ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Sociales y Humanísticas

Aplicación de machine learning para el control interno de la identificación de anomalías y proyección de mermas en la producción de papel higiénico: un enfoque de auditoría

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Licenciado en Auditoría y Control de Gestión

Presentado por:

Anibal Samuel Flores de Valgas Williams
Victor Andrés García Moreno

GUAYAQUIL – ECUADOR

Año: 2023

DEDICATORIA

Dedico este logro con gratitud a Dios, fuente inagotable de fortaleza. Dedicada a mi amada madre, Ruth, cuyo amor incondicional y sacrificio incansable han sido mi inspiración constante. A mi querido padre, Ángel, cuyo apoyo inquebrantable y sabias palabras me han inspirado a superar obstáculos y a crecer en cada paso. A mi querida hermanita Dora y a mi hermano mayor Elian.

Y esta dedicación es para todos aquellos que iluminaron mi camino con su amistad, aliento y paciencia.

Victor Andres Garcia Moreno

DEDICATORIA

Dedico este trabajo, en primer lugar, a Dios, que ha iluminado mi camino con su orientación.

A mis queridos padres Jimmy y Maria, cuyo amor incondicional y apoyo constante han cimentado mi éxito. A mi apreciada Abuelita Mercy, cuyo respaldo constante ha sido un faro de esperanza en este proceso. A mi abuelito, aunque ya no esté entre nosotros, su deseo de presenciar mi prosperidad continúa impulsando cada uno de mis pasos.

Y finalmente, a mis adoradas hermanas, mi motor inspirador. Espero que este proyecto no solo refleje mi esfuerzo, sino también la promesa de un futuro que les haga sentir orgullosas y felices.

Anibal Samuel Flores de Valgas Williams

AGRADECIMIENTO

Quiero tomar un momento para expresar mi sincero agradecimiento a dos pilares fundamentales en mi vida, mi querido padre Ángel y mi amada madre Ruth. Vuestra constante dedicación, apoyo inquebrantable y amor desinteresado han sido la fuerza motriz detrás de todos mis logros. Vuestras palabras de aliento en los momentos de duda, vuestra orientación sabia en los momentos de incertidumbre y vuestro cariño en cada paso del camino me han dado la confianza y la determinación para alcanzar este hito en mi vida.

No podría estar más agradecido a mi familia extensa y amigos, quienes me han brindado su apoyo cuando más lo necesitaba.

Victor Andres Garcia Moreno

AGRADECIMIENTO

En este momento tan significativo, deseo expresar mi profundo agradecimiento a quienes han sido los cimientos de mi trayectoria. A mis padres, les agradezco por su apoyo inquebrantable y amor constante, brindándome de guía y fuerza a lo largo de este camino. A mis abuelos, mi gratitud por su sabiduría y afecto que han iluminado mi sendero.

Agradezco a mis amigos, sus risas, apoyo y presencia en cada capítulo de mi travesía. También quiero agradecer a todas las personas que han sido parte de mi vida, cada encuentro y experiencia ha contribuido a mi desarrollo. Siendo este logro el resultado de la sinergia de todas estas relaciones.

Anibal Samuel Flores de Valgas Williams

RESUMEN

La industria ecuatoriana de fabricación de papel higiénico se enfrenta al importante desafío de controlar y reducir las mermas durante el proceso de fabricación. Las variaciones significativas en los porcentajes de desperdicio en la empresa "PapelGo" han llevado a costos innecesarios y una disminución de la rentabilidad. Una falta de control en la producción de mermas en 2022 generó mermas superiores al 5.5% permitido y aceptable, lo que hizo que se necesitara una solución.

Se recomienda la implementación de métodos de machine learning, particularmente el modelo de memoria de larga duración LSTM. En la producción de papel higiénico, este método permite la detección precisa de anomalías y la proyección de mermas eficientes. El modelo LSTM identifica patrones complejos y relaciones causales entre múltiples variables interrelacionadas al analizar datos históricos y en tiempo real. Esto brinda información útil para la optimización de procesos y la toma de decisiones inteligentes.

A medida que se obtienen más datos, el modelo LSTM aprovecha su capacidad de capturar relaciones temporales y su aprendizaje continuo, lo que resulta en una detección y proyección de mermas más precisas. Esta nueva solución no solo mejora la eficiencia operativa y reduce costos, sino que también actualiza las prácticas de auditoría y control de la empresa "PapelGo", lo que aumenta su competitividad y sostenibilidad en el mercado.

Palabras Claves: Desperdicio, Machine Learning, LSTM, Detección de Anomalías, Proyección de Merma.

ABSTRACT

The Ecuadorian toilet paper manufacturing industry faces the important challenge of controlling and reducing waste during the manufacturing process. Significant variations in waste percentages in the company "PapelGo" have led to unnecessary costs and a decrease in profitability. A lack of control in the production of shrinkage in 2022 generated shrinkage in excess of the 5.5% allowed, which led to the need for a solution.

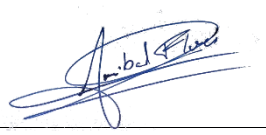
The implementation of machine learning methods is recommended, particularly the LSTM long-time memory model. In toilet paper production, this method enables accurate anomaly detection and efficient shrinkage projection. The LSTM model identifies complex patterns and causal relationships between multiple interrelated variables by analyzing historical and real-time data. This provides useful information for process optimization and intelligent decision making.

As more data is obtained, the LSTM model leverages its ability to capture temporal relationships and continuous learning, resulting in more accurate shrink detection and projection. This new solution not only improves operational efficiency and reduces costs, but also updates the audit and control practices of the "PapelGo" company, increasing its competitiveness and sustainability in the market.

Key words: Wastage, Machine Learning, LSTM, Anomaly Detection, Shrinkage Projection.

Declaración Expresa

"Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Flores de Valgas Williams Anibal Samuel y García Moreno Victor Andres y damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"



Flores de Valgas Williams
Anibal Samuel

García Moreno Victor
Andrés

Evaluadores

Firmado electrónicamente por:
BENIGNO ALFREDO
ARMIJOS DE LA CRUZ

Msc. Benigno Armijos De La Cruz

Profesor de Materia

Msc. Jessica Espinoza Tóala

Tutor de proyecto

Contenido

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

CAPÍTULO I

1. INTRODUCCIÓN

En un entorno empresarial altamente competitivo, lograr una eficiente gestión de costos es fundamental para garantizar la rentabilidad y la continuidad de cualquier organización. La producción de papel higiénico, como uno de los productos de consumo masivo más demandados, no es ajena a esta realidad. Las empresas del sector se enfrentan a diversos desafíos, entre ellos, la identificación y corrección oportuna de anomalías que puedan surgir durante el proceso productivo.

La auditoría y control de gestión desempeñan un papel crucial en este contexto, brindando una perspectiva objetiva y analítica de los procesos internos de una organización. En particular, la aplicación de herramientas avanzadas como el Machine Learning ha demostrado su capacidad para detectar patrones y anomalías en grandes volúmenes de datos, proporcionando insights valiosos y facilitando la toma de decisiones informadas.

En este sentido, el presente proyecto se centra en la implementación de una solución basada en Machine Learning para identificar y analizar las anomalías de producción en la fabricación de papel higiénico que generan mermas de producción, con el fin de optimizar los costos asociados a estas irregularidades. Para ello, se utilizarán técnicas de análisis de datos y algoritmos de aprendizaje automático que permitan identificar patrones anómalos, establecer relaciones causales y proponer acciones correctivas.

El plazo establecido para la realización de este proyecto es de tres meses, durante los cuales se llevará a cabo un exhaustivo análisis de los datos proporcionados, se implementarán los modelos de Machine Learning adecuados y se realizarán pruebas para validar la eficacia y la aplicabilidad de la solución propuesta. A lo largo del desarrollo del proyecto, se prestará especial atención a la ética y confidencialidad de la información manejada, garantizando la protección de los datos sensibles de la empresa.

Se espera que los resultados obtenidos aporten conocimientos relevantes sobre la detección y prevención de anomalías en la producción de mermas de papel higiénico, y que las recomendaciones propuestas contribuyan a la optimización de los costos asociados a estas irregularidades.

Se explorará la aplicación del Machine Learning como una herramienta innovadora y eficiente en la identificación de anomalías de producción. A través de la implementación de técnicas analíticas y algoritmos avanzados, se pretende brindar soluciones prácticas y concretas a una problemática relevante en el sector empresarial, contribuyendo así al desarrollo y la competitividad de las organizaciones.

1.1. Descripción del problema

En Ecuador, la industria de producción de papel higiénico cuenta con un número considerable de empresas dedicadas a esta actividad. Aunque no se dispone de datos precisos sobre el número exacto de industrias de producción de papel higiénico en el país, se estima que existen varias decenas de empresas dedicadas a esta actividad.

La fabricación de papel higiénico es un proceso complejo que involucra múltiples etapas y factores, desde la elección de materia prima hasta la producción del producto final. La industria de papel higiénico o “PapelGo” maneja cuatro categorías de papel higiénico que son: PH CLÁSICO DH, PH CLÁSICO TH, PH KIND DH, PH MCPROPIA DH y PH MCPROPIA TH. Durante el proceso de producción de estos productos surgen diversas anomalías que afectan a la calidad final del producto o que producen mermas de papel. Estas anomalías pueden incluir desperdicios excesivos de materiales, tiempos de inactividad prolongados de maquinaria, fluctuaciones en la eficiencia energética o mala planificación.

El problema por resolver, en cuestión se relaciona con la gestión ineficiente de las mermas en la producción de papel higiénico. Estas mermas son los sobrantes de papel generados durante el proceso de corte y deben mantenerse dentro de un margen aceptable del 5.5%, establecido por PapelGo. Sin embargo, se observa una variación significativa en los porcentajes de mermas de producción, se puede observar en el Grafico 1, el número de ordenes de producción que han generado mermas mayores al 5.5% por mes del 2022.

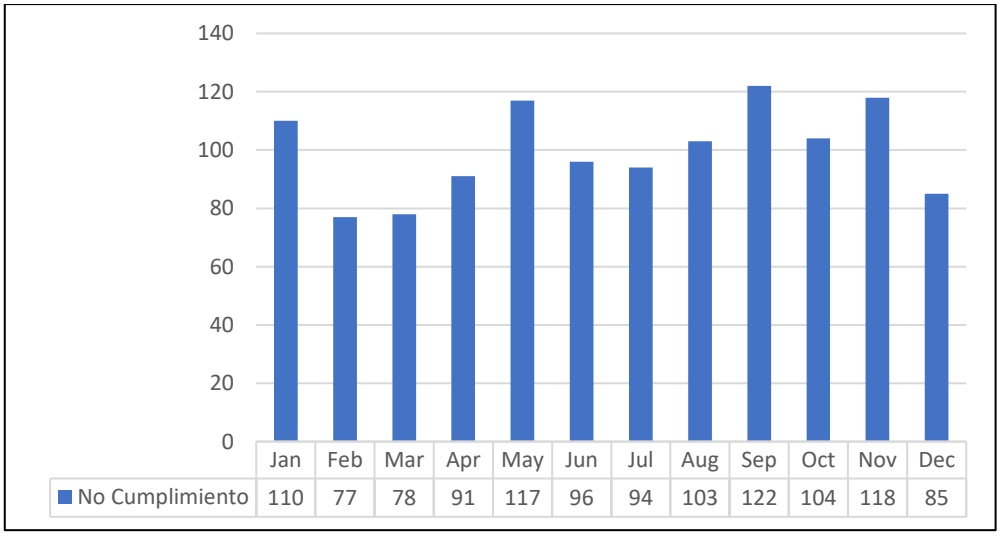


Gráfico 1 Mermas de Producción mayor al 5.5% permitido en el 2022

Esta variación supera el margen aceptable y afectar negativamente la rentabilidad y los costos de producción. Las causas de estas variaciones pueden ser diversas como la dirección inadecuada del supervisor de la maquinaria, los horarios de cortes poco eficientes, la configuración inapropiada de la maquinaria hasta el desempeño irregular de las máquinas utilizadas en el proceso de producción.

Este problema representa un desafío importante para las industrias de producción de papel higiénico en Ecuador, ya que la gestión eficiente de las mermas es esencial para garantizar la rentabilidad y la competitividad en el mercado. La falta de control y reducción de las mermas puede resultar en costos innecesarios, desperdicios de materiales y una disminución en la eficiencia operativa.

1.2. Justificación del problema

La implementación de técnicas de Machine Learning en este proyecto proporciona una solución efectiva y es fundamental para abordar el problema de gestión de mermas en la producción de papel higiénico en Ecuador debido a su capacidad para analizar grandes volúmenes de datos y descubrir patrones complejos. Mediante algoritmos avanzados, el Machine Learning puede identificar las causas subyacentes de las variaciones en los porcentajes de mermas, considerando múltiples variables interrelacionadas. Esto permite tomar decisiones más informadas, implementar medidas preventivas y optimizar los procesos de producción, lo que resulta en una reducción de costos y una mejora en la eficiencia operativa de las industrias de papel higiénico.

El Machine Learning permite analizar grandes cantidades de datos en tiempo real y descubrir patrones complejos que podrían pasar desapercibidos para los métodos tradicionales. Al entrenar algoritmos con datos históricos y en tiempo real, es posible identificar anomalías, establecer relaciones de causa y efecto, y predecir posibles problemas en la producción que influyan en el aumento de mermas en la producción.

Además, el Machine Learning permite el aprendizaje continuo y la adaptación a medida que se obtienen más datos, lo que mejora aún más la capacidad de detección y prevención de anomalías en la producción para optimizar las mermas. Al utilizar técnicas avanzadas como la detección de anomalías, las redes neuronales o el aprendizaje supervisado, se pueden desarrollar modelos precisos y confiables para la identificación y clasificación de anomalías en tiempo real.

1.3. Alcance

El proyecto se centrará en el análisis de datos históricos y en tiempo real de la producción de papel higiénico, con el objetivo de identificar anomalías y proyectar los residuos de mermas. Se utilizarán herramientas y técnicas de Machine Learning para analizar factores que influyen en la producción de merma como la maquinaria, supervisores, horarios de producción. El alcance se limita a la identificación de anomalías y la propuesta de soluciones basadas en Machine Learning.

1.4. Objetivos

1.4.1. Objetivo General

Realizar un análisis de los datos de las mermas de producción aplicando algoritmos de machine learning en Python para poder identificar anomalías y proyectar mermas que nos puedan ayudar a identificar opciones de mejora y tomar decisiones preventivas y estos resultados sean presentados en un dashboard.

1.4.2. Objetivos Específicos

- Realizar una depuración de los datos de Registros de Mermas en Excel para obtener una data limpia para el análisis.
- Realizar un análisis exploratorio de Registros de Mermas en RStudio para poder identificar relaciones entre variables de la data.
- Detectar anomalías en la producción de papel higiénico utilizando un modelo de machine learning que analizando datos secuenciales podrá encontrar patrones irregulares lo que ayudará a mitigar mermas.
- Proyectar las categorías de mermas en la producción de papel higiénico, utilizando un modelo de machine learning analizando patrones temporales y relaciones en los datos para mejorar el control y reducir las pérdidas durante el proceso de fabricación.
- Desarrollar un dashboard de control de mermas de producción utilizando la herramienta de Power BI para presentar los resultados de los análisis de los modelos de identificación de anomalías y proyección de categorías de mermas.

2. MARCO TEÓRICO

El marco teórico que se presenta a continuación permite tener un mejor entendimiento de los conceptos utilizados a lo largo de este proyecto. Por medio del presente marco teórico se definirán conceptos claves partiendo desde un enfoque global como general permitiéndonos conocer el auge de una de las profesiones que busca y promueve la importancia de manejar un control en las empresas y por ende en los residuos que los mismos generan, así como su correcta proyección en uno de los mercados de productos masivos con mayor participación en el mercado.

Donde posteriormente se mencionarán técnicas de predicción utilizando herramientas de Machine Learning de forma que se ahondará aspectos claves, sus modelos y su importancia en la actualidad de las empresas.

2.1. Marco Conceptual

El auge de la auditoria se remonta a principio del siglo XX, cuando las empresas comenzaron a aumentar de dimensiones debido a la incorporación de técnicas y tecnologías que permitían la fabricación en masa lo que trajo la necesidad de perfeccionar las técnicas contables. Es así como surgen las primeras asociaciones profesionales, como el instituto de Contadores públicos de Inglaterra y Gales que establecieron los primeros estándares para las prácticas contables de forma que se pueda garantizar la competencia y ética de los contadores. El concepto moderno de auditoría surge de los fracasos financieros y económicos de las sociedades nacidas en la ya mencionada revolución industrial. La poca seriedad y profesionalismo provocó la quiebra de un gran número de empresas situación que dio lugar a la imposición legal de la revisión de los estados financieros es así como se establece la misión del auditor por evaluar y verificar su exactitud. (Villardefrancos Álvarez, 2006)

2.1.1. Control Interno y el enfoque de la Auditoría Interna

El control interno es un proceso que se lleva a cabo en todos los niveles de una organización para asegurarse de que se cumplan adecuadamente los objetivos de la empresa. Este proceso es esencial para proteger los activos, verificar la precisión y veracidad de la información administrativa y financiera, así como para fomentar la eficiencia en las operaciones para lograr el cumplimiento de los objetivos y metas corporativas. En una empresa, la falta de estos controles puede resultar en pérdidas económicas, eficiencia y razonabilidad de la información contable ocasionando toma de decisiones incorrectas. (Tapia, 2017)

La auditoría interna surge de la misma necesidad de fortalecer las áreas de control interno de las organizaciones para reducir y evitar riesgos, proporcionar una evaluación completa y constructiva de las actividades y prácticas de una organización, identificar áreas de mejora, sugerir soluciones y verificar que las acciones correctivas se implementan efectivamente. Los auditores internos desempeñan un papel clave en la supervisión y evaluación independiente de las actividades de una organización ayudando a garantizar el cumplimiento de los objetivos estratégicos y su eficacia. (Tapia, 2017)

2.1.2. Los Costos y su importancia

La comprensión de los costos es fundamental para la gestión eficiente y rentable de cualquier negocio. Los costos en las empresas son los gastos que incurren para producir bienes o brindar servicios y estos pueden ser de dos tipos:

- Costos directamente relacionados con la producción.
- Costos indirectos asociados con la administración.

Los costos tienen dos categorías principales: costos fijos y costos variables. Aquellos costos fijos son los que no varían en función al nivel de producción como el alquiler del local o los salarios del personal administrativo. En cambio, los costos variables están relacionados con el nivel de producción como la materia prima.

La importancia de entender los costos radica en las oportunidades al momento de tomar decisiones por la gerencia para la fijación de precios, el determinar los volúmenes producidos y la identificación de áreas de eficiencias y ahorros. Una gestión eficiente de los costos puede proporcionar una ventaja competitiva significativa para alcanzar el éxito a mediano y largo plazo. (Mera, 2019)

2.1.3. Mermas

En toda actividad económica en donde se manejan significativos volúmenes de bienes resulta prácticamente inevitable el tener que afrontar situaciones de mermas cuando de activos de consumos se refiere, los cuales tienden a ser consecuencia de procesos de comercialización o del proceso de producción. (Ferrer, 2017)

En cuanto a la normativa tributaria, define a la merma como la pérdida física en el volumen, peso o cantidad de las existencias, ocasionada por causas inherentes a su naturaleza o al proceso productivo.

2.1.3.1. Merma normal y merma anormal

En los procesos productivos las empresas fijan porcentajes de pérdidas por mermas normales o inherentes del proceso de las materias primas o suministros que son consumidos en el momento de producir bienes. Estas mermas se suman al costo de los productos elaborados en proporción al volumen de producción.

En cuanto a las mermas anormales son aquellas que sobrepasan al porcentaje preestablecido en los diversos procesos de producción. Estas pérdidas no son consideradas en el costo del producto sino deberán ser cargadas como parte de los gastos del periodo.

2.1.4. Anomalías

Las anomalías o datos atípicos hacen referencia a datos que difieren significativamente del patrón general de un conjunto de datos. Estas observaciones pueden ser el resultado de errores al ingresar los datos, eventos inusuales que ocasionaron un aumento o disminución en valores o simplemente pueden ser motivo de una situación específica que no siguen el comportamiento de los datos.

Las anomalías pueden ser presentadas en diversas formas ya sea como fue mencionado por un aumento o diminución en valores que difieren significativamente del patrón que han llevado los datos en las series de tiempo.

2.1.4.1. Detección de anomalías en grandes volúmenes de datos.

Actualmente las empresas registran toda su información y transacciones en sus sistemas informáticos mismas que generan una gran cantidad de datos también conocida con el termino de Big Data. Esta basta cantidad de datos sobrepasan las capacidades humanas para su procesamiento y análisis manual. Siendo capacidades que limitan a la persona a identificar posibles datos desviados del flujo normal que lleven a problemas en la gestión o procesamiento de estos. (Domínguez, 2018)

Una de las soluciones planteadas en la identificación de estos valores que se desvían del flujo o tendencia de los valores son las ya mencionadas en el presente trabajo como anomalías o datos atípicos. Para alcanzar con este objetivo es necesario el usar herramientas informáticas que permitan identificar patrones y relaciones de datos que conlleven a identificar comportamientos inusuales como el que se presenta en el siguiente grafico en donde se puede apreciar que un valor en el índice 44 se aleja significativamente de la tendencia o patrón que sigue la serie que para efectos de este ejemplo representa el peso de cierto objeto.

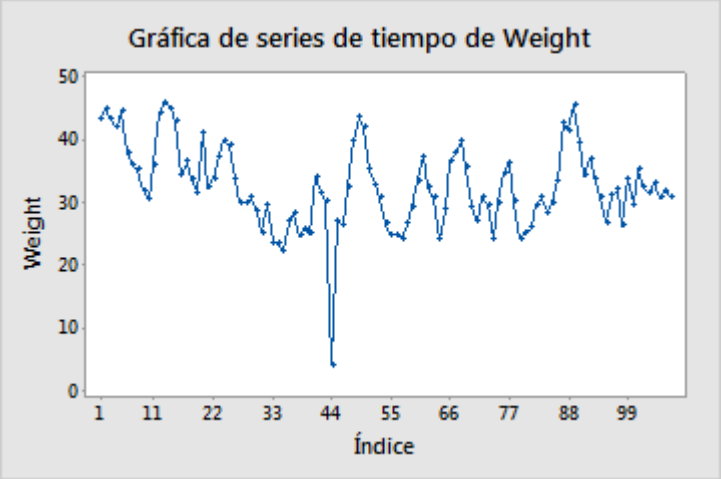


Gráfico 2 Series de tiempo de Weight *Fuente: Minitab Support*

La detección de estos comportamientos inusuales es importante en campos como la minería de datos, el aprendizaje automático, las industrias financieras, la detección de fraudes, así como también para manejar la calidad y control de sistemas. Entre los métodos más comunes para la identificación de las anomalías se encuentran los siguientes;

Métodos estadísticos	Enfoque basado en distancia	Métodos basados en series temporales	Modelos de aprendizaje autónomos
Este tipo de métodos utilizan medidas como la desviación estándar, los valores Z o incluso los percentiles para identificar puntos atípicos que comparen variables con los demás datos.	En este método se calcula la distancia entre los valores o puntos de datos y sus vecinos cercanos para determinar aquellos que tengan una distancia considerable y puedan ser catalogados como datos atípicos.	Se centra en detectar patrones inusuales en datos que siguen una secuencia en relación con el tiempo.	Estos algoritmos construyen modelos que simulen el comportamiento normal de los datos. Esto puede ayudar al momento en que el sistema identifique valores anormales alejado al comportamiento o tendencia establecido en el modelo detectando y enviando alertas informando la presencia de valores atípicos.

Tabla 1 Métodos comunes para identificar anomalías *Fuente: Elaboración propia*

2.1.5. Mermas y Anomalías

Una vez establecido lo que representan ambos términos se puede establecer la relación entre merma y los datos atípicos como la influencia en la magnitud de estos desperdicios en un sistema o proceso productivo causando una serie de problemas y causas raíz que requieren el accionar respectivo por parte de los miembros de una organización.

La relación entre ambos conceptos se puede entender de mejor manera de las siguientes formas:

- Los datos atípicos pueden llegar a distorsionar la medición de mermas al ser valores extremos que no representan el comportamiento típico del proceso lo que puede generar una sobreestimación o subestimación de las mermas lo que consecuentemente dificultaría el encontrar causas raíz en estas pérdidas.
- Los datos atípicos pueden ocultar o exagerar factores influyentes en las mermas de producción lo que puede generar una percepción errónea de aquellos factores influyentes en la producción de desperdicio.
- Los datos atípicos también pueden apreciarse como indicadores de problemas en los procesos que generaron estos desperdicios.

Cabe recalcar que la presencia de datos atípicos en las mermas afecta la toma de decisiones en relación con su reducción puesto que, si no son identificadas correctamente y se comprenden los valores atípicos, será poco probable que las acciones correctivas para mitigar los desperdicios tengan el alcance deseado.

2.2. Marco Metodológico

2.2.1. CRISP-DM

CRISP-DM conocido por sus siglas en inglés (Cross Industry Standard Process for Data Mining) proporciona el orden y la duración de un proyecto de data science o de análisis de datos. Este tipo de modelo cubre las fases del proyecto, las tareas que se llevarán a cabo y las relaciones que existes entre tareas.

Esta metodología mira al proceso de análisis como un proyecto profesional estableciendo un contexto superior influyendo en la elaboración de los estándares. Este modelo toma en cuenta la existencia de un usuario o cliente ajeno al equipo de desarrollo, así como el hecho de que el proyecto no solo abarca el lograr un modelo idóneo que se apegue a la estructura de los datos, sino que está relacionado con otros proyectos y es indispensable manejar una documentación exhaustica para que el equipo de desarrollo utilice el conocimiento adquirido. (Wirth, 2019)

El ciclo de vida del proyecto de análisis de datos consiste en seis fases mostradas a continuación;

2.2.1.1. Fase I. Definición de las necesidades del cliente.

Durante esta fase inicial se deberá enfocar en la comprensión de los objetivos del proyecto en base a necesidades identificadas con el usuario o cliente. Después el conocimiento es convertido en la definición de un problema de análisis de datos y en el plan preliminar diseñado para alcanzar los objetivos planteados.

2.2.1.2. Fase II. Estudio y comprensión de los datos.

La fase de entendimiento de datos comienza con la recopilación de datos y continúa con actividades que permiten lograr un mejor entendimiento de los datos, identificando problemas de calidad para obtener conocimiento preliminar sobre los datos y/o descubrir relaciones interesantes para formar hipótesis sobre la información oculta.

2.2.1.3. Fase III. Análisis de datos y selección de características.

Todas las actividades necesarias para crear el conjunto final de datos, es decir, aquella información que se utilizarán en las herramientas de modelado inicia con datos en bruto que formará parte de la fase de preparación. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de la información para las herramientas que modelan.

2.2.1.4. Fase IV. Modelado

En esta etapa, se seleccionan y aplican las técnicas de modelado más relevantes para el problema (las más efectivas) y se calibran los parámetros para que sean ideales. La minería de datos típicamente presenta una variedad de enfoques. Algunas técnicas requieren una forma de datos específica. Por lo tanto, prácticamente todos los proyectos terminan en la fase de preparación de datos.

2.2.1.5. Fase V. Evaluación

Para esta fase se han construido uno o varios modelos que parecen tener calidad suficiente desde la perspectiva del análisis de datos en esta etapa del proyecto. Antes de proceder al despliegue final del modelo, es crucial evaluarlo a fondo y revisar los pasos que se tomaron para crearlo, así como comparar el modelo que se obtuvo con los objetivos de la empresa. Un objetivo clave es determinar si hay alguna cuestión comercial importante que no haya sido considerada adecuadamente. Al final de esta etapa, se debe tomar una decisión sobre qué hacer con los resultados del proceso de análisis de datos.

2.2.1.6. Fase VI. Despliegue o puesta en producción

La creación del modelo no es el último paso en un proyecto. Si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento debe organizarse y presentarse para que el cliente pueda usarlo. La etapa de desarrollo puede ser tan sencilla como la creación de un informe o tan compleja como la realización regular y posiblemente automatizada de un proceso de análisis de datos en la organización, dependiendo de los requisitos.

En el presente proyecto se abordarán para las primeras fases del es decir para el estudio y la comprensión de los datos, así como el análisis de los mismos sistemas ya existentes que permiten alcanzar insights valiosos mediante técnicas estadísticas y por su versatilidad de funcionamientos como lo es RStudio y Python.

2.2.2. RStudio

RStudio es un Entorno de Desarrollo Integrado, (IDE por sus siglas en ingles Integrated Development Environment), utilizado en el lenguaje de programación R (Allaire, 2011). Este programa sirve para tratar datos, importarlos, transformarlos, modelarlos. R es uno de los principales lenguajes para la ciencia de datos. Proporciona herramientas, como funciones de visualización, para explorar y analizar los datos antes de ser utilizados en cualquier aprendizaje automático machine Learning. También es útil para evaluar los resultados del algoritmo de predicción y evaluar por medio de técnicas estadísticas el comportamiento que rigen a los datos.

2.2.3. Machine Learning

El machine Learning es un área de la ingeniería artificial que engloba un conjunto de técnicas para logar un aprendizaje automático a través del entrenamiento de grandes cantidades de datos. En la actualidad existen diferentes modelos que usan esta técnica para lograr una precisión y confianza superior. La construcción de estos modelos requiere adaptaciones propias debido a la naturaleza de los datos esto implica que debería ser capaz de modificar su comportamiento en base a los objetivos deseados.

Es por ello por lo que actualmente existen tres grandes grupos de algoritmos de machine Learning los cuales son:

Algoritmos supervisados	Algoritmos semi - supervisados	Algoritmos no supervisados
Estos algoritmos usan un conjunto de datos de entrenamiento preclasificados, los cuales son procesados para realizar predicciones sobre los mismos, haciendo correcciones cuando estas son equivocadas hasta que el modelo haya alcanzado nivel deseado de precisión.	Este tipo de algoritmos combinan un conjunto de datos tanto preclasificados como no clasificados para generar una función deseada. En este tipo de modelos deben aprender las estructuras para organizar los datos, así como para realizar predicciones.	El conjunto de datos no se encuentra clasificado y no tienen un resultado conocido. Es por ello por lo que se debe deducir la estructura de los datos de entrada lo que se puede conseguir mediante un proceso matemático para reducir redundancias sistemáticas.

Tabla 2 Grupo de algoritmos de Machine Learning *Fuente: Elaboración propia*

Y Entre cada grupo de algoritmo existen elementos principales del aprendizaje autónomo se pueden resaltar los siguientes:

2.2.3.1. Datos

Todos los modelos de aprendizaje autónomo requieren de una base de datos o también conocidos como conjunto de entrenamiento mismos que pueden ser manipulados y recolectados por el ser humano o ser almacenados de forma automática por sistemas informáticos o softwares. (Michael Paluszek, 2019)

2.2.3.2. Modelos

Estos modelos son esenciales porque capturan las tendencias y el comportamiento de los datos, mismos que pueden ser de origen humano basándose en la experiencia y de las mismas observaciones, sin embargo, existen algunas técnicas del aprendizaje autónomo que crean o desarrollan sus propios modelos sin una estructura de origen humana. (Michael Paluszek, 2019)

2.2.3.3. Entrenamiento

De forma similar que el ser humano que por medio del entrenamiento logra aprender nuevas tareas, los modelos de aprendizaje autónomo requieren el pasar por una etapa de entrenamiento para ajustar el modelo a las tendencias que sigan los datos. Esta etapa consiste en ingresar al sistema un conjunto de datos de entrada para que la máquina pueda hacer predicciones. (Michael Paluszek, 2019)

2.2.4. Redes Neuronales

Una red Neuronal (RNA) es un modelo de nodos interconectados que simulan de cierta forma el funcionamiento del cerebro humano. Este modelo computacional de interconexiones tiene una capa de entrada, una de salida y al menos una capa oculta las cuales tienen interconexiones entre ellas como se puede observar en la Ilustración 1.

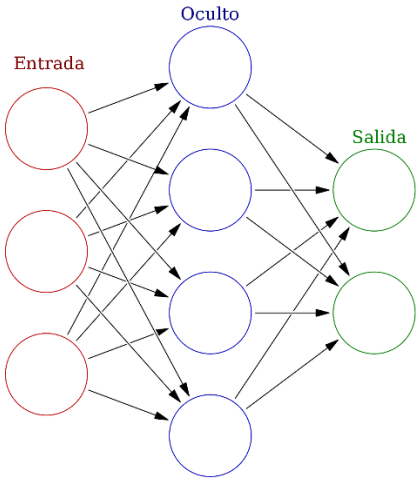


Ilustración 1 Modelo red neuronal artificial *Fuente: IBM Documentación*

Resulta indispensable para entender el funcionamiento de un RNA el introducir la definición de neurona.

2.2.5. Neurona

Las neuronas artificiales son la base del funcionamiento de las redes neuronales en el machine learning al imitar el funcionamiento de una neurona biológica. En donde recibe señales de entrada que se suman y si el resultado supera un umbral determinado por la función de activación, esta neurona al igual que las biológicas dispara una señal de salida.

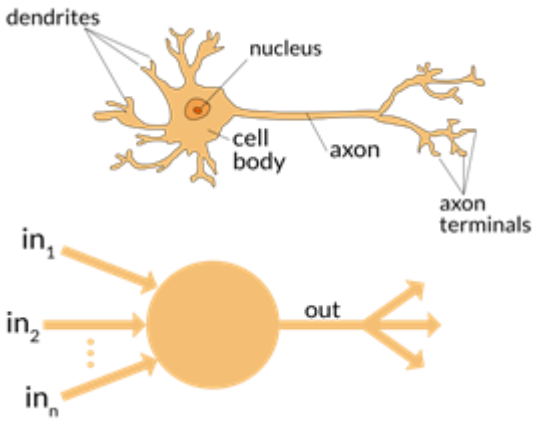


Ilustración 2 Funcionamiento de las neuronas *Fuente: Magiquo Software*

Una neurona artificial consta de tres componentes claves:

Entradas ponderadas	Sumados	Función de activación
Cada neurona recibe múltiples entradas las cuales están ponderadas por un peso específico que influyen en que tan relevante es cierta información para ser procesada midiendo su impacto en la variable de salida. Durante el proceso de entrenamiento se puede ajustar estos pesos para mejores rendimientos del modelo.	Después de aplicar los pesos a las entradas, la neurona realiza una suma ponderada de todas las entradas de forma que pueda calcular la "activación" que representa la fuera de la señal recibida.	La activación representa la capacidad de aprendizaje no lineal del modelo que permite capturar relaciones complejas de datos y no limitar a soluciones lineales.

Tabla 3 Componentes claves de una neurona artificial *Fuente: Elaboración propia*

2.2.6. Modelos de Regresión y su valor de predicción

Los modelos de regresión son utilizados para estimar o predecir un valor numérico a partir de un conjunto de datos de entrada funcionando como el caso de la neurona explicada anteriormente para establecer una relación entre variables y poder de cierta forma predecir un output o una variable de respuesta.

El objetivo principal de un modelo de regresión es encontrar la mejor relación entre las variables independientes y las dependientes para hacer relaciones precisas sobre nuevos datos. Estos tipos de modelos buscan identificar tendencias y patrones

de los datos de entrenamiento para generar y hacer predicciones.

El presente proyecto se centrará en el modelo de regresión LSTM que es un tipo de red neuronal diseñado para problemas de regresión:

2.2.6.1. Modelo de LSTM

El tipo de red neuronal LSTM fue desarrollado a partir de redes neuronales recurrentes (RNN) en donde entre sus principales características evita la dependencia a largo plazo debido a su unidad de almacenamiento única y a que ayuda a predecir series de tiempo financieras. Esta red neuronal se ha convertido en un aprendizaje de regresión famoso debido a su gran capacidad de aproximación no lineal y a su aprendizaje autónomo adaptativo. (Grégoire Montavon, 2012)

Sepp Hochreiter y Jurgen Schmidhuber crearon este algoritmo con el objetivo de solucionar el problema del desvanecimiento del gradiente de las RNN mediante la inclusión de una celda de memoria, lo que permitiría el traspaso de información desde el pasado hasta el momento t. En su trabajo (Hochreiter, 1997)

Debido a que el LSTM es una modificación de una RNN, primero debe definirse como una RNN:

-Una RNN es una red neuronal diseñado para procesar información ordenadamente. Se pueden apreciar tres componentes, predictores xt, capa interna A donde se ponen algunos pesos y luego una emisión ht. Si se explora más en detalle, se puede observar cómo se dividen estas partes en un proceso secuencial en el que los datos fluyen de la primera secuencia a la siguiente. En resumen, La salida del momento t depende fundamentalmente de la salida en t-1 y del predictor del momento t. (Grégoire Montavon, 2012)

Como se mencionó anteriormente, el desvanecimiento del gradiente y la complejidad de entrenamiento de las RNN son las razones por las que el LSTM surge.

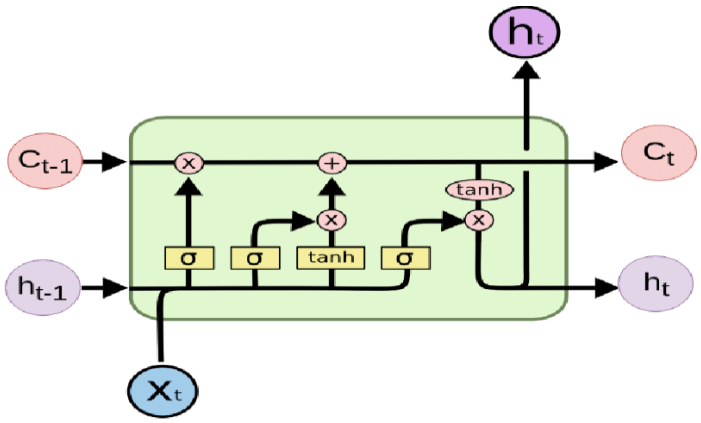


Ilustración 3 Estructura de una Red LSTM *Fuente: ResearchGate, Alejandro Casallas*

Una capa LSTM consta de tres puertas y el estado de la celda, como se muestra en la ilustración 3. La información fluye por el estado de la celda; Ct. La puerta de olvido es la primera posible modificación del estado de la celda. La puerta de olvido indica que la información del pasado debe olvidarse o recordarse. En donde la puerta de salida finalmente decide cómo queda la información final.

El flujo de información se explica ahora paso a paso. La información sobre el estado de la celda se transmite de Ct-1 a Ct. La Ecuación de olvido ocurre en las puertas donde σ (·) es una función sigmoidea. Si el valor de la función σ (·) es igual a 1, se mantiene toda la información del pasado y el estado de la celda no se altera.

Como consecuencia de lo mencionado, dos operaciones matemáticas más se realizan en la puerta de salida. En primer lugar, se aplica una función tangente hiperbólica a los valores del estado de la celda. Esta sección es donde el algoritmo determina si incluye o no el pasado en la salida y por otro lado queda saber qué información de los predictores y de la salida de t-1 procesar en donde por medio

de la función ht El algoritmo puede expulsar la salida del momento t. y a su vez servirá como input para t+1. (Grégoire Montavon, 2012)

Para un mejor entendimiento, la celda LSTM cuenta de forma resumida con la siguiente arquitectura en donde se puede resaltar cuatro componentes importantes:

Puerta de entrada	**Puerta del Olvido**	**Puerta de salida**	**Célula de memoria**
Permite el control de las entradas en la célula de memoria. Utiliza una función de activación sigmoidea para identificar qué información considerar y cual debería olvidar.	Controla el flujo de información de la puerta de entrada que debería olvidar. Y al igual que la puerta anterior usa una activación sigmoidea para decidir qué información debe olvidar.	En esta puerta se controla la salida de la cédula de memoria. Utiliza una función de sigmoidea y una función tangente para decidir qué información debe presentar.	Este es el componente principal del modelo en donde su estructura permite almacenar información a largo plazo y puede olvidar datos de forma selectiva o añadir nueva información.

Tabla 4 Componentes del modelo LSTM *Fuente: Elaboración propia*

2.2.7. Procesamiento

Los siguientes pasos se incluyen en el preprocesamiento utilizado para el LSTM. Primero, se divide los conjuntos de datos en dos particiones. Esto significa que el primer setenta por ciento de las observaciones permanecerán en el conjunto de entrenamiento y el treinta por ciento de las observaciones posteriores formarán parte del conjunto de validación. Esta partición nueva es

necesaria para que el LSTM pueda ajustar los pesos que asigna durante su proceso de aprendizaje sin experimentar sobreajuste. Para lograrlo, utilizará el conjunto de entrenamiento para asignar pesos y el conjunto de validación para evaluar y corregir los pesos.

Además, es importante considerar que en el procesamiento los tres parámetros que tienen el mayor impacto o peso en el ajuste del LSTM son el tipo de optimizador, el número de neuronas y la tasa de aprendizaje;

Parámetros	**Detalle**
Optimizados	Un optimizador es un algoritmo que ajusta automáticamente los parámetros de un modelo para que las predicciones se acerquen más a los valores reales, reduciendo así el error durante el entrenamiento.
Neurona	la cantidad de neuronas indica la cantidad de unidades de procesamiento en una capa. Esto tiene un impacto en la capacidad del modelo para comprender patrones en los datos, ya que pocas limitan el aprendizaje y muchas pueden provocar sobreajuste.
Tasa de aprendizaje	Se refiere a la cantidad de parámetros que se ajustan en cada paso de aprendizaje. Equilibra la rapidez y la precisión del ajuste del modelo.

Tabla 5 Parámetros con mayor impacto LSTM *Fuente: Elaboración propia*

Tener en cuenta que existen más parámetros que se pueden añadir con la finalidad de mejorar el modelo LSTM a la tendencia de los datos.

2.2.7.1. Optimizador Adam

En relación con lo mencionado. El algoritmo de optimización Adams será usado durante este proyecto. Por sus siglas en ingles “Adaptive Moment Estimation” es un algoritmo conocido en el entrenamiento de redes neuronales y en el aprendizaje automático en general. La ventaja de este optimizador radica en su eficiencia y adaptabilidad, debido a que ajusta automáticamente la tasa de aprendizaje para cada parámetro, lo que permite trabajar con problemas de gradientes dispersos o ruidosos. Este método es un optimizador adaptativo que combina conceptos de otros optimizadores muy usados como: RMS Root Mean Square y el descenso de gradiente estocástico. El objetivo de Adam es mejorar la velocidad de convergencia y la eficiencia en la optimización en parámetros de machine learning.

2.2.8. Comparación de LSTM con otras técnicas

Estas redes LSTM tiene una amplia variedad de aplicaciones en IA, entre las que permite el reconocimiento y detección de anomalías. Su capacidad para seleccionar y olvidar información es útil para el resultado actual basado en datos de entrenamiento. Obteniendo buenas estimaciones y decisiones basadas en datos históricos.

2.2.9. Python

Python es uno de los mejores lenguajes de programación con mayor popularidad en el mundo de la ciencia de datos, mantiene una sintaxis simple, clara y sencilla, así como también maneja un gran número de librerías complementarias desarrolladas para una amplia gama de tareas. Es importante resaltar que este lenguaje a pesar de que para tareas de cálculo intensivos puede mostrar un menor rendimiento a otro tipo de lenguajes con C+ se han desarrollado librerías como NumpPy y SciPy para la implementación de capas inferiores para operaciones rápidas y vectoriales en matrices del tipo multidimensionales. (Sebastian Raschka, 2019)

2.2.9.1. Paquetes para cálculos científicos, ciencia de datos y aprendizaje autónomo

A lo largo de este proyecto se ha usado librerías como NumPy para almacenar y manipular datos. Así como también se ha usado la librería Pandas, librería creada sobre NumPy para proporcionar herramientas de alto nivel en la manipulación de datos permitiendo el trabajar con valores tabulados de formas más sencillas y es útil para la limpieza y preparación de los datos antes de su entrenamiento.

2.2.9.2. Python y su aplicación con el machine learning

Python se ha convertido en uno de los lenguajes más utilizados para ser conectado con el machine learning debido a su rica colección de bibliotecas y herramientas especiales que facilita el desarrollo de modelos y algoritmos de machine learning.

Así como también gracias a la comunidad de desarrollo activa que ha ayudado a crear tutoriales, cursos y recursos educativos que hacen que el aprendizaje automático sea accesible para todos.

Sin duda, la versatilidad de Python hace que los científicos de datos y los ingenieros de machine learning pueden experimentar con algoritmos y soluciones, creando prototipos rápidamente y luego implementar de manera eficiente modelos sofisticados en producción.

2.2.10. Google Colaboratory

Google Colaboratory, también conocido como Google Colab, es una herramienta de desarrollo y análisis en la nube que permite programar en Python con muchas ventajas. En primer lugar, tiene accesos a recursos en la nube sin necesidad de configurar entornos locales. Esto nos ayuda a aprovechar de manera sencilla y eficiente la potencia del computador y almacenamiento de Google para proyectos.

Además, permite el acceso a unidades de procesamiento gráfico (GPU) o unidades de procesamiento tensorial (TPU), lo que acelera significativamente el entrenamiento de modelos de aprendizaje automático. Otra ventaja es que la plataforma tiene preinstalada una gran grama de librerías populares lo que simplifica la configuración y el uso de herramientas esenciales para el aprendizaje automático.

2.2.11. Power BI

Power BI es un conjunto de herramientas de análisis empresarial que proporcionan información a toda la organización. Permite el enlace de cientos de diferentes fuentes de datos, lo que permite una preparación sencilla de las ya mencionadas bases. Hace que todos los datos e información importantes para el control y la dirección sean fáciles de ver a todo nivel organización. (Microsoft, 2020)

Power BI, la nueva herramienta de Business Inteligence (BI) que se integra con la suite de productividad Office 365, permite el análisis y la interacción con una gran cantidad de datos que se encuentran en Excel por medio de paneles e informes dinámicos gracias a la amplia recopilación de información y datos. (Softeng, 2015)

Ventaja	Detalle
Flexibilidad	Permite obtener datos cruciales para una amplia gama de situaciones.
Optimización, limpieza, transformación y combinación	Estos datos de múltiples fuentes permiten un análisis detallado y la detección de patrones.
Innovación	Con visualizaciones de datos interactivas, podrás crear informes sorprendentes.
Personalización	Utiliza herramientas de creación de temas, formato y diseño para crear informes variados.
Multiuso	Crear informes optimizados para teléfonos inteligentes

Tabla 6 Ventajas de Power BI Fuente: Elaboración propia

2.2.12. Mejora del modelo

Resulta indispensable para mejorar el modelo de machine learning medir la calidad del conjunto de muestras es decir de las pruebas. Es importante recordar que los proyectos profesionales usan conjuntos de validación.

Primero se usa el conjunto de datos de entrenamiento para crear el modelo y obtener sus medidas de calidad en relación con las muestras de entrenamiento. En este sentido si el modelo cuenta con muchos errores, las medidas de calidad devolverán valores muy bajos. Lo que puede representar que hace falta mayor cantidad de datos para que el modelo se ajuste, las características del modelo no resultan relevantes o que el algoritmo de aprendizaje no es el indicado para el tipo de datos. (Robadilla, 2020)

Una de las causas por las que el modelo tienda a dar variables de respuesta poco acertadas es debido a que el modelo sufra de un sub-ajuste (underfitting). Para ilustrar el caso, se presenta la siguiente figura en donde se puede apreciar un modelo lineal muy básico que no se ajusta a la distribución de los datos.

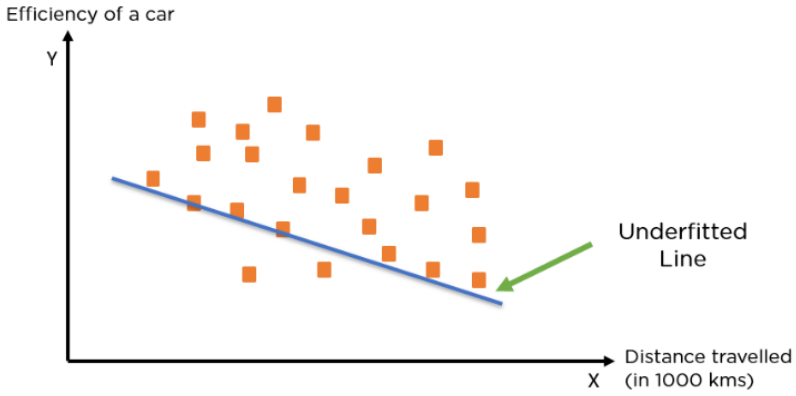


Gráfico 3 Underfitting line *Fuente: Medium, Anup Bhande*

En cambio, también pudiese ocurrir la situación inversa en donde el modelo se encuentre sobre ajustado (overfitting) es decir cuando puede predecir con demasiada exactitud el modelo de entrenamiento, pero no las de prueba. Por medio de la siguiente visualización se puede ejemplificar el tipo de sub-ajuste en el gráfico de la izquierda como se explicó anteriormente, en el centro un gráfico con un modelo robusto que se alinea a la tendencia de los datos y en la derecha un modelo sobre ajustado en donde se memorizó los valores de entrenamiento y no es capaz de generalizar con otro tipo de datos.

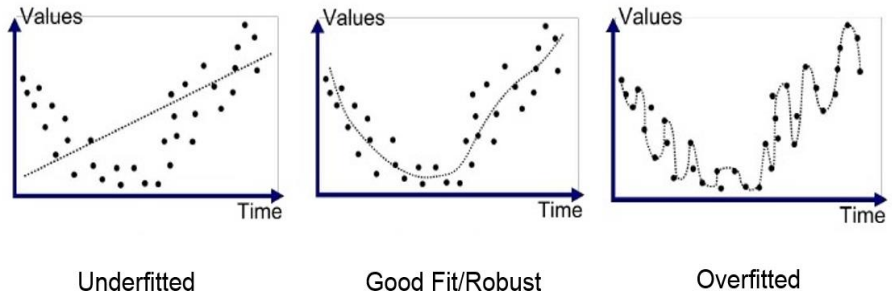


Gráfico 4 Procesamiento del modelo *Fuente: Medium, Anup Bhande*

Para evitar este sobreajuste se puede entrenar el modelo usando más datos para dificultar la memorización de los datos o en caso de no disponer con un mayor base de datos se puede optar por usar un modelo más simple a este método se lo conocer como la regularización de los datos mismo que a pesar de ser la solución más simple tiende a ser la más eficiente. (Robadilla, 2020)

2.2.13. Calidad y Rendimiento

Se utilizan una variedad de métodos y métricas para evaluar la calidad y el rendimiento de los modelos en estadística y aprendizaje automático los que para objeto de este proyecto se analizarán 3: (Juan R. Camarillo-Peñaranda, 2013)

2.2.13.1. R cuadrado (R2)

Es una medida que muestra cómo las predicciones de un modelo se ajustan a los valores de los datos reales. R2 ofrece una puntuación de 0 a 1 indicando que el modelo no explica la variabilidad de los datos y que se ajusta perfectamente a los datos. En pocas palabras, R2 calcula la proporción de la variabilidad en los valores de respuesta que el modelo explica. (Juan R. Camarillo-Peñaranda, 2013)

2.2.13.2. El error absoluto medio (MAE)

Se define como la magnitud promedio de los errores entre los valores reales y las predicciones del modelo. Calcule la diferencia absoluta promedio entre los valores reales y las predicciones. Dado que no considera el cuadrado de errores, MAE es más resistente a valores atípicos. (Juan R. Camarillo-Peñaranda, 2013)

2.2.13.3. El error cuadrático medio (MSE)

Es una medida adicional que mide la precisión de las predicciones del modelo. Calcula la media de los cuadrados de las diferencias entre los valores reales y las predicciones. MSE es más sensible a valores imprevistos porque penaliza más los errores grandes. (Juan R. Camarillo-Peñaranda, 2013)

2.3. Marco Referencial

En una sociedad globalizada en donde las revoluciones tecnológicas han facilitado significativamente la forma con la que las personas hacen negocios en cualquier parte del mundo sin importar la hora ni el lugar de la transacción; así mismo, al existir una mayor cantidad de transacciones, métodos y sistemas para llevar a cabo una actividad, estas nuevas herramientas han provocado que los métodos de control evolucionen y se adapten a esta nueva era digital.

Los nuevos avances en el control interno y gestión de riesgos involucran directamente al rol del auditor por su capacidad para reforzar la transparencia de los procesos y la correcta evaluación del control interno mientras gestiona el riesgo; sin duda la evolución de la tecnología será el factor determinante en el cambio del rol del auditor del futuro. (BDO, 2020).

Los conceptos de auditoría tradicional cambiaran significativamente debido a la cantidad de datos que mantienen las empresas; Big Data es un término usado con frecuencia desde los años 90 y hace referencia al análisis de grandes cantidades de datos basado en la programación de fórmula de procesamiento que ya existen en la actualidad. Se estima que el tiempo dedicado a los nuevos procesos de auditoría será reducido en comparación a los procesos tradicionales. (Acalá, 2020)

En el contexto de la auditoría el Big Data implica tener acceso a una amplia gama de datos financieros y operativos provenientes de sistemas internos de las empresas o fuentes externas. Al tener grandes conjuntos de datos es indispensable el buen procesamiento de estos, una de las mejores herramientas para este trabajo es el Machine Learning proveniente de la inteligencia artificial en donde permite aprender y mejorar automáticamente a partir de datos sin la necesidad de ser programadores explícitamente.

En el área de auditoría el Machine Learning es una herramienta que puede aplicarse para analizar, extraer información relevante de grandes conjuntos de datos de forma que se pueda identificar tendencias, patrones y anomalías para generar insights valiosos para los auditores.

CAPÍTULO II

3. METODOLOGÍA

3.1. Definiciones, Siglas y Abreviaciones.

Definiciones generales

PH: Papel higiénico

DH: Papel higiénico doble hoja

TH: Papel higiénico triple hoja

Btos: Bultos

ML: Machine Learning

R: lenguaje de programación enfocado en el análisis estadístico.

Python: lenguaje de programación

MSE: Mean Squared Error

MAE: Mean Absolute Error

Modelo IA: Modelo de Identificación de Anomalías

Modelo PCM: Modelo de Proyección de Categorías de Mermas

3.2. Procedimiento

Para el desarrollo de este proyecto se utilizó la metodología de CRIP-DM, para resolver el problema antes descrito. Siguiendo una serie de pasos, como se puede ver en la ilustración.

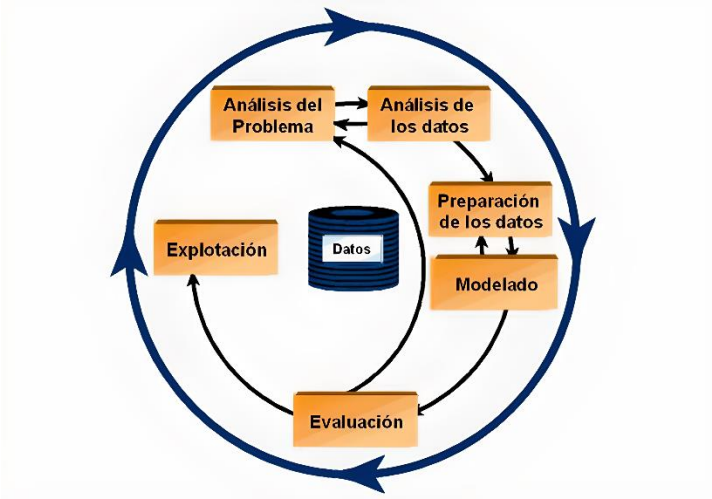


Ilustración 4 Metodología CRIP-DM *Fuente: HERMINWEB, Yoanni Ordoñez*

Este proyecto es ambicioso y busca obtener como resultados no solo una proyección de mermas futuras si no también la relación de características que influyen en la producción de merma para identificar anomalías por lo cual se evaluará la implementación de un modelo de machine learning para el análisis de este objetivo.

3.3. Herramientas para el desarrollo

Las siguientes son algunas de las herramientas que se aportaron durante el proyecto para desarrollar e implementar los modelos de Machine Learning:

3.3.1. Servidor de Prueba

Google Colaboratory

- RAM del Sistema: 12.7 GB
- Disco: 107.7GB

3.3.2. Software

Los análisis estadísticos se realizaron en RStudio

- Software RStudio versión 2023.06.0 Build 421
- Lenguaje de Programación R versión 4.3.1

Librerías de RStudio usadas.

- Ggplot2
- Corrplot

3.4. Proceso

Como se mencionó anteriormente, para el desarrollo del proyecto se siguió la metodología de CRIP-DM y paso a paso, cabe aclarar que, por forma, las evaluaciones de los modelos se encontrarán en el siguiente capitulo.

3.4.1. Análisis del Problema

Como se planteó en el Capítulo 1, el problema se encuentra en la falta de control de la producción de mermas en la fabricación de papel higiénico. Partiendo de ello, se comenzó a dar el seguimiento de la producción de papel higiénico dentro de la industria para poder evidenciar como era el proceso de la trata de la merma y su registro en el sistema.

Se ha realizado entrevistas al asistente de costos para obtener la información requerida para profundizar sobre el problema y se ha determinado los siguientes hallazgos.

3.4.1.1. Hallazgos

- Dentro del proceso de la producción el supervisor a cargo de la orden de fabricación debe llevar un registro sobre lo que se produce y los desperdicios consecuentes a ello. Ahora bien, este registro se realiza mediante un Excel el cual tienen acceso todos los supervisores a crear, modificar y eliminar registros de OF, esta base de datos no es revisada en la veracidad de los datos reportados ni contiene evidencia que lo que reportan es la realidad de producción.
- No todos los registros de producción tienen una orden de fabricación, ni todos los registros para realizar pruebas tienen actas que validen o confirmen la orden.
- Se evidencio que hay una diferencia entre el registro de planta y lo registrado en la plataforma del SAP acerca de las mermas de producción del mes de enero a junio del 2023. Como se puede observar en los gráficos de la línea de producción 1, mostrando los porcentajes promedio por mes entre lo registrado en planta (RP) y lo registrado en SAP (RS).

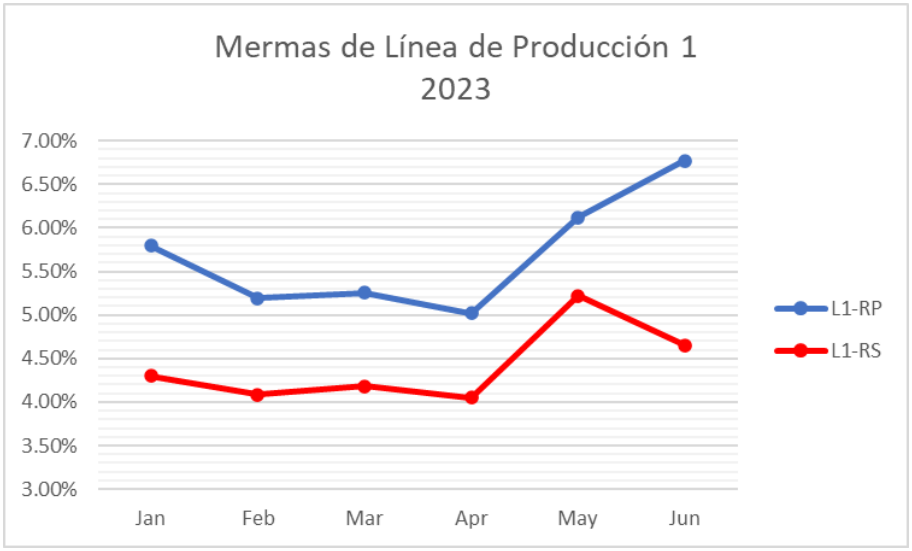


Gráfico 5 Mermas de Línea de Producción 1 2023 Fuente: Propia

- El control que llevan los supervisores, no se lleva un registro el peso del rollo antes de producción lo cual no permite hacer un match entre el Peso Producción y el Peso Desperdicio.
- Se reconoce que una maquinaria genera más desperdicios que otra, esto en relación con la vida útil de cada una, se evidencia que la maquinaria relativamente más nueva que la otra genera menos mermas.

3.4.2. Análisis de los Datos

Una vez el cliente nos facilitó la base de datos se acordó la confidencialidad por motivos de divulgación de costos internos de producción, por lo cual los valores de los datos no son los reales. Estos datos fueron entregados en una plantilla de Excel, que fue descargado de un compartido que manejan los supervisores de producción. La base de datos contiene registros históricos desde el 2020 de forma diaria incluidos los registros de mermas, fechas, líneas de producción, turnos, líderes, categorías de desperdicio y otras variables relevantes.

Para el proyecto se ha seleccionado las variables relevantes que están relacionadas a la producción de merma y creado variables que se creyeron correspondientes y necesarias para el análisis:

#	Column	Non-Null Count	Dtype
0	Fecha	3232 non-null	datetime64[ns]
1	Linea	3232 non-null	object
2	Turno	3232 non-null	object
3	LIDER	3232 non-null	object
4	Enfoque	3232 non-null	object
5	OF	3232 non-null	object
6	SKU	3232 non-null	object
7	Descripción	3232 non-null	object
8	Papel Higienico	3232 non-null	object
9	Metraje (mts)	3232 non-null	float64
10	Unidades Paquete	3232 non-null	int64
11	PaquetesxBTOs	3232 non-null	int64
12	Cantidad de Rollos	3232 non-null	object
13	Cantidad de Rollos X	3232 non-null	object
14	Pallet	3232 non-null	int64
15	BTO´s Sueltos	3232 non-null	int64
16	BTO´s Pallet	3232 non-null	int64
17	Peso Teórico	3232 non-null	float64
18	Peso Produccion	3232 non-null	float64
19	LOG	3232 non-null	float64
20	REBABA (PUNTAS)	3232 non-null	float64
21	ROLLOS	3232 non-null	float64
22	SABANAS DEL PROCESO	3232 non-null	float64
23	ROLLOS RECHAZADOS POR CALIDAD	3232 non-null	float64
24	RECORTE POR PRUEBA	3232 non-null	float64
25	DESMANTE X CALIDAD	3232 non-null	float64
26	DESMANTE POR MANIPULACION	3232 non-null	float64
27	DESCARTE PRODUCCION (KG)	3232 non-null	float64
28	DESCARTE PRUEBA (KG)	3232 non-null	float64
29	DESCARTE OTROS (KG)	3232 non-null	float64
30	TOTAL DESPERDICIO (KG)	3232 non-null	float64
31	PESO TOTAL	3232 non-null	float64
32	MERMA (%)	3232 non-null	float64
33	OBS	3232 non-null	int64

dtypes: datetime64[ns](1), float64(17), int64(6), object(10)

Tabla 7 Información de las variables *Fuente: Propia*

3.4.3. Preparación de los Datos

Una vez analizada la base de datos y entendiendo su estructura y variables se realizó la depuración de los datos, eliminando datos no relacionados a la producción de PH, corrigiendo errores de tipeo y por lo general transformaciones y agregaciones de datos necesarios.

Fecha	Línea	Turno	Líder	Enfoque	OF	SKU	…
3/8/2020	L20	1T	Supervisor 1	Producción	SOF99992	95252825	…
3/8/2020	L10	2T	Supervisor 2	Prueba	SOF99992	59825984	…
3/8/2020	L10	3T	Supervisor 1	Producción	SOF99996	85512684	…
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Tabla 8 Fragmento de los Datos de Entrada *Fuente: Propia*

Una vez lista y depurada la base de datos se comenzó a realizar un análisis exploratorio. Se aplicó todos los análisis en RStudio como revisar los resúmenes estadísticos de cada una de las variables como la media, máximo, mínimos e identificación de cuartiles y verificar que no haya datos atípicos.

DESCARTE OTROS (KG)	TOTAL DESPERDICIO (KG)	PESO TOTAL	MERMA (%)
Min. : 0.00	Min. : 0.0	Min. : 46.72	Min. :0.00000
1st Qu.: 0.00	1st Qu.: 194.0	1st Qu.: 3416.10	1st Qu.:0.03900
Median : 35.00	Median : 329.0	Median : 6648.30	Median :0.04879
Mean : 57.75	Mean : 383.9	Mean : 7159.18	Mean :0.06472
3rd Qu.: 85.00	3rd Qu.: 486.1	3rd Qu.:10131.67	3rd Qu.:0.06398
Max. :826.00	Max. :45309.1	Max. :48648.70	Max. :0.95326

Tabla 9 Estadísticas de las variables *Fuente: Propia*

Con la función summary () se puede identificar el problema planteado, en vista que el promedio general de MERMA (%) es de 0.06472 superior al establecido como aceptable de 0.055, lo cual también se puede observar en el gráfico de distribución.

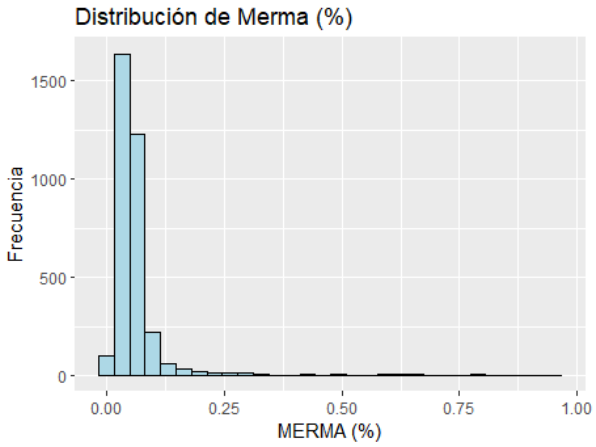


Gráfico 6 Distribución de Merma (%) *Fuente: Propia*

Se analizó las mermas en relación con las líneas de producción para a través de la comparación poder determinar valores atípicos, se utilizó el grafico de caja para identificar estos datos, como se muestra en la gráfica.

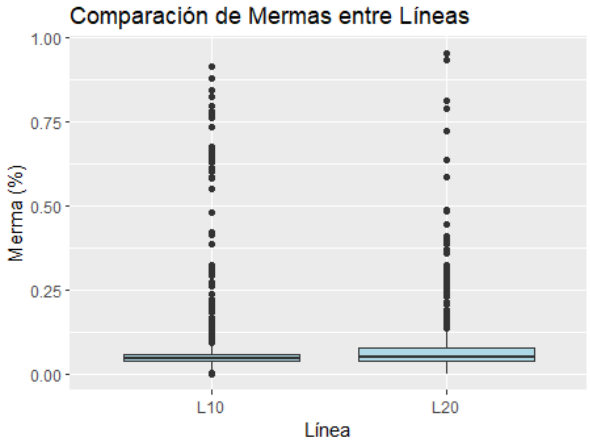


Gráfico 7 Comparación de mermas entre líneas de producción *Fuente: Propia*

Depurar la base de datos y eliminar datos atípicos es esencial para mejorar la calidad y confiabilidad del modelo LSTM. Al eliminar datos ruidosos y sesgados, se logra un conjunto de datos más limpio y representativo, lo que permite al modelo aprender patrones más precisos y hacer predicciones más acertadas. Además, la eliminación de datos atípicos reduce el riesgo de sobreajuste y mejora la estabilidad del modelo en nuevos datos. Esta práctica también ayuda a obtener un conjunto de datos homogéneo y coherente, lo que facilita el aprendizaje de relaciones significativas. Sin embargo, es importante realizar la depuración de datos con cuidado y basada en el conocimiento del dominio, para evitar la eliminación de datos relevantes y valiosos para el análisis.

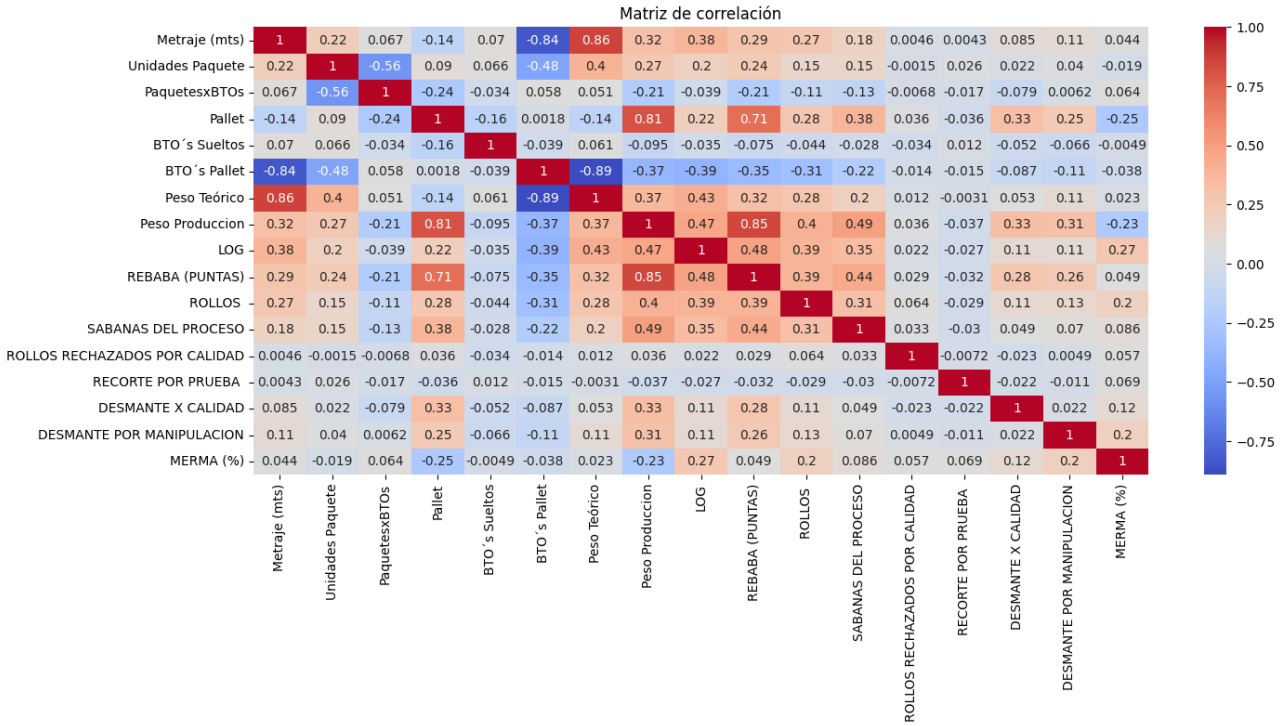
3.4.4. Modelado

El proceso se llevará a cabo en la plataforma de Google Colaboratory para iniciar el desarrollo del modelo de Redes Neuronales de Memoria a Largo Plazo (LSTM). Esto aprovechará su capacidad de cómputo y su facilidad de acceso a las librerías esenciales para el análisis de datos y el diseño de gráficos. Se cargará la base de datos asignada, que proporcionará la información necesaria para el análisis y la construcción del modelo.

Primero, se importarán las librerías de procesamiento y manipulación de datos básicos, como NumPy y Pandas. Estos permiten realizar operaciones numéricas y organizar los datos en estructuras de datos adecuadas para el análisis. Además, se utilizará Scikit-learn para dividir los datos en conjuntos de prueba y entrenamiento, así como MinMaxScaler para escalar los datos para garantizar que todas las variables estén en el rango adecuado para el entrenamiento del modelo.

Antes de comenzar el análisis de LSTM, se utilizarán las características de la librería Matplotlib para crear gráficos y visualizaciones, lo que permitirá explorar la distribución de los datos, detectar patrones y comprender las relaciones entre las variables.

Se llevará a cabo un análisis de correlación completo durante el proceso de modelado para determinar las relaciones existentes entre las diferentes variables del conjunto de datos. Anterior a este análisis, se llevó a cabo una depuración minuciosa. Se eliminó la inclusión de variables que eran la suma o combinación lineal de otras, ya que su inclusión podría sesgar la interpretación de las relaciones entre las variables de interés. Este paso de depuración fue crucial para garantizar que las variables seleccionadas fueran independientes y reflejaran adecuadamente la complejidad del sistema subyacente.



La correlación nos ayuda a comprender cómo se relacionan las variables en los datos. Las variables que están fuertemente relacionadas con la variable objetivo (en este caso, las mermas) tienen un impacto significativo en su comportamiento al analizar la correlación.

3.4.4.1. Implementación de LSTM

El análisis detallado de las características del problema y las ventajas únicas del modelo LSTM conduce a la elección de este modelo como la solución para abordar el desafío de gestión de mermas en la producción de papel higiénico. La capacidad de LSTM para capturar relaciones temporales complejas en datos secuenciales la distingue de otros modelos.

Este modelo es ideal para analizar las interacciones entre estas variables a lo largo del tiempo porque la producción de papel higiénico implica un proceso secuencial con varias etapas interdependientes.

La producción de papel higiénico requiere una comprensión profunda de los patrones de producción y las tendencias históricas. Dado que puede aprender de los datos pasados y utilizar esa información para predecir y detectar patrones anómalos en las mermas de producción futuras, LSTM brilla en este contexto.

En un contexto multidimensional como este, es esencial que el modelo sea flexible para manejar múltiples entradas. El modelo puede usar las diversas variables involucradas en la producción como entradas, lo que permite una representación completa de la información relevante.

El LSTM es capaz de comprender y modelar las relaciones no lineales y secuenciales en los datos, a diferencia de modelos menos apropiados, como los lineales o los de árboles de decisión. La detección efectiva de patrones normales y anomalías en la producción es posible gracias a la capacidad del LSTM para reconstruir secuencias de entrada con baja pérdida. Al implementar este modelo, la gerencia y los auditores internos pueden controlar las mermas, reducir costos innecesarios y mejorar la eficiencia operativa.

3.4.4.2. Modelo para Identificación de Anomalías

El modelo de red neuronal LSTM (Long Short-Term Memory) se utiliza para detectar anomalías en datos secuenciales. La variante LSTM de las redes neuronales recurrentes permite capturar patrones y relaciones en secuencias de datos a lo largo del tiempo. El modelo aprende secuencias normales con poca pérdida y luego evalúa secuencias nuevas en tiempo real. Una anomalía ocurre cuando el error de reconstrucción supera un umbral. Esta técnica permite la detección efectiva de patrones complejos en datos secuenciales, lo que conduce a un control de mermas de producción más efectivo.

Importan librerías clave para la manipulación y análisis de datos. MinMaxScaler es un escalador de datos, mientras que pandas y NumPy se utilizan para la manipulación de datos. Además, para construir el modelo de red neuronal, las clases secuenciales, LSTM y densas de Keras son esenciales.

La estructura secuencial de Keras se utiliza para construir el modelo LSTM. Incluye una capa LSTM con función de activación "relu" y una capa densa con función de activación "sigmoid".

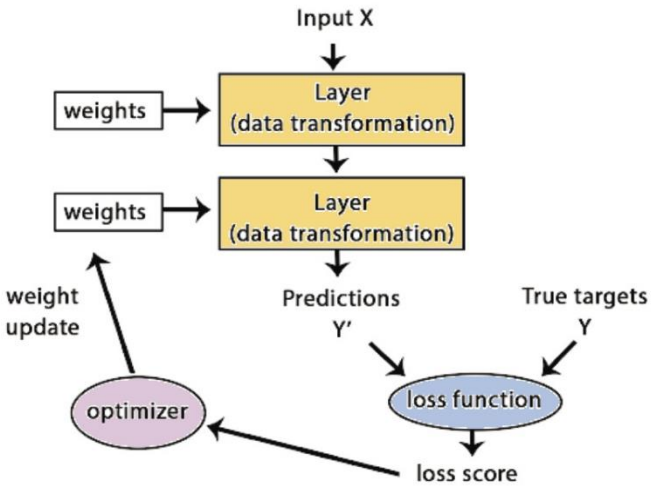


Ilustración 5 Keras Template *Fuente: Sixx*

El modelo se compila con el optimizador "Adam", que es apropiado para entrenar redes neuronales, esta capacidad de adaptación dinámica ayuda al modelo a encontrar anomalías y relaciones subyacentes en los datos, lo que mejora la detección precisa de comportamientos inusuales durante el proceso de producción. Además, se indica la función de pérdida "MSE", también conocida como error cuadrado medio, que mide la discrepancia entre las salidas reales y las predicciones del modelo. El modelo puede detectar patrones y comportamientos normales en la información secuencial. Los datos se dividen en conjuntos de prueba y entrenamiento. Se reformatean las secuencias de entrenamiento para crear secuencias de tiempo y sus respectivas salidas esperadas. El modelo se entrena repetidamente para reducir la pérdida entre las predicciones y las salidas anticipadas.

3.4.4.3. Modelo para Proyección de Categorías de Mermas

El modelo LSTM se basa en las "celdas de memoria", que son unidades especializadas de memoria. Estas celdas pueden almacenar y acceder a datos a lo largo de secuencias de tiempo, lo que permite al modelo retener y aprender patrones a diferentes escalas temporales. Las celdas LSTM están diseñadas para abordar este problema, esencial para el análisis de datos secuenciales en procesos de producción, a diferencia de las redes neuronales recurrentes convencionales, que con frecuencia tienen problemas para recordar patrones a largo plazo.

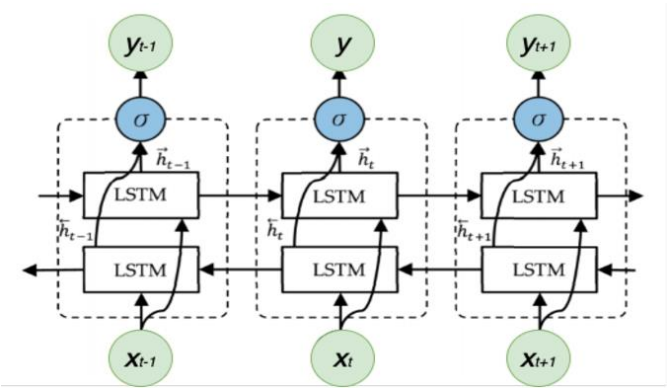


Ilustración 6 Celdas de memoria LSTM *Fuente: MAQSOOD 2021*

El código proporcionado utiliza una capa bidireccional de LSTM para implementar una arquitectura bidireccional de LSTM. Con esta configuración, el modelo puede capturar información futura y anterior en las secuencias de datos. Esto es particularmente útil en situaciones donde los patrones anómalos pueden estar relacionados con eventos futuros. La utilización de la función de salida entre las capas LSTM ayuda a regularizar el modelo y evitar el sobreajuste, lo que aumenta su capacidad de generalización y detección de anomalías precisas.

La capa más baja del modelo es una capa densa en la que la cantidad de neuronas es igual a la cantidad de variables objetivo que se desean predecir. En este caso, las variables LOG, REBABA (PUNTAS) y ROLLOS se utilizaron como objetivos. La función de activación de esta capa se configura en su forma predeterminada, que es la función de activación "lineal". Esta función es adecuada para generar predicciones continuas en el rango real porque el objetivo es predecir valores numéricos.

3.4.5. Explotación (Dashboard de Control)

Se recopilará y visualizará la información obtenida de la base de datos de Control de Mermas de la empresa y todos los datos derivados de los modelos antes descritos. Esta herramienta es la ideal para que actúe de manera interactiva para la gestión y control de mermas del proceso de producción.

CAPÍTULO III

4. RESULTADOS

El capítulo de resultados de este proyecto presenta un análisis detallado de los hallazgos obtenidos mediante dos métodos clave: el uso de un modelo LSTM para Detectar Anomalías y el uso de otro modelo LSTM para proyectar Categorías de Mermas

El primer método es útil para encontrar desviaciones significativas en datos secuenciales, lo que lo convierte en una valiosa herramienta para el monitoreo y mantenimiento predictivo en entornos industriales. Por otro lado, el segundo método permite predecir con gran precisión las categorías de mermas, lo que facilita la planificación estratégica y la toma de decisiones informadas para reducir costos y mejorar la eficiencia operativa.

4.1. Análisis Exploratorio

El aumento en la producción de la empresa de papel higiénico desde 2020 hasta ahora se debe a una creciente demanda del mercado. Este aumento en la producción, como se registra en la tabla, demuestra la capacidad de la empresa para adaptarse a las necesidades cambiantes de los consumidores.

Meses	Años			
	2020	2021	2022	2023
Ene	-	69	71	155
Feb	-	64	68	116
Mar	-	65	83	129
Abr	-	69	91	164
May	-	64	108	144
Jun	-	59	106	118
Jul	-	72	125	83
Ago	60	72	134	-
Sep	59	90	122	-
Oct	63	93	138	-
Nov	45	97	143	-
Dic	43	65	146	-

Tabla 10 Demanda de PH *Fuente: Propia*

El aumento significativo en la producción de papel higiénico está directamente relacionado con la mayor cantidad de mermas generadas durante el proceso. Dado que las complejidades del proceso y las interdependencias entre variables pueden amplificar errores y desperdicios, este aumento en las mermas es una consecuencia lógica del aumento en la actividad productiva.

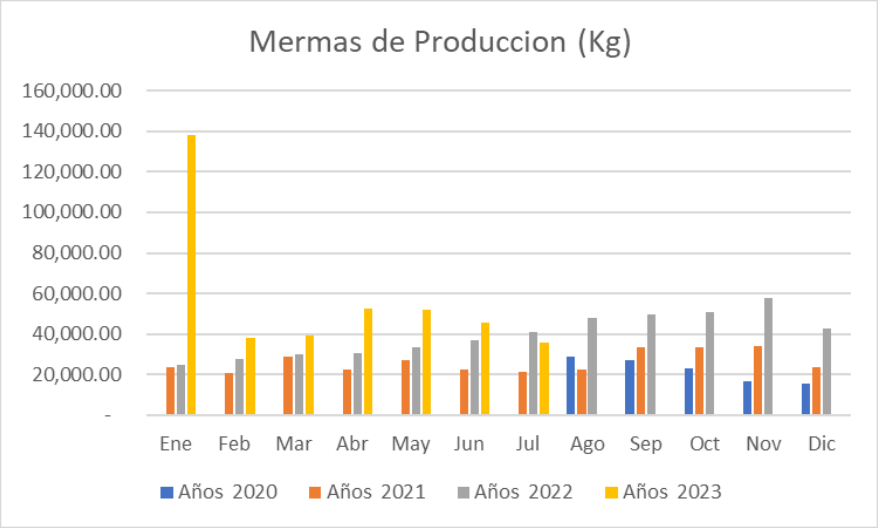


Gráfico 8 Mermas de Producción (Kg) *Fuente: Propia*

En el proceso de fabricación de papel higiénico, hay una variedad de factores interrelacionados que contribuyen a la producción de desperdicios y daños en el producto final, son variables clave que se muestran en los datos "Metraje (mts)", "Unidades Paquete", "Pallet", "BTO´s Pallet", "Peso Teórico".

4.2. Identificación de Anomalías

El modelo de identificación de anomalías utiliza técnicas de aprendizaje automático para detectar patrones inusuales en los datos para distinguir comportamientos normales de anómalos en una variedad de aplicaciones, como la detección de fraudes o el mantenimiento predictivo.

4.2.1. Curva de Pérdida durante el entrenamiento

A medida que se entrena el modelo, se puede ver cómo disminuye la pérdida (MSE). La pérdida disminuye a intervalos regulares, lo que indicaría que el modelo está aprendiendo y mejorando con los datos de entrenamiento.

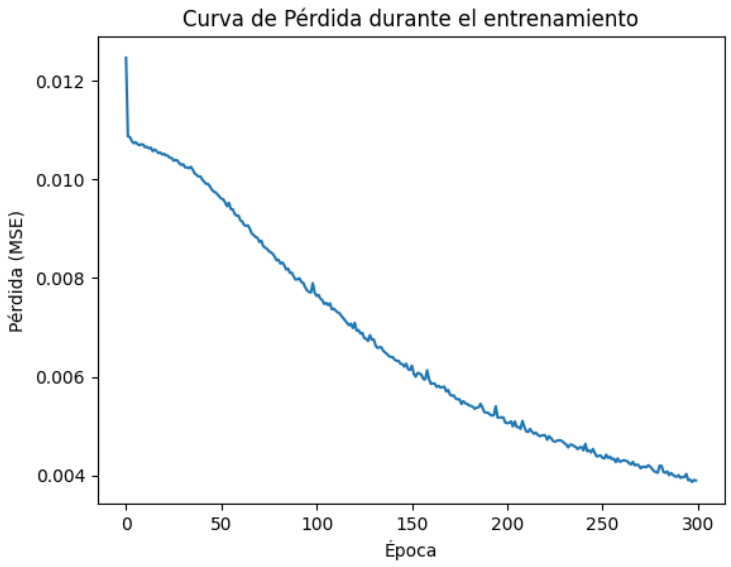


Gráfico 9 Modelo IA-Curva de Pérdida durante el entrenamiento *Fuente: Propia*

El modelo tiene un rendimiento favorable en la predicción de las variables objetivo, según los valores de Mean Absolute Error (MAE), Mean Squared Error (MSE) y Root Mean Squared Error (RMSE).

El modelo tiene un bajo error promedio en las predicciones en valor absoluto con un MAE de 0,0433. Además, un RMSE de 0,0678 y un MSE de 0.0046 indican que los errores cuadráticos son bajos en relación con la magnitud de las variables objetivo.

4.2.2. Evaluación del Modelo

El coeficiente de determinación R2 es del 43.63%, a pesar de los valores extremadamente bajos de MAE y MSE, podría indicar que el modelo no está explicando adecuadamente la variabilidad de los datos. Sin embargo, tener un R2 relativamente bajo podría ser razonable y justificable en el contexto de la identificación de anomalías.

Las anomalías naturales son eventos raros y atípicos, por lo que el modelo podría tener problemas para capturar su variabilidad, lo que podría resultar en un R2 más bajo. Además, en algunos casos de detección de anomalías, el enfoque puede estar en la precisión de las predicciones anómalas en lugar de la explicación de la variabilidad general.

Debido a su capacidad para capturar patrones que se desarrollan con el tiempo y su entrenamiento previo para aprender y reconstruir secuencias de entrada, las redes LSTM son ideales para identificar anomalías en datos secuenciales.

El objetivo es monitorear nuevas situaciones en tiempo real para encontrar anomalías potenciales. Para evaluar la calidad de la predicción, se calcula el Mean Squared Error (MSE) y se reconstruye la secuencia seleccionada utilizando el modelo LSTM.

Se puede observar y representar la secuencia en el gráfico si el error supera el umbral de 0.03, lo que indica que la secuencia es anómala. En general, esta estrategia basada en el error de reconstrucción es útil para identificar eventos inusuales en tiempo real, como se observa en el gráfico.

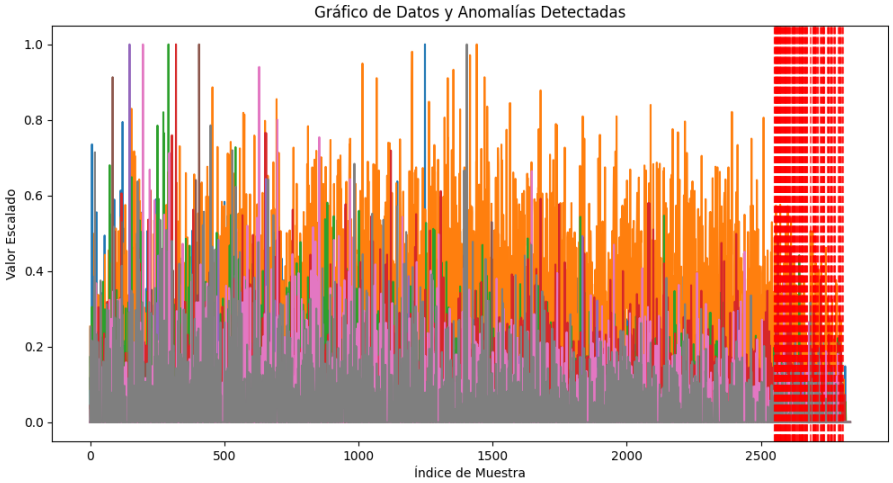


Gráfico 10 Modelo IA-Identificación de Anomalías en los Datos *Fuente: Propia*

4.2.3. Resultados obtenidos

De la base de datos de registros de mermas, se han identificado 284 registros en el periodo enero 2020- julio 2023, que presentan anomalías con base al modelo y sus análisis. Estas anomalías detectadas como se observan en la tabla deberán ser revisadas debido a que, con base en el análisis este grupo de 284 registros contienen datos que difieren significativamente de la distribución general del modelo.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fecha	Linea	Turno	LIDER	Enfoque	OF	SKU	Papel Higienico	Metraje (mts)	Unidades Paquete	Paquetes xBTOs	Pallet	BTO´s Sueltos	BTO´s Pallet
2	2023-05-02 00:00:00	L10	1T	Lider 1	Producción	18008935	1TTEC001020	PH KIND DH	15	6	8.00	55.00	19.00	36.00
3	2021-08-10 00:00:00	L10	1T	Lider 1	Producción	SOF99999	1TTEC001052	PH CLASICO DH	17	12	4.00	120.00	-	36.00
4	2023-04-14 00:00:00	L20	2T	Lider 1	Producción	18008916	1TTEC001019	PH KIND DH	15	4	12.00	28.00	-	40.00
5	2022-10-27 00:00:00	L10	3T	Lider 1	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	140.00	-	36.00
6	2021-02-09 00:00:00	L10	1T	Lider 2	Producción	SOF99999	1TTEC000941	PH CLASICO DH	18	12	4.00	165.00	2.00	36.00
7	2021-04-13 00:00:00	L20	1T	Lider 2	Producción	SOF99999	1TTEC001000	PH KIND DH	17	2	1.00	25.00	-	36.00
8	2021-08-10 00:00:00	L10	1T	Lider 1	Producción	SOF99999	1TTEC001052	PH CLASICO DH	17	12	4.00	120.00	-	36.00
9	2022-10-27 00:00:00	L10	3T	Lider 1	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	140.00	-	36.00
10	2020-12-01 00:00:00	L10	1T	Lider 2	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	95.00	-	36.00
11	2021-02-09 00:00:00	L10	1T	Lider 2	Producción	SOF99999	1TTEC000941	PH CLASICO DH	18	12	4.00	165.00	5.00	36.00
12	2021-08-10 00:00:00	L10	1T	Lider 1	Producción	SOF99999	1TTEC001052	PH CLASICO DH	17	12	4.00	120.00	-	36.00
13	2023-04-14 00:00:00	L20	2T	Lider 1	Producción	18008916	1TTEC001019	PH KIND DH	15	4	12.00	28.00	-	40.00
14	2020-12-01 00:00:00	L10	1T	Lider 2	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	95.00	-	36.00
15	2022-10-27 00:00:00	L10	3T	Lider 1	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	140.00	345.00	36.00
16	2021-04-13 00:00:00	L20	1T	Lider 2	Producción	SOF99999	1TTEC001000	PH KIND DH	17	2	1.00	25.00	-	36.00
17	2023-04-14 00:00:00	L20	2T	Lider 1	Producción	18008916	1TTEC001019	PH KIND DH	15	4	12.00	28.00	-	40.00
18	2021-08-10 00:00:00	L10	1T	Lider 1	Producción	SOF99999	1TTEC001052	PH CLASICO DH	17	12	4.00	120.00	-	36.00
19	2021-08-10 00:00:00	L10	1T	Lider 1	Producción	SOF99999	1TTEC001052	PH CLASICO DH	17	12	4.00	120.00	-	36.00
20	2021-02-17 00:00:00	L10	1T	Lider 3	Producción	SOF99999	1TTEC000869	PH CLASICO TH	35	12	4.00	75.00	-	27.00
21	2022-10-27 00:00:00	L10	3T	Lider 1	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	140.00	-	36.00
22	2021-02-17 00:00:00	L10	1T	Lider 3	Producción	SOF99999	1TTEC000869	PH CLASICO TH	35	12	4.00	75.00	-	27.00
23	2022-10-27 00:00:00	L10	3T	Lider 1	Producción	SOF99999	1TTEC001020	PH KIND DH	15	6	8.00	140.00	-	36.00
24	2022-12-20 00:00:00	L20	1T	Lider 1	Producción	SOF99999	1TTEC001019	PH KIND DH	15	4	12.00	15.00	-	40.00
25	2021-02-09 00:00:00	L10	1T	Lider 2	Producción	SOF99999	1TTEC000941	PH CLASICO DH	18	12	4.00	165.00	2.00	36.00

Tabla 11 Anomalías Detectadas *Fuente: Propia*

4.3. Proyección de categorías de mermas de producción

El modelo LSTM para predecir múltiples variables objetivo en series temporales. Las LSTM pueden capturar patrones y dependencias a largo plazo en los datos, lo que las hace adecuadas para tareas de predicción en secuencias.

4.3.1. Curva de Pérdida

La evolución de la función de pérdida (MSE) durante el entrenamiento se ilustra en la curva de pérdida. El enfoque LSTM es adecuado para capturar patrones complejos y dependencias temporales, lo que mejora la capacidad de generalización del modelo y permite realizar predicciones precisas en nuevos datos. Esto se demuestra por una disminución suave y estable.

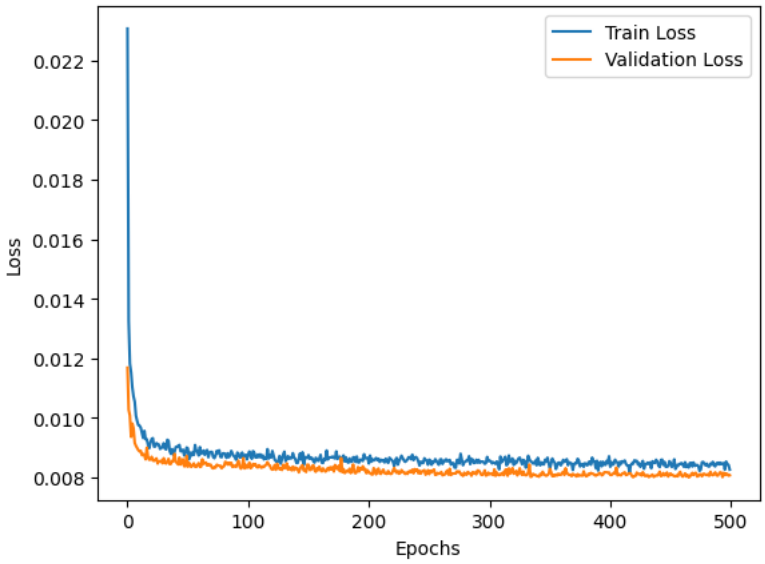


Gráfico 11 Modelo PCM-Curva de Pérdida durante el entrenamiento

Considerando el contexto del problema y las características de los datos, los valores MSE y MAE obtenidos, con un MSE de 1012.25 y un MAE de 25.93, parecen ser razonablemente buenos.

El MSE muestra que el modelo tiene cierta dispersión en sus predicciones, lo cual es común en problemas con una amplia gama de valores objetivo. Sin embargo, el MAE indica que las predicciones tienen un error absoluto promedio de aproximadamente 27.95 unidades, lo cual es aceptable dependiendo de la escala y precisión requeridas en la aplicación.

Además, es importante tener en cuenta que las métricas pueden variar según los datos y el tipo de problema. En general, estos resultados indican que el modelo es adecuado para la tarea en cuestión y puede hacer predicciones precisas.

4.3.2. Evaluación del Modelo

Según los resultados, los datos parecen ser útiles y de calidad para tu modelo. El coeficiente de determinación R2 del 40.01% es moderado, pero los valores MSE y MAE son razonablemente bajos, lo que indica que el modelo tiene una buena capacidad para hacer predicciones precisas con un error absoluto promedio bajo. Además, la curva de pérdida durante el entrenamiento muestra una convergencia estable, lo que indica que el modelo ha aprendido a partir de los datos de entrenamiento de manera efectiva.

4.3.3. Resultados Obtenidos

En el siguiente grafico se puede visualizar las predicciones en comparación con los valores reales para cada variable objetivo. Se puede comparar visualmente la precisión del modelo en la predicción de cada variable objetivo al representar las curvas correspondientes a las predicciones y los valores reales para cada variable objetivo en subplots distintos.

Si las curvas de predicción están muy cerca de las curvas de los valores reales, significa que el modelo está haciendo predicciones precisas y siguiendo las tendencias de los datos reales de manera adecuada.

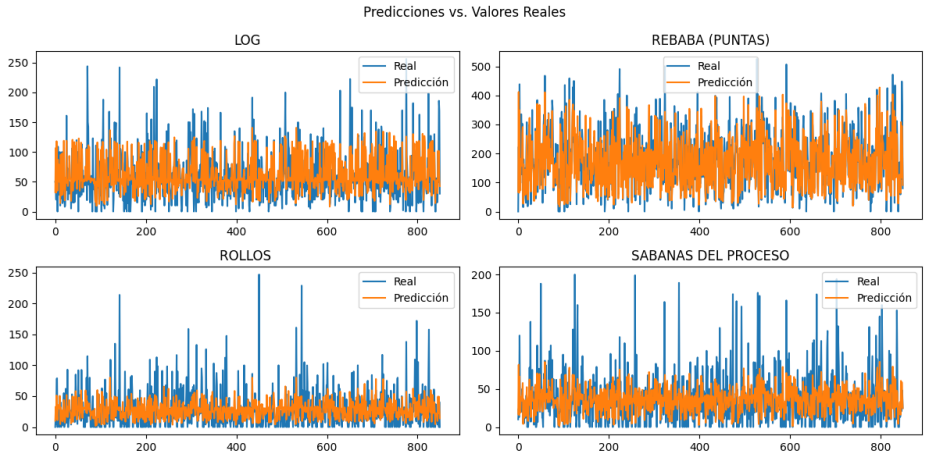


Gráfico 12 Modelo PCM-Ajuste del modelo *Fuente: Propia*

4.4. Dashboard de Control

El desarrollo e implementación de un dashboard efectivo para controlar las mermas de producción. Este dashboard se convierte en una herramienta crucial para la presentación y el análisis de los resultados obtenidos a través de la plataforma Power BI, utilizando los modelos de identificación de anomalías y proyección de categorías de mermas. El dashboard, Anexo 1, crea un espacio interactivo y visualmente intuitivo donde los insights de los modelos se presentan de manera clara y comprensible, lo que permite una toma de decisiones informada y la identificación temprana de tendencias y patrones relevantes en los datos.

CAPÍTULO IV

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En Ecuador, la aplicación de métodos de machine learning para controlar las mermas en la producción de papel higiénico podría ser una solución exitosa y prometedora. El modelo LSTM ha permitido el análisis de grandes cantidades de datos, la identificación de patrones complejos y la proyección de categorías de merma con un ajuste óptimo. Esto podría dar resultado a una disminución significativa del porcentaje de merma, lo que se traduce alrededor de 250,000.00 toneladas de desperdicios y costos innecesarios relacionados con las mermas.

Un resultado significativo es que estos modelos brindan una ventaja competitiva a las industrias de papel higiénico del país. Las empresas pueden tomar decisiones informadas que mejoran la eficiencia operativa y la calidad del producto al detectar anomalías, optimizar procesos de producción y predecir categorías de merma. Las industrias pueden aumentar su competitividad en un mercado que se enfoca cada vez más en la eficiencia y la innovación con la implementación adecuada de estas herramientas.

En este contexto, el papel del auditor interno se vuelve crucial. El auditor puede aprovechar estas tecnologías ya existentes para controlar las mermas de producción de la empresa, aunque no es necesario que sea un experto en programación. Al trabajar con estos modelos ya existentes, el auditor puede obtener información útil para monitorear el proceso de producción, identificar áreas de mejora y sugerir soluciones. La adopción de estas herramientas de análisis de datos permite al auditor tener una visión más profunda y precisa de los patrones de merma, lo que facilita la toma de decisiones informadas para reducir y reducir la cantidad de

desperdicio en la producción.

Además de los beneficios financieros y operativos, el uso de estos modelos tiene un efecto beneficioso en el medio ambiente. Un mejor control y optimización reduce las mermas de producción, lo que conduce a una producción más sostenible y a la preservación del medio ambiente.

Por último, los modelos de machine learning utilizados en la producción de papel higiénico ofrecen una solución efectiva para las industrias del sector. Estas herramientas son útiles para el auditor interno para monitorear y controlar las mermas, lo que reduce significativamente los costos, los desperdicios y el impacto positivo en el medio ambiente. El uso de tecnologías avanzadas y la innovación son el camino hacia una gestión más eficiente y sostenible en el mundo empresarial, y estos modelos son un ejemplo claro de cómo estas herramientas pueden beneficiar a las empresas de papel higiénico y a la sociedad en su conjunto.

5.2. Recomendaciones

- Brindar capacitaciones a los supervisores y personal involucrado en la captura de datos sobre la importancia de registros precisos y su impacto en la gestión de mermas.
- Implemente un sistema de registro automatizado de la producción, idealmente integrado con el sistema de control implementado en este proyecto. Esto reducirá la posibilidad de errores manuales en los registros y garantizará la veracidad de los datos.
- Establezca normas y procedimientos claros para los registros de producción y pruebas. Esto incluye asegurarse de que cada producción y prueba esté vinculada a una orden de fabricación y tenga su acta de validación correspondiente.
- Realizar una reconciliación periódica y exhaustiva entre los datos registrados en la plataforma SAP y los registrados en planta. Esto ayudará a identificar, corregir las discrepancias o determinar en qué instancias se están registrando contablemente las mermas.
- Agregar al registro de producción columnas indispensables que no solo ayudan al modelo a entender y aprender de él, sino que también permiten un control más eficiente. Se sugiere agregar las columnas de Categoría de Papel Higiénico, Metraje (mts), Peso de Producción, Peso Total, Merma (%), como se sugiere en el Anexo 2.
- Realizar un mapeo del proceso de producción para poder determinar políticas del manejo de mermas y sus controles.
- Revisar el modelaje planteado en el proyecto cada cierto periodo de tiempo para realizar ajustes pertinentes, de ser necesario, para que el modelo trabaje coherentemente al cambio de la situación de la empresa.

6. REFERENCIAS

Acalá, U. d. (28 de 02 de 2020). *Master Big Data, Data Science y Business Intelligence* . Obtenido de Master Big Data, Data Science y Business Intelligence : https://master-bigdata.com/origen-big-data/

Allaire, J. J. (2011). RStudio: Integrated Development Environment for R . *Book of Contributed Abstracts,* 14.

Arevalo, D., & Padilla, C. (2016). Medición de la Confiabilidad del Aprendizaje del Programa Rstudio Mediante Alfa de Cronbach. *Revista Politecnica* .

BDO, A. (28 de 02 de 2020). *S.L.P. y BDO Abogados y Asesores Tributarios*. Obtenido de S.L.P. y BDO Abogados y Asesores Tributarios: https://www.bdo.es/es-es/blogs-es/blog-coordenadas-bdo/abril-2018/el-futuro-de-la-auditoria

datacamp. (Septiembre de 2022). *datacamp*. Obtenido de RStudio Tutorial: https://www.datacamp.com/tutorial/r-studio-tutorial

Domínguez, O. T. (2018). *Detección de anomalías en grandes volúmenes de datos.* Bogota: Revista Facultad de Ingeniería .

Ferrer, A. (2017). *Mermas y Desmedros Criterios Contables y Tributarios.* Lima: Universidad Peruana de Ciencias Aplicadas .

Grégoire Montavon, G. O. (2012). *Neural Networks: Tricks of the Trade* . Springer.

Hochreiter, S. &. (1997). *Long Short-Term Memory. Neural Computation.* Canada.

Juan R. Camarillo-Peñaranda, A. J.-M. (2013). *Recomendaciones para Seleccionar Índices para la Validación de Modelos.* Tecno.Lógicas.,ISSN 0123-7799.

MAQSOOD, S. (2021). *School of Information and Communications Technology, University of Tasmania, Hobart,.* IEEE Access.

Mera, P. D. (2019). *Importancia de los costos y el control en la gestión de la calidad de bienes y servicios.* Guayaquil: Revista Científica.

Michael Paluszek, S. T. (2019). *MATLAB Machine Learning Recipes: A Problemsolution Approach.* . Nueva Jersey: Apress.

Microsoft. (2020). *Microsoft BI.* Obtenido de Microsoft BI.: https://powerbi.microsoft.com/es-

es/what-is-power-bi/

Robadilla, J. (2020). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras.* Bogotá. Colombia: Ra-ma Editorial.

Sebastian Raschka, V. M. (2019). *Aprendizaje automático y aprendizaje profundo con Python, scikit-learning y TensorFlow*. Barcelona, España: Marcombo.

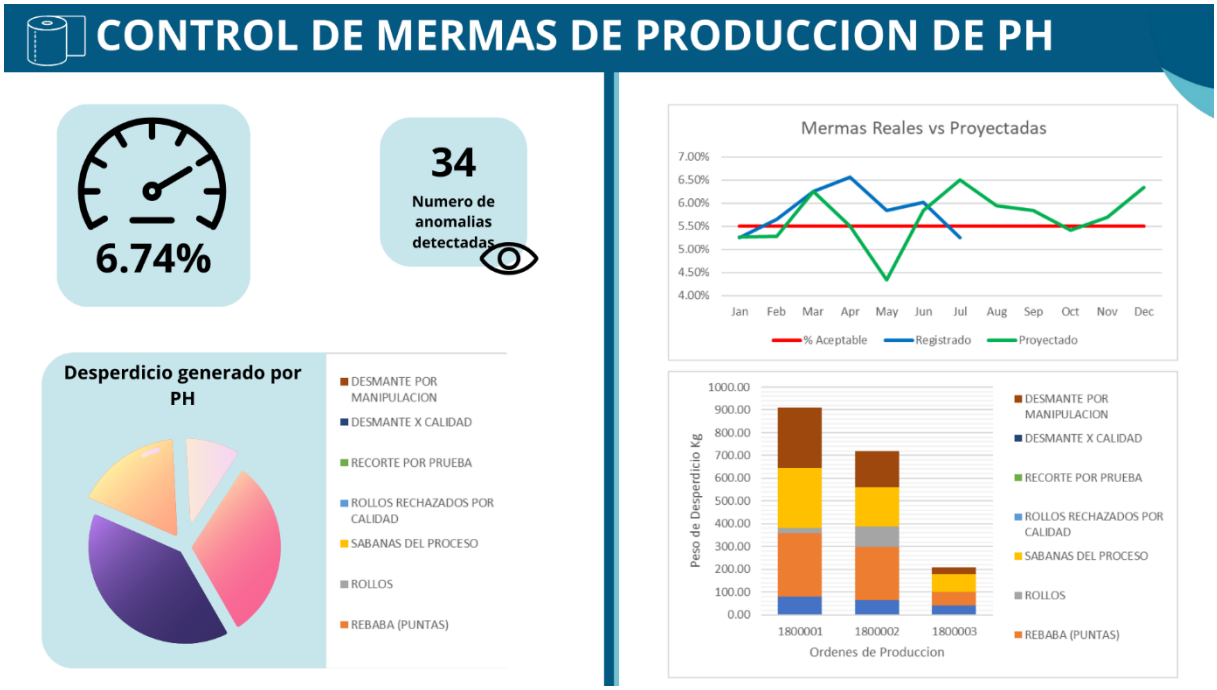
Softeng. (2015). *Softeng*. Obtenido de Softeng: https://www.softeng.es/blog/la-nueva-herramienta-de-office-365-para-trabajar-con-tus-datos-power-bi/

Tapia, C. K. (2017). *Auditoria Interna, Perspectiva de vanguardia 1ra edición.* Ciudad de Mexico: Instituto Mexicano de contadores Público.

Villardefrancos Álvarez, M. (2006). *Ciencias de la Información Vol. 37.* La Habana, Cuba: Instituto de Información Científica y Tecnológica.

Wirth, R. (2019). *CRISP-DM: Hacia un modelo de proceso estándar para la mineria de datos.* Alemania: DaimlerChrysler Research & Technology.

7. ANEXOS



Anexo 1 Dashboard Control de Merma
NOTA: Esta es una opción de como poder presentar los resultados obtenidos, el dashboard cambiara de las necesidades del cliente, empresa.

Fecha	Linea	Turno	LIDER	Enfoque	OF	SKU	Descripción	Papel Higienico	Metraje (mts)	Unidades Paquete	PaquetesxBTOs	Cantidad de Rollos	Cantidad de Rollos X	Pallet	BTO's Sueltos	BTO's Pallet	Peso Teórico	Peso Produccion	LOG	REBABA (PUNTAS)	ROLLOS	SABANAS DEL PROCESO	ROLLOS RECHAZADOS POR	RECORTE POR PRUEBA	DESMANTE X CALIDAD	DESMANTE POR MANIPULACIO	DESCARTE PRODUCCION (KG)	DESCARTE PRUEBA (KG)	DESCARTE OTROS (KG)	TOTAL DESPERDICIO (KG)	PESO TOTAL	MERMA (%)	OBS

Anexo 2 Registro de Mermas de Producción