, el servicio de almacenamiento en la nube de Amazon, conocido por su capacidad para manejar grandes volúmenes de datos de manera segura. Paralela- mente, se emplea el concepto de bases de datos federadas, que permite gestionar y acceder a estos datos como si provinieran de una única fuente. Este enfoque facilita significativamente el análisis y la consulta de información que proviene de diferentes fuentes.

El último paso en este proceso es el uso de Python, un lenguaje ampliamente reconocido en el campo del análisis de datos, para dos tareas fundamentales: el entrenamiento de modelos de series de tiempo y la generación de predicciones. Estos modelos, entrenados con los datos almacenados, son luego utilizados para predecir tendencias futuras. Las predicciones generadas se reintegran al sistema para su análisis y para facilitar la toma de decisiones, completando así el ciclo de integración de datos. Este proceso integral representa un flujo de trabajo moderno en el ámbito de big data y análisis predictivo, combinando tecnologías avanzadas y procedimientos de gestión de datos.

6 Automatización de Datos

La automatización en el escenario descrito se manifiesta en varios aspectos clave del flujo de trabajo de integración de datos, mejorando la eficiencia y la eficacia del proceso completo.

Primero, la fase de ETL (Extracción, Transformación y Carga) representa un componente crítico de automatización. Al automatizar la extracción de datos de una URL específica y su posterior transformación y carga, se elimina la necesidad de intervención manual en estas etapas. Esto no solo acelera el proceso sino que también reduce los errores humanos, garantizando una mayor consistencia y fiabilidad en los datos recopilados y procesados.

En segundo lugar, el uso de almacenamiento en AWS S3 y bases de datos federadas incorpora otro nivel de automatización. El almacenamiento de datos en S3 se gestiona de forma automática, proporcionando escalabilidad y seguridad sin intervención manual. Las bases de datos federadas, por su parte, automatizan el acceso y la gestión de datos de múltiples fuentes, permitiendo que se traten como si fueran una única entidad. Esto simplifica enormemente las consultas y el análisis de datos, reduciendo la complejidad y el tiempo necesario para gestionar diversas fuentes de datos.

Finalmente, la automatización alcanza su punto culminante en el uso de Python para el entrenamiento de modelos de series de tiempo y la generación de predicciones. Este paso implica la utilización de algoritmos avanzados que pueden procesar grandes conjuntos de datos, aprender de ellos y realizar predicciones sin intervención humana continua. Esta capacidad no solo agiliza el proceso de análisis sino que también permite descubrir insights y tendencias que podrían no ser evidentes a través de métodos de análisis tradicionales.

En conjunto, estos elementos de automatización transforman el proceso de integración de datos en una operación más eficiente, precisa y escalable, permitiendo a la organización concentrarse en la interpretación y utilización de los insights generados en lugar de en los aspectos operativos de la gestión de datos.

7 IA

El uso más destacado de la IA se encuentra en el entrenamiento de modelos de series de tiempo utilizando Python. Estos modelos son un ejemplo de aprendizaje automático, una rama de la IA, donde el sistema aprende a identificar patrones y tendencias en los datos históricos. Este aprendizaje permite al modelo hacer predicciones o estimaciones sobre datos futuros. Al aplicar técnicas de aprendizaje automático, el modelo puede ajustarse continuamente a medida que recibe

nuevos datos, mejorando su precisión y relevancia a lo largo del tiempo.

Además, la generación de predicciones basada en estos modelos entrenados es otro aspecto donde la IA agrega valor significativo. Al analizar los datos actuales y su contexto histórico, los modelos de IA pueden predecir tendencias futuras y resultados potenciales con un alto grado de precisión. Esta capacidad es crucial para la toma de decisiones informadas y proactivas en diversos sectores, desde la predicción de precios hasta la anticipación de cambios en el mercado.

Por último, aunque de manera menos directa, la IA también puede estar implicada en la optimización de los procesos de ETL y en la gestión de bases de datos federadas. Algoritmos inteligentes pueden ser utilizados para mejorar la eficiencia de la extracción y transformación de datos, así como para la gestión eficaz de las consultas y el acceso a datos federados. Esto incluye la clasificación inteligente de los datos, la identificación de información relevante y la optimización de consultas para acelerar el acceso y el análisis.

8 Próximos pasos

- Fuente de Datos: La URL proporcionada sirve como una fuente de datos primaria, que valida las ventas de una gasolinera en un período determinado. Este es un paso crucial porque asegura que los datos recogidos son relevantes y actualizados, lo cual es fundamental para cualquier análisis posterior. En términos de mejoras futuras, se podría ampliar la recopilación de datos para incluir variables externas como condiciones climáticas o eventos locales, que podrían influir en las ventas de la gasolinera.

- ETL Local y en la Nube: Actualmente, el proceso ETL se lleva a cabo tanto localmente como en la nube. Este enfoque híbrido permite flexibilidad y escalabilidad. En el futuro, se podría automatizar aún más este proceso utilizando IA para identificar y corregir anomalías en los datos en tiempo real, mejorando así la calidad y la fiabilidad de los datos.

- Almacenamiento en AWS S3: El uso de AWS S3 para el almacenamiento de datos garantiza seguridad y escalabilidad. Una mejora futura podría ser la implementación de técnicas de compresión y cifrado de datos más avanzadas para optimizar el almacenamiento y mejorar la seguridad de los datos.

- Bases de Datos Federadas: Las bases de datos federadas permiten una gestión unificada de los datos. Una posible mejora sería la integración de capacidades de aprendizaje automático para optimizar las consultas de datos, haciendo que el proceso sea más eficiente y rápido.

- Entrenamiento de Modelo de Series de Tiempo con Python: El uso de Python para entrenar modelos de series de tiempo es esencial para entender las tendencias de ventas. Una mejora aquí podría ser la incorporación de modelos más sofisticados de IA, como redes neuronales recurrentes, que pueden capturar mejor las dependencias temporales complejas en los datos.

- Generación de Predicciones con Python y Carga a la Base de Datos Federadas: Después de generar predicciones con Python, cargar estas predicciones en las bases de datos federadas permite un acceso y análisis centralizado. Se podría mejorar este paso implementando algoritmos de IA que ajusten automáticamente los modelos en función del rendimiento de las predicciones pasadas.

- Visualización con Power BI: Finalmente, la visualización de datos con Power BI facilita la interpretación de los resultados. Una mejora aquí podría ser la integración de dashboards interactivos y en tiempo real, permitiendo a los usuarios explorar diferentes escenarios predictivos y comprender mejor el impacto de diversas variables en las ventas.

9 Lecciones aprendidas

- Importancia de la Calidad de los Datos: Una lección clave es la importancia de la calidad de los datos. La precisión de las predicciones y análisis depende directamente de la calidad de los datos recogidos de la fuente. Esto subraya la necesidad de mecanismos efectivos de validación y limpieza de datos en el proceso ETL.
- Flexibilidad del Almacenamiento en la Nube: El uso de AWS S3 para el almacenamiento de datos destaca la flexibilidad y escalabilidad que ofrece la nube. Este aspecto es crucial para gestionar volúmenes de datos en aumento y asegura que el sistema pueda adaptarse a las necesidades cambiantes del proyecto.
- Beneficios de la Automatización en ETL: La automatización del proceso ETL ahorra tiempo y reduce errores, lo que es esencial para mantener la eficiencia operativa. Esta lección es aplicable en muchos contextos donde el manejo de grandes conjuntos de datos es una necesidad.
- Valor del Aprendizaje Automático en Predicciones: El proyecto demuestra el valor del aprendizaje automático y la IA en la generación de predicciones precisas. Los modelos de series de tiempo, por ejemplo, pueden revelar tendencias ocultas y ayudar en la toma de decisiones proactivas.
- Importancia de la Visualización de Datos: La integración de Power BI para la visualización de datos resalta la importancia de presentar la información de manera comprensible. Las visualizaciones efectivas son cruciales para interpretar correctamente los análisis y compartir insights con aquellos que no son expertos en datos.
- Necesidad de Integración y Flexibilidad de Sistemas: El proyecto enseña la importancia de tener sistemas que puedan integrarse eficientemente. Las bases de datos federadas son un buen ejemplo de cómo distintas fuentes de datos pueden ser combinadas para un análisis más holístico.
- Adaptabilidad y Escalabilidad: Finalmente, el proyecto ilustra la necesidad de adaptabilidad y escalabilidad en el análisis de datos. A medida que la empresa crece o cambian las condiciones del mercado, la capacidad de ajustar rápidamente los modelos y procesos es crucial para mantener la relevancia y eficacia del análisis.

10 Bibliografía

References

Baron, D. (2018). *Cloud Storage Forensics*. Syngress.

Chollet, F. (2017). "Keras: Deep Learning for Humans". In: *GitHub Repository*.

Dean, J. and S. Ghemawat (2008). "MapReduce: Simplified Data Processing on Large Clusters". In: *Communications of the ACM* 51.1, pp. 107–113.

Hyndman, R.J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice*. OTexts.

Kimball, R. and M. Ross (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley.

Mehta, A. and S. Bhattacharya (2019). *Pro Power BI Desktop*. Apress.

Molinaro, A. (2021). *Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases*. Springer.