

 | UNIVERSIDAD
SEÑOR DE SIPÁN

**FACULTAD DE INGENIERÍA, ARQUITECTURA Y
URBANISMO**

**ESCUELA PROFESIONAL DE INGENIERÍA DE
SISTEMAS**

TESIS

**ANÁLISIS COMPARATIVO DE LAS TÉCNICAS
DE MINERÍA DE DATOS PARA LA ESTIMACIÓN
DE CONSUMOS DE ENERGÍA ELÉCTRICA EN LA
EMPRESA ELECTRONORTE S.A.**

**PARA OPTAR EL TÍTULO PROFESIONAL DE
INGENIERO DE SISTEMAS**

Autor:

Bach. Jhong Guillen Shirley Julissa

Asesor:

Mg. Requejo Chaname Walter Juan

Línea de Investigación:

Infraestructura, tecnología y medio ambiente

Pimentel, Perú

2019



DEDICATORIA

A mi esposo y mis hijas, que gracias a su apoyo e incentivo estoy logrando mis metas.

A mis padres, que son un digno ejemplo a seguir y por la fortaleza que los caracteriza.

A mi hermana, por todos sus valores y dedicación en hacer lo que nos apasiona.

Y a todas las personas que compartieron sus conocimientos conmigo.

AGRADECIMIENTO

En primera instancia, agradezco a Dios por su infinito amor y permitir culminar este proyecto.

A mi familia, quienes son la fuerza que necesito día a día para seguir adelante.

Al personal docente, por su gran sabiduría quienes se han esforzado por ayudarme a lograr una de mis metas.

ÍNDICE

DEDICATORIA.....	2
AGRADECIMIENTO	3
RESUMEN.....	7
ABSTRACT	8
INTRODUCCIÓN.....	9
CAPITULO I: PROBLEMA DE INVESTIGACION.....	10
1.1. Situación problemática	10
1.2. Formulación del Problema.....	12
1.3. Delimitación de la investigación.....	12
1.4. Justificación e importancia de la investigación.....	12
1.5. Limitaciones de la investigación	13
1.6. Objetivos de la investigación	14
CAPITULO II: MARCO TEORICO.....	15
2.1. Antecedentes de la investigación	15
2.2. Estado del arte	18
2.3. Bases teórico-científicas	21
2.4. Definición de términos básicos	34
CAPITULO III: MARCO METODOLOGICO	38
3.1. Tipo y diseño de la investigación.....	38
3.2. Población y muestra.....	39
3.3. Hipótesis	48
3.4. Variables	48
3.5. Operacionalización de variables.....	49
3.6. Abordaje Metodológico, técnicas e instrumentos de recolección de datos	50
3.7. Procedimiento para la recolección de datos.....	51
3.8. Análisis estadístico e interpretación de los datos	51
3.9. Principios éticos	52
3.10. Criterios de rigor científico.....	53
CAPITULO IV: ANALISIS E INTERPRETACION DE LOS RESULTADOS	54
4.1. Resultados en tablas y gráficos	54
4.2. Discusión de resultados	67
CAPITULO V: DESARROLLO DE LA PROPUESTA.....	68
5.1. Generalidades de la propuesta.....	68
5.1.1. Marco General de Trabajo.....	68
5.1.2. Desarrollo de la propuesta	70
CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES.....	110
6.1. Conclusiones	110
6.2. Recomendaciones.....	112
Referencias Bibliográficas	113



ÍNDICE DE FIGURAS

Figura 1. Clusterización de Datos.....	22
Figura 2. Redes Neuronales.....	24
Figura 3. Metodología XP	30
Figura 4. Metodología SCRUM.....	31
Figura 5. Metodología CRISP-DM.	31
Figura 6. Metodología SEMMA.....	32
Figura 7. Método de desarrollo de la investigación	69
Figura 8. Flujoograma actual del proceso de facturación	72
Figura 9. Modelo propuesto de trabajo	73
Figura 10. Extracción de datos a nivel ficheros CSV y RDS.....	75
Figura 11. Tabla HechoConsumo.....	76
Figura 12. Tabla Dimensión PeriodoFacturacion	76
Figura 13. Tabla Dimensión Administrativo	77
Figura 14. Tabla Dimensión Empresa	77
Figura 15. Tabla Dimensión PeriodoFacturacion	78
Figura 16. Tabla Dimensión PeriodoFacturacion	78
Figura 17. Tabla hecho.....	79
Figura 18. Formato de análisis series de tiempo.....	80
Figura 19. Tiempo obtenido de procesamiento de una tabla Hecho	81
Figura 20. Matriz de suministros por periodo y consumo	81
Figura 21. Tratamiento de nulos (Directo)	82
Figura 22. Tratamiento de Nulos (Indirecto).....	83
Figura 23. Agregación del valor NULL	84
Figura 24. Extracción de los datos.....	87
Figura 25. Algoritmo en R – Holtwinters Biblioteca Forecast R.....	89
Figura 26. Código del modelo con Holtwinters.....	90
Figura 27. Algoritmo R – Nnetar	91
Figura 28. Código Algoritmo R - Nnetar.....	93
Figura 29. Formulas algoritmo ARIMA	94
Figura 30. Código Fuente Algoritmo ARIMA en R.....	95
Figura 31. Código Algoritmo ARIMA	96
Figura 32. Algoritmo SVM en R.....	97
Figura 33. Conceptualización de SVM.....	98
Figura 34. Implementación de SVM en R.....	99
Figura 35. Script Web Laboratorio con R y Shiny	104
Figura 36. Pantalla Laboratorio 1 – Realizar pronósticos en APP.....	105
Figura 37. Pantalla Laboratorio 2 – Pantalla para extraer Muestreo	106
Figura 38. Pantalla Laboratorio 3 – Resultados y calculo MAPE.....	107



ÍNDICE DE TABLAS

Tabla 1	40
Tabla 2	41
Tabla 3	43
Tabla 4	49
Tabla 5	50
Tabla 6	52
Tabla 7	53
Tabla 8	55
Tabla 9	60
Tabla 10	61
Tabla 11	66
Tabla 12	79
Tabla 13	85
Tabla 14	86
Tabla 15	90
Tabla 16	92
Tabla 17	93
Tabla 18	96
Tabla 19	99
Tabla 20	100
Tabla 21	101
Tabla 22	102
Tabla 23	102
Tabla 24	103



RESUMEN

La investigación denominada “Análisis comparativo de las técnicas de minería de datos para la estimación de consumos de energía eléctrica en la empresa ELECTRONORTE S.A.” tiene como objetivo analizar los diversos algoritmos de aprendizaje de minería de datos para el diseño de modelos con tendencia predictiva en los sistemas de Inteligencia de Negocios.

La minería de datos es una herramienta que permita a través de la exploración y análisis de datos extraer patrones de comportamiento en el histórico de datos de determinado fenómeno (Institución, Empresa, etc.), según la naturaleza del fenómeno y los datos se aplican diversas técnicas, siendo relevante mencionar a técnicas como Regresión, Clasificación, Asociación y Agrupación; esta exploración se realiza mediante algoritmos computacionales de aprendizaje.

El ámbito de aplicación de éstas técnicas, en esta investigación se localiza en los datos de los clientes de la empresa ELECTRONORTE S.A., empresa del rubro eléctrico de servicios públicos, siendo el objetivo de la misma analizar los datos con respecto al comportamiento de consumo eléctrico de los clientes para determinados periodos comerciales.

Por lo tanto, el tipo de modelo se asocia con las técnicas de Regresión o series de tiempo, ya que se pretende realizar estimaciones de los consumos de energía eléctrica de los clientes. Evidentemente es un caso de pronósticos de series de tiempo, para lo cual se conocen diversos algoritmos como es el caso de Holtwinters, ARIMA, ETS, Redes Neuronales, entre otros, con los cuales se puede realizar el pronóstico según el análisis histórico de la serie de tiempo. Se sabe también que cada algoritmo puede ser aplicado a una realidad específica, siendo el ámbito energético al cual se someterán a evaluación los mencionados algoritmos para establecer cual tiene mejor precisión en el desarrollo de modelos predictivos orientado a este problema.

Palabras claves: Minería de datos, Pronósticos.



ABSTRACT

The present investigation named "Comparative Analysis Techniques Data Mining for the estimation of consumption Electric Power Company ElectroNorte SA" has the objective to analyze the various learning algorithms used in mining techniques Data design predictive Models in Business Intelligence Systems.

Data mining is a tool that allows for a through exploration and Data Analysis extract patterns of behavior in the historic Data certain phenomenon (Institution, Company, etc.), according to the nature of the phenomenon and apply data Various techniques being relevant technicians as mentioned a regression, classification, grouping Association and Exploration. This scan is performed using computational learning algorithms.

The scope of the Technical Data Mining in this research is located in the customer data Business ElectroNorte SA, Company electrical category of Utilities, being the m objective of the Same · analyze the data regarding the power consumption behavior of customers for certain trading periods.

It's like that, the model type is associated with regression techniques or series of Time, and intends to perform estimates of consumption Electric Customers Energy. Obviously it's a case Forecasting time series, for which is known are various algorithms as in the case of Holtwinters, ARIMA, ETS, Neural Networks, among others, with whom the prognosis may be performed according to historical analysis time series. This is also known that each algorithm can be applied to a specific reality, being the energy field to which an evaluation will undergo the above algorithms to establish which has better accuracy Oriented Development of predictive models this problem.

Keywords: Data Mining, Forecast.



INTRODUCCIÓN

El crecimiento de las TIC's, con soporte de automatización a los procesos operativos de las organizaciones en las últimas décadas ha sido exponencial, existiendo, hoy en día, una variedad de sistemas informáticos especializados en la gestión y control de los datos en los procesos transaccionales en diversas plataformas, siendo la web y móvil las que recientemente lideran este ámbito, o grandes sistemas denominados ERP, CRM y otras variedades de soluciones han tenido como objetivo la transacción.

Luego, en un segundo nivel se encuentran los sistemas de inteligencia de negocios, aquellos que gestionan con eficiencia y eficacia reportes y consolidan la información que brinda soporte a la toma de decisiones, tanto tácticas como estratégicas; sin embargo, estos sistemas solo se preocupaban por obtener la información y almacenarla, haciendo un mínimo de análisis, sin tener en cuenta, si dicha información puede ser de utilidad o no.

A raíz de esto, es que nace la minería de datos, que hoy evoluciona, junto con la ciencia de datos, a ser una herramienta que pretende mediante una exploración científica sobre los datos encontrar patrones de comportamiento no perceptibles en los típicos reportes de sistemas hasta hoy conocidos.

La minería de datos emplea técnicas que al realizar un análisis masivo sobre el histórico de información logra extraer reglas de comportamiento que pueden ser usados para anticipar situaciones de un determinado fenómeno.

Dentro del entorno de las empresas de energía eléctrica, se requiere utilizar estas técnicas para realizar un análisis especializado sobre los datos históricos de consumo de energía eléctrica de los clientes.

Determinar que algoritmo de aprendizaje para la detección de patrones de comportamiento que emita una mejor precisión sobre las estimaciones de consumo de energía eléctrica es propósito principal de esta investigación.



CAPITULO I: PROBLEMA DE INVESTIGACION

1.1. Situación problemática

A nivel internacional

Las técnicas estadísticas para realizar el cálculo de pronósticos sobre series de tiempo son conocidas, desde métodos estocásticos como el ARIMA, hasta los que se usan para finanzas como Holtwinters, así también, si se emplea el método de regresión, se puede determinar la ecuación que permite saber el cálculo del monto siguiente en una serie.

A lo largo de los años y con el avance de la informática estas técnicas estadísticas se automatizaron en algoritmos computacionales de procesamiento de datos, el mayor uso que se hace en los últimos años es agrupar todas estas técnicas dentro de una corriente denominada como Minería de Datos, que es una herramienta matemática computacional que explota los datos a fin de reconocer patrones de comportamiento que sirven para explicar un fenómeno y realizar una predicción en base a estos datos encontrados.

Cuando se trata de pronósticos, en la minería de datos, se asocia rápidamente a los modelos de regresión, también a la aplicación de modelos de series de tiempo. Hoy en día con el avance de la inteligencia artificial se han complementado nuevas herramientas a las técnicas, por ejemplo, se combina la regresión logística con el uso de las redes neuronales. Sin embargo, aquí radica el problema. (Vega, 2012).

En el tema del análisis de las técnicas para la estimación de consumos, se aplican algoritmos para obtener el pronóstico de consumos, de todos aquellos suministros (viviendas) que no hayan registrado ninguna lectura.

El consumo estimado es calculado por el análisis inteligente que realiza el modelo detectando patrones en la evolución histórica de los consumos, donde el



modelo aprende que un suministro puede incrementar o decaer su consumo en ciertos meses del año (festividades, cambios de estación, entre otros), mientras que el modelo actual calcula el consumo en función al promedio de los últimos seis meses anteriores.

A nivel Nacional

Sandoval (2014) en su proyecto de investigación plantea como objetivo principal el desarrollo de un Aplicativo que permita realizar pronósticos sobre el inventario de repuestos, haciendo uso de modelos estadísticos. El proyecto se centra en determinados procesos relacionados con la gestión del pronóstico de importación, gestión del pedido de importación y gestión de canibalización de motos por garantías. Asimismo, se identifica como el problema principal, las demoras existentes en el proceso de importación desde la India, aproximadamente 4 meses, ocasionando rupturas en el stock de algunos productos y bajas en los ingresos percibidos en los últimos años. Por lo que, la propuesta se enfoca en lograr mayor precisión en los resultados, minimizar las rupturas de stock y por consecuencia ofrecerles a los clientes un servicio de máxima calidad.

A nivel Regional

De acuerdo con Díaz (2013), en su investigación presenta un esfuerzo por desarrollar una herramienta de software que, haciendo un adecuado pre procesamiento de datos y técnicas sobre series de tiempo logre obtener un pronóstico cercano al valor real con el que se puedan realizar estimaciones mejorando el cálculo de consumos de lecturas.

A nivel Institucional

En la actualidad ENSA posee un total de 225374 suministros (clientes a quienes se les comercializa el servicio de energía eléctrica); de los cuales 215374 poseen sistema de medición activo (clientes que cuentan con medidor instalado),



los mismos que son lecturados de acuerdo al cronograma de facturación, que incluye la fecha de toma de lectura aprobado por la Gerencia Comercial.

La problemática radica en que la cantidad de consumos inconsistentes va en aumento, ocasionando que los supervisores de facturación se vean obligados a estimar el mismo tomando como referencia un consumo promedio estimado, originando, a su vez que, gran parte de los suministros con estos consumos estimados presenten reclamos; lo que conlleva a realizar “re facturaciones” o modificaciones sobre los recibos ya emitidos, afectando los índices de insatisfacción al cliente.

Como determinar que algoritmo de análisis de series de tiempo o regresión servirá o brindará el mejor resultado al problema de los consumos, representa el objetivo principal de la presente tesis.

1.2. Formulación del Problema

¿Qué algoritmo de minería de datos será el óptimo, para realizar estimaciones de consumo de energía eléctrica?

1.3. Delimitación de la investigación

Se efectúa un análisis comparando las diferentes técnicas de minería existentes; además se plantea la construcción de un portal web para la visualización de los datos, simulaciones y análisis.

1.4. Justificación e importancia de la investigación

A nivel Tecnológico

Se analizan los diversos algoritmos y fundamentos matemáticos del tratamiento de series de tiempo, así como las técnicas empleadas para la generación de conocimiento basadas en el proceso de minería de datos

A nivel Social

Permite analizar la conducta o patrones por estratos de los clientes en el consumo de energía eléctrica, asumiendo variables como dispersión por situación geográfica, tipo de cliente, etc.

A nivel Económico

Permite optimizar el proceso de facturación minimizando costos por procesos operativos y penalizaciones por errores de facturación.

A nivel Científico

Permite tener un fundamento matemático estadístico para las decisiones que se toman a partir de la falta de datos, donde se establece un criterio básico de promedios que no ha dado buenos resultados hasta la fecha.

1.5. Limitaciones de la investigación

Se tiene como dominio de exploración geográfica toda la ciudad de Chiclayo. El alcance para el desarrollo web comprende este espacio, y se dará en el periodo mayo 2016–setiembre 2019 para mostrar la implementación del prototipo y posterior análisis de los resultados.

La solución comprende un analizador de estimaciones para determinado periodo comercial, así como un simulador para comprender los diversos algoritmos analizados.



1.6. Objetivos de la investigación

1.6.1. Objetivo General

Realizar el análisis comparativo de las técnicas de minería de datos para la estimación de consumos de energía eléctrica en la empresa ElectroNorte S.A.

1.6.2. Objetivos Específicos

1. Analizar los datos para estimación de consumo de energía eléctrica.
2. Seleccionar las técnicas adecuadas que se adapten a la realidad de los datos recopilados.
3. Evaluar e identificar el algoritmo óptimo para la estimación de consumo.
4. Diseñar un aplicativo que realice el análisis de estimación de consumos con las técnicas utilizadas.



CAPITULO II: MARCO TEORICO

2.1. Antecedentes de la investigación

2.1.1. A nivel internacional

- A. Fernández (2007) en su investigación toma como base las pruebas realizadas haciendo uso de 4 modelos de series temporales: ARIMA multiplicativo estacional, UC, wavelets y SVM. Con los resultados obtenidos, el autor afirma, que ARIMA y UC, se caracterizan por mostrar datos más exactos, clasificándolas como métodos tradicionales; sin embargo, wavelets y SVM brindan información adicional a la dada por ARIMA y UC. Por lo que señala que, las combinaciones lineales pueden dar mejores resultados que los pronósticos individuales. Con esto, se llegan a las siguientes conclusiones: El horizonte de tiempo seleccionado es muy importante al momento de decidir qué modelo se deberá usar o cual combinación lineal se tomará para lograr la máxima calidad del pronóstico. Además, es necesario resaltar la utilidad de la información brindada por SVM al combinarse con wavelets. No obstante, SVM puede llegar a superar a los algoritmos UC y ARIMA.
- B. Calvo (2008) en su trabajo de investigación emplea los modelos aditivos de regresión múltiple y el algoritmo backfitting para plantear un algoritmo de predicción en series de tiempo y búsqueda de intervalos de confianza. Para esto recurrió a métodos estadísticos no paramétricos que permitieran el trabajo de funciones unidimensionales. Los resultados obtenidos arrojaron un margen de error del 2.77% en el modelo, concluyendo en su correcto funcionamiento y en su capacidad para solucionar problemas de dimensionalidad.
- C. Madrigal (2006) plantea en su investigación el desarrollo de un método de pronóstico, haciendo uso de la técnica de suavización exponencial con el método de Holt-Winters, incorporando intervalos de predicción.



Los criterios seleccionados arrojaron márgenes de error muy bajos, con un 86% de confiabilidad; sin embargo la situación cambia al ir aumentando la cantidad de periodos a estudiar.

2.1.2. A nivel nacional

- D.** Ramírez (2007) en su investigación compara diferentes técnicas, como redes neuronales, análisis discriminante, máquinas vectoriales de soporte, árboles de decisiones y regresión logística; con el propósito de identificar patrones de comportamiento similares en un cliente con score crediticio. Los resultados arrojaron que la regresión lineal, como técnica tradicional, muestra una predicción más exacta. No obstante, las redes neuronales, tienen mayor exactitud en cada ejecución realizada. En los casos de buenos créditos, el modelo que mejor predice es una red neuronal probabilística; mientras que, para los casos de créditos malos, es mejor una red neuronal multicapa.

Del mismo modo, de acuerdo con la investigación, para construir un buen modelo, se debe prestar especial atención al proceso de selección de los datos, buscando siempre la máxima calidad y sobre todo, tener siempre en cuenta la adecuada elección de las variables, cuya influencia es determinante para el éxito del modelo

2.1.3. A nivel regional

- E.** Díaz (2013) en su tesis plantea parte del problema de facturación del consumo eléctrico por debajo del real o en exceso, generado por la toma de lectura en campo errada por parte de la empresa PEXPORT S.A.C. y el parámetro de consumo genérico manejado por la concesionaria; ante esto, el propósito es la implementación de un sistema web, que haciendo uso de técnicas de minería de datos, contribuya a mejorar la gestión del proceso de facturación de la empresa PEXPORT S.A.C. reflejándose en una mejora de sus resultados económicos y operativos,



por ende, esto repercute, no solo en mejorar la facturación del servicio de energía eléctrica a los clientes finales de ElectroNorte S.A., , si no también, en generar mayores beneficios para los trabajadores de la empresa, automatizando procesos muy importantes que garanticen un buen servicio.

En relación a la investigación, este trabajo brinda información útil en cuanto a la aplicación de la metodología CRISP-DM, mostrando el proceso de desarrollo realizado.

2.2. Estado del arte

La minería de datos no es un término nuevo. Desde tiempos atrás, diversos estadísticos han venido manejando y perfeccionando el uso de éstas técnicas, pasando por denominaciones como Data Fishing, Data Mining o Data Archaeology; con la finalidad de encontrar relaciones y patrones ocultos en las bases de datos e información, sin el planteamiento de una hipótesis previa.

Fue en los ochenta, algunos estudiosos como *Rakesh Agrawal*, *Gio Wiederhold* y otros, adecuaron el término a *Minería de Datos* y *KDD*.

Este proceso de crecimiento y las herramientas desarrolladas en este tiempo pueden dividirse en 4 etapas:

- Colección de los Datos (1960).
- Acceso a los Datos (1980).
- Almacén de Datos y Apoyo a las Decisiones (inicios de los 1990).
- Minería de Datos Inteligente. (finales de la década de 1990).

De acuerdo con Daylan (2002) en el “International Review on Computers and Software” la minería de datos es una muestra del progreso exponencial de las tecnologías de la información, donde los datos son tratados con un gran nivel de detalle, extraídos de diversas fuentes externas como bases de datos y almacenes de datos.

Para el 2003, Guil, Bosch y Marin (2003) en su proyecto, diseñó modelos de minería de datos basado en restricciones temporales para su posterior integración en el sistema de soporte a decisiones.

Ésta técnica dispone de un gran número de operadores (procesos):

- Árbol de decisión
- Reglas de asociación
- K- Medias

- SVM
- Redes Neuronales
- Series de Tiempo

Según Merian (2011) en el Big Data. Computer World, en el 2010, implementar minería de datos en las empresas es un reto, que los lleva a sacar el máximo provecho en cuanto a infraestructura de almacenamiento y procesamiento se refiere; además del proceso de captación de profesionales capacitados que puedan realizar el trabajo de adaptación e innovación específica de acuerdo a cada caso en particular. Según los estudios realizados, en el año 2018, solo en los Estados Unidos de América serán necesarios un aproximado de 140 000 a 190 000 expertos con estos conocimientos.

En 2012, como lo muestra Vega (2012) en su estudio, la alternativa confiable que permite descubrir conocimiento a partir de la exploración de bases de datos, es el uso de técnicas de minería de datos, que integradas, posibilitan la obtención de patrones globales de análisis.

En la actualidad, de acuerdo con Ramírez (2015), la minería de datos se puede utilizar para los siguientes casos:

- Filtración de información
- Elección de variables
- Algoritmos de extracción de conocimiento
- Interpretación y evaluación
- Generación de recomendaciones
- Detección de anomalías
- Administración de riesgos
- Segmentación de clientes
- Publicidad dirigida

Del mismo modo, se divide en cuatro etapas principales:



- Determinar los objetivos
- Pre procesar los datos
- Determinar el modelo
- Analizar los resultados

En la actualidad, de acuerdo con (Montalvo, 2016) en su investigación sobre comparación de algoritmos para predicción de ventas, *“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.”*, esta premisa es enunciada, dado que en la actualidad existen bibliotecas con los algoritmos construidos, pero que deben pasar una fase de modelado de datos.

2.3. Bases teórico-científicas

2.3.1. Minería de Datos

A. Concepto

Proceso que acepta como entrada datos y entrega como salida información valiosa y precisa. Los diversos algoritmos que existen permiten descubrir patrones de comportamiento similares entre los datos analizados para luego plantear solución a determinadas situaciones.

La minería de datos forma parte de la familia de Business Intelligence (BI), integrada también, por el procesamiento analítico en línea, los informes empresariales y ETL.

En los últimos años, la cantidad de datos manejados en las empresas se han duplicado, recurriendo a su almacenamiento en grandes bases de datos y almacenes de datos. Como resultado, las empresas se han convertido en fuentes ricas de información, pero, pobres en conocimiento. Es en este punto donde entra a relucir la minería de datos, posibilitando la extracción de patrones ocultos de entre los datos y su transferencia al conocimiento.

B. Métodos de Minería de Datos

Álvarez (2012) plantea que la minería de datos tiene dos enfoques principales, la predicción y la descripción. En ambos casos se dispone de una gran variedad de métodos y técnicas para el descubrimiento del conocimiento. Entre los predictivos, destacan la clasificación y la regresión, mientras que por los descriptivos, están el *clustering* y las reglas de asociación.

Clasificación: Consiste en la identificación de ciertos atributos que vinculan a un elemento a un conjunto o grupo de acuerdo a



determinados patrones de datos; para predecir su comportamiento en nuevas situaciones.

Regresión: es una función que permite asignarle un valor real a un elemento dado. Se usa para la predicción de valores futuros en casos específicos, como el comportamiento de una demanda futura, empleando las ventas o el consumo como fuente de datos

Clustering: Divide al conjunto de datos en pequeños grupos con características similares entre los elementos del mismo grupo, pero diferentes entre ellos. Permite la identificación de un grupo de características o clústeres para la descripción de los datos.

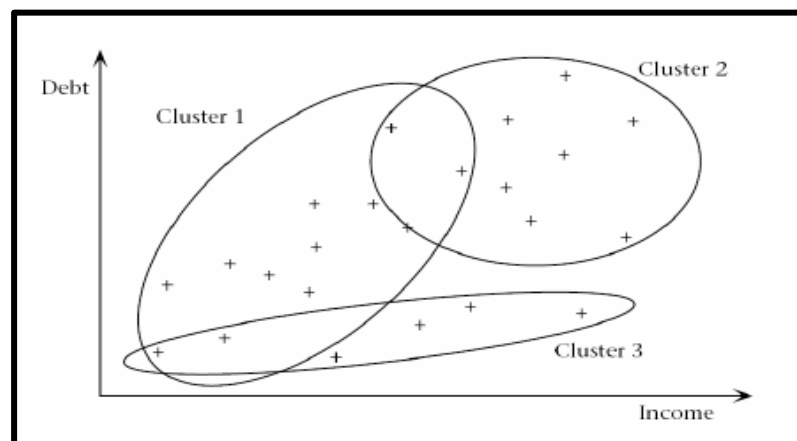


Figura 1. Clusterización de Datos.

Reglas de asociación: instrumento descriptivo, cuya finalidad es encontrar relaciones entre los elementos estudiados. Un ejemplo muy usado es el caso de la canasta de compras en una tienda, con referencia al análisis de los productos que forman parte de ésta.

Series de Tiempo: Son los valores que se estudian tomando en cuenta los periodos secuenciales y ordenados temporalmente en los cuales se desarrolla. Está representada por una curva que crece a los largo del tiempo. Un clásico ejemplo, son las ventas diarias de un producto, las cuales pueden crecer o decrecer en el tiempo.



Trabajar con series de tiempo significa realizar un pronóstico en el futuro de los valores no conocidos en el tiempo tomando como base los valores históricos; con el objetivo de optimizar algún proceso dentro de una determinada área, como analizar los inventarios, la producción o el personal.

Se debe tener en cuenta el uso de dos variables estructurales:

- El **período**, representado por meses, semanas y días.
- El **horizonte**, representa el número de períodos que serán pronosticados.

Métodos de suavizamiento y pronóstico para series de tiempo

Métodos de promedios móviles

Este método supone la sucesión de valores con respecto a medias aritméticas.

Empleando de forma correcta los movimientos medios, se eliminan las variaciones estacionales o cíclicas, quedando el movimiento de tendencia.

$$\frac{Y_1 + Y_2 + \dots + Y_N}{N}, \frac{Y_2 + Y_3 + \dots + Y_{N+1}}{N}, \frac{Y_3 + Y_4 + \dots + Y_{N+2}}{N}, \dots$$

Suavización exponencial

Este método es un caso especial de promedios móviles ponderados de los valores anteriores y actuales, donde las ponderaciones van disminuyendo exponencialmente. Se usa para suavizar y para realizar pronósticos. Su fórmula es:



$$Y_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot Y_t$$

Donde:

Y_{t+1} = pronóstico para cualquier período futuro.

α = constante de suavización, a la cual se le da un valor entre 0 y 1.

X_t = valor real para el período de tiempo.

Y_t = pronóstico hecho previamente para el período de tiempo

Otros métodos de minería de datos

Algoritmo de redes neuronales

Se usan principalmente para casos de reconocimiento de patrones y sistemas de clasificación.

Representan neuronas artificiales, es decir, simulan la actuación de las neuronas del cerebro humano. Éstas son entrenadas para generalizar patrones de predicción y clasificación. Cada neurona procesa los datos de forma independiente y eleva los resultados a la siguiente capa de la red.

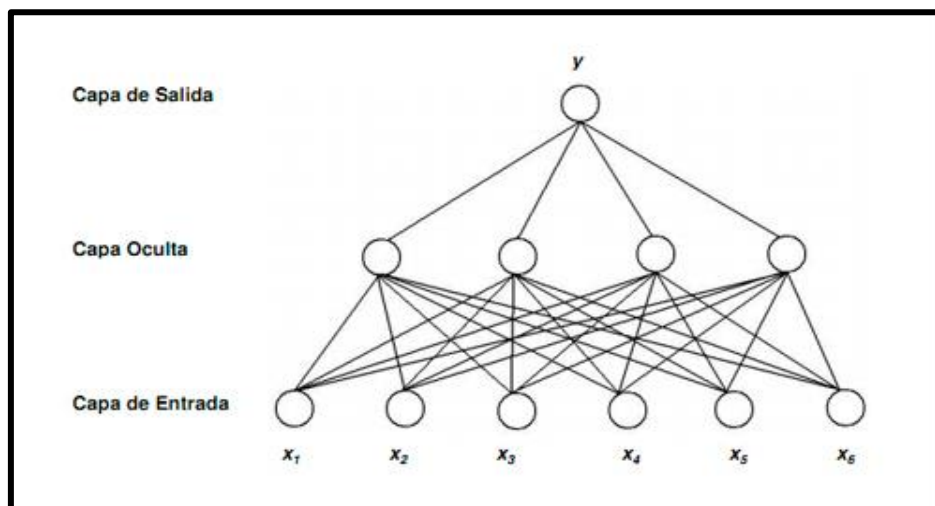


Figura 2. Redes Neuronales.



Algoritmo SVR o SVM

Las máquinas de vectores de soporte son algoritmos de aprendizaje supervisado.

Se adecúan a los problemas de clasificación y regresión. El modelo SVM muestra los puntos en estudio en el espacio, separándolos lo más posible. En función a la proximidad de cada uno, se pueden clasificar a una u otra clase.

Algoritmo ARIMA

Este modelo utiliza la metodología Box-Jenkins (B-J) para determinar características en una serie, tales como son la tendencia y la estacionalidad, proporcionando nuevos valores a la serie a través del uso de los valores históricos relevantes de entre todos los demás datos.

2.3.2. Técnicas para validar los modelos de minería de datos

En el mundo práctico y real, las predicciones perfectas y exactas son imposibles, por lo que la persona encargada de tomar las decisiones deben saber prever y lidiar con un grado alto de error.

Para medir el grado de error de un pronóstico es preciso comparar los resultados de por lo menos dos técnicas; además del nivel de confiabilidad de cada una; para determinar la más óptima.

Las técnicas más empleadas son:

A. Error del pronóstico: diferencia entre el valor real y el pronosticado en el período correspondiente. Su fórmula es:

$$E_t = F_t - Y_t$$

$$EA_t(\%) = \frac{|F_t - Y_t|}{F_t} * 100$$



E_t : Error

$EA_t(\%)$: Error Absoluto Porcentual

F_t : Valor real de la serie

Y_t : Valor pronosticado de la serie

B. Precisión del pronóstico (Forecast accuracy): medida que permite visualizar la cercanía existente entre el pronóstico y el valor real de la serie:

$$\text{Precisión}(\%) = 1 - EA(\%)$$

Notas:

- Valor real = pronóstico, precisión del 100%.
- Error > al 100%, precisión del 0%.
- Los límites de precisión están entre 0 y 1.
- Fórmulas de medición del error en el pronóstico (las más utilizadas).
- Fórmulas de selección

Fórmulas de interpretación

A. Porcentaje de error medio absoluto (MAPE): media de los errores porcentuales en valor absoluto, no considera el signo del error sólo la magnitud.

$$MAPE = \frac{\sum_{t=1}^n \frac{|F_t - Y_t|}{F_t}}{n}$$

- MAPE = Porcentaje de error medio absoluto
- Sumatoria de valores absolutos (F_t son las observaciones actuales de las series de tiempo – Y_t son las series de tiempo



estimadas o pronosticadas) entre F_t son las observaciones actuales de las series de tiempo

c. n = Numero de observaciones

Esta técnica no debe ser usada cuando se trabaja con poca cantidad de datos. Ya que el valor real está como denominador en la ecuación de la formula, el MAPE resulta indefinido cuando el valor real de la serie es cero. Del mismo modo, cuando valor real no es cero, pero representa una cantidad pequeña, el MAPE toma valores extremos.

2.3.3. Herramientas de desarrollo

A. Herramientas de Desarrollo de Minería de datos

a. Rapid Miner

Es un software desarrollado por Rapid- I, empresa basada en Dortmund, Alemania. Es el líder de código abierto del sistema de minería de datos y gestión del descubrimiento de conocimiento. Permite analizar los datos de forma sencilla y cuenta con un motor que puede ser integrado en otros productos.

Está desarrollado en java, dispone de una gran variedad de operadores que mediante su encadenamiento, permiten desarrollar los procesos a través de un entorno gráfico y amigable al usuario.

b. R Project

Es un entorno y un lenguaje para el cálculo estadístico y generación de gráficos. Provee una gran variedad de técnicas estadísticas (modelos lineales, test estadísticos, series de tiempo, algoritmos de clasificación y agrupamiento, etc.) y graficas; además de ofrecer un lenguaje de programación completo con el que se puede añadir nuevas técnicas mediante la definición de funciones. R se integra a diferentes gestores de bases de datos y



posee un gran número de bibliotecas y paquetes que hacen su uso más sencillo, adaptándose a diferentes lenguajes de programación.

B. Herramientas de desarrollo Web

a. HTML5

Elemento clave en la construcción de sitios web y aplicaciones. Es una mejora de la combinación entre JavaScript, HTML y CSS. HTML5 brinda estándares para el desarrollo de cada componente de una web, flexibilidad en el uso de elementos y estructuras; además de involucrar una serie de importantes tecnologías. La función de HTML es proveer elementos estructurales; CSS se concentra en convertir esta estructura en amigable y atractiva a la vista del usuario; mientras que, JavaScript le aporta el dinamismo y la funcionalidad al sitio web.

b. JavaScript

Lenguaje de desarrollo de aplicaciones web cliente/ servidor. Lo construido en JavaScript se inserta directamente en el documento HTML que se presenta al usuario.

Se diseñó como un lenguaje para crear scripts que serían incrustados en los archivos HTML y le agregarían dinamismo a las páginas web.

c. CSS

Lenguaje que sirve como complemento y ayuda a reducir lo complejo de HTML. Potencia la estructura visual de la página, con variados estilos y elementos, como tamaño, color, fondos, bordes, etc.



d. PHP

Leguaje de código abierto interpretado del lado del servidor. Resalta por su potencia, versatilidad, modularidad y robustez.

La ventaja de PHP frente a ASP es su característica multiplataforma. Además, resulta un lenguaje que le da a los programas agilidad, estabilidad y rapidez, contrario a los desarrollados bajo ASP.

Actualmente, PHP viene creciendo rápidamente, siendo considerado uno de los lenguajes más potentes y confiables para desarrollar aplicaciones web.

2.3.4. Metodologías de Desarrollo de Software

A. RUP

Proceso Unificado de Rational orientado al desarrollo de software. Es propiedad de IBM. Al lado del Lenguaje Unificado de Modelado UML, forman una de las metodologías más completas y ampliamente utilizadas en el proceso de desarrollo de aplicaciones orientadas a objetos.

RUP no implica la secuencia de fases firmemente definidas; representa una metodología que se adapta al entorno y realidad de casa empresa.

B. XP

La programación extrema es una metodología ágil de desarrollo de software, planteada por Kent Beck. Su principal diferencia del resto, es su enfoque en la adaptabilidad; es decir, los cambios durante la ejecución del proyecto son considerados naturales y aceptables; son tratados como puntos claves para entender mejor el trabajo y lograr mejores resultados.



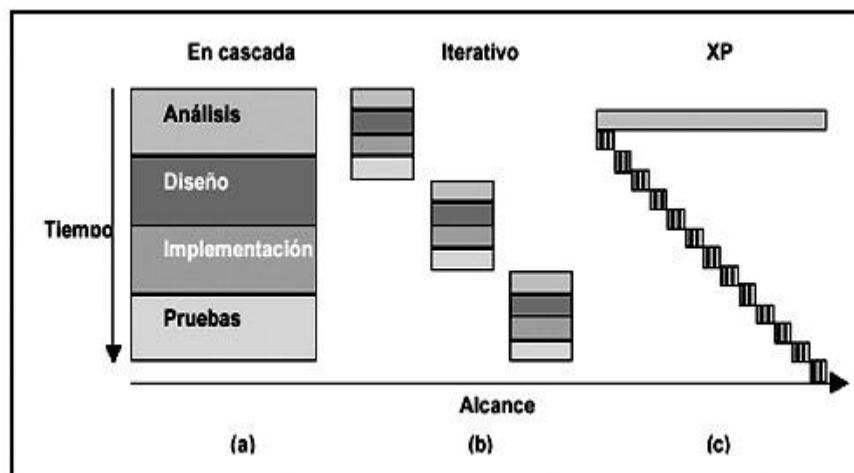


Figura 3. Metodología XP

C. METODOLOGIA SCRUM

Metodología de desarrollo ágil; se fundamenta en la creación de ciclos conocidos como iteraciones, a las cuales se denominan “SPRINTS”.

FASES DE LA METODOLOGIA:

- a. Concepto: definición de las características del producto.
- b. Especulación: plantea los límites a tomarse en cuenta para desarrollar el producto; tales como costes y agendas.
- c. Exploración: se añaden funcionalidades definidas en la fase de especulación.
- d. Revisión: se realiza una exploración de lo construido con la finalidad de identificar errores; además de realizar una contrastación con el objetivo planteado.
- e. Cierre: se entrega en la fecha planificada una versión del producto. De acuerdo a esto se realizan cambios, mantenimiento, hasta que se acerque a la versión final del producto.





Figura 4. Metodología SCRUM

2.3.5. Metodologías de Desarrollo de Modelos de Minería de Datos

A. CRISP – DM

Ésta metodología abarca cuatro niveles de abstracción, que se ordenan jerárquicamente, desde el nivel más general a lo más específico. Consta de seis fases, que recorren el proyecto de minería de datos de forma total; desde la definición de los objetivos del negocio, hasta el control y mantenimiento del modelo propuesto. Cada fase contempla tareas, que se plantean desde la forma más general a lo específico.

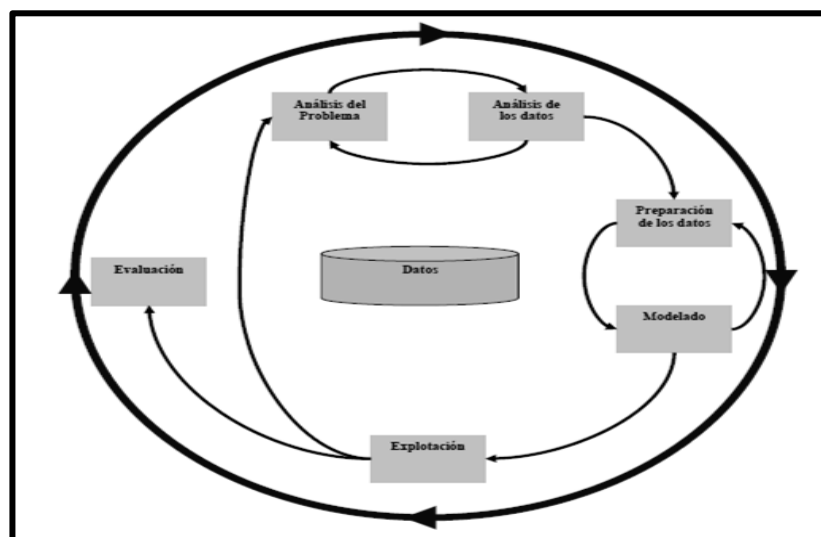


Figura 5. Metodología CRISP-DM.



B. SEMMA

Desarrollado por SAS Institute. Aplica para procesos de selección, exploración y modelado en grandes cantidades de datos, con el propósito de encontrar patrones de negocio no conocidos.

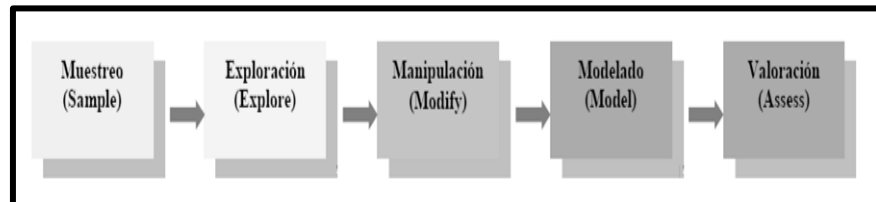


Figura 6. Metodología SEMMA

Todo inicia con la extracción de la población muestral a analizar. Luego se exploran los datos para simplificar el problema y optimizar el modelo. Para esto, se utilizan herramientas estadísticas. En la tercera fase se manipulan los datos, con el fin de definir el formato correcto para introducirlos en el modelo; para después, proseguir con el análisis y modelado de los datos con técnicas como análisis discriminante, métodos de agrupamiento y análisis de regresión, estableciendo las relaciones entre las variables explicativas y las del objeto del estudio. En la fase final, se valoran los resultados con un análisis de bondad a través de métodos.

2.3.6. Herramientas de Desarrollo de Software

Base de Datos

a. Microsoft SQL Server

Gestor de bases de datos de tipo relacional, propiedad de Microsoft. Tiene a T-SQL y ANSI SQL como lenguajes para la estructuración de consultas.



b. Postgres

Sistema de gestión de base de datos relacional, libre y orientado a objetos. Publicado bajo la licencia BSD.

Su desarrollo no es responsabilidad de una empresa en particular, sino de toda una comunidad de desarrolladores, que trabajan de forma conjunta y colaborativa, a la cual se denomina PostgreSQL Global Development Group.

c. MySQL

Sistema de gestión de base de datos relacional, multihilo y multiusuario. Actualmente es desarrollada como software libre bajo un esquema de licenciamiento dual entre Sun Microsystems y Oracle Corporation.



2.4. Definición de términos básicos

A. Almacén de datos

Colección de datos que se enfoca a un ámbito de la empresa u organización. Puede cambiar en el tiempo y está integrada con otros elementos de la empresa; permite ordenar los datos para dar soporte a toma de decisiones. (Kimball, 1998).

B. Análisis prospectivo de datos

Análisis que se realiza sobre un conjunto de datos con la finalidad de predecir tendencias a futuro o comportamientos con base en datos históricos (Lezcano, 2010).

C. Árbol de decisión

Estructura de decisiones definida que toma la forma de un árbol, que establecen reglas para clasificar los datos con los cuales se viene trabajando (Asencios, 2004).

D. Método

Modo de trabajo que sintetiza un fin u objetivo, a través del seguimiento de un proceso organizado y sistemático (Getoor y Ben, 2007).

E. Metodología

Conjunto de métodos que se fundamentan en una disciplina científica (Grudnitsky, 1992).

F. Minería de datos

Conjunto de técnicas orientadas al descubrimiento de conocimiento o patrones de comportamiento presentes en grandes volúmenes de datos. Supone el uso de diferentes algoritmos y herramientas que contribuyen al correcto funcionamiento del proceso (Valcárcel, 2004).

G. Modelo predictivo

Estructura construida con el fin de realizar predicciones con base en un conjunto de datos históricos (Lezcano, 2010).



H. Técnicas de Predicción

Representan métodos cuya aplicación permite realizar estimaciones o pronósticos sobre una serie de tiempo, con base en información histórica presente en la serie tomada como objeto de análisis (Getoor y Ben, 2007).

I. Predicción de ventas

Son las operaciones realizadas por el equipo del área comercial para predecir las ventas del próximo año. Una correcta previsión de las ventas, supone la eficiente tarea de elaboración del presupuesto completo de la empresa (Schaefer, 2012).

J. Desviación absoluta media (MAD)

Determina la precisión del pronóstico realizado, a través del promedio de la magnitud de los errores del pronóstico, para seleccionar el de menor valor.

$$MAD = \frac{\sum_{t=1}^n |F_t - Y_t|}{n}$$

Notas:

- a. MAD = desviación absoluta media
- b. Sumatoria de valores absolutos (Ft son las observaciones actuales de las series de tiempo – Yt son las series de tiempo estimadas o pronosticadas)
- c. n = Numero de observaciones

K. Error cuadrático medio (ECM)

Dado por el promedio de los cuadrados del error calculado en cada periodo. Es utilizado para realizar las comparaciones de precisión entre los métodos de pronósticos, seleccionando el de menor ECM.

$$ECM = \frac{\sum_{t=1}^n (F_t - Y_t)^2}{n}$$

- a. ECM = Error cuadrático medio



- b. Sumatoria de valores (Ft son las observaciones actuales de las series de tiempo – Yt son las series de tiempo estimadas o pronosticadas) al cuadrado
- c. n = Numero de observaciones

L. Raíz del error cuadrático medio (RECM)

Es la raíz del promedio de los cuadrados del error dado en cada periodo. También se usa al comparar y seleccionar la precisión de los diferentes métodos de pronóstico; pero se diferencia del anterior, ya que muestra los resultados en las unidades originales.

$$ECM = \sqrt{\frac{\sum_{t=1}^n (F_t - Y_t)^2}{n}}$$

- a. RECM = Raíz del Error cuadrático medio
- b. Raíz de la Sumatoria de valores (Ft son las observaciones actuales de las series de tiempo – Yt son las series de tiempo estimadas o pronosticadas) al cuadrado
- c. n = Numero de observaciones

M. Porcentaje de error medio (MPE)

Media del error porcentual. Es una medida simple, que muestra si el error del pronóstico tiende hacia arriba o hacia abajo. Puede ser positivo o negativo.

$$MPE = \frac{\sum_{t=1}^n \frac{(F_t - Y_t)}{F_t}}{n}$$

- a. MPE = Porcentaje de error medio
- b. Sumatoria de (Ft son las observaciones actuales de las series de tiempo – Yt son las series de tiempo estimadas o pronosticadas) entre Ft son las observaciones actuales de las series de tiempo



c. n = Numero de observaciones

N. Relación MAD/MEDIA

Representa una alternativa para el MAPE, ya que, puede adaptarse mejor a datos con características intermitentes y de bajo volumen.

Cuando el valor de la serie es igual a cero o toma valores extremos, es imposible hacer uso del MAPE; por lo que lo más adecuado es emplear la relación MAD/Media, es decir, dividir el MAD entre la Media.

CAPITULO III: MARCO METODOLOGICO

3.1. Tipo y diseño de la investigación

3.1.1. Tipo de Investigación

Explicativa: busca describir y explicar detalladamente la influencia que existente de la variable independiente hacia la variable dependiente.

3.1.2. Diseño de la investigación

Pre Experimental – Propositiva

Pre Experimental: tiene como finalidad demostrar la validez de la hipótesis a través del uso de métodos experimentales. No existe un Grupo de Control para realizar las comparaciones pertinentes, ya que se sólo se aplicará a una entidad en particular.

Propositiva: Se fundamenta en el planteamiento de una propuesta de solución para el problema definido.

Pre Experimental - Propositivo

	T₁		T₂
M	O₁	x	O₂

Donde:

M: Es las muestras que se está observando: transacciones, procesos, etc. (Y)

O₁: PRE TEST: Entrevista, encuesta, observación, análisis documentario, etc. (Y)

X: Aplicación a nivel de prueba de la propuesta de especialidad: Técnicas Data Mining (X)

T₁: Tiempo de medición inicial con información actual.

T₂: Tiempo de medición posterior a la simulación de la propuesta de solución X.



O₂: Es la observación luego de la simulación de la propuesta de solución X – POST TEST. (Y)

3.2. Población y muestra

3.2.1. Población

La totalidad de suministros e histórico de consumos de la empresa ELECTRONORTE S.A el cual hasta el periodo 201909 resulta en 326383 suministros, los cuales se dividen por la jerarquía Empresa – Unidad de Negocio – Ciclo de Facturación.



Tabla 1
Suministros por Ciclos de Facturación

Empresa	Unidad De Negocio	Ciclo	Suministros
Electronorte S.A.	Cajamarca Centro	I - Cajamarca Centro	1093
Electronorte S.A.	Cajamarca Centro	II - Cajamarca Centro	5157
Electronorte S.A.	Cajamarca Centro	III - Cajamarca Centro	6988
Electronorte S.A.	Cajamarca Centro	IV - Cajamarca Centro	993
Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358
Electronorte S.A.	Chiclayo	Chiclayo 0	3070
Electronorte S.A.	Chiclayo	Chiclayo 01	8266
Electronorte S.A.	Chiclayo	Chiclayo 01A	7371
Electronorte S.A.	Chiclayo	Chiclayo 02	7293
Electronorte S.A.	Chiclayo	Chiclayo 02A	7536
Electronorte S.A.	Chiclayo	Chiclayo 03	14081
Electronorte S.A.	Chiclayo	Chiclayo 04	13082
Electronorte S.A.	Chiclayo	Chiclayo 05	14060
Electronorte S.A.	Chiclayo	Chiclayo 05A	7003
Electronorte S.A.	Chiclayo	Chiclayo 06	9111
Electronorte S.A.	Chiclayo	Chiclayo 07	14665
Electronorte S.A.	Chiclayo	Chiclayo 08	14305
Electronorte S.A.	Chiclayo	Chiclayo 09	14541
Electronorte S.A.	Chiclayo	Chiclayo 10	14322
Electronorte S.A.	Chiclayo	Chiclayo 11	13404
Electronorte S.A.	Chiclayo	Chiclayo 12	13574
Electronorte S.A.	Sucursales	Sucursales 01	3731
Electronorte S.A.	Sucursales	Sucursales 02	7914
Electronorte S.A.	Sucursales	Sucursales 03	7233
Electronorte S.A.	Sucursales	Sucursales 04	9046
Electronorte S.A.	Sucursales	Sucursales 04A	1941
Electronorte S.A.	Sucursales	Sucursales 05	12930
Electronorte S.A.	Sucursales	Sucursales 06	12709
Electronorte S.A.	Sucursales	Sucursales 07	18320
Electronorte S.A.	Sucursales	Sucursales 08	12887
Electronorte S.A.	Sucursales	Sucursales 09	5939
Electronorte S.A.	Sucursales	Sucursales 10	8520
Electronorte S.A.	Sucursales	Sucursales 11	4454
Electronorte S.A.	Sucursales	Sucursales 14	8843
Electronorte S.A.	Sucursales	Sucursales 12	6208
Electronorte S.A.	Sucursales	Sucursales 13	6435
Total			326383

Fuente: Elaborado por el autor



3.2.2. Muestra

Para fines de la investigación se realizó un muestreo estratificado sobre la población total de suministros divididos por la jerarquía mencionada en el ítem anterior, considere que se aplica un factor de corrección para reducir la muestra y se corrige factorización a 0 muestra, reemplazándola por 1, dado que no puede haber valores 0 para un determinado ítem como muestra.

Tabla 2
Muestreo estratificado a la población de suministros

Empresa	Unidad De Negocio	Ciclo	Suministros	Total Sum	% Rep	C. Muestral	Factor 1%
Electronorte S.A.	Cajamarca Centro	I - Cajamarca Centro	1093	326383	0.0033	0	1
Electronorte S.A.	Cajamarca Centro	II - Cajamarca Centro	5157	326383	0.0158	103	1
Electronorte S.A.	Cajamarca Centro	III - Cajamarca Centro	6988	326383	0.0214	140	1
Electronorte S.A.	Cajamarca Centro	IV - Cajamarca Centro	993	326383	0.003	0	1
Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	326383	0.0287	281	3
Electronorte S.A.	Chiclayo	Chiclayo 0	3070	326383	0.0094	31	1
Electronorte S.A.	Chiclayo	Chiclayo 01	8266	326383	0.0253	248	2
Electronorte S.A.	Chiclayo	Chiclayo 01A	7371	326383	0.0226	147	1
Electronorte S.A.	Chiclayo	Chiclayo 02	7293	326383	0.0223	146	1
Electronorte S.A.	Chiclayo	Chiclayo 02A	7536	326383	0.0231	151	2
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	326383	0.0431	563	6
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	326383	0.0401	523	5
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	326383	0.0431	562	6
Electronorte S.A.	Chiclayo	Chiclayo 05A	7003	326383	0.0215	140	1
Electronorte S.A.	Chiclayo	Chiclayo 06	9111	326383	0.0279	273	3



Empresa	Unidad De Negocio	Ciclo	Suministros	Total Sum	% Rep	C. Muestral	Factor 1%
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	326383	0.0449	587	6
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	326383	0.0438	572	6
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	326383	0.0446	582	6
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	326383	0.0439	573	6
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	326383	0.0411	536	5
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	326383	0.0416	543	5
Electronorte S.A.	Sucursales	Sucursales 01	3731	326383	0.0114	37	1
Electronorte S.A.	Sucursales	Sucursales 02	7914	326383	0.0242	158	2
Electronorte S.A.	Sucursales	Sucursales 03	7233	326383	0.0222	145	1
Electronorte S.A.	Sucursales	Sucursales 04	9046	326383	0.0277	271	3
Electronorte S.A.	Sucursales	Sucursales 04A	1941	326383	0.0059	19	1
Electronorte S.A.	Sucursales	Sucursales 05	12930	326383	0.0396	517	5
Electronorte S.A.	Sucursales	Sucursales 06	12709	326383	0.0389	508	5
Electronorte S.A.	Sucursales	Sucursales 07	18320	326383	0.0561	1099	11
Electronorte S.A.	Sucursales	Sucursales 08	12887	326383	0.0395	515	5
Electronorte S.A.	Sucursales	Sucursales 09	5939	326383	0.0182	119	1
Electronorte S.A.	Sucursales	Sucursales 10	8520	326383	0.0261	256	3
Electronorte S.A.	Sucursales	Sucursales 11	4454	326383	0.0136	45	1
Electronorte S.A.	Sucursales	Sucursales 14	8843	326383	0.0271	265	3
Electronorte S.A.	Sucursales	Sucursales12	6208	326383	0.019	124	1
Electronorte S.A.	Sucursales	Sucursales13	6435	326383	0.0197	129	1
Total							113

Fuente: Elaborado por el autor



Se aplicó un algoritmo de ordenamiento simple para determinar que suministros dentro del grupo muestra obtenido por ciclo debe escogerse, obteniendo la siguiente muestra final a detalle por cantidad de suministros por ciclo y que suministro ha sido seleccionado para la presente investigación.

Tabla 3
Muestreo Detalle

Empresa	UUNN	Ciclo	Total Suministros	Total Muestra	N° Suministro
Electronorte S.A.	Cajamarca Centro	I - Cajamarca Centro	1093	1	28361839
Electronorte S.A.	Cajamarca Centro	II - Cajamarca Centro	5157	1	28217192
Electronorte S.A.	Cajamarca Centro	III - Cajamarca Centro	6988	1	28179959
Electronorte S.A.	Cajamarca Centro	IV - Cajamarca Centro	993	1	25610289
Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	3	25602080
Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	3	25602563
Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	3	25602581
Electronorte S.A.	Chiclayo	Chiclayo 0	3070	1	25778262
Electronorte S.A.	Chiclayo	Chiclayo 01	8266	2	25602302
Electronorte S.A.	Chiclayo	Chiclayo 01	8266	2	25602320
Electronorte S.A.	Chiclayo	Chiclayo 01A	7371	1	25667871
Electronorte S.A.	Chiclayo	Chiclayo 02	7293	1	25403069
Electronorte S.A.	Chiclayo	Chiclayo 02A	7536	2	25191969
Electronorte S.A.	Chiclayo	Chiclayo 02A	7536	2	25606240
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017151
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017170
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017198



Empresa	UUNN	Ciclo	Total Suministros	Total Muestra	N° Suministro
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017204
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017213
Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017231
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190783
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190792
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190809
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190818
Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190827
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005820
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005830
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005849
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005858
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005867
Electronorte S.A.	Chiclayo	Chiclayo 05	14060	6	25005876
Electronorte S.A.	Chiclayo	Chiclayo 05A	7003	1	25300268
Electronorte S.A.	Chiclayo	Chiclayo 06	9111	3	25191996
Electronorte S.A.	Chiclayo	Chiclayo 06	9111	3	25300114
Electronorte S.A.	Chiclayo	Chiclayo 06	9111	3	25300123
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25041961
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25041970
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25068179
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25068188
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25068197
Electronorte S.A.	Chiclayo	Chiclayo 07	14665	6	25068689



Empresa	UUNN	Ciclo	Total Suministros	Total Muestra	N° Suministro
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25191904
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25202777
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25202786
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25202795
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25202801
Electronorte S.A.	Chiclayo	Chiclayo 08	14305	6	25202810
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25191833
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25450309
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25450318
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25450327
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25450336
Electronorte S.A.	Chiclayo	Chiclayo 09	14541	6	25450345
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25104631
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25191851
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25582943
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25583234
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25583270
Electronorte S.A.	Chiclayo	Chiclayo 10	14322	6	25584115
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	5	25191735
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	5	25583038
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	5	25583477
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	5	25583495
Electronorte S.A.	Chiclayo	Chiclayo 11	13404	5	25583548
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	5	25000029



Empresa	UUNN	Ciclo	Total Suministros	Total Muestra	N° Suministro
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	5	25000047
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	5	25000056
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	5	25000065
Electronorte S.A.	Chiclayo	Chiclayo 12	13574	5	25000083
Electronorte S.A.	Sucursales	Sucursales 01	3731	1	25685673
Electronorte S.A.	Sucursales	Sucursales 02	7914	2	25623071
Electronorte S.A.	Sucursales	Sucursales 02	7914	2	25660469
Electronorte S.A.	Sucursales	Sucursales 03	7233	1	25600667
Electronorte S.A.	Sucursales	Sucursales 04	9046	3	25623151
Electronorte S.A.	Sucursales	Sucursales 04	9046	3	25623160
Electronorte S.A.	Sucursales	Sucursales 04	9046	3	25668046
Electronorte S.A.	Sucursales	Sucursales 04A	1941	1	26903991
Electronorte S.A.	Sucursales	Sucursales 05	12930	5	25604085
Electronorte S.A.	Sucursales	Sucursales 05	12930	5	25676988
Electronorte S.A.	Sucursales	Sucursales 05	12930	5	25677009
Electronorte S.A.	Sucursales	Sucursales 05	12930	5	25677027
Electronorte S.A.	Sucursales	Sucursales 05	12930	5	25677054
Electronorte S.A.	Sucursales	Sucursales 06	12709	5	25602062
Electronorte S.A.	Sucursales	Sucursales 06	12709	5	25603909
Electronorte S.A.	Sucursales	Sucursales 06	12709	5	25613048
Electronorte S.A.	Sucursales	Sucursales 06	12709	5	25647832
Electronorte S.A.	Sucursales	Sucursales 06	12709	5	25647959
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25159818
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25159845



Empresa	UUNN	Ciclo	Total Suministros	Total Muestra	N° Suministro
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25159872
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25159890
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25160087
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25160096
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25160130
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25608039
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25608146
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25608155
Electronorte S.A.	Sucursales	Sucursales 07	18320	11	25608164
Electronorte S.A.	Sucursales	Sucursales 08	12887	5	25758671
Electronorte S.A.	Sucursales	Sucursales 08	12887	5	25760428
Electronorte S.A.	Sucursales	Sucursales 08	12887	5	25760437
Electronorte S.A.	Sucursales	Sucursales 08	12887	5	25777247
Electronorte S.A.	Sucursales	Sucursales 08	12887	5	25817567
Electronorte S.A.	Sucursales	Sucursales 09	5939	1	25607363
Electronorte S.A.	Sucursales	Sucursales 10	8520	3	25605681
Electronorte S.A.	Sucursales	Sucursales 10	8520	3	25612120
Electronorte S.A.	Sucursales	Sucursales 10	8520	3	25612130
Electronorte S.A.	Sucursales	Sucursales 11	4454	1	25160022
Electronorte S.A.	Sucursales	Sucursales 14	8843	3	26414755
Electronorte S.A.	Sucursales	Sucursales 14	8843	3	26925809
Electronorte S.A.	Sucursales	Sucursales 14	8843	3	26925818
Electronorte S.A.	Sucursales	Sucursales12	6208	1	25685637
Electronorte S.A.	Sucursales	Sucursales13	6435	1	25685655

Fuente: Elaborado por el autor



3.3. Hipótesis

El análisis de las técnicas de minería de datos me permitirá identificar la mejor herramienta para estimaciones de consumo de energía eléctrica.

3.4. Variables

3.4.1. Variable Independiente

Técnicas de minería de datos

Orientadas al descubrimiento de conocimiento y la búsqueda de patrones de comportamiento similares en grandes conjuntos de datos. Los algoritmos suponen la extracción de información precisa que de soporte a la toma de decisiones, que permita predecir determinadas situaciones presentes en las actitudes comunes en diversos clientes (Valcárcel, 2004).

3.4.2. Variable Dependiente

Estimaciones de consumo de energía eléctrica

3.5. Operacionalización de variables

Tabla 4
Variable Independiente: Técnicas de minería de datos

Dimensión	Indicador	Ecuación
Confiabilidad	Índice de error de algoritmos evaluados	<p>MAPE</p> $MAPE = \frac{\sum_{t=1}^n \frac{ F_t - Y_t }{F_t}}{n}$ <p>a. MAPE = Porcentaje de error medio absoluto b. Sumatoria de valores absolutos (Ft son las observaciones actuales de las series de tiempo – Yt son las series de tiempo estimadas o pronosticadas) entre Ft son las observaciones actuales de las series de tiempo c. n = Numero de observaciones</p> <p>ACC</p> <p>Precisión es Exactitud – Error (MAPE)</p>
	Tiempo	<p>Índice de tiempo de entrenamiento para cada algoritmo</p> <p>PROMEDIO DE TIEMPO EN SEGUNDOS PARA CASOS DE MUESTRA</p>
Longitud	<p>Numero de meses mínimo soportado para generar entrenamiento y pronósticos</p> <p>CANTIDAD DE OBSERVACIONES (PERIODOS DE FACTURACION)</p>	

Fuente: Elaborado por el autor



Tabla 5
Variable Dependiente: Estimaciones de consumo de energía eléctrica

Dimensión	Indicador	Pregunta	Ecuación	Tipo	UM	Categoría	Ítem
Estimación	Eficiencia del modelo	¿Cuántos errores genera el modelo?	Basado en ACC Tabla 4	Cuantitativo	%	MALO: 0 a 84%	1
						BUENO: 85 a 100%	
Tiempo	Tiempo de ejecución del modelo	¿Cuánto tiempo tarda en procesar el modelo los pronósticos?	Basado en tiempo Tabla 4	Cuantitativo	Seg.	MALO: más de 1 minuto	2
						REGULAR: 30 seg a 60 seg	

Fuente: Elaborado por el autor

3.6. Abordaje Metodológico, técnicas e instrumentos de recolección de datos

3.6.1. Abordaje Metodológico

El método usado es el inductivo, pues se parte de una situación específica hacia lo general; es decir, a través de técnicas se logra captar una determinada situación problemática en la empresa, la misma que será descrita y analizada durante el proceso de investigación, para dar como resultado un producto.

3.6.2. Técnicas de recolección de datos

Entrevistas

Representan los diálogos entablados con las personas que forman parte importante del proceso en estudio, es decir, los involucrados en las



distintas áreas; los cuales están en la capacidad de brindar información confiable y valiosa que conlleve a la realización exitosa del proyecto

Uso de Metodología para diseño de modelos predictivos

Representan las metodologías usadas para el desarrollo de la investigación. Para el caso en particular, se utiliza la metodología CRISP- DM para la aplicación de técnicas de minería de datos y la creación de los modelos y, la metodología de desarrollo ágil XP, para la construcción del sistema web.

3.6.3. Instrumentos de recolección de datos

Guía de entrevista: contiene la lista de preguntas a realizarse, de acuerdo al número de participantes determinados previamente y que se relacionan directamente con el proceso de investigación de la tesis.

3.7. Procedimiento para la recolección de datos

Extracción de datos del Datamart de facturación provisto por ElectroNorte S.A.

3.8. Análisis estadístico e interpretación de los datos

Tabulación de datos

- A través de tablas y gráficos estadísticos
- Organizado por preguntas

Análisis de datos

- La interpretación es en base a los indicadores planteados
- Los indicadores de tendencia central a usar son: Media, mediana, moda

- Los indicadores de tendencia no central a usar son: Cuartiles, deciles, percentiles
- Los indicadores de medida de dispersión a utilizar son: Rango, varianza, desviación típica, coeficiente de variación.
- Utilización del software SPSS y MS Excel para el análisis de los datos

3.9. Principios éticos

Tabla 6
Criterios éticos

Confiabilidad	Características científicas del criterio
Confiabilidad	Se aplicaron técnicas estadísticas para medir y apreciar los niveles de consistencia de cada uno de los instrumentos de recolección de los datos.
Validación	La propuesta de solución planteada y los distintos instrumentos de recolección de datos diseñados y aplicados fueron validados a través de la técnica de Juicio de Expertos.

Fuente: Elaborado por el autor



3.10. Criterios de rigor científico

Tabla 7
Criterios de rigor científico

Criterios	Características éticas del criterio
Confidencialidad	Se asegura el anonimato total y la protección de los datos personales de cada uno de los participantes de la investigación.
Objetividad	Se garantiza el uso de criterios técnicos y morales, orientados a brindar imparcialidad en cada aspecto de la situación encontrada.
Originalidad	Se citan cada una de las fuentes bibliográficas consultadas y sobre las cuales se basó la investigación, como prueba de la inexistencia de plagio intelectual.
Veracidad	Los datos recopilados y mostrados son totalmente verdaderos, cuidando la transparencia y la confiabilidad de los mismos.
Derechos laborales	Se asegura el respeto por los derechos laborales de cada uno de los trabajadores y colaboradores de la entidad en estudio.

Fuente: Elaborado por el autor



CAPITULO IV: ANALISIS E INTERPRETACION DE LOS RESULTADOS

4.1. Resultados en tablas y gráficos

A. Confiabilidad de la Predicción

En el siguiente análisis, de un total de 113 suministros, se ha validado el valor del pronóstico con el valor real, para poder hallar el error de cada suministro, utilizando la fórmula del Error Porcentual Absoluto Medio (MAPE).

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$



Tabla 8
Confiabilidad de predicción por algoritmo

UUNN	Ciclo	Mes	Suministro	Periodo	Real	Arima	Error Ar	Hw	Error Hw	Nnetar	Error Nnetar	Svm	Error Svm
Cajamarca Centro	I - Cajamarca Centro	1	28361839	201909	20	20	0	20	0	20	0	20	0
Cajamarca Centro	II - Cajamarca Centro	1	28217192	201909	68	71	4.41	65	4.41	70	2.94	67	1.47
Cajamarca Centro	III - Cajamarca Centro	1	28179959	201909	184	183	0.54	246	33.7	181	1.63	179	2.72
Cajamarca Centro	IV - Cajamarca Centro	1	25610289	201909	25	25	0	24	4	25	0	25	0
Cajamarca Centro	V - Cajamarca Centro	3	25602080	201909	16	16	0	7	56.25	16	0	17	6.25
Cajamarca Centro	V - Cajamarca Centro	3	25602563	201909	165	163	1.21	173	4.85	164	0.61	169	2.42
Cajamarca Centro	V - Cajamarca Centro	3	25602581	201909	43	43	0	48	11.63	43	0	44	2.33
Chiclayo	Chiclayo 0	1	25778262	201909	87	107	22.99	42	51.72	87	0	122	40.23
Chiclayo	Chiclayo 01	2	25602302	201909	6454	0	100	7620	18.07	4889	24.25	2069	67.94
Chiclayo	Chiclayo 01	2	25602320	201909	113	131	15.93	132	16.81	111	1.77	97	14.16
Chiclayo	Chiclayo 01A	1	25667871	201909	92	89	3.26	74	19.57	94	2.17	102	10.87
Chiclayo	Chiclayo 02	1	25403069	201909	42	41	2.38	44	4.76	43	2.38	35	16.67
Chiclayo	Chiclayo 02A	2	25191969	201909	0	0	0	0	0	0	0	0	0
Chiclayo	Chiclayo 02A	2	25606240	201909	132	122	7.58	137	3.79	132	0	107	18.94
Chiclayo	Chiclayo 03	6	25017151	201909	196	184	6.12	180	8.16	193	1.53	234	19.39
Chiclayo	Chiclayo 03	6	25017170	201909	78	81	3.85	77	1.28	80	2.56	75	3.85
Chiclayo	Chiclayo 03	6	25017198	201909	176	175	0.57	182	3.41	173	1.7	158	10.23
Chiclayo	Chiclayo 03	6	25017204	201909	329	330	0.3	335	1.82	334	1.52	294	10.64
Chiclayo	Chiclayo 03	6	25017213	201909	524	525	0.19	547	4.39	532	1.53	576	9.92
Chiclayo	Chiclayo 03	6	25017231	201909	1460	1440	1.37	1448	0.82	1483	1.58	1308	10.41
Chiclayo	Chiclayo 04	5	25190783	201909	1285	2580	100.78	2271	76.73	1155	10.12	958	25.45
Chiclayo	Chiclayo 04	5	25190792	201909	179	199	11.17	173	3.35	180	0.56	214	19.55



UUNN	Ciclo	Mes	Suministro	Periodo	Real	Arima	Error Ar	Hw	Error Hw	Nnetar	Error Nnetar	Svm	Error Svm
Chiclayo	Chiclayo 04	5	25190809	201909	242	241	0.41	230	4.96	240	0.83	236	2.48
Chiclayo	Chiclayo 04	5	25190818	201909	282	284	0.71	302	7.09	278	1.42	284	0.71
Chiclayo	Chiclayo 04	5	25190827	201909	275	336	22.18	314	14.18	266	3.27	221	19.64
Chiclayo	Chiclayo 05	6	25005820	201909	0	0	0	0	0	0	0	0	0
Chiclayo	Chiclayo 05	6	25005830	201909	591	0	100	1378	133.16	464	21.49	0	100
Chiclayo	Chiclayo 05	6	25005849	201909	20	20	0	17	15	21	5	19	5
Chiclayo	Chiclayo 05	6	25005858	201909	2355	2369	0.59	2355	0	2627	11.55	2338	0.72
Chiclayo	Chiclayo 05	6	25005867	201909	195	198	1.54	201	3.08	199	2.05	176	9.74
Chiclayo	Chiclayo 05	6	25005876	201909	200	165	17.5	300	50	212	6	131	34.5
Chiclayo	Chiclayo 05A	1	25300268	201909	193	233	20.73	233	20.73	217	12.44	161	16.58
Chiclayo	Chiclayo 06	3	25191996	201909	0	0	0	0	0	0	0	0	0
Chiclayo	Chiclayo 06	3	25300114	201909	55	56	1.82	53	3.64	56	1.82	51	7.27
Chiclayo	Chiclayo 06	3	25300123	201909	23	25	8.7	22	4.35	26	13.04	0	100
Chiclayo	Chiclayo 07	6	25041961	201909	235	246	4.68	243	3.4	236	0.43	233	0.85
Chiclayo	Chiclayo 07	6	25041970	201909	321	309	3.74	301	6.23	318	0.93	318	0.93
Chiclayo	Chiclayo 07	6	25068179	201909	197	218	10.66	185	6.09	211	7.11	239	21.32
Chiclayo	Chiclayo 07	6	25068188	201909	37	34	8.11	33	10.81	34	8.11	28	24.32
Chiclayo	Chiclayo 07	6	25068197	201909	98	98	0	115	17.35	96	2.04	90	8.16
Chiclayo	Chiclayo 07	6	25068689	201909	410	416	1.46	447	9.02	403	1.71	371	9.51
Chiclayo	Chiclayo 08	6	25191904	201909	3948	19596	396.35	9254	134.4	5410	37.03	1189	69.88
Chiclayo	Chiclayo 08	6	25202777	201909	220	219	0.45	225	2.27	214	2.73	207	5.91
Chiclayo	Chiclayo 08	6	25202786	201909	49	51	4.08	56	14.29	44	10.2	51	4.08
Chiclayo	Chiclayo 08	6	25202795	201909	126040	95042	24.59	1E+05	8.33	208366	65.32	22094	82.47
Chiclayo	Chiclayo 08	6	25202801	201909	155	153	1.29	172	10.97	153	1.29	159	2.58



UUNN	Ciclo	Mes	Suministro	Periodo	Real	Arima	Error Ar	Hw	Error Hw	Nnetar	Error Nnetar	Svm	Error Svm
Chiclayo	Chiclayo 08	6	25202810	201909	762	0	100	296	61.15	1350	77.17	330	56.69
Chiclayo	Chiclayo 09	6	25191833	201909	6852	19027	177.69	34351	401.33	32315	371.61	0	100
Chiclayo	Chiclayo 09	6	25450309	201909	72	70	2.78	99	37.5	72	0	3	95.83
Chiclayo	Chiclayo 09	6	25450318	201909	40	43	7.5	31	22.5	68	70	45	12.5
Chiclayo	Chiclayo 09	6	25450327	201909	401	401	0	349	12.97	409	2	372	7.23
Chiclayo	Chiclayo 09	6	25450336	201909	130	132	1.54	123	5.38	131	0.77	134	3.08
Chiclayo	Chiclayo 09	6	25450345	201909	1407	1397	0.71	1348	4.19	1378	2.06	1449	2.99
Chiclayo	Chiclayo 10	6	25104631	201909	196	196	0	198	1.02	193	1.53	207	5.61
Chiclayo	Chiclayo 10	6	25191851	201909	0	0	0	0	0	0	0	0	0
Chiclayo	Chiclayo 10	6	25582943	201909	113	113	0	110	2.65	114	0.88	118	4.42
Chiclayo	Chiclayo 10	6	25583234	201909	227	224	1.32	212	6.61	223	1.76	183	19.38
Chiclayo	Chiclayo 10	6	25583270	201909	89	85	4.49	80	10.11	89	0	98	10.11
Chiclayo	Chiclayo 10	6	25584115	201909	32	31	3.12	24	25	28	12.5	32	0
Chiclayo	Chiclayo 11	5	25191735	201909	689	799	15.97	722	4.79	692	0.44	664	3.63
Chiclayo	Chiclayo 11	5	25583038	201909	249	248	0.4	212	14.86	255	2.41	286	14.86
Chiclayo	Chiclayo 11	5	25583477	201909	32	32	0	30	6.25	38	18.75	26	18.75
Chiclayo	Chiclayo 11	5	25583495	201909	68	67	1.47	71	4.41	71	4.41	15	77.94
Chiclayo	Chiclayo 11	5	25583548	201909	102	97	4.9	99	2.94	102	0	101	0.98
Chiclayo	Chiclayo 12	5	25000029	201909	249	242	2.81	235	5.62	250	0.4	260	4.42
Chiclayo	Chiclayo 12	5	25000047	201909	3691	2943	20.27	0	100	2956	19.91	1570	57.46
Chiclayo	Chiclayo 12	5	25000056	201909	194	217	11.86	196	1.03	252	29.9	183	5.67
Chiclayo	Chiclayo 12	5	25000065	201909	55	55	0	51	7.27	54	1.82	57	3.64
Chiclayo	Chiclayo 12	5	25000083	201909	3445	3689	7.08	0	100	2665	22.64	1085	68.51
Sucursales	Sucursales 01	1	25685673	201909	40	37	7.5	33	17.5	39	2.5	41	2.5



UUNN	Ciclo	Mes	Suministro	Periodo	Real	Arima	Error Ar	Hw	Error Hw	Nnetar	Error Nnetar	Svm	Error Svm
Sucursales	Sucursales 02	2	25623071	201909	18	18	0	16	11.11	18	0	19	5.56
Sucursales	Sucursales 02	2	25660469	201909	1	1	0	1	0	1	0	1	0
Sucursales	Sucursales 03	1	25600667	201909	177	185	4.52	225	27.12	175	1.13	143	19.21
Sucursales	Sucursales 04	3	25623151	201909	3764	3728	0.96	3990	6	3815	1.35	3568	5.21
Sucursales	Sucursales 04	3	25623160	201909	0	0	0	0	0	0	0	0	0
Sucursales	Sucursales 04	3	25668046	201909	55	55	0	62	12.73	55	0	48	12.73
Sucursales	Sucursales 04A	1	26903991	201909	86	89	3.49	109	26.74	88	2.33	84	2.33
Sucursales	Sucursales 05	5	25604085	201909	112	113	0.89	109	2.68	113	0.89	122	8.93
Sucursales	Sucursales 05	5	25676988	201909	0	0	0	0	0	0	0	0	0
Sucursales	Sucursales 05	5	25677009	201909	26463	29293	10.69	22863	13.6	25764	2.64	37013	39.87
Sucursales	Sucursales 05	5	25677027	201909	17	18	5.88	17	0	20	17.65	18	5.88
Sucursales	Sucursales 05	5	25677054	201909	56	57	1.79	60	7.14	62	10.71	46	17.86
Sucursales	Sucursales 06	5	25602062	201909	169	161	4.73	147	13.02	189	11.83	147	13.02
Sucursales	Sucursales 06	5	25603909	201909	309	323	4.53	275	11	280	9.39	457	47.9
Sucursales	Sucursales 06	5	25613048	201909	0	0	0	0	0	0	0	0	0
Sucursales	Sucursales 06	5	25647832	201909	65	62	4.62	67	3.08	65	0	36	44.62
Sucursales	Sucursales 06	5	25647959	201909	525	554	5.52	214	59.24	520	0.95	673	28.19
Sucursales	Sucursales 07	11	25159818	201909	1907	1952	2.36	1819	4.61	1901	0.31	1924	0.89
Sucursales	Sucursales 07	11	25159845	201909	25147	20459	18.64	22721	9.65	24436	2.83	23888	5.01
Sucursales	Sucursales 07	11	25159872	201909	18	18	0	23	27.78	21	16.67	9	50
Sucursales	Sucursales 07	11	25159890	201909	1131	1120	0.97	1138	0.62	1195	5.66	954	15.65
Sucursales	Sucursales 07	11	25160087	201909	152	111	26.97	153	0.66	108	28.95	116	23.68
Sucursales	Sucursales 07	11	25160096	201909	5817	2717	53.29	2173	62.64	9726	67.2	3340	42.58
Sucursales	Sucursales 07	11	25160130	201909	1859	1846	0.7	1702	8.45	1725	7.21	2125	14.31



UUNN	Ciclo	Mes	Suministro	Periodo	Real	Arima	Error Ar	Hw	Error Hw	Nnetar	Error Nnetar	Svm	Error Svm
Sucursales	Sucursales 07	11	25608039	201909	71	84	18.31	80	12.68	73	2.82	71	0
Sucursales	Sucursales 07	11	25608146	201909	57	57	0	58	1.75	55	3.51	54	5.26
Sucursales	Sucursales 07	11	25608155	201909	96	95	1.04	103	7.29	96	0	94	2.08
Sucursales	Sucursales 07	11	25608164	201909	114	113	0.88	127	11.4	117	2.63	106	7.02
Sucursales	Sucursales 08	5	25758671	201909	36	36	0	43	19.44	37	2.78	35	2.78
Sucursales	Sucursales 08	5	25760428	201909	29	29	0	24	17.24	29	0	31	6.9
Sucursales	Sucursales 08	5	25760437	201909	32	36	12.5	38	18.75	32	0	31	3.12
Sucursales	Sucursales 08	5	25777247	201909	97	93	4.12	88	9.28	98	1.03	103	6.19
Sucursales	Sucursales 08	5	25817567	201909	103	105	1.94	118	14.56	104	0.97	84	18.45
Sucursales	Sucursales 09	1	25607363	201909	1609	1873	16.41	1720	6.9	1690	5.03	1541	4.23
Sucursales	Sucursales 10	3	25605681	201909	91	5	94.51	1351	1384.62	144	58.24	60	34.07
Sucursales	Sucursales 10	3	25612120	201909	4020	681	83.06	0	100	5120	27.36	2901	27.84
Sucursales	Sucursales 10	3	25612130	201909	30	37	23.33	28	6.67	31	3.33	31	3.33
Sucursales	Sucursales 11	1	25160022	201909	293	163	44.37	92	68.6	214	26.96	244	16.72
Sucursales	Sucursales 14	3	26414755	201909	10	9	10	4	60	11	10	12	20
Sucursales	Sucursales 14	3	26925809	201909	24	23	4.17	26	8.33	27	12.5	23	4.17
Sucursales	Sucursales 14	3	26925818	201909	11	11	0	12	9.09	11	0	12	9.09
Sucursales	Sucursales12	1	25685637	201909	27	39	44.44	32	18.52	28	3.7	18	33.33
Sucursales	Sucursales13	1	25685655	201909	23	25	8.7	29	26.09	25	8.7	16	30.43

Fuente: Elaborado por el autor



Tabla 9
Indicador I - Estimación

	Medida	ARIMA	HW	NNETAR	SVM
1	MAPE	15.94672566	33.389646	11.26884956	18.00646018
2	ACC	84.05327434	66.610354	88.73115044	81.99353982

Fuente: Elaborado por el autor

MAPE: Error Porcentual Absoluto Medio, es una medida de la precisión de predicción de un método de pronóstico.

ACC: Exactitud y precisión de un pronóstico, obtenido a partir del error calculado (MAPE).

Se obtuvo un desempeño ACC (Accuracy) para Arima de 84.05 % dado que el error MAPE resultó en 15.94%. Para el caso de HoltWinter el ACC es de 66.61 y el MAPE obtenido es de 33.38 %, siendo el peor modelo según los resultados, en el caso de la Red Neuronal Autoregresiva NNETAR el ACC obtiene un 88.73 % y un MAPE de 11.26 % el cual tiene el mejor desempeño de los modelos, para el caso del SVM el valor del ACC es 81.99% y un MAPE de 18.00 %, ordenando estos algoritmos por su desempeño obtenemos a la RED NEURONAL AUTOREGRESIVA, ARIMA, SVM y HOLTWINTERS. Según nuestro cuadro de evaluación de la métrica de estimación Tabla 5, el único modelo que podría entrar a la categoría de BUENO es la RED NEURONAL AUTOREGRESIVA.



B. Tiempo de Procesamiento del Modelo

Tabla 10
Indicador II – Tiempo de procesamiento Detallado

UUNN	Ciclo	Suministros Total	Muestra	Suministro	Periodo Pronosticado	Tiempo ARIMA	Tiempo HW	Tiempo NNETAR	Tiempo SVM
Cajamarca Centro	I - Cajamarca Centro	1093	1	28361839	201909	0.4	0.1	0.6	0.06
Cajamarca Centro	II - Cajamarca Centro	5157	1	28217192	201909	0.28	0.19	0.6	0.07
Cajamarca Centro	III - Cajamarca Centro	6988	1	28179959	201909	0.2	0.12	0.59	0.08
Cajamarca Centro	IV - Cajamarca Centro	993	1	25610289	201909	0.3	0.14	0.36	0.05
Cajamarca Centro	V - Cajamarca Centro	9358	3	25602080	201909	0.19	0.16	1.1	0.04
Cajamarca Centro	V - Cajamarca Centro	9358	3	25602563	201909	0.22	0.14	0.59	0.05
Cajamarca Centro	V - Cajamarca Centro	9358	3	25602581	201909	0.26	0.11	1.12	0.07
Chiclayo	Chiclayo 0	3070	1	25778262	201909	0.14	0.15	0.26	0.06
Chiclayo	Chiclayo 01	8266	2	25602302	201909	0.28	0.11	0.25	0.18
Chiclayo	Chiclayo 01	8266	2	25602320	201909	0.25	0.16	0.39	0.05
Chiclayo	Chiclayo 01A	7371	1	25667871	201909	0.27	0.22	0.35	0.05
Chiclayo	Chiclayo 02	7293	1	25403069	201909	0.21	0.11	0.27	0.07
Chiclayo	Chiclayo 02A	7536	2	25191969	201909	0	0	0	0
Chiclayo	Chiclayo 02A	7536	2	25606240	201909	0.14	0.09	0.29	0.06
Chiclayo	Chiclayo 03	14081	6	25017151	201909	0.22	0.1	0.66	0.12
Chiclayo	Chiclayo 03	14081	6	25017170	201909	0.25	0.12	0.41	0.14
Chiclayo	Chiclayo 03	14081	6	25017198	201909	0.24	0.18	0.24	0.06
Chiclayo	Chiclayo 03	14081	6	25017204	201909	0.19	0.11	0.48	0.06
Chiclayo	Chiclayo 03	14081	6	25017213	201909	0.2	0.11	0.64	0.07
Chiclayo	Chiclayo 03	14081	6	25017231	201909	0.3	0.13	0.44	0.06



UUNN	Ciclo	Suministros Total	Muestra	Suministro	Periodo Pronosticado	Tiempo ARIMA	Tiempo HW	Tiempo NNETAR	Tiempo SVM
Chiclayo	Chiclayo 04	13082	5	25190783	201909	0.27	0.14	0.38	0.07
Chiclayo	Chiclayo 04	13082	5	25190792	201909	0.2	0.14	0.32	0.08
Chiclayo	Chiclayo 04	13082	5	25190809	201909	0.31	0.17	0.32	0.09
Chiclayo	Chiclayo 04	13082	5	25190818	201909	0.38	0.15	0.44	0.07
Chiclayo	Chiclayo 04	13082	5	25190827	201909	0.33	0.11	0.27	0.06
Chiclayo	Chiclayo 05	14060	6	25005820	201909	0	0	0	0
Chiclayo	Chiclayo 05	14060	6	25005830	201909	0.23	0.14	0.59	0.08
Chiclayo	Chiclayo 05	14060	6	25005849	201909	0.19	0.12	0.43	0.05
Chiclayo	Chiclayo 05	14060	6	25005858	201909	0.21	0.11	0.37	0.05
Chiclayo	Chiclayo 05	14060	6	25005867	201909	0.2	0.27	0.27	0.04
Chiclayo	Chiclayo 05	14060	6	25005876	201909	0.33	0.26	0.29	0.06
Chiclayo	Chiclayo 05A	7003	1	25300268	201909	0.39	0.12	0.55	0.07
Chiclayo	Chiclayo 06	9111	3	25191996	201909	0	0	0	0
Chiclayo	Chiclayo 06	9111	3	25300114	201909	0.2	0.13	0.25	0.06
Chiclayo	Chiclayo 06	9111	3	25300123	201909	0.2	0.12	0.53	0.06
Chiclayo	Chiclayo 07	14665	6	25041961	201909	0.25	0.16	0.39	0.05
Chiclayo	Chiclayo 07	14665	6	25041970	201909	0.32	0.11	0.28	0.06
Chiclayo	Chiclayo 07	14665	6	25068179	201909	0.16	0.1	0.63	0.05
Chiclayo	Chiclayo 07	14665	6	25068188	201909	0.1	0.11	0.71	0.07
Chiclayo	Chiclayo 07	14665	6	25068197	201909	0.25	0.14	0.49	0.07
Chiclayo	Chiclayo 07	14665	6	25068689	201909	0.2	0.13	0.63	0.06
Chiclayo	Chiclayo 08	14305	6	25191904	201909	0.27	0.16	0.39	0.06
Chiclayo	Chiclayo 08	14305	6	25202777	201909	0.25	0.13	0.25	0.06
Chiclayo	Chiclayo 08	14305	6	25202786	201909	0.2	0.13	0.25	0.05



UUNN	Ciclo	Suministros Total	Muestra	Suministro	Periodo Pronosticado	Tiempo ARIMA	Tiempo HW	Tiempo NNETAR	Tiempo SVM
Chiclayo	Chiclayo 08	14305	6	25202795	201909	0.25	0.14	0.66	0.07
Chiclayo	Chiclayo 08	14305	6	25202801	201909	0.17	0.1	0.6	0.07
Chiclayo	Chiclayo 08	14305	6	25202810	201909	0.2	0.11	0.34	0.1
Chiclayo	Chiclayo 09	14541	6	25191833	201909	0.14	0.25	0.71	0.06
Chiclayo	Chiclayo 09	14541	6	25450309	201909	0.27	0.11	0.58	0.05
Chiclayo	Chiclayo 09	14541	6	25450318	201909	0.25	0.15	0.62	0.06
Chiclayo	Chiclayo 09	14541	6	25450327	201909	0.2	0.17	0.25	0.06
Chiclayo	Chiclayo 09	14541	6	25450336	201909	0.17	0.12	0.3	0.14
Chiclayo	Chiclayo 09	14541	6	25450345	201909	0.23	0.17	0.29	0.16
Chiclayo	Chiclayo 10	14322	6	25104631	201909	0.25	0.19	0.29	0.07
Chiclayo	Chiclayo 10	14322	6	25191851	201909	0	0	0	0
Chiclayo	Chiclayo 10	14322	6	25582943	201909	0.25	0.15	0.57	0.07
Chiclayo	Chiclayo 10	14322	6	25583234	201909	0.22	0.09	0.61	0.06
Chiclayo	Chiclayo 10	14322	6	25583270	201909	0.2	0.13	0.61	0.05
Chiclayo	Chiclayo 10	14322	6	25584115	201909	0.23	0.13	0.95	0.04
Chiclayo	Chiclayo 11	13404	5	25191735	201909	0.25	0.11	0.36	0.04
Chiclayo	Chiclayo 11	13404	5	25583038	201909	0.19	0.16	0.28	0.08
Chiclayo	Chiclayo 11	13404	5	25583477	201909	0.21	0.12	0.4	0.08
Chiclayo	Chiclayo 11	13404	5	25583495	201909	0.2	0.25	0.6	0.08
Chiclayo	Chiclayo 11	13404	5	25583548	201909	0.28	0.24	0.57	0.08
Chiclayo	Chiclayo 12	13574	5	25000029	201909	0.39	0.15	0.5	0.1
Chiclayo	Chiclayo 12	13574	5	25000047	201909	0.38	0.27	0.55	0.14
Chiclayo	Chiclayo 12	13574	5	25000056	201909	0.31	0.22	0.97	0.14
Chiclayo	Chiclayo 12	13574	5	25000065	201909	0.54	0.29	0.52	0.1



UUNN	Ciclo	Suministros Total	Muestra	Suministro	Periodo Pronosticado	Tiempo ARIMA	Tiempo HW	Tiempo NNETAR	Tiempo SVM
Chiclayo	Chiclayo 12	13574	5	25000083	201909	0.35	0.22	0.92	0.12
Sucursales	Sucursales 01	3731	1	25685673	201909	0.35	0.17	0.42	0.07
Sucursales	Sucursales 02	7914	2	25623071	201909	0.26	0.14	0.35	0.06
Sucursales	Sucursales 02	7914	2	25660469	201909	0.22	0.28	0.34	0.13
Sucursales	Sucursales 03	7233	1	25600667	201909	0.24	0.19	0.85	0.08
Sucursales	Sucursales 04	9046	3	25623151	201909	0.18	0.22	0.29	0.1
Sucursales	Sucursales 04	9046	3	25623160	201909	0	0	0	0
Sucursales	Sucursales 04	9046	3	25668046	201909	0.28	0.14	0.57	0.06
Sucursales	Sucursales 04A	1941	1	26903991	201909	0.34	0.17	0.38	0.06
Sucursales	Sucursales 05	12930	5	25604085	201909	0.17	0.15	0.51	0.04
Sucursales	Sucursales 05	12930	5	25676988	201909	0	0	0	0
Sucursales	Sucursales 05	12930	5	25677009	201909	0.14	0.14	0.4	0.06
Sucursales	Sucursales 05	12930	5	25677027	201909	0.28	0.12	0.47	0.05
Sucursales	Sucursales 05	12930	5	25677054	201909	0.19	0.13	0.51	0.06
Sucursales	Sucursales 06	12709	5	25602062	201909	0.16	0.13	0.67	0.04
Sucursales	Sucursales 06	12709	5	25603909	201909	0.28	0.12	0.25	0.05
Sucursales	Sucursales 06	12709	5	25613048	201909	0	0	0	0
Sucursales	Sucursales 06	12709	5	25647832	201909	0.22	0.11	0.3	0.05
Sucursales	Sucursales 06	12709	5	25647959	201909	0.28	0.15	0.25	0.05
Sucursales	Sucursales 07	18320	11	25159818	201909	0.33	0.13	0.29	0.05
Sucursales	Sucursales 07	18320	11	25159845	201909	0.29	0.12	0.52	0.05
Sucursales	Sucursales 07	18320	11	25159872	201909	0.21	0.12	0.25	0.05
Sucursales	Sucursales 07	18320	11	25159890	201909	0.17	0.11	0.6	0.08
Sucursales	Sucursales 07	18320	11	25160087	201909	0.11	0.11	1.17	0.06



UUNN	Ciclo	Suministros Total	Muestra	Suministro	Periodo Pronosticado	Tiempo ARIMA	Tiempo HW	Tiempo NNETAR	Tiempo SVM
Sucursales	Sucursales 07	18320	11	25160096	201909	0.27	0.11	0.42	0.06
Sucursales	Sucursales 07	18320	11	25160130	201909	0.21	0.11	0.63	0.06
Sucursales	Sucursales 07	18320	11	25608039	201909	0.34	0.14	0.34	0.04
Sucursales	Sucursales 07	18320	11	25608146	201909	0.19	0.12	0.6	0.06
Sucursales	Sucursales 07	18320	11	25608164	201909	0.2	0.13	0.27	0.05
Sucursales	Sucursales 08	12887	5	25758671	201909	0.17	0.11	0.63	0.05
Sucursales	Sucursales 08	12887	5	25760428	201909	0.21	0.13	0.51	0.07
Sucursales	Sucursales 08	12887	5	25760437	201909	0.22	0.16	0.26	0.04
Sucursales	Sucursales 08	12887	5	25777247	201909	0.2	0.13	0.25	0.05
Sucursales	Sucursales 08	12887	5	25817567	201909	0.23	0.11	0.27	0.06
Sucursales	Sucursales 09	5939	1	25607363	201909	0.33	0.12	0.25	0.06
Sucursales	Sucursales 10	8520	3	25605681	201909	0.21	0.12	0.36	0.08
Sucursales	Sucursales 10	8520	3	25612120	201909	0.12	0.1	0.41	0.06
Sucursales	Sucursales 10	8520	3	25612130	201909	0.37	0.12	0.25	0.06
Sucursales	Sucursales 11	4454	1	25160022	201909	0.25	0.13	0.36	0.09
Sucursales	Sucursales 14	8843	3	26414755	201909	0.21	0.35	0.29	0.06
Sucursales	Sucursales 14	8843	3	26925809	201909	0.19	0.14	0.56	0.05
Sucursales	Sucursales 14	8843	3	26925818	201909	0.24	0.12	0.22	0.04
Sucursales	Sucursales12	6208	1	25685637	201909	0.21	0.11	0.47	0.06
Sucursales	Sucursales13	6435	1	25685655	201909	0.24	0.16	0.25	0.05

Fuente: Elaborado por el autor



Tabla 11
Indicador II – Tiempo de procesamiento Global por Algoritmos

	Medida	ARIMA	HW	NNETAR	SVM
1	TIEMPO PROMEDIO	0.227168142	0.137433628	0.433539823	0.064247788
2	TIEMPO TOTAL	25.67	15.53	48.99	7.26

Fuente: Elaborado por el autor

4.2. Discusión de resultados

Para el primer indicador de medición, orientado a la medición de la correcta estimación de consumo, se obtuvo un desempeño ACC (Accuracy) para Arima de 84.05 % dado que el error MAPE resultó en 15.94%. Para el caso de HoltWinter el ACC es de 66.61 y el MAPE obtenido es de 33.38 % %, siendo el peor modelo según los resultados, en el caso de la Red Neuronal Autoregresiva NNETAR el ACC obtiene un 88.73 % y un MAPE de 11.26 % el cual tiene el mejor desempeño de los modelos, para el caso del SVM el valor del ACC es 81.99% y un MAPE de 18.00 %, ordenando estos algoritmos por su desempeño obtenemos a la RED NEURONAL AUTOREGRESIVA, ARIMA, SVM y HOLTWINTERS. Según nuestro cuadro de evaluación de la métrica de estimación Tabla 55, el único modelo que podría entrar a la categoría de BUENO es la RED NEURONAL AUTOREGRESIVA. Este resultado es interesante, dado que HOLTWINTERS y ARIMA esquemas convencionales tienen inconvenientes según la naturaleza de la serie de tiempo y es de difícil optimización en los cálculos que se realiza en cuanto a sus coeficientes, en el caso de Holtwinters se observa buen detalle en cuanto a series con componentes estacionales, sin embargo para componentes cíclicos o estacionarios ARIMA representa una gran ventaja, es por esto que los algoritmos de Redes Neuronales y SVM comprenden un mejor desempeño.

Y es aquí que el segundo indicador entra en detalle, extrayendo información real de todos los refactorados con concepto de error de lectura, es decir que se verificó que la lectura de campo era errónea, a este grupo de suministros se les debe comparar donde existe similitud o igualdad en los resultados de los suministros inconsistentes del modelo, porque si un suministro es declarado error de lectura, el modelo debe detectarlo como inconsistente. Obteniendo finalmente un total del 62.77 % de suministros que denotan error de lectura. Estableciendo la cantidad de inconsistencias, y las validas que obtiene el modelo, el tercer indicador hace referencia al tiempo ejercido para resolver dicha inconsistencia, obteniendo finalmente un total del 77 % de reducción en el tiempo que se empleaba sin los datos proporcionados por el modelo, al proceso actual.



CAPITULO V: DESARROLLO DE LA PROPUESTA

5.1. Generalidades de la propuesta

5.1.1. Marco General de Trabajo

Para el desarrollo de ésta investigación, se ha esquematizado la siguiente estrategia metodológica de trabajo que permite lograr los objetivos específicos planteados y por ende el objetivo general de la investigación.

La investigación está dividida en dos etapas, la primera abarca la generación del modelo predictivo, la cual realiza el análisis de la realidad del negocio, análisis de datos, generación del modelo, implementación de los algoritmos a comparar y por último, la generación de resultados. En la segunda etapa, se describe el proceso de desarrollo de la solución que permite visualizar los resultados, el acoplamiento entre los datos originales, el modelo y resultados y la dinámica web del sistema.

Para generar el modelo predictivo debe considerarse una fase previa de extracción de información del sistema transaccional de la empresa, sin embargo esta fase es omitida, ya que ELECTRONORTE S.A. cuenta con un datawarehouse de Facturación, provisto como backup, el cual contiene la información vital para la investigación.

La estrategia planteada se visualiza en el siguiente diagrama:

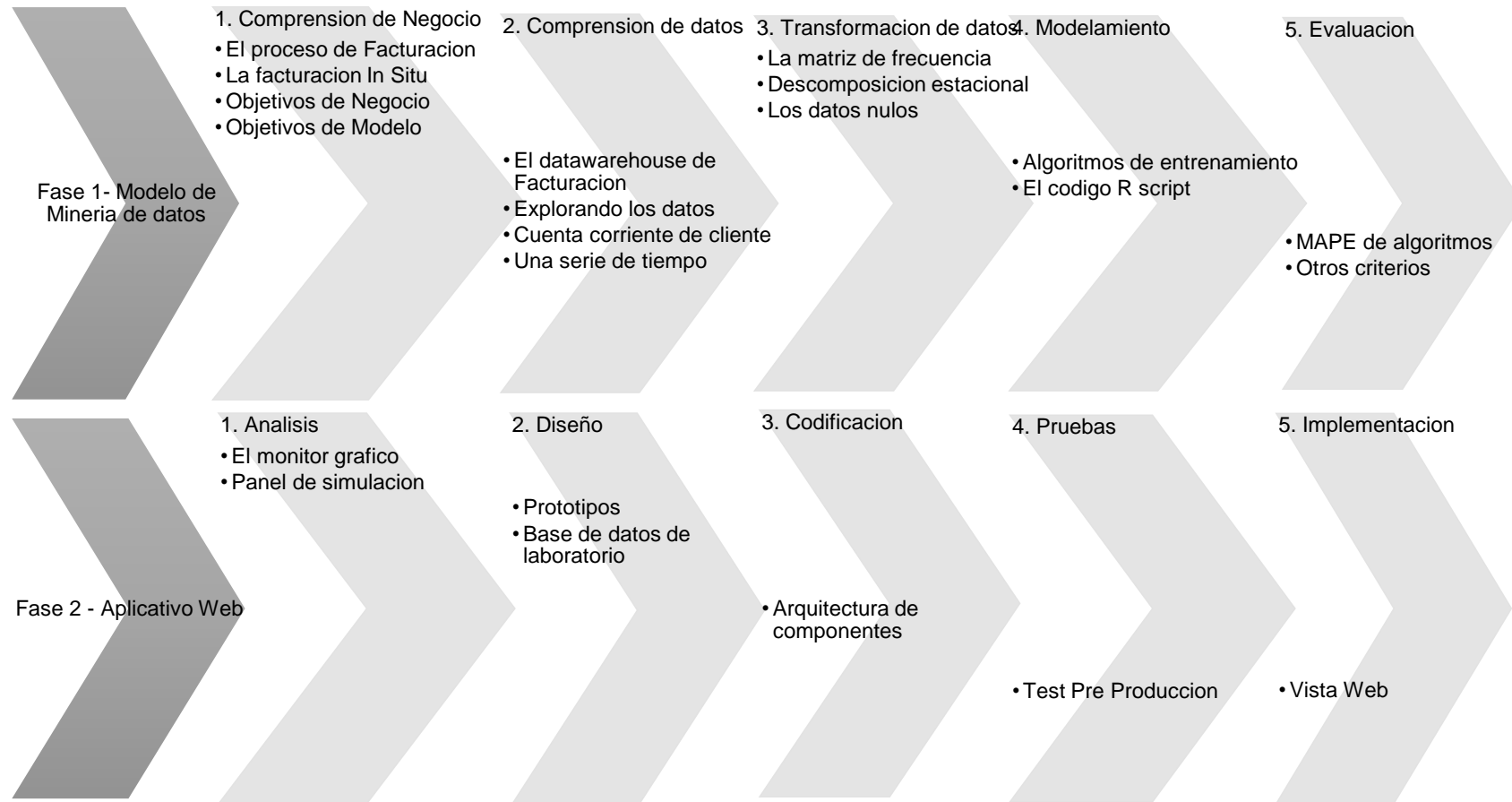


Figura 7. Método de desarrollo de la investigación



5.1.2. Desarrollo de la propuesta

FASE 1: MINERIA DE DATOS – METODOLOGIA CRISP DM

A. Comprensión de Negocio

a. El proceso de facturación

El modelo actual de Facturación que se aplica en ENSA, se basa en un esquema empírico propio de las empresas de servicios públicos del sector, es decir sigue un flujo, el mismo que se observa en la figura siguiente.

Debe considerarse que la figura 9 muestra el proceso de facturación estructurado en pasos secuenciales, desde la generación de los parámetros que sirven para el cálculo de los consumos, hasta la ejecución de la lectura de medidor y entrega póstuma del recibo.

Esta secuencia de pasos es única y estándar para todo proceso de facturación, su ejecución es similar en cualquier empresa de servicios públicos de energía eléctrica. Sin embargo, con éste proyecto de investigación se plantea evaluar el flujograma de este proceso y como se resuelve (Tiempo y Costos).

El análisis del flujograma actual del proceso de facturación genera el siguiente enunciado:

Se han identificado al menos 06 actores o roles que participan como elementos del proceso de facturación, como es el caso del facturador, supervisor, lectorista, imprenta, compaginación y repartidor de recibos; estos actores se vinculan en todos los subprocesos de la facturación donde se presenta la siguiente problemática:

1. Manejo de pliegos y parámetros de facturación de manera manual y propensa a errores.
2. Limitada supervisión a los procesos de lectura y reparto ejecutados por las empresas terciarizadas.
3. Proceso de consistencia de datos no automatizado.
4. Generación de errores por la estimación de consumos en los casos de suministros sin lectura.
5. Problemas operativos con la impresión de los recibos.
6. Esquema no óptimo donde actividades requieren un doble trabajo del personal (Lectura de medidor y realizar reparto de recibo)
7. En el esquema actual de entrega posterior de los recibos, el cliente no cuenta con el plazo de 15 días para realizar el pago.
8. Elevados índices de reclamos por errores en la facturación.
9. Generación de ciclos repetitivos de visita a campo (Reparto de recibo).

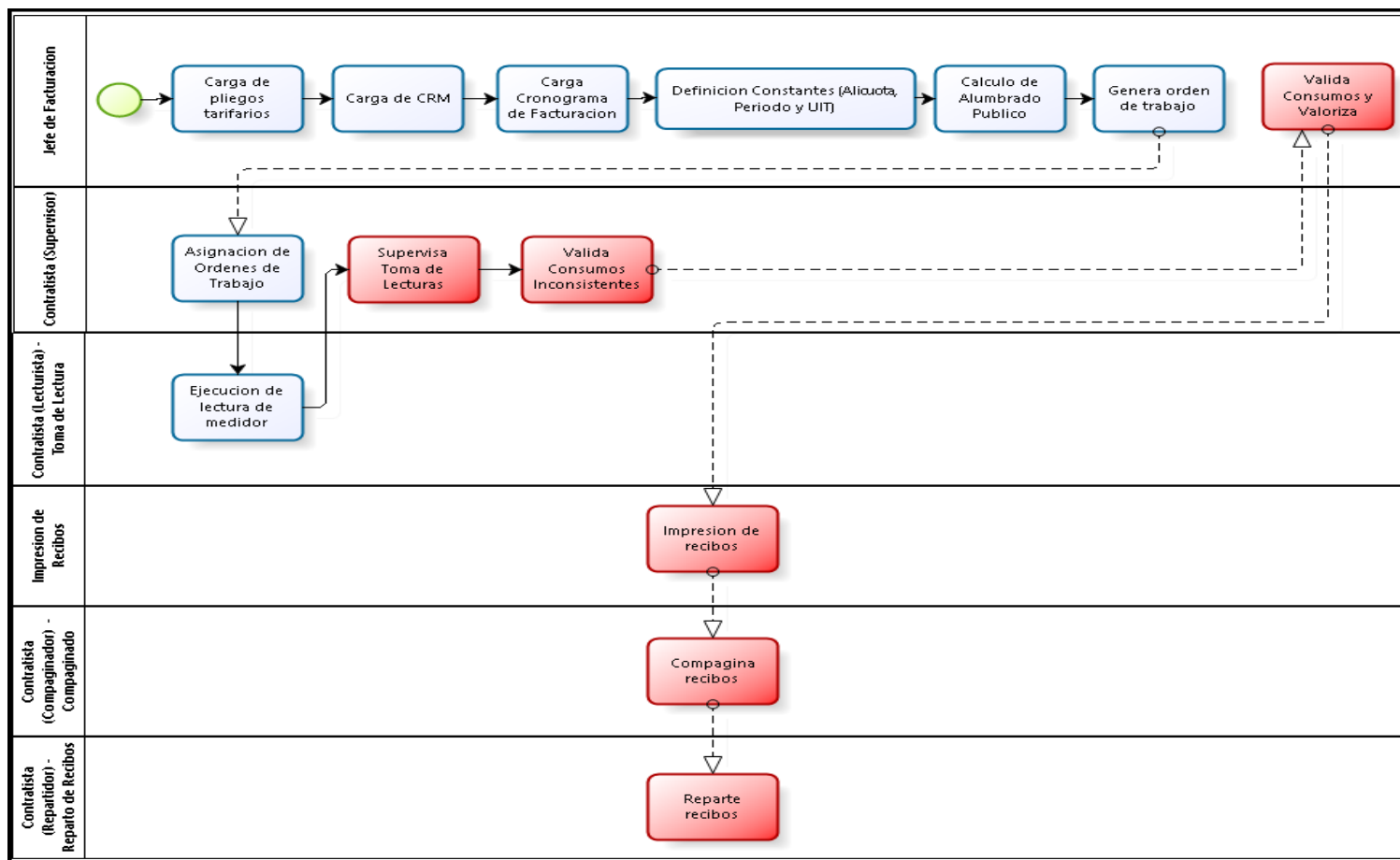


Figura 8. Flujograma actual del proceso de facturación



b. La facturación InSitu

Se propone a la Facturación In-Situ como un modelo de facturación inmediata, empleando tecnologías de fácil acceso y bajo costo. Esta metodología proporcionaría la oportunidad de integrar en una sola actividad los procesos de toma de lectura, análisis de las inconsistencias, facturación, impresión de recibo y reparto del recibo. De esta manera se reducen los tiempos de ejecución y costos operativos de los procesos, elevando la calidad del servicio.

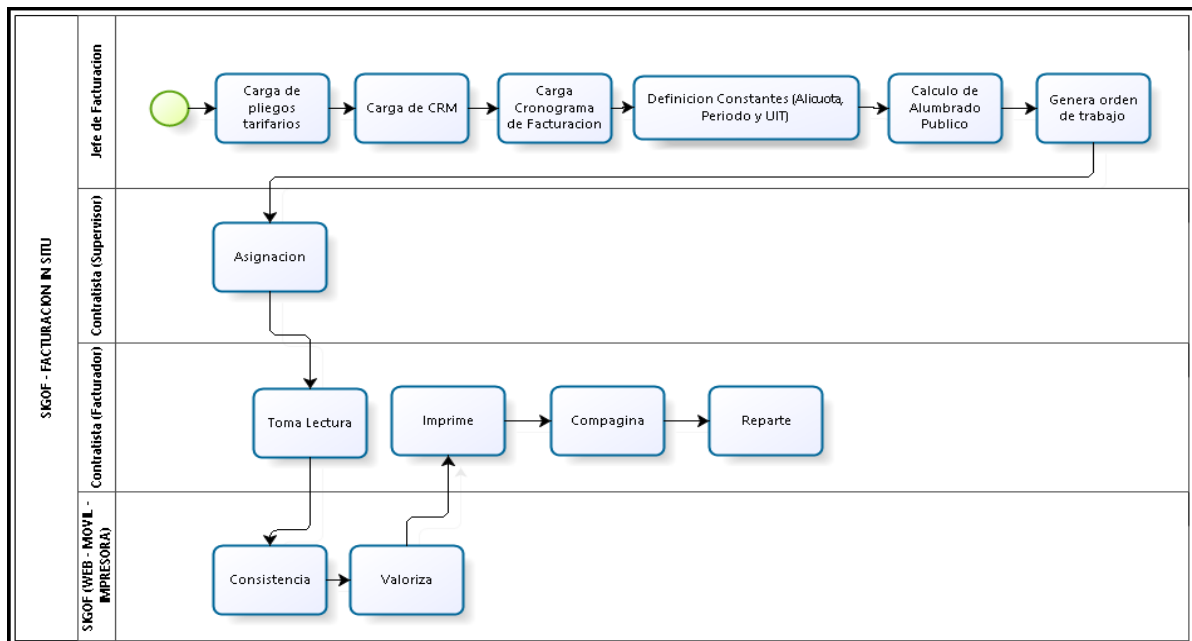


Figura 9. Modelo propuesto de trabajo

c. Objetivos de Negocio

A partir de lo expuesto se presenta la problemática actual en ELECTRONORTE S.A., la cual hace referencia a que se requiere un modelo de trabajo IN-SITU, donde debe considerarse el uso de un aplicativo informático para los casos en los que no se puede capturar la lectura; estos casos se pueden dar por diversos motivos,



uno de los cuales podría ser la existencia de medidores interiores que imposibilitan que el personal capture la lectura de medidor,

En primera instancia se podría resolver este problema emitiendo el promedio en función a los meses anteriores del medidor para que se imprima el recibo, sin embargo, este método pone en riesgo la calidad de facturación, lo que podría generar incremento en los reclamos por montos de facturación mal calculados.

Por lo tanto, el objetivo de Negocio principal es:

Generar las estimaciones de consumo para un determinado periodo comercial en el proceso de facturación.

d. Objetivos de Modelo

El modelo por lo tanto requiere:

- Diseñar un modelo de minería de datos que permita obtener las estimaciones de consumo de energía eléctrica de los clientes para un determinado periodo comercial.
- Comparar que algoritmo tiene un mayor grado de precisión y mínimo margen de error.

B. Comprensión de los datos

La empresa de energía eléctrica cuenta con un Datawarehouse donde se procesan reportes pertinentes con respecto a la facturación, la base de datos está bajo el gestor SQL SERVER 2014 alojado en servidores IBM.



Por razones de seguridad y confidencialidad de los datos y del DATAWAREHOUSE se realizó una extracción de datos a nivel ficheros CSV y RDS, que comprende parte del esquema del DWH orientado a Facturación, se enmascaró los campos.

```

1  setwd("C:\\Program Files (x86)\\Zend\\Apache2\\htdocs\\gcelab\\script\\dataset")
2
3  library(RODBC)
4
5
6  cnx <- odbcDriverConnect('driver={SQL Server};server=localhost;database=DWH_Distribucion')
7
8
9  dataset1<- sqlQuery(cnx,"select * from NROSERVICIO")
10 saveRDS(dataset1,file="DIM_SUMINISTRO")
11
12 dataset2<- sqlQuery(cnx,"select * from OBSERVACION")
13 saveRDS(dataset2,file="DIM_OBSERVACION")
14
15 dataset3<- sqlQuery(cnx,"select * from ADMINISTRATIVO")
16 saveRDS(dataset3,file="DIM_ADMINISTRATIVO")
17
18 dataset4<- sqlQuery(cnx,"select * from MEDIDOR")
19 saveRDS(dataset4,file="DIM_MEDIDOR")
20
21 dataset5<- sqlQuery(cnx,"select * from PERIODOFACTURACION")
22 saveRDS(dataset5,file="DIM_PERIODOFACTURACION")
23
24 dataset6<- sqlQuery(cnx,"select * from ConsumoHist")
25 saveRDS(dataset6,file="PIVOT_HECHOCONSUMO")
26
27
28 res_set1<- sqlQuery(cnx,"select * from DetalleModelo2")
29 saveRDS(res_set1,file="RES_MODEL_EVA")
30
31 res_set2<- sqlQuery(cnx,"select * from LaboratorioResultado")
32 saveRDS(res_set2,file="RES_MODEL_PER")
33
34
35 group_per<- sqlQuery(cnx,"SELECT idPeriodoFacturacion FROM FACT_VENTASERVICIOCONSUMO
36 GROUP BY idPeriodoFacturacion ORDER BY idPeriodoFacturacion")
37
38 nrow(group_per)
39
40 for (i in 1:nrow(group_per)) {
41   per<-group_per[i,1]
42   query<-paste("SELECT * FROM FACT_VENTASERVICIOCONSUMOS where idPeriodoFacturacion=",per,sep="")
43   fac_ven<-sqlQuery(cnx,query)
44   file_g<-paste("FACT_CONSUMO_",per,sep="")
45   saveRDS(fac_ven,file=file_g)
46 }
47
48 readRDS(file="DIM_OBSERVACION")

```

Figura 10. Extracción de datos a nivel ficheros CSV y RDS

HechoConsumo: Tabla con el hecho del proceso de toma de lecturas, consumos calculados; de esta tabla se extrae el vector histórico de consumos.



```

> Hecho_Consumo<-read.csv("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\
FACT_CONSUMO_CSV\\FACT_CONSUMO_201801.csv"
,nrows=1000)
tipo<-sapply(Hecho_Consumo, class)
as.data.frame(tipo)
+ > >
X integer tipo
id integer
suministro integer
seriefab factor
pfactura integer
tipolectura factor
cliente factor
foto1 integer
latitud1 numeric
longitud1 numeric
fechaejecucion1 factor
montoconsumo1 numeric
lectura01 integer
glomas_unidadnegocio_id integer
resultadoevaluacion factor
idciclo integer
nombruta factor
direccion factor
pinicio integer
idempresa integer
idproveedor integer
app_last_tipo factor
sector integer
ruta integer
obs1 integer
> |

```

Figura 11. Tabla HechoConsumo

Dimensión PeriodoFacturacion: es la dimensión que define el periodo de facturación del registro, su frecuencia es mensual, además considera los valores de la fecha de toma de lectura.

```

> Dim_Periodo<-readRDS("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\DIM_PERIODO")
tipo<-sapply(Dim_Periodo, class)
as.data.frame(tipo)
> >
pfactura character
>

```

Figura 12. Tabla Dimensión PeriodoFacturacion

Dimensión Administrativo: es la dimensión con información de la ubicación del suministro, jerarquizado por la estructura ubigeo de Ensa, denotado por Empresa, Unidad de Negocio, Ciclo. Donde una empresa contiene 1 o varias Unidades de Negocio, una Unidad de Negocio contiene 1 o varios Ciclos de Facturación.



```
> Dim_Administrativo<-read.csv("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa
\\ADMINISTRATIVO\\ADMINISTRATIVO.csv"
,nrows=1000)
tipo<-sapply(Dim_Administrativo, class)
as.data.frame(tipo)
+ > >
X integer tipo
id integer
suministro integer
idciclo integer
glomas_unidadnegocio_id integer
idempleado integer
> |
```

Figura 13. Tabla Dimensión Administrativo

Dimensión Empresa: es la dimensión que define la empresa, en este caso ENSA pertenece al GRUPO DISTRILUZ, que también contempla las empresas del mismo rubro que son HIDRANDINA, ENOSA y ELECTROCENTRO.

```
> Dim_Empresa<-readRDS("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\
DIM_EMPRESA")
tipo<-sapply(Dim_Empresa, class)
as.data.frame(tipo)
+ > idempresa nombreempresa abreviaempresa nombrecompleto codigoidentificacion
1 integer character character character character
2 integer character character character character
| direccion iddireccioncodificada direccioncomplementaria idorganizacion
1 character integer character integer
2 character integer character integer
| ruc empidcodigo_sap idtipoidentidadrepresentante
1 character character integer
2 character character integer
| nroidentidadrepresentante cargorepresentante estado oslastlogin oslastdate
1 character character character integer character POSIXct
2 character character character integer character POSIXt
| oslastapp osfirstlogin osfirstdate osfirstapp codigocentralriesgo
1 character character POSIXct character character
2 character character POSIXt character character
| codigoosinergmin logo representante representante_firma email
1 character character character character character character
2 character character character character character character
| web telefono serviluz
1 character character character
2 character character character
> |
```

Figura 14. Tabla Dimensión Empresa

Dimensión Unidad Negocio: Una empresa puede tener 1 o varias unidades de negocio.



```
> Dim_UUNN<-readRDS("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\
DIM_UNIDADNEGOCIO")
tipo<-sapply(Dim_UUNN, class)
as.data.frame(tipo)
> >
          tipo
iduunn      integer
idempresa   integer
nombreunidadnegocio character
abreviaunidadnegocio character
iddireccioncodificada integer
direccion   character
direccioncomplementaria character
unidcodigo_sap character
essede      integer
estado      integer
telefonocentral character
> |
```

Figura 15. Tabla Dimensión PeriodoFacturacion

Dimensión Unidad Ciclo: Una unidad de negocio puede tener 1 o varios ciclos, los ciclos son divisiones geográficas expresadas por el calendario de facturación, por ejemplo, el CICLO 1 es un distrito político de una ciudad y su calendario se ejecuta los días 20 de cada mes.

```
> Dim_Ciclo<-readRDS("H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\
DIM_CICLO_2")
tipo<-sapply(Dim_Ciclo, class)
as.data.frame(tipo)
> >
          tipo
idciclo      character
nombreciclo  character
glommas_unidadnegocio_id integer
idempresa    integer
> |
```

Figura 16. Tabla Dimensión PeriodoFacturacion

C. Transformación de los datos

Al explorar los datos en la tabla Hecho, contiene un promedio de 54 000 000 millones de registros hacia abajo por cada suministro en todos sus periodos de facturación desde el periodo 201605 (mayo 2016) hasta el 201909 (setiembre 2019).



Al explorar por SQL el dataset del DWH en la tabla hecho se encuentra el siguiente esquema:

	idPartition	idPeriodoFacturacion	idAdministrativo	idNroServicio	idMedidor	idOrdenTrabajo	idObservacion	idContratoEnergia	idEmision	idVencimiento	idFechaLecturaAnterior	idFechaLecturaActual
1	3201001	8432	266	45000803	249349	0	0	27	8436	8456	8402	8433
2	3201001	8432	266	45000812	245173	0	0	27	8436	8456	8402	8433
3	3201001	8432	266	45000821	244669	0	0	27	8436	8456	8402	8433
4	3201001	8432	266	45000830	245174	0	0	27	8436	8456	8402	8433
5	3201001	8432	266	45000840	805226	0	0	27	8436	8456	8402	8433
6	3201001	8432	266	45000859	244769	0	0	27	8436	8456	8402	8433
7	3201001	8432	266	45000868	840563	0	0	27	8436	8456	8402	8433
8	3201001	8432	266	45000877	247393	0	0	27	8436	8456	8402	8433
9	3201001	8432	266	45000886	245096	0	0	27	8436	8456	8402	8433
10	3201001	8432	266	45000895	841076	0	0	27	8436	8456	8402	8433
11	3201001	8432	266	45000901	245175	0	0	27	8436	8456	8402	8433
12	3201001	8432	266	45000910	244670	0	0	27	8436	8456	8402	8433
13	3201001	8432	266	45000920	840597	0	0	27	8436	8456	8402	8433
14	3201001	8432	266	45000939	840598	0	0	27	8436	8456	8402	8433
15	3201001	8432	266	45000948	846558	0	0	27	8436	8456	8402	8433
16	3201001	8432	266	45000957	247840	0	0	27	8436	8456	8402	8433
17	3201001	8432	266	45000966	843640	0	0	27	8436	8456	8402	8433
18	3201001	8432	266	45000984	245282	0	0	27	8436	8456	8402	8433
19	3201001	8432	266	45000992	245392	0	0	27	8436	8456	8402	8433

Figura 17. Tabla hecho

En este punto se procede a preparar los datos:

Tabla 12
Preparación de datos

Años/ Meses	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
2011	-	-	290	277	297	292	293	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								

Fuente: Elaborado por el autor

Utilizando el software para modelos analíticos, R Project se realiza el script siguiente de conexión:



```

> cvreal <- sqlQuery(cnx,vreal)
>
>
> atomize <- ts(cvreal,frequency=12,start=c(2010,01))
> atomize
      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2010  28  22  27  61  84  57  40  46  57  72  56  32
2011  27  31  41  29  27  29  33  31  25  23  21  21
2012  14  21  27  22  29  26  23  20  25  22  23  22
2013  21  22  33  24  25  20  16  37  21  23  34  30
2014  29  34  32  29  31  31  22  20  24  24  28  28
2015  17  11   9  12
> |

```

Figura 18. Formato de análisis series de tiempo

Para ello se diseñan matrices de resumen y condensación de datos especializadas para que el modelo puede procesarlas sin inconveniente. Los motivos por el cual se crean las matrices analíticas se deben a:

Gestión de recursos informáticos

Uno de los problemas habituales al tratar bases de datos históricas (Datamart o Datawarehouse), es el tratamiento a la gran cantidad de información con la que se trabaja, por más que los datos numéricos optimicen el recorrido en una tabla “Hecho”.

A continuación, se presenta una imagen con el tiempo obtenido de procesamiento de una tabla Hecho del datawarehouse y una segunda imagen con el tiempo procesado en una tabla matriz consolidada.



	idPartition	idPeriodoFacturacion	idAdministrativo	idNro Servicio	idMedidor	idOrden Trabajo	idObs
1	3201001	8432	266	45000803	249349	0	0
2	3201001	8432	266	45000812	245173	0	0
3	3201001	8432	266	45000821	244669	0	0
4	3201001	8432	266	45000830	245174	0	0
5	3201001	8432	266	45000840	805226	0	0
6	3201001	8432	266	45000859	244769	0	0

Figura 19. Tiempo obtenido de procesamiento de una tabla Hecho

El tiempo para procesar la tabla Hecho es de 9 minutos para 40 601877 registros.

	idNroServicio	201001	201002	201003	201004	201005	201006	201007	201008	201009
1	45393729	76	68	77	72	77	77	72	80	78
2	45314402	0	2	1	2	2	3	3	3	2
3	45325809	5	10	22	16	28	39	73	72	64
4	46834327	14	13	15	13	14	13	15	15	17
5	46760573	127	76	91	81	71	78	73	83	78
6	48280498	38	38	50	34	34	34	42	31	41

Figura 20. Matriz de suministros por periodo y consumo

Sin embargo, una tabla matriz de suministros por periodo y consumo tarda 21 segundos para 743762 (Transformación de los 40 millones de registros de la tabla Hecho).

Tratamiento de nulos (Directo)

Estos casos particulares pueden darse por las siguientes causas:



- Resultó imposible tomar la lectura del medidor.
En el nulo directo, aun así, cuando en el proceso de lecturas no se ha podido extraer el monto de lectura del suministro, esta pasa con un código de observación a la empresa de energía eléctrica, y es el facturador quien se encarga de estimar un consumo, ya sea por promedios u otros criterios en función de este.

	idPeriodoFacturacion	idNroServicio	idObservacion
1	8432	45002399	48
2	8432	45010864	3
3	8432	45018942	45
4	8432	45034098	43
5	8432	45072366	43
6	8432	45078280	3
7	8432	45090506	3
8	8432	45173654	45
9	8432	45176549	45
10	8432	45176861	45
11	8432	45178409	5

Figura 21. Tratamiento de nulos (Directo)

Tratamiento de Nulos (Indirecto)

No siempre los datos tendrán una salida óptima, es decir aun con el Datamart se tienen inconsistencias en los datos como las descritas a continuación:

- Meses en los que no se generó ningún registró por corte de servicio.
- Suministros nuevos

En la siguiente imagen se describe un nulo Indirecto u oculto.



14	8825	3201102	0
15	8856	3201103	22
16	8886	3201104	27
17	8917	3201105	33
18	8947	3201106	31
19	8978	3201107	23
20	9009	3201108	16
21	9039	3201109	8
22	9070	3201110	39
23	9131	3201112	13
24	9162	3201201	21
25	9191	3201202	19
26	9222	3201203	4
27	9252	3201204	5

Consulta ejecutada correctamente.

Figura 22. Tratamiento de Nulos (Indirecto)

Nótese en los registros que no hay existencia del periodo 201111, esto puede darse debido a un corte, por el cual el suministro queda excepto al proceso de toma de lecturas.

Para identificar estos nulos, se debe crear una matriz que describa los periodos como columnas y los suministros como filas, en la intersección se debe especificar el detalle de la ocurrencia (El registro consumo si es que existe registro para ese periodo, o un Nulo)

Esta matriz genera un índice para el periodo correlativo, y agrega el valor de consumo detectado, en el caso de no encontrar ningún registro (Ni siquiera Null), agrega un Null.



```

> pivot <- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\PIVOT\\PIVOT")
head(pivot)
> suministro 201603 201604 201605 201606 201607 201608 201609 201610 201611
1 5000095 NA 235 245 190 180 184 192 194 198
2 5000110 NA 387 404 394 330 332 354 348 355
3 5000139 NA 478 253 229 214 143 151 156 178
4 5000184 NA NA NA NA 115 119 120 121 124
5 5000460 NA NA NA 40 19 12 14 14 21
6 5000550 NA 98 73 48 41 43 37 38 43
.....
201612 201701 201702 201703 201704 201705 201706 201707 201708 201709 201710
1 210 224 201 226 188 180 170 164 182 181 190
2 411 512 418 433 380 413 356 349 336 352 349
3 179 249 230 251 189 247 240 198 209 205 251
4 139 134 132 NA 169 154 140 132 128 126 123
5 27 34 31 21 29 24 18 27 27 22 13
6 49 71 77 91 93 85 91 81 61 55 57
.....
201711 201712 201801 201802 201803 201804 201805 201806 201807 201809 201810
1 197 204 191 186 243 212 NA 213 NA NA NA
2 346 454 486 418 446 422 NA 381 NA NA NA
3 239 272 229 253 396 411 NA 164 NA NA NA
4 120 137 135 NA 130 150 NA 127 NA NA NA
5 18 22 21 22 23 25 NA 29 NA NA NA
6 50 64 86 73 87 89 NA 100 NA NA NA
.....
201811 201812 201901 201902 201903 201904 201905 201906 201907 201908 201909
1 NA NA 260 232 271 262 345 251 NA NA NA
2 NA NA 465 402 468 491 619 427 NA NA NA
3 NA NA 154 185 206 199 257 171 NA NA NA
4 NA NA 193 175 155 145 384 239 NA NA NA
5 NA NA 22 19 20 17 26 17 NA NA NA
6 NA NA 101 99 73 80 113 82 NA NA NA
> |

```

Figura 23. Agregación del valor NULL

Por lo tanto, se estará pasando de una tabla de 56 000 000 millones de registros con 39 periodos para suministros como registros, a una tabla resumen de 2 000 000 registros con 39 columnas pertenecientes a todos los periodos, esta matriz permite agilizar el tiempo de consulta, asimismo como un mejor panorama de análisis de datos.

D. Modelamiento

Se propone la construcción de un modelo de minería de datos utilizando técnicas de pronósticos.

A continuación, se presenta la tabla que contiene el proceso de comparación que se realizó para seleccionar técnicas más adecuadas.



Tabla 13
Selección de las técnicas a utilizarse

NOMBRE DE TÉCNICA	DESCRIPCIÓN	ALGORITMOS	¿ES ADECUADO PARA LA INVESTIGACIÓN?
REGRESIÓN		Redes Neuronales, SVM	SI
SERIES TEMPORALES		Holtwinters, ARIMA, entre otras	SI
CLASIFICACION AD HOC	Basado en reglas por construcciones lógicas múltiples variables	Árbol de decisiones, Redes Bayesianas	NO

Fuente: Elaborado por el autor

Para esto, se le asignó a cada algoritmo criterios de evaluación, los mismos que son detallados a continuación:



Tabla 14
Criterios de evaluación de los algoritmos seleccionados

MODELO DE MINERÍA DE DATOS PARA LA PREDICCIÓN				
	REDES			
	HOLTWINTER	ARIMA	NEURONALES	SVM
S				
Fundamento teórico				
Modelo parametrizado	X	X		
Datos estacionales	X	X	X	
Método estadístico	X	X		
Capacidad iterativa (Aprendizaje)			X	X
Cantidad de datos de la serie	24	40	3	
Fundamento computacional				
Procesamiento CPU	Mínimo		Medio	
Consumo RAM	Mínimo		Medio	
Tiempo computacional	Mínimo		Medio	
Fundamento objetivo del modelo				
Confiabilidad de precisión pronostico	Después de pruebas		Después de pruebas	
Confiabilidad de precisión consistencias	Después de pruebas		Después de pruebas	

Fuente: Elaborado por el autor

El procedimiento se inicia con la extracción de los datos transformados en la tabla matriz:



```

216 # imprime Tabla x5 Analisis de Forecast con Algoritmos TAB - PANTALLA GO LIVE
217 observeEvent(input$x3_rows_selected,{
218   output$title1 = renderText({HTML("Pronóstico Generado con multiples algoritmos")})
219   output$x5 = DT::renderDataTable(selection = 'single',{
220
221     dum_ciclo <- input$thirdselection
222     dum_ciclo <- paste(input$thirdselection, collapse = '\',\')
223
224
225     table1_query <-paste("select da.suministro,dc.nombreciclo,du.nombreunidadnegocio,de.nombre
226     from d_administrativo da
227     join d_ciclo dc on da.idciclo = dc.idciclo
228     join d_uunn du on dc.glomas_unidadnegocio_id = du.iduunn
229     join d_empresa de on du.idempresa = de.idempresa
230     where dc.nombreciclo in ('",dum_ciclo,"') order by da.suministro",sep="")
231     table1 <- sqlDF(table1_query)
232
233
234     s1<-input$x3_rows_selected
235
236     s = table1[s1,1]
237
238     if (length(s)) {
239       #df <- pivot[s,]
240       df <- pivot %>%
241       | filter(suministro %in% s)
242
243
244       dx <- df
245       df <- subset( df, select = -c(1,2) )
246       #df[is.na(df)] <- 0
247       df <- t(df)
248
249       perpronosticado<-"201909"
250
251       for (i in 1:ncol(df)) {
252         sum<-dx[i,1]
253         dfp<-df[,i]
254
255         if(all(is.na(dfp))==FALSE){
256
257
258           #Aplicando la media para rellenar valores faltantes
259           dfp[is.na(dfp)] <- round(mean(dfp, na.rm = TRUE),0)
260
261           last<-tail(dfp, 1)
262           con_real<-last[1]

```

Figura 24. Extracción de los datos

El algoritmo procesa y extrae los suministros a ser evaluados, luego por cada suministro extra su información histórica y lo transforma al formato de frecuencias para inicializar el entrenamiento para los siguientes algoritmos.

Modelo A - HOLTWINTERS

Descripción del Modelo A

En el modelo A, se utiliza la técnica de Holt-Winters, que abarca ecuaciones de pronóstico y suavizado, tanto para el nivel tendencia, el estacional y el último para la variación cíclica.

Hay dos variaciones en esta técnica, que nacen como producto de la oscilación del componente estacional. Se emplea el método aditivo, si las variaciones estacionales mantienen cierta constancia a través de la



serie; aquí el componente estacional se expresa en términos absolutos y en la ecuación de nivel de la serie se ajusta estacionalmente restando el componente estacional. Mientras que, se opta por el método multiplicativo, cuando las variaciones estacionales presentan cambios muy notables en la serie. Con este método el componente estacional se expresa en términos relativos (porcentajes) y la serie se ajusta estacionalmente.

Para el estudio se emplea el componente aditivo en las fórmulas Holtwinters que a continuación se describen:

Fórmulas para Alpha, Beta y Gamma

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

A continuación, se visualiza el código del algoritmo Holtwinters en R:




```

21 HoltWinters <-
22 function (x,
23
24     # smoothing parameters
25     alpha = NULL, # level
26     beta  = NULL, # trend
27     gamma = NULL, # seasonal component
28     seasonal = c("additive", "multiplicative"),
29     start.periods = 2,
30
31     # starting values
32     l.start = NULL, # level
33     b.start = NULL, # trend
34     s.start = NULL, # seasonal components vector of length `period`
35
36     # starting values for optim
37     optim.start = c(alpha = 0.3, beta = 0.1, gamma = 0.1),
38     optim.control = list()
39 )
40 {
41   x <- as.ts(x)
42   seasonal <- match.arg(seasonal)
43   f <- frequency(x)
44
45   if(!is.null(alpha) && (alpha == 0))
46     stop ("cannot fit models without level ('alpha' must not be 0 or FALSE)")
47   if(!all(is.null(c(alpha, beta, gamma))) &&
48       any(c(alpha, beta, gamma) < 0 || c(alpha, beta, gamma) > 1))
49     stop ("'alpha', 'beta' and 'gamma' must be within the unit interval")
50   if((is.null(gamma) || gamma > 0)) {
51     if (seasonal == "multiplicative" && any(x == 0))
52       stop ("data must be non-zero for multiplicative Holt-Winters")
53     if (start.periods < 2)
54       stop ("need at least 2 periods to compute seasonal start values")
55   }
56 }

```

Figura 25. Algoritmo en R – Holtwinters Biblioteca Forecast R

Para el análisis, la frecuencia de distribución es de 12, por los meses de cada año; tal y como se estableció en la matriz de ingreso en la fase de Preparación de los Datos.

Evaluación del Modelo A

A continuación, se visualiza el código del modelo con Holtwinters



```
##### PRONOSTICANDO CON HOLTWINTERS
fre<- ts(dfp, frequency=12,start=c(2016,04))
aplicahw <- HoltWinters(fre)
pronosticohw<- forecast(aplicahw, h=1)
pro_hw<-round(as.numeric(pronosticohw[4]), digits = 0)
```

Figura 26. Código del modelo con Holtwinters

En R se trabaja con el histórico del suministro; el modelo realiza el entrenamiento de la serie, y determina los valores que formarán parte de ésta de forma automática; éstos son: el Alpha (variación), Beta (tendencia) y Gamma (estacionalidad). Su valor oscila entre 0 a 1. Con los valores obtenidos, se procede a extraer los pronosticos requeridos; interesando, en este caso, el próximo valor futuro, calculándose también, la mínima y máxima como rango esperado.

Tabla 15
Resultados obtenidos con HoltWinters

UUNN	Ciclo	Mues	Suministro	Periodo	Real	HW	Error HW
Cajamarca Centro	I - Cajamarca Centro	1	28361839	201909	20	20	0
Cajamarca Centro	II - Cajamarca Centro	1	28217192	201909	68	65	4.41
Cajamarca Centro	III - Cajamarca Centro	1	28179959	201909	184	246	33.7
Cajamarca Centro	IV - Cajamarca Centro	1	25610289	201909	25	24	4
Cajamarca Centro	V - Cajamarca Centro	3	25602080	201909	16	7	56.25
Cajamarca Centro	V - Cajamarca Centro	3	25602563	201909	165	173	4.85
Cajamarca Centro	V - Cajamarca Centro	3	25602581	201909	43	48	11.63
Chiclayo	Chiclayo 0	1	25778262	201909	87	42	51.72
Chiclayo	Chiclayo 01	2	25602302	201909	6454	7620	18.07
Chiclayo	Chiclayo 01	2	25602320	201909	113	132	16.81
Chiclayo	Chiclayo 01A	1	25667871	201909	92	74	19.57
Chiclayo	Chiclayo 02	1	25403069	201909	42	44	4.76
Chiclayo	Chiclayo 02A	2	25191969	201909	0	0	0
Chiclayo	Chiclayo 02A	2	25606240	201909	132	137	3.79
Chiclayo	Chiclayo 03	6	25017151	201909	196	180	8.16
Chiclayo	Chiclayo 03	6	25017170	201909	78	77	1.28
Chiclayo	Chiclayo 03	6	25017198	201909	176	182	3.41

Fuente: Elaborado por el autor



Modelo B - NNETAR

Descripción del Modelo B

Nnetar es una Red neuronal autoregresivo, ésta analiza el comportamiento de diversas variables con el fin de determinar un estado objetivo.

```

98 nnetar <- function(y, p, P=1, size, repeats=20, xreg=NULL, lambda=NULL, model=NULL, subset=NULL)
99   useoldmodel <- FALSE
100   yname <- deparse(substitute(y))
101   if (!is.null(model)) {
102     # Use previously fitted model
103     useoldmodel <- TRUE
104     # Check for conflicts between new and old data:
105     # Check model class
106     if (!is.nnetar(model)) {
107       stop("Model must be a nnetar object")
108     }
109     # Check new data
110     m <- max(round(frequency(model$x)), 1L)
111     minlength <- max(c(model$p, model$P * m)) + 1
112     if (length(x) < minlength) {
113       stop(paste("Series must be at least of length", minlength, "to use fitted model"))
114     }
115     if (tsp(as.ts(x))[3] != m) {
116       warning(paste("Data frequency doesn't match fitted model, coercing to frequency =", m))
117       x <- ts(x, frequency = m)
118     }
119     # Check xreg
120     if (!is.null(model$xreg)) {
121       if (is.null(xreg)) {
122         stop("No external regressors provided")
123       }
124       if (NCOL(xreg) != NCOL(model$xreg)) {
125         stop("Number of external regressors does not match fitted model")
126       }
127     }
128     # Update parameters with previous model
129     lambda <- model$lambda
130     size <- model$size
131     p <- model$p
132     P <- model$P
133     if (P > 0) {
134       lags <- sort(unique(c(1:p, m * (1:P))))

```

Figura 27. Algoritmo R – Nnetar

En la investigación, solo se ingresa el vector numérico de series de tiempo.

Es así, que el análisis de la serie a través de una red neuronal se trata con un método previo, la teoría de ventanas, que es un algoritmo que permite expandir y generar atributos (columnas) con los datos iniciales del vector, para luego explicar la relación de cada uno de éstos a partir de un modelo regresivo.

Evaluación del Modelo B

Haciendo uso de R se aplica el algoritmo sobre el histórico del suministro. El modelo realiza el entrenamiento de la serie, determinando en forma automática e interpretativa los valores que formarán parte de la serie de tiempo.

Tabla 16
Red Neuronal

Periodo	V Original	V-1	V-2	V-3
201001	45	¿	¿	¿
201002	65	45	¿	¿
201003	55	65	45	¿
201004	75	55	65	45
201005	89	75	55	65
201006	13	89	75	55
201007	X	¿	¿	¿

(Objetivo)

Fuente: Elaborado por el autor



En la tabla anterior se muestra el formato a analizado por la red neuronal y explica el fenómeno que se obtuvo para cada mes real; además se observa la distribución dada según la información que contiene el vector original. A continuación, se presenta el Algoritmo Nnetar desarrollado en código original:

```
##### PRONOSTICANDO CON NNETAR - REDES NEURANALES AUTOREGRESIVAS
aplicaNn <- nnetar(dfp)
pronosticoNn<- forecast(aplicaNn,h=1)
pro_nnetar <- round(as.numeric(pronosticoNn$mean[1]), digits = 0)
```

Figura 28. Código Algoritmo R - Nnetar

Luego de la implementación del algoritmo, se aprecia su funcionamiento en la siguiente tabla:

Tabla 17
Resultados obtenidos con Nnetar

UUNN	Ciclo	Mues	Suministro	Periodo	Real	NNetar	Error NNetar
Cajamarca Centro	I - Cajamarca Centro	1	28361839	201909	20	20	0
Cajamarca Centro	II - Cajamarca Centro	1	28217192	201909	68	70	2.94
Cajamarca Centro	III - Cajamarca Centro	1	28179959	201909	184	181	1.63
Cajamarca Centro	IV - Cajamarca Centro	1	25610289	201909	25	25	0
Cajamarca Centro	V - Cajamarca Centro	3	25602080	201909	16	16	0
Cajamarca Centro	V - Cajamarca Centro	3	25602563	201909	165	164	0.61
Cajamarca Centro	V - Cajamarca Centro	3	25602581	201909	43	43	0
Chiclayo	Chiclayo 0	1	25778262	201909	87	87	0
Chiclayo	Chiclayo 01	2	25602302	201909	6454	4889	24.25
Chiclayo	Chiclayo 01	2	25602320	201909	113	111	1.77
Chiclayo	Chiclayo 01A	1	25667871	201909	92	94	2.17
Chiclayo	Chiclayo 02	1	25403069	201909	42	43	2.38
Chiclayo	Chiclayo 02A	2	25191969	201909	0	0	0
Chiclayo	Chiclayo 02A	2	25606240	201909	132	132	0
Chiclayo	Chiclayo 03	6	25017151	201909	196	193	1.53
Chiclayo	Chiclayo 03	6	25017170	201909	78	80	2.56
Chiclayo	Chiclayo 03	6	25017198	201909	176	173	1.7
Chiclayo	Chiclayo 03	6	25017204	201909	329	334	1.52
Chiclayo	Chiclayo 03	6	25017213	201909	524	532	1.53

Fuente: Elaborado por el autor



Modelo C - ARIMA

Descripción del Modelo C

Arima disminuye la tendencia y la pendiente, ya que emplea una constante de atenuación que es diferente para cada una de ellas.

Es un modelo de suavizado exponencial que utiliza directamente la tendencia al obtener la diferencia entre los valores sucesivos, para pronosticar “n” periodos.

Fórmulas de ARIMA

A non-seasonal ARIMA model can be written as

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \dots + \theta_q B^q) \varepsilon_t, \quad (8.4)$$

or equivalently as

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d (y_t - \mu t^d / d!) = (1 + \theta_1 B + \dots + \theta_q B^q) \varepsilon_t, \quad (8.5)$$

where $c = \mu(1 - \phi_1 - \dots - \phi_p)$ and μ is the mean of $(1 - B)^d y_t$. R uses the parameterisation of Equation (8.5).

Figura 29. Formulas algoritmo ARIMA



```

> Arima
function (y, order = c(0, 0, 0), seasonal = c(0, 0, 0), xreg = NULL,
  include.mean = TRUE, include.drift = FALSE, include.constant,
  lambda = model$lambda, biasadj = FALSE, method = c("CSS-ML",
    "ML", "CSS"), model = NULL, x = y, ...)
{
  series <- deparse(substitute(y))
  origx <- y
  if (!is.null(lambda)) {
    x <- BoxCox(x, lambda)
    lambda <- attr(x, "lambda")
    if (is.null(attr(lambda, "biasadj"))) {
      attr(lambda, "biasadj") <- biasadj
    }
  }
  if (!is.null(xreg)) {
    if (!is.numeric(xreg))
      stop("xreg should be a numeric matrix or a numeric vector")
    xreg <- as.matrix(xreg)
    if (is.null(colnames(xreg))) {
      colnames(xreg) <- if (ncol(xreg) == 1)
        "xreg"
      else paste("xreg", 1:ncol(xreg), sep = "")
    }
  }
  if (!is.list(seasonal)) {
    if (frequency(x) <= 1) {
      seasonal <- list(order = c(0, 0, 0), period = NA)
      if (length(x) <= order[2L])
        stop("Not enough data to fit the model")
    }
    else {
      seasonal <- list(order = seasonal, period = frequency(x))
      if (length(x) <= order[2L] + seasonal$order[2L] *
        seasonal$period)
        stop("Not enough data to fit the model")
    }
  }
  if (!missing(include.constant)) {
    if (include.constant) {
      include.mean <- TRUE
      if ((order[2] + seasonal$order[2]) == 1) {
        include.drift <- TRUE
      }
    }
    else {
      include.mean <- include.drift <- FALSE
    }
  }
  if ((order[2] + seasonal$order[2]) > 1 & include.drift) {
    warning("No drift term fitted as the order of difference is 2 or more.")
    include.drift <- FALSE
  }
}

```

Figura 30. Código Fuente Algoritmo ARIMA en R



Evaluación del Modelo C

Algoritmo ARIMA en líneas de código.

```
##### PRONOSTICANDO AUTOREGRESION

aplicaar <- arima(dfp,order=c(0,3,3))
pronosticoar<- forecast(aplicaar, h=1)
pro_ar<-round(as.numeric(pronosticoar[4]), digits = 0)
```

Figura 31. Código Algoritmo ARIMA

Luego de la implementación del algoritmo, se aprecia su funcionamiento en la siguiente tabla:

Tabla 18
Resultados obtenidos con ARIMA

UUNN	Ciclo	Mues	Suministro	Periodo	Real	ARIMA	Error AR
Cajamarca Centro	I - Cajamarca Centro	1	28361839	201909	20	20	0
Cajamarca Centro	II - Cajamarca Centro	1	28217192	201909	68	71	4.41
Cajamarca Centro	III - Cajamarca Centro	1	28179959	201909	184	183	0.54
Cajamarca Centro	IV - Cajamarca Centro	1	25610289	201909	25	25	0
Cajamarca Centro	V - Cajamarca Centro	3	25602080	201909	16	16	0
Cajamarca Centro	V - Cajamarca Centro	3	25602563	201909	165	163	1.21
Cajamarca Centro	V - Cajamarca Centro	3	25602581	201909	43	43	0
Chiclayo	Chiclayo 0	1	25778262	201909	87	107	22.99
Chiclayo	Chiclayo 01	2	25602302	201909	6454	0	100
Chiclayo	Chiclayo 01	2	25602320	201909	113	131	15.93
Chiclayo	Chiclayo 01A	1	25667871	201909	92	89	3.26
Chiclayo	Chiclayo 02	1	25403069	201909	42	41	2.38
Chiclayo	Chiclayo 02A	2	25191969	201909	0	0	0
Chiclayo	Chiclayo 02A	2	25606240	201909	132	122	7.58
Chiclayo	Chiclayo 03	6	25017151	201909	196	184	6.12
Chiclayo	Chiclayo 03	6	25017170	201909	78	81	3.85
Chiclayo	Chiclayo 03	6	25017198	201909	176	175	0.57
Chiclayo	Chiclayo 03	6	25017204	201909	329	330	0.3
Chiclayo	Chiclayo 03	6	25017213	201909	524	525	0.19

Fuente: Elaborado por el autor



Modelo D - SVM

Descripción del Modelo SVM

```

1  svm <-
2  function (x, ...)
3    UseMethod ("svm")
4
5  svm.formula <-
6  function (formula, data = NULL, ..., subset, na.action = na.omit, scale = TRUE)
7  {
8    call <- match.call()
9    if (!inherits(formula, "formula"))
10     stop("method is only for formula objects")
11    m <- match.call(expand.dots = FALSE)
12    if (inherits(eval.parent(m$data), "matrix"))
13     m$data <- as.data.frame(eval.parent(m$data))
14    m$... <- NULL
15    m$scale <- NULL
16    m[[1L]] <- quote(stats::model.frame)
17    m$na.action <- na.action
18    m <- eval(m, parent.frame())
19    Terms <- attr(m, "terms")
20    attr(Terms, "intercept") <- 0
21    x <- model.matrix(Terms, m)
22    y <- model.extract(m, "response")
23    attr(x, "na.action") <- attr(y, "na.action") <- attr(m, "na.action")
24    if (length(scale) == 1)
25     scale <- rep(scale, ncol(x))
26    if (any(scale)) {
27     remove <- unique(c(which(labels(Terms) %in%
28                           names(attr(x, "contrasts"))),
29                       which(!scale)
30                           )
31                       )
32     scale <- !attr(x, "assign") %in% remove
33   }
34   ret <- svm.default (x, y, scale = scale, ..., na.action = na.action)
35   ret$call <- call
36   ret$call[[1]] <- as.name("svm")
37   ret$terms <- Terms
38   if (!is.null(attr(m, "na.action")))

```

Figura 32. Algoritmo SVM en R

Recientemente, las máquinas de soporte vectorial (Support Vector Machine) se han propuesto como técnica novedosa en el pronóstico de series de tiempo. Siendo un tipo específico de algoritmos de aprendizaje, que se caracteriza por el control de capacidad de la función de decisión (Similar a lo que ocurre con una red neuronal).



Sus funciones núcleo se orientan en una teoría de minimización de riesgo estructural para estimar una función al minimizar un límite superior (umbral). Para los problemas de series de tiempo son dos los factores clave que trata SVM, uno es el ruido y el otro es la no estacionalidad, lo que dificulta el aprendizaje con otros algoritmos sobre los datos históricos para captar la dependencia entre futuro y pasado.

Similar a la red neuronal este modelo puede ser utilizado para resolver problemas de clasificación, utilizando regresión también se pueden resolver problemas para pronósticos de series de tiempo.

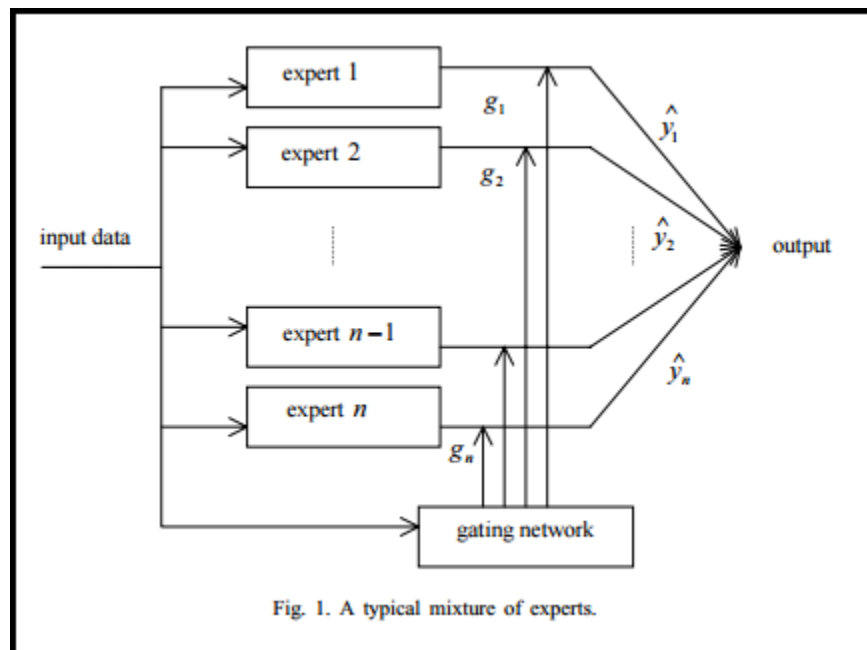


Figura 33. Conceptualización de SVM

Evaluación del Modelo SVM

A continuación, se muestra la implementación de SVM en R con el código para el modelo de la investigación.



```
##### PRONOSTICANDO CON SVM - MAQUINA DE SOPORTE VECTORIAL
b <- c(1:40)
c <- as.vector(t(dfp))
test<- c(1:40)

#test <- c(36:36)
mod <- svm(c ~ b, kernel = "linear", gamma = 1, cost = 20, type="eps-regressior
pro_svm<-predict(mod, newdata = data.frame(x = test))
pro_svm<-round(as.numeric(pro_svm[1]), digits = 0)
```

Figura 34. Implementación de SVM en R

Capturando los datos en la entidad para pruebas se puede visualizar la siguiente información:

Tabla 19
Resultados de SVM

UUNN	Ciclo	Mues	Suministro	Periodo	Real	SVM	Error SVM
Cajamarca Centro	I - Cajamarca Centro	1	28361839	201909	20	20	0
Cajamarca Centro	II - Cajamarca Centro	1	28217192	201909	68	67	1.47
Cajamarca Centro	III - Cajamarca Centro	1	28179959	201909	184	179	2.72
Cajamarca Centro	IV - Cajamarca Centro	1	25610289	201909	25	25	0
Cajamarca Centro	V - Cajamarca Centro	3	25602080	201909	16	17	6.25
Cajamarca Centro	V - Cajamarca Centro	3	25602563	201909	165	169	2.42
Cajamarca Centro	V - Cajamarca Centro	3	25602581	201909	43	44	2.33
Chiclayo	Chiclayo 0	1	25778262	201909	87	122	40.23
Chiclayo	Chiclayo 01	2	25602302	201909	6454	2069	67.94
Chiclayo	Chiclayo 01	2	25602320	201909	113	97	14.16
Chiclayo	Chiclayo 01A	1	25667871	201909	92	102	10.87
Chiclayo	Chiclayo 02	1	25403069	201909	42	35	16.67
Chiclayo	Chiclayo 02A	2	25191969	201909	0	0	0
Chiclayo	Chiclayo 02A	2	25606240	201909	132	107	18.94
Chiclayo	Chiclayo 03	6	25017151	201909	196	234	19.39
Chiclayo	Chiclayo 03	6	25017170	201909	78	75	3.85
Chiclayo	Chiclayo 03	6	25017198	201909	176	158	10.23
Chiclayo	Chiclayo 03	6	25017204	201909	329	294	10.64
Chiclayo	Chiclayo 03	6	25017213	201909	524	576	9.92

Fuente: Elaborado por el autor

E. Evaluación

Este apéndice ha sido presentado en el capítulo IV del informe.



FASE 2: APLICACIÓN WEB – METODOLOGIA XP

A. Análisis

En el caso de la aplicación, se trata del desarrollo de un portal de visualización de datos donde el usuario puede evaluar los distintos algoritmos que se utilizaron para desarrollar el modelo. Por lo tanto se han logrado extraer las siguientes historias de usuario como objetivo de la aplicación:

Tabla 20
Historia de usuario

HISTORIA DE USUARIO			
Código	Nombre	Descripción	Actores
HU 1	Gestión de pronósticos	El usuario puede visualizar los pronósticos resultantes del modelo propuesto	Usuario Analista, Jefe de Facturación, Supervisor de Facturación
HU 2	Gestión de Histórico	El usuario puede visualizar el registro histórico de los datos	Usuario Analista
HU 3	Gestión de Estadística y Muestreo	El usuario puede visualizar el proceso de muestreo de la data, así como generar los suministros propios del universo de muestreo	Usuario Analista
HU 4	Gestión de Resultados Investigación	El usuario puede procesar los resultados de las diversas técnicas para obtener las estimaciones, así también puede calcular el MAPE	Administrador de sistemas, Jefe de Facturación

Fuente: Elaborado por el autor



Especificación de Historias de usuario

Cada historia de usuario presentada se detalla de manera explícita según el siguiente formato:

Tabla 21
Historia de usuario 01

Historia de Usuario	HU 01 – Gestión de pronósticos
Actores	Analista, Jefe de Facturación, Supervisor de Facturación
Descripción	<ol style="list-style-type: none"> 1. El usuario ingresará al módulo gestión de pronósticos presionando el link que debe estar ubicado en el menú de sistema. 2. El sistema presentará opciones de selección para que el usuario elija la empresa, unidad de negocio y ciclo para buscar los suministros, puede filtrarse como opción múltiple. 3. El usuario selecciona sus filtros en los 03 niveles 4. El sistema presentará el detalle de los resultados en la página donde mostrara la matriz de suministros en la tabla 1. 5. El usuario seleccionará uno o varios suministros a criterio. 6. El sistema presentará una tabla 2 donde se visualice los pronósticos para el último periodo, conformen se seleccionen suministros en la primera tabla se generarán más registros en la tabla 2. 7. Si el usuario pulsa un registro suministro de la tabla 2 que contiene los pronósticos podrá visualizar el histórico en gráfico.
Observación	Ninguna

Fuente: Elaborado por el autor



Tabla 22
Historia de usuario 02

Historia de Usuario	HU 02 – Gestión de histórico
Actores	Usuario Analista
Descripción	<ol style="list-style-type: none"> 1. El usuario analista puede consultar el histórico que alimenta el modelo en el link presentado por el menú del sistema. 2. El sistema cargará un menú desplegable para visualizar los filtros mencionados en la HU01. 3. El usuario seleccionará los filtros 4. En la tabla 1 se cargará el listado de suministros según los filtros, al seleccionar los suministros en una tabla 3 se cargará el valor histórico de cada suministro. 5. El usuario verificara los datos.
Observación	Ninguna
Fuente: Elaborado por el autor	

Tabla 23
Historia de usuario 03

Historia de Usuario	HU 03 – Gestión de Estadística y muestreo
Actores	Analista
Descripción	<ol style="list-style-type: none"> 1. El usuario podrá visualizar la composición del muestreo de la investigación basada en la fórmula del muestro estratificado. 2. El sistema solicita al usuario ingresar la empresa a la que se aplicará el muestreo. 3. El usuario selecciona la empresa. 4. El sistema presenta en una tabla los resultados. Estos resultados se generan en el archivo CSV maestro del sistema. 5. El usuario puede también generar un CSV para descargar a partir de la tabla.
Observación	
Fuente: Elaborado por el autor	

Tabla 24
Historia de usuario 04

Historia de Usuario	HU 04 – Gestión de resultados de investigación
Actores	Todos los usuarios de sistema
Descripción	<ol style="list-style-type: none"> 1. El usuario podrá visualizar los resultados según el muestreo aplicado. 2. El sistema ofrece opciones al usuario, como importar un archivo CSV para la muestra, procesar la muestra por defecto del aplicativo y calcular el MAPE de los resultados. 3. El usuario ejecuta el método de procesamiento a discreción, el sistema muestra una tabla que contiene los suministros del CSV de muestra y obtiene los pronósticos para las distintas técnicas. El sistema procesa los datos aplicando esquema de colores según el indicador de estimaciones del proyecto. 4. Si el usuario desea conocer el MAPE global debe presionar la opción “calcular MAPE”, el sistema mostrara en una pequeña tabla los cálculos obtenidos.
Observación	

Fuente: Elaborado por el autor

B. Diseño

Se utilizará el framework Shiny para R por lo cual la propuesta será una web de conexión directa al motor R para procesamiento dividida en 03 pantallas (Consulta – Muestreo – Resultados)

C. Codificación

La codificación está realizada mediante R, se agregarán unas capturas de código y pantallas resultantes.



```

1 library("sqldf")
2 library("forecast")
3 library("e1071")
4 library("dplyr")
5
6 pathf1<- "H:/xampp/htdocs/appLabEnsa"
7
8 d_empresa<- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\DIM_EMPRESA")
9 d_uunn<- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\DIM_UNIDADNEGOCIO")
10 d_ciclo<- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\DIM_CICLO_2")
11 #d_ciclo_g <- paste("select ciclo as idciclo,ciclo as nombreciclo,glomas_unidadnegocio_id
12 #from d_ciclo group by ciclo,glomas_unidadnegocio_id",sep="")
13 #d_ciclo <- sqldf(d_ciclo_g)
14
15 d_administrativo <- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\ADMINISTRATIVO\\ADMINISTRATIVO")
16 pivot <- readRDS(file="H:\\xampp\\htdocs\\appLabENSA\\dataset\\DWH_Ensa\\PIVOT\\PIVOT")
17
18 if (interactive()) {
19
20   shinyApp(==
21     ui = fluidPage(theme = shinytheme("cerulean"),
22       #shinythemes::themeSelector(), # <--- Add this somewhere in the UI
23       navbarPage("LabEnsa - Panel de Laboratorio - Tesis USS",
24         tabPanel("Go Live!",
25           sidebarPanel(==
26             ),
27           mainPanel(
28             tabsetPanel(
29               tabPanel("Tabla General de Suministros", dataTableOutput(outputId = 'x3')),
30               #tabPanel("Tabla de suministros por busqueda",verbatimTextOutput('x4'))
31               tabPanel("Historial de Consumo Por Suministro", dataTableOutput(outputId = 'x4'))
32             )
33           ),
34           hr(),
35           fluidRow(
36             column(7, style='margin-left:15px;',htmlOutput("title1"),
37               dataTableOutput(outputId = 'x5')
38             ),
39             column(4,htmlOutput("title2"),
40               plotOutput("g1", click = "plot_click")
41               #verbatimTextOutput('g1')
42             )
43           )
44         ),
45         tabPanel("Estadística",
46           sidebarPanel(
47             uiOutput("viz2firstSelection")
48           ),
49           mainPanel(
50             tabsetPanel(id = "inTabset2",
51               tabPanel("Muestreo",value = "panel31", dataTableOutput(outputId = 'x6'),
52                 downloadButton("downloadData21", "Descargar"),actionButton("gen21", "Generar Suministros", class = "btn-succ
53               )
54             tabPanel("Detalle Muestreo",value = "panel32", dataTableOutput(outputId = 'x7'),downloadButton("downloadData22
55             )
56           )
57         ),
58         tabPanel("Resultados Investigación",|
59           sidebarPanel(==
60
61
62
63

```

Figura 35. Script Web Laboratorio con R y Shiny



Se muestra la pantalla Laboratorio 1 – Realizar pronósticos de consumo en APP de cada algoritmo y por cada suministro seleccionado.

The screenshot displays the 'Laboratorio 1' interface with the following components:

- Navigation Bar:** LabEnsa - Panel de Laboratorio - Tesis USS | Go Live | Estadística | Resultados Investigación | Análisis Tiempos
- Filters:**
 - Empresa: Electronorte S.A.
 - Unidad de Negocio: Chiclayo
 - Ciclo de Facturación: Chiclayo 01
- Table General de Suministros:**

Suministro	Ciclo	Unidad Negocio	Empresa	
1	25602302	Chiclayo 01	Chiclayo	Electronorte S.A.
2	25602320	Chiclayo 01	Chiclayo	Electronorte S.A.
3	25700410	Chiclayo 01	Chiclayo	Electronorte S.A.
4	25742107	Chiclayo 01	Chiclayo	Electronorte S.A.
5	25743534	Chiclayo 01	Chiclayo	Electronorte S.A.
6	25754654	Chiclayo 01	Chiclayo	Electronorte S.A.
7	25755508	Chiclayo 01	Chiclayo	Electronorte S.A.
8	25815099	Chiclayo 01	Chiclayo	Electronorte S.A.
9	25834970	Chiclayo 01	Chiclayo	Electronorte S.A.
10	25862660	Chiclayo 01	Chiclayo	Electronorte S.A.
- Pronóstico Generado con multiples algoritmos:**

Suministro	Periodo Pronosticado	Monto Real	Pronostico AR	Pronostico HW	Pronostico NNetar	Pronostico SVM	
201909	25834970	201909	71	72	69	71	70
201909	25602320	201909	113	131	132	111	97
201909	25754654	201909	79	80	80	78	74
- Imagen proyectada por suministro seleccionado:**

25834970

Figura 36. Pantalla Laboratorio 1 – Realizar pronósticos en APP



Pantalla Laboratorio 2 – Pantalla de Estadística, aplicando el método de Muestreo Estratificado, se seleccionaron 113 suministros como muestra.

LabEnsa - Panel de Laboratorio - Tesis USS Go Live! Estadística Resultados Investigación Análisis Tiempos

Empresa:

Muestreo [Detalle Muestreo](#)

Show 25 entries

	Empresa	Unidad de Negocio	Ciclo	Suministros	Total Sum	% Rep	C. Muestral
1	Electronorte S.A.	Cajamarca Centro	I - Cajamarca Centro	1093	326383	0.0033	0
2	Electronorte S.A.	Cajamarca Centro	II - Cajamarca Centro	5157	326383	0.0158	103
3	Electronorte S.A.	Cajamarca Centro	III - Cajamarca Centro	6988	326383	0.0214	140
4	Electronorte S.A.	Cajamarca Centro	IV - Cajamarca Centro	993	326383	0.003	0
5	Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	326383	0.0287	281
6	Electronorte S.A.	Chiclayo	Chiclayo 0	3070	326383	0.0094	31
7	Electronorte S.A.	Chiclayo	Chiclayo 01	8266	326383	0.0253	248
8	Electronorte S.A.	Chiclayo	Chiclayo 01A	7371	326383	0.0226	147
9	Electronorte S.A.	Chiclayo	Chiclayo 02	7293	326383	0.0223	146
10	Electronorte S.A.	Chiclayo	Chiclayo 02A	7536	326383	0.0231	151
11	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	326383	0.0431	563
12	Electronorte S.A.	Chiclayo	Chiclayo 04	13082	326383	0.0401	523
13	Electronorte S.A.	Chiclayo	Chiclayo 05	14060	326383	0.0431	562
14	Electronorte S.A.	Chiclayo	Chiclayo 05A	7003	326383	0.0215	140

Figura 37. Pantalla Laboratorio 2 – Pantalla para extraer Muestreo



Pantalla donde muestra el pronóstico de consumo con cada técnica utilizada, mostrando resultados del MAPE (Porcentaje de Error Medio Absoluto) y ACC (Precisión del Pronóstico – Forecast Accuracy).

LabEnsa - Panel de Laboratorio - tesis USS

Co Nivel Estadística Resultados Investigación Análisis Tiempos

Importe el fichero si desea cargar nueva muestra

Browse... No file selected

Análisis Muestra Subida Análisis Muestra Lab Calcular MAPE

Resultado General

Show 10 entries

Search:

Medida	ARIMA	HW	NNETAR	SVM
1 MAPE	15.946725637160	33.3690490170991	0.59620310504071	10.0064001709912
2 ACC	84.0532743382832	68.6103530823000	91.4137168141593	81.9035305230083

Showing 1 to 2 of 2 entries

Previous 1 Next

Resultado Detallado

Show 26 entries

Search:

Empresa	UUNN	Ciclo	Total.Suministros	Total.Muestra	Suministro	Periodo Pronosticado	Monto Real	Pronostico ARIMA	Error%AbsAR	Pronostico HW	Error%AbsHW	Pronostico NNetar	Error%AbsNNetar	Pronostico SVM	Error%AbsSVM
1	Electronorte S.A.	Cajamarca Centro	I - Cajamarca Centro	1093	1	20361039	201909	20	0	20	0	20	0	20	0
2	Electronorte S.A.	Cajamarca Centro	II - Cajamarca Centro	5157	1	28217192	201909	60	4.41	65	4.41	70	2.94	67	1.47
3	Electronorte S.A.	Cajamarca Centro	III - Cajamarca Centro	6988	1	28179859	201909	164	0.54	246	0.22	181	1.90	179	2.70
4	Electronorte S.A.	Cajamarca Centro	IV - Cajamarca Centro	903	1	25610299	201909	25	0	24	4	25	0	25	0
5	Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9308	3	25602090	201909	16	0	7	23.25	16	0	17	6.25
6	Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	3	25802583	201909	165	1.31	173	4.88	184	0.81	180	3.40
7	Electronorte S.A.	Cajamarca Centro	V - Cajamarca Centro	9358	3	25602591	201909	43	0	40	11.63	43	0	44	2.32
8	Electronorte S.A.	Chiclayo	Chiclayo 0	3070	1	25778282	201909	87	27.90	42	21.75	87	0	122	30.21
9	Electronorte S.A.	Chiclayo	Chiclayo 01	8266	2	25602302	201909	6454	0.00	7520	10.07	4809	24.05	2069	17.84
10	Electronorte S.A.	Chiclayo	Chiclayo 01	8266	2	25602320	201909	113	13.61	132	14.01	111	1.77	97	14.16
11	Electronorte S.A.	Chiclayo	Chiclayo 01A	7371	1	25667871	201909	92	3.26	74	18.77	94	3.17	102	19.87
12	Electronorte S.A.	Chiclayo	Chiclayo 02	7293	1	25403099	201909	42	2.36	44	4.76	43	2.36	35	19.05
13	Electronorte S.A.	Chiclayo	Chiclayo 02A	7538	2	25191989	201909	0	0	0	0	0	0	0	0
14	Electronorte S.A.	Chiclayo	Chiclayo 02A	7536	2	25606240	201909	132	7.56	137	3.75	130	1.92	107	18.91
15	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017151	201909	196	6.10	190	6.76	193	1.65	234	19.70
16	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017170	201909	78	3.85	77	1.26	80	2.96	75	3.85
17	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017196	201909	176	0.57	182	3.81	173	1.7	156	19.23
18	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017204	201909	329	0.7	335	1.92	333	1.90	294	19.84
19	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017213	201909	524	0.19	547	4.36	543	3.93	576	9.92
20	Electronorte S.A.	Chiclayo	Chiclayo 03	14081	6	25017231	201909	1460	1.37	1448	0.92	1483	1.50	1368	19.91
21	Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190793	201909	1285	100.72	2271	23.73	1155	10.12	958	24.40
22	Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190792	201909	179	11.17	173	3.30	190	0.96	214	19.32
23	Electronorte S.A.	Chiclayo	Chiclayo 04	13082	5	25190806	201909	242	0.41	230	4.96	240	0.91	238	0.46

Figura 38. Pantalla Laboratorio 3 – Resultados y calculo MAPE



De acuerdo a la imagen de la página anterior, se asignó el color verde a aquellos suministros donde el MAPE (Porcentaje de Error Medio Absoluto) era menor o igual al 15% y el color rojo a los que tenían el MAPE mayor al 15%.

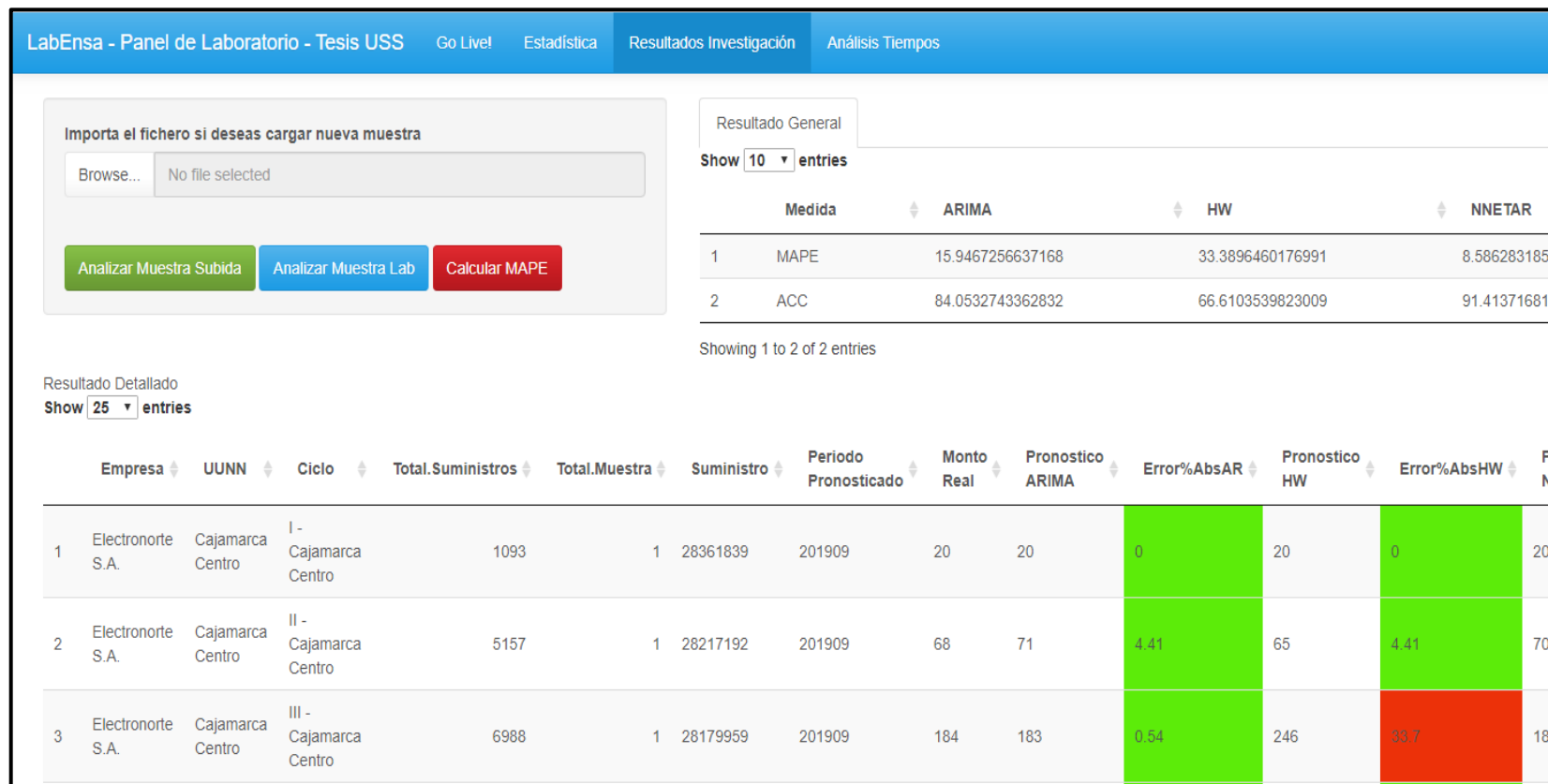


Figura 39. Pantalla Laboratorio 3 – Resultados y calculo MAPE



D. Pruebas

La fase de pruebas está documentada en el capítulo IV.

CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Se realizó la recopilación y análisis de la información brindada por ELECTRONORTE S.A. determinando así los procesos críticos del problema, se realizó el análisis del datawarehouse actual de Electronorte con los datos de los procesos de facturación mensual, determinando que esta base de datos analítica contenía todos los campos necesarios para realizar la investigación, así mismo para aplicar las técnicas, se determinó realizar un muestreo estratificado obteniendo como resultado 113 medidores como muestra representativa por ser una base con 56 000 000 de registros, comprendiendo los consumos históricos desde el periodo mayo 2016 hasta setiembre 2019.

Se comparó y seleccionó las técnicas predictivas de minería de datos, determinándose como modelo a utilizar, uno de series de tiempo, por la naturaleza de los datos trabajados en el datawarehouse que comprende las series numéricas que tienen como marcador de índice al periodo comercial en que se facturó el consumo, por lo tanto, se realizó el análisis de las técnicas y algoritmos que intervienen en este modelo.

Dentro de las técnicas predictivas se determinó utilizar los algoritmos de ARIMA, Holtwinters, Redes Neuronales y SVM (Maquina de soporte vectorial), ya que al realizar el análisis se fueron descartando algunas técnicas por no cumplir con los criterios establecidos y requeridos para ser implementados en el modelo a desarrollar.

Al evaluar los resultados, orientado a la medición de la correcta estimación de consumo, se obtuvo un desempeño ACC (Accuracy) para Arima de 84.05 % dado que el error MAPE resulto en 15.94%. Para el caso de HoltWinter el ACC es de 66.61 y el MAPE obtenido es de 33.38 % %, siendo el peor modelo según los



resultados, en el caso de la Red Neuronal Autoregresiva NNETAR el ACC obtiene un 88.73 % y un MAPE de 11.26 % el cual tiene el mejor desempeño de los modelos, para el caso del SVM el valor del ACC es 81.99% y un MAPE de 18.00 %, ordenando estos algoritmos por su desempeño obtenemos a la RED NEURONAL AUTOREGRESIVA, ARIMA, SVM y HOLTWINTERS. Según nuestro cuadro de evaluación de la métrica de estimación Tabla 5, el único modelo que podría entrar a la categoría de BUENO es la RED NEURONAL AUTOREGRESIVA. Este resultado es interesante, dado que HOLTWINTERS y ARIMA esquemas convencionales tienen inconvenientes según la naturaleza de la serie de tiempo y es de difícil optimización en los cálculos que se realiza en cuanto a sus coeficientes, en el caso de Holtwinters se observa buen detalle en cuanto a series con componentes estacionales, sin embargo para componentes cíclicos o estacionarios ARIMA representa una gran ventaja, es por esto que los algoritmos de Redes Neuronales y SVM comprenden un mejor desempeño. En el caso de segundo indicador se determinó que el mejor algoritmo en ACC tarda en promedio 0.43 segundos en realizar el entrenamiento con un tiempo total de 48.99 segundos para los 113 sujetos de prueba, el segundo mejor algoritmo, es decir ARIMA, tarda 25.67 casi un 50 % menor a Nnetar, un caso peculiar se concentra en SVM con un tiempo promedio de 7 segundos. Este factor es importante cuando solo se han tratado 113 suministros, recordando que toda la población es de 356 000 suministros solo para ENSA aproximadamente. Por lo tanto, nnetar debe ser tratado con otras estrategias de procesamiento de datos.

Se construyó una aplicación web para realizar simulaciones extrayendo el histórico de consumos de los suministros del datawarehouse, para evaluar el comportamiento con las distintas técnicas utilizadas durante esta investigación.



6.2. Recomendaciones

Para bases de datos con excesiva cantidad de información, incluyendo de naturaleza Datawarehouse, se recomienda aplicar técnicas de matrices consolidadas para el tratamiento de series de tiempo, en esta investigación al aplicar esta técnica se redujo una consulta de 9 minutos a 21 segundos.

Los tratamientos de valores nulos cuando se trabaja con este tipo de datos deben ser de la manera más minuciosa y cuidadosa posible. Gracias al uso de una matriz consolidada, se logró identificar los valores faltantes en la base de datos, los mismos que podrían ocasionar daños al momento de realizar los cálculos para la serie de tiempo.

La técnica SVM usa los principios de regresión que también se usan en Redes Neuronales, por lo que obtiene buenos resultados sobre todo en una naturaleza de series donde no siempre existe estacionalidad para casos particulares de suministros eléctricos.

Referencias Bibliográficas

- Alvarez, C. A. (2012). *Aplicacion de tecnicas de mineria de datos para mejorar el proceso de control de gestion en ENTEL*. Santiago de Chile: Universidad de Chile.
- Asencios, V. V. (2004). Datamining y el Descubrimiento del Conocimiento. *Revista de la Facultad de Ingenieria Industrial*, (7), p. 83-86.
- Bustos, J. M. (2011). *Diseño e implementacion de un modelo predictivo para detectar patrones de fuga en los servicios de telefonía del sur*. Valdivia: Universidad Austral de Chile.
- Calvo Rodríguez, A. (2008). *Predicción en series de Tiempo con Modelos Aditivos*. España: Universidade da Coruña.
- Díaz, J. A. (2013). *Implementacion de una aplicacion web utilizando mineria de datos para mejorar la gestion de facturacion en la empresa PEXPORT SAC*. Chiclayo: Universidad Señor de Sipan.
- Dongre, J., Prajapati, G. L., & Tokekar, S. (2014). *El papel del algoritmo Apriori para encontrar las reglas de Asociacion de mineria de datos*. Indore: Universidad de Indore.
- Fernández Maturana, V. P. (2007). Wavelet-and SVM-based forecasts: An analysis of the U.S. metal and materials. *Resources Policy*, 1-2.
- Getoor, L., & Ben, T. (2007). *Introducción a estadística de relación de aprendizaje*. MIT.
- Grudnitsky, B. J. (1992). *Diseño de sistemas de información. Teoría y Práctica*. México: Megabyte Grupo Noriega.
- Guil, F., Bosch, A., & Marin, R. (2003). *Una Propuesta para la Minería de Patrones Temporales Borrosos*. Murcia: Universidad de Murcia.
- Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit*. Wiley India.
- Lezcano, R. (2010). *Minería de datos* (Trabajo de investigación bibliográfica). Universidad Nacional del Nordeste, Corrientes, Argentina.
- 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.
- Montalvo, I. R. (2016). *ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE VENTAS*. Chiclayo.
- Ramírez A., A. Y. (2007). Técnicas de Minería de Datos Aplicadas a la Construcción de Modelos de Score Crediticio. *Mathematical Problems in Engineering*.
- Sandoval Vicente, J. F. (2014). *Sistema de pronóstico de inventario basado en modelos estadísticos para la distribución de repuestos del sector motos*. Lima: Universidad Peruana de Ciencias Aplicadas - UPC.
- Schaefer, J. M. (2012). El deporte, los artículos deportivos y la industria del deporte. *OMPI - Organización mundial de la propiedad intelectual*.



- Valcárcel Asencios, V. (2004). Data Mining y el descubrimiento del conocimiento. *Industrial Data*, 83-86.
- Vega, D. M. (2012). *Integración de modelos de agrupamiento y reglas de asociación obtenidos de múltiples fuentes de datos*. La Habana: Instituto Superior Politecnico Jose Antonio Echeverría.
- Vega, H. W. (2012). *Minería de Datos aplicados a las ventas con Tarjeta de Credito realizados en las tiendas Saga Falabella*. Lima: Universidad Tecnologica del Peru.

