

# Design and implementation of forecasting models to estimate the beverage demand in convenience stores

José D. Morcillo, Ph.D.<sup>1</sup>, Ana Paula Burgos, Ing.<sup>1</sup>, Martha Garza, Ing.<sup>1</sup>, Stephany Obregón, Ing.<sup>1</sup>, and Ángel Galindo, Ing.<sup>1</sup>

<sup>1</sup>Escuela de Ingeniería y Tecnologías, Universidad de Monterrey, México.

<sup>1</sup>[jose.morcillo@udem.edu.mx](mailto:jose.morcillo@udem.edu.mx), [ana.burgosn@udem.edu](mailto:ana.burgosn@udem.edu), [martha.garzal@udem.edu](mailto:martha.garzal@udem.edu), [stephany.obregon@udem.edu](mailto:stephany.obregon@udem.edu), [angel.galindo@udem.edu](mailto:angel.galindo@udem.edu)

*Abstract— This research was carried out for a Commercial Chain of Convenience Stores in order to increase the potential of its demand forecasting models and thus improve its logistics planning and decision making. Additionally, in 2018, the company noticed a significant increase in the level of its sales during a heat wave that was experienced in the country. For this reason, it was necessary to improve the forecasting models and validate the importance of external variables. Different univariate and multivariate forecasting models were built and validated, and then two were selected to propose the most suitable forecasting method and a final application.*

*Keywords—Demand forecasting, time series, univariate and multivariate models, feature selection, convenience store.*

Digital Object Identifier (DOI):  
<http://dx.doi.org/10.18687/LACCEI2021.1.1.444>  
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

# Diseño e implementación de modelos de pronóstico para estimar la demanda de bebidas en tiendas de conveniencia

José D. Morcillo, Ph.D.<sup>1</sup>, Ana Paula Burgos, Ing.<sup>1</sup>, Martha Garza, Ing.<sup>1</sup>, Stephany Obregón, Ing.<sup>1</sup>, and Ángel Galindo, Ing.<sup>1</sup>

<sup>1</sup>Escuela de Ingeniería y Tecnologías, Universidad de Monterrey, México.

<sup>1</sup>*jose.morcillo@udem.edu.mx, ana.burgosn@udem.edu, martha.garzal@udem.edu, stephany.obregon@udem.edu, angel.galindo@udem.edu*

**Resumen—** Esta investigación se realizó para una Cadena Comercial de Tiendas de Conveniencia con el propósito de incrementar el potencial de sus modelos de pronóstico de demanda, y así mejorar su planeación logística y toma de decisiones. Adicionalmente, en 2018, la empresa notó un aumento significativo en el nivel de sus ventas durante una ola de calor que se vivió en el país. Por este motivo, fue necesario mejorar los modelos de pronóstico y validar la importancia de las variables externas. Se construyeron y validaron diferentes modelos de pronóstico univariados y multivariados, y luego se seleccionaron dos para proponer el método de pronósticos más adecuado y una aplicación final.

**Palabras clave:—** Pronóstico de demanda, series de tiempo, modelos univariados y multivariados, selección de características, tienda de conveniencia.

**Abstract—** This research was carried out for a Commercial Chain of Convenience Stores in order to increase the potential of its demand forecasting models and thus improve its logistics planning and decision making. Additionally, in 2018, the company noticed a significant increase in the level of its sales during a heat wave that was experienced in the country. For this reason, it was necessary to improve the forecasting models and validate the importance of external variables. Different univariate and multivariate forecasting models were built and validated, and then two were selected to propose the most suitable forecasting method and a final application.

**Keywords—** Demand forecasting, time series, univariate and multivariate models, feature selection, convenience store.

## I. INTRODUCCIÓN

Dentro del entorno empresarial, las ventas son esenciales para el desarrollo de las organizaciones y son consideradas como una de las principales fuentes de ingresos. La capacidad de anticiparse al comportamiento de estas es un aspecto que contribuye a una toma de decisiones estratégicas. Por tal razón, la elaboración de pronósticos de demanda es una actividad importante que tiene un impacto directo en los costos y la rentabilidad [20].

Hoy en día, las compañías abordan distintos métodos para la generación de sus pronósticos de demanda. Los métodos tradicionales de pronósticos están basados en series de tiempo, los cuales predicen la demanda futura basada en datos

históricos [10][20]. Una serie de tiempo se define como una secuencia de observaciones ordenadas de manera cronológica medidas en un determinado momento [15]. Los valores que toman las observaciones de una serie de tiempo suelen ser la consecuencia de tres componentes: tendencia, estacionalidad y aleatoriedad [37].

Para realizar un análisis sobre el tipo de modelo a emplear, se utiliza la función de autocorrelación simple (ACF) para determinar los componentes de la serie [37]. Para esta investigación se encontró una prevalencia de tendencia y estacionalidad en los productos bajo estudio, por tanto, se optó por construir modelos que tomen en cuenta estas dos componentes, tales como Holt-Winters, Descomposición Aditiva y Multiplicativa, SARIMA y SARIMAX.

El método de Holt-Winters toma en cuenta tres componentes: nivel, tendencia y estacionalidad [5], para los cuales existe una constante de suavizado, ( $\alpha$ ,  $\beta$ ,  $\gamma$ ), respectivamente, que toma valores entre cero y uno. A medida que sus valores sean más grandes, significan más peso para los datos más recientes y viceversa [35].

Por otro lado, los modelos de descomposición suelen descomponer la serie de tiempo, reensamblando sus componentes para generar pronósticos. Se utiliza el método aditivo cuando existe una variación estacional constante y el multiplicativo cuando existe una variación estacional creciente o decreciente [31].

El modelo de SARIMA(p,d,q)(P, D, Q) modela la dependencia regular, de tendencia y estacional de la serie, respectivamente [37]. Mientras el modelo de SARIMAX sigue esta misma estructura, sin embargo, tiene la habilidad de capturar información sobre variables exógenas que puedan ayudar a entender y pronosticar la variable bajo estudio [24].

De igual manera, en esta investigación se implementó un modelo de Redes Neuronales Artificiales (RNA), ya que, a diferencia de los métodos clásicos, son capaces de predecir series de tiempo lineales y no lineales, y capturar cualquier interrelación en los datos [12].

A pesar de que los métodos de series de tiempo son los más utilizados, una de sus limitaciones principales es el hecho de que no consideran el efecto de variables externas [29]. Por esta razón, se ha descubierto que los algoritmos de aprendizaje de máquina son muy útiles para abordar este tipo de problemas, ya que tienen la habilidad de procesar grandes cantidades de datos y detectar patrones causados por factores externos [20].

Entre los algoritmos más utilizados se encuentran los Árboles de Regresión y Random Forest. El algoritmo de Random Forest se genera a partir de múltiples árboles de regresión individuales que se van entrenando mediante un conjunto de observaciones aleatorias de los datos de entrenamiento por medio de bootstrap y las salidas de todos los árboles se combinan en una salida final conocida como ensamblado [13].

El algoritmo de Gradient Boosting Regression Tree (GBRT) construye árboles de decisión individuales de manera secuencial y la salida de los árboles se combina a lo largo del proceso, ya que cada árbol busca mejorar los errores de los árboles anteriores. Otra forma de implementar este tipo de modelos con mayor rapidez y rendimiento es a través del XGBoost que busca reducir el espacio de búsqueda observando la distribución de las variables de una manera generalizada [34].

Por otra parte, se tienen las máquinas de soporte vectorial que son un tipo de algoritmo de aprendizaje supervisado que construye un hiperplano óptimo en forma de superficie de decisión que busca encontrar el margen máximo que mejor se aproxime a la variable continua bajo estudio [4].

Para obtener resultados precisos y confiables, existen diferentes técnicas como la integración y generalización de modelos para lo cual se pueden implementar métodos de ensamble, que combinan las predicciones del conjunto de diferentes modelos mediante una votación para realizar nuevas predicciones. La precisión que se obtiene con esta combinación generalmente supera la precisión de cada componente individual [7].

Para poder implementar los modelos de aprendizaje de máquina, se necesita evaluar el efecto que tienen distintas variables en las ventas. Mediante un estudio de literatura, se descubrió que la temperatura es una fuente significativa de fluctuaciones en la demanda del consumidor influyendo en alrededor de un 3.4% de variación en ventas [22]. Además, se deben considerar otras variables meteorológicas como la precipitación, duración de luz solar y humedad relativa [26].

Otras variables que tienen incidencia en las ventas son la inclusión de eventos y días festivos [21]. Así mismo, se debe considerar la variación entre días de la semana, haciendo

énfasis los fines de semana y la estación del año, ya que la reactividad del modelo puede diferir [2]. En cuanto a las variables promocionales, se debe considerar el tipo de promoción, su posicionamiento en tienda, la duración de la promoción y el precio del producto [19].

En el caso de las categorías bajo estudio se descubrió que, para el caso de agua, la marca, la presencia en el mercado y la densidad poblacional tienen un impacto directo en las ventas [28]. Mientras que, para el caso de isotónicos, la marca, el sabor, el precio, el contenido en mililitros, la ubicación de la tienda y el clima tienen gran incidencia en las ventas [38].

Otro grupo de variables que se han tomado en cuenta para estudios de demanda, son las variables macroeconómicas, tales como el Índice de Precios al Consumidor y el Índice de Precios al Producto [3], así como la inflación, el poder adquisitivo, la paridad del peso, etc., [11]. Sin embargo, se ha descubierto que no tienen gran influencia en el comportamiento de las ventas.

Es fundamental que las organizaciones tengan noción sobre la relevancia que tiene la construcción de modelos de estimación que mejor se adapten a sus necesidades. A pesar de que hay una gran cantidad de técnicas de pronósticos, existen muchas organizaciones que no consideran esta clase de procedimientos por distintas razones. Algunas de ellas son la falta de conocimiento sobre estos temas y su aplicación, así como la complejidad en cuanto a la selección de las técnicas más favorables [1].

La presente investigación fue realizada en una Cadena de Tiendas de Conveniencia y su objetivo principal fue el desarrollo de un método de pronósticos que mejore la asertividad en los pronósticos de dos categorías específicas de bebidas. Para la realización de este estudio, se comenzó con un análisis profundo de la información de ventas proporcionada por la organización, con el fin de entender su comportamiento a través del tiempo, así como el proceso de generación de pronósticos de la compañía. A partir de lo anterior, se notó que era muy probable que las ventas se vieran influenciadas por otro tipo de variables.

Posteriormente, se exploraron distintos modelos de pronóstico, dividiendo el análisis en dos fases: la Fase I con modelos de series de tiempo (univariados), y la Fase II perteneciente a modelos multivariados utilizando algoritmos de aprendizaje de máquina, para lo cual se llevó a cabo un estudio de literatura identificando la incidencia que tienen distintas variables en el comportamiento de las ventas. De esta manera, se hizo una evaluación de los modelos construidos mediante métricas como MAPE, RMSE y  $R^2$ , para identificar aquellos que fueran más certeros y se ajustaran mejor a las necesidades de la empresa.

## II. DESCRIPCIÓN DEL PROBLEMA

Las tiendas de conveniencia normalmente usan la herramienta RDF (Retail Demand Forecasting) de Oracle para calcular sus pronósticos de venta e iniciar su proceso logístico 42 días antes del lanzamiento de las promociones. En este periodo lo que se hace es acordar los códigos que entrarán en promoción y la dinámica que tendrán. Así mismo, el equipo a cargo de la generación de pronósticos toma en cuenta el comportamiento histórico de las ventas de los productos con promociones similares, la temporada del periodo promocional y los acuerdos con sus proveedores, lo que quiere decir que los resultados arrojados por RDF son manipulados de manera manual según el criterio del equipo a cargo.

En el año 2018, México presenció una ola de calor que provocó un incremento inesperado en el consumo de bebidas [36]. De acuerdo con [8], los analistas estimaron que marcas importantes de bebidas tendrían una expansión de hasta 3% en dicho año por el elevado consumo de estos productos debido a las altas temperaturas. Sin embargo, a pesar del incremento en ventas, esta ola de calor generó una baja asertividad en los pronósticos de las mismas, representando un error de aproximadamente 0.4 millones de productos. Esto se debió a que la herramienta RDF únicamente maneja métodos de series de tiempo, implicando que no se incluyen variables externas, por lo que no pudieron anticipar este suceso.

Un análisis hecho por Actinver detectó que el volumen de ventas de las refresqueras aumenta un 0.5 por ciento por cada dos grados que aumenta la temperatura, tomando en cuenta el promedio regular [8]. Es por esto que la cadena de tiendas de conveniencia está interesada en mejorar su proceso de pronósticos, incluyendo variables externas y buscando un efecto significativo en el desempeño de sus pronósticos.

## III. METODOLOGÍA

Se escogió la metodología CRISP-DM (por sus siglas en inglés Cross Industry Standard Process for Data Mining) para abarcar la problemática, ya que se adapta a cualquier tipo de proyecto de ciencia de datos que se quiera analizar [23], es iterativa y permite utilizar las herramientas que mejor se acoplen a las empresas [27]. Los 6 pasos de la metodología se describen en [9].

## IV. RESULTADOS

### A. Comprensión del Negocio

La cadena de tiendas de conveniencia bajo estudio cuenta con un catálogo de más de 2,500 productos diferentes. Las áreas de Planeación de la Demanda, Abastecimiento y Comercial trabajan en conjunto para generar los pronósticos de venta mediante el uso de la herramienta RDF, la cual emplea diferentes métodos de series de tiempo, optando por el

modelo que arroje el menor error de manera semanal. Un área de oportunidad que presentan sus pronósticos es la exclusión de variables externas, tales como precio, promociones o variables meteorológicas, entre otras, las cuales pueden tener un efecto directo en las ventas y en el incremento de la asertividad de sus pronósticos.

Se recolectaron los registros de ventas diarias del año 2019 de toda la familia de líquidos a nivel nacional y acotando el estudio a las plazas de Tijuana, Puebla y Toluca, tomando en cuenta la mayor contribución a las ventas, baja asertividad en sus pronósticos y rangos de temperatura (alta, templada y baja). Posteriormente, se hizo un análisis por categoría y se seleccionaron un total de 22 artículos pertenecientes a las categorías de agua sin sabor e isotónicos, ya que ambos representan aproximadamente el 60% de las ventas y un 63.5% de los errores totales en sus pronósticos. Al elegir estos productos, se obtuvo el coeficiente de determinación  $R^2$  de los pronósticos del 2019, así como el MAPE, los cuales se muestran en la Tabla 1.

TABLA 1. ANÁLISIS DEL COEFICIENTE DE DETERMINACIÓN  $R^2$  Y MAPE.

COEFICIENTE DE DETERMINACIÓN $R^2$				
Categoría	Mínimo	Máximo	Promedio	Mediana
Agua sin sabor	-4.24	0.88	0.05	0.33
Isotónicos	-8.32	0.65	-0.77	-0.50
MAPE				
	Mínimo	Máximo	Promedio	Mediana
Agua sin sabor	10%	53%	27%	21%
Isotónicos	15%	87%	30%	27%

Como se puede observar en la Tabla 1, el valor promedio del  $R^2$  es muy bajo y el MAPE muy alto en ambas categorías, lo que significa que los errores de los pronósticos son muy grandes en ciertas semanas. Dicho esto, se considera necesario incrementar mejorar estos indicadores para mejorar el desempeño de los pronósticos de venta de las tiendas de conveniencia.

### B. Comprensión de los datos

Después de seleccionar los 22 productos, se hizo una exploración de los datos de ventas diarias del año 2019, obteniendo un total de más de 24,000 datos. Para iniciar con el análisis, se construyeron diagramas de caja a nivel producto/plaza con el propósito de ver la distribución y dispersión de las ventas, así como identificar valores atípicos (ver Fig. 1).

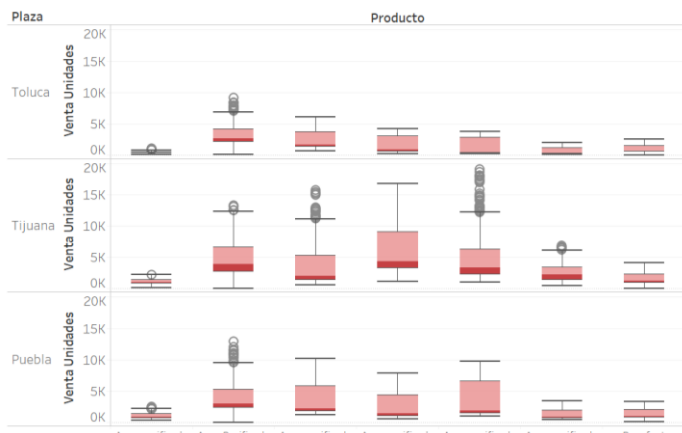


Fig. 1. Diagrama de cajas de la categoría agua sin sabor por plaza.

En la Fig. 1 se puede notar que existe una alta dispersión en los datos a través de la asimetría entre las cajas, las cuales muestran un sesgo positivo. Así mismo, puede verse que la mayor parte de los productos presentan valores atípicos; sin embargo, revisando con la empresa, se determinó que estos datos corresponden a fechas donde se tienen presente una promoción. Este mismo análisis se llevó a cabo para la categoría de Isotónicos, en donde se llegaron a las mismas conclusiones, con una mayor presencia de datos atípicos.

Posteriormente, se graficó el comportamiento de ventas por mes y día de la semana en cada una de las categorías (ver Figs. 2 - 5).

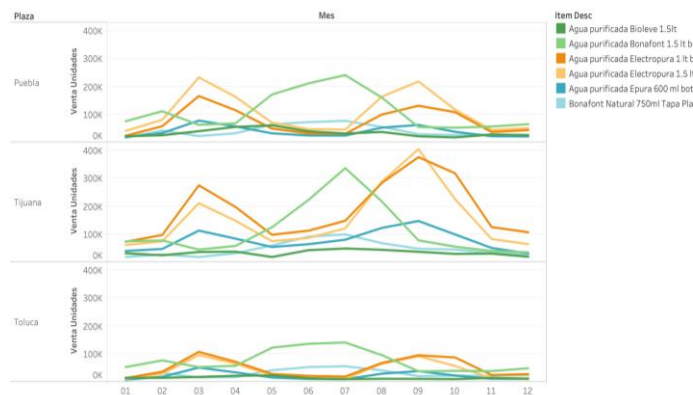


Fig 2. Ventas de la categoría agua sin sabor por mes.

A través de la Fig. 2, se llegó a la conclusión de que los productos de la categoría de agua sin sabor se comportan de la misma manera según su marca, presentando picos de venta durante los mismos periodos, sobre todo en los meses de verano. Estas observaciones fueron discutidas con la empresa, quien agregó que esto podía ser causado por el lanzamiento de promociones, las cuales son precisamente lanzadas por marca.

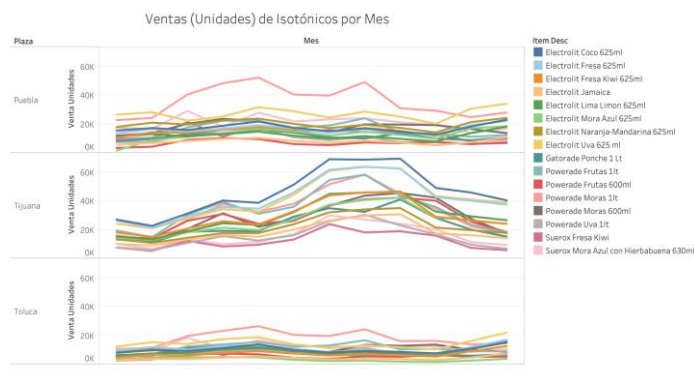


Fig 3. Ventas de la categoría isotónicos por mes.

Este mismo análisis se hizo para la categoría de isotónicos y se encontró que el comportamiento de los productos es similar durante el año para las plazas de Puebla y Toluca (ver Fig. 3). Por otra parte, Tijuana presenta un comportamiento diferente, mostrando un alza en sus ventas durante los meses de Julio a septiembre.

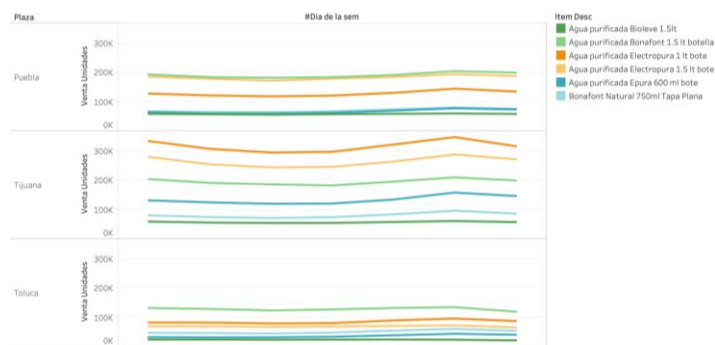


Fig. 4. Ventas de la categoría agua sin sabor por día de la semana.

La Fig. 4 fue de utilidad para concluir que los productos de agua sin sabor suelen tener un comportamiento relativamente constante. En el caso de Puebla y Toluca, se puede notar un incremento mínimo en las ventas de estos productos durante los sábados, y en Tijuana existe un aumento más notable de las ventas en los días lunes y sábado.

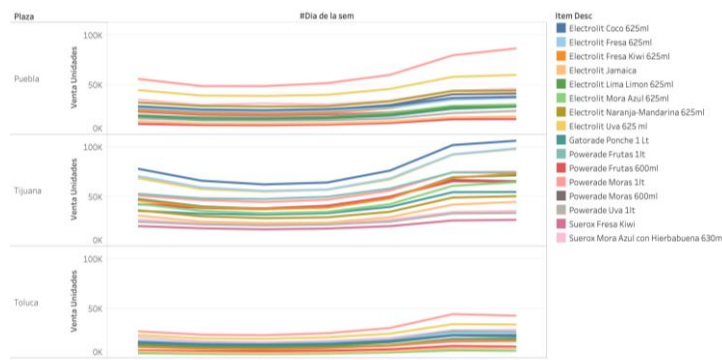


Fig 5. Ventas de la categoría isotónicos por día de la semana.

A diferencia del agua sin sabor, la categoría de isotónicos presenta un incremento notorio los fines de semana para todas las plazas (ver Fig. 5).

### C. Preparación de los datos

Al realizar la limpieza de la base de datos, se detectaron datos faltantes en el mes de marzo, por lo que se decidió aplicar una técnica de imputación determinística para llenar estos espacios con valores plausibles. Esto se realizó mediante la elaboración de diagramas de caja y la obtención de las medidas de tendencia central media y mediana del mes, semana y día de la semana correspondientes, dejando un total de 24,090 registros para los 22 artículos.

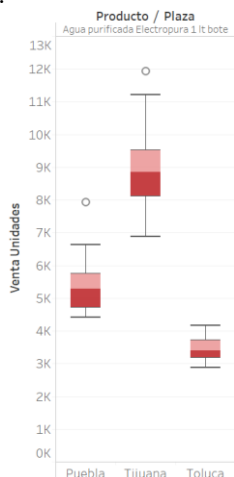


Fig. 6. Box plot de ventas en el mes de marzo del 2019 de Electropura 1 litro.

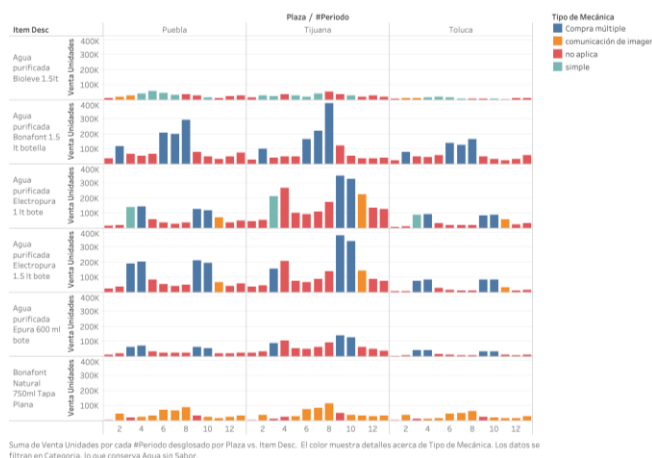
En la Fig. 6 se observa el comportamiento de uno de los productos que presentaba ausencia de datos. En el caso de Puebla y Tijuana se utilizó la mediana debido a la presencia de valores atípicos, mientras que para Toluca se utilizó la media.

Después de haber llevado a cabo el estudio de literatura, se decidió integrar a la base de datos un total de 44 variables distribuidas en cinco segmentos diferentes: variables meteorológicas, de fechas, promociones, características del producto e indicadores globales.

Dentro de las variables consideradas, la compañía proporcionó un consolidado de promociones de cada uno de sus artículos durante el año 2019, el cual contiene información relacionada con el tipo de promoción, el espacio destinado para difundir las promociones y los periodos promocionales, entre otros.

Se decidió analizar también el tipo de mecánica en cada uno de los productos, la cual se refiere a la promoción que adopta un producto, siendo estas: compra múltiple, comunicación de imagen, simple, o no aplica cuando no se tiene promoción. Para esto, se generaron gráficas para observar las ventas de los

productos según el tipo de mecánica que se maneja dentro de sus promociones (ver Figs. 7 y 8).



Suma de Venta Unidades por cada #Periodo desglosado por Plaza vs. Item Desc. El color muestra detalles acerca de Tipo de Mecánica. Los datos se filtran en Categoría, lo que conserva Agua sin Sabor.

Fig. 7. Venta en unidades de agua sin sabor por tipo de mecánica promocional.

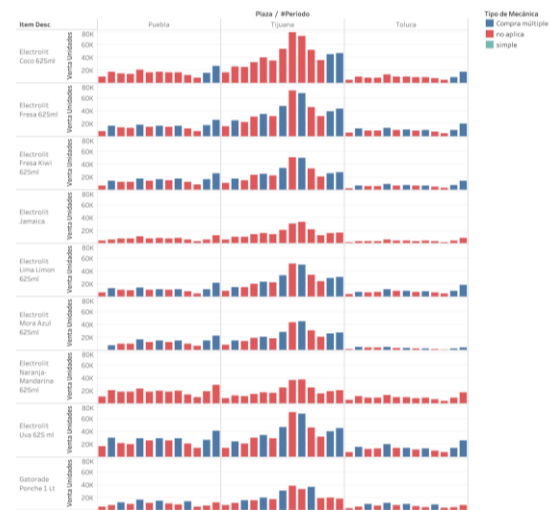


Fig. 8. Venta en unidades de isotónicos por tipo de mecánica promocional.

Para la categoría de agua sin sabor (ver Fig. 7), se puede identificar que durante los periodos donde hay promoción de compra múltiple, las ventas son mayores. Por su parte, en isotónicos hay menos promociones, que además, no necesariamente implican un alza en las ventas (ver Fig. 8).

### D. Modelado

Para seleccionar la técnica de modelado adecuada en cada uno de los artículos bajo estudio, se decidió dividir esta etapa en dos fases. La fase I consistió en llevar a cabo modelos univariados de series de tiempo y la fase II modelos multivariados.

#### Fase I - Modelos Univariados

La fase I se llevó a cabo en RStudio y comenzó con la transformación de los datos a una serie de tiempo con la función *ts*, así como la construcción de correlogramas para determinar los componentes de la serie (ver Figs. 9 y 10).

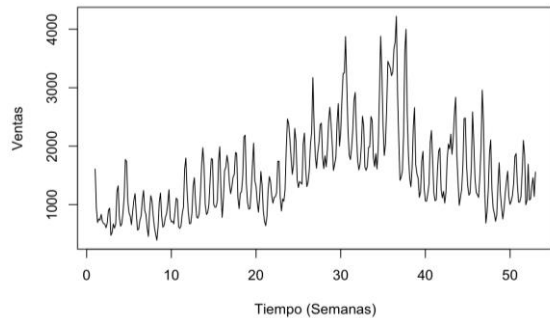


Fig. 9. Serie de tiempo de Electrolit Coco – Tijuana.

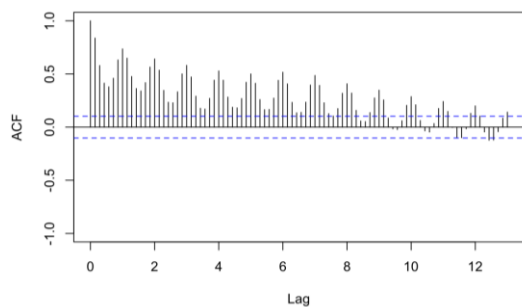


Fig. 10. Correlograma Electrolit Coco – Tijuana.

Al analizar las Figs. 9 y 10 para su correspondiente producto, se descubrió que prevalecen la tendencia y una estacionalidad de 7 días, por lo que se optó por construir modelos que tomen en cuenta estos dos componentes. En estos modelos se incluyeron, *Descomposición Aditiva y Multiplicativa*, *Holt-Winters*, *SARIMA*, *SARIMAX* y *Redes Neuronales Regresivas*.

Los modelos de descomposición aditiva y multiplicativa se llevaron a cabo con la función *decompose*, la cual asigna valores a los parámetros  $\alpha$ ,  $\beta$ ,  $\gamma$  y  $\phi$  con el fin de encontrar la combinación de valores que contribuya al mejor desempeño de los modelos.

El modelo de Holt-Winters, se implementó con la función *HoltWinters*, que busca encontrar los valores óptimos de las constantes de suavizado  $\alpha$ ,  $\beta$ , y  $\gamma$  de manera que se minimice el error cuadrático medio [16].

El modelo *SARIMA* se realizó con la función *auto.arima*, la cual evalúa entre todos los posibles modelos aquel que se ajuste mejor a los datos considerando diferentes criterios [25]. Por otra parte, el modelo *SARIMAX* se llevó a cabo con la misma función, utilizando el argumento *xreg* para poder incorporar variables externas.

Para implementar estos modelos, se llevó a cabo una serie de pasos:

1. Construir el modelo con su respectiva función.
2. Obtener la serie ajustada mediante el argumento *fitted* y compararla contra la serie real.
3. Pronosticar valores futuros con la función *forecast*.
4. Obtener las métricas de error mediante *summary* para comparar modelos e identificar el que tiene un mejor desempeño.

En la Fig. 13 se muestra un ejemplo de la gráfica resultante al comparar el comportamiento del modelo generado con descomposición aditiva y comparada con la serie de datos real.

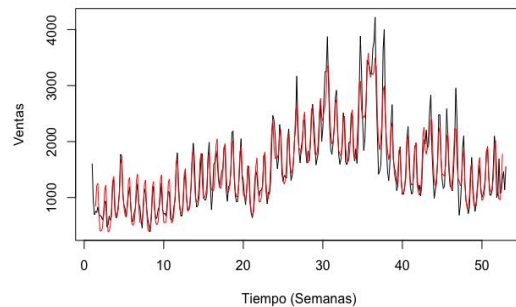


Fig. 13. Serie real vs serie ajustada (D. Aditiva) Electrolit Coco Tijuana.

Para concluir la fase I se probó el modelo de *Redes Neuronales* con la librería *nnetar*, la cual funciona estableciendo parámetros como el número de *lags* estacionales con el parámetro  $P=7$ , y el parámetro de *size*, que determina la cantidad de nodos en la capa oculta.

Se decidió dividir los datos en entrenamiento y prueba, con una proporción de 80% y 20%, respectivamente. Así mismo, se creó un bucle *while*, el cual se puede observar en la Fig. 16, con el propósito de optimizar el valor del parámetro *size*, que va del 1 al 35 y calcula el error para los datos de entrenamiento y prueba, así como la diferencia entre ellos.

```
while (s < 36) {
  set.seed(20)
  sales.nnetar <- nnetar(train.ts, P = 7, size = s, lambda= "auto")
  sales.nnetar
  y_fit <- sales.nnetar$fitted

  #ts.plot(train.ts, y_fit, gpars = list(col=c("black","red")))

  RMSE(train.ts[50:292], y_fit[50:292])
  RMSE_train <- RMSE(train.ts[50:292], y_fit[50:292])

  sales_pred <- forecast(sales.nnetar, h=valid)
  sales_pred$fitted

  RMSE(sales_pred$fitted[220:292], valid.ts[1:73])
  RMSE_test <- RMSE(sales_pred$fitted[220:292], valid.ts[1:73])

  diferencia <- RMSE_test - RMSE_train

  a <- append(a,s)
  b <- append(b,RMSE_train)
  c <- append(c,RMSE_test)
  d <- append(d,diferencia)

  s <- s+1
}
```

Fig. 16. Código del bucle *while* en RStudio para Electrolit Coco – Tijuana.

## Fase II - Modelos Multivariados

La fase II consistió en implementar modelos multivariados mediante algoritmos de aprendizaje de máquina y se llevó a cabo en *Jupyter Notebook* con el lenguaje de programación de *Python*.

Esta fase inició con un estudio de significancia estadística, debido a la presencia de alta correlación entre algunas variables independientes con las ventas. Esto se llevó a cabo con el paquete de *Statsmodels* en *Python*, con el cual se realizó una prueba de hipótesis, una vez eliminada la presencia de multicolinealidad entre las variables independientes.

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.522e+04	1.69e+04	1.492	0.137	-8025.807	5.85e+04
Q("T. MEDIA")	111.9707	8.995	12.449	0.000	94.280	129.661
Q("Unemployment rate")	1227.7235	244.907	5.013	0.000	746.049	1709.398
Q("Population total (K)")	-0.1778	0.100	-1.774	0.077	-0.375	0.019
Q("#Dia de la sem")	124.9703	9.245	13.517	0.000	106.787	143.154
Q("Vacaciones y puentes")	161.4652	58.239	2.772	0.006	46.923	276.008
Q("#dias")	14.6712	7.045	2.083	0.038	0.816	28.527
Q("Velocidad Viento (Máx)")	0.3697	1.227	0.301	0.763	-2.043	2.782
Q("Componentes num")	-21.8524	7.291	-2.997	0.003	-36.192	-7.513
Q("Dia.1")	-2.9216	2.139	-1.366	0.173	-7.129	1.286
Q("Dias festivos")	-76.6680	86.447	-0.887	0.376	-246.690	93.354
Q("Humedad Maxima")	374.9012	163.215	2.297	0.022	53.895	695.907
Q("Humedad Minima")	-908.3291	146.518	-6.199	0.000	-1196.495	-620.163
Q("#Estacion ")	26.0667	20.338	1.282	0.201	-13.934	66.067
Q("Presion Maxima")	-7.6398	7.332	-1.042	0.298	-22.059	6.780

Fig. 18. Resultados de la prueba de hipótesis con *Statsmodels* de *Electrolit Coco - Tijuana*.

En la primera columna de la Fig. 18 se observan las 14 variables ingresadas y en la quinta columna su *p-value*. Para el caso de *Electrolit Coco* en Tijuana, 8 de ellas tuvieron un valor menor a 0.05, lo que asegura una confiabilidad del 95% y las establece como significativas.

Con los resultados obtenidos con *Statsmodels* se implementó el modelo de Regresión Lineal Múltiple, siguiendo el método de *Forward Selection*, empezando con la variable que tuviera el menor *p-value* y mayor correlación con la variable dependiente. Con cada variable ingresada se evaluó si la  $R^2$  ajustada incrementaba y el valor del RMSE disminuía, y de no ser así, esta era descartada.

Posteriormente, se realizó un estudio de Feature Engineering para definir la importancia de variables mediante la construcción de árboles de regresión o random forest, considerando la ganancia de información con la función de *feature importance* en Python [32]. Esta otorga un score de importancia a cada variable dependiendo de las veces que se utiliza para hacer subnodos, por lo que, mientras más se use, mayor es su importancia [6].

La Fig. 17 muestra un ejemplo de los resultados obtenidos para un producto en el modelo de Random Forest.

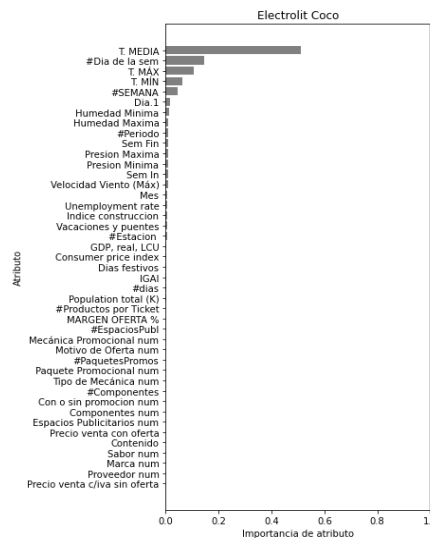


Fig. 17. Feature importance según Random Forest.

En este caso, la variable más importante fue la de temperatura media, lo que quiere decir que fue la variable que más se utilizó para la construcción de los árboles y es la que mejor explica el comportamiento de los datos. En las posiciones subsiguientes a la de esta variable se encuentran algunas otras como temperatura máxima, día de la semana y número de semana, por lo que, al momento de implementar los modelos, se debería probar inicialmente con estas variables, ya que son las que presentan una mayor importancia.

Una vez que se identificaron las variables más importantes para cada producto, se implementaron los modelos de *Random Forest*, *GBRT*, *XGBoost* y *SVR*. Para su implementación y entrenamiento, se llevó a cabo la siguiente serie de pasos:

1. Definir la variable dependiente (Ventas).
2. Definir las variables independientes, siguiendo el método de *Forward Selection* (agregando variables de mayor a menor importancia).
3. Dividir los datos en entrenamiento y prueba, en una proporción de 80-20.
4. Optimizar hiperparámetros con la función *GridSearchCV*.
5. Evaluar el  $R^2$  de los datos de entrenamiento y prueba para detectar presencia de sobreajuste (overfitting).
6. Obtener MAPE y RMSE para comparar los modelos.

Para el modelo de *Random Forest* y *XGBoost*, se ajustaron los parámetros de '*n\_estimators*', los cuales determinan la cantidad de árboles a construir, '*max\_depth*', que determina la máxima profundidad de cada árbol y '*max\_samples*', que mide la cantidad de muestras bootstrap de los datos de entrenamiento que se utilizan para construir cada árbol [30].



En el caso de *GBRT*, se contemplaron algunos otros como *'learning\_rate'*, que controla el nivel de mejora de un árbol respecto al anterior [18], y *'subsamples'*, que ayuda al crecimiento de las ramas de los árboles para mejorar el desempeño de los modelos [17].

Por otra parte, para el método de *SVR* se optimizó el parámetro de penalización *'C'*, que controla la compensación entre los errores de entrenamiento y los márgenes rígidos y el *'kernel'* escogido, que permite transformar a un espacio de mayor dimensionalidad los datos que no se pueden separar linealmente [14].

Para cada uno de estos hiperparámetros se definió un rango de valores y con la función de *GridSearchCV* se evaluaron todas las posibles combinaciones y se obtuvo el mejor modelo. Así mismo, se determinaron otros parámetros como un *cv=5*, que divide los datos de forma aleatoria en 5 grupos y de manera iterativa se va entrenando y probando el modelo [33]. Además, se fijó un *random\_state=0*, que permite obtener el mismo resultado donde sea que se corra el modelo.

Para implementar los diferentes algoritmos se utilizaron las funciones de *RandomForestRegressor* y *GradientBoostingRegressor* de la librería *sklearn.ensemble*; *XGBRegressor* de la librería *xgboost* y *SVR* de *sklearn.svm* [33].

Tras observar que los árboles regresivos y las máquinas de soporte vectorial obtuvieron buenos resultados, se decidió implementar un modelo de ensamble con los 4 modelos previamente mencionados con la función de *VotingRegressor* de la librería *sklearn.ensemble*. Para ejecutar este modelo, se llevaron a cabo los mismos pasos que los otros modelos, sin embargo, se evitó el doble esfuerzo computacional seleccionando aquellas variables que más se repitieron en los modelos anteriores y colocando los parámetros previamente optimizados.

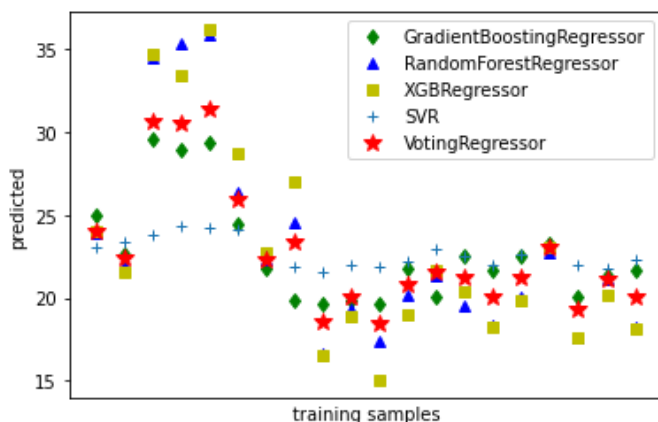


Fig. 18. Generación de pronósticos con el modelo de ensamble.

La Fig. 18 muestra el funcionamiento del modelo de ensamble y se puede observar cómo combina todos los modelos, promediando las predicciones individuales para generar una predicción final.

Por último, se construyó un modelo de Redes Neuronales en el cual se incluyeron todas las variables independientes, se escalaron utilizando la función *StandardScaler* y se dividió un 80% de los datos para utilizarlos como datos de entrenamiento, dejando el 20% restante para los datos de prueba. Posteriormente, se definió una primera versión de la red y se entrenó con algunas capas y un número aleatorio de nodos en cada una. Una vez entrenada se evaluaron las gráficas del error (*MAPE* y *MSE*), como la que se observa en la Fig. 19.

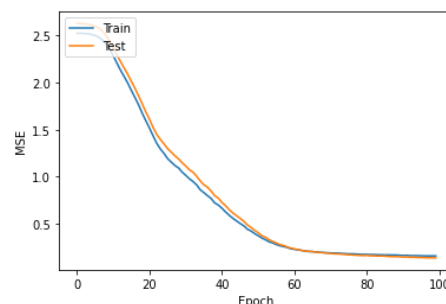


Fig. 19. Comportamiento del MSE según las iteraciones de la red.

La Fig. 19 permite confirmar que no existe un sobreajuste en los datos, debido a que el error se comporta de manera muy similar para los datos de entrenamiento y los de validación. Así mismo, se puede observar que entre más iteraciones haga la red, más disminuye el error. Después de este análisis, se generan los pronósticos de las ventas para el siguiente periodo y se calculó su  $R^2$  ajustada, así como su *MSE* y *RMSE*.

Una complejidad de los modelos de Redes Neuronales, es que hay muchos parámetros que pueden ser optimizados para obtener mejores resultados. En este caso, únicamente se optimizó la cantidad de nodos en cada capa con apoyo de la función *GridSearchCV*, debido a que la optimización de los parámetros de este modelo puede demorar horas. Al encontrar los valores óptimos para la cantidad de nodos, se definió una segunda versión de la red y se entrenó con los nuevos parámetros, se analizó nuevamente la gráfica del MSE, se generaron los pronósticos y se calcularon sus métricas de desempeño.

### E. Evaluación

Con el objetivo de validar los modelos, se debe medir la confiabilidad y exactitud de ellos, al mismo tiempo que se identifican los elementos que presentan un área de oportunidad para modificarlos y mejorar su desempeño. En este estudio se analizaron los resultados a través de la métrica *RMSE* en los

modelos de series de tiempo y en los multivariados, así como el  $R^2$  de los datos de entrenamiento y los de prueba en el caso de los modelos multivariados. Además, se deben buscar los valores óptimos de los parámetros incluidos en los modelos, garantizando los mejores resultados a través de la función *GridSearchCV*.

En primer lugar, se obtuvo un registro del error que presenta cada uno de los modelos univariados, reportando los indicadores del RMSE y del MAPE con apoyo de la función *summary*.

Al ver detenidamente los errores obtenidos por el modelo de Redes Neuronales, así como el ajuste de la línea a los datos, se observó que existía un sobreajuste u *overfitting*. A partir de los resultados del bucle en el modelo, se pudo concluir que entre más nodos existan en la capa oculta, la métrica del error para los datos de entrenamiento disminuye, provocando una diferencia mayor entre los RMSE's, ya que este no disminuye para los datos de validación. Los resultados obtenidos para Electrolit Coco en la plaza de Tijuana se pueden observar en la Fig. 20.

Size	RMSE train	RMSE test	Diferencia
1	270.278398	1121.096	850.8176
2	238.138974	1108.685	870.5458
3	207.971385	1116.864	908.8922
4	166.887411	1117.586	950.6988
5	146.534684	1122.013	975.4782
6	117.380107	1126.153	1008.7733
7	97.462482	1126.482	1029.0200
8	83.471414	1142.043	1058.5711
9	68.508063	1142.624	1074.1164
10	54.206833	1146.255	1092.0486

Fig. 20. Valores del RMSE y su diferencia para Electrolit Coco Tijuana.

Como se puede observar en la Fig. 20, la diferencia entre los RMSE's incrementa conforme lo hace el valor del *size*. Así mismo, se puede concluir a partir de esta tabla que aun teniendo el parámetro de *size* con un valor de 1 (valor que produce el menor valor del RMSE para los datos de entrenamiento), la diferencia entre los RMSE's resulta significativamente alta, lo que confirma el sobreajuste que obtiene este modelo, por lo cual se procede a descartar como un modelo válido para la empresa.

Después de haber descartado este modelo, se llevó a cabo un análisis excluyendo los resultados de Redes Neuronales (ver Tabla A en el apartado de Anexos), en el cual se encontró que el *de Descomposición Aditiva* se desempeñó mejor para el 75% de los productos de isotónicos. En la categoría de agua sin sabor, se encontró que el 66.67% de los productos también mostró un mejor resultado con modelos de *Descomposición Aditiva*. Lo anterior significa que, de manera general, los modelos de *Descomposición Aditiva* son más adecuados para

los productos, logrando mejores resultados en 48 de los 66 SKUs bajo estudio.

En segundo lugar, se realizaron las evaluaciones de los modelos multivariados en donde, una vez finalizados, se decidió hacer el cálculo y la evaluación del promedio de los errores y precisión de cada uno de los tipos de modelos bajo estudio, analizándolos por categoría y plaza

Al obtener la Tabla B (ver Anexos), y comparar el desempeño de cada producto en cada modelo, se observó que el 31.82% de los productos se desempeñó mejor con el modelo de Ensamble. Así mismo, se observó que un 19.70% se adecuó mejor al modelo de Random Forest, otro 19.70% al XGBoost, y el último 28.78% se distribuye en el resto de los modelos.

Después de haber realizado el análisis tanto univariado como multivariado, se decidió hacer una comparación de los resultados del RMSE de los mejores modelos de cada fase con el fin de encontrar el modelo con mejor desempeño. Los resultados de esta comparación se muestran en la Tabla C (ver Anexos). Esta tabla, muestra el promedio del RMSE de los productos pertenecientes a cada una de las categorías en su respectiva plaza. Comparando estos resultados, se observa que la columna resaltada, correspondiente al modelo de descomposición aditiva, tiene el menor resultado de RMSE, indicando que este modelo es el que mejor desempeño tuvo de los 8 que se han realizado.

#### F. Implementación

Para la etapa final de la metodología, se identificaron las variables que fueron utilizadas en los mejores modelos para cada producto, buscando explorar la posibilidad de agrupar la mayor cantidad de productos en un solo modelo. Al hacer esto, se encontró que los productos de las mismas marcas utilizaban las mismas variables. Las conclusiones de este análisis se encuentran en la Tabla D del apartado de Anexos.

A partir de las conclusiones de la Tabla D, se realizaron las agrupaciones de los productos, dejando 3 grupos finales. Esto se realizó con el fin de encontrar el modelo que sea capaz de dar el menor error para la agrupación de productos encontrada por las variables que más se repetían entre sí. En primera instancia se realizó este análisis para la categoría de agua sin sabor. Para el caso de la Plaza de Tijuana los resultados pueden verse en la Tabla E del apartado de Anexos.

Al hacer este mismo análisis para los 3 grupos en cada una de las plazas, se encontró que el 83.3% de los productos de agua sin sabor presentan mejores resultados con el modelo SARIMAX. Así mismo, se encontró que este mismo modelo se ajusta mejor en más del 60% de los productos de Gatorade, Powerade y Suerox. Por último, se encontró que para los

productos de Electrolit, más del 50% se adecúan mejor al modelo Ensamble.

Dicho esto, se construyeron dos modelos de *SARIMAX* en *RStudio*, uno para los productos de agua sin sabor y otro para los productos de isotónicos incluyendo las marcas de Gatorade, Powerade y Suerox. Estos modelos toman en cuenta las variables de temperatura media y temperatura máxima, y a través de un bucle, son capaces de calcular los pronósticos para cada uno de los productos de manera individual en cada una de las plazas. Al final del código, se incluyó una función para exportar los resultados a Excel, incluyendo las métricas del error y los pronósticos a nivel diario. Un ejemplo de esto se muestra en la Fig. A en el apartado de Anexos.

Así mismo, se creó un modelo de Ensamble en Python para los productos de Electrolit, contemplando variables promocionales, de temperatura y las relacionadas con las fechas. Este modelo también incluye un bucle que genera los pronósticos a nivel producto/plaza, incluyendo también un código que permite exportar los datos a Excel como se observa en la Fig. B en Anexos.

Después de haber construido los modelos finales, se generó una tabla, la cual se encuentra en Anexos (Tabla F y G), con las métricas del error de cada uno a nivel plaza/categoría. Esta tabla sirvió para garantizar que los modelos tengan un buen desempeño y sean de utilidad para la empresa.

Así mismo, se crearon dos documentos como entregables para la empresa. El primero de ellos es un manual de uso de los modelos de pronóstico, incluyendo un modelo para calcular la importancia de las variables. Esto se hizo con el fin de que la empresa pueda replicar este método en el resto de sus productos a nivel nacional. Por último, se creó un reporte de hallazgos en el cual se incluyeron todos los resultados obtenidos a lo largo de este estudio, incluyendo algunas observaciones y recomendaciones para la empresa.

#### IV. CONCLUSIONES Y RECOMENDACIONES

A lo largo de este estudio, se probaron un total de 12 modelos diferentes para evaluar el desempeño con cada uno de los productos estudiados. Al evaluar los resultados obtenidos, se encontró que los modelos de series de tiempo fueron los que mostraron un mejor desempeño, obteniendo los valores del error con las métricas del RMSE y el MAPE más bajos. Tomando en cuenta lo anterior, se recomendó a la empresa que utilice estos modelos. No obstante, los modelos multivariados servirán para que la empresa pueda reaccionar y ajustar sus pronósticos ante los cambios de temperatura y las otras variables externas que se evaluaron. Es por esto que se hace entrega de ambos modelos, para darle libertad en la toma de decisiones a la empresa.

Por otro lado, es importante mencionar que los resultados y las conclusiones descritas en este trabajo son únicamente para los productos estudiados. Esto se hace con la intención de dejar claro que, si se desea aplicar estos modelos a otros productos o a los mismos en otras plazas, se debe replicar todo el proceso de análisis que se llevó a cabo, ya que de otro modo no se pueden garantizar los mismos resultados.

Para facilitar el proceso de réplica al resto de los productos y garantizar que la empresa pueda hacer un uso correcto de los modelos, se elaboró un manual con la serie de pasos a seguir de manera detallada, incluyendo una explicación de lo que hace cada línea de código. Este manual permitirá que la empresa pueda identificar las partes del código que debe cambiar, y lo que deben incluir en sus bases de datos para que los modelos creados por el equipo puedan ser utilizados para el resto de los productos de las categorías analizadas, e inclusive para cualquier otra familia de productos.

Así mismo, se recomienda que, si se decide replicar este método para otros productos, también se lleve a cabo un estudio de la literatura sobre variables externas, debido a que las variables que se estudiaron para este caso fueron particularmente seleccionadas para productos de bebidas.

#### AGRADECIMIENTOS

Agradecemos a la Universidad de Monterrey por el apoyo en la realización de este trabajo investigativo.

#### REFERENCES

- [1] Acosta, M. et al (2013). Validation Study on Artificial Neural Network-Based Selection of Time Series Forecasting Techniques.
- [2] Ali-Vehmas, A. (2017) Towards a Weather-proactive Supply Chain Recuperado de: <https://www.relexsolutions.com/resources/towards-a-weather-proactive-supply-chain/>
- [3] Arteaga, E. (2010) Construcción de un Modelo Econométrico para Estimar las Ventas Mensuales de las Cuatro Marcas Principales de Bebidas Gaseosas de la Empresa Ecuador Bottling Company Corp Recuperado de: <https://bibdigital.epn.edu.ec/bitstream/15000/1824/1/CD-2781.pdf>
- [4] Awad M., Khanna R. (2015) Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA. [https://doi.org/10.1007/978-1-4302-5990-9\\_4](https://doi.org/10.1007/978-1-4302-5990-9_4)
- [5] Banda, H. & Garza, R. (2014) Aplicación teórica del método Holt Winters al problema de credit scoring de las instituciones de microfinanzas Volumen 15, núm. 2, julio-diciembre 2014: 5-21
- [6] Browniee, J. (2016) Feature Importance and Feature Selection With XGBoost in Python Recuperado de: <https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/>
- [7] Cadenas, J., Garrido, C. & Díaz, R. (2010) Mejorando el comportamiento de ensambles basados en boosting, bagging y random forest mediante Soft computing SISTEMAS, CIBERNÉTICA E INFORMÁTICA VOLUMEN 7 - NÚMERO 1 - AÑO 2010 ISSN: 1690-8627
- [8] Castañares, G. (2018). Coca-Cola, 'Pepsi' y Oxxo se 'prenderán' con la ola de calor. Recuperado de: <https://www.elfinanciero.com.mx/empresas/coca-cola-pepsi-y-oxxo-se-prenderan-con-la-ola-de-calor>

- [9] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. DaimlerChrysler.
- [10] Elizondo, I. (2002). Desarrollo Metodológico y Técnico de un Pronóstico de Ventas para un Producto con Patrones Cíclicos [Universidad Autónoma de Nuevo León]. <http://cdigital.dgb.uanl.mx/te/1020146946.PDF>
- [11] Escobar, E., Díaz, J. & Taracena, L. (2009) Modelo para el ajuste de pronósticos agregados utilizando lógica difusa ISSN 1405-7743
- [12] Escobar, L., Valdes, J. & Zapata, S. (s.f.) Redes Neuronales Artificiales en predicción de Series de Tiempo. Una aplicación a la industria Recuperado de: <https://www.palermo.edu/ingenieria/Pdf2010/CyT9/02.pdf>
- [13] Espinosa, J. (2020) Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito volumen XXI (número 3) <https://doi.org/10.22201/fi.25940732e.2020.21.3.022>
- [14] Fernández, J.L. (2014) Clasificación de Respuestas de Pulsos de un ADS por Análisis Discriminante Actas de las XXXV Jornadas de Automática, 3-5 de septiembre de 2014, Valencia ISBN-13: 978-84-697-0589-6
- [15] González, P. (2009) Análisis de series temporales: Modelos ARIMA ISBN: 978-84-692-3814-1 Recuperado de: <https://perhuaman.files.wordpress.com/2014/06/econometria-modelos-arima.pdf>
- [16] González, R. (2017) Método Holt Winters R Pubs by RStudio Recuperado de: <https://rpubs.com/rjuares/ResumenFinalMCAA>
- [17] IBM. (2016). Opciones de Generación de nodo XGBoost Tree. Recuperado de: [https://www.ibm.com/support/knowledgecenter/es/SS3RA7\\_sub/modele\\_r\\_mainhelp\\_client\\_ddita/clementine/python\\_nodes\\_xgboost\\_tree\\_build.html?view=embed&fbclid=IwAR2Iteyw5Ry1Qq-3fiyqpbfLZ1ao5BvIQ7Bdehvb95j\\_GnKj3WXJcvpO31U](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modele_r_mainhelp_client_ddita/clementine/python_nodes_xgboost_tree_build.html?view=embed&fbclid=IwAR2Iteyw5Ry1Qq-3fiyqpbfLZ1ao5BvIQ7Bdehvb95j_GnKj3WXJcvpO31U)
- [18] InteractiveChaos. (s.f.). Gradient Boosting. Recuperado de: <https://www.interactivechaos.com/manual/tutorial-de-machine-learning/gradient-boosting>
- [19] Kaleva, H. & Smaros, J. (s.f.) The Complete Guide to Machine Learning in Retail Demand Forecasting Recuperado de: <https://hub.relexsolutions.com/hubfs/The-Complete-Guide-to-Machine-Learning-in-Retail-Demand-Forecasting-links.pdf>
- [20] Kilimci, Z., Akyuz, A., Uysal, M., Akyokus, S., Bulbul, B., & Ekmiş, M. (2019). An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain. <https://doi.org/10.1155/2019/9067367>
- [21] Krylstedt, J. & Weidlertz, A. (2016) A Study of Weather's Impact on Consumption of Goods: For Certain Weather-Dependent Products at a Small Grocery Store SCI School of Engineering Sciences ISRN-KTH/MAT/K--16/36--SE
- [22] Lazo, Jeffrey K., Megan Lawson, Peter H. Larsen, and Donald M. Waldman. (2011) U.S. economic sensitivity to weather variability. Bulletin of the American Meteorological Society 92: 709–720. doi:10.1175/2011BAMS2928.1.
- [23] Leonard, E. & Castro, Y. (2013) Metodologías para desarrollar Almacén de Datos. Recuperado de: <https://www.redalyc.org/pdf/1939/193930080003.pdf>
- [24] Lozano Forero, S. (2018). Propuesta de un índice SIPSA para el pronóstico de la inflación de Alimentos. Evidencia empírica.(Especialización). Fundación Universitaria los Libertadores.
- [25] Meca, I. & Agullo, O. (2018) Modelos ARIMA R Pubs by RStudio Recuperado de: <https://rpubs.com/Meca/386432#:~:text=La%20funci%C3%B3n%20aut o.,estacionalidad%2C%20diferencias%2C%20entre%20otras>
- [26] Mirasgedis, S., Georgopoulou, E., Sarafidis, Y., Papagiannaki, K. & Lalas, D.P. (2013) The Impact of Climate Change on the Pattern of Demand for Bottled Water and Non-Alcoholic Beverages John Wiley & Sons, Ltd and ERP Environment DOI: 10.1002/bse
- [27] Moine, J., Haedo, A. & Gordillo, S. (2011) Estudio comparativo de metodologías para minería de datos Recuperado de: [https://digital.cic.gba.gob.ar/bitstream/handle/11746/3525/11746\\_3525.pdf-PDFA.pdf?sequence=1&isAllowed=y](https://digital.cic.gba.gob.ar/bitstream/handle/11746/3525/11746_3525.pdf-PDFA.pdf?sequence=1&isAllowed=y)
- [28] Ortega, A. (2016) Los factores determinantes del aumento del consumo de agua embotellada den México Análisis desde el enfoque de Políticas Públicas Recuperado de: <http://repositorio-digital.cide.edu/bitstream/handle/11651/1440/153342.pdf?sequence=1&isAllowed=y>
- [29] Pavlyshenko, B. (2019). Machine-Learning Models for Sales Time Series Forecasting. Data 2019,4, 15; doi:10.3390/data4010015
- [30] Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830. Recuperado de: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [31] Pérez, P. (2012) Análisis en series de tiempo de los registros de la humedad relativa en la zona de Texcoco, México Recuperado de: <https://www.fcfm.buap.mx/assets/docs/docencia/tesis/matematicas/PedroPerezCortez.pdf>
- [32] Plonski, P. (2020) The 3 Ways To Compute Feature Importance in the Random Forest Recuperado de: <https://towardsdatascience.com/the-3-ways-to-compute-feature-importance-in-the-random-forest-96c86b49e6d4>
- [33] Scikit-learn. (2020). 3.2.4.3.6. sklearn.ensemble.GradientBoostingRegressor. Recuperado de: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>
- [34] Tseng, G. (2018). Gradient Boosting and XGBoost. Recuperado de: <https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5>
- [35] Turquie, Y. (2015) “Metodología de Inventarios y Pronósticos para Distribuidores de Productos de Telecomunicación” Recuperado de: <http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/9642/Informe.pdf?sequence=4>
- [36] Velázquez, M. (2019). Temperatura a nivel nacional aumentó 1.2 grados. Recuperado de: <https://www.economista.com.mx/arteseideas/Temperatura-a-nivel-nacional-aumento-1.2-grados-20190524-0016.html>
- [37] Villavicencio, J. (2010). Introducción a series de tiempo. Puerto Rico.
- [38] Villón, A. (2019) “Análisis del comportamiento de compra de bebidas isotónicas en la ciudad de Guayaquil” Recuperado de: <http://192.188.52.94:8080/bitstream/3317/12284/1/T-UCSG-POS-MGM-120.pdf>

## Anexos

TABLA A

Promedio del error por plaza y categoría para cada modelo (sin tener en cuenta las redes neuronales).

Plaza	Categoría	Métrica	Modelos Univariados				
			Holt Winters	Descomposición Multiplicativo	Descomposición Aditivo	SARIMA	SARIMAX
Puebla	Isotónicos	RMSE	94.62	21.09	18.41	93.09	86.7
		MAPE	13.25	10.13	2.56	13.30	12.05
	Agua Sin Sabor	RMSE	289.12	73.53	64.62	269.57	336.84
		MAPE	10.18	2.95	2.26	9.35	10.15
Tijuana	Isotónicos	RMSE	187.51	39.76	37.92	158.98	128.29
		MAPE	16.92	3.07	3.30	13.83	12.22
	Agua Sin Sabor	RMSE	604.60	141.62	123.82	531.12	434.73
		MAPE	15.55	20.10	3.11	12.78	11.61
Toluca	Isotónicos	RMSE	55.06	10.46	10.26	51.87	48.90
		MAPE	14.86	2.64	2.80	13.88	13.05
	Agua Sin Sabor	RMSE	210.12	56.02	50.28	199.52	182.96
		MAPE	13.06	2.97	3.00	12.26	11.71

TABLA B

Promedio del RMSE en modelos multivariados por plaza y categoría.

Plaza	Categoría	Métrica	Modelo						
			R.L.	Random Forest	GBRT	XG Boost	SVR	Ensamble	Redes N.
Puebla	Isotónicos	RMSE	117.37	89.46	87.26	80.14	81.09	79.62	114.88
		R <sup>2</sup> test	0.61	0.78	0.80	0.82	0.81	0.83	0.6
	Agua Sin Sabor	RMSE	582.03	428.25	507.44	489.37	548.12	489.98	503.12
		R <sup>2</sup> test	0.82	0.91	0.90	0.89	0.88	0.9	0.86
Tijuana	Isotónicos	RMSE	227.08	149.61	141.23	139.41	156.8	136.47	213.54
		R <sup>2</sup> test	0.73	0.88	0.89	0.88	0.86	0.89	0.79
	Agua Sin Sabor	RMSE	1125.26	590.00	649.82	633.71	800.79	642.44	1040.03
		R <sup>2</sup> test	0.78	0.90	0.88	0.89	0.85	0.89	0.7
Toluca	Isotónicos	RMSE	76.59	56.07	51.94	52.61	55.1	50.37	59.9
		R <sup>2</sup> test	0.73	0.78	0.82	0.80	0.78	0.82	0.73
	Agua Sin Sabor	RMSE	292.78	206.77	274.50	266.47	264.89	289.87	274.15
		R <sup>2</sup> test	0.83	0.93	0.91	0.90	0.91	0.89	0.89

TABLA C  
Promedio del RMSE en modelos univariados y multivariados por plaza y categoría.

Plaza	Categoría	Métrica	Modelos de Series de Tiempo			Modelos Multivariados		
			Descomp. Multiplicativo	Descomp. Aditivo	SARIMAX	Random Forest	XGBoost	Ensamble
Puebla	Isotónicos	RMSE	21.09	18.41	86.76	93.21	80.14	79.62
	Agua Sin Sabor	RMSE	73.53	64.62	336.84	428.25	489.36	489.98
Tijuana	Isotónicos	RMSE	39.76	37.92	158.98	158.04	139.41	136.47
	Agua Sin Sabor	RMSE	141.62	123.82	434.73	590.00	633.71	642.44
Toluca	Isotónicos	RMSE	10.46	10.26	48.90	58.73	52.61	50.37
	Agua Sin Sabor	RMSE	56.02	50.28	182.96	206.77	266.46	289.87

TABLA D  
Variables utilizadas en los mejores modelos multivariados.

	Tijuana	Toluca	Puebla
Agua sin Sabor	Las variables que más contribuyen al desempeño de los modelos de pronóstico son #Semana, las de temperatura y las promocionales.		
Isotónicos	<b>Electrolit:</b> variables promocionales, temperatura media y máxima, #Dia de la Semana		
	<b>Gatorade, Powerade &amp; Suerox:</b> variables de temperatura media y máxima		

TABLA E  
Selección de modelo para agua sin sabor según el RMSE.

	Bioleve	Electropura			Bonafont	
	Bioleve 1.5 lt	Electropura 1 lt	Electropura 1.5 lt	Epura 600 ml	Bonafont 750 ml	Bonafont 1.5 lt
Random Forest	169.729	890.477	1073.733	409.621	260.821	735.625
GBRT	148.7657	1002.0434	1104.0213	464.1048	452.0356	727.9384
XGBoost	145.159	951.406	888.807	497.465	326.8	992.637
SARIMAX	<b>145.0274</b>	<b>706.4271</b>	<b>725.8048</b>	<b>294.8268</b>	<b>212.3661</b>	<b>523.9088</b>
SVR	176.23	1246.365	1693.66	465.5	290.88	932.09
Ensamble	145.744	891.366	892.556	479.026	309.495	988.958
Redes	176.72	2374.28	1921.86	774.58	313.59	679.17

TABLA F  
RMSE y MAPE del modelo SARIMAX.

Puebla		Tijuana		Toluca	
RMSE	MAPE	RMSE	MAPE	RMSE	MAPE

<b>Agua sin Sabor</b>	336.10	10.98	436.48	11.96	182.89	12.48
<b>Isotónicos</b>	99.15	12.53	108.91	14.16	53.98	13.43

TABLA G  
RMSE y R<sup>2</sup> ajustada del modelo de ensamble.

	Puebla		Tijuana		Toluca	
	RMSE	R2 Ajustada	RMSE	R2 Ajustada	RMSE	R2 Ajustada
<b>Electrolit</b>	70.20	0.89	120.94	0.95	42.72	0.90

FIGURA A  
Output del modelo SARIMAX.

	A	B	C	D	E	F	G	H	I	J
1	Plaza	Producto	RMSE	MAPE	Pronosticos	NA.	NA.1	NA.2	NA.3	NA.4
2	Puebla	Agua purificada Bioleve 1	123.745177	8.91395166	440.4232	468.35808	610.34071	528.88774	480.30581	710.43767
3	Puebla	Agua purificada Bonafont	517.573584	8.05930727	1270.701	1312.8628	1559.096	1377.1921	1259.6173	1718.8326
4	Puebla	Agua purificada Electropu	399.569010	12.1110112	730.56487	869.98507	1159.062	1044.7825	977.02921	1347.7771
5	Puebla	Agua purificada Electropu	576.830577	10.3508753	763.80995	725.75512	1226.4225	992.58737	874.87449	1488.1689
6	Puebla	Agua purificada Epura 60	206.547774	10.4510550	342.66489	438.75095	544.98587	529.96924	496.19896	607.66182
7	Puebla	Bonafont Natural 750ml	1192.347990	16.0179145	650.80432	794.97857	871.71636	826.2805	759.23126	893.54947
8	Tijuana	Agua purificada Bioleve 1	145.027360	11.1697615	609.94027	727.02525	839.27429	764.84949	699.96775	971.65856
9	Tijuana	Agua purificada Bonafont	535.219125	11.0346512	931.95083	1127.4655	1276.1161	1139.7882	1023.4586	1478.7007
10	Tijuana	Agua purificada Electropu	706.427090	10.6553990	2574.706	3244.4312	3788.4162	3604.6809	3189.1754	4387.4905
11	Tijuana	Agua purificada Electropu	725.804778	12.0902948	1855.3006	2185.4566	2733.1171	2547.5467	2296.888	3289.9998
12	Tijuana	Agua purificada Epura 60	294.229819	10.7775459	898.43376	1150.1918	1285.363	1288.3539	1160.2318	1507.6141
13	Tijuana	Bonafont Natural 750ml	212.203233	16.0460156	769.9415	925.71607	1054.6837	1061.4899	948.50186	1098.6574
14	Toluca	Agua purificada Bioleve 1	55.1801428	9.75261285	323.08093	382.32709	416.19548	381.43132	346.48408	492.15538
15	Toluca	Agua purificada Bonafont	306.872192	9.40409007	1194.4676	1401.8359	1645.1481	1465.6898	1304.8282	1797.3634
16	Toluca	Agua purificada Electropu	244.316355	13.6268097	523.19436	671.79344	765.85762	710.71709	617.63088	842.18752
17	Toluca	Agua purificada Electropu	224.533149	12.7976541	271.28265	351.70819	355.47784	350.35767	335.96164	442.37892
18	Toluca	Agua purificada Epura 60	121.593112	13.9213516	236.67316	283.38981	354.97656	354.34456	308.30862	401.63833
19	Toluca	Bonafont Natural 750ml	144.868959	15.3744831	362.77054	505.85406	602.51746	528.15529	495.45976	698.53355
20										

FIGURA B  
Output del modelo de ensamble.

	A	B	C	D	E	F	G	H	I	J	K
1		Plaza	Producto	ac_train	ac_test	rmse	mape	r2_ajustada	0	1	2
2	0	Puebla	Electrolit Coco 625ml	0.853472734	0.710496418	83.06526624	12.3468903	0.85059965	591.7612	543.7897	570.2973
3	1	Puebla	Electrolit Fresa 625ml	0.894328177	0.780704033	69.2541654	10.77284621	0.892256181	694.9502	625.3151	612.0817
4	2	Puebla	Electrolit Fresa Kiwi 625ml	0.914848553	0.848245621	65.99827799	12.13254269	0.913178917	504.8428	474.8307	517.4161
5	3	Puebla	Electrolit Jamaica	0.887755449	0.780846703	38.83409031	14.0023433	0.885554575	271.2179	244.7809	259.9571
6	4	Puebla	Electrolit Lima Limon 625ml	0.918899153	0.804277151	57.7832443	12.85886095	0.91730894	478.8893	460.193	454.9588
7	5	Puebla	Electrolit Mora Azul 625ml	0.950624235	0.912241931	54.60688919	88.49441197	0.949656083	300.6625	316.5011	357.3794
8	6	Puebla	Electrolit Naranja-Mandarina 625ml	0.802696266	0.732901047	93.70617237	12.70228709	0.798827565	709.2516	654.0481	660.2623
9	7	Puebla	Electrolit Uva 625 ml	0.887177448	0.836061764	98.32778381	8.92441499	0.884965273	1100.589	963.034	960.1022
10	8	Tijuana	Electrolit Coco 625ml	0.95696584	0.940504045	155.2923823	10.18144292	0.956122033	1109.82	1146.864	1207.01
11	9	Tijuana	Electrolit Fresa 625ml	0.952114298	0.938220805	142.3883866	8.799770708	0.951175362	1150.561	1161.888	1245.149
12	10	Tijuana	Electrolit Fresa Kiwi 625ml	0.960297096	0.927929951	121.0467182	12.27849489	0.959518608	792.0608	771.0395	911.4394
13	11	Tijuana	Electrolit Jamaica	0.953906987	0.918758363	76.69249412	12.27584679	0.953003203	477.8562	479.5825	504.4124
14	12	Tijuana	Electrolit Lima Limon 625ml	0.959073149	0.931009927	123.9396817	13.15704215	0.958270662	728.7022	763.6402	849.7489
15	13	Tijuana	Electrolit Mora Azul 625ml	0.95583051	0.948131245	89.65036635	9.428773587	0.954964441	693.9466	723.7699	768.274
16	14	Tijuana	Electrolit Naranja-Mandarina 625ml	0.946251578	0.914742698	90.23588598	11.7762051	0.945197687	544.5497	537.2821	580.7656
17	15	Tijuana	Electrolit Uva 625 ml	0.959586412	0.922585776	168.2420396	11.05325357	0.958793989	1156.222	1202.689	1298.136
18	16	Toluca	Electrolit Coco 625ml	0.884237365	0.766418451	50.61008357	13.08581828	0.88206847	315.5377	319.4384	328.2595