Implementación de un modelo analítico para la predicción de la venta del portafolio de productos OTC de un Laboratorio Farmacéutico

Molina Rea, Karina Gabriela

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de Titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios

Msc. Jaramillo Vinueza, Byron Orlando

15 de septiembre del 2020

URKUND

Urkund Analysis Result

Document Information

Analyzed document	TESIS KARINA MOLINA - ANALISIS URKUND.docx (D79429885)
Submitted	9/19/2020 1:01:00 AM
Submitted by	Gualotuña Alvarez Tatiana Marisol
Submitter email	tmgualotunia@espe.edu.ec
Similarity	1%
Analysis address	tmgualotunia.espe@analysis.urkund.com

Firma:

...

Msc. Jaramillo Vinueza, Byron Orlando

DIRECTOR

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Implementación de un modelo analítico para la predicción de la venta del portafolio de productos OTC de un Laboratorio Farmacéutico**" fue realizado por la Ing. **Molina Rea, Karina Gabriela** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí,15 de septiembre de 2020

Firma:

.....

Msc. Jaramillo Vinueza, Byron Orlando

Director

C.C.:1714555255

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo **Molina Rea, Karina Gabriela**, con cédula de ciudadanía n°1721512711, declaro que el contenido, ideas y criterios del trabajo de titulación**: Implementación de un modelo analítico para la predicción de la venta del portafolio de productos OTC de un Laboratorio Farmacéutico** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 15 de septiembre de 2020

Firma

.....

Molina Rea, Karina Gabriela

C.C.: 1721512711

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo **Molina Rea, Karina Gabriela** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Implementación de un modelo analítico para la predicción de la venta del portafolio de productos OTC de un Laboratorio Farmacéutico** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 15 de septiembre de 2020

Firma

.....

Molina Rea, Karina Gabriela

C.C.: 1721512711

DEDICATORIA

El presente trabajo de investigación está dedicado a todas las personas que me brindaron su apoyo incondicional en cada etapa de este proyecto.

AGRADECIMIENTOS

A Dios, por brindarme la sabiduría y por guiarme en cada instante de este camino.

A mi familia, por siempre estar presentes con su apoyo, dándome la fortaleza necesaria para concluir mis objetivos propuestos.

A mi Director de tesis, Msc Byron Jaramillo quien con sus conocimientos supo guiarme, para la realización exitosa del presente proyecto.

A todas las personas que no dudaron de mí y me respaldaron día a día.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

Resumen

En la industria farmacéutica la estimación de las ventas proyecta un análisis de la demanda futura para la toma de decisiones del negocio y un panorama amplio para la asignación de recursos sobre este. El presente proyecto presenta la búsqueda del mejor modelo analítico para la predicción de las ventas de la línea OTC de un Laboratorio Farmacéutico, a través de la metodología de investigación Design Science Research y también la aplicación de la metodología KDD. Para la construcción del modelo analítico se emplearon los modelos ARIMAX, redes neuronales y Holt Winters a través del método de regresión y series temporales, ya que se utilizaron datos históricos de las ventas en unidades a nivel de subcategoría de producto, para el entrenamiento de datos en la herramienta analítica Rapidminer. El forecast mostró que cada subcategoría tiene un comportamiento de la demanda diferente por lo que un modelo no necesariamente se ajusta a todas las subcategorías. La comparación con los pronósticos de los modelos obtuvo una precisión menor al 10%, lo cual fue aceptable y contribuyó a la disminución de error que manejaba el Laboratorio Farmacéutico, el cual era del 15% en promedio, es decir que las unidades proyectadas con los modelos construidos se ajustan a la venta real del periodo pronosticado.

Palabras Clave:

- **FARMACÉUTICA**
- **ARIMAX**
- **REDES NEURONALES**
- **HOLT WINTERS**
- **FORECAST**

Abstract

In the pharmaceutical industry, sales estimates project an analysis of future demand for business decisions and a broad picture for the allocation of resources to the business. This project presents the search of the best analytical model for the prediction of the sales of the OTC line of a Pharmaceutical Laboratory, through the research methodology Design Science Research and also the application of the KDD methodology. For the construction of the analytical model, ARIMAX, neural networks and Holt Winters models were used through the regression method and time series, since histórical data of sales in units at the level of product subcategory were used for data training in the analytical tool Rapidminer. The forecast showed that each subcategory has a different demand behavior so a model does not necessarily fit all subcategories. The comparison with the models' forecasts obtained an accuracy lower than 10%, which was acceptable and contributed to the decrease of error handled by the Pharmaceutical Laboratory, which was 15% on average, that is, the projected units with the built models adjust to the real sales of the forecast period.

Keywords:

- **PHARMACEUTICAL**
- **ARIMAX**
- **NEURAL NETWORKS**
- **HOLT WINTERS**
- **FORECASTING**

Capítulo I

Introducción

Antecedentes

Las empresas de la industria farmacéutica requieren obtener una visión precisa de la venta de sus productos para su pronóstico y consecuentemente para el cumplimiento de sus objetivos comerciales. “En la actualidad, el mercado farmacéutico ecuatoriano se caracteriza por un complejo sistema de producción, compra, distribución y dispensación de medicamentos” (ORTIZ, GALARZA, CORNEJO, FERNANDO, & PONCE, 2014). De acuerdo con Chris Tauton y Jonathan Feinbaum; existen cambios en el plazo de la demanda de los productos farmacéuticos continuamente, por lo que la implementación de pronósticos de la demanda es un punto de partida clave para que la cadena de abastecimiento mejore la sinergia entre ventas, distribución y producción.

Las complejidades de los pronósticos tienen varios niveles de datos requeridos, las previsiones tienen diferentes construcciones estructurales. En el área comercial los pronósticos se utilizan para tomar decisiones de asignación de recursos, es decir que la previsión crea presión en el constructo del modelo y los análisis, por lo que el desafío clave para la previsión es crear un proceso en el que las necesidades de la función se puedan cumplir sin comprometer la integridad del enfoque de pronóstico.

La empresa que se denominará por efectos de confidencialidad como “El Laboratorio Farmacéutico” tomado como caso de estudio, ha logrado un crecimiento sostenido a nivel nacional, llegando a ubicarse en posiciones importantes en la industria farmacéutica ecuatoriana. Además, cuenta con una planta de producción nacional, con el objetivo de poner a disposición de la comunidad médica y de todos los ecuatorianos una amplia gama de productos farmacéuticos

orientados a satisfacer las necesidades médicas existentes. El portafolio de productos cuenta con medicamentos OTC (Over the Counter), que por sus siglas en inglés significa “sobre mostrador” y se refiere a los medicamentos que no requieren de receta médica, es decir que son de venta libre.

En el último año se ha desabastecido un producto de la clasificación de vitaminas que tiene alta rotación en el mercado OTC. Por otro lado, en el mismo año se generó sobre stock de un producto de la clasificación de digestivos que tiene ventas bajas, razón por el cual esto ocasiona incumplimiento del objetivo de venta mensual con respecto a las unidades de venta, perjudicando los indicadores de la gestión comercial.

Figura 1

Comparación de Ventas y Stock

		UNIDADES					
		MES					
LOP	PRODUCTO	1	2	3	PROMEDIO VENTA	STOCK	MESES INVENTARIO
OTC	ITEM 1	840	240	240	440	5,000	11.36

		UNIDADES					
		MES					
LOP	PRODUCTO	1	2	3	PROMEDIO VENTA	STOCK	MESES INVENTARIO
OTC	ITEM 2	10,000	9,500	12,000	10,500	5,439	0.52

Tomado de (Laboratorio Farmacéutico, 2019)

Como se aprecia en la figura 1, el stock del ítem 1 tiene 11.36 meses de inventario, el cual corre el riesgo de pasar a la bodega de los productos menor a un año de vencimiento y por ende a proceso de destrucción, ocasionando pérdida en valores y perjudicando la rentabilidad de la línea de productos OTC. Por el contrario, el ítem 2 tiene un stock menor a un mes, 0.52

respectivamente, es decir que probablemente se deje de atender órdenes de compra ya que el stock actual no abastece al promedio de unidad de venta mensual.

Justificación e Importancia

El presente proyecto tiene como fin recuperar la venta sell out[1] pérdida en la línea de productos OTC, es decir que con el análisis predictivo de la venta se logrará cubrir las unidades faltantes y que el área de planificación pueda realizar un trabajo más preciso con respecto a la producción de estos productos.

Según el ALFE[2], la industria farmacéutica es un referente de calidad, eficiencia y desarrollo que expande su campo de acción para beneficio de la salud de los ecuatorianos.

"En los últimos 5 años, la inversión total de la industria farmacéutica nacional superó los 90 millones de dólares. Esta industria busca la generación del mayor valor agregado posible, por lo que la inversión en nueva tecnología es permanente, esto contribuye a garantizar el acceso a medicamentos hechos en Ecuador". (ALFE, 2017) Consecuentemente, la tecnología es un área importante donde la industria farmacéutica enfoca sus objetivos empresariales. El Laboratorio Farmacéutico necesita tener una visión amplia de la producción de los productos de la línea OTC, en base a una buena planificación de la venta.

La demanda de medicamentos en el país está determinada por la prescripción médica, sin embargo, los productos OTC que conforman el mercado de consumo o popular tienen otra figura de regulación, ya que no dependen de la prescripción médica, por lo que son catalogados de venta libre.

[1] Adquisición del producto en el punto de venta (ventas que podrían beneficiar a más personas, además de nuestro cliente directo)

[2] ALFE (Asociación de Laboratorios Farmacéuticos Ecuatorianos)

La importancia con respecto a la congruencia entre la demanda y la capacidad de oferta debe incidir en el análisis de predicción de la venta de los productos de la línea, para el abastecimiento adecuado y oportuno en el mercado de consumo.

Este proyecto permitirá al laboratorio farmacéutico tener una visión correcta de la planificación de ventas y optimización de recursos de producción, con la construcción del modelo analítico de predicción de la venta.

Situación Actual del Laboratorio Farmacéutico

De acuerdo con la metodología de investigación aplicada en el presente proyecto, Design Science Research (DSR)[3]. Se articulan inicialmente las características para resolver el problema de investigación en base a la construcción y aplicación del diseño de la solución.

Bajo esta premisa, la solución debe resolver un problema importante, en este caso para la organización. La situación actual del laboratorio farmacéutico no analiza un modelo predictivo de la venta y realiza los pronósticos de ventas de la línea OTC de forma cualitativa, basándose en una evaluación subjetiva de las unidades de venta, es decir que los cálculos realizados se basan simplemente en el histórico de ventas, calculando la venta promedio mensual sin considerar los diversos factores que influyen en la demanda de los productos. La estimación de ventas se realiza a través del denominado forecast, dicho procesamiento es una técnica predictiva que utiliza datos históricos para provisionar la venta en un periodo determinado de tiempo, en este caso el

[3] DSR (Design Science Research) metodología de investigación de tecnología de la información basada en resultados, que ofrece pautas específicas para la evaluación e interacción dentro de los proyectos de investigación.

laboratorio farmacéutico lo realiza de manera mensual proyectando para los siguientes tres meses de venta:

Figura 2

Forecast Laboratorio Farmacéutico

Línea	Producto	ene-20	feb-20	mar-20	abr-20	may-20	jun-20	YTD	PROMEDIO	STOCK	jul-20	ago-20	sep-20
OTC	ITEM 1	1.055	1.055	1.206	1.206	1.206	1.356	7.084	1.256	1.000	1.356	1.206	1.356
OTC	ITEM 2	8.439	8.252	9.002	8.815	8.815	10.127	53.450	9.252	15.020	9.940	9.002	10.127
OTC	ITEM 3	14.116	14.116	16.133	16.133	16.133	18.150	94.781	16.805	10.000	18.150	16.133	18.150
OTC	ITEM 4	6.215	6.215	7.104	7.104	7.104	7.992	41.734	7.400	1.000	7.992	7.104	7.992
OTC	ITEM 5	5.298	5.298	6.056	6.056	6.056	6.813	35.577	6.308	0	6.813	6.056	6.813

Tomado de (Laboratorio Farmacéutico, 2019)

Como se aprecia en la figura 2, la empresa solo realiza el forecast en base a la técnica de media móvil, es decir la venta promedio, desconociendo su demanda real, ocasionando innumerables pérdidas en ventas y desperdicios del producto, afectado además a los tiempos que toma la realización del forecast mensual.

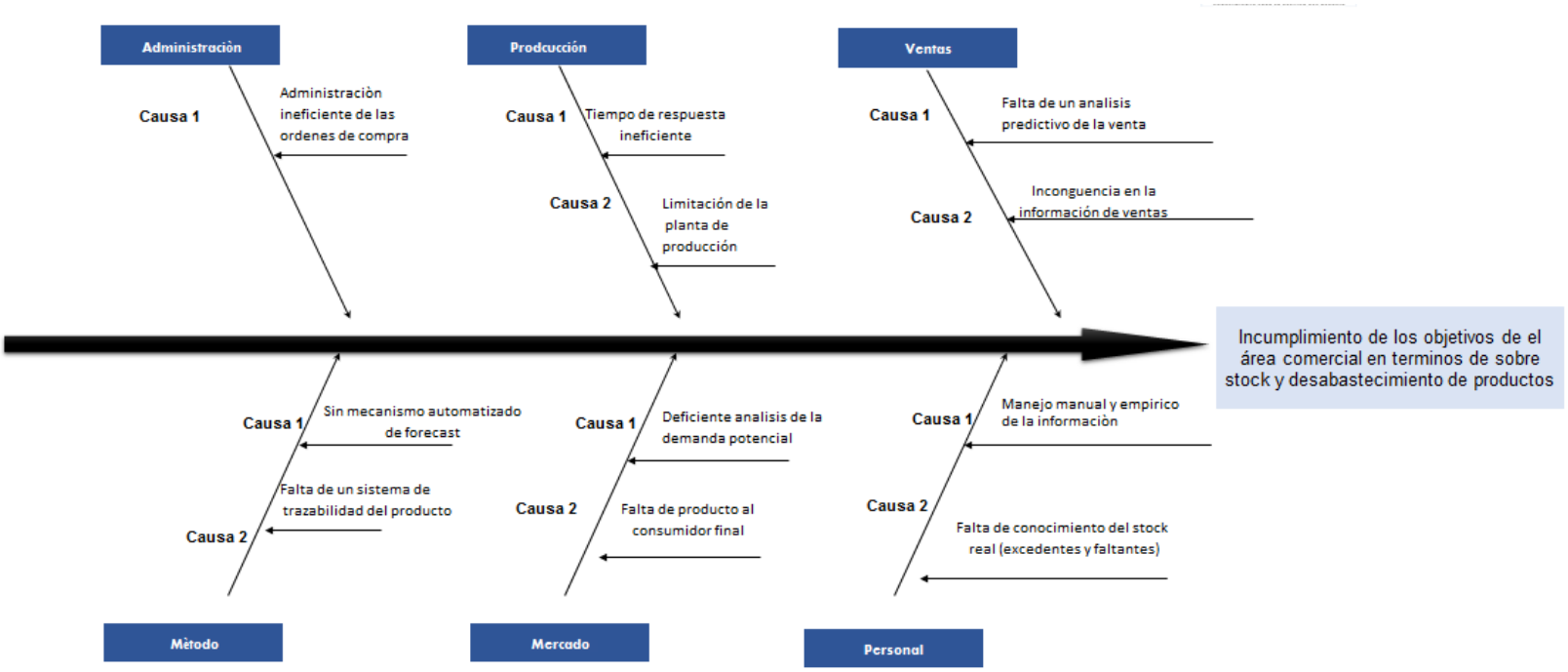
Planteamiento del Problema

El laboratorio farmacéutico, tiene sobre stock de productos con baja rotación y, por otro lado, tiene un nivel bajo de stock de productos con alta rotación en el mercado. Este problema genera ineficiencia y falta de reacción en el área comercial para el cumplimiento de sus objetivos en términos de unidad de venta, las devoluciones de productos con fechas de expiración cortas producen una pérdida importante en los costos de incineración y ventas del laboratorio. En el año 2018 se dejó de atender el 11% de órdenes de compra por falta de stock y las devoluciones representaron el 10% de la venta bruta en el mismo año. (LABORATORIO, 2018)

A continuación, se presenta el diagrama de Ishikawa donde se analiza el problema presentado a causa de la falta de un análisis predictivo de la venta con respecto a la falta de stock y sobre stock:

Figura 3

Diagrama de Ishikawa



En la figura 3 se puede apreciar que, el Laboratorio Farmacéutico aborda específicamente como problema el incumplimiento de los objetivos de venta en unidades del área comercial de la línea de productos OTC, a causa de una inadecuada planificación de la venta, limitada capacidad de producción, incorrecto análisis del mercado, una mala administración de las órdenes de compra, inadecuada aplicación del método de análisis de forecast, y un incorrecto manejo de la información de stock por parte del personal involucrado. Cada causa analiza las dificultades que atraviesa el problema central del laboratorio farmacéutico dentro de la línea de productos OTC. Los problemas dentro de la gestión comercial están enmarcados por dos factores específicos:

Factores Tangibles, que incluyen datos de los objetivos del negocio, información histórica, ventas (KYOCERA, 2016)

Factores Intangibles, que incluyen estacionalidad de venta de un medicamento (KYOCERA, 2016)

Sin modelos analíticos que permitan obtener una predicción de la venta, el área comercial del Laboratorio Farmacéutico no puede anticiparse a sus acciones comerciales, incurriendo en pérdida de ventas y recursos que impiden el crecimiento de la línea de productos, por tal razón una buena planificación de la venta permitirá solventar los problemas de la gestión comercial.

Objetivos

Objetivo general

Construir un modelo analítico basado en minería de datos, que permita predecir la venta de productos de la línea OTC (productos sin receta médica) de un Laboratorio Farmacéutico tomado como caso de estudio, para llegar a los niveles mínimos de unidad de venta establecidos por el área comercial.

Objetivos específicos.

OE1: Realizar el estudio del estado actual utilizando la técnica de la entrevista a los responsables de cada área involucrada y el método de la observación sistemática para delimitar las causas y variables del contexto del problema.

OE2: Identificar las técnicas predictivas más adecuadas en la gestión de información para el pronóstico de ventas.

OE3: Determinar el modelo analítico a través del análisis de patrones de comportamiento, depuración de datos y evaluación de las técnicas predictivas, para resolver el contexto del problema identificado.

OE4: Evaluar el modelo analítico a través del método descriptivo aplicando el análisis de los resultados, para definir la solución del problema con respecto al sobre stock y desabastecimiento de los productos.

OE5: Validar la solución implementada a través de indicadores de precisión como el MAPE (error porcentual absoluto medio)[4], para determinar si el modelo se aprueba o se rechaza.

Hipótesis

La construcción del modelo analítico basado en minería de datos, para la predicción de la venta del portafolio de productos OTC, permitirá reasignar la distribución de ventas con un error que no supere el 5% de la predicción.

Alcance

El alcance del presente proyecto es construir el modelo analítico para predecir la venta de los productos de la línea OTC del Laboratorio Farmacéutico. Esto contribuirá a la planificación de la demanda y a tener un correcto pronóstico de ventas considerando los factores que influyen en la proyección de ventas.

Para la construcción del modelo se busca realizar con herramientas analíticas que se ajusten a la predicción de ventas de los productos de la empresa, procurando realizar posteriormente la evolución y validación del modelo

El modelo a construir utilizará un histórico de los datos de ventas del laboratorio farmacéutico de 24 meses, para obtener la previsión de datos que permitan encontrar el pronóstico de ventas más acertado.

[4] MAPE (error porcentual absoluto medio): promedio de los errores porcentuales absolutos del pronóstico

Metodología

La metodología utilizada para la presente investigación es Design Science Research (DSR). Dicha metodología es aplicable ya que reúne las características para resolver el problema de investigación en base a la construcción y aplicación del artefacto diseñado, para lo cual se siguió las siguientes fases:

Fase 1. Conocimiento del problema: La solución debe resolver un problema importante en este caso para la organización. La situación actual del laboratorio farmacéutico no analiza un modelo predictivo de la venta, por lo cual tiene un efecto negativo en el área comercial de la línea OTC. En este sentido se realizó un análisis descriptivo a través de la entrevista a la Jefe de Planificación de la Producción y Gerente de Línea OTC del Laboratorio Farmacéutico; quienes aportarán con la información necesaria para el conocimiento del problema.

Fase 2. Sugerencia: sugiere el diseño de la construcción del artefacto. En este punto se estableció la posible solución a través de las series de tiempo y recolección de datos iniciales, utilizando la técnica de observación usando fuentes documentales y estadísticas.

Fase 3. Desarrollo: se define la implementación del artefacto, en este paso se aplicó la metodología de minería de datos KDD, que se revisará más adelante, mediante el cual se desarrolló el proceso de minería de datos, permitiendo encontrar el modelo que mejor se ajuste a la predicción de la venta de la línea de productos OTC.

Para la extracción de conocimiento en base de datos se utilizó la metodología KDD que propone 5 fases para su desarrollo:

- Selección de Datos: en este paso se define qué datos van a ser trabajados para su recolección, tipo de extracción y atributos de entrada y salida. En este punto es importante tener los datos objetivos para su selección.

- Procesamiento y Limpieza: se realiza el análisis de la calidad de los datos, aplicando operaciones básicas que implica la eliminación de ruido y campos de datos vacíos.
- Transformación: en esta etapa se busca la reducción de las dimensiones, este método simplifica las tablas de una base de datos. Esta reducción permite la eliminación de atributos que no son relevantes o a la vez redundantes.
- Minería de Datos: se realiza el análisis exploratorio, hipótesis y el modelo de selección, para esto se efectúa la técnica de series de tiempo con el fin de buscar los patrones de datos de las ventas para obtener la predicción.
- Evaluación e Interpretación: en esta etap a se interpretan los patrones minados y descubiertos

Fase 4. Evaluación: se evaluó los resultados de la solución, es decir si los modelos predictivos son óptimos. La técnica para evaluación es descriptiva y exploratoria, a través de la entrevista a la Gerente de la Línea OTC y también el análisis de los resultados del modelo a través de métricas de precisión.

Fase 5. Conclusión: se estableció la afirmación o negación de la hipótesis, en base a los patrones de comportamiento de los modelos predictivos.

Capítulo II

Marco Teórico

En esta sección se realizó la revisión de la literatura, para un análisis exhaustivo de la identificación de las herramientas y técnicas a utilizar. Además, se describieron las referencias que caracterizan los términos utilizados en la industria farmacéutica como mercados, productos y evaluación de ventas. Del mismo modo se detallan los conceptos que engloba la utilización de herramientas y modelos predictivos, que deben comprenderse como un sistema de términos básicos y ser entendidos para la investigación del proyecto.

Para orientar los fundamentos del presente proyecto es importante exponer los conceptos que se presentan dentro de la información. A continuación, se recopilan los términos fundamentales del desarrollo del proyecto, que permiten la explicación del estudio y sustento teórico.

Mercado OTC (Over the Counter)

Dentro de la industria farmacéutica, el Mercado OTC lo conforman todos los productos y/o medicamentos disponibles para la venta, que son categorizados por el ARCSA como “venta libre”, es decir que no requieren de receta médica para adquirirlos.

“Tradicionalmente, el mercado de OTC se divide en 4 grandes categorías, que son las especialidades farmacéuticas publicitarias, los productos de consumo, los productos llamados semi éticos --aquellos que todavía son de prescripción, pero que tiene poco sentido que lo sigan siendo, porque todo el mundo los pide sin receta-- y, finalmente, la parafarmacia, que comprende a su vez los productos de cuidado personal como la cosmética y los productos de higiene o belleza; la nutrición y alimentación, y los productos sanitarios auténticos, como apósitos, vendas, artículos para ostomía, incontinencia, etc.” (GRANDA, 2003)

Indicadores de Ventas

Dentro de una organización los indicadores de gestión son importantes para la medición de los objetivos establecidos dentro de un periodo. En el caso de los indicadores de ventas; aportan al control de los objetivos cuantitativos del área comercial, por lo que se puede tener una visión clara de las ventas y también comparar con datos históricos que contribuyen al mejoramiento del desempeño y productividad.

"La función comercial conecta la empresa con el mercado. Dentro de la función comercial, la investigación comercial es la primera etapa que realiza la empresa y la venta es la última. La investigación comercial identifica las necesidades existentes en el mercado, informa a la empresa y esta adapta su producción a dichas necesidades." (JIMÉNEZ & MATÍNEZ, 2016)

"Cuando hablamos de control, necesariamente tenemos que hablar de otra actividad de dirección inseparable, la primera de la definición de Fayol de la función de administración, la planificación. Efectivamente, el principio de la dirección es la planificación, el final el control. Quien tiene derecho a planificar, organizar, dirigir y coordinar, también lo tiene a comprobar si las previsiones se han realizado conforme a lo estipulado, es decir a controlar." (ARTAL, 2007)

Productos OTC

Los laboratorios farmacéuticos tienen varias líneas de negocio, como la línea de productos éticos (requieren de receta médica) y la línea de productos OTC (venta libre, sin receta médica). Esta última es clasificada como Over The Counter, ya que son productos que pueden ser de exhibición sobre mostrador.

"Cuya denominación viene dada por la inicial Over – The – Counter Drugs, concepto norteamericano que incluirá a todos los productos de venta de mostrador y que abarca fundamentalmente a parafarmacia, línea blanca, pañales, etc." (ORDUÑA, 2004)

Predicciones de la Demanda

Es imprescindible que los laboratorios farmacéuticos cuenten con una gestión adecuada de la predicción de la demanda de sus productos, ya que así pueden asegurar el abastecimiento para cubrir las necesidades del mercado.

"El concepto de previsión de la demanda se refiere a las actividades, estrategias y herramientas que se utilizan en la empresa para hacer estimaciones sobre la cantidad de ventas que puede tener la empresa en un futuro, y que se utilizan para poder ajustar lo máximo posible los procesos de decisión y planificación estratégica con el objetivo de ahorrar costes.

De este modo se reduce la incertidumbre en los procesos de gestión corporativa de toma de decisiones basándolos en datos reales, aunque de carácter estimativo. Esto hace que ningún método de los que se pueda emplear sea perfecto ni seguro al 100%, sin embargo, ayudan a que los cálculos sean más fiables, y a medida que vaya disponiendo de más datos, estos se pueden ir incorporando al proceso decisorio con el fin de prever o ajustar posibles desviaciones presentes o futuras." (MARTÍNEZ, 2014)

Algoritmo

Dentro del contexto informático, los algoritmos son representados por una serie de pasos ordenados con el fin de dar solución a un problema.

"Un algoritmo es un método para resolver un problema. En otras palabras, un algoritmo es una fórmula para la resolución de un problema.

La resolución de un problema se basa en identificar y definir los siguientes pasos:

1. Definición o análisis del problema

2. Diseño del algoritmo

3.Transformación del algoritmo en un programa

4. Ejecución y validación del problema" (SANCHEZ)

Minería de Datos

La exploración de grandes volúmenes de datos es aplicada a través de la minería de datos, esta técnica permite tener una visión de las tendencias de patrones de comportamiento, en este sentido aporta al pronóstico de ventas a partir de una construcción de modelos predictivos.

"Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos." (PÉREZ, 2008)

Técnicas de Aprendizaje Supervisado

El aprendizaje supervisado utiliza técnicas que permiten predecir los valores futuros, basados en datos históricos. Emplea problemas de clasificación y de regresión.

"Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos como otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se induce un modelo que relaciona la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado. En este grupo se encuentran, por una parte, algoritmos que resuelven problemas de clasificación debido a que trabajan con etiquetas discretas (árboles de decisión, tablas de decisión, inducción neuronal, etc.) y, por otra, algoritmos que se utilizan en la predicción de valores continuos como son la regresión o las series temporales." (TUYA, RAMOS, & DOLADO, 2017)

Técnicas de Aprendizaje No Supervisado

El aprendizaje no supervisado no emplea datos etiquetados, es decir, que no emplean datos históricos para su entrenamiento, pero sí utiliza modelos donde emplea problemas de clustering agrupando los datos para su análisis.

"Los algoritmos no supervisados o de descubrimiento del conocimiento realizan tareas descriptivas como el descubrimiento de patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio científico o de negocio de ellas." (TUYA, RAMOS, & DOLADO, 2017)

Series de Tiempo

Para el pronóstico de ventas son importantes los intervalos de tiempo, ya que al utilizar datos históricos es necesario emplear la estacionalidad, los ciclos y las tendencias dentro de períodos determinados.

"Una serie temporal consiste en una secuencia de valores de varias variables que evolucionan (van cambiando) en el tiempo. Se trata de predecir el comportamiento futuro del fenómeno o sistema dinámico que genera esos valores basándose en una colección de datos históricos. Por ejemplo, la predicción del consumo de energía eléctrica o la predicción del número de vacunas contra la gripe que se van a demandar en una región determinada. La mejor manera de resolver estos problemas es encontrando la ley subyacente que genera dichos procesos. Esta ley se puede obtener mediante métodos analíticos, como puede ser un conjunto de ecuaciones diferenciales. Sin embargo, la información que vamos a tener del proceso va a ser generalmente parcial o incompleta y; por lo tanto, la predicción no se puede hacer mediante un modelo analítico conocido. Se intentará descubrir alguna regularidad empírica fuerte en las observaciones de las series temporales. En muchos problemas del mundo real algunas regularidades, como la periodicidad, aparecen enmascaradas por ruidos, e incluso algunos procesos dinámicos se

describen por series de tiempo caóticas, donde los datos parecen aleatorios sin periodicidades aparente.” (PÉREZ, 2008)

Redes Neuronales

La técnica de redes neuronales permite la extensa variedad de aplicaciones de modelos y además la capacidad de aprendizaje que comprende, es decir que procesan información capaz de resolver funciones no lineales y además se basan en redes de entrada a partir de escenarios, es decir que la técnica procesa una amplia información ya que aprende de datos complejos.

> “Las ANN[5] al margen de "parecerse" al cerebro presentan una serie de características propias del cerebro. Por ejemplo, las ANN aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos y abstraen las características principales de una serie de datos.
>
> - Aprender: adquirir el conocimiento de una cosa por medio del estudio, ejercicio o experiencia. Las ANN pueden cambiar su comportamiento en función del entorno. Se les muestra un conjunto de entradas y ellas mismas se ajustan para producir unas salidas consistentes.
> - Generalizar: extender o ampliar una cosa. Las ANN generalizan automáticamente debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.
> - Abstraer: aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas ANN son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos.” (Olabe)

[5] Artificial Neural Networks (ANN)

Regresión

El análisis de regresión es un método estadístico utilizado para estimar valores futuros, y analiza la relación entre una variable dependiente y otra independiente. De acuerdo con Keat&Young, "El propósito básico del análisis de regresión es el de estimar la relación cuantitativa entre variables. El primer paso en este procedimiento estadístico es el de especificar el modelo de regresión (también llamado ecuación de regresión). El segundo consiste en obtener datos acerca de las variables especificadas en el modelo. El tercero es estimar el impacto cuantitativo que cada una de las variables independientes tiene en la variable dependiente. El cuarto paso es probar la significancia estadística de los resultados de regresión. Finalmente, los resultados del análisis de regresión resultan útiles como material de apoyo en la elaboración de políticas y en la toma de decisiones de negocios." (Young&Keat, 2004)

Holt Winters

"El filtro lineal conocido como método de Holt-Winters es una variante del alisado exponencial doble de Holt diseñado para realizar predicciones en series con tendencia aproximadamente lineal y con clara influencia de la componente estacional. Dependiendo del esquema de agregación elegido para la tendencia y la componente estacional, se habla del método de Holt-Winters multiplicativo o aditivo. En ambos casos, la componente irregular interviene aditivamente en el modelo" (Lorenzo, 2007)

Metodología KDD

Para la aplicación de la técnica de minería de datos se aplicará la metodología KDD[6] que permitirá encontrar el modelo válido para la predicción de la venta de la línea de productos OTC.

[6] KDD (Descubrimiento de Conocimiento en Bases de Datos): proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información

"Metodología KDD proveniente de sus siglas en inglés: extracción de conocimiento a partir de base de datos (Knowledge Discovery in Databases); KDD es un nuevo paradigma que se centra en la exploración informatizada de grandes cantidades de datos y en el descubrimiento de relevantes patrones interesantes dentro de ellos, es utilizado como un mecanismo que resume y analiza los contenidos de los conjuntos de conceptos que aportan a las bases de datos (Feldman y Dagan, 1995)." (Barriga & Castillo, 2018)

Herramientas y técnicas de predicción utilizadas en el mercado

Las herramientas de usabilidad buscan el manejo de grandes volúmenes de datos basados en mejorar las capacidades de las empresas. Varias de las herramientas son una combinación entre funcionalidades de software para analítica y suites. El software de analítica es capaz de preparar, modelar, analizar, y proyectar la visualización de los datos. La necesidad de la información de la demanda a través de la proyección de ventas focaliza varias técnicas predictivas las cuales están orientadas en una estimación de variables cuantitativas. Los modelos tienen como objetivo entrenar al modelo y así poder predecir los datos, en este caso las ventas del Laboratorio Farmacéutico. Dentro de las técnicas predictivas como: ARIMA[7], ALGORITMO; se tiene los análisis de árboles de decisión, regresión lineal, redes neuronales, series temporales.

La necesidad de buscar la precisión sobre el análisis predictivo se convierte en un ámbito importante, con el fin de obtener un modelo más fiable y eficiente para la predicción de ventas de los productos dentro de la industria, por tal motivo la importancia de analizar las herramientas y técnicas disponibles en el mercado para la ejecución del presente proyecto.

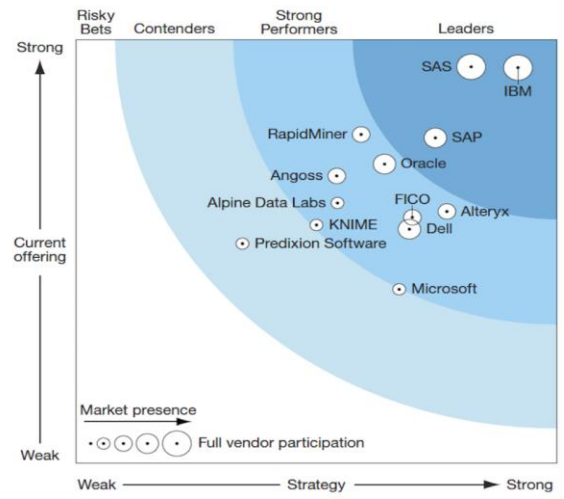
[7] ARIMA (AutoRegressive Integrated Moving Average)

La transformación de la analítica avanzada ha dado lugar al uso y desarrollo del análisis predictivo, “El análisis predictivo es un término paraguas para referirnos al conjunto de procesos que implican aplicar diferentes técnicas computacionales con el objetivo de realizar predicciones sobre el futuro basándose en datos pasados” (Iñaki, 2017), aportando de forma positiva al desempeño de las compañías en la toma de decisiones. En el mercado se puede obtener diversas soluciones para el análisis predictivo, dentro de las más importantes se encuentran: IBM, SAS, SAP, Oracle, Microsoft, Alteryx o KNIME. Cada solución puede ser empleada de acuerdo con las necesidades del negocio.

“El uso de APIs: las compañías de sectores como el financiero, el retail o el energético usan interfaces de programación de aplicaciones para construir modelos predictivos y extraer valor de los datos para: sacar conclusiones de los datos para tomar decisiones, predecir el comportamiento de los clientes para ajustar oferta y precios, también conocer su opinión sobre productos o servicios, conocer cómo se puede aumentar la productividad y el rendimiento, prevenir o detectar el fraude” (JUAN, 2015)

Figura 4

Clasificación anual de las mejores soluciones de análisis predictivo



Nota. la figura representa el análisis de Forrest Research 2015 donde se resume la clasificación de las mejores soluciones de análisis predictivo. Tomada de BBVA API_Market, 2015 (https://bbvaopen4u.com/es/actualidad/el-ranking-de-las-mejores-soluciones-de-analisis-predictivo-para-empresas)

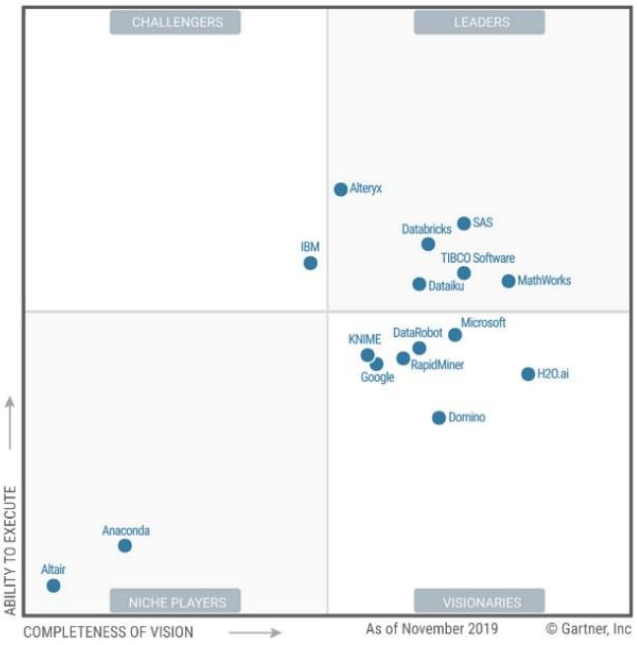
Por otro lado, en el cuadrante de Gartner 2020, se valoraron a 16 herramientas comparadas en 4 cuadrantes: líderes, retadores, visionarios y jugadores de nicho. Una característica importante es que, en la evaluación solo se consideraron las herramientas que cuentan con licencia comercial

En este sentido tenemos las siguientes clasificaciones:

- **Líderes:** Alteryx, Dataiku, Databricks, MathWorks, SAS, TIBCO
- **Retadores:** IBM
- **Visionarios:** DataRobot, Domino, Google, H2O.ai, KNIME, Microsoft, RapidMiner
- **Jugadores de Nicho:** Anaconda, Altair

Figura 5

Cuadrante mágico de Gartner 2020 para plataformas de ciencia de datos y aprendizaje automático



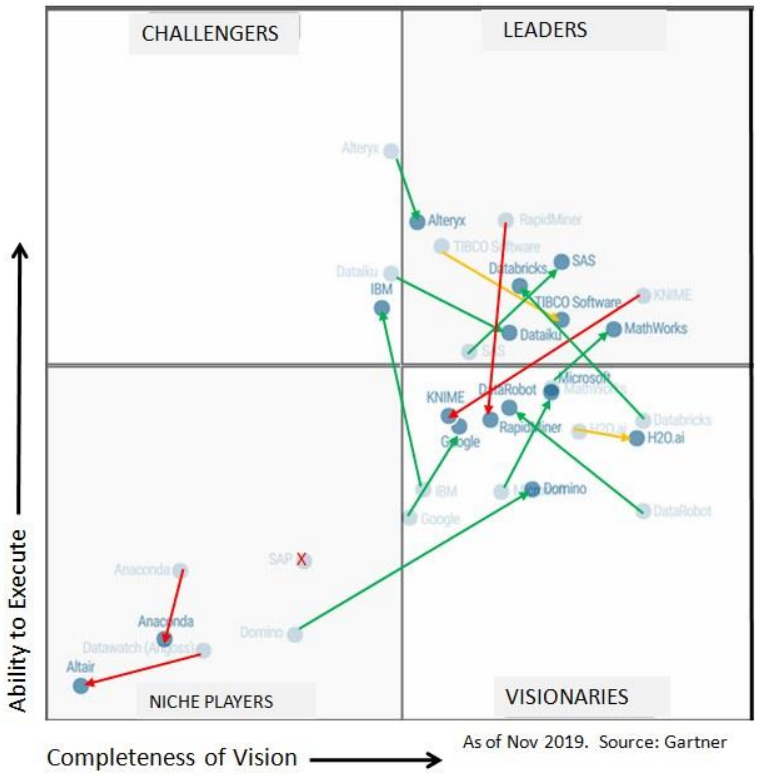
Nota. la figura muestra el cuadrante con las plataformas de ciencia de datos y aprendizaje automático. Adaptado de alteryx, 2020, (https://www.alteryx.com/es-419/third-party-content/gartner-2020-mq-data-science-machine-learning-thank-you)

En la figura 5 podemos revisar que las herramientas líderes son: Alteryx, Dataiku, Databricks, MathWorks, SAS, TIBCO. Las cuales podemos evaluar dentro de los factores críticos que nos permitirán posteriormente elegir la mejor herramienta para el desarrollo del modelo

Gartner no solo muestra un análisis actual de los proveedores dentro sus 4 cuadrantes. En la siguiente figura 6, se puede apreciar la comparación de los dos últimos años (2020 vs 2019), “Las flechas son de color verde si la posición firme mejora significativamente (más lejos del origen), rojo si la posición se debilita. Círculos verdes indican 2 nuevas empresas (Google y DataRobot), mientras que el proveedor de marcas rojas X cayó este año (Teradata).” (KDnuggets, 2019)

Figura 6

Cuadrante mágico de Gartner, cambios 2020 vs 2019

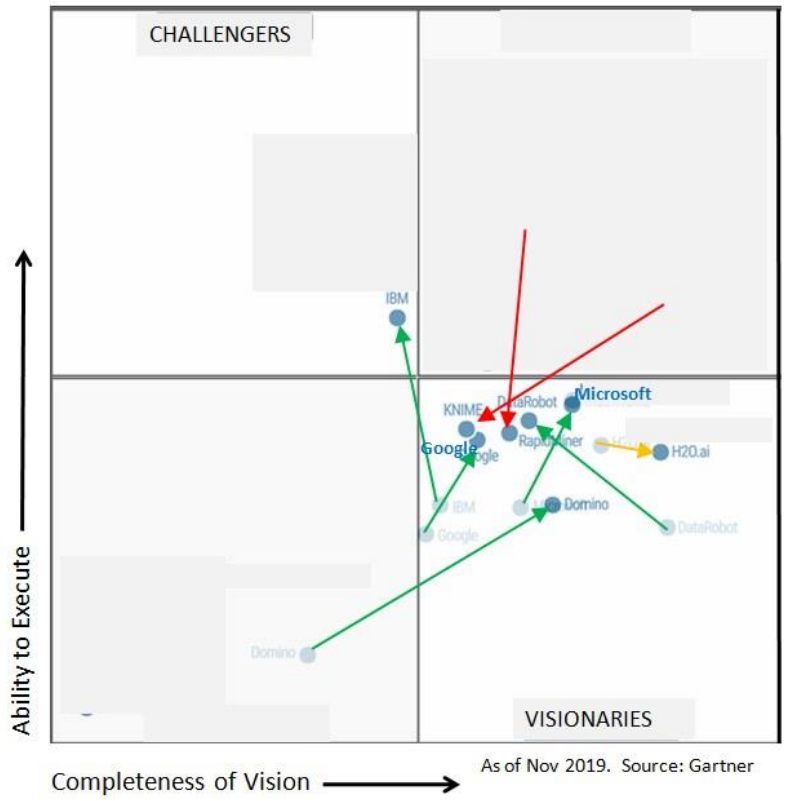


Nota. esta figura indica el cambio de posición en el cuadrante mágico de Gartner de las plataformas de ciencias de datos y aprendizaje automático. Tomada de KDnuggets, 2020, (https://www.kdnuggets.com/2020/02/gartner-mq-2020-data-science-machine-learning.html)

La comparación de los dos años (2019 y 2020), se puede citar a los líderes en cuanto a cambios en el 2020, donde han podido conseguir un equilibrio entre un uso más práctico y datos más sofisticados, como es el caso de RAPIDMINER, KANIME que se encuentran en el cuadrante como visionarios; por otro lado, tiene una capacidad muy amplia, ya que tiene un código abierto con una cobertura en cuanto a capacidades que alcanzan un 85%, haciéndolo competitivo frente a otras herramientas

Figura 7

Cuadrantes mágicos de Gartner Desafiadores y Visionarios



Nota. esta figura muestra los cuadrantes de retadores y visionarios donde se ubican las plataformas de ciencia de datos y aprendizaje automático. Tomada de KDnuggets, 2020, (https://www.kdnuggets.com/2020/02/gartner-mq-2020-data-science-machine-)learning.html

En el mercado se puede obtener una amplia gama de productos disponibles, los cuales ofrecen una profunda capacidad y enfoques variados para desarrollar, operacionalizar y administrar modelos. En este sentido es importante evaluar las necesidades específicas al determinar una herramienta analítica de acuerdo con la necesidad de la construcción del modelo y el pronóstico de ventas para los productos OTC.

Revisión de la Literatura

Para complementar la investigación con respecto a la identificación, evaluación y determinar las mejores herramientas y técnicas existentes en el mercado de la utilización de

minería de datos para la implementación de modelos analíticos, se realizó la revisión sistemática de la literatura.

Criterios de Inclusión

Los artículos para considerar deben incluir:

- Modelos predictivos de otras empresas de producción de medicamentos
- Ventajas que proporciona la técnica de minería de datos al control y planificación de inventarios
- Técnicas de predicción para forecast de las empresas farmacéuticas para planificar la venta
- Ventajas que se obtiene en el área comercial al planificar la venta
- La relación entre un modelo analítico para la predicción de la venta con el cumplimento de objetivos comerciales

Criterios de Exclusión

En contraste a lo anterior, se excluyeron los artículos:

- Modelos predictivos relacionados a empresas de servicios
- Información no relacionada a venta de producto terminado
- Técnicas de predicción de empresas de productos industriales

Resultados. En la realización de la búsqueda de estudios relacionados se obtuvieron los siguientes resultados:

Tabla 1

Resultados de búsqueda

CADENAS	Nº ESTUDIOS	ESTUDIOS GRUPO DE CONTROL
CD1	14	EC1,EC10
CD2	16	EC2, EC4
CD3	201	EC3,EC5,EC7
CD4	10	EC4,EC9
CD5	**53**	**EC5,EC7**
CD6	215	EC6,EC10
CD7	26	EC7, EC10
CD8	177	EC8EC10,EC9
CD9	213	EC9, EC3
CD10	559	EC10, EC4, EC3

En la tabla 1 se puede apreciar los resultados de las cadenas del grupo de control y la selección de la cadena la CD5. Dentro del conjunto de estudios se descartaron 33, a través de una depuración orientada a definir la relación y valor con el objetivo de búsqueda.

Tabla 2

Estudios Seleccionados

#	ARTICLE	SELECT
1	Sales prediction for a pharmaceutical distribution company: A data mining based approach. (Ribeiro, Seruca, & Durão, 2016)	X
2	Fundamental Analysis of Stock Trading Systems using Classification Techniques. (Cheng & Chen, 2007)	X
3	Model of the new sales planning optimization and sales force deployment ERP business intelligence module for direct sales of the products and services with temporal characteristics. (Velić, Padavić, & Lovrić, 2012)	X
4	Research and Development on Lean Collaborative Software System for Sales Activity Management. (Yuewei, Shuangyu, & Binchao, 2009)	X
5	Analysis of channel of sales promotion under consignment contract with revenue sharing. (Wang S. , 2010)	X
6	Implementation of mobile-based monitoring sales system in Semi Tani Shop. (Utomo, Sayyidati, & Rahmanto, 2018)	X

#	ARTICLE	SELECT
7	A Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry. (Marcoux & Selouani, 2009)	X
8	Frame discussion on a general sales management system. (Kangping & Yanhong, 2010)	X
9	Effectiveness of OLAP-Based Sales Analysis in Retail Enterprises. (Ju & Han, 2008)	X
10	Sales Resource Management Training: A Guide to Developing Effective Salespeople. (Wang, Lee, & Timothy, 2010)	X
11	Knowledge discovery, analysis and prediction in healthcare using data mining and analytics. (Raul, Patil, Raheja, & Sawant, 2017)	X
12	Newspaper Vendor Sales Prediction Using Artificial Neural Networks. (Fakharudin, Mohamad, & Johan, 2009)	X
13	A methodology of predicting automotive sales trends through data mining. (Shahid & Manarvi, 2009)	X
14	Towards great challenge in sales and operation planning. (Tudorie & Borangiu, 2011)	X
15	Improving Sales Process of an Automotive Company with Fuzzy Miner Techniques. (Koosawad, Saguansakdiyotin, Palangsantikul, Porouhan, & Premchaiswadi, 2018)	X
16	DDB and B/S Model-Based Special Steel Sales Management System. (Li & Li, 2009)	X
17	Sales data management system of chain enterprises based on NFC technology. (Yiqun, Zhenzhen, & Longjun, 2008)	X
18	Supply Chain Revenue-sharing Coordination with Sales Effort Effects. (Ye & Zeng, 2011)	X
19	Application of decision support system based on data warehouse in sales management. (Zhang, He, & Xu, 2012)	X
20	A retail-competition supply chain with promotion effort and sales learning curve. (Tsao, 2008)	X

Los estudios relacionados confirman que la técnica de la minería de datos tiene un amplio alcance bibliográfico en la industria farmacéutica y una estrecha relación en la utilización orientado al análisis de ventas predictiva.

A continuación, se concluye la identificación de las técnicas y herramientas más relevantes de los estudios revisados:

- En el estudio (Ribeiro, Seruca, & Durão, 2016), se confirma los beneficios de la aplicación de la técnica de series temporales, para mejorar la predicción de ventas de productos de una empresa de distribución farmacéutica.
- La aplicación del sistema de soporte basado en data warehouse es importante para la gestión de ventas. De acuerdo al estudio de (Zhang, He, & Xu, 2012), el esquema de diseño del sistema de soporte de decisiones de gestión de ventas y la importancia de la construcción de un modelo de datos, como condición para la toma de decisiones.
- En el estudio de (Shahid & Manarvi, 2009), se menciona que la recopilación de los datos históricos de la venta permite examinar las tendencias de las ventas a través de la técnica de minería de datos, por lo que tener la información histórica es un punto de partida para el análisis de las ventas.
- En el estudio de (Raul, Patil, Raheja, & Sawant, 2017), se hace referencia a que se puede lograr una mejor distribución de los medicamentos a través de la minería de datos y minería web, ya que pudieron analizar y agrupar medicamentos en función de la acción objetivo.

Se puede rescatar de los estudios revisados, que la técnica de minería de datos es un referente para la aplicación en el análisis de ventas y que el modelo de series temporales permite mejorar los pronósticos de ventas.

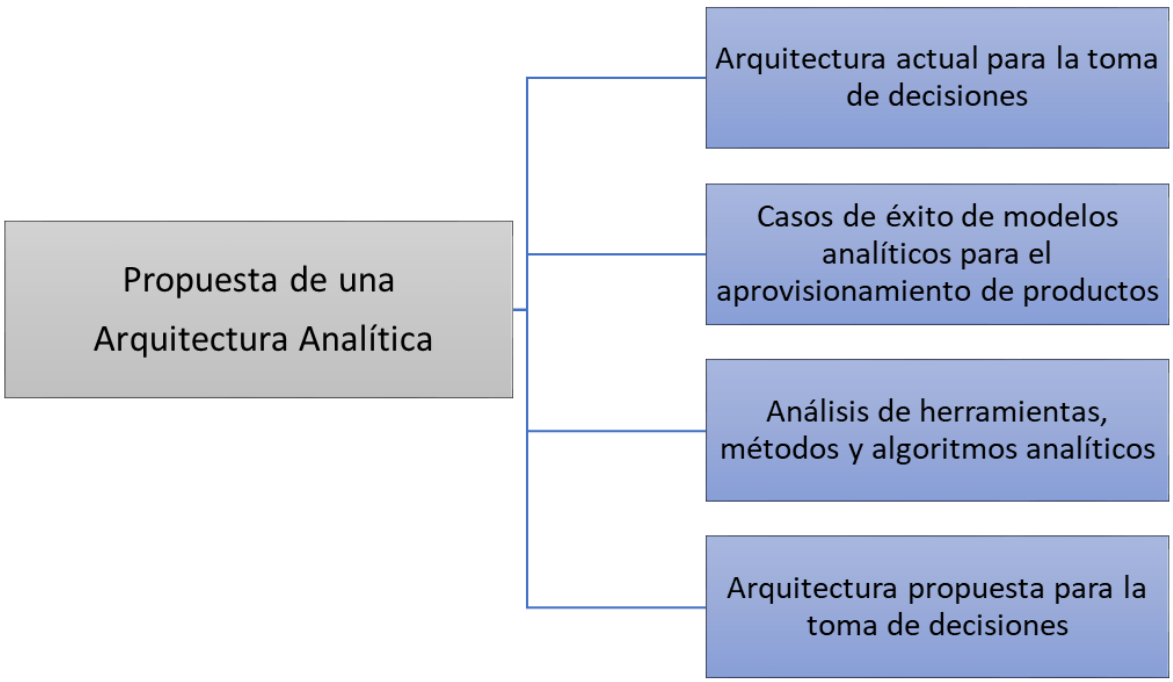
Capítulo III

Propuesta de una Arquitectura Analítica

En este apartado, se presenta la arquitectura para la toma de decisiones que tiene actualmente el Laboratorio Farmacéutico, también se analizaron tres casos de éxito de empresas en las que se ha implementado modelos predictivos para la realización de forecast y toma de decisiones. Además, se realizó un análisis de las herramientas, métodos y algoritmos disponibles en el mercado, para el desarrollo de la propuesta de la arquitectura analítica, con el fin de realizar la construcción del modelo analítico.

Figura 8

Esquema de Temas de Análisis para la Propuesta de una Arquitectura Analítica



Nota: la figura indica el esquema de los temas de análisis a seguir, para establecer la propuesta de una arquitectura analítica.

Arquitectura Actual Para la Toma de Decisiones

El Laboratorio Farmacéutico utiliza como sistema de gestión de base de datos relacional el sistema de gestión MySQL SERVER de Oracle. Este sistema permite la administración de los datos del Laboratorio Farmacéutico en términos de procesamiento transaccional, y operaciones empresariales, además tiene una alta disponibilidad y escalabilidad, que permite una capacidad más rápida de los datos. consecuentemente utilizan cubos OLAP para estructurar los datos realizando la consolidación de estos, posteriormente utilizan herramientas de visualización como por ejemplo Microsoft Excel, que permite tabular la información y crear gráficas, para que el usuario pueda acceder a la información de análisis de ventas. Una de las principales complicaciones es el manejo de grandes volúmenes de datos que se debe exportar a un archivo.

La gestión de base de datos en MySQL permite la creación de base de datos donde se almacena la información

En el siguiente flujo se muestra, que la gestión de bases de datos relacional se realiza en MySQL versión 5.2, para estructurar los datos en cubos OLAP a partir de tablas

relacionales. Estas relaciones ayudan a estructurar los datos para crear, cambiar y extraer los datos. Las consultas de la información se realizan a través de archivos xls, se tabulan los datos de forma manual para obtener la información histórica y realizar el forecast.

Figura 9

Arquitectura Analítica del Laboratorio Farmacéutico

Adaptado de Laboratorio Farmacéutico, 2019

En la figura 9 se puede apreciar que la arquitectura analítica para la toma de decisiones del Laboratorio Farmacéutico sigue tres procesos básicos. El primer proceso consiste en extraer los datos del repositorio MySQL para crear los cubos OLAP. La base de datos relacional de las ventas se exporta a la herramienta Microsoft Excel para analizar y visualizar la información a través de la manipulación de los datos de la venta histórica. Por consiguiente, el pronóstico de ventas se realiza en base a una metodología muy sencilla utilizando la técnica de media móvil, ocasionando una precisión baja, con un promedio del 15% de error del pronóstico, causando problemas en el abastecimiento de los productos de la línea OTC.

Casos de éxito de modelos analíticos para el aprovisionamiento de productos relacionados a la industria farmacéutica

Los modelos predictivos sin duda son de gran utilidad en varios sectores, en la industria farmacéutica se obtiene un gran volumen de datos históricos que permiten generar datos estadísticos y modelos predictivos, con una inmensa oportunidad de mejora para las empresas, ya

que optimizan los recursos y mejoran la toma de decisiones en base a información que permite el incremento del volumen del negocio.

Se ha considerado los siguientes casos, ya que plantean un desarrollo de implementación de modelos que han permitido mejorar el giro de negocio desde una perspectiva de abastecimiento de productos. A continuación, se presentan tres casos importantes de éxito de las compañías: La Favorita, Empresa "Farmaceútica", Operador logístico (Suiza).

Es importante también señalar que los casos de éxito presentados contribuyeron al presente estudio de investigación, ya que permitió realizar un análisis desde el conocimiento del problema de cada negocio hasta la solución de implementación, a través de los métodos y técnicas utilizadas, para la construcción del modelo predictivo. Además, el análisis de los casos sirvió para precisar la propuesta de la arquitectura analítica, para la toma de decisiones del Laboratorio Farmacéutico.

Caso 1: Corporación la Favorita. La Favorita es una empresa retail[8] que maneja una cadena de abastecimiento compleja para las tiendas a nivel nacional. Hace tres años presentó un problema similar a la del caso de estudio del presente trabajo, con respecto al análisis de los pronósticos de compras y ventas. A pesar de que no se trata de la misma industria, este caso está relacionado a la venta de los productos OTC (Over the Counter), ya que ambos casos dependen de la precisión de los pronósticos de ventas, para una adecuada provisión de los productos.

A continuación, el caso presentará el problema, y el desafío de desarrollar un modelo para el pronóstico de las ventas.

[8] Retail: es una palabra de origen inglés que se usa para referir el comercio al detalle, es decir, la venta de productos al consumidor final.

Corporación Favorita es una cadena comercial ecuatoriana, que cuenta con empresas filiales en áreas comercial, industrial e inmobiliaria, con presencia en el Ecuador y varios países de Latinoamérica. Nació en 1952, como una bodega que comercializaba artículos para el hogar, tanto nacionales como importados. En 1957 abrió Supermercados La Favorita, el primer autoservicio del país (Favorita, s.f.). Actualmente confían en métodos de pronóstico subjetivos con muy poca información para respaldarlos y muy poca automatización para ejecutar planes, para lo cual se enfocaron en construir un modelo que pronostique con mayor precisión las ventas de sus productos a través de un concurso presentado en la página web kaggle[9], el mismo que se lanzó hace dos años, para poder acceder al portal del concurso se puede dirigir al siguiente LINK (https://www.kaggle.com/c/favorita-grocery-sales-forecasting).

El concurso fue lanzado a nivel mundial, con el objetivo de mejorar los pronósticos de venta de Corporación La Favorita, por parte de cualquier persona o equipo de personas inscritas en dicho concurso.

Ingeniería de características

- Características básicas
- Características de la categoría: tienda, artículo, familia, clase, clúster ...
- Promoción
- Características estadísticas: utilizamos algunos métodos para estadificar algunos objetivos para diferentes claves en diferentes ventanas de tiempo
- Ventanas de tiempo
- Cálculos de variables acumuladas a: [1,3,5,7,14,30,60,140], serie de tiempo días

[9] Kaggle: es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático

- Agregadas por: tienda x artículo, artículo, tienda x clase
- Variable Objetivo: unit_sales
- Media, mediana, máxima, mínima, desviación estándar (Aplicadas a las variables utilizadas)
- Días desde la última aparición
- Diferencia de valor medio entre ventanas de tiempo adyacentes (solo para ventanas de tiempo igual)
- Características de bajo impacto
- Días festivos
- Otras claves como: grupo x elemento, tienda x familia

Caso 2: Análisis de series de tiempo univariadas y pronóstico de datos de ventas de productos farmacéuticos a pequeña escala. El objetivo de la investigación detrás del documento fue validar diferentes métodos y enfoques relacionados con la preparación, análisis y pronóstico de datos de series temporales de ventas, con el objetivo de facilitar la recomendación de estrategias de ventas y marketing basadas en los efectos de tendencia / estacionalidad y pronosticar las ventas de ocho grupos diferentes de productos farmacéuticos con diversas características, como estacionalidad, cantidad de residuos y variación de datos de ventas. Todos estos análisis y pronósticos se realizan a pequeña escala, para un único distribuidor, cadena de farmacias o incluso farmacias individuales. Cada uno de los métodos se complementa con dos enfoques de optimización y validación, relevantes para el pronóstico a corto plazo (denominado escenario de pronóstico continuo) y el pronóstico a largo plazo.

El conjunto de datos se construye a partir del conjunto de datos inicial que consta de 600.000 datos transaccionales recopilados en 6 años (período 2014-2019), que indican la fecha y hora de

venta, el nombre de la marca del medicamento farmacéutico y la cantidad vendida, exportados desde el sistema de Punto de Venta de una farmacia. El grupo seleccionado de fármacos del conjunto de datos (57 fármacos) se clasifica en las siguientes categorías del Sistema de Clasificación de Químicos Terapéuticos Anatómicos (ATC): - M01AB - Productos antiinflamatorios y antirreumáticos, no esteroides, derivados de ácido acético y sustancias relacionadas - M01AE - Anti -productos inflamatorios y antirreumáticos, no esteroides, derivados del ácido propiónico - N02BA - Otros analgésicos y antipiréticos, ácido salicílico y derivados - N02BE / B - Otros analgésicos y antipiréticos, pirazolonas y anilidas - N05B - Medicamentos psicolépticos, Fármacos ansiolíticos - N05C - Fármacos psicolépticos, hipnóticos y sedantes - R03 - Fármacos para enfermedades obstructivas de las vías respiratorias - R06 - Antihistamínicos para uso sistémico Los datos de ventas se vuelven a muestrear en los períodos por hora, diario, semanal y mensual. Los datos ya están preprocesados, donde el procesamiento incluyó detección y tratamiento de valores atípicos y la imputación de datos faltantes. El conjunto de datos se construye a partir del conjunto de datos inicial que consta de 600.000 datos transaccionales recopilados en 6 años (período 2014-2019), que indican la fecha y hora de venta, el nombre de la marca del medicamento farmacéutico y la cantidad vendida, exportados desde el sistema de Punto de Venta en el individuo farmacia. El grupo seleccionado de fármacos del conjunto de datos (57 fármacos) se clasifica en las siguientes categorías del Sistema de Clasificación de Químicos Terapéuticos Anatómicos (ATC): - M01AB - Productos antiinflamatorios y antirreumáticos, no esteroides, derivados de ácido acético y sustancias relacionadas - M01AE - Anti -productos inflamatorios y antirreumáticos, no esteroides, derivados del ácido propiónico - N02BA - Otros analgésicos y antipiréticos, ácido salicílico y derivados - N02BE / B - Otros analgésicos y antipiréticos, pirazolonas y anilidas - N05B - Medicamentos psicolépticos, Fármacos ansiolíticos - N05C - Fármacos psicolépticos, hipnóticos y

sedantes - R03 - Fármacos para enfermedades obstructivas de las vías respiratorias - R06 - Antihistamínicos para uso sistémico.

Los datos de ventas se vuelven a muestrear en los períodos por hora, día, semana y mes. Los datos ya están pre procesados, donde el procesamiento incluyó detección y tratamiento de valores atípicos y la imputación de datos faltantes.

Para el pronóstico continuo, el método ARIMA (Auto-ARIMA para series con carácter estacional) supera a Prophet, un modelo para pronosticar datos de series temporales basados en un modelo aditivo en el que las tendencias no lineales se ajustan a la estacionalidad pronóstico de ventas. Todos los métodos en todos los casos (con excepción del Prophet N02BE) superan los puntos de referencia

Para concluir, los análisis y pronósticos de series temporales han guiado conclusiones y recomendaciones potencialmente útiles a la farmacia. Se demostró que el análisis de estacionalidad diario, semanal y anual era útil para identificar los períodos en los que se podían implementar campañas especiales de ventas y marketing, a excepción de las categorías de medicamentos N05B y N05C que no exhibían regularidades significativas. Los pronósticos han demostrado ser mejores que los métodos Naive[10] en intervalos aceptables para la planificación a largo plazo. Es muy probable que los pronósticos se puedan mejorar significativamente al ampliar el alcance del problema a pronósticos de series de tiempo multivariadas e incluir variables explicativas, tales como:

[10] Naive Bayes: Se basan en una técnica de clasificación estadística llamada “teorema de Bayes”, para construir clasificadores

- Datos del tiempo. Las ventas de medicamentos antirreumáticos en las categorías M01AB y M01AE podrían verse afectadas por los cambios en la presión atmosférica. La disminución repentina en todas las categorías podría explicarse por condiciones climáticas extremas, como fuertes lluvias, tormentas eléctricas y tormentas de nieve.
- Precio de las drogas. Los picos de ventas pueden explicarse por los descuentos, aplicados a corto plazo. La introducción de esta función puede facilitar el análisis de pronóstico de rendimiento de ventas durante las campañas de marketing que implican reducciones de precios.
- Fechas del pago de la pensión. Los picos de ventas son visibles en las fechas de pago de las pensiones estatales.
- Feriados nacionales, ya que se espera que los días no laborables con patrones estacionales similares a los domingos interrumpan las ventas diarias.

El trabajo futuro en el pronóstico de series de tiempo univariadas incluye aumentar el número de datos, explorar otras métricas de precisión diferentes, la optimización de hiper parámetros para modelos LSTM y probar otras arquitecturas, como CNN LSTM y ConvLSTM. Sin embargo, se espera que las mejoras clave en el pronóstico de ventas reduzcan la incertidumbre de los modelos al expandirse al problema de pronóstico de series temporales multivariadas, como se explicó anteriormente. (Kaggle, s.f.)

Caso 3. Operador Logístico (Suiza). El siguiente caso fue desarrollado para un operador logístico de productos farmacéuticos, la contribución del caso son las variables que utilizan para poder focalizar la conservación de antibióticos en el transporte de cadena de frío.

Descripción del cliente. Esta empresa es una de las empresas logísticas líderes en el sector farmacéutico. Esta empresa posee una instalación de 20.000 m2 de 4 plantas, de 30 metros de altura. El suministro de los fármacos lo realiza desde fábrica o desde el almacén de los proveedores.

Descripción del proyecto. Este operador logístico tiene más de 6.000 clientes finales, con un almacén automático que trata aproximadamente 6.000 palets y 66.000 cajas. La cadencia de entrada es de 100 palets al día, y la cadencia de salida es de 2.500 líneas de pedido al día, con un conteo de traspasos internos de 1.000 líneas al día.

Cadena de frío: El proyecto se focaliza en la conservación y transporte de vacunas antibióticas que deben conservarse en el rango de -20ºC y + 8ºC.

Todo lo que salga de ese rango, merma la calidad del medicamento o vacuna. Se disponen de cubetas que garantizan el rango de temperatura ideal durante 24 horas, y son utilizadas para el transporte de fármacos.

Problemática para resolver. Se desconoce la fiabilidad de la cadena de frío. Existen reclamaciones por roturas en la misma, por lo que se necesita conocer el origen de estas roturas y el posible impacto que genera en los clientes, con el fin de poder reclamar el producto en el menor tiempo posible, es decir, garantizar la trazabilidad.

Desarrollo de proyecto. Dada la complejidad del análisis y la cantidad de variables que intervienen (+100) en la problemática, además del cuadro de mando implementado con los

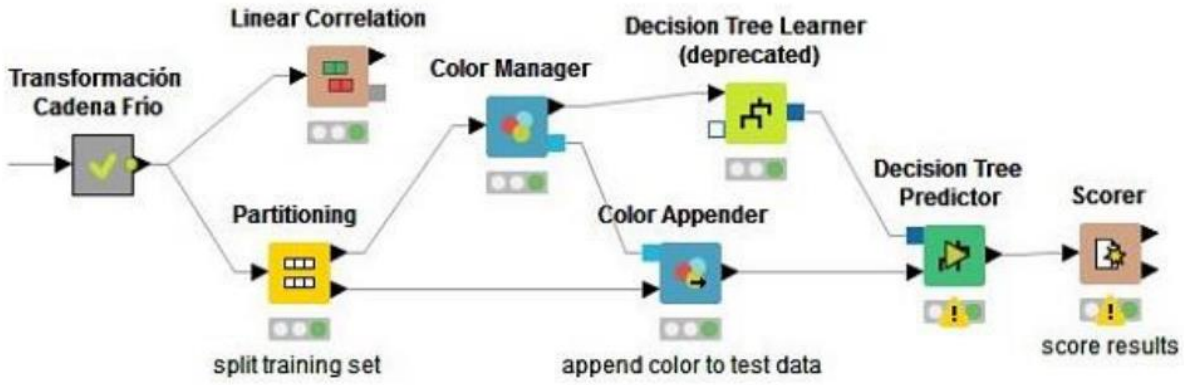
indicadores necesarios (tecnologías Business Intelligence), deciden realizar un estudio con técnicas de minería de datos para reducir las variables que intervienen.

Definen la pérdida de cadena de frío como “todo aquel movimiento que saliendo del almacén de frío supera las 24 horas antes de la expedición”, es decir, antes de abandonar el almacén interno.

Uso de árboles de decisión. El objetivo era definir las variables que intervienen en el proceso de la cadena de frío, por lo que analizaron diferentes muestras de las que se disponía, cada una de ellas definida por una serie de características como el cliente, tipo de almacenamiento, mes o día de la semana, y categorizaron a las mismas en función de si habían participado en la cadena de frío directamente o no.

Figura 10

Árbol de decisión en el proceso de la cadena de frío



Nota. la figura representa el flujo del árbol de decisión de la cadena de frío de la empresa logística. Tomado de LIS SOLUTIONS, 2016

Dividieron el conjunto de datos en dos subconjuntos: unos de entrenamiento y otros de test, para poner a prueba el modelo, generalizando a la hora de clasificar datos nuevos.

Figura 11

Árbol de decisión en el proceso de la cadena de frío

Nota. La figura muestra el árbol de decisión, donde intervienen los días donde se rompe la cadena de frío en el proceso logístico. Tomado de LIS SOLUTIONS, 2016

Tal y como se puede observar en el árbol de decisión anterior, la mayoría de los casos en los que sucedía la rotura de la cadena de frío (141/143) era en viernes o sábado, por lo que la variable "día" participaba directamente. Por otro lado, según el mismo razonamiento se relacionaba de la misma forma las variables de Prioridad y Operario.

Esta información la llevaron al cuadro de mando de exploración donde les permitió tener una total visibilidad de la trazabilidad de los productos que rompían la cadena de frío y la solución de manipular de forma distinta los productos de este tipo.

Mejoras. Gracias a la aplicación de algoritmos de minería de datos y herramientas de visualización (Business Intelligence) consiguieron mantener un proceso crítico controlado, por lo que fueron evitados los errores que podrían haber sido fatales para esta empresa. Con este proyecto se hizo visible la problemática, a través de una monitorización en tiempo real de la

situación de la cadena de frío y aplicando alarmas de riesgo a nivel de línea de pedido, por lo que la prevención fue absoluta (LIS SOLUTIONS, 2016)

Recomendaciones y estrategias usadas en los casos de éxito

La siguiente tabla 3, resume el método de análisis utilizado y las características relevantes empleadas como las variables consideradas para la construcción del modelo:

Tabla 3

Resumen casos de éxito de modelos analíticos

Caso de Estudio	Técnicas Utilizadas	Solución Adoptada	Resultados Alcanzados	Descripción General
Aplicación de Forecast: Empresa La Favorita	LGBM (Gradient Boosted Machines): basado en modelos de árboles de decisión, utilizada en la herramienta Python NN (K-Nearest-Neighbor):clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos)	model_1: 0.506 / 0.511, 16 modelos LGBM entrenados para el código fuente de cada día model_2: 0.507 / 0.513, 16 NN modelos entrenados para el código fuente de cada día model_3: 0.512 / 0.515, 1 modelo LGBM durante 16 días con las mismas características que model_1 model_4: 0.517 / 0.519; 1 NN modelo	De los 4.100 artículos, 3.114 son no perecederos y 986 son perecederos. Las páginas de la competencia establecen que los artículos perecederos pesan más que los no perecederos. Una vida útil más corta significa un margen de error menor al pronosticar las ventas.	1. Características básicas utilizadas en el modelo: • Tienda, artículo, familia, clase, clúster • Promociones 2. Características estadísticas generadas • Ventanas de tiempo • Variables objetivo: unit_sales • Estadísticas: media, mediana, máximo, mínimo, desviación estándar (Aplicadas a las características básicas) 3. Características irrelevantes Días festivos

Caso de Estudio	Técnicas Utilizadas	Solución Adoptada	Resultados Alcanzados	1. Descripción General
Análisis de Series de Tiempo: Productos Farmacéuticos de un Punto de Venta	Prophet: El modelo facilita la personalización y los pronósticos confiables con configuraciones predeterminadas. ARIMA: utilizada para pronosticar series de tiempo estacionarias Redes Neuronales: arquitecturas de aprendizaje profundo que se caracterizan por el uso de unidades LSTM en capas ocultas.	Utilizaron el conjunto de bibliotecas de Python, pyramid-arima, para poder utilizar el método Auto-ARIMA El análisis de series de tiempo incluyó análisis de estacionalidad, auto correlación, regularidad y distribución de datos.	El análisis de estacionalidad diario, semanal y anual demostró ser útil para identificar los períodos en los que se podrían implementar campañas especiales de ventas y marketing, excepto para las categorías de medicamentos N05B y N05C que no mostraron regularidades significativas.	2. El conjunto de datos inicial consistió en 600.000 datos transaccionales recopilados en 6 años (período 2014-2019) 3. Características: fecha, hora de venta, marca, cantidad venta 4. Muestra de 58 drogas clasificada en 8 categorías (ATC) Sistema de Clasificación Química Terapéutica Anatómica
Operador Logístico, cadena de abastecimiento	Utilizaron técnicas de minería de datos y método de árboles de decisión para reducir las variables que intervienen en el proceso de cadena de frío.	Definir variables que intervienen y realizar la trazabilidad de productos que rompen el proceso de cadena de frío	Seguimiento en tiempo real del proceso de la cadena de frío, aplicando alarmas de riesgo a nivel de línea de pedido	1. Elección de 100 variables 2. Análisis de un estudio con técnica de minería de datos para reducir las variables que intervienen 3. Variables relacionadas: Días, Prioridad y Operario 4. Lis Solutions, es la consultora que trabajó para las mejoras del proyecto, ésta utiliza herramientas como: tableau, SAP, Qlik

Resultados

- Los tres casos reúnen características que aportan al desarrollo del modelo analítico a

realizar, por un lado, podemos recopilar que los “efectos de tendencia / estacionalidad y

pronosticar las ventas de ocho grupos diferentes de productos farmacéuticos con diversas características, como estacionalidad, cantidad de residuos y variación de datos de ventas" funcionan para la efectividad de los métodos de pronósticos ARIMA y redes neuronales.

- Podemos considerar que el procesamiento incluye detección y tratamiento de valores atípicos y la imputación de datos faltantes, además los análisis y pronósticos de series temporales han guiado conclusiones y recomendaciones potencialmente útiles a la farmacia o punto de venta.
- Se demostró que el análisis de estacionalidad diario, semanal y anual son útiles para identificar los períodos en los que se podían implementar campañas especiales de ventas y marketing, estas actividades son imprescindibles para poder garantizar la rotación adecuada de stock.
- Se puede rescatar que el proceso de minería de datos permite la aplicación de algoritmos. En el caso del operador logístico podemos constatar que la variable "días de la semana" es importante al momento de definir el análisis predictivo. (LIS SOLUTIONS, 2016)
- El modelo ARIMA realiza pronósticos utilizando series temporales, por lo que se consideraría como una de las mejores opciones para el pronóstico de ventas, con el fin de encontrar patrones para analizar el pronóstico. (Zdravkivic, s.f.)
- El método Light GBM, ayuda a obtener un mejor ajuste del modelo, por lo que se obtiene mayor precisión en el pronóstico, dejando de lado una explicación clara de cómo internamente el algoritmo selecciona las variables más importantes. (Kaggle, 2018)

Herramientas Analíticas

Para la selección de herramientas y algoritmos a utilizar se analizará las diversas alternativas en base a los análisis del cuadrante de Gartner, Forrester Research y además a la encuesta que realiza KDnuggets, un sitio web de ciencia de datos que ejecuta encuestas a expertos en análisis de datos para determinar las herramientas y plataformas más populares utilizadas en el mercado.

Selección de la herramienta analítica

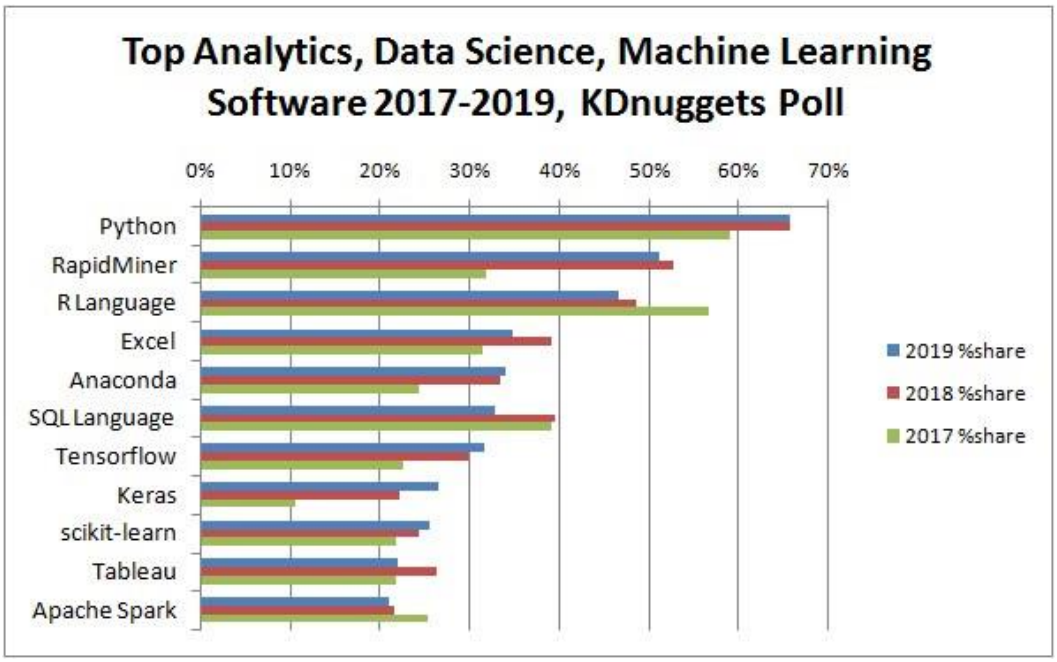
De acuerdo con Gregory Piatetsky-Shapiro quien es el presidente de KDnuggets, (es un conocido experto en Business Analytics, Data Mining y Data Science, 3 y uno de los principales influyentes en el campo; la encuesta 2019 realizada de software de KDnuggets[11]) la participación en la encuesta de herramientas analíticas tuvo más de 1.800 votantes y el porcentaje promedio de herramientas por encuestado fue 6.7%.

De acuerdo con la figura 11, Phyton se mantiene desde el año pasado como líder dentro de las 11 principales herramientas de aprendizaje automático y ciencia de datos, detrás le sigue RapidMiner que mantiene su participación en alrededor del 51%, lo que fue un reflejo de una gran base de usuarios y una exitosa campaña para motivar a sus usuarios. Y en tercer lugar se encuentra Lenguaje R, que ha disminuido dos años seguidos, varios de los usuarios resaltaron que se debe incluir RStudio dentro de la encuesta.

[11] (KDnuggets) es un sitio líder en inteligencia artificial, análisis, big data, minería de datos, ciencia de datos y aprendizaje automático y está editado por Gregory Piatetsky-Shapiro y Matthew Mayo https://www.kdnuggets.com/about/index.html

Figura 12

Herramientas principales en 2019 y su participación en las encuestas de 2017 y 2018



Nota. la figura muestra el ranking 2017-1019 de las principales herramientas utilizadas por los científicos de datos. Tomado de KDnuggets 2020, (https://www.kdnuggets.com)

Para la selección de herramientas, se han evaluado las diferentes opciones y sus características, en este sentido se determinó RAPIDMINER como herramienta de análisis para la construcción del modelo analítico. La herramienta tiene un buen posicionamiento en el mercado y además es de fácil acceso y manejo.

Para la selección de la herramienta analítica se determinó a través del análisis de las opciones que ofrece el mercado y a través del análisis del cuadro mágico de Gartner a Rapidminer, “Los científicos expertos en datos y otros profesionales que trabajan en roles de ciencia de datos requieren capacidades para obtener datos, construir modelos y poner en práctica conocimientos de aprendizaje automático. El crecimiento significativo de los proveedores, el desarrollo de

productos y una mirada de visiones competitivas reflejan un mercado saludable que está madurando rápidamente.” (Peter Krensky, 2020)

Rapidminer está disponible con una edición gratuita que permite y facilita el análisis y trabajo de los modelos analíticos.

Tabla 4

Tabla comparativa para la selección de la herramienta

				PONDERACIÓN		
HERRAMIENTA	**COSTO**	**FORRESTER RESEARCH**	**GARTNER**	**FORRESTER RESEARCH**	**GARTNER**	**PROMEDIO**
EXCEL	Código abierto	NA	NA	0	0	0
R LANGUAGE	Código abierto	NA	NA	0	0	0
RAPIDMINER	**Código abierto**	**Fuerte desempeño**	**Visionario**	**3**	**2**	**2,5**
PYTHON	Código abierto	NA	NA	0	0	0
SAS	Compra de licencia	Líder	Líder	4	4	4
IBM	Compra de licencia	Líder	Retadores	4	3	3,5
ALTERYX	Compra de licencia	Fuerte desempeño	Líder	3	4	3,5
KNIME	Código abierto	Fuerte desempeño	Visionario	3	2	2,5
ANGOSS	Compra de licencia	Fuerte desempeño	Jugadores de nicho	3	1	2
MICROSOFT	Compra de licencia	Contenedores	Visionario	2	2	2
ORACLE	Compra de licencia	Fuerte desempeño	NA	3	0	1,5

FICO	Compra de licencia	Fuerte desempeño	NA	3	0	1,5
DELL	Compra de licencia	Fuerte desempeño	NA	3	0	1,5
ALPINE DATA LABS	Compra de licencia	Fuerte desempeño	NA	3	0	1,5
PREDIXION SOFTWARE	Compra de licencia	Contenedores	NA	2	0	1

Realizando el análisis de la Tabla N°4, se han mencionado las principales herramientas que califican las nuevas tendencias para análisis y minería de datos; Gartner, Forrester Research y KDnuggets permiten analizar y determinar a los proveedores tecnológicos tanto en su desempeño como en su visión, comparándolas entre sí con sus diferentes características.

Dentro de la ponderación realizada en la Tabla N°4, se estableció un peso de calificación en base a los cuadrantes de cada empresa (Gartner y Forrester Research), para posteriormente obtener un peso promedio que nos permita identificar la herramienta con mejor puntaje.

Tabla 5

Top Analytics / Data Science / ML Software en 2019 KDnuggets Poll

Herramienta	2019 % share	2018 % share	2017 % share
Python	**65.8%**	65.6%	59.0%
RapidMiner	**51.2%**	52.7%	31.9%
R Language	**46.6%**	48.5%	56.6%
Excel	**34.8%**	39.1%	31.5%
Anaconda	**33.9%**	33.4%	24.3%
SQL Language	**32.8%**	39.6%	39.2%
Tensorflow	**31.7%**	29.9%	22.7%
Keras	**26.6%**	22.2%	10.7%
scikit-learn	**25.5%**	24.4%	21.9%
Tableau	**22.1%**	26.4%	21.8%
Apache Spark	**21.0%**	21.5%	25.5%

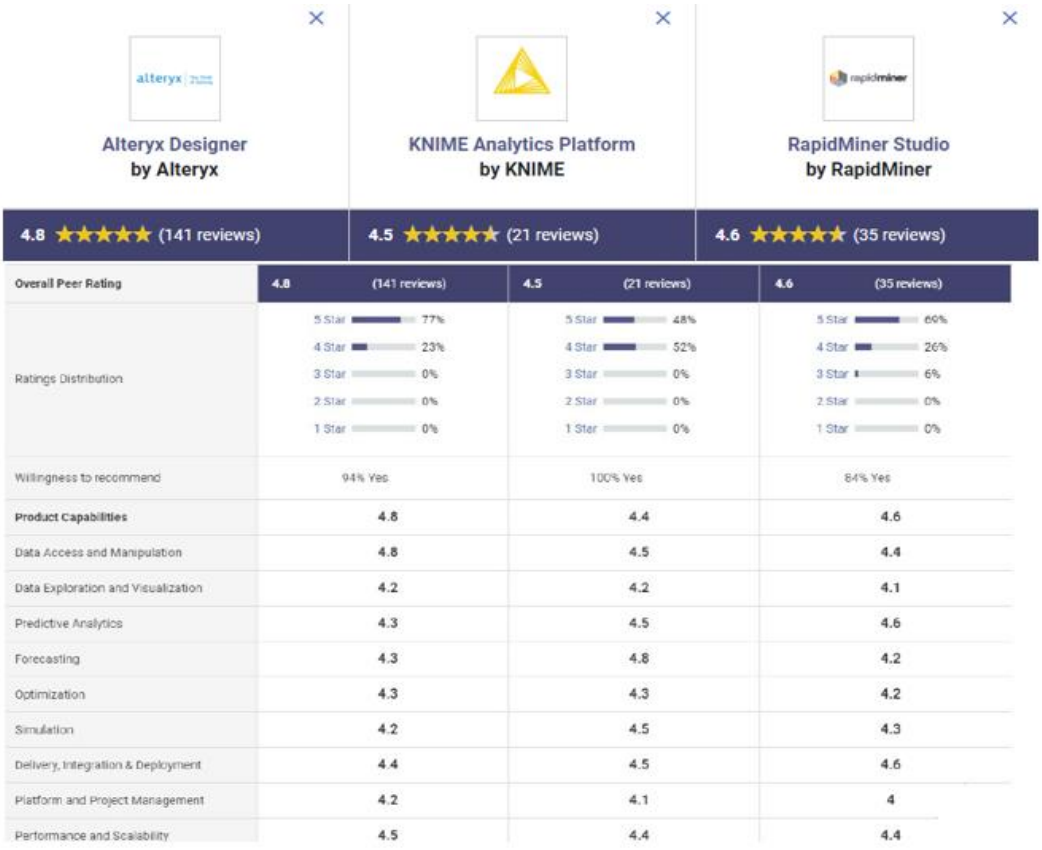
El análisis realizado con las encuestas de KDnuggets, se basó en las herramientas con mejor porcentaje/top de “Encuesta sobre el software KDnuggets Analytics / Data Science 2019”.

Podemos observar que la Herramienta Rapidminer es evaluada por las dos consultoras y considerada en las encuestas del sitio web KDnuggets, a pesar de que para Gartner, la herramienta Rapidminer pasó de ser Líder el año pasado a ser Visionario en el reporte del 2020, por un crecimiento reducido a comparación de las demás herramientas, a pesar de ser Visionario mantiene un gran enfoque en innovación, en cuanto a agilidad y escalabilidad. Por otro lado, KDnuggets a través de sus encuestas realizadas mantiene a Rapidminer con el 51% de participación de una gran base de usuarios.

Adicionalmente se investigó a través de la página web Gartner Peer Insights (es una fuente de información para compradores de tecnología), la comparación de las herramientas open source bajo ciertas comparaciones y calificaciones que realizan los propios usuarios en base a la experiencia:

Figura 13

Comparación Alteryx Designer, KNIME Analytics Platform, RapidMiner Studio



	Alteryx Designer by Alteryx	KNIME Analytics Platform by KNIME	RapidMiner Studio by RapidMiner
	4.8 ★★★★★ (141 reviews)	4.5 ★★★★★ (21 reviews)	4.6 ★★★★★ (35 reviews)
Overall Peer Rating	4.8 (141 reviews)	4.5 (21 reviews)	4.6 (35 reviews)
Ratings Distribution	5 Star 77% 4 Star 23% 3 Star 0% 2 Star 0% 1 Star 0%	5 Star 48% 4 Star 52% 3 Star 0% 2 Star 0% 1 Star 0%	5 Star 69% 4 Star 26% 3 Star 6% 2 Star 0% 1 Star 0%
Willingness to recommend	94% Yes	100% Yes	84% Yes
Product Capabilities	4.8	4.4	4.6
Data Access and Manipulation	4.8	4.5	4.4
Data Exploration and Visualization	4.2	4.2	4.1
Predictive Analytics	4.3	4.5	4.6
Forecasting	4.3	4.8	4.2
Optimization	4.3	4.3	4.2
Simulation	4.2	4.5	4.3
Delivery, Integration & Deployment	4.4	4.5	4.6
Platform and Project Management	4.2	4.1	4
Performance and Scalability	4.5	4.4	4.4

Nota. La figura muestra la comparación de las características y puntuación entre Alteryx Designer, KNIME Analytics Platform, RapidMiner Studio. Tomado de Gartner peer insights 2020, (https://www.gartner.com/en/products/peer-insights)

La figura 13, muestra que las herramientas Alteryx Designer y Rapidminer Studio tienen una mejor calificación con respecto a las opiniones de los usuarios, 77% y 69% respectivamente. Las capacidades de cada producto muestran que tienen una calificación similar, sin embargo, Rapidminer Studio tiene una mejor puntuación con respecto a la capacidad versátil de analítica predictiva, esta característica nos permitirá una mejor y óptima construcción del modelo analítico para la predicción de ventas del Laboratorio Farmacéutico.

Por otro lado, en el espacio G2 Crowd (mercado tecnológico más grande del mundo donde las empresas pueden descubrir, revisar y administrar la tecnología que necesitan para alcanzar su potencial), se compara a Rapidminer con Knime, donde la última plataforma mencionada tiene una calificación de 4.3/5 a comparación de Rapidminer que tiene un puntaje de 4.5/5. Estos resultados se calculan a través de datos en tiempo real de los usuarios.

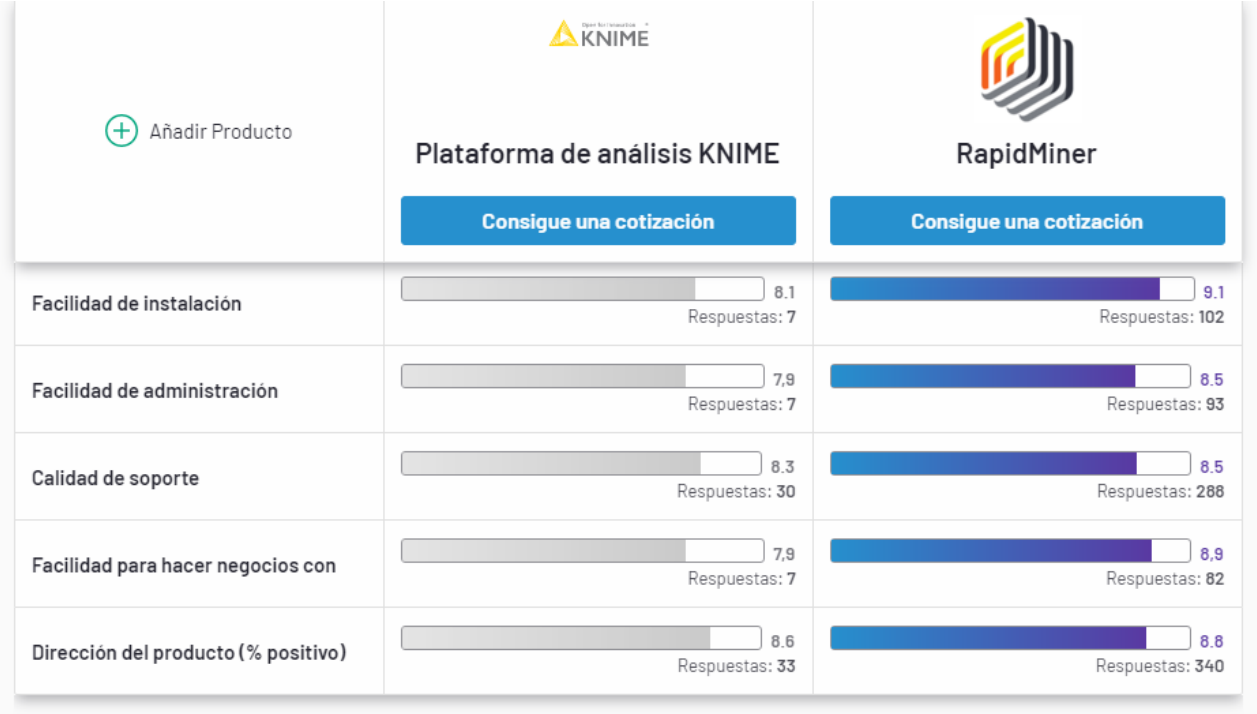
Figura 14

Comparación KNIME Analytics Platform vs RapidMiner

Nota. La figura indica la calificación de usuarios entre las herramientas de analítica KNIME y RapidMiner. Tomado de G2 Crowd 2020, (https://www.g2.com/compare/knime-analytics-platform-vs-rapidminer-studio).

Figura 15

Comparación y característica KNIME Analytics Platform vs RapidMiner



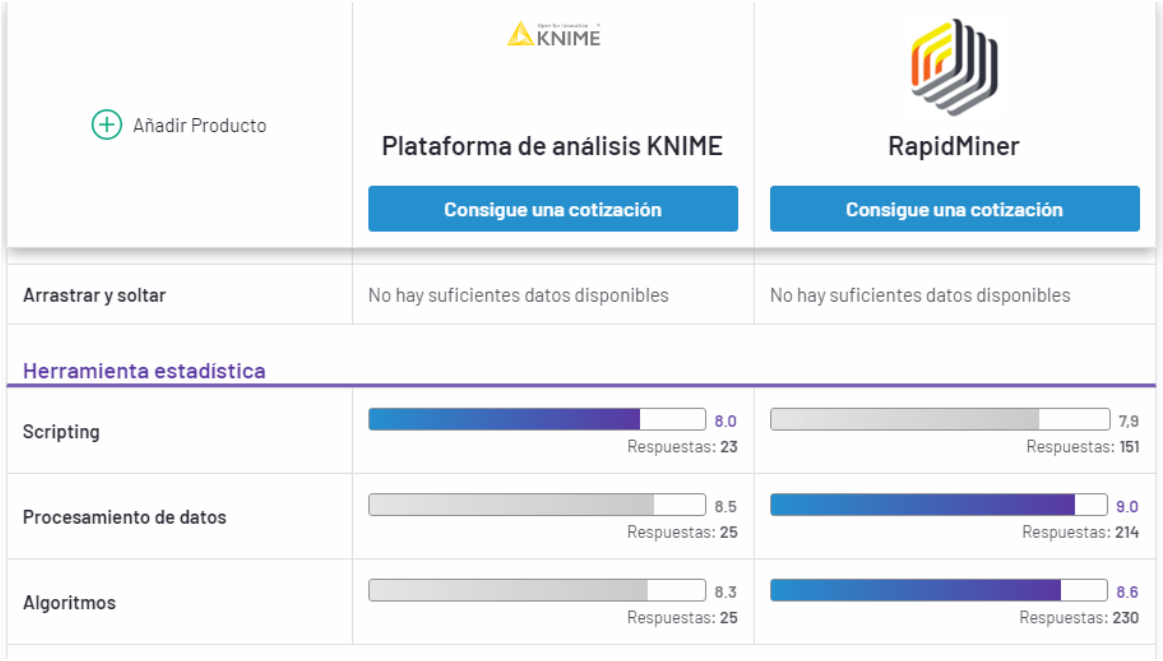
Tomado de G2 Crowd, 2020, (https://www.g2.com/compare/knime-analytics-platform-vs-rapidminer-studio).

En la figura 15, se detalla las características calificadas por los usuarios sobre ambas herramientas; por un lado, podemos observar que Knime tiene un promedio de 8,16 con respecto a 5 características calificadas por los usuarios, mientras que Rapidminer tiene un promedio de 8,76; entres las características se destaca la facilidad de instalación y la facilidad de uso.

En la siguiente figura N°16, Rapidminer tiene una calificación con respecto al procesamiento de datos de 9,0, Scripting 7,9 y de algoritmos de 8,6; frente a un número de usuarios que oscila entre 151 y 230.

Figura 16

Comparación herramienta estadística KNIME Analytics Platform vs RapidMiner



Nota. La figura muestra la puntuación de los usuarios sobre las herramientas, en las características de Scripting, procesamiento de datos y algoritmos. Tomado de G2 Crowd, 2020, (https://www.g2.com/compare/knime-analytics-platform-vs-rapidminer-studio).

De acuerdo con el análisis realizado se seleccionó la herramienta Rapidminer Studio, por su importante participación en el mercado, y opiniones en base a la experiencia de los usuarios. La compatibilidad y versatilidad permitirá los análisis técnicos de los datos de ventas para la construcción del modelo.

Las calificaciones y puntuaciones destacadas permiten visualizar que la herramienta Rapidminer tiene varias características positivas por la facilidad de uso que les motiva a los usuarios a aplicar un amplio conjunto de datos y la aplicación de una variedad de técnicas de minería de datos y aprendizaje automático que permite construir modelos predictivos.

La funcionalidad de la herramienta Rapidminer permitirá la extracción de datos, aprendizaje automático, construcción del modelo predictivo y la exploración y visualización de datos de las ventas del Laboratorio Farmacéutico.

Justificación de selección de herramienta

De acuerdo con la clasificación de soluciones de análisis predictivo de Forrester Research, se analizaron las firmas más representativas dentro del mercado que se pueden implementar para realizar el análisis predictivo de la venta del laboratorio farmacéutico:

SAS: crea de manera rápida y fácil mejores modelos, sin embargo, es una herramienta costosa, "además el uso de cada solución la venden por paquetes" (Rayon, 2015); esta es una característica que pone en desventaja la herramienta a pesar de que su base de operación con respecto a la herramienta tenga una ventaja competitiva en el mercado. (SAS, 2019)

IBM: tiene un robusto conjunto de características y soluciones a través de IBM SPSS, tiene código abierto Python y R que pueden integrarse. (IBM, s.f.)

SAP: automatiza los procesos de selección del algoritmo de predicción, además valida el modelo creado. Una desventaja es que los modelos no son tan precisos específicamente y además no permite cambiar parámetros de los análisis realizados. (Ircio, 2017)

RAPIDMINER: es una herramienta de Business Intelligence que integra aprendizaje automático para minería de datos. Incluye los procesos de carga y transformación de datos, procesamiento y visualización, análisis predictivo y modelos estadísticos además de la evaluación y despliegue. Una ventaja de la herramienta es que tiene código abierto donde se crean nuevas tendencias y se ajusta a las necesidades del mercado. Además, dispone de una versión gratuita, que permite realizar pruebas gratis.

ORACLE: proporciona a los usuarios el descubrimiento de datos, informes y paneles. Oracle Analytics se enfoca en incorporar, consumir y capacitar modelos de ML para enriquecer su preparación, descubrimiento y colaboración de datos.

ALTERYX: la funcionalidad es la buena integración en lenguaje R análisis de ubicación (location analytics) y geoespaciales, además puede integrarse a Tableau y QlikView y de este modo Alteryx supera sus carencias en visualización interactiva. Pero por el otro lado, Tableau y Qlik superan sus limitaciones en preparación de datos y análisis.

Cloudera (Hadoop), Databricks (Spark) y Revolution Analytics (Lenguaje R), esta última, recientemente adquirida por Microsoft, son otras alianzas estratégicas de este fabricante. Una desventaja es que Alteryx es utilizada sólo como una solución departamental y como complemento de otros productos específicos, por otro lado, tiene limitaciones en reporting y visualización de contenido BI móvil.

Una vez realizada la comparación de las diferentes herramientas posicionadas en el mercado como soluciones analíticas, se obtuvo que, de 13 herramientas, SAS se posiciona como líder en la clasificación de Forrest Research, sin embargo, requiere de una licencia pagada lo cual dificulta la adquisición de la herramienta. En el cuadrante de Gartner 2020, por el contrario, Rapidminer y KNIME, son herramientas posicionadas como visionarios, una ventaja es que ambas herramientas tienen código abierto, esto facilitará la implementación y trabajo del proyecto. Adicionalmente tienen compatibilidad con varias fuentes de datos como Excel, Acces, Oracle, IBM DB2, Microsoft SQL, SAP Sybase, Ingres, MySQL, Postgres, SPSS, dBase y también cualquier fuente de texto plano. Por lo que se seleccionará una herramienta que tenga como ventaja un gran peso de posicionamiento en la clasificación de plataformas, potencial velocidad de desarrollo para el análisis de datos sin requerir de la compra de licencias.

Para obtener una revisión exhaustiva de las capacidades funcionales de cada plataforma, se analizarán las características de cada herramienta según el cuadrante mágico de Gartner en la siguiente tabla:

Tabla comparativa de herramientas analíticas

Tabla 6

Herramientas Posicionadas en el Mercado

Herramienta	Soluciones	Costo	Forrester Research	Gartner
Ibm	Analítica de clientes: detecta tendencias de mercados y anticipación a grados de satisfacción de clientes Analítica operacional: evaluar costes operativos, velocidad, flexibilidad y calidad a través de la recolección, almacenamiento y análisis de datos	Compra de licencia	Líder	Retadores
Sas	Técnicas interactivas de modelado descriptivo. Construcción de modelos predictivos con técnicas como la regresión lineal, modelos lineales generalizados, regresión logística y árboles de clasificación	Compra de licencia	Líder	Líder
Oracle	Oracle r enterprise ofrece bibliotecas de R Oracle data mining ofrece algoritmos de minería de datos de gran alcance	Compra de licencia	Fuerte desempeño	NA
Rapidminer	Permite la carga, la transformación y el modelado de grandes cantidades de datos a partir de fuentes como excel, access, oracle, ibm db2, microsoft sql, sap sybase, ingres, mysql, postgres, spss, dbase y también cualquier fuente de texto plano	Código abierto	Fuerte desempeño	Visionario
Fico	Una gran variedad de algoritmos, incluidos árboles de decisiones y modelos de conjuntos segmentados, que aprenden a reconocer patrones complejos dentro de datos relevantes y voluminosos. Se integra con plataformas analíticas e informáticas populares para que sus herramientas existentes pasen a un nivel totalmente nuevo; esto incluye hadoop, mapreduce, sas y r.	Compra de licencia	Fuerte desempeño	NA

Alteryx	Es una plataforma que permite el acceso, la gestión y el análisis predictivo de los datos en la misma herramienta. Puede manejar grandes volúmenes de datos ser capaz de interpretarlos gracias al análisis espacial	Compra de licencia	Fuerte desempeño	Líder

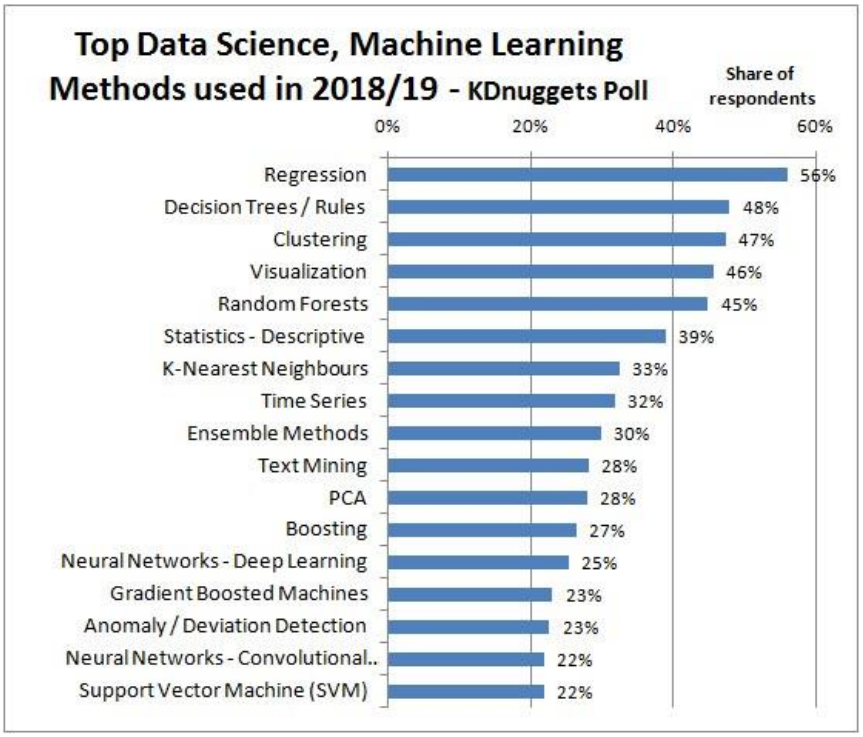
Herramienta	**Soluciones**	**Costo**	**Forrester Research**	**Gartner**
Dell	Ofrece soluciones desde la base de datos relacionales tradicionales hasta clusteres de hadoop Las soluciones de análisis pueden utilizarse con datos estructurados, semiestructurados o no estructurados	Compra de licencia	Fuerte desempeño	NA
Angoss	Es una de las mejores herramientas basadas en el análisis predictivo. Lo que más gusta a los usuarios es que reúne los tres idiomas r, sas y sql. Es fácil de usar en comparación con otros competidores.	Compra de licencia	Fuerte desempeño	Jugadores de nicho
Alpine data labs	Proporciona un entorno visual colaborativo para crear e implementar flujos de trabajo analíticos y modelos predictivos Es una interfaz de análisis avanzada que funciona con apache hadoop y big data	Compra de licencia	Fuerte desempeño	NA
Knime	Es una plataforma de minería de datos para la creación de modelos predictivos visuales	Código abierto	Fuerte desempeño	Visionario
Predixion software	Es una plataforma de análisis basada en la nube que proporciona análisis avanzados en tiempo real en el punto de decisión	Compra de licencia	Contenedores	NA
Microsoft	Sql server usa un conjunto de herramientas para implementar y administrar bases de datos en la nube y en entornos locales que permite a sus clientes el diseño de soluciones de análisis predictivo	Compra de licencia	Contenedores	Visionario

Métodos y Algoritmos Analíticos

Para la construcción del modelo se evaluarán dos tipos de métodos de aprendizaje automático, supervisado y no supervisado. De acuerdo con la publicación “Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey”, los modelos se agrupan en diversas redes poco profundas y grandes conjuntos de datos. Otro reporte importante es el de KDnuggets, que ha identificado a través de una encuesta realizada en el 2019, la comparación de diversos algoritmos basado en una encuesta, que determinó los 17 principales métodos de aprendizaje automático utilizados por los científicos de datos:

Figura 17

Top Data Science, Métodos de aprendizaje automático utilizados, 2018/2019



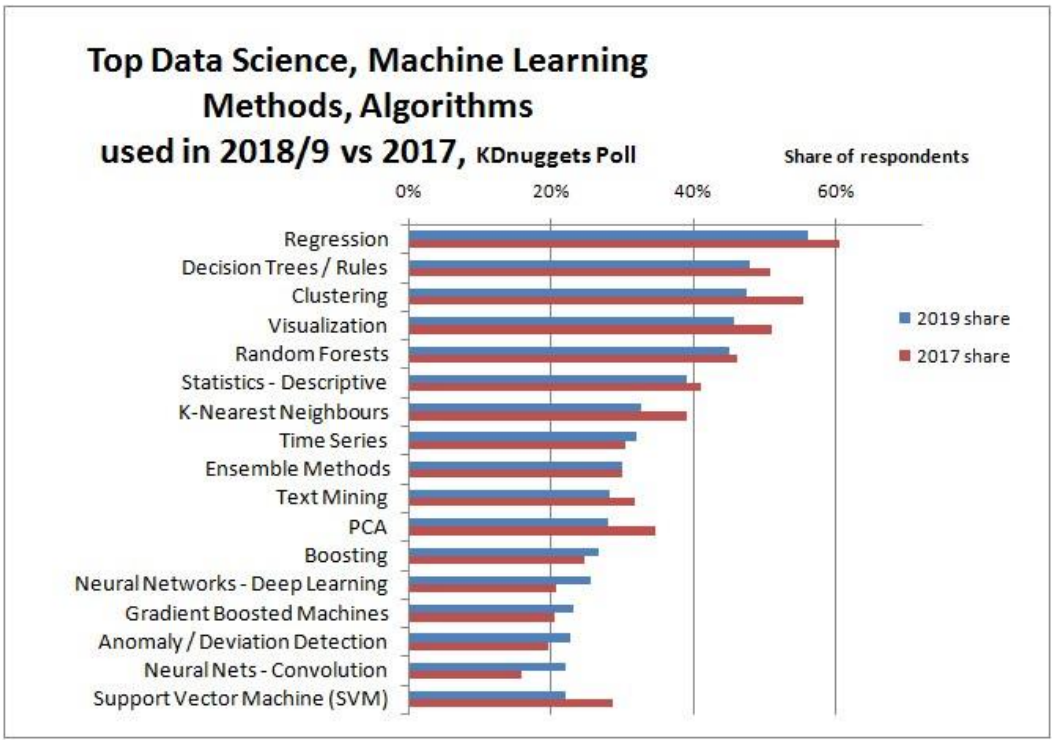
Nota. La figura indica el ranking de los modelos analíticos utilizados por los científicos de datos. Tomado de KDnuggets,2019, (https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html)

A comparación de las encuestas realizadas en el año 2017, el encuestado utilizó 7,4 métodos y algoritmos, de acuerdo con la figura 17 Top Data Science, Métodos de aprendizaje automático utilizados, 2018/9 vs 2017; hay un gran aumento en el uso de tecnologías basadas en redes neuronales. Se pude determinar los siguientes incrementos:

- Generative Adversarial Networks (GAN) 101.8% más, de 2.3% en 2017 a 4.7% en 2018/9
- Redes neuronales – Redes neuronales recurrentes (RNN) 56.5% más, del 10.5% al 16.5%
- Aprendizaje de refuerzo 56.1% más, de 4.2% a 6.6%
- Redes neuronales – Convolución 38.8% más, del 15.8% al 22%
- Otros métodos 27.1% más, de 6.1% a 7.8%
- Algoritmos y métodos genéticos / evolutivos 25.7% más, de 4.8% a 6.0%
- Redes neuronales – Aprendizaje profundo 23.5% más, del 20.6% al 25%

Figura 18

Top Data Science, Métodos de aprendizaje automático utilizados, 2018/9 vs 2017



Nota. La figura muestra el ranking comparativo 2017 y 2019 de los métodos analíticos utilizados por los científicos de datos. Tomado de KDnuggets,2019, (https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html)

Selección del algoritmo analítico

En base al análisis realizado sobre los casos de éxito de modelos analíticos para el aprovisionamiento de productos relacionados a la industria farmacéutica y además los resultados arrojados por KDnuggets, se puede determinar que las técnicas y métodos de algoritmo más utilizados con éxito son en las industrias relacionadas a los productos farmacéuticos son: método de regresión, series temporales, redes neuronales.

En el caso 2, se pudo constatar que el método ARIMA es un modelo integrado de media móvil que permite un mejor pronóstico de ventas a corto plazo, en este sentido se pueden agrupar o

clarificar los productos por ATC (Anatomical Therapeutic Chemical[12]) para obtener un mejor alcance del pronóstico.

Una de las ventajas del modelo ARIMA es que permite descubrir valores de forma lineal a partir de datos anteriores, en este caso puede partir del histórico de ventas, para la realización del pronóstico. Sin embargo, se trabajará con el modelo de Regresión, ya que se encuentra con mejor ranking en las encuestas de métodos de aprendizaje automático utilizado en el 2019, realizadas por KDnuggets. Este método, además, incluye factores cíclicos o también estacionales ya que las ventas de los productos OTC pueden tener un comportamiento que depende además de las series de tiempo.

De acuerdo a la evaluación realizada, los modelos seleccionados permiten evidenciar que los científicos de datos lo utilizan en mayor porcentaje (67%) la regresión, tomando como ventaja que este modelo utiliza la relación entre varias variables dependientes (Y) e independientes (X), por otro lados las series temporales encajan con las variables seleccionadas ya que tienen como característica, la forma secuencial de las series temporales, tendencia y estacionalidad como en el caso de CORP. LA FAVORITA, el modelo de regresión que utilizaron obtuvo un menor error en todo el concurso, el cual contribuye satisfactoriamente a los resultados del desarrollo del proyecto, lograron pronosticar con mayor precisión las ventas de los productos.

El método de series temporales, en el caso 2 de la distribuidora farmacéutica, demostró que los análisis de estacionalidad son útiles para identificar los períodos en los que se pueden

[12] ATC: es un sistema europeo de codificación de sustancias farmacéuticas y medicamentos en cinco niveles con arreglo al sistema u órgano efector y al efecto farmacológico, las indicaciones terapéuticas y la estructura química de un fármaco. (María Verónica Saladrigas, marzo 2014)

implementar campañas especiales de ventas y marketing, lo cual permite tener mayor precisión con el análisis del forecast de ventas a nivel de negocio.

Figura 19

Comparación Top Ranking KDnuggets vs Caso de Éxito

TOP RANKING
1. Regresión
2. Árboles / Reglas de Decisión
3. Agrupación
4. Visualización
5. Bosques al azar
6. Estadística – Descriptiva
7. K-vecinos más cercanos
8. Series de Tiempo
9. Métodos de conjunto
10. Minería de textos
11. PCA
12. Impulso
13. Redes neuronales – Aprendizaje profundo
14. Máquinas de refuerzo gradual
15. Detección de anomalía / desviación
16. Redes neuronales – Redes neuronales convolucionales (CNN)
17. Máquina de vectores de soporte (SVM)

CASOS DE ÉXITO
1. Regresión
8. Series Temporales
13. Redes neuronales

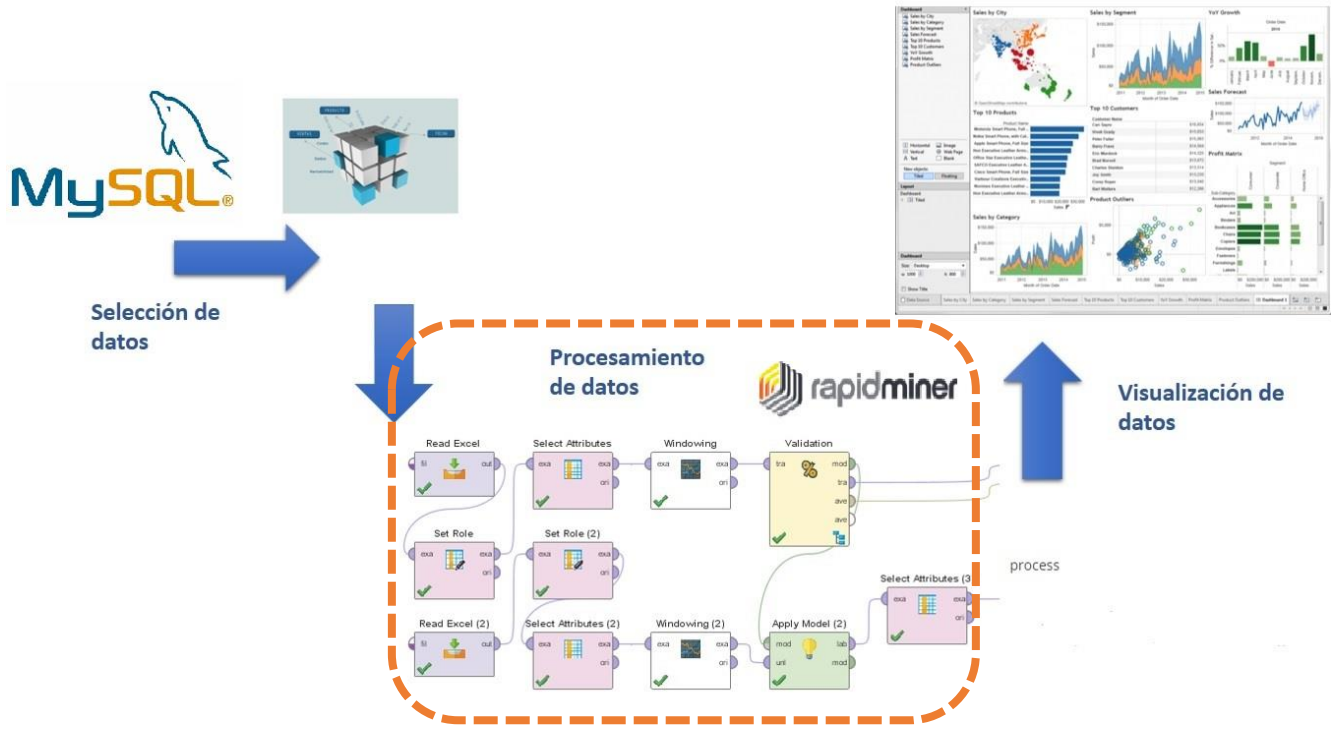
Nota: la figura muestra la lista de los métodos analizados y evaluados con respecto a los casos de éxito revisados.

Arquitectura Propuesta Para la Toma De Decisiones

En base al análisis realizado, se plantea el diseño de la construcción como lo sugiere la metodología de investigación Design Science Research en la fase II, para dar la solución de acuerdo a la necesidad que requiere el Laboratorio Farmacéutico, en obtener un modelo analítico para la predicción de la demanda, que le permita tomar decisiones asertivas, se realiza la siguiente propuesta como arquitectura de flujo de datos para la toma de decisiones:

Figura 20

Propuesta para el flujo de datos



La figura 20, indica los componentes claves para el flujo de datos y la arquitectura propuesta para la toma de decisiones:

La fuente de datos: que recopila el Laboratorio Farmacéutico es una mezcla entre archivos xlsx y csv, como también de motores de bases de datos existentes, los cuales serán ingestados en MySQL para su tratamiento y posterior procesamiento.

La calidad de datos: Rapidminer permite realizar el procesamiento y limpieza de datos, como: eliminar datos faltantes, agrupar datos, crear nuevos atributos, transformar datos y detectar errores de los valores atípicos.

Analítica Avanzada: para realizar este proceso se aplicará a través de la herramienta Rapidminer distintos modelos, los mismos que se definieron en la sección de la selección del algoritmo analítico, para predecir los valores futuros de las ventas del Laboratorio Farmacéutico donde implica el siguiente proceso:

- **Entrada**
  - modelo de pronóstico *(IOObject):* El modelo de pronóstico para el que se generarán nuevos valores. También contiene los valores de la serie temporal original en la que se entrenó el modelo.
- **Salida**
  - conjunto de ejemplos *(tabla de datos)*: *ExampleSet* que contiene los valores de serie temporal previstos. Dependiendo de los parámetros seleccionados, también se incluyen los valores originales.
  - original *(IOObject)* El modelo que se proporciona como entrada se pasa sin cambios.
- **Parámetros**
  - Forecast_horizon: este parámetro especifica la duración del pronóstico. Indica cuántos puntos nuevos genera el modelo.
  - add_original_time_series: Si este parámetro se establece en verdadero, se agrega al resultado un atributo adicional que contiene los valores originales de la serie temporal
  - add_combined_time_series: si este parámetro se establece en verdadero, se agrega un atributo adicional al resultado. Este atributo contiene los datos de series temporales originales y los valores pronosticados.

Exploración y visualización de la información; Rapidminer también incluye herramientas y gráficos de visualización interactivas de datos que permitirá evaluar el estado la calidad e integración de los datos. Estos incluyen comprender los patrones de los datos, la distribución de diagramas y gráficas lineales.

Capítulo IV

Desarrollo del Modelo Predictivo

En esta sección, se realizó la implementación de la solución de acuerdo con la metodología de investigación aplicada Design Science Research (DSR). Además, se utilizó la metodología KDD como parte del proceso de desarrollo para el descubrimiento de datos y desarrollo de los modelos predictivos.

Para definir las fases del desarrollo del proyecto se utilizó la metodología KDD (Knowledge Discovery in Database), que permitió a través de un esquema ordenado el descubrimiento y el análisis de los datos de las ventas. Este proceso permite identificar patrones válidos y potencialmente útiles. Para esto se realizó el seguimiento de las 5 fases que son parte de KDD.

Desarrollo Basado en la Metodología KDD

"El Descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados (Agrawal y Srikant, 1994) (Chen, Han y Yu,1996) (Piatetsky Shapiro, Brachman y Khabaza, 1996) (Han y Kamber, 2001)." (Timarán-Pereira, y otros, 2016)

La metodología KDD es una herramienta que permite de manera ordenada la realización de minería de datos, la cual se aplicó dentro de la metodología de investigación Design Science Research en la fase del desarrollo, en este sentido el descubrimiento de los datos y el manejo de la base de datos de las ventas de la línea OTC nos permite seguir un proceso metodológico más ordenado y eficiente para descubrir el modelo predictivo de las ventas.

Conocimiento del negocio

La comprensión del conocimiento del negocio es importante por lo que se debe entender los objetivos de la aplicación de minería de datos en el presente proyecto y aprender acerca de la situación actual e importancia con respecto a la proyección de ventas del Laboratorio Farmacéutico.

Para obtener más información sobre la situación actual del negocio, se realizó una entrevista a los responsables involucrados en realizar el forecast de productos de la línea OTC, quienes son: el Gerente de Planificación y la Gerente de la línea OTC. Las preguntas fueron estructuradas con la intención de ratificar el contexto del problema y objeto de estudio

A continuación, se detallan las preguntas y respuestas realizadas referente a la realización de los pronósticos del Laboratorio Farmacéutico:

¿Qué departamento/área de trabajo lidera la obtención del pronóstico de venta?

Actualmente el pronóstico lo realiza el departamento comercial y planificación, sin embargo, la información la obtiene el área comercial a través de los históricos de ventas.

¿De qué forma se está calculando el pronóstico?

El pronóstico se calcula en base al promedio de ventas del último trimestre, es decir se aplica el método de media móvil.

¿Cuál es el nivel de precisión de los pronósticos realizados?

La precisión que se tiene es del 15% en promedio de error del pronóstico

¿Considera importante la construcción de un modelo predictivo para el pronóstico de ventas?

Los responsables de cada departamento respondieron que, sí es importante apoyarse en un modelo predictivo, que permita obtener información sistematizada para realizar el forecast de los productos, reduciendo el error del pronóstico.

¿Cree usted que un modelo predictivo para el pronóstico de ventas mejore la ejecución de los objetivos del departamento comercial y de producción?

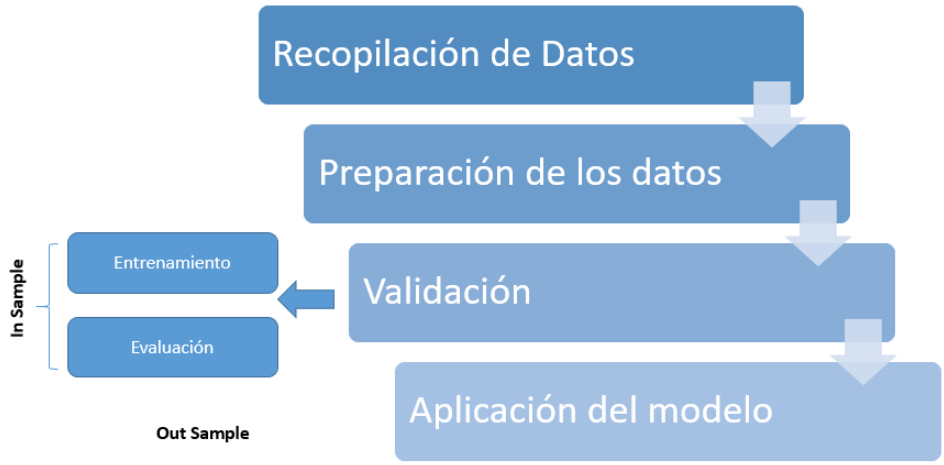
El Gerente de Planificación y la Gerente línea OTC, están de acuerdo en que mejorar la precisión de los pronósticos a través de un modelo predictivo les permitirá tomar mejores decisiones, relacionado con los objetivos de cada departamento.

La aplicación de la metodología KDD requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada, debido a que en muchas ocasiones los datos provienen de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido. Por otra parte, es necesario interpretar y evaluar los resultados obtenidos.

Para la extracción del conocimiento de la base de datos de las ventas de los productos OTC se siguió las siguientes fases:

Figura 21

Esquema del Constructo



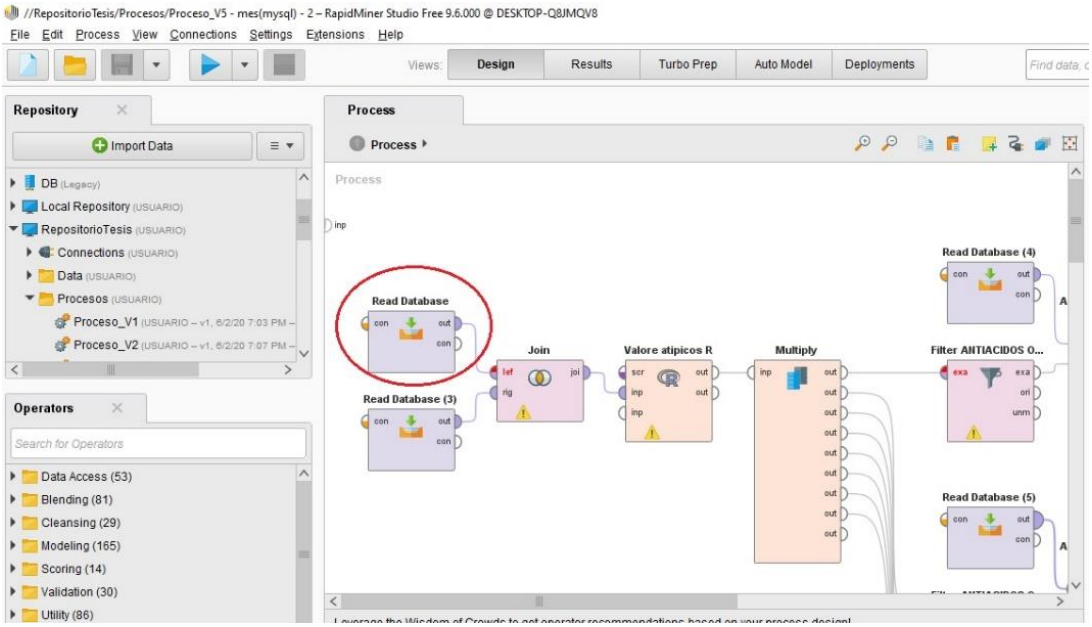
Fase I selección de datos

En esta etapa se recopilan todos los datos históricos de las ventas de la línea OTC, de los años 2018 y 2019. A través de la fuente de información de la base de datos MySQL, que permitirá estructurar el esquema de información*.*

Los datos están compuestos de variables internas relacionadas al: producto, subcategorías, cliente, valores y tiempo; reuniendo 3086 registros. Y como variables externas los datos del PIB, precio de petróleo y días festivos. Estos datos se identificaron como relevantes para la realización de minería de datos y posterior construcción del modelo predictivo.

Figura 22

Lectura de la Base de Datos



Nota. La figura muestra el flujo del proceso que construye modelo analítico predictivo, iniciando con el operador de lectura de base de datos.

Además, se consideraron variables correlacionadas a la venta de los productos OTC, como el PIB (Producto Interno Bruto), precio del petróleo y días festivos; en la misma estacionalidad de las ventas realizadas.

Se consideró el PIB, ya que es un indicador que está directamente relacionado con las ventas de una empresa, en este caso el Laboratorio Farmacéutico puede esperar un aumento o decrecimiento en el consumo de sus ventas.

Para cada caso se determinó el precio del petróleo como una variable relacionada, ya que la significancia de este indicador puede afectar las variaciones de la demanda de consumo.

Y adicionalmente se consideraron los días correspondientes al feriado nacional, ya que esta variable también influye en la tendencia de demanda, debido a la fluctuación de poder adquisitivo.

Fase II exploración y limpieza de datos

En esta fase se realizó la exploración de los datos recopilados, la revisión de la información duplicada y datos incompletos, es decir la revisión de la calidad de los datos de las ventas de los 24 últimos períodos.

En la revisión realizada se detectó que en tres categorías solo se clasifica un producto, por lo tanto, los datos se considerarán a partir de la subcategoría y producto.

Para la validación de datos se verificó además las series de tiempo por cada subcategoría:

- Visualización de las series de tiempo
- Verificación de la estacionalidad

Figura 23

Subcategorías Productos OTC

SUBCATEGORIA	PRODUCTO
ANTIACIDOS OTC	SKU1
	SKU2
CALCIO	SKU3
DESCONGESTIVOS Y ANTIALERGICOS	SKU4
	SKU5
DIGESTIVOS ENZIMAS	SKU6
HEPATOPROTECTORES OTC	SKU7
	SKU8
LAGRIMAS NATURALES	SKU9
MULTIVITAMINICOS ADULTOS	SKU10
VITAMINAS OTC	SKU11

Nota. La figura muestra las subcategorías utilizadas en la construcción de los modelos analíticos y el número de productos

Además, con el objetivo de obtener resultados de calidad en el proceso de minería de datos y que estos sean más útiles, se consideran aquellas variables o atributos que van a ser relevantes para la realización de la predicción.

Figura 24

Exploración de Datos

Data Frame Summary

data

Dimensions: 22122 x 23
Duplicates: 28

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	COD.CLI [character]	1. 300030 2. 300208 3. 305924 4. 300040 5. 306279 6. 300090 7. 300000 8. 300794 9. 303392 10. 307018 [109 others]	3782 (17.1%) 2787 (12.6%) 2704 (12.2%) 2614 (11.8%) 1974 (8.9%) 1091 (4.9%) 1046 (4.7%) 1003 (4.5%) 927 (4.2%) 741 (3.3%) 3453 (15.6%)		22122 (100%)	0 (0%)
2	CLIENTE [character]	1. DISTRIBUIDORA FARMACEUTIC 2. CORPORACION EL ROSADO S.A 3. MEGA SANTAMARIA S.A. 4. PHARMABRAND S.A 5. QUIFATEX S.A. 6. FARMACIAS Y COMISARIATOS 7. ECONOFARM S.A.	3856 (17.4%) 2787 (12.6%) 2704 (12.2%) 2614 (11.8%) 1974 (8.9%) 1091 (4.9%) 1046 (4.7%)		22122 (100%)	0 (0%)

Para la limpieza de datos se utilizó un operador en Rapidminer que detecta datos faltantes y en blanco, además de valores duplicados que pueden encontrarse en la base de datos de las ventas.

A pesar de que inicialmente no se obtuvo datos perdidos en la base de datos, la información cruzada con las variables relacionadas (PIB, precio del petróleo, días feriado nacional) generaron datos perdidos, ya que se consideró la frecuencia de venta por semana, en este sentido no todas las variables tuvieron un valor o un dato a este nivel, por lo que se realizó una interpolación[13] de los datos.

Fase III transformación

"En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos (Fayyad et al., 1996)." (Timarán-Pereira, y otros, 2016)

La transformación consistió en simplificar la tabla de datos, para esto se transformaron los atributos según las necesidades del algoritmo aplicado. Es decir, se realizó la transformación de variables con respecto a la reducción de forma horizontal a la variable "fecha", ya que esta no tenía formato único.

Como complemento, en la etapa de transformación de datos se realizó un script a través del operador Execute R, para corregir los valores atípicos, para el Q1 = 0,25 y Q3=0,75.

En ciertos casos, para utilizar técnicas que necesitaban otro tipo de configuración o algoritmos más complejos se hizo uso del motor estadístico R para implementar script acorde a las

[13] Interpolación: es un proceso para estimar los valores que se encuentran entre los puntos (datos) conocidos. (Amos, 2006)

necesidades, esta es una de las muchas ventajas que nos brinda RapidMiner y una de las razones por la cual se trabajó con dicha herramienta.

1. Para la transformación de datos de las variables externas, se realizó una interpolación de las series para cada fecha
2. Para la transformación de las series de ventas:

- Se verificaron los valores faltantes: imputación por el promedio de valores cercanos
- Se verificaron los valores atípicos: imputación por el percentil 95

Fase IV minería de datos

La realización de minería de datos implica la descripción y predicción, el primer caso se enfoca en encontrar patrones que puedan describir a los datos y sean interpretables y en el segundo donde las variables permitan predecir los valores desconocidos. En esta fase se ejecutaron modelos con estacionalidad semanal y mensual para obtener de esta manera los resultados de la venta.

En este sentido para el descubrimiento de patrones que interesan en las ventas de la línea de productos OTC, se realizó la técnica de minería de datos basado en correlaciones y patrones secuenciales.

En esta fase se identifican patrones para poder explicar los datos, ya sea realizando un análisis descriptivo que incluyen correlaciones, reglas de asociación, patrones secuenciales y clustering.

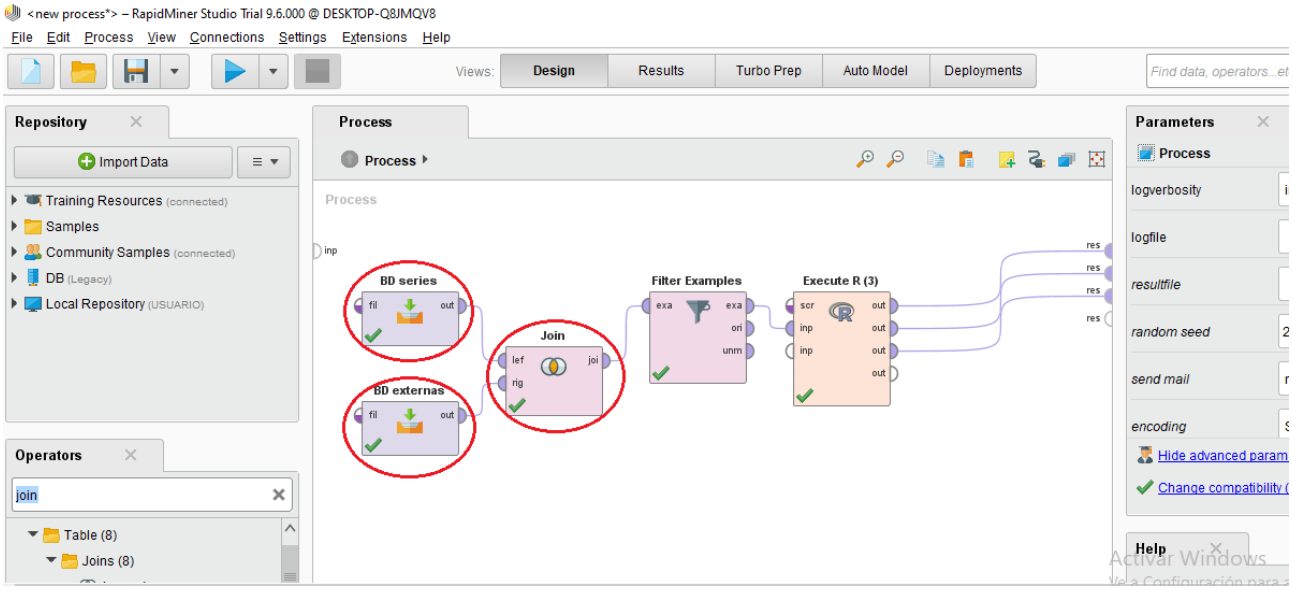
Modelamiento de las Series de Tiempo.

***Unificación de Todas las Bases*.** Para la unificación de todas las bases de datos (Histórico de ventas, PIB, precio del petróleo, días de feriado nacional), se realizó un join entre variables

Se utilizó las variables de las series de tiempo y variables externas como entrada, utilizando atributos clave de acuerdo con el ID, para obtener el conjunto de datos.

Figura 25

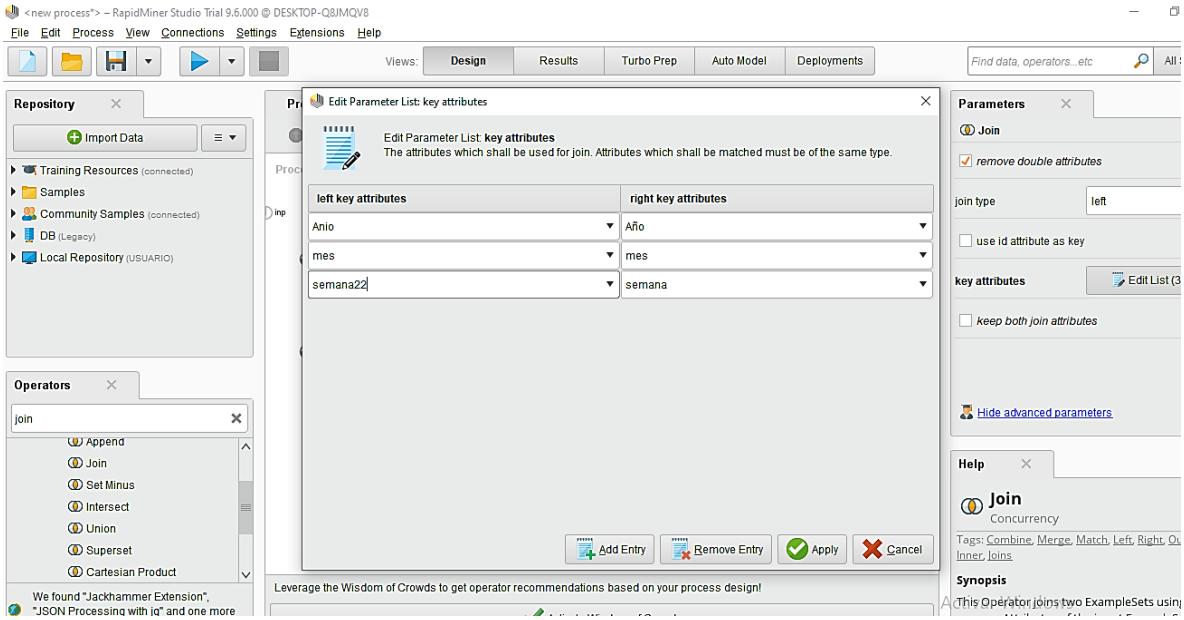
Join BD series de tiempo /BD externas



En la figura 25 se puede apreciar el flujo del proceso y la unión de los conjuntos de datos (BD_series) y (BD_externas) con el operador “Join”. La serie de tiempo de cada base de datos es el registro en relación a nivel semana

Figura 26

Join Atributos

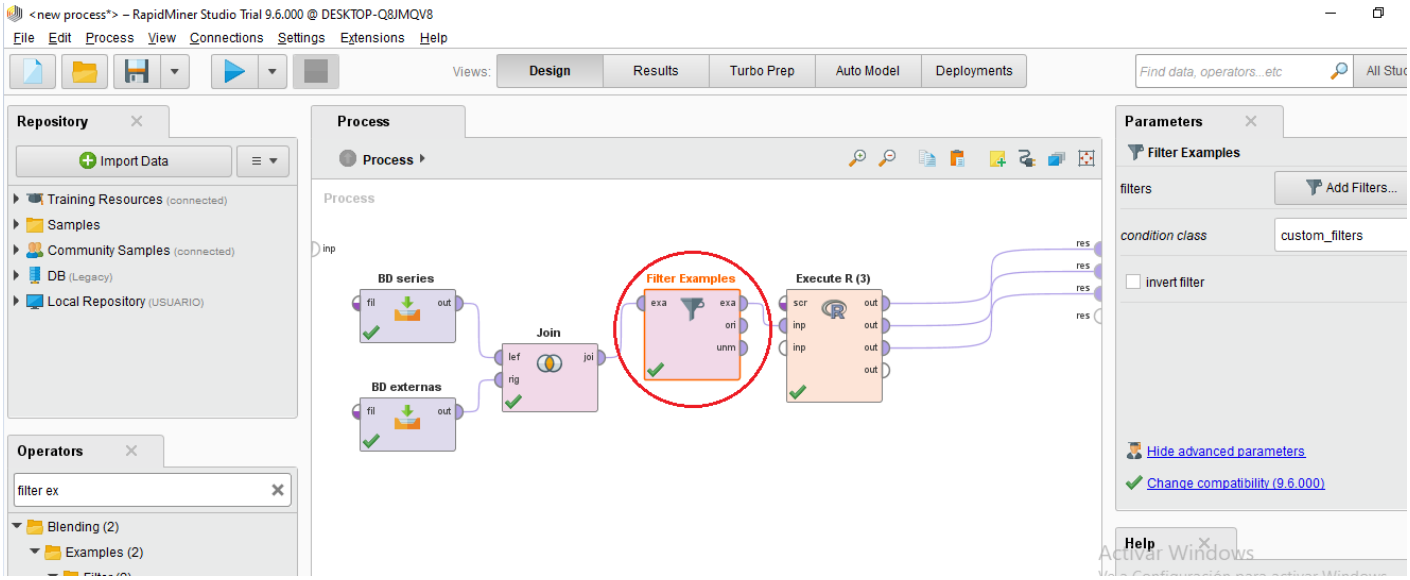


Nota. La figura muestra los atributos de las bases internas y externas, con las que se utiliza el operador Join en la herramienta RapidMiner

Separación de Cada Serie de Tiempo Por la Subcategoría Respectiva. Dentro de la base de datos, se obtuvieron 8 subcategorías, las cuales fueron filtradas con el operador “Filter Examples”. Este operador facilita la filtración de los atributos de la base de datos, en este caso se filtró cada subcategoría para posteriormente aplicar los modelos: ARIMAX, REDES NEURONALES y HOLT WINTERS

Figura 27 *Filter Examples Subcategorías*

Filter Examples Subcategorías



Nota. La figura muestra el operador que filtra las subcategorías para la construcción del modelo.

Para la construcción del modelo se utilizaron los datos de las ventas a nivel de subcategorías de producto, por lo que este es el parámetro que permitió la definición de la condición de filtro personalizado. Esta condición consta de un atributo y un valor para que coincidan las series de tiempo.

Comparación de las Variables Internas y Externas. De acuerdo con los resultados obtenidos, en la siguiente figura 28, se puede observar que las ventas tienen una débil correlación con el valor del PIB, específicamente 0,002. Lo mismo ocurre entre las ventas y los días de feriado, con una correlación de 0,002; y por último el precio del petróleo tiene un valor de 0,003.

Estos valores muestran que las correlaciones entre las variables externas y las ventas en unidades tienen una correlación débil en el periodo 2018 y 2019.

Unidad de Venta-Valor del PIB

Figura 28

Correlación Venta-PIB

Nota. La figura muestra la correlación de 0,002 entre las variables, ventas del laboratorio farmacéutico y el PIB.

Figura 29

Series Temporales y PIB

ID	SerieTiempo	Valor_pib_interp
1	36.090,00	0,10
2	23.772,00	-0,50
3	16.731,00	-1,10
4	18.381,00	-0,40
5	22.611,00	0,30
6	26.784,00	1,00
7	29.673,00	0,93
8	29.145,00	0,87
9	44.065,50	0,80
10	34.011,00	0,53
11	23.928,00	0,27

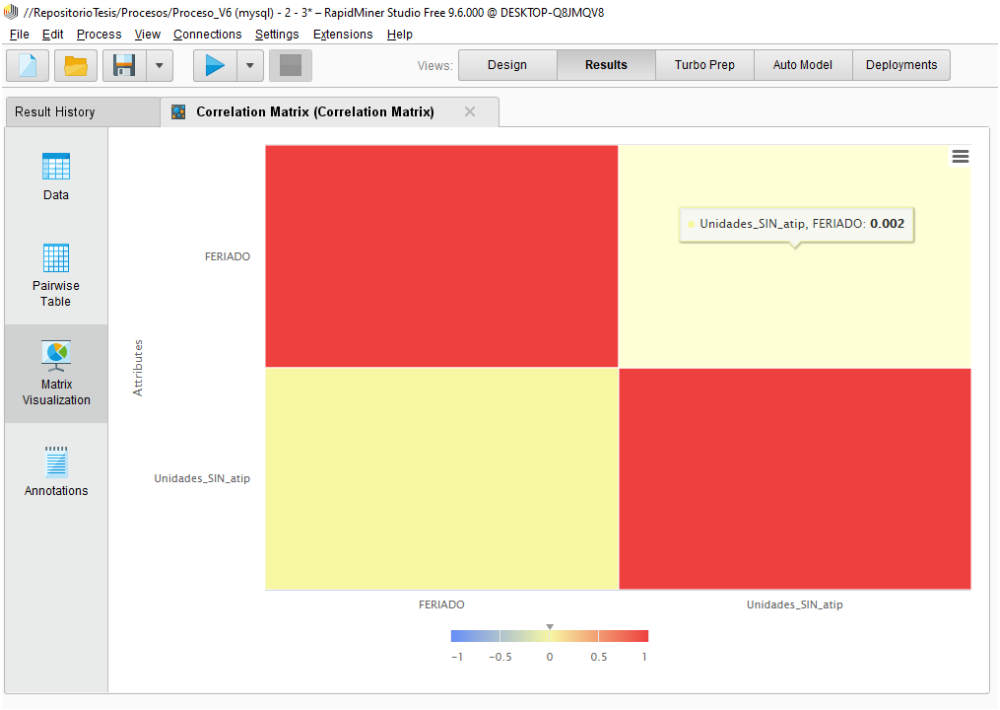
Nota. La figura muestra la tendencia de las series de tiempo, entre variables ventas del laboratorio farmacéutico y el PIB

Unidad de Venta – Feriados

El número de feriados tuvo una correlación de 0.002 con respecto a las ventas, es decir una débil correlación. Y la tendencia muestra que el número de unidades vendidas no aumenta en los días de feriado.

Figura 30

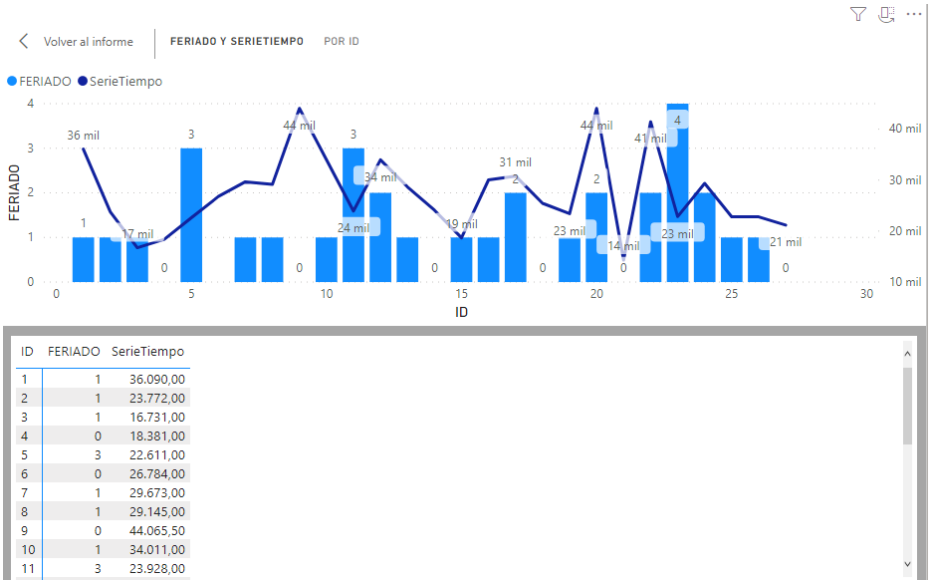
Correlación Venta-Feriados



Nota. La figura muestra la correlación de 0,002 entre las variables ventas del laboratorio farmacéutico y día de feriado

Figura 31

Series Temporales y Número de Feriados



ID	FERIADO	SerieTiempo
1	1	36.090,00
2	1	23.772,00
3	1	16.731,00
4	0	18.381,00
5	3	22.611,00
6	0	26.784,00
7	1	29.673,00
8	1	29.145,00
9	0	44.065,50
10	1	34.011,00
11	3	23.928,00

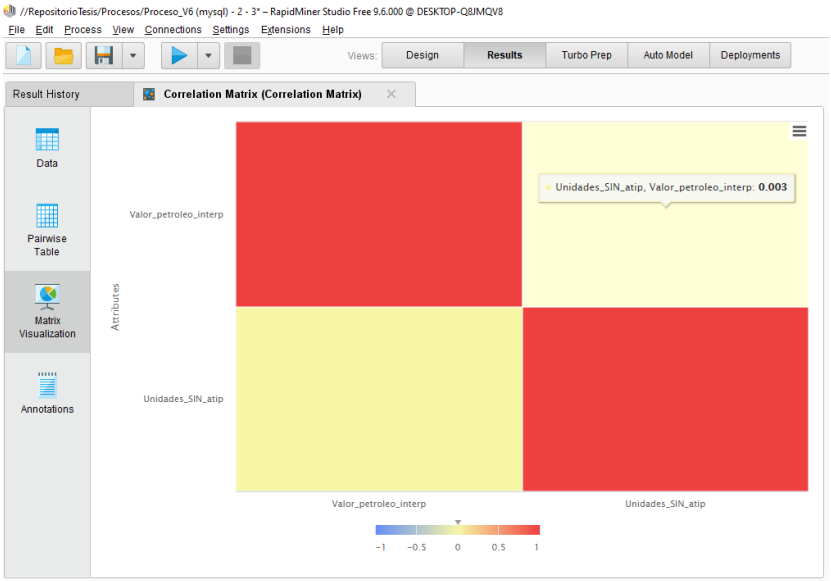
Nota. La figura muestra la tendencia de las series de tiempo de las ventas del laboratorio farmacéutico y los días de feriado

Unidad de Venta – Precio del Petróleo

La variable “precio del petróleo” comparada con la unidad de venta, arrojó una correlación de 0.003, la cual es débil, es decir que no están relacionadas estrechamente, la tendencia muestra que las series de tiempo no están estrechamente relacionadas con el precio del petróleo, solo desde el periodo 10 se observa que mientras disminuye el precio del petróleo también disminuye la unidad de venta.

Figura 32

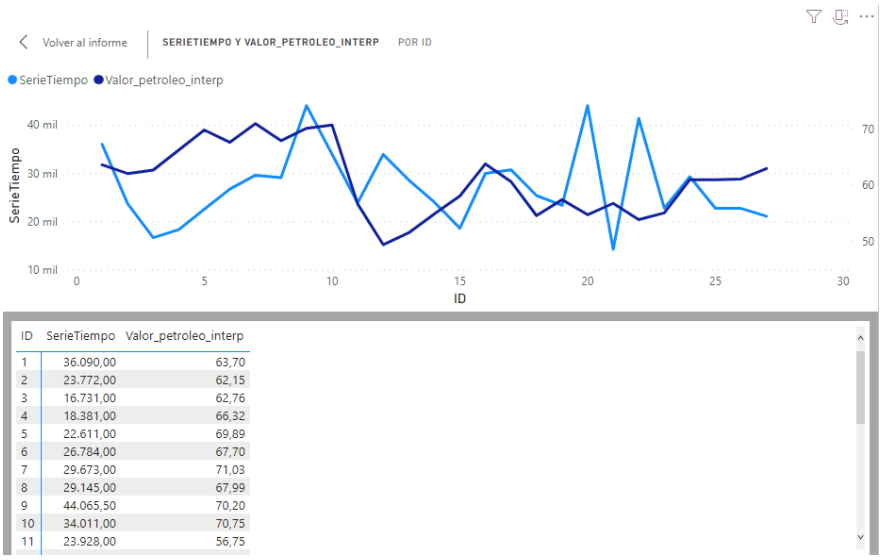
Correlación Venta - Precio del Petróleo



Nota. La figura muestra la correlación de 0,003 entre la variable ventas del laboratorio farmacéutico y el precio del petróleo

Figura 33

Series de Tiempo y Valor del Petróleo



ID	SerieTiempo	Valor_petroleo_interp
1	36.090,00	63,70
2	23.772,00	62,15
3	16.731,00	62,76
4	18.381,00	66,32
5	22.611,00	69,89
6	26.784,00	67,70
7	29.673,00	71,03
8	29.145,00	67,99
9	44.065,50	70,20
10	34.011,00	70,75
11	23.928,00	56,75

Nota. La figura muestra la tendencia entre la serie de tiempo de las ventas del laboratorio farmacéutico y el precio del petróleo

Aplicación del Modelo y Obtención de los Pronósticos Respectivos. Según lo definido en el capítulo III, se consideraron tres modelos a utilizar: ARIMAX, REDES NEURONALES y HOLT WINTERS. El entrenamiento de datos se realizó a través del flujo de cada operador seleccionado que permitió obtener los algoritmos para la realización del pronóstico de ventas de cada subcategoría, y además la métrica MAPE para la validación de los modelos.

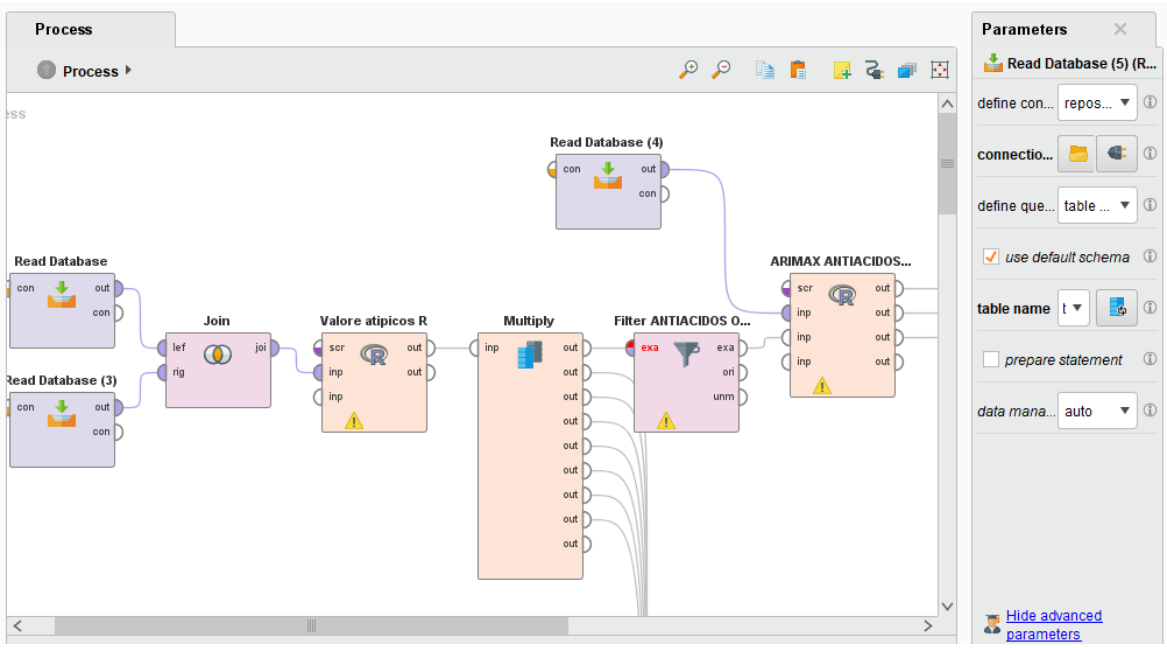
Para la realización del pronóstico por subcategoría, se entrenaron los diferentes modelos, obteniendo como resultado una nueva serie con valores de las respectivas predicciones, al periodo enero-marzo 2020.

Implementación de un Modelo ARIMAX. Para el pronóstico móvil integrado se aplicó el operador ARIMAX, este modelo en estadística y econometría, en particular en series temporales, es un modelo autorregresivo integrado de promedio móvil que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro.

Este modelo de regresión múltiple permitió pronosticar los datos estacionarios, donde se consideraron las variables externas.

Figura 34

Parámetros Modelo ARIMAX



Nota. La figura muestra los parámetros requeridos para la construcción del modelo ARIMAX en la herramienta Rapidminer

En la salida de los datos se obtuvo la predicción del modelo ARIMAX ajustado, para los próximos 3 períodos:

Figura 35

Modelo Arimax

```
modeloST <- auto.arima(serie, xreg=xreg)

xregT <- as.matrix(dataT[, c("Valor_pib_interp", "Valor_petroleo_interp", "FERIADO")])
pronosticos <- forecast(modeloST, h=3, xreg=xregT[1:3,])
pronosticos <- as.data.frame(pronosticos)
pronosticos$ID <- 1:3

asas <- summary(modeloST)
df_metricas <- data.frame(
METRICA = c("MAE", "MAPE", "MRSE"),
VALOR   = c(asas[3], asas[5], asas[2])
)
```

Nota. La figura muestra el modelo predictivo ARIMAX de las unidades de ventas del laboratorio farmacéutico

En la figura 35, el código ARIMAX muestra el algoritmo del modelo donde se consideran las variables internas (series temporales) y las variables externas (PIB, precio del petróleo y feriados), para pronosticar los siguientes 3 períodos mensuales.

Forecast Modelo ARIMAX. A continuación, se presentan los datos de la predicción realizada con el modelo ARIMAX, exportada en formato Excel, de los siguientes 3 períodos.

Tabla 7

Forecast Ventas Subcategoría Antiácidos, Modelo ARIMAX

SUBCATEGORÍA	SerieTiempo	Tipo	Año	ID
ANTIÁCIDOS OTC	20.132	predicción	2020	1
ANTIÁCIDOS OTC	23.422	predicción	2020	2
ANTIÁCIDOS OTC	20.510	predicción	2020	3

En la tabla 7, se presenta la predicción obtenida con el modelo ARIMAX. La estructura está compuesta por las subcategorías, series de tiempo (valores en unidad de venta), el tipo que se refiere a la predicción de los valores y el ID correspondiente al periodo pronosticado.

Figura 36

Gráfica forecast ventas, Modelo ARIMAX

Nota. La figura muestra el forecast de los tres períodos proyectados de las unidades de venta del laboratorio farmacéutico, con el modelo ARIMAX.

Implementación de un Modelo Redes Neuronales. Se aplicó el operador Deep Learning para la construcción del modelo redes neuronales, este operador se basa en una red neuronal artificial, que emplea múltiples capas para el entrenamiento del modelo, capaz de aprender del modelo mediante los datos históricos de las ventas.

El modelo de redes neuronales se basa en multicapas que se entrenan con el descenso de gradiente estocástico utilizando el método de propagación hacia atrás (backpropagation). La red puede contener una gran cantidad de capas ocultas que consisten en neuronas con funciones de activación de tanh. Las características avanzadas como la tasa de aprendizaje adaptativo, el recocido de la tasa y otras regularizaciones permiten una alta precisión predictiva al disminuir el error de predicción probabilístico. Para los parámetros del operador:

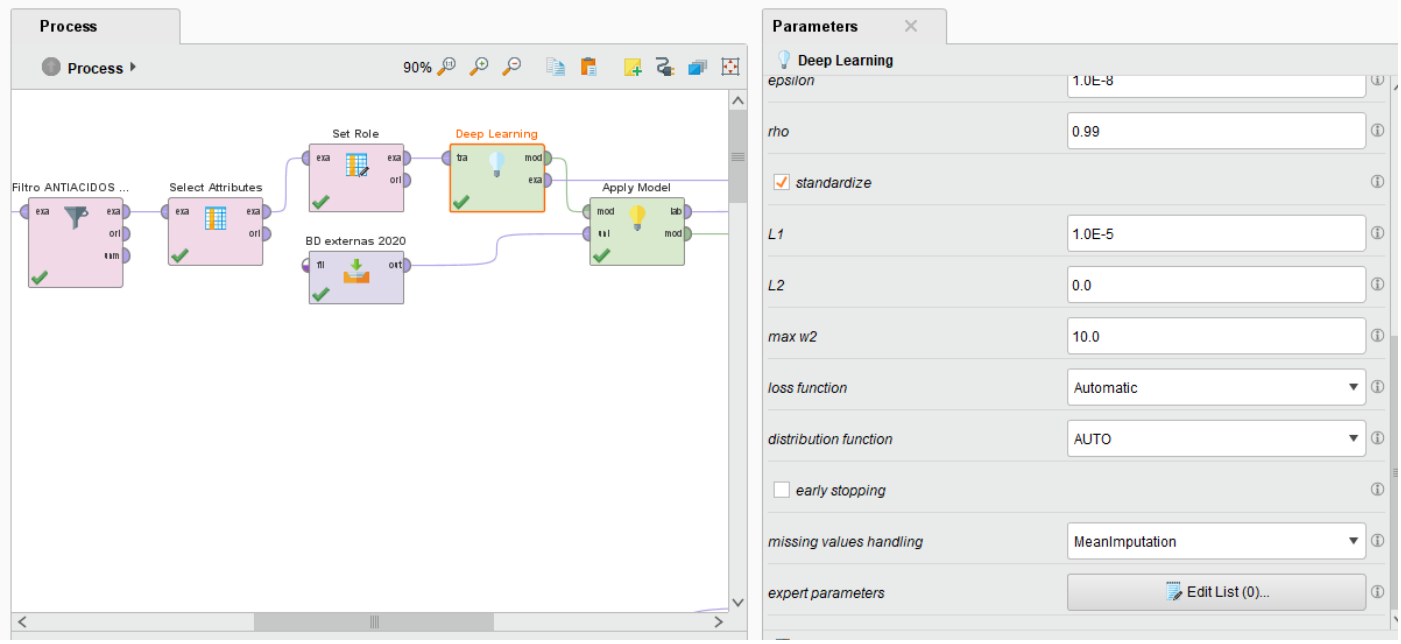
- Hidden layers: se aplicaron 2 capas ocultas por 50 neuronas artificiales, creando así modelo inicial sencillo
- Loss_function y distribution_function: La función perdida y de distribución se establece por defecto es decir es automática.
- Max w2: el valor máximo de la suma de las raíces cuadradas de las entradas de una neurona, por defecto se establece en el valor 10. Se utilizó para el número de intentos permitidos para lograr conjuntos de elementos mínimos.

El operador inicial es un clúster H2O[14] que ejecuta el algoritmo, se utilizó para predecir el atributo del conjunto de datos.

[14] H2O: es un producto creado por la compañía H2O.ai con el objetivo de combinar los principales algoritmos de machine learning y aprendizaje estadístico con el Big Data (abril 2020, Joaquin Amat)

Figura 37

Parámetros Modelo Redes Neuronales



Nota. La figura muestra los parámetros requeridos para la creación del modelo redes neuronales en la herramienta Rapidminer.

El operador Deep Learning se usó la opción de tasa de aprendizaje adaptativo (predeterminada). El algoritmo determinó automáticamente la tasa de aprendizaje/performance en función de los parámetros determinados.

Forecast Modelo Redes Neuronales. Para el desarrollo del forecast se procedió a realizar el entrenamiento de los datos del histórico de ventas por subcategoría, se seleccionaron los atributos (feriado, unidades sin valores atípicos, valor del petróleo y valor del PIB). Los parámetros del modelo fueron construidos con 2 capas ocultas de 50 neuronas cada una por defecto.

Se obtuvo la siguiente estructura de datos: la subcategoría, las series de tiempo (valor de unidad de venta) que corresponde al tipo de predicción, los valores de las variables externas, y el ID que corresponde al periodo de tiempo.

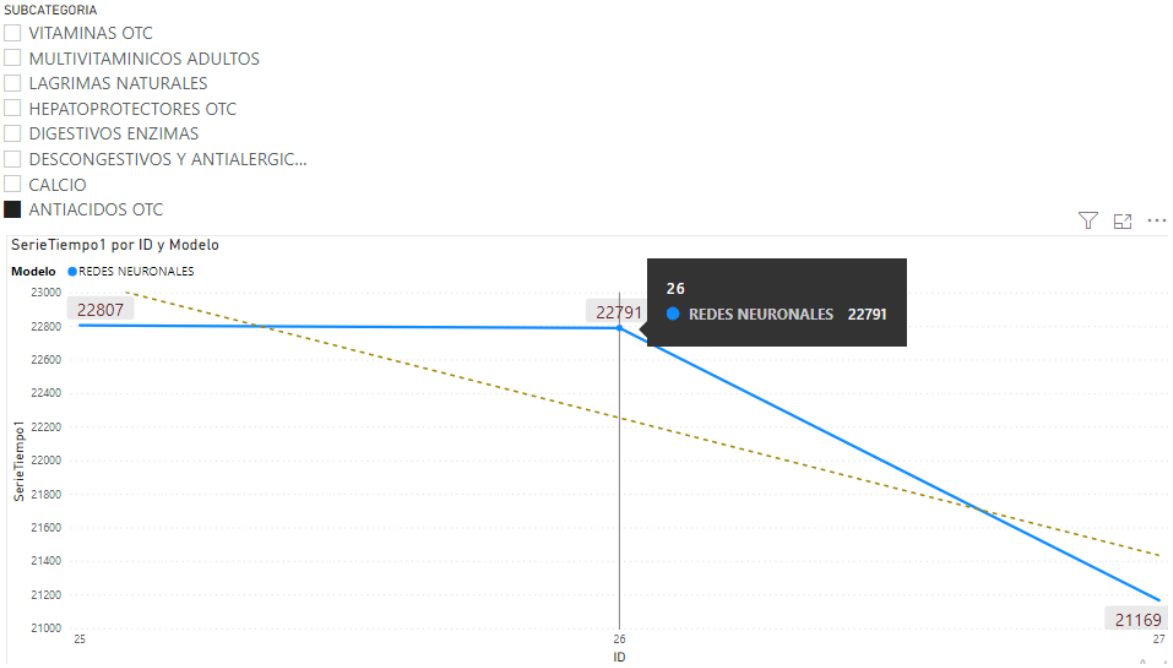
Tabla 8

Forecast Ventas Subcategoría Antiácidos, Modelo Redes Neuronales

SUBCATEGORÍA	SerieTiempo	Tipo	ID
ANTIÁCIDOS OTC	22.807	predicción	25
ANTIÁCIDOS OTC	22.791	predicción	26
ANTIÁCIDOS OTC	21.169	predicción	27

Figura 38

Gráfica forecast ventas, Modelo Redes Neuronales



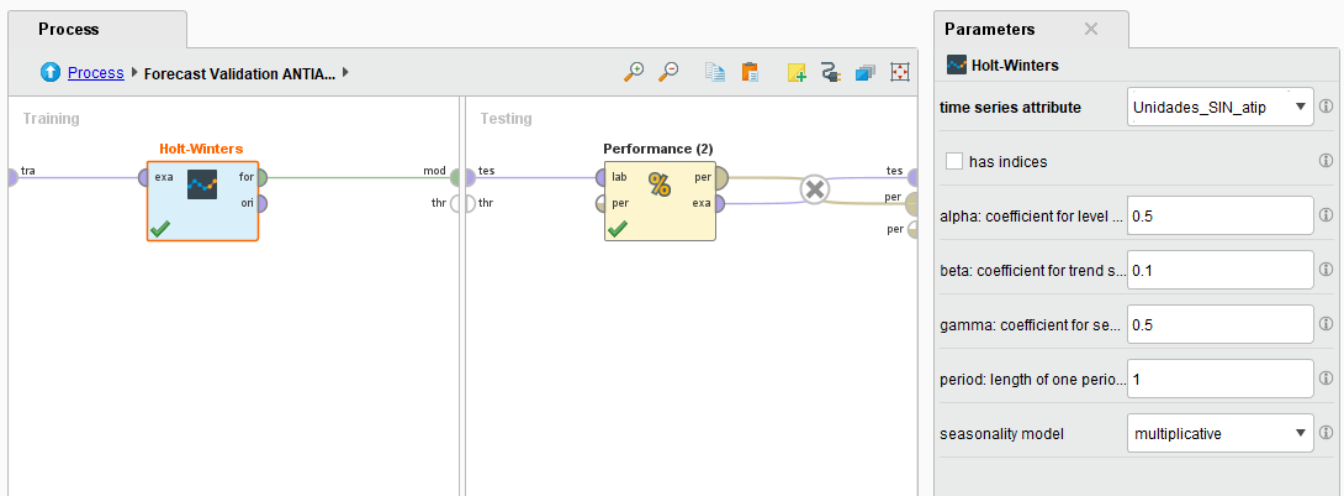
Nota. La figura muestra el forecast de los tres períodos proyectados de las unidades de venta del laboratorio farmacéutico, con el modelo redes neuronales.

Implementación De Un Modelo Holt Winters. El modelo Holt Winters, se aplicó para un suavizado exponencial triple, considerando la tendencia y la estacionalidad de las series de tiempo. Además, se utilizó el modelo multiplicativo estacional, esto supone que a medida que se incrementan los datos también incrementa el patrón estacional.

En la figura 39 se puede apreciar que, la primera fase para la construcción del modelo es parametrizar los componentes: alfa (nivel), beta (suavizado de tendencias) y gamma (coeficiente para el suavizado estacional).

Figura 39

Parámetros Modelo Holt Winters



Nota. La figura muestra los parámetros requeridos para la construcción del modelo Holt Winters en la herramienta Rapidminer.

Forecast Holt-Winters. El entrenamiento de los datos realizado con el modelo Holt Winters para el pronóstico de los siguientes 3 períodos presentó los siguientes datos, los mismos que fueron exportados a través del operador Write Excel.

Tabla 9

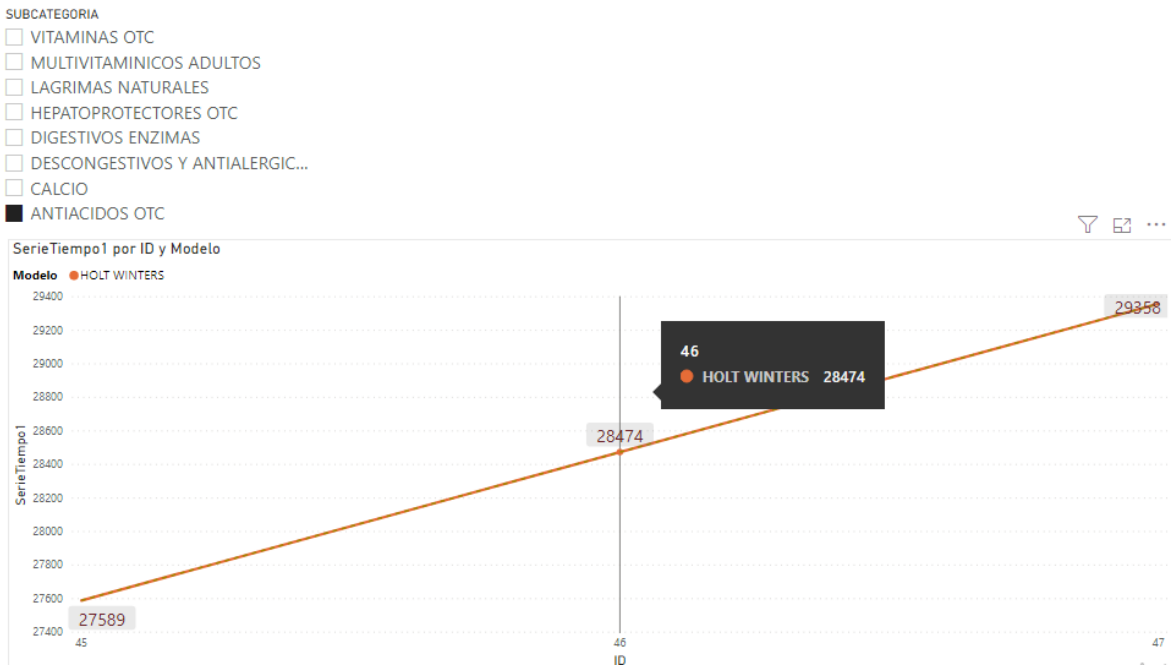
Forecast Ventas Subcategoría Antiácidos, Modelo Holt Winters

SUBCATEGORÍA	SerieTiempo	Tipo	ID
ANTIÁCIDOS OTC	27.589	predicción	45
ANTIÁCIDOS OTC	28.474	predicción	46
ANTIÁCIDOS OTC	29.358	predicción	47

La estructura muestra la subcategoría, las series de tiempo (unidad de venta) que corresponden a la predicción, y el ID (periodo de la serie de tiempo).

Figura 40

Gráfica forecast ventas, Modelo Holt Winters



Nota. La figura muestra el forecast de los tres períodos proyectados de las unidades de venta del laboratorio farmacéutico, con el modelo Holt Winters.

Evaluación de las Métricas de Exactitud de los Modelos. Las métricas permiten evaluar la precisión y performance de los resultados generados. El MAE (error absoluto medio) es la media del error absoluto, el KPI MAPE (error porcentual absoluto medio) contribuye a la medición de la precisión del pronóstico, y es la suma de los errores absolutos individuales divididos por la demanda. Y por último el MASE (error medio absoluto escalado) devuelve los errores elevados absolutos medios.

En primera instancia se consideraron las métricas MAE, MASE y MAPE para la evaluación de los resultados, por lo que se utilizó el operador “Execute” para obtener el performance del modelo ARIMAX.

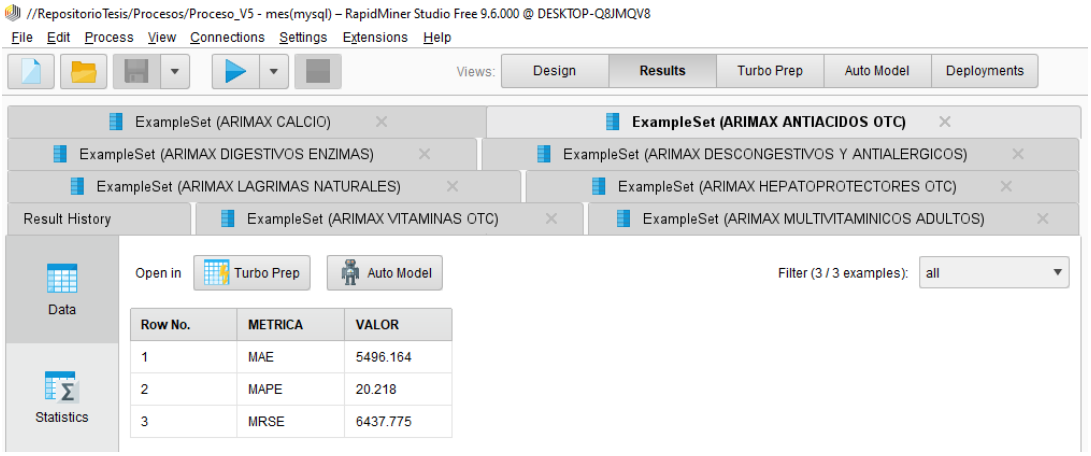
Para el modelo Redes Neuronales se utilizó el operador Cross Validation, que permite una validación cruzada para estimar el rendimiento del modelo, los parámetros a elegidos para el performance fueron: MRS (error cuadrático medio), MAE (error absoluto), MAPE (error absoluto medio).

En el modelo Holt Winters se utilizó el operador Forecast Validation, donde se definieron los mismos parámetros comparables con los demás modelos: MRS (error cuadrático medio), MAE (error absoluto), MAPE (error absoluto medio).

A continuación, las siguientes figuras muestran los resultados de cada performance por subcategoría:

Figura 41

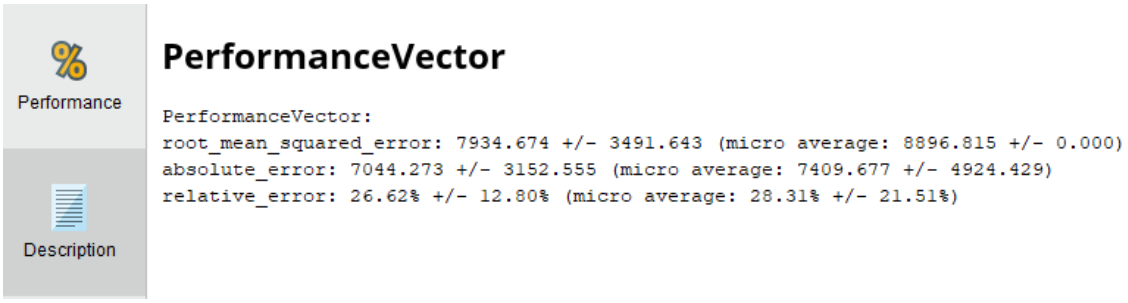
Métricas Modelo ARIMAX



Nota. La figura muestra las métricas utilizadas para el modelo ARIMAX.

Figura 42

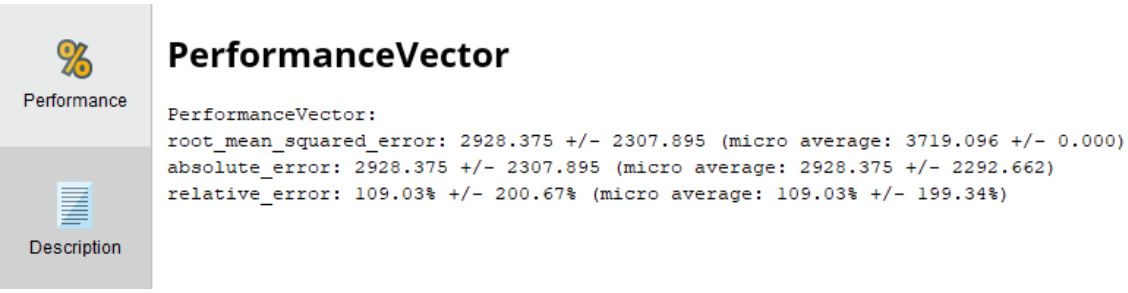
Métricas Modelo Redes Neuronales



Nota. La figura muestra las métricas del modelo redes neuronales, estos valores permitieron comparar los errores del modelo.

Figura 43

Métricas Modelo Holt Winters



Nota. La figura muestra las métricas del modelo Holt Winters, son los valores que permitieron comprar los errores del modelo

Se realizaron varios flujos por subcategoría con distintas agrupaciones de las variables a predecir, con series de tiempo semanales y mensuales. El operador devuelve la evaluación del desempeño del entrenamiento del modelo y con este se determinó automáticamente los cálculos, para posteriormente realizar la validación de cada modelo.

Fase V evaluación e interpretación

En la etapa de interpretación, se consolidó el conocimiento descubierto para realizar el análisis de forecast por producto de la línea OTC. Es decir que se recopilaron todos los datos predictivos de los modelos por subcategoría para realizar un análisis comparativo de los resultados en base al comportamiento de cada subcategoría. Esta revisión se muestra más adelante en el forecast del laboratorio.

De igual forma, la fase de evaluación e interpretación se lo desarrollará en la siguiente sección del trabajo denominado “Pruebas y validación del modelo”

Pruebas y Validación del Modelo

La medición y precisión de los pronósticos fueron realizados en base al análisis de las métricas obtenidas de cada modelo y también a la comparación de la venta realizada en el periodo

proyectado, desde la fecha 01 de enero 2018 hasta el 12 de diciembre 2019, pronosticando enero-marzo 2020.

En este sentido, la precisión mide la diferencia que tiene los valores pronosticados y el valor real, por lo que también se obtuvo un panorama de la magnitud del error, donde se aplicó el pronóstico menos la demanda real de los productos de la línea OTC.

Métricas de evaluación de modelos

Como se mencionó anteriormente, la métrica utilizada que permitió comparar los modelos aplicados fue: el MAPE (error porcentual absoluto medio), "este KPI es el más utilizado para medir la precisión del pronóstico, el cual suma los errores absolutos divididos por la demanda" (Vandeput, 2019); esta métrica fue útil y de relevancia para poder comparar cada modelo predictivo y determinar la precisión de cada forecast. La métrica fue definida además por los responsables de la ejecución del forecast del Laboratorio Farmacéutico. La elección del indicador se realizó conjuntamente con el área de planificación para estimar la tendencia en términos porcentuales y comprender de una mejor manera la métrica del modelo.

Evaluación de resultados de los modelos

Para la evaluación de los modelos predictivos, se evaluó desde la parte técnica, por lo que se consolidaron los datos de las métricas obtenidas a través de la herramienta Rapidminer, por cada subcategoría, con el objetivo de poder comparar los márgenes de los errores.

Como se observa en la tabla 10, los errores determinaron el análisis del modelo más preciso, por lo que se analizaron los porcentajes en base al KPI MAPE (error porcentual absoluto medio). La estacionalidad implicó además la creación de 3 modelos por cada SUBCATEGORÍA, ya que se consideró la estacionalidad por semanas y por mes.

Además, observamos en la tabla 10, que se determinaron los porcentajes como aceptables a comparación con los demás resultados, esto lo definió el Gerente de Planificación de la producción conjuntamente con la Gerente de la línea de productos OTC del Laboratorio Farmacéutico.

Por ejemplo, la subcategoría de ANTIÁCIDOS OTC tiene un MAPE de 13,82%; es decir que los errores porcentuales promedio del pronóstico tienen un error de 13,82% en términos absolutos.

Tabla 10

Análisis MAPE (error porcentual absoluto medio)

SUBCATEGORÍA	MODELO	ESTACIONALIDAD	MAPE
ANTIÁCIDOS OTC	ARIMAX	MES	13,82%
CALCIO	ARIMAX	MES	67,00%
DESCONGESTIVOS Y ANTIALÉRGICOS	REDES NEURONALES	MES	6,45%
DIGESTIVOS ENZIMAS	ARIMAX	MES	357,10%
HEPATOPROTECTORES OTC	REDES NEURONALES	MES	9,78%
LAGRIMAS NATURALES	HOLT WINTERS	MES	17,32%
MULTIVITAMINICOS ADULTOS	HOLT WINTERS	MES	13,21%
VITAMINAS OTC	REDES NEURONALES	MES	5,02%

Los resultados pueden mostrar también, que un modelo no aplica para todas las subcategorías, ya que cada una tiene un comportamiento diferente de la venta mensual de la línea OTC. Dentro del mercado OTC las especialidades farmacéuticas son diferentes, por lo que la medida cuantitativa de la venta depende de las características propias de cada producto y su indicación terapéutica, que pertenece a una subcategoría de productos y que responden a diferentes patologías y necesidades de mercado. Por esta razón cada subcategoría tiene un comportamiento distinto de demanda y por ende de la venta, esto permitió realizar el análisis de cada modelo, al cual se ajusta a cada una de las subcategorías.

Validación del Modelo con el Negocio

En base a la necesidad del negocio se realizó la validación de los modelos seleccionados y entrenados, comparando la proyección con la venta real de los períodos pronosticados, considerando la estacionalidad de semanas y meses para el entrenamiento de más modelos.

Criterios de aceptación del modelo con el negocio.

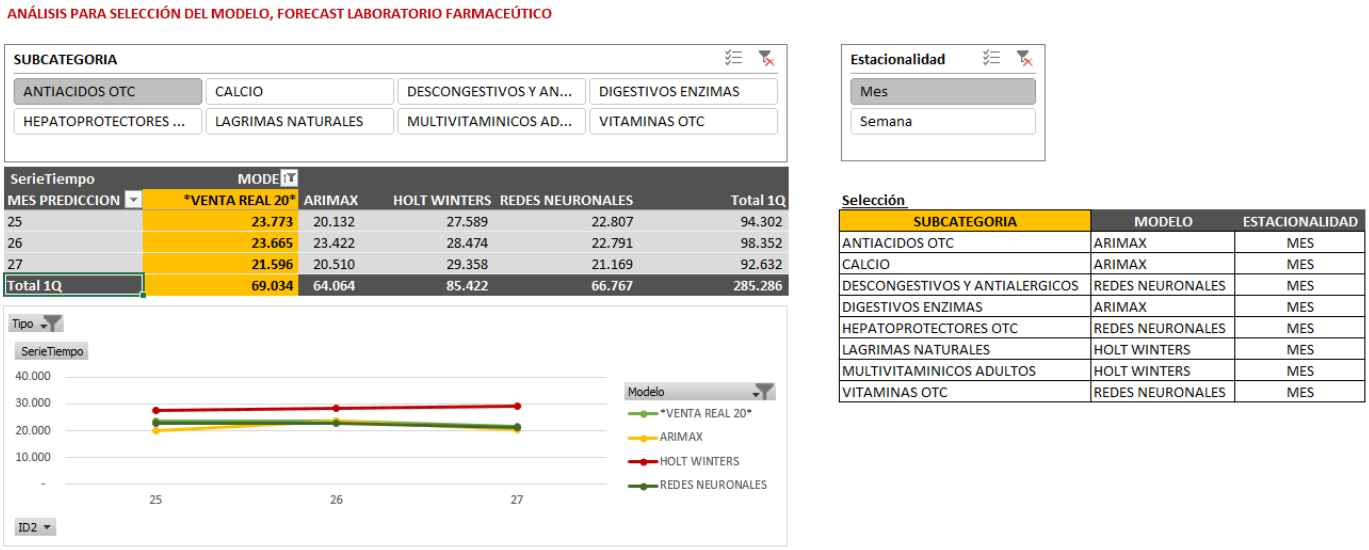
De acuerdo a una entrevista previamente realizada a la Gerente de la Línea OTC y al Gerente de Planificación del Laboratorio Farmacéutico, la planificación de la demanda o pronóstico debe estar basada en los datos históricos, precisión del modelo predictivo y datos actuales del negocio, que permitan minimizar los errores que se generan por la falta de stock, que impiden el cumplimiento de los objetivos comerciales y la disminución de gasto del producto que envían a destrucción por el sobre stock en bodega.

Proceso de validación del modelo con el negocio.

En esta etapa se realizó la validación de los modelos en base a un análisis de selección comparado con la venta real del mismo periodo. Se consideraron las subcategorías y validación por cada modelo entrenado. En el Anexo 1. se demuestra de forma cuantitativa, el análisis de cada modelo por subcategoría y el error a comparación de la venta real.

Figura 44

Tablero para análisis selección del modelo



Nota. La figura muestra el tablero utilizado para el análisis de los modelos ARIMAX, redes neuronales y Holt Winters, para la selección de modelo por la estacionalidad mes y semana.

De acuerdo con la comparación del pronóstico con la venta real se determinó los porcentajes de error y la precisión de cada modelo seleccionado.

Tabla 11

Forecast Laboratorio Farmacéutico Mensual

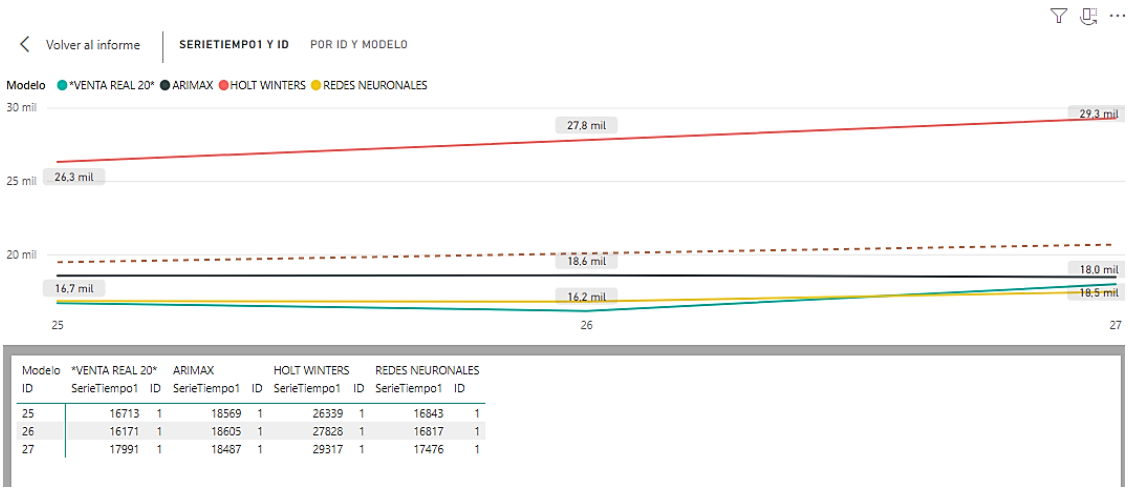
SUBCATEGORÍA	MODELO	01	02	03	1Q
ANTIÁCIDOS OTC	ARIMAX	20.132	23.422	20.510	64.064
CALCIO	REDES NEURONALES	13.479	13.458	14.596	41.533
DESCONGESTIVOS Y ANTIALÉRGICOS	REDES NEURONALES	16.843	16.817	17.476	51.136
DIGESTIVOS ENZIMAS	ARIMAX	2.984	3.253	3.568	9.804
HEPATOPROTECTORES OTC	ARIMAX	50.363	50.192	50.708	151.263
LAGRIMAS NATURALES	HOLT WINTERS	28.670	30.511	32.352	91.533
MULTIVITAMINICOS ADULTOS	HOLT WINTERS	39.469	41.069	42.669	123.206

VITAMINAS OTC	ARIMAX	15.319	15.279	16.984	47.581

Nota. La tabla muestra el forecast para los tres períodos proyectados, por subcategoría de acuerdo con el modelo seleccionado

Figura 45

Forecast Descongestivos y Antialérgicos

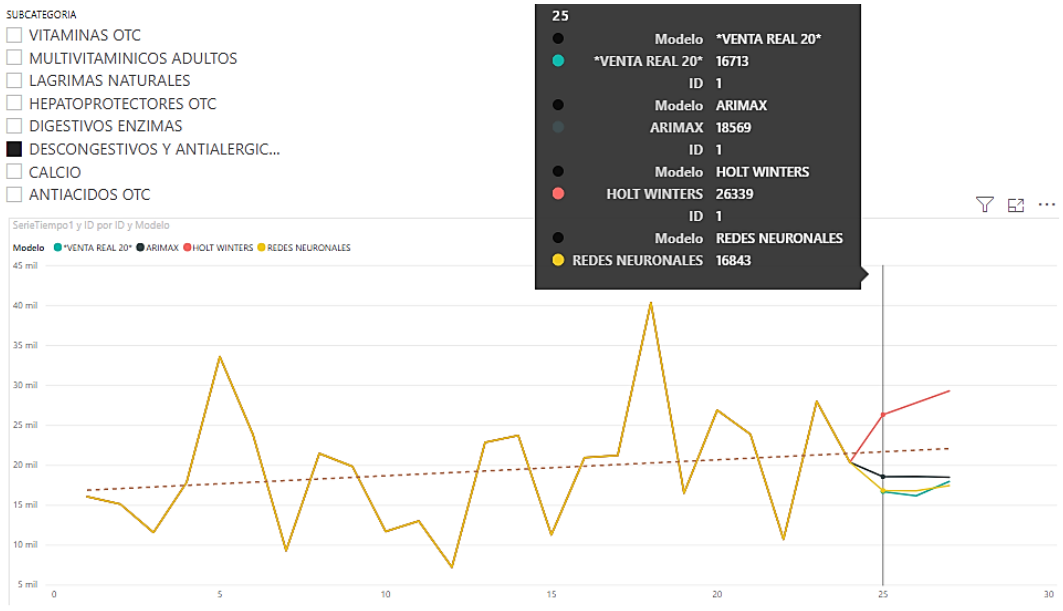


Modelo ID	*VENTA REAL 20* SerieTiempo1	ID	ARIMAX SerieTiempo1	ID	HOLT WINTERS SerieTiempo1	ID	REDES NEURONALES SerieTiempo1	ID
25	16713	1	18569	1	26339	1	16843	1
26	16171	1	18605	1	27828	1	16817	1
27	17991	1	18487	1	29317	1	17476	1

Nota. La figura muestra el forecast obtenido para la subcategoría descongestivos y antialérgicos, utilizando los tres modelos predictivos: ARIMAX, redes neuronales y Holt Winters

Figura 46

Tendencia de Ventas Descongestivos y Antialérgicos



En la figura 45, se muestra el resultado del pronóstico realizado con los modelos ARIMAX, REDES NEURONALES y HOLT WINTERS, para la subcategoría descongestivos y antialérgicos. Se puede observar que la venta real del mismo periodo se ajusta con el modelo de redes neuronales y mantienen una misma tendencia, en el trimestre pronosticado se obtuvo 51,136 unidades proyectadas y en la venta real se tuvo 50,875 unidades vendidas. Por otro lado, el modelo Holt Winters tiene una tendencia diferente a la venta real con una precisión en promedio de -64,21% y el modelo Arimax -9,64%. (Ver Anexo 1)

En la figura 46 se visualiza la tendencia de cada modelo predictivo con respecto a la venta real. El periodo 1-24 presenta la venta del Laboratorio Farmacéutico, la proyección para el periodo 25-27 demuestra la comparación entre modelos y el comportamiento de ventas proyectadas para cada uno. La asociación entre los valores en unidades de las ventas reales y los proyectados con el modelo de redes neuronales sintetizan el análisis de selección del modelo, dando lugar a proyectar planes futuros para la línea de negocio, de acuerdo con el área de planificación y comercial.

Tabla 12

Distribución de forecast porcentual por producto

SUBCATEGORÍA	PRODUCTO	% PESO DE VENTA	01	02	03	1Q
ANTIÁCIDOS OTC	SKU1	42%	8.456	9.837	8.614	26.907
ANTIÁCIDOS OTC	SKU2	58%	11.677	13.585	11.896	37.157
CALCIO	SKU3	100%	13.479	13.458	14.596	41.533
DESCONGESTIVOS Y ANTIALÉRGICOS	SKU4	8%	1.347	1.345	1.398	4.091
DESCONGESTIVOS Y ANTIALÉRGICOS	SKU5	92%	15.496	15.472	16.078	47.045
DIGESTIVOS ENZIMAS	SKU6	100%	2.984	3.253	3.568	9.804
HEPATOPROTECTORES OTC	SKU7	34%	17.123	17.065	17.241	51.430
HEPATOPROTECTORES OTC	SKU8	66%	33.240	33.127	33.467	99.834
LAGRIMAS NATURALES	SKU9	100%	28.670	30.511	32.352	91.533
MULTIVITAMINICOS ADULTOS	SKU10	100%	39.469	41.069	42.669	123.206
VITAMINAS OTC	SKU11	100%	15.319	15.279	16.984	47.581

Nota. La tabla muestra la distribución porcentual del forecast para los productos de las subcategorías, para los tres períodos proyectados de venta en unidades.

Para la validación también se realizó la distribución de forecast en unidades de acuerdo con la participación de venta por producto (SKU), proyectado en el mismo periodo.

Tabla 13

Análisis de la Precisión del Modelo

SUBCATEGORÍA	VENTA REAL 1Q	PRONÓSTICO			%PRECISIÓN DEL MODELO		
		ARIMAX	HOLT WINTERS	REDES NEURONALES	ARIMAX	HOLT WINTERS	REDES NEURONALES
ANTIÁCIDOS OTC	69.034	64.064	85.422	66.767	7%	-24%	3%
CALCIO	45.004	44.549	56.579	41.533	1%	-26%	8%
DESCONGESTIVOS Y ANTIALERGICOS	50.875	55.660	83.485	51.136	-9%	-64%	-1%
DIGESTIVOS ENZIMAS	9.377	9.804	20.470	10.447	-5%	-118%	-11%
HEPATOPROTECTORES OTC	153.642	151.263	161.839	153.914	2%	-5%	0%
LAGRIMAS NATURALES	91.044	31.990	91.533	31.176	65%	-1%	66%
MULTIVITAMINICOS ADULTOS	122.492	91.678	123.206	91.024	25%	-1%	26%
VITAMINAS OTC	46.937	47.581	48.975	47.469	-1%	-4%	-1%

En la tabla 13 se muestra la suma del forecast de los tres períodos pronosticados. Se utilizaron todas las subcategorías y la venta real de los tres períodos proyectados para comparar con los modelos predictivos, y determinar el porcentaje de precisión de cada uno obteniendo así: para ANTIÁCIDOS OTC un 7,20%, en la subcategoría CALCIO se obtuvo 1,01%, DESCONGESTIVOS Y ANTIALÉRGICOS tiene una precisión de -0,51%, DIGESTIVOS Y ENZIMAS tiene un porcentaje de -4,56%, los HEPATOPROTECTORES -0,18% y VITAMINA C -1,13%.

En promedio se obtuvo con el modelo ARIMAX el 1,22% de precisión, el modelo REDES NEURONALES -0,61% y con el modelo HOLT WINTERS -0,56%.

Actualmente el Laboratorio Farmacéutico tiene una precisión en promedio del 15% en la realización del forecast para cada una de las subcategorías de acuerdo con el área de planificación,

Por lo que los resultados generados, ayudan en el proceso de automatización (Ahorro de recursos) y mejora al área de planificación de producción y gestión comercial, con una precisión de forecast menor al 10% de la diferencia entre lo real y el forecast.

Síntesis de las Preguntas de Investigación

OE1: Realizar el estudio del estado actual utilizando la técnica de la entrevista a los responsables de cada área involucrada y el método de la observación sistemática para delimitar las causas y variables del contexto del problema.

OE1- RQ1: ¿El estudio del estado actual a través de la observación delimita las causas y variables del contexto del problema?

Se recopiló información con respecto a indicadores de gestión de las ventas y stock, para el planteamiento del problema delimitando causas y variables que se detallan en el diagrama de Ishikawa.

OE1- RQ2: ¿Se puede determinar la situación actual mediante la técnica de la entrevista?

Se realizó una entrevista a los responsables del departamento comercial y de planificación, quienes corroboraron que actualmente el Laboratorio Farmacéutico no cuenta con un modelo predictivo que les permita pronosticar las ventas para mejorar la toma de decisiones.

OE2: Identificar las técnicas predictivas más adecuadas en la gestión de información para el pronóstico de ventas.

OE2- RQ1: ¿Cuáles son las técnicas más adecuadas para el análisis de la venta de productos farmacéuticos?

De acuerdo con la literatura revisada y análisis realizado las técnicas más adecuadas para el pronóstico de ventas son: la regresión, series temporales y redes neuronales

OE2- RQ2: ¿Las técnicas predictivas permitieron mejorar el análisis de la venta?

Las técnicas utilizadas en el desarrollo permitieron emplear un proceso de exploración y minería de datos a través de la metodología KDD, por lo que mejoró el análisis de la venta.

OE3: Determinar el modelo analítico a través del análisis de patrones de comportamiento, depuración de datos y evaluación de las técnicas predictivas, para resolver el contexto del problema identificado.

OE3- RQ1: ¿Se resolvió el problema planteado a través de un modelo analítico?

Se resolvió el problema planteado, ya que, a través de cada modelo analítico empleado por cada subcategoría, se obtuvo una mejor precisión menor al 10%, de la diferencia entre la venta real y el forecast.

OE3- RQ2: ¿la determinación de patrones de comportamiento, la depuración de datos, y la evaluación de las técnicas predictivas definió el mejor modelo analítico?

A través de la aplicación de la metodología KDD, se construyeron los modelos analíticos y en base a la validación de análisis de selección del modelo, se definieron los mejores modelos por subcategoría y estacionalidad.

OE4: Evaluar el modelo analítico a través del método descriptivo aplicando el análisis de los resultados, para definir la solución del problema con respecto al sobre stock y desabastecimiento de los productos.

OE4- RQ1: ¿El método descriptivo sirvió para evaluar el modelo analítico a implementarse?

Se realizó la validación del modelo con el negocio utilizando criterios de aceptación, que incluyen datos históricos, precisión del modelo predictivo y datos actuales del negocio, definidos por los responsables del departamento comercial y de planificación, por lo que el método descriptivo definió la evaluación del modelo analítico.

OE4- RQ2: ¿El modelo analítico resolvió el problema con respecto a el sobre stock y desabastecimiento de los productos?

Los modelos analíticos resolvieron el problema, ya que se obtuvo una mejor precisión del pronóstico, esto se evidencia en el análisis de forecast vs la venta real.

OE5: Validar la solución implementada a través de indicadores de precisión como el MAPE (error porcentual absoluto medio) para determinar si el modelo se aprueba o se rechaza.

OE5- RQ1: ¿El indicador MAPE (error porcentual absoluto medio) determinara la aprobación o rechazo de la solución implementada?

A través del indicador MAPE se realizó el análisis del error porcentual absoluto medio, esta métrica definió la aceptación o rechazo de cada modelo

OE5- RQ2: ¿La validación de la solución implementada representó el error mínimo definido por el área comercial del Laboratorio Farmacéutico?

El Laboratorio Farmacéutico tenía una precisión de error del 15% en promedio y los modelos validados con el negocio tuvieron una precisión de error menor al 10%.

Capítulo V

Conclusiones y Recomendaciones

Conclusiones

- El estudio del estado actual permite la recopilación de información con respecto a los indicadores de gestión de las ventas y stock, para el planteamiento del problema delimitando causas y variables que se detallan en el diagrama de Ishikawa.
- A través de la entrevista a los responsables del departamento comercial y de planificación, se pudo corroborar que actualmente el Laboratorio Farmacéutico no cuenta con un modelo predictivo, que les permita pronosticar las ventas para mejorar la toma de decisiones.
- De acuerdo con la literatura revisada y análisis realizado las técnicas más adecuadas para el pronóstico de ventas son: la regresión, series temporales y redes neuronales
- Las técnicas utilizadas en el desarrollo permitieron emplear un proceso de exploración y minería de datos de las ventas.
- Se resolvió el problema planteado ya que, a través de cada modelo analítico empleado por cada subcategoría, se obtuvo una mejor precisión menor al 10% de la diferencia entre la venta real y el forecast.
- A través de la aplicación de la metodología KDD se construyeron los modelos analíticos y en base a la validación de análisis de selección del modelo, se definieron los mejores modelos por subcategoría y estacionalidad.
- El método descriptivo aportó para la validación del modelo con el negocio, utilizando criterios de aceptación que incluyen datos históricos, precisión del modelo predictivo y

datos actuales del negocio, definidos por los responsables del departamento comercial y de planificación.

- Los modelos analíticos construidos resolvieron el problema ya que se obtuvo una mejor precisión, esto se evidencia en el análisis de forecast vs la venta real.
- A través del indicador MAPE se realizó el análisis del error porcentual absoluto medio, esta métrica definió la aceptación o rechazo de cada modelo
- La validación de la solución con el negocio representó una precisión de error del 10% mínimo, definido por el área comercial del Laboratorio Farmacéutico

Recomendaciones

- Se recomienda seguir entrenado los modelos establecidos con nuevos datos para continuar estabilizando los pronósticos de cada subcategoría del negocio
- Al momento de realizar la evaluación de los pronósticos se debe realizar el seguimiento respectivo para la validación de la venta real para medir los errores, con el objetivo de mejorar el modelo
- El Laboratorio Farmacéutico no solo debe suponer los criterios del negocio para la realización del forecast sino también la evaluación de la implementación de las herramientas analíticas para la construcción de los modelos predictivos.
- Se recomienda considerar otras variables externas al momento de entrenar el modelo, como la participación del mercado por cada subcategoría para ajustar los pronósticos a la demanda del mercado farmacéutico.

Bibliografía

ALFE. (MAYO de 2017). *CLUSTER FARMA*.

Amos, G. (2006). *MATLAB.* Barcelona: REVERTÉ.

ARTAL, M. (2007). *DIRECCIÓN DE VENTAS .* MADRID.

Barriga, A. D., & Castillo, C. D. (2018). *INTERPRETACIÓN: UN RETO EN LA INVESTIGACIÓN EDUCATIVA.* México.

BBA API_Market. (s.f.). Obtenido de https://bbvaopen4u.com/es/actualidad/el-ranking-de-las-mejores-soluciones-de-analisis-predictivo-para-empresas

Cheng, C.-H., & Chen, Y.-S. (2007). Fundamental Analysis of Stock Trading Systems using Classification Techniques. *2007 International Conference on Machine Learning and Cybernetics.* Hong Kong: IEEE.

Dicovskyi Luis, P. H. (2006). *Sistema de Analisis Estadistico con SPSS.* Managua.

Fakharudin, A. S., Mohamad, M. A., & Johan, M. U. (2009). Newspaper Vendor Sales Prediction Using Artificial Neural Networks. *2009 International Conference on Education Technology and Computer.* Singapore: IEEE.

Favorita, C. (s.f.). *Corporación Favorita*. Obtenido de Corporación Favorita: http://www.corporacionfavorita.com/acerca-de/quienes-somos/

GRANDA, E. (2003). MERCADO OTC EN EL MOSTRADOR . *ECONOMIA Y SALUD* .

Iñaki, L. (31 de OCTUBRE de 2017). *BAOOS ANALYTICS EVERYWHERE*. Obtenido de https://www.baoss.es/analisis-predictivo-que-es/

JIMENÉZ, C., & MATÍNEZ, A. (2016). *ORGANIZACIÓN DE EQUIPO DE VENTAS .* ESPAÑA.

Ju, C., & Han, M. (2008). Effectiveness of OLAP-Based Sales Analysis in Retail Enterprises. *2008 ISECS International Colloquium on Computing, Communication, Control, and Management.* Guangzhou: IEEE.

JUAN, F. (12 de MAYO de 2015). *API_MARKET*. Obtenido de API_MARKET: https://bbvaopen4u.com/es/actualidad/el-ranking-de-las-mejores-soluciones-de-analisis-predictivo-para-empresas

Kaggle. (2018). Obtenido de Kaggle: https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47582

Kangping, W., & Yanhong, Z. (2010). Frame discussion on a general sales management system. *2010 2nd International Conference on Future Computer and Communication.* Wuha: IEEE.

Koosawad, K., Saguansakdiyotin, N., Palangsantikul, P., Porouhan, P., & Premchaiswadi, W. (2018). Improving Sales Process of an Automotive Company with Fuzzy Miner Techniques. *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE).* Bangkok: IEEE.

Li, X., & Li, X. (2009). DDB and B/S Model-Based Special Steel Sales Management System. *2009 Second International Symposium on Knowledge Acquisition and Modeling.* Wuhan: IEEE.

LIS SOLUTIONS. (2016). *Lis Solution.*

Lorenzo, J. M. (2007). *Estadistica Descriptiva.* Madrid: Clara Ma de la Fuente Roja.

Marcoux, J., & Selouani, S.-A. (2009). A Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry. *2009 WRI World Congress on Computer Science and Information Engineering.* Los Angeles: IEEE.

MARTÍNEZ, A. (2014). *PLANIFICACIÓN Y GESTIÓN DE LA DEMANDA .* ESPAÑA.

Olabe, X. B. (s.f.). *Redes Neuronales y sus Aplicaciones .* Bilbao: UPV-EHU.

ORANGE. (s.f.). BIG DATA. *ORANGE*.

ORDUÑA, F. A. (2004). *MANUAL DE VISITADOR MEDICO .* ESPAÑA.

ORTIZ, E., GALARZA, C., CORNEJO, FERNANDO, & PONCE, J. (2014). ACCESO A MEDICAMENTOS Y SITUACIÓN DEL MERCADO FARMACEUTICO EN EL ECUADOR . *PANAM SALUD PUBLICA* , 57.

PÉREZ, C. (2008). *MINERÍA DE DATOS.* ESPAÑA.

Peter Krensky, P. d. (2020). *Cuadrante mágico para plataformas de ciencia de datos y aprendizaje automático.*

Rapidminer. (2020). Obtenido de https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/deep_learning.html

Raul, A., Patil, A., Raheja, P., & Sawant, R. (2017). Knowledge discovery, analysis and prediction in healthcare using data mining and analytics. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT).* Dehradun: IEEE.

Ribeiro, A., Seruca, I., & Durão, N. (15-18 de Junio de 2016). Sales prediction for a pharmaceutical distribution company: A data mining based approach. *2016 11th Iberian Conference on Information Systems and Technologies (CISTI).* Las Palmas: IEEE.

SANCHÉZ, E. (s.f.). *MANUAL DE JAVA .*

Shahid, S., & Manarvi, I. (2009). A methodology of predicting automotive sales trends through data mining. *2009 International Conference on Computers & Industrial Engineering.* Troyes: IEEE.

Timarán-Pereira, S.R., Hernández-Arteaga, I., Caicedo-Zambrano, J., S., . . . C., J. (2016). El proceso de descubrimiento de conocimiento en bases de datos. Bogotá: Ediciones Universidad Cooperativa de Colombia.

Tsao, Y.-C. (2008). A retail-competition supply chain with promotion effort and sales learning curve. *2008 IEEE International Conference on Service Operations and Logistics, and Informatics.* Beijing: IEEE.

Tudorie, C. R., & Borangiu, T. (2011). Towards great challenge in sales and operation planning. *Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems.* Prague: IEEE.

TUYA, J., RAMOS, I., & DOLADO, J. (2017). *TÉCNICAS CUANTITATIVAS PARA LA GESTIÓN EN LA INGENIERÍA DE SOFTWARE.* ESPAÑA.

Utomo, H. S., Sayyidati, R., & Rahmanto, O. (2018). Implementation of mobile-based monitoring sales system in Semi Tani Shop. *2017 International Conference on Sustainable Information Engineering and Technology (SIET).* Malang: IEEE.

Vandeput, N. (5 de julio de 2019). *Analytics Vidhia*. Obtenido de https://medium.com/analytics-vidhya/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Velić, M., Padavić, I., & Lovrić, Z. (2012). Model of the new sales planning optimization and sales force deployment ERP business intelligence module for direct sales of the products and services with temporal characteristics. *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces.* Cavtat: IEEE.

Wang, J., Lee, N., & Timothy, A. (2010). Sales Resource Management Training: A Guide to Developing Effective Salespeople. *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering.* Kunming: IEEE.

Wang, S. (2010). Analysis of channel of sales promotion under consignment contract with revenue sharing. *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM).* Harbin: IEEE.

Ye, W., & Zeng, J. (2011). Supply Chain Revenue-sharing Coordination with Sales Effort Effects. *2011 International Conference on Information Management, Innovation Management and Industrial Engineering.* Shenzhen: IEEE.

Yiqun, X., Zhenzhen, H., & Longjun, W. (2008). Sales data management system of chain enterprises based on NFC technology. *2008 2nd International Conference on Anti-counterfeiting, Security and Identification.* Guiyang: IEEE.

Young&Keat. (2004). *ECONOMIA DE EMPRESA.* México: Pearson Educación.

Yuewei, B., Shuangyu, W., & Binchao, L. (2009). Research and Development on Lean Collaborative Software System for Sales Activity Management. *2009 International Forum on Information Technology and Applications.* Chengdu: IEEE.

Zdravkivic, M. (s.f.). *Kaggle*. Obtenido de Kaggle: https://www.kaggle.com/milanzdravkovic/pharma-sales-data-analysis-and-forecasting

Zhang, C., He, W., & Xu, X. (2012). Application of decision support system based on data warehouse in sales management. *2012 International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA).* Sanya: IEEE.

Anexos