**Comparativa de los modelos de predicción de series de tiempo y machine learning para la demanda efectiva de productos envasados**

**Cynthia Vanessa Ospina Campos**

20011304@aloe.ulima.edu.pe

Universidad de Lima

**Resumen:**

Hoy en día la necesidad de los negocios por no desaparecer hace que estos busquen ofrecer productos diferenciados a sus clientes, en ese sentido el apoyo en tecnología es bastante importante para agregar valor y potenciar el negocio, a pesar de ser aprovechado por la gran mayoría encontramos en el sector minorista que las bodegas no aprovechan de estos avances, de las más de 414000 bodegas en lima muy pocas llevando control de su información y dejan de ganar esa ventaja competitiva. Hoy en día las bodegas necesitan apoyarse en pronósticos para ofertar mejor su carta de productos envasados, la economía en la que vivimos además obliga a generar esta demanda de capacidades para no desaparecer. En esta investigación el propósito es el de apoyar a las bodegas a generar esa ventaja competitiva por media de tecnologías como son los a modelos de series de tiempo que permiten pronosticar de manera sencilla la demanda de los productos envasados. Para esto la metodología seguida ha sido la de entender el perfil de la bodega, la información que guardan, las variables más relevantes y aplicar las técnicas de ARMIA y Holt-Winters para evaluar su efectividad. Como resultados en base al análisis de los pronósticos resultantes podemos concluir que ambas técnicas tienen relevancia para el modelo siendo la técnica de Holt-Winters más relevante debido a la exactitud de sus pronósticos.

**Palabras Clave: Modelos de pronóstico, Series de tiempo, Holt-Winters, ARIMA, Demanda, Retail, Exactitud, Bodegas, Máquinas de Soporte de Regresión, Aprendizaje de Máquinas**

**Abstract:**

Today, the need for businesses to not disappear makes them seek to offer differentiated products to their customers, in that sense, support in technology is quite important to add value and enhance the business, despite being taken advantage of by the vast majority of us. In the retail sector that wineries do not take advantage of these advances, of the more than 414,000 wineries in Lima, very few keep control of their information and fail to gain that competitive advantage. Today wineries need to rely on forecasts to better offer their packaged products menu, the economy in which we live also forces us to generate this demand for capacities so as not to disappear. In this research, the purpose is to support wineries to generate that competitive advantage through technologies such as time series models that allow simple forecasting of the demand for packaged products. For this, the methodology followed has been to understand the winery's profile, the information they keep, the most relevant variables and apply the ARMIA and Holt-Winters techniques to evaluate its effectiveness. As results based on the analysis of the resulting forecasts, we can conclude that both techniques are relevant to the model, being the Holt-Winters technique more relevant due to the accuracy of its forecasts.

**Keywords: Forecast Models, Time Series, Holt-Winter, ARIMA, Demand, Retail, Accuracy, Grocery**

# INTRODUCCIÓN

Actualmente se vive en una época en la que el acceso a los servicios de tecnología es cada vez más común, podemos ver como las empresas aprovechan nuevas tecnologías y hacen de su uso para generar valor, reducir ineficiencias y optimizar sus operaciones. En general podemos decir que con la llegada de las nuevas tecnologías los negocios están evolucionando a plataformas digitales, sistemas móviles, Marketplace y otros de los cuales aprovechar para sacar ventaja competitiva.

Esta ventaja competitiva no se está viendo reflejada en los minoristas del sector Retail, en las bodegas y puestos de mercado no se aprovecha de estas ventajas, al no usar sistemas transaccionales no llevan un control de su negocio lo que resulta en no tener información sobre lo que venden o compran, por lo general la reposición de productos se da en base a lo que el dueño considera; al no aprovechar las bondades de las nuevas tecnologías y la minería de datos dejan de agregar valor a su negocio.

La situación de las bodegas y puestos de mercados ocurre dado que, al ser negocios informales, están metidos por sobre todo en su día a día, dejando de lado otras variables que podrían generar un impacto positivo en su negocio. El poco conocimiento en tecnología es otro de los factores por los que no llegan a explotar estás nuevas tecnologías, y por último la edad media de los bodegueros no permite que terminen de adaptarse (situándose entre 40 y 50 años según estudio de Kantar). Los bodegueros son los grandes perjudicados al no aprovechar de esto, elaborar estos pronósticos de manera sencilla y que sea traspasable serán elementos claves para que entiendan de su potencia y empiecen a aprovecharlos.

Explicado el problema que enfrentan los bodegueros, en esta investigación planteamos un modelo de pronóstico que nos permita responder a esa demanda de productos envasados. Los bodegueros así podrán reorientar mejor su cartera de productos a lo que sus clientes de verdad quieren y necesitan, aprovechar la minería de datos en este caso les dará esa ventaja competitiva sobre su competencia y permitirá que sus negocios crezcan.

En el ámbito académico buscamos comparar dos técnicas dentro de un sector (bodegas del sector de Retail) en el que el empleo de este tipo de técnicas es complicado debido a la poca información que se puede encontrar.

Dentro de las posibilidades de técnicas que se pueden aplicar tenemos la regresión lineal como primera opción, sin embargo, para su uso se tendría que pronosticar las variables externas las cuales tienen cierto grado de error, siendo así no tan efectiva para este caso.

Las técnicas de series de tiempo son las adecuadas para estos casos dado que se trabaja con información estacional, en ese sentido, ARIMA es un modelo estadístico para analizar y pronosticar datos de series temporales. Es un método simple pero preciso considerado así para ser uno de los estándares en la predicción de series temporales.

El segundo método dentro de las series de tiempo es Holt-Winters, se basa en un algoritmo iterativo en el que cada cierto tiempo (mes o semana) se hace un pronóstico del comportamiento de la serie en base a promedios ponderados de los datos anteriores.

“El desarrollo del método de pronóstico de datos de series temporales es bastante rápido, lo que resulta en muchas opciones de métodos que pueden usarse para predecir los datos por necesidad” (Aziz et al., 2018) y dado que las técnicas de series de tiempo aíslan de mejor manera el factor estacional podemos decir que para estos casos las técnicas de series de tiempo solucionan lo que otras técnicas como la regresión lineal no pueden.

Esta investigación contribuye para resolver este problema dado que se busca dar respuesta a las variables más influyentes para la venta de productos envasados de una bodega, por medio de los modelos de pronóstico de series de tiempo ARIMA y Holt-Winters, y la técnica de máquina de soporte de regresión se busca diseñar modelos de predicción para predecir la demanda de productos envasados para bodegas y por último se busca implementar estás tres técnicas para evaluar su desempeño y exactitud con respecto a sus pronósticos.

# ESTADO DEL ARTE

2.1 ° Modelos de pronóstico tradicionales

En este primer artículo se menciona que durante la época del *Black Friday* se eleva la complejidad al momento de gestionar la cantidad de personas que se tiene dentro de las tiendas, adicional a contar con personal limitado se vuelve aún más complejo buscar a los clientes ideales. Los autores creen que aplicando técnicas de predicción se podría llegar a tener ese resultado exitoso que todos esperan durante el Black Friday. El estudio se centra en el desarrollo de algoritmos para predicción que sean precisos y eficientes para analizar el gasto del cliente en el pasado y predecir el gasto futuro de estos tomando en cuenta las mismas características. En la investigación se presentan diferentes algoritmos y en diferentes plataformas para encontrar la mejor predicción, se implementaron siete diferentes algoritmos de aprendizaje automático. Adicional a esto, la investigación también visualiza técnicas de preprocesamiento y visualización de datos para lograr los resultados óptimos.

Durante la investigación citada se aplicaron las siguientes técnicas y se obtuvieron los siguientes resultados:

* Primero, aplicando la regresión lineal por medio de Python como lenguaje de programación y apoyado en la librería sklearn se obtuvo un error cuadrático medio del orden de 4800. (RMSE, mide el error que hay entre dos conjuntos de datos)
* Segundo, aplicando de una red neuronal, se obtuvo un indicador de RSME de 6000 que terminó siendo la peor de toda la investigación.
* Tercero, aplicando una red neuronal con *Deep Learning* se obtuvo un RMSE de 4200, sin embargo, la red no podía predecir debajo de $6000.
* Cuarto, se aplicó un árbol de decisión por el que se obtuvo un RSME de 3800 para el modelo.
* Quinto, se aplicó un árbol de decisión con *bagging* se obtuvo un RSME de 2900.
* Por último, se aplicó un modelo con XGBoost por el que se obtuvo un RSME de 2400. (C. S. M. Wu et al., 2019),

Según Gunaseelan, Patil y Wu concluyeron que el uso de modelos complejos como lo son la red neuronal pueden ser una exageración, los modelos más simples podrían funcionar bien si se aplica una correcta limpieza de datos, es así como concluyen que el número de compradores en Black Friday va a aumentar. El estudio está de acuerdo que las técnicas de aprendizaje automático producen mejores modelos de predicción que se pueden usar en las tiendas y los propietarios de las tiendas pueden analizar su base de clientes para apuntar mejor a los clientes y aumentar las ventas en un Black Friday. El estudio también está de acuerdo en que los datos deben procesarse previamente para lograr un conjunto de datos efectivo para desarrollar el modelo de predicción. Se discutieron varias técnicas en este estudio para lograr el mejor modelo. Sin embargo, todavía no hay una solución definitiva sobre cuál es la técnica correcta para lograr un modelo con alta precisión.

Para mejorar los resultados, se obtiene datos con muchas características y aumentamos la cantidad. Se deben realizar más investigaciones para mejorar las técnicas de aprendizaje automático existentes para trabajar en tiempo real y desarrollar un modelo eficiente.

En mi opinión, considero que la complejidad presentada en el *Black Friday* hace bastante interesante ver que algoritmo podría o no ser el ideal para este caso, al final vemos que hay varios con buenos rendimiento, yo opino igual que los autores con respecto a que las técnicas más convencionales pueden tener un excelente performance siempre y cuando la data que les entreguemos sea de calidad, sin necesidad de buscar algo más rebuscado.

En la investigación “Modeling Individual User’s Responsiveness to Maximize Recommendation Impact” los autores nos hablan de que los sistemas de recomendación proporcionan información personalizada basada en las preferencias de los usuarios, las diferencias en las preferencias entre los usuarios se estiman en base a registros anteriores, ya sea registros de clics o registros de compra. En general, los sistemas de recomendación parten del supuesto que los usuarios responderán a las recomendaciones, siempre que sus elementos favoritos estén correctamente seleccionados.

Los autores señalan que, a pesar de eso, no están tomando en cuenta que la capacidad de respuesta a las recomendaciones depende del tipo de usuarios; mientras que algunos usuarios pueden ser fácilmente persuadidos para que tomen medidas, otros pueden dudar más. Los autores proponen un modelo de predicción de compra que incorpora las diferencias en la capacidad de respuesta, se deriva la capacidad de respuesta de los usuarios individuales de una combinación de registros de compras y registros de recomendaciones, para verificar la mejora de la precisión de predicción de compra se verificó utilizando un conjunto de datos de compra. (Sato et al., 2016)

El impacto del modelo presentado por los autores excede el de un modelo convencional que ignora la capacidad de respuesta de los usuarios individuales. Estos resultados demuestran la importancia de modelar la capacidad de respuesta de los usuarios individuales, para la casuística en la que los registros de recomendaciones son insuficientes, la capacidad de respuesta debe estimarse a partir de otras fuentes. (Sato et al., 2016)

En el trabajo “Modeling Individual User’s Responsiveness to Maximize Recommendation Impact” se investigó la correlación de la capacidad de respuesta con los atributos del usuario y los atributos del elemento, las estimaciones de la capacidad de respuesta de los atributos correlacionados superaron a las estimaciones medias, además, el impacto de la recomendación del modelo estimado a partir de los atributos correlacionados fue casi comparable al modelo estimado a partir de los registros de recomendación. Se evaluaron los impactos de las recomendaciones obtenidos de la capacidad de respuesta estimada, usando el usuario y al ítem específico, en los casos evaluados se obtuvo: LI-CR de 0.39, LI-UISR de 0.593 y LI-UISR-E de 0.534. En el modelo explicado se propuso un modelo de predicción de compra que incorporara diferencias individuales en la espontaneidad de las recomendaciones, el modelo final mejoró la precisión de la predicción de compra y el impacto de las recomendaciones, estos resultados confirmaron la importancia de modelar la capacidad de respuesta individualizada. (Sato et al., 2016)

En mi opinión, esta investigación parte de un muy punto que anteriormente no había sido debatido, las recomendaciones aún carecen de mayor personalización y la forma en que lo enfocan los autores me parece buena, los resultados demuestran que aplicar esta técnica aportará más.

2.2 ° Modelos de pronóstico basados en series de tiempo

En la investigación “Sales Forecasting Using Holt-Winters in Enterprise Resource Planning At Sales and Distribution” de los investigadores Chrystian Sugiarto, V., Sarno, R., & Sunaryono, D. (2016) nos muestran que desde la data que se genera de los módulos de distribución y ventas de un ERP (*Enterprise Resource Planning*) se puede implementar un mecanismo de predicción de venta mensual con un porcentaje de error bien bajo, para esto los autores nos comentan que se utilizó el pronóstico de ventas Holt-winters, método multiplicativo y método aditivo.

Para el método de Holt-Winters Multiplicativo; la ecuación de predicción, nivel, tendencia y estacionalidad son mostradas a continuación en los puntos (6,7 y 8 respectivamente): (Sugiarto et al., 2017)

Para la generación de los pronósticos, se tomaron en cuenta tres componentes básicos en el pronóstico de la ecuación: (Sugiarto et al., 2017)

Nivel de ventas: datos derivados de las ventas de los cuales se han eliminado patrones estacionales, con el objetivo de obtener los datos apropiados sin la influencia de otras variables.

Tendencia de ventas: cambio que ocurre o se espera entre este mes y el mes que viene.

Índice estacional: Usado para el mes en el que se deseado pronosticar.

El estudio “Sales Forecasting Using Holt-Winters in Enterprise Resource Planning At Sales and Distribution” concluyó que el uso de pronósticos de ventas en un ERP para pronosticar clientes tiene una alta y muy efectiva demanda. Con el pronóstico de ventas en un ERP, la producción de bienes se puede configurar según los requisitos del cliente, así como la capacidad de un almacén para tener bienes, por lo que no se abastecen, ni subestiman. Asimismo, podrían actuar como planificación anticipada para la compra de materias primas, presupuestos y el número de empleados necesarios para la producción. Los métodos de Holt-Winters pueden usarse en el pronóstico de ventas cuando los datos reales tienen una tendencia y estacionalidad. La diferencia de Holt-Winter Multiplicative Method y Holt-Winters método aditivo se da cuando el contenido se da en el uso para una temporada particular de variaciones. El método multiplicativo Holt-Winter se usa mejor si las variaciones estacionales varían según la serie de datos.

Sarno, Sugiarto y Sunaryono (2016) explican que mientras Holt-Winter se usa mejor cuando la estacionalidad de las variaciones es constante en las series de datos y Holt-Winters tiene tres parámetros de suavizado y pueden capturar la estacionalidad. Los parámetros de suavizado, α, β y γ se utilizan como parámetros que pueden ajustarse para obtener un resultado de error mínimo. La medida de error utilizado para seleccionar el mejor pronóstico es MAPE, MSE y MAE. Método Multiplicativo Holt-Winters MAPE es 4%, MSE es1,711,066, y MAE es 989.44. Método aditivo Holt-Winters MAPE es 4%, MSE es 1,591,608 y MAE es 954.93. Esta muestra que ambos métodos tienen el valor de error no es mucho diferente, pero el método aditivo Holt-Winters tiene un error valor menor que el error en Holt-Winters Multiplicativo Método en este caso. Los resultados de la implementación del pronóstico de ventas en el sistema ERP se puede ejecutar bien. Cuando se crea el pronóstico, entonces las materias primas que se comprarán deberán ser informado inmediatamente al módulo de compras y en el módulo de producción, el resultado del pronóstico de ventas informará a hacer productos terminados de acuerdo con la cantidad de pronóstico dada. Con el método Holt-Winters para pronósticos de ventas en un ERP sistema, facilitará la determinación de la cantidad de crudo materiales que se comprarán y la cantidad de artículos que deben ser producido, de modo que la distribución de bienes al cliente Estar siempre disponible.

En mi opinión, el empleo de estos métodos para la predicción de ventas en el ERP me parece adecuado ya que en general separan los efectos que podría causar la estacionalidad dentro de las predicciones, además en base a los resultados podemos rescatar que ambos métodos son válidos.

En la investigación “Applying Statistical Learning Methods for Forecasting Prices and Enhancing the Probability of Success in Logistics Tenders” se explora la aplicación de pronósticos y aprendizaje estadístico. Usando series temporales que son para: pronosticar el costo a lo largo del servicio logístico y construir un intervalo “consciente del riesgo” para los precios que se ofrecerán en la oferta; también usa los datos históricos para desarrollar modelos estadísticos de aprendizaje. Estos modelos pueden predecir la probabilidad de éxito en una licitación basada en los valores reales de diferentes variables. (Nataraj et al., 2020).

La selección de proveedores se ha convertido en un concurso en el que las empresas de transporte compite brindando sus mejores ofertas, sostienen Alvarez, Bayliss, Juan, Nataraj, Panadero, Sada (2020). En los procesos de subasta tradicionales, los vendedores cotizan la tarifa más baja y brinda servicios de valor agregado para ganar la oferta, una propuesta incluye: el precio más bajo posible, el servicio más rápido y servicios adicionales como despacho de aduana, servicio de consolidación, etc. El objetivo principal de su trabajo es mejorar la precisión del pronóstico del análisis tradicional de series de tiempo. La precisión de sus modelos de pronóstico se evalúa con el uso del error cuadrático medio, utilizando datos históricos.

En nuestro caso, confiamos en métodos avanzados de suavizado exponencial para pronosticar la evolución de los precios a lo largo del año. Además, utilizamos un modelo de regresión logística para predecir la probabilidad de éxito en una licitación logística basada en datos históricos. En primer lugar, un análisis descriptivo de los datos disponibles es realizado. Este análisis nos permite identificar aquellas variables que muestran una alta correlación con la probabilidad de Éxito. Cada herramienta solo proporciona un costo parcial para la ruta dada, pero los costos de diferentes herramientas necesitan integrarse en un paso posterior para obtener el costo total de transporte; también faltan herramientas que permitan identificar la causa del éxito o el fracaso en las licitaciones como métodos que pueden predecir la probabilidad de éxito según el precio ofrecido. Se propone un enfoque de etapas múltiples como metodología global (Nataraj et al., 2020).

Etapa 1: automatización del análisis de costos de transporte: El objetivo es obtener estimaciones confiables de los costos de transporte entre un origen y un destino. Es posible que se necesiten métodos de análisis de datos como modelos de regresión de variables múltiples en caso de falta de datos.

Etapa 2 - Predicción de la evolución de los costes y el establecimiento de largo plazo: se enfoca en el uso de métodos de aprendizaje estadístico para pronosticar la evolución temporal de estos costos.

Etapa 3 - Modelado de probabilidades de éxito en licitaciones: esta etapa propone el uso de datos históricos y estadísticos.

Etapa 4 - Optimización de la cartera de licitaciones: propone combinar los modelos probabilísticos desarrollados durante la Etapa 3 con métodos de inteligencia artificial para optimizar la selección de licitaciones y tarifas con el objetivo de maximizar los beneficios esperados para la empresa, consideramos que los algoritmos aleatorios sesgados que pueden ser excelentes opciones para esta etapa.

Con respecto a nuestro caso, hemos empleado el método de suavizado exponencial de Holt-Winter que ya está implementado en lenguajes de programación como Python o R, así como en estadísticas software como Minitab, SPSS, etc. Tanto los métodos de Holt-Winter como los de ARIMA son métodos bien reconocidos en el análisis de series de tiempo, y ambos se siguen utilizando en múltiples aplicaciones. La razón principal de esto es que ninguno de estos métodos es claramente superior al otro. Se selecciona el método de Winter, el que proporciona los valores más bajos para los errores MSD, MAD y MAPE los parámetros de suavizado se han obtenido después de un proceso relativamente rápido de prueba y error. Esto es una ventaja que ofrece el método Holt-Winters sobre el método ARIMA.

Según la investigación “*Biomedical applications of time series analysis*” el autor habla de los métodos de análisis de series temporales y que estos son divididos en dos categorías: (Ferenti, 2018).

Métodos de dominio de tiempo: Se pueden dividir en métodos de regresión de series temporales “clásicas” que emplean regresores exógenos y métodos estocástico componente. En epidemiologia, los modelos de regresión son complicados por el hecho de que la variable de respuesta es conteo de datos dando lugar a modelos lineales generalizados.

Dominio de frecuencia: Conversión de series de tiempo, usando la transformada de Fourier la serie temporal se representa como la suma ponderada de sinusoides. Este análisis nos permite obtener conocimiento de los componentes periódicos de las series de tiempo es posible investigar la ciclicidad/ estacionalidad de los datos originales.

La transformada de Fourier nos permite que esta evolucione con el tiempo, pero hay métodos que hacen compensación entre resolución de tiempo y de frecuencia como el análisis wavelet (técnica para procesamiento de señales). La investigación de periodicidad en los datos epidemiológicos también es utilizada en el análisis de señales biomédicas, como el análisis de grabaciones de ECG. Los métodos de sobre dispersión y no linealidad son de uso frecuente desde la epidemiología ambiental hasta el modelado de enfermedades infecciosas. Un problema en el modelado de series de tiempo es la presencia de autocorrelación. Para capturar las series temporales, los modelos ARIMA son usados en otras áreas, que también aparecen en medicina. Los métodos de la serie incluyen filtrado / suavizado, el análisis de series de tiempo multivariables y más modelos complejos de espacio de estado. (Ferenti, 2018)

En mi opinión, esta investigación nos demuestra que el empleo de modelos ARIMA tiene un campo acción muy amplio y no solamente ser visto desde el lado comercial (ventas), ya que en el caso de la biotecnología apoya bastante para la administración de la salud que es donde se generan grandes cantidades de información.

En la tesis “Propuesta de un modelo de pronósticos de demanda y gestión de inventarios para la planeación de demanda en prendas de vestir juvenil” (Chapa & Mireilli, 2018) el autor nos comenta que se empleó la estrategia de clasificación ABC, en la que las líneas de producto se pasan a clasificar en base a 3 criterios: impacto que se tiene en el inventario, impacto que se tiene en el valor total, impacto que se tienen de venta o de costo; en estos casos se permite conocer a la empresa que el 20% de sus existencias concentra el 80% del valor de su inventario.

En Gestión de Compras hay una herramienta que hace seguimiento de compras y comprueba si lo que se compra cumple con lo que se proyecta, debido a que no se hace un buen seguimiento la empresa tiene un sobre stock de productos. La empresa tiene cuatro líneas de cluster juvenil: blusa, polo, vestido y jean que presenta la mayor demanda del cluster, por eso es importante ejecutarlas adecuadamente con el uso de pronósticos. Se hizo un análisis de evaluación económica financiera, donde los costos de las mejoras de las propuestas para la implementación, cuantifica ahorros, dando como resultado que es recomendable implementar las mejoras por el beneficio obtenido. Usa métodos de pronósticos cuantitativos que son más asertivos; los métodos actuales son generales basados en lo que el área comercial crea vender sin analizar la venta histórica. El método estadístico de Holt – Winters permite ahorrar S/. 6000000 para la temporada PV2018 ya que considera nivel, tendencia y estacionalidad de la demanda; este método tiene la ventaja de adaptar la información disponible actualizada; realizando un triple suavizamiento exponencial donde necesita calibrar la información mediante los índices: alfa, beta y gamma que varían entre 0 y 1. Se asigna cero cuando el último dato real es accidental y el patrón de las ventas debe mantener los niveles, tendencias o factores de estacionalidad históricos. Se asigna uno cuando el último nivel de ventas es relevante, estableciendo un nuevo nivel, tendencia o estacionalidad en las ventas. Hay que implementar una herramienta de gestión de compras, que permite hacer un seguimiento de las compras hechas y verificar si la compra cumple con el plan para la temporada. En la actualidad, la empresa no hace un seguimiento, obteniendo compras no planeadas que incrementan el inventario innecesariamente. (Chapa & Mireilli, 2018)

En la investigación un análisis comparativo de Regresión Lineal y Regresión Vectorial de Soporte de los autores Kavitha S, Varuna S, Ramya R nos comentaran que para predecir el futuro de los negocios es necesario tener el conocimiento del comportamiento de los consumidores, el producto y las ganancias. Las técnicas estadísticas se evalúan para un modelo predictivo basándose en los datos de series de tiempo, combinando datos históricos con datos actuales para una mayor precisión. Los datos de la serie temporal son analizados con modelos de regresión. Aquí vamos a comparar la regresión lineal y el modelo de regresión de vectores usando los datos de entrenamiento para usar el mejor modelo para una buena predicción y precisión.

El análisis ayuda a identificar el problema de la fuente de datos, las soluciones se basan en el procesamiento analítico en línea (OLAP). El análisis predictivo usa algoritmos de aprendizaje automático y estadístico para obtener futuras predicciones.

Se usa el aprendizaje automático, para mejorar la predicción ya que hace una combinación de informática y análisis estadístico. En la investigación se analizan vectores lineales y de soporte.

El aprendizaje de máquina supervisado infiere una función por datos de entrenamiento etiquetados que consiste en tuplas, estos se encuentran mapeados con un valor de datos de entrada y produce resultados precisos. La tupla es un vector que contiene valores de atributo; los datos destino tienen más de un resultado o un valor continuo. Este algoritmo de aprendizaje supervisado se utiliza para la clasificación y regresión.

En el caso de clasificación; predice etiquetas basadas en categorías. Por ejemplo, clasificar un estudiante como aprobado o reprobado.

Regresión; es un análisis estadístico para identificar relación entre las variables dependientes que pueden ser univariante y que identifica la dependencia entre una sola variable o multivariante. Se identifica la dependencia entre varias variables

Los modelos de regresión en base a las variables x generan el resultado y siendo este dependiente de x.

Regresión lineal: es el método predictivo más común, puede ser lineal simple ya que predice usando una sola variable independiente.

Regresión de vectores de soporte (Support Vector Machine): es un programa de aprendizaje supervisado, puede ser lineal o no lineal dependiendo del kernel que se configure.

Regresión de vectores de soporte lineal; utiliza kernel lineal que son funciones para la regresión de Kernel lineal SMOReg, implementa la regresión vectorial de soporte con varios núcleos.

Regresión de vectores de soporte no lineal, las funciones de kernel no lineales se utilizan para procesar datos de entrenamiento en el espacio de “n” dimensiones.

La configuración experimental; se lleva a cabo usando una herramienta de análisis de datos existente llamada Weka donde los datos se han recopilado del repositorio público. Los datos de evaluación de los estudiantes se toman para análisis multivariante que consta de 1000 instancias que a su vez tienen 5 atributos. El desempeño se evalúa con las métricas obtenidas que son: Media error absoluto y error cuadrático medio para evaluación.

En la investigación, la función LinearMedSq se adapta mejor a las ecuaciones lineales, el modelo se compara más con la función SMOReg que con la lineal.

En la investigación Feature Selection Base on Twin Support Vector Regression se nos explica sobre la regresión vectorial de soporte gemelo (TSVR) que es un algoritmo basado en el vector de soporte de regresión (SVR) y el vector de soporte (TWSVM). Hay algunos algoritmos de regresión vectorial de soporte que no se pueden aplicar a TSVR, por lo que un método propuesto es analizando los pesos y las funciones superior e inferior en TSVR; fusionando los dos vectores de peso, luego ordenando y eliminando las características (RFE). Los resultados demuestran la eficacia del algoritmo mejorando el rendimiento de la regresión.

En el campo del aprendizaje automático, la selección de variables en una de las búsquedas más importantes. La selección de variables de los datos no solo reduce dimensionalmente el espacio, simplificando la complejidad del modelo, sino también para remover montones de variables redundantes y repetitivas. La eliminación recursiva de características (RFE) está basado en la tecnología de clasificación de características para la selección de características específicas de clasificación de entidades.

En comparación con otros métodos de aprendizaje, SVR tiene la capacidad de generalización y fue usado en la economía, la ingeniería y la bioinformática, para mejorar la capacidad de generalización de SVR, Peng extendió el uso de la máquina de vectores de soporte gemelo (TWSVM) para el campo de regresión y la regresión de Vector de Soporte Doble (TSVR). TSVR-RFE fusionando así dos vectores de peso en un vector, teniendo un rendimiento superior en características

Para controlar y revisar las predicciones, trabajamos con la exactitud de la predicción, en ese sentido los indicadores usados son:

SSE es la suma del error cuadrado del valor verdadero y valor de predicción.

SSE denotada como precisión de ajuste, cuanto más pequeña sea la SSE, más adecuado será el valor estimado.

SST es la desviación de la suma al cuadrado de la prueba.

SSR es la suma de los errores al cuadrado del valor medio de las muestras de prueba y para representar el rendimiento de la regresión.

NMSE es el error cuadrado medio normalizado. Cuanto más pequeño sea el NMSE mejor para el modelo.

Para problemas de selección de características, una selección de características método TSVR-RFE se propone en este documento. En TSVR-RFE, fusionamos dos vectores de peso en un vector, a continuación, los vectores de peso combinado se ordenan, y la selección de características se logra eliminando las Funciones. Los resultados experimentales de varios UCI conjuntos de datos verifican que el TSVR-RFE propuesto no es sólo tiene una buena capacidad de generalización, pero también más rápido que las obras existentes.

En la investigación “E-commerce Time Series Forecasting using LSTM Neural Network and support Vector Regression” de los autores Ghassen Chniti, Houda Bakir y Hedi Zaher nos exponen su investigación basada en los pronósticos de series de tiempo de comercio electrónico utilizando LSTM Neuronal Regresión de vector de red y soporte. Los autores nos mencionan lo siguiente: para predecir los precios de los teléfonos en los mercados europeos se usa Long Short Term Memory Neural Network (LSTM) y Vector de soporte Regresión (SVR). Los vectores de soporte de regresión son un método para la clasificación y regresión. La red neuronal SVR predice el precio del día siguiente con RMSE 35, 43 euros con univariante y LSTM son los más los precisos. El modelo SVR predice el precio del día siguiente con RMSE 35, 43 euros con modelo univariante. Pero, al usar modelos multivariados, la red neuronal LSTM da la predicción más precisa para el próximo precio del día con RMSE de 24.718 Euros.

Según el informe de comercio electrónico global B2C de 2016 elaborado por la Ecommerce Foundation. La facturación global de B2C es de unos 2.272.700 millones de dólares en 2015, representan un incremento del 19,9% sobre la facturación del anterior año, que se puede explicar por la tasa de penetración en el mercado.

Las series de tiempo estudia las observaciones pasadas para desarrollar un modelo apropiado; comprendiendo el pasado para predecir los valores futuros. Los investigadores han trabajado en diferentes modelos estadísticos lineales como ARIMA para mejorar la precisión del pronóstico, se han usado varios modelos como el modelo bilineal, el modelo autorregresivo de umbral y el heterocedástico condicional autorregresivo (ARCH) también hay modelos basados en datos como LSTM y SVR.

El aprendizaje por refuerzo se usa para simular el comportamiento como en Robótica. Martin y col, presentó una red para predecir valores de series de tiempo futuras. Mukherjee y Col. En su artículo usó Support Vector Machine para tiempo no lineal, predicción de series. Lee Giles y col, utilizó Neural recurrente Redes e inferencia gramatical para predecir el cambio al día siguiente, Cao y Tay demostraron que una Máquina de Vector de Soporte con base radial regularizada. La función puede realizar una mejor predicción que la de backpropagation.

Assaad y col presentó un algoritmo para pronosticar series de tiempo con neuronales recurrentes, Kuremotoa y col utilizó Deep Belief Nets red neuronal probabilística generativa compuesta por 3 capas de máquina de Boltzman restringida (RBM) para el pronóstico de series de tiempo.

Primero se introdujo en el análisis de series de tiempo como un método para detectar anomalías de Bontemps et al. en. Gers et al utilizó un enfoque de ventana de tiempo para que el LSTM prediga series de tiempo.

Memoria a corto plazo (LSTM); es un tipo de red neuronal que es adecuada cuando se trata de Modelado de dependencias de largo alcance. La memoria contiene bloques en su lugar de unidades ocultas. Reducen los parámetros compartiendo las puertas.

Se prueban los modelos en series de tiempo desde el 03 de Abril de 2016 hasta el 23 de Marzo de 2017 . Siendo el objetivo una serie de tiempo disponible desde el 3 de abril de 2016 hasta el 23 de marzo de 2017, dando los precios relacionados con el Samsung Galaxy S7. Usaron Python 3.6 con diferentes API. Para el aprendizaje automático y el aprendizaje. El cálculo de entrenamiento y predicción se realizó con la plataforma informática e interfaz de programación de aplicaciones (API) CUDA de NVIDIA.

Se usará el error de predicción de la raíz cuadrada media ya que proporciona una imagen completa de la distribución de errores. El RMSE concede un peso relativamente alto a los grandes errores [23] que se adapta bien a nuestro enfoque. Se presenta una comparación de SVR y LSTM en la predicción de series de tiempo considerando los enfoques: unviarios y multivariados. Los experimentos con datos de comercio electrónico relacionados con los mercados telefónicos muestran que LSTM multivariante es más precisa que multivariante SVR en la predicción de tales series de tiempo. Donde el SVR es mejor que el LSTM.

El problema de la estimación de posición de una aeronave dentro de un espacio aéreo tiene la solución en un multisensor seguimiento de objetivos, siendo requisito para el monitoreo como el Control del Tráfico Aéreo (ATC) donde la posición de la aeronave se usa para apoyar las decisiones de control, estimando la ubicación de los objetivos usando mediciones de sensores como el radar de vigilancia secundario (SSR). El radar ya no es el único sensor que puede garantizar la vigilancia del tráfico aéreo. El error de radar hace que sea difícil obtener con precisión la ubicación del avión, proveniente de dos aspectos: error del sistema y error de medición; llevando a una gran diferencia entre la posición de detección original del objetivo y la verdadera posición del objetivo. Por eso hay que encontrar maneras de reducir el error de detección de radar porque el error del sistema es fijo y puede eliminarse. La transmisión de vigilancia dependiente (ADS-B) permite obtener la transmisión de información de posición de alta precisión de la aeronave por el avión usando sensores aerotransportados basados en un Sistema de Posicionamiento Global (GPS) para proporcionar Tráfico Aéreo. Sin embargo, la estabilidad y la continuidad de los datos no son tan buenas como el radar secundario.

Por eso, elegimos fusionar datos SSR de varias fuentes en lugar de utilizar datos ADS-B.

La Regresión Vectorial de Apoyo (SVR) puede resolver el problema de la alta dimensión de los datos con baja complejidad computacional, lo aplicaremos para datos multisensor, fusión para mejorar la capacidad de seguimiento.

Función de base radial (RBF) se utilizó como funciones del kernel SVR.

Los datos originales recibidos de SSR están en forma de flujo, tenemos que procesarlo antes de que pueda ser introducido en nuestro; el sistema de entrenamiento de modelos toma el mensaje ADSB como datos de etiqueta, mientras que la predicción del sistema solo toma datos SSR multifuente como entrada.

Los datos SSR adoptan el euro control radar de intercambio de información (ASTERIX). Un avión puede ser detectado por varios radares en el mismo tiempo. El Módulo de análisis de mensajes ASTERIX analiza los mensajes cat048 y cat021.

La Máquina de Vectores de Soporte (SVM) se desarrolló para solucionar problemas de clasificación, pero también se puede aplicar en regresión; la idea es mapear los datos de entrenamiento del espacio de entrada en un espacio de características de mayor dimensión a través del núcleo función y luego construir un hiperplano de separación con margen máximo en el espacio de entidades.

La función Kernel calcula el producto interno en el original espacio de características. RBF se utiliza comúnmente como el núcleo para la regresión.

Usar SVR para reducir el error de medición de los radares a la aeronave, separar la predicción de x de la predicción de y nos permite minimizar los errores en ambas direcciones.

En mi opinión, esta investigación podría ser base para el planteamiento que le estoy dando a la mía, dado que tienen la misma orientación a la demanda de productos, sin embargo, el sector al que se está aplicando es otro.

# ANTECEDENTES

La investigación parte de la aplicación de técnicas de minería de datos que son Arima y Holt-Winters, la selección de estos modelos de series de tiempo se da debido a que nos permiten aislar de mejor manera las variables externas y reducir el error de estás. En el sector minorista (bodegas) las series de tiempo resultan de mayor importancia dado que se ajustan mejor al flujo de artículos vendidos o comprados. Lo que se busca con el pronóstico de las series de tiempo es extender los valores históricos al futuro, que es donde generalmente no hay mediciones disponibles en este sector.

3.1 ° Modelo ARIMA

El modelo ARIMA es representado por la siguiente formula:

Dónde:

* d corresponde a las d diferencias que son necesarias para convertir la serie original en estacionario.
* , … son los parámetros que pertenecen a la parte autorregresiva del modelo
* son los parámetros pertenecientes a la parte de medias móviles del modelo
* es una constante
* es el término del error

El modelo ARIMA como paso previo a su ejecución necesita de la identificación de los coeficientes y números de regresión que estos usarán, la expresión del modelo se da de la siguiente manera: ARIMA (*p, d, q*). En donde p es la componente autorregresiva, d es la componente integrada y q la componente de media móvil

ARIMA permite la búsqueda de patrones para la predicción del futuro, se explican en base a los datos del pasado estás estimaciones que se hacen a futuro y se evita tener las variables independientes. ARIMA es muy sensible a la precisión con que se determinan sus coeficientes.

3.2 ° Método de pronóstico Holt-Winters

El método de Holt-Winters es representado por las siguiente formula:

Donde:

* representa la constante de atenuación del promedio de los datos (0 < < 1).
* representa la constante de atenuación de la estimación de tendencia (0 < < 1).
* representa la constante de atenuación de la estacionalidad (0 < < 1).
* es el valor atenuado en el periodo t.
* es la estimación de la tendencia del periodo t.
* es la estimación de la estacionalidad del periodo t.
* *L* es la longitud de la estacionalidad.
* *P* es el número de periodos a pronosticar en el futuro.

Este es un método para obtener pronóstico que se compone del triple exponente suavizado y presenta la ventaja versus otras técnicas de ser más sencillo y fácil de adaptase a medida que se va suministrando más información, al ser un extendido del método Holt (dos exponentes suavizados) a este se le añade una consideración de nivel, tendencia y estacionalidad para una determinada serie de tiempo.

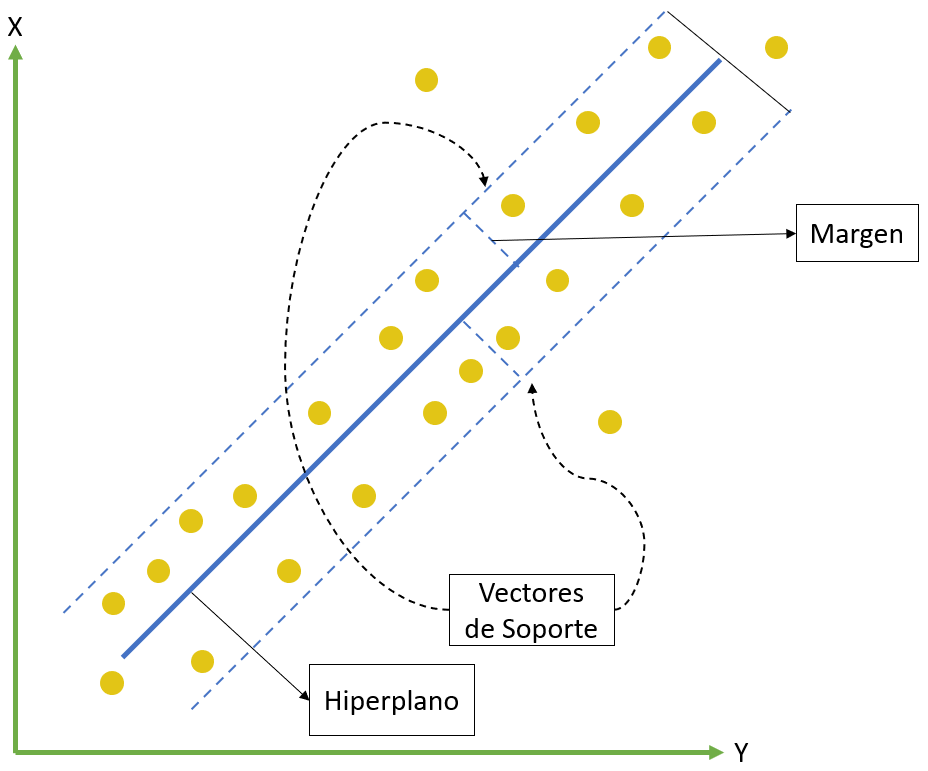
3.3 ° Vectores de Soporte Regresión

“Las máquinas de vectores de soporte son un conjunto de algoritmos de aprendizaje supervisado, los métodos se encuentran relacionados con problemas de clasificación y regresión” (Kavitha et al., 2017), en base a data de entrenamiento podemos etiquetar las clases y entrenar una máquina de soporte para construir un modelo que prediga la clase de una nueva muestra. SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión.

“SVM (*Support Vector Machine*) como técnica nos permite realizar clasificaciones. Clasifica en base a variables categóricas discretas y en base a eso es que se centra su poder predictivo” (Kavitha et al., 2017), en cambio, cuando nos referimos a SVR (*Support Vector Regresion*) indicamos que trabajamos con regresores, regresores que realizan regresiones en base a la predicción de variables continúas ordenadas. Ambos usan algoritmos muy similares, pero predicen diferentes tipos de variables.

Como podemos ver en la figura 1, los vectores de soporte son los puntos de datos más cercanos al hiperplano, un hiperplano es un plano de decisión que separa entre un conjunto de objetos, para la figura 1 se está trabajando con datos líneas dado que es más fácil de ver. El margen es un espacio entre las dos líneas (vectores de soporte) en los puntos más cercanos de la clase.

Figura 1: Máquina de Soporte de Regresión



Fuente: Desarrollo Propio

SVM se soporta en la función del kernel para identificar el tipo de recta que necesita graficar, en ese sentido, la manera más simple de realizar la separación es mediante una línea recta, un plano recto o hiperplano N-dimensional.

El truco del kernel consiste en gráficar los puntos de datos en los ejes X y Z, en el que Z es la suma cuadrada de X y Y. A partir de ahí se puede realizar la separación lineal fácilmente. Es decir, un kernel transforma un espacio de datos de entrada en la forma requerida.

Desafortunadamente los universos a estudiar no se suelen presentar de esta manera, sino más bien se debe tratar con más de dos variables predictoras o curvas no lineales o casos donde los conjuntos de datos no pueden ser completamente separados.

Para estos casos, encontramos los siguientes tipos de funciones kernel: polinomial, perceptrón, función de base radial gaussiana y sigmoide.

Donde:

* W es la magnitud del vector o hiperplano
* C es una constante y debe ser mayor a 0, determina el equilibrio entre la regularidad de la función y la cuantía hasta la cual toleramos desviaciones mayores que las bandas de soporte
* son las variables que controlan el error cometido por la función de regresión al aproximar a las bandas

El parámetro de regularización (C) es la constante que determina el equilibrio entre la regularidad de la función y la cuantía hasta la cual toleramos desviaciones mayores que las bandas de soporte

3.4 ° Perfil del consumidor

Según un estudio titulado “¿Cuáles son los tipos de consumidores peruanos?” desarrollado por Kantar Worldpanel (2018) y publicado en la web Peru-Retail, nos menciona que “los tipos de consumidores peruanos son: ahorradores, malabaristas, planificadores, despreocupados y escaladores”. Dado que nos encontramos en un mercado donde los productos y servicios son parecidos, segmenta a su público objetivo para tener una compra personalizada.

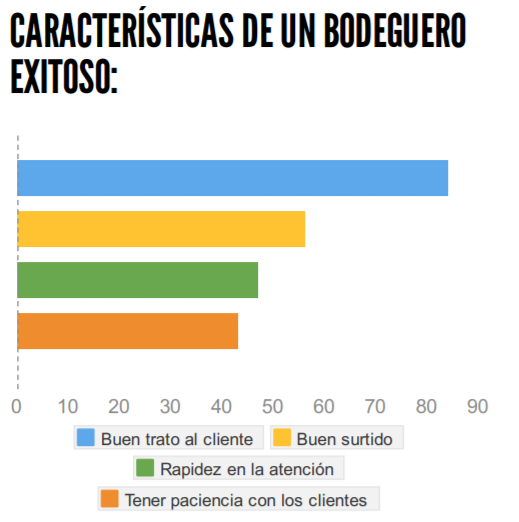
En la publicación se pasa a explicar cada uno de los tipos de consumidores donde tenemos al 10% de la población que pertenece a los consumidores ahorradores, enfocados en la compra de bebidas y alimentos para mascotas, prefieren las marcas Premium pero ahorran en formatos grandes; consumidores malabaristas compran en bodegas y en formatos pequeños, cuentan con un gran presupuesto por lo que gastan menos; consumidores planificadores hacen la planificación de sus compras tratando de ahorrar, priorizando alimentos y el cuidado del hogar, buscan una variedad de formatos y tienen la relación calidad-precio; consumidores despreocupados que priorizan marcas premium en alimentos y bebidas para personas y también para mascotas, también consumen marcas económicas en lácteos y cuidado personal, tienen como canal preferido los autoservicios, bodegas y mercados gastando 40% por encima del promedio; consumidores escaladores, son digitales, utilizan *cash and carry* y el 20% de sus compras se realizan en *discounters*, Kantar Worldpanel (2018)

3.5 ° Perfil de la bodega

Para entender el perfil de los bodegueros en el Perú tenemos que ir hasta el último estudio que se realizó, este lo encontramos en el portal de Ipsos que nos detalla cierta información tomada el 2014.

Ipsos (2014) dice que el canal tradicional en el Perú es la mejor para venta de productos, en especial los de consumo masivo. Como se puede ver en la figura 1 de lo más resaltante del estudio podemos mencionar que en el 2014 tuvieron ingresos del punto de venta, siendo el principal sustento de más del 40% de hogares de los bodegueros en lima, además de que tenían un incremento en ventas del 10%. Del perfil sociodemográfico podemos mencionar por un lado que 73% son mujeres, están casadas y tienen 48 años, 47% de estas mujeres alcanzó secundaria completa / técnica incompleta y el 67% nació en provincia; en hombres se puede mencionar que el 47% son jefes del hogar, 58% pertenece al nivel socio económico C y que tienen 11 años de experiencia como bodegueros. Con respecto a la gestión de la bodega, podemos mencionar que el apoyo publicitario es importante para la venta de productos, 11% hace la compra y gastos de las bodegas al contado, 26% tiene alguna deuda con un banco, 54% cuenta con teléfono público en la bodega, 81% de bodegas operan en locales propios.

Figura 2. Característica de un Bodeguero Exitoso



Fuente: (IPSOS, 2014) Perfil de la bodega y el Bodeguero.

El perfil de la bodega puede ser definido de la siguiente manera: “las bodegas son locales independientes donde se venden principalmente abarrotes; es decir, productos como alimentos, artículos de aseo personal, limpieza, menestras, gaseosas y golosinas, entre otros” (Esparza et al., 2017), los bodegueros atienden a los clientes, es usual que el bodeguero sea el dueño del negocio o sus familiares, además usualmente tienen jornadas largas de trabajo.

3.6 ° Indicadores

Para la validación de los modelos propuestos se va a evaluar la predicción que hagan, para eso se propone el uso de indicadores que nos permitirá medir si la predicción es acertada o no, a la vez comparar entre los dos modelos cuál de ellos es mejor. En general, el empleo de indicadores ayuda a tener mayor claridad sobre los cálculos obtenidos.

Una vez tengamos aplicadas las técnicas de Holt-Winters y ARIMA, procederemos a evaluar la exactitud de los pronósticos, para eso hemos definido que la exactitud del pronóstico sea revisada en base a los siguientes indicadores:

MAE: Media del error absoluto, calcula el promedio de los errores en unidades.

MSE: Error cuadrático medio, es un estimador que evalúa el promedio de los errores al cuadrado, en conclusión, la diferencia de un estimador versus lo que se estima.

MAPE: Media del error absoluto en porcentaje, establece el promedio del error en porcentaje.

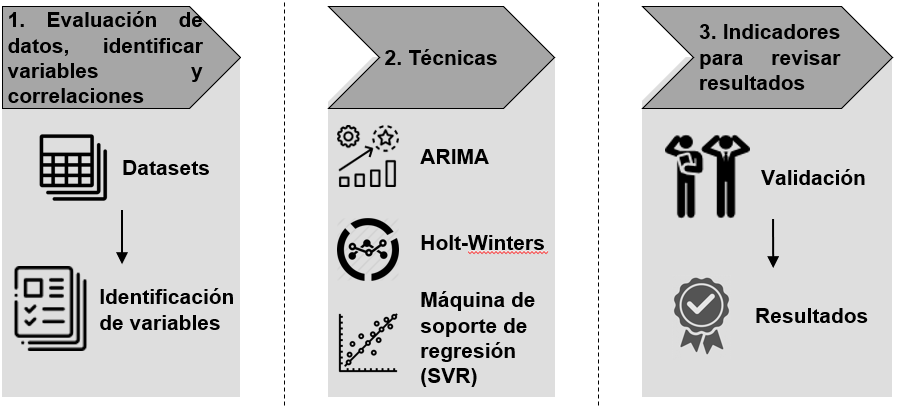
Exactitud de pronóstico: Es el porcentaje de exactitud de los pronosticado versus el real, indicador que se obtiene en base al resultado del MAPE.

# METODOLOGÍA

De la revisión de los antecedentes y el estado del arte hemos podido evaluar cuales son las técnicas que mejor se adaptan para el pronóstico de la demanda en las bodegas, dentro de la revisión evaluamos diferentes técnicas de series de tiempo y algoritmos de machine learning como regresión lineal, de todas estas técnicas resaltamos como las más relevantes para esta investigación las técnicas de series de tiempo. En especial hemos visto que la aplicación del modelo Arima y el método de Holt-winters, ambas pasaran por la metodología explicada en la figura 1, según la figura hemos dividido en tres etapas la ejecución, la primera referente a la data (evaluación del set de datos y la identificación de variables relevantes), la segunda etapa referente a las técnicas (construcción del modelo) y la tercera referente a los resultados (evaluación de los modelos y resultados).

Adicional a las técnicas de series de tiempo también nos apoyaremos en la técnica de máquina de soportes de regresión (SVR) que en base a la data de los productos de las bodegas nos ayudará a realizar este tipo de predicciones. La principal diferencia entre esto se da que emplearemos dos técnicas de minería de datos y una de *machine learning* en el que aprovecharemos el data set con las diferentes variables que tenemos para el modelo

Figura 1: Metodología de desarrollo



4.1  Datos

4.1.1   Información de la bodega

La información que se necesita de las bodegas por lo general la tienen en sus facturas de compra, esto permite tener una primera visión sobre todos los productos que adquieren sin embargo cuando se quiere ir a las ventas es necesario que ellos registren esto como parte de sus actividades, ahí el primer problema que se enfrenta el sector minorista cuando se quiere mayor información analizar y sacar conclusiones de cara a evaluarse con respecto a sus clientes, se ha recopilado información de compras, ventas y atención que se puede observar en la tabla .

Hay 10 categorías de productos que se venden dentro de las bodegas, de un total de 150 productos. El precio de venta y precio de compra son registrados durante la compra de la mercadería y durante la venta que se le hace al cliente, además se tiene información sobre la modalidad en que paga el cliente, el tiempo de atención y si el cliente es preferente (viene a comprar seguido).

Tabla 1

*Resumen de variables utilizadas para la clasificación*

|  |  |  |  |
| --- | --- | --- | --- |
| VARIABLES | tIPO DE DATO | VALORES | TOTAL DE VALORES |
| tipo de producto | CATEGÓRICO | Aguas, Cervezas, Salsas, Fideos, Detergentes, Conservas y Galletas | 10 |
| PRODUCTO | DISCRETO |  | 150 |
| PRECIO VENTA | CONTINUO | Entre 1 a 20000 | 20000 |
| PRECIO COMPRA | CONTINUO | Entre 1 a 5000 | 5000 |
| MODALIDAD PAGO | DISCRETO | 1 o 2 | 3 |
| CLIENTE PREFERENTE | DISCRETO | 1 o 2 | 2 |
| TIEMPO ATENCIÓN | CONTINUO | Tiempo | 5000 |

Una vez capturada la información pasamos a revisar los datos para validar que puedan ser usados en el modelo, lo primero que se debe de evaluar es verificar que no se tengan valores atípicos o que la información no tenga data sucia, en caso sucede deberíamos limpiar la información para generar un set de datos limpio y óptimo para su análisis.

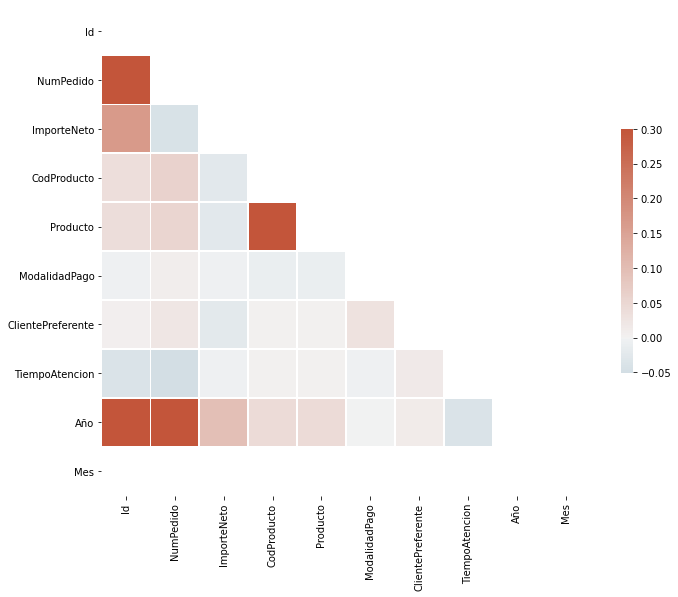
A la par necesitamos identificar las variables que deberíamos usar para el modelo, es decir, variables que podamos añadir en base a la información leída de investigaciones anteriores, además de nuevas variables que considere interesantes para el modelo.

Una vez completamos esto necesitamos evaluar la correlación de la información, lo que se quiere es evitar el encontrar dependencia entre variables para que la predicción que obtengamos sea la correcta, lo ideal en estos casos es que no existe dependencia de las variables para que los pronósticos sean precisos, en caso suceda es posible aplicar otro tipo de técnicas para evitar la correlación.

4.1.2  Variables de modelo

Para la evaluación de las variables se aplicó un análisis de correlación de Pearson, en este podemos revisar si existe grado de relación entre 2 variables, la idea de aplicar este tipo de análisis es validar la independencia de las variables para que el resultado de la predicción no sea erróneo por estar mal influencia de las variables que ingresan al modelo.

Figura 3: análisis de correlación



Fuente: Desarrollo Propio

A continuación, explicamos la descripción de las variables, tipo de datos y los valores que puede tener:

La variable categoría viene a ser un agrupador del tipo de producto que se está vendiendo, es un dato de tipo texto y como valores pueden ser los que se observan en la tabla 1

Tabla 2

*Valores de tipo de producto*

|  |
| --- |
| vARIABLE DE tipo de producto |
| Cerveza |
| salsas |
| aguas |
| fideos |
| HARINAS |
| detergentes |
| galletas |

La variable de precio de venta representa la venta real por la que se vende un producto, es un dato de tipo decimal y como valores puede ser cualquier monto, los valores se pueden en la tabla 4 donde se puede ver el precio por producto

La variable de precio de compra representa el costo de obtener el producto del proveedor, es un dato de tipo decimal y como valores puede ser cualquier monto.

La modalidad de pago es la forma en la que ha pagado el cliente, es un dato de tipo texto y como valores puede contener: efectivo, por transferencia o con tarjeta.

El cliente preferente hace referencia a si el cliente que viene a comprar es un cliente que viene seguido a la bodega, el tipo de dato es texto y como valores se tiene: si o no.

El tiempo de atención nos indica cuanto se ha tardado el cliente en realizar su compra dentro de la bodega, es un dato de tipo tiempo y como valores tiene la cantidad de tiempo en minutos que se atendió el cliente.

De la evaluación de las variables hemos podido ver aplicando un análisis de correlación hemos podido evaluar esto.

4.2  Modelo ARIMA

Corroborador los pasos explicados párrafos anteriores, se ejecuta las técnicas, para este caso se procede con la aplicación del modelo ARIMA, lo primero que se hace es en base a la serie de datos obtenido se convierten a una serie estacional, para esto se genera un segundo set de datos basado en el primero. En segunda instancia, se ejecuta el modelo autorregresivo (AR) para obtener los primeros resultados de las predicciones, con esta información recién se procede a la ejecución del modelo ARIMA, es necesario indicar la cantidad periodos que se desea predecir, el orden integrado y la media móvil.

En la figura 3 podemos ver el pseudocódigo para la ejecución del modelo ARIMA, en base a la definición de los parámetros p, d y q se ejecutan bucles para dentro hacer que el modelo se ajuste en base a la fórmula de AR, posteriormente este ajuste es pasado a una optimización y si se cumple con la condición el modelo continúa ajustando.

Pseudocódigo del modelo ARIMA

Para p 🡨 0 a 5 Entonces

Para d 🡨 0 a 1 Entonces

Para q 🡨 0 a 1 Entonces

Modelo 🡨 ajustar (ARIMA(p,d,q))

Aic\_curr 🡨 Compute\_AIC(model)

Si Aic\_curr < aic Entonces

Modelo\_opt 🡨 model

Aic 🡨 aic\_curr

Retornar model\_opt

4.3  Método Holt-Winters

Para la técnica de Holt-Winters indicamos la cantidad de intervalos que queremos pronosticar (un ejemplo es la cantidad de meses), indicar los valores del alfa beta y gama, a partir de eso tiendo las variables se procede a calcular el valor atenuado en el periodo (), con estos resultados se pasar a calcular la estimación de la tendencia del periodo (). A continuación, se calcula la estimación de la estacionalidad del periodo ().

Una vez estos valores son calculados dentro del algoritmo se obtiene el pronóstico para cada uno de los periodos, es decir, un pronóstico ajustado a la demanda real. El pronóstico ajustado a la demanda real es una de las características del modelo de Winters (se adapta con facilidad a los cambios drásticos en el comportamiento de los datos). Para terminar, se calcula el error (diferente entre el pronóstico y el valor real).

En la figura 4 podemos ver cómo es que se genera el pseudocódigo para la ejecución del método de Holt-Winters en el que podemos ver cómo es que se calcula el valor atenuado en el periodo, la estimación de la tendencia en el periodo y la estimación de la estacionalidad del periodo.

Pseudocódigo del método Holt-Winters

Para i 🡨 0 a 32:

Si i = Tamaño(Y) Entonces

Y 🡨 A + B \* S

A 🡨Alpha \* (Y / S ) + ( 1 – Alpha) \* (A+B)

B 🡨Beta \* A + ( 1 – Beta) \* B

S 🡨Gamma \* (Y / (A + B) ) + ( 1 – Gamma) \* S

Y 🡨(A + B) \* S

4.4  Técnica de Soporte de Vectores de Regresión

El método de soporte de vector de regresión nos permitirá desde la aplicación de una técnica de *Machine Learning* añadir variables categóricas a la información para ver que tanto afectan los factores externos entorno a la información que hemos podido obtener de la bodega en evaluación.

Como paso previo a la aplicación de la técnica de SVR tenemos que pasar por la revisión de la data, se necesita que no exista correlación entre la información. Para esto aplicaremos una correlación sencilla a la información y obtendremos un gráfico que nos dirá el nivel de correlación (directa o inversa) de existir. Sí estuviéramos en un caso de correlación entre variables aplicaríamos el análisis de componentes (PCA) principales en el que el número de variables resultantes estaría dado por el porcentaje de varianza que explica el modelo.

Para la técnica de soporte vectorial primero necesitamos definir algunos parámetros con los que el algoritmo podrá comenzar a entrar, para esto primero necesitamos definir el kernel, acá indicamos el tipo de recta con la que vamos a trabajar: lineal polinomial, rbf, sigmoide y pre-computed. Una vez definido el kernel, pasamos a indicar el valor de parámetro de regularización (C) y de épsilon para la ecuación.

El parámetro de regularización C, que nos indica el equilibrio entre la regularidad de la función y la cuantía hasta la cual toleramos desviaciones mayores que las bandas de soporte, por defecto las librerías nos lo colocan como 1, con esto aseguramos el equilibrio entre la regularidad de la función y la cuantía hasta la cual podemos tolerar desviaciones mayores que las bandas de soporte.

La épsilon son las variables que controlan el error cometido por la función de regresión al aproximar las bandas, en este caso estamos tomando en cuenta como valor 0.2, para obtener este valor hemos aplicado el margen máximo de clasificación que nos ayuda a obtener este valor.

Pseudocódigo para Máquina de Soportes de Regresión

C 🡨100

Epsilon 🡨 0.1

Kernel 🡨 [‘linear’,’RBF’,’Polynomial’]

Para i🡨1 a 3

SVR[i] 🡨 SVR(Kernel[i], C, Epsilon)

# EXPERIMENTACIÓN

* 1. ° Base de Datos

Tabla 3

Detalle Base de Datos

|  |  |  |
| --- | --- | --- |
| VARIABLES | tIPO DE DATO | Cantidad registros |
| Id | INT | 14759 |
| NumERO DE Pedido | INT | 14759 |
| fecHA DE entrega | DATE | 14759 |
| importe neto | FLOAT | 14759 |
| codIGO DE producto | STRING | 9277 |
| Detalle | STRING | 14759 |
| Producto | STRING | 14759 |
| produc descRIPCIÓN | STRING | 9277 |
| fecha | DATE | 14759 |
| modaLIDAD DE PAGO | STRING | 14759 |
| CLIENTE PREFERENTE | INT | 14759 |
| TIEMPO ATENCIÓN | TIME | 14759 |

La base de datos cuenta con la información que podemos ver en la tabla 3, la información se comenzó a capturar desde el 2013 desde las boletas y facturas de compra que se han podido recolectar desde esa época, poco después del inicio la información se almaceno periódicamente dentro de un Excel en el que se tiene toda la información posible recolectada.

* 1. Limpieza, análisis y normalización de los datos

Para la aplicación de la serie de tiempo se definió aplicar un primer filtro de fecha en el que se acotó de 2016 a 2018 (3 años) para ingresar a los algoritmos de series de tiempo por producto. Además, otros filtros que se tuvieron que aplicar para tener únicamente la información de venta fue del detalle igual a venta, descartando la parte de compra y descuentos.

Posterior a la aplicación de los filtros dentro de la información necesitábamos identificar los productos que poseen toda la data necesaria para entrar a los algoritmos, para eso se realizo un cruce de la información en base a la venta de los productos en los 3 años a revisar. En la figura 3 podemos encontrar a los 10 productos con mayor cantidad de venta durante este periodo.

Figura 4: 10 productos con mayor venta por año



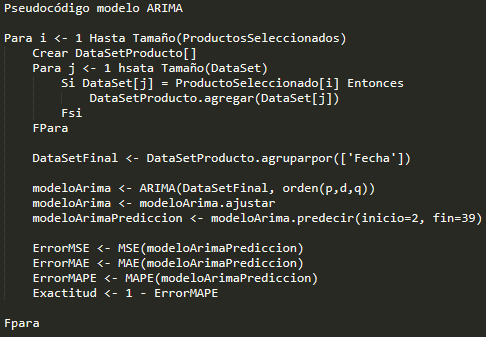
Fuente: Desarrollo Propio

* 1. Aplicación ARIMA

ARIMA al ser una técnica de serie de tiempo el input principal que recibe es la venta del producto para identificar la tendencia y posterior a eso poder realizar la predicción. En ese sentido, de la información revisada el modelo de ARIMA recibe como inputs de información fecha, venta y producto para realizar la proyección.

En la figura 4 podemos ver las librerías necesarias para la ejecución y la lógica para poder generar el análisis de series de tiempo para los productos. A grandes rasgos se realiza un filtro para quedarnos con las variables mencionados anteriormente y almacenarlas en un dataframe que nos permitirá generar el modelo ARIMA por cada producto.

Figura 5: Pseudocódigo Modelo ARIMA



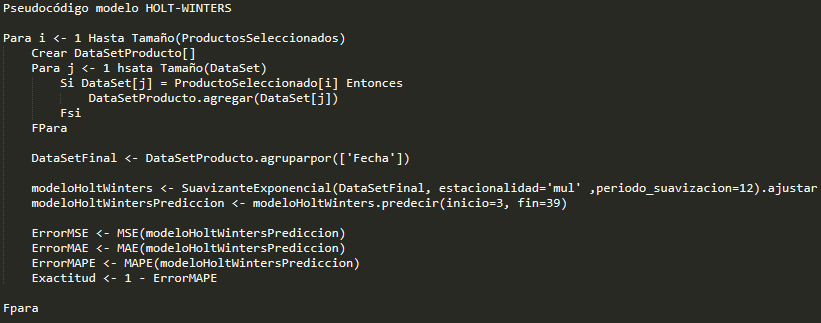
Fuente: Desarrollo Propio

* 1. Aplicación Holt-Winters

Holt-Winters al ser una técnica de serie de tiempo el input principal que recibe es la venta del producto para identificar la tendencia y posterior a eso poder realizar la predicción. En ese sentido, de la información revisada el modelo de Holt-Winters recibe como inputs de información fecha, venta y producto para realizar la proyección.

En la figura 5 podemos ver las librerías necesarias para la ejecución y la lógica para poder generar el análisis de series de tiempo para los productos. A grandes rasgos se realiza un filtro para quedarnos con las variables mencionados anteriormente y almacenarlas en un dataframe que nos permitirá generar el modelo Holt-Winters por cada producto.

Figura 6: Pseudocódigo Modelo Holt-Winters



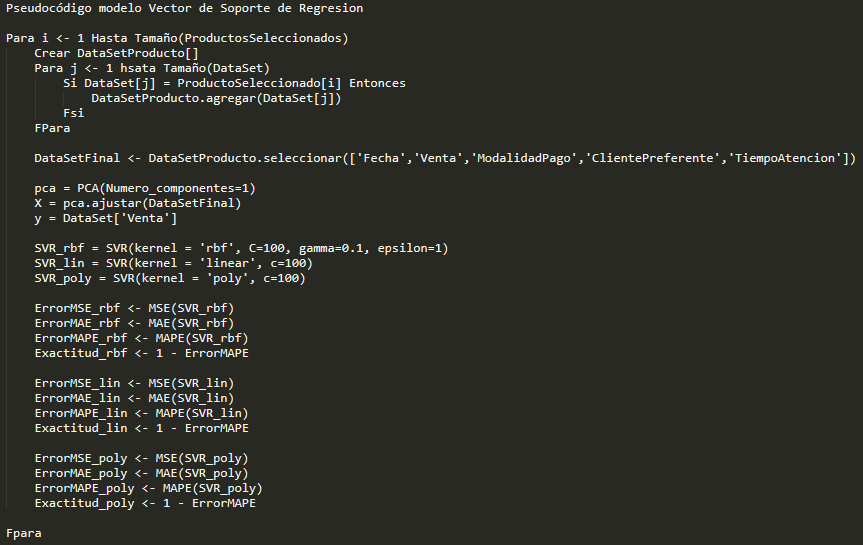
Fuente: Desarrollo Propio

* 1. Aplicación Máquina de Soporte de Regresión

A diferencia de las técnicas de series de tiempo, las máquinas de soporte regresión si necesitan data categórica para tener un poder predictivo más potente, en ese sentido las variables que estamos añadiendo al modelo son las de fecha, venta, modalidad de pago, cliente preferente y tiempo de atención.

Una vez aplicados los filtros necesarios, esta información la almacenamos en un dataframe para su posterior ejecución, como podemos ver en la figura 6, antes de la ejecución del SVR hemos aplicado la técnica de PCA para reducir los componentes a 1 y que la información se pueda graficar de una manera lineal dentro de un gráfico, además con esto nos aseguramos de eliminar la correlación de las variables en caso existiera.

Figura 6: Pseudocódigo Modelo SVR



Fuente: Desarrollo Propio

# RESULTADOS

5.1 ° ARIMA

Aplicando el modelo ARIMA para la información evaluada, (tomando como valor p = 5, d = 1 y q = 1) se obtiene como resultado los pronósticos que se ven en la figura 7, figura 8 y figura 9, de las gráficas se puede observar que la línea azul (pronóstico) busca tener la misma tendencia de la línea naranja (real), se ha realizado un pronóstico general para la bodega y pronóstico para 2 productos de la bodega.

Figura 7. Pronóstico del modelo ARIMA para la bodega

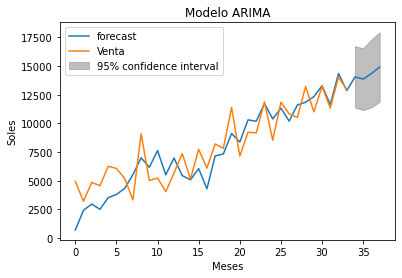
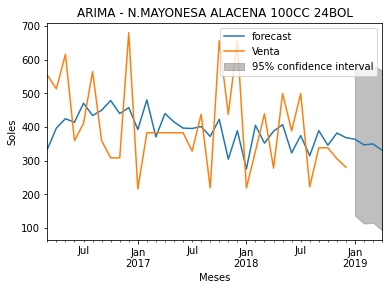


Figura 8. Pronóstico del modelo ARIMA para producto Harina Nicolini Premium



Figura 9. Pronóstico del modelo ARIMA para Mayonesa Alacena



5.2 ° Holt-Winters

Aplicando el método Holt-Winters para la información evaluada, se obtiene como resultado los pronósticos que se ven en la figura 10, figura 11 y figura 12, de la gráfica se puede observar que la línea azul (pronóstico) busca tener la misma tendencia de la línea naranja (real), en este caso la línea de tendencia mostrada es casi idéntica al real dando como resultado del pronóstico una caída para los siguientes 2 meses y después un alza.

Figura 10. Pronóstico de método de Holt-Winters para la bodega

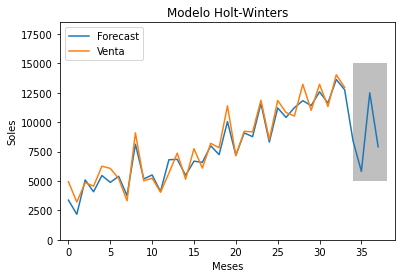


Figura 11. Pronóstico del método de Holt-Winters para producto Harina Nicolini Premium

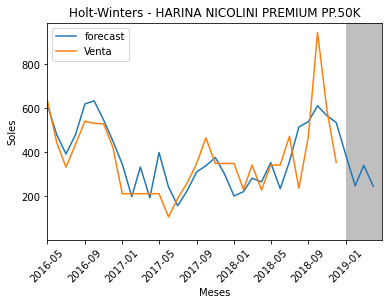
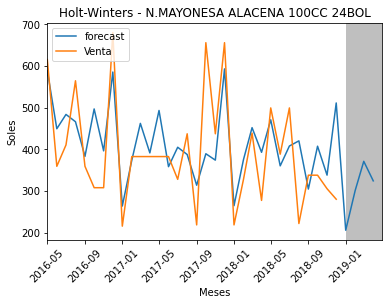


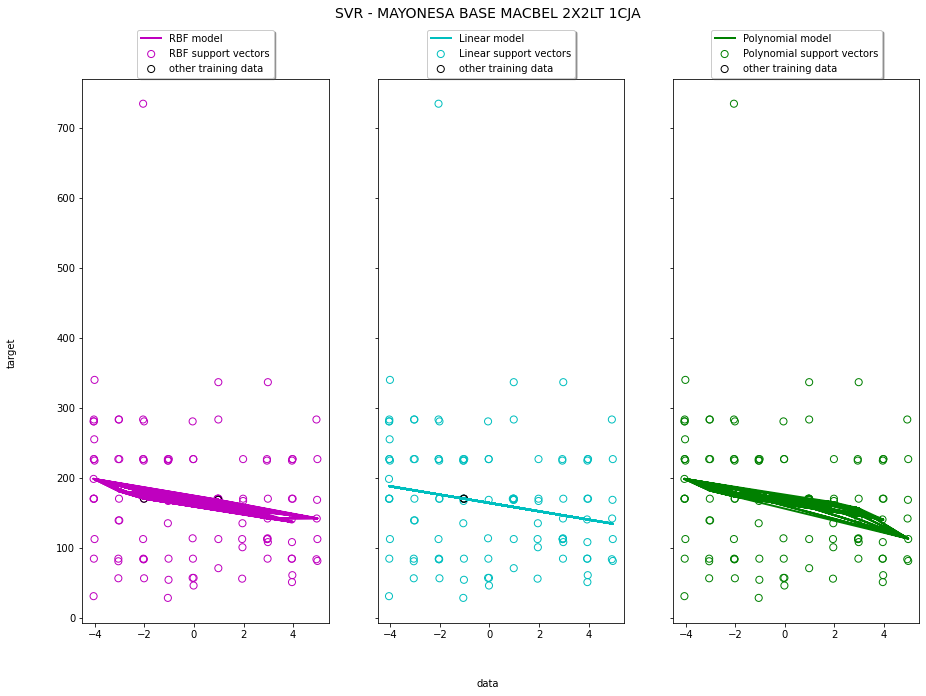
Figura 12. Pronóstico del método de Holt-Winters para Mayonesa Alacena



5.3 Soporte de Vectores de Regresión

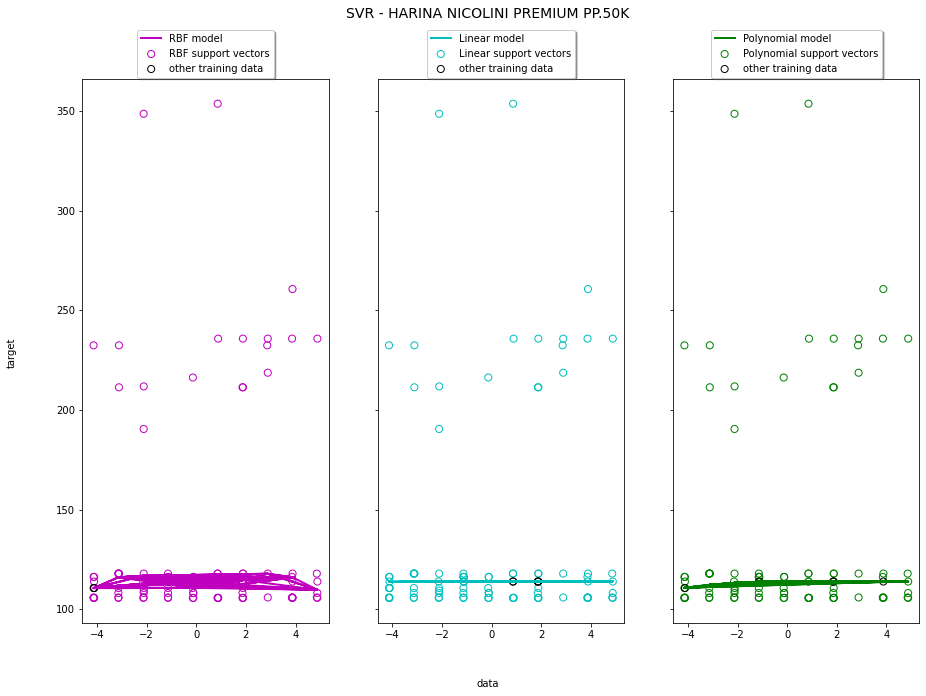
Aplicando la técnica de SVR para los mismos productos hemos podido encontrar los siguientes resultados para 3 opciones del kernel.

Figura 13. Pronóstico del método de SVR para Mayonesa Macbel



En la figura 13 se puede observar una caída con respecto a la proyección para los 3 escenarios aplicados, en ese sentido podemos decir que el producto parece que esta dejando de tener ventas o en su defecto a futuro seguirá vendiendo menos.

Figura 14. Pronóstico del método de SVR para Nueva Manteca Costa



En la figura 14 podemos observar que el producto se mantiene más estable respecto por ejemplo al anterior, en ese sentido, podemos decir que las ventas a futuro de este producto se mantendrán casi iguales en los 3 escenarios.

5.4 ° Evaluación de indicadores

Tabla 4

*Resultados de las técnicas para ARIMA y Holt-Winters para los pronósticos generales.*

|  |  |  |
| --- | --- | --- |
| Técnica | ARIMA | Holt-Winters |
| MAE | 1,650.18 | 883.72 |
| MSE | 4,019,589.75 | 1,164,982.4 |
| MAPE | 0.25 | 0.21 |
| EXACTITUD | 0.75 | 0.79 |

En ambos casos se aplicaron las mismas técnicas para evaluar y comparar los resultados, en este caso de la tabla 4 podemos observar 4 indicadores para ARIMA y para Holt-Winters, revisando los resultados de los indicadores podemos ver que la diferencia entre los pronósticos es poca, solamente 4% en la exactitud que es el indicador principal por el que podemos evaluar los errores en ambos casos.

Con respecto a los demás indicadores podemos ver que el modelo de Holt-Winters los números son inferiores indicando así que los errores presentados para Holt-Winters son más chicos versus ARIMA, esto nos indica mejor precisión.

Evaluando las gráficas 8 y 11 podemos ver que el pronóstico en el caso de Holt-Winters se ajusta más al real, esto a la vez se puede ver con mejor detalle en la tabla 4 con la técnica del MSE que nos dice que el error es menor en Holt-Winters al contrario del modelo ARIMA.

Tabla 5

*Resultado de las técnicas para ARIMA y Holt-Winters para el producto Harina Nicolini Premium*

|  |  |  |
| --- | --- | --- |
| Técnica | ARIMA | Holt-Winters |
| MAE | 350.12 | 86.124 |
| MSE | 49,234.22 | 13,328.636 |
| MAPE | 0.286 | 0.173 |
| EXACTITUD | 0.714 | 0.827 |

Con respecto a los resultados de las figuras 6 y 9 podemos analizar en base a los indicadores obtenidos que en general, el método de Holt-Winters presenta mejores resultados, basados en la exactitud en este resultado si encontramos una diferencia entre ambos dado que el pronóstico obtenido para la exactitud en el caso de Holt-Winters en mayor (7% de diferencia).

Tabla 7

Resultado de las técnicas para ARIMA y Holt-Winters para el producto Mayonesa Alacena

|  |  |  |
| --- | --- | --- |
| Técnica | ARIMA | Holt-Winters |
| MAE | 1,650.18 | 49.877 |
| MSE | 4,019,589.75 | 3,818.511 |
| MAPE | 0.195 | 0.036 |
| EXACTITUD | 0.805 | 0.964 |

Con respecto a los resultados de las figuras 7 y 10, podemos analizar en base a los indicadores obtenido, el método de Holt-Winters presenta resultados mejores resultados en todos los indicadores, la exactitud que se presenta en este caso es alta (15% de diferencia),

# DISCUSIÓN

Basados en la tabla 5 se ve una comparativa de los errores del MSE, MAE y la exactitud, de estos indicadores aplicados a la técnica de Holt-Winters obtenemos un menor error al momento de hacer las predicciones y una mejor exactitud, sin embargo, las diferencias que encontramos versus el modelo ARIMA no son tantas, únicamente tenemos un 4% de diferencia en la exactitud.

“Este método de pronóstico (Holt-Winters) tiene la ventaja de ser fácil de adaptar fácilmente información disponible actualizada” (Chapa & Mireilli, 2018), tal como menciona el autor en el desarrollo de la investigación fue más sencillo implementar este método. Tomando la exactitud con la que se debe de evaluar estos modelos hemos conseguido tener una mayor exactitud usando el método de Holt-Winters, para una investigación parecida el autor nos comenta que “el método de Holt-Winters permite un ahorro de S/. 6,000,000 para la temporada PV 2018” (Chapa & Mireilli, 2018), analizando la gráfica 8 podemos ver que el pronóstico irá de caída en los siguientes 2 meses, en adelante irá subiendo hasta recuperarse.

En el caso de los resultados del modelo ARIMA se nos dice que “los resultados del pronóstico utilizando el método ARIMA tienen una precisión bastante buena en el pronóstico a corto plazo, mientras que para el pronóstico a largo plazo generalmente tenderá a ser plano / constante” (Siregar et al., 2018), en ese sentido los resultados que vemos del modelo ARIMA presentan un mayor error dados los meses que están intentando predecir, además la exactitud que estamos evaluando es menor en comparación que el Holt-Winters. Lo que se observa es que para ARIMA en los meses iniciales se tienen mayor error y conforme se va acercando a los meses 25 a 35 recién ese error se reduce.

Tomando en cuenta el MAPE de la tabla 5 para ambas técnicas, tenemos bajo error para técnicas de series de tiempo, en ambos casos podríamos decir que las dos técnicas cumplen dado que según investigaciones previas se acepta a partir del error menor al 30% y una exactitud mayor a 70%, sin embargo, dada la proximidad de los resultados de exactitud no podemos asegurar que una técnica sea mejor que otra, deberíamos proceder a evaluar a un nivel más bajo para tener resultados más precisos.

Las tablas 6 y 7 responden a esa necesidad de evaluar más a fondo los resultados de precisión entre ARIMA y Holt-Winters, en los 2 productos evaluados podemos ver mejores indicadores en el Holt-Winters, siendo la exactitud para el caso de Holt-Winters alta (superior al 70%) y en ambos casos superior. Es necesario para continuar evaluando esta comparativa reproducir el análisis a más productos.

# CONCLUSIONES

Basados en los resultados obtenidos y lo discutido en el punto anterior podemos concluir lo siguiente:

* El empleo de modelos de series de tiempos para predicciones de demanda en bodegas es el más adecuado debido a que se aísla la estacionalidad que existe dentro de la información, en ese sentido los pronósticos que se obtienen son más adecuados y presentan una mejor exactitud versus otros modelos existentes.
* Durante el análisis de los resultados, respecto a la aplicación de las técnicas de series de tiempo de ARIMA y Holt-Winters (en el modelo de la bodega general) podemos decir en base a la exactitud de ambos que el modelo de Holt-Winters presenta una mejor exactitud (4% arriba respecto ARIMA), sin embargo, ambos son aceptados dado que la exactitud en los dos casos es mayor al 70%. En ese sentido y apoyo por los resultados de exactitud que se obtienen de los productos vemos que el modelo de Holt-Winters en ambos sacos presenta mejor exactitud, sin embargo, para continuar evaluando esta comparativa es necesario reproducir este análisis a más productos.
* El modelo de pronóstico de series de tiempo más efectivo para la demanda de productos envasados de una bodega es el método de Holt-Winters con una exactitud en sus pronósticos del 79% (4% más que el modelo ARIMA), sin embargo, la diferencia al ser poca no es posible asegurar que el modelo sea mejor que el otro, para continuar evaluando esta comparativa es necesario reproducirla a nivel de los productos.
* Se implementó el método de Holt-Winters y el modelo ARIMA y se obtuvo para ambas técnicas la predicción de productos envasados siendo el método de Holt-Winters 4% más exacto que el modelo ARIMA.
* El empleo de modelo de Holt-Winters llevado a la práctica tiene la ventaja de ser de mayor adaptabilidad a la información que se encuentre, además, para su desarrollo es mucho más sencillo, esto es un plus durante el desarrollo de la investigación dado que facilita su traspaso a diferentes situaciones.

(Kavitha et al., 2017)(J. Y. Wu, 2017)(Q. Wu et al., 2019)

Agradecimientos especiales al profesor José Taquia Gutiérrez por el apoyo y las consultas absueltas durante el desarrollo de esta investigación, a la profesora Rosario Guzmán Jiménez por ser una muy buena supervisora académica en el proceso de desarrollo y, además, por su ayuda para poder captar todos mensajes. Asimismo, al docente Guillermo Zevallos Luna-Victoria por ser mi guía en el proceso de desarrollo del presente trabajo.

**REFERENCIAS**

Aziz, A., Isti Khomah, K. N., & Yohanes, S. P. (2018). Prediction the Price of National Groceries Using Average Based Fuzzy Time Series with Song - Chissom and Markov Chain Approach. *Proceedings of ICAITI 2018 - 1st International Conference on Applied Information Technology and Innovation: Toward A New Paradigm for the Design of Assistive Technology in Smart Home Care*, 129–134. https://doi.org/10.1109/ICAITI.2018.8686737

Chapa, P., & Mireilli, J. (2018). *Demanda Y Gestión De Inventarios Para La Planeación De Demanda En Prendas De Vestir Juvenil*. 0–143.

Ferenti, T. (2018). *Biomedical applications of time series analysis*. 000083–000084. https://doi.org/10.1109/nc.2017.8263256

Kavitha, S., Varuna, S., & Ramya, R. (2017). A comparative analysis on linear regression and support vector regression. *Proceedings of 2016 Online International Conference on Green Engineering and Technologies, IC-GET 2016*. https://doi.org/10.1109/GET.2016.7916627

Nataraj, S., Alvarez, C., Sada, L., Juan, A. A., Panadero, J., & Bayliss, C. (2020). Applying Statistical Learning Methods for Forecasting Prices and Enhancing the Probability of Success in Logistics Tenders. *Transportation Research Procedia*, *47*, 529–536. https://doi.org/10.1016/j.trpro.2020.03.128

Sato, M., Izumo, H., & Sonoda, T. (2016). Modeling individual users’ responsiveness to maximize recommendation impact. *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 259–267. https://doi.org/10.1145/2930238.2930259

Siregar, B., Nababan, E. B., Yap, A., Andayani, U., & Fahmi. (2018). Forecasting of raw material needed for plastic products based in income data using ARIMA method. *Proceeding - 2017 5th International Conference on Electrical, Electronics and Information Engineering: Smart Innovations for Bridging Future Technologies, ICEEIE 2017*, *2018*-*Janua*, 135–139. https://doi.org/10.1109/ICEEIE.2017.8328777

Sugiarto, V. C., Sarno, R., & Sunaryono, D. (2017). Sales forecasting using Holt-Winters in Enterprise Resource Planning at sales and distribution module. *Proceedings of 2016 International Conference on Information and Communication Technology and Systems, ICTS 2016*, 8–13. https://doi.org/10.1109/ICTS.2016.7910264

Wu, C. S. M., Patil, P., & Gunaseelan, S. (2019). Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *2018*-*Novem*, 16–20. https://doi.org/10.1109/ICSESS.2018.8663760

Wu, J. Y. (2017). Housing Price prediction Using Support Vector Regression. *Master’s Projects*, 540. https://doi.org/10.31979/etd.vpub-6bgs

Wu, Q., Zhang, H., Jing, R., & Li, Y. (2019). Feature Selection Based on Twin Support Vector Regression. *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019*, *0*(3), 2903–2907. https://doi.org/10.1109/SSCI44817.2019.9003001