



**FACULTAD DE INGENIERIA, ARQUITECTURA Y
URBANISMO**

**ESCUELA ACADÉMICO PROFESIONAL DE
INGENIERIA DE SISTEMAS**

TESIS

**APLICACIÓN DE TÉCNICAS DE MINERÍA DE
DATOS PARA PRONÓSTICO DE PRODUCCIÓN
DE ESPÁRRAGOS**

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

Autor:

Fernández Rojas Luis Humberto

Asesor:

Mg. Tuesta Monteza Víctor Alexci

Línea de Investigación:

Ciencia de la Computación

Pimentel – Perú

2019

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PRONÓSTICO DE PRODUCCIÓN DE ESPÁRRAGOS

Aprobación de Tesis

M.SC. Carlos Alberto Chirinos Mundaca
Presidente del jurado de tesis

Ing. Heber Iván Mejía Cabrera
Secretario del jurado de tesis

Mg. Víctor Alexci Tuesta Monteza
Vocal del jurado de tesis

DEDICATORIA

Este proyecto está dedicado a mis padres, pilares fundamentales en mi vida. Sin ellos, jamás hubiese podido conseguir lo que hasta ahora logré. Su tenacidad y lucha insaciable han hecho de ellos el gran ejemplo a seguir y destacar, no solo para mí, sino para mis hermanos y familia en general, que, sin ellos, no hubiese podido ser realizado.

Luis Humberto Fernández Rojas

AGRADECIMIENTO

Los resultados de este proyecto, están dedicados a todas aquellas personas que, de alguna forma, son parte de su culminación.

A mis padres quienes a lo largo de toda mi vida han apoyado y motivado mi formación académica, creyeron en mí en todo momento y no dudaron de mis habilidades. A mis profesores a quienes les debo gran parte de mis conocimientos, gracias a su paciencia y enseñanza y finalmente un eterno agradecimiento a esta prestigiosa universidad la cual abrió abre sus puertas a jóvenes como nosotros, preparándonos para un futuro competitivo y formándonos como personas de bien.

Mi sincero agradecimiento hacia el ingeniero Víctor Alexi Tuesta Monteza por su ayuda desinteresada, su apoyo a cada instante para lograr desarrollar esta investigación.

Luis Humberto Fernández Rojas

INDICE GENERAL

| | |
|--|----|
| DEDICATORIA | 3 |
| AGRADECIMIENTO | 4 |
| RESUMEN | 14 |
| ABSTRACT..... | 15 |
| INTRODUCCIÓN | 15 |
| CAPÍTULO I:..... | 16 |
| PROBLEMA DE LA INVESTIGACIÓN..... | 16 |
| 1.1. Situación Problemática..... | 17 |
| 1.2. Formulación del problema..... | 18 |
| 1.3. Delimitaciones de la investigación | 18 |
| 1.4. Justificación e Importancia de la investigación | 19 |
| 1.5. Limitaciones de la Investigación | 19 |
| 1.6. Objetivos de la Investigación | 21 |
| 1.6.1. Objetivo general | 21 |
| 1.6.2. Objetivos específicos | 21 |
| CAPITULO II..... | 22 |
| 2. Marco teórico | 23 |
| 2.1 Antecedentes de la Investigación | 23 |
| 2.2 Estado del arte..... | 25 |
| 2.3 Bases Teórico Científicas | 27 |
| 2.3.1. Aplicación Web | 44 |
| 2.3.1.1. Metodologías de Desarrollo de Software..... | 45 |
| 2.3.1.2. Proceso XP..... | 45 |
| 2.3.1.3. Metodologías de Desarrollo de Modelos de Minería de Datos | 47 |
| 2.3.1.5. Definición de términos básicos..... | 50 |



| | |
|--|-----------|
| CAPÍTULO III: | 52 |
| 3. Marcometodológico | 54 |
| 3.1. Tipo y diseño de la investigación | 54 |
| 3.2. Población y muestra | 55 |
| 3.2.1. Población | 55 |
| 3.2.2. Muestra | 56 |
| 3.2.3. Hipótesis | 56 |
| 3.3. Variables | 57 |
| 3.3.1. Variable independiente: | 57 |
| 3.3.2. Variable dependiente: | 57 |
| 3.4. Operacionalización | 57 |
| 3.5. Métodos, técnicas e instrumentos de recolección de datos | 58 |
| 3.5.1. Método de investigación: | 58 |
| 3.5.2. Técnicas | 59 |
| 3.5.3. Procedimientos para la recolección de datos | 59 |
| 3.5.4. Análisis estadístico e interpretación de los datos | 59 |
| 3.6. Principios éticos | 61 |
| 3.7. Criterios de rigor científico | 62 |
| 3.8. Evaluación económica del software | 62 |
| CAPÍTULO IV: | 65 |
| 4. Análisis e interpretación de los resultados | 66 |
| 4.1. Resultados de las tablas y graficos | 66 |
| 4.2. Contrastación de la hipótesis | 77 |
| 4.3. Discusión de resultados | 77 |
| CAPÍTULO V: | 79 |
| 5. Desarrollo de la propuesta | 81 |
| 5.1. Generalidades | 81 |



| | |
|--|-----|
| 5.2. metodología de desarrollo..... | 81 |
| 5.3. Etapas..... | 86 |
| 5.2.1 Etapa I diseño del modelado de minería de datos..... | 86 |
| 5.2.1.1. Comprensión de Negocio..... | 86 |
| APÍTULO VI: | 136 |
| CONCLUSIONES | 137 |
| RECOMENDACIONES | 138 |
| Bibliografía..... | 139 |
| ANEXOS | 142 |



ÍNDICE DE FIGURAS

| | |
|---|-----------|
| Figura N° 01: Proceso de Minería de Datos..... | 29 |
| Figura N°02: Cauterización de Datos..... | 30 |
| Figura N° 03: Gráfico de Tendencia de un conjunto de datos de los años 1974-1989..... | 30 |
| Figura N° 04: Gráfica de valores en el tiempo, donde se observa la estacionalidad..... | 31 |
| Figura N° 05: Proceso ideal de Minería de Datos..... | 32 |
| Figura N° 06: Red Neuronal..... | 38 |
| Figura N° 07: Técnicas de Data Mining..... | 39 |
| Figura N° 08: producción de espárragos..... | 43 |
| Figura N° 09: XP en las fases de desarrollo de software..... | 45 |
| Figura N° 10: Fases de la Metodología SCRUM..... | 46 |
| Figura N° 11: Fases del proceso de modelado metodología CRISP-DM. | |
| Figura N° 12: Fases de la Metodología SEMMA..... | 47 |
| Figura N° 13: Cliente- Servidor..... | 48 |
| Figura N° 14: Etapas de metodologías..... | 51 |
| Figura N° 15: Lista de clientes que entregan producción a la empresa..... | 84 |
| Figura N° 16: Series de Tiempo..... | 86 |
| FiguraN°17: Archivos de Data original..... | 89 |
| Figura N° 18: Datos extraídos originales | 90 |
| Figura N° 19: Producción de espárrago..... | 91 |
| Figura N°20: Extracción de datos a repositorio de análisis..... | 92 |
| Figura N° 21: Archivos Excel data original | 94 |
| Figura N° 22: Lotes de archivos generados para procesar..... | 94 |
| Figura N° 23: Base de datos poblada SQL Server..... | 95 |
| Figura N° 24: Base de datos poblada..... | 96 |
| Figura N° 25: Esquema de datos generados en SQL Server..... | 96 |
| Figura N° 26: ETL SQL Server..... | 96 |
| Figura N° 27: ETL Selección SQL Server..... | 97 |



| | |
|---|------------|
| Figura N° 28 : Script SQL para consulta que genera el formato deseado..... | 99 |
| Figura N°29: Script ejecutado desde R Project..... | 100 |
| Figura N° 30: Vista de datos obtenidos desde SQL en R con formato Series de Tiempo..... | 101 |
| Figura N° 31: ETL Script en R | 103 |
| Figura N° 32: Script para leer CSV..... | 105 |
| Figura N° 33: Librería para conexión ODBC..... | 105 |
| Figura N° 34: Datos antes de volcar a SQL Server..... | 106 |
| Figura N° 35: Datos listos en R para SQL Server | 106 |
| Figura N° 36: Modelo de procesos | 109 |
| Figura N° 37: construcción de los Modelos en R..... | 110 |
| Figura N°38: Librería Forecast código abierto en Github..... | 111 |
| Figura N°39: Librería Forecast código abierto en Github | 112 |
| Figura N° 40: Representación de la arquitectura capas de la red..... | 120 |
| Figura N° 41: Aalgoritmo para análisis y comparación de resultados entre tecnicas – ver anexo n° 3: plan de pruebas..... | 128 |
| Figura N° 42: Resultado de los algoritmos aplicados en los el modelado..... | 128 |
| Figura N° 43: Script HTML Y PHP..... | 133 |
| Figura N° 44: Ingreso a la aplicación a través de un usuario y una contraseña..... | 133 |
| Figura N° 45: evaluaciones de producción en porcentajes..... | 134 |
| Figura N°46: Pantalla de resultado de los modelos | 134 |
| Figura N°47: Resultados del análisis de los algoritmos..... | 135 |
| Figura N° 48: Muestra las simulaciones por meses..... | 135 |
| Figura N° 49: Data original en Excel..... | 143 |
| Figura N° 50: Ingreso a la aplicación a través de un usuario y una contraseña | 144 |
| Figura N° 51: Evaluaciones de producción en porcentajes..... | 144 |
| Figura N° 52: Pronósticos por periodos y tiempo..... | 145 |



| | |
|---|------------|
| Figura N° 53: Consolidado histórico..... | 145 |
| Figura N° 54: Resultados del análisis de los algoritmos..... | 146 |
| Figura N° 55: Se muestra los pronósticos..... | 146 |
| Figura N° 56: Muestra las simulaciones por periodos y meses..... | 147 |
| Figura N° 57: Entrenamientos de algoritmos..... | 147 |



INDICE DE GRAFICOS

| | |
|--|-----------|
| Gráfico N° 01: Evaluación con HW | 68 |
| Gráfico N° 02: Evaluación con ARNA..... | 70 |
| Gráfico N° 03: Evaluación con ARIMA | 72 |
| Gráfico N° 04: Tiempo empleado de algoritmos..... | 74 |
| Gráfico N° 05: CPU uso de algoritmo..... | 75 |
| Gráfico N° 06: Tiempo de generación de pronósticos en aplicación web..... | 76 |

INDICE TABLAS

| | |
|---|---------------|
| Tabla N° 01: Descripción de técnicas y algoritmos..... | 30 |
| Tabla N° 02: Modelos de minería de datos..... | 30 |
| Tabla N° 03: Producción de espárragos Periodo - Año 2012..... | 34 |
| Tabla N° 04: Producción de espárragos Periodo - Año 2013..... | 45 |
| Tabla N° 05: Producción de espárragos Periodo - Año 2014..... | 45 |
| Tabla N° 06: Producción de espárragos Periodo - Año 2015..... | 46 |
| Tabla N° 07: Determinación de variables..... | 47 |
| Tabla N° 08: Métodos y Técnicas de investigación..... | 59 |
| Tabla N° 09: Criterios de para los principios éticos..... | 60 |
| Tabla N° 10: Criterios de rigor científico..... | 61 |
| Tabla N° 11: Indicadores /Factores por Medida de Proyecto..... | 63 |
| Tabla N°12: Distribución de esfuerzo y tiempo de desarrollo por etapas..... | 64 |
| Tabla N°13: Tabla de Validación Holtwinters..... | 67 |
| Tabla N° 14: Tabla de Validación ARNA..... | 69 |
| Tabla N° 15: Tabla de Validación Arima..... | 71 |
| Tabla N° 16: Comparación de Resultados..... | 72 |
| Tabla N° 17: Procesamiento de modelo | 73 |
| Tabla N° 18: Tiempo de uso de CPU..... | 75 |
| Tabla N° 19: Tiempo de Procesamiento del aplicativo Web..... | 76 |
| Tabla N° 20: Comparación de Metodologías de Desarrollo de Modelo de Minería de Datos..... | 83 |
| Tabla N° 21: Periodo – Producción..... | 87 |
| Tabla N° 22: Descripción de la tabla de la base de datos..... | 91 |
| Tabla N° 23: Cantidad de cajas y Kg de espárragos..... | 93 |
| Tabla N° 24: Evaluación de las técnicas a utilizar..... | 107 |
| Tabla N° 25: Criterios de evaluación de las técnicas a utilizar..... | 108 |
| Tabla N° 26: representación de la técnica Holt-winters..... | 115 |



| | |
|--|------------|
| Tabla N° 27: Entrenando serie de tiempo..... | 116 |
| Tabla N° 28: Entrenamiento..... | 117 |
| Tabla N° 29: Verificando residuales..... | 117 |
| Tabla N° 30: Verificación de Coeficientes..... | 118 |
| Tabla N° 31: Entrenamiento de modelo | 118 |
| Tabla N° 32: Resultado de pronóstico | 118 |
| Tabla N° 33: Entrenamiento de modelo..... | 121 |
| Tabla N° 34: Verificación de entrenamiento..... | 122 |
| Tabla N° 35: Verificación de residuales..... | 122 |
| Tabla N° 36: Red Neuronal capas ocultas..... | 122 |
| Tabla N° 37: Plot Entrenamiento..... | 123 |
| Tabla N° 38 Resultado de Pronostico..... | 123 |
| Tabla N° 39: Serie de Tiempo con Arima..... | 125 |
| Tabla N° 40: Verificación de procedimiento Arima..... | 125 |
| Tabla N° 41: Verificación de Arima..... | 125 |
| Tabla N° 42: Verificación de grafico de Arima – Ploteo..... | 126 |
| Tabla N° 43: Resultado de pronóstico..... | 126 |
| Tabla N° 44: Prioridad y Dificultad de Historia de Usuario..... | 129 |
| Tabla N° 45: Esquema de Diario de Actividades..... | 130 |
| Tabla N° 46: Requerimiento 01..... | 130 |
| Tabla N° 47: Requerimiento 02 decisiones..... | 130 |
| Tabla N° 48: Requerimiento 03..... | 131 |
| Tabla N° 49: Requerimiento 04..... | 131 |



RESUMEN

Las técnicas en la Minería de Datos tiene un papel muy importante como tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento adquirido de grandes volúmenes de datos, así como para identificar tendencias y comportamientos en la información, que faciliten una mejor comprensión de los avances tecnológicos que nos rodean y ayudan en la toma de decisiones.(Molero, 2008). En este estudio, se evaluó las técnicas de minería de datos, como una herramienta computacional para predecir la producción de espárragos para fines del negocio y agrícolas en la localidad de Jayanca Lambayeque. Las técnicas predictivas evaluadas fueron: ARIMA, Holwinters, Redes Neuronales el análisis y comparación de los patrones en series temporales arrojó como resultados.

Se obtuvieron comparaciones significativas al contrastar los resultados de las predicciones de las técnicas estudiadas, con los datos observados. Los coeficientes de correlación más altos se obtuvieron con la técnica Arima. En conclusión, es viable el uso de técnicas de predicción para aplicarse en la producción de espárragos en la región Lambayeque

Palabras claves: *Tecnología, descubrimiento, conocimientos, análisis, producción*



ABSTRACT

Techniques in Data Mining have a very important role as support technology to explore, analyze, understand and apply the knowledge acquired from large volumes of data, as well as to identify trends and behaviors in information, which facilitate a better understanding of the technological advances that surround us and help in making decisions. (Molero, 2008). In this study, data mining techniques were evaluated, as a computational tool to predict the production of asparagus for business purposes and agricultural in the locality of Jayanca Lambayeque. The predictive techniques evaluated were: ARIMA, Holwinters, Neural Networks The analysis and comparison of the patterns in time series yielded as results.

Significant comparisons were obtained by comparing the results of the predictions of the techniques studied with the observed data. The highest correlation coefficients were obtained with the Arima technique. In conclusion, the use of prediction techniques to apply in the production of asparagus in the region is possibleLambayeque

keywords: Technology, discovery, knowledge, analysis, production

INTRODUCCIÓN

Hoy en día en todo el mundo en su mayoría de empresas producen gran cantidad de información cada día, de modo que se está enfrentando a la paradoja de que, cuantos más datos están disponibles, menos información se tiene. Para llegar a superar este problema en los últimos años se han generado técnicas y/o métodos que facilitan el procesamiento de los datos y permiten realizar un análisis más detallado de estos de forma automática; es aquí donde la minería de datos nos ofrece según (Rodríguez Rodríguez, 2010), “un conjunto de técnicas para el análisis de datos”.

El propósito de esta investigación es realizar un estudio de evaluación de las técnicas de minería de datos para los pronósticos de producción de espárragos, se analizó determino las variables utilizadas en los datos histórico de la producción de espárragos, se seleccionarán las técnicas predictivas para luego compararlas; así mismo también se construyó una aplicación web usando las técnicas de predicción, para luego evaluar los resultados obtenidos de la investigación.

En el lugar de aplicación se realizó en la localidad de Jayanca una empresa agro exportadora, en lo referente a la producción de espárragos aún existen inconsistencias de la producción, los cuales se pueden lograr solucionar tales inconsistencias utilizando las técnicas de pronóstico de minería de datos más eficientes.



CAPÍTULO I:
PROBLEMA DE LA INVESTIGACIÓN

1.1. Situación Problemática

Ante el crecimiento de la tecnología y la competencia, las empresas se preocupan por brindar el mejor servicio, que llene las necesidades del cliente y su satisfacción sea la más alta. Por lo que la calidad representa el factor más importante dentro del servicio ofrecido, en la actualidad las empresas agrícolas, poseen una gran cantidad de información registrada, en una base de datos de los cuales no poseen un modelo de predicción activo para obtener resultados y evaluar los meses a futuro. (Bagurskas, 2015)

La problemática radica en que la cantidad de producción o el aumento o disminución, ocasionando que los productores se vean obligados a mantener constante o aumentar su producción.

En donde presentan gran dificultad para saber qué cantidad producirán en los meses siguientes, debido a la variación de producción es por ello que solo se realiza un cálculo aproximado, con imaginaciones de producciones anteriores. En cuanto cada productor predice para saber su producción a futuro realizando cálculos por cantidad de área a cultivar, también a medida de la variación del tiempo en donde está generando dificultades para obtener una predicción exacta. En cuanto a la producción se realizó pronósticos por la fertilidad de los suelos, por lo que también hace que la producción aumente o disminuya, tomando también como un acceso a predecir la producción a través del clima, en la variación de la temperatura hace que la producción varíe. (González, 2005)

El problema que se pretende resolver es la predicción inexacta de la cantidad de producción de espárrago, el cálculo del pronóstico dentro de un ambiente imprevisto que proyecta una tasa de error relativamente alta ante los resultados obtenidos en la realidad, ciertas cantidades adquiridos por las empresas respecto a la producción real que genera inventarios en exceso o en falta, lo cual rebaja los niveles de servicio a los compradores, las exportaciones, siendo este un factor influyente para el éxito.



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 Holtwinter, así también si se emplea la regresión logística se puede determinar la ecuación que permite saber el cálculo del monto siguiente en la 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.

1.2. Formulación del problema

¿De qué manera se predice la cantidad de producción de esparrago?

1.3. Delimitaciones de la investigación

La investigación consiste en el desarrollo de un módulo analítico que permita estimar la variación de producción a través de un modelo de una serie de tiempos, en el proceso de consistencia de la producción, tomando como ámbito la una empresa agro exportadora, que comprende los productores (clientes) de una gran parte de Jayanca Lambayeque siendo un aproximado de 22000 clientes bajo análisis, se centra en la observación.

1.4. Justificación e Importancia de la investigación

a. CIENTÍFICA

Conociendo que la Minería de Datos se fundamenta en la búsqueda de patrones dentro de grandes bases de datos, utilizando diversos métodos tanto de estadística como de inteligencia artificial, haciendo uso de recursos informáticos y tecnológicos, en el presente proyecto se busca aprovechar los beneficios de la misma con el fin de extraer información y generar conocimiento con el fin de pronosticar la producción de espárragos.

b. INSTITUCIONAL

El sistema basado en inteligencia de negocios para analizar los datos para pronosticar la Producción de Espárragos con la Aplicación de técnicas de minería de datos en el Perú, no se desarrolla para una empresa específica, sino para todas las empresas de la región Lambayeque que buscan este tipo de información.

c. SOCIAL

El desarrollo de una investigación basado en inteligencia de negocios para analizar los datos de producción de Espárragos con la Aplicación de técnicas de minería de datos en el Perú, resulta importante para nuestro país debido a que se obtendrán los pronósticos en la producción del mencionado alimento.

1.5. Limitaciones de la Investigación

La investigación consiste en el estudio de técnicas de minería de datos, a través de un modelo de series de tiempo con las técnicas de holt-winter Arrima y Red Neuronal, con la producción de espárragos en la zona de Jayanca Lambayeque. El alcance para el desarrollo del modelo comprende en este espacio, y se dará en el periodo Agosto 2016 – Diciembre 2016 para mostrar la implementación del prototipo y posterior análisis de resultado.

La solución comprende un modelo predictivo comparativo, mostrando resultado en una aplicación web.

1.6 Objetivos de la Investigación

1.6.1. Objetivo general

Aplicar técnicas de minería de datos para el pronóstico de la producción de espárragos.

1.6.2. Objetivos específicos

- a. Recopilar información histórica del estado del proceso para el pronóstico de producción de espárragos
- b. Analizar los algoritmos en la fase de modelado predictivo.
- c. Desarrollar el modelo para la predicción de la producción de espárragos.
- d. Implementar una aplicación web para mostrar resultados de pronósticos

CAPITULO II.
MARCO TEÓRICO.

2. Marco teórico

2.1 Antecedentes de la Investigación

Métodos y técnicas de inteligencia computacional y minería de datos para la toma de decisiones en explotación de campos maduros.

Autor: Rafael Rueda Reyes, Año: 2014, Institución: Instituto Mexicano Del Petróleo, México.

Se trata del Modelado Inverso Inteligente de Yacimientos Naturalmente Fracturados (MIIDY), un modelo predictivo de yacimiento que parte del comportamiento histórico, aplicando técnicas de minería de datos.

Esta metodología se basa en modelos dirigidos por datos y utiliza técnicas de minería de datos, ya que el comportamiento se aprende de los datos reales e históricos y no de los modelos matemáticos acerca de nuestro conocimiento actual. (Leonid Sheremetov, 2014).

La relación de esta tesis con la que se desarrolló es que se trata de demostrar la importancia de la aplicación de técnicas de minería de datos para pronosticar la producción de espárragos en el tiempo, así como en la predicción de otras variables dinámicas.

Minería de Datos como soporte a la toma de decisiones empresariales.

Autor: Yelitza Josefina Marcano Aular y Rosalba Talavera Pereira, Año: 2014, Institución: Universidad del Zulia, Venezuela.

La tarea por mejorar el acceso a la información está cobrando cada vez más fuerza, especialmente en los negocios actuales, donde se requiere principalmente de procesos basados en el recurso información, de manera automatizada y reutilizable. Entre los beneficios que ofrece la técnica está en la posibilidad de elevar los niveles de competencia de los negocios, basándose en la rapidez para identificar, procesar y extraer la información que realmente es importante, descubriendo conocimiento y patrones en bases de datos.

La relación está en que la minería de datos nos permite obtener información de manera automatizada y reutilizable de forma rápida para identificar, procesar y extraer la información que realmente es importante, descubriendo conocimiento y patrones importantes en el pronóstico de la producción de espárragos.

Minería de Datos aplicados a las ventas con Tarjeta de Crédito realizados en las tiendas Saga Falabella.

Autor: Hober Willy Siccha Vega. Año 2012, Institución: Universidad Tecnológica del Perú.

Esta investigación se centra con enormes datos que actualmente se generan para su análisis y búsqueda de fundamentos, probar hipótesis, el muestreo, la teoría de límite central, la teoría de la estimación, la regresión, el análisis de varianza, el diseño de experimentos.

Las cantidades de información en la actualidad son tan enormes que es prácticamente imposible su asimilación por una sola persona, por lo que se hace necesario contar con nuevos métodos de procesamiento de datos, nuevas tecnologías que nos permitan y nos faciliten el proceso de búsqueda del conocimiento escondido al interior de los enormes datos históricos y de estos datos existentes nos proporcionen la esencia contenida en la base de datos.

Esta investigación es determinar el comportamiento a futuro y la naturaleza de los datos históricos de ventas con tarjeta de crédito en las tiendas de Saga Falabella de la ciudad de Lima a través de la explotación de las técnicas de minería de datos, con la finalidad de ayudar a los miembros de la alta dirección a analizar los hábitos de los clientes a fin de satisfacer mejor su demanda, mejorar la administración de los inventarios de los productos que están asociados a las transacciones de ventas y mejorar los volúmenes de ventas.



Implementación de un modelo de Minería de Datos para mejorar la toma de decisiones comerciales en la empresa Star Perú S.A.C.

Autor: López López Gustavo & Vélez Rojas Emerson, Año: 2009, Institución: Universidad Nacional del Santa, Chimbote.

En esta investigación se logró recolectar los datos necesarios, se analizaron para luego construir el modelo de minería de datos identificando, los atributos de entrada y de predicción de empleados, así como de clientes además de la identificación de patrones para a través de los resultados ayudar a mejorar la toma de decisiones en la empresa.

La relación con esta tesis son los métodos utilizados para obtener la información necesaria para el pronóstico de la producción de espárragos aplicando las técnicas de la minería de datos a través del DataWarehouse.

Pronóstico de la demanda de bienes usando redes neuronales.

Autor Arturo Jesús Gálvez Aguilar, año 2012.

La presente investigación tuvo como finalidad construir un modelo de red neuronal para pronosticar la demanda de los productos de una empresa industrial o comercializadora dado que ellos necesitan estos resultados para poder planificar sus actividades en otras áreas tales como: Marketing, Producción, Finanzas, etc.; para el caso de la investigación los resultados se presentan en un horizonte de pronóstico a corto plazo; a su vez si se llega a demostrar la hipótesis que el modelo pronostica con error mínimo.

En el caso de la investigación es Limpiar y seleccionar la base de datos. Recolectar datos históricos del consumo de los clientes, Realizar una exploración de los datos recolectados, detectando y corrigiendo anomalías en la data recolectada (valores extremos, ruidosos, inconsistentes). Construir modelos basados en Redes neuronales del tipo de Retro propagación, Entrenar y ajustar los parámetros de cada red neuronal.

2.2 Estado del arte

La minería de datos no es un concepto nuevo, es más, data de la década del 70, pero si ha cobrado un gran interés y avance en los últimos años, a través de la mejora de las técnicas que utiliza para el análisis de los datos.

Año (2014) PRONOSTICO DE LA DEMANDA DE BIENES USANDO REDES NEURONALES (ARTURO JESUS GALVEZ 2014) La presente tesis tiene como finalidad construir un modelo de red neuronal para pronosticar la demanda de los productos de una empresa industrial o comercializadora dado que ellos necesitan estos resultados para poder planificar sus actividades en otras áreas tales como: Marketing, Producción, Finanzas.

Para la realización de este proceso se ha subdividido en tres subprocesos: Eliminación de valores vacíos Eliminación de valores redundantes Ajuste de datos

Limpieza de datos. En este proceso es eliminar las filas que tienen en alguna de sus filas valores nulos.

Se calcula el min Max de los datos de entrada; es decir el valor mínimo y máximo de cada dupla, Se crea una red neuronal cuya arquitectura cuenta con 10 neuronas en la capa intermedia y 1 neurona en la capa de salida; este número de neuronas es constantemente modificado por el tesista para lograr el óptimo modelo, Se entrena a la red para obtener resultados que nos permitan aseverar si nuestra red llega a aprender o no y cuanto es el error final, Se calcula el error de los datos pronosticados comparándolos con el valor real de la siguiente manera:

La hipótesis de la investigación pues el MAPE es menor al 1%. Las redes neuronales del tapón de oído y las del antejo 3M de los experimentos anteriores son equivalentes en cuanto a arquitectura pues ambas tienen 10 neuronas en la capa intermedia, y además los datos de entrada de los dos experimentos pasan sobre el mismo proceso de desarrollo de la red pero se diferencian en que la primera llega más rápida y eficientemente a la meta porque tiene datos históricos

con un comportamiento más predecible que la segunda tanto es así que tiene una pequeñísima cantidad de días sin ventas, aproximadamente el 2\% del total de la data histórica tomada en el experimento

En el año 2015 “Modelo de Minería de Daros para Identificación de Patrones Que Influye En el Aprovechamiento Mexicano”(Jaime A H.C 2015) realiza una investigación de pronósticos utilizando redes neuronales de SQL server en donde identifica factores de alumnos con buenas o malas perspectivas de aprovechamiento académico, calcula la probabilidad de alumno y clasifica los diferentes atributos de los alumnos

A través de red neuronal de sql server la técnica arboles de decisiones de sql server, algoritmo de clasificación y de regresión con el modelo de predicción de atributos con una regresión, también utiliza la técnica de agrupamiento o clustering de SQL server que realiza una segmentación de datos.

El modelo más efectivo para esta investigación fue el modelo de red neuronal ya que tiene un mejor comportamiento respecto a los modelos de árboles y cluster k- mediana se asocia a la metodología CRISP-DM

Año 2016“Aplicación de técnicas de minería de datos e inteligencia artificial a datos de espectrometría de masas para el descubrimiento de conocimiento” (H. López-Fernández 2016) MALDI-TOF fueron estudiados y comparados. Además, se desarrolló un algoritmo de emparejamiento de picos llamado Forward, el cual fue utilizado en todos los desarrollos y colaboraciones. Incluye la comparación de más librerías disponibles públicamente, así como la inclusión de más conjuntos de datos. Durante el curso de la investigación, la técnica de agrupamiento doble o biclustering se aplicó para en análisis de datos de MALDI-TOF, siendo capaz de extraer información útil y generar nuevas hipótesis. Su adecuación fue evaluada comparándola contra el agrupamiento jerárquico empleando dos conjuntos de datos reales. Aunque los resultados fueron prometedores, se debe continuar trabajando en esta línea en el futuro para profundizar y expandir este estudio.

Su adecuación fue evaluada comparándola contra el agrupamiento jerárquico empleando dos conjuntos de datos reales. Aunque los resultados fueron prometedores, se debe continuar trabajando en esta línea en el futuro para profundizar y expandir este estudio.

2.3 Bases Teórico Científicas

Las técnicas de minería de datos se emplean para mejorar el rendimiento de los procesos del negocio o industriales en los que se maneja grandes volúmenes de información estructurada y almacenada en grandes bases de datos. Por ejemplo, se usan con éxito en aplicaciones de control de procesos productivos, como herramienta de ayuda a la planificación y a la decisión en marketing, finanzas, etc (Rodríguez, 2013)

La minería de datos lo define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.(Wintten & Frank, 2000)

Por lo tanto, dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil con todo, la minería de datos será un paso en el proceso de descubrimiento de conocimiento, consistiendo en la aplicación de algoritmos particulares (métodos) que bajo algún objetivo aceptable, para producir una enumeración de patrones (modelos) sobre los datos. Se aplican para ello técnicas estadísticas y de inteligencia artificial (algoritmos) para descubrir patrones e irregularidades en los grandes volúmenes de datos. Es, por tanto, una tecnología que utiliza técnicas conocidas.(Trujillo, Mozon, & Pardo, 2011).

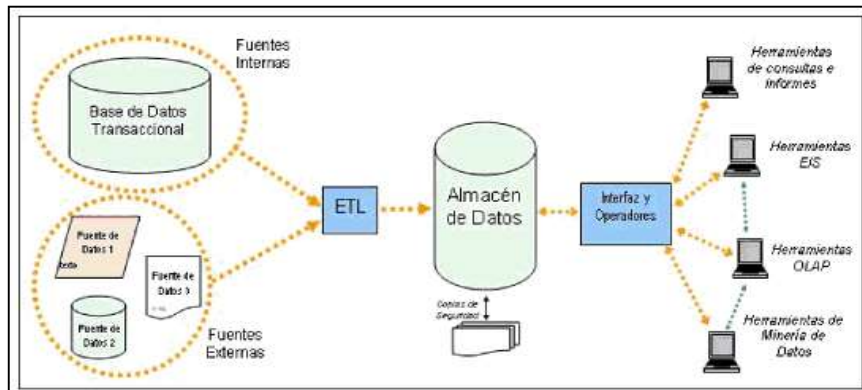
2.3.1. Minería de Datos

Es una actividad de extracción cuyo objetivo es el de descubrir hechos contenidos en las bases de datos históricos.

Es un proceso no trivial que tiene como entrada datos y como salida Información, en este proceso se hace un análisis detallado a través del uso de algoritmos para descubrir patrones o comportamiento de los datos.

La minería de datos es un miembro clave en la familia de productos de Business Intelligence (BI), junto con el procesamiento analítico en línea (OLAP), los informes empresariales y ETL (Cargas, transformación y extracción de datos). La minería de datos trata de analizar los datos y la búsqueda de patrones ocultos utilizando métodos automáticos o semiautomáticos. Durante la última década, grandes volúmenes de datos se han almacenado en las bases de datos, gran parte de estos datos proviene de software de negocios, tales como Aplicaciones Financieras, Planificación de Recursos Empresariales (ERP), Gestión de la Relación con Clientes (CRM), y Registros web. El resultado de este proceso ha convertido a las organizaciones ricas en datos e información, pero pobres en conocimientos, llegando a alcanzar colecciones de datos tan grandes que el uso práctico de estos almacenes se ha convertido en limitada. El objetivo principal de la Minería de Datos es extraer patrones ocultos a partir de estos datos, aumentando su valor intrínseco y la transferencia de los datos al conocimiento.(Bustos, 2011)

Figura N° 1: Proceso de minería de datos



Fuente: “Minería de Datos: Técnicas y Herramientas”(DMC.SAC, 2010).

2.3.1.1. Métodos de Minería de Datos

Los dos caminos principales de la minería de datos hacen referencia a la predicción y a la descripción. Para ambos existen una variedad de métodos que se pueden utilizar, con el fin de descubrir conocimiento. Dentro de los métodos predictivos se encuentran la clasificación y regresión, por otra parte en los descriptivos se tienen el *clustering* y las reglas de asociación.(Alvarez, 2012)

a. Clasificación

Se define como la identificación de características o atributos que hacen que un elemento se vincule a un grupo siguiendo un patrón de datos. Este último se puede utilizar para predecir cómo se comportarán nuevas instancias.

b. Regresión

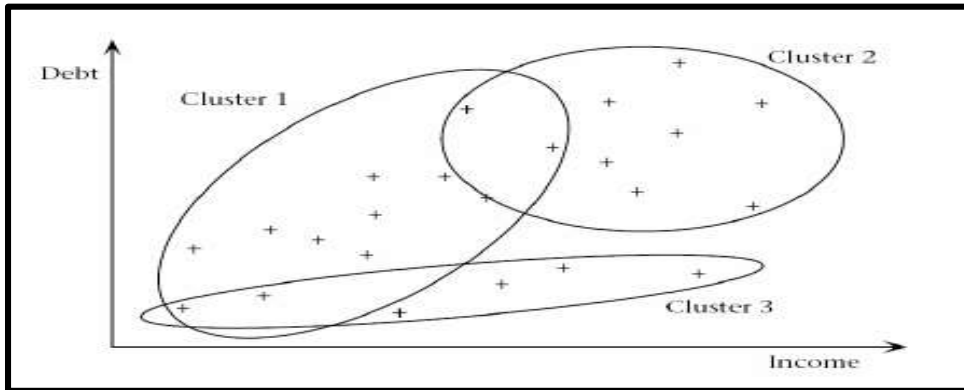
Función que le asigna a un elemento un valor real, utilizando valores existentes para predecir datos futuros. Las regresiones se pueden utilizar por ejemplo para predecir comportamiento de la demanda futura, utilizando las ventas o el consumo pasado.

c. Clustering

Divide el conjunto de datos en grupos que son muy diferentes unos de otros, pero cuyos elementos sean muy similares entre sí. Es un método descriptivo que identifica un grupo de categorías o “*clústeres*” para describir los datos.



Figura N° 02: Cauterización de Datos



Fuente: "Técnicas de Minería de Datos para la Retención de Clientes en el Sector Asegurador" (Rodríguez, 2013)

d. Reglas de asociación

Son otro instrumento descriptivo, donde el objetivo es encontrar relaciones significativas entre los datos, utilizando probabilidades de ocurrencia de dos objetos. Un claro ejemplo es el análisis de los artículos o productos de una canasta de compras en una tienda. (García, 2010)

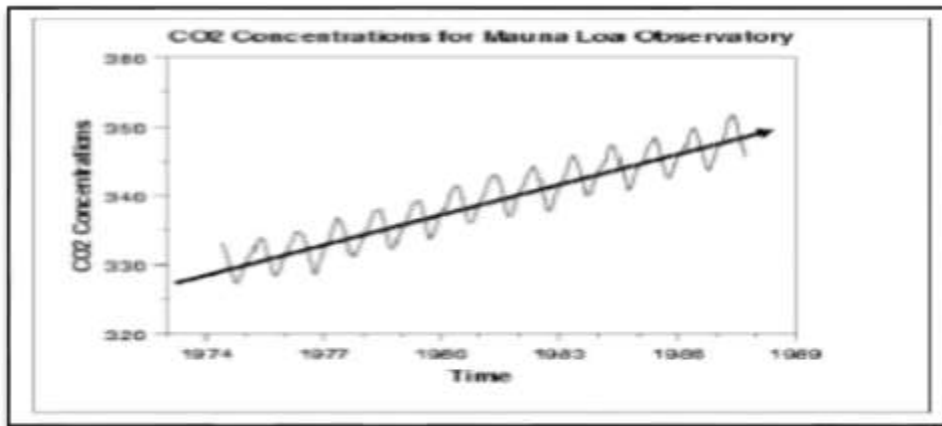
e. Series de Tiempo

Representan el conjunto de observaciones hechas con respecto a una variable en períodos de tiempo determinados. Permite realizar pronósticos en el tiempo con respecto a la variable evaluada. Es una secuencia ordenada de valores de una variable en intervalos de tiempo periódicos y consecutivos. Algunas definiciones que se usan con esta técnica son:

f. Tendencia

Patrón de comportamiento de los elementos en un entorno particular durante un periodo de tiempo. Si los datos muestran una tendencia, se ajustan con algún tipo de curva o recta y modelar los residuales. (Hossein, 1994-2015)

Figura N° 03: Gráfico de Tendencia de un conjunto de datos de los años 1974-1989

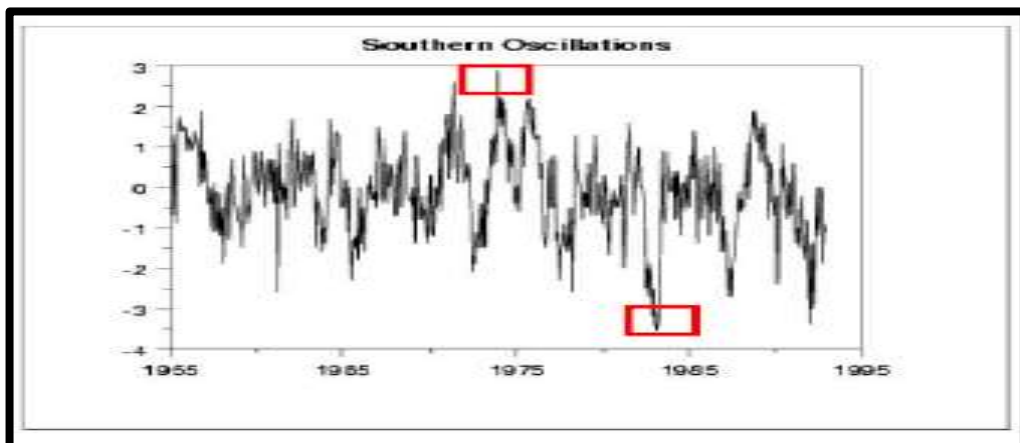


Fuente: Minería de datos con Aplicaciones (Cruz Arrela Liliana, 2010)

g. Estacionalidad

Son fluctuaciones periódicas, cuando ocurren picos o bajas en determinados períodos. La serie de tiempo muestra un incremento en algunos meses y una declinación en otros.(Hossein, 1994-2015)

Figura N° 04: Gráfica de valores en el tiempo, donde se observa la estacionalidad



Fuente: Minería de datos con Aplicaciones (Cruz Arrela Liliana, 2010,).



2.3.1.2. Técnicas de Minería de Datos

Se dice comúnmente que el proceso de minería de datos convierte datos en conocimiento, tal cual un alquimista pudiera convertir espigas de trigo en lingotes de oro. Por si esto no fuera poco, en algunos casos se llega a decir que el objetivo es extraer “verdad a partir de basura”.

Figura N° 05: Proceso ideal de Minería de Datos



Fuente: “Introducción a la Minería de Datos”.(Ruelas Santoyo & Laguna González, 2013)

En la figura las técnicas de minería de datos aparecen como una especie de colador pasapurés que, al introducirle los datos (en forma de vista minable junto a ciertos elementos asociados) produce, sin grumos ni atasco alguno, una serie de patrones lustrosos y relucientes.

Ciertamente las cosas no son tan simples como en la figura. Hemos de pensar que el colador pasapurés debe ser algo mágico o, al menos, muy sofisticado en su interior. Como la magia y la prestidigitación requieren muchas más horas de aprendizaje que la minería de datos, hemos de afrontar la realidad: en general, los procesos que extraen patrones a partir de datos son computacionalmente costosos y, lo que es más importante, son más costosos cuanto más expresivos, novedosos, comprensibles e interesantes queramos que sean los patrones extraídos.(Ruelas Santoyo & Laguna González, 2013)

De hecho, cuando algún algoritmo obtiene malos resultados no se debe a que la mayoría de investigadores de numerosas universidades y centros de investigación que han trabajado durante las últimas décadas en realizarlo y

perfeccionarlo sean unos ineptos. Es mucho más probable que se trate de que, o bien no existe un patrón en los datos, o bien no estemos utilizando la herramienta adecuada para encontrarlos, o bien el patrón sea realmente difícil de encontrar. (Dongre, Prajapati, & Tokekar, 2014)

Existen tareas, presentaciones de tareas, instancias de tareas, que son más sencillas que otras. Por ejemplo, la extracción de reglas de asociación es un problema más sencillo, computacionalmente hablando, que la clasificación. Esto quiere decir que, para los mismos datos, generalmente deberemos esperar menos tiempo a que un algoritmo de extracción de reglas de asociación acabe que un algoritmo de clasificación. Del mismo modo, dos problemas de clasificación, con el mismo número de ejemplos, tipos de atributos y números de clases pueden diferir en dificultad. Dependerá de los intrincados que estén los patrones en los datos o si realmente existen patrones plausibles en ellos. (Vega, 2012)

Además de los datos y de la tarea, existen otros aspectos que influyen en el aprendizaje, que suelen denominarse conjuntamente vías. Quizás las vías que más influye en esta complejidad sea la manera de expresar o definir los patrones (vías del lenguaje). Por ejemplo, no es lo mismo una regresión lineal que una regresión realizada por una red neuronal multicapa. Ambos métodos permiten realizar la misma tarea, pero la expresividad y la mayor capacidad de la segunda se paga, de alguna manera, con un mayor tiempo de espera para obtener el modelo. El conocimiento previo es otro tipo de bias, que puede ayudar a refinar el espacio de búsqueda, la clasificación inicial de las técnicas de minería de datos distingue entre técnicas predictivas, en las que las variables pueden clasificarse inicialmente en dependientes e independientes (similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante), técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus (similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante) y técnicas auxiliares. (Pérez López & Santén González, 2008)

2.3.1.2.1. Técnicas Predictivas

Especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido.

Formalmente, la aplicación de todo modelo debe superar las fases de identificación objetiva(a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos), estimación(proceso de cálculo de los parámetros del modelo elegido para que los datos en la fase de identificación), diagnóstico(proceso de contraste de la Valdez del modelo estimado) y predicción (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes).(Tello, Eslava, & Tobias , 2012)

En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después del DataMining y también debe contrastarse antes de aceptarse como válido.(Pérez López & Santén González, 2008).

- a. **ARIMA:** (Modelo autorregresivo integrado de media móvil). Los modelos autor regresivos o de medias móviles es un proceso estocástico es una sucesión de variables aleatorias Y_t ordenadas, pudiendo tomar t cualquier valor entre. Por ejemplo, la siguiente sucesión de variables aleatorias puede ser considerada como proceso estocástico: Programa Citius.- Técnicas de Previsión de variables financieras

$$Y_5, Y_4, Y_3, Y_2, \dots, Y_3, Y_4$$

El subíndice t no tiene, en principio, ninguna interpretación a priori, aunque si hablamos de proceso estocástico en el contexto del análisis de series temporales este subíndice representará el paso del tiempo. Cada una de las variables Y_t que configuran un proceso estocástico tendrán su propia función de distribución con sus correspondientes momentos. Así mismo, cada par de

esas variables tendrán su correspondiente función de distribución conjunta y sus funciones de distribución marginales. Esto mismo ocurrirá, ya no para cada par de variables, sino para conjuntos más amplios de las mismas. De esta forma, para caracterizar un proceso estocástico deberíamos especificar las funciones de distribución conjunta de cualquier conjunto de variables:(Rafael Arce)

b. Método de HOLT-WINTERS

El filtro lineal conocido como método de Holt-Winters es una variante del alisado exponencial doble de Holt diseñado para realizar predicciones en series con tendencia aproximadamente lineal y con clara influencia de la componente estacional. Dependiendo del esquema de agregación elegido para la tendencia y la componente estacional, se habla del método Holt-Winters multiplicativo o aditivo. En ambos casos, la componente irregular interviene aditivamente en el modelo.(MONTERO, 2007)

En el caso multiplicativo tiene la forma $y_t = T_t E_t$, y considerando la tendencia aproximadamente lineal $T_t = a_1 + b_1 t$, el modelo conjunto resulta ser:

$$y_t = (a_1 + b_1 t) \cdot E_t + 1_t$$

Dónde:

a_1 es la ordenada en el origen de la serie.

b_1 es la pendiente.

E_t es el factor estacional multiplicativo.

El método de Holt-Winters consiste en tres ecuaciones de alisado, una por cada parámetro:

- La ordenada en el origen se estima mediante la ecuación:



$$a_t = \alpha \frac{Y_t}{E_{t-L}} + (1 - \alpha)(a_{t-1} + b_{t-1})$$

Que es semejante a la del método del Holt, salvo que en lugar del valor original y_t se utiliza el valor <<desestacionalizado>> $\frac{y_t}{E_{t+L}}$ 9

- La pendiente se estima mediante la ecuación

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$$

que coincide con la del método de Holt.

- Por último, el factor estacional se estima mediante la ecuación

$$E_t = \gamma \frac{y_t}{a_t} + (1 - \gamma)E_{t-L}$$

Donde se utiliza la serie sin tendencia, para que esta última no afecte a la estimación de los factores. Es decir, el factor estacional se obtiene a partir de una serie en la que se ha eliminado la tendencia. (Caridad y Otero, 2013).

El modelo Holt-Winters incorpora un conjunto de procedimientos que conforman el núcleo de la familia de series temporales de alisado exponencial. Holt-Winters puede adaptarse fácilmente a cambios y tendencias, así como a patrones estacionales. En comparación con otras técnicas, como ARIMA, el tiempo necesario para calcular el pronóstico es considerablemente más rápido. Esto significa que cualquier usuario puede poner en práctica la técnica de Holt-Winters. Más allá de sus características técnicas, su aplicación en entornos de negocio es muy común. De hecho, Holt-Winters se utiliza habitualmente por muchas compañías para pronosticar la demanda a corto plazo cuando los datos de venta contienen tendencias y patrones estacionales de un modo subyacente. (Luna, 2002)

Esta técnica se basa en la atenuación de los valores de la serie de tiempo, obteniendo el promedio de estos de manera exponencial; es decir, los datos se



ponderan dando un mayor peso a las observaciones más recientes y uno menor a las más antiguas. La expresión para realizar el cálculo de la suavización exponencial es:(Acosta Cervantes, Villareal Marroquín , & Cabrera Ríos, 2013)

Dónde:

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

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

P_{t+1} : Pronóstico o predicción para el periodo “t+1”

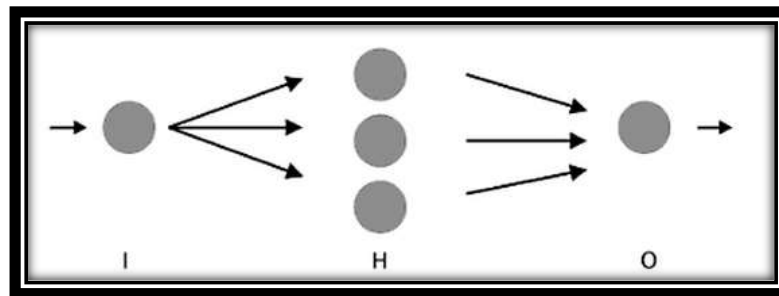
P_t : Pronóstico o predicción en el periodo “t”.

α : Factor de suavización ($0 \leq \alpha \leq 1$)

- c. Redes Neuronales:** Son métodos inspirados en el funcionamiento del cerebro humano, en particular en la forma cómo las neuronas reaccionan y propagan estímulos formando una red neuronal, o neuronal. La capacidad humana de reconocer patrones e identificar clases justificó el desarrollo de una metodología general para la identificación de patrones (clasificación supervisada o predicción de clase). Hay tres tipos básicos de redes neuronales: perceptrón, función de base radial y mapas auto organizables. (Vieria Braga, Ortiz Valencia, & Ramirez Carvajal, 2009)



Figura N° 06: Red Neuronal



Fuente: Introducción a la Minería de Datos

Hay dos tipos principales de aprendizaje en RNA:

- Aprendizaje supervisado:** Con este tipo de aprendizaje, proporcionamos a la red un conjunto de datos de entrada y la respuesta correcta. El conjunto de datos de entrada es propagado hacia adelante hasta que la activación alcanza las neuronas de la capa de salida. Entonces podemos comparar la respuesta calculada por la red de aquella que se desea obtener, el valor real, objetivo o “blanco” (de target, en inglés). Entonces se ajustan los pesos para asegurar que la red produzca de una manera más probable una respuesta correcta en el caso que se vuelva a presentar el mismo o similar patrón de entrada. Este tipo de aprendizaje será útil especialmente para las tareas de regresión y clasificación. (Sanchez, 2010)
- Pre procesamiento:** El análisis y limpieza de los datos son las líneas principales a seguir en esta sección, donde se produce el tratamiento de valores ausentes (missing), los valores fuera de rango (outliers). Para ello, se emplean distintas técnicas de imputación de datos que van desde un tratamiento valor a valor (simple imputation) hasta un reemplazo contemplando múltiples variables y sus valores (multiple imputation). (Rios, 2013)

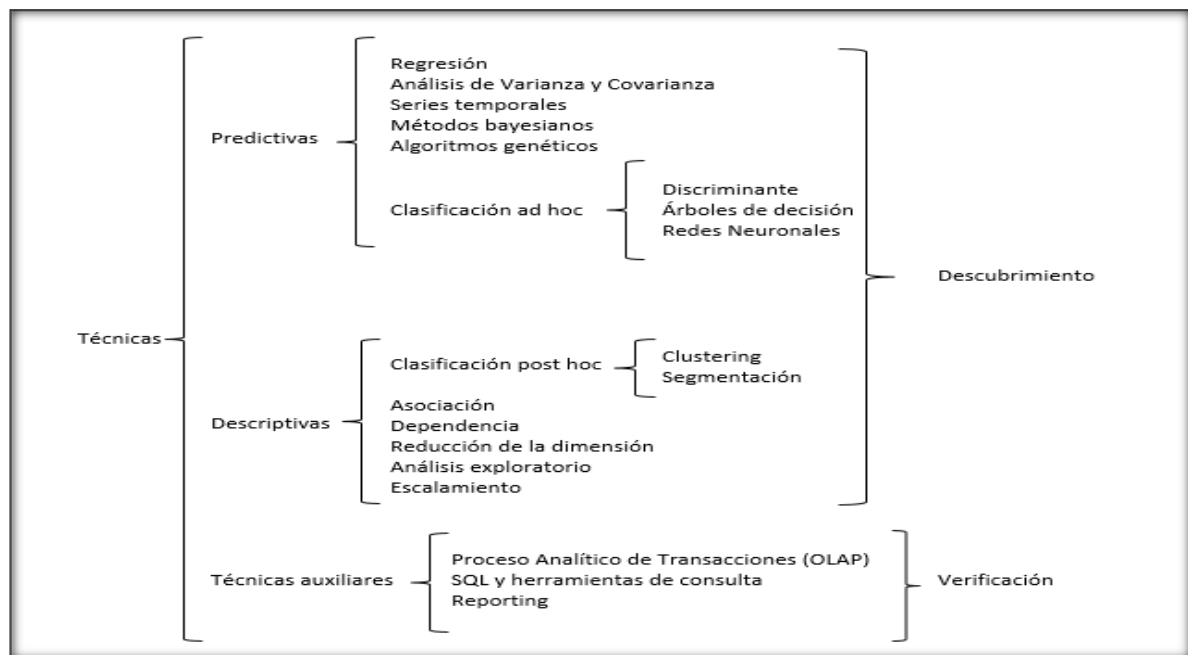
2.3.1.2.2. Técnicas Descriptivas

En estas técnicas no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones.(Pérez López & Santén González, 2008)

2.3.1.2.3. Técnicas Auxiliares

Son herramientas de apoyo más superficiales y limitadas. Se trata de nuevos métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la verificación.(Pérez López & Santén González, 2008).

Figura N° 7: Técnicas de Data Mining



Fuente: Minería de Datos – Técnicas y Herramientas

También las Técnicas de clasificación pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, arboles de decisión y redes neuronales) como a las descriptivas (clustering y segmentación).



Las técnicas de clasificación predictivas suelen denominarse técnicas de clasificación ad hoc ya que clasifican individuos u observaciones dentro de grupos previamente definidos. Las técnicas de clasificación descriptivas se denominan técnicas de clasificación post hoc porque realizan clasificación sin especificación previa de los grupos. (Pérez López & Santén González, 2008).

2.3.1.3. Evaluación de las técnicas de minería de datos

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

Tabla N° 1: Descripción de técnicas y algoritmos

| Técnicas de Minería | Descripción | Algoritmos | ¿Apto para investigación? |
|----------------------|---|--|---|
| Regresión | Modelos de 2 variables | Holtwinter Arima Máquinas de Vectores de Soporte | Si. Solo se usa consumo y periodo en análisis |
| Clasificación | Basado en reglas por construcciones lógicas múltiples variables | Árbol de decisiones | No |
| Asociación | Hechos en común para determinado grupo de datos múltiples variables | A priori FP-Growth Éclat | No |
| Agrupación | Agrupación de series de vectores en un mapa de dispersión | K means | No |

Fuente: Elaboración Propia



Tabla N° 2: Modelos de minería de datos

| Modelo de minería de datos para pronósticos con series de tiempo (Modelos de Regresión) | Holtwinter | Máquinas de Vectores de Soporte (SVM) | Arima |
|--|--------------------|--|--------------|
| Evaluación fundamento teórico | | | |
| <i>Modelo parametrizado</i> | SI | NO | SI |
| <i>Datos estacionales</i> | SI | SI | SI |
| <i>Método estadístico</i> | SI | NO | SI |
| <i>Capacidad iterativa (Aprendizaje)</i> | NO | SI | NO |
| <i>Cantidad de datos de la serie</i> | 24 | 3 | 80 |
| Evaluación fundamento computacional | | | |
| <i>Procesamiento CPU</i> | Mínimo | Medio | Mínimo |
| <i>Consumo RAM</i> | Mínimo | Medio | Mínimo |
| <i>Tiempo computacional</i> | Mínimo | Medio | Mínimo |
| Evaluación fundamento objetivo del modelo | | | |
| <i>Confiabilidad de precisión pronostico</i> | Después de pruebas | Después de pruebas | |
| <i>Confiabilidad de precisión consistencias</i> | Después de pruebas | Después de pruebas | |

Fuente: Elaboración Propia

- **Data Warehousing**, en la primera etapa del KDD (Integración), se requieren fuentes de información consolidadas, por ello, es que generalmente se aplica este procedimiento posterior a la implementación de un data warehouse (DWH) en la compañía.



Este concepto se define como la colección de tecnologías de soporte decisivo que permite al trabajador tomar buenas y rápidas decisiones; Esta colección debe ser orientada al sujeto, integrada, variante en el tiempo y estable, por ende, generalmente, se mantiene apartada de las bases de datos operacionales, pues se busca la consolidación de los datos, por lo tanto, es lógico pensar que un data warehouse contiene datos consolidados a partir de múltiples bases de datos operacionales, durante extensos períodos de tiempo, por lo que es común que su tamaño alcance varios gigabytes o terabytes. Es importante destacar que cada estructura en un data warehouse posee una dimensión temporal (Kimball, 1998)

2.3.2. Producción de espárragos

Según la FAO en el año 2015), en el mundo, sólo Perú y Tailandia logran producir espárrago durante todo el año. En el resto de países, la producción es muy estacional, concentrándose en numerosos países entre abril y junio. En los meses de septiembre hasta febrero son pocos los países abastecedores, la producción de espárrago se incrementa notoriamente en los meses de agosto a marzo, disminuyendo un poco en los meses restantes por la baja de temperatura (debido al invierno). Sin embargo, cabe precisar que se tienen dos campañas para el espárrago verde, una inicial de enero a junio y la principal de septiembre a diciembre realizándose las exportaciones de acuerdo con las ventanas en los mercados de destino y el saldo de la producción es envasado en conservas o congelado.(Carmer Tejada, 2015)

En los años 2007 y 2008 la producción de espárragos ha ido creciendo en el porcentaje del 6.8%, aunque las exportaciones cayeron en 2.1%, la producción de nuestro país, que alcanzó su record en el 2011 (con 392,306 mil toneladas), se concentra en la costa, principalmente en los departamentos de Ica y La Libertad, y está constituida en su mayor parte por el espárrago verde. El producto peruano es de gran calidad, debido a las características del clima, que resulta



óptimo para su cultivo y que además permite obtener altos rendimientos durante todo el año. 2012

Si bien China es el primer productor mundial, no es un país exportador, dado que consume toda su inmensa producción. El primer exportador es el Perú, que el año 2012 le vendió al mundo 225,320 toneladas (el 60% de su producción) por un total de US\$ 531 millones (y el 2013 acaba de venderle 168 mil toneladas, por un valor de US\$ 547 millones). Del monto total exportado en el 2012, el 64% correspondió a espárragos frescos o refrigerados, 26% a preparados o en conserva y 10% a congelados. El principal mercado para nuestro producto es Estados Unidos, país al que le siguen en importancia varios europeos, entre ellos España, Holanda, Francia y Reino Unido (COMEXPERU, 2010)

Figura N° 08: producción de espárragos



Fuente: Cultivo de espárragos (HACH América 2012)

2.3.3. Aplicación Web

En los primeros días de la Web, los sitios Web consistían de páginas estáticas, permitiendo una interacción limitada con el usuario. Al comienzo de los años 90, estas limitaciones fueron superadas cuando los servidores Web fueron reemplazados para permitir comunicaciones a través del desarrollo de

fragmentos de código que eran ejecutados del lado del servidor. A partir de entonces las aplicaciones dejaron de ser estáticas y solamente editadas por aquellos “gurúes” del HTML y se permitieron a usuarios normales interactuar con las aplicaciones por primera vez, también se dice que una aplicación web es un tipo especial de aplicación cliente/servidor, donde el cliente (navegador, explorador o visualizador), el servidor (el servidor web) y el protocolo mediante el que se comunican, están estandarizados y no han de ser creados por el programador de aplicaciones. La característica común que comparte todas las aplicaciones web es el hecho de centralizar el software para facilitar las tareas de mantenimiento y actualización de grandes sistemas. Es decir, evitar tener copias de las aplicaciones en todos los puestos de trabajo, lo que puede convertirse en una pesadilla a la hora de distribuir actualizaciones y garantizar que en todos los puestos de trabajo funcione correctamente. (Mora, 2001)

2.3.3.1. Metodologías de Desarrollo de Software

A. XP

La programación extrema o extreme Programming (XP) es una metodología de desarrollo de la ingeniería de software formulada por Kent Beck, autor del primer libro sobre la materia, *Extreme Programming Explained: Embrace Change* (1999). Es el más destacado de los procesos ágiles de desarrollo de software, al igual que éstos, la programación extrema se diferencia de las metodologías tradicionales principalmente en que pone más énfasis en la adaptabilidad que en la previsibilidad. Los defensores de la XP consideran que los cambios de requisitos sobre la marcha son un aspecto natural, inevitable e incluso deseable del desarrollo de proyectos. Creen que ser capaz de adaptarse a los cambios de requisitos en cualquier punto de la vida del proyecto es una aproximación mejor y más realista que intentar definir todos los requisitos al comienzo del proyecto e invertir esfuerzos después en controlar los cambios en los requisitos, en la programación extrema todos los requerimientos se expresan como escenarios llamados historias de usuario los cuales se implementan directamente como una serie de tareas. Los programadores trabajan en parejas



y desarrollan pruebas para cada tarea antes de escribir el código. Todas las pruebas se deben ejecutar satisfactoriamente cuando el código nuevo se integre al sistema. Existe un pequeño espacio de tiempo entre las entregas del sistema. La programación extrema implica varias prácticas que se ajustan a los principios de los métodos ágiles, el desarrollo incremental se lleva a cabo a través de entregas del sistema pequeñas y frecuentes y por medio de un enfoque para la descripción de requerimientos basados en las historias de cliente o escenarios que pueden ser la base para el proceso de planificación, la participación del cliente se lleva a cabo a través del compromiso a tiempo completo del cliente en el equipo de desarrollo.(Flores, 2016)

El interés en las personas, en vez de en los procesos, se lleva a cabo a través de la programación en parejas, la propiedad colectiva del código del sistema, y un proceso de desarrollo sostenible que no implique excesivas jornadas de

- **Ventajas:**

Programación organizada.

Menor tasa de errores.

Satisfacción del programador.

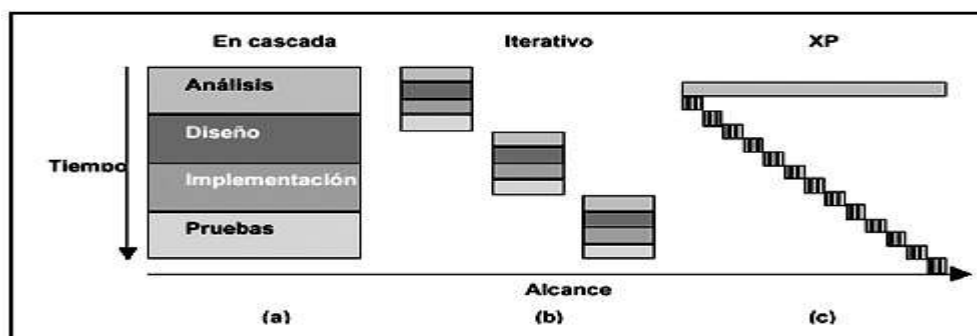
- **Desventajas:**

Es recomendable emplearlo solo en proyectos a corto plazo.

Altas comisiones en caso de fallar.

2.3.3.2. Proceso XP

Figura N° 09: XP en las fases de desarrollo de software



Fuente: Extrem Programming. (2010). Recuperado de Extrem

b- SCRUM

SCRUM es una metodología de desarrollo ágil; tiene como base la creación de ciclos breves para el desarrollo, que se conocen como iteraciones, pero en SCRUM se denominan “SPRINTS”. No se basa en el seguimiento de un plan, sino en la adaptación continua a las circunstancias de la evolución del proyecto. Como método ágil:

Es un modo de desarrollo adaptable, antes que predictivo

Orientado a las personas, más que a los procesos.

Emplea el modelo de construcción incremental basado en iteraciones y revisiones.

Scrum es un proceso en el que se aplican de manera regular un conjunto de mejores prácticas para trabajar en equipo y obtener el mejor resultado posible de un proyecto. Estas prácticas se apoyan unas a otras y su selección tiene origen en un estudio de la manera de trabajar de equipos altamente productivos.

En Scrum se realizan entregas parciales y regulares del resultado final del proyecto, priorizadas por el beneficio que aportan al receptor del proyecto. Por ello, Scrum está especialmente indicado para proyectos en entornos complejos, donde se necesita obtener resultados pronto, donde los requisitos son cambiantes o poco definidos, donde la innovación, la competitividad y la productividad son fundamentales.

- **Ventajas**

Programación organizada.

Menor tasa de errores.

Satisfacción del programador.

- **Desventajas:**

Es recomendable emplearlo solo en proyectos a corto plazo.

Altas comisiones en caso de fallar.

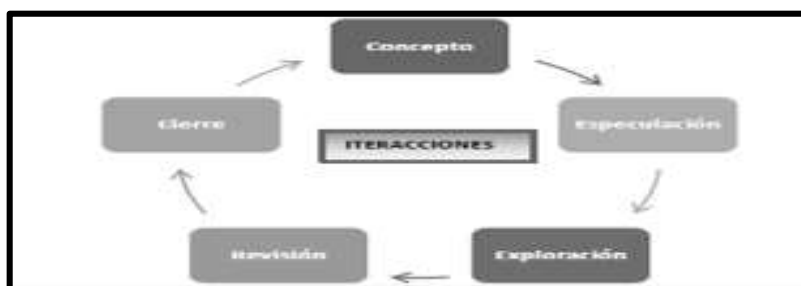
- **Fases de la metodología:**

- **Concepto:** definición de las características del producto.



- **Especulación:** se establecen límites para el desarrollo del producto, como costes y agendas.
- **Exploración:** se añaden funcionalidades definidas en la fase de especulación.
- **Revisión:** Se revisa todo lo construido y se contrasta con el objetivo definido.
- **Cierre:** se entrega en la fecha planificada una versión del producto deseado. De acuerdo a esto se realizan cambios, mantenimiento, hasta que se acerque a la versión final del producto.

Figura N° 10: Fases de la Metodología SCRUM



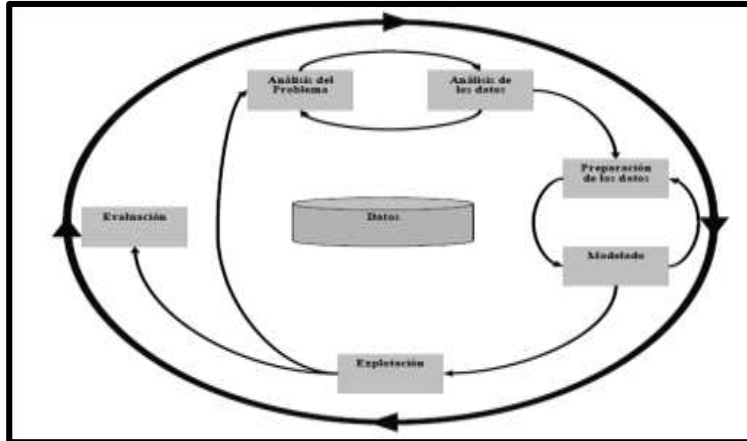
Fuente: “Gestión de Proyectos Informáticos: Metodología SCRUM”
(Manuel Trigos Gallego, p. 34).

2.3.3.3. Metodologías de Desarrollo de Modelos de Minería de Datos

A. CRISP – DM

La metodología CRISP- DM consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. El proceso está organizado en seis fases, que recorren toda la vida del proyecto de datamining, desde la definición de los objetivos del negocio, hasta la vigilancia y mantenimiento del modelo que se propone. Cada fase está estructurada en tareas generales, que se proyectan a tareas más específicas, con resultados concretos.

Figura N° 11: Fases del proceso de modelado metodología CRISP-DM

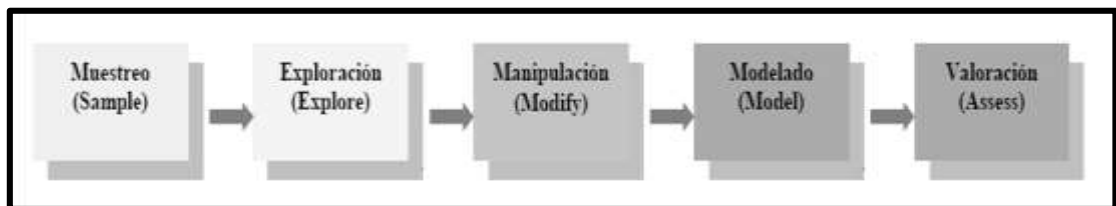


Fuente: “Detección de Patrones de Daños y Averías en la Industria Automotriz” (Ing. Hugo Daniel Flores, 2009, p. 17.)

B. SEMMA

SAS Institute es el desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.

Figura N° 12: Fases de la Metodología SEMMA



Fuente: “Detección de Patrones de Daños y Averías en la Industria Automotriz”. (Ing. Hugo Daniel Flores, 2009, p. 15).

2.3.3.4. Herramientas de Desarrollo de Minería de Datos

a. R-Project



Es un entorno de trabajo basado en los entornos de programación S y S-PLUS desarrollados a principios de los años noventa del pasado siglo por Bill Venables y David M. Como señalan Venables et al. (2011), es un entorno integrado de facilidades informáticas para la manipulación de datos, el cálculo y la generación de gráficos. R-Project pretende convertirse en un sistema internamente coherente que se caracterizaría por un desarrollo basado en la contribución relativamente altruista de la comunidad científica. (López Puga, 2010)

b. Rapidminer

RapidMiner (anteriormente, YALE, Yet Another Learning Environment) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación y en aplicaciones empresariales. (Beltran & Poveda, 2010)

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, reprocesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en Weka. (Beltran & Poveda, 2010)

c. WEKA

Hoy en día existen muchas herramientas de Minería de Datos que ayudan a las empresas a extraer, resumir, mejorar y analizar sus datos almacenados, con el fin de saber que es lo que venden, lo que funciona y lo que no, quién está comprando y quién no. Entre estas herramientas encontramos a Weka. Esta es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. Weka está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla. Este programa es de libre distribución y difusión. (García Molina, 2006)



d. SPSS CLEMENTINE

SPSS Clementine es una herramienta integrada de minería de datos que incluye diversas fuentes de datos (ASCII, XLS, ODBC, etc.), un interfaz visual basado en procesos/flujos de datos (streams), distintas herramientas de minería de datos (correlación, reglas de asociación, regresión, segmentación, clasificación, redes neuronales, reglas y árboles de decisiones, etc.), manipulación de datos (pick & mix, muestreo, combinación y separación, etc.), combinación de modelos, visualización de datos, exportación de modelos a distintos lenguajes (C, SPSS, SAS, etc.), exportación de datos integrada a otros programas (XLS) y generación de informes.

El entorno del Clementine está basado en nodos que se van disponiendo y conectando para formar un flujo, o stream, traducido por Clementine también como “ruta”. Los streams pueden alojarse en ficheros separados (.sir) o se pueden organizar en proyectos (.cpj). De hecho, tanto los streams como los proyectos de minería de datos se almacenan en ficheros separados que se puede cargar, guardar, modificar, reejecutar o reorganizar y que son independientes de las fuentes de datos.

2.3.3.5. Definición de términos básicos

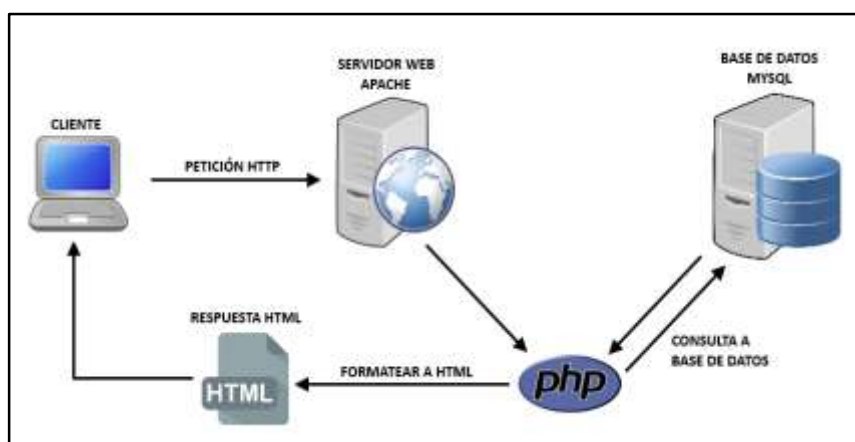
2.3.3.5.1. Aplicación Web

Una aplicación web es un tipo especial de aplicación cliente/servidor, donde el cliente (navegador, explorador o visualizador), el servidor (el servidor web) y el protocolo mediante el que se comunican, están estandarizados y no han de ser creados por el programador de aplicaciones.

Suelen distinguirse en tres niveles (como en las arquitecturas cliente/servidor de tres niveles): el nivel superior que interacciona con el usuario (el cliente web, normalmente un navegador), el nivel inferior que proporciona los datos (la base de datos) y el nivel intermedio que procesa los datos (el servidor web).



Figura N° 13: Cliente- Servidor



Fuente: "Procedimientos de cliente servidor."(MAKERS, 2014)

- **CRISP – DM**

Cross Industry Standard Process for Data Mining. Se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos.

- **Semma:**

Muestra, Explorar, Modificar, Modelar y Evaluar. Es una lista de pasos secuenciales desarrollados por SAS Institute Inc., uno de los mayores productores de estadísticas y de inteligencia de negocios de software. Orienta la aplicación de minería de datos de aplicaciones. A pesar de que SEMMA se considera a menudo una metodología general de minería de datos, SAS afirma que es una organización lógica del sistema de herramienta funcional de uno de sus productos, SAS Enterprise Miner, para la realización de las tareas principales de la minería de datos.

a. Definición de la terminología

Son hechos, medidas u observaciones, que pueden presentarse (o no) en un contexto dado. Datos sin contexto.

Un conjunto de soluciones y servicios que permiten crear u consolidar proyectos de manera inteligente destacando elaboración de portales web.



CAPÍTULO III: MARCOMETODOLÓGICO

3. Marco metodológico

3.1. Tipo y diseño de la investigación

La presente investigación es:

De tipo Tecnológica – Propositiva. La elaboración de la investigación es

Tecnológica, porque tiene como objetivo la implementación de una solución basada en Minería de Datos.

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

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

Aplicada: Porque aplica teorías especializadas con el tema de investigación

Explicativa: Porque busca explicar la forma en que la variable independiente influye en la dependiente

Pre Experimental – Propositiva

| | | | |
|-----|----|---|----|
| | T1 | | T2 |
| M → | O1 | X | O2 |

Dónde:

M: Proceso del pronóstico de la producción de espárragos en la Región.

O1: Es la observación a desarrollar en la muestra – PRE TEST: análisis documentario.

X: Sistema web para la aplicación de técnicas de minería de datos basada en soluciones OLAP de Base de datos multidimensionales

T1: Es el tiempo de medición inicial con información actual.

T2: Es el tiempo de medición posterior a la simulación de la propuesta de solución.

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

M: (T1, T2, O1, O2) M

3.2. Población y muestra

3.2.1. Población

La población está compuesta por los 22000 registros de producción de espárragos en la localidad de Jayanca Lambayeque de los años 2012 hasta el 2015

Producción de espárragos que genera la empresa. Representa los registros que se origina por la producción de espárragos, que generan almacenamientos de grandes cantidades que representa producción, como indicador de avance por periodo.

Tabla N° 3: Producción de espárragos Periodo - Año 2012

| CdigoCliente | Nombres y Apellidos | Localidad | Cant Cajas | Kg Por Caja | Total Kg por caja |
|-----------------|---------------------------------------|-----------|-------------|-------------|-------------------|
| 190-11-00035740 | OLIVOS SANTA CRUZ JULIO | JAYANCA | 220.00 | 40.00 | 8800.00 |
| 190-11-00038921 | ORTEGA CAMPOS KATHIA DEL PILAR | JAYANCA | 320.00 | 40.00 | 12800.00 |
| 190-11-00194820 | MACO EFUS BERTILA | JAYANCA | 320.00 | 40.00 | 12800.00 |
| 190-11-00460179 | GAONA GOMES LUZ NELITA | JAYANCA | 620.00 | 40.00 | 24800.00 |
| 190-11-00757132 | VARGAS CHUQUIMANGO LUZ LIDIA | JAYANCA | 420.00 | 40.00 | 16800.00 |
| 190-11-00763576 | REQUELME CAMPOS TEODOLINDA | JAYANCA | 520.00 | 40.00 | 20800.00 |
| 190-11-00956234 | ALCAS GOMEZ MARILU PATRICIA | JAYANCA | 420.00 | 40.00 | 16800.00 |
| 190-11-01041270 | GONZALES ROMERO YANINA DEL ROSARIO | JAYANCA | 520.00 | 40.00 | 20800.00 |
| 190-11-01041300 | FLORES FARFAN HERMINIA MADALEYNE | JAYANCA | 620.00 | 40.00 | 24800.00 |
| 190-11-01041474 | CORONEL LLATAS JUAN CARLOS | JAYANCA | 620.00 | 40.00 | 24800.00 |
| 190-11-01041476 | TORRES CASTRO VICTOR | JAYANCA | 520.00 | 40.00 | 20800.00 |
| 190-11-01041487 | RIMARACHIN MEDINA ANGELICA | JAYANCA | 520.00 | 40.00 | 20800.00 |
| 190-11-01041525 | MORALES GONZALES VIRGINIA DEL PILAR | JAYANCA | 220.00 | 40.00 | 8800.00 |
| 190-11-01041662 | MOGOLLON GUARDERAS BEATRIZ VICTORIA | JAYANCA | 220.00 | 40.00 | 8800.00 |
| 190-11-01041674 | QUISPE LOZANO ENMA DEL ROSARIO | JAYANCA | 424.00 | 40.00 | 16960.00 |
| 190-11-01041757 | ZAPATA DE LA CRUZ JESSICA DEL ROSARIO | JAYANCA | 324.00 | 40.00 | 12960.00 |
| 190-11-01041759 | REQUELME BAUTISTA MARIA DELICIA | JAYANCA | 220.00 | 40.00 | 8800.00 |
| 190-11-01041848 | FERNANDEZ ROJAS LILIANA | JAYANCA | 220.00 | 40.00 | 8800.00 |
| 190-11-01041954 | MORANTE PURIZACA PAULA VIRGINIA | JAYANCA | 420.00 | 40.00 | 16800.00 |
| 190-11-01041979 | VILLALOBOS VASQUEZ MARIA ISABEL | JAYANCA | 520.00 | 40.00 | 20800.00 |
| 190-11-01041988 | GASTELO GAMARRA MARIA AURORA | JAYANCA | 520.00 | 40.00 | 20800.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Total | | | 1809 | | |

Fuente: Base de Datos de la Empresa, Enero 2012 - Diciembre 2012



Tabla N° 4: Producción de espárragos Periodo - Año 2013

| Codigo de Cliente | Apellidos y Nombres | Localidad | Cant Cajas | Kg Por Caja | Total Kg por caja |
|-------------------|-------------------------------------|-----------|-------------|-------------|-------------------|
| 190-11-00195274 | BECERRA GUEVARA HILDA | JAYANCA | 156.00 | 40.00 | 6240.00 |
| 190-11-00269692 | VELASQUEZ GUTIERREZ ROSA ELIZABETH | JAYANCA | 118.00 | 40.00 | 4720.00 |
| 190-11-00955126 | PAICO CHANAME JORGE LUIS | JAYANCA | 156.00 | 40.00 | 6240.00 |
| 190-11-00955256 | GALVEZ CHUQUIMANGO GREYS ESTEFANIA | JAYANCA | 158.00 | 40.00 | 6320.00 |
| 190-11-00955676 | NIÑO BRAVO MARIA ANGELICA | JAYANCA | 218.00 | 40.00 | 8720.00 |
| 190-11-01041449 | PULIDO MONTENEGRO ZANDRA ROXANA | JAYANCA | 356.00 | 40.00 | 14240.00 |
| 190-11-01041526 | CASTILLO BARRAGAN ACARINI ESMERALDA | JAYANCA | 424.00 | 40.00 | 16960.00 |
| 190-11-01041818 | YAMUNAQUE SILVA ANA YSABEL | JAYANCA | 324.00 | 40.00 | 12960.00 |
| 190-11-01041964 | SUYON DIAZ OLENKA VICTORIA | JAYANCA | 418.00 | 40.00 | 16720.00 |
| 190-11-01041983 | ORTIZ PARICAHUA LUCIA LUZMILA | JAYANCA | 318.00 | 40.00 | 12720.00 |
| 190-11-01042064 | DIAZ SAAVEDRA MELISSA ESTHER | JAYANCA | 654.00 | 40.00 | 26160.00 |
| 190-11-01042181 | TARRILLO BULNES ALEXANDRA MARILU | JAYANCA | 618.00 | 40.00 | 24720.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Total | | | 7048 | | |

Fuente: Base de Datos de la Empresa, Enero 2013-
Diciembre 2013

Tabla N°5: Producción de espárragos Periodo - Año 2014

| CdigoCliente | Nombres y Apellidos | Localidad | Cant Cajas | Kg Por Caja | Total Kg por caja |
|-----------------|---|-----------|-------------|-------------|-------------------|
| 190-11-01236937 | VELASQUEZ BUSTAMANTE GIAN MARCO | JAYANCA | 756.00 | 40.00 | 30240.00 |
| 190-13-00051065 | BUSTAMANTE CIEZA VERONICA | JAYANCA | 856.00 | 40.00 | 34240.00 |
| 190-13-00051224 | VIDARTE CASTRO MARLENY VIDALINA DEL PILAR | JAYANCA | 724.00 | 40.00 | 28960.00 |
| 190-13-00051230 | RAMIREZ BACILIO BRIGIDA TERESA | JAYANCA | 318.00 | 40.00 | 12720.00 |
| 190-13-00269345 | CHUPILLON YOVERA YESENIA KATHERINE | JAYANCA | 518.00 | 40.00 | 20720.00 |
| 190-13-00486843 | MACALOPU VIDARTE JENIFER NATALY | JAYANCA | 624.00 | 40.00 | 24960.00 |
| 190-13-00698009 | VALERA VEGA ITAMAR | JAYANCA | 424.00 | 40.00 | 16960.00 |
| 190-13-00698012 | RAMOS GARCIA ANALLY KELLY | JAYANCA | 424.00 | 40.00 | 16960.00 |
| 190-13-00698017 | URIARTE NUÑEZ ZOILA ROSA | JAYANCA | 409.00 | 40.00 | 16360.00 |
| 190-13-00698023 | FLORES CHIROQUE HECTOR | JAYANCA | 724.00 | 40.00 | 28960.00 |
| 190-13-00698093 | FLORES SOBRINO CARMEN LASTENIA | JAYANCA | 549.00 | 40.00 | 21960.00 |
| 190-13-00698098 | FARFAN ALEJANDRIA ESTHERFILIA | JAYANCA | 518.00 | 40.00 | 20720.00 |
| 190-13-00698100 | SANCHEZ BAUTISTA BERBELINDA | JAYANCA | 424.00 | 40.00 | 16960.00 |
| 190-13-00698123 | DUCEP SALDAÑA DORILA | JAYANCA | 456.00 | 40.00 | 18240.00 |
| 190-13-00698125 | SANCHEZ ACOSTA MARIA YOLANDA | JAYANCA | 324.00 | 40.00 | 12960.00 |
| 190-13-00698126 | BRAVO VALLEJOS ZENAIDA | JAYANCA | 624.00 | 40.00 | 24960.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Total | | | 8603 | | |

Fuente: Base de Datos de la Empresa, Enero 2014 - Diciembre 2014



Tabla N°6: Producción de espárragos Periodo - Año 2015

| CodigoCliente | Nombres y Apellidos | Localidad | Cant Cajas | Kg Por Caja | Total Kg por caja |
|----------------------|-------------------------------------|------------------|-------------------|--------------------|--------------------------|
| 190-13-00051060 | ROMERO DIAZ ADELA DEL MILAGRO | JAYANCA | 956.00 | 40.00 | 38240.00 |
| 190-13-00486148 | CORREA MIO GLORIA DELMIRA | JAYANCA | 456.00 | 40.00 | 38240.00 |
| 190-14-00013028 | CARHUAJULCA ANGELES ROSA EVANGELINA | JAYANCA | 856.00 | 40.00 | 34240.00 |
| 190-14-00013375 | FACHO ARNAO ORLANDO BALTAZAR | JAYANCA | 856.00 | 40.00 | 34240.00 |
| 190-14-00218500 | GARCIA MENDOZA CINTHIA MARGARITA | JAYANCA | 656.00 | 40.00 | 26240.00 |
| 190-14-00225873 | ANGELES MORI ANALI DEL MILAGRO | JAYANCA | 556.00 | 40.00 | 22240.00 |
| 190-14-00573190 | EFFIO CAMPOS LORENA ANAMARIA | JAYANCA | 656.00 | 40.00 | 26240.00 |
| 190-14-00861018 | ESPINOZA DE ESQUEN MARIA CRUZ | JAYANCA | 656.00 | 40.00 | 26240.00 |
| 190-14-01357350 | SANCHEZ GONZALES ANA CECILIA | JAYANCA | 956.00 | 40.00 | 38240.00 |
| 190-14-02108301 | RUIZ HUAMAN ROSA LUZ | JAYANCA | 556.00 | 40.00 | 22240.00 |
| 190-14-02109217 | ELIAS HORNA MARIA MAGDALENA | JAYANCA | 456.00 | 40.00 | 18240.00 |
| 190-14-02109388 | CHUPILLON YOVERA YESENIA KATHERINE | JAYANCA | 556.00 | 40.00 | 22240.00 |
| 190-14-02109746 | FARRO ORDOÑEZ MARIA DEL MILAGRO | JAYANCA | 656.00 | 40.00 | 26240.00 |
| 190-14-02110449 | GALVEZ POZO CONSUELO DEL PILAR | JAYANCA | 956.00 | 40.00 | 38240.00 |
| 190-14-02110787 | TANTARICO MEJIA LIZET YOJANA | JAYANCA | 756.00 | 40.00 | 30240.00 |
| 190-14-02110899 | ACUÑA MONTENEGRO MARGORIE JUANITA | JAYANCA | 856.00 | 40.00 | 34240.00 |
| 190-14-02110949 | BECERRA RUESTAS MARIA ROXANA | JAYANCA | 956.00 | 40.00 | 38240.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Total | | | 4725 | | |

Fuente: Base de Datos de la Empresa, Enero 2015 - Junio 2015

Comprendiendo un total 22000 registros de producción de espárragos en la empresa - localidad de Jayanca Lambayeque de los años 2012 hasta el 2015

3.2.2. Muestra

La muestra es poblacional con un total de 22000 registros.

3.2.3. Hipótesis

Mediante la comparación de técnicas de minería de datos se podrá elegir la técnica para la predicción de la producción de espárragos



3.3. Variables

3.3.1. Variable independiente:

Técnicas de minería de datos

3.3.2. Variable dependiente:

Pronóstico de producción de espárragos.

3.4. Operacionalización

En la propuesta de investigación, utilizamos la variable independiente y la dependiente, las cuales se indican en la siguiente tabla. Se muestran las dimensiones, indicadores y técnicas de recolección de datos.

Tabla N° 07: Determinación de variables

| Variable independiente | Dimensiones | Indicadores | Fórmula |
|------------------------------|-------------|------------------------------------|--|
| Técnicas de Minería de Datos | Tiempo | Tiempo de procesamiento del modelo | $\sum \frac{TP}{N}$ <p>TP: Tiempo procesamiento N: Número de meses</p> |
| | Datos | Uso de CPU por procesamiento | $\sum \frac{CPU}{N}$ <p>CPU: Uso de CPU N: Número de meses</p> |



| Variable Dependiente | Dimensiones | Indicadores | Formula |
|---|------------------------|--|---|
| Pronósticos de Producción de espárragos | Grado de confiabilidad | Confiabilidad de los pronósticos generados (Mide la confiabilidad del modelo con respecto a las predicciones realizadas) | $PCPV = 100 - \left(\frac{\sum \frac{MR - MP}{MR}}{N} \right) * 100$ <p>PCPV: Porcentaje de confiabilidad de predicción de espárragos. MP: Monto pronosticado MR: Monto real N: Número de meses</p> |

Fuente: Elaboración propia

3.5. Métodos, técnicas e instrumentos de recolección de datos

3.5.1. Método de investigación:

El método empleado es el deductivo, reflejado desde el análisis general de la problemática de estudio, pasando por las bases teóricas hasta desagregar de forma particular los indicios o indicadores de las variables de estudio, más precisamente en términos de los indicadores de la variable dependiente.

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

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



3.5.2. Técnicas

Las técnicas empleadas para la recopilación de la información son:

Tabla N° 08: Métodos y Técnicas de investigación

| Método/Técnica | Descripción |
|----------------|--|
| Observación | La información histórica se encuentra en un repositorio de datos la cual será Materia de estudio para la presente investigación de tesis. |

Fuente: Elaboración Propia

3.5.3. Procedimientos para la recolección de datos

- a. **Recolección de datos:** Observación Directa
- b. **Análisis de Resultados:** Excel 2010

Para la recolección de datos se hará una revisión de las variantes de consumos que permitan saber qué suministros pueden estar aptos para aplicar la técnica de minería de datos en sus consumos.

Para ello es importante realizar una limpieza de los datos, para reducir los ruidos y valores nulos, así como seleccionar características eliminando atributos irrelevantes o duplicados.

3.5.4. Análisis estadístico e interpretación de los datos

El Análisis estadísticos de los datos se basa en el:

- a. Uso de tablas, para evaluar resultados de las técnicas de minería de datos
- b. Uso de gráficos estadísticos, para evaluar resultados de las técnicas de minería de datos.
- c. Uso de técnicas estadísticas como la media y desviación estándar.



El Análisis estadísticos de los datos se basa en el uso de las siguientes herramientas:

Gestor de Base de datos MS SQL SERVER 2012 R2.

MS Integration Service

MS Analysis Services

R – PROJECT

En este proceso se hará una clasificación de los datos materia de estudio, para lo cual se creará un DATAMART con las dimensiones necesarias para gestionar la información.

3.6. Principios éticos

La presente investigación se realiza siguiendo los principios éticos que debe tener todo investigador que se tomarán en cuenta.

Tabla N° 09: Criterios de para los principios éticos

Fuente: Elaboración Propia

| Criterios | Características éticas del criterio |
|---------------------------|---|
| Confidencialidad | Asegura la protección de la identidad de la institución y las personas que participan como informantes de la investigación. |
| Objetividad | El análisis de la situación encontrada se basa en criterios técnicos e imparciales. |
| Originalidad | Se citan las fuentes bibliográficas de la información. |
| Veracidad | La información mostrada es verdadera, cuidando la confidencialidad de ésta. |
| Derechos laborales | La propuesta de solución propicia el respeto a los derechos laborales en la entidad de estudio. |



La presente investigación se realiza siguiendo los principios éticos que debe tener todo investigador que se tomarán en cuenta

3.7. Criterios de rigor científico

La presente investigación se realiza siguiendo los juicios científicos establecidos, estos permiten garantizar la calidad de la propuesta de investigación.

Tabla N° 10: Criterios de rigor científico

| Criterios | Características científicas del criterio |
|---------------|--|
| Confiabilidad | Se realizan cálculos estadísticos para la determinación del nivel de consistencia interna de los instrumentos de recolección de datos. |
| Validación | Se validan los instrumentos de recolección de datos y la propuesta de solución a través de Juicio de Expertos. |

Fuente: Elaboración Propia

Así, seguimos la coherencia metodológica durante el desarrollo de la propuesta de la investigación, realización apropiada del muestreo de datos, los cuales son al azar para ser totalmente imparcial en el recojo de datos.

3.8. Evaluación económica del software

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

ANÁLISIS PRELIMINAR

DEFINICIÓN DE REQUERIMIENTOS:

Donde:

RS = Responsabilidades del Sistema

Se considera la siguiente lista, siendo seis:



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

$$RS = 6$$

F = Funciones de Sistema:

$$F = 280 * RS$$

$$F = 1680$$

MF = Miles de Funciones

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

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

$$MF = 1.68$$

ESF = Esfuerzo.

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

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

$$ESF = 4.13795714$$

TDES = Tiempo de Desarrollo

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

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

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



CH = Cantidad de Hombres por MES

$$CH = ESF/TDES$$

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

$$CH = 0.9645$$

CH = 1 personas por mes

CHM = Costo Hombre por Mes

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

$$CHM = 1 * 2400$$

$$\mathbf{CHM = 2400}$$

CD = Costo de Desarrollo

$$CD = ESF * CHM$$

$$CD = 4.138 * 2400$$

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

Por las características del proyecto, los siguientes indicadores son:

Tabla N° 11: Indicadores /Factores por Medida de Proyecto

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

Fuente: Elaboración propia



Tabla N° 12 Distribución de esfuerzo y tiempo de desarrollo por etapas

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

Fuente: Elaboración propia



CAPÍTULO IV:
ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

4. Análisis e interpretación de los resultados.

4.1. Resultados en tablas y gráficos

Siendo el objetivo general de este estudio:

“Realizar un análisis comparativo de las técnicas de minería de datos para la predicción de producción de espárragos en una empresa de la localidad de Jayanca - Lambayeque

Este análisis se realizará midiendo tres indicadores

A. Confiabilidad de la Predicción generados por el modelo

En el siguiente indicador se medirá la confiabilidad del modelo a partir de los resultados obtenidos por los pronósticos generados por cada algoritmo de entrenamiento, contrastando este monto con meses reales para la precisión y error del algoritmo.

- **Pronóstico Holtwinters**

PCPV: Porcentaje de confiabilidad de predicción de ventas realizadas

MP: Monto Pronosticado

MR: Monto Real

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

N: Numero de meses Probados

La precisión se encuentra aplicando la formula el MP/MRy en el margen de error se aplica la fórmula para calcular PCV es igual a 100 menos la sumatoria de MR-MP Dividido con MR Dividido con N multiplicado por 100

Tabla N° 13: Tabla de Validación Holtwinters

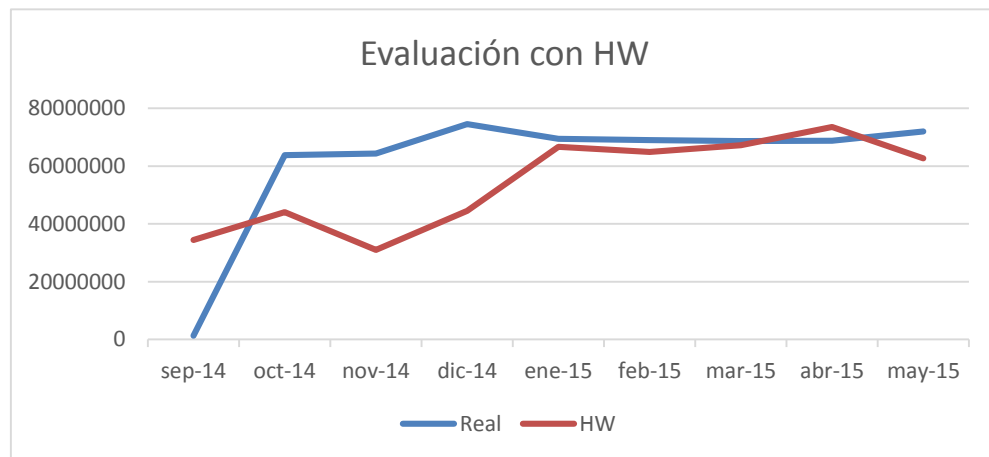
| | Periodo | Real | HW | Precisión | Error |
|---|---------|----------|----------|-----------|-------|
| 1 | Set-14 | 1319220 | 34382116 | 26.0625 | 26.06 |
| 2 | Oct-14 | 63779240 | 44038049 | 0.6905 | 30.95 |
| 3 | Nov-14 | 64395520 | 30997970 | 0.4814 | 51.86 |
| 4 | Dic-14 | 74496100 | 44479126 | 0.5971 | 40.29 |
| 5 | Ene-15 | 69463660 | 66668569 | 0.9598 | 4.02 |
| 6 | Feb-15 | 68978000 | 64927521 | 0.9413 | 5.87 |
| 7 | Mar-15 | 68726560 | 67257766 | 0.9786 | 2.14 |
| 8 | Abr-15 | 68789280 | 73500212 | 1.0685 | 1.07 |
| 9 | May-15 | 72008860 | 62672573 | 0.8703 | 12.97 |
| | | | | | 19.47 |
| | | | | Confianza | 80.53 |

Fuente: Elaboración Propia

Tabla de evaluación de la técnica de woltwinter cuenta con periodos producción de los meses de Setiembre 2014 a Mayo 2015 en la columna periodo describe los meces y al año que corresponde, en la columna Real muestra la cantidad de producción, en la columna HW, muestra la cantidad de producción pronostico en evaluación con la técnica HoltWinter, en la columna Precisión muestra el porcentaje de la evaluación realizada por la técnica HW, en la columna Error muestra el margen de error de los resultados evaluados por cada periodo por producción, al final mostrando un 80.50 de confianza.



Gráfico N° 01: Evaluación con HW



Fuente: elaboración propia

En este grafico el pronóstico de woltwinternos muestra que de setiembre a octubre 2014 aumenta la producción y en noviembre baja la producción y en diciembre hay un aumento de producción, luego en enero abril 2015 la producción se mantiene, en donde tambien nos muestra que en mayo 2015 existe una baja de producción.

- **Pronóstico Red Neuronal Autorregresiva**

PCPV: Porcentaje de confiabilidad de predicción de ventas realizadas

MP: Monto Pronosticado

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

MR: Monto Real

N: Numero de meses Probados

La precisión se encuentra aplicando la formula el MP/MR y en el margen de error se aplica la fórmula para calcular PCV es igual a 100 menos la sumatoria de MR-MP Dividido con MR Dividido con N multiplicado por 100



Tabla N° 14: Tabla de Validación ARNA

| | Periodo | Real | ARNA | % | Error |
|---|---------|----------|----------|------------------|--------------|
| 1 | Set-14 | 1319220 | 46868733 | 35.5276 | 35.53 |
| 2 | Oct-14 | 63779240 | 40132884 | 0.6292 | 37.08 |
| 3 | Nov-14 | 64395520 | 29972821 | 0.4654 | 53.46 |
| 4 | Dic-14 | 74496100 | 77931397 | 1.0461 | 1.05 |
| 5 | Ene-15 | 69463660 | 68423582 | 0.9850 | 1.50 |
| 6 | Feb-15 | 68978000 | 68988864 | 1.0002 | 1.00 |
| 7 | Mar-15 | 68726560 | 68241792 | 0.9929 | 0.71 |
| 8 | Abr-15 | 68789280 | 70095415 | 1.0190 | 1.02 |
| 9 | May-15 | 72008860 | 72369333 | 1.0050 | 1.01 |
| | | | | | 14.70 |
| | | | | Confianza | 85.30 |

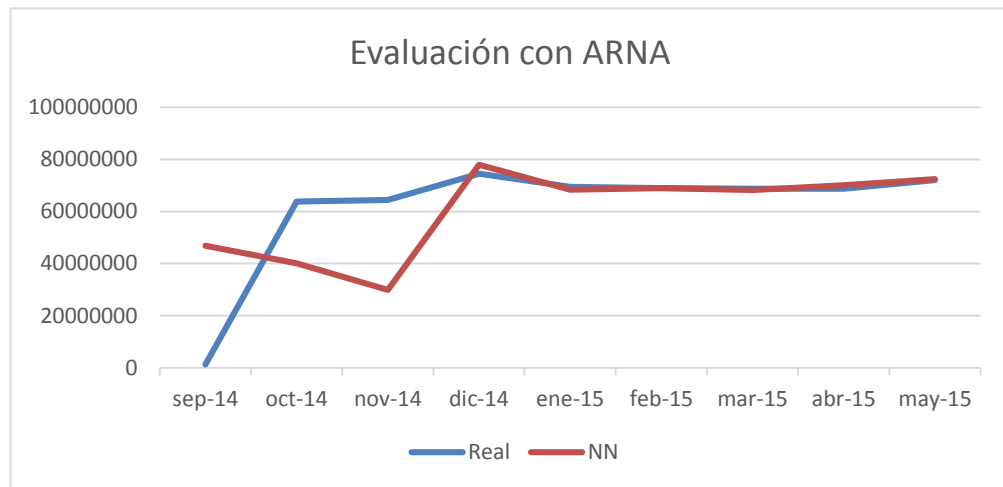
Fuente: Elaboración Propia

En el caso de la red neuronal se puede apreciar mejor aproximación con respecto a los resultados obtenidos por HW.

Tabla de evaluación de la técnica de Red Neuronal ARNA cuenta con periodos producción de los meses de Setiembre 2014 a Mayo 2015 en la columna periodo describe los meses y al año que corresponde, en la columna Real muestra la cantidad de producción, en la columna ARNA, muestra la cantidad de producción pronostico en evaluación con la técnica Red Neuronal Auto regresiva , en la columna Precisión muestra el porcentaje de la evaluación realizada por la técnica ARNA, en la columna Error muestra el margen de error de los resultados evaluados por cada periodo por producción, al final mostrando un 85.30 de confianza.



Gráfico N°02: Evaluación con ARNA



Fuente: Elaboración Propia

En este grafico el pronóstico de Red Neuronal Auto regresivos muestra que de setiembre a octubre 2014 aumenta la producción y en noviembre baja la producción y en diciembre hay un aumento de producción, luego en enero a mayo 2015 la producción se mantiene.

- **Pronóstico Arima**

PCPV: Porcentaje de confiabilidad de predicción de ventas realizadas

MP: Monto Pronosticado

MR: Monto Real

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

N: Numero de meses Probados

La precisión se encuentra aplicando la formula el MP/MR y en el margen de error se aplica la fórmula para calcular PCV es igual a 100 menos la sumatoria de MR-MP Dividido con MR Dividido con N multiplicado por 100



Tabla N° 15: Tabla de Validación Arima

| | Periodo | Real | AR | % | Error |
|---|---------|----------|----------|------------------|--------------|
| 1 | Set-14 | 1319220 | 14123796 | 10.7062 | 10.71 |
| 2 | Oct-14 | 63779240 | 32558299 | 0.5105 | 48.95 |
| 3 | Nov-14 | 64395520 | 36136675 | 0.5612 | 43.88 |
| 4 | Dic-14 | 74496100 | 67940768 | 0.9120 | 8.80 |
| 5 | Ene-15 | 69463660 | 70279731 | 1.0117 | 1.01 |
| 6 | Feb-15 | 68978000 | 69567912 | 1.0086 | 1.01 |
| 7 | Mar-15 | 68726560 | 69020027 | 1.0043 | 1.00 |
| 8 | Abr-15 | 68789280 | 68807426 | 1.0003 | 1.00 |
| 9 | May-15 | 72008860 | 70668538 | 0.9814 | 1.86 |
| | | | | | 13.14 |
| | | | | Confianza | 86.86 |

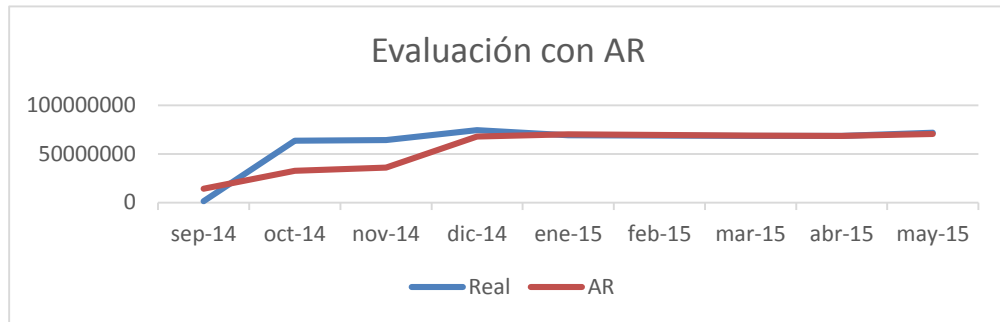
Fuente: Elaboración Propia

Similar al caso con la Red Neuronal, ARIMA también tiene una mejor precisión en los pronósticos obtenidos.

Tabla de evaluación de la técnica de ARIMA cuenta con periodos producción de los meses de Setiembre 2014 a Mayo 2015 en la columna periodo describe los meses y al año que corresponde, en la columna Real muestra la cantidad de producción, en la columna AR(Arima) muestra la cantidad de producción pronostico en evaluación con la técnica ARIMA, en la columna Precisión muestra el porcentaje de la evaluación realizada por la técnica de ARIMA, en la columna Error muestra el margen de error de los resultados evaluados por cada periodo por producción, al final mostrando un 86.86 de confianza



Gráfico N° 03: Evaluación con ARIMA



Fuente: Elaboración propia

En este grafico el pronóstico de ARIMA nos muestra que de setiembre a octubre 2014 aumenta la producción y en noviembre baja la producción y en diciembre hay un aumento de producción, luego en enero abril 2015 la producción se mantiene, en donde también nos muestra que en mayo 2015 existe una baja de producción.

En esta tabla se realiza la comparación de los resultados de cuanto de confianza se aproxima cada técnica

**Tabla N° 16:
Resultados**

| Confianza | Tec | % |
|-----------|------|-------------|
| | HW | 80.52906656 |
| | ARNA | 85.29655977 |
| | AR | 86.86369785 |

Comparación de

Fuente: Elaboración Propia

Descripción. En esta tabla nos muestra los resultados de cuanto de confianza ofrece cada técnica al pronóstico en donde HotWinter nos brinda un 80.5 % de



confianza, Red Neuronal Autor regresiva nos brinda un 85. 3 % de confianza y ARIMA nos muestra un 86. 9 % de confianza

B. Tiempo de procesamiento de modelo para obtener la estima

Como segundo indicador, se medirá el tiempo de procesamiento del modelo en función a cada algoritmo de entrenamiento

$$\sum \frac{TP}{N}$$

TP: Tiempo procesamiento

N: Número de meses

Tabla N° 17: Procesamiento de modelo

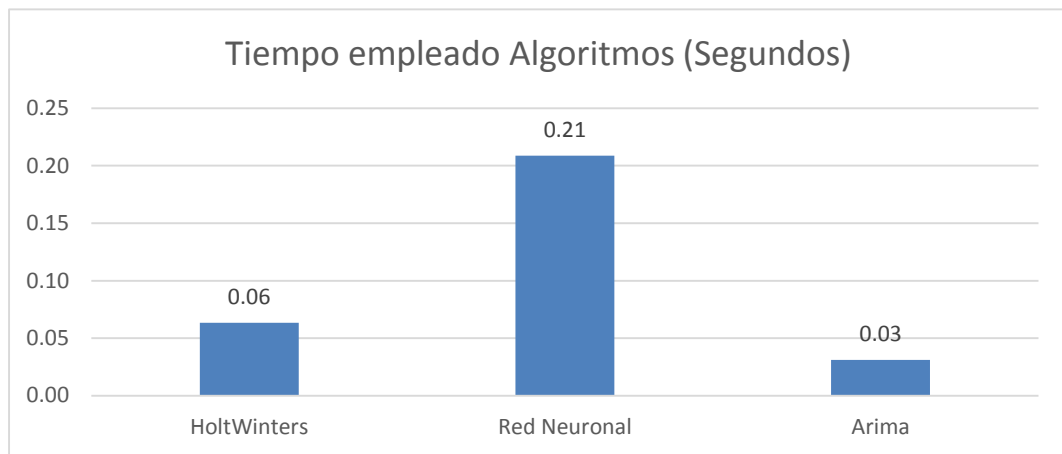
| Ítem | Tipo | Periodo | HoltWinters | Red Neuronal | Arima |
|------|-------------|---------|-------------|--------------|-------|
| 171 | Laboratorio | Set-14 | 0.07 | 0.19 | 0.03 |
| 172 | Laboratorio | Oct-14 | 0.05 | 0.19 | 0.05 |
| 173 | Laboratorio | Nov-14 | 0.06 | 0.2 | 0.03 |
| 174 | Laboratorio | Dic-14 | 0.06 | 0.2 | 0.03 |
| 175 | Laboratorio | Ene-15 | 0.06 | 0.2 | 0.02 |
| 176 | Laboratorio | Feb-15 | 0.09 | 0.21 | 0.03 |
| 177 | Laboratorio | Mar-15 | 0.06 | 0.22 | 0.03 |
| 178 | Laboratorio | Abr-15 | 0.06 | 0.23 | 0.03 |
| 179 | Laboratorio | May-15 | 0.06 | 0.24 | 0.03 |
| | | | 0.06 | 0.21 | 0.03 |

Fuente: Elaboración Propia

Descripción. En esta tabla se muestra el tiempo en segundos a espera al proceso de cada algoritmo para brindar los resultados de pronósticos de producción.



Gráfico N° 04: Tiempo empleado de algoritmos



Fuente: Elaboración propia

Descripción de gráfico. Nos muestra que HoltWinter demora 0.06 segundos en mostrar los resultados. Red neuronal 0.21 segundos y ARIMA 0.03 segundos

C. Uso de CPU, número de puntos mínimos para el vector que procesará el modelo

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

Como tercer indicador, se medirá el uso de cpu del modelo en función a cada algoritmo de entrenamiento

$$\sum \frac{CPU}{N}$$

CPU: Uso de CPU

N: Número de meses



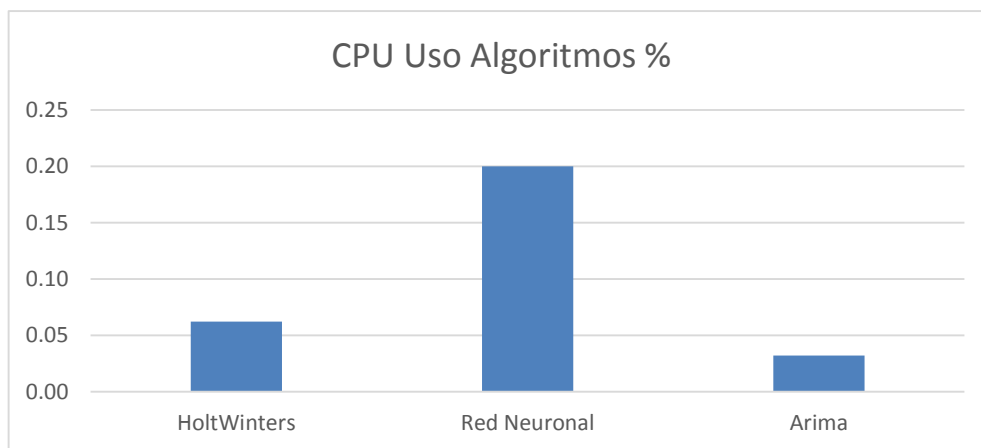
Tabla N° 18: Tiempo de uso de CPU

| Ítem | Tipo | Periodo | HoltWinters | Red | |
|------|-------------|---------|-------------|----------|-------|
| | | | | Neuronal | Arima |
| 171 | Laboratorio | Set-14 | 0.07 | 0.17 | 0.04 |
| 172 | Laboratorio | Oct-14 | 0.05 | 0.19 | 0.05 |
| 173 | Laboratorio | Nov-14 | 0.06 | 0.2 | 0.03 |
| 174 | Laboratorio | Dic-14 | 0.06 | 0.19 | 0.03 |
| 175 | Laboratorio | Ene-15 | 0.06 | 0.18 | 0.02 |
| 176 | Laboratorio | Feb-15 | 0.08 | 0.2 | 0.03 |
| 177 | Laboratorio | Mar-15 | 0.06 | 0.22 | 0.03 |
| 178 | Laboratorio | Abr-15 | 0.06 | 0.21 | 0.03 |
| 179 | Laboratorio | May-15 | 0.06 | 0.24 | 0.03 |
| | | | 0.06 | 0.20 | 0.03 |

Fuente: elaboración propia

Descripción. En esta tabla se muestra el tiempo de uso de CPU en segundos para el proceso de cada algoritmo para brindar los resultados de pronósticos de producción.

Gráfico N° 05: CPU uso de Algoritmo



Fuente: elaboración propia



Descripción de gráfico. Nos muestra que HotWinter ocupo 0.06 segundos en CPU. Red neuronal 0.21 segundos y ARIMA 0.03 segundos

D. Tiempo para generar estimación en el aplicación Web

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

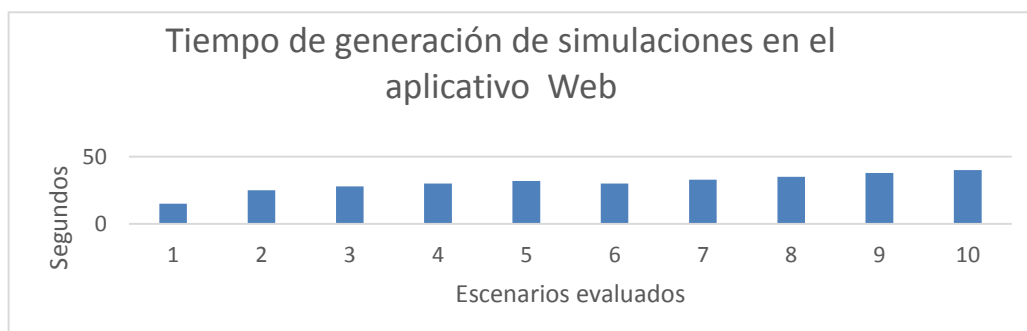
Tabla N° 19: Tiempo de Procesamiento del aplicativo Web

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

Fuente: Elaboración Propia

En la Tabla N° 19 se observa que el tiempo promedio de generación de estimaciones en el sistema web es de 30,60 segundos.

Gráfico N° 06: Tiempo de generación de pronósticos en aplicación web



Fuente: Elaboración Propia

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



4.2. Contrastación de la hipótesis.

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

4.3. Discusión de resultados

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

Con respecto al primer indicador comparando las tres técnicas, es decir Holtwinters, Red Neuronal y ARIMA. Se puede decir que ARIMA obtuvo el nivel de confianza más elevado en comparación a HoltWinters y Red Neuronal, esto se denota en los valores obtenidos al calcular la razón (valor calculado entre el monto real y el monto pronóstico para saber el grado de relación que existe uno con respecto del otro), obteniendo para HoltWinters unos 80.52 % de confiabilidad, contra un Red Neuronal con 85.29 % y un ARIMA con un 86.86 % que lo sitúa como el mejor de los tres algoritmos.

B. Tiempo de procesamiento para obtener la estimación

En el tiempo de procesamiento al evaluar estas técnicas se obtuvo que con el método Arima el tiempo promedio de ejecución de 0.03 segundos, el menor de todos los tiempos comparando contra una Red Neuronal de 0.21 y HW de 0.06 segundos.

C. Uso de CPU

En el uso de CPU al evaluar estas técnicas se obtuvo que con el método Arima el uso de CPU es 0.03 segundos, el menor de todos comparando contra una Red Neuronal de 0.21 y HW de 0.06 %.

D. Tiempo para generar estimación en el sistema

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

**CAPÍTULO V:
DESARROLLO DE LA PROPUESTA**

5. Desarrollo de la propuesta

5.1. Generalidades

Es una solución informática que pretende validar certeramente la estimación de producción de esparrago de la empresa en la localidad de Jayanca Lambayeque, a partir del descubrimiento de patrones de producción, los cuales serán analizados aplicando para ello Minería de Datos.

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

A. Características del producto

Nombre del Producto: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PRONÓSTICO DE PRODUCCIÓN DE ESPÁRRAGOS

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

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

Adaptabilidad: Esta solución es fácilmente adaptable a empresas dedicadas a generar gran cantidad de producción.

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

B. Funciones del Sistema Web

Módulo ingreso – El usuario ingresará a la aplicación mediante un usuario y una contraseña

Módulo de análisis – La aplicación web tendrá la función en realizar el análisis con cada una de las técnicas asignadas

Módulos pronósticos – Este módulo permite dar los resultados de los pronósticos realizados con sus gráficos

Módulo de simulaciones. Se mostrará las simulaciones por periodos de la predicción

C. Usos del producto

Evaluación de usuarios: Brindará la información necesaria para evaluar y confirmar la producción calculadas en base a las simulaciones.

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

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

5.2. Metodología de desarrollo

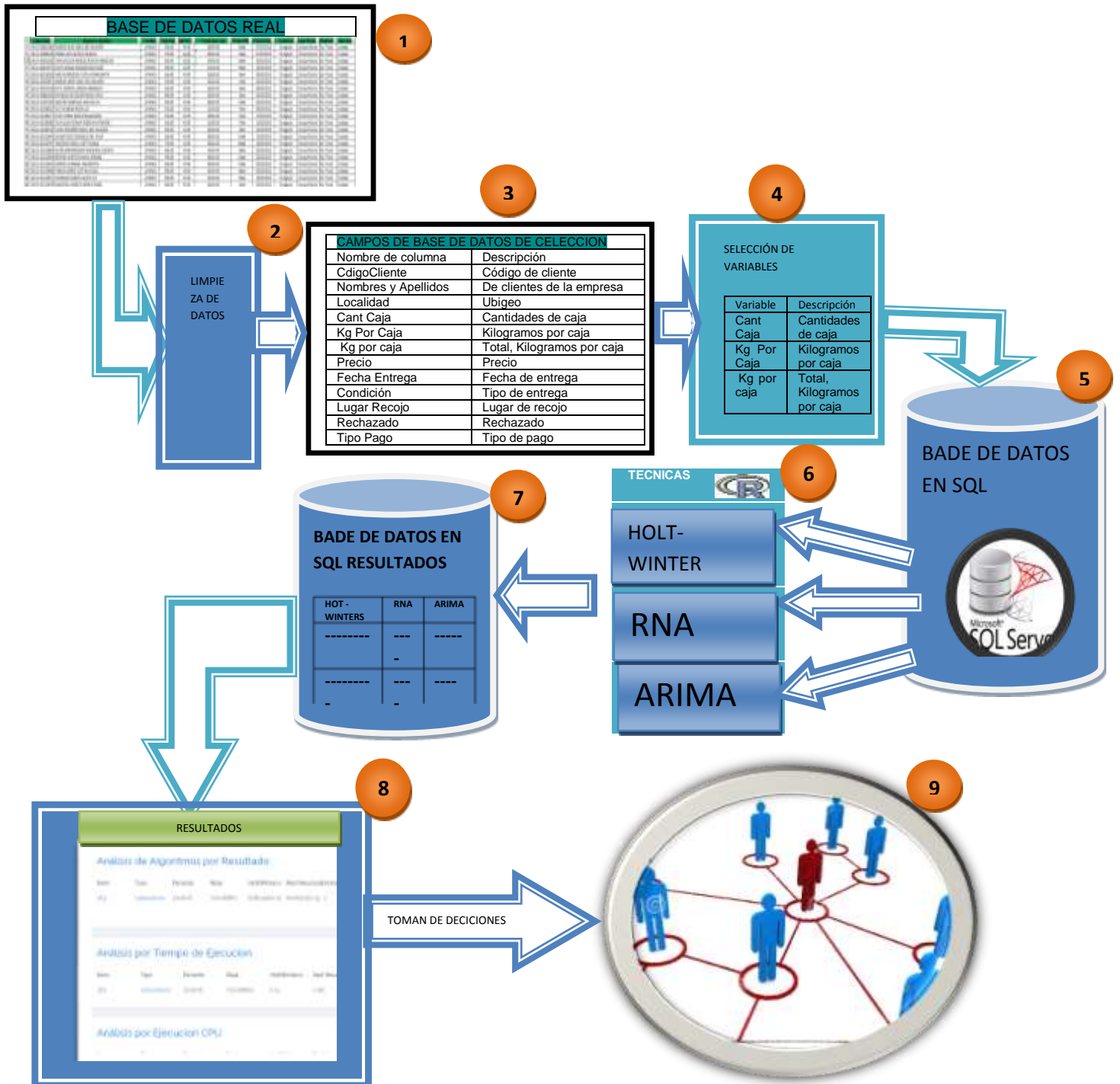
Para la siguiente investigación se ha propuesto dividirla en dos etapas:

Etapa I: Desarrollo de modelos de predicción usando la minería de datos, desde la comprensión del negocio, datos iniciales, transformación de datos, modelado y aplicación del algoritmo, evaluación de performance.

Etapa II: Detallar las fases para el diseño y construcción del sistema web, y como se mostrarán los resultados que permitan a los supervisores y analistas mejorar el análisis de producción e incluyendo las ventas. Resaltar que en esta etapa se empleará la Metodología de desarrollo ágil XP.



MODELO DE DESARROLLO



Se aplica el siguiente marco conceptual para el desarrollo de esta investigación:

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

MATRIZ DE SELECCIÓN DE METODOLOGIA A UTILIZAR

Se realiza a través de dos metodologías **CRISP-DM** y **SEMMA**, se selecciona de acuerdo a los procesos, facilidades y flexibilidad que ofrece cada metodología y que se adecua a los procesos del desarrollo.

Descripción.

Si = 2 (Cuenta con una elección de herramientas gratuitas)

No = 0 (las fases que ofrece la metodología) (Que si ofrece con una Comercial – Licencias - Privativa)

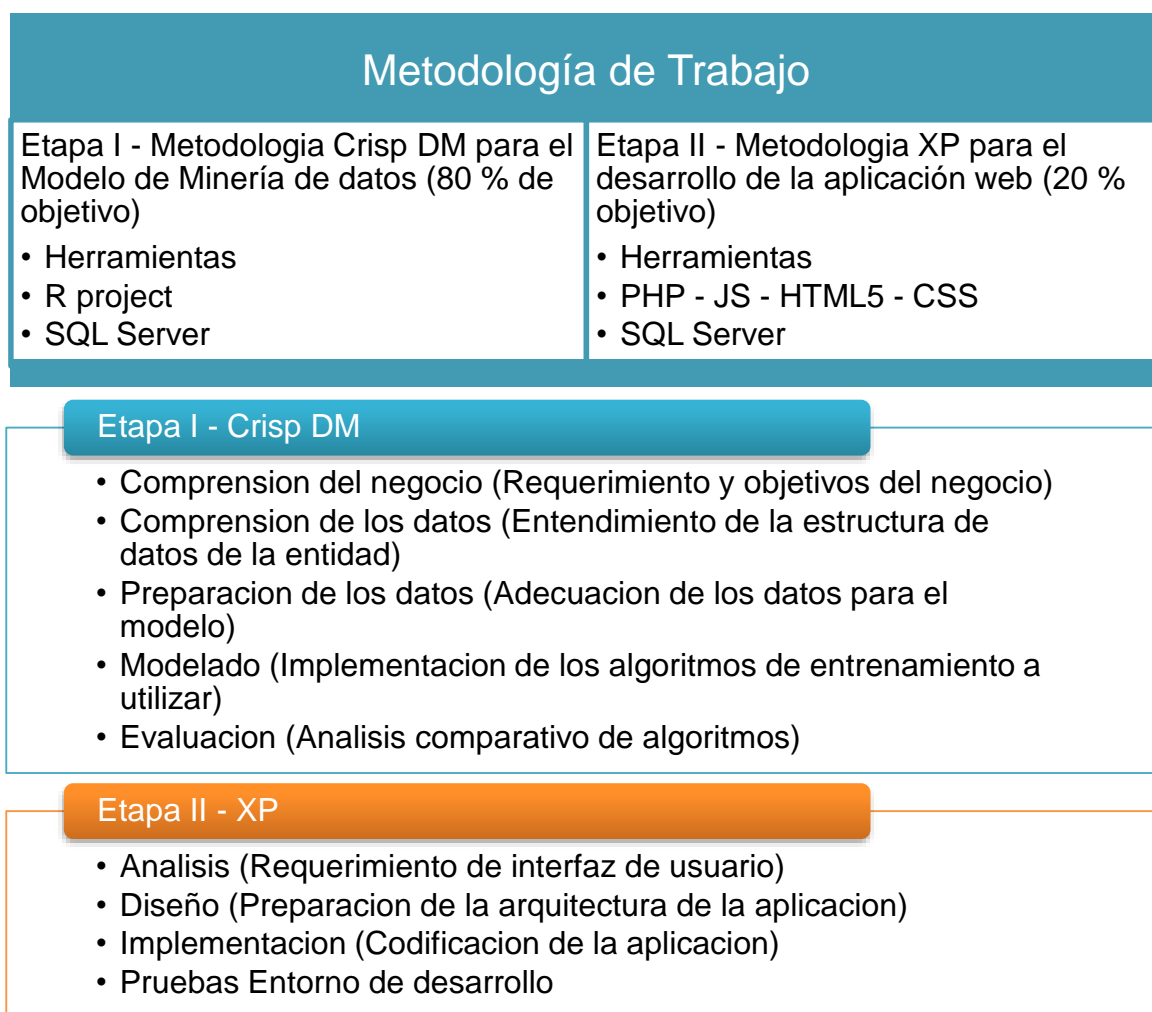
Tabla N° 20: Comparación de Metodologías de Desarrollo de Modelo de Minería de Datos

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

Fuente: Elaboración Propia

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

Figura N° 14: Etapas de metodologías



Fuente: Elaboración propia

Cabe recalcar que ambas etapas no son consecutivas, aunque si depende una de la otra para su funcionamiento, su desarrollo puede ser dado en un escenario



de paralelismo, es decir si bien la aplicación web necesita tener un modelo funcionando con datos para poder visualizar, la construcción de la interfaz web se puede dar desde el momento en que se determinan los objetivos del negocio del modelo de minería

5.3. Etapas

5.2.1 Etapa I diseño del modelado de minería de datos

A. En este objetivo se recopiló información histórica del estado del proceso para el pronóstico de producción de espárragos

5.3.1.1. Comprensión de Negocio

A. Descripción del problema

La empresa agro exportadora se dedica a la actividad agroindustrial a través del cultivo, empaque y exportación de productos frescos como el espárrago, uva, palta, mandarina. Para ello cuenta con personal altamente calificado y se esfuerza para que los productos lleguen a su destino en perfectas condiciones de higiene, seguridad y acorde con las exigencias del mercado nacional e internacional.

La empresa se encuentra ubicada en la región Lambayeque, donde se empacan la producción de espárragos que cumplen con los más altos estándares internacionales de calidad.

Su ubicación estratégica hace posible el desplazamiento oportuno de los productos, manteniendo a salvo su frescura.

El espárrago verde es de un color verde-violeta, y cuenta con una yema comestible totalmente exquisita que, generalmente, no excede del 20% de su tamaño total.

a) Producción:

El paso inicial del proceso de facturación de una empresa agroexportadora se define por considerar las variables de ponderación y cálculo del valor a facturar, estos cálculos están sujetos a cambios por lo que es necesario realizar



una actualización de los pliegos de producción de espárragos antes de generar los valores de cálculos de monto.

b) Trabajo de recojo de producción de espárrago

Es el paso por el cual se establece el cronograma de actividades la verificación de la cantidad de producción de espárrago que entrega cada cliente o agricultor en cada periodo o cada campaña, para lo cual se tienen un tiempo asignado para realizar dicho proceso.

c) Clientes o agricultores

Es la actividad por la cual personal de recojo del espárrago, y extrae las cantidades de espárrago recogida

Figura N° 15: Lista de clientes que entregan producción a la empresa

| Nombre y Apellidos | Lugar | Stock Lit | No. Pkg | Valor Stock Lit | Precio Kg | Fecha entrega |
|---------------------------------------|---------|-----------|---------|-----------------|-----------|---------------|
| ROMERO DIAZ ADILA DEL MILAGRO | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 07/05/2013 |
| ECORREA MRO GLORIA DELMIRA | JAYANCA | 454.00 | 40.00 | 34240.00 | 13344 | 07/05/2013 |
| CARRUAJULLA ANGELES ROSA EVANGELINA | JAYANCA | 854.00 | 40.00 | 34240.00 | 13344 | 30/05/2013 |
| PACHO ARNADO ORLANDO BALTAZAR | JAYANCA | 854.00 | 40.00 | 34240.00 | 13344 | 23/05/2013 |
| SANCIA MENDOZA CINTHA MARGARITA | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 26/05/2013 |
| ANGELES MORI ANALI DEL MILAGRO | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 14/05/2013 |
| EFFIO CAMPOS LORENA ANAMARIA | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 09/05/2013 |
| ESPINOZA DE ESQUEW MARIA CRUZ | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 04/05/2013 |
| SANCHEZ GONZALEZ ANA CECLIA | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 05/05/2013 |
| RUIZ HUMÁN ROSA LUZ | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 09/05/2013 |
| ELIAS HORNIA MARIA MAGDALENA | JAYANCA | 454.00 | 40.00 | 18240.00 | 43940 | 27/05/2013 |
| CHURILLON YOVERA YESENIA KATHERINE | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 12/05/2013 |
| TABRO OROÑOZ MARIA DEL MILAGRO | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 04/05/2013 |
| SALVEZ POZO CONSUELO DEL PILAR | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 11/05/2013 |
| TANTARICO MESA LIZET YOVANA | JAYANCA | 754.00 | 40.00 | 30240.00 | 103940 | 16/05/2013 |
| ACUNA MONTENEGRO MARGORIE SUANITA | JAYANCA | 854.00 | 40.00 | 34240.00 | 13344 | 05/05/2013 |
| BECERRA RUESTAS MARIA ROSANA | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 05/05/2013 |
| CAMPOS CHANAME ANA BERTHA | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 05/05/2013 |
| PEREDA MUÑOZ JUSTINA HILDA | JAYANCA | 854.00 | 40.00 | 34240.00 | 13344 | 05/05/2013 |
| CARRANZA RAMOS SAIDY LUZ | JAYANCA | 854.00 | 40.00 | 34240.00 | 13344 | 18/05/2013 |
| SANDOVAL ASENCIO MARIA ISABEL | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 05/05/2013 |
| DIAZ CAMPOS SHEYLA MARQUEL | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 05/05/2013 |
| FLORES MORENO LILIANA | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 05/05/2013 |
| VARGAS VERGARA LUIS ALBERTO | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 11/05/2013 |
| INFRO CUSQUE JESUS AMALIA | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 23/05/2013 |
| QUINTANA MONTENEGRO DIANA MARIBEL | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 27/05/2013 |
| SEGURA PEREZ MAIRA JOHANNA | JAYANCA | 454.00 | 40.00 | 18240.00 | 43940 | 27/04/2013 |
| RIVADENDRA YANAVACO JASTIN | JAYANCA | 354.00 | 40.00 | 22240.00 | 77940 | 11/05/2013 |
| CUYMA GUEVARA ISIDRA | JAYANCA | 854.00 | 40.00 | 28240.00 | 81940 | 23/04/2013 |
| BALCA RODRIGUEZ VDA DE MONTALVO DORIS | JAYANCA | 954.00 | 40.00 | 38240.00 | 13344 | 08/05/2013 |
| CANTINA LIGONTOP MARIA DEL PILAR | JAYANCA | 754.00 | 40.00 | 30240.00 | 103940 | 23/04/2013 |

Fuente: Elaboración propia

Se demuestra la cantidad de producción por meses, se calcula la cantidad total de espárrago del mes por el precio que cuesta cada kg acumulando un total en soles, se genera una boleta (Los montos que figuran son referenciales).



En general es como determinar cuándo la producción es variante o inconsistente al mes, a partir de que principios se debe considerar que la producción es mayor o menor a la anterior o desde el punto de vista hacer un pronóstico para en mes o periodo que viene, un primer criterio de análisis determina que si se usa el promedio de producción se puede identificar la correlación de la producción, para ello actualmente se usa criterios como producción o cantidad promedio, producción del mes anterior y producciones actuales.

Tabla N° 21: Periodo - Producción

| PERIODO | REGISTROS |
|------------------------|--------------|
| ENERO - DICIEMBRE 2012 | 1809 |
| ENERO - DICIEMBRE 2013 | 7048 |
| ENERO - DICIEMBRE 2014 | 8603 |
| ENERO - DICIEMBRE 2015 | 4725 |
| TOTAL | 22185 |

Fuente: Elaboración Propia

Es uno de los principales procesos que realizan

b. Necesidades y Expectativas

b.1. Búsqueda de la mejora en las predicciones con respecto a la producción de espárragos en un periodo de tiempo determinado.

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

B. Objetivos del Negocio



- a) Analizar tendencias de predicción con respecto a la producción de espárragos.
- b) Realizar pronósticos de producción de forma anual, mensual y trimestral, con base en un nivel de confianza previamente definido en un periodo determinado.

I. Criterios de Éxito

- a) Confiabilidad de los pronósticos arrojados en un determinado periodo.
- b) Facilidad de acceso en la interacción del usuario al portal web.

II. Evaluación de la situación

e.1 Se cuenta con la base de datos de producción anual registrada en Excel por empresa.

Esta información es usada como fuente principal en el ingreso de los datos necesarios para la creación del modelo de series de tiempo.

III. Requerimientos

f.1 El sistema debe permitir generar la visualización de las predicciones de la producción de esparrago en tiempos anuales, mensuales y por periodos.

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

IV. Restricciones

- a) Se requiere la base de datos de toda desde hace 5 años de antigüedad como mínimo para los procesos de entrenamiento y testeo del modelo.
- b) De la información obtenida, los datos deben estar libre de errores y valores en valores en blanco.



V. Determinación de los Objetivos de minería de datos

a. Objetivos del proyecto

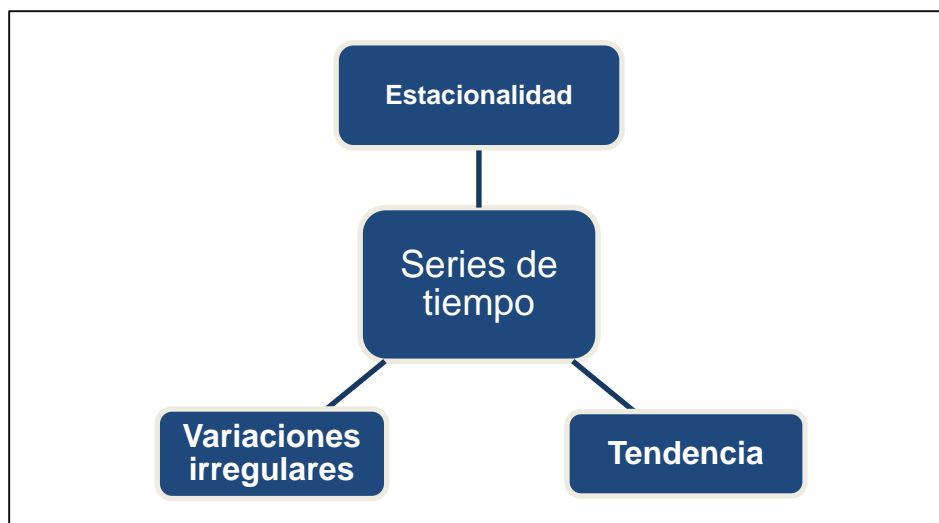
- a) Generar pronósticos de series de tiempo.
- b) Entrenar el modelo para su mejor eficiencia.
- c) Testear el modelo para el resultado.
- d) Criterios de éxito del proyecto
- e) Confiabilidad del modelo diseñado e implementado.
- f) Optimización del tiempo para la generación de reportes.

C. Comprensión de datos

a) Análisis descriptivo de una serie de tiempo

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

Figura N° 16: Series de Tiempo



Fuente: Elaboración propia

D. Recolección de los Datos del Negocio Iniciales

Proceso de Adquisición

Los datos obtenidos corresponden a la producción de espárrago de forma anual y mensual. En registros Excel

Figura N° 17: Archivos de data original

| Nombre | Fecha de modifica... | Tipo | Tamaño |
|--------------------|-----------------------|----------------------|----------|
| ~SAÑO 2015 -- 4725 | 16/06/2015 12:11 a... | Hoja de cálculo d... | 1 KB |
| AÑO 2012 -- 1809 | 05/06/2018 09:40 a... | Hoja de cálculo d... | 165 KB |
| AÑO 2013 -- 7048 | 15/11/2016 07:30 a... | Hoja de cálculo d... | 645 KB |
| AÑO 2014 -- 8603 | 15/11/2016 07:31 a... | Hoja de cálculo d... | 2,110 KB |
| AÑO 2015 -- 4725 | 15/11/2016 07:32 a... | Hoja de cálculo d... | 424 KB |

Fuente: Elaboración propia

D. Selección de las Variables a utilizar

Tiempo: Atributo principal para el proceso de predicción, ya que permite el ordenamiento de la serie. Se tomarán datos desde el año 2012-2015.

Producción: Atributo que contiene la información de los resultados de producción anual y mensual. Datos guardados en Microsoft Excel y que serán migrados manualmente a la base de datos creada para el proceso de entrenamiento del modelo creado con el software R-Project.

E. Datos y métodos de captura

Los datos son extraídos de la base transaccional de producción, datos guardados en Microsoft Excel

Figura N° 18: Datos extraídos originales

| PERIODO Año - 2012 | | | | | | | | | |
|--------------------|-----------------|--|-----------|-------------|-----------|--------|---------------|-----------|--------------|
| ID | DESCRIPCION | LOCALIDAD | CANT CAJA | KG POR CAJA | TOTAL, KG | PRECIO | FECHA ENTREGA | CONDICION | LUGAR RECOJO |
| 4 | 120-11-0002746 | SUNYÓL SANTA CRUZ JHAH | 220.00 | 40.00 | 8800.00 | 39000 | 11/12/2012 | Entregado | CampoPlanta |
| 5 | 120-11-0018881 | DETESA CAMPOS SATHA DEL PLAZ | 300.00 | 40.00 | 12000.00 | 44800 | 02/12/2012 | Entregado | CampoPlanta |
| 6 | 120-11-0019420 | MAGO EFUS BERTHA | 300.00 | 40.00 | 12000.00 | 44800 | 11/12/2012 | Entregado | CampoPlanta |
| 7 | 120-11-00860176 | SAGOMA GÓMEZ LUZ NEILTA | 600.00 | 40.00 | 24000.00 | 86800 | 10/12/2012 | Entregado | CampoPlanta |
| 8 | 120-11-00797132 | MARVAS CHUGUIMANHO LUZ LUISA | 420.00 | 40.00 | 16800.00 | 58800 | 03/12/2012 | Entregado | CampoPlanta |
| 9 | 120-11-00703076 | REGULMI CAMPOS TECOCOLINDA | 500.00 | 40.00 | 20000.00 | 71800 | 21/12/2012 | Entregado | CampoPlanta |
| 10 | 120-11-00956234 | ALCAS GÓMEZ MABEL PATRICIA | 420.00 | 40.00 | 16800.00 | 58800 | 24/12/2012 | Entregado | CampoPlanta |
| 11 | 120-11-00041330 | GONZALEZ BOMBIO KRASNA DEL ROSARIO | 500.00 | 40.00 | 20000.00 | 72800 | 25/09/2012 | Entregado | CampoPlanta |
| 12 | 120-11-00041380 | PLORES FARRAN HERMINIA MADALEYNE | 600.00 | 40.00 | 24000.00 | 86800 | 25/09/2012 | Entregado | CampoPlanta |
| 13 | 120-11-00041434 | LORONEL LLATAS JUAN CARLOS | 600.00 | 40.00 | 24000.00 | 86800 | 25/09/2012 | Entregado | CampoPlanta |
| 14 | 120-11-00041476 | POBRES CARRIO VICTOR | 500.00 | 40.00 | 20000.00 | 72800 | 25/09/2012 | Entregado | CampoPlanta |
| 15 | 120-11-00041487 | BHARACCHIN MEDINA ANGELICA | 500.00 | 40.00 | 20000.00 | 72800 | 25/09/2012 | Entregado | CampoPlanta |
| 16 | 120-11-00041523 | MIRALLES GONZALEZ VIRGINIA DEL PLAZ | 220.00 | 40.00 | 8800.00 | 30200 | 21/12/2012 | Entregado | CampoPlanta |
| 17 | 120-11-00041882 | MOSCOLONI GUARTERAS NAATRE VICTORIA | 200.00 | 40.00 | 8000.00 | 28000 | 11/12/2012 | Entregado | CampoPlanta |
| 18 | 120-11-00041874 | GILSEP LUDIANO ENAYA DEL ROSARIO | 404.00 | 40.00 | 16160.00 | 55360 | 14/12/2012 | Entregado | CampoPlanta |
| 19 | 120-11-00041957 | SAPATA DEL LA CRUZ JESSICA DEL ROSARIO | 324.00 | 40.00 | 12960.00 | 45360 | 14/12/2012 | Entregado | CampoPlanta |
| 20 | 120-11-00041978 | REGULMI BAUTISTA MARIA CELICIA | 220.00 | 40.00 | 8800.00 | 30600 | 14/12/2012 | Entregado | CampoPlanta |
| 21 | 120-11-00041988 | PERMANEZ ROJAS LILIANA | 200.00 | 40.00 | 8000.00 | 28800 | 14/12/2012 | Entregado | CampoPlanta |
| 22 | 120-11-00041954 | MIRANTE PIRUZACA PAULA VIRGINIA | 420.00 | 40.00 | 16800.00 | 58800 | 14/12/2012 | Entregado | CampoPlanta |
| 23 | 120-11-00041979 | VILLALBOS VARGAS MARIA ISABEL | 500.00 | 40.00 | 20000.00 | 72800 | 11/12/2012 | Entregado | CampoPlanta |
| 24 | 120-11-00041989 | GASTILO GÁMARRA MARIA AURORA | 500.00 | 40.00 | 20000.00 | 72800 | 11/12/2012 | Entregado | CampoPlanta |
| 25 | 120-11-00041989 | IBAN GASTILO MIRA LUMITH | 420.00 | 40.00 | 16800.00 | 58800 | 03/12/2012 | Entregado | CampoPlanta |
| 26 | 120-11-00042023 | POJAS SANCHEZ JOSE GABRIEL | 424.00 | 40.00 | 16960.00 | 57360 | 04/12/2012 | Entregado | CampoPlanta |
| 27 | 120-11-00042051 | MEJIA TRAZADO LINDA | 312.00 | 40.00 | 12480.00 | 44160 | 11/12/2012 | Entregado | CampoPlanta |
| 28 | 120-11-00042084 | PIMÓN MUÑOZ JUSTINA SILVIA | 312.00 | 40.00 | 12480.00 | 44160 | 11/12/2012 | Entregado | CampoPlanta |
| 29 | 120-11-00042094 | SANCES GUPTANA ESTHER | 413.00 | 40.00 | 16520.00 | 58160 | 11/12/2012 | Entregado | CampoPlanta |
| 30 | 120-11-00042098 | CAMACHO CHUGUIMANHO SAMARA LUISA | 312.00 | 40.00 | 12480.00 | 44160 | 11/12/2012 | Entregado | CampoPlanta |
| 31 | 120-11-00042113 | HERNANDEZ HUAYANA ROSA ANGELICA | 413.00 | 40.00 | 16520.00 | 58160 | 09/12/2012 | Entregado | CampoPlanta |
| 32 | 120-11-00042113 | DE LA CRUZ SUCOBIA ROSA MARIA | 418.00 | 40.00 | 16720.00 | 58520 | 04/12/2012 | Entregado | CampoPlanta |
| 33 | 120-11-00042154 | VALQUEZ ALCAS BIANCA ROSA | 309.00 | 40.00 | 12360.00 | 43680 | 06/12/2012 | Entregado | CampoPlanta |
| 34 | 120-11-00042155 | SANCHEZ BAUTISTA BEBELINDA | 304.00 | 40.00 | 12160.00 | 43160 | 11/12/2012 | Entregado | CampoPlanta |

Fuente: Elaboración propia

Esta información está compuesta por las siguientes columnas

Tabla N° 22: Descripción de la tabla de la base de datos

| | |
|---------------------|--------------------------------|
| Nombro de columna | Descripción |
| CdigoCliente | Código de cliente |
| Nombres y Apellidos | Nombres y apellidos de cliente |
| Localidad | Ubigeo |
| Cant Caja | Cantidades de caja |
| Kg Por Caja | Kilogramos por caja |
| Total, Kg | Total, Kilogramos por caja |
| Precio | Precio |
| Fecha Entrega | Fecha de entrega |
| Condición | Tipo de entrega |
| Lugar Recojo | Lugar de recojo |
| Rechazado | Rechazado |
| Tipo Pago | Tipo de pago |

Fuente: Elaboración Propia

DETALLE DE CAMPOS DE BASE DATOS REALES

Código de cliente: Código generado de cada cliente según su base datos origina

Nombre y apellidos: Contiene los nombre y apellido de los clientes que entregan producción a la empresa



Localidad: Lugar de donde la empresa recoge la producción de cada cliente

Cant Caja: Cantidad de cajas a recoger

Kg Por Caja: La cantidad de Kg que contiene cada caja empacada

Total, Kg: El total de kg de producción

Precio: El precio total a pagar a cada cliente por su producción

Fecha de entrega: contiene la fecha exacta de entrega de producción

Condición: Es el tipo de producto de entrega ya sea esparrago verde o blanco

Lugar de recojo: Contiene el lugar o el campo o chacra de regajo.

Rechazado: Contiene los datos de los kilos rechazados por control de calidad

Tipo de pago: Contiene los datos que si fue pagado al contado o al final de la cosecha

F. Exploración de Datos

La construcción del modelo de predicción se desarrolla con información obtenida desde el año 2012 hasta el año 2015. Estos datos son los que ingresan en una pequeña base de datos obtenida por la migración de datos en repositorios ofimáticos a la base de datos en el gestor SQL Server 2014 para que realice el entrenamiento del modelo; de los cuales se utiliza el 70% para el entrenamiento y el 30% para las pruebas de predicciones.

Figura N° 19: producción de espárragos

| Id | Localidad | Cant. Caja | Kg Por Caja | Total, Kg | Precio | Fecha de entrega | Condición | Lugar de recojo | Rechazado | Tipo de pago |
|----|-----------|------------|-------------|-----------|--------|------------------|-----------|-----------------|-----------|--------------|
| 1 | JAYAPUCA | 1.10 | 100 | 110 | 110 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 2 | JAYAPUCA | 1.20 | 100 | 120 | 120 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 3 | JAYAPUCA | 1.30 | 100 | 130 | 130 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 4 | JAYAPUCA | 1.40 | 100 | 140 | 140 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 5 | JAYAPUCA | 1.50 | 100 | 150 | 150 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 6 | JAYAPUCA | 1.60 | 100 | 160 | 160 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 7 | JAYAPUCA | 1.70 | 100 | 170 | 170 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 8 | JAYAPUCA | 1.80 | 100 | 180 | 180 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 9 | JAYAPUCA | 1.90 | 100 | 190 | 190 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 10 | JAYAPUCA | 2.00 | 100 | 200 | 200 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 11 | JAYAPUCA | 2.10 | 100 | 210 | 210 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 12 | JAYAPUCA | 2.20 | 100 | 220 | 220 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 13 | JAYAPUCA | 2.30 | 100 | 230 | 230 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 14 | JAYAPUCA | 2.40 | 100 | 240 | 240 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 15 | JAYAPUCA | 2.50 | 100 | 250 | 250 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 16 | JAYAPUCA | 2.60 | 100 | 260 | 260 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 17 | JAYAPUCA | 2.70 | 100 | 270 | 270 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 18 | JAYAPUCA | 2.80 | 100 | 280 | 280 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 19 | JAYAPUCA | 2.90 | 100 | 290 | 290 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 20 | JAYAPUCA | 3.00 | 100 | 300 | 300 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 21 | JAYAPUCA | 3.10 | 100 | 310 | 310 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 22 | JAYAPUCA | 3.20 | 100 | 320 | 320 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 23 | JAYAPUCA | 3.30 | 100 | 330 | 330 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 24 | JAYAPUCA | 3.40 | 100 | 340 | 340 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 25 | JAYAPUCA | 3.50 | 100 | 350 | 350 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 26 | JAYAPUCA | 3.60 | 100 | 360 | 360 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 27 | JAYAPUCA | 3.70 | 100 | 370 | 370 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 28 | JAYAPUCA | 3.80 | 100 | 380 | 380 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 29 | JAYAPUCA | 3.90 | 100 | 390 | 390 | 2012-01-01 | Verde | Chacra | 0 | Contado |
| 30 | JAYAPUCA | 4.00 | 100 | 400 | 400 | 2012-01-01 | Verde | Chacra | 0 | Contado |

Fuente: Elaboración propia



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

Tabla N° 23: cantidad de cajas y Kg de espárragos

| Años | Cant Cajas | Kg Por Caja | Total Kg por caja | Precio s/0.00 |
|-------------|-------------------|--------------------|--------------------------|----------------------|
| 2012 | 931439.00 | 72360.00 | 37257560.00 | 130401460.00 |
| 2013 | 3307797.00 | 281920.00 | 132311880.00 | 549703660.00 |
| 2014 | 4513857.00 | 343640.00 | 180554280.00 | 631939980 |
| 2015 | 2490629.00 | 189000.00 | 99645160.00 | 348758060 |

Fuente: Elaboración Propia

i. Preparación de los datos

G. Datos Seleccionados

De la base de datos obtenida, se obtienen diferentes tipos de información con respecto a la producción de espárrago, lo cual son datos relevantes, para ello, se ha realizado un análisis de la data a utilizar para el correcto funcionamiento del modelo.

Debe considerarse además que se ha analizado y utilizado los campos Anulada, para el proceso de limpieza de datos.

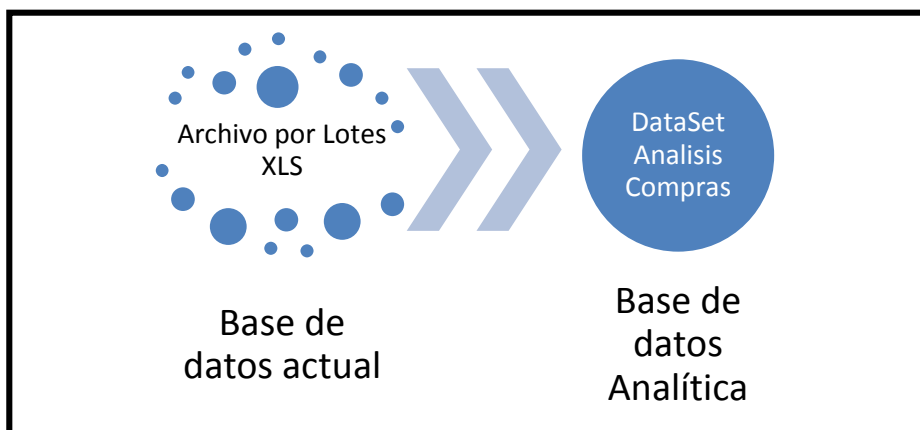
Esta proporcionada por la entidad está en función a archivos ofimáticos de tipo Excel con los registros de compra por los clientes.

Al explorar los datos en la tabla Hecho, contiene un total de 22 000 registros desde el periodo 201201 (enero 2012) hasta el 201505 (mayo 2015).



Mediante esta información recopilada fue muy relevante para extraer y transformar estos datos para realizar pronósticos. Cuenta con la cantidad de registros históricos

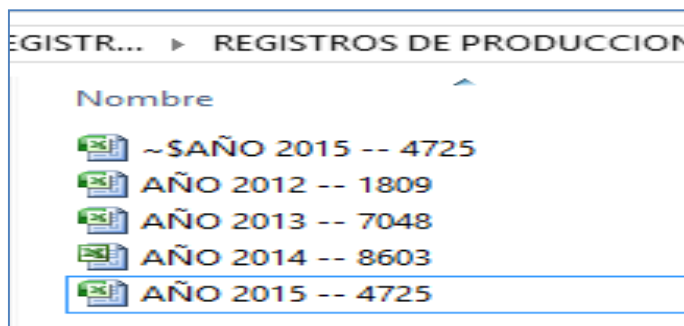
Figura N° 20: Extracción de datos a repositorio de análisis



Fuente: Elaboración Propia

Por lo tanto, se ha considerado como estrategia la migración de estos datos ofimáticos Excel al motor de base de datos SQL Server, donde se realizará el procesamiento analítico, para lo cual se ha desarrollado un pequeño proceso ETL de migración utilizando el lenguaje de programación R.

Figura N° 21: Archivos Excel data original



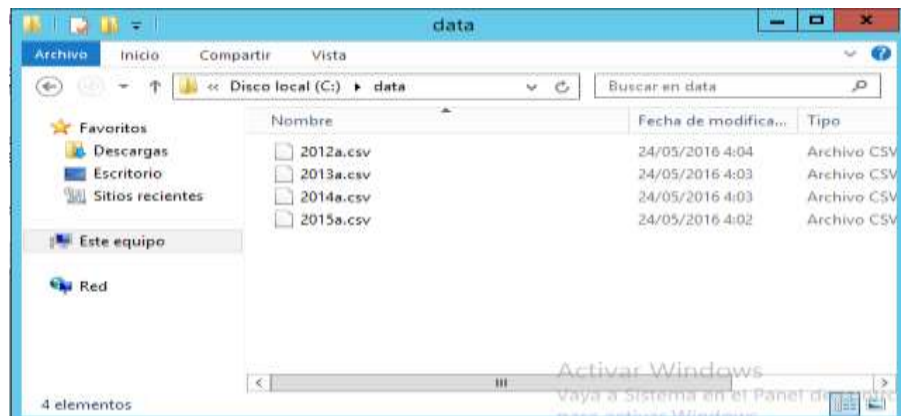
Fuente: Elaboración propia

Se ha diseñado el siguiente procedimiento de ejecución:

Se ha solicitado que se generen los archivos Excel en formato CSV para cada año de la transacción compras con el esquema antes mencionado, dichos archivos se almacenan en una carpeta temporal para ser procesadas por el lenguaje.



Figura N° 22: Lotes de archivos generados para procesar



Fuente: Elaboración Propia

Luego se procede a diseñar el script de extracción y migración al SQL Server.

Figura N° 23: Base de datos poblada SQL Server

Fuente: Elaboración Propia

Vista en SQL Server migración de 22174 registros de Excel

Figura N° 24: Base de datos poblada

| | | | | | |
|----|----|-----------------|---------------------------------------|---------|--------|
| 11 | 14 | 190-11-01041476 | TORRES CASTRO VICTOR | JAYANCA | 520.00 |
| 12 | 15 | 190-11-01041487 | RIMARACHIN MEDINA ANGELICA | JAYANCA | 520.00 |
| 13 | 16 | 190-11-01041525 | MORALES GONZALES VIRGINIA DEL PILAR | JAYANCA | 220.00 |
| 14 | 17 | 190-11-01041662 | MÓGOLLÓN GUARDERAS BEATRIZ VICTORIA | JAYANCA | 220.00 |
| 15 | 18 | 190-11-01041674 | QUISPE LOZANO ENMA DEL ROSARIO | JAYANCA | 424.00 |
| 16 | 19 | 190-11-01041757 | ZAPATA DE LA CRUZ JESSICA DEL ROSARIO | JAYANCA | 324.00 |
| 17 | 20 | 190-11-01041759 | REQUELMÉ BAUTISTA MARIA DELICIA | JAYANCA | 220.00 |

Consulta ejecutada correcta... WIN-3AF20B655FC (12.0 RTM) | Luis (51) | BIEspa | 00:00:00 | 22174 filas

Fuente: Elaboración Propia

Figura N° 25: Esquema de datos generados en SQL Server

| Nombre de columna | Tipo de datos | Permitir val... |
|-------------------|---------------|-------------------------------------|
| rownames | varchar(255) | <input checked="" type="checkbox"/> |
| CodigoCliente | varchar(255) | <input checked="" type="checkbox"/> |
| Nombres | varchar(255) | <input checked="" type="checkbox"/> |
| Localidad | varchar(255) | <input checked="" type="checkbox"/> |
| CantCajas | varchar(255) | <input checked="" type="checkbox"/> |
| KgPorCaja | varchar(255) | <input checked="" type="checkbox"/> |
| TotalKgPorCaja | varchar(255) | <input checked="" type="checkbox"/> |
| Precio | varchar(255) | <input checked="" type="checkbox"/> |
| FechaEntrega | varchar(255) | <input checked="" type="checkbox"/> |
| Condicion | varchar(255) | <input checked="" type="checkbox"/> |
| LugarRecojo | varchar(255) | <input checked="" type="checkbox"/> |
| Rechazado | varchar(255) | <input checked="" type="checkbox"/> |
| TipoPago | varchar(255) | <input type="checkbox"/> |

Fuente: Elaboración Propia

Figura N° 26: ETL SQL Server

```

SELECT
    CASE WHEN ([Fecha entrega]) <= '2012-01-01' THEN '2012-01-01' ELSE [Fecha entrega] END AS [Fecha entrega],
    [Codigo Cliente] AS [Codigo Cliente],
    [Nombres] AS [Nombres],
    [Localidad] AS [Localidad],
    [CantCajas] AS [CantCajas],
    [KgPorCaja] AS [KgPorCaja],
    [TotalKgPorCaja] AS [TotalKgPorCaja],
    [Precio] AS [Precio],
    [FechaEntrega] AS [FechaEntrega],
    [Condicion] AS [Condicion],
    [LugarRecojo] AS [LugarRecojo],
    [Rechazado] AS [Rechazado],
    [TipoPago] AS [TipoPago]
FROM [dbo].[HistCompra]
WHERE ([Fecha entrega] <= '2012-01-01')

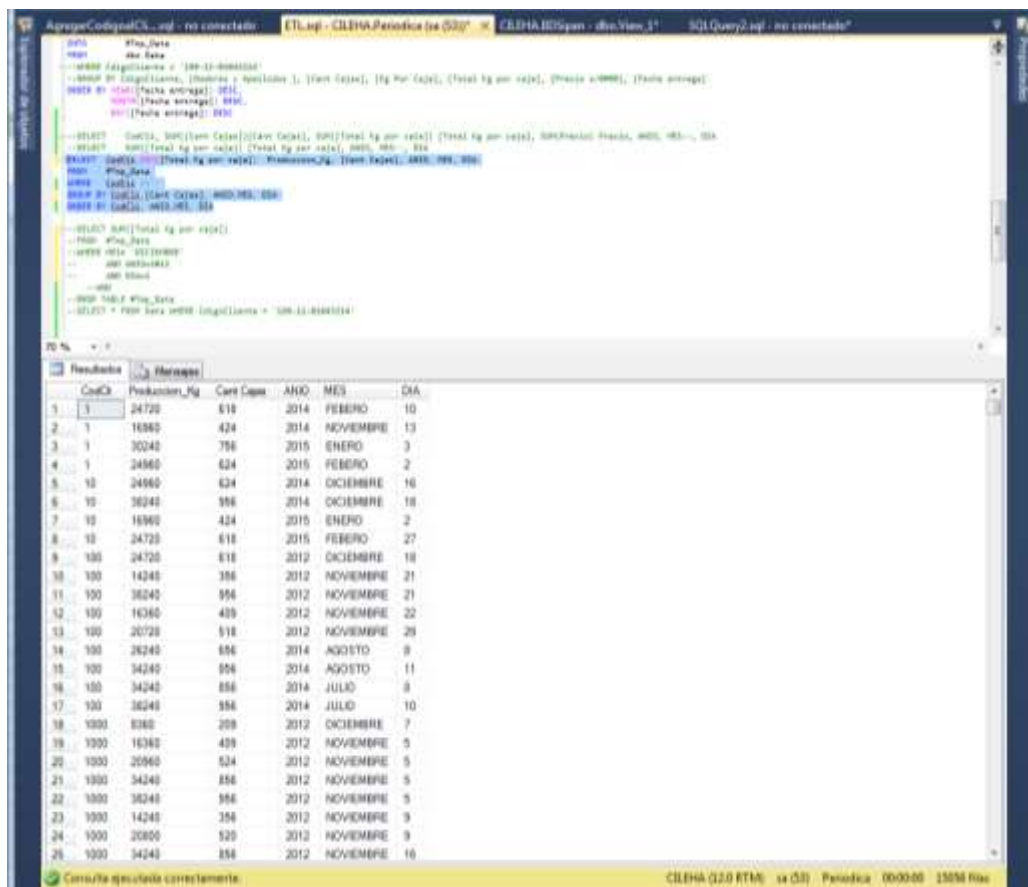
```



Fuente: Elaboración Propia

ETL en SQL Server donde se realiza la limpieza de datos, de los datos de fecha se ha generado el script para obtener el año, el mes en que se genero la producción y se filtrado a los registros con datos nulos quedando de esta manera:

Figura N° 27: ETL SQL Server



Fuente: Elaboración Propia

En este proceso los Clientes no tenían Códigos que los diferenciaba por persona, los códigos que existen en los datos eran de clientes por cada vez que entregaban el producto, entonces se desarrolló un script para asignarles a cada uno:



Figura N° 28: Script SQL para consulta que genera el formato deseado

```

use BIEspa

select * from HistCompra order by rownames

select * from HistCompra
where fechaentrega like '%'

delete from Histcompra where rownames=41001

select top 3359 precio,fechaentrega from HistCompra
where condicion ='Etregado' order by fechaentrega asc

select * from (
select cast(mes as varchar(2))+ '-' + cast(anio as CHAR(4)) as periodo,mes,anio,total from (
select MONTH(fechem) as mes,year(fechem) as anio, sum(pre) as total from (
select convert(date, fechaentrega, 103) as fechem,CAST(precio AS DECIMAL(10, 2)) as pre from H
group by MONTH(fechem),year(fechem)) as fin) as mining
order by anio,mes
) %

```

| periodo | mes | anio | total |
|---------|-----|------|-------------|
| 9-2012 | 9 | 2012 | 1425200.00 |
| 10-2012 | 10 | 2012 | 22288540.00 |
| 11-2012 | 11 | 2012 | 72709480.00 |
| 12-2012 | 12 | 2012 | 44212440.00 |
| 1-2013 | 1 | 2013 | 71785500.00 |
| 2-2013 | 2 | 2013 | 60576120.00 |
| 3-2013 | 3 | 2013 | 56979180.00 |
| 4-2013 | 4 | 2013 | 55681520.00 |
| 5-2013 | 5 | 2013 | 63765720.00 |
| 6-2013 | 6 | 2013 | 41962760.00 |
| 7-2013 | 7 | 2013 | 38081080.00 |
| 8-2013 | 8 | 2013 | 37851400.00 |
| 9-2013 | 9 | 2013 | 35846640.00 |
| 10-2013 | 10 | 2013 | 43852700.00 |
| 11-2013 | 11 | 2013 | 50166140.00 |
| 12-2013 | 12 | 2013 | 33802440.00 |
| 1-2014 | 1 | 2014 | 41332900.00 |
| 2-2014 | 2 | 2014 | 58269820.00 |
| 3-2014 | 3 | 2014 | 51521820.00 |
| 4-2014 | 4 | 2014 | 54815740.00 |
| 5-2014 | 5 | 2014 | 42847700.00 |
| 6-2014 | 6 | 2014 | 43000440.00 |

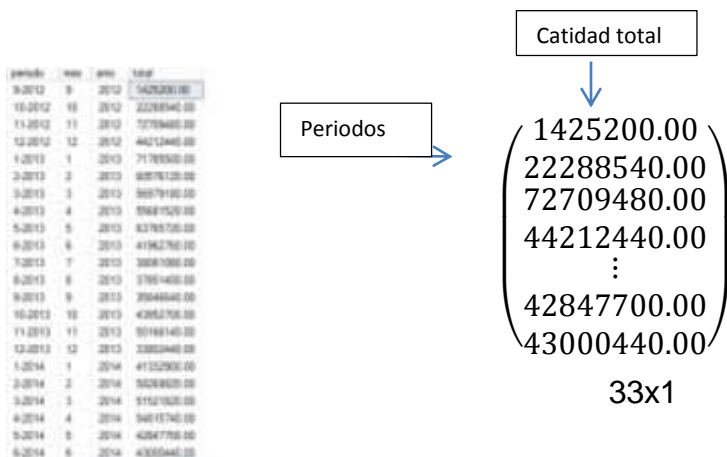
Consultas ejecutadas correctamente. | WIN-3AF20B655FC (12.0 RTM) | Luis (51) | BIEspa | 00:00:00 | 33 filas

Fuente: Elaboración Propia

Y obtenemos un set consolidado de datos que se puede tratar para análisis predictivo, es decir un modelo de series de tiempo que permita predecir el comportamiento de compras de esparrago, denotado en la forma de análisis por frecuencias mensuales.

Representamos los datos en la figura N° 28 en matriz columna de orden 33x1 las cantidades de espárragos de la siguiente manera

MATRIZ COLUMNA DE ORDEN 33X1



Verificamos que este set este transformado en dicha matriz para lo cual utilizaremos código R:

Figura N° 29: Script ejecutado desde R project

```
vec<-sqlQuery(bd, "select total from (
select cast(mes as varchar(2))+ '-' + cast(anio as CHAR(4)) as
periodo,mes,anio,total from (
select MONTH(fechem) as mes,year(fechem) as anio, sum(pre) as total from (
select convert(date, fechaentrega, 103) as fechem,CAST(precio AS
DECIMAL(10, 2)) as pre from HistCompra) as q
group by MONTH(fechem),year(fechem)) as fin) as mining
order by anio,mes")

vec<-as.vector(t(vec))
vec <- as.numeric(vec)

fre<- ts(vec, frequency=12,start=c(2012,09))
```

Fuente: Elaboración Propia

Obteniendo la matriz para que el modelo de minería pueda procesarla.



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

a) Estructuración de los datos

Para la creación del modelo con series de tiempo, los atributos utilizados son identificados de la siguiente manera: al atributo periodo se denota como año y mes; y al atributo monto venta como total, ya que representa el objetivo a predecir, como se muestra en la siguiente imagen. En esta fase preparamos los datos para tener la forma.

Obteniendo la matriz para que el modelo de minería pueda procesarla.

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

Figura N° 30: Vista de datos obtenidos en la matriz desde SQL en R con formato Series de Tiempo

```
> fre<- ts(vec, frequency=12,start=c(2012,09))
> fre
      Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2012
2013 71785500 60576120 56979180 55681520 63765720 41962760 38081080 37851400
2014 41332900 58269820 51521820 54815740 42847700 43000440 43997100 41953940
2015 69463660 68978000 68726560 68789280 72008860
      Sep      Oct      Nov      Dec
2012 1425200 22288540 72709480 44212440
2013 35846640 43852700 50166140 33802440
2014 1319220 63779240 64395520 74496100
2015
> fre
      Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2012
2013 71785500 60576120 56979180 55681520 63765720 41962760 38081080 37851400 35846640 43852700 50166140 33802440
2014 41332900 58269820 51521820 54815740 42847700 43000440 43997100 41953940 1319220 63779240 64395520 74496100
2015 69463660 68978000 68726560 68789280 72008860
> |
```

Fuente: Elaboración Propia

Para ello se diseño matriz de resumen y condensación de datos especializadas para que el modelo puede procesarlas sin inconveniente. Los motivos por el cual se crea la matriz analítica.



Los datos obtenidos en la figura N°30 se representan en una matriz de orden en filas representan los años y en las columnas los meses

MATRIZ DE ORDEN EN FILAS REPRESENTAN LOS AÑOS Y EN LAS COLUMNAS LOS MESES

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 2012 | | | | | | | | | | | | |
| 2013 | 71785500 | 60576120 | 56979180 | 59881520 | 63765720 | 41862760 | 38021080 | 37851400 | 35846640 | 43852700 | 50168140 | 33802440 |
| 2014 | 41332900 | 58249820 | 51521820 | 54815740 | 42847700 | 43000440 | 43997100 | 41853940 | 1318220 | 63779240 | 64385520 | 74496100 |
| 2015 | 69463660 | 68978000 | 68726560 | 68789280 | 72008860 | | | | | | | |

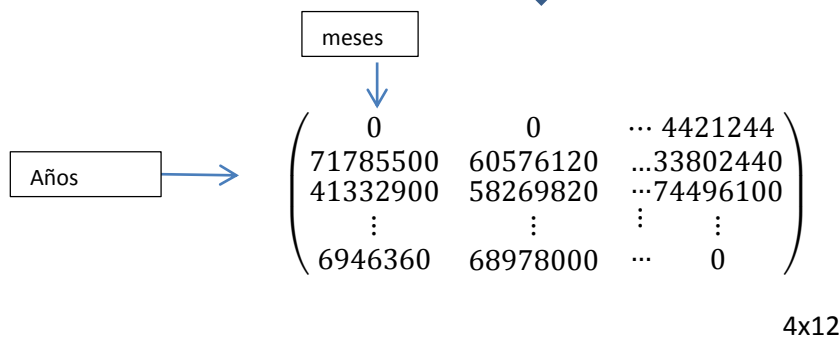


Figura N° 31: ETL Script en R

```
setwd("c:/data")
options(max.print = 99999999)

data1 <- read.csv(file="c:/data/2012a.csv", header=TRUE, sep=",")

data1 <- as.data.frame(data1)
nr <- nrow(data1)
nc <- ncol(data1)
head(data1)

dataset1 <- data1[-c(1:3),]
head(dataset1)
nrow(dataset1)
dataset1[1812,]

dataset1 <- dataset1[c(1:1809),]
head(dataset1)

data2 <- read.csv(file="c:/data/2013a.csv", header=TRUE, sep=",")

data2 <- as.data.frame(data2)
nr <- nrow(data2)
nc <- ncol(data2)
head(data2)

dataset2 <- data2[-c(1:3),]
head(dataset2)
nrow(dataset2)
dataset2[7048,]

dataset2 <- dataset2[c(1:7048),]
head(dataset2)

data3 <- read.csv(file="c:/data/2014a.csv", header=TRUE, sep=",")

data3 <- as.data.frame(data3)
nr <- nrow(data3)
nc <- ncol(data3)
head(data3)

dataset3 <- data3[-c(1:3),]
head(dataset3)
nrow(dataset3)
dataset3[8592,]

dataset3 <- dataset3[c(1:8592),]
head(dataset3)
```

```
data4 <- read.csv(file="c:/data/2015a.csv", header=TRUE, sep=",")

data4 <- as.data.frame(data4)
nr <- nrow(data4)
nc <- ncol(data4)
head(data4)

dataset4 <- data4[-c(1:3),]
head(dataset4)
nrow(dataset4)
dataset4[4725,]

dataset4 <- dataset4[c(1:4725),]
head(dataset4)

setfinal <- rbind(dataset1,dataset2,dataset3,dataset4)

col <-
c("CodigoCliente","Nombres","Localidad","CantCajas","KgPorCaja","TotalKgPor
Caja","Precio","FechaEntrega","Condicion","LugarRecojo","Rechazado","TipoPa
go")

names(setfinal) <- col
head(setfinal)
nrow(setfinal)

setfinal <- setfinal[-c(1:3),]

library("RODBC")
bd <- odbcDriverConnect('driver={SQL
Server};server=localhost;database=BIEspa;trusted_connection=true')

sqlSave(bd,data.frame(setfinal),"HistCompra",safer=FALSE,append=TRUE)

del<-sqlQuery(bd, "delete from Histcompra where rownames=41001")
```

Fuente: Elaboración Propia



Figura N° 32: Script para leer CSV

```

RGui (64-bit)
Archivo  Editar  Visualizar  Misc  Paquetes  Ventanas  Ayuda

R Console

setwd("c:/data")
data1 <- read.csv(file="c:/data/2012a.csv", header=TRUE, sep=",")
data1 <- as.data.frame(data1)
nr <- nrow(data1)
nc <- ncol(data1)
h <- head(data1)
h

```

| BETA.S.A.REGISTRO.GENERAL.DE.COMPRAS.DE.ESPARRAGO..FUNDO.SEDE.JAYANCA.LAMBAYEQUE.PERU. | | PERIODO Año - 2012 | | | |
|--|--------------------------------|--------------------|-----------|--------|---------------|
| | | CodigoCliente | | | |
| | | 190-11-00035740 | | | |
| | | 190-11-00038921 | | | |
| | | 190-11-00194820 | | | |
| | X | X.1 | X.2 | X.3 | |
| 1 | | | | | |
| 2 | | | | | |
| 3 | Nombres y Apellidos | | Localidad | Cant | |
| 4 | OLIVOS SANTA CRUZ JULIO | | JAYANCA | 20.00 | |
| 5 | ORTEGA CAMPOS KATHIA DEL PILAR | | JAYANCA | 320.00 | |
| 6 | MACO EFUS BERTILA | | JAYANCA | 320.00 | |
| | X.4 | X.5 | X.6 | X.7 | X.8 |
| 1 | | | | | |
| 2 | | | | | |
| 3 | Total Kg | por caja | Precio | s/0.00 | Fecha entrega |
| 4 | 8800.00 | | 30800 | | 11/12/2012 |
| 5 | 12800.00 | | 44800 | | 22/11/2012 |
| 6 | 12800.00 | | 44800 | | 11/12/2012 |
| | X.9 | X.10 | | | |
| 1 | | | | | |
| 2 | | | | | |
| 3 | Rechazadas | Tipo_Pago | | | |
| 4 | No-Tiene | Contado | | | |
| 5 | No-Tiene | Contado | | | |
| 6 | No-Tiene | Contado | | | |

```

Total Kg por caja Precio s/0.00 Fecha entrega Condicion Lugar Recojo
8800.00 30800 11/12/2012 Etregado CampoPlanta
12800.00 44800 22/11/2012 Etregado CampoPlanta
12800.00 44800 11/12/2012 Etregado CampoPlanta

```

Fuente: Elaboración Propia

En esta figura se muestra la extracción de de los datos de SQL Server a R para su análisis de los algoritmos

- Establecemos conexión con BD mediante RODBC

Figura N° 33: Librería para conexión ODBC

```

> library("RODBC")
> bd <- odbcDriverConnect('driver={SQL Server};server=localhost;database=BIEspa;trusted_connection=true')
> bd
RODBC Connection
Details:
 case=nochange
 DRIVER=SQL Server
 SERVER=localhost
 UID=Administrador
 Trusted_Connection=Yes
 WSID=WIN-3AF20B6S5FC
 DATABASE=BIEspa
>

```

Fuente: Elaboración propia

- Se procede al diseño del algoritmo



Figura N° 34: Datos antes de volcar a SQL Server

```
> head(setfinal)
CodigoCliente      Nombres Localidad CantCajas KgPorCaja
4 190-11-00035740  OLIVOS SANTA CRUZ JULIO  JAYANCA  220.00  40.00
5 190-11-00038921  ORTEGA CAMPOS KATHIA DEL PILAR  JAYANCA  320.00  40.00
6 190-11-00194820  MACO EFUS BERTILA  JAYANCA  320.00  40.00
7 190-11-00460179  GAONA GOMES LUZ NELITA  JAYANCA  620.00  40.00
8 190-11-00757132  VARGAS CHUQUIMANGO LUZ LIDIA  JAYANCA  420.00  40.00
9 190-11-00763576  REQUELME CAMPOS TEODOLINDA  JAYANCA  520.00  40.00
>
```

Fuente: Elaboración Propia.

- Vista previa en R

Figura N° 35: Datos listos en R para SQL Server

```
> head(setfinal)
CodigoCliente      Nombres Localidad CantCajas KgPorCaja
4 190-11-00035740  OLIVOS SANTA CRUZ JULIO  JAYANCA  220.00  40.00
5 190-11-00038921  ORTEGA CAMPOS KATHIA DEL PILAR  JAYANCA  320.00  40.00
6 190-11-00194820  MACO EFUS BERTILA  JAYANCA  320.00  40.00
7 190-11-00460179  GAONA GOMES LUZ NELITA  JAYANCA  620.00  40.00
8 190-11-00757132  VARGAS CHUQUIMANGO LUZ LIDIA  JAYANCA  420.00  40.00
9 190-11-00763576  REQUELME CAMPOS TEODOLINDA  JAYANCA  520.00  40.00
>
```

Fuente: Elaboración Propia

B. Mediante este objetivo es analizar los algoritmos en la fase de modelado predictivo.

i. Modelado análisis de los algoritmos

En la investigación se propone construir un modelo de minería de datos utilizando técnicas de pronósticos, a continuación, se presenta la tabla que se realizó para la selección de las técnicas adecuadas.

Evaluación de técnicas del modelo de minería de datos



Tabla N° 24: Evaluación de las técnicas a utilizar

| TÉCNICA DE MINERÍA DE DATOS | DESCRIPCIÓN DE LA TÉCNICA | ALGORITMOS | ¿ES ADECUADO PARA LA INVESTIGACIÓN ? | DESCRIPCION EL PORQUE |
|-----------------------------|---|--|--------------------------------------|--|
| REGRESIÓN | | Redes Neuronales, Maquinas de soporte Vectorial(SVM) | SI | Porque tiene 3 componentes Estacionalidad, Serie de Tiempo y Tendencia. Resultados aleatorios |
| SERIES TEMPORALES | | Holtwinters, Arima, entre otras | SI | Porque tiene 4 componentes Estacionalidad, Serie de Tiempo y Tendencia métodos estadísticos Resultados fijos |
| CLASIFICACION AD HOC | Basado en reglas por construcciones lógicas múltiples variables | Árbol de decisiones, Redes bayesianas SVM | NO | Generalmente son de búsqueda binaria |

Fuente: Elaboración propia

En la evaluación de las técnicas a utilizar se realiza con una descripción de la técnica y si es adecuada a esta investigación

SI. (Si cumple y es adecuada para aplicar como técnica de minería de datos)

NO. (No cumple y no es adecuada para aplicar como técnica de minería de datos)



Descripción. En la evaluación realizada se selecciona las técnicas de HoltWinter, Arima y Red Neuronal Autorregresiva, quedando descartado la técnica de redes Bayesianas y árbol de decisiones, SVM

Evaluación de técnicas del modelo de minería de datos

Para lo cual se han establecido los siguientes criterios de evaluación de los algoritmos a utilizar.

Descripción.

X. (Representa 1 marcación de columna indicando con que característica cuenta) y en la cantidad de datos de la serie se realiza una sumatoria de los campos marcados según su evaluación de cada técnica

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

Tabla N° 25: Criterios de evaluación de las técnicas a utilizar

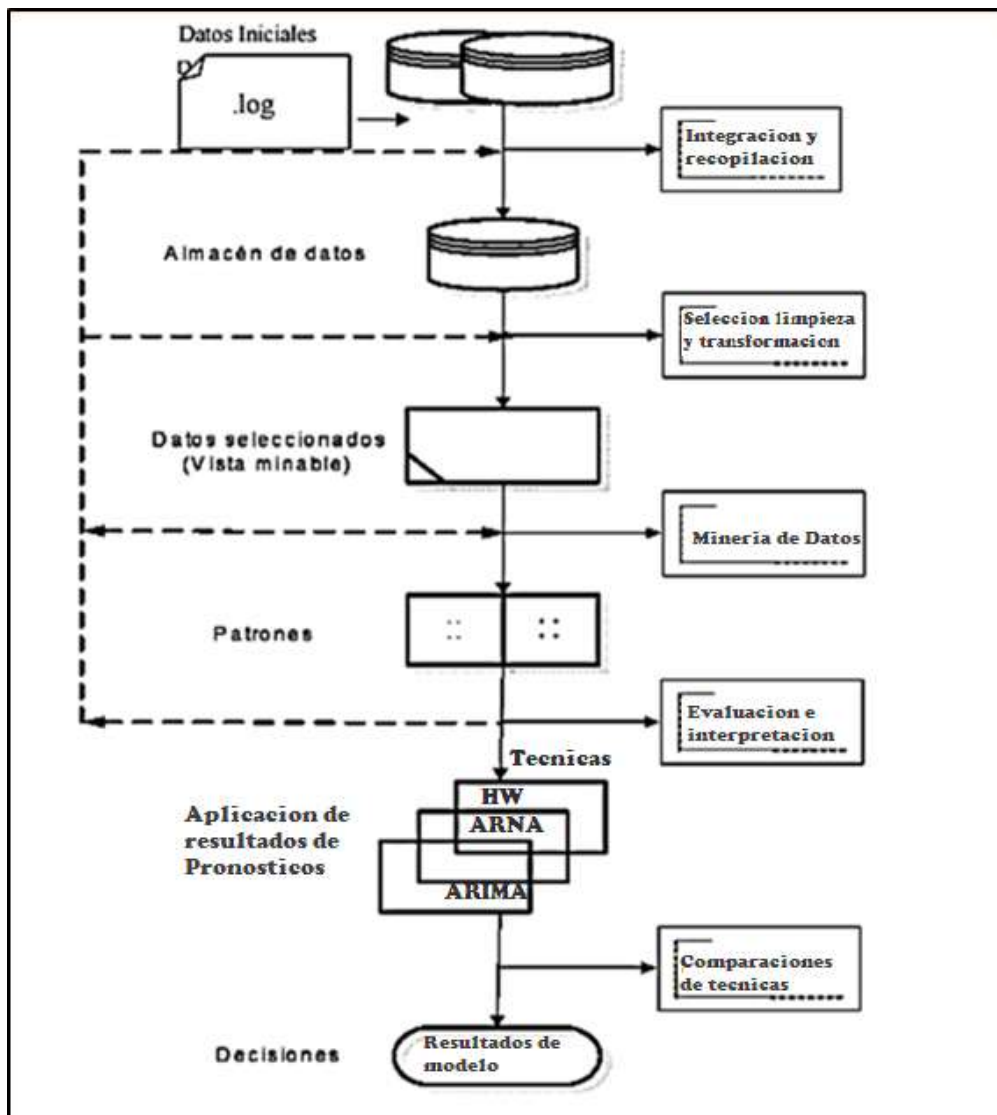
| MODELO DE MINERÍA DE DATOS PARA LA PREDICCIÓN | | | | |
|---|-------------|----------|------------------|-----|
| | HOLTWINTERS | ARIMA | REDES NEURONALES | SVM |
| Evaluación fundamento teórico | | | | |
| Modelo parametrizado | X | X | | |
| Datos estacionales | X | X | X | |
| Método estadístico | X | X | | |
| Capacidad iterativa (Aprendizaje) | | | X | |
| Cantidad de datos de la serie | 3 | 3 | 2 | |

Fuente: Elaboración propia



Descripción. En la evaluación realizada se selecciona las técnicas de HoltWinter, Arrima y Red Neuronal Auto regresiva, quedando descartado la técnica de redes Bayesianas y árbol de decisiones, SVM

Figura 36: Modelo de procesos



Fuente: Elaboración propia

Modelo en la cual las técnicas, 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 ya propuestos.

Figura N° 37: Construcción de los Modelos en R

```

model.r
#CONEXION A BASE DE DATOS
library("RODBC")
bd <- odbcDriverConnect('driver={SQL Server};server=localhost;database=BIEspa;trusted_connection=true')
#del1<-sqlQuery(bd, "delete from Evaluacion")
#del2<-sqlQuery(bd, "delete from DetalleEvaluacion")
#CONSULTA PARA EXTRAER DATOS
vec<-sqlQuery(bd, "select total from (
select cast(mes as varchar(2))+'-'+cast(anio as CHAR(4)) as periodo,mes,anio,total from (
select MONTH(fe cham) as mes,year(fe cham) as anio, sum(pre) as total from (
select convert(date, fechaentrega, 103) as fe cham,CAST(precio AS DECIMAL(10, 2)) as pre from HistCompra) as q
group by MONTH(fe cham),year(fe cham)) as fin) as mining
order by anio,mes")
vec<-as.vector(t(vec))
vec <- as.numeric(vec)
real <-vec[length(vec)]
tam <-length(vec)-1
vec <-vec[1:tam]
#LIBRERIA FORECAST CONTIENE ALGORITMOS PARA PRONOSTICOS
library("forecast")
fre<- ts(vec, frequency=12,start=c(2012,09))
#Arima
Arima
fit <- Arima(fre,c(3,1,0))
fit <- forecast.Arima(fit, h=1)
plot(forecast(fit))
#Holtwinters
fit <- Holtwinters(fre, alpha=0.3, beta=0.1, gamma=0.1)
fit <- Holtwinters(fre)
pro1 <- forecast.Holtwinters(fit, h=1)
texto <- paste("c:/program Files (x86)/Zend/Apache24/htdocs/datamining/plot/pronosticof.jpg",sep="")
imagen<-jpeg(texto, width= 800, height=400)
plot(pro1,main="Pronostico Establecido")
dev.off()
#red neuronal
fit <- nnetar(fre, decay=0.6, maxit=100)
pro2 <- forecast.nnetar(fit, h=1)
pro2 <- as.numeric(pro2$mean[1])
log2<- system.time( replicate(10, nnetar(fre)))
cpu2<- as.numeric(log2[1])
t2<- as.numeric(log2[3])
#plot(forecast(fit,h=1))
#lines(fre)
#GRABA RESULTADOS EN TABLA DETALLE EVALUACION
imprime<- paste("insert into DetalleEvaluacion (IdEvaluacion,Periodo,ValorReal,ValAlgoritmo1,ValAlgoritmo2,TiempoSeg1,TiempoSeg2,CPUseg1,CPU2
inserta <- sqlQuery(bd,imprime)

```

Fuente: Elaboración propia

Fuente: Elaboración propia

Se muestra la codificación del modelo en para el análisis de cada uno de los algoritmos



Figura N° 39: Librería Forecast código abierto en Github

```

robjhyndman Made y the standard first argument for modelling functions.
4 contributors
424 lines (186 sloc) 12.8 KB
1 # Modelled on the HoltWinters() function but with more conventions
2 # Initialization
3 # Written by George Zimis, 23 October 2012
4
5 HoltWintersZZ <- function(x,
6
7     # smoothing parameters
8     alpha = NULL, # level
9     beta = NULL, # trend
10    gamma = NULL, # seasonal component
11    seasonal = c("additive", "multiplicative"),
12    exponential = FALSE, # exponential
13    phi = NULL, # damp
14    lambda = NULL, # box-cox
15    biasadj = FALSE # adjusted back-transformed mean for box-cox
16 )
17 {
18     x <- as.TS(x)
19     seasonal <- match.arg(seasonal)
20     m <- frequency(x)
21     lam <- length(x)
22
23     if(!is.null(lamoda)){
24         x <- BoxCox(x, lamoda)
25     }
26
27     if(is.null(phi) || !is.numeric(phi))
28         phi <- 0
29     if(!is.null(alpha) && !is.numeric(alpha))
30         stop("cannot fit models without level ('alpha' must not be 0 or FALSE).")
31     if(!all(is.null(c(alpha, beta, gamma))) &&
32         any(c(alpha, beta, gamma) < 0 || c(alpha, beta, gamma) > 1))
33         stop("'alpha', 'beta' and 'gamma' must be within the unit interval.")
34     if(is.null(gamma) || gamma > 0) {
35         if (seasonal == "multiplicative" && any(x <= 0))
36             stop("data must be positive for multiplicative Holt-Winters.")
37     }
38 }

```

Fuente: Elaboración Propia

Los pasos para aplicar este método son:

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



TECNICA - A

a. Descripción de la técnica Holt-winters A

El método de “Holt-Winter” realiza un pronóstico de la producción de espárragos como una serie temporal, a partir de la base de datos de la empresa agro exportadora de Jayanca

Este método está basado en un algoritmo iterativo que a cada tiempo (mes), realiza un pronóstico sobre el comportamiento de la serie en base a promedios ponderados de los datos anteriores.

El algoritmo tiene tres parámetros, cada uno de ellos asociado a diferentes componentes de la serie.

El valor de estos parámetros se ajusta, comparando la serie real con la pronosticada para ese mismo lapso. Una vez realizado los ajustes se procede a hacer el pronóstico para el periodo en donde no hay datos.

Cada una de estas componentes estacionales está asociada a un parámetro, general llamado α , β , γ . Además estos valores pueden estar fijados o escogerse de manera que minimicen el error cuadrático medio comparando el **comportamiento de la serie real** y de la serie pronosticada en la zona en la que se superponen.

El modelo de Holt-Winters se aplica, en un período t , que según la producción de espárragos que esta representados por 2 periodos al año. según nuestro trabajo de investigación.

El método de Holt-Winter se aplica en esquema aditivo.

Donde se describe paso a paso la representación de la formula aplicada para este modelo.

- y_t Producciones registradas en el periodo t
- $\hat{y}_{t+k/t}$ previsión de producción para el periodo $t + k$ basada en datos hasta t
- L_t nivel medio desestacionalizado de la serie en el periodo t
- T_t tendencia de la serie en el periodo t , es decir, incremento o decremento del nivel medio desestacionalizado durante un periodo
- S_t componente estacional en el periodo t
- L Longitud de la estacionalidad
- p numero de periodos a pronosticar a futuro

Cuando se dispone de una nueva observación los tres términos que intervienen (L_t, T_t y S_t) se actualizan de forma iterativa mediante alisado exponencial. Las ecuaciones de actualización son las siguientes:

Aplicando la fórmula matemática, se toma como referencia las cantidades de producción real de espárragos por meses que comprende de enero a diciembre de los años 2013, 2014 y 2015 para pronosticar los 4 meses siguientes del año.

La ecuación (E1) proporciona un valor para el nivel medio en el momento t .

y_t corregido de estacionalidad y combinándolo con la suma entre el nivel medio y el incremento (o decremento) esperados para el mes inmediatamente anterior.

Donde se

$$L_t = \alpha \frac{y_t}{S_{t-L}} + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad \mathbf{E1}$$

En la ecuación (E2) aproxima el valor de la tendencia en t tomando por un lado la diferencia entre los niveles medios en y_t y_{t-1} y, por otro, el valor de la tendencia en el periodo anterior.

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad \mathbf{E2}$$



En la ecuación (E3) combina un acercamiento al efecto estacional en el mes t , que se consigue dividiendo el valor de la serie por la estimación del nivel medio, con el valor del factor estacional del mismo periodo del año anterior.

Igualmente es optar por los valores de los parámetros seleccionados por el programa, en general, elegidos como los que minimizan los errores cuadráticos medios. Por lo visto en la librería

$$S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-L} \tag{E3}$$

En la ecuación (E4) y_{t+p} es el valor del t , sumando los valores de L_t más el valor T_t que multiplica S_{t-1+p}

$$y_{t+p} = (L_t + pT)S_{t-1+p}$$

Donde $0 < \alpha, \beta, \gamma < 1$

Tabla N° 26: representación de la técnica Holt-winters

| MESES | t | Yt | Lt | Tt | St | | Yt-1' | Error |
|-------|----|-------------|------------|-------------|------------|---------|------------|-------------|
| | -2 | | | | 1 | St-st-1 | | |
| | -1 | | | | 1 | | | |
| | 0 | | | | 1 | | | |
| ENE | 1 | 6137697 | 6137697 | 0 | 1 | | | |
| FEB | 2 | 8296353 | 6785293.8 | 64759.68 | 1.02226962 | | 6137697 | 2158656 |
| MAR | 3 | 7617369 | 7080248.14 | 87779.1456 | 1.00758619 | | 6850053.48 | 767315.52 |
| ABR | 4 | 17459844 | 10255572.3 | 396533.647 | 1.07024739 | | 7168027.28 | 10291816.72 |
| MAY | 5 | 540005 | 7618475.66 | 93170.6188 | 0.9070881 | | 10652105.9 | 10112100.94 |
| JUN | 6 | 6718050 | 7369662.61 | 58972.2519 | 1.01120083 | | 7883381.72 | 1165331.718 |
| JUL | 7 | 7312377 | 7377240.88 | 53832.8541 | 1.00594833 | | 7484989.88 | 172612.8762 |
| AGO | 8 | 1671212 | 5670207.42 | -122253.778 | 0.99269621 | | 7953087.26 | 6281875.262 |
| SEP | 9 | 13383110 | 8309745.13 | 153925.371 | 0.97743248 | | 5032482.72 | 8350627.284 |
| OCT | 10 | 8712670 | 8509417.89 | 158500.11 | 1.01246931 | | 8558470.67 | 154199.3279 |
| NOV | 11 | 5840357 | 7809289.22 | 72637.2323 | 0.9801408 | | 8719477.59 | 2879120.593 |
| DIC | 12 | 12132460 | 9183865.99 | 202831.186 | 1.02553283 | | 7824358.52 | 4308101.481 |
| ENE | 13 | | | | | | 9174862.73 | |
| FEB | 14 | | | | | | 9298382.43 | |
| MAR | 15 | | | | | | 9001481.75 | |
| ABR | 16 | PRONOSTICOS | | | | | 9418356.07 | |
| | | | | | | | | 4240159.793 |

L 4
 α 0.3
 β 0.1
 γ 0.1



Fuente: Elaboración propia

En la tabla N° 26 se aplica una representación aplicando cada una de las fórmulas para obtener un resultado que conforma el **Alpha Beta y Gama** toda la fórmula, así como también el y_{t+p} .

En Aplicación de la formula en código fuente se realiza para la técnica HOLTWINTERS

Considerando Alpha = 0.3, beta = 0.1, gamma = 0.1

1. Una visualización se entrena la serie con holtwinters

Tabla N° 27: Entrenando serie de tiempo

| | |
|---------|--|
| ENTRADA | fit <- HoltWinters(fre, alpha=0.3, beta=0.1, gamma=0.1) |
| SALIDA | <pre> > fit <- HoltWinters(fre, alpha=0.3, beta=0.1, gamma=0.1) > fit Holt-Winters exponential smoothing with trend and additive seasonal component. Call: HoltWinters(x = fre, alpha = 0.3, beta = 0.1, gamma = 0.1) Smoothing parameters: alpha: 0.3 beta : 0.1 gamma: 0.1 Coefficients: [,1] a 23062674.4 b 1023501.2 s1 -6137697.7 s2 -8296353.0 s3 -7617369.8 s4 -12452884.0 s5 340085.2 s6 6718050.6 s7 -7312377.3 s8 -1671212.6 s9 13883110.1 s10 8712670.3 s11 8840857.8 s12 12192460.6 > </pre> |

Fuente: Elaboración Propia



2. Se está verificando entrenamiento.

Tabla N° 28: Entrenamiento

| | |
|----------|---|
| ENTRAD A | fit <- HoltWinters(fre, alpha=0.3, beta=0.1, gamma=0.1) |
| SALIDA | <pre>> fit\$fitted xhat level trend season Sep 2013 40845772 51164517 -616698.9 -9702046 Oct 2013 46848884 49048078 -766672.9 -1432522 Nov 2013 52314569 47382550 -856558.4 5788577 Dec 2013 35213676 45881463 -921011.2 -9746776 Jan 2014 41067680 44537081 -963348.3 -2506053 Feb 2014 56711334 43653299 -955391.7 14013427 Mar 2014 51215140 43165453 -908637.1 8958324 Apr 2014 46777314 42348820 -899436.7 5327931 May 2014 56655557 43860911 -658284.0 13452930 Jun 2014 31010776 39060270 -1072519.7 -6976974 Jul 2014 31715774 41584649 -712829.8 -9156046 Aug 2014 36191056 44556218 -344390.0 -8020772 Sep 2014 35717204 45940693 -171503.4 -10051985 Oct 2014 32604097 35449794 -1203443.0 -1642255 Nov 2014 48968892 43598894 -268188.7 5638187 Dec 2014 38307742 47958694 194610.2 -9845562 Jan 2015 57802585 59009812 1280260.9 -2487488 Feb 2015 79541009 63788395 1630093.2 14122521 Mar 2015 72542580 62249586 1313202.9 8979792 Apr 2015 69507325 62417982 1198722.3 5890621 May 2015 77064852 63401291 1177180.9 12486380 > </pre> |

Tabla: Elaboración Propia

3. Se realiza una verificación de residuales

Tabla N° 29: Verificando residuales

| | |
|----------|---|
| ENTRAD A | fit[2] |
| SALIDA | <pre>> fit[2] \$X Jan Feb Mar Apr May Jun Jul Aug 2012 2013 71785500 60576120 56979180 55681520 63765720 41962760 38081080 37851400 2014 41332900 58269820 51521820 54815740 42847700 43000440 43997100 41953940 2015 69463660 68978000 68726560 68789280 72008860 Sep Oct Nov Dec 2012 1425200 22288540 72709480 44212440 2013 35846640 43852700 50166140 33802440 2014 1319220 63779240 64395520 74496100 2015</pre> |



Fuente: Elaboración Propia

4. Realiza la verificación de coeficientes

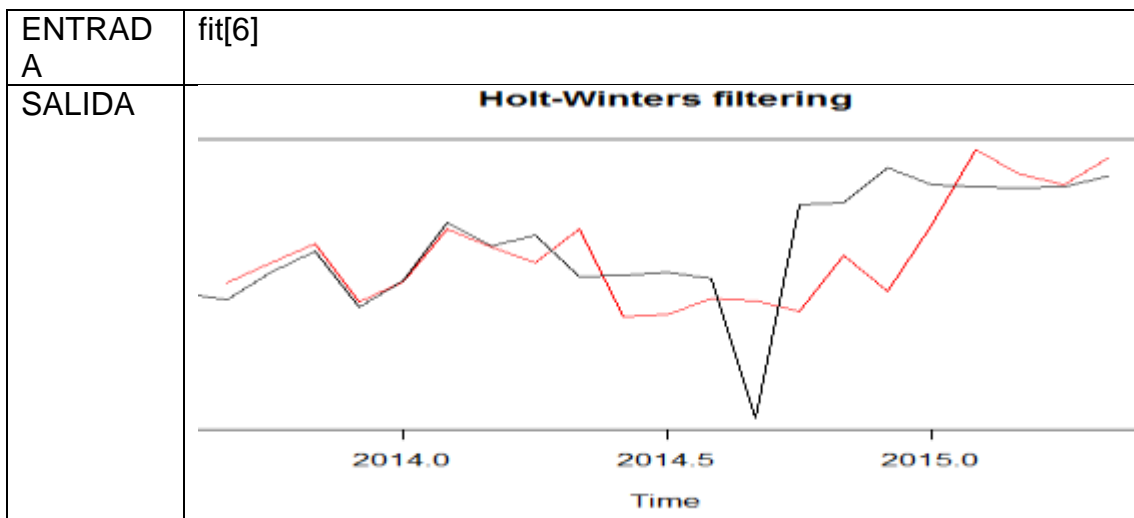
Tabla N° 30: Verificación de Coeficientes

| | |
|----------|---|
| ENTRAD A | fit[6] |
| SALIDA | <pre>> fit[6] \$coefficients a b s1 s2 s3 s4 63061674.4 1025501.2 -6137697.7 -8296353.0 -7617369.8 -12459844.0 s5 s6 s7 s8 s9 s10 540005.5 6718050.6 -7312377.3 -1671212.6 13383110.1 8712670.3 s11 s12 5840357.5 12132460.6</pre> |

Fuente: Elaboración Propia

5. Resultado, verificando ploteando entrenamiento

Tabla N° 31: Entrenamiento de modelo



Fuente: Elaboración Propia

6. Se realiza la verificación de conteniendo pronostico

Tabla N° 32: Resultado de pronostico

| | |
|----------|--|
| ENTRAD A | pro1 <- forecast.HoltWinters(fit, h=1) |
|----------|--|



| | |
|--------|--|
| SALIDA | > <code>pro1</code> |
| | <pre> Point Forecast Lo 80 Hi 80 Lo 95 Hi 95 Jun 2015 57949478 38780888 77118068 28633649 87265307 </pre> |
| | > |

Fuente: Elaboración propia

TECNICA - B

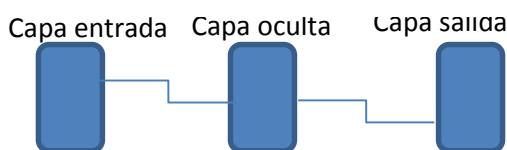
b. Evaluación de la técnica Red Neuronal autorregresiva (ARNA) B

Su diseño y arquitectura computacional está basado profundamente en la librería paquete ‘forecast’. El paquete utilizado es el “*nnetar*”

La arquitectura

Esta red neuronal puede considerarse como una red de “neuronas” que se organizan en **3 capas**.

1 capa de los predictores (o entradas) 1 capa oculta, y 1 capa de los pronósticos (salidas) mientras más capas ocultas más preciso puede ser el algoritmo



En representación real

$VR - 1$

| | | | | | |
|----|----|----|----|----|----|
| M1 | VR | V1 | V2 | V3 | V4 |
| M2 | | ? | ? | ? | ? |
| M3 | | | ? | ? | ? |
| M4 | | | | ? | ? |
| M5 | | | | | ? |
| M6 | | | | | |

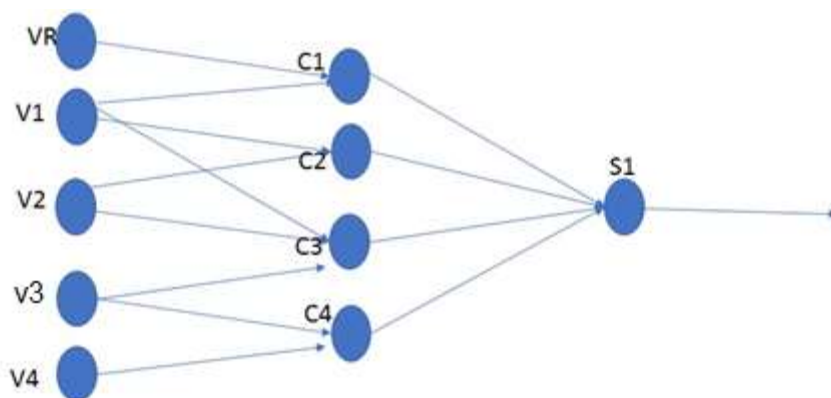
VR. Es el valor real, V1. Es el valor 1, V2. Es el valor 2, V3. Es el valor 3, V4. Es el valor 4, cada uno de ellos representa un valor, M1, M2, M3, M4, M5, Representan a los meses

Por lo tanto, está definido que la red neuronal que se aplica es la autorregresiva con 3 capas.



- En la capa de entrada cuenta con 5 neuronas de entrada incluyendo la serie original que representa a los meses que se va pronosticar en este caso se pronosticara los 4 primeros meses del año
- En la capa oculta cuenta con 4 neuronas
- En la capa de salida cuenta con una neurona que representa al pronostico

Figura Ni 40: Representación de la arquitectura capas de la red



Fuente: Elaboración propia

conjunto de entradas $x_j(t)$ producción de espárragos en el periodo.

Unos pesos sinápticos w_{ij} asociados a las entradas (frecuencia el valor de producción acumulado que ingresa genera un valor más próximo en cuanto al valor de pronóstico resultante).

Una regla de propagación $h_i(t) = \sigma(w_{ij}, x_j(t))$. La más común suele ser $h_i(t) = \sum w_{ij}x_j$.

Una **función de activación** $y_i(t) = f_i(h_i(t))$ que representa simultáneamente la salida de la neurona y su estado de activación.

Con frecuencia se añade al conjunto de pesos de la neurona un parámetro adicional, θ_i , que denominaremos umbral, que se resta del potencial postsináptico, por lo que el argumento de la función de activación queda

$$\sum_j w_{ij}x_j - \theta_i$$

De forma equivalente, si hacemos que los índices i y j comiencen por 0 y definiendo $w_{i0} = \theta_i$ y $x_0 = -1$ (constante) podemos obtener el comportamiento de la neurona a través de:

$$y_i(t) = f_i \left(\sum_{j=0}^n w_{ij}x_j \right)$$

La función de activación que tendría la neurona es:

$$y_i = f \left(\sum w_{ij}X_j - \theta_i \right) = \sum w_{ij}X_j - \theta_i$$

Por lo tanto se denota en R con la siguiente formula

$$Nnttar(y, p, P =, Size, repearts = ?, decay = 0.5)$$

Y = serie de tiempo (free), P = tamaño de vueltas, p = desfase de entre observaciones de la ventana, **Size** = tamaño de la capa oculta, **repats**= repartición de la red neuronal, **decay** limitante para los pesos generados para la interacción de la red.

1. Se realiza el entrenamiento de la serie con red neuronal.

Tabla Ni 33: Entrenamiento de modelo

| | |
|---------|---|
| ENTRADA | fit <-nnetar(fre, 5, P = 1, 4, repeats = 500, decay =0.5) |
| SALIDA | <pre> > library("forecast") Loading required package: zoo Attaching package: 'zoo' The following objects are masked from 'package:base': as.Date, as.Date.numeric Loading required package: timeDate This is forecast 7.1 > fit <-nnetar(fre, 5, P = 1, 4, repeats = 500, decay =.5) > fit Series: fre Model: NNAR(5,1,4) [12] Call: nnetar(x = fre, p = 5, P = 1, size = 4, repeats = 500, decay = 0.5) Average of 500 networks, each of which is a 6-4-1 network with 33 weights options were - linear output units decay=0.5 sigma^2 estimated as 1.765e+14 > </pre> |

Fuente: Elaboración Propia



2. Se está verificando entrenamiento

Tabla N° 34: Verificación de entrenamiento

| | |
|----------|---|
| ENTRAD A | Fit\$fitted |
| SALIDA | <pre>> fit\$fitted Jan Feb Mar Apr May Jun Jul Aug 2012 2013 NA NA NA NA NA NA NA NA 2014 44533375 46714761 57200049 52270175 55703581 45602350 45307489 45735575 2015 64575480 63954936 63057580 63268302 62040516 Sep Oct Nov Dec 2012 NA NA NA NA 2013 41178167 41852494 49460227 49866759 2014 44667760 39802142 59968871 58275649 2015 > </pre> |

Fuente: Elaboración propia

3. Se está verificando residuales

Tabla N° 35: Verificación de residuales

| | |
|----------|---|
| ENTRAD A | Fit\$residuals |
| SALIDA | <pre>> fit\$residuals Jan Feb Mar Apr May Jun 2012 2013 NA NA NA NA NA NA 2014 -3200474.9 11555058.9 -5678229.1 2545565.0 -12855881.0 -2601910.0 2015 4888180.5 5023063.8 5668979.9 5520978.5 9968343.9 Jul Aug Sep Oct Nov Dec 2012 NA NA NA NA NA NA 2013 NA NA -5331526.9 2000205.9 705913.3 -16064319.0 2014 -1310388.5 -3781635.2 -43348540.0 23977097.6 4426649.5 16220450.0 2015 > </pre> |

Fuente: Elaboración Propia

4. Verificando tipo de red neuronal y capa oculta.

Tabla N° 36: Red Neuronal capas ocultas

| | |
|---------|--------|
| ENTRADA | fit[8] |
|---------|--------|

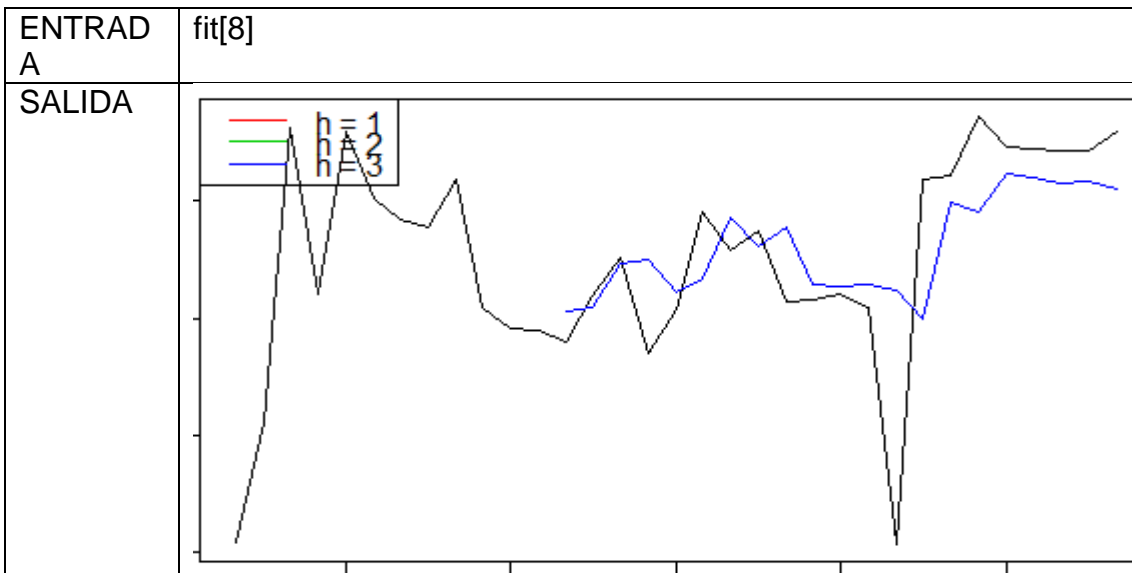


| | |
|--------|---|
| SALIDA | <pre>> fit[8] \$model Average of 500 networks, each of which is a 6-4-1 network with 33 weights options were - linear output units decay=0.5 > </pre> |
|--------|---|

Fuente: Elaboración Propia

5. Verificando plot entrenamiento

Tabla N° 37: Plot Entrenamiento



Fuente: Elaboración Propia

6. Pronosticando con modelo

Tabla N° 38: Resultado de Pronostico

| | |
|---------|---|
| ENTRADA | pro2 <- forecast.nnetar(fit, h=1) |
| SALIDA | <pre>> pro2 <- forecast.nnetar(fit, h=1) > pro2 May 2015 60959471 > </pre> |

Fuente: Elaboración Propia



TECNICA: C

c. Descripción de la técnica Arima C

Los modelos auto regresivos es un proceso estocástico es una sucesión de variables aleatorias Y_t ordenadas, pudiendo tomar t cualquier valor entre. Por ejemplo, la siguiente sucesión de variables aleatorias puede ser considerada como proceso estocástico.

El modelo ARIMA cuenta con tres parámetros principales:

Parámetro p : Asociado a la parte autorregresiva (AR) del modelo

Parámetro d : Asociado a la parte integrada (I) del modelo

Parámetro Q : Asociado a la parte del promedio móvil (MA) del modelo

El diseño ARIMA se denota de forma ARIMA (p, d, q) indicando el rezago del modelo con la que se realiza en modelo. Por ejemplo, un ARIMA (3,1,0) (o equivalente, AR (1)) está dado por la ecuación:

$$Z_t = \phi_1 Y_{t-1} + A_t$$

A_t Es un ruido blanco aleatorio

Z_{t-1} Es la observación pasada de la variable Z_t

La forma general de un modelo autorregresivo de promedios móviles estacionales (ARMA). Combina los procesos autorregresivos (AR(p)) y de promedios móviles (MA(q)) y se le acompaña en su definición con los órdenes correspondientes (p, q).

$$\text{Proceso ARIMA (3,1,0)} = Z_t = \phi_1 Z_{t-1} + a_t - \theta_1 a_{t-1}$$



en el cuál, el valor actual de la serie, a_t , (error aleatorio), puede explicarse en función de p valores pasados Z_{t-1}, a_t, a_{t-p}

$$Z_t = \Phi_1 Y_{t-12} + a_t$$

$\Phi_1 Z_{t-12}$ Es primer rezago estacional de 12 meses de la variable

$$\text{Proceso ARIMA } (3, 1, 0)_{12}: Z_t = \Phi_1 Z_{t-12} + a_t - \theta_1 a_{t-12}$$

Cuenta con una parte estacional dentro del modelo se tiene que verificar que los datos que se va a trabajar cuenten con datos estacionales. En este caso se esta utilizando en los parámetros $pq = 3, 1, 0$ en la cual es es un AR 3

Entrenando serie con la técnica Arima

Tabla N° 39: Serie de Tiempo con Arima

| | |
|---------|--|
| ENTRADA | fit <- Arima(fre,c(3,1,0)) |
| SALIDA | <pre>> fit Series: fre ARIMA(3,1,0) Coefficients: ar1 ar2 ar3 -0.4158 -0.1275 -0.0257 s.e. 0.1930 0.2355 0.2088 sigma^2 estimated as 3.51e+14: log likelihood=-579.79 AIC=1167.58 AICc=1169.06 BIC=1173.44 > </pre> |

Fuente: Elaboración Propia

1. Verificando procedimiento Arima – entrenamiento

Tabla N° 40: Verificación de procedimiento Arima

| | |
|---------|---|
| ENTRADA | fit <- Arima(fre,c(3,1,0)) |
| SALIDA | <pre>> fitted(fit) Jan Feb Mar Apr May Jun Jul Aug 2012 2013 49097434 62658851 62453210 59196152 56967586 60661932 50031492 42267643 2014 39596283 40125934 50687017 51974853 53871571 47577578 44378365 43870498 2015 68613632 70252546 69562267 69022261 68807731 Sep Oct Nov Dec 2012 1423775 3041895 14994851 49453755 2013 39001696 36809227 40785103 46571498 2014 42672533 18451067 43040526 57218516 2015 > </pre> |

Fuente: Elaboración Propia

2. Verificando procedimiento arima – residuales



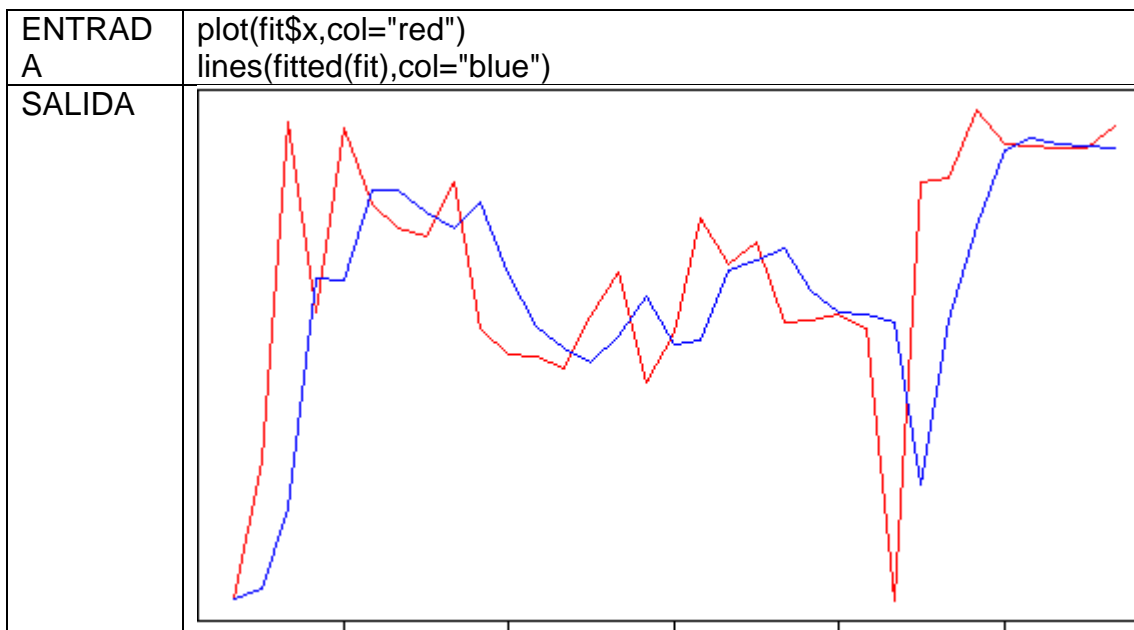
Tabla N° 41: Verificación de Arima

| | |
|---------|--|
| ENTRADA | fit[8] |
| SALIDA | <pre>> fit[8] \$residuals Jan Feb Mar Apr M 2012 22688066.392 -2082731.099 -5474030.376 -3514632.477 6798134.3 2013 1736617.460 18143885.790 834803.090 2040887.038 -11023871.4 2014 850027.823 -1274545.967 -835707.042 -232980.821 3201129.4 Jun Jul Aug Sep O 2012 1425.199 19246645.3 2013 -18699172.054 -11950411.762 -4416243.154 -3155056.286 7043473.4 2014 -4577138.024 -381265.361 -1916557.864 -41353312.532 45328172.5 2015 Nov Dec 2012 57714629.284 -5241315.474 2013 9381037.454 -12769058.075 2014 21354993.759 17277584.303 2015</pre> |

Fuente: Elaboración Propia

3. Verificando procedimiento arima – ploteando entrenamiento

Tabla N° 42: Verificación de grafico de Arima - Ploteo



Fuente: elaboración Propia

4. Se está verificando procedimiento arima – obteniendo horizonte

Tabla N° 43: Resultado de pronóstico

| | |
|---------|----------------------------------|
| ENTRADA | pro3 <- forecast.Arima(fit, h=1) |
| A | |



| | |
|--------|---|
| SALIDA | <pre>> pro3 Point Forecast Lo 80 Hi 80 Lo 95 Hi 95 Jun 2015 70668538 46658666 94678410 33948607 10738846 > </pre> |
|--------|---|

Fuente: Elaboración Propia

C. Este objetivo es realizar el desarrollo del modelo para la predicción de la producción de espárragos.

ii. Evaluación

5.4.3.1. Evaluar los resultados

Objetivos – Criterios de Evaluación del Negocio

a) Analizar tendencias de producción:

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

b) Realizar pronósticos de producción.

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

c) Confianza del Modelo:

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

d) Objetivos - Criterios de Evaluación del Proyecto

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

El modelo permite entrenar la data para obtener un porcentaje de confianza del mismo.



Figura N° 41: ALGORITMO PARA ANALISIS Y COMPARACION DE RESULTADOS ENTRE TECNICAS – VER ANEXO N° 3: PLAN DE PRUEBAS

```

#PASO 1 - INSTANCIAR DE LIBRERIAS
library(RDBC)
library(forecast)
options(nav.print = 99999999)

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

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

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

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

ar <- Historico[1]
ar2<-as.vector(t(ar))
vaniel<-as.numeric(ar2[1])

br <- Historico[2]
br2<-as.vector(t(br))
vmes<-as.numeric(br2[1])

#PASO 6 - DECLARA PESES INICIA ATRAS A PARTIR DEL ULTIMO MES HISTORICO REGISTRADO EN EL VECTOR
chestest <- 0
    
```

Fuente: elaboración propia

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

Figura N° 42: Resultado de los algoritmos aplicados en los el modelado

| IdDe. | MEvaluacion | Periodo | ValorReal | ValAlgoritmo1 | ValAlgoritmo2 | ValAlgoritmo3 | TiempoSeq1 | TiempoSeq2 | TiempoSeq3 | CPUSeq1 | CPUSeq2 | CPUSeq3 |
|-------|-------------|---------|-----------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| 171 | 2 | 9-2014 | 1319220 | 34382115.96257... | 46888732.80504... | 14123796.36528... | 0.0699999999999... | 0.1899999999999... | 0.0300000000000... | 0.0700000000000... | 0.1700000000000... | 0.0399999999999... |
| 172 | 2 | 10-2014 | 6379240 | 44038048.81590... | 40132884.08704... | 3253298.30803... | 0.0500000000000... | 0.1899999999999... | 0.0500000000000... | 0.0500000000000... | 0.1899999999999... | 0.0500000000000... |
| 173 | 2 | 11-2014 | 64395520 | 30997970.06595... | 29972820.90156... | 36136674.61144... | 0.0600000000000... | 0.1999999999999... | 0.0299999999999... | 0.0599999999999... | 0.1999999999999... | 0.0299999999999... |
| 174 | 2 | 12-2014 | 74496100 | 44479126.15977... | 77931397.00426... | 67940767.726854 | 0.0600000000000... | 0.2000000000000... | 0.0300000000000... | 0.0599999999999... | 0.1899999999999... | 0.0300000000000... |
| 175 | 2 | 1-2015 | 69463660 | 66668568.84418... | 68423581.65873... | 70279731.389267 | 0.0600000000000... | 0.2000000000000... | 0.0199999999999... | 0.0600000000000... | 0.18 | 0.0200000000000... |
| 176 | 2 | 2-2015 | 68978000 | 64827521.34338... | 68888864.38623... | 68567911.686827 | 0.0900000000000... | 0.2099999999999... | 0.0300000000000... | 0.0799999999999... | 0.2000000000000... | 0.0300000000000... |
| 177 | 2 | 3-2015 | 68726560 | 67257766.24484... | 68241792.39122... | 69020027.413072 | 0.0600000000000... | 0.2199999999999... | 0.0300000000000... | 0.0600000000000... | 0.2200000000000... | 0.0300000000000... |
| 178 | 2 | 4-2015 | 68789280 | 73500212.377518 | 70895415.46148... | 68807426.483882 | 0.0600000000000... | 0.2300000000000... | 0.0300000000000... | 0.0600000000000... | 0.2100000000000... | 0.0300000000000... |
| 179 | 2 | 5-2015 | 72008860 | 62672572.88014... | 72389333.02206... | 70668538.162124 | 0.0600000000000... | 0.2400000000000... | 0.0299999999999... | 0.0600000000000... | 0.2400000000000... | 0.0300000000000... |

Fuente: Elaboración propia



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

A. Fase I: Gestión de Proyecto

a.1 Planificación del Proyecto

Tabla N° 44: Prioridad y Dificultad de Historia de Usuario

| HISTORIA DE USUARIO | PRIORIDAD | N° ITERACIONES |
|--|-----------|----------------|
| 1. CONSULTAR PRONOSTICO POR PERIODOS Y ANUAL | ALTA | 3 |
| 2. GENERAR SIMULACIONES | ALTA | 3 |
| 3. GESTION DE USUARIOS. | MEDIA | 2 |
| 4. MOSTRAR RESULTADOS | MEDIA | 2 |

Fuente: Extraído de la Metodología XP

La prioridad es definida por el aspecto del sistema, es decir, la función principal en este caso está representado por las dos primeras historias de usuario, que hacen referencia al tratamiento de los datos, dejando de lado en menor grado a las siguientes historias como la gestión de usuarios o la de simuladores. Por lo que es necesario al finalizar la primera iteración que los entregables cuenten con un avance satisfactorio para el cliente ofreciendo un producto funcional.

a.2 Diario de Actividades

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



Tabla N° 45: Esquema de Diario de Actividades

| ACTIVIDADES | TIEMPO | | | | | | | | | | | | | |
|-----------------------|--------|----|------|----|----|----|------|----|----|----|----|----|----|----|
| | MES1 | | MES2 | | | | MES3 | | | | | | | |
| HISTORIA DE USUARIO 1 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| HISTORIA DE USUARIO 2 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | | | | |
| HISTORIA DE USUARIO 3 | | | | | | | 11 | | | | 14 | | | |
| HISTORIA DE USUARIO 4 | | | | | | | 11 | | | | 14 | | | |

Fuente: Extraído de la Metodología XP

a.3 Historia de usuario detallado

Tabla N° 46: Requerimiento 01

| Historia de Usuario | |
|--|------------------------------|
| Número: 1 | Usuario: Ingresar al Sistema |
| Nombre historia: CONSULTAR SIMULAR POR PERIODOS Y ANUAL | |
| Prioridad en negocio: Alta | Riesgo en desarrollo: Baja |
| Entrevistado: | |
| Descripción: El Jefe de departamento podrá acceder al módulo simulaciones de la información anual. | |
| Observaciones: | |

Fuente: Elaboración Propia

Tabla N° 47: Requerimiento 02 decisiones

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

Fuente: Elaboración Propia

Tabla N° 48: Requerimiento 03

| Historia de Usuario | |
|--|--|
| Número: 3 | Usuario: Administrador del aplicativo Web |
| Nombre historia: GESTIÓN DE USUARIOS | |
| Prioridad en negocio: Alta | Riesgo en desarrollo: Baja |
| Entrevistado: | |
| Descripción: El aplicativo contará con niveles de usuario: Administrador y Operarios. Cada uno de ellos tendrá restricciones al aplicativo. Administrador: Acceso a todos los módulos del aplicativo. El aplicativo debe permitir, visualizar y estructurar nuevos simulaciones. | |
| Observaciones: | |

Fuente: Elaboración Propia

Tabla N° 49: Requerimiento 04

| Historia de Usuario | |
|--|-----------------------------------|
| Número: 4 | Usuario: Gerente General |
| Nombre historia: Administración del Aplicativo | |
| Prioridad en negocio: Alta | Riesgo en desarrollo: Baja |
| Entrevistado: Gerente General | |
| Descripción: El administrador del aplicativo tiene potestad de dar de alta, edición o baja a los las simulaciones. | |
| Observaciones: | |

Fuente: Elaboración Propia

a.4 Requerimientos no funcionales

- **Facilidad de Uso:** Se pretende que el aplicativo se muestre en un entorno amigable y de fácil uso. De modo que el impacto que sufrirán los usuarios para comprender la información mostrada sea de fácil acceso.



- **Escalabilidad:** La herramienta permitirá generar nuevas simulaciones según los nuevos requerimientos.
- **Portabilidad:** La solución usa herramientas de Microsoft con el fin de adaptarse a las herramientas utilizadas en la entidad.
- **Seguridad:** Gestionar el acceso del usuario al aplicativo Web

a.5 Identificación de Stakeholders

A.5.1 Jefe de Producción, encargado del área de los procesos dentro de este. Es quien se encarga de los procedimientos para el análisis, estrategias y ver los las simulaciones de Pronostico.

A. FASE II: Diseño

d) En este objetivo es Implementar una aplicación web para mostrar resultados de los pronósticos

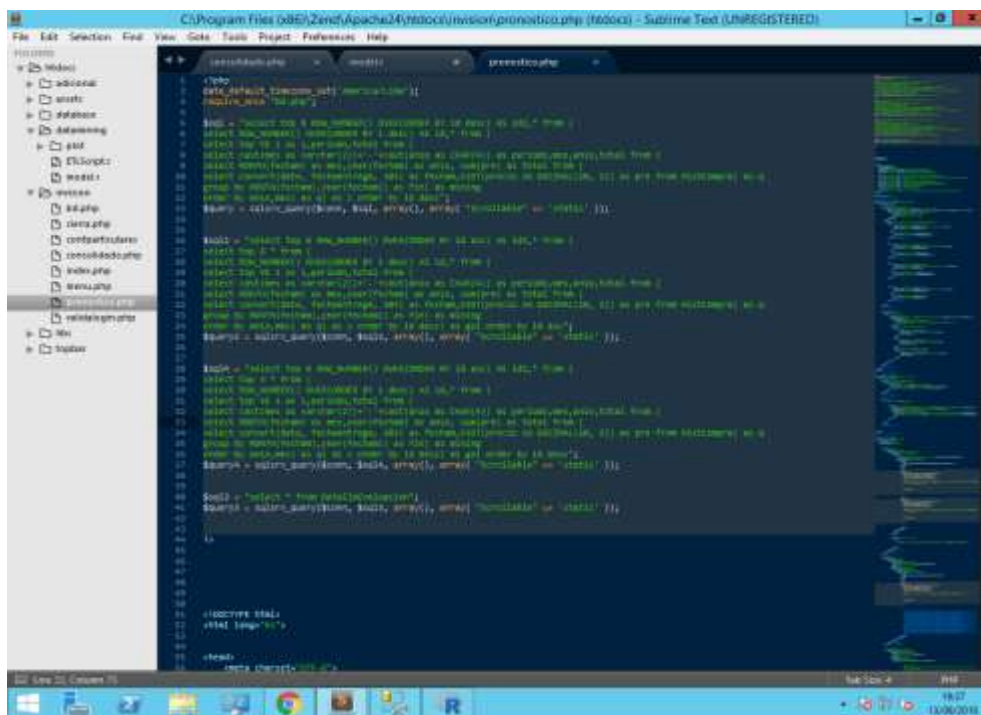
Base de datos

El aplicativo está diseñado para mostrar propósitos, la base de datos esta ara almacenar de los datos de las simulaciones y de los pronósticos realizados desde el aplicativo web, siendo este la mínima unidad representativa de tiempo registrado, por lo tanto, el aplicativo contempla esta captura de datos y el almacenamiento de información por parte de la ejecución del modelo, así como los datos de usuarios.

Implementación

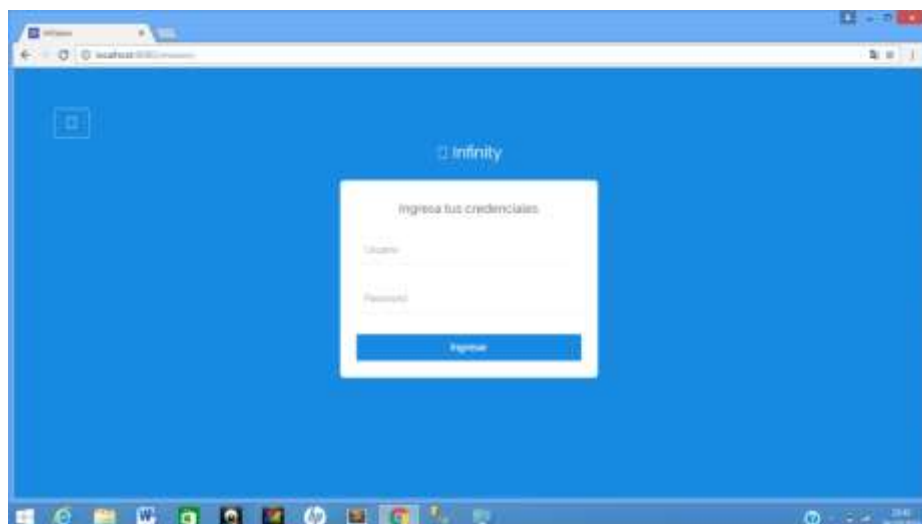
La codificación se realizó en PHP con el propósito de obtener los resultados generados por el modelo en R y que fueron impresos en la base de datos.

Figura N° 43: Script HTML Y PHP



Fuente: Elaboración Propia

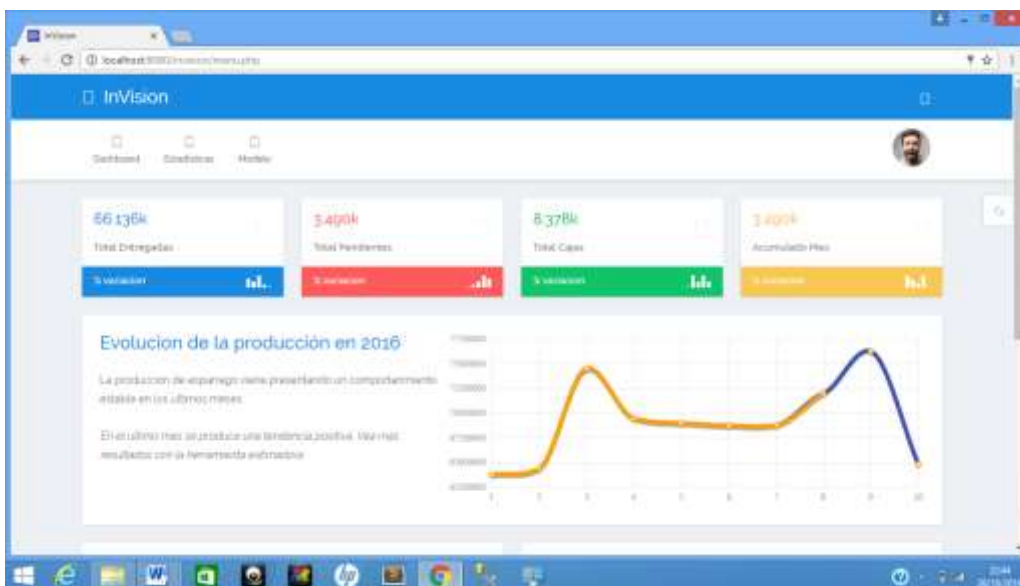
Figura N°44: Ingreso a la aplicación a través de un usuario y una contraseña



Fuente: Elaboración Propia

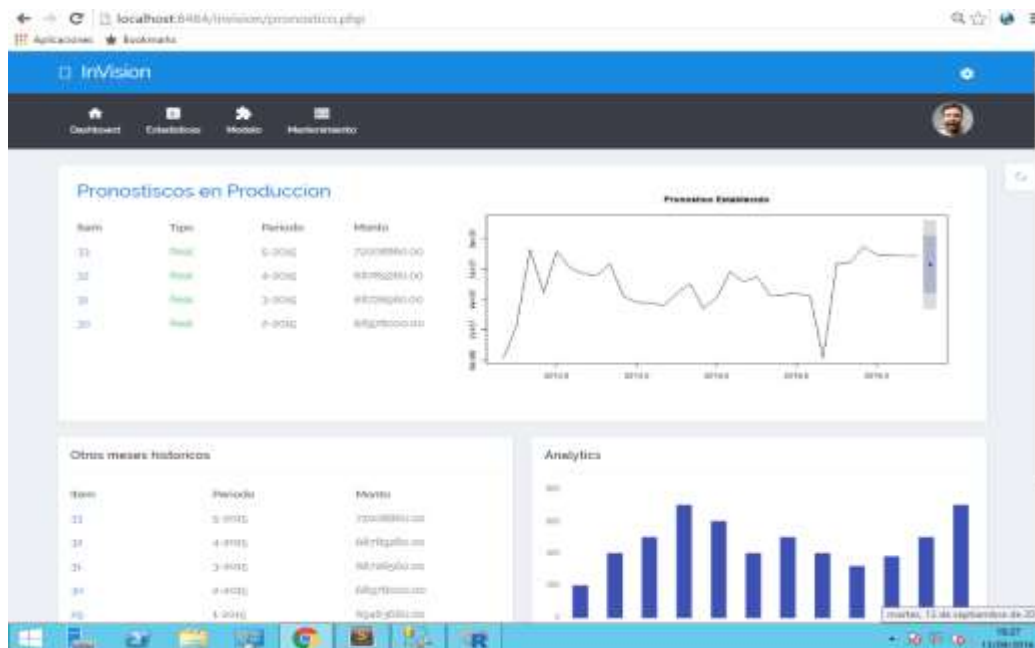


Figura N° 45: evaluaciones de producción en porcentajes



Fuente: Elaboración Propia

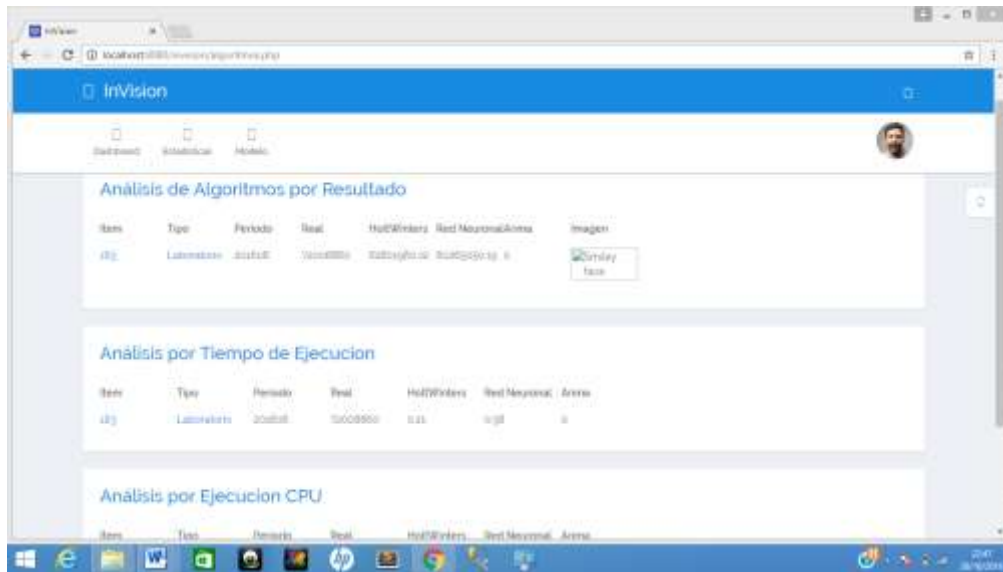
Figura N° 46: Pantalla de resultado de los modelos



Fuente: Elaboración Propia

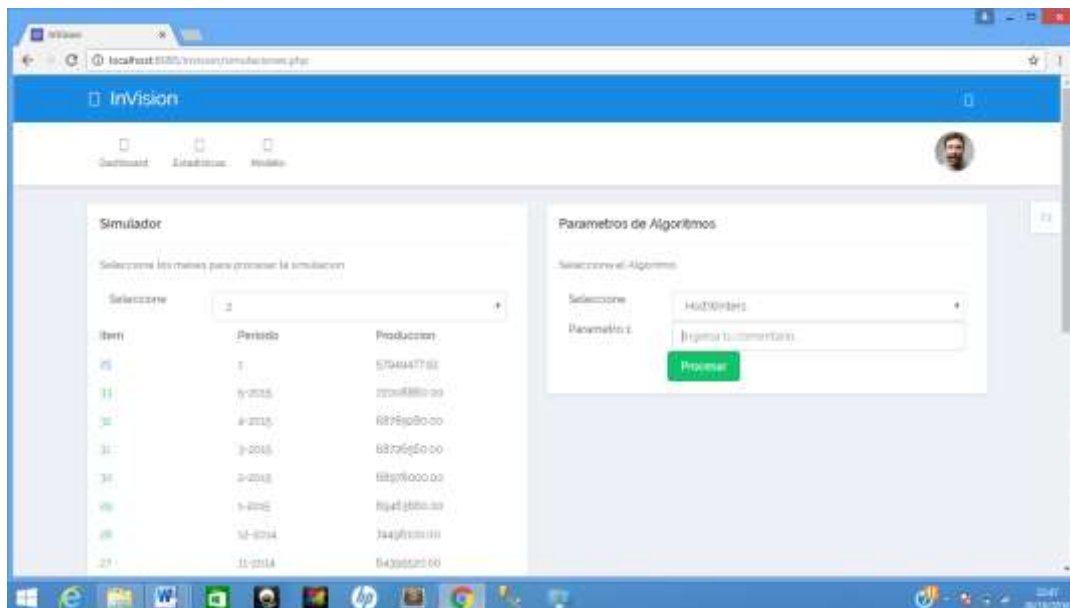


Figura N°47: Resultados del análisis de los algoritmos



Fuente: Elaboración Propia

Figura N° 48: Muestra las simulaciones por meses



Fuente: Elaboración Propia

APÍTULO VI:
CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

d. Se procedió a realizar la recopilación de información para la investigación, el cual consistió en reportes de tipo Excel que contienen el consolidado anual detallado en días, de la producción de esparrago de una empresa agro exportadora en la localidad de Jayanca la Lambayeque. Por lo tanto, se diseñó un proceso de extracción de datos para unificar la información, trasladando el 100 % de datos originales al repositorio de base de datos para su respectivo análisis.

e. En función a los datos obtenidos en el proceso de extracción, al analizar los tipos de datos, se determinó utilizar el modelado por series de tiempo, ya que en los datos recopilados se cuenta con los tipos fecha, monto, los cuales al transformarlo se convierte en un vector de montos por tiempo, por lo tanto, los algoritmos de pronósticos son los que se adaptan a este tipo de problema, para lo cual se utilizó Holtwinters (HW), Arima y RedNeuronal Autoregresiva.

ARIMA obtuvo el nivel de confianza más elevado en comparación a HoltWinters y Red Neuronal autorregresiva, esto se denota en los valores obtenidos al calcular la razón (valor calculado entre el monto real y el monto pronóstico para saber el grado de relación que existe uno con respecto del otro), obteniendo para HoltWinters unos 80.52 % de confiabilidad, contra un Red Neuronal Autoregresiva con 85.29 % y un ARIMA con un 86.86 % que lo sitúa como el mejor de los tres algoritmos.

f. Se procedió a construir el modelo para el pronóstico de la producción de espárragos, a través de un simulador, el cual se realizó para soportar los tres algoritmos de aprendizaje elegidos obteniendo el tiempo de procesamiento al evaluar estas técnicas, que con el método Arima el tiempo promedio de ejecución de 0.03 segundos, el menor de todos los tiempos comparado contra una Red Neuronal Autorregresiva de 0.21 y HW de 0.06 segundos. En el uso de CPU al evaluar estas técnicas se obtuvo que con el método Arima el uso de CPU es 0.03 segundos, el menor de todos comparando contra una Red Neuronal de 0.21 y HW de 0.06 %.

g. Se diseñó un aplicativo web que permitió la visualización de los datos del modelo, en una arquitectura de componentes, donde php es el lenguaje servidor, html5 es la vista y R Project es el motor analítico. Para la gestión de resultados se utilizó el motor SQL Server.



RECOMENDACIONES

- a. se recomienda que para realizar el pronóstico de producción de espárragos es indispensable que la información sea precisa y obtenida del área correspondiente en la empresa.
- b. Según el trabajo de investigación se sugiere aplicar la librerías forescat que utiliza Paquetes para predicción en la cual permiten obtener resultados en una menor tiempo.
- c. Se recomienda a la empresa agroindustrial que se implemente un área o se designe a una persona con conocimientos de elaboración e interpretación de pronósticos, procesos y operaciones de producción y control y administración con la finalidad de realizar el pronóstico y planificación de la producción de espárragos.



Bibliografía

- Caridad y Otero, J. M. (2013). *Econometría. Modelos econométricos y series temporales con los paquetes uTSP y TSP*.
- López Puga, J. (2010). *INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R Y R COMMANDER EN PSICOLOGÍA Y EDUCACIÓN*. Bogotá, Colombia.
- Acosta Cervantes, M. C., Villareal Marroquín, M. G., & Cabrera Ríos, M. (marzo de 2013). Estudio de validación de un método para seleccionar Técnicas de Pronóstico de series de tiempo mediante redes neuronales artificiales. México.
- 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.
- Alzahani, S., Althopity, A., Alghamdi, A., Alshehri, B., & Alhuaid, S. (4 de Diciembre de 2014). An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction. Arabia.
- Bagurskas. (2015). Con la base de datos de una empresa se puede predecir el comportamiento futuro de de sus clientes". *Datalab Consulting*, <http://revistaganamas.com.pe/con-la-base-de-datos-de-una-empresa-se-puede-predecir-el-comportamiento-futuro-de-sus-clientes/>.
- Beltran, D., & Poveda, D. (2010). *RAPIDMINER*. Bogotá, Colombia.
- Bhatla, N., & Jyoti, K. (2012). *An Analysis of Heart Disease Prediction using Different Data Mining Techniques*.
- 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.
- Carmer Tejada, O. N. (2015). *PROYECTO PRIVADO DE EXPORTACIÓN DE CONSERVA DEESPARRAGO*. Recuperado el 2 de octubre de 2016, de <http://documents.mx/documents/proyecto-de-exportacion-de-conserva-de-esparrago.html#>
- Coghlan, A. (2015). *A Little Book of R for Times Series*. Cambridge: Trust Sanger Institute, Cambridge, U.K. .
- COMEXPERU. (2010). *semanariocomexperu.wordpress.com*. Recuperado el 2 de OCTUBRE de 2016, de esparragos-oro-verde-y-blanco: <https://semanariocomexperu.wordpress.com/esparragos-oro-verde-y-blanco/>
- DMC.SAC, D. M. (2010). *Herramientas de Tomas de Decisiones*. Recuperado el 03 de Octubre de 2016, de <http://es.slideshare.net/dataminingperu/presentacin-minera-de-datos>
- 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.



- Flores, E. (2016). *PROGRAMACION EXTREMA XP* . Recuperado el 01 de octubre de 2016, de ingenieriaesoftware.mex: http://ingenieriaesoftware.mex.tl/52753_XP---Extreme-Programing.html
- García Molina, H. (2006). *Avances en Informática y Sistemas Computacionales*.
- García, A. &. (2010). "Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies". Recuperado el 5 de Noviembre de 2016, de <http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/1339/006312G216.pdf;jsessionid=36911815BA5E92BCF8596DA89E916FB3?sequence=1>
- González. (2005). <https://addi.ehu.es>. Recuperado el 3 de Noviembre de 2016, de bitstream: <https://addi.ehu.es/bitstream/10810/12493/1/05-09pil.pdf>
- Guil, F., Bosch, A., & Marin, R. (2003). *Una Propuesta para la Minería de Patrones Temporales Borrosos*. Murcia: Universidad de Murcia.
- Hossein, P. (1994-2015). "Toma de Decisiones con Periodos de Tiempo Crítico en Economía y Finanzas". Recuperado el 6 de Mayo de 2016, de <http://home.ubalt.edu/ntsbarsh/Business-stat/stat-data/Forecasts.htm>
- Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit*. Wiley India.
- Luna, G. L. (2002). *Busqueda de Patrones de Comportamiento de Cubos de Datos*. Ciudad de Mexico: Instituto Politecnico Nacional.
- MAKERS. (07 de agosto de 2014). *Procedimiento de cliente Cervidor* . Recuperado el 2 de octubre de 2016, de diymakers.es: <http://diymakers.es/raspberry-pi-como-servidor-web/>
- Molero, G. (1 de Octubre de 2008). "Dasarrollo de un Modelo Basado en Tecnicas de Minería de Datos Para Clasificar Zonas climatologicamente similares en el estado de Michoacan". Recuperado el 1 de enero de 2015
- MONTERO, J. M. (2007). *Metodo de Holtwinters*. Recuperado el 3 de Noviembre de 2016, de Estadística descriptiva: <https://books.google.com.pe/books?isbn=8497325141>
- Mora, L. (8 de octubre de 2001). *Programacion en Internet: Clientes Web*. Recuperado el 2 de octubre de 2016, de <http://gplsi.dlsi.ua.es/>: <http://gplsi.dlsi.ua.es/~slujan/materiales/pi-cliente-muestra.pdf>
- Pérez López, C., & Santén González, D. (2008). *Minería de Datos, técnicas y herramientas*. Madril: Thomson Ediciones.
- Rafael Arce, R. (s.f.). *MODELOS ARIMA*. Recuperado el 05 de octubre de 2016, de https://www.uam.es/personal_pdi/economicas/anadelsur/pdf/Box-Jenkins.PDF
- Rios, B. (2013). "Aplicación de minería de datos para predecir fuga de clientes en la industria de telecomunicaciones". Recuperado el 10 de noviembre de 2016, de <http://www.dii.uchile.cl/~ris/RIS2013/rios.pdf>



- Rodríguez Rodríguez, J. E. (2010). *Fundamentos de minería de datos*. Bogotá: Universidad Distrital Francisco José de Caldas.
- Rodriguez, M. (2013).
- Ruelas Santoyo, E., & Laguna González, J. (2013). *Comparación de predicción basada en redes neuronales contra métodos estadísticos en el pronóstico de ventas*. Recuperado el 05 de octubre de 2016, de <http://www.redalyc.org/pdf/2150/215037911008.pdf>
- Sanchez, M. (2010).
- Tello, M. L., Eslava, H. J., & Tobias, L. B. (Agosto de 2012). Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos. Colombia.
- Trujillo, J. C., Mozon, J. N., & Pardillo, J. (2011). *Diseño y explotación de almacenes de datos*. Club Universitario.
- Vallejos, R. (2012). *Introducción a las Series Cronológicas*. Universidad Técnica Federico Santa María.
- 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.
- Verjel Ibañez, A. (2014). *Generación de un modelo para predecir la demanda del servicio aéreo en la ciudad de Ocaña aplicando técnicas de minería de datos*. Obtenido de <http://www.eumed.net/rev/tlatemoani/16/energia.pdf>
- Vieria Braga, L. P., Ortiz Valencia, L. I., & Ramirez Carvajal, S. S. (2009). *Introducción a la Minería de Datos*. Brasil.
- Wintten, & Frank. (2000). *Data Mining: Practica Machine Learning tools and Techniques*.



ANEXOS

Anexo N° 1: Registros de producción de espárrago

Figura N° 49: Data original en excel

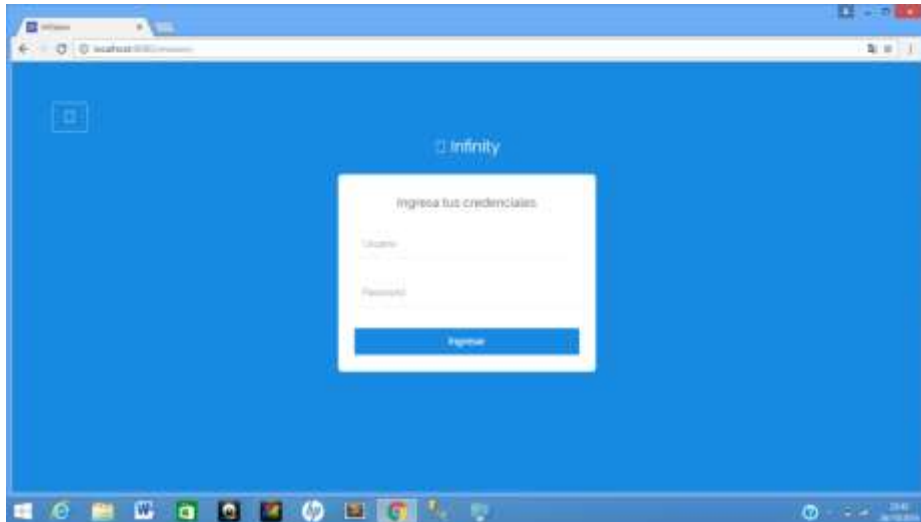
| ID | Identificación | Apellido y Nombre | Comuna | Superficie | Alt. Par. | Temperatura | Hum. Rel. | Fecha Inicio | Comienza | Finaliza | Estado | Superficie | Producción |
|----|----------------|--------------------------------------|---------|------------|-----------|-------------|-----------|--------------|-----------|-------------|----------|------------|------------|
| 5 | 130-14-0011190 | RIVERO DIAZ ARIEL DEL MILAGRO | JAYANCA | 954.00 | 40.00 | 30340.00 | 13340 | 07/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 6 | 130-14-0011191 | CORDIA MHO SLORIA DELMIRA | JAYANCA | 424.00 | 40.00 | 30340.00 | 13340 | 07/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 7 | 130-14-0011192 | CARHUALLCA ANGELES ROSA EVANGELINA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 30/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 8 | 130-14-0011193 | FACHO ARRADO ORLANDO BALTAZAR | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 25/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 9 | 130-14-0011200 | GARCIA MENDOZA CYNTHIA MARGARITA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 28/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 10 | 130-14-0011201 | ANGELIS MORA ANA DEL MILAGRO | JAYANCA | 304.00 | 40.00 | 22340.00 | 7140 | 18/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 11 | 130-14-0011210 | OFFIO CAMPOS LORENA ANAMARIA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 09/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 12 | 130-14-0011211 | ESPINOZA DE ENQUEN MARIA CRUZ | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 04/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 13 | 130-14-0011212 | SANCHEZ GONZALES ANA CELIA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 14 | 130-14-0011213 | RUIZ HUAMAN ROSA LUZ | JAYANCA | 324.00 | 40.00 | 22340.00 | 7140 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 15 | 130-14-0011217 | BLAS HERNAN MARIA MARGALENA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 27/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 16 | 130-14-0011218 | CHUPELON YUVERA RESENA KATHERINE | JAYANCA | 324.00 | 40.00 | 22340.00 | 7140 | 12/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 17 | 130-14-0011219 | FARRO ORDÓÑEZ MARIA DEL MILAGRO | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 04/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 18 | 130-14-0011249 | SALVEZ POZO CONSUELO DEL PILAR | JAYANCA | 724.00 | 40.00 | 30340.00 | 13340 | 11/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 19 | 130-14-0011277 | TANTARICO MEJIA LUIS YOLANA | JAYANCA | 724.00 | 40.00 | 30340.00 | 13340 | 14/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 20 | 130-14-0011289 | JACUÑA MONTENEGRO MARGORIE JANITA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 21 | 130-14-0011298 | BELTRERA RUISTEZ MARIA ROSANA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 22 | 130-14-0011299 | CKAMPES CHANARE ANA BERTHA | JAYANCA | 724.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 23 | 130-14-0011299 | PINEDA MUÑOZ JUSTINA HELIA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 24 | 130-14-0011297 | CRAPANZA RAMOS ALEJON LUZ | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 14/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 25 | 130-14-0011297 | SANCHEZ AMENCO MARIA ISABEL | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 26 | 130-14-0011298 | DIAZ CAMPOS SHYLLA MARISOL | JAYANCA | 324.00 | 40.00 | 22340.00 | 7140 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 27 | 130-14-0011320 | FLORES MORENO LUCIANA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 05/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 28 | 130-14-0011340 | VASQUEZ VERGARA LUIS ALBERTO | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 11/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 29 | 130-14-0011340 | PIÑO CLODER JESUS AMALIA | JAYANCA | 724.00 | 40.00 | 34340.00 | 13840 | 15/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 30 | 130-14-0011323 | QUINTANA MONTENEGRO DIANA MARIBEL | JAYANCA | 324.00 | 40.00 | 22340.00 | 7140 | 27/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 31 | 130-14-0011326 | RODRIGA PEREZ MARIA JERONIMA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 25/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 32 | 130-14-0011388 | RIVADENEIRA YAVAFACO JASTIN | JAYANCA | 324.00 | 40.00 | 22340.00 | 7140 | 11/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 33 | 130-14-0011392 | CURMA SUYHARA IDELSA | JAYANCA | 424.00 | 40.00 | 34340.00 | 13840 | 25/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 34 | 130-14-0011395 | BACA RODRIGUEZ VDA DE MONTALVO DORIS | JAYANCA | 724.00 | 40.00 | 34340.00 | 13840 | 08/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |
| 35 | 130-14-0011399 | SANCAYA LIONTOP MARIA DEL PILAR | JAYANCA | 724.00 | 40.00 | 34340.00 | 13840 | 25/01/2015 | Estragado | CampoPlanta | No Tieme | Cotado | |

Fuente: Elaboración propia

ANEXO N° 2: Manual de usuario Ingreso a la aplicación web



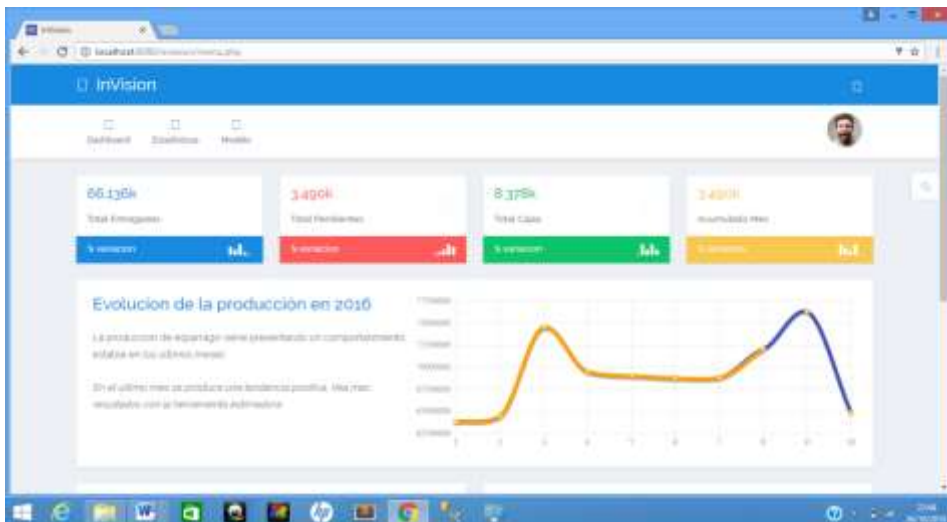
Figura N° 50: Ingreso a la aplicación a través de un usuario y una contraseña



Fuente: Elaboración Propia

Se muestra para ingresar un usuario y una contraseña para ingresar al aplicativo Web

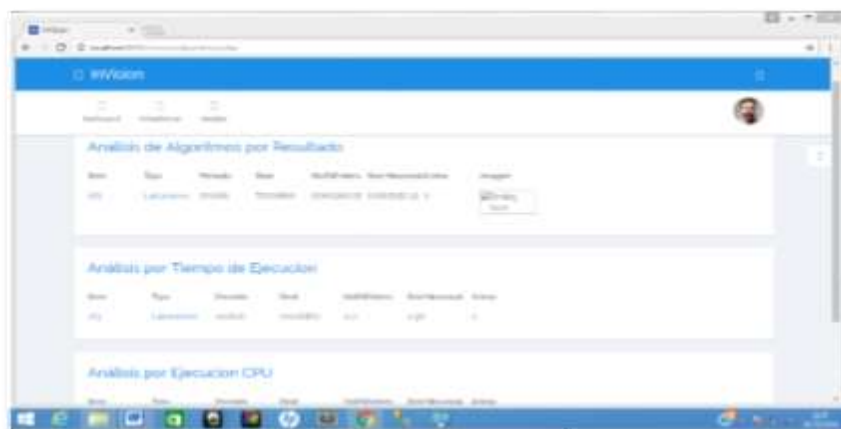
Figura N° 51: evaluaciones de producción en porcentajes



Fuente: Elaboración Propia



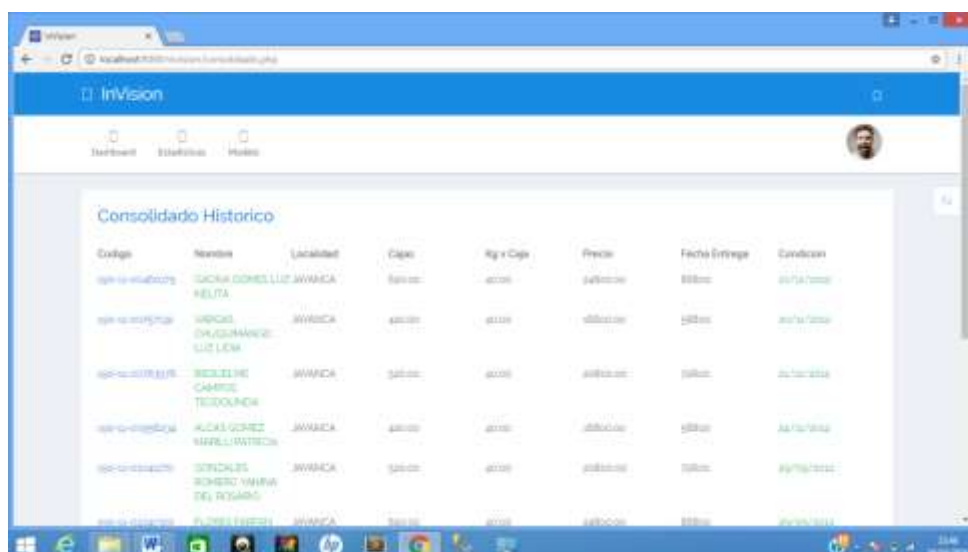
Figura N° 52: Pronósticos por periodos y tiempo



Fuente: Elaboración Propia

Se muestra las variaciones de la producción realizadas

Figura N° 53: Consolidado histórico

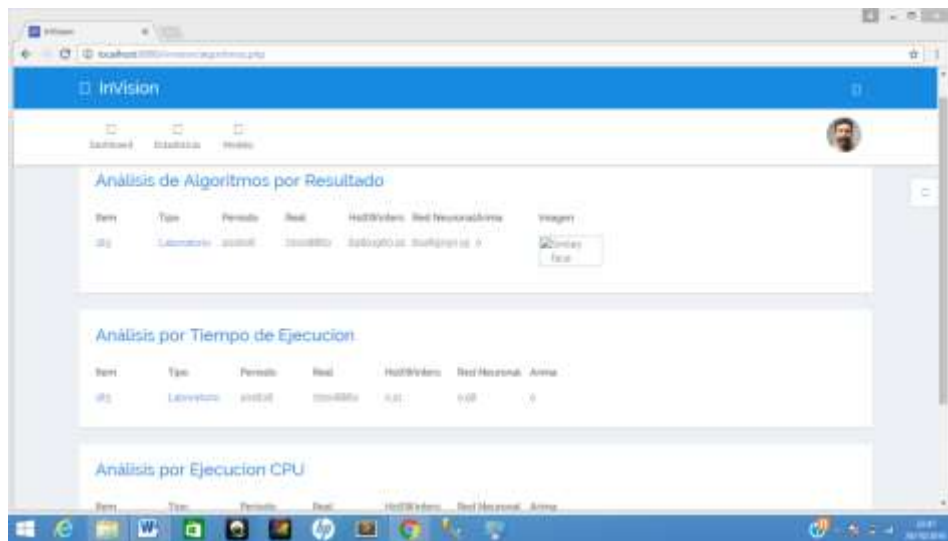


Fuente: Elaboración Propia

Se muestra los datos históricos de la producción desde la base de datos



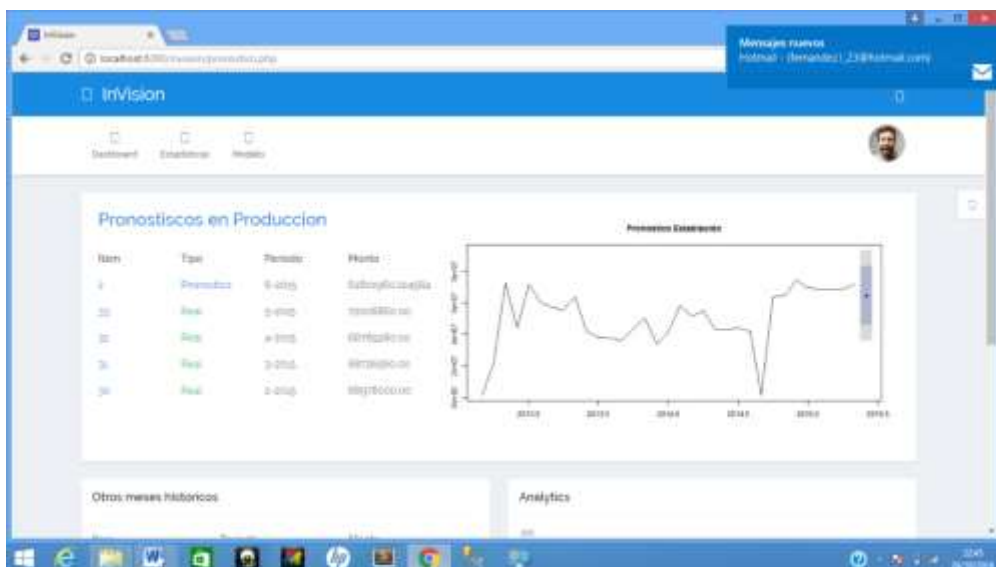
Figura N° 54: Resultados del análisis de los algoritmos



Fuente: Elaboración Propia

Ingresamos para mostrar el análisis de los modelos realizados

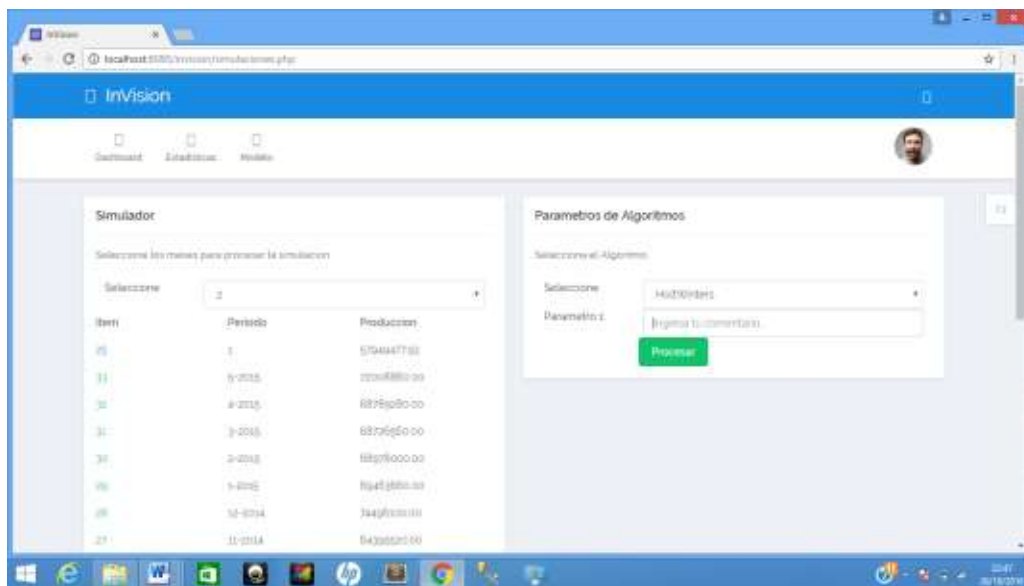
Figura N° 55 Se muestra los pronósticos.



Fuente: Elaboración Propia

Se ingresa para realizar los pronósticos

Figura N° 56: Muestra las simulaciones por periodos y meses



Fuente: Elaboración Propia

Ingresamos para realizar las simulaciones con los algoritmos

ANEXO N° 3: Plan de pruebas de técnicas

Figura N° 57: Entrenamientos de algoritmos

