FACULTAD DE INGENIERIA, ARQUITECTURA Y URBANISMO

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERIA DE SISTEMAS

TESIS

ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE VENTAS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

AUTOR

Bach. IRENE LEYDI ROQUE MONTALVO

Pimentel, 09 de noviembre de 2016

ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE VENTAS

Aprobación de la tesis

Irene Leydi Roque Montalvo
Autora

Ing. Heber Iván Mejía Cabrera
Presidente del jurado de tesis

Mg. Rosa América Cobeñas Sánchez
Secretario del jurado de tesis

Ing. Denny John Fuentes Adrianzén
Asesor Especialista/Vocal del jurado de tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DEDICATORIA

A:

Dios*, por darme la oportunidad de vivir y por estar conmigo en cada paso que doy, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo el periodo de estudio.*

Mis padres, *por haberme apoyado en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido ser una persona de bien, pero más que nada, por su amor.*

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

AGRADECIMIENTOS

A mis padres, por su comprensión, paciencia y ánimo.

A mi asesor, por su orientación y colaboración en el desarrollo de este trabajo de investigación.

Al Dr. Jorge Narváez Restelli, Jefe del Departamento de Ayuda al diagnóstico y tratamiento del Hospital II "Luis Heysen Incháustegui", por su apoyo incondicional y desinteresado durante toda mi formación académica.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Contenido

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Figuras

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Gráficos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Tablas

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

RESUMEN

El presente trabajo denominado "**ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE VENTAS**" realiza un análisis comparativo entre las distintas técnicas usadas en la minería de datos para el diseño de modelo de pronósticos de series de tiempo. En la actualidad existen diversas técnicas para la generación de pronósticos de series de tiempo, desde los modelos de tipo estadísticos, o los más avanzados que usan algoritmos computacionales basados en inteligencia artificial como es el caso de las redes neuronales o las máquinas de soporte vectorial.

El problema no trata sobre la construcción de un modelo de minería de datos, si no de evaluar que algoritmo y técnica sirve o tiene un mejor performance para un problema determinado, ya que no es lo mismo aplicar criterios de pronósticos a series de tipo ventas, que, para series de clima, u otros. Donde cada algoritmo tiene un grado de influencia según el problema a enfocarse.

El ámbito de estudio de esta investigación se centra en la empresa "El Astro S.A.C." para determinar las estimaciones de ventas según el volumen que genera mensual o trimestral, uno de los algoritmos más usados para pronósticos de ventas desde el punto de vista estadístico es el Holtwinters, en esta investigación se realizará un análisis de los datos para comparar este algoritmo contra otros métodos como son: Holt y ETS.

PALABRAS CLAVES

Minería de Datos, Pronósticos, Holtwinters, Holt, ETS.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ABSTRACT

This paper called **"COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR PREDICTING SALES"** makes a comparative analysis between the different techniques used in data mining to design forecasting model time series. At present, there are various techniques for generating time series forecasts from statistical type models, or using the most advanced computational algorithms based on artificial intelligence such as neural networks or support vector machines.

The problem is not about building a data mining model, if not assess that algorithm and technique serves or has a better performance for a given problem, since it is not the same criteria applied to series-type forecasts sales that for series of weather, or other. Where each algorithm has a degree of influence by the problem into focus.

The scope of this research focuses on the company "El Astro SAC" to determine estimates of sales by volume generated monthly or quarterly, one of the widely used algorithms for sales forecasts from a statistical point of view is the Holtwinters, this research data analysis is performed to compare this algorithm with other methods such as: Holt and ETS.

KEYWORDS

Data Mining, Forecasts, Holtwinters, Holt, ETS.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INTRODUCCIÓN

El desarrollo de sistemas informáticos en la actualidad ha tenido un crecimiento exponencial, en los últimos años con los nuevos dispositivos tecnológicos ha sido necesario dividir el ámbito del desarrollo de software, términos como el desarrollo web, desarrollo móvil, soluciones de inteligencia de negocios son cada vez más vistos, dentro del software de inteligencia de negocios destaca el surgimiento de una técnica denominada “Minería de Datos”.

La minería de datos es una técnica matemático-computacional que usa principios estadísticos para hacer la explotación y análisis de datos en sistemas informáticos, los cuales están en bases de datos, el objetivo de la minería de datos es detectar patrones de comportamiento en estos datos para generar probabilidades, predicciones, pronósticos y análisis descriptivo avanzado, que permitan explicar el fenómeno de negocio de la empresa.

Uno de estos sistemas informáticos es el software comercial de la empresa “El Astro S.A.C.”, esta empresa se dedica a la venta de artículos deportivos, su base de datos básicamente engloba el proceso de ventas, por lo tanto, es necesario para la tienda conocer los consolidados de la misma, así como realizar estimaciones o pronósticos.

Este problema puede ser resuelto empleando las técnicas de minería de datos diseñando un modelo de pronósticos de series de tiempo, sin embargo existen múltiples algoritmos para diseñar este modelo, cada algoritmo tiene particularidades que influyen en el pronóstico, por lo tanto es necesario evaluar estos algoritmos con el fin de determinar cuál es el que tiene mejor performance para obtener un pronóstico que se asemeje a lo que ocurre cuando se da el fenómeno de artículos deportivos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO I

EL PROBLEMA DE INVESTIGACIÓN

CAPÍTULO I: PROBLEMA DE LA INVESTIGACIÓN

1.1. Situación Problemática

Desde hace algunas décadas, el hombre se ha visto en la necesidad de administrar sus actividades, la mayoría de éstas comerciales, como lo son el uso al que se le da al dinero tanto en el hogar como a nivel empresarial, por lo tanto le es necesario almacenar un historial de algunas o la mayoría de sus actividades comerciales, lo que lo obliga a llevar de manera ordenada el cómo y en qué ha gastado su dinero, hasta el punto de ser necesario contar con una persona que se dedique a administrar, almacenar y vigilar dichas actividades a nivel empresarial.

Con el paso del tiempo se ha visto que para dar una adecuada administración de todas esas actividades y con el fin de evitar muchos conflictos, en la mayoría de los lugares como por ejemplo en los hospitales, se realizan historial de visitas, entradas y salidas de pacientes; en las estaciones de policía se registran con hora y fecha exactas los hechos sucedidos; en almacenes grandes se registran las transacciones en facturas con fecha de compra, entre otros ejemplos; por lo que se comienza a formar una generación masiva de datos los cuales llevan a la creación de almacenes o bodegas de datos, algunos con un crecimiento tan exagerado que hasta para las consultas realizadas por lenguajes como SQL es imposible lograr resultados eficientes (García Bermúdez & Acevedo Ramírez, 2010).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Los algoritmos de Minería de Datos realizan en general tareas de predicción de información desconocida que puede estar contenida en los datos, como también puede realizar la labor de describir patrones de comportamiento de los datos.

1.2. Formulación del problema

En este campo de la minería de datos se han realizado importantes avances específicamente en técnicas de algoritmos de predicción, sin embargo se están realizando investigaciones sobre técnicas de minería de datos para predecir demanda o determinar patrones de comportamiento. (Fernández Maturana, 2007) en su investigación Wavelet-and SVM-based forecasts: An analysis of the U.S. metal and materials manufacturing industry" estudio el desempeño de los pronósticos para lo cual utilizo cuatro técnicas de estimación: ARIMA multiplicativo estacional, componentes no observables (UC), wavelets (ondas cortas) y support vector machines (SVM) determinando que ARIMA y UC superan en exactitud a Wavelets y SVM, así como también (Calvo Rodríguez, 2008) en su investigación Predicción en Series de Tiempo con Modelos Aditivos utilizarón modelos aditivos para regresión múltiple y el algoritmo backfitting para su estimación y búsqueda de intervalos de confianza, obteniendo como resultado que Modelo Aditivo tiene un margen de error del 2,77% llegando a la conclusión que este modelo sirve para solucionar el problema de dimensionalidad; por otro lado (Madrigal Espinoza, 2006) en su tesis Modelos de espacio de estados subyacentes al método multiplicativo de HoltWinters desarrolló un método de pronóstico basado en suavización exponencial que incorpore el cálculo de intervalos de predicción, para una serie de tiempo que presenta tendencia aditiva y

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

múltiples estacionalidades multiplicativas en el cual se obtuvo un 85% de confiabilidad al utilizar el método multiplicativo de HoltWinters.

A pesar de que estos algoritmos obtuvieron en promedio hasta un 85 % de efectividad para pronósticos aún resta mejorar esas tasas de acierto es por ello que en esta investigación se pretende realizar un análisis de rendimiento comparativo de las diferentes técnicas de minería de datos como son las series temporales, seleccionando la que mejor se adapta para la predicción de ventas orientado en el sector comercialización de artículos deportivos tomando como referencia la empresa El Astro SAC ubicada en la calle Alfredo Lapoint Nro. 1189 Chiclayo. Las técnicas propuestas en este caso son ARIMA, HoltWinters, Holt y ETS por ser las que más se adaptan a este trabajo.

1.3. Delimitación de la Investigación

Esta investigación está enfocado al análisis comparativo de técnicas de minería de datos para predicción usando algoritmos de series de tiempo como son Holtwinters, Holt y ETS.

El sistema contempla dos módulos, uno para visualizar los datos estadísticos del proceso de ventas y otro modulo que tiene la característica de presentar los datos del modelo de minería, así como también una interfaz de simulación para evaluar el comportamiento con distintos algoritmos propios de la investigación.

1.4. Justificación e Importancia

El motivo de desarrollo de esta investigación es que existe un problema real en la capacidad de procesar grandes cantidades

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

de datos, los cuales generan las áreas operativas de cada empresa; problema que puede ser resuelto con la aplicación de algoritmos de minería de datos. (Asencios, 2004).

Desde el punto de vista social, las técnicas para minería de datos devienen necesarias en diferentes áreas de la vida diaria, tales como economía, el nivel empresarial, la salud, la investigación científica, etcétera; en estas áreas generalmente existe una gran cantidad de datos generados permanentemente por los sistemas de información transaccionales, y que no son analizados o en su defecto han sido analizados solo parcialmente; conteniendo gran cantidad de información que aún no ha sido debidamente evaluada y que constituye la base para una correcta definición empresarial.

Desde el punto de vista tecnológico, la presente investigación se justifica entonces por el impacto que representa estudiar los algoritmos de minería de datos en la solución de problemas donde se requiere el uso de grandes repositorios de datos para convertirlos en información útil, usando para ello técnicas avanzadas de minería de datos; las cuales serán evaluadas para medir su grado de efectividad.

Desde el punto de vista académico, la presente investigación es importante por su aporte al estudio y conocimiento de las técnicas de minería de datos.

1.5. Limitaciones de la Investigación

Los datos históricos extraídos de la base de datos del sistema comercial de “El Astro S.A.C. comprenden el periodo Agosto 2011 – Junio 2014 por lo tanto existe información cuantificada mensual equivalente a 35 meses.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

1.6. Objetivos de la Investigación

1. Objetivos de la Investigación

Realizar un análisis comparativo del rendimiento de las técnicas de minería de datos para la predicción de ventas orientado a la comercialización de artículos deportivos.

2. Objetivos específicos

a) Realizar una evaluación de las técnicas de minería de datos existentes en la actualidad.

b) Seleccionar las técnicas de minería de datos para la predicción de ventas para realizar el análisis comparativo.

c) Seleccionar un modelo de evaluación para elaborar la comparación entre las técnicas de minería de datos para la predicción de ventas.

d) Desarrollar la aplicación que muestre los resultados de las técnicas de minería de datos seleccionada.

e) Analizar los resultados.

f) Realizar la evaluación económica de la propuesta.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO II

MARCO TEÓRICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes de estudios:

A. WAVELET-AND SVM-BASED FORECASTS: AN ANALYSIS OF THE U.S. METAL AND MATERIALS MANUFACTURING INDUSTRY" (Fernández Maturana, 2007) esta investigación estudia el desempeño de los pronósticos de cuatro modelos de series temporales: ARIMA multiplicativo estacional, componentes no observables (UC), wavelets (ondas cortas) y support vector machines (SVM), obteniendo como resultados que los métodos tradicionales de ARIMA y UC suelen ser más exactos sin embargo, wavelets y SVM pueden contener información adicional a aquella contenida en los pronósticos de ARIMA y UC. En consecuencia, las combinaciones lineales de pronósticos pueden ser preferibles a los pronósticos individuales. En términos generales, los resultados muestran que ARIMA y UC suelen superar, en valor de predicción del 5%, a SVM y wavelets, mientras que UC es superior a ARIMA en un nivel de significancia del 1%. Llegando a las siguientes conclusiones: En primer término el horizonte temporal escogido es un elemento clave para decidir qué modelo o combinación lineal de modelos es preferible, en términos de la calidad de pronóstico. En especial, para pronósticos de mediano y más largo plazo, parece ser que el test de Harvey, Leybourne y Newbold discrimina mejor entre dos modelos competitivos, sobre la base del error cuadrático medio y del error medio en valor absoluto. En segundo término, en general, ARIMA y UC suelen superar a wavelets y SVM. Sin embargo, en varios casos los dos últimos pueden contener información adicional a aquella contenida en los pronósticos de los dos primeros. En consecuencia, las combinaciones lineales de

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

pronósticos pueden ser preferibles a los pronósticos individuales. Por último, en general, la información proporcionada por SVM resulta de mayor utilidad cuando se consideran combinaciones lineales de SVM y wavelets. No obstante, SVM puede superar a UC y ARIMA.

B. MODELOS DE ESPACIO DE ESTADOS SUBYACENTES AL MÉTODO MULTIPLICATIVO DE HOLT-WINTERS CON MÚLTIPLE ESTACIONALIDAD (Madrigal Espinoza, 2006) Esta investigación tuvo como objetivo desarrollar un método de pronóstico basado en suavización exponencial que incorpore el cálculo de intervalos de predicción, para una serie de tiempo que presenta tendencia aditiva y múltiples estacionalidades multiplicativas, aplicando el método multiplicativo de Holt-Winters los resultados obtenidos al minimizar la función objetivo propuesta para el método multiplicativo de Holt-Winters de 2 y 8 semestres para el caso de errores aditivos (= 0) y multiplicativos (= 1). Ambos criterios de selección favorecieron al mismo modelo, que en este caso es el de errores aditivos; el margen de error arrojado por este método es cinco veces menor al de errores multiplicativos llegando a la conclusión que, el modelo con errores multiplicativos muestra muy buen ajuste para los primeros datos un 86% de confiabilidad,. Sin embargo, a más cantidad de datos (24 periodos), deja de ajustarse con la precisión que lo hizo al principio y de hecho, conforme pasa el tiempo, el ajuste parece ir empeorando.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

C. TÉCNICAS DE MINERÍA DE DATOS APLICADAS A LA CONSTRUCCIÓN DE MODELOS DE SCORE CREDITICIO

(Ramírez A., 2007) en su trabajo realizo la comparación de diferentes técnicas como son las de redes neuronales, análisis discriminante, máquinas vectoriales de soporte, árboles de decisiones y regresión logística, las cuales son más empleadas de score crediticio para así determinar patrones de comportamiento de un cliente como resultados se obtuvo que dentro de las técnicas tradicionales, la Regresión Lineal es la que mejor exactitud en predicción muestra. Sin embargo, las Redes Neuronales muestran una mayor exactitud en cada una de las ejecuciones que los autores realizan. En general, los modelos presentados predicen mejor los buenos créditos que los malos con excepción de dos casos de Análisis Discriminante (DA). El mejor predictor para la clase buena (i.e. buen crédito) es una red neuronal probabilística (Probabilistic Neural Networks, PNN) mientras que una Red Neuronal Multicapa Feed-Forward (Multi-Layer Feed Forward Network, MLFN) predice mejor los créditos malos Por otra parte, se ha realizado una comparación entre los distintos modelos en donde los modelos sobresalientes son las redes neuronales y las máquinas vectoriales de soporte (SVM). Sin embargo, no se desconoce el esfuerzo y la precisión presentada por modelos híbridos. Estos modelos híbridos han mostrado mejores resultados frente a modelos tradicionales, sin embargo dichos modelos no son contundentes y la diferencia entre estos y los tradicionales suele ser muy baja.

Por otra parte, de acuerdo a lo que han reportado varios autores, las consideraciones importantes para construir un buen modelo radica en la calidad de los datos escogidos y en la selección adecuada de las variables que influyen en los modelos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

D. **APLICACIÓN DE MINERÍA DE DATOS PARA LA EXPLORACIÓN Y DETECCIÓN DE PATRONES DELICTIVOS EN ARGENTINA** (Perversi, 2007) en su investigación realizo la identificación y detección de patrones de homicidios dolosos vinculados con el tipo de arma empleada, para tal cometido utilizaron los algoritmos K-means (para agrupar los hechos según su similitud), Inducción, ID3, C4.5 (para identificar reglas de pertenencia). Obtuvo como resultado que el 98,8% de instancias bien clasificadas confirma que los clústeres determinados por K-means responden a un criterio determinado subyacente a los datos.

E. **La investigación MINERÍA DE DATOS APLICADA EN LA DETECCIÓN DE INTRUSOS realizado por** (Vallejo Pérez, 2012) partiendo de una base de datos con captura de conexiones realizada para el "The Third International Knowledge Discovery and Data Mining Tools Competition", y basados en la metodología CRISP-DM, pretendieron mostrar mediante diferentes técnicas de modelado que se basan en cálculos estadísticos, como la aplicación de la minería de datos sirve para analizar y procesar grandes volúmenes de datos de una manera predictiva, con miras a entregar unos resultados de manera oportuna, eficiente, eficaz y confiable, entregando información que ayuda en la toma de decisiones en la prevención de posibles intrusiones.

Utilizaron las técnicas de modelado que se basan en el uso de algoritmos, los modelos evaluados o utilizados en su caso de estudio son: El modelo CHAID, el modelo ARBOL C&R y el modelo C 5.0.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Luego de evaluar cada uno de los modelos según la necesidad esperada, se procede a revisar los resultados entregados por el modelo, es de esta manera como se obtiene un mejor acercamiento a los datos entregados. El modelo que mejor comportamiento tiene en evaluar la variable “Ataque” vs las otras variables de entrada fue el modelo CHAID. Las conclusiones a las que llegaron es que la minería de datos basada en una metodología adecuada, puede ser muy útil en el proceso de exploración de datos, toda vez que mediante tecnologías analíticas y procesos estadísticos nos permitió generar reglas que a partir de datos históricos de capturas, para generar reglas y patrones que permiten predecir intrusiones.

F. La investigación APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL presentada por (Martínez Álvarez, 2012) tuvo por objetivo reducir el tiempo de cálculo de los indicadores de servicios privados de ENTEL Chile, para lo cual aplicó la técnica de minería de datos Clustering usando los algoritmos K-means y técnicas de clasificación como árbol de decisiones y redes neuronales; con la finalidad de identificar relaciones entre los datos de clientes y servicios, generando conocimiento para la empresa a través de la caracterización de las preferencias de cada uno.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.2. Estado del arte

A. **WAVELET-AND SVM-BASED FORECASTS: AN ANALYSIS OF THE U.S. METAL AND MATERIALS MANUFACTURING INDUSTRY"** (Fernández Maturana, 2007) En esta investigación se estudia el desempeño de los pronósticos de cuatro modelos de series temporales: ARIMA multiplicativo estacional, componentes no observables (UC), wavelets (ondas cortas) y support vector machines (SVM), obteniendo como resultados que los métodos tradicionales de ARIMA y UC suelen ser más exactos sin embargo, wavelets y SVM pueden contener información adicional a aquella contenida en los pronósticos de ARIMA y UC. En consecuencia, las combinaciones lineales de pronósticos pueden ser preferibles a los pronósticos individuales. En términos generales, los resultados muestran que ARIMA y UC suelen superar, en valor de predicción del 5%, a SVM y wavelets, mientras que UC es superior a ARIMA en un nivel de significancia del 1%. Llegando a las siguientes conclusiones: En primer término el horizonte temporal escogido es un elemento clave para decidir qué modelo o combinación lineal de modelos es preferible, en términos de la calidad de pronóstico. En especial, para pronósticos de mediano y más largo plazo, parece ser que el test de Harvey, Leybourne y Newbold discrimina mejor entre dos modelos competitivos, sobre la base del error cuadrático medio y del error medio en valor absoluto. En segundo término, en general, ARIMA y UC suelen superar a wavelets y SVM. Sin embargo, en varios casos los dos últimos pueden contener información adicional a aquella contenida en los pronósticos de los dos primeros. En consecuencia, las combinaciones lineales de pronósticos pueden ser preferibles a los pronósticos individuales. Por

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

último, en general, la información proporcionada por SVM resulta de mayor utilidad cuando se consideran combinaciones lineales de SVM y wavelets. No obstante, SVM puede superar a UC y ARIMA.

B. **PREDICCIÓN EN SERIES DE TIEMPO CON MODELOS ADITIVOS** (Calvo Rodríguez, 2008), esta investigación presenta un algoritmo para la predicción en series de tiempo y búsqueda de intervalos de confianza, utilizarón los modelos aditivos para regresión múltiple y el algoritmo backfitting para su estimación. El método utilizó la estadística no paramétrica de forma que sólo se trabaje con funciones unidimensionales, ya que se sabe que cuando se trabaja con dimensiones altas los métodos son costosos y, en ocasiones, dan problemas. Se obtuvo como resultado que Modelo Aditivo tiene un margen de error del 2,77% llegando a la conclusión que este modelo sirve para solucionar el problema de dimensionalidad.

C. **UNA INVESTIGACIÓN BASADA EN MINERÍA DE DATOS: PREDICCIÓN DE VENTAS PARA COMPAÑÍAS FARMACÉUTICAS DE DISTRIBUCIÓN** realizado por (Zadeh, 2014) en esta investigación propone determinar un método de minería de datos como combinación de herramientas para análisis de red y series de tiempo; con la finalidad de controlar los niveles de inventario en compañías farmacéuticas de distribución PDCs para prevenir el costo excesivo de inventario y prevenir además la pérdida de clientes. Los modelos se construyen a partir de la metodología ARIMA.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

D. MODELOS DE ESPACIO DE ESTADOS SUBYACENTES AL MÉTODO MULTIPLICATIVO DE HOLT-WINTERS CON MÚLTIPLE ESTACIONALIDAD (Madrigal Espinoza, 2006) Esta investigación tuvo como objetivo desarrollar un método de pronóstico basado en suavización exponencial que incorpore el cálculo de intervalos de predicción, para una serie de tiempo que presenta tendencia aditiva y múltiples estacionalidades multiplicativas, aplicando el método multiplicativo de Holt-Winters los resultados obtenidos al minimizar la función objetivo propuesta para el método multiplicativo de Holt-Winters de 2 y 8 semestres para el caso de errores aditivos (= 0) y multiplicativos (= 1). Ambos criterios de selección favorecieron al mismo modelo, que en este caso es el de errores aditivos; el margen de error arrojado por este método es cinco veces menor al de errores multiplicativos llegando a la conclusión que, el modelo con errores multiplicativos muestra muy buen ajuste para los primeros datos. Sin embargo, a más cantidad de datos (24 periodos), deja de ajustarse con la precisión que lo hizo al principio y de hecho, conforme pasa el tiempo, el ajuste parece ir empeorando.

2.3. Bases Teórico Científicas

Se presentan los conocimientos o bases teóricas que serán empleadas a lo largo de la investigación.

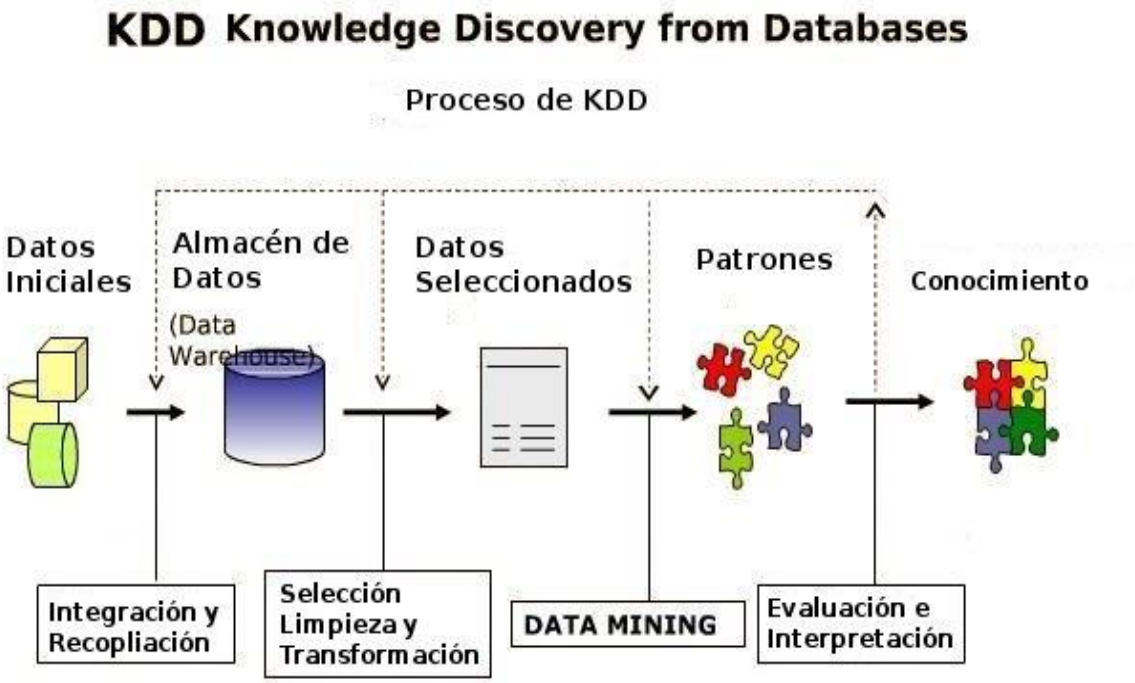
A. Proceso KDD para la obtención del conocimiento

Según (Molina Neyra & Murakami de la Cruz, 2008) KDD (Knowledge Discovery in Databases) es una metodología genérica para encontrar información en un gran conjunto de datos y con ello generar conocimiento. Se define como un proceso no trivial de extracción de información a partir de los

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

datos, la cual se encuentra presente de forma implícita, previamente desconocida y potencialmente útil para el usuario o para el negocio. El objetivo principal de esta metodología es automatizar el procesamiento de los datos, permitiendo a los usuarios dedicar más tiempo a las tareas de análisis y al descubrimiento de relaciones entre los datos. El KDD es un proceso que consta de una serie de etapas consecutivas, y funciona de forma iterativa e interactiva. Iterativa, ya que es posible regresar desde cualquier etapa a una anterior para ajustar los parámetros o supuestos previos, e interactiva pues el usuario experto del negocio tiene que estar presente para aportar con su conocimiento en la preparación de los datos y en la validación de los resultados que se obtengan durante el proceso.

Figura 1: Etapas del Proceso KDD



Fuente: *(Molina Neyra & Murakami de la Cruz, 2008)*

Las etapas de este proceso son:

Identificación del problema en estudio, teniendo un objetivo claro para el problema a resolver, entendiendo las metas del proceso y cuáles son las preguntas que se quieren responder.

Selección e integración de los datos, para contar con un conjunto objetivo desde el cual obtener el conocimiento. Se obtienen los datos desde los sistemas operacionales, los cuales pueden venir en diferentes formatos y en algunas oportunidades con errores, por lo cual es importante realizar una etapa de procesamiento.

Preparación de los datos (limpieza y pre-procesamiento), ya que en general, como se dijo en la etapa anterior, los datos provienen desde varias fuentes y en diferentes formatos. En esta etapa se escogen técnicas y estrategias para corregir errores en el conjunto de datos seleccionado, tratar la información faltante y unificar formatos.

Transformación y almacenamiento de los datos, punto en el que se pueden reducir o agrupar los datos en las características de interés. Se consolida la información y escoge una arquitectura acorde a las necesidades del problema que permita almacenarla, por ejemplo, un Data Mart.

Selección y aplicación de algoritmos de Data Mining, utilizando técnicas adecuadas según la hipótesis planteada y el análisis que se quiera hacer. Las técnicas seleccionadas permitirán generar modelos de minería de datos, y con ello descubrir patrones de información implícitos en los datos.

Interpretación y evaluación de los patrones encontrados, identificando los nuevos conocimientos y apoyándose en los expertos del negocio para ver si se pueden tomar acciones con estos resultados. Para interpretarlos, es necesario visualizarlos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

de diversas formas, validando los patrones y modelos de datos, documentando los procedimientos y consideraciones de manera que se generen propuestas de valor para el negocio.

B. Minería de datos

Según (Dandretta, 2002) dice:

La minería de datos está conformada por un conjunto de técnicas y algoritmos que sirven para hacer análisis de conjuntos de datos, extrayendo patrones y relaciones entre ellos, convirtiéndolos en información valiosa y útil para quienes toman las decisiones.

El uso potencial del Data Mining en las empresas es identificar nuevas oportunidades de negocio, adaptar los productos ofrecidos o encontrar los clientes más valiosos con el fin de retenerlos, y de esta manera aumentar los ingresos y reducir las pérdidas o costos. Al determinar las características de los buenos clientes, las empresas pueden enfocarse en aquellos de características similares y diseñar productos o servicios acordes a sus necesidades. Dentro de la industria de telecomunicaciones por ejemplo, las áreas de interés donde aplicar minería de datos son: detección de fraude, asignación de recursos para instalaciones o servicios técnicos, análisis de clientes, pronósticos de demanda, proyecciones de crecimiento de la industria y predicción de fallas en la red. Los algoritmos que destacan en estos casos son los de regresión, clustering y clasificación. El uso de minería de datos se debe entender como un apoyo para los analistas, y no reemplaza al conocimiento que tienen los expertos del negocio, ni elimina la necesidad de entender los datos. El Data Mining no funciona por sí sólo, ya que los patrones que se encuentren en los datos deben ser

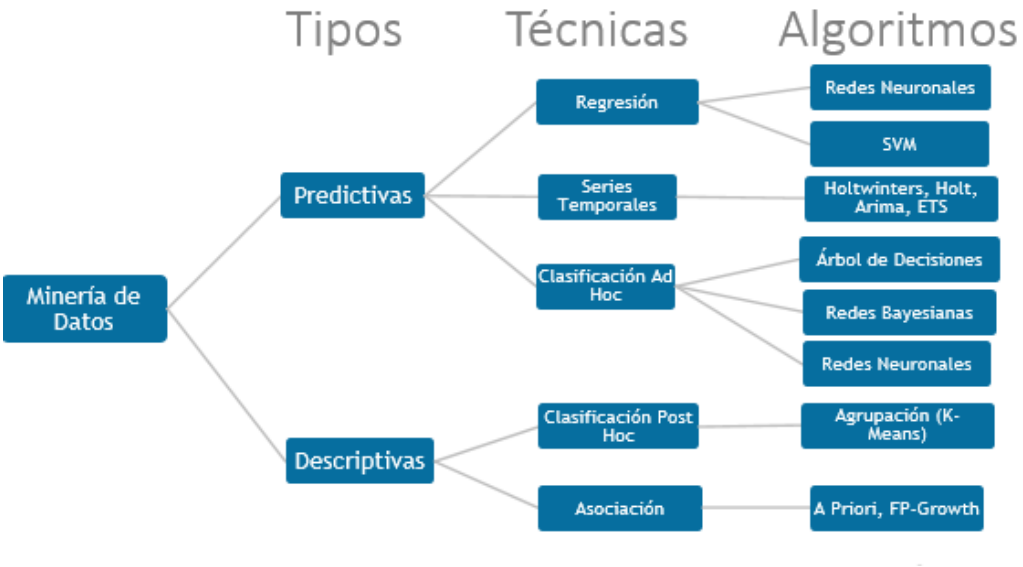
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

interpretados y validados para ver si responden a las consultas del negocio, y si son aplicables en el mundo real.

C. Técnicas de minería de datos:

En la figura N° 2 nos describe las técnicas de la minería de datos.

Figura 2: Técnicas de Minería de Datos



Fuente (Molina López & García Herrero, 2006)

(Molina López & García Herrero, 2006) Las técnicas de Minería de Datos tratan de obtener patrones o modelos a partir de los datos recopilados. Una técnica constituye el enfoque conceptual para extraer la información de los datos.

Cada algoritmo representa, la manera de desarrollar una determinada técnica paso a paso. Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones. De esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos, por lo que la figura N° 2 representa para qué propósito son más utilizadas las técnicas. Por ejemplo, las redes de neuronas pueden servir para predicción, clasificación e incluso para aprendizaje no supervisado.

El aprendizaje inductivo no supervisado estudia el aprendizaje sin la ayuda del maestro; es decir, se aborda el aprendizaje sin supervisión. Según (Valcárcel Asencios, 2004) dice: La aplicación de técnicas de data mining en grandes bases de datos persiguen los siguientes resultados:

Regresión: (Valcárcel Asencios, 2004) Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística). Las regresiones se pueden utilizar por ejemplo para predecir comportamiento de la demanda futura, utilizando las ventas pasadas.

Series Temporales: (Ortíz Farro, 2015) Es el conocimiento de una variable a través del tiempo, para que a partir de ese conocimiento y con el supuesto de que no se producirán cambios, poder realizar predicciones. Llamamos Serie de Tiempo a un conjunto de mediciones de cierto fenómeno o experimento registradas secuencialmente en el tiempo. Estas observaciones serán denotadas por:

{x(t1), x(t2), ..., x(tn)} = {x(t) : t Î T Í R} con x(ti) el valor de la variable x en el instante ti.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Si T = Z se dice que la serie de tiempo es discreta y si T = R se dice que la serie de tiempo es continua.

Los métodos de análisis de series de tiempo consideran el hecho que los datos tomados en diversos periodos de tiempo pueden tener algunas características de auto correlación, tendencia o estacionalidad.

MODELOS CLASICOS DE SERIES DE TIEMPO

MODELOS DE DESCOMPOSICIÓN

(Ortíz Farro, 2015) Un modelo clásico para una serie de tiempo, supone que una serie x(1), ..., x(n) puede ser expresada como suma o producto de tres componentes: tendencia, estacionalidad y un término de error aleatorio.

Existen tres modelos de series de tiempos, que generalmente se aceptan como buenas aproximaciones a las verdaderas relaciones, entre los componentes de los datos observados. Estos son:

1. Aditivo: $X(t) = T(t) + E(t) + A(t)$
2. Multiplicativo: $X(t) = T(t) \cdot E(t) \cdot A(t)$
3. Mixto: $X(t) = T(t) \cdot E(t) + A(t)$

Donde:

- X(t) serie observada en instante t
- T(t) componente de tendencia
- E(t) componente estacional
- A(t) componente aleatoria (accidental)

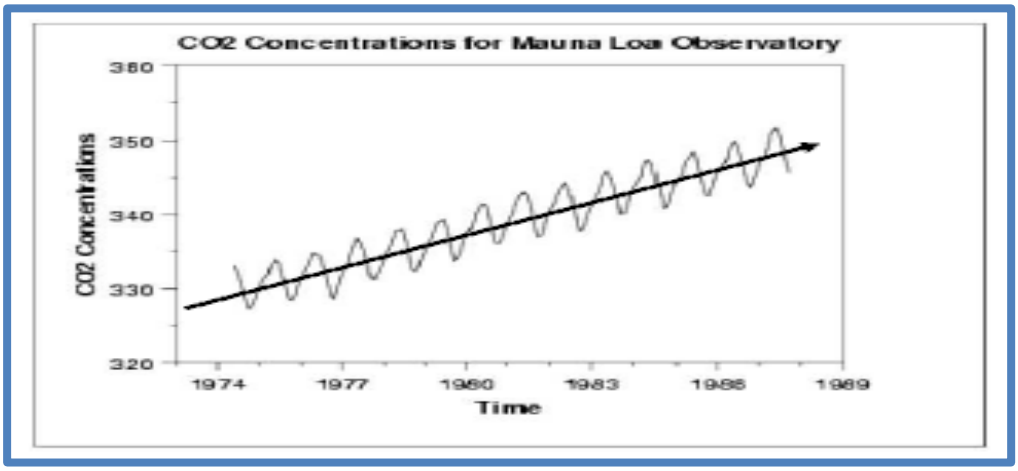
Un modelo aditivo, es adecuado, por ejemplo, cuando E(t) no depende de otras componentes, como T(t), sí por el contrario la estacionalidad varía con la tendencia, el modelo más adecuado

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

es un modelo multiplicativo . Es claro que este modelo puede ser transformado en aditivo, tomando logaritmos. El problema que se presenta, es modelar adecuadamente las componentes de la serie.

ESTIMACIÓN DE LA TENDENCIA: (Cruz Arrela, 2010) Patrón de comportamiento de los elementos en un entorno particular durante un periodo de tiempo.

Figura 3: Gráfico de Tendencia de un conjunto de datos de los años 1974-1989



Fuente: (Cruz Arrela, 2010)

Supondremos aquí que la componente estacional E(t) no está presente y que el modelo aditivo es adecuado, esto es:

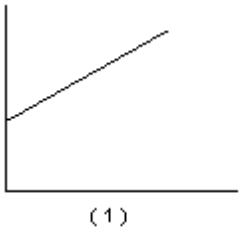
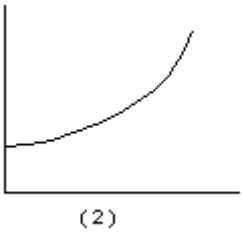
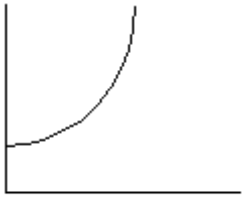
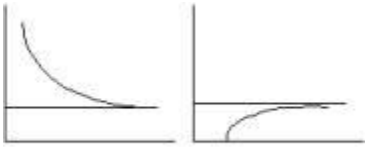
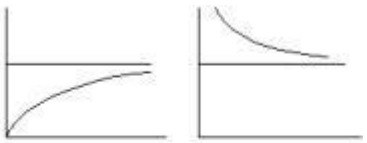
$$X(t) = T(t) + A(t),$$ donde A(t) es ruido blanco.

Hay varios métodos para estimar T(t). Los más utilizados consisten en:

AJUSTE DE UNA FUNCIÓN

Los siguientes gráficos ilustran algunas de las formas de estas curvas.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

1. $T(t) = a + bt$ (Lineal)  (1)	2. $T(t) = a\ e^{bt}$ (Exponencial)  (2)	3. $T(t) = a + b\ e^{bt}$ (Exponencial modificada)  (3)
4. $T(t) = b_0 + b_{1t} ,...,+ b_{mt}^m$ (Polinomial)	5. $T(t) = \exp(a + b(rt))$ (Gompertz $0 < r < 1$)  (5)	6. $T(t) = \frac{1}{a + b(r^t)}, 0 < r < 1$ (Logística)  (6)

ESTIMACIÓN DE LA ESTACIONALIDAD: (Ortíz Farro, 2015) La estimación de la estacionalidad no sólo se realiza con el fin de incorporarla al modelo para obtener predicciones, sino también con el fin de eliminarla de la serie para visualizar otras componentes como tendencia y componente irregular que se pueden confundir en las fluctuaciones estacionales.

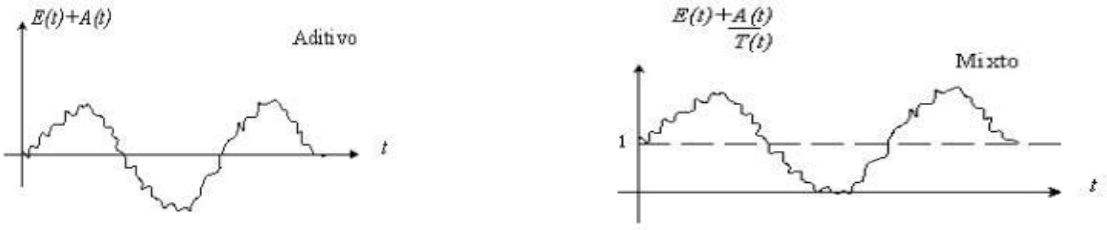
De acuerdo con los modelos de descomposición, se asume el siguiente modelo para T(t),

a) $X(t) - T(t) = E(t) + A(t)$ Aditivo

b) $$\frac{X(t)}{T(t)} = \underbrace{E(t) + \frac{A(t)}{T(t)}}_{Estacional+accidental}$$ Mixto

Una vez removida la tendencia se obtiene los siguientes gráficos, donde aparece el modelo aditivo y en la el modelo mixto

Figura 4: Modelo aditivo y mixto



Fuente: (Cruz Arrela, 2010)

Pues si no hay tendencia, se espera

$$\forall t, \quad E(t) = 0$$
$$\forall t, \quad T(t) = 1$$

Como $E(t) = E(t+12) = E(t+24) = \cdots$ para serie mensual, entonces basta estimar E(1), E(2), E(3), ... , E(12). Para una serie trimestral, bastaría conocer: E(1), E(2), E(3) y E(4).

Suponga que se ha estimado la tendencia por alguno de los métodos vistos en la sección previa. Sea $\hat{T}(t)$ la estimación de la tendencia ya sea mediante una curva o filtros lineales. Entonces,

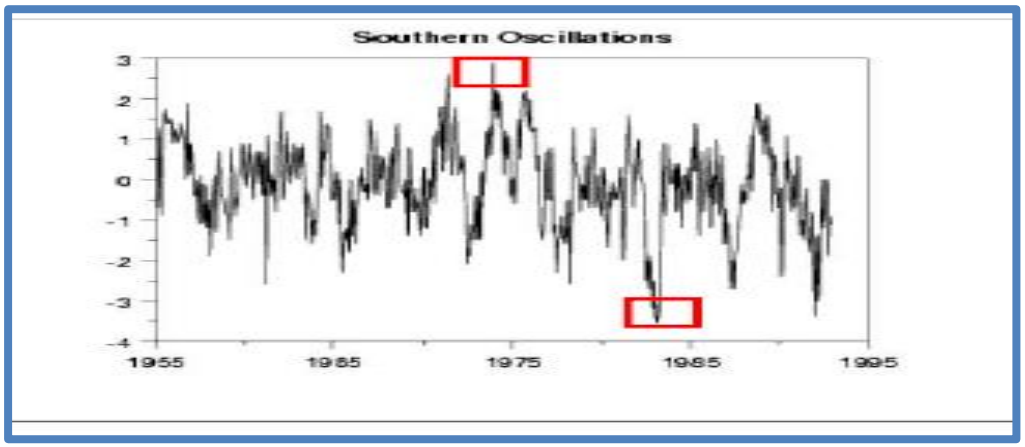
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

- Si el modelo es aditivo $R(t) = X(t) - \hat{T}(t), \quad t = 1,\ldots,n$ representa la serie con los efectos de tendencia removidos.

- Análogamente, si el modelo es mixto $W(t) = \frac{X(t)}{\hat{T}(t)}$ representa la serie, una vez removidos los efectos de tendencia.

Estas series generadas a partir de la original por eliminación de la tendencia se denominan “series de residuos” y deberán contener predominantemente fluctuaciones estacionales. Para estimar la estacionalidad se requiere haber decidido el modelo a utilizar (mixto o aditivo), lamentablemente esto no es siempre claro, ya sea porque no contamos con información a priori para suponerlo o porque el gráfico no ha dejado evidencia suficientemente clara como para decidirnos por alguno de ellos. En tal situación se propone calcular ambas series residuales y elegir aquella cuyos valores correspondientes a una estación dada oscilen menos en torno a su promedio. Son fluctuaciones periódicas, cuando se observan picos en determinados periodos, por ejemplo, ventas en navidad o fiestas patrias.

Figura 5: Gráfica de valores en el tiempo, donde se observa la estacionalidad



Fuente: (Cruz Arrela, 2010)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Para analizar la estacionalidad de una serie introduciremos un concepto de gran interés en el análisis de series temporales: la función de autocorrelación.

La función de autocorrelación mide la correlación entre los valores de la serie distanciados un lapso de tiempo k.

Recordemos la fórmula del coeficiente de correlación simple, dados N pares de observaciones y, x:

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

De igual forma, dada una secuencia temporal de N observaciones x1...xN, podemos formar N-1 parejas de observaciones contiguas (x1, x2), (x2, x3), ...(xN-1, xN) y calcular el coeficiente de correlación de estas parejas. A este coeficiente lo denominaremos coeficiente de autocorrelación de orden 1 y lo denotamos como r1. Análogamente se pueden formar parejas con puntos separados por una distancia 2, es decir (x1, x3), (x2, x4), etc. y calcular el nuevo coeficiente de autocorrelación de orden 2. De forma general, si preparamos parejas con puntos separados una distancia k, calcularemos el coeficiente de autocorrelación de orden k.

Al igual que para el coeficiente de correlación lineal simple, se puede calcular un error estándar y por tanto un intervalo de confianza para el coeficiente de autocorrelación.

La función de autocorrelación es el conjunto de coeficientes de autocorrelación rk desde 1 hasta un máximo que no puede

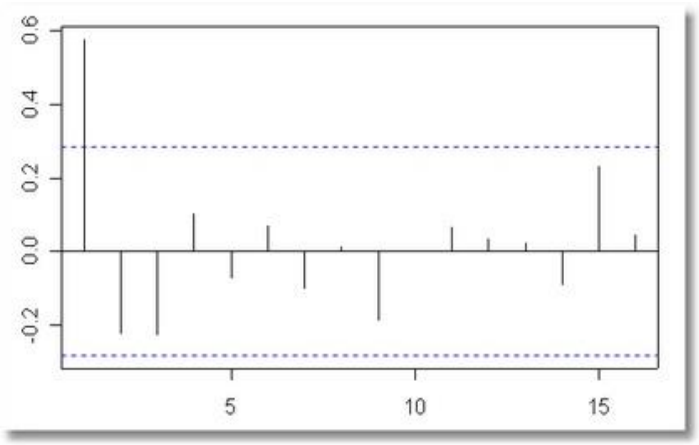
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

exceder la mitad de los valores observados, y es de gran importancia para estudiar la estacionalidad de la serie, ya que si ésta existe, los valores separados entre sí por intervalos iguales al periodo estacional deben estar correlacionados de alguna forma. Es decir que el coeficiente de autocorrelación para un retardo igual al periodo estacional debe ser significativamente diferente de 0.

Relacionada con la función de autocorrelación nos encontramos con la función de autocorrelación parcial. En el coeficiente de autocorrelación parcial de orden k, se calcula la correlación entre parejas de valores separados esa distancia pero eliminando el efecto debido a la correlación producida por retardos anteriores a k.

En la figura 5 vemos una gráfica típica de la función de autocorrelación parcial, en la que se marcan los intervalos de confianza para ayudar a detectar los valores significativos y cuya posición en el eje X nos indicará la probable presencia de un factor de estacionalidad para ese valor de retardo.

Figura 6: Función de autocorrelación parcial



Fuente: (Cruz Arrela, 2010)

Técnicas de Series de Tiempo en Minería de Datos

Si bien existen diversas familias de Minería de Datos predictivas, los datos con los que se cuenta para esta investigación son series de tiempo. Por ende, los modelos a aplicar en primer lugar son los de series de tiempo.

Aunque existen más métodos de pronóstico, por simplicidad se presentan sólo los considerados más usuales y sencillos de llevar a cabo

a. **Promedios Móviles:** se construye sustituyendo cada valor de la serie por la media obtenida y algunos valores inmediatamente anteriores o posteriores. Se considera el promedio móvil a partir de las tres observaciones más recientes. La ecuación es la siguiente:

$$Promedio\ Móvil = \sum \frac{n\ valores\ más\ recientes\ de\ datos}{n}$$

b. **Suavización Exponencial:** Este método emplea un promedio ponderado de la serie de tiempo pasada como pronóstico. Se selecciona solo un factor de ponderación, el de la observación más reciente. El modelo es el siguiente:

$$Ft + 1 = aYt + (1 - \propto)Ft$$

Dónde:

$F_{t+1:}$ Pronóstico de la serie de tiempo en el periodo t + 1

Y_t: valor real de la serie de tiempo en el periodo t

F_t: Pronósticos de la serie de tiempo para el periodo t

α: Constante de suavizamiento, $0 \leq \alpha \leq 1$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c. ARIMA: (Modelo autorregresivo integrado de media móvil). Está compuesto por tres componentes:

C: la variable que define la serie temporal y t depende de una constante C.

c.1 Componente autorregresivo: número de retrasos de la serie Yt que se introducen en el modelo, denotado con la letra ***p***. Se modela el comportamiento de la variable como una regresión lineal múltiple, con valores de la serie temporal, retrasados un periodo de muestreo.

c.2 Componente de media móvil: dependencia de la serie temporal Yt con valores pasados de los errores. Numero de errores introducidos al modelo, se denota con la letra ***q***.

Fórmula: arima (p, d,q)

$$Y_t^{(d)} = C + \underbrace{\phi_1 \cdot Y_{t-1}^{(d)} + \ldots + \phi_p \cdot Y_{t-p}^{(d)}}_{Comp.\ \text{Autorregresiva}} + \underbrace{\theta_1 \cdot \varepsilon_{t-1}^{(d)} + \ldots + \theta_q \cdot \varepsilon_{t-q}^{(d)}}_{Comp.\ \text{de Media Móvil}} + \varepsilon_t^{(d)}$$

D: orden de diferenciación.

Proceso Estocástico: Es una sucesión de variables aleatorias Yt ordenadas, donde t puede tomar cualquier valor. Se da la evolución de una variable en función de otra (por lo general el tiempo).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

d. HOLTWINTERS:

(Coghlan, 2015) El modelo Holt-Winters incorpora un conjunto de procedimientos que conforman el núcleo de la familia de series temporales de alisado exponencial. Holt-Winters puede adaptarse fácilmente a cambios y tendencias, así como a patrones estacionales. En comparación con otras técnicas, como ARIMA, el tiempo necesario para calcular el pronóstico es considerablemente más rápido. Esto significa que cualquier usuario puede poner en práctica la técnica de Holt-Winters. Más allá de sus características técnicas, su aplicación en entornos de negocio es muy común. De hecho, Holt-Winters se utiliza habitualmente por muchas compañías para pronosticar la demanda a corto plazo cuando los datos de venta contienen tendencias y patrones estacionales de un modo subyacente.

Esta técnica se basa en la atenuación de los valores de la serie de tiempo, obteniendo el promedio de estos de manera exponencial; es decir, los datos se ponderan dando un mayor peso a las observaciones más recientes y uno menor a las más antiguas.

La expresión para realizar el cálculo de la suavización exponencial es:

$$P_{t+1} = \alpha Y_t + \alpha(\alpha-1)Y_{t-1} + \alpha(\alpha-1)^2 Y_{t-2} + \ldots \alpha(\alpha-1)^{n-1} Y_{t-(n-1)}$$

Donde:

Yt : Valor de la serie en el periodo “t”.

Pt+1: Pronóstico o predicción para el periodo “t+1”

Pt : Pronóstico o predicción en el periodo “t”.

α : Factor de suavización ($o \leq \alpha \leq 1$)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Método aditivo Holt-Winters (Hyndman & Athanasopoulos, 2015)**.** La forma de componentes para el método aditivo es:

$$\begin{aligned}\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t-m+h_m^+}\\ \ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})\\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}\\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},\end{aligned}$$

Donde h+m=⌊(h−1)mod m⌋+1, asegura que las estimaciones de los índices estacionales utilizados para pronosticar provienen del último año de la muestra. La ecuación de nivel muestra una media ponderada entre la observación desestacionalizado (yt−st−m) y el pronóstico no estacional (ℓt−1+bt−1) para el tiempo de t. La ecuación de tendencia es idéntico al método lineal de Holt. La ecuación de temporada muestra una media ponderada entre el actual índice de estacionalidad, (yt−ℓt−1−bt−1), y el índice estacional de la misma temporada del año pasado (es decir, tiempo, periodos, años).

La ecuación para el componente estacional suele expresarse como:

$$s_t = \gamma^*(y_t - \ell_t) + (1-\gamma^*)s_{t-m}.$$

Si sustituimos ℓt de la ecuación de suavizado para el nivel de la forma componente anterior, obtenemos:

$$s_t = \gamma^*(1-\alpha)(y_t - \ell_{t-1} - b_{t-1}) + [1-\gamma^*(1-\alpha)]s_{t-m}$$

La forma de corrección de errores de las ecuaciones de suavizado es:

$$\begin{aligned}\ell_t &= \ell_{t-1} + b_{t-1} + \alpha e_t\\ b_t &= b_{t-1} + \alpha\beta^* e_t\\ s_t &= s_{t-m} + \gamma e_t.\end{aligned}$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Holt-Winters método multiplicativo (Hyndman & Athanasopoulos, 2015)**.** La forma de componentes para el método multiplicativo es:

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)s_{t-m+h_m^+}.$$
$$\ell_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}$$

Y la representación de corrección de errores es:

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha \frac{e_t}{s_{t-m}}$$
$$b_t = b_{t-1} + \alpha\beta^* \frac{e_t}{s_{t-m}}$$
$$s_t = s_t + \gamma \frac{e_t}{(\ell_{t-1} + b_{t-1})}$$

where $e_t = y_t - (\ell_{t-1} + b_{t-1})s_{t-m}$.

e. Método de tendencia lineal de Holt

(Hyndman R. J., 2015) Holt (1957) extendió suavización exponencial simple para permitir la predicción de los datos con una tendencia. Este método implica una ecuación de predicción y dos ecuaciones de suavizado (uno para el nivel y uno para la tendencia):

Ecuación Pronóstico $\hat{y}_{t+h|t} = \ell_t + hb_t$

Ecuación nivel $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$

Ecuación de tendencia. $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$

Donde ℓt denota una estimación del nivel de la serie en el tiempo t, bt denota una estimación de la tendencia (pendiente)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

de la serie en el tiempo t, α es el parámetro de suavizado para el nivel, 0≤α≤1 y β * es el parámetro de suavización de la tendencia, 0≤β * ≤1.

La ecuación de nivel aquí muestra que ℓt es un promedio ponderado de observación yt y la muestra dentro de la previsión de un paso por delante de tiempo t, aquí se da por ℓt-1 + bt-1. La ecuación de tendencia muestra que bt es una media ponderada de la tendencia estimada en el momento t basado en ℓt-ℓt-1 y bt-1, la estimación anterior de la tendencia.

La función de previsión ya no es plana, sino de tendencias. La previsión h-paso por delante es igual al último nivel estimado plus h veces el último valor tendencia estimada. Por lo tanto las previsiones son una función lineal de h.

La forma de corrección de errores del nivel y las ecuaciones de tendencia muestran los ajustes en función de los errores de predicción de un paso dentro de la muestra.

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha e_t$$
$$b_t = b_{t-1} + \alpha\beta^* e_t$$

Donde: $e_t = y_t - (\ell_{t-1} + b_{t-1}) = y_t - \hat{y}_{t|t-1}.$

f. ETS - Exponential smoothing state

(Hyndman R. J., 2015) Métodos de suavización exponencial han existido desde la década de 1950, y son los métodos de

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

pronóstico más populares utilizados en los negocios y la industria. Recientemente, suavizado exponencial ha revolucionado con la introducción de un marco de modelización completa incorporando innovaciones modelos de estado espacio, cálculo de probabilidades, los intervalos de predicción y los procedimientos para la selección del modelo. Los métodos de suavización exponencial son algoritmos que generan predicciones puntuales. Un modelo estadístico es un estocástico (o aleatorio) proceso generador de datos que puede producir una distribución de toda previsión.

ETS(A, N, N) Suavización exponencial simple con errores aditivos (Hyndman & Athanasopoulos, 2015): La forma de corrección de errores de suavización exponencial simple es dado por:

$$\ell t=\ell t-1+\alpha et$$

Donde $et=yt-\ell t-1$ and $y\hat{}t|t-1=\ell t-1.$ Así, $et=yt-y\hat{}t|t-1$ representa un error de pronóstico de un solo paso y podemos escribir $y_t=\ell_{t-1}+e_t.$

Entonces las ecuaciones del modelo se pueden escribir:

$$y_t=\ell_{t-1}+\varepsilon_t \quad (1.1)$$

$$\ell_t=\ell_{t-1}+\alpha\varepsilon_t. \quad (1.2)$$

Nos referimos a (1.1) como la ecuación de medición (u observación) y (1.2) como la ecuación de estado (o de transición). Estas dos ecuaciones, junto con la distribución estadística de los errores, forman un modelo estadístico completamente especificado. En concreto, estos constituyen un modelo de espacio de innovaciones estado subyacente suavización exponencial simple.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El término "innovaciones" proviene del hecho de que todas las ecuaciones en este tipo de especificación utilizan el mismo proceso de error aleatorio, εt. Por la misma razón también se conoce esta fórmula como una "fuente única de error" modelo en contraste con múltiples fuentes alternativas de formulaciones de error.

La ecuación de medición muestra la relación entre las observaciones y los estados no observados. En este caso yt observación es una función lineal de la ℓt-1 nivel, la parte predecible de yt, y εt error aleatorio, la parte impredecible de yt. Para otros modelos espaciales innovaciones estatales, esta relación puede ser no lineal.

La ecuación de transición muestra la evolución de la situación a través del tiempo. La influencia del parámetro de suavizado α es el mismo que para los métodos discutidos anteriormente. Por ejemplo, α regula el grado de cambio en los niveles sucesivos. Cuanto mayor sea el valor de α, más rápido los cambios en el nivel; cuanto menor es el valor de α, más suave los cambios. En el extremo más bajo, donde α = 0, el nivel de la serie no cambia con el tiempo. En el otro extremo, donde α = 1, el modelo se reduce a un modelo de paseo aleatorio, yt = yt-1 + εt.

ETS (M, N, N) Suavización exponencial simple con errores multiplicativos (Hyndman & Athanasopoulos, 2015): Se puede especificar modelos con errores multiplicativos escribiendo los errores aleatorios de un solo paso como errores relativos:

$$\varepsilon_t = \frac{y_t - \hat{y}_{t|t-1}}{\hat{y}_{t|t-1}}$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$. Substituting $\hat{y}_{t|t-1} = \ell_{t-1}$ gives $y_t = \ell_{t-1} + \ell_{t-1}\varepsilon_t$ and $e_t = y_t - \hat{y}_{t|t-1} = \ell_{t-1}\varepsilon_t$.

Entonces podemos escribir la forma multiplicativa del modelo de espacio de estado como:

$$y_t = \ell_{t-1}(1 + \varepsilon_t)$$
$$\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t).$$

g. Redes Neuronales

(Hyndman & Athanasopoulos, 2015) Las redes neuronales artificiales están pronosticando métodos que se basan en modelos matemáticos simples del cerebro. Permiten relaciones no lineales complejas entre la variable de respuesta y sus predictores

Arquitectura de red neuronal

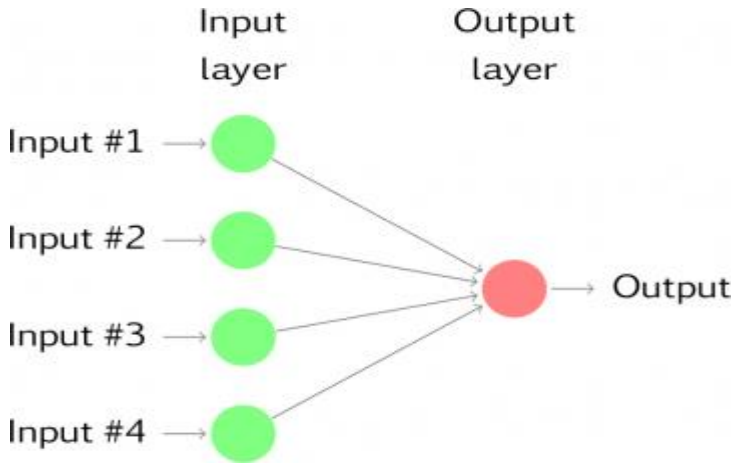
Una red neural puede ser pensada como una red de "neuronas" organizadas en capas. Los predictores (o entradas) de la capa inferior, y las previsiones (o salidas) forman la capa superior. Puede haber capas intermedias que contienen neuronas ocultas.

Las redes muy simples no contienen capas ocultas y son equivalentes a la regresión lineal. La figura 7 muestra la versión de la red neural de una regresión lineal con cuatro predictores. Los coeficientes adjuntos a estos predictores son llamados "pesos". Las previsiones se obtienen mediante una combinación lineal de las entradas. Los pesos son seleccionados en el marco de redes neuronales utilizando un

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

"algoritmo de aprendizaje" que minimiza una "función de costos" como MSE. Por supuesto, en este sencillo ejemplo, podemos utilizar la regresión lineal que es un método mucho más eficiente para la formación del modelo.

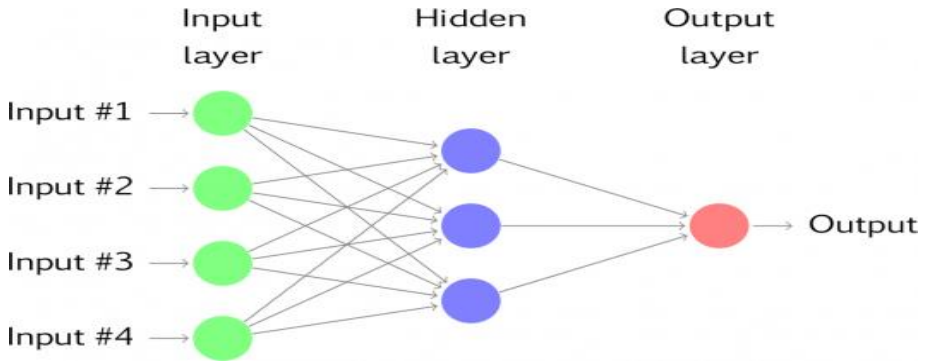
Figura 7: Una simple red neuronal equivalente a una regresión lineal



Fuente: (Hyndman & Athanasopoulos, 2015)

Una vez que añadimos una capa intermedia con neuronas ocultas, la red neuronal se convierte en no-lineal. Un ejemplo sencillo se muestra en la Figura 8.

Figura 8: Una red neuronal con cuatro entradas y una capa oculta con tres neuronas ocultas.



Fuente: (Hyndman & Athanasopoulos, 2015)

Esto se conoce como una red de alimentación hacia adelante de múltiples capas, donde cada capa de nodos recibe entradas de las capas anteriores. Las salidas de los nodos de una capa son entradas a la capa siguiente. Las entradas a cada nodo se combinan utilizando una combinación lineal ponderada. El resultado es entonces modificado por una función no lineal antes de ser salida. Por ejemplo, las entradas a la neurona oculta j en la figura 8 se combinan linealmente para dar:

$$z_j = b_j + \sum_{i=1}^{4} w_{i,j} x_i.$$

En la capa oculta, este es entonces modificado usando una función no lineal tal como una sigmoide,

$$s(z) = \frac{1}{1+e^{-z}},$$

Para dar la entrada para la siguiente capa. Esto tiende a reducir el efecto de valores de entrada extremos, con lo que la red un poco robusta a los valores atípicos.

Los parámetros b1, b2, b3 y w1,1, ..., w4,3 son "aprendidas" de los datos. Los valores de los pesos son a menudo restringidos para evitar que se conviertan en demasiado grandes. El parámetro que restringe los pesos se conoce como el "parámetro de decaimiento" y con frecuencia se ajusta para que sea igual a 0,1.

Los pesos toman valores aleatorios, para empezar, que luego son actualizados utilizando los datos observados. En consecuencia, hay un elemento de aleatoriedad en las predicciones producidas por una red neural. Por lo tanto, la red está normalmente entrenado varias veces utilizando diferentes puntos de partida al azar, y los resultados se promedian.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El número de capas ocultas, y el número de nodos en cada capa oculta, se deben especificar con antelación.

Características de una red neuronal artificial

Paralela: Las RNA cuentan con una gran cantidad de neuronas, cada una de ellas trabajando simultáneamente con una parte del problema mayor.

Distribuida: Las RNA cuentan con una gran cantidad de neuronas, a través de las cuales distribuyen su memoria. Esto los diferencia de los sistemas computacionales tradicionales, los que sólo cuentan con un procesador y memoria fija.

Adaptativa: Las RNA tienen la capacidad de adaptarse al entorno modificando sus pesos sinápticos, aprendiendo de las experiencias, consiguiendo generalizar conceptos a partir de casos particulares, permitiéndole encontrar una solución aceptable al problema.

Ventajas de las redes neuronales artificiales

Aprendizaje adaptativo: capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.

Auto-organización: una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

Tolerancia a fallas: la destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Operación en tiempo real: los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.

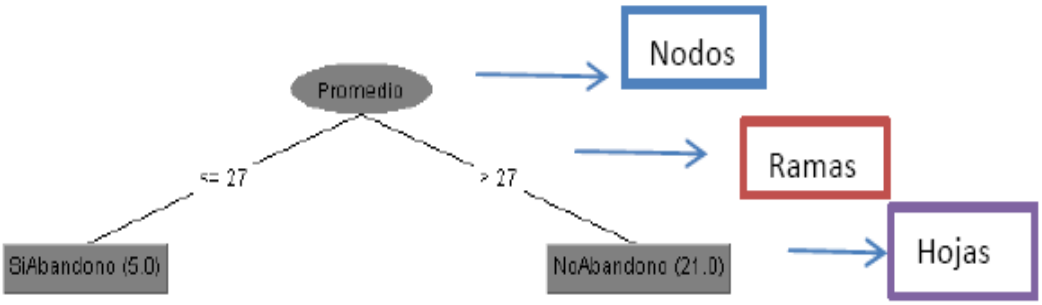
Fácil inserción dentro de la tecnología existente: se pueden obtener circuitos integrados especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Esto facilitará la integración modular en los sistemas existentes.

Clasificación: (Valcárcel Asencios, 2004) Se define como la identificación de características o atributos que hacen que un elemento se vincule a un grupo siguiendo un patrón de datos. Este último se puede utilizar para predecir cómo se comportarán nuevas instancias.

Los algoritmos de clasificación realizan cortes sobre una variable (lo cual limita su expresividad, pero facilita su comprensión). Generalmente se usan técnicas heurísticas en su construcción

El algoritmo puede generar un árbol de decisión a través de los datos ingresados, seleccionando el atributo que clasifique a los datos.

Figura 9: Técnica de clasificación



Fuente: *(Valcárcel Asencios, 2004)*

Agrupamiento (clustering): Hace corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similaridad. Es decir, agrupan a los datos bajo diferentes métodos y criterios. Las técnicas más usadas son las clásicas (distancia mínima) y las redes neuronales (método de Kohonen o método de Neural-Gas).

El propósito de K-Means es ubicar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares; todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con estos y asociados a aquel que sea el más próximo, en términos de una distancia previamente elegida.

El objetivo que se busca mediante el algoritmo es minimizar la varianza total intragrupo o la función de error cuadrático, para que el algoritmo pueda generar mejores resultados

$$V = \sum_{i=0}^{k} \sum_{j \in s_1} |x_j - u_i|$$

Resumen: Se obtienen representaciones compactas para subconjuntos de los datos de entrada (análisis interactivo de datos, generación automática de informes, visualización de datos).

Modelado de Dependencias: Se obtienen descripciones de dependencias existentes entre variables. El análisis de relaciones (por ejemplo las reglas de asociación), en el que se determinan relaciones existentes entre elementos de una base de datos, podría considerarse un caso particular de modelado de dependencias.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

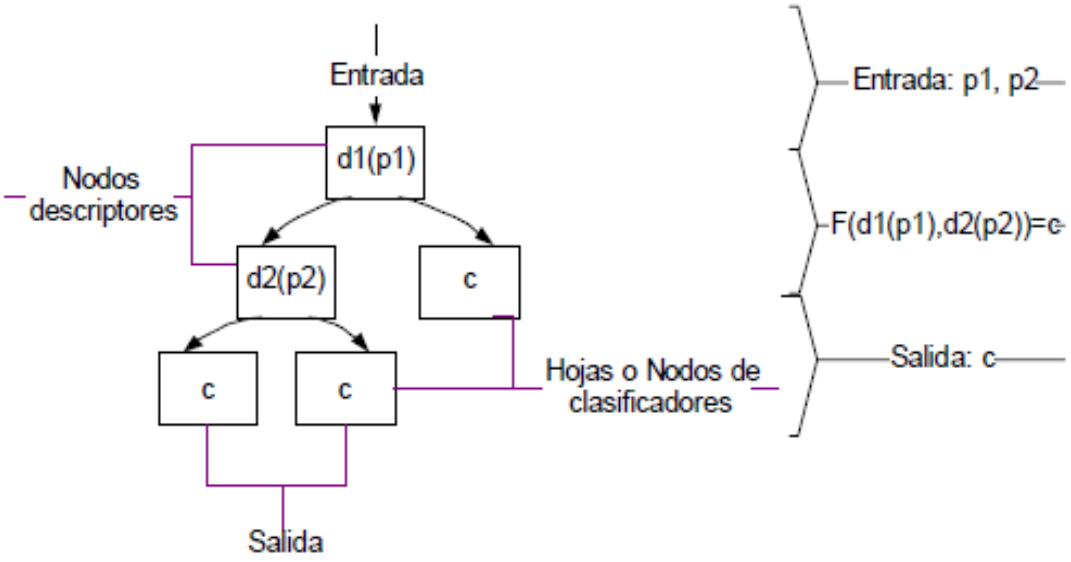
Análisis de Secuencias: Se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos (redes neuronales multicapas).

Redes bayesianas: Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones. Una de las características principales de los métodos bayesianos es el uso de distribuciones de probabilidad para cuantificar incertidumbre de los datos que se desea modelar. Una de las desventajas de los métodos bayesianos es que no pueden realizar predicciones con pocos datos, ya que no podría proporcionar un modelo correcto con poca cantidad de información. (Aluja, 2001).

Árboles de decisión: Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación. La técnica basada en árboles de decisión es quizás el método más fácil de utilizar y entender. Un árbol de decisión es una estructura jerárquica conformada por un conjunto de nodos, en donde cada nodo establece una condición o regla la misma que puede retornar verdadero o falso.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Figura 10: Estructura de un árbol de decisión



Fuente: (Edison, 2010)

D. Metodología para minería de datos CRISP-DM

Según (IBM, 2012) dice: CRISP-DM son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar trabajos de minería de datos.

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

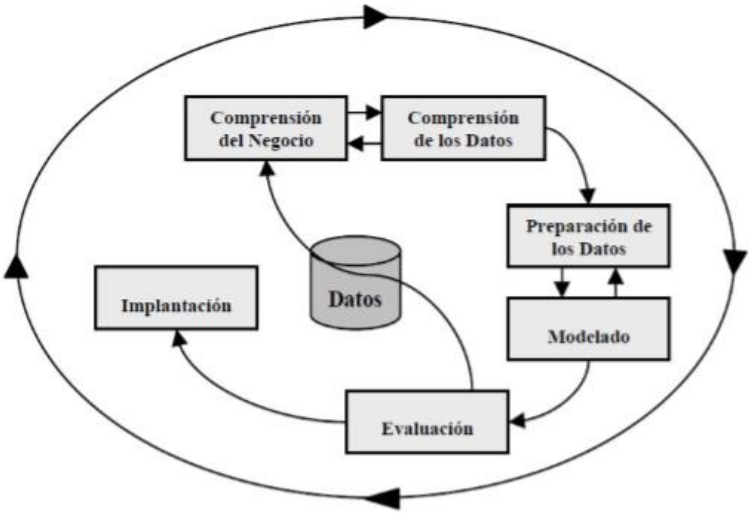
Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

Según (The CRISP-DM, 2013) la metodología CRISP – DM, como lo muestra la Fig. 2, está descrita en términos de un modelo de proceso jerárquico, que consiste en una serie de tareas descritas en cuatro niveles de abstracción (de lo general a

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

lo específico): Fases, tareas genéricas, tareas especializadas e instancias de proceso.

Figura 11: Fases del Proceso de Metodología CRISP-DM



Fuente: (The CRISP-DM, 2013)

Entendimiento del negocio, (The CRISP-DM, 2013)esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio, para convertir este conocimiento en un problema de definición de minería de datos y un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos, (The CRISP-DM, 2013) esta fase inicia con una colección inicial de datos y procede con actividades para familiarizarse con ellos, identificar problemas de calidad en los mismos, descubrir una primera idea de estos o detectar conjuntos interesantes que permitan formar hipótesis en la búsqueda de información escondida.

Preparación de los datos, (The CRISP-DM, 2013) cubre todas las actividades para construir la base final de datos (datos que

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

serán el alimento de las herramientas de modelado) desde una base en bruto. Es preferible que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Estas tareas incluyen tabulación, documentación y selección de atributos, también como transformación y limpieza de datos para las herramientas de modelado.

Modelado, (The CRISP-DM, 2013) se seleccionan y aplican varias técnicas, y sus parámetros son calibrados a los valores óptimos. Por lo general hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por lo tanto será a menudo necesario devolverse a la fase de preparación de datos.

Evaluación, (The CRISP-DM, 2013) al llegar a esta fase se ha construido un modelo (s) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la entrega final del modelo es importante evaluarlo más a fondo y revisar los pasos ejecutados para construirlo, de tal forma que este lo más cercano posible de alcanzar los objetivos del negocio. Un objetivo clave es determinar si hay algún evento importante del negocio que no haya sido considerado lo suficiente. Al final de esta fase, se debe tener una decisión sobre el uso de los resultados de minería de datos.

Despliegue, (The CRISP-DM, 2013) la creación del modelo por lo general no es el final del proyecto. Incluso si el propósito del modelo es incrementar conocimiento sobre los datos, el conocimiento ganado necesitará ser organizado y presentado de una manera que el cliente lo pueda usar. A menudo implica aplicar modelos en vivo dentro del proceso de toma de decisiones de una organización, por ejemplo, en la

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

personalización en tiempo real de las páginas web o la puntuación repetida en bases de datos de mercadeo. Sin embargo, dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso repetible de minería de datos a través de la empresa. En muchos casos es el cliente, no el analista de datos, quien realiza los pasos de despliegue. Sin embargo, incluso si el analista no carga con el esfuerzo de despliegue, es importante que el cliente entienda que acciones deben ser llevadas a cabo para hacer uso de los modelos creados.

E. Herramientas para minería de datos

Según (Microsoft, 2013) existen diversos tipos de herramientas para apoyar las actividades de minería de datos:

Asistentes de minería de datos: facilita la creación de estructuras y de modelos de minería de datos, usando orígenes de datos relacionales o datos multidimensionales en cubos.

Visores de modelos: para explorar los modelos de minería de datos una vez creados; con ellos, se puede examinar los modelos mediante visores adaptados a cada algoritmo o analizar con mayor profundidad utilizando el visor de contenido del modelo.

Generador de consultas de predicción: ayuda a crear consultas de predicción, también puede probar la exactitud de los modelos respecto a un conjunto de datos de exclusión o datos externos, o utilizar validación cruzada para evaluar la calidad del conjunto de datos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Herramientas para limpiar datos: contiene herramientas que puede utilizar para limpiar datos, automatizar tareas como la creación de predicciones y actualización de modelos y para crear soluciones de minería de datos de texto.

F. Librerías para minería de datos

Según (Edison, 2010) dice:

Algunas de las librerías más importantes para minería de datos son:

XELOPES: Es una librería con licencia pública GNU para el desarrollo de aplicaciones de minería de datos. Entre sus principales características tenemos acceso a datos, modelos de redes neuronales, métodos de agrupación, métodos de reglas de asociación, árboles lineales, árboles no lineales y exportación de datos.

MLC++: Es un conjunto de librerías que fueron desarrolladas por la Universidad de Stanford, basadas en lenguaje C++. Sus principales características son acceso a datos incluyendo archivos de formato plano, transformación de datos y métodos de aprendizaje mediante objetivos.

Lenguaje R: R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de implementación GNU del premiado lenguaje S. R y S-PLUS versión comercial de S- son, probablemente, los dos lenguajes más utilizados en

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

investigación, por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, bioinformática y las matemáticas financieras.

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, test estadísticos, análisis de seriales temporales, algoritmos de clasificación y agrupamiento, etc.).

G. Predicción

Según (Bunge, 2001) dice:

El término predicción puede referirse tanto a la «acción y al efecto de predecir» como a «las palabras que manifiestan aquello que se predice»; en este sentido, predecir algo es «anunciar por revelación, ciencia o conjetura algo que ha de suceder».

La predicción constituye una de las esencias claves de la ciencia, de una teoría científica o de un modelo científico. Así, el éxito se mide por el éxito o acierto que tengan sus predicciones.

La predicción en el contexto científico es una declaración precisa de lo que ocurrirá en determinadas condiciones especificadas. Se puede expresar a través del silogismo: "Si A es cierto, entonces B también será cierto."

El método científico concluye con la prueba de afirmaciones que son consecuencias lógicas del corpus de las teorías científicas. Generalmente esto se hace a través de experimentos que deben poder repetirse o mediante estudios observacionales rigurosos.

Una teoría científica cuyas aseveraciones no son corroboradas por las observaciones, por las pruebas o por experimentos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

probablemente será rechazada. Las teorías que generan muchas predicciones que resultan de gran valor (tanto por su interés científico como por sus aplicaciones) se confirman o se falsean fácilmente y, en muchos campos científicos, las más deseables son aquéllas que, con número bajo de principios básicos, predicen un gran número de sucesos.

H. Técnicas de predicción con minería de datos

Según (Aluja, 2001) cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y solo puede crecer en el futuro. La siguiente es una lista de técnicas:

Análisis Factoriales Descriptivos: Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

«MarketBasketAnalysis» o análisis de la cesta de la compra: Permite detectar que productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes del de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

Técnicas de «clustering»: Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí,

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

2.4. Definición de términos básicos

A. Almacén de datos

Es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. (Kimball, 1998).

B. Análisis prospectivo de datos

Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos. (Lezcano, 2010)

C. Árbol de decisión

Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. (Asencios, 2004)

D. Método

Modo ordenado y sistemático de proceder para lograr un fin / conjunto de reglas. (Getoor & Ben, 2007)

E. Metodología

Conjunto de métodos que se siguen en una disciplina científica / ciencia del método y de la sistematización científica. (Grudnitsky, 1992)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

F. Minería de datos

Descubrimiento de relaciones en grandes conjuntos de datos. Conjunto de técnicas aplicadas al proceso de extracción y presentación de conocimiento que yace implícito en grandes conjuntos de datos, que es desconocido y útil en términos de negocios, y que permite predecir en forma automatizada el comportamiento de los clientes. (Valcárcel Asencios, 2004)

G. Modelo predictivo

Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos (Lezcano, 2010)

H. Técnicas de Predicción

Métodos que tienen por finalidad obtener estimaciones o pronósticos de valores futuros de una serie temporal a partir de la información histórica contenida en la serie observada hasta el momento actual. (Getoor & Ben, 2007)

I. Predicción de ventas

Se llama previsión de ventas al cálculo que hace el departamento comercial de una compañía del volumen de ventas que realizará el año próximo. La realización de una correcta previsión de ventas es vital para una empresa pues de ella se deriva el presupuesto de ingresos y de gastos y por consiguiente, las previsiones de fabricación, aprovisionamiento, logística, recursos humanos. (Schaefer, 2012).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO III

MARCO METODOLÓGICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO III: MARCO METODOLÓGICO

3.1. Tipo y Diseño de la Investigación

La presente investigación es:

De tipo Tecnológica – Propositiva

La elaboración de la investigación es tecnológica, porque tiene como objetivo la implementación de una solución basada en Minería de Datos.

Y propositiva, porque los resultados obtenidos en función de los indicadores son estimaciones que se podrían generar al implementar dicha aplicación.

Su diseño Cuasi-Experimental: Porque consiste en seleccionar los grupos de la muestra en los que se prueba la variable sin ningún tipo de selección aleatoria.

3.2. Población y muestra

3.2.1 Población

La población está compuesta por el número registros de ventas de periodos 2011-2014 de la Empresa El Astro S.A.C.

Comprobante emitido. Representa el documento que se origina por la compra de un artículo, que generan los ingresos por ventas como indicador de avance en un periodo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 1: Ventas - Año 2011

AÑO	MES	COMPROBANTES EMITIDOS
2011	**AGOSTO**	287
	SEPTIEMBRE	991
	OCTUBRE	983
	NOVIEMBRE	841
	DICIEMBRE	2012
	TOTAL	**5114**

Fuente: Base de Datos de la Empresa de El Astro S.A.C Agosto 2011 - Diciembre 2011

Tabla 2: Ventas - Año 2012

AÑO	MES	COMPROBANTES EMITIDOS
2012	**ENERO**	1162
	FEBRERO	1270
	MARZO	1491
	ABRIL	1114
	MAYO	1120
	JUNIO	1249
	JULIO	1273
	AGOSTO	1287
	SEPTIEMBRE	1166
	OCTUBRE	1121
	NOVIEMBRE	1003
	DICIEMBRE	2007
	TOTAL	**15263**

Fuente: Base de Datos de la Empresa de El Astro S.A.C Enero 2012 - Diciembre 2012

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 3: Ventas - Año 2013

AÑO	MES	COMPROBANTES EMITIDOS
2013	**ENERO**	1613
	FEBRERO	1396
	MARZO	1273
	ABRIL	1075
	MAYO	1383
	JUNIO	1339
	JULIO	1537
	AGOSTO	1242
	SEPTIEMBRE	1148
	OCTUBRE	1214
	NOVIEMBRE	1311
	DICIEMBRE	2404
	TOTAL	**16935**

Fuente: Base de Datos de la Empresa de El Astro S.A.C Enero 2013 - Diciembre 2013

Tabla 4: Ventas - Año 2014

AÑO	MES	COMPROBANTES EMITIDOS
2014	**ENERO**	1612
	FEBRERO	1550
	MARZO	1755
	ABRIL	1227
	MAYO	1368
	JUNIO	1298
	TOTAL	**8810**

Fuente: Base de Datos de la Empresa de El Astro S.A.C Enero 2014 - Junio 2014

Comprendiendo un total de 46122 registros de la base de datos de ventas de la empresa El Astro S.A.C.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.2.2 Muestra

Como es una base de datos la información es accesible; debido a eso se trabajará con los 46122 registros de las ventas por lo tanto no se requiere muestreo.

3.3. Hipótesis

Mediante la comparación de técnicas de minería de datos se podrá elegir la técnica para la predicción de ventas en el sector comercial de artículos deportivo.

3.4. Variables – Operacionalización

3.4.1. Variable Independiente

Técnicas de Minería de Datos

3.4.2. Variable Dependiente

Predicción de ventas

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.5. Operacionalización

Tabla 5: Operacionalización de Variables

Variable independiente	Dimensiones	Indicadores	Fórmula
Técnicas de Minería de Datos	Tiempo	Tiempo de procesamiento del modelo	T1 / T2
	Datos	Número de puntos mínimos para el procesamiento de estimaciones	MP1 / MP2
Variable Dependiente	**Dimensiones**	**Indicadores**	**Formula**
Predicción de ventas	Tiempo	Tiempo para generar estimación (Sistema)	TP
	Grado de confiabilidad	Confiabilidad de los pronósticos generados (Mide la confiabilidad del modelo con respecto a las predicciones realizadas)	$PCPV = 100 - (\frac{\sum \frac{MR - MP}{MR}}{N} * 100)$ PCPV: Porcentaje de confiabilidad de predicción de Ventas. MP: Monto pronosticado MR: Monto real N: Número de observaciones

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.6. Métodos, técnicas e instrumentos de recolección de datos

3.6.1 Métodos:

Observación del comportamiento del sistema, para conocer los procesos que se realizan en la predicción de ventas.

Pruebas en la data simulando escenarios, utilizando herramientas de minería de datos que permitan evaluar cada una de las técnicas de minería de datos seleccionada.

3.6.2 Técnicas e instrumentos de recolección de datos:

Ficha de observación de pruebas y comportamiento del sistema

3.7. Procedimiento para la recolección de datos

Para la recolección de datos se usará las siguientes técnicas.

3.7.1 Observación

Mediante esta técnica, se pretende observar el comportamiento del sistema implementado con las técnicas de minería de datos.

3.7.2 Prueba del sistema

Las técnicas implementadas serán puestas en marcha procesadas para determinar su nivel de desempeño de acuerdo con los indicadores.

3.8. Análisis estadístico e Interpretación de los datos

El Análisis estadísticos de los datos se basa en el:

3.8.1 Uso de tablas, para evaluar resultados de las técnicas de minería de datos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.8.2 Uso de gráficos estadísticos, para evaluar resultados de las técnicas de minería de datos.

3.8.3 Uso de técnicas estadísticas como la media y desviación estándar.

3.9. Principios éticos

3.9.1 Confidencialidad: Asegura la protección de la identidad de la institución y las personas que participan como informantes de la investigación.

3.9.2 Objetividad: El análisis de la situación encontrada se basa en criterios técnicos e imparciales.

3.9.3 Originalidad: Se citan las fuentes bibliográficas de la información.

3.9.4 Veracidad: La información mostrada es verdadera, cuidando la confidencialidad.

3.10. Criterios de rigor científico

3.10.1 Confiabilidad: Se realizan cálculos estadísticos para la determinación del nivel de consistencia interna de los instrumentos de recolección de datos.

3.10.2 Validación: Se validan los instrumentos de recolección de datos y la propuesta de solución a través de Juicio de Expertos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.11. Evaluación económica del software

Para calcular el costo del software se utilizó la formulación de Barry W. Boehm.

ANÁLISIS PRELIMINAR

DEFINICIÓN DE REQUERIMIENTOS:

Donde:

RS = Responsabilidades del Sistema

Se considera la siguiente lista, siendo seis:

a. Generar modelo de series de tiempo
b. Entrenar modelo
c. Monitorear actividades
d. Realizar estimaciones
e. Generar reportes
f. Visualizar comparación de modelos predictivos

$$\mathbf{RS = 6}$$

F = Funciones de Sistema:

$$\mathrm{F} = 280 * \mathrm{RS}$$

$$\mathbf{F = 1680}$$

MF = Miles de Funciones

$$\mathrm{MF} = \frac{\mathrm{F}}{1000}$$

$$\mathrm{MF} = \frac{1680}{1000}$$

$$\mathbf{MF = 1.68}$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ESF = Esfuerzo.

$$\text{ESF} = 2.4(\text{MF})^{1.05}$$

$$\text{ESF} = 2.4(1.68)^{1.05}$$

$$\mathbf{ESF = 4.13795714}$$

TDES = Tiempo de Desarrollo

$$\text{TDES} = 2.5(\text{ESF})^{0.38}$$

$$\text{TDES} = 2.5(4.13795714)^{0.38}$$

$$\mathbf{TDES = 4.29\ meses}$$

CH = Cantidad de Hombres por MES

$$\text{CH} = \text{ESF}/\text{TDES}$$

$$\text{CH} = \frac{4.13795714}{4.29}$$

$$\text{CH} = 0.9645$$

$$\mathbf{CH = 1\ personas\ por\ mes}$$

CHM = Costo Hombre por Mes

$$\text{CHM} = \text{CH} * \text{SPM (Salario Promedio Mensual)}$$

$$\text{CHM} = 1 * 2400$$

$$\mathbf{CHM = 2400}$$

CD = Costo de Desarrollo

$$\text{CD} = \text{ESF} * \text{CHM}$$

$$\text{CD} = 4.138 * 2400$$

$$\mathbf{CD = S/.\ 9{,}931.20}$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Por las características del proyecto, los siguientes indicadores son:

Tabla 6: Indicadores /Factores por Medida de Proyecto

Indicadores	**Modo**	**Pequeño**
		2 MF
Esfuerzo	Orgánico	5.00
Productividad		400.00
Tiempo de Desarrollo		4.60
Personal		1.10

Fuente: Elaboración propia

Tabla 7: Distribución de esfuerzo y tiempo de desarrollo por etapas

Indicador / Modo	**Fases**		**2 MF**
Esfuerzo			
Orgánico	Estudio Preliminar		6%
	Análisis		16%
	Diseño y Desarrollo		68%
		Diseño	26%
		Desarrollo	42%
	Prueba e Implantación		16%
Tiempo de Desarrollo			
Orgánico	Estudio Preliminar		10%
	Análisis		19%
	Diseño y Desarrollo		63%
	Prueba e Implantación		18%

Fuente: Elaboración propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO IV

ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO IV: ANÁLISIS E INTERPRETACIÓN DE LOS DATOS

4.1. Resultados en tablas y gráficos

Siendo el objetivo general de este estudio:

"Realizar un análisis comparativo de las técnicas de minería de datos para la predicción de ventas orientado a la comercialización de artículos deportivos".

Este análisis se realizará midiendo tres indicadores

A. Confiabilidad de los pronósticos generados por el modelo

Este indicador mide la capacidad de confianza de cada algoritmo evaluado, dado los objetivos de esta investigación donde se debe evaluar las técnicas, en este caso son: Holtwinters, Holt y ETS para seleccionar la técnica que tenga mejor confianza.

$$PCPV = 100 - (\frac{\sum \frac{MR - MP}{MR}}{N} * 100)$$

PCPV: Porcentaje de confiabilidad de predicción de Ventas.

MP: Monto pronosticado

MR: Monto real

N: Número de observaciones

Tabla 8: Generación de los Pronósticos

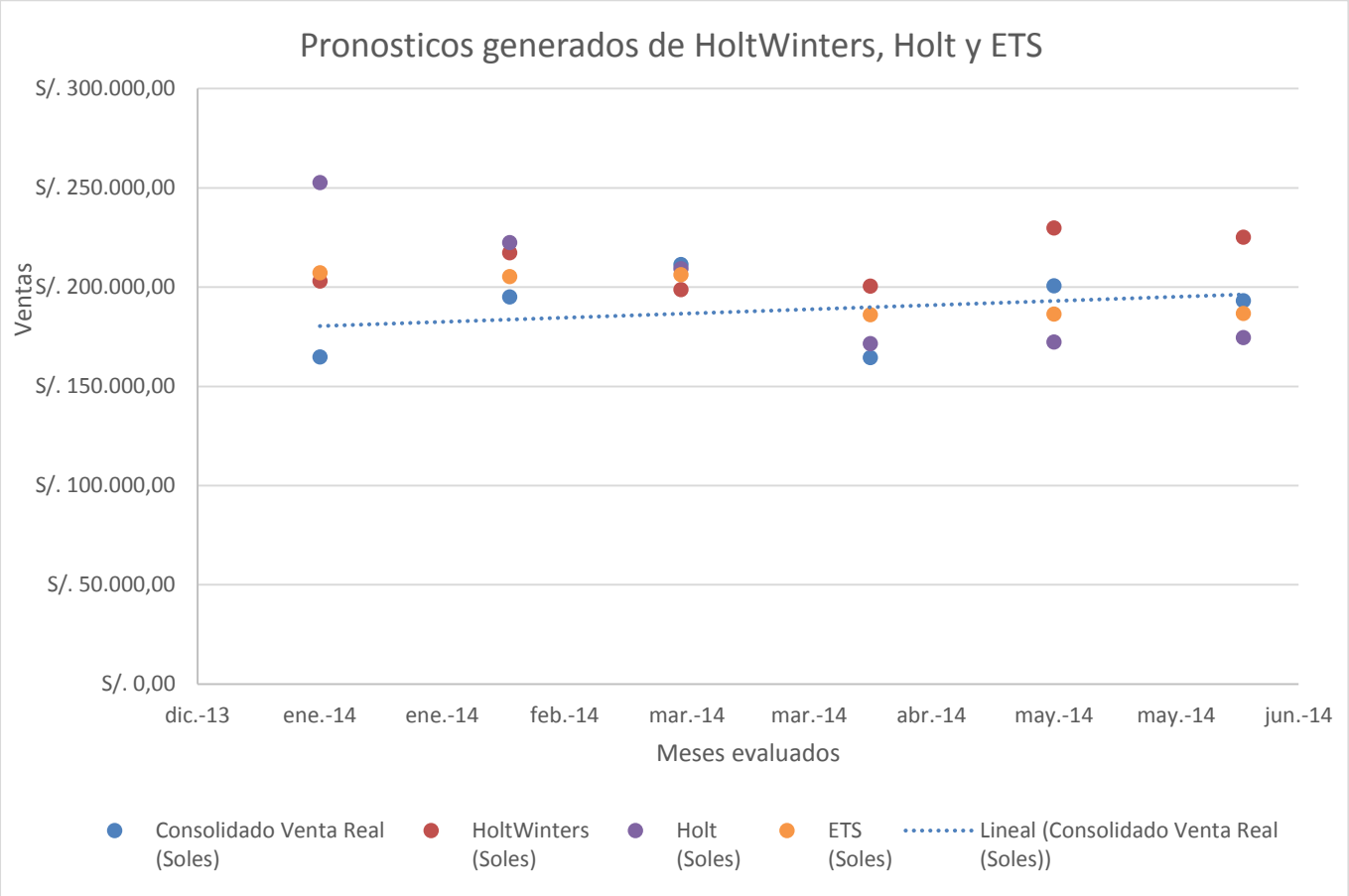
Meses Evaluados	Consolidado Venta Real (Soles)	HoltWinters (Soles)	Holt (Soles)	ETS (Soles)
ene-14	S/. 164.891,50	S/. 203.112,45	S/. 252.664,36	S/. 207.161,92
feb-14	S/. 194.987,90	S/. 217.362,01	S/. 222.369,47	S/. 205.319,15
mar-14	S/. 211.407,09	S/. 198.673,48	S/. 209.334,76	S/. 206.194,54
abr-14	S/. 164.574,00	S/. 200.435,65	S/. 171.514,71	S/. 186.049,64
may-14	S/. 200.738,60	S/. 229.804,80	S/. 172.390,20	S/. 186.415,82
jun-14	S/. 193.210,80	S/. 225.109,57	S/. 174.608,26	S/. 186.763,31

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En la Tabla N° 08 se muestra los resultados obtenidos del pronóstico generado con los algoritmos HoltWinters, Holt y ETS en comparación al consolidado de la venta real de los últimos seis meses.

Gráfico 1: Pronósticos de Ventas de HoltWinters, Holt y ETS



Fuente: Elaboración Propia

En el gráfico N° 1 podemos observar que los puntos que se acercan a la línea de tendencia de consolidado real corresponden a ETS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 9: Resultados obtenidos con la fórmula aplicada

Meses Evaluados	HoltWinters	Holt	ETS
ene-14	23,18	53,23	25,64
feb-14	11,47	14,04	5,30
mar-14	6,02	0,98	2,47
abr-14	21,79	4,22	13,05
may-14	14,48	14,12	7,14
jun-14	16,51	9,63	3,34
Total	**15,58**	**16,04**	**9,49**

Fuente: Elaboración Propia

PCPV= 100 %- Total = Grado de confianza

Grado de Confianza	HW= 84,42 %	Holt= 83,96 %	ETS=90,51 %

En la tabla N° 9, observamos que de acuerdo a la fórmula aplicada el porcentaje de confiabilidad del modelo con respecto a los pronósticos arrojados en los meses determinados en la muestra obteniendo un grado de confianza en HoltWinters 84.42 %, Holt con un 83.96% y ETS obtuvo el 90.51%. Por lo tanto el nivel de confianza más elevado corresponde a ETS.

B. Tiempo de procesamiento para obtener la estimación

Este indicador mide el tiempo que le toma a cada técnica calcular u obtener la estimación requerida. Según los objetivos iniciales de la investigación.

$$T1 \; / \; T2 \; / \; T3$$

T1: TIEMPO DE PROCESAMIENTO DE ALGORITMO 1
T2: TIEMPO DE PROCESAMIENTO DE ALGORITMO 2
T3: TIEMPO DE PROCESAMIENTO DE ALGORITMO 3

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

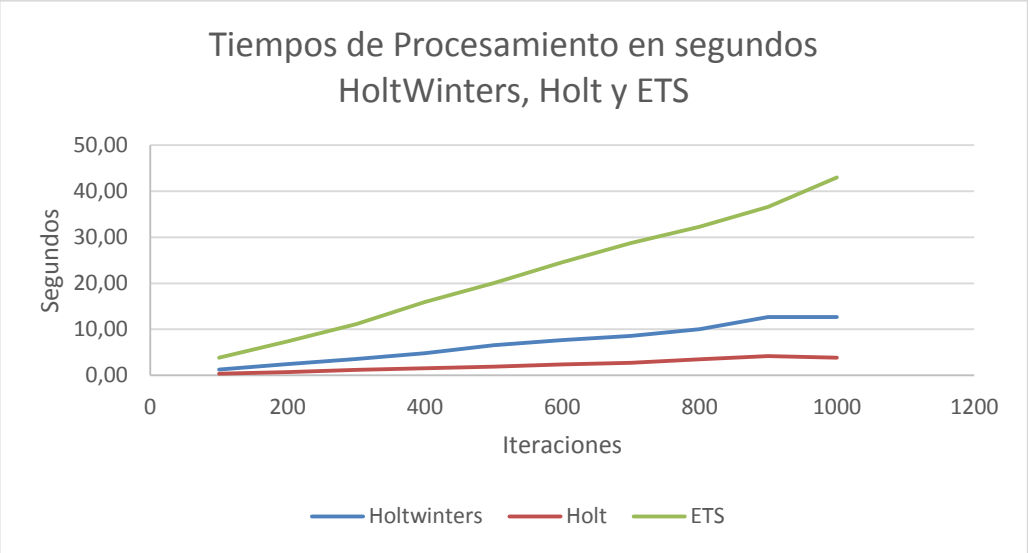
Tabla 10: Tiempo de Procesamiento entre HoltWinters, Holt y ETS

Iteraciones realizadas	Holtwinters (Segundos)	Holt (Segundos)	ETS (Segundos)
100	1,23	0,37	3,81
200	2,42	0,74	7,39
300	3,57	1,18	11,16
400	4,79	1,52	15,91
500	6,54	1,86	20,01
600	7,66	2,39	24,54
700	8,55	2,70	28,70
800	10,00	3,47	32,29
900	12,69	4,17	36,56
1000	12,64	3,85	42,95
Promedios	**7,01**	**2,23**	**22,33**

Fuente: Elaboración Propia

En la Tabla N° 10 podemos apreciar las iteraciones y el tiempo en segundos que demoran los algoritmos para el procesamiento de los datos. El algoritmo Holt tiene mejor tiempo de procesamiento en cada iteración equivalente a un promedio de 2.23 segundos.

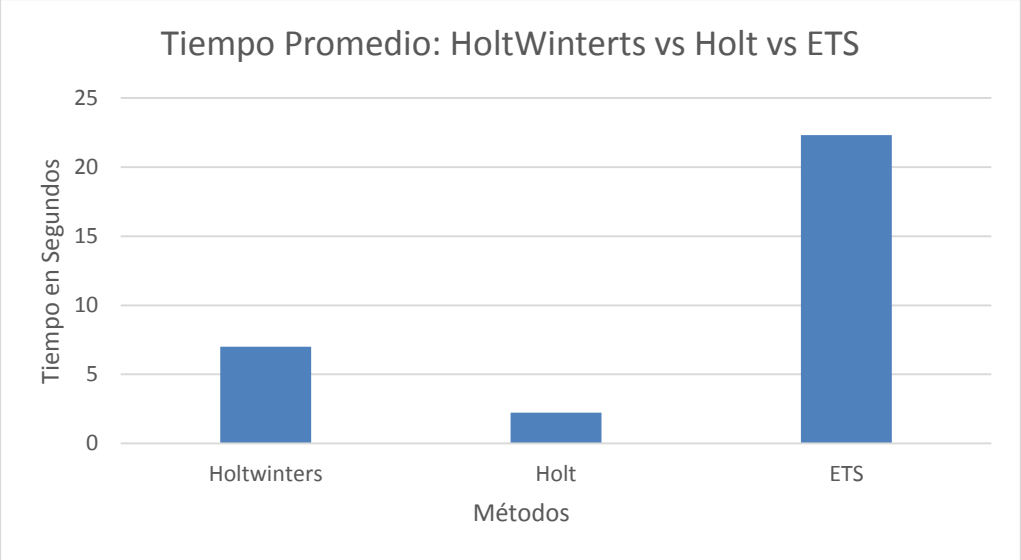
Gráfico 2: Tiempo de Procesamiento entre HoltWinters, Holt y ETS



Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 3: Tiempo Promedio entre HoltWinters, Holt y ETS



Fuente: Elaboración Propia

El gráfico N° 3 representa el promedio de los algoritmos usados para el procesamiento de los datos, en el cual podemos observar que el algoritmo de ETS es el que mayor tiempo demoró para dicho procesamiento.

C. Número de puntos mínimos para el vector que procesará el modelo

Este indicador mide la capacidad de las técnicas y la variabilidad de sus resultados en función al número de datos históricos que ingresa en el vector de series de tiempo, donde se evalúa gradualmente la disminución de los mismos para evaluar el rendimiento de los algoritmos, hasta determinar cuál es el mínimo de datos que puede tratar el modelo.

$$MP1 \ / \ MP2 \ / \ MP3$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

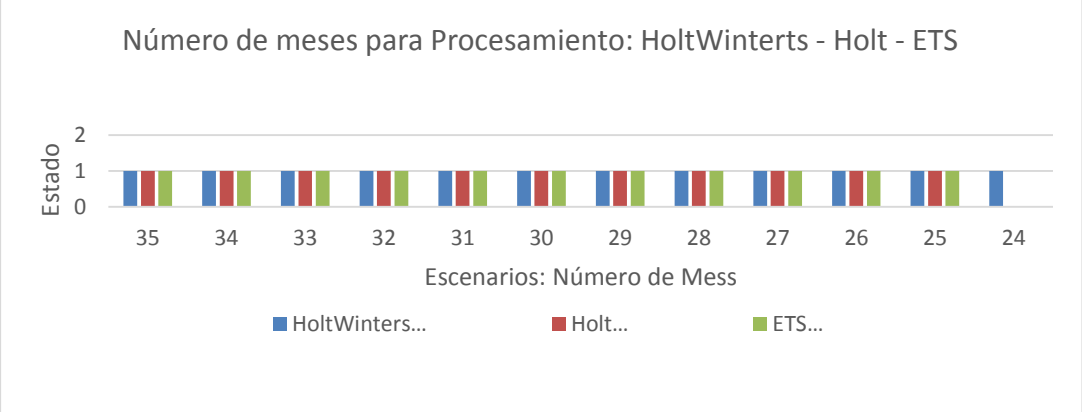
Tabla 11: Número de Meses Mínimos para el Procesamiento de Estimaciones para HoltWinters, Holt y ETS

Escenario Número de meses	HoltWinters (Estado)	Holt (Estado)	ETS (Estado)
35	1	1	1
34	1	1	1
33	1	1	1
32	1	1	1
31	1	1	1
30	1	1	1
29	1	1	1
28	1	1	1
27	1	1	1
26	1	1	1
25	**1**	**1**	**1**
24	**1**	**0**	**0**

Fuente: Elaboración Propia

Podemos apreciar en la Tabla N° 11 que HoltWinters trabaja con datos acumulados.

Gráfico 4: Cantidad de meses mínimos para el procesamiento HoltWinters, Holt y ETS



Fuente: Elaboración Propia

En el gráfico N° 4 podemos observar que no ha existido decaimiento en cuanto a la cantidad mínima requerida para el procesamiento en los algoritmos utilizados.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

D. Tiempo para generar estimación en el sistema

Este indicador mide el tiempo en la solución diseñada, con respecto a la usabilidad del usuario en el simulador del sistema web para generar un análisis que obtenga una estimación requerida.

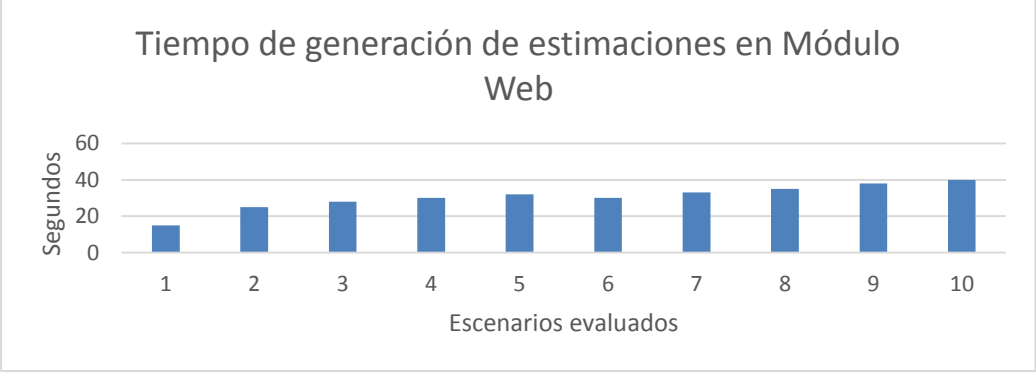
Tabla 12: Tiempo de Procesamiento del Sistema Web

Escenarios Evaluados	Sistema Web
1	15
2	25
3	28
4	30
5	32
6	30
7	33
8	35
9	38
10	40
Promedio	30,60 seg.

Fuente: Elaboración Propia

En la Tabla N° 12 se observa que el tiempo promedio de generación de estimaciones en el sistema web es de 30,60 segundos.

Gráfico 5: Tiempo de generación de pronósticos en Módulo



Fuente: Elaboración Propia

El gráfico N°5 nos permite observar la variación del tiempo de generación de pronósticos en el Módulo Web.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

4.2. Contrastación de la hipótesis.

En cuanto a la contrastación de la hipótesis podemos afirmar que se pudieron ejecutar los objetivos planteados y realizar un análisis comparativo del rendimiento de las técnicas de minería de datos utilizadas en esta investigación para la predicción de ventas. Los objetivos alcanzados fueron realizar una evaluación de las técnicas de minería de las cuales se seleccionó los modelos que mejor se adaptan para este tipo de predicción como son las series de tiempo utilizando los algoritmos de HoltWinters, Holt y ETS debido a que en dichos métodos tiene presentes los componentes de nivel, tendencia y estacionalidad la cual se adapta a la data proporcionada, logrando desarrollar la aplicación que muestra los resultados generados al comparar los métodos utilizados, con el fin de poder analizar los resultados obtenidos.

4.3. Discusión de los resultados.

A. Grado de confiabilidad de los pronósticos generados por el modelo

Con respecto al primer indicador comparando las tres técnicas, es decir Holtwinters, Holt y ETS. Podemos decir que ETS obtuvo el nivel de confianza más elevado en comparación a HoltWinters y Holt, esto se denota en los valores obtenidos al aplicar la fórmula de MAPE (valor real y el monto pronóstico para saber el grado de relación que existe uno con respecto del otro), obteniendo un 84.42% de grado de confianza de HoltWinters, el 83,96 % de grado de confianza de Holt y el 90,51% de grado de confianza de ETS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

lo que significa que en esta comparación ETS es el que mayor grado de confianza obtuvo.

B. Tiempo de procesamiento para obtener la estimación

En el tiempo de procesamiento al evaluar estas técnicas se obtuvo que con el método Holt el tiempo promedio de ejecución de 2.23 segundos siendo superior a diferencia de HoltWinters, que tiene 7,01 segundos y ETS con 22,33 segundos.

C. Número de puntos mínimos para el vector que procesara el modelo

Para el caso del tercer indicador, se evaluó distintos escenarios donde se mide la cantidad de meses que puede procesar cada técnica, en esta evaluación se obtuvo el resultado que HoltWinters, Holt y ETS soportaron la ejecución para todos los escenarios planteados, es decir desde 35 meses hasta vectores pequeños de 25 meses.

D. Tiempo para generar estimación en el sistema

Para el último indicador se obtuvo que, en la usabilidad del sistema web, se generó un tiempo promedio de 30.6 segundos para generar una estimación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V

DESARROLLO DE LA PROPUESTA

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V: DESARROLLO DE LA PROPUESTA

5.1. Generalidades

Es una solución informática que pretende validar certeramente la estimación de ventas de la empresa Astro S.A.C., a partir del descubrimiento de patrones de ventas de cada cliente, los cuales serán analizados aplicando para ello Minería de Datos.

Esta tesis además plantea un análisis descriptivo – comparativo, de las técnicas a utilizar en la creación del modelo predictivo, analizando en primer orden el problema y las variables que se consideran de ingreso y como estas técnicas se utilizarán, además de evaluar los resultados de las mismas.

A. Características del producto

Nombre del Producto: ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE VENTAS

Plataforma y Arquitectura: La Aplicación Web estará desarrollada en el lenguaje de programación PHP junto con el gestor de base de datos SQL SERVER 2014.

Facilidad de Uso: La aplicación permite interactuar con los datos por medio de visualización de reportes.

Adaptabilidad: Esta solución es fácilmente adaptable a empresas dedicadas a la comercialización de artículos deportivos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Grado de Confianza: Con la comparación de los modelos predictivos se obtendrán los resultados específicos y se estimará el margen de error mínimo con las predicciones arrojadas, se escogerá el mejor modelo que brinde mejores resultados.

B. Funciones del Sistema Web

Módulo de Monitoreo de Actividades – El sistema web tendrá la función de monitorear el estado de las actividades (ventas, compras).

Módulo de Estimaciones – El sistema web tendrá la función en realizar estimaciones ventas y compras mensuales y trimestrales seleccionando los algoritmos adecuados.

Módulo de Estudio – Este módulo permite dar origen a un estudio, y establecer los valores necesarios para tal.

C. Usos del producto

Evaluación de usuarios: Brindará la información necesaria para evaluar y confirmar las ventas calculadas en base a las simulaciones en el cual debe fluctuar la nueva estimación de venta.

Ayuda a la toma de decisiones: Permitirá fortalecer la toma de decisiones para mejorar el crecimiento y mejora del área de ventas.

Gráficas de análisis: Se analizará a través de gráficas estadísticas el comportamiento de ventas.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.2. Metodología de desarrollo

Para la siguiente investigación se ha propuesto dividirla en dos etapas, una que comprende todo lo relacionado al desarrollo de modelos de predicción usando la minería de datos, en esta etapa se contemplará todas las fases que se utilizan en la Metodología de desarrollos de modelos de minería de datos, desde la comprensión del negocio, datos iniciales, transformación de datos, modelado y aplicación del algoritmo, evaluación de performance. Para la segunda etapa este informe detallará las fases para el diseño y construcción del sistema web, y como se mostrarán los resultados que permitan a los supervisores y analistas mejorar el análisis de ventas, cabe resaltar que en esta etapa se empleará la Metodología de desarrollo ágil XP. Se aplica el siguiente marco conceptual para el desarrollo de esta investigación:

Dado que la investigación tiene como esquema principal, el modelo de minería de datos se ha resuelto determinar brevemente un cuadro comparativo para la determinación de la metodología que permita resolver esta etapa.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 13: Comparación de Metodologías de Desarrollo de Modelo de Minería de Datos

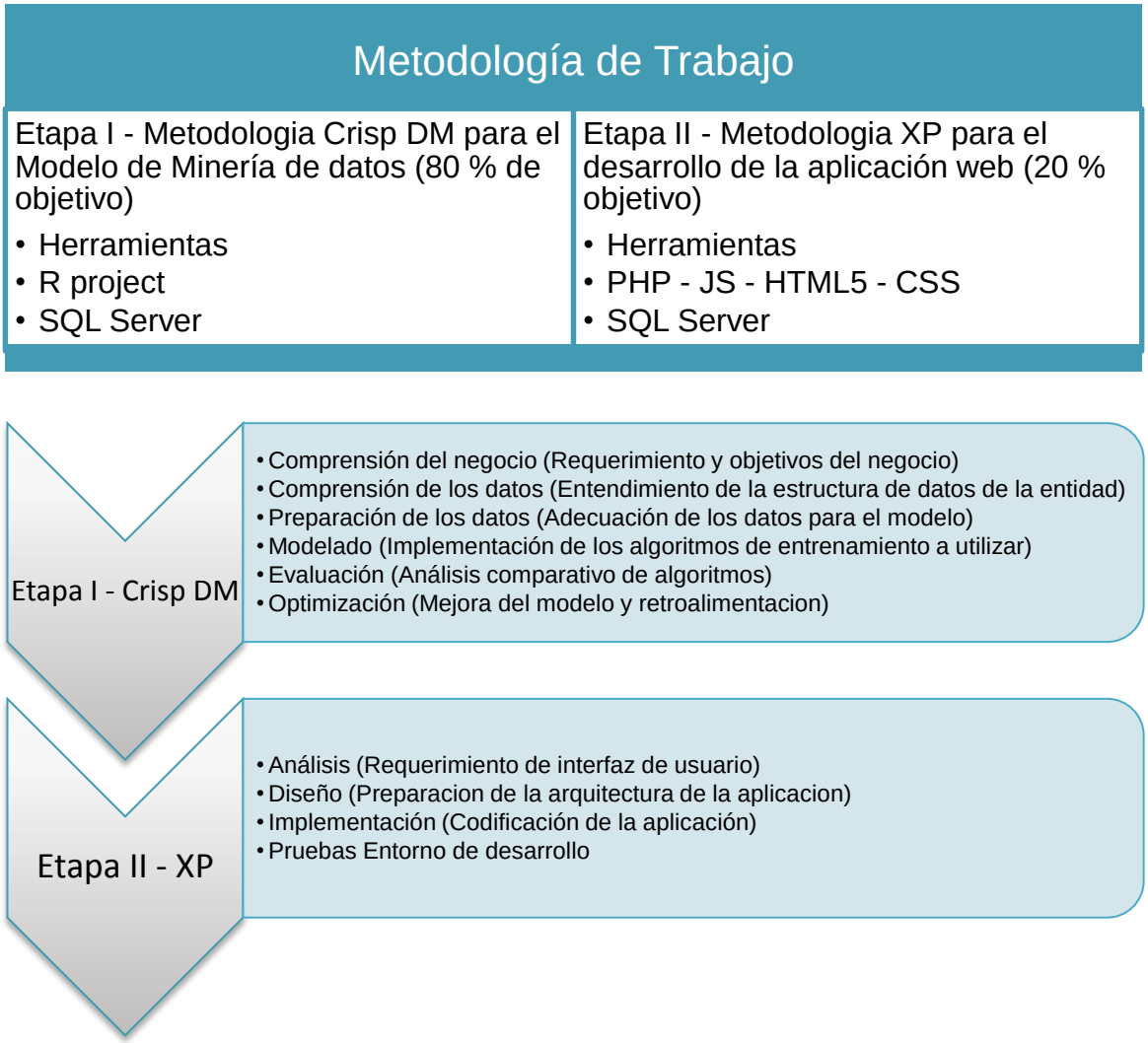
Metodologías de Desarrollo de Modelo de Minería de datos	CRISP-DM	SEMMA
Libre elección de herramientas	2	0
Todas las fases pueden relacionar	2	0
Procesos de Inteligencia de Negocios	2	0
Comercial – Licencias - Privativa	0	2
Técnicas de ETL	2	2
Módulo de referencia para el usuario	2	0
Total	10	4

Fuente: Elaboración Propia

Se asignó pesos a los diferentes criterios para las metodologías (Si=2 y NO=0). Se determina usar CRISP-DM, por ser una metodología flexible en cuanto a herramientas, además que integra el proceso de comprensión de negocio (Gestión del proyecto por objetivos empresariales) y obtuvo el mayor peso, SEMMA es una buena alternativa siempre y cuando se use en proyecto con tecnologías SAS.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Figura 12: Metodología de Trabajo



Fuente: Elaboración Propia

Cabe recalcar que ambas etapas no son consecutivas, aunque si depende una de la otra para su funcionamiento, su desarrollo puede ser dado en un escenario de paralelismo, es decir si bien la aplicación web necesita tener un modelo funcionando con datos para poder visualizar, la construcción de la interfaz web se puede dar desde el momento en que se determinan los objetivos del negocio del modelo de minería.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

1. Etapa I – Diseño del modelo de Minería de datos

A. Comprensión del negocio

a. Descripción del problema

El Astro S.A.C es una empresa de comercialización de artículos deportivos, uno de sus procesos críticos es la venta, cuya frecuencia es diaria. La venta es un proceso por el cual se cuantifica el valor del costo de servicio, para cuantificar este valor es necesario realizar una serie de procesos detallados de la siguiente manera:

Tabla 14: Periodo - Ventas

PERIODO	VENTAS
ABRIL	145 874
MAYO	145 874
JUNIO	158 745

Fuente: Elaboración Propia

Es uno de los principales procesos que realizan

b. Necesidades y Expectativas

b.1.Búsqueda de la mejora en las predicciones con respecto a las ventas de artículos deportivos en un periodo de tiempo determinado.

b.2. Implementar una nueva y mejor técnica para la automatización del proceso de predicción.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c. Objetivos de Negocio

c.1 Analizar tendencias de predicción con respecto a las ventas de artículos deportivos.

c.2 Realizar pronósticos de ventas de forma anual, mensual y trimestral, con base en un nivel de confianza previamente definido en un periodo determinado.

d. Criterios de Éxito

d.1 Confiabilidad de los pronósticos arrojados en un determinado periodo.

d.2 Facilidad de acceso en la interacción del usuario al portal web.

e. Evaluación de la situación

e.1 Se cuenta con la base de datos de ventas de artículos deportivos anual registrada por empresa El Astro S.A.C desde el año 2011. Esta información es usada como fuente principal en el ingreso de los datos necesarios para la creación del modelo de series de tiempo.

f. Requerimientos

f.1 El sistema debe permitir generar reportes para la visualización de las predicciones de las ventas de artículos deportivos en tiempos anuales, mensuales y trimestrales.

f.2 Visualizar la comparación de modelos predictivos y utilizando el mejor para beneficios de la empresa.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

g. Restricciones

g.1 Se requiere la base de datos de todas las ventas desde hace 5 años de antigüedad como mínimo para los procesos de entrenamiento y testeo del modelo.

g.2 De la información obtenida, los datos deben estar libre de errores y valores en valores en blanco.

h. Determinación de los Objetivos de minería de datos

h.1 Objetivos del Proyecto

h.1.1 Generar un modelo de series de tiempo, que arroje predicciones con un alto grado de confianza en un tiempo determinado.

h.2.2 Entrenar el modelo para su mejor eficiencia.

h.3.3 Testear el modelo para el resultado.

h.2 Criterios de éxito del proyecto

h.2.1 Confiabilidad del modelo diseñado e implementado.

h.2.2 Optimización del tiempo para la generación de reportes.

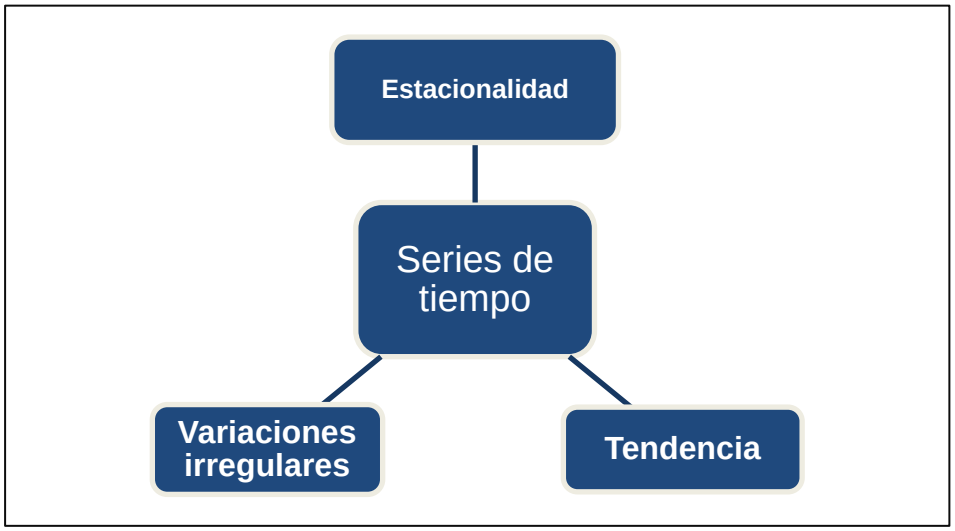
B. Comprensión de los datos

Análisis descriptivo de una serie de tiempo

En esta fase se busca la determinación de existencia de comportamiento estacional en la serie de ventas para lo cual como primer paso será graficarla, esto permitirá identificar su tendencia, estacionalidad y variaciones irregulares.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Figura 13 : Series de Tiempo



Fuente: Elaboración propia

B.1 Recolección de los Datos del Negocio Iniciales

a. Proceso de Adquisición

Los datos obtenidos corresponden a la venta de artículos deportivos de forma anual y mensual.

b. Selección de las Variables a utilizar

- **Tiempo:** Atributo principal para el proceso de predicción, ya que permite el ordenamiento de la serie. Se tomarán datos desde el año 2011-2014.

- **Monto_Ventas:** Atributo que contiene la información de los resultados de las ventas de artículos deportivos en forma anual y mensual. Datos guardados en Microsoft Excel y que serán migrados manualmente a la base de datos creada para el

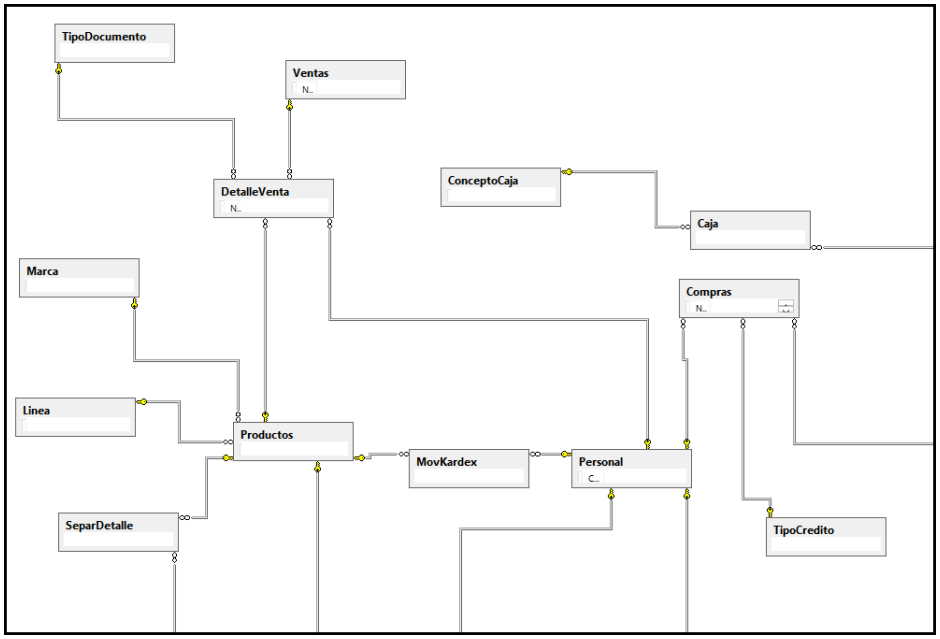
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

proceso de entrenamiento del modelo creado con el software R-Project.

c. Datos y métodos de captura

Los datos son extraídos de la base transaccional de ventas existente en la empresa en SQL 2014.

Figura N° 14 : Diagrama E-R Esquema Ventas



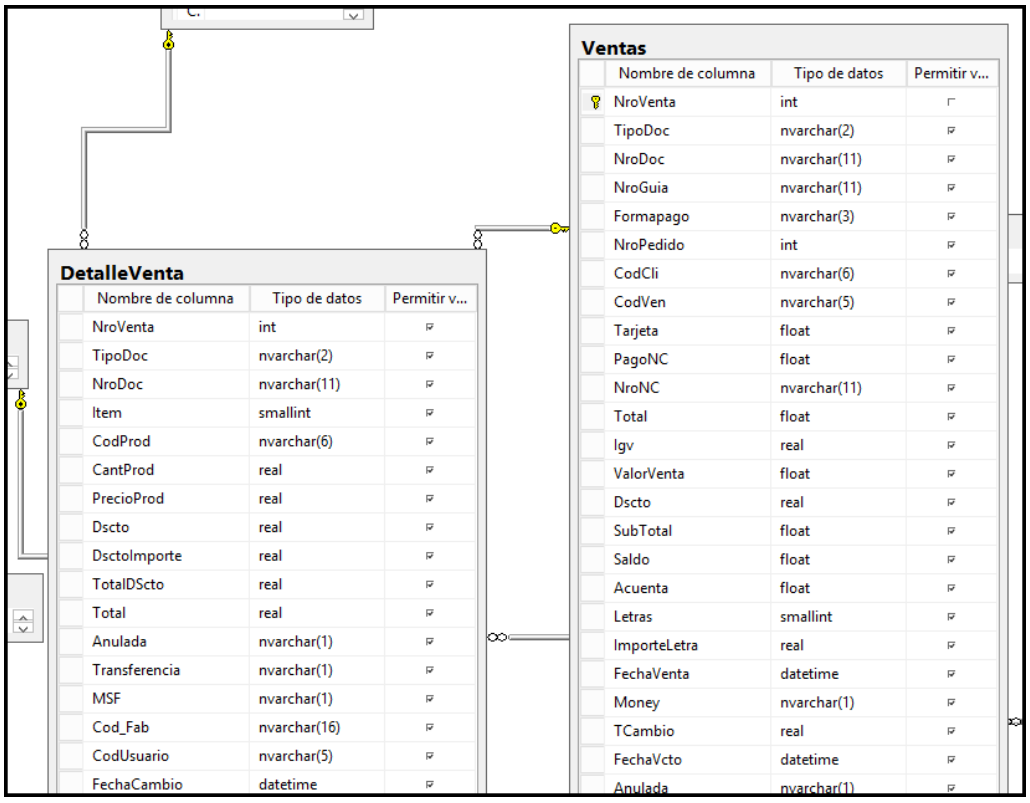
Fuente: Elaboración propia

d. Exploración de Datos

La construcción del modelo de predicción se desarrolla con información obtenida desde el año 2011 hasta el año 2014. Estos datos son los que ingresan en una pequeña base de datos obtenida por la migración de datos en repositorios ofimáticos a la base de datos en el gestor SQL Server 2014 para que realice el entrenamiento del modelo; de los cuales se utiliza el 70% para el entrenamiento y el 30% para las pruebas de predicciones.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Figura 15: Entidades Ventas - DetalleVenta



Fuente: Elaboración propia

Al realizar este proceso de aprendizaje en el modelo se obtiene un valor aproximado que medirá el rendimiento del modelo mostrando el porcentaje de error, el cual deberá ser mínimo para demostrar que el modelo está bien creado con un alto grado de certeza.

Tabla 15: Ventas de Artículos Deportivos 2011-2014

Años / Meses	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
2011	-	-	-	-	-	-	-	290	290	290	290	290
2012	290	290	290	290	290	290	230	309	221	272	269	346
2013	300	269	310	285	306	303	295	295	298	296	298	297
2014	296	296	149	288	290	277						

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

C. Preparación de los datos

c.1 Datos Seleccionados

De la base de datos obtenida, se obtienen diferentes tipos de información con respecto a la venta de artículos deportivos, lo cual son datos relevantes, para ello, se ha realizado un análisis de la data con los atributos a utilizar para el correcto funcionamiento del modelo. Debe considerarse además que se ha analizado y utilizado los campos Anulada, para el proceso de limpieza de datos.

Figura 16 : Scripts SQL para Ventas

```
/*CONFRONTANDO COMPRAS VS VENTAS*/

select anio,mes,sum(total) as totalventas from (
select year(FechaVenta)as anio, MONTH(FechaVenta) as mes,Total from ventas where anulada = 'N') as irene
group by anio,mes order by anio, mes


select top 35 anio,mes,sum(total) as totalventas from (
select year(FechaDoc)as anio, MONTH(FechaDoc) as mes,Total from Compras where Cancelada = 'N') as irene
group by anio,mes order by anio, mes


select sum (totalventas) from (
select sum(total) as totalventas from (
select year(FechaVenta)as anio, MONTH(FechaVenta) as mes,Total from ventas where anulada = 'N'
and year(FechaVenta)=2014 and MONTH(FechaVenta)=06) as irene
group by anio,mes ) as golventas

select count(*) as totaltic from (
select nrodoc from ventas
where anulada = 'N'
and year(FechaVenta)=2014 and MONTH(FechaVenta)=06
group by NroDoc) as gg



select top 10 p.Descripcion, count(*) from DetalleVenta dv
join Productos p on dv.CodProd = p.CodProd
where year(FechaOferta)=2014 and MONTH(FechaOferta)=06
group by p.Descripcion order by count(*) desc


select top 5 p.Descripcion, count(*) from DetalleVenta dv
join Productos p on dv.CodProd = p.CodProd
where year(FechaOferta)=2014 and MONTH(FechaOferta)=06
group by p.Descripcion order by count(*) desc
```

Fuente: Elaboración propia

c.2 Estructuración de los datos

Para la creación del modelo con series de tiempo, los atributos utilizados son identificados de la siguiente manera: al atributo periodo se denota como año y mes; y al atributo monto venta como total, ya que representa el objetivo a predecir, como se

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

muestra en la siguiente imagen. En esta fase preparamos los datos para tener la forma:

Tabla 16: Serie de tiempo – preparación de datos

Años / Meses	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
2011	-	-	-	-	-	-	-	290	290	290	290	290
2012	290	290	290	290	290	290	230	309	221	272	269	346
2013	300	269	310	285	306	303	295	295	298	296	298	297
2014	296	296	149	288	290	277						

Fuente: Elaboración propia

D. Modelado

Tabla 17: Evaluación de las técnicas de minería de datos

Técnicas de Minería	Descripción	Algoritmos	¿Adecuados para esta investigación?
Regresión	Modelos de 2 variables	HoltWinters Holt ETS	SI. Solo se usa ventas y periodo en análisis
Clasificación	Basado en reglas por construcciones lógicas múltiples variables	Árbol de decisiones	No
Asociación	Hechos en común para determinado grupo de datos múltiples variables	A priori FP-Growth Éclat	No
Agrupación	Agrupación de series de vectores en un mapa de dispersión.	K means	No
Redes Neuronales	Modelos matemáticos	Backpropagation Red Neuronal R - nnetar.R	No

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En este caso se propone construir un modelo de minería de datos de pronósticos usando series de tiempo, por lo que se evaluarán las siguientes técnicas usadas en este rubro:

Tabla 18: Modelos de Minería de Datos

Modelo de minería de datos para pronósticos con series de tiempo (Modelos de Regresión)	HoltWinters	Holt	ETS	Arima
Evaluación fundamento teórico				
Modelo parametrizado	SI	SI	SI	SI
Datos estacionales	SI	SI	SI	SI
Método estadístico	SI	SI	SI	SI
Capacidad iterativa (Aprendizaje)	NO	NO	NO	NO
Cantidad de datos de la serie	24	25	25	80
Evaluación fundamento computacional				
Procesamiento CPU	Mínimo	Mínimo	Mínimo	Mínimo
Consumo RAM	Mínimo	Mínimo	Mínimo	Mínimo
Tiempo computacional	Mínimo	Mínimo	Mínimo	Mínimo
Evaluación fundamento objetivo del modelo				
Confiabilidad de precisión pronóstico	Después de pruebas	Después de pruebas	Después de pruebas	Después de pruebas

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Se ha considerado usar HoltWinters, Holt y ETS por requerir un número adecuado de meses, con la que se dispone en el histórico de ventas, sin embargo para este caso debido a la cantidad de datos se cuenta no es factible emplear el algoritmo ARIMA.

HoltWinters – Suavizado exponencial

Holt (1957) y Winters (1960) extendieron el método de Holt para capturar estacionalidad. El método de temporada Holt-Winters comprende la ecuación de pronóstico y tres ecuaciones de suavizado - uno para el nivel tendencia, estacional y variación cíclica (Coghlan, 2015).O

Hay dos variaciones a este método que difieren en la naturaleza del componente estacional. Se prefiere el método aditivo cuando las variaciones estacionales son más o menos constante a través de la serie, mientras que se prefiere el método multiplicativo cuando las variaciones estacionales están cambiando proporcional al nivel de la serie. Con el método aditivo, el componente estacional se expresa en términos absolutos en la escala de la serie observada, y en la ecuación de nivel de la serie se ajusta estacionalmente restando el componente estacional. Dentro de cada año, el componente estacional se suma a aproximadamente cero. Con el método multiplicativo, el componente estacional se expresa en términos relativos (porcentajes) y la serie se ajusta estacionalmente dividiendo a través por el componente estacional. Dentro de cada año, el componente estacional se suma a aproximadamente m.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

MÉTODO – HOLTWINTERS

Los pasos para aplicar este método son:

1. Obtener la serie de tiempo objetivo
2. Si existieran datos nulos/vacíos (completar la serie con el promedio o mediana, según criterio del investigador)
3. Calcular el promedio por año.
4. Identificar y determinar si la serie presenta un esquema aditivo o multiplicativo, para lo cual se emplea las siguientes fórmulas (Ver anexo N° 1: Prueba de Laboratorio – Funcionamiento del HoltWinters ¿Ad o mul).

   - Cálculo de las diferencias estacionales y de los cocientes estacionales (p=frecuencia):

     $$d_t = y_t - y_{t-p}$$
     $$c_t = c_t / c_{t-p}$$

   - Cálcluo del coeficiente de variación

     $$C.V.d_t = Desvest(d_t) / \bar{d}_t$$
     $$C.V.c_t = Desvest(c_t) / \bar{c}_t$$

   C.V. d_t < C.V. c_t → Esquema aditivo
   C.V. c_t < C.V. d_t → Esquema multiplicativo

5. Una vez determinado el esquema, aplicamos el entrenamiento de la serie (dejando el último valor de la serie, para validar el modelo) con las fórmulas correspondientes (según esquema)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Para el Modelo Multiplicativo:

$$F_t = \alpha\left(\frac{D_t}{I_{t-p}}\right) + (1-\alpha)(F_{t-1} + b_{t-1})$$
$$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$$
$$I_t = \gamma\left(\frac{D_t}{F_t}\right) + (1-\gamma)I_{t-p}$$

Para pronosticar periodos futuros se define:

$$\boldsymbol{f_{t+k} = (F_t + kb_t)I_{t+k-p}}$$

Para el caso de Modelo Aditivo:

$$F_t = \alpha(D_t - I_{t-p}) + (1-\alpha)(F_{t-1} + b_{t-1})$$
$$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$$
$$I_t = \gamma(D_t - F_t) + (1-\gamma)I_{t-p}$$

Para pronosticar periodos futuros se define:

$$\boldsymbol{f_{t+k} = F_t + kb_t + I_{t+k-p}}$$

Donde:

D_t Es la observación en el periodo t.

F_t Es el nivel medio desestacionalizado de la serie en el periodo t.

b_t Es la tendencia de la serie en el periodo t, es decir incremento o decremento del nivel medio desestacionalizado durante un periodo.

I_t Es el componente estacional en el periodo t

$\boldsymbol{f_{t+k}}$ Pronóstico para el periodo t+k basado en datos hasta t

$\alpha, \beta, \mathbf{y}\ \gamma$ $(0 < \alpha, \beta, \gamma < 1)$ Son los parámetros de suavizado, asociados con el nivel medio, la tendencia y la estacionalidad respectivamente, siendo p el número de periodos que componen el ciclo estacional.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

6. Los valores de los parámetros anteriormente descritos, pueden ser asignados manualmente por el investigador (ver columnas Dt, Ft, bt, It-p, It de anexo N° 2)
7. Realizar el pronóstico. (ver columna: f t+k de anexo N° 2)
8. Determinar el Error de estimación. (Ver columna: Dt - ft+k de anexo N°2).

Para esta investigación se usa el componente aditivo en las formulas Holtwinters.

Gráfico 6: Algoritmo HoltWinters

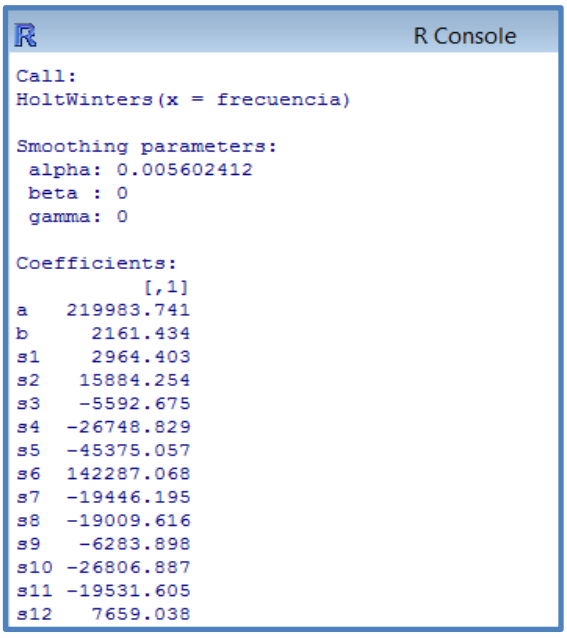
```
HoltWinters <-
function (x,

          # smoothing parameters
          alpha    = NULL, # level
          beta     = NULL, # trend
          gamma    = NULL, # seasonal component
          seasonal = c("additive", "multiplicative"),
          start.periods = 2,

          # starting values
          l.start  = NULL, # level
          b.start  = NULL, # trend
          s.start  = NULL, # seasonal components vector of length `period'

          # starting values for optim
          optim.start = c(alpha = 0.3, beta = 0.1, gamma = 0.1),
          optim.control = list()
          )
{
    x <- as.ts(x)
    seasonal <- match.arg(seasonal)
    f <- frequency(x)

    if(!is.null(alpha) && (alpha == 0))
        stop ("cannot fit models without level ('alpha' must not be 0 or FALSE)")
    if(!all(is.null(c(alpha, beta, gamma))) &&
        any(c(alpha, beta, gamma) < 0 || c(alpha, beta, gamma) > 1))
        stop ("'alpha', 'beta' and 'gamma' must be within the unit interval")
    if((is.null(gamma) || gamma > 0)) {
        if (seasonal == "multiplicative" && any(x == 0))
            stop ("data must be non-zero for multiplicative Holt-Winters")
        if (start.periods < 2)
            stop ("need at least 2 periods to compute seasonal start values")
    }
```

Fuente: (Hyndman R. J., 2015)

El algoritmo inicial será el de Holtwinters, esta es una técnica usada en pronósticos en negocios, da importancia a los meses recientes, para el estudio la frecuencia de distribución será de 12 (Equivalente a los 12 meses de cada año, como se determinó la matriz de ingreso en la preparación de los datos).

Al aplicar el algoritmo Holtwinters al histórico de las ventas el modelo realiza el entrenamiento de la serie donde determina de manera automática e interpretativa los valores de componentes de la serie de tiempo, estos valores son el alpha (variación), beta (tendencia) y gamma (estacionalidad). Su valor oscila entre 0 a 1 Los coeficientes se calculan y también los valores de entrenamiento y residuales:

Gráfico 7: Aplicación de Algoritmo

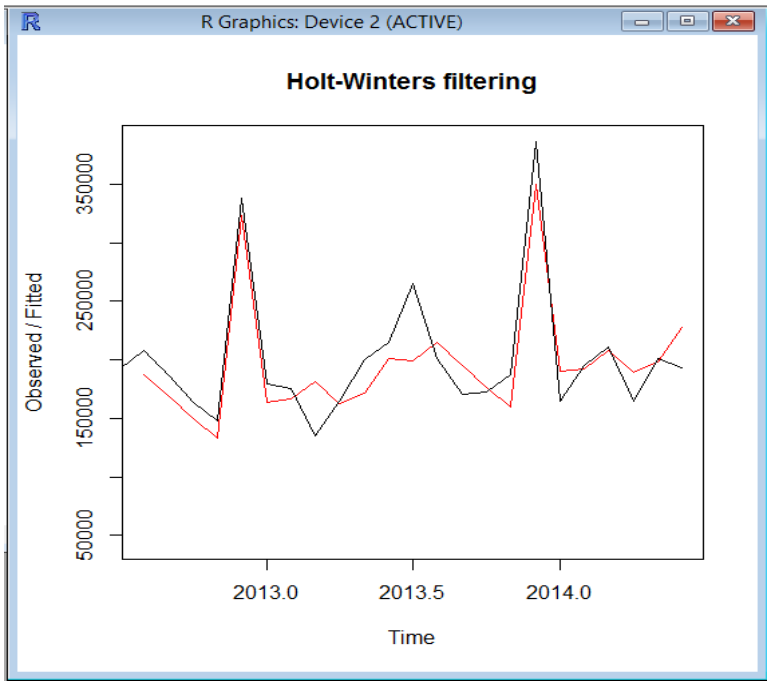


Fuente: (Consola de Ejecución)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En el gráfico N° 8 Se observa la representación del histórico real (color negro) y el entrenamiento generado por el modelo (línea roja), siendo que mientras la línea roja tienda a aproximarse a la real, el pronóstico que genere este aprendizaje tendrá mayor confiabilidad.

Gráfico 8: Valores de entrenamiento vs Valor Real usando Holt-Winters



Fuente: (Consola de Ejecución)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En el gráfico N° 9, podemos observar valores de entrenamiento nivel, tendencia y estacionalidad.

Gráfico 9: Valores de Entrenamiento usando Holt-Winters

R Console

	xhat	level	trend	season
Aug 2012	187651.7	169606.0	2161.434	15884.254
Sep 2012	168452.6	171883.8	2161.434	-5592.675
Oct 2012	149558.6	174146.0	2161.434	-26748.829
Nov 2012	133175.7	176389.3	2161.434	-45375.057
Dec 2012	323079.4	178630.9	2161.434	142287.068
Jan 2013	163588.5	180873.2	2161.434	-19446.195
Feb 2013	166276.8	183125.0	2161.434	-19009.616
Mar 2013	181215.1	185337.5	2161.434	-6283.898
Apr 2013	162593.0	187238.5	2161.434	-26806.887
May 2013	172041.2	189411.4	2161.434	-19531.605
Jun 2013	201551.1	191730.7	2161.434	7659.038
Jul 2013	199093.4	193967.5	2161.434	2964.403
Aug 2013	214546.3	196500.6	2161.434	15884.254
Sep 2013	195158.0	198589.3	2161.434	-5592.675
Oct 2013	176027.9	200615.3	2161.434	-26748.829
Nov 2013	159545.8	202759.5	2161.434	-45375.057
Dec 2013	349522.6	205074.1	2161.434	142287.068
Jan 2014	190155.9	207440.7	2161.434	-19446.195
Feb 2014	192612.4	209460.5	2161.434	-19009.616
Mar 2014	207512.8	211635.3	2161.434	-6283.898
Apr 2014	189173.1	213818.5	2161.434	-26806.887
May 2014	198472.0	215842.2	2161.434	-19531.605
Jun 2014	227836.8	218016.3	2161.434	7659.038

Fuente: *(Consola de Ejecución)*

En el grafico N° 10, se muestra los coeficientes con los que trabaja Holt-Winters:

Gráfico 10: Coeficientes de Holt-Winters

```
$coefficients
         a          b         s1         s2         s3         s4         s5
219983.741   2161.434   2964.403  15884.254  -5592.675 -26748.829 -45375.057
        s6         s7         s8         s9        s10        s11        s12
142287.068 -19446.195 -19009.616  -6283.898 -26806.887 -19531.605   7659.038
```

Fuente: (Consola de Ejecución)

Una vez obtenido estos valores se aplica la extracción del pronóstico requerido, en este caso solo interesa el próximo valor futuro, además se calcula la mínima y máxima como

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

rango esperado, en el grafico N° 11, se aplica el entrenamiento para sacar pronóstico de los próximos tres meses.

Gráfico 11: Predicciones aplicando Holtwinters

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2014	225109.6	192811.9	257407.3	175714.5	274504.7
Aug 2014	240190.9	207892.6	272489.1	190795.0	289586.7
Sep 2014	220875.4	188576.6	253174.1	171478.7	270272.0

Fuente: (Consola de Ejecución)

Método de Holt - Alisado Exponencial

(De la Fuente Fernández, 2011) El método propuesto por Holt es un método de alisado exponencial que utiliza dos parámetros de alisado en lugar de uno. Es aplicable también a series que tengan una tendencia aproximadamente lineal y ha dado muy buenos resultados en la previsión de distintas áreas de la economía empresarial: gestión de stocks, financiación, ventas, etc.

Las dos variables alisadas que se calculan en este método tienen una relación directa e inmediata con los parámetros del modelo lineal

$$X_t = b_0 + b_1 t + \varepsilon_t$$

La estimación del término dependiente utilizando t observaciones da la ordenada o nivel de la tendencia para ese punto.

En el método Holt se calculan directamente dos variables de alisado por cada momento del tiempo:

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

S_t : estimación del *nivel* de la serie en t

➢ La ecuación de *predicción en el método de Holt:* $\hat{X}_{(t+m)/t} = S_t + b_{1t}m$ (5)

na

🕮 Para la aplicación de las ecuaciones $\begin{cases} b_{1t} = \beta (S_t - S_{t-1}) + (1-\beta) b_{1t-1} \\ \hat{X}_{(t+m)/t} = S_t + b_{1t}m \end{cases}$ es necesario conocer los valores iniciales de S_0 y b_{10}, así como los coeficientes de alisado de α y β

y

Los valores iniciales para comenzar la recursión se puede obtener directamente a partir de los coeficientes ($\hat{b}_0$ y $\hat{b}_1$) obtenidos en el ajuste de una recta de regresión por mínimos cuadrados utilizando toda la información disponible, haciendo: $\begin{cases} S_0 = \hat{b}_0 \\ b_{10} = \hat{b}_1 \end{cases}$

primera ecuación o ecuación de nivel: $S_t = \alpha X_t + (1-\alpha)(S_{t-1} + b_{1t-1})$ (3)

Se observa que esta ecuación es una media ponderada entre X_t y $(S_{t-1} + b_{1t-1})$, siendo $(S_{t-1} + b_{1t-1})$ una estimación de $[S_{t-1} = b_0 + b_1(t-1)]$ donde se ha sustituido el parámetro de la pendiente b_1 por b_{1t-1}.

Así pues, la *primera ecuación o ecuación de nivel* proporciona directamente el valor del nivel de la tendencia en el momento t, teniendo el mismo papel que b_{0t} en el método del AED.

➢ La segunda ecuación del método de Holt permite, a su vez, calcular la pendiente b_{1t} de forma recursiva, mediante la ecuación de alisado:

segunda ecuación o ecuación de nivel: $b_{1t} = \beta (S_t - S_{t-1}) + (1-\beta) b_{1t-1}$ (4)

Como estimación de la *pendiente* se toma la diferencia entre el nivel de la tendencia en t y en $(t-1)$: $(S_t - S_{t-1})$. En el segundo término, como ocurre en todas las ecuaciones de alisado exponencial, aparece la variable alisada con un retardo.

Con objeto de determinar el valor de la pendiente b_{1t} sobre el eje de ordenadas, se toma como referencia la paralela al eje de abscisas que pasa por S_{t-1}.

En el corte realizado en t se observa que b_{1t} está situado entre $(S_t - S_{t-1})$ y b_{1t-1}, siendo este último término la *pendiente* que ya se había obtenido en $(t-1)$.

1.

Fuente: (De la Fuente Fernández, 2011)

En esta investigación se usa el método Holt suavizado exponencial, con el siguiente código:

Gráfico 12: Script del algoritmo Holt

```
holt <- function (x, h = 10, damped = FALSE, level = c(80, 95), fan = FALSE,
          initial=c("optimal","simple"), exponential=FALSE, alpha=NULL, beta=NULL, ...)
{
  initial <- match.arg(initial)
  if(initial=="optimal" | damped)
  {
        if(exponential)
            fcast <- forecast(ets(x, "MMN", alpha=alpha, beta=beta, damped = damped, opt.crit="mse"), h, level = level, fan = fan, ...)
          else
            fcast <- forecast(ets(x, "AAN", alpha=alpha, beta=beta, damped = damped, opt.crit="mse"), h, level = level, fan = fan, ...)
        }
  else
    fcast <- forecast(HoltWintersZZ(x, alpha=alpha, beta=beta, gamma=FALSE, exponential=exponential),
        h, level = level, fan = fan, ...)
  if (damped)
  {
          fcast$method <- "Damped Holt's method"
    if(initial=="simple")
      warning("Damped Holt's method requires optimal initialization")
  }
        else
          fcast$method <- "Holt's method"
  if(exponential)
          fcast$method <- paste(fcast$method,"with exponential trend")
  fcast$model$method <- fcast$method
  fcast$model$call <- match.call()
  return(fcast)
}
```

Al aplicar el algoritmo Holt al histórico de las ventas obtenemos el siguiente resultado:

Gráfico 13: Aplicación del algoritmo Holt

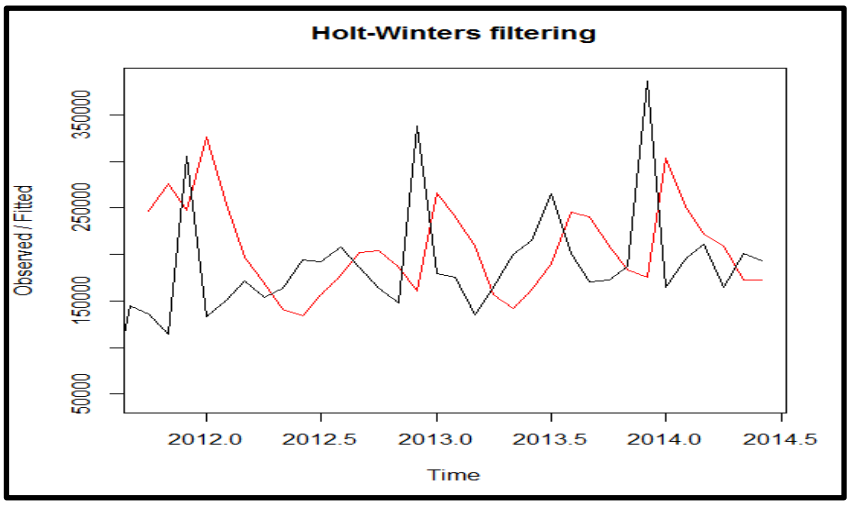
```
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = frecuencia, gamma = FALSE)

Smoothing parameters:
 alpha: 0.4268688
 beta : 0.5224288
 gamma: FALSE

Coefficients:
         [,1]
a 181607.887
b  -6999.623
```

Fuente: (Consola de Ejecución)

En el gráfico N° 14 de similar manera en el gráfico de datos reales y entrenamiento de Holtwinters, se pueden apreciar los valores de la serie al aplicar Holt, en este caso se observa una tendencia no ajustada entre las dos representaciones.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 14: Valores de entrenamiento vs Valor Real usando Holt



Fuente: (Consola de Ejecución)

En el gráfico N° 15 se muestra los valores de entrenamiento nivel y tendencia.

Gráfico 15: Valores de Entrenamiento – Holt

R Console

	xhat	level	trend
Oct 2011	246174.9	145014.5	101160.4400
Nov 2011	275939.0	199278.5	76660.4265
Dec 2011	247760.9	207076.3	40684.5736
Jan 2012	326128.4	272512.7	53615.6507
Feb 2012	254151.7	243633.7	10518.0612
Mar 2012	197361.8	209940.8	-12578.9885
Apr 2012	168229.6	186488.9	-18259.3163
May 2012	140887.9	162263.9	-21375.9940
Jun 2012	134570.3	150778.9	-16208.6340
Jul 2012	156946.6	159914.6	-2968.0081
Aug 2012	177099.5	172133.5	4966.0479
Sep 2012	202426.2	190473.3	11952.8987
Oct 2012	203988.8	195601.4	8387.3929
Nov 2012	186496.2	186989.7	-493.4181
Dec 2012	160646.4	169841.0	-9194.5906
Jan 2013	266401.5	236150.6	30250.9536
Feb 2013	240312.4	229394.8	10917.5645
Mar 2013	209046.5	212604.4	-3557.8823
Apr 2013	157185.5	177318.8	-20133.3236
May 2013	141899.3	160369.3	-18470.0256
Jun 2013	161324.7	166790.7	-5466.0088
Jul 2013	190751.1	184243.6	6507.4934
Aug 2013	245791.1	222629.4	23161.6365
Sep 2013	240207.6	226910.0	13297.6192
Oct 2013	208519.4	210658.9	-2139.5024
Nov 2013	183262.8	193335.0	-10072.2396
Dec 2013	175552.5	184814.2	-9261.7309
Jan 2014	303139.7	265441.1	37698.6547
Feb 2014	250994.0	244125.9	6868.1252
Mar 2014	221465.0	227086.7	-5621.7107
Apr 2014	209306.9	217171.6	-7864.7169
May 2014	172371.3	190211.8	-17840.5333
Jun 2014	172966.0	184480.4	-11514.3799

Fuente: (Consola de Ejecución)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En el grafico N° 16, se muestra los coeficientes con los que trabaja Holt.

Gráfico 16: Coeficientes de Holt

```
$coefficients
           a           b
181607.887  -6999.623
```

Fuente: (Consola de Ejecución)

Una vez obtenido estos valores se aplica la extracción del pronóstico requerido, en el grafico N° 11, se aplica el entrenamiento para sacar pronóstico de los próximos tres meses

Gráfico 17: Predicciones aplicando Holt

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2014	174608.3	65650.4577	283566.1	7971.679	341244.9
Aug 2014	167608.6	37663.4719	297553.8	-31125.346	366342.6
Sep 2014	160609.0	-422.7374	321640.8	-85667.800	406885.8

Fuente: (Consola de Ejecución)

ETS – EXPONENTIAL SMOOTHING STATE

Es una regla de la técnica general para suavizar los datos de series de tiempo, sobre todo para aplicar de forma recursiva hasta tres filtros de paso bajo con funciones de la ventana exponenciales. Estas técnicas tienen amplia aplicación que no está destinado a ser rigurosamente exactos o fiables para cada situación. Es un

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

procedimiento fácil de aprender y fácil de aplicar para el cálculo aproximado o recordar algún valor, o para hacer alguna determinación sobre la base de suposiciones previas por parte del usuario, tales como la estacionalidad.

En este caso supondremos que la serie Z1, Z2,..., Zn exhibe una tendencia cuadrática por consiguiente los métodos de suavizamiento exponencial simple y doble no son aplicables a series con este tipo de comportamiento. La idea es considerar un triple suavizamiento siguiendo los lineamientos con los cuales fue construido el suavizamiento exponencial doble. Es decir, el modelo en cuestión es:

$$Z_t = \beta_0 + \beta_1 t + \beta_2 t^2, t \geq 1.$$

las ecuaciones del suavizamiento son las siguientes:

$$\begin{aligned}
\overline{\overline{Z}}_t &= \alpha \overline{\overline{Z}}_t - (1-\alpha)\overline{Z}_{t-1}, \\
\overline{\overline{Z}}_t &= \alpha \overline{Z}_t + (1-\alpha)\overline{\overline{Z}}_{t-1}, \\
\overline{Z}_t &= \alpha Z_t + (1-\alpha)\overline{Z}_{t-1}.
\end{aligned}$$

La ecuación de predicción para $t = n + k$ es

$$\widehat{Z}_n(k) = a_n + b_n k + c_n k^2, \ k \geq 1$$

donde a_n, b_n y c_n son estimaciones de los parámetros β_1, β_2 y β_3 respectivamente, dadas por

$$\begin{aligned}
a_n &= 3\overline{Z}_n - 3\overline{\overline{Z}}_n + \overline{\overline{\overline{Z}}}_n, \\
b_n &= \frac{\alpha^2}{2(1-\alpha)}\left((6-5\alpha)\overline{Z}_n - 2(5-4\alpha)\overline{\overline{Z}}_n + (4-3\alpha)\overline{\overline{\overline{Z}}}_n\right), \\
c_n &= \frac{\alpha^2}{2(1-\alpha)}\left(\overline{Z}_n - 2\overline{\overline{Z}}_n + \overline{\overline{\overline{Z}}}_n\right).
\end{aligned}$$

Para estimar la constante α procedemos como antes. Para iniciar el algoritmo podemos usar por ejemplo $\overline{Z}_1 = \overline{\overline{Z}}_1 = \overline{\overline{\overline{Z}}}_1 = Z_1$. Otra forma de inicializar el proceso recursivo es resolviendo el sistema

$$\begin{aligned}
a_0 &= 3\overline{Z}_0 - 3\overline{\overline{Z}}_0 + \overline{\overline{\overline{Z}}}_0, \\
b_0 &= \frac{\alpha^2}{2(1-\alpha)}\left((6-5\alpha)\overline{Z}_0 - 2(5-4\alpha)\overline{\overline{Z}}_0 + (4-3\alpha)\overline{\overline{\overline{Z}}}_0\right), \\
c_0 &= \frac{\alpha^2}{2(1-\alpha)}\left(\overline{Z}_0 - 2\overline{\overline{Z}}_0 + \overline{\overline{\overline{Z}}}_0\right).
\end{aligned}$$

Los valores de a_0, b_0 y c_0 se pueden obtener usando mínimos cuadrados en el modelo

$$Z_t = a + bt + ct^2 + \epsilon_t, \ t = 1, 2, \ldots, n_0,$$

donde $n_0 << n$.

Fuente (Vallejos, 2012)

Para esta investigación se usa el algoritmo suavizamiento exponencial simple aditivo, con el siguiente código:

Gráfico 18: Script del Algoritmo ETS

```
ets <- function(y, model="ZZZ", damped=NULL,
    alpha=NULL, beta=NULL, gamma=NULL, phi=NULL, additive.only=FALSE, lambda=NULL,
    lower=c(rep(0.0001,3), 0.8), upper=c(rep(0.9999,3),0.98),
    opt.crit=c("lik","amse","mse","sigma","mae"), nmse=3, bounds=c("both","usual","admissible"),
    ic=c("aicc","aic","bic"),restrict=TRUE, allow.multiplicative.trend=FALSE,
    use.initial.values=FALSE, ...)
{
  #dataname <- substitute(y)
  opt.crit <- match.arg(opt.crit)
  bounds <- match.arg(bounds)
  ic <- match.arg(ic)

  #if(max(y,na.rm=TRUE) > 1e6)
  #    warning("Very large numbers which may cause numerical problems. Try scaling the data first")

  if(any(class(y) %in% c("data.frame","list","matrix","mts")))
    stop("y should be a univariate time series")
  y <- as.ts(y)

  # Check if data is constant
  if (is.constant(y))
    return(ses(y, alpha=0.99999, initial='simple')$model)

  # Remove missing values near ends
  ny <- length(y)
  y <- na.contiguous(y)
  if(ny != length(y))
    warning("Missing values encountered. Using longest contiguous portion of time series")
```

Fuente: (Hyndman R. J., 2015)

Al aplicar el algoritmo ETS al histórico de las ventas obtenemos el siguiente resultado:

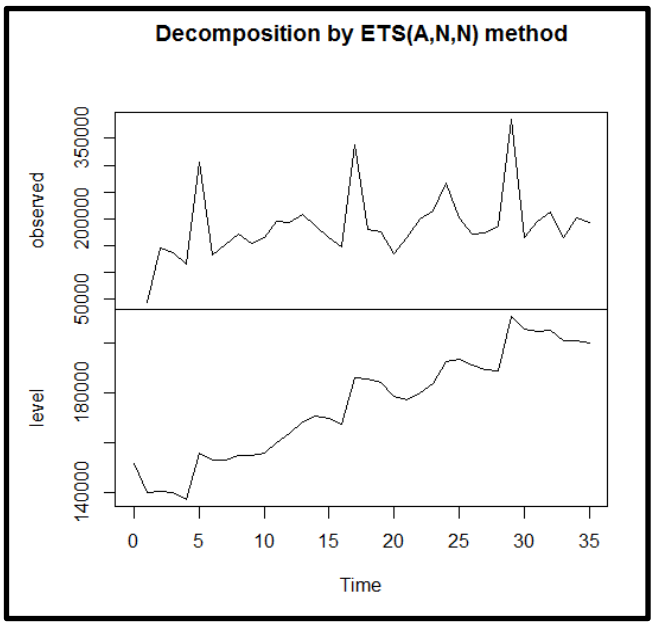
Gráfico 19: Aplicación del algoritmo ETS

```
ETS(A,N,N)

Call:
 ets(y = vectormonto, model = "ANN")

  Smoothing parameters:
    alpha = 0.1097

  Initial states:
    l = 151920.8444

  sigma:  61184.73

     AIC     AICc      BIC
899.9529 900.3279 903.0636
```

Fuente: (Consola de Ejecución)

En el gráfico N° 20 se puede apreciar la descomposición estacional que realiza el algoritmo ETS para el entrenamiento de la serie.

Gráfico 20: Gráfico de entrenamiento vs Valor Real usando ETS



Fuente: (Consola de Ejecución)

En el gráfico N° 21 muestra los valores de entrenamiento de las ventas teniendo en cuenta la frecuencia.

Gráfico 21: Valores de Entrenamiento – ETS

```
Time Series:
Start = 1
End = 35
Frequency = 1
 [1] 186763.6 186749.3 186745.1 186740.1 186732.8 186744.7 186739.4 186735.7
 [9] 186734.2 186731.0 186728.7 186729.4 186730.0 186732.2 186732.2 186729.9
[17] 186726.0 186741.1 186740.4 186739.2 186734.0 186731.8 186733.2 186736.0
[25] 186743.9 186745.4 186743.8 186742.4 186742.4 186762.4 186760.2 186761.0
[33] 186763.5 186761.3 186762.7
```

Fuente: (Consola de Ejecución)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En el grafico N° 22, se aplica el entrenamiento para obtener el pronóstico de los próximos tres meses

Gráfico 22: Predicciones aplicando ETS

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
36	186763.3	109173.9	264352.7	68100.57	305426.1
37	186763.3	109173.9	264352.7	68100.57	305426.1
38	186763.3	109173.9	264352.7	68100.56	305426.1

Fuente: (Consola de Ejecución)

E. Evaluación

e.1 Evaluar los resultados

e.2 Objetivos – Criterios de Evaluación del Negocio

e.2.1 Analizar tendencias de ventas:

El modelo permite generar gráficos de series donde se puede apreciar la tendencia de la variable “ventas” en los años y meses que se han pronosticado.

e.2.2 Realizar pronósticos de producción para el próximo trimestre:

El modelo muestra los pronósticos generados para los próximos meses y años.

e.2.3 Confianza del Modelo:

Se puede llegar a determinar la confianza de los pronósticos haciendo comparaciones con valores reales, pues mientras menor sea el error, mayor será la confianza del modelo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

e.3 Objetivos - Criterios de Evaluación del Proyecto

e.3.1 Se ha generado un modelo de series de tiempo, que permite realizar pronósticos futuros a 3 saltos siguientes.

e.3.2 El modelo permite entrenar y testear la data para obtener un porcentaje de confianza del mismo.

ALGORITMO PARA ANALISIS Y COMPARACION DE RESULTADOS ENTRE TECNICAS – VER ANEXO N° 3: PLAN DE PRUEBAS

```
#PASO 1 - INSTANCIA DE LIBRERIAS
library(RODBC)
library(forecast)
options(max.print = 99999999)

#PASO 2 - ESTABLECER CONEXION CON BASE DE DATOS SQL SERVER
consql <- odbcDriverConnect('driver={SQL Server};server=localhost;database=comvenSQL;trusted_connection=true')

#PASO 3 - BORRAR LOS DATOS DE LA TABLA DETLAB OPCIONAL
del<-sqlQuery(consql, "delete from detlab")


#PASO 4 - REALIZAR CONSULTA DE EXTRACCION DE DATOS HISTORICOS
historico<-sqlQuery(consql, paste("select anio,mes,sum(total) as totalventas from (
select year(FechaVenta)as anio, MONTH(FechaVenta) as mes,Total from ventas where anulada = 'N') as irene group by anio,mes order by anio,


#PASO 5 - TRANSFORMANDO LOS DATOS
colmonto <- historico[3]
vectormonto <-as.vector(t(colmonto))
cont <- length(vectormonto)

ar <- historico[1]
ar2<-as.vector(t(ar))
vanioi<-as.numeric(ar2[1])

br <- historico[2]
br2<-as.vector(t(br))
vmesi<-as.numeric(br2[1])


#PASO 6 - DECLARA MESES HACIA ATRAS A PARTIR DEL ULTIMO MES HISTORICO REGISTRADO EN EL VECTOR

cmestest <- 6
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Datos Generados, simulación entre datos históricos y estimaciones realizadas con ambos algoritmos

	IdDetalleLab	IdLaboratorio	IdPeriodoDestino	ValorReal	A1HoltWinters	A2Holt	A3ETS
1	25	1	20141	164891,5	203112,456114073	252664,364193807	207161,9270972
2	26	1	20142	194987,9	217362,012060682	222369,47732274	205319,154232777
3	27	1	20143	211407,09	198673,48561586	209334,765425924	206194,540304057
4	28	1	20144	164574	200435,65539188	171514,717579959	186049,64401071
5	29	1	20145	200738,6	229804,804866049	172390,203781113	186415,820725618
6	30	1	20146	193210,8	225109,577607358	174608,264934083	186763,318092522

2. Etapa II – Metodología XP para el desarrollo de aplicación web

A. Fase I: Gestión de Proyecto

a.1 Planificación del Proyecto

Tabla 19: Prioridad y Dificultad de Historia de Usuario

HISTORIA DE USUARIO	PRIORIDAD	N° ITERACIONES
1. **CONSULTAR Y GENERAR REPORTES DE ACTIVIDAD MENSUAL Y ANUAL**	**ALTA**	**3**
2. **GENERAR PROYECCIONES Y ESTIMACIONES.**	**ALTA**	**3**
3. **GESTION DE USUARIOS.**	**MEDIA**	**2**
4. **GESTION DE REPORTES**	**MEDIA**	**2**

Fuente: Extraído de la Metodología XP

La prioridad es definida por el aspecto del sistema, es decir, la función principal en este caso está representado por las dos primerias historias

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

de usuario, que hacen referencia al tratamiento de los datos, dejando de lado en menor grado a las siguientes historias como la gestión de usuarios o la de reportes. Por lo que es necesario al finalizar la primera iteración que los entregables cuenten con un avance satisfactorio para el cliente ofreciendo un producto funcional.

a.2 Diario de Actividades

Una vez obtenida la prioridad por historia de usuarios, y en función a la dificultad y número de iteraciones, se establece el siguiente cronograma de actividades que permite tener un mejor control sobre las iteraciones y los documentos entregables de esta.

Tabla 20: Esquema de Diario de Actividades

ACTIVIDADES	TIEMPO											
	MES1				MES2				MES3			
HISTORIA DE USUARIO 1	I1	I1	I1	I1	I1	I1	I1	I1	I1	I		
HISTORIA DE USUARIO 2	I1	I1	I1	I1	I1	I1	I1	I1	I1			
HISTORIA DE USUARIO 3							I1			I4		
HISTORIA DE USUARIO 4							I1			I4		

Fuente: Extraído de la Metodología XP

a.3 Historia de usuario detallado

Tabla 21: Requerimiento 01

Historia de Usuario	
Número: 1	**Usuario: Supervisor de Ventas**
Nombre historia: CONSULTAR Y GENERAR REPORTES DE ACTIVIDAD MENSUAL Y ANUAL	
Prioridad en negocio: **Alta**	**Riesgo en desarrollo:** **Baja**
Entrevistado:	
Descripción: **El Jefe de departamento podrá acceder al módulo de monitoreo de información anual.**	
Observaciones:	

Fuente: Elaboración Propia

Tabla 22: Requerimiento 02

Historia de Usuario	
Número: 2	**Usuario: Gerente General**
Nombre historia: GENERAR PROYECCIONES Y ESTIMACIONES	
Prioridad en negocio: **Alta**	**Riesgo en desarrollo:** **Baja**
Entrevistado: Gerente General y Jefe de Ventas	
Descripción: **El Gerente y jefe de departamento podrán acceder al módulo de proyecciones y estimaciones donde podrán simular con los datos cualquier escenario posible que le permita el sistema de análisis, puede visualizar el modelo por defecto o generar nuevos valores a partir de simulaciones.**	
Observaciones:	

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 23: Requerimiento 03

Historia de Usuario	
Número: 3	**Usuario: Administrador del Sistema**
Nombre historia: GESTIÓN DE USUARIOS	
Prioridad en negocio: **Alta**	**Riesgo en desarrollo:** **Baja**
Entrevistado:	
Descripción: El sistema contará con 2 niveles de usuario: Administrador y Operarios. Cada uno de ellos tendrá restricciones en el sistema. Administrador: Acceso a todos los módulos del sistema. Operarios: Acceso a la visualización de reportes del modelo, que son los resultados de las predicciones. El sistema debe permitir, visualizar y estructurar nuevos reportes.	
Observaciones:	

Fuente: Elaboración Propia

Tabla 24: Requerimiento 04

Historia de Usuario	
Número: 4	**Usuario: Gerente General**
Nombre historia: Administración del Sistema	
Prioridad en negocio: **Alta**	**Riesgo en desarrollo:** **Baja**
Entrevistado: Gerente General	
Descripción: **El administrador de sistema tiene potestad de dar de alta, edición o baja a los reportes del portal y establecer los permisos según negocio donde sean dirigidos.**	
Observaciones:	

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

a.4 Requerimientos no funcionales

- **Facilidad de Uso:** Se pretende que el sistema se muestre en un entorno amigable y de fácil uso. De modo que el impacto que sufrirán los usuarios para comprender la información mostrada sea de fácil acceso.
- **Escalabilidad:** La herramienta permitirá generar nuevos reportes según los nuevos requerimientos.
- **Portabilidad:** La solución usa herramientas de Microsoft con el fin de adaptarse a las herramientas utilizadas en la entidad.
- **Seguridad:** Gestionar el acceso de los usuarios y los permisos a los módulos del sistema.

a.5 Identificación de Stakeholders

a.5.1 Jefe de Ventas, encargado del área de facturación y los procesos dentro de este. Es quien se encarga de los procedimientos para el análisis, estrategias y ver los reportes de las ventas.

a.5.2 Supervisor de Ventas es ayudante del Jefe de ventas, encargado de velar por el proceso de ventas.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

B. FASE II: Diseño

Base de Datos Relacional

El sistema está diseñado para cumplir dos propósitos, la captura de los datos que viene a ser la migración de los documentos ofimáticos en función al consolidado de producción mensual, siendo este la mínima unidad representativa de tiempo registrado, por lo tanto, el sistema contempla esta captura de datos y el almacenamiento de información por parte de la ejecución del modelo, así como los datos administrativos del sistema.

C. FASE III: fase de Implementación

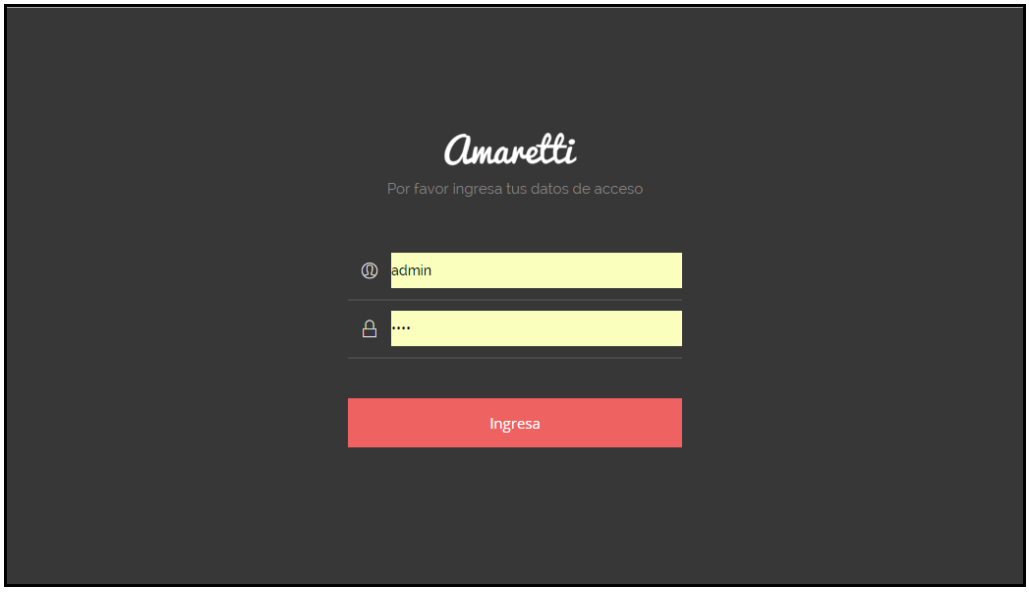
La codificación se realizó en lenguaje PHP con HTML5.

Gráfico 23: Codificación HTML5 y PHP

Fuente: Elaboración Propia

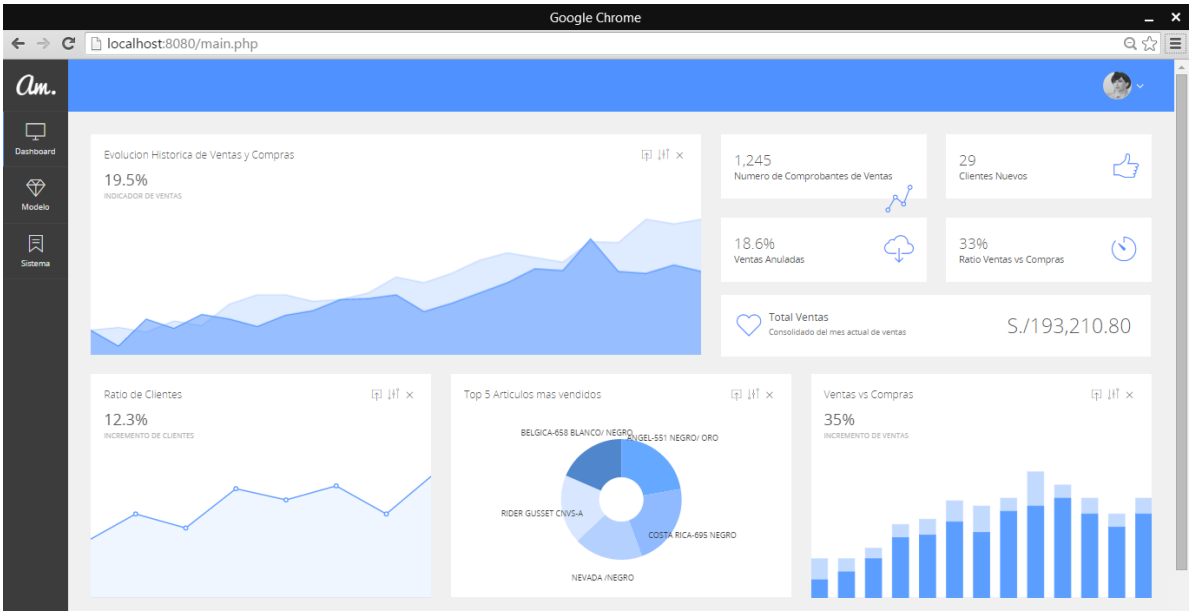
Con lo cual se obtuvo las siguientes interfaces.

Gráfico 24: Inicio de sistema



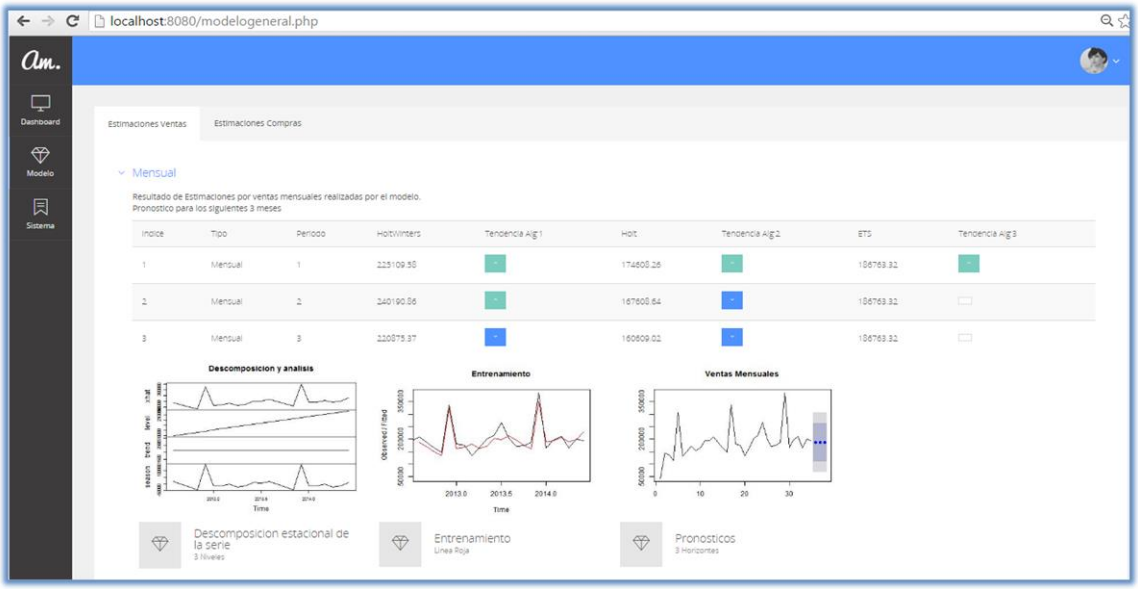
Fuente: Elaboración Propia

Gráfico 25: Indicadores consolidados de Ventas



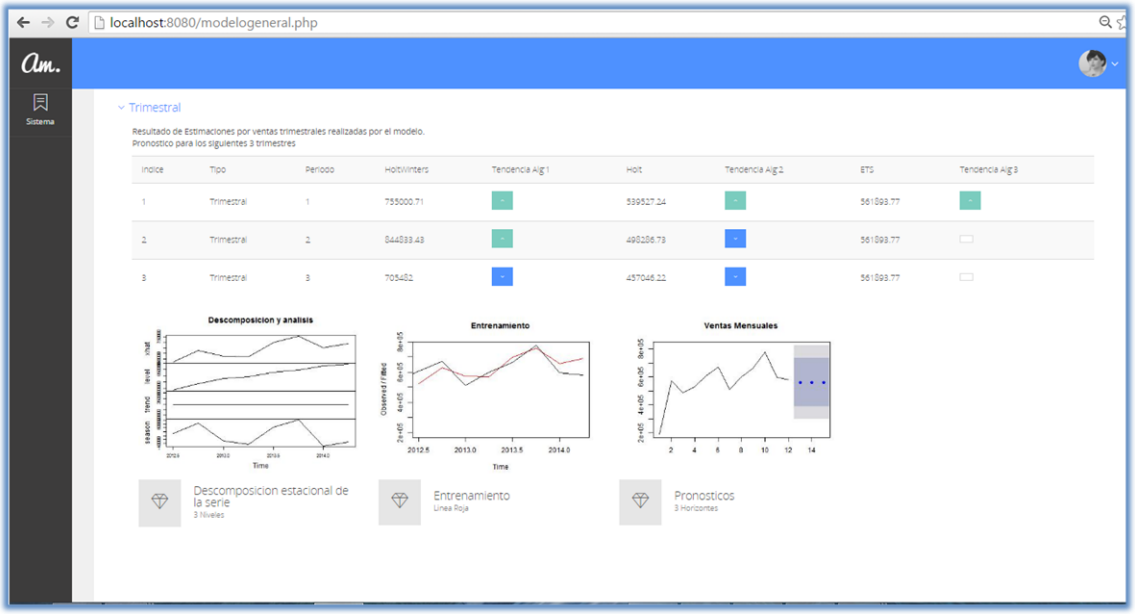
Fuente: Elaboración Propia

Gráfico 26: Resultados del modelo



Fuente: Elaboración Propia

Gráfico 27: Generador de estimaciones y proyecciones



Fuente: Elaboración Propia

CAPITULO VI

CONCLUSIONES Y RECOMENDACIONES

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

a. Se realizó la evaluación de las técnicas de minería de datos Regresión, Series temporales, Redes Neuronales, Agrupamiento teniendo como resultado que las técnicas más adecuadas para el ámbito de ventas en esta investigación son las de series de tiempo, siendo muy utilizado por las ciencias de econometría, los métodos de Holtwinters (sector financiero) como lo afirma (Calvo Rodríguez, 2008) en su investigación, cabe destacar que el uso de redes neuronales para la determinación de pronósticos financiero no ha sido del todo aceptado por la comunidad, ya que al ser un proceso iterativo variante, no hay un principio exacto que permita demostrar obtener dicho resultado.

b. Se realizó el análisis comparativo de técnicas de minería de datos con lo cual se demostró que para esta investigación las de series temporales se ajusta a nuestro estudio para lo cual dicha comparación y análisis se muestra en la tabla N° 17, de acuerdo a los criterios de selección se obtuvo que para el presente trabajo de investigación las técnicas más adecuadas son HoltWinters, Holt y ETS como se muestra en la tabla N° 18. Siendo ETS el que mejor confiabilidad presenta, Podemos decir que ETS obtuvo el nivel de confianza más elevado en comparación a HoltWinters y Holt, esto se denota en los valores obtenidos al aplicar la fórmula de MAPE (valor real y el monto pronóstico para saber el grado de relación que existe uno con respecto del otro), obteniendo un 84.42% de grado de confianza de HoltWinters,

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

el 83,96 % de grado de confianza de Holt y el 90,51% de grado de confianza de ETS lo que significa que en esta comparación ETS es el que mayor grado de confianza obtuvo.

En el tiempo de procesamiento al evaluar estas técnicas se obtuvo que con el método Holt el tiempo promedio de ejecución es de 2.23 segundos siendo superior a diferencia de HoltWinters, que tiene 7,01 segundos y ETS con 22,33 segundos.

c. Se seleccionó el modelo de evaluación para medir o comparar técnicas mediante los indicadores los cuales fueron la confiabilidad que obtuvo el modelo al predecir las ventas y el tiempo que demoro en generar los pronósticos. En el caso de los resultados al evaluar los algoritmos se obtuvo que ETS tiene mejor confiabilidad, sin embargo, genera un tiempo elevado (22 segundos) para procesar los datos, analizando el ámbito de aplicación del algoritmo donde se trazan valores consolidados, es un tiempo aceptable que no atañe a la elección de este algoritmo. También podemos decir que Holt y HoltWinters pueden estimar a más horizontes aunque no son tan precisos, pero ETS aunque es preciso no es tan bueno si alejas el horizonte.

d. Se desarrolló la aplicación para evaluar los resultados obtenidos. El sistema se diseñó en PHP obteniendo una interfaz capaz de interactuar con el servidor a fin de ejecutar los modelos, ya sean reales o de simulación para que el usuario realice las pruebas pertinentes.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

6.2. Recomendaciones

a. Se recomienda establecer un mecanismo de análisis en función a los requerimientos y performance de la técnica, en esta investigación se analizó los factores que conllevan a una técnica en la incidencia de su nivel de confiabilidad, como por ejemplo el tamaño de vector histórico que se requiere.

b. Se recomienda que para realizar pronósticos utilizando algoritmos de series de tiempo como son los métodos Holt-Winters, Holt y ETS se debe emplear como mínimo 25 meses de datos históricos los cuales ingresan en el vector de series de tiempo, además debe considerarse el objetivo de los pronósticos, en la investigación se demostró que ETS es un buen algoritmo cuando se trata de valores cercanos, sin embargo al alejar el horizonte de pronósticos genera ruido que no permite obtener buenos pronósticos como en un horizonte 3.

c. Otro de los factores a recomendar para futuras investigaciones es que los criterios de análisis del algoritmo como es el caso del tiempo, estén sujetos a la naturaleza del negocio, en este caso aceptar que ETS es un buen algoritmo por que se aplica a consolidados de ventas. Sin embargo para casos en los cuales se necesite realizar procedimiento de minería masiva (como el perfil de carga de un determinado cliente), ETS no sería una buena alternativa por el tiempo de procesamiento que demandaría.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

BIBLIOGRAFÍA

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

BIBLIOGRAFÍA

Aluja. (2001). *La Minería de Datos, entre la estadística y la inteligencia artificial*. http://www.idescat.cat/sort/questiio/questiiopdf/25.3.4.Aluja.pdf

Asencios, V. V. (2004). Datamining y el Descubrimiento del Conocimiento. *Revista de la Facultad de Ingenieria Industrial*, 5.

Barrientos, F. (Setiembre de 2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. *Revista Ingeniería de Sistemas*.

Bunge, M. (2001). *La investigación científica: Su estrategia y su filosofía.* Siglo XXI Editores. Obtenido de http://es.wikipedia.org/wiki/Predicci%C3%B3n

Calvo Rodríguez, A. (2008). *Predicción en series de Tiempo con Modelos Aditivos.* España: Universidade da Coruña.

Coghlan, A. (2015). *A Little Book of R for Times Series.* Cambrigde: Trust Sanger Institute, Cambridge, U.K. .

Consola de Ejecución, R.-P. (s.f.). Consola de ejecución en R - PROJECT.

Cruz Arrela, L. (2010). Minería de Datos con Aplicaciones. En L. Cruz Arrela.

Dandretta, G. H. (2002). *Web mining: implementando técnicas de data minning en un servidor web.* Universidad de Belgrano.

De la Fuente Fernández, S. (2011). *Alisado Series Temporales.* Universidad Autónoma de Madrid: Departamento de Economía Aplicada.

E.Kendall, K., & Kendall, J. E. (2005). *Análisis y diseño de sistemas.* México: PEARSON EDUCACIÓN.

Edison, R. (2010). *Minería de datos.*

Española, R. A. (s.f.). *Diccionario de la Real Academia Española.*

Fernández Maturana, V. P. (2007). Wavelet-and SVM-based forecasts: An analysis of the U.S. metal and materials. *Resources Policy*, 1-2.

García Bermúdez, J. A., & Acevedo Ramírez, Á. M. (2010). Análisis para Predicción de ventas utilizando minería de datos en Almacenes de grandes superficie. Pereira: Universidad Tecnologica de Pereira.

Getoor, L., & Ben, T. (2007). *Introducción a estadística de relación de aprendizaje.* MIT.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Grudnitsky, B. J. (1992). *Diseño de sistemas de información. Teoría y Práctica.* México: Megabyte Grupo Noriega.

http://www.r-project.org/. (s.f.). Obtenido de http://www.r-project.org/: http://www.r-project.org/

Hyndman, R. J. (2015). *Github*. Obtenido de Github: https://github.com/robjhyndman/forecast

Hyndman, R., & Athanasopoulos, G. (2015). *Texts Online, Open - Access Textbooks*. Obtenido de Texts Online, Open - Access Textbooks: https://www.otexts.org/

IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler.* EEUU: IBM.

Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit.* Wiley India.

Lezcano, R. D. (2010). *Minería de datos.*

Ma, N. (2013). Neural network algorithm based method for stock price trend prediction. Beijing, China: Asian Network for Scientific Information.

Madrigal Espinoza, S. D. (2006). *MODELOS DE ESPACIO DE ESTADOS SUBYACENTES AL MÉTODO MULTIPLICATIVO DE HOLT-WINTERS CON MÚLTIPLE ESTACIONALIDAD.* San Nicolás de los Garza, Nuevo León - México.

Martínez Álvarez, C. (Octubre de 2012). Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en ENTEL. Santiago, Santiago, Chile.

Microsoft. (2013). *MSDN Library.*

Molina López, J. M., & García Herrero, J. (2006). Técnicas de Minería de Datos basadas en Aprendizaje Automático. *Técnicas de Análisis de datos*. Universidad Carlos III de Madrid.

Molina Neyra, C. A., & Murakami de la Cruz, S. E. (2008). Implementación de una solución informática basado en M-Commerce aplicado a sistemas de distribución comercial. Lima, Lima, Lima.

Nojek, S., Britos, P., Rossi, B., & García, M. R. (2008). Prónostico de Ventas: Comparación de Predicción basada en Redes Neuronales versus Método Estadístico. *Reportes Técnicos en Ingenieria del Software*, 1-12.

Ortíz Farro, P. (03 de 2015). MINERÍA DE DATOS CON SERIES DE TIEMPO EN EL DESARROLLO E IMPLEMENTACIÓN DEL SISTEMA INTELIGENTE QUE PREDICE LA PRODUCCIÓN DE ARROZ EN EL ÁMBITO DE LA GERENCIA REGIONAL DE AGRICULTURA – LAMBAYEQUE. Universidad Señor de Sipán.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Perversi, I. (2007). Aplicación de Minería de Datos para la Explotación y Detección de Patrones Delictivos en Argentina. XIII Congreso Argentino de Ciencias de la Computación.

Ramírez A., A. Y. (2007). Técnicas de Minería de Datos Aplicadas a la Construcción de Modelos de Score Crediticio. *Mathematical Problems in Engineering*.

Rowley. (1994). *The Basics of Systems Análisis and Design for Information Managers.* Londres: Clive Bingley.

Schaefer, J. M. (2012). El deporte, los artículos deportivos y la industria del deporte. *OMPI - Organización mundial de la propiedad intelectual*.

Senn, J. A. (1992). *Análisis y diseño de sistemas de información.* México: McGraw-Hill.

Servente, M. (2002). Algoritmos TDIDT Aplicados a la Minería Inteligente. Argentina: Facultad de Ingeniería - Universidad de Buenos Aires.

Siccha Vega, H. W. (2012). Minería de Datos aplicados a las ventas con Tarjeta de Crédito realizados en las tiendas Saga Falabella. Lima, Lima, Perú: Universidad Tecnológica del Perú.

Std, I. (1993). *IEEE Software Engineering Standard: Glossary of Software Engineering Terminology.* IEEE Computer Society Press.

The CRISP-DM, c. (2013). *CRISP-DM 1.0.* Step Data Mining guide.

Valcárcel Asencios, V. (2004). Data Mining y el descubrimiento del conocimiento. *Industrial Data*, 83-86.

Vallejo P., D., & Tenelanda V., G. (2012). Minería de datos aplicada en detección de intrusos. Medellín: USBMed Vol. 3, N° 1.

Vallejo Pérez, D. &. (2012). *Minería de datos aplicada en detección de intrusos.* Medellin: Ubicación en Biblioteca USB Medellín (San Benito): CD-2031t.

Vallejos, R. (2012). *Introducción a las Series Cronológicas.* Universidad Técnica Federico Santa María .

Vidaurre Siadén, Y. (Abril de 2012). Aplicación de redes neuronales artificiales para el pronóstico de la demanda de agua potable en la empresa EPSEL S.A de la ciudad de Lambayeque. Chiclayo, Chiclayo, Lambayeque.

Wang, X. (2013). Predicción del estado de la red de abastecimiento basada en un modelo óptimo de combinación de minería de datos. California, EEUU: Journal of Applied Sciences.

Zadeh, N. K. (2014). Una investigación basada en minería de datos: Predicción de ventas para compañías farmacéuticas de distribución. *Hindawi*, 15.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ANEXOS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe