PROPUESTA DE MODELO DE PRONÓSTICO DE DEMANDA

Juan David Pino Galvis

Informe de práctica como requisito para optar al título de:

Ingeniero Industrial

Asesor

Luz Marcela Restrepo Tamayo, Magíster (MSc) en Ciencias - Estadística

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Industrial

Medellín, Colombia

2022

Cita	(Pino Galvis, 2022)
Referencia **Estilo APA 7 (2020)**	Pino Galvis, J. D. (2022). *Propuesta de modelo de pronóstico de demanda* [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.

Asesor interno de la universidad de Antioquia: Luz Marcela Restrepo Tamayo
Alimentos cárnicos Zenú: Área monitoreo

Dedicatoria

A mis padres, por su amor, trabajo y sacrificio en todos estos años, gracias a ustedes he logrado llegar hasta aquí y ser la persona que soy. A mi hermano por estar siempre presente, acompañándome y brindándome apoyo moral, que me brindo a lo largo de esta etapa de mi vida. A todos mis amigos que me han apoyado y han hecho que en momentos difíciles me trasmitieron tranquilidad de que las cosas iban a salir de la mejor manera.

Agradecimientos

Agradezco a mi familia por todo el apoyo moral y económico brindado a lo largo de mi proceso académico, a los docentes que me formaron con los conocimientos técnicos para afrontar los retos profesionales, a mis compañeros por aportar a mi crecimiento personal durante toda la universidad.

Tabla de contenido

Lista de tablas

Lista de figuras

Resumen

Una de las aplicaciones actuales de la ciencia de datos y que ha cobrado gran relevancia, es el aprendizaje y posterior predicción del comportamiento de la demanda de productos, que puede estudiarse a partir de diferentes metodologías estadísticas enfocadas en el tiempo. Sin embargo, debido a la variabilidad de la demanda y de las variables que la afectan, es cada vez más complejo en el día de hoy realizar una proyección similar a la realidad.

Considerando lo anterior, fue necesario construir una solución asertiva para mejorar el entendimiento del fenómeno y para el desarrollo de un procedimiento basado en un modelo de pronóstico. En primer lugar, se identificaron las variables que afectan la demanda de los productos de línea, luego, se obtuvo la información pertinente de las variables con sus históricos, posteriormente se organizó y se estructuró adecuadamente esta información, de acuerdo con los lineamientos requeridos para la correcta predicción y con el fin de aumentar la precisión del modelo a la realidad. Posteriormente, se implementó el modelo y se realizaron pruebas de verificación y validación de los resultados para evaluar el rendimiento del modelo con la metodología actual. Todo lo anterior con el apoyo del equipo comercial y de monitoreo de la cadena de suministro de la empresa manufacturera de alimentos cárnicos.

Los resultados del proyecto fueron aceptados por los expertos del negocio realizando una documentación de las actividades y se tomó la decisión de continuar perfeccionando la solución.

Palabras clave: Series de tiempo, Pronóstico de demanda, Algoritmo, Análisis de datos

Abstract

One of the current applications of data science has gained great relevance lately. It is basically the learning and subsequent prediction of the behavior of product demand, which can be studied from different statistical methodologies focused on time. However, due to the variability of the demand and the aspects that affect it, it has become increasingly complex nowadays to make a forecast close to reality.

Therefore, it was necessary to build an assertive solution to improve the understanding of the phenomenon and to develop a solution based on a forecasting model. First of all, the variables that affect the demand of the line products were identified. Then, the pertinent information of the variables with their historical data was obtained. After that, the information was organized and well structured, according to the guidelines required for the correct prediction in order to increase the accuracy of the model to reality. Finally, the model was implemented as well as verification tests and validation of the results were performed to evaluate the execution of the model with the current methodology. All of the above was supported by the company's commercial and supply chain monitoring team.

The results of the project were accepted by the business experts by documenting activities and the decision was made to continue improving the solution.

Keywords: Time series, Demand forecasting, Algorithm, Data analysis

Introducción

Debido al aumento de la información y los datos dentro y fuera de las organizaciones en los últimos años, las empresas han tenido que afrontar el reto de aprovechar esta información para su mejoramiento en diversas áreas, impactando en la manera como toman decisiones en los niveles estratégicos, tácticos y operativos dentro de esta.

Una de las aplicaciones actuales de la ciencia de datos y que ha cobrado gran relevancia, es el aprendizaje y posterior pronóstico del comportamiento de la demanda de productos y servicios, elemento que puede estudiarse a partir de diferentes metodologías estadísticas enfocadas en el tiempo. Sin embargo, debido a la variabilidad de la demanda y de las variables que la afectan, como el precio, la época del año, cantidades disponibles, productos de la competencia, entre otros, es cada vez más complejo realizar una proyección que represente de manera adecuada el comportamiento real. En consecuencia, las empresas tienen la necesidad de desarrollar nuevas herramientas que les permita conocer el comportamiento de la nueva demanda, para generar una producción en la cual no se subestime o sobreestime la misma.

Teniendo en cuenta lo anterior, el presente trabajo se desarrolla con el objetivo de proponer un modelo de pronóstico de ventas, que permita estimar de manera asertiva el comportamiento de las ventas de las referencias otorgadas por la empresa, tomando como punto de partida, la realidad y contexto de una empresa manufacturera de alimentos cárnicos de Panamá.

Para la realización de este estudio se utilizarán métodos de aprendizaje automático con algoritmos supervisados mediante la aplicación de modelos de series de tiempo, siendo esta una metodología pertinente para el problema a abordar y los datos suministrados por la empresa. El desarrollo del modelo se realizará en Rstudio, el cual tiene uso frecuente en aplicaciones de Machine Learning.

Para el desarrollo de la propuesta de modelo de pronóstico, en primer lugar, se identifican las variables que afectan la demanda de los productos. Luego, se obtiene la información pertinente de las variables con sus históricos, se organizan, se limpia y se estructura adecuadamente de

acuerdo con los lineamientos requeridos para la correcta predicción con el fin de aumentar la precisión del modelo a la realidad. Por último, se implementa el modelo y se realizan pruebas de verificación y validación de los resultados con la realidad para evaluar el rendimiento del modelo con la metodología actual. Todo lo anterior con el apoyo del equipo de cadena de suministro en Panamá y monitoreo de la cadena de suministro de la compañía en Colombia.

1. Objetivos

1.1 Objetivo general

Proponer un modelo de pronóstico de ventas que permita estimar de manera más asertiva y próxima a la realidad del comportamiento de las ventas de las referencias de línea en la sucursal de Panamá.

1.2 Objetivos específicos

- Identificar las variables que afectan la demanda de las referencias de línea.
- Obtener, limpiar, estructurar y analizar la información existente de las variables definidas con el histórico correspondiente.
- Evaluar un pronóstico de demanda basado en varios modelos para series de tiempo utilizando machine Learning de pronóstico.
- Validar el resultado del modelo en comparación con lo propuesto por el equipo de Panamá y monitoreo.

2. Marco teórico

- **Machine learning**

Machine learning es una rama de la inteligencia artificial que busca mejorar el análisis de datos por medio de herramientas estadísticas, probabilísticas y de optimización, en pro de una predicción futura, ya sea por la implementación de nuevos sistemas o simplemente el mejoramiento de los ya existentes, mediante el uso de algoritmos basados en información antigua o reciente que permita el funcionamiento óptimo del sistema a trabajar (Cárdenas, 2018).

Existen tres diferentes tipos de machine learning: aprendizaje supervisado, aprendizaje no supervisado y, aprendizaje reforzado. El aprendizaje supervisado comienza típicamente con un conjunto conocido de datos y una cierta comprensión de cómo se clasifican los mismos, en el aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar y por último el aprendizaje de refuerzo difiere de otros tipos de aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo(IBM, n.d.).

- **Pronóstico de demanda**

Basándose en (Cuba et al., 2017) se puede decir que históricamente en el contexto empresarial, los responsables de procesos y la alta dirección, centran gran parte de sus esfuerzos en conocer el estado futuro de sus ventas, demanda e insumos, entre otros. De lo anterior se deriva la vital importancia que presenta la certera realización de pronósticos para la gestión de la cadena de abastecimiento, ya que es una de las premisas para planificar, organizar, implementar y controlar logísticamente un conjunto de actividades o procesos, coordinados para aprovechar los factores productivos de la forma más efectiva posible. Tomando como punto de partida el pronóstico, los analistas puedes determinar variables como la capacidad para satisfacer la demanda pronosticada, así como administrar con antelación el balance de las capacidades con el objetivo de evitar subutilizaciones o cuellos de botella. A diferencia de elegir cualquier método cuantitativo tales como series de tiempo, método causal, o cualitativo (juicio), es importante entender que para una

correcta realización de pronósticos es importante no elegir solo uno ya que al usar y combinar varios, sus resultados se pueden complementar para llegar a una meta más acertada.

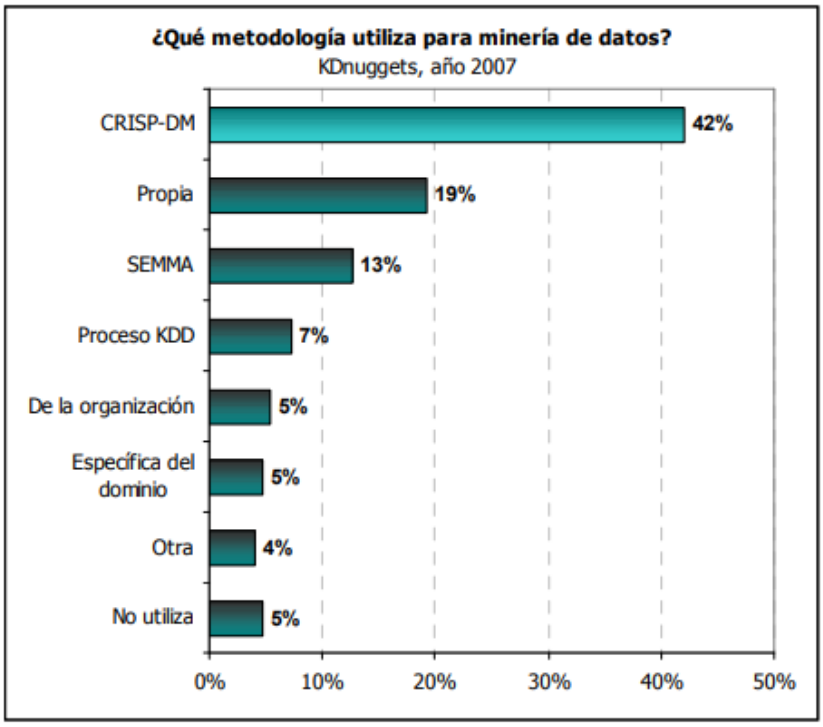
Como se mencionó el pronóstico es muy importante para la gestión de la cadena de abastecimiento, además se debe tener en cuenta que en la demanda de algún bien de consumo depende de algunos factores como (Morales Castro, Ramirez Reyes, & Rodríguez Albor, 2019):

- El crecimiento en el ingreso.
- Cambios en los precios.
- Incremento neto en el crecimiento de la población.
- Cambios en los patrones de consumo (cambios en gustos y preferencias).
- Cambios en la composición de la familia.
- Cambios en la distribución del ingreso.

3. Metodología

Para el desarrollo de la propuesta del modelo de pronóstico se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Es un modelo estándar abierto propuesto en 1999 por IBM para proyectos relacionados con minería de datos (IDECA, n.d.). Esta metodología ha sido la más utilizada por grupos de analítica como se puede evidenciar en la **¡Error! No se encuentra el origen de la referencia.** publicada por la comunidad de KDnuggets (Moine, Haedo, & Gordillo, 2011).

Figura 1
Metodologías usadas para la analítica de datos

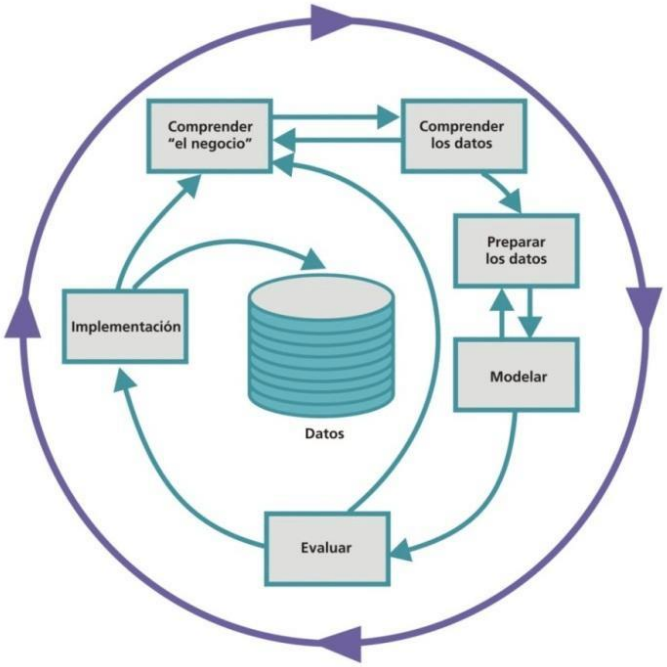


Nota. Fuente (Moine et al., 2011)

Hasta la actualidad, es una de las metodologías más utilizadas. El modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre ellas. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre ellas si es necesario. En la Figura 2 se puede observar las fases de la metodología y las

posibles secuencias entre ellas (Módulo, Diplomado Por Por Elizabeth León Guzmán, León Guzmán, & Profesora, n.d.).

Figura 2
Fases de la metodología CRISP-DM



Nota. Fuente (Módulo et al., n.d.)

Cada una de las fases del modelo se explican a continuación (Wirth, 2000):

Fase1: Es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos no se podrá garantizar resultados fiables.

Fase 2: Comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis, principalmente se realizan actividades como recolección de datos iniciales, descripción de los datos, exploración de datos y, verificación de la calidad de los datos.

Fase 3: En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de procesamiento de estos que se utilizarán posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de estos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Fase 4: Se seleccionan las técnicas de modelado más apropiadas para el proyecto de procesamiento de datos específico. Las técnicas para utilizar en esta fase deben estar alineadas al caso de uso, como también en el problema que se desea resolver, la naturaleza y tipología de las variables y los datos, además del conocimiento y tiempo que requiere el analista para desarrollar el modelo.

Fase 5: En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse, además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error.

Fase 6: Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, es decir, documentar y comunicar de alguna forma (diagrama de flujo, infografía, etc.) para permitir que cualquier usuario que utilice el algoritmo lo realice de forma asertiva.

De acuerdo con la metodología anterior que se va a desarrollar para el cumplimiento de los objetivos, se propone el siguiente desarrollo del informe.

- Entendimiento del negocio
- Entendimiento de los datos
- Preparación de los datos / Modelamiento

- Evaluación
- Implementación

4. Resultados

4.1 Entendimiento del negocio

La empresa desde hace varios años está incursionando en el mercado de Panamá. El equipo comercial define las cantidades y productos que se ofrecen según el histórico de ventas que se tiene de cada una de las referencias con modelos basados en promedios que, en ocasiones, no identifican algunos componentes de la serie de tiempo. La necesidad del negocio radica en que algunas veces, el pronóstico no sigue el comportamiento de la serie de tiempo, lo que genera un resultado poco acertado y, como consecuencia, ocasiona faltantes o sobrantes de productos.

Considerando lo anterior, la empresa junto con el departamento de planeación de la cadena de suministro, se han fijado el objetivo de implementar un proyecto aprovechando las nuevas tecnologías como la inteligencia artificial para obtener información significativa y valiosa de los datos, que permitan estudiar el fenómeno de pronóstico de la demanda de manera más asertiva y así mejorar los procesos e indicadores.

Teniendo en cuenta la dimensión de la necesidad a resolver y el tiempo estimado en el desarrollo del proyecto, el equipo de Panamá y de monitoreo, se fijan un alcance de un modelo que logre pronosticar las referencias utilizadas en Panamá cada una asociada a su respectivo canal.

4.1.1 Identificación de variables

En primera instancia, para la identificación, definición y tipo de variables, se tomó en consideración variables que, junto a un equipo interdisciplinario de la compañía como variables importantes a considerar por su alta probabilidad de impacto en el pronóstico como también variables descritas en diversos artículos académicos. De acuerdo con lo anterior y teniendo en cuenta la información histórica disponible de ventas y oferta, se realizaron los ajustes pertinentes para organizar la información.

La Tabla 1 presenta las variables identificadas, el tipo y su descripción. Las variables fueron evaluadas y, de acuerdo con su disponibilidad, entendimiento y requerimientos, serán seleccionadas o no como insumo para el modelo.

Tabla 1
Variables analizadas para el modelo

Variable	Tipo	Descripción
Referencia	Cualitativo	Referencia de línea la cual se le realiza el pronóstico.
Canal	Cualitativo	Clasificación donde se encuentran los clientes las cuales son: Grandes cadenas, Autoservicios, Alternativo, Institucional y tradicional.
Semana	Cuantitativo	Semana desde donde se inicia el pronóstico.
Duración	Cuantitativo	Es la duración del pronóstico en el futuro.
Pronóstico	Cuantitativo	Periodos los cuales se realiza la proyección de la demanda según cada modelo de series de tiempo.
Modelo	Cualitativo	Modelo de serie temporal elegido para pronosticar según los parámetros del algoritmo.

De la Tabla 1, las variables Referencia, Canal, Semana y Duración son los insumos para el modelamiento, ya que con esta información se pueden hacer variaciones como desde qué semana se pueden tomar las ventas para pronóstico, qué canal específico se desea pronosticar, cuántos periodos pronosticar y cuáles referencias específicas pronosticar. Por otro lado, las variables Pronóstico y Modelo se usan después de correr el modelo; la primera muestra el pronóstico de cada referencia y, la segunda, qué modelo de serie de tiempo se selecciona para una referencia específica. La inclusión de esta variable permite realizar análisis para mejorar el modelo.

4.2 Entendimiento de los datos

De acuerdo con el entendimiento de la necesidad y a la disponibilidad de la información, se obtuvo la siguiente información de bases de datos de la compañía usando SQL Server principalmente (Tabla 2).

Tabla 2
Estructura de las bases de datos en SQL

CodMat	Canal	Cantidad	SemanaCalendario
1001813	Alternativo	0	2021.04
1001813	Tradicional	419	2016.09

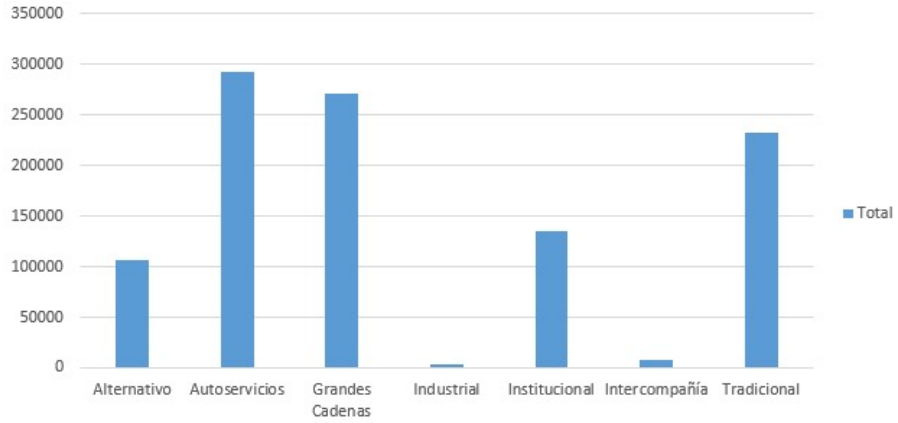
CodMat	Canal	Cantidad	SemanaCalendario
1001981	Autoservicios	543.111	2021.04
1001981	Tradicional	0	2017.02
1002006	Institucional	24	2021.04
1002022	Institucional	0	2021.04
1002057	Institucional	2	2020.03
1002090	Alternativo	0	2021.04
1002090	Institucional	2	2021.04
1003535	Alternativo	0	2021.04
1003535	Tradicional	9	2021.04
1003570	Institucional	0	2021.04
1003570	Tradicional	7	2021.04
1004547	Tradicional	219	2021.04

Como se puede identificar en la tabla, se tiene el historial de ventas de la referencia de línea correspondiente por semana agrupado por tipo de canal. Es importante mencionar que existen registros de ventas desde el 2016 y actualmente existen 528 referencias activas.

En la Figura 3 se pueden evidenciar los siete tipos de canales que existen y la representación en ventas que registra cada uno. Se puede inferir que las cinco categorías principales son Alternativo, Autoservicios, Grandes cadenas, Institucional y Tradicional, ya que representan más frecuencia en las ventas totales.

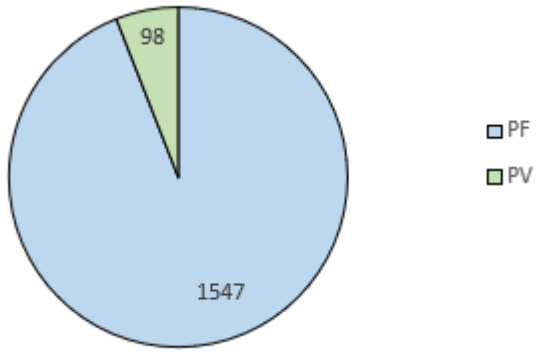
Figura 3

Frecuencias de ventas por canal



También es importante mencionar que las referencias tienen composiciones diferentes y que se diferencian entre peso fijo y peso variable, esto conlleva a que el registro de ventas debe operarse de manera diferente para el peso fijo son ventas por unidad y para peso variable ventas por kilogramos, según la composición que tenga. El comportamiento mencionado se puede evidenciar en la Figura 4.

Figura 4
Tipo de peso por referencia



4.3 Preparación de los datos / Modelamiento

Posterior al entendimiento de los datos, es importante preparar o realizar una limpieza de estos, ya que actualmente las organizaciones toman decisiones, cada vez más, con base en el conocimiento derivado de los datos almacenados. Por tanto, es de vital importancia que los datos contengan la menor cantidad de errores posible, a causa de que es bastante común que una base de datos tenga del 60% al 90% problemas en la calidad de los datos (Huang, Ran, & Blume, 2017).

Aunque idealmente, los datos almacenados no deberían contener errores, es casi inevitable que existan y merecen toda la atención para poder hacer inferencias válidas y apoyar la toma de decisiones acertadas. Basado en lo anterior, se analizaron qué comportamientos erróneos tenían los datos y se identificaron los problemas que podían afectar el pronóstico.

4.3.1 Preparación de los datos

Desde este punto se comenzó a programar todo el algoritmo con la preparación de estos datos en el programa Rstudio. Primero y como se mencionó en el punto anterior, los datos se encuentran en SQL Server, así que para optimizar tiempo y recursos, se utilizó el paquete RODBC de Rstudio para traer ese Dataset directamente y realizar el debido tratamiento, lo que se puede evidenciar en el siguiente código:

```
#crea conexion sql
conn<-odbcDriverConnect('Driver={SQL Server};
                         Server=10.89.xxxx.11;
                         Uid=xxxxx;
                         Pwd=xxxxx+;
                         Database=AcpcsMl;
                         MARS_Connection=Yes;')
```

Como se puede evidenciar en el código anterior, se abre una conexión hacía la base de datos que se llama conn. Esto ahora es un objeto que contiene toda la información para hacer una conexión usando ODBC, incluyendo el tipo de conexión (SQL), la dirección donde está la base de datos (Servidor), el usuario y la contraseña según las credenciales de la empresa y el nombre de la base de datos.

A nivel de tablas se trabaja al hacer búsquedas, entre varias formas que permite RODBC, puede definirse la búsqueda usando el canal (en este caso conn) y la consulta en SQL. En la siguiente línea de código, se muestra la consulta en SQL que sirve para saber qué referencias se ingresarán en el algoritmo de limpieza y pronóstico.

```
sentenciaSqlglobal<-paste("select codmat,canal
                          from (
                              select codmat,canal,sum(case when cant>0 then 1 else 0 end) positivas
                              ,sum(case when cant=0 then 1 else 0 end)  ausentes
                            from   [AcpcsMl].[Ml].[Pronostico_Panama_Vta_Sem_V]
                              where 1=1
                              and semanacalendario<='2021.35'
                              group by  codmat,canal
                          )base
                          where convert(float,ausentes)/(positivas+ausentes)<0.3
                          and positivas>24
                          group by codmat,canal
                          order by 1,2

                        ", sep = "")

dfresumen<-sqlQuery(conn,sentenciaSqlglobal)
```

La anterior consulta tiene la funcionalidad de elegir el conjunto de datos iniciales, la cual cuenta con dos restricciones. La primera consiste en que los datos positivos deben ser mínimo 25 datos (en algunos libros muestra que deben ser mayor a 30 pero, por el comportamiento de los datos, se decidió desde el equipo de monitoreo tener esta medida de 25) y, la segunda, es que los datos ausentes no pueden superar el 30%de los datos totales, ya que esto genera errores en algunos modelos de pronóstico. Para estas restricciones se tiene en cuenta la tabla en SQL ([AcpcsMl].[Ml].[Pronostico_Panama_Vta_Sem_V]) y desde qué semana se quieren tomar las ventas de cada referencia (semanacalendario).

Por último, la función sqlQuery de Rstudio permite generar el data frame denominado dfresumen gracias a la conexión conn y a la consulta sentenciaSqlglobal, que imprime todas las referencias con su respectivo canal que cumplen con las restricciones planteadas como se evidencia en la Figura 5.

Figura 5
Referencias para pronosticar

	codmat	canal
1	1001813	Autoservicios
2	1001813	Grandes Cadenas
3	1001813	Tradicional
4	1001938	Autoservicios
5	1001938	Grandes Cadenas
6	1001938	Tradicional
7	1001981	Autoservicios
8	1001981	Grandes Cadenas
9	1001981	Tradicional
10	1001994	Autoservicios
11	1001994	Grandes Cadenas
12	1001994	Tradicional
13	1002006	Autoservicios
14	1002006	Grandes Cadenas
15	1002006	Tradicional
16	1002022	Autoservicios
17	1002022	Grandes Cadenas
18	1002022	Tradicional
19	1002057	Autoservicios
20	1002057	Grandes Cadenas
21	1002057	Tradicional
22	1002069	Autoservicios
23	1002069	Grandes Cadenas
24	1002069	Tradicional
25	1002081	Autoservicios
26	1002081	Grandes Cadenas
27	1002081	Tradicional
28	1002090	Autoservicios

A continuación, se crea una función encargada de todo el proyecto e ingresa una por una las referencias de la Figura 5. En ella se preparan los datos y se ejecutan los pronósticos. Esta función solicita el material y el canal, el cual es suministrado por el dfresumen, se ingresa la semana en la cual se quieren tener las ventas históricas por referencia y una identificación, la cual registra que día se corrió el algoritmo.

```
fPronosticarMatCanal<- function(material,canal,semana,idEjecucion){
```

Esta función tiene varios procesos para llegar al resultado final. Inicialmente se obtiene el conjunto de datos de ventas históricas según la referencia y el canal con ayuda de una consulta de

SQL. Posteriormente a este, se le realizan los procesos de limpieza y pronóstico que se detallan a continuación.

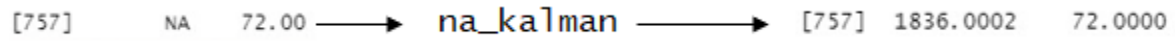
Primero se identificó que en las ventas históricas de línea existen semanas que no hubo ventas, por lo cual ese dato es cero y puede sesgar el comportamiento de los datos en el momento de realizar un pronóstico (Tabla 3).

Tabla 3
Referencias sin ventas

Referencia	Fecha	Canal	Cantidad
10XXX13	2016-10-31	Grandes Cadenas	0
10XXX13	2017-11-27	Grandes Cadenas	0
10XXX13	2018-12-17	Grandes Cadenas	0
10XXX13	2019-04-01	Grandes Cadenas	0

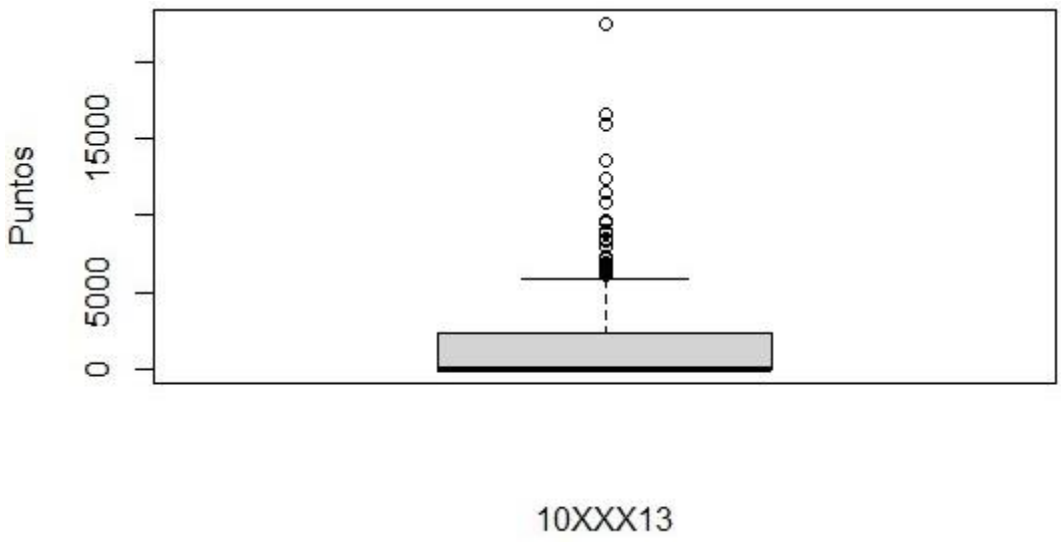
Luego se tomó la decisión en conjunto del equipo de monitoreo de reemplazar estas semanas sin ventas, ya que estos datos son considerados como datos perdidos y pueden tener un efecto significativo en los comportamientos del modelo (Kang, 2013) y para ello se utilizó la función na_kalman de Rstudio para imputar y reemplazar las ventas inexistentes como se puede evidenciar en la Figura 6

Figura 6
Función na_kalman



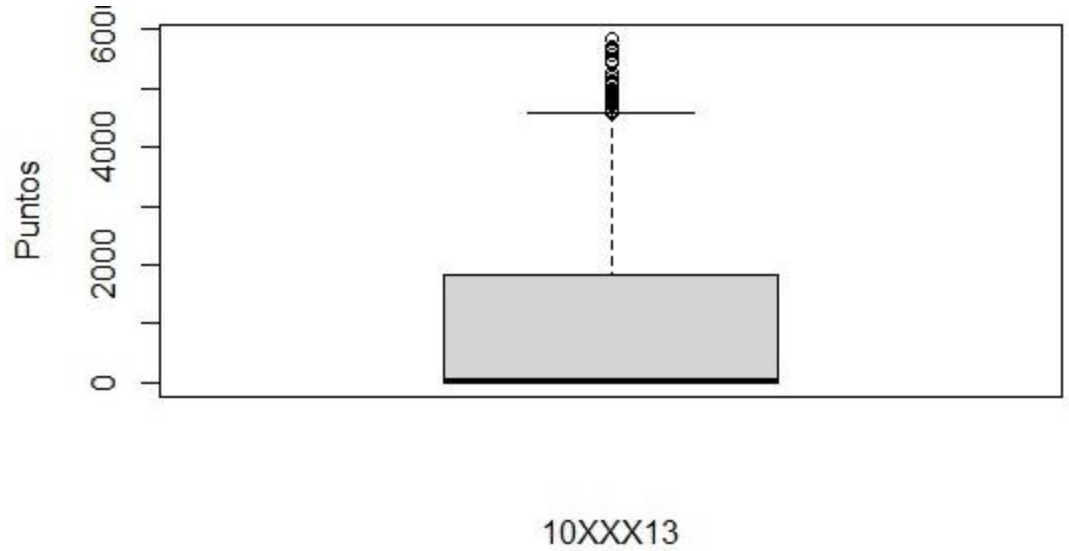
Segundo y luego de reemplazar las ventas en cero, se realizó una identificación para los datos atípicos con un método univariante de detección de outliers, el cual fue el boxplot o diagrama de caja. Este gráfico suministra información sobre los valores máximos y los mínimos, los cuartiles Q1, Q2 o mediana y Q3, además de la existencia de outliers y la simetría de la distribución (MorenoCastellanosJuanGabriel2012). Usando la función de boxplot se identifican los datos atípicos (Figura 7).

Figura 7
Boxplot del conjunto de datos



Después de identificar los datos atípicos del conjunto de datos, se usa la función tsclean de Rstudio para reemplazarlos, y la serie de tiempo quedaría como se aprecia en la Figura 8.

Figura 8
Boxplot datos reemplazados



4.3.2 Generación de pronósticos

Una vez obtenido el DataSet limpio y preparado, se procede a realizar la técnica para modelar y pronosticar las referencias de línea, para ello se utilizan varios modelos de series de tiempo, ya que una serie de tiempo puede tener varios componentes (“Series de Tiempo,” n.d.) y según el o los componentes hay modelos que se suavizan mejor. Por ejemplo: un modelo holt-winters suaviza mejor una serie que tenga componentes de tendencia y estacionalidad (Ramón, n.d.)todo el proceso se evidencia en la Figura 9.

Figura 9

Diagrama de flujo del proceso de pronóstico

Luego de ingresar el Dataset se realizan 9 tipos de pronósticos los cuales son:

- Prophet
- Snaive
- Arima
- Redes neuronales

- Logaritmo (Snaive, Arima, Redes neuronales)
- Regresión lineal
- Holt-winters

Después de realizar los 9 pronósticos, el algoritmo debe decidir cuál es el mejor. Junto con el equipo de monitoreo, se decidió usar la medida estadística del error cuadrático medio (RMSE) siendo de utilidad para comparar el pronóstico de n periodos con las ultimas n ventas (Julián Vélez Correa, 2016) y así elegir el pronóstico más acertado según estos dos conjuntos de datos. Esto se puede evidenciar en las siguientes líneas de código, ya que el dfTest representa las ultimas n ventas seleccionadas para la comparación y el pn es el pronóstico de cada modelo.

```
rmse1<-rmse(dfTest$cant, p1[1:Np,1:1])
rmse2<-rmse(dfTest$cant, p2[1:Np,1:1])
rmse3<-rmse(dfTest$cant, p3[1:Np,1:1])
rmse4<-rmse(dfTest$cant, p4[1:Np,1:1])
rmse5<-rmse(dfTest$cant, p5[1:Np,1:1])
rmse6<-rmse(dfTest$cant, p6[1:Np,1:1])
rmse7<-rmse(dfTest$cant, p7[1:Np,1:1])
rmse8<-rmse(dfTest$cant, p8[1:Np,1:1])
rmse9<-rmse(dfTest$cant, p9[1:Np,1:1])
```

Por último, el modelo que resulta ganador es registrado en la base de datos de SQL de la empresa el cual se ejecuta con el siguiente código:

```
for(i in 1:nrow(dfinsertar)) {

  lcodmat<-dfinsertar[i,1,1]
  lcanal<- dfinsertar[i,2,2]
  lperiodo<-  dfinsertar[i,3,3]
  lpronostico<-  dfinsertar[i,4,4]
  insertarSqlSentencia<-paste("insert into [AcpcsMl].[Ml].[Pronostico_Panama_Pronosticos] ( [IdEjecucion]
                                                        ,[CodMat]
                                                        ,[Canal]
                                                        ,[Periodo]
                                                        ,[Pronostico]
                              ) values('",idEjecucion,"','",lcodmat,"','",lcanal,"',",lperiodo,",",lpronostico,")")

  sqlQuery(conn,insertarSqlSentencia)

}
```

4.4 Evaluación

Según la Tabla 4, se puede evidenciar que se pronosticó una gran cantidad de referencias siendo un 98% de referencias pronosticadas versus las referencias activas, las que no entraron en

el algoritmo significan que no tienen la cantidad necesaria de datos para realizar un pronóstico. También se muestra que el algoritmo tuvo en cuenta cinco canales de siete ya que los otros, al tener tan pocas ventas, no sería muy fiable realizar un pronóstico y, por último, el tiempo de ejecución puede ser bastante largo, por ende, se deben realizar con antelación la corrida que se solicita este tiempo es así de largo ya que el algoritmo tiene muchos datos y utiliza, como se mencionó, varios modelos de pronóstico.

Tabla 4

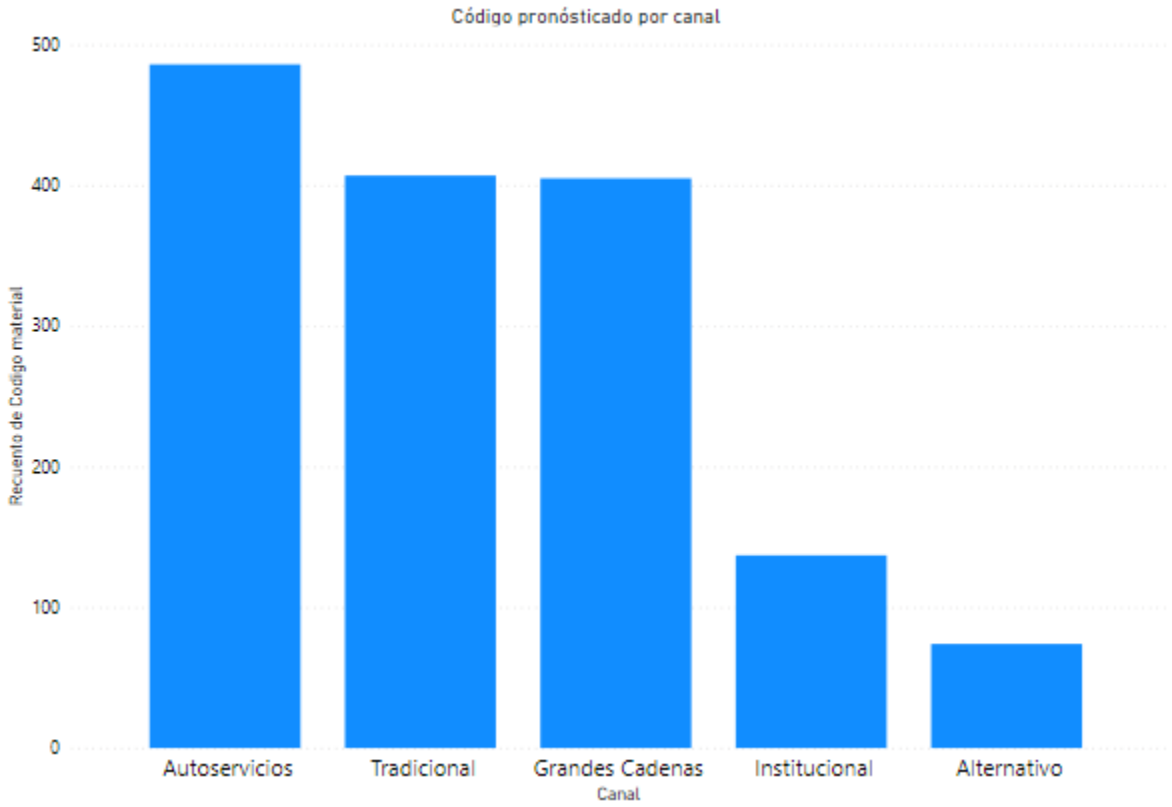
Resumen de resultados

Resultado	Valores
Referencias pronosticadas por canal	1594
Referencias pronosticadas	518
Referencias activas	528
Canales	5
Pronóstico (Semanas)	14
Tiempo de ejecución (Minutos)	6376

En la Figura 10 se puede evidenciar cuáles son los canales que representan más referencias dentro de la organización. Se puede apreciar que los tres canales más representativos en función del número de referencias de línea son Autoservicios, Tradicional y Grandes cadenas, resaltando clientes como Justo y bueno, Super xtra, Pricesmart y Super 99.

Figura 10

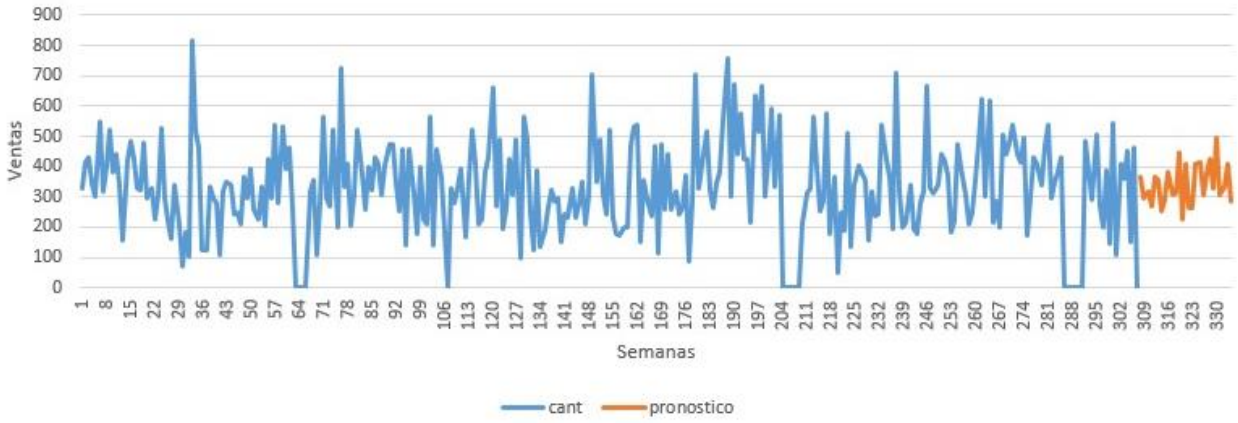
Diagrama de barras de los canales pronosticados



En la **¡Error! La autoreferencia al marcador no es válida.** se puede evidenciar el resultado del pronóstico en comparación con las ventas reales. Se puede notar que el comportamiento de esta referencia en específico es ruido blanco, es decir, tiene un comportamiento de serie temporal estacionaria en media, las cuales son normales en la realidad. En este caso, se espera que el pronóstico no tenga componentes de tendencia ni estacionalidad como se puede evidenciar en la gráfica, además se disminuyó la varianza que tenía la cantidad real para tener un pronóstico más estable que permitiera dejar de subestimar o sobrestimar la producción.

Figura 11

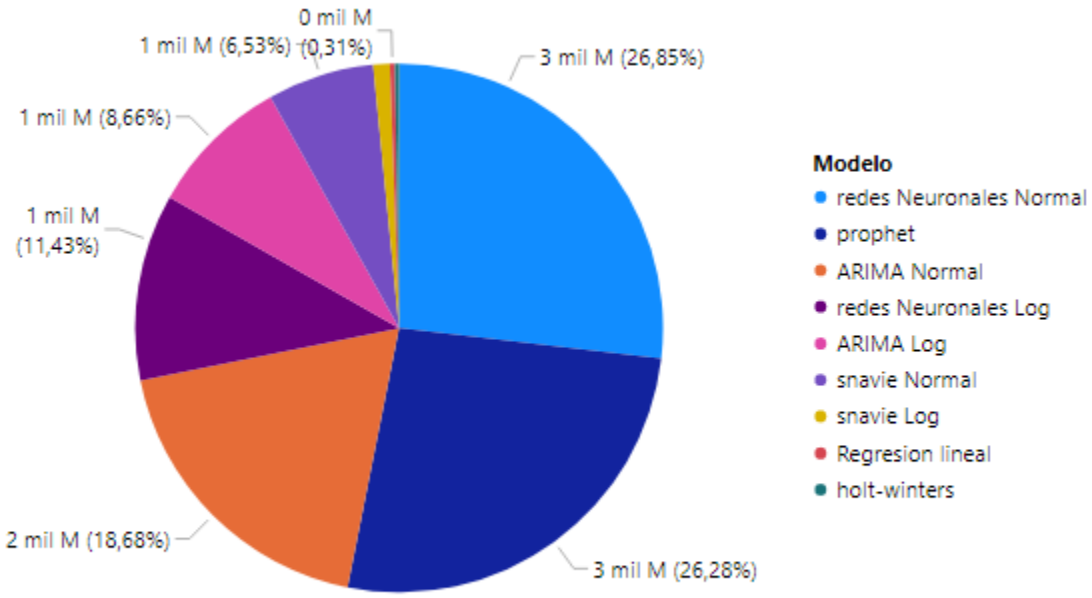
Comparación de las ventas reales contra el pronostico



En la Figura 12 se muestra el porcentaje de modelos usados en una corrida. Se puede evidenciar que los modelos más usados son las redes neuronales, Prophet y Arima ya que representan más del 70% de modelos usados en la corrida. Esto permite optimizar el tiempo de modelamiento.

Figura 12

Diagrama de torta con los modelos de series de tiempo más usados



4.5 Implementación

Desde el equipo de Panamá se analizaron los pronósticos realizados por el modelo en tres corridas y se concluyó que funcionan para predecir comportamientos de las referencias en un futuro a corto plazo, además estos pronósticos ya se están utilizando para la planeación de la demanda cada mes, por ende, se firmaron acuerdos de servicio por parte del equipo de Panamá y del equipo de monitoreo que sirven para que este modelo siga funcionando ya que ha sido de gran ayuda en este proceso.

En la Figura 13 se pueden ver las actividades que debe realizar el equipo de Panamá para la corrida normal del modelo, el cual consta de actualizar referencias que vayan entrando y saliendo del portafolio, ventas de referencias y tipo de peso.

Figura 13
Acuerdos de servicio del equipo en Panamá

Diseño Organizacional
Herramienta: Acuerdos Interproceso

SE RECOMIENDA DILIGENCIAR SOLO LOS ESPACIOS QUE CONSIDEREN LES AGREGA VALOR

Proceso/ Subproceso involucrado: Planeación de demanda			
Cliente	Analista de Monitorio	Proveedor	Analista de Planeación de Demanda
Cargo Responsable	Analista de Monitorio	Cargo Responsable	Analista de Planeación de Demanda
Nombre	Juan David / Esteban Jimenez	Nombre	Jonathan Vega

Acuerdos		
ID	Entregable	Características
1	Actualización de referencias activas	Se realiza actualización de referncias que vayan entrando o saliendo del portafolio cada vez que se den situaciones
2	Actualización de histórico de ventas	Actualizar histórico de ventas mensualmente los 15 de cada mes
3	Actualización de tipo de material	Materiales peso fijo o peso variable

En la Figura 14 se evidencian las actividades que debe realizar el equipo de monitoreo para cumplir con las necesidades que solicitan el equipo de planeación de demanda de Panamá.

Figura 14
Acuerdos de servicio del equipo de monitoreo

Diseño Organizacional
Herramienta: Acuerdos Interproceso

SE RECOMIENDA DILIGENCIAR SOLO LOS ESPACIOS QUE CONSIDEREN LES AGREGA VALOR

Proceso/ Subproceso involucrado: Planeación de demanda			
Cliente	Analista de Planeación de Demanda	Proveedor	Analista de Monitorio
Cargo Responsable	Analista de Planeación de Demanda	Cargo Responsable	Analista de Monitorio
Nombre	Jonathan Vega	Nombre	Juan David Pino / Esteban Jimenez

Acuerdos		
ID	Entregable	Características
1	Corrida de pronósticos de venta	Se envíe el pronóstico de ventas para el mes correspondiente los 25 de cada mes
2	Explosión de referencias	Se envía referencias explosionadas los 11 de cada mes

5. Análisis

Se puede evidenciar que la metodología usada para la realización de todo el algoritmo fue de gran ayuda, ya que induce a realizar un análisis de la organización y del negocio como tal, para así entender todo el sistema de manera objetiva y ejecutar una identificación de las necesidades que tiene dicha organización, además al ser una metodología sin dependencias entre las fases se pueden hacer correcciones en cualquier momento.

Segundo, es importante mencionar el desarrollo de todo el algoritmo apoyado en el programa Rstudio. Fue una herramienta vital ya que fue diseñado para realizar análisis estadísticos y al tener gran conjunto de programas integrados para el manejo de datos, simulaciones, cálculos y realización de gráficos, aportó a la conexión con las bases de datos de la empresa y así llevar a cabo los pronósticos, en cifras esto se puede evidenciar que Rstudio optimizó el tiempo, ya que se pasó de un tiempo de realización de pronósticos de 20160 minutos a 6376.

Tercero, al usar nueve tipos de modelos de series de tiempo diferentes se está asegurando que el algoritmo analice la mayor cantidad de componentes de estas, con el fin de realizar pronósticos más asertivos, ya que si se compara con el método inicial con este solo usaban el cálculo medio móvil para interpretar la demanda.

Cuarto, es importante mencionar que el algoritmo engloba gran cantidad de referencias con un 98% de estas e incluyendo los canales de distribución. Es importante mencionar que las referencias que no ingresaron en el algoritmo no cumplen con la mínima cantidad de datos solicitados, con esto se obtiene un gran avance en cobertura ya que en la sucursal de Panamá solo tenían en cuentas las referencias sin sus respectivos canales.

6. Conclusiones

Las fases de entendimiento del negocio y de los datos, permitieron identificar los comportamientos más importantes para tener en cuenta en el estudio de pronóstico de productos, para así entender que se necesitaban cierto número de registros mínimos en ventas para tener un pronóstico asertivo.

Se logro aplicar técnicas estadísticas de limpieza de datos para los datos iniciales apoyados en el programa Rstudio, con el fin de obtener resultados más asertivos sin pronósticos sobrestimados o subestimados.

Se evaluaron diferentes modelos de series de tiempo en las corridas apoyados en el programa Rstudio, con el fin de que estos analizaran comportamientos que normalmente tienen estos tipos de datos y que así el modelo tuviera varias opciones para elegir y no centrarse en solo una.

La medida estadística error cuadrático medio (RMSE) fue de gran utilidad para elegir los modelos ganadores ya que comparaba los pronósticos con las ventas reales.

El algoritmo puede usarse de forma flexible, actualmente se está usando para pronosticar de manera semanal, pero se puede usar para varias formas según el registro de datos en tiempos regulares ya sea mensual, anual, semestral, entre otros.

Se uso el programa Rstudio para todo el proceso de programación, ya que es un software especializado en estadística y manejo de datos y se puede usar de manera libre.

Los resultados del modelo se compartieron con el equipo en Panamá y se aprobó la utilidad de este para la planeación de demanda y se definieron actividades requeridas para su constante actualización.

7. Recomendaciones

Para darle continuidad a este proyecto, se recomienda una persona que tenga conocimientos en análisis de datos y en diferentes herramientas como Rstudio, Python, entre otras.

Referencias

Cárdenas, J. M. (2018). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD DENNIYE HINESTROZA RAMÍREZ*.

Cuba, O., Lao, Y. ;, Rivas-Méndez, A. ;, Pérez-Pravia, M., Caridad, ;, & Marrero-Delgado, F. (2017). Ciencias Holguín. *Ciencias Holguín*, *23*(1), 1–18. Retrieved from http://www.redalyc.org/articulo.oa?id=181549596004

Huang, J. L., Ran, S., & Blume, B. D. (2017). Understanding training transfer from the adaptive performance perspective. *The Cambridge Handbook of Workplace Training and Employee Development*, *6*(1), 75–97. https://doi.org/10.1017/9781316091067.006

IBM. (n.d.). ¿Qué es Machine Learning? - Colombia | IBM. Retrieved August 13, 2021, from https://www.ibm.com/co-es/analytics/machine-learning

IDECA. (n.d.). Metodología para la Analítica de datos. Retrieved from www.ideca.gov.co

Julián Vélez Correa, P. N. F. (2016). VALIDACIÓN DE MEDIDAS DE EVALUACIÓN PARA EL PRONÓSTICO DE LA TASA DE CAMBIO EN COLOMBIA, 14–16.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/KJAE.2013.64.5.402

Módulo, M., Diplomado Por Por Elizabeth León Guzmán, D., León Guzmán, E., & Profesora, P. (n.d.). *Minería de Datos -Agrupamiento Minería de Datos Minería de Datos*. Retrieved from http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion11_Agrupacion.pdf

Moine, J. Mi., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*, 278–281. Retrieved from http://sedici.unlp.edu.ar/handle/10915/20034

Morales Castro, A., Ramirez Reyes, E., & Rodríguez Albor, G. (2019). Pronóstico de ventas de las empresas del sector alimentos: una aplicación de redes neuronales. *Semestre Económico*, *22*(52), 161–177. https://doi.org/10.22395/seec.v22n52a7

Ramón, J. G. (n.d.). RPubs - Holt-Winters. Retrieved December 19, 2021, from

https://rpubs.com/nanrosvil/283121

Series de Tiempo. (n.d.).

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39.