UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE GUAYAQUIL

CARRERA DE COMPUTACIÓN

AUTOMATIZACIÓN DE UN PROCESO ETL PARA LA INDUSTRIA FARMACÉUTICA MEDIANTE UN MODELO PREDICTIVO BASADO EN POWER BI

Trabajo de titulación previo a la obtención del
Título de Ingeniero en Ciencias de la Computación

AUTORES: PAMELA KERLY CAJAMARCA PALMA

JONATHAN TOMMY HAZ GUAMÁN

TUTOR: GALO ENRIQUE VALVERDE LANDIVAR

Guayaquil – Ecuador

2024

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Pamela Kerly Cajamarca Palma con documento de identificación N° 0951451327 y Jonathan Tommy Haz Guamán con documento de identificación N° 0954182317; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Guayaquil, 05 de febrero del año 2024

Atentamente,

Pamela Kerly Cajamarca Palma

0951451327

Jonathan Tommy Haz Guamán

0954182317

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Pamela Kerly Cajamarca Palma con documento de identificación No. 0951451327 y Jonathan Tommy Haz Guamán con documento de identificación No. 0954182317, expresamos nuestra voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: "Automatización de un proceso ETL para la industria farmacéutica mediante un modelo predictivo basado en Power BI", el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

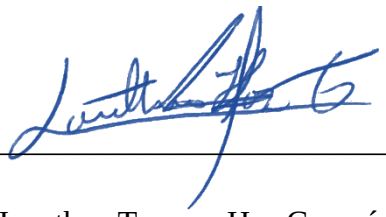
En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Guayaquil, 05 de febrero del año 2024

Atentamente,

Pamela Kerly Cajamarca Palma

0951451327



Jonathan Tommy Haz Guamán

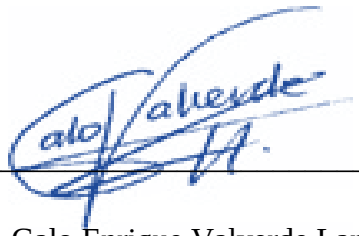
0954182317

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Galo Enrique Valverde Landivar con documento de identificación N° 0912511532, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: AUTOMATIZACIÓN DE UN PROCESO ETL PARA LA INDUSTRIA FARMACÉUTICA MEDIANTE UN MODELO PREDICTIVO BASADO EN POWER BI, realizado por Pamela Kerly Cajamarca Palma con documento de identificación N° 0951451327 y Jonathan Tommy Haz Guamán con documento de identificación N° 0954182317, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Guayaquil, 05 de febrero del año 2024

Atentamente,



Ing. Galo Enrique Valverde Landivar, Msc
0912511532

DEDICATORIA

Este trabajo se la dedico a mi Dios quién supo guiarme por el buen camino, darme fuerzas para seguir adelante y no desmayar en los problemas que se presentaban, enseñándome a encarar las adversidades sin perder nunca la dignidad ni desfallecer en el intento. A mi familia quienes por ellos soy lo que soy. Para mis padres y mi prometida por el apoyo de cada uno de ellos con sus consejos, comprensión, amor, ayuda en los momentos difíciles, y por ayudarme con los recursos necesarios para estudiar. Me han dado todo lo que soy como persona, mis valores, mis principios, mi carácter, mi empeño, mi perseverancia, mi coraje para conseguir mis objetivos.

Jonathan Tommy Haz Guamán

Dedico este trabajo a mis padres que no me abandonaron en ningún momento y no falto ese sustento económico para que yo pudiera seguir en mis estudios, a mis hermanos por haberme ayudado en cada momento en mi trayecto profesional y ser la persona quien soy ahora.

Pamela Kerly Cajamarca Palma

AGRADECIMIENTO

Primeramente, agradezco a la Universidad Politécnica Salesiana por haberme aceptado ser parte de ella y abierto las puertas para poder estudiar mi carrera, así como también a los diferentes docentes que brindaron sus conocimientos y su apoyo para seguir adelante día a día. Agradezco también a mi tutor el Ing. Galo Valverde por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento, así como también haberme tenido toda la paciencia del mundo para guiarme durante todo el desarrollo. Y para finalizar, también agradezco a todos los que fueron mis compañeros de clase: Pamela, Douglas, Richard y Karen por haberme acompañado durante todos los niveles de Universidad ya que gracias al compañerismo, amistad y apoyo moral han aportado en un alto porcentaje o mis ganas de seguir adelante en mi carrera profesional.

Jonathan Tommy Haz Guaman

Agradezco a mis padres por haber invertido en mi vida estudiantil de forma económica, sin ellos estaría pasando momentos difíciles, así que agradezco tenerlos en mi vida. Doy gracias a mis hermanos que fueron mis consejeros y no perdiera mi rumbo en mi trayecto profesional. Por último, doy gracias a mis amigos de la universidad, trabajo y a mis mejores amigas Daysi y Mary por su apoyo moral para seguir con la carrera. No puedo estar mas agradecidas con las personas que me rodean, en serio, gracias.

Pamela Kerly Cajamarca Palma

RESUMEN

Este estudio presenta un análisis sobre la automatización del proceso ETL en la industria farmacéutica y su integración con modelos predictivos en Power BI. La investigación se llevó a cabo para determinar el impacto tecnológico de una solución digital en un esquema empresarial, con el uso de indicadores para conocer su desempeño durante la toma de decisiones gerenciales y su mejora del tiempo de espera en el procesamiento de datos. Con la metodología propuesta, se abarcó desde la recopilación de datos hasta la conexión con Power BI y la evaluación de la eficacia en el sistema desarrollado. Se empleó Python para la extracción y transformación de los datos, asegurando la coherencia y eficiencia en el manejo de grandes conjuntos de información, en conjunto con la carga de datos, se realizó en un DataWarehouse, y la conexión con Power BI permitió un análisis detallado de patrones y tendencias. Y, por último, se integró la implementación de modelos predictivos de la plataforma, que proporcionó indicadores sobre la demanda de medicamentos en la industria. Los resultados obtenidos, destacó mejoras significativas en la eficiencia operativa al reducir tareas manuales, que permitió al personal de salud enfocarse en tareas más específicas y que no sean repetitivas. Las pruebas exhaustivas demostraron su funcionalidad y precisión del sistema, respaldando su impacto positivo en la agilidad operativa y en la toma de decisiones gerenciales.

Palabras Clave: ETL, Datawarehouse, Inteligencia de Negocios, Python, Machine Learning, Power BI, Industria Farmacéutica

ABSTRACT

This study presents an analysis of the automation of the ETL process in the pharmaceutical industry and its integration with predictive models in Power BI. The research was carried out to determine the technological impact of a digital solution in a business scheme, with the use of indicators to know its performance during management decision making and its improvement of the waiting time in data processing. The proposed methodology covered from data collection to the connection with Power BI and the evaluation of the efficiency of the developed system. Python was used for data extraction and transformation, ensuring consistency and efficiency in the handling of large sets of information, in conjunction with data loading, was performed in a DataWarehouse, and the connection with Power BI allowed a detailed analysis of patterns and trends. And finally, the implementation of predictive models of the platform was integrated, which provided indicators on the demand for drugs in the industry. The results obtained highlighted significant improvements in operational efficiency by reducing manual tasks, allowing healthcare personnel to focus on more specific, non-repetitive tasks. Extensive testing demonstrated the functionality and accuracy of the system, supporting its positive impact on operational agility and management decision making.

Keywords: ETL, Datawarehouse, Business Intelligence, Python, Machine Learning, Power BI, Pharmaceutical Industry

ÍNDICE DE CONTENIDO

1. INTRODUCCIÓN

La industria farmacéutica ha experimentado un rápido avance tecnológico y una creciente dependencia de la recopilación y análisis de datos para respaldar la investigación, el desarrollo, producción y distribución de medicamentos de una manera muy factible, permitiendo así establecer varias tecnologías para el apoyo en la toma de decisiones, como es el caso de los procesos ETL (Extracción, Transformación y Carga). Este proceso se ha convertido en una herramienta fundamental en el manejo de datos garantizando la calidad y consistencia de los datos, así como prepararlos para su uso en informes, análisis de negocios y otras aplicaciones (Acosta Diaz, 2020).

A pesar de la eficacia de este proceso, su ejecución manual conlleva una carga administrativa significativa y se enfrenta a desafíos sustanciales que impactan la integridad y confiablidad de la información crítica para la producción y distribución de los productos farmacéuticos. La manipulación manual de grandes volúmenes de datos implica la intervención humana en cada fase del proceso ETL, aumentando la probabilidad de errores críticos y resultando en omisiones, inserciones incorrectas o interpretaciones erróneas, comprometiendo la calidad y precisión de información procesada.

Entre otras tecnologías que serviría de apoyo seria en la incorporación de herramientas de análisis avanzado, como Power BI. Es una plataforma de análisis de negocios que permite visualizar datos de manera interactiva, lo que permite a los profesionales de la salud visualizar datos complejos de manera clara y concisa. Esto puede concluir a una compresión más profunda en los resultados y la identificación de oportunidades y áreas de mejora. Otra característica que tiene la plataforma es la implementación de modelos predictivos para aprovechar datos históricos y en tiempo real para anticipar eventos futuros, como la demanda de medicamentos, el comportamiento de pacientes, o la eficacia de ciertos tratamientos (García Estrella et al., 2021).

Con la implementación de la automatización en el proceso ETL, se reduce el tiempo necesario para preparar los datos, lo que acelera la toma de decisiones basada en información precisa y actualizada. Con ello se reduce la probabilidad de errores humanos mejorando la eficiencia operativa, permitiendo a los profesionales de salud centrarse en otras tareas especializadas, como el análisis de resultados y la toma de decisiones

estratégicas. En cuanto a la integridad y manejo de los datos, estos aseguran que estén conformes con los estándares regulatorios, lo que es crucial para cumplir con las normativas de la industria.

En este estudio, se busca resaltar el cómo estas tecnologías pueden facilitar en un esquema empresarial para conocer su desempeño durante la toma de decisiones gerenciales y su mejora del tiempo de espera en el procesamiento de datos. Para ello, el objetivo principal de este articulo consiste en proponer una automatización de proceso con ETL con la finalidad de desarrollar esa solución digital e integrarlo a un modelo predictivo de Machine Learning en Power Bi.

2. REVISIÓN DE LITERATURA

2.1. Big Data

El término "Big Data" se utiliza recientemente para referirse al procesamiento de grandes volúmenes de datos, diseñados para elevar la relevancia del conjunto de información convirtiéndolo en el activo más importante de una empresa. Uno de los factores más fundamentales en el proceso de desarrollo es la capacidad de convertir datos en información, así como la cantidad adecuada de datos pertinentes. Según (Araque González & Giampietro Torres, 2023) “este tipo de innovación permite un mayor control sobre los procesos, así como una mejor eficacia en el desempeño del conjunto de información”. Para (Ardagna et al., 2021) “Big data no solo significa una gran cantidad de datos, sino que apunta a escenarios donde los datos son diversos y que deben demostrarse que son dignos de confianza”. Antes de utilizar los datos para tomar decisiones cruciales, es necesario asegurarse de que los datos sean precisos y de alta calidad, porque en el contexto empresarial puede tener un impacto significativo en los resultados. Por ello Big Data se ha convertido en unas de las herramientas con mayor potencial. Una de las características que tiene Big Data es su capacidad de realizar análisis de datos predictivos, identificar tendencias en los comportamientos de los datos. “Permitir a las empresas a reducir costos y con ello el desarrollo de nuevos productos y servicios en base a las necesidades de los clientes” (Aversa et al., 2020). El uso del Big Data en el mundo digital es la combinación y el uso del machine learning, análisis de datos e inteligencia artificial con el fin de optimizar los diversos procesos posibles tanto operativos como de producción, mejorando la competitividad de las empresas con el uso de las herramientas. “Las empresas u organizaciones se tornan cada vez más digitales y como resultado, ocasiona un gran volumen de datos durante el proceso” (Dubey et al., 2020). Según datos recientes (Liang et al., 2023) “su aplicación se divide en tres partes: recopilación de datos desde diferentes fuentes, estructuración y limpieza de esta para su facilidad de análisis y por último la extracción de conocimiento”. Teniendo en cuenta estos procesos, esta investigación ayuda a manejar estos problemas comunes que maneja el área operativa y de producción al realizar estos procesos manuales manejando unos grandes volúmenes de datos en la transformación de la información.

Sin embargo, este tipo de propuestas plantea desafíos en las empresas en cuanto a niveles de desarrollo de valor que se esperan, haciendo que surjan necesidades como: la de rediseñar procesos para alcanzar la eficiencia y calidad de los datos, plantear esquemas de gestión en la recopilación de información requerida, contar con una infraestructura tecnológica que permita el almacenamientos de grandes volúmenes de datos, por ultimo personal capacitado para el análisis y gestión de los datos históricos, protección y velocidad en la entrega de informes. Para facilitar su uso según (Nti et al., 2022) "se deben precisar las fuentes de datos incompletas que perjudican la productividad de los datos, por esos motivos se necesita ser eficiente en el tema de automatizar los procesos de análisis de datos con el menor esfuerzo humano". Al hacerlo se implementa algoritmos que manejen automáticamente la información faltante. "Por lo tanto, este campo de la analítica en Big Data recibe mayor atención en el ámbito empresarial, porque se invierten más en tecnologías emergentes relacionadas con este tema para obtener una ventaja competitiva". (Dubey et al., 2020)

2.2. ETL y su importancia en la ciencia de datos

El modelo ETL como sus siglas lo dicen Extracción, Transformación y Carga, según (Acosta Diaz, 2020) "en el proceso de extracción, los datos pueden provenir desde una gran variedad de fuentes, tales como documentos de textos, hojas de cálculo, datos de la web, etc." Se requiere de herramientas para la extracción de los datos, estos pueden incluir como base de datos, sitios web y mucho más. Posteriormente, los datos en bruto pasan a la fase de transformación, eliminando datos nulos o filas y columnas innecesarias, corrección de errores, conversión de tipos de datos y agregación de datos como sumar o promediar valores. También se pueden aplicar filtros para que se muestren solo los datos relevantes para la empresa. En el último proceso, una vez que los datos son transformados, se cargan en un Data Warehouse que es un repositorio donde contiene datos estructurados para su análisis.

La analítica basada en datos es fundamental para las empresas en las tomas de decisiones estratégicas y operativas. Las organizaciones deben contar con un proceso para capturar datos desde múltiples fuentes de forma casi instantánea y cargarlo en un almacén de datos (Data Warehouse). Por esto automatizar el proceso ETL facilita en la toma de decisiones y mejora la eficiencia de las actividades operativas. Para (Biswas et al., 2022) "establecer

un proceso ETL automatizado para un modelo predictivo ayuda a las empresas a administrar los riesgos, capturar cambios en tiempo real". Además (Hira & Deshpande, 2023) "el modelo propuesto para la integración de datos en tiempo real ofrece una serie de ventajas, entre las que se incluyen una reducción significativa del tiempo y el coste de la construcción de modelos multidimensionales" lo que lo convierte en una opción atractiva para las empresas que necesitan tomar decisiones basadas en datos de forma rápida y eficaz.

2.3. Modelos De Datos Predictivos Con Machine Learning

Para (Dogan & Birant, 2021) "el aprendizaje automático o machine learning es un campo de la inteligencia artificial que ayuda a los ordenadores a aprender de los datos sin ser programados explícitamente". Los algoritmos de ML se entrenan para identificar patrones y hacer predicciones. Se han aplicado ampliamente para una gama de aplicaciones como: recomendación de productos, detección de fraudes, diagnóstico médico, conducción autónoma, etc. Al permitir a las empresas tomar decisiones más informadas y optimizar sus procesos, el ML puede ayudar a las empresas a mejorar su eficiencia y calidad. Dado que tiene un impacto directo en la productividad y en los costos.

La investigación hecha por (Urra-González & Ramos-Maldonado, 2023) "el ML se utiliza para encontrar patrones y tendencias de grandes volúmenes de datos, se entrenan con datos históricos que enseñan a los algoritmos a reconocer patrones". Una vez que los algoritmos son entrenados, se utiliza para predecir eventos futuros. "El procesamiento y limpieza de datos son tareas que se deben hacer antes de aplicar algoritmos de aprendizaje automático (Tarik et al., 2021)". Los datos sin filtrar son generalmente menos fiables e incompletos y su uso en la modelización puede conducir a resultados equivocados. Por ejemplo, si un modelo de aprendizaje automático se entrena con datos que contienen errores, el modelo puede aprender a identificar los errores como patrones. Esto puede dar lugar a predicciones inexactas o engañosas. Para evitar estos problemas, es importante filtrar los datos antes de utilizarlos en procesos de modelización. El filtrado de datos es el proceso de identificar y eliminar los errores, inconsistencias y valores atípicos de los datos.

Existen diferentes técnicas de ML que pueden utilizarse como una alternativa ideal y aparte tienen muchos beneficios en comparación con los modelos estadísticos: permiten hacer predicciones más precisas, se pueden entrenar en grandes conjuntos de datos mucho más eficientes, suelen ser más robustos ya que no dependen de suposiciones específicas, "se puede implementar sin la necesidad de un conocimiento profundo de la estadística, haciendo más accesibles para los analista de negocios que no tienen una formación del tema de estadística. (Viljanen et al., 2022)"

2.4. Power BI

Power BI es una poderosa plataforma de análisis de datos y visualización desarrollada por Microsoft que permite a las organizaciones transformar sus datos en información significativa y visualmente atractiva. Este conjunto de herramientas y servicios se ha convertido en una solución integral para la toma de decisiones basada en datos, ya que capacita a los usuarios para recopilar datos de diversas fuentes, modelarlos, realizar análisis avanzados y crear informes interactivos y paneles de control.

La principal fortaleza de Power BI radica en su capacidad para conectarse a una amplia gama de fuentes de datos, desde bases de datos relacionales hasta servicios en la nube como Azure, Excel, SharePoint, y muchas otras fuentes locales o en línea. Una vez que los datos se han importado, Power BI permite a los usuarios modelar y transformar esta información según sus necesidades mediante la creación de relaciones entre tablas, la aplicación de cálculos personalizados y la limpieza de datos.

Power BI proporciona herramientas para limpiar, transformar y modelar datos antes de utilizarlos en los informes. Esto es útil para asegurarse de que los datos sean precisos y estén listos para su análisis. Existe la necesidad de simplificar los procesos de BI, reducir el tiempo de respuesta y aumentar la eficiencia y eficacia. Se muestran varias formas de ejecutar la inteligencia de negocios, desde la fuente, es decir, la minería de datos, hasta la etapa final del proceso relacionada con la toma de decisiones.

En este sentido, se ofrecen alternativas innovadoras que, tras las investigaciones necesarias, han ido más allá de lo que hoy conocemos como BI, "permitiendo no sólo tomar decisiones, sino también ofrecer propuestas que cuentan con soporte automatizado

y que permiten procesar realmente los datos. por separado y procesar informes más realistas basados en datos de diferentes fuentes". (García Estrella et al., 2021)

Power BI es una herramienta versátil que puede ayudar a las empresas farmacéuticas a gestionar y analizar una amplia variedad de datos, desde ventas y marketing hasta investigación clínica y cumplimiento normativo. Esto puede conducir a una toma de decisiones más informada y eficiente en la industria farmacéutica. No se limita solo a la creación de informes estáticos; también permite la exploración de datos en tiempo real y el análisis ad hoc. Los usuarios pueden realizar consultas y realizar análisis exploratorios sobre los datos para descubrir patrones, tendencias y relaciones ocultas.

2.5. Datawarehouse

Un almacén de datos es una estructura de datos centralizada que almacena información histórica y actuales de una organización. Los datos se recopilan de diversas fuentes, como sistemas transaccionales, bases de datos relacionales y otras aplicaciones y se almacenan en un formato estandarizado para facilitar su análisis, donde los usuarios pueden acceder desde un almacén de datos a través de herramientas analíticas, como herramientas de inteligencia de negocios, clientes SQL y otras aplicaciones.

Las herramientas analíticas permiten a los usuarios realizar consultas, informes y análisis avanzados de los datos, siendo esta una herramienta fundamental para la toma de decisiones. "Al proporcionar una visión integral de los datos de una organización, permiten a los usuarios identificar tendencias, patrones y oportunidades que de otro modo serían invisibles". (Acosta Diaz 2020).

Su propósito según (Martinez Sanchez et al., 2020) "es para contener información histórica que ha sido previamente recolectada y transformada con el fin de que los datos sean cargados para su consulta" facilitando su administración y recuperación de la información. Para (Loja-Tepán et al., 2021) "un DataWarehouse no es más que una base de datos empresarial o corporativa, que se compone de diversas fuentes que la empresa cuenta". Estas deben ser confiables y su almacenamiento debe ser el adecuado para permitir su análisis desde diferentes entornos y perspectivas.

2.6. Business Intelligence

"Inteligencia de Negocios, es un campo interdisciplinario que combina tecnología, procesos y análisis de datos para ayudar a las organizaciones a comprender mejor su funcionamiento interno, el entorno empresarial y tomar decisiones más informadas" (Martínez Zabaleta & Rodríguez Luna, 2023). En esencia, el BI se enfoca en la recopilación, análisis y presentación de datos de manera significativa y accesible, lo que permite a las empresas transformar información en conocimiento. El concepto de BI abarca una amplia gama de actividades, desde la extracción y transformación de datos (ETL), donde se recopilan y preparan los datos de diversas fuentes, hasta el análisis y la visualización de datos, que permiten a los usuarios identificar patrones, tendencias y oportunidades clave. Estos datos pueden provenir de múltiples fuentes, como bases de datos empresariales, aplicaciones en la nube, redes sociales y mucho más.

"Es importante resaltar que un componente principal para la toma de decisiones son los Key Performance Indicator (KPI) que son valores que muestran el resultado de las acciones, que una organización, ejecutó" (Mendoza-Rivera, 2022). Permiten conocer, a manera resumida, el desempeño de una organización.

El flujo y la gestión adecuados de datos e información son esenciales para un proceso de toma de decisiones exitoso. Transferir esta estrategia al ámbito universitario supone dotar a profesores y administradores de sistemas que apoyen su toma de decisiones respecto de su actividad docente. La utilidad del BI se extiende a todas las áreas de una organización, desde la gestión de operaciones y finanzas hasta el marketing y la toma de decisiones estratégicas.

"Al brindar a los usuarios herramientas para explorar datos de manera interactiva y crear informes personalizados, el BI facilita la identificación de áreas de mejora y la optimización de procesos empresariales". (Mora, 2020)

Un aspecto esencial del BI es la creación de paneles de control y cuadros de mando, que ofrecen una representación visual en tiempo real de los datos clave de una organización. Estos paneles permiten a los tomadores de decisiones supervisar el rendimiento, evaluar el impacto de las decisiones y tomar medidas correctivas cuando sea necesario. En un entorno empresarial en constante cambio, el BI desempeña un papel crítico al proporcionar información oportuna y relevante que ayuda a las organizaciones a adaptarse y tomar decisiones estratégicas basadas en datos. Además, con la evolución de

tecnologías como el aprendizaje automático y la inteligencia artificial, el BI está avanzando hacia el análisis predictivo y prescriptivo, lo que permite a las empresas anticipar tendencias y tomar decisiones proactivas.

2.7. Python

Python es un lenguaje de programación de alto nivel, interpretado, orientado a objetos y con una semántica dinámica integrada. Cuenta con varias características, disponiendo de una amplia gama de bibliotecas y módulos que garantiza un rendimiento bastante alto y se lo puede compilar en varias plataformas. El propósito que tiene Python como lenguaje de programación son muchas, pero en las que más destaca es en el tema de análisis de datos, ya que se lo puede utilizar para analizar grandes volúmenes de datos haciendo notorio su eficiencia y rapidez. Según (Blank & Deb, 2020) Python se ha convertido en el lenguaje de programación preferido para proyectos de desarrollo que tengan que ver con análisis de datos, aprendizaje automático y profundo. Muchas áreas científicas, como la ingeniería, el análisis de datos y el aprendizaje profundo, dependen de la optimización. Estos campos están creciendo rápidamente y sus ideas se utilizan para una variedad de propósitos, como recopilar datos de un gran volumen o ajustar modelos de predicción precisos. Es crucial que un algoritmo se implemente de manera eficiente en un lenguaje de programación adecuado siempre que tenga que manejar una gran cantidad de datos.

Las librerías esenciales que se usan en análisis de datos son Numpy, que ayuda a realizar cálculos científicos y una estructura de matrices multidimensionales. Pandas es otra librería esencial que permite almacenar y estructurar datos de manera rápida, lo que facilita la manipulación de datos y proporciona una serie de operaciones matemáticas al igual que Numpy. Por último, tenemos la librería Selenium, es una herramienta poderosa que se usa para la automatización de diversas tareas como: realizar pruebas automatizadas de aplicaciones web, rastrear sitios web y recopilar datos, gestionar cuentas, etc. Por estas características Python es una herramienta ideal para muchos proyectos de investigación e industrias donde la investigación puede ser bastante complicada debido a estas características.

3. METODOLOGÍA

El método empírico analítico, de tipo experimental, permitió la evaluación sistemática de la implementación propuesta de la automatización del proceso ETL y la integración con modelos predictivos en Power BI en el contexto especifico de la industria farmacéutica.

3.1. Recopilación de Datos

Para la selección de los datos, se llevó a cabo una búsqueda exhaustiva para recopilar información crucial sobre la industria farmacéutica desde diversas fuentes. Entre las bases de datos, se encuentra Kaggle, una valiosa plataforma que ofrece conjuntos de datos diversificados que permiten a los profesionales acceder a información relevante. El repositorio seleccionado constó de más de 2000 registros, que indican la fecha y hora de venta, la marca del medicamento y la cantidad vendida.

3.2. Extracción de los datos

En esta fase, se implementó un sistema de extracción de datos utilizando Python como la herramienta principal debido a su flexibilidad, amplia gama de bibliotecas y eficacia en el manejo de tareas relacionadas con la recopilación de datos (Mukhopadhyay, 2018). Con los datos que se recopiló anteriormente, fueron ubicadas en un disco local como un archivo de valores separados por comas (csv) y se ejecutó la extracción con la librería Pandas que es especializada para el manejo y análisis de estructuras de datos.

3.3. Transformación de Datos

Después de la extracción de la base de datos internas u otro tipo de repositorio sobre la industria farmacéutica, los datos pasaron por un proceso de transformación. Se identificó y manejó datos nulos, duplicados o inconsistentes según sea necesario para garantizar la coherencia en el análisis posterior (Acosta Diaz, 2020). Y se realizó otros tipos de transformaciones adicionales según los requisitos específicos del dominio farmacéutico, esto puede incluir la normalización de unidades, cálculos específicos o agregaciones.

3.4. Carga de Datos

Cuando la información fue procesada en el formato correcto, se llevó a cabo la carga de datos en un repositorio que será nuestro DataWarehouse (Loja-Tepán et al., 2021), con información esencial de la industria farmacéutica. Se aseguró de que los datos estén disponibles de manera eficiente y estructurada para análisis posteriores. Nuestro

DataWarehouse procesada se ubicó en la base de datos MySQL Workbench que como se muestra en la Figura 1 tuvo como nombre pharma_sales.

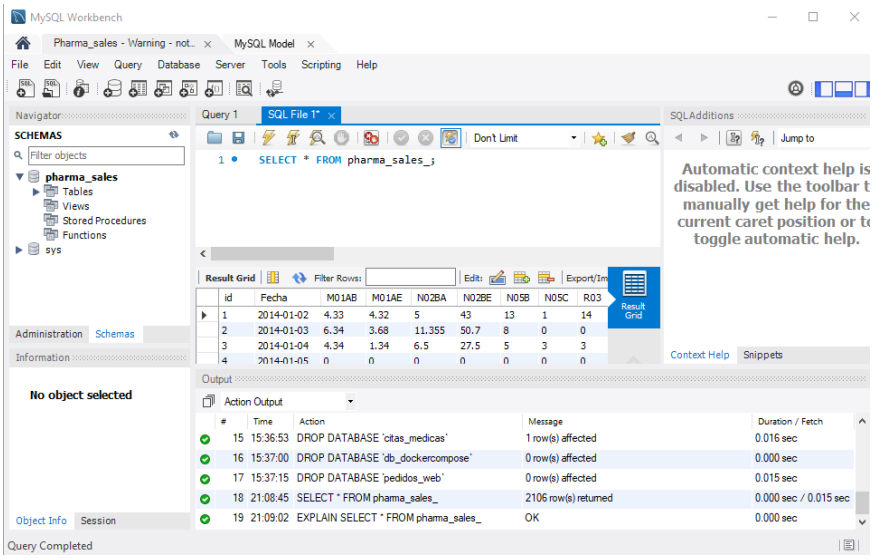


Figura 1. Repositorio de Datos Pharma_sales

3.5. Conexión entre el repositorio de Datos y Power Bi

Al ingresar a la plataforma, se estableció una integración entre nuestro DataWarehouse y Power BI mediante credenciales de conexión. En la figura 2, después de ingresar las credenciales, se importó de manera correcta los datos y permitió ver la tabla o vista previa del repositorio.

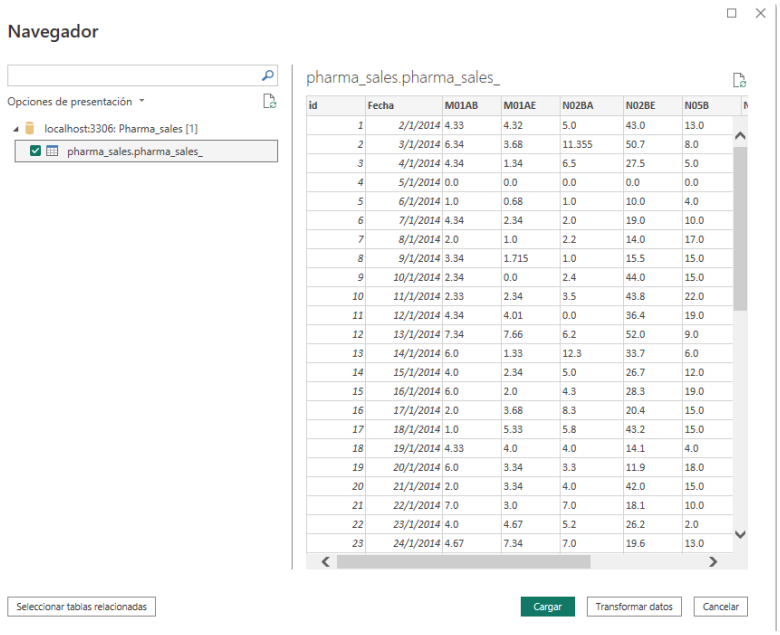


Figura 2. Conexión entre MySQL y Power BI

3.6. Análisis de Datos

Este paso implicó un análisis detallado de los datos procesados en Power BI. Se incluyó la exploración de patrones, tendencias y relaciones ocultas de los datos, que proporcionaran información valiosa para la toma de decisiones en la industria farmacéutica, como seguimiento de ventas, análisis de tendencias y evaluación de la eficacia de ciertos productos. Que proporcionó a los profesionales de la salud los conocimientos necesarios para mejorar la eficiencia operativa y la toma de decisiones estratégicas.

3.7. Integración de modelos predictivos en Power BI

Se implementó modelos compatibles en la plataforma de Power BI para la anticipación de eventos futuros, como la demanda de medicamentos o tendencias en la industria farmacéutica. Esto se incluyó como un forecast que es una predicción acerca de un valor, permitió el análisis predictivo a través de utilizar patrones y tendencias pasadas en los datos para predecir eventos futuros.

3.8. Evaluación de los datos

Se realizó pruebas exhaustivas para validar la funcionalidad y eficacia en el sistema desarrollado. Por ello, se centró en la precisión, confiabilidad y eficiencia en el proceso ETL (Hira & Deshpande, 2023). Además, se validó la precisión de modelos predictivos implementados, comparando con predicciones de datos reales para evaluar la confiabilidad de los modelos.

Es importante mencionar que la disponibilidad y calidad de los datos pueden variar, lo que podría afectar la generalización de los resultados. Por ello, la adaptación a cambios rápidos en los requisitos del negocio farmacéutico deberá incluir procesos agiles y prácticas de desarrollo, que permitirán ajustes rápidos en respuestas a nuevas necesidades comerciales o descubrimientos en la investigación farmacéutica, como las tendencias del mercado, avances tecnológicos y modificaciones en las regulaciones gubernamentales que afectan a la industria. Este enfoque garantizará que la solución siga siendo valiosa y eficaz en un entorno empresarial dinámico.

4. RESULTADOS

La eficiencia operativa tuvo una mejora significativa al reducir las tareas manuales asociadas con el modelo ETL sobre la industria farmacéutica, la automatización de este proceso mediante Python simplifica la extracción y procesamiento de datos, eliminando la necesidad de intervenciones manuales repetitivas. A través de esta implementación, permite al personal centrarse en tareas más específicas en lugar de actividades rutinarias y propensa a errores.

Durante las pruebas, la solución demuestra la capacidad que tiene de extraer una gran carga de datos sin ningún tipo de demora, demostrando que Python es la elección predominante para entornos de análisis de datos. También se evidencia la habilidad para manejar datos nulos, duplicados o inconsistentes, incluyendo la identificación y gestión de datos redundantes y así como otras transformaciones necesarias para garantizar la coherencia. En conjunto, la carga de data procesada en el repositorio ubicado en MySQL se lleva a cabo una vez que la información ha sido adecuadamente transformada y estructurada.

Con la integración de Power BI, la conexión entre el repositorio de datos y la plataforma es estable, y posibilita la representación visual clara de los datos relevantes para la industria farmacéutica, utilizando herramientas como gráficos, tablas dinámicas y otros elementos visuales intuitivos.

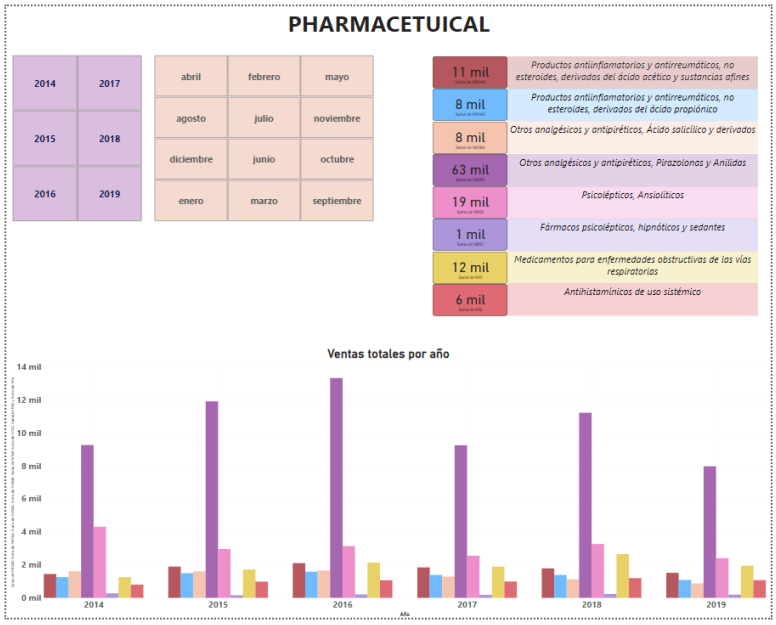


Figura 3. Indicadores de farmacéuticos vendidos por año

En la figura 3 se muestran capturas de pantalla con los indicadores principales acerca de las ventas totales por año, aplicando filtros por año y mes e incluso una descripción de cada venta total sobre los farmacéuticos vendidos.

Además de la visualización de los datos, se implementa modelos predictivos compatibles con la plataforma, lo que permite anticipar eventos futuros, aspectos como la demanda de medicamentos, todo ello puede ser previstos mediante herramientas de análisis como forecast o previsión

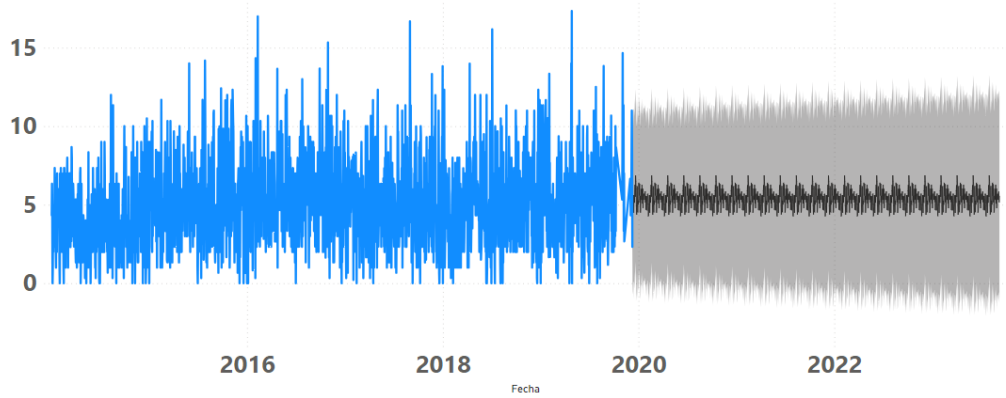


Figura 4. Análisis predictivo en Power BI

A continuación, en la figura 4, el grafico de líneas muestra los valores pronosticados para los años posteriores en función de los datos históricos sobre un producto farmacéutico. El pronóstico generado es una estimación basada en la información dada en los años anteriores.

Una vez concluida con la implementación, se lleva a cabo pruebas exhaustivas para validad la funcionalidad y eficacia del sistema desarrollado, centrándose en múltiples aspectos clave. Esto incluye la validación de la integridad de los datos, la detección de posibles errores y la confirmación de que los datos se han procesado de manera precisa y completa.

5. CONCLUSIÓN

Al evaluar las limitaciones esta área sobre la automatización de procesos ETL y modelos predictivos, el modelo ha proporcionado una visión clara que identificó y evaluó desafíos y restricciones para la implementación de esta solución. Por ello, el articulo permitió conocer de mejor manera el tema y así el desarrollo de soluciones efectivas.

Durante las pruebas, desarrollar una solución digital basada en el lenguaje de programación Python con la finalidad de automatizar un modelo ETL e integrarlo a un modelo predictivo de Machine Learning en Power BI; la solución demostró su capacidad para extraer grandes volúmenes de datos de manera eficiente, sin demoras significativas, también la habilidad para manejar datos nulos, duplicados o inconsistentes. La integración exitosa con Power BI facilitó la representación visual clara de los datos relevantes para la industria, como la demanda de medicamentos.

Sobre el impacto tecnológico de la solución digital en un esquema empresarial, con el uso de indicadores para conocer su desempeño durante la toma de decisiones gerenciales y su mejora del tiempo de espera en el procesamiento de datos, se ha demostrado que la solución ha tenido un impacto positivo al reducir el tiempo de espera, esta capacidad de preprocesamiento mejora la agilidad operativa al garantizar que los datos estén disponibles para el análisis casi en tiempo real.

6. REFERENCIAS

Acosta Diaz, F. J. D. Á. J. L. A. (2020). *Solución de inteligencia de negocios para automatizar el proceso de selección y evaluación del proyecto "Lectores de Paso"* [Universidad Peruana de Ciencias Aplicadas (UPC)]. https://doi.org/10.19083/tesis/653633

Araque González, G. A., & Giampietro Torres, V. J. (2023). El Big Data aplicado en la industria 4.0 : un caso en el sector textil colombiano con un enfoque en la inteligencia de negocios. *Cuaderno Activa*, *14*(1). https://doi.org/10.53995/20278101.1176

Ardagna, C. A., Bellandi, V., Bezzi, M., Ceravolo, P., Damiani, E., & Hebert, C. (2021). Model-Based Big Data Analytics-as-a-Service: Take Big Data to the Next Level. *IEEE Transactions on Services Computing*, *14*(2), 516–529. https://doi.org/10.1109/TSC.2018.2816941

Aversa, J., Hernandez, T., & Doherty, S. (2020). Spatial Big Data and Business Location Decision-Making: Opportunities and Challenges. In *Regional Intelligence* (pp. 205–224). Springer International Publishing. https://doi.org/10.1007/978-3-030-36479-3_11

Biswas, N., Mondal, A. S., Kusumastuti, A., Saha, S., & Mondal, K. C. (2022). Automated credit assessment framework using ETL process and machine learning. *Innovations in Systems and Software Engineering*. https://doi.org/10.1007/s11334-022-00522-x

Blank, J., & Deb, K. (2020). Pymoo: Multi-Objective Optimization in Python. *IEEE Access*, *8*, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., & Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, *226*, 107599. https://doi.org/10.1016/j.ijpe.2019.107599

García Estrella, C. W., Barón Ramírez, E., & Sánchez Gárate, S. K. (2021). La inteligencia de negocios y la analítica de datos en los procesos empresariales. *Revista Científica de Sistemas e Informática*, *1*(2), 38–53. https://doi.org/10.51252/rcsi.v1i2.167

Hira, S., & Deshpande, P. S. (2023). Automated heuristic based context dependent <scp>ETL</scp> process to generate multi-dimensional model for tabular data. *Concurrency and Computation: Practice and Experience*, *35*(2). https://doi.org/10.1002/cpe.7459

Liang, R., Huang, C., Zhang, C., Li, B., Saydam, S., & Canbulat, I. (2023). Exploring the Fusion Potentials of Data Visualization and Data Analytics in the Process of Mining Digitalization. *IEEE Access*, *11*, 40608–40628. https://doi.org/10.1109/ACCESS.2023.3267813

Loja-Tepán, M. G., Bermeo-Pazmiño, K. V., & Cisneros-Quintanilla, D. P. (2021). Inteligencia de Negocios aplicado al área técnica en una empresa de Telecomunicaciones. *CIENCIAMATRIA*, *7*(12), 147–177. https://doi.org/10.35381/cm.v7i12.424

Martinez Sanchez, C. A., Murcia López, A. F., Cortes Querales, M. F., & Posada Charum, M. C. (2020). *Solución de inteligencia de negocios para el aprovechamiento de los datos del servicio médico en una entidad bancaria* [Pontificia Universidad Javeriana]. https://doi.org/10.11144/Javeriana.10554.62064

Martínez Zabaleta, M. E., & Rodríguez Luna, R. E. (2023). Inteligencia empresarial y su rol en la generación de valor en los procesos de negocios. *Tendencias*, *24*(1), 226–251. https://doi.org/10.22267/rtend.222302.222

Mendoza-Rivera, R. D. (2022). *Inteligencia de Negocios para Agilizar la Toma de Decisiones en la Gestión de Pacientes de Policlínicos de Salud*. 187–191. https://doi.org/10.54808/CISCI2022.01.187

Mora, G. (2020). Influencia de la inteligencia de negocios en los procesos de toma de decisiones dentro de las instituciones financieras. *Realidad Empresarial*, *10*, 21–24. https://doi.org/10.5377/reuca.v0i10.10574

Mukhopadhyay, S. (2018). *Advanced Data Analytics Using Python*. Apress. https://doi.org/10.1007/978-1-4842-3450-1

Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Mining and Analytics*, *5*(2), 81–97. https://doi.org/10.26599/BDMA.2021.9020028

Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, *184*, 835–840. https://doi.org/10.1016/j.procs.2021.03.104

Urra-González, C., & Ramos-Maldonado, M. (2023). Un enfoque de machine learning para la predicción de la calidad de tableros contrachapados. *Maderas. Ciencia y Tecnología*, *25*. https://doi.org/10.4067/S0718-221X2023000100436

Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, *21*(1), 4. https://doi.org/10.1186/s12942-022-00304-5